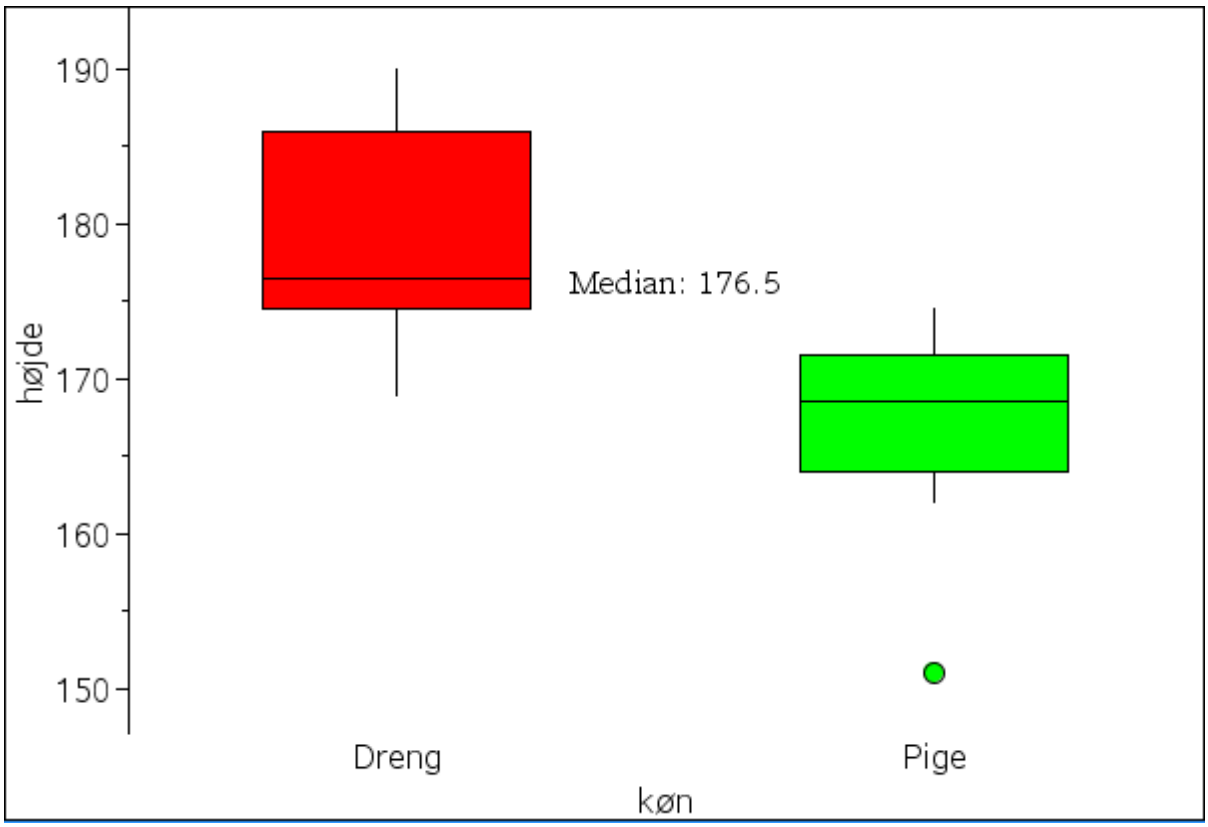


Statistik med TI-Nspire CAS version 3.2

Bjørn Felsager

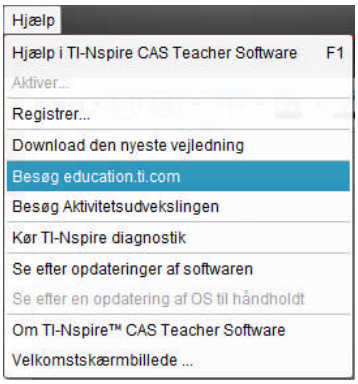
September 2012

[Fjerde udgave]

Indholdsfortegnelse

Forord

TI-Nspire CAS er et dynamisk databehandlingsprogram, der egner sig til undervisning i databehandling på mange niveauer: Folkeskolens ældste klasser, gymnasiet og HF samt de videregående uddannelser. **Statistik med TI-Nspire CAS** er skrevet som introduktion til **TI-Nspire CAS** til brug for dels undervisningen i statistik på det *indledende niveau i gymnasiet*, dvs. **Mat C**, hvor det især er den første del omkring den beskrivende statistik, der vil være af interesse, dels undervisningen i statistik på de *to fortsætterniveauer i gymnasiet*, dvs. **Mat B og Mat A**, hvor det især er den anden del omkring den bekræftende statistik, der vil være af interesse. Begge dele er skrevet så de kan indgå i et tværfagligt samarbejde med dels samfundsfag, dels de naturvidenskabelige fag – ikke mindst biologi. Hæftet kan benyttes uafhængigt af andre introduktionshæfter, men for at få et bedre kendskab til **TI-Nspire CAS** kan det anbefales også at gennemarbejde et introducerende hæfte om variabelsammenhænge, fx eksempelsamlingen, der kan hentes på Texas hjemmeside. Alle de datasæt, der omtales i hæftet kan tilsvarende hentes som rådata på Texas hjemmeside – hvor man fx kan få adgang via **Hjælp**-menuen:



Første del omhandler den **beskrivende statistik**: *Hvordan trækker man information ud om et datasæt*. I det indledende afsnit indføres de vigtigste graftyper og deskriptorer: *prikplot, histogram* og *boksplot*, *kvartilsæt* og *middelværdi*. Dernæst følger et afsnit, der kommer rundt om de vigtigste kendetegn ved en statistisk fordeling: *niveauet*, *spredningen* og *formen*. Her introduceres også **gamma-graden** som en karakterisering af ordinale variable. Da mange datasæt i praksis optræder som grupperede data (som også er et centralt emne i kernestoffet for MatC) afsluttes der med et længere afsnit om *grupperede data* med en gennemgang af de tilhørende graftyper og deskriptorer: *søjlediagram*, *histogram*, *sumkurve* og *boksplot*, *kvartilsæt* og *middelværdi*.

Anden del omhandler den **bekræftende statistik**: Hvad er en statistisk hypotese? Hvordan sandsynliggør man en hypotese? Der lægges vægt på eksperimentelle metoder, som i langt højere grad er tilgængelige for undervisningen på dette indledende niveau. I det første indledende afsnit indføres en række centrale begreber, som fx nulhypotesen H_0 og p-værdi. De følgende fire afsnit omhandler derefter forskellige aspekter af de vigtigste hypotesetest, chi2-testen, der nu er kernestof på matematik B og A: Der lægges ud med to simple eksempler på henholdsvis Goodness-of-fit testen og testen for uafhængighed. Derefter gennemgås et større eksempel, der egner sig for projektsamarbejder med samfundsfag: Spørgeskemaanalyser. Hovedvægten ligger på *forståelsen* af den statistiske metode og lægger derfor først og fremmest op til projektsamarbejder og den mundtlig eksamen i matematik. Men i hvert eksempel gennemgås også den indbyggede kanoniske test, der især egner sig til skriftlige eksamen – men først når man har forstået metoden. Til slut udvides det statistiske testarsenal med gamma-testen, som egner sig til at undersøge sammenhængen mellem to ordinale variable.

I forhold til tredje udgave af **Introduktion til statistik med TI-Nspire CAS** er der i denne udgave kun sket mindre ændringer. I afsnit 8 er eksemplet med **McCain versus Obama** udskiftet med et mere tidssvarende eksempel baseret på YouTube videoen **Er piger venstreorienterede**? Fra Gladsakse Gymnasium. Dels er der taget hensyn til forbedringer i programmet, herunder indførelsen histogrammer med ulige store intervaller.

Bjørn Felsager, september 2012

Beskrivende statistik

1 Grundlæggende TI-Nspire CAS-teknikker

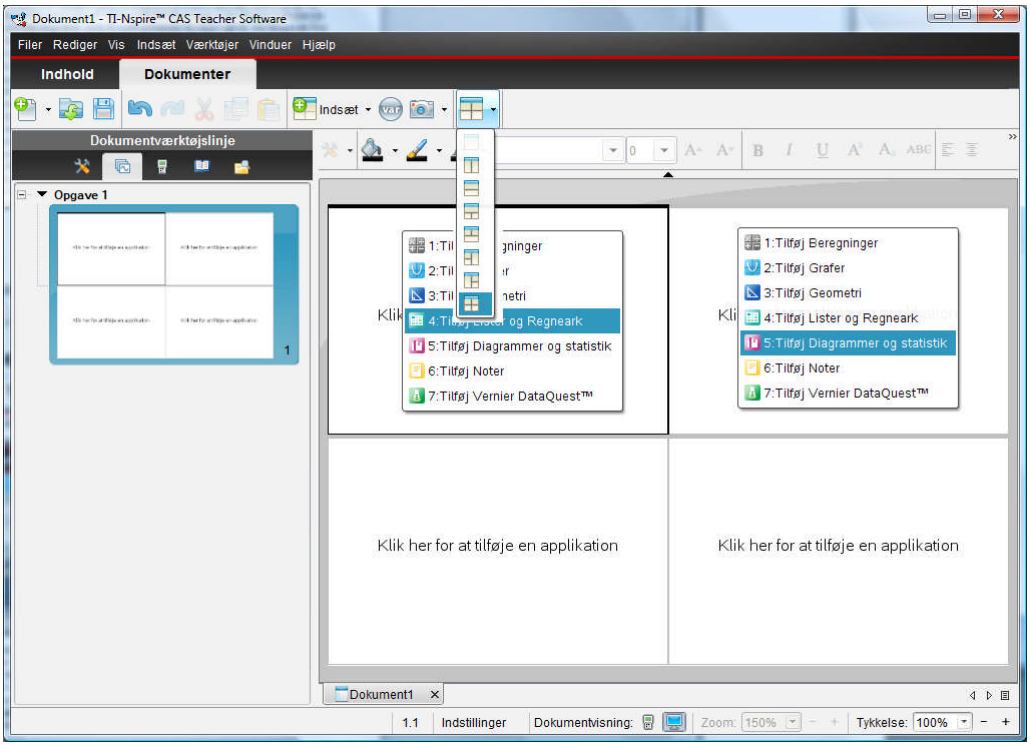
1.1 Velkommen til TI-Nspire CAS

Når du åbner for **TI-Nspire CAS** viser der sig først et velkomstskærmbillede

Her vælger du **Lister og Regneark**-værkstedet i **Lynstart**-menuen

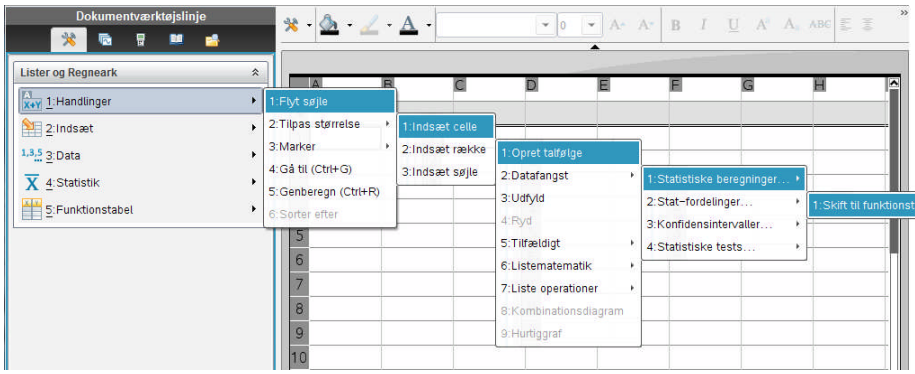
Her vælger du **Nyt dokument** ved at lukke velkomst-skærmbilledet!

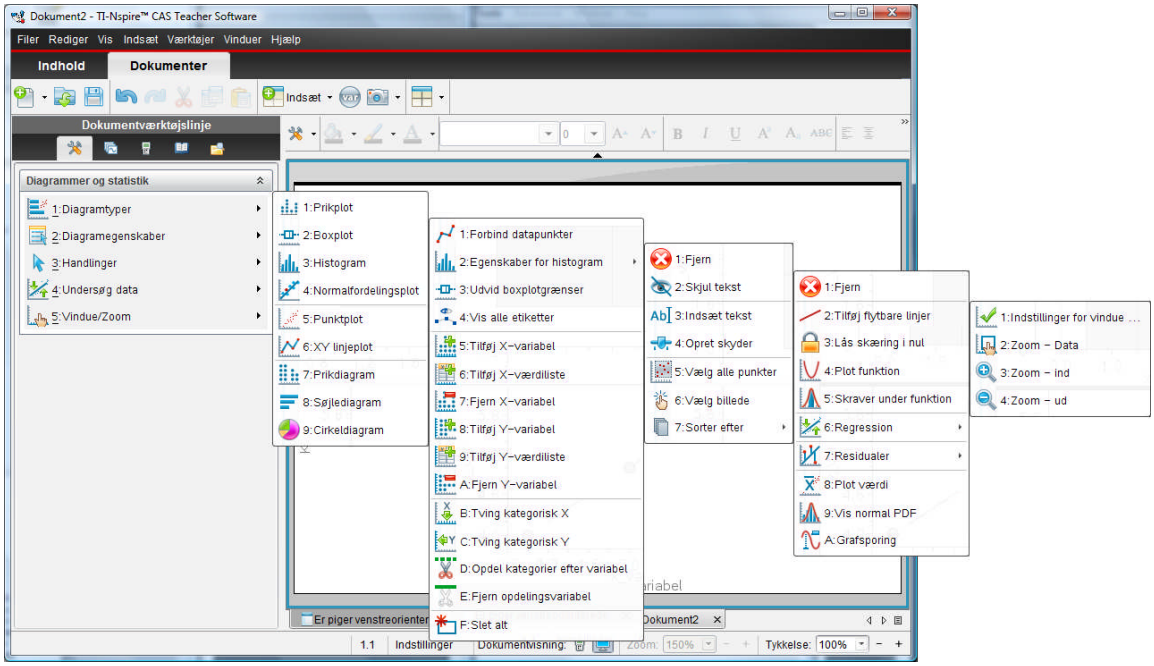
Lukkes velkomstskærmbilledet fås et skærmbillede opdelt i en **overordnet menubjælke** (med de sædvanlige Windows-menuer: **Filer**, **Rediger** osv.), et **sidepanel** med bl.a. en **sidesorterer**, der giver overblik over siderne i det enkelte dokument (samt et lommeregnertastatur, en avanceret stifinder og diverse muligheder for at indsætte formler, symboler og kommandoer). Dertil kommer et **dokumentvindue** med et arbejdsområde for det aktuelle dokument, samt en særskilt menubjælke, der afspejler det aktive værksted. De tilgængelige dokumenter vises i faner forneden. **Arbejdsområdet** kan som vist opdeles i op til fire forskellige værksteder. I statistikken vil vi især benytte værkstederne **Lister og Regneark** og **Diagrammer og Statistik**.



1.2 Lister og regneark

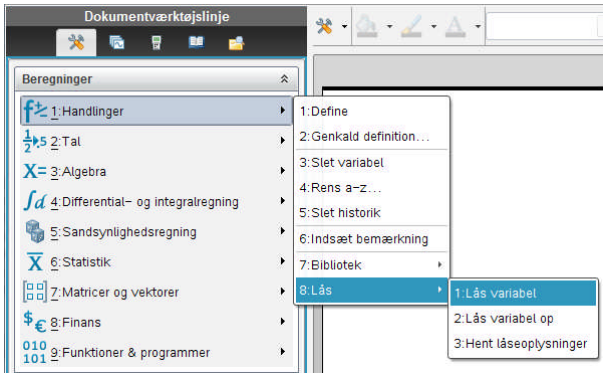
Værkstederne er tilknyttet deres egne menubjælker, der viser hvad man kan arbejde med i det pågældende værksted:





I **Lister og Regneark**-værkstedet kan man arbejde med datatabeller, simuleringer og forskellige numeriske statistiske beregninger og målinger. I **Diagrammer og Statistik**-værkstedet kan man arbejde med grafiske fremstillinger af data og foretage forskellige simple grafiske undersøgelser af data, herunder kan man arbejde med **skydere** (parametre), ligesom man kan indsætte **tekstbokse**. I **Lister og Regneark**-værkstedet finder man specielt i statistik-menuen rimeligt avancerede statistiske værktøjer til fx at udregne **konfidensintervaller** og efterprøve kanoniske **statistiske hypoteser**. Dem vil vi kun kort omtale i dette hæfte, da den fulde brug af disse teoretiske værktøjer kræver et indgående kendskab til statistisk teori.

De to værksteder – **Lister og Regneark** og **Diagrammer og Statistik** – er uløseligt forbundne: De fremviser blot to forskellige repræsentationer af de samme grundlæggende data. Ændrer man i data i regnearket følger graferne automatisk med, og hvis man omvendt trækker i datapunkterne i **Diagrammer og Statistik**-værkstedet følger cellerne i **Lister og Regneark**-værkstedet automatisk med. Man kan dog godt låse en variabel, så det ikke er muligt at trække i datapunkterne. Det sker ved hjælp af kommandoen **Lock**, der skal udføres i et **Beregninger**-værksted.



2 Oprettelse af datasæt: Håndtering af variable

Men lad os komme i gang med et eksempel: Kernen i **TI-Nspire CAS** er dets unikke evne til at håndtere **variable**, så lad os gøre nogle observationer og knytte variable til dem. Det kunne være om *klassen*, hvor vi kunne se på datasættet bestående af de enkelte elever karakteriseret ved forskellige egenskaber, såsom navn, køn, alder, højde osv. Det kan I imidlertid selv lege med. Her vil vi i stedet se på et eksempel, der også er så simpelt, at vi selv nemt kan taste data ind. Der er givet følgende oplysninger om fodboldklubberne i superligaen for sæsonen 2009-2010 (kilde: www.Tipsbladet.dk):

Klub	**Anslået spillerbudget 2009-10 i millioner kr.**	**Points 2009-10**
FC København	125	68
OB	47.5	59
Brøndby IF	67.5	52
Esbjerg fB	27.5	50
AaB	37.5	48
FC Midtjylland	37.5	47
FC Nordsjælland	21	43
Silkeborg	21	43
SønderjyskE	17	41
Randers FC	32.5	40
AGF	27.5	38
HB Køge	12.5	19

I den ovenstående tabel er der tre variable: **Klub**, **budget** og **points**. Vi åbner derfor for et **Lister og Regneark**-værksted og starter med at give et navn til den første variabel – klubbens **navn** – ved at klikke på titelfeltet A og skrive løs. Herefter skriver vi navne ind i de følgende celler, idet vi husker at der nu skal gåseøjne " " omkring navne (for at adskille dem fra symbolske udtryk, idet Lister og regneark også er et symbolsk regneark). Læg mærke til at gåseøjne automatisk sættes parvis i indtastningsfeltet, ligesom med parenteser osv.

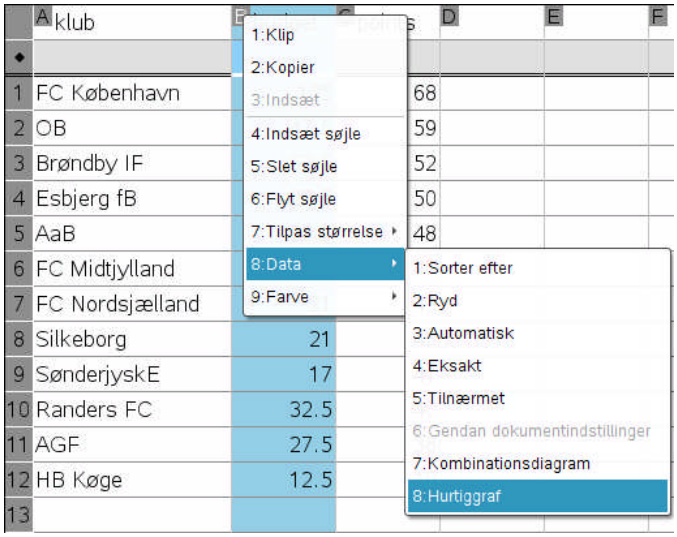
Alternativt kan man simpelthen kopiere data fra tabellen ind i **TI-Nspire CAS**, hvis man fx har en elektronisk Wordudgave til rådighed. Den første variabel - **klub** - er nem nok at indtaste, da den bare består af en tekst. Sådanne **tekstvariable** kaldes også for **kategoriske variable**, fordi de opdeler klubberne i forskellige kategorier, fx efter navn eller geografisk placering. Tekstvariable er venstrestillede, ligesom i et almindeligt regneark, og de er karakteriseret ved at de er omsluttet af gåseøjne " ". I det hele taget minder tabellen i sin struktur meget om et regneark med nummererede rækker, idet hver klub har fået tildelt sit eget **indeks**, der angiver klubbens plads i tabellen. Læg også mærke til at søjlerne, dvs. de enkelte variable, alle er født med bogstavnavne – **A**, **B**, **C** osv. – som i et almindeligt regneark. Det gør det nemt også at referere til en enkelt celle. I **TI-Nspire CAS** – som er et dynamisk regneark – kan vi selv vælge om vi vil arbejde med hele søjler/lister ad gangen (sådan som du måske også kender det fra din grafregner) eller om vi vil arbejde med enkeltceller (sådan som du kender det fra fx **Excel**).

	A klub	B budget	C points
◆			
1	FC København	125	68
2	OB	47.5	59
3	Brøndby IF	67.5	52
4	Esbjerg fB	27.5	50
5	AaB	37.5	48
6	FC Midtjylland	37.5	47
7	FC Nordsjælland	21	43
8	Silkeborg	21	43
9	SønderjyskE	17	41
10	Randers FC	32.5	40
11	AGF	27.5	38
12	HB Køge	12.5	19

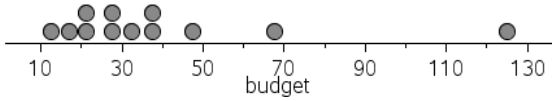
De to næste variable, **budget** og **points** er **talvariable**. De kaldes også for **numeriske variable**. Talvariable er højrestillede, ligesom i et almindeligt regneark.

2.1 Oprettelse af grafer: Prikplot og histogram

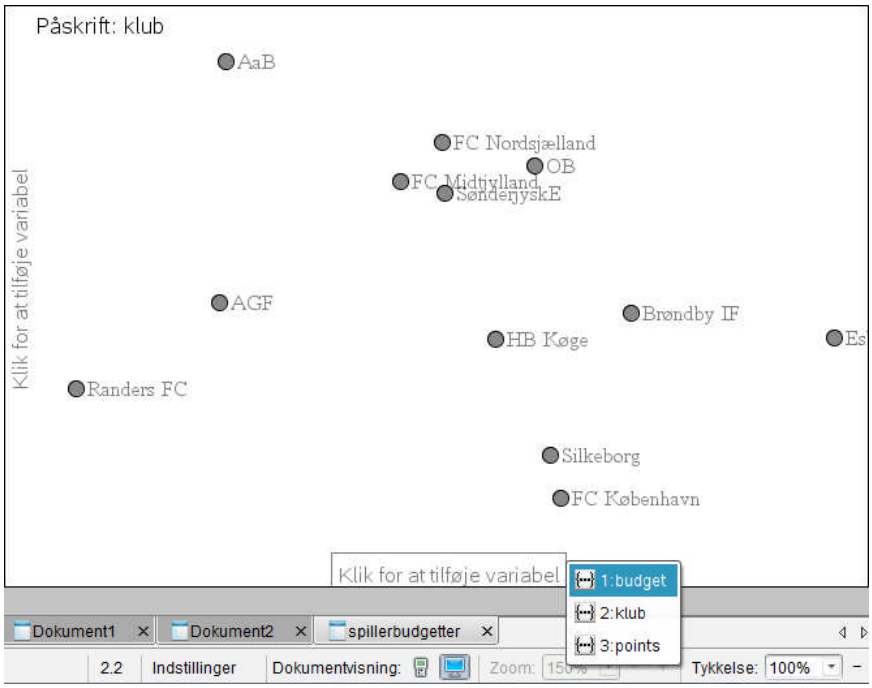
Hvordan kan vi nu danne os et grafisk overblik over denne tabel? Det kan gøres på flere forskellige måder. Her vil vi nu først se på nogle grafiske metoder til at danne sig et overblik over de *numeriske variable* (vi vil senere også se nærmere på graftyperne for kategoriske variable).



Vi kan nu oprette en graf over budgetterne ved at markere søjlen for **budget** (klik på søjlenavnet B i titelfeltet) og derefter højreklikke i den markerede søjle. Vi kan da vælge menupunktet **Data** > **Hurtig-graf**, hvorved der automatisk oprettes et **Diagrammer og Statistik**-værksted med den ønskede graf:



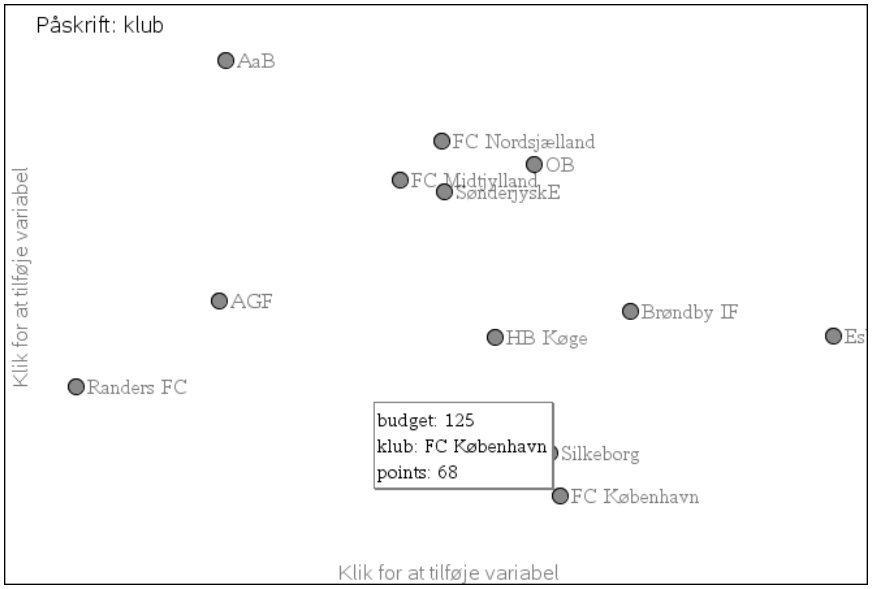
Man kan også selv oprette et **Diagrammer og statistik**-værksted. Der åbnes da for en ustruktureret afbildning af datasættet med en tilfældig fordeling af datapunkterne svarende til den første kategoriske variabel (hvis der er en og ellers den første numeriske variabel). Derefter kan man som vist klikke i variablen **budget** på førsteaksen i grafrummet.



Datapunkterne struktureres da i en glidende bevægelse efter værdien af variablen **budget** og slutresultatet er det samme som den ovenstående **hurtig-graf**. Herved fremkommer altså et **prikplot**, der giver en god fornemmelse for budgetfordelingen, som er tydeligt **højreskæv** og

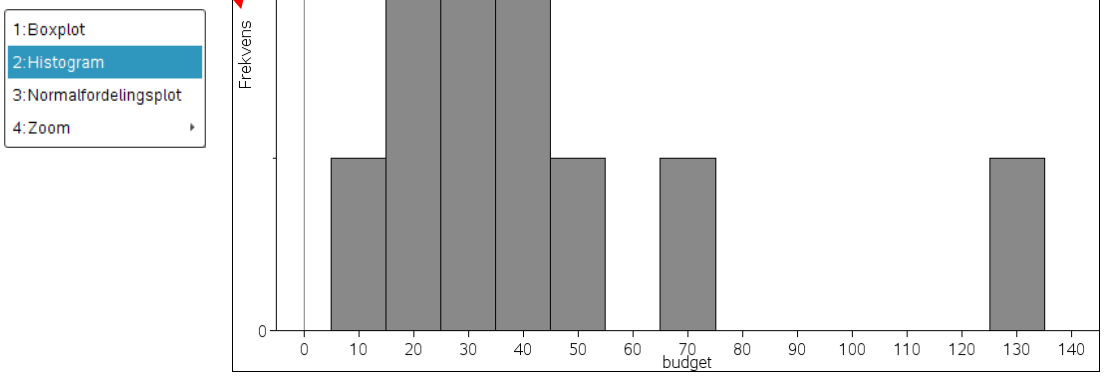
har en **central klump** omkring 30 millioner kr. og en lang **hale** til højre med de to storklubber Brøndby IF og FC København. Læg mærke til at prikplottet er **stakket**, dvs. at prikkerne anbringes oven på hinanden, så vi får en tydelig fornemmelse af fordelingens form.

Selv om klubbernes navne ikke optræder på prikplottet kan vi nemt finde ud af hvem der gemmer sig bag prikkerne. Hertil oprettes fx en **ustruktureret graf** (dvs. vi åbner for et **diagrammer og statistik**-værksted, men afsætter ikke variable). Klikker vi på et datapunkt, blinker det nemlig i alle andre grafer. Flytter vi musen hen til det blinkende punkt i den ustrukturerede graf fås en oplysningsseddel om værdierne for alle variablene knyttet til datapunktet:

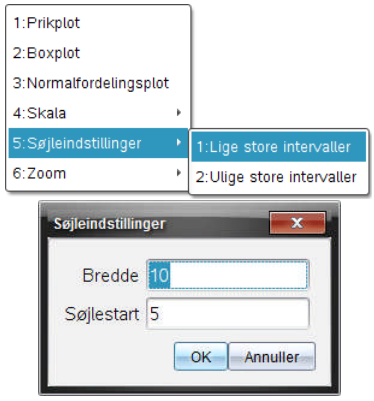


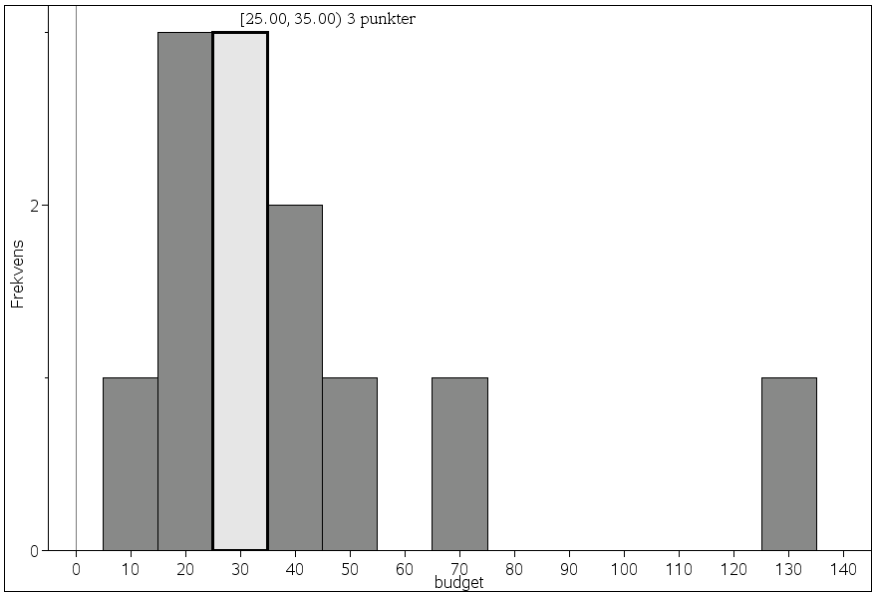
Vi kan også skifte graftype ved at højreklikke og afsætte budgetterne i et **histogram**:

Betegnelsen **frekvens** er det samme som **hyppighed**.



Histogrammet viser det samlede overordnede mønster som prikplottet. Vi kan nu tilpasse histogrammet ved at trække i histogramboksenes kanter eller ved at højreklikke i grafrummet og derved få adgang til menupunktet **Søjleindstillinger > Lige store intervaller**. Her kan vi selv sætte **intervalbredden** og regulere **intervalstarten**, dvs. placeringen af det første intervalendepunkt. Vi vender tilbage til histogrammer med ulige store intervaller, der kræver en separat liste over intervalendepunkterne.

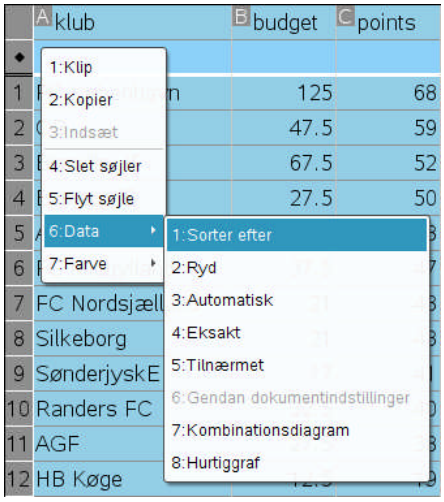


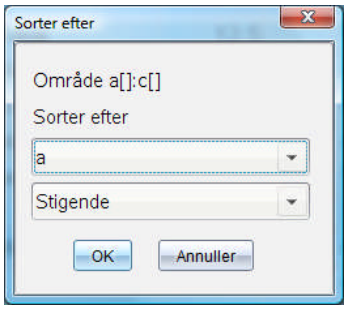


Læg mærke til at hvert af intervallerne starter i det *venstre* endepunkt. (Der er tale om et tilfældigt valg. I andre undervisningstraditioner kan man derfor møde det modsatte valg, hvor det er højre endepunkt, der regnes med). Hvis vi fx som vist markerer det ene af **typeintervallerne**, kan vi se at det drejer sig om tre klubber med budgetter fra 25 millioner kr. (inklusive) op til 35 millioner kr. (eksklusive). Læg også mærke til at intervallet angives med en lidt anden konvention end den danske, idet der både benyttes kantede og runde parenteser: kantet når parentesen er lukket og rund når parentesen er åben.

2.2 Oprettelse af grafer: Kvartilsæt og boksplot

Vi kan ordne budgetterne i rækkefølge ved at markere alle variablene **Klub**, **budget** og **points** og højreklikke for at vælge kommandoen **Sorter** (herved sikres at alle de markerede søjler ordnes samtidigt). Herefter vælger vi dels at sortere efter kolonne **b[]** (dvs. i vores tilfælde variablen **budget**) og dels at sortere stigende. Læg også mærke til områdebetegnelsen a[]:c[], dvs. området fra søjle a til søjle c:





Vi finder da - hvor man på den følgende figur skal lægge mærke til at når budgetterne er ens følges den oprindelige orden, dvs. placeringen efter points, jfr. Esbjerg og AGF med det samme årsbudget på 27.5 millioner kr. Her kommer Esbjerg foran AGF fordi Esbjerg på forhånd stod over AGF.

Median

	A klub	B budget	C points
◆			
1	HB Køge	12.5	19
2	SønderjyskE	17	41
3	FC Nordsjælland	21	43
4	Silkeborg	21	43
5	Esbjerg fB	27.5	50
6	AGF	27.5	38
7	Randers FC	32.5	40
8	AaB	37.5	48
9	FC Midtjylland	37.5	47
10	OB	47.5	59
11	Brøndby IF	67.5	52
12	FC København	125	68

Vi kan nu aflæse det mindste budget (HB Køge), dvs. **minimum**, det midterste budget (midt mellem AGF og Randers FC), dvs. **medianen**, samt det største budget (FC København), dvs. **maksimum**

Min = 12.5 Med = 30 Max = 125

Medianen er den *midterste observation*. Hvis der er et *lige antal* observationer, vil der i stedet være to midterobservationer. Man har da vedtaget, at medianen i dette tilfælde er *gennemsnittet af de to midterste observationer*.

Medianen deler nu det ordnede datasæt i to halvdele. Vi kan derfor fortsætte med at fastlægge medianerne for hver af disse to halvdele. De kaldes **første og tredje kvartil** Q_1 og Q_3 (jfr. ordet kvart, som står for en fjerdedel, idet kvartilerne deler datasættet i fjerdedele). Hvis der er et lige antal observationer er det oplagt, hvordan datasættet splittes i to halvdele. Med et ulige antal er det lidt mere indviklet, fordi man i princippet både kan medtage og udelukke den midterste observation i de to halvdele. I **TI-Nspire CAS** har man nu vedtaget at man aldrig medtager medianen, dvs. de to halvdele består af de datapunkter, der går forud for medianen og de datapunkter der følger efter medianen:

De to **kvartiler** udgør *medianerne for de to halvdele* af det ordnede datasæt. Hvis der er et ulige antal observationer regnes midterobservationen *ikke* med til de to halvdele.

Q_1

	A klub	B budget	C points
◆			
1	HB Køge	12.5	19
2	SønderjyskE	17	41
3	FC Nordsjælland	21	43
4	Silkeborg	21	43
5	Esbjerg fB	27.5	50
6	AGF	27.5	38
7	Randers FC	32.5	40
8	AaB	37.5	48
9	FC Midtjylland	37.5	47
10	OB	47.5	59
11	Brøndby IF	67.5	52
12	FC København	125	68

	A klub	B budget	C points
◆			
1	HB Køge	12.5	19
2	SønderjyskE	17	41
3	FC Nordsjælland	21	43
4	Silkeborg	21	43
5	Esbjerg fB	27.5	50
6	AGF	27.5	38
7	Randers FC	32.5	40
8	AaB	37.5	48
9	FC Midtjylland	37.5	47
10	OB	47.5	59
11	Brøndby IF	67.5	52
12	FC København	125	68

Q_3

De to halvdele ser altså således ud. Første kvartil Q_1 er derfor givet ved 21 millioner kr. i årsbudget, mens tredje kvartil Q_3 er givet ved 42.5 millioner kr. i årsbudget.

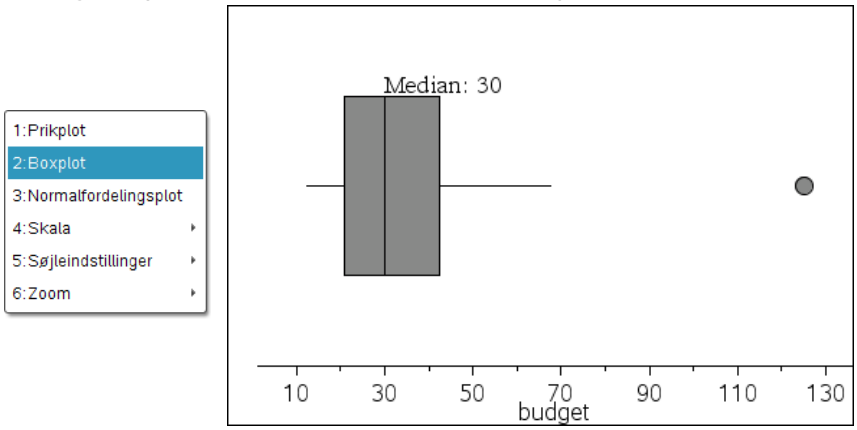
Kvartilsættet bestående af den første kvartil, medianen og den tredje kvartil (hvor medianen kan opfattes som den anden kvartil) deler datasættet i fire dele, som hver for sig rummer (ca!) en fjerdedel af observationerne. Tilføjer vi ydermere minimum og maksimum (den nulte kvartil og den fjerde kvartil) til kvartilsættet kaldes det for **det udvidede kvartilsæt** eller **de fem statistiske nøgletal**. De fem nøgletal styres af kommandoen FiveNumSummary. Der findes i kataloget. Den kan anvendes i **Beregninger**-værkstedet eller som her **Noter**-værkstedet:

FiveNumSummary **budget** ▸ *Udført*

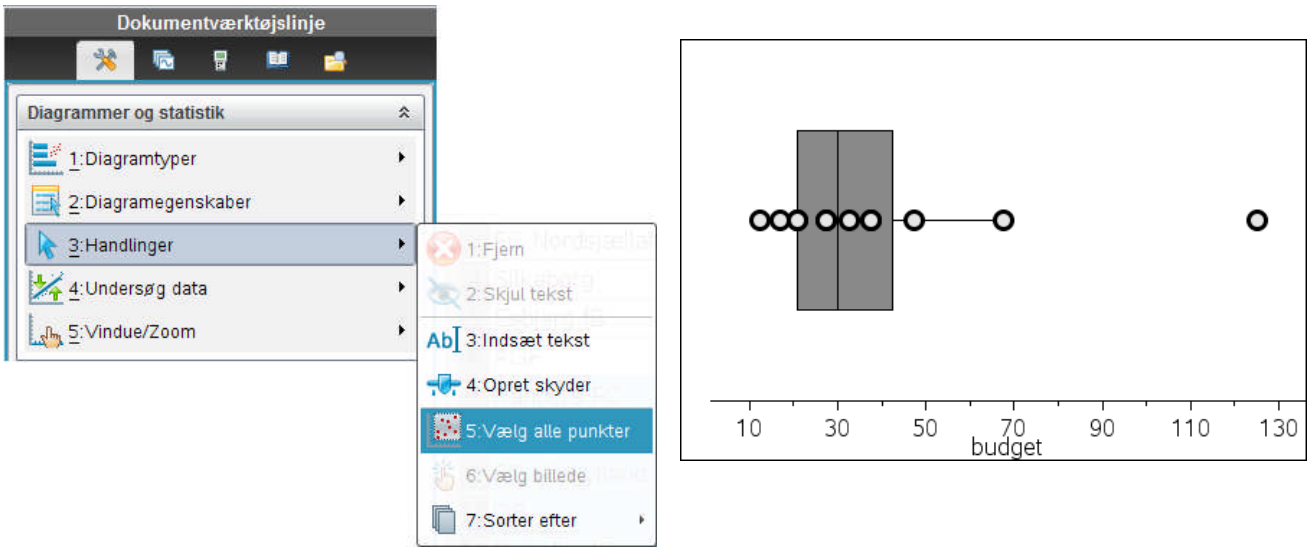
stat.results ▸

"Titel"	"Fem-nummer sammendrag"
"MinX"	12.5
"Q₁X"	21.
"MedianX"	30.
"Q₃X"	42.5
"MaxX"	125.

Vi samler ofte de fem nøgletal i et såkaldt **boksplot**, hvor boksens to ender angiver første og tredje kvartil, mens medianen markeres med en lodret streg inde i boksen. Boksen indeholder altså (mindst!) halvdelen af observationerne. Dertil føjer vi vandrette streger, der i princippet rækker helt ud til den mindste observation og den største observation. Men som udgangspunkt vil **TI-Nspire CAS** i stedet vælge at skille de yderste observationer ud, hvis de ligger markant langt væk fra de øvrige. Flytter vi markøren hen over boksplottet kan vi aflæse de fem nøgletal



Boksplottet giver et klart indtryk af den **skævhed** der er i budgetterne for klubberne i superligaen. Halvdelen af budgetterne ligger i den venstre hale og den venstre halvdel af boksen, som begge er små i forhold til den højre halvdel og den meget lange højre hale. Boksplottet giver derimod ikke noget indtryk af hvordan de enkelte individuelle data fordeler sig. Skyldes den højre hale fx blot nogle få observationer eller er der tale om en hel stribe af observationer.



Vælger vi menupunktet **Vælg alle punkter** fra **Handlinger**-menuen, ser vi at boksplottet er udspændt af 9 værdier (hvoraf en del af de mindre værdier optræder flere gange).

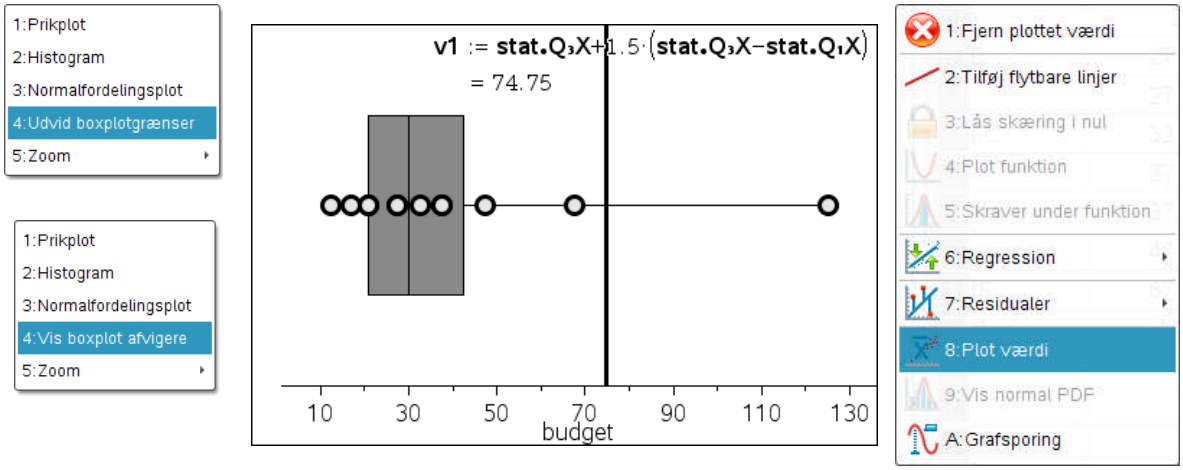
2.3 Afvigere: Tukeys regel

Her kan vi nu få glæde af begrebet en **afviger** eller **perifer observation**. En observation kaldes en **afviger** eller siges at være **perifer**, hvis den ligger usædvanligt langt ude i forhold til den centrale boks. Det er selvfølgelig et skøn, hvornår noget ligger så langt ude at det må betragtes som ekstremt. **TI-Nspire CAS** benytter en tommelfingerregel opstillet af den amerikanske statistiker Tukey – der også er ophavsmand til boxplottet – der har vist sig i praksis at være yderst nyttig til at spotte afvigerne. Tukey tager udgangspunkt i boksens bredde, den såkaldte **kvartilbredde**, dvs. tallet $Q_3 - Q_1$:

> **Tukeys regel**: Hvis en observation ligger længere væk end halvanden kvartilbredde fra den centrale kasse, anses den for at være en afviger. Grænserne for linjestykkerne, dvs. nedre kvartil minus halvanden kvartilbredde og øvre kvartil plus halvanden kvartilbredde, kaldes Tukeys **hegn**.

I vores tilfælde er kvartilbredden 21.5 millioner kr. (= 42.5 – 21). Halvanden kvartilbredde er derfor 32.25 millioner kr. Trækker vi 32.25 millioner kr. fra den venstre kant, dvs. Q_1 = 21 millioner kr., fås et negativt budget. Dvs. vi skal ned på et negativt budget for at have et ekstremt lavt budget. Det er der ingen klubber, der har! Tilsvarende skal vi lægge 32.25 millioner kr. til den højre kant, dvs. Q_3 = 42.5 millioner kr. Vi skal altså op over et årsbudget på 74.75 millioner kr. for at have et ekstremt højt budget. Det er der kun én klub, der har!

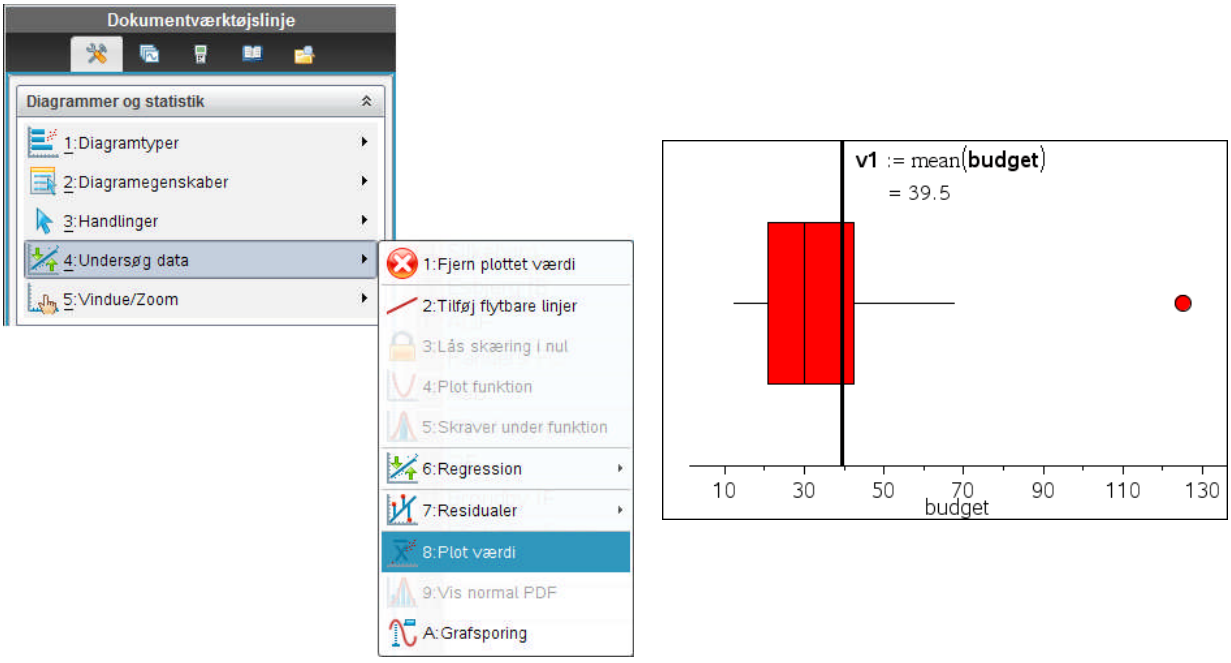
Bemærkning: Man behøver ikke få vist de perifere data i et boksplot. Hvis man højreklikker på boksplottet kan man vælge menupunktet **Udvid boxplotgrænser** og derved udstrække halerne til minimum- og maksimum-værdierne helt uafhængigt af om der er perifere målinger til stede. Tilsvarende kan man få afvigerne tilbage ved at højreklikke og vælge **Vis Boxplot afvigere**:



Bemærkning: Men hvis man vil inddrage Tukeys hegn i diagrammet, skal man som vist anvende **Plot værdi** fra **Undersøg data**-menuen. Der er ingen direkte kommando til udregning af kvartiler, så her må man i stedet benytte stat-variablene, der først lagres når man udfører en statistisk beregning, fx via fivenumsummary-kommandoen. De fås som vist ved at taste **stat punktum** og så vælge deskriptoren fra rullegardinsmenuen.

2.4 Middelværdien versus medianen

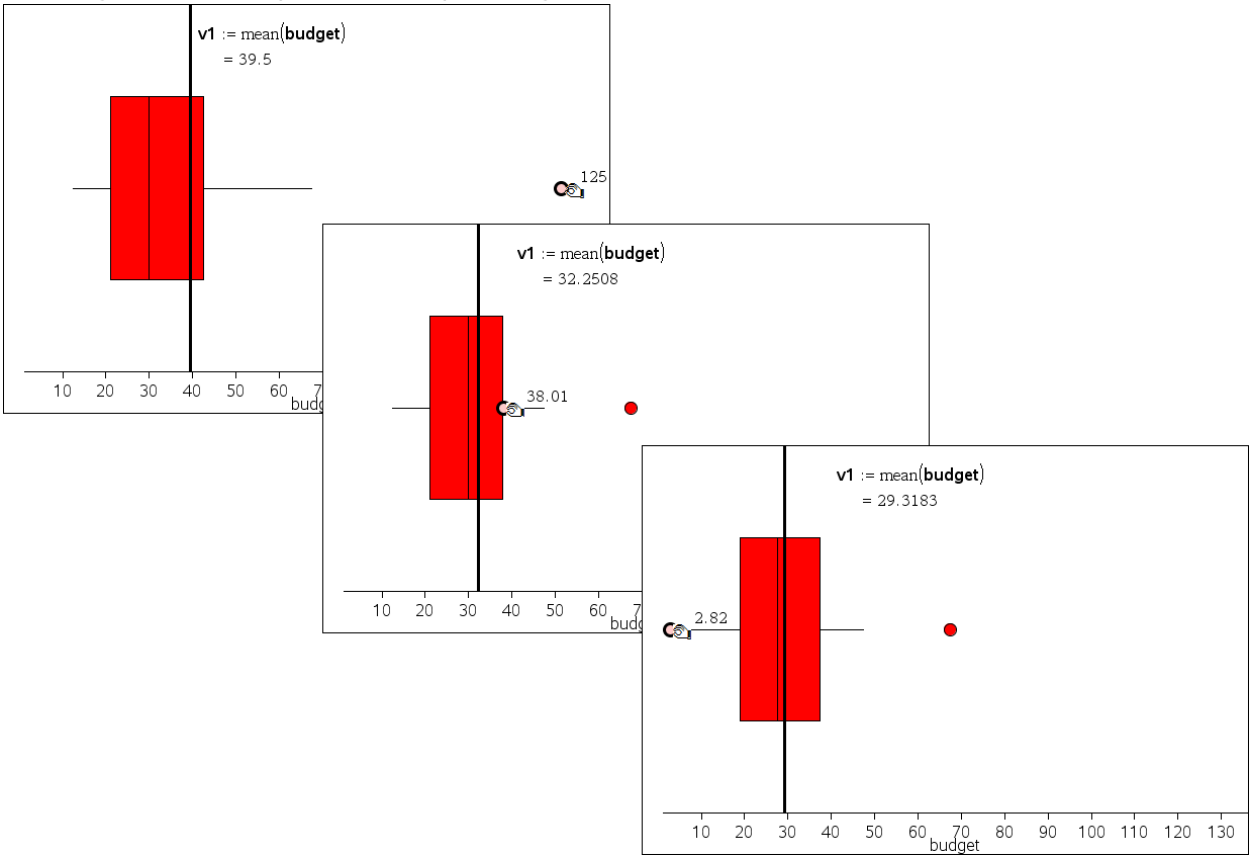
Inden vi forlader eksemplet med superligaen vil vi se på endnu en statistisk deskriptor som kan være af stor nytte til beskrivelse af data. Som et mål for den centrale eller typiske værdi har vi indtil videre benyttet medianen. Men i mange sammenhænge vil vi foretrække **middelværdien** eller **gennemsnittet**. I det ovenstående tilfælde skal vi altså finde det samlede budget og dividere det med antallet af klubber, dvs. 12.



Grafisk tilføjer vi middelværdien ved at vælge menupunktet **Plot værdi** fra **Undersøg data**-menuen. Herefter indskriver vi formlen **mean(budget)** for at få tegnet middelværdien (og tilsvarende formlen **median(budget)** for at få tegnet medianen)

$$middelbudget = \frac{12.5 + \cdots + 125}{12} = 39.5$$

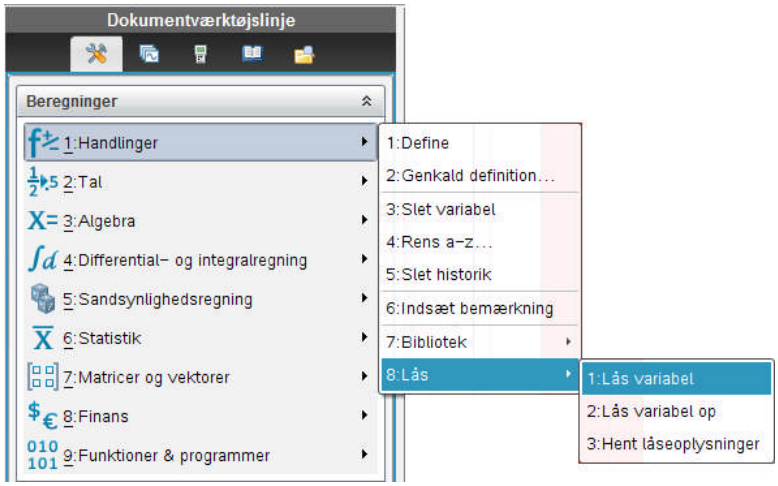
Gennemsnitsbudgettet i superligaen er altså 39.5 millioner kr., hvilket ligger et godt stykke over medianen på 30 millioner kr. Det afspejler den uforholdsmæssige store indflydelse afvigeren har på gennemsnittet. Det er derfor det er godt også at have medianen til rådighed. Medianen ligger altid i den centrale klump. Ydermere er medianen **robust** dvs. påvirkes ikke af tilstedeværelsen af en enkelt eller nogle få afvigere. Medianen er derfor et bedre mål for det *typiske årsbudget i superligaen*. For at undersøge den indflydelse det ekstreme årsbudget har på middelværdien kan vi bare gribe fat i det perifere datapunkt og trække i det.



Man kan da netop se, hvordan middelbudgettet hele tiden flytter sig, mens medianbudgettet ligger stille indtil vi kører forbi medianen og selv da giver den kun et lille ryk.

På samme måde rykker kvartilerne selvfølgelig også, når vi passerer dem. Læg dog mærke til, at vi ikke kan trække middelbudgettet forbi medianbudgettet i det ovenstående eksempel. Fordelingen forbliver altså højreskæv.

> *Bemærkning*: En sådan dynamisk tilpasning af modellen er fin til at demonstrere, hvilken indflydelse udvalgte data har på modellen. Men vi kan også ved et uheld komme til at flytte rundt på datapunkterne. Sker uheldet alligevel må man fortryde et passende antal gange indtil den oprindelige værdi er genoprettet!



Bemærkning: For at undgå at man utilsigtet kommer til at trække i et datapunkt og dermed ændre dets værdi kan man *låse* en variabel, i dette tilfælde **budget**-variablen. Det sker ved hjælp af kommandoen **lock**, der dog ikke er tilgængelig i **Lister og Regneark**-værkstedet. For at låse en variabel må men derfor åbne **Beregninger**-værkstedet, der er det eneste værksted, hvor man kan låse variable. Som det ses kan man også få oplyst status for en variabel ved hjælp af **Hent låseoplysninger**, dvs. **getLockInfo**()-kommandoen. Hvis variablen er ulåst returneres 0, hvis den er låst returneres 1. På denne måde kan man styre hvilke variable man kan trække i og hvordan man kan trække i dem i **Diagrammer og Statistik**-værkstedet.

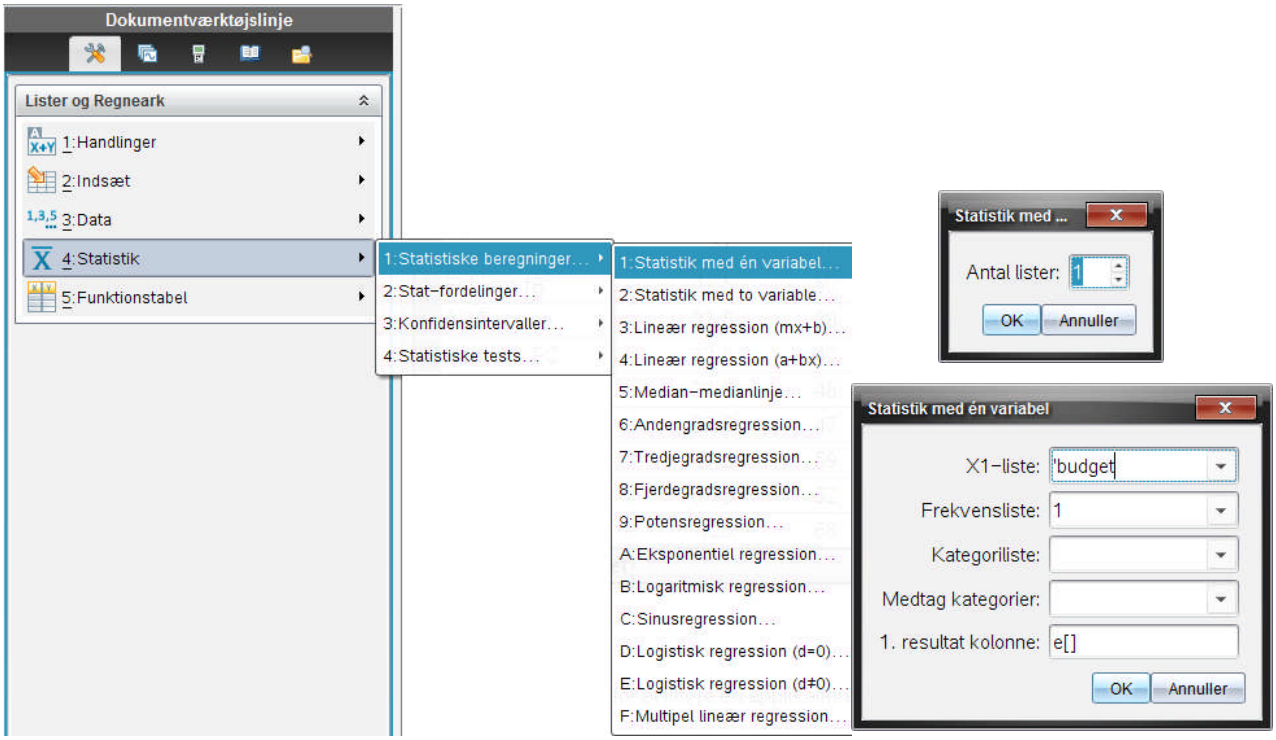
2.5 Statistiske beregninger

Til slut vil vi se lidt på mulighederne for at udføre simple statistiske **beregninger**. Vi kan da dels udnytte muligheden for at skrive formler i cellerne fra regnearket. Formler er karakteriseret ved at de starter med et lighedstegn =. Så for at udregne middelbudget og medianbudget benytter vi som vist formlerne D1 =mean(**budget**) og D2 =median(**budget**):

	A klub	B budget	C points	D	E
◆					
1	HB Køge	12.5	19	39.5	
2	SønderjyskE	17	41	30.	
3	FC Nordsjælland	21	43		
4	Silkeborg	21	43		
5	Esbjerg fB	27.5	50		
6	AGF	27.5	38		
7	Randers FC	32.5	40		
8	AaB	37.5	48		
9	FC Midtjylland	37.5	47		
10	OB	47.5	59		
11	Brøndby IF	67.5	52		
12	FC København	125	68		

D2 =median(**budget**)

Men det afgørende er da, at vi holder os fri af de søjler vi har brugt til de variable. Vi skal altså holde os 'under dobbeltstregen', dvs. indenfor celleregnearket.



Dels kan vi som vist på figuren vælge menupunktet **Statistiske beregninger > Statistik med én variabel** fra **Statistik**-menuen. Dermed har vi adgang til alle de statistiske deskriptorer som vi har introduceret i det foregående.

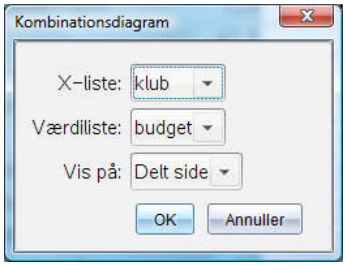
Her skal vi først og fremmest være opmærksomme på at vi både skal angive den liste, som vi vil udregne statistikken for, og den søjle som vi vil have resultaterne skrevet i (og hvor det er vores eget ansvar at den ikke overskriver andre resultater). De øvrige muligheder, frekvenslisten og kategorilisten giver mulighed for mere detaljerede statistiske beregninger. Hvis vi fx har indtastet en hyppighedstabel, så er det listen over hyppighederne, der skal noteres i frekvenslisten. Hvis vi ønsker at splitte resultaterne på kategorier, fx klubber i Jylland og klubber på øerne, så kan det også lade sig gøre at udregne de såkaldte subtotaler. Men her udregner vi bare statistikken for én samlet liste uden at tage hensyn til diverse filtre. Her har vi markeret de fem statistiske nøgletal (der ligger til grund for boksplottet). Vi ser også middelværdien (der noteres som $\bar{x}$), samt de to spredningsmål, stikprøvespredningen sx og populationsspredningen σx. Endelig oplyses antallet af data, i dette tilfælde n = 12.

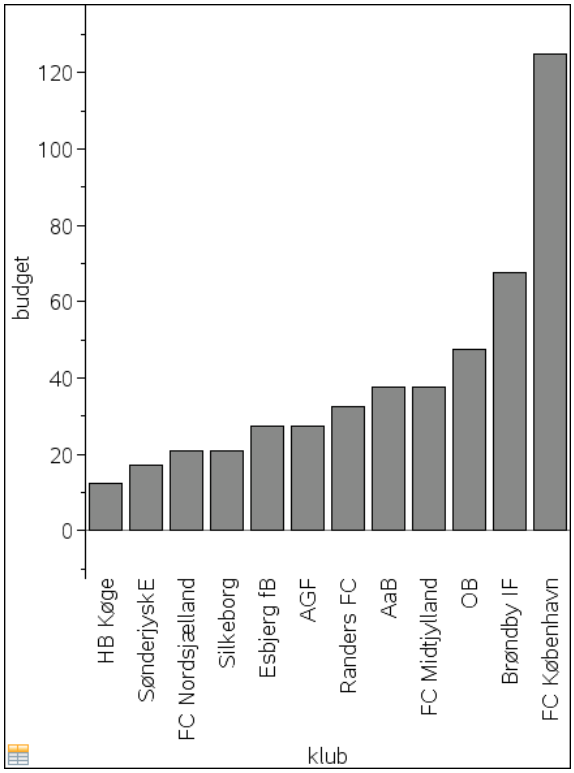
E	F
	=OneVar('budget,1): Copy\
Titel	Statistik med én variabel
$\bar{x}$	39.5
Σx	474.
Σx^2	29146.
$sx := sn{-}1x$	30.7822
$\sigma x := \sigma nx$	29.4717
n	12.
MinX	12.5
Q_1X	21.
MedianX	30.
Q_3X	42.5
MaxX	125.
$SSX := \Sigma(x-\bar{x})^2$	10423.

Øvelse: Prøv nu selv at undersøge pointsfordelingen for de 12 klubber i superligaen.

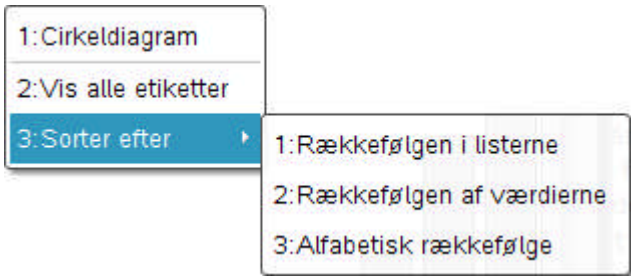
2.6 Kombinationsdiagrammer

Vi har hidtil håndteret budget som en numerisk variabel, men de enkelte budgetter er jo tilknyttet klubberne og det kan i forskellige sammenhænge være hensigtsmæssigt at få klubberne med ind i diagrammet. Man kan da anvende et kombinationsdiagram, som kombinerer en kategorisk variabel, her **klub**, med en numerisk variabel, her **budget**. Diagrammet oprettes da principielt som et kategorisk diagram, dvs. et søjlediagram eller et cirkeldiagram, men søjlens højde afspejler budgettets størrelse, ligesom cirkeludsnittets størrelse afspejler budgettets størrelse. Det ser således ud: Marker de to første søjler med oplysninger om **klub** og **budget** – højreklik i det markerede område og vælg **kombinationsdiagram**. Du føres da til en dialogboks, der beder dig præcisere navnet på den kategoriske liste (X-listen) og den tilhørende værdiliste. Der oprettes da automatisk et vindue med et **Diagrammer og statistik**-værksted:





Som udgangspunkt vises klubberne i samme rækkefølge som i kategorilisten (som her var ordnet fra mindst til størst budget). Men rækkefølgen kan man selv bestemme. Højreklikkes får man adgang til at sortere efter rækkefølgen i listerne, rækkefølgen efter værdierne (fra størst til mindst) og alfabetisk rækkefølge.



Men man kan også bare gribe en etiket for en af kategorierne og flytte rundt på den med musen efter forgodtbefindende. Der er altså frit slag!

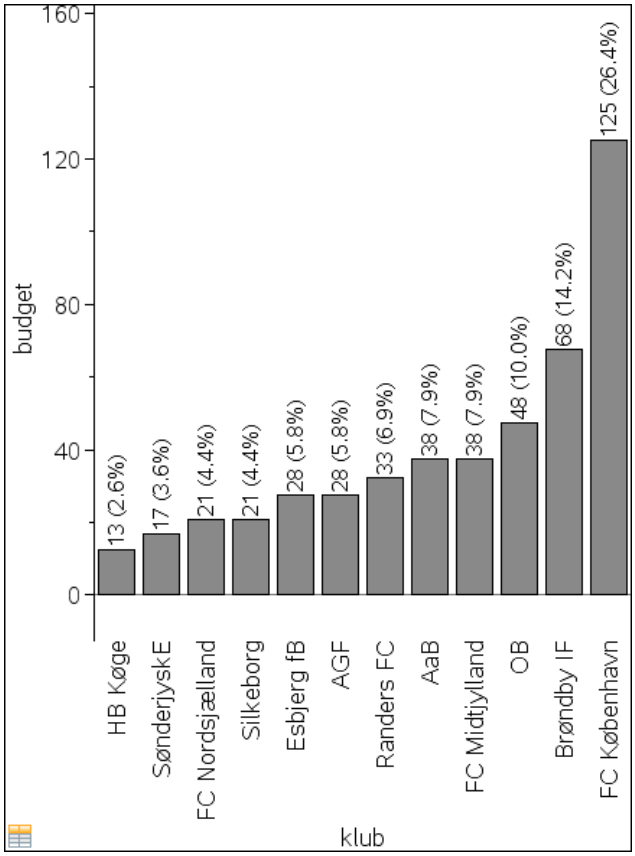
Som det ses er der også mulighed for at få vist alle etiketter, altså oplysninger om de enkelte søjler på én gang. Det kan da betale sig at justere den lodrette akse, så der bliver plads over søjlerne til at vise oplysningerne:

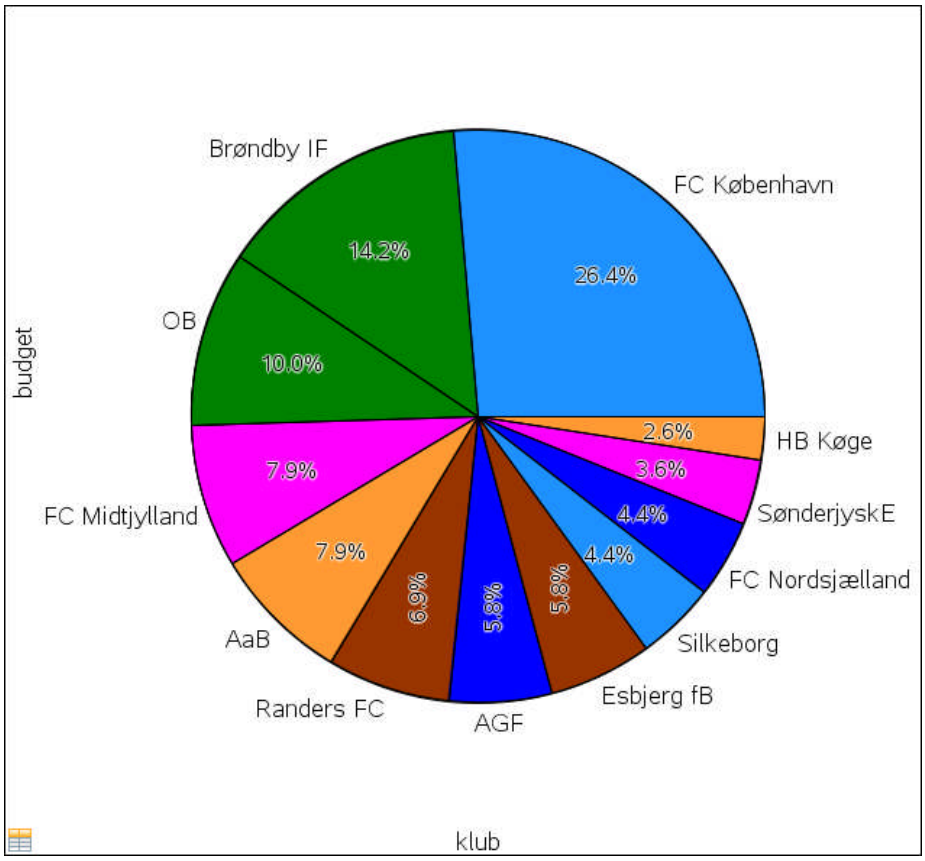
Vi ser da fx at FC København bruger mere end ¼ af det samlede budget i superligaen og næsten dobbelt så meget som sin nærmeste konkurrent Brøndby IF. Det er altså ikke uden grund at FC København blev skilt ud som en afviger!

De samme tendenser kan selvfølgelig også ses på cirkeldiagrammet vist nedenunder med **Vis alle etiketter** slået til. Denne gang får vi kun oplyst procentfordelingen, som netop er cirkeldiagrammets særkende.

Man kan godt selv oprette kombinationsdiagrammer i diagrammer og statistik. I så fald skal man blot højreklikke i diagramområdet og vælge **Tilføj X-variabel med værdiliste** eller **Tilføj Y-variabel med værdiliste**, hvis man foretrækker vandrette søjler.

Bemærkning: Man kan også godt tilføje flere værdilister ved at højreklikke og derved oprette et grupperet søjlediagram.





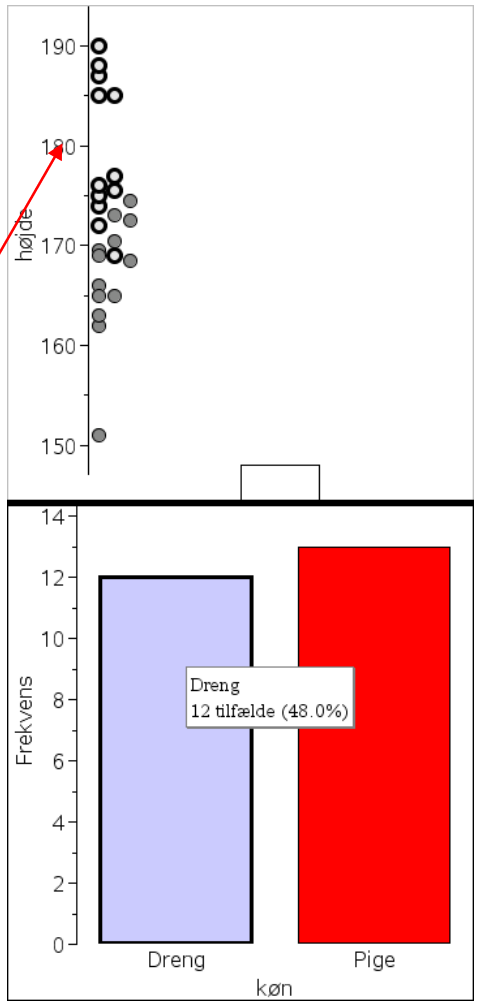
Afslutningsvis bemærker vi at kombinationsdiagrammet har sit eget logo , så man aldrig behøver være i tvivl om hvordan diagrammet er oprettet; Og at den numeriske variabel for et søjlediagram godt kan antage negative værdier. Derimod kræver et cirkeldiagram positive værdier.

3 Numeriske variable opdelt på kategorier

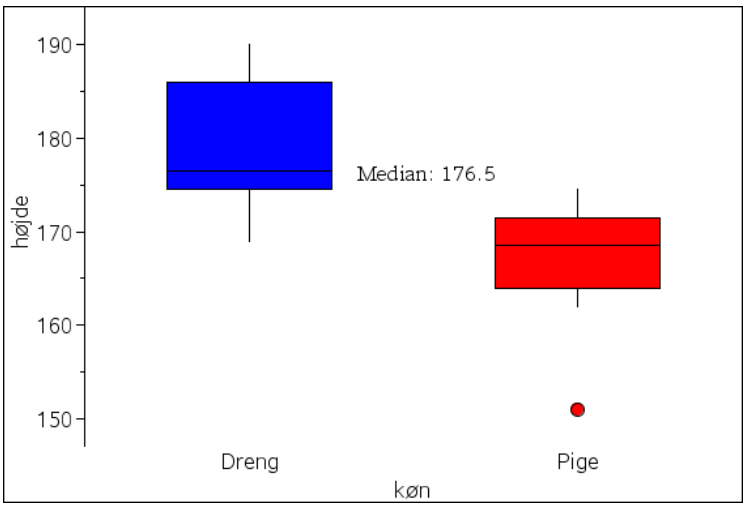
3.1 Grafisk sammenligning af to fordelinger fra ét datasæt

Vi ser nu på den situation, hvor vi har brug for at sammenligne to fordelinger. Som udgangspunkt kan vi fx se på følgende datasæt, hvor vi har målt højderne for en bestemt 1g-klasse:

	A elev	B køn	C højde
◆			
1	Ali	Dreng	172
2	Andreas	Dreng	187
3	Camilla	Pige	162
4	Cecilie	Pige	173
5	Christian	Dreng	174
6	Christina	Pige	163
7	Cilie	Pige	169.5
8	Heidi	Pige	169
9	Jakob	Dreng	175
10	Jamal	Dreng	169
11	Jane	Pige	166
12	Julian	Dreng	176
13	Kasper L	Dreng	185
14	Kasper O	Dreng	190
15	Louise K	Pige	165
16	Louise R	Pige	168.5
17	Maria	Pige	170.5
18	Martin	Dreng	175.5
19	Marzia	Pige	151
20	Mohamed	Dreng	185
21	Nikita	Dreng	177
22	Paw	Dreng	188
23	Rose	Pige	174.5
24	Sofie K	Pige	165
25	Sofie T	Pige	172.5



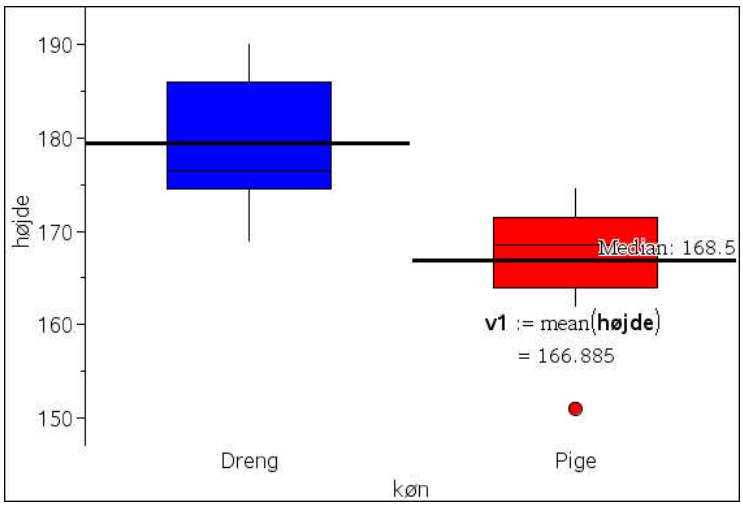
Af prikdiagrammet ses, at højderne falder i to større klumper, en stor central klump omkring 170 cm og en mindre klump øverst omkring 187 cm. Men opretter vi også et søjlediagram for køn i klassen og markerer vi drengene kan vi tydeligt se i prikdiagrammet at drengenes højder typisk ligger øverst i den samlede fordeling og pigernes tilsvarende nederst i den samlede fordeling. Det gør det nærtliggende at undersøge sammenhængen mellem **køn** og **højde**, dvs. sammenligne drengenes højde og pigernes højde. Vi afsætter da den uafhængige variabel **køn** ud af førsteaksen og den afhængige variabel **højde** op af andenaksen (og skifter til boksplot). Derved får vi netop *kombineret* de to grafer for oven.



Det ses da tydeligt at drengenes boksplot ligger væsentligt højere end pigernes boksplot, idet de to kasser er skilt helt ad (dvs. drengenes første kvartil ligger over pigernes tredje kvartil,

ligesom drengenes median ligger over pigernes maksimum og tilsvarende ligger pigernes median under drengenes minimum). De højeste 50% af drengene er altså højere end samtlige piger ligesom de laveste 50% af pigerne er lavere end samtlige drenge.

Vi ser også at drengenes fordeling ser ud til at være højreskæv, mens pigernes fordeling ser ud til at være venstreskæv. Det kan vi bekræfte ved at tilføje middelværdierne til boksplottet, idet vi vælger **Plot værdi** fra **Undersøg data**-menuen:



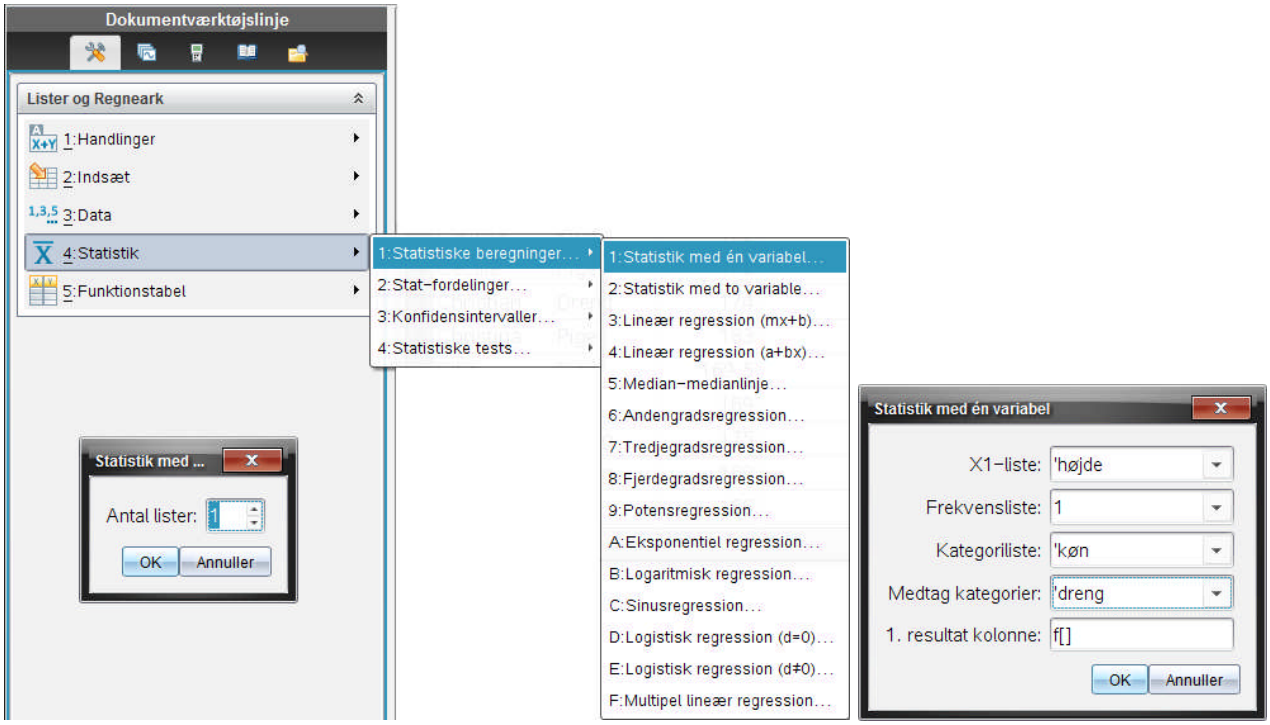
Vi ser da netop at middelhøjden for drengens tilfælde ligger over medianen, mens den for pigernes tilfælde ligger under medianen.

3.2 Statiske beregninger for to fordelinger fra ét datasæt: Subtotaler*

Vi kan også udføre **statistiske beregninger** på de to køn hver for sig, dvs. finde **subtotaler**. Det er lidt kompliceret, så afsnittet kan overspringes i første gennemlæsning. Det kræver at vi først opretter særskilte kategorilister for de to køn, hvor vi skriver "Dreng" i listen for **dreng** og "Pige" i listen for **pige**:

	A navn	B køn	C højde	D dreng	E pige
◆					
1	Ali	Dreng	172	Dreng	Pige
2	Andreas	Dreng	187		
3	Camilla	Pige	162		
4	Cecilie	Pige	173		
5	Christian	Dreng	174		
6	Christina	Pige	163		
7	Cilie	Pige	169.5		
8	Heidi	Pige	169		
9	Jakob	Dreng	175		
10	Jamal	Dreng	169		

Vi placerer så markøren i den sidste kolonne, her e[] og vælge menupunktet **Statistik>Statistiske beregninger>Statistik med en én variabel**: Vi svarer herefter ja til en enkelt liste (nemlig listen over højder), men i dialogboksen for statistik med én variabel tilføjer vi nu kategorilisten **køn**, dvs. vi meddeler at den statistiske beregning for **højde** skal afhænge af hvad der står i kategorilisten **køn**. Herefter sætter vi et *filter* på i **Medtag kategorier**, idet vi kun medtager **dreng**! Resultatet er da netop at vi får udregnet drengens middelhøjde osv. Bagefter gentager vi det samme men nu med **pige** i **Medtag kategorier**, dvs. denne gang er det kun pigerne, der inddrages i den statistiske beregning.



	F	G	H	I
◆		=OneVar('højde,1,'køn,'dreng)		=OneVar('højde,1,'køn,'pige)
1	Titel	Statistik med én variabel	Titel	Statistik med én variabel
2	$\bar{x}$	179.458	$\bar{x}$	166.885
3	Σx	2153.5	Σx	2169.5
4	Σx^2	387014.	Σx^2	362509.
5	sx := Sn-...	7.07575	sx := Sn-...	6.14462
6	σx := σn...	6.77452	σx := σn...	5.90356
7	n	12.	n	13.
8	MinX	169.	MinX	151.
9	Q_1X	174.5	Q_1X	164.
10	MedianX...	176.5	MedianX...	168.5
11	Q_3X	186.	Q_3X	171.5
12	MaxX	190.	MaxX	174.5
13	SSX := Σ...	550.729	SSX := Σ...	453.077

I =OneVar('højde,1,'køn,'pige): CopyVar Stat., Stat2.

Bemærkning: De statistiske resultater af de to beregninger gemmes i hver sin statistiske variabel. Det fremgår af formelfelterne, hvoraf vi kun ser det sidste i det ovenstående skærmbillede:

Man kan ikke referere direkte til statistiske variable i **Lister og regneark**-værkstedet, men man kan gøre det i **Noter**-værkstedet eller **Beregninger**-værkstedet. Her viser vi det i **Noter**-værkstedet, hvor vi opretter et matematikfelt (**CTRL M**), skriver **stat1.** og vælger fx $\bar{x}$

Læg mærke til at variablen $\mathbf{stat1.\bar{x}}$ har et fornavn **stat1** og et efternavn $\bar{x}$ adskilt af et punktum. Så snart man taster punktummet, dukker som vist et rullevindue op med de mulige variabelnavne hørende til **stat1**. Det gør det meget nemt at fiske værdierne for de statistiske beregninger!

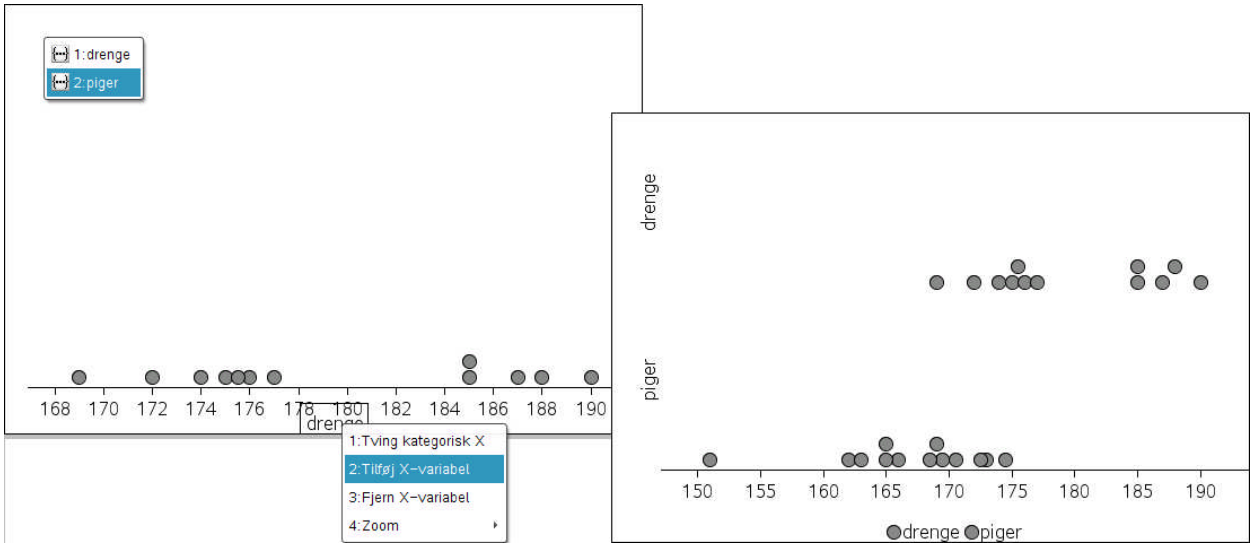
3.3 Grafisk sammenligning af to adskilte/ustakkede datasæt

I det ovenstående eksempel med klassens højder var alle data samlet i et enkelt datasæt. Men sådan er det desværre ikke altid. Ofte præsenteres man for data der på forhånd er skilt ud i to separate datasæt. Man kan da enten samle de to adskilte datasæt i et enkelt datasæt (vi viser senere hvordan) eller man kan arbejde direkte med de to adskilte datasæt.

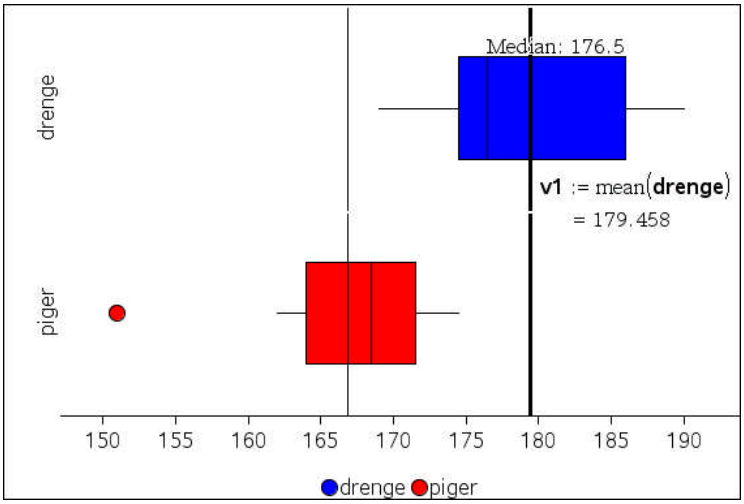
	A drenge	B piger
◆		
1	172	162
2	187	173
3	174	163
4	175	169.5
5	169	169
6	176	166
7	185	165
8	190	168.5
9	175.5	170.5
10	185	151
11	177	174.5
12	188	165
13		172.5

Vi viser teknikkerne til det sidste med det samme eksempel med højderne, men nu blot præsenteret som to datasæt, ét for drengenes højder og ét for pigernes højder (i modsætning til før hvor hver række handlede om ét bestemt individ og angav navnet, kønnet og højden for dette individ er der nu ikke længere nogen umiddelbar tilknytning mellem dataene i de forskellige søjler, der oven i købet har forskellig længde).

Det er nu trivielt at få oprettet et prikplot over drengenes højder ved at afsætte dem langs førsteaksen. Men hvordan får vi nu pigernes højder med ind i det samme diagram, så vi umiddelbart kan sammenligne dem? Det gør vi ved at højreklikke på aksefeltet og vælge **Tilføj X-variabel**, hvorefter vi tilføjer pigernes højder.



Derefter kan de selvfølgelig omdannes til boksplot på sædvanlig vis. Men denne gang kan vi altså ikke afsætte boksplottene lodret som i det foregående tilfælde, hvor vi havde et frit valg:



Hvis vi nu vil tilføje middelhøjderne ved hjælp af **Plot værdi** fra **Undersøg data**-menuen må vi også gøre det af to omgange, én for drenge og én for piger.

3.4 Statistiske beregninger for to adskilte/ustakkede datasæt

Det er altså lidt mere klodset at håndtere sammenligningen af to fordelinger grafisk, når datasættet som udgangspunkt er splittet på to adskilte datasæt. Til gengæld er det nemmere at udføre de statistiske beregninger. Som før vælger vi **Statistik med én variabel** fra **Statistik>stat-beregning** menuen, men denne gang vælger vi bare de to lister!

	A drenge	B piger	C	D	E
◆				=OneVar('drenge,1)	=OneVar('piger,1)
1	172	162	Titel	Statistik med én v...	Statistik med én...
2	187	173	$\bar{x}$	179.458	166.885
3	174	163	Σx	2153.5	2169.5
4	175	169.5	Σx^2	387014.	362509.
5	169	169	sx := Sn-...	7.07575	6.14462
6	176	166	σx := σn...	6.77452	5.90356
7	185	165	n	12.	13.
8	190	168.5	MinX	169.	151.
9	175.5	170.5	Q_1X	174.5	164.
10	185	151	MedianX...	176.5	168.5
11	177	174.5	Q_3X	186.	171.5
12	188	165	MaxX	190.	174.5
13		172.5	SSX := Σ...	550.729	453.077

E =OneVar('piger,1): CopyVar Stat., Stat4.

En anden fordel med de adskilte datasæt er at det er nemt at farvelægge regnearket i overensstemmelse med grafernes farvelægning:

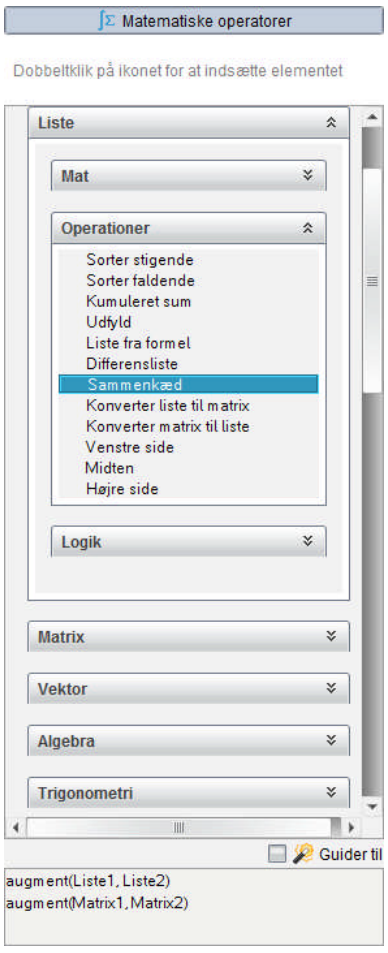
	A drenge	B piger	C	D	E
◆				=OneVar('drenge,1)	=OneVar('piger,1)
1	172	162	Titel	Statistik med én v...	Statistik med én...
2	187	173	$\bar{x}$	179.458	166.885
3	174	163	Σx	2153.5	2169.5
4	175	169.5	Σx^2	387014.	362509.
5	169	169	sx := Sn-...	7.07575	6.14462
6	176	166	σx := σn...	6.77452	5.90356
7	185	165	n	12.	13.
8	190	168.5	MinX	169.	151.
9	175.5	170.5	Q_1X	174.5	164.
10	185	151	MedianX...	176.5	168.5
11	177	174.5	Q_3X	186.	171.5
12	188	165	MaxX	190.	174.5
13		172.5	SSX := Σ...	550.729	453.077

C

3.5 Om samling og opdeling af datasæt*

Det følgende afsnit er lidt teknisk og kan roligt overspringes i første omgang! Men i projektarbejder vil man ofte komme ud for at det er praktisk at kunne opdele et enkelt datasæt i flere delsæt og tilsvarende samle adskilte datasæt i et enkelt datasæt. Først samler vi det ovenstående adskilte datasæt. Vi tilføjer da først yderligere to lister svarende til variablen med det 'skjulte køn'. Derefter *stabler* vi variablene ved at bruge **augment**-kommandoen for sammenkædning af lister

	A drenge_køn	B drenge	C piger_køn	D piger
◆				
1	Dreng	172	Pige	162
2	Dreng	187	Pige	173
3	Dreng	174	Pige	163
4	Dreng	175	Pige	169.5
5	Dreng	169	Pige	169
6	Dreng	176	Pige	166
7	Dreng	185	Pige	165
8	Dreng	190	Pige	168.5
9	Dreng	175.5	Pige	170.5
10	Dreng	185	Pige	151
11	Dreng	177	Pige	174.5
12	Dreng	188	Pige	165
13			Pige	172.5
A1	"Dreng"			



	A drenge_køn	B drenge	C piger_køn	D piger	E køn	F højde
◆					=augmen	=augmen
1	Dreng	172	Pige	162	Dreng	172
2	Dreng	187	Pige	173	Dreng	187
3	Dreng	174	Pige	163	Dreng	174
4	Dreng	175	Pige	169.5	Dreng	175
5	Dreng	169	Pige	169	Dreng	169
6	Dreng	176	Pige	166	Dreng	176
7	Dreng	185	Pige	165	Dreng	185
8	Dreng	190	Pige	168.5	Dreng	190
9	Dreng	175.5	Pige	170.5	Dreng	175.5
10	Dreng	185	Pige	151	Dreng	185
11	Dreng	177	Pige	174.5	Dreng	177
12	Dreng	188	Pige	165	Dreng	188
13			Pige	172.5	Pige	162
14					Pige	173
15					Pige	163
16					Pige	169.5
17					Pige	169
18					Pige	166
19					Pige	165
20					Pige	168.5
21					Pige	170.5
22					Pige	151
23					Pige	174.5
24					Pige	165
25					Pige	172.5
F	højde:=augment('drenge,'piger)					

Dermed er det oprindelige adskilte datasæt baseret på listerne **drenge** og **piger** omdannet til et enkelt samlet datasæt baseret på listerne **køn** og **højde**. Disse to nye lister kan nu evt. kopieres over i en ny opgave eller et nyt dokument for videre bearbejdning.

Vender vi i stedet tilbage til det oprindelige datasæt baseret på hele klassens højdetal må vi nu først sortere datasættet efter **køn**:

	A elev	B køn	C højde
1	Ali	Dreng	172
2	Andreas	Dreng	187
3	Camilla	Pige	162
4	Cecilie	Pige	173
5	Christian	Dreng	174
6	Christina	Pige	163
7	Cilie	Pige	169.5
8	Heidi	Pige	169
9	Jakob	Dreng	175
10	Jamal	Dreng	169
11	Jane	Pige	166
12	Julian	Dreng	176
13	Kasper L	Dreng	185
14	Kasper O	Dreng	190
15	Louise K	Pige	165
16	Louise R	Pige	168.5
17	Maria	Pige	170.5
18	Martin	Dreng	175.5
19	Marzia	Pige	151
20	Mohamed	Dreng	185
21	Nikita	Dreng	177
22	Paw	Dreng	188
23	Rose	Pige	174.5
24	Sofie K	Pige	165
25	Sofie T	Pige	172.5

	A elev	B køn	C højde
1	Ali	Dreng	172
2	Andreas	Dreng	187
3	Christian	Dreng	174
4	Jakob	Dreng	175
5	Jamal	Dreng	169
6	Julian	Dreng	176
7	Kasper L	Dreng	185
8	Kasper O	Dreng	190
9	Martin	Dreng	175.5
10	Mohamed	Dreng	185
11	Nikita	Dreng	177
12	Paw	Dreng	188
13	Camilla	Pige	162
14	Cecilie	Pige	173
15	Christina	Pige	163
16	Cilie	Pige	169.5
17	Heidi	Pige	169
18	Jane	Pige	166
19	Louise K	Pige	165
20	Louise R	Pige	168.5
21	Maria	Pige	170.5
22	Marzia	Pige	151
23	Rose	Pige	174.5
24	Sofie K	Pige	165
25	Sofie T	Pige	172.5

Derefter er det nemt at sværte drengenes højder til og overføre dem til en ny liste og tilsvarende for pigernes højde:

	A elev	B køn	C højde	D drenge
				=c1:c12
1	Ali	Dreng	172	
2	Andreas	Dreng	187	
3	Christian	Dreng	174	
4	Jakob	Dreng	175	
5	Jamal	Dreng	169	
6	Julian	Dreng	176	
7	Kasper L	Dreng	185	
8	Kasper O	Dreng	190	
9	Martin	Dreng	175.5	
10	Mohamed	Dreng	185	
11	Nikita	Dreng	177	
12	Paw	Dreng	188	
13	Camilla	Pige	162	
14	Cecilie	Pige	173	

	A elev	B køn	C højde	D drenge	E piger
				=c1:c12	=c13:c25
1	Ali	Dreng	172	172	162
2	Andreas	Dreng	187	187	173
3	Christian	Dreng	174	174	163
4	Jakob	Dreng	175	175	169.5
5	Jamal	Dreng	169	169	169
6	Julian	Dreng	176	176	166
7	Kasper L	Dreng	185	185	165
8	Kasper O	Dreng	190	190	168.5
9	Martin	Dreng	175.5	175.5	170.5
10	Mohamed	Dreng	185	185	151
11	Nikita	Dreng	177	177	174.5
12	Paw	Dreng	188	188	165
13	Camilla	Pige	162		172.5
14	Cecilie	Pige	173		

4 På opdagelse i data

Vi har nu fået en første fornemmelse for strukturen af et datasæt. Kernen i den beskrivende statistik er de grafiske fremstillinger af data. Øjet er vores bedste mønstergenkender, så når vi vil danne os et overblik over strukturen for et datasæt, dvs. fordelingen af de variable, er det langt det nemmeste, hvis vi begynder med at visualisere fordelingen før vi kigger dybt i tabeller og begynder at udføre indviklede beregninger. Det udtrykkes ofte med sloganet: 'Du skal tegne, før du kan regne'.

Den mest grundlæggende graftype for en numerisk variabel hørende til et datasæt er prikplottet. Men dertil kommer så de supplerende graftyper: histogram, boksplot og normalfordelingsplot, som hver for sig er gode til at fremhæve forskellige sider af strukturen.

Hvad er det så man kan hæfte sig ved, når man forsøger at danne sig et indtryk af strukturen for en variabel? Der er første og fremmest følgende tre kendetegn: *Form*, *niveau* og *spredning.*

1. **Form**: Ligger dataene jævnt fordelt ud over et interval eller er de fleste data samlet i en eller flere klumper? Ligger dataene symmetrisk eller ligger de skævt. Som hjælp til det sidste kan man også se på forskellen mellem middelværdien og medianen, der kan opfattes som et mål for skævheden.
2. **Niveau**: Hvad er den typiske værdi for variablen? Hvis dataene med tilnærmelse ligger symmetrisk, vil vi ofte foretrække middelværdien som den typiske værdi, men hvis dataene ligger skævt vil vi ofte foretrække medianen som den typiske værdi.
3. **Spredning**: Ligger dataene meget tæt på den typiske værdi eller spreder de sig ud over et stort område? For en jævn fordeling, vil vi ofte benytte **variationsbredden** (forskellen mellem den mindste og den største værdi) som et mål for spredningen. For en fordeling med en central pukkel og lange haler ud til siden vil vi ofte benytte kvartilbredden (tykkelsen af kvartilboksen) som et mål for spredningen.

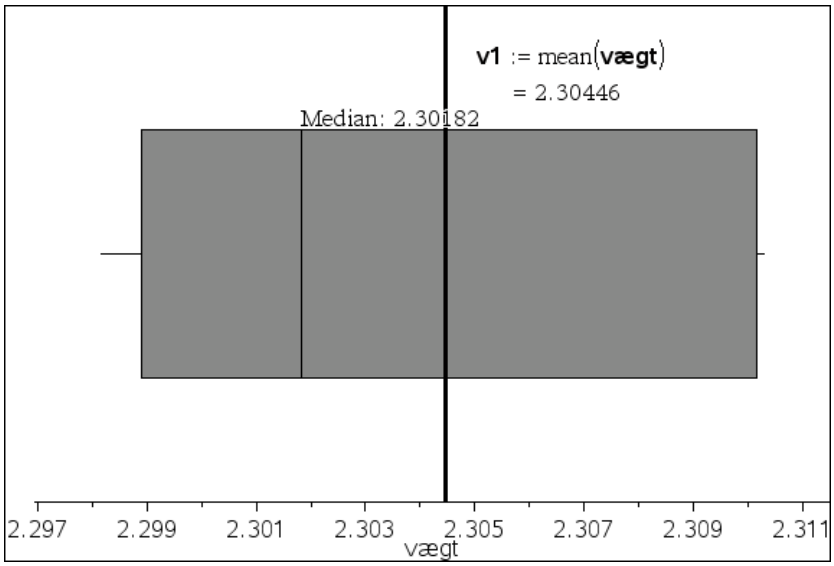
Men andre kendetegn kan også falde i øjnene: Er der fx tydelige huller? Er der tydelige perifere/afvigende observationer osv.

4.1 Rayleigh og densiteten for kvælstof

Som et typisk eksempel på en opdagelsesrejse i et datasæt vil vi se på et berømt historisk eksempel (http://web.lemoyne.edu/~GIUNTA/RAYLEIGH.HTML): Rayleighs undersøgelse af densiteten for kvælstof N_2, som udgør den vigtigste komponent i atmosfærisk luft. Den næst vigtigste er ilt O_2. Ved at fjerne ilten fra atmosfærisk tør luft kunne han isolere kvælstoffet. Tilsvarende kunne han frembringe rent kvælstof ved at nedbryde forskellige simple kemiske forbindelser. Derved fandt han frem til følgende eksperimentelle data (der kan hentes på Texas hjemmeside)

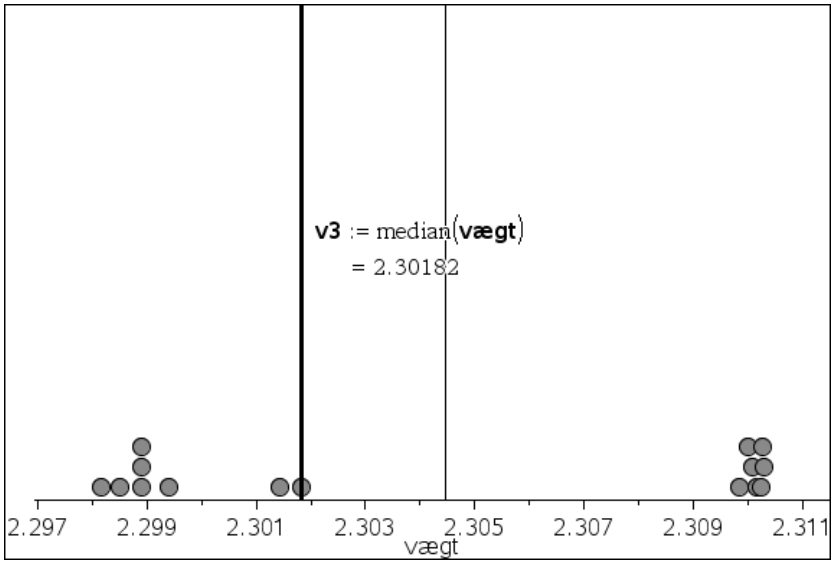
	A dato	B kilde	C metode	D vægt
◆				
1	29-nov-93	Nitrogenoxid	Varmt jern	2.30143
2	05-dec-93	Nitrogenoxid	Varmt jern	2.29816
3	06-dec-93	Nitrogenoxid	Varmt jern	2.30182
4	08-dec-93	Nitrogenoxid	Varmt jern	2.2989
5	12-dec-93	Atmosfærisk luft	Varmt jern	2.31017
6	14-dec-93	Atmosfærisk luft	Varmt jern	2.30986
7	19-dec-93	Atmosfærisk luft	Varmt jern	2.3101
8	22-dec-93	Atmosfærisk luft	Varmt jern	2.31001
9	26-dec-93	Dinitrogenoxid	Varmt jern	2.29889
10	28-dec-93	Dinitrogenoxid	Varmt jern	2.2994
11	09-jan-94	Ammoniumnitrogendioxid	Varmt jern	2.29849
12	13-jan-94	Ammoniumnitrogendioxid	Varmt jern	2.29889
13	27-jan-94	Atmosfærisk luft	Jernhydrat	2.31024
14	30-jan-94	Atmosfærisk luft	Jernhydrat	2.3103
15	01-feb-94	Atmosfærisk luft	Jernhydrat	2.31028

For at danne sig et indtryk af fordelingen for de målte vægte afbildes de i et prikdiagram henholdsvis et boksplot.

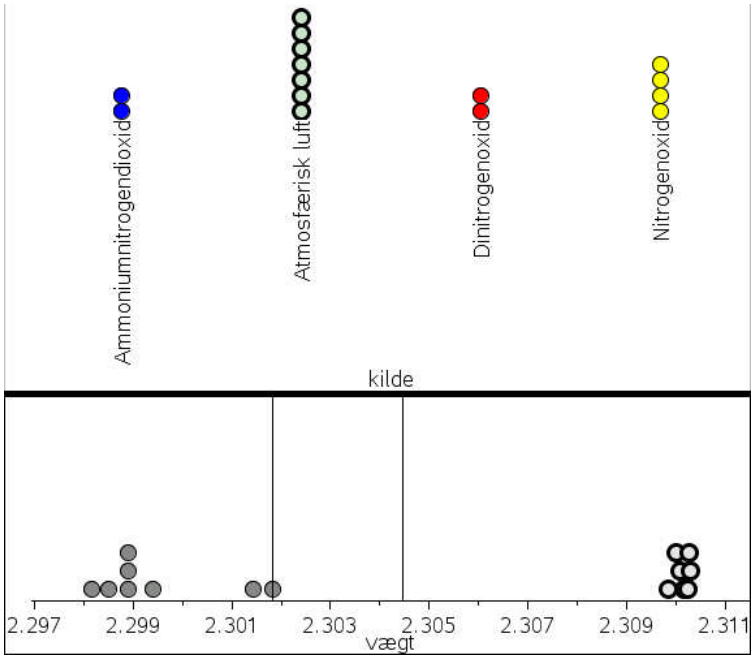


Prøver vi først at se på boksplottet er boksen usædvanlig bred i forhold til de to haler. Samtidigt er fordelingen tydeligt højreskæv, idet den højre del af boksen er meget større end den venstre del; dette bekræftes yderligere af at middelværdien ligger langt inde i den højre del. Men der ud over er det svært at se på boksplottet, hvad det egentlig er, der gør fordelingen så usædvanlig.

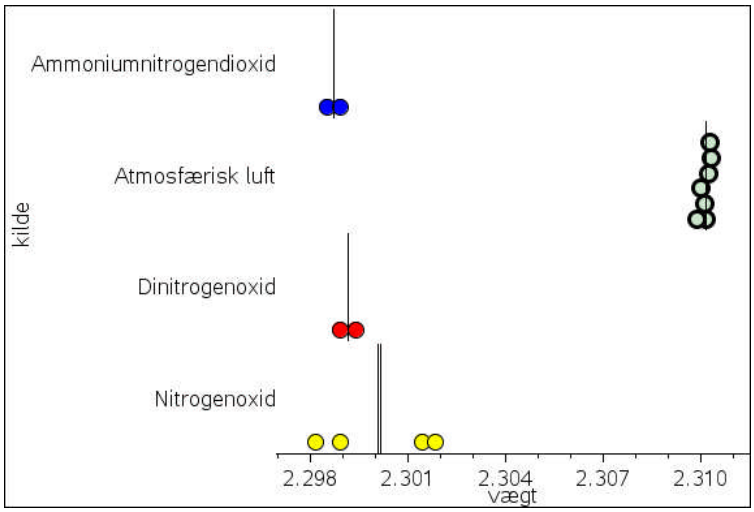
Kigger vi derimod på prikplottet falder det tydeligt i øjnene at fordelingen er skilt ad i to (måske endda tre) klumper: En snæver klump yderst til højre omkring massen 2.310g og en tilsvarende bredere klump til venstre omkring 2.299g (og måske er der endda tegn på en tredje klump omkring 2.3015g).



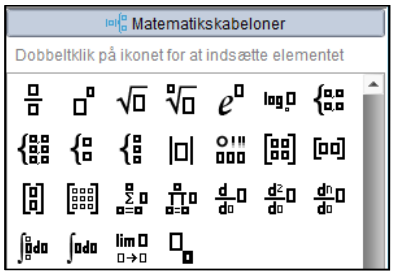
Hvad kan nu være årsagen til denne opsplitning af datasættet? Opretter vi et prikdiagram for listen **kilde** og markerer vi efterfølgende den snævre klump omkring 2.310g i prikdiagrammet for **vægt** ses det tydeligt i **kilde**-diagrammet, at den er koblet til de målinger, der stammer fra atmosfærisk luft.

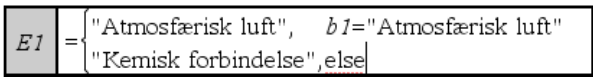


Der er altså tydeligvis en **skjult variabel**, der giver anledning til en systematisk forskel på den kvælstof, der isoleres fra den atmosfæriske luft og den kvælstof, der isoleres fra forskellige kemiske forbindelser. Det samme kan ses tydeligt på grafen, hvis vi benytter variablen **kilde** til at splitte prikplottet:



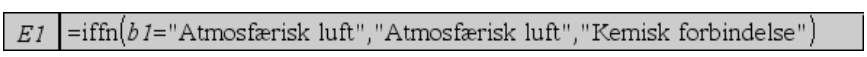
I virkeligheden er der altså tale om en sammenblanding af to adskilte datasæt. Det kan vi se endnu tydeligere, hvis vi indfører en sammensat variabel, **Oprindelse**, der skelner mellem de målinger, der stammer fra kemiske forbindelser og de målinger, der stammer fra atmosfærisk luft. Vi indskriver derfor den følgende betingede celleformel





hvor skabelonen hentes i skabelonpaletten i det venstre sidepanel.

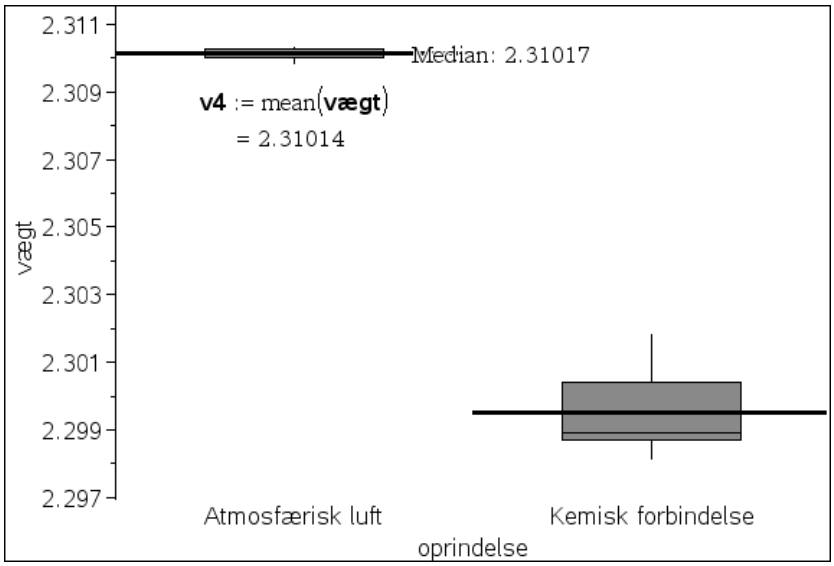
Bemærkning: Den kan også skrives ved hjælp af kommandoen IfFn (if-funktionen):



Ved at indføre **Oprindelse** som uafhængig variabel og **Vægt** som afhængig får vi netop tydeligt adskilt de to grupper af data:

	A dato	B kilde	C metode	D vægt	E oprindelse
◆					
1	29-nov-93	Nitrogenoxid	Varmt jern	2.30143	Kemisk forbindelse
2	05-dec-93	Nitrogenoxid	Varmt jern	2.29816	Kemisk forbindelse
3	06-dec-93	Nitrogenoxid	Varmt jern	2.30182	Kemisk forbindelse
4	08-dec-93	Nitrogenoxid	Varmt jern	2.2989	Kemisk forbindelse
5	12-dec-93	Atmosfærisk luft	Varmt jern	2.31017	Atmosfærisk luft
6	14-dec-93	Atmosfærisk luft	Varmt jern	2.30986	Atmosfærisk luft
7	19-dec-93	Atmosfærisk luft	Varmt jern	2.3101	Atmosfærisk luft
8	22-dec-93	Atmosfærisk luft	Varmt jern	2.31001	Atmosfærisk luft
9	26-dec-93	Dinitrogenoxid	Varmt jern	2.29889	Kemisk forbindelse
10	28-dec-93	Dinitrogenoxid	Varmt jern	2.2994	Kemisk forbindelse
11	09-jan-94	Ammoniumnitrogendioxid	Varmt jern	2.29849	Kemisk forbindelse
12	13-jan-94	Ammoniumnitrogendioxid	Varmt jern	2.29889	Kemisk forbindelse
13	27-jan-94	Atmosfærisk luft	Jernhydrat	2.31024	Atmosfærisk luft
14	30-jan-94	Atmosfærisk luft	Jernhydrat	2.3103	Atmosfærisk luft
15	01-feb-94	Atmosfærisk luft	Jernhydrat	2.31028	Atmosfærisk luft

E1 =iffn(*b1*="Atmosfærisk luft","Atmosfærisk luft","Kemisk forbindelse")



Tilbage stod så bare at identificere den sande natur af den skjulte variabel: Hvorfor adskilte den kvælstof, der blev udskilt af atmosfærisk luft sig fra den kvælstof, der blev isoleret fra en kemisk forbindelse? Rayleigh gættede på at den atmosfæriske luft måske indeholdt et ukendt stof, der forstyrrede målingerne. I så fald repræsenterer målingerne fra de kemiske forbindelser den rene kvælstof, mens målingerne fra den atmosfæriske luft er 'forurenet' af det skjulte stof. Hvis dette skjulte stof havde en højere densitet end kvælstof, ville det netop kunne trække målingerne en anelse i vejret, så de kom til at ligge på et højere niveau. Rayleigh gik på jagt efter det ukendte stof, hvorved han netop opdagede den første ædelgas, argon, hvilket udløste en nobelpris. Bemærk i øvrigt hvordan hans opdagelse kun kunne lade sig gøre, fordi han dels målte meget præcist, dels benyttede flere af hinanden uafhængige metoder til fremstillingen af kvælstof.

4.2 På jagt efter variabelsammenhænge

Hvis vi er i gang med at undersøge et datasæt vil vi typisk kigge på mulige sammenhænge mellem to variable. Der bliver da forskellige muligheder, idet der er to fundamentalt forskellige variable, de kategoriske variable og de numeriske variable. De tre typer sammenhænge vi kan se efter er derfor:

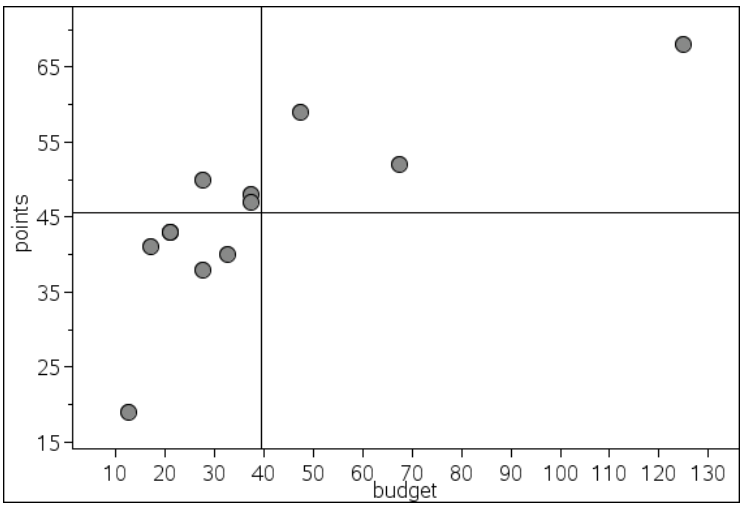
1) Sammenhængen mellem to numeriske variable
2) Sammenhængen mellem en kategorisk og en numerisk variabel
3) Sammenhængen mellem to kategoriske variable

Vi ser her på det første tilfælde:

Num-Num:
Her kan vi fx vende tilbage til datasættet med superligaklubberne

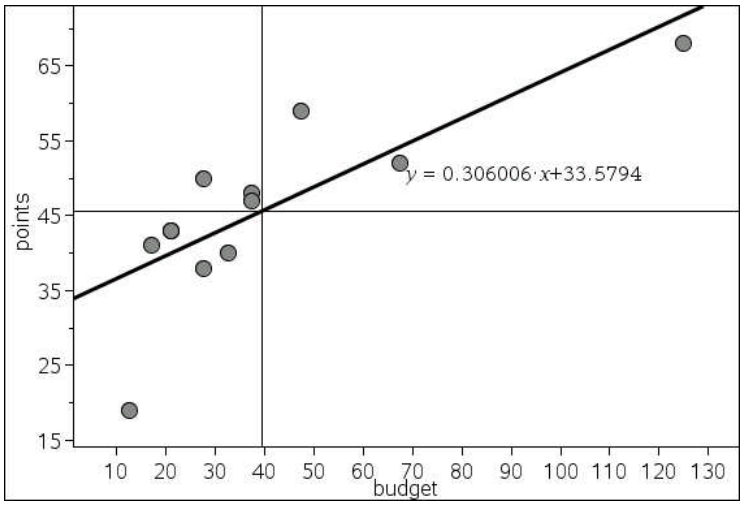
	A klub	B budget	C points
◆			
1	HB Køge	12.5	19
2	SønderjyskE	17	41
3	FC Nordsjælland	21	43
4	Silkeborg	21	43
5	Esbjerg fB	27.5	50
6	AGF	27.5	38
7	Randers FC	32.5	40
8	AaB	37.5	48
9	FC Midtjylland	37.5	47
10	OB	47.5	59
11	Brøndby IF	67.5	52
12	FC København	125	68

Her har vi nu netop to numeriske variable: **budget** og **points**. Det er da ikke svært at forestille sig at der er en sammenhæng mellem spillerbudgettet og så det antal points man får i kampen med de andre. Jo større budget, jo dyrere spillere og dermed forhåbentligt også bedre spillere. Alt andet lige ville man derfor nok forvente at klubberne med de store budgetter klarer sig bedst. Vi afbilder derfor **budget** ud af førsteaksen som den uafhængige variabel og **points** op af andenaksen som den afhængige variabel:

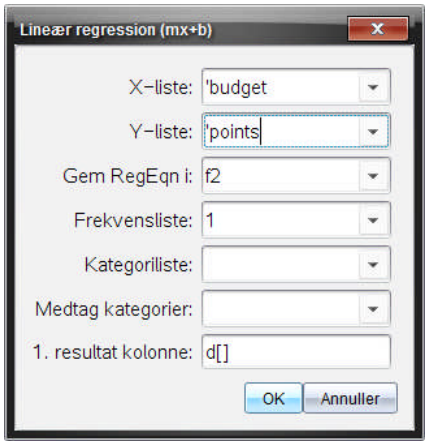


Her kan man jo nok genkende en stigende tendens! Vi har tydeliggjort tendensen ved at tilføje middelværdierne for **budget** (lodret) og **points** (vandret) ved hjælp af menupunkterne **Plot værdi** og **Plot funktion** fra **Undersøg data**-menuen. I forhold til tyngdepunktet (middel_budget, middel_points))ligger 3 af klubberne i 'første kvadrant' (+,+) og de 6 af klubberne i 'tredje kvadrant' (−,−) svarende til en voksende sammenhæng.

For at undersøge den voksende sammenhæng nærmere kan vi nu tilføje en lineær regression fra **Undersøg data**-menuen. Læg mærke til at regressionslinjen netop går gennem tyngdepunktet!



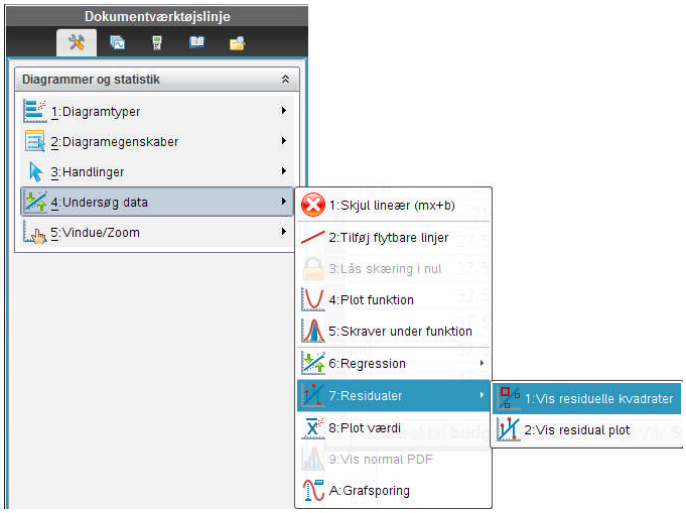
Da hældningen er ca. 0.3, ser vi altså at alt andet lige så vil en ekstra investering på 10 millioner i spillerbudgettet føre til at øget pointsantal med 3 points. Men modellen er jo ikke meget præcis, for datapunkterne ligger ikke så overbevisende tæt på tendenslinjen. Udfører vi som vist den lineære regression i **Lister og Regneark**-værkstedet fås da også kun en *forklaringsgrad* på 62%.

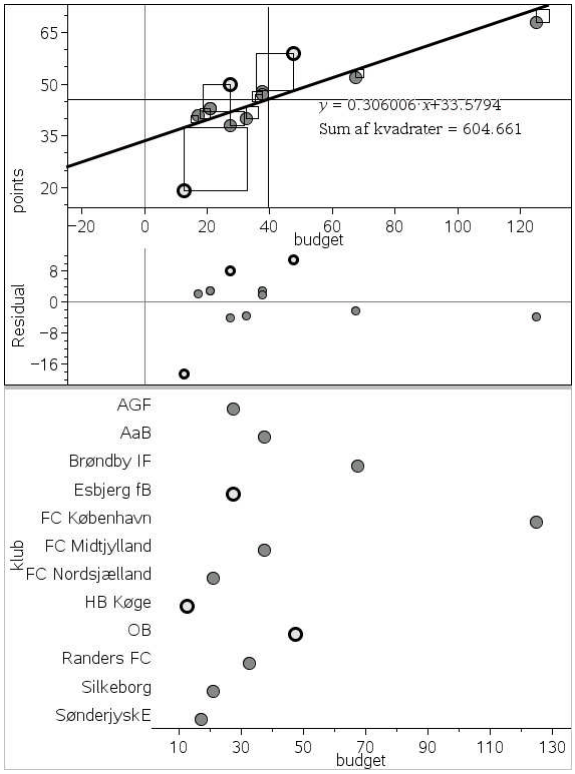


	A klub	B budget	C points	D	E
◆					=LinRegMx('budget,'points,1)
1	HB Køge	12.5	19	Titel	Lineær regression (mx+b)
2	SønderjyskE	17	41	RegEqn	m*x+b
3	FC Nordsjælland	21	43	m	0.306006
4	Silkeborg	21	43	b	33.5794
5	Esbjerg fB	27.5	50	r²	0.617465
6	AGF	27.5	38	r	0.785789
7	Randers FC	32.5	40	Resid	{-18.404506060315,2.21846...
8	AaB	37.5	48		
9	FC Midtjylland	37.5	47		
10	OB	47.5	59		
11	Brøndby IF	67.5	52		
12	FC København	125	68		

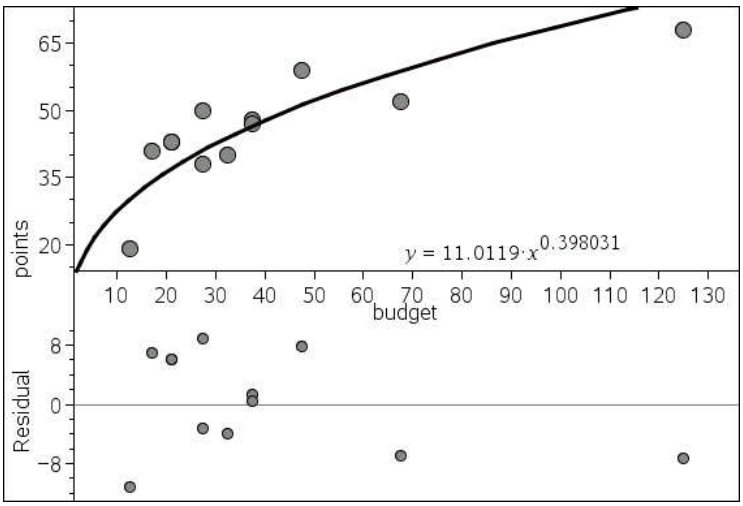
E =LinRegMx('budget,'points,1) : CopyVar Stat.RegEqn,'f2: CopyVar Stat.,

Vi kan også supplere med såvel et **residualplot** som de **residuelle kvadrater** fra **Undersøg data**-menuen. Vi ser da på den følgende figur, at de typiske residualer ligger mellem -5 og +5 og at summen af de residuelle kvadrater er 640.661. Den rette linje er fastlagt ud fra mindste kvadraters metode, dvs. den rette linje er lagt, så summen af kvadraterne er mindst mulig. Vi ser også, at der ikke er noget tydeligt mønster i residualerne, om end det kunne se ud som om der er en tendens til at de starter og slutter lavt. Det tyder på at den lineære model er en rimelig model. De tre klubber der afviger mest: OB og Esbjerg til den positive side og Køge til den negative har derfor blot klaret sig overraskende godt henholdsvis overraskende dårligt i forhold til deres investeringer.





Den lineære model er selvfølgelig ikke den eneste mulige. En anden mulighed er potensmodellen, som forudsætter en form for proportionalitet mellem indsatsen (budgettet) og udbyttet (pointene). Vi skal da blot i stedet vælge en **Potensiel regression** i **Undersøg data**-menuen:



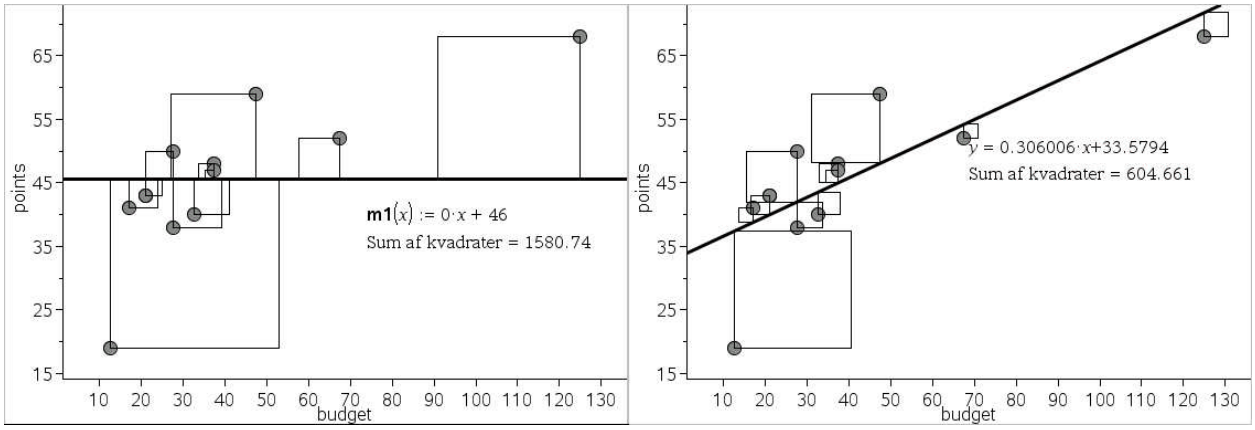
Summen af residualkvadraterne oplyses kun for lineære modeller! Residualplottet minder om det foregående, så også denne model kan med rimelighed bruges. I modsætning til den lineære model flader den ud, dvs. det er nærmest indbygget i modellen at penge spiller en mindre og mindre rolle, jo større budgettet er. Vi er heller ikke her i nærheden af en ligefrem proportionalitet, så selv om FX København har et spillerbudget, der er stort set dobbelt så stort som den nærmeste konkurrent Brøndby IF, så får de bestemt ikke dobbelt så mange points.

I potensmodellen er det nu eksponenten på ca. 0.40, der er den afgørende i beskrivelsen af væksten, som er en procent-procent-vækst. Dvs. hvis vi investerer 10% mere i spillerbudgettet bør vi forvente en stigning på ca. 4% i pointene.

Bemærkning: Det er ikke muligt objektivt at vælge mellem de to modeller, den lineære og potensmodellen i det ovenstående tilfælde. Afhængigt af hvad det er man vil fremhæve kan man derfor bruge den ene eller den anden efter behov. Man kunne nu måske tro at man kunne vælge mellem de to modeller ved at inddrage forklaringsgraderne for de to modeller. Men det giver desværre ingen mening, dels fordi potensregressionen ikke direkte benytter sig af mindste kvadraters metode, dels fordi forklaringsgraden for potensregressionen ikke udregnes direkte på grundlag de originale rå data. Vi vil dog ikke forfølge dette nærmere her.

4.3 Forklaringsgraden for en lineær regression*

Det følgende afsnit er lidt teknisk og kan roligt overspringes i første omgang! Men i projektarbejder vil man ofte komme ud for at det er praktisk selv at kunne beregne en forklaringsgrad og det er derfor vigtigt at kende teknikken. Her forklarer vi den med udgangspunkt i eksemplet med sammenhængen mellem budget og points for superligaklubberne. Udgangspunktet er summen af kvadraterne, dvs. 604.661. Hvis det er et lille tal, vil vi opfatte modellen som god. I modsat fald vil vi opfatte modellen som dårlig. Men et tal som 604.661 er jo ikke i sig selv stort eller lille. Det bliver det kun, når vi sammenligner det med en standard. Vi skal derfor have en passende standard vi kan sammenligne den lineære regression med. Udgangspunktet var vores antagelse om at datasættet alt andet lige kunne beskrives med en lineær sammenhæng. Alternativet er at der slet ikke er nogen sammenhæng, dvs. pointene afhænger i virkeligheden slet ikke af budgetterne. I så fald svinger pointene altså rent tilfældigt op og ned omkring deres middelværdi. Det kan vi beskrive med en vandret linje med ligningen y = mean(**points**). Det er ikke svært at tegne en sådan linje med **Plot funktion** fra **Undersøg data**-menuen, men vi kan ikke få oplyst de residuelle kvadrater umiddelbart. I stedet må vi så tilføje en flytbar linje og regulere på den så den bliver vandret ved at dreje den ud i siderne og flytte den op og ned midtpå. Denne flytbare linje er nemlig tilknyttet en sum af kvadrater og flytter vi den hen til den ligger oven på middelværdien ser vi netop, at vi får den mindst mulige kvadratsum blandt alle de vandrette linjer:



Det fører til den tilnærmede værdi 1580.74 for den vandrette linje y = mean(**points**), hvor datapunkterne antages at svinge rent tilfældigt omkring deres middelværdi. Hvis vores antagelse om den lineære tendens skal være rimelig må det nu føre til en væsentlig bedre model end antagelsen om, at der slet ikke er nogen sammenhæng. Dvs. kvadratsummen 604.661 for den lineære regression må være væsentligt mindre end kvadratsummen 1580.74 for den bedste vandrette linje. Vi kan derfor se på hvor mange procent de 640.661 udgør af 1580.74:

$$\frac{604.661}{1580.74} = 38.25\%$$

I sammenligning med antagelsen om at der slet ikke er nogen sammenhæng er det altså lykkedes os at reducere summen af kvadrater med 61.75%, dvs. det er lykkedes os at forklare 61.75% af variationen ved at bruge den lineære model. Dette er forklaringen på forklaringsgraden r^2 = 61.75% ☺

Hvis vi selv skal udføre beregningen i regnearket skal vi have fat i de to kvadratsummer hørende til den bedste vandrette linje og den bedste skrå linje. De kan dels udregnes med celleformler dels kan de trækkes ud af **statistik for en enkelt variabel**, idet det allersidste punkt $SSX := \sum(x - \bar{x})^2$ netop svarer til kvadratsummen (variationen).

Først celleformlerne, hvor vi udnytter at vi ved, hvor forskriften for den lineære regressionsfunktion er gemt:

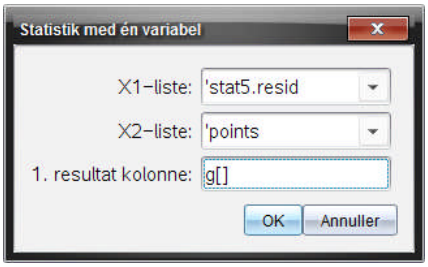
F
Kvadratsum lineær regr. =
604.661
Kvadratsum vandret =
1580.67
Forklaringsgrad =
61.7465

F2	=sum((**points**−*f2*(**budget**))²)
F4	=sum((**points**−mean(**points**))²)·1.
F6	=(*f4*−*f2*)/*f4*·100.

Læg mærke til at denne formelgymnastik kan bruges på en vilkårlig model til beskrivelse af sammenhængen uanset om den kommer fra en regressionsmodel eller ej.

Dernæst statistik med en enkelt variabel, hvor vi udnytter at de statistiske variable for den lineære regression er gemt i stat5:

E	=LinRegMx(**'budget**,**'points**,1) : CopyVar **Stat.RegEqn**,*'f2*: CopyVar **Stat.**, **Stat5.**



G	H	I
	=OneVar('stat5.resid,1):	=OneVar('points,1):
Titel	Statistik med én variab...	Statistik med én v...
$\bar{x}$	1.33333E-13	45.6667
Σx	1.6E-12	548.
Σx²	604.661	26606.
sx := sn-...	7.41412	11.9874
σx := σn...	7.09848	11.477
n	12.	12.
MinX	-18.4045	19.
Q₁X	-3.6774	40.5
MedianX...	2.08191	45.
Q₃X	2.99444	51.
MaxX	10.8853	68.
SSX := Σ...	604.661	1580.67

SSX for residualerne, dvs. 604.661, er da netop summen af kvadraterne hørende til den lineære regression, mens SSX for variablen **points**, dvs. 1580.67, netop er summen af kvadraterne hørende til den vandrette linje gennem middelpointene. Forklaringsgraden fås nu på sædvanlig vis

12	MaxX	10.8853	68.	Forklaringsgrad =
13	SSX := Σ...	604.661	1580.67	61.7465
J13	=(*i13*−*h13*)/*i13*·100.			

Læg mærke til at denne teknik kun kan bruges, hvis vi kender residualerne for modellen, dvs. i praksis kan den kun anvendes på de indbyggede regressionsmodeller.

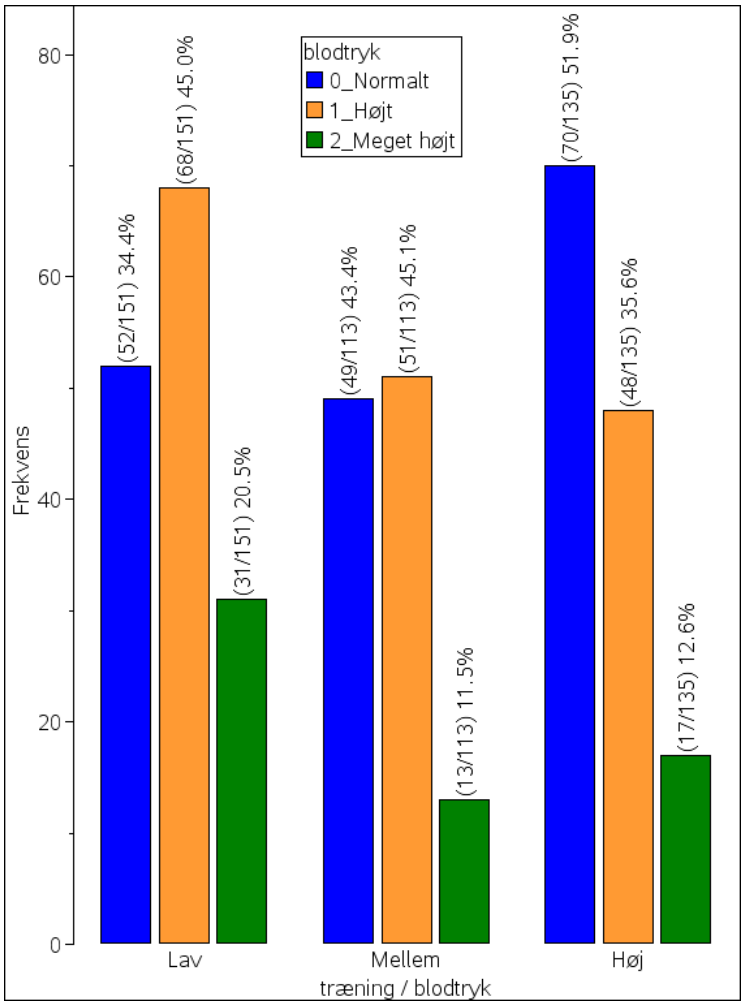
4.4 Sammenhængen mellem ordinale variable: Gamma-graden*

Det følgende afsnit handler om en speciel problemstilling af særlig interesse for biologi og samfundsfag og kan overspringes i første gennemlæsning.

I biologi og samfundsfag vil de enkelte variable ikke altid være numeriske, men de kan stadigvæk rangordnes. I spørgeskemaer kan man fx spørge om respondenten er meget uenig, lidt uenig, lidt enig eller meget enig. I sådanne tilfælde opererer man altså med kategoriske variable, der kan rangordnes entydigt. Sådanne variable kaldes **ordinale variable** (dvs. variable, der kan ordnes i rækkefølge).

Som et eksempel kan vi se på patienter, der undersøges for deres blodtryk. Samtidigt registrerer man forskellige variable, fx deres træningstilstand og deres alkoholforbrug. Resultaterne er indsat i et regneark, hvoraf vi kun viser de første 20:

	A blodtryk	B træning	C alkoholforbrug
◆			
1	1_Højt	Lav	lavt
2	0_Normalt	Høj	lavt
3	0_Normalt	Mellem	Middel
4	0_Normalt	Lav	Middel
5	0_Normalt	Høj	Middel
6	0_Normalt	Mellem	Højt
7	2_Meget højt	Mellem	Højt
8	0_Normalt	Høj	Middel
9	0_Normalt	Mellem	lavt
10	1_Højt	Lav	Højt
11	0_Normalt	Høj	lavt
12	1_Højt	Lav	Højt
13	1_Højt	Lav	Højt
14	1_Højt	Høj	Højt
15	2_Meget højt	Lav	Højt
16	1_Højt	Lav	Middel
17	0_Normalt	Lav	lavt
18	1_Højt	Lav	lavt
19	2_Meget højt	Lav	Højt
20	0_Normalt	Mellem	lavt



De to forklarende variable, **træning** og **alkoholforbrug**, antager tre værdier: Lav, mellem og høj, mens responsvariablen **blodtryk** er kodet lidt mere omhyggeligt, som 0_normalt, 1_højt og 2_meget højt. Vi skal nemlig arbejde med grupperede søjlediagrammer, hvor vi ikke direkte kan styre rækkefølgen af responsvariablen. Vi opretter et **Diagrammer og statistik**-værksted og indsætter først fx den forklarende variabel **træning**. Derefter højreklikker vi og vælger **Opdel kategorier efter variabel** og tilføjer responsvariablen blodtryk, dvs. vi splitter variablen træning ved hjælp af variablen blodtryk. Til sidst omsætter vi diagrammet til et søjlediagram og vælger menupunktet **Vis alle etiketter** i menuen for **Diagramegenskaber**. Vi får da talt alle kombinationer af de to variable op og kan aflæse deres værdier i diagrammet. Vi lægger mærke til tendensen til at blodtrykket falder, når man træner meget: Lav træning har fx den højeste forekomst af Meget højt blodtryk. Men tendensen er ikke helt entydig, og fremstår ikke helt så klart som vi måske kunne ønske os.

Det er her **gamma-graden** kan komme i spil. Vi skal da have omsat data til en krydstabel. Der findes snedige kommandoer til at frembringe en sådan krydstabel, men langt det nemmeste

er at aflæse tallene fra det ovenstående diagram. Læg mærke til at tabellen er opbygget ligesom i et koordinatsystem, dvs. den lodrette akse peger opad!

D	E	F	G
↑ Meget højt	31	13	17
↑ Højt	68	51	48
↑ Normalt	52	49	70
blodtryk/træning	↦ Lav	↦ Mellem	↦ Høj

Ideen er nu at man kigger på hvad der sker med den afhængige variabel, når den uafhængige variabel ændres: Der er tre muligheder, den kan vokse, den kan aftage og den kan forblive konstant. For hvert eneste par af datapunkter registreres hvad der sker. Derefter tæller man det samlede antal par, hvor de følges ad, dvs. når den ene vokser, vokser den anden også. Sådanne par vil vi kalde **voksende par** (concordant pairs). Tilsvarende tæller man det samlede antal par, hvor de reagerer modsat, dvs. når den ene vokser aftager den anden. Sådanne par vil vi kalde **aftagende par** (discordant pairs). Endelig ignorerer alle de par, hvor enten den uafhængige eller den afhængige variabel forbliver konstant. Sådanne par vil vi kalde **konstante par** (ties). I praksis foregår det således: For en given celle, deles skemaet op af den vandrette række gennem cellen og den lodrette søjle gennem cellen:

↑ Meget højt	31	13	17
↑ Højt	68	51	48
↑ Normalt	52	49	70
blodtryk/træning	↦ Lav	↦ Mellem	↦ Høj

Her er det midtercellen 51 vi tager udgangspunkt i. Den vandrette række og den lodrette søjle ignoreres, da de genererer konstante par. Resten af cellerne fordeles på fire kvadranter i forhold til den valgte celle: Første og tredje kvadrant, dvs. 17 og 52 svarer til voksende par (hældningen er positiv). Anden og fjerde kvadrant, dvs. 31 og 70 svarer til aftagende par (hældningen er negativ). Midtercellen involverer altså 51·17+51·52 = 51·69 voksende par og tilsvarende 51·31+51·70 = 51·101 aftagende par (hvorfor der er flere aftagende end voksende par). Men for at få alle mulige kombinationer talt op systematisk begynder vi i nederste venstre hjørne med de voksende par og øverste venstre hjørne med de aftagende par:

Voksende par: Der er i alt 9 kombinationer med voksende par: 12800

31	13	17
68	51	48
52	49	70

52·(13+17+51+48)

31	13	17
68	51	48
52	49	70

68·(13+17)

31	13	17
68	51	48
52	49	70

49·(17+48)

31	13	17
68	51	48
52	49	70

51·17

Aftagende par: Der er i alt 9 kombinationer med aftagende par: 19954

31	13	17
68	51	48
52	49	70

52·(13+17+51+48)

31	13	17
68	51	48
52	49	70

68·(13+17)

31	13	17
68	51	48
52	49	70

49·(17+48)

31	13	17
68	51	48
52	49	70

51·17

Der er altså 19954 aftagende par mod kun 12800 voksende par, hvilket bekræfter den aftagende tendens: Jo mere træning man foretager, jo mindre vil blodtrykket alt andet lige være. Gammagraden defineres så som forskellen mellem antallet af voksende og aftagende par, divideret med summen af voksende og aftagende par, dvs. brøken:

$$\gamma = \frac{V - A}{V + A}$$

Gammagraden vil altid ligge mellem -1 og 1 og den vil være negativ for aftagende sammenhænge, positiv for voksende sammenhænge. Den vil kun være -1, hvis der er tale om en rent aftagende sammenhæng, dvs. der er slet ingen voksende par: V = 0. Den vil kun være +1, hvis der er tale om en rent voksende sammenhæng, dvs. der er slet ingen aftagende par: A = 0. Endelig vil den være 0, når der er lige mange voksende som aftagende par, dvs. sammenhængen er hverken voksende eller aftagende.

Men kan nemt udregne gamma-graden i regnearket:

	D	E	F	G	H
◆					
1	↑ Meget højt	31	13	17	
2	↑ Højt	68	51	48	
3	↑ Normalt	52	49	70	
4	blodtryk/træning	↦ Lav	↦ Mellem	↦ Høj	
5	Voksende par =	12800			
6	Aftagende par =	19954			
7	Gamma =	-0.218416			

E5 =e3·sum(f1:g2)+e2·sum(f1:g1)+f3·sum(g1:g2)+f2·g1

Gammagraden er altså i dette tilfælde -0.218 eller -21.8%. Gammagraden kan fortolkes lidt på samme måde som forklaringsgraden: Vi ignorerer de konstante par og ser på et vilkårligt par af observationer, hvor x-rangen vokser. Vi kan da forudsige at y-rangen vil aftage. Men forudsigelsen er ikke perfekt, men behæftet med en fejl, da der også vil forekomme tilfælde, hvor y-rangen vokser, selv om gamma-graden som her er negativ. Men det at vi kender gamma-graden tillader os at sige at usikkerheden på y-rangen er blevet reduceret med 21.8%. De voksende par udgør kun 39.1 % af alle ikke-konstante par, mens de aftagende par udgør 60.9% af alle ikke-konstante par. Det er altså mere sandsynligt at trække et tilfældigt aftagende par end et tilfældigt voksende par. Men forskellen er trods alt kun 21.8%, så det er ikke nogen stærk aftagende sammenhæng.

Styrken ved gamma-graden er at den er nem at beregne og nem at forstå. Svagheden er især at den ikke tage hensyn til konstante par. Hvis der er rigtigt mange konstante par i forhold til de voksende og aftagende par giver den derfor ikke megen mening. Den er også symmetrisk, dvs. den fortæller ikke noget om hvilken vej en eventuel kausal sammenhæng mellem de to variable peger: Vi får præcis den samme gamma-grad, selv om vi vender tabellen.

Øvelse: Prøv nu selv at undersøge sammenhængen mellem blodtryk og alkoholforbrug. Du skulle da gerne finde en gamma-grad på 0.205190.

5 Grupperede observationer

5.1 Gruppering af data: Hyppigheder/frekvenser

Når vi indsamler data, så er de indsamlede data, de **rå data**, den kilde som vi senere kan udvinde alle informationerne fra. De rå data udgør statistikkens hellige gral (på samme måde som de eksperimentelle data i naturvidenskaberne) og vi bør altid så vidt muligt arbejde direkte med de rå data, fordi de kan vendes og drejes og dermed ses fra alle synsvinkler, hvorved vi i et rigt datasæt kan blive ved med at gå på opdagelse og opdage nye sammenhænge.

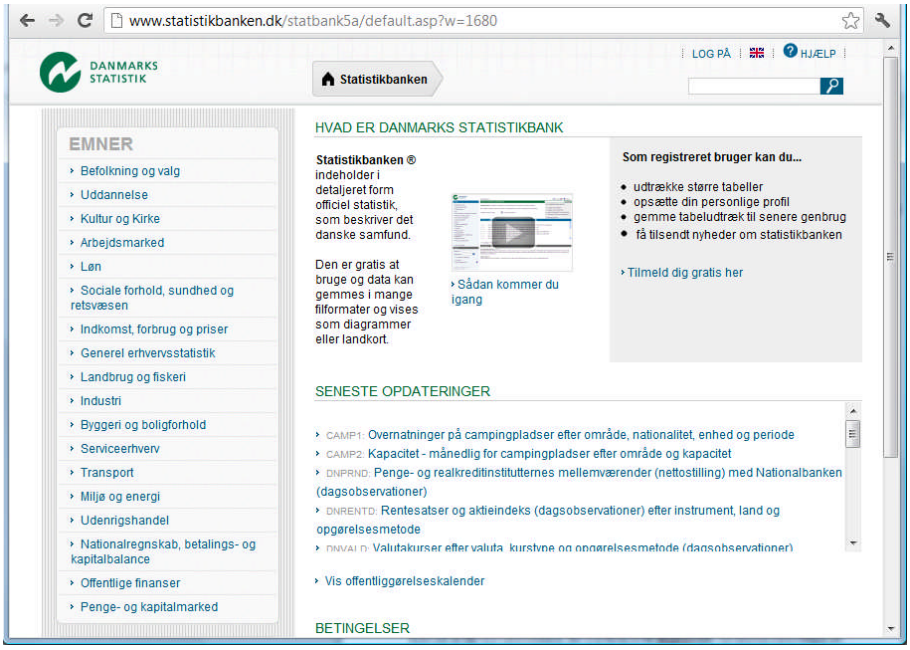
Men når man præsenterer sine konklusioner vil man ofte forenkle situationen og kun vise de **forarbejdede data**, der umiddelbart understøtter ens konklusioner. Typisk vil man **gruppere** data, dvs. slå dem samme i et mindre antal grupper, hvor man så ikke længere skelner mellem de individuelle data. Det kan være i form af et boksplot, hvor datasættet deles i fire lige store grupper efter størrelse, eller det kan være i form af et histogram, hvor man har valgt en passende intervalinddeling for at fremhæve nogle typiske træk ved fordelingen. Når man på denne måde grupperer sine data mister man altså information: De forarbejdede data repræsenterer halvfabrikata. Hvis vi kun har adgang til de grupperede data kan vi derfor ikke længere drage præcise konklusioner, men må nøjes med tilforladelige skøn. Vi kan sammenligne det med tilberedning af mad: Hvis vi har alle råvarerne til rådighed kan vi lave alle mulige varianter af retter, men hvis råvarerne først er hældt sammen i en stor gryde og kogt sammen til en grød er der ikke så meget mere at stille op, selvom der stadigvæk kan tilføjes forskellige krydderier.

Alligevel er det vigtigt at have kendskab til de vigtigste metoder til at trække informationer ud af grupperede data, da det ofte er på den form vi vil møde data i andres undersøgelser, når de fremlægges i fx avisartikler eller hentes på nettet, og kun ved at kende til sådanne teknikker vil vi kunne forholde os kritisk til de påstande, der er knyttet til undersøgelsen.

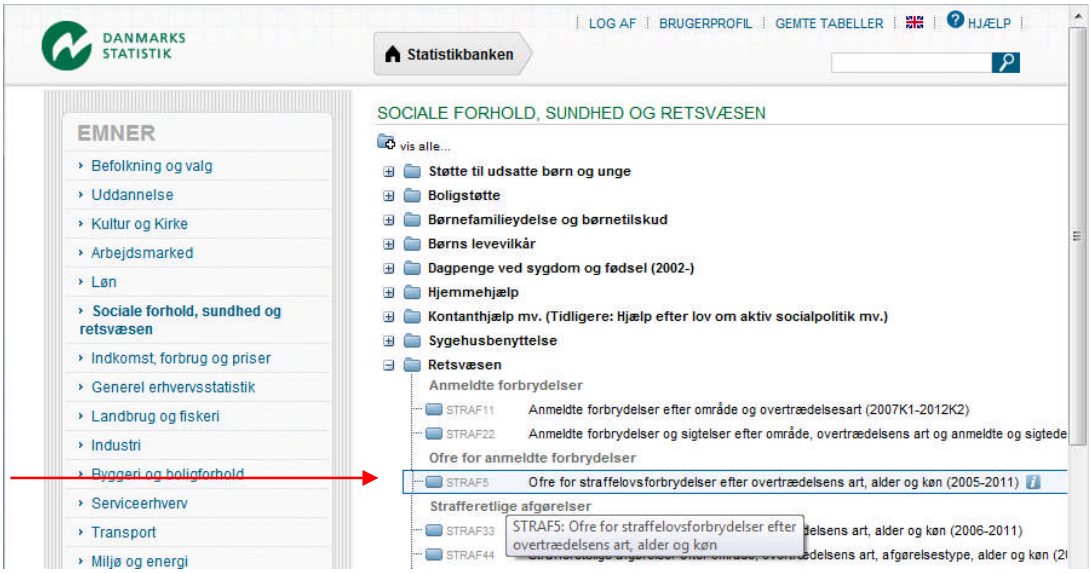
I det følgende afsnit ser vi derfor på i hvor høj grad det stadigvæk i forbindelse med grupperede data er muligt at skønne *troværdigt* over størrelsen af medianer, kvartiler, middelværdier osv., ligesom vi ser nærmere på hvilke graftyper vi kan tilnærme med grupperede data.
Centralt for de grupperede data står **hyppighedstabellerne**, hvor observationerne er inddelt i passende intervaller, hvorefter man har talt op hvor mange observationer der falder i de enkelte observationsintervaller. Det er på basis af disse hyppighedstabeller vi skal forsøge at uddrage troværdige informationer. Da vi ikke kender de individuelle data i de enkelte intervaller gør vi nu følgende antagelse:

> De individuelle data i et givet observationsinterval antages at være tilfældigt fordelt i intervallet og antages derfor at ligge jævnt fordelt indenfor intervallets grænser.

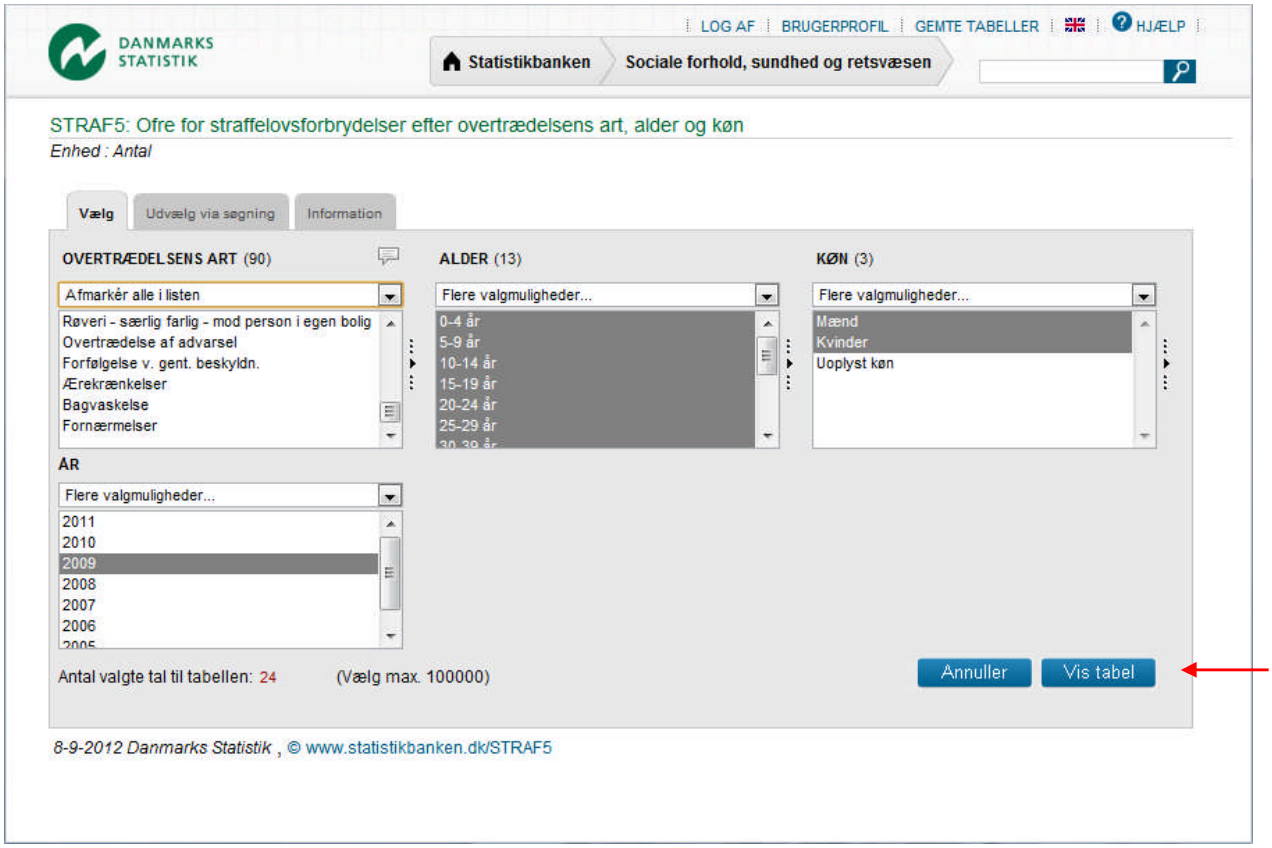
Det behøver ikke være tilfældet i virkeligheden, men hvis ellers data er udvalgt rimeligt tilfældigt er det ret usandsynligt, hvis de fx systematisk alle ligger i den venstre halvdel af intervallet.



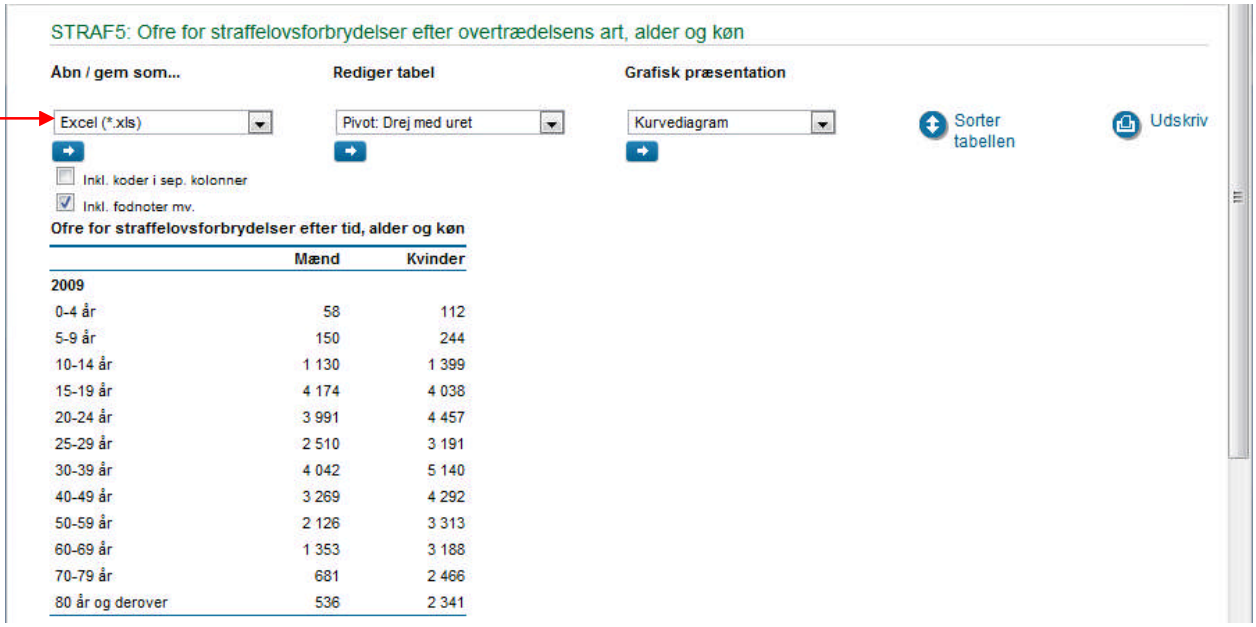
Lad os nu hente nogle data på nettet. Det kan fx være data fra Danmarks statistik, der har samlet deres data på webstedet www.statistikbanken.dk . Vi kan fx hente data fra retsvæsenet:



Her kan vi så interessere os for ofrene for de anmeldte forbrydelser



Som man kan se er der 90 forskellige slags forbrydelser, der bliver registreret startende med sædelighedsforbrydelser. Vi slår alle forbrydelserne sammen ved at undlade at markere i **overtrædelsens art** og vælger tilsvarende at opdele data i alle aldersgrupper (men ikke **Uoplyst alder**), begge køn (men ikke **Uoplyst køn**) samt årstallet 2009. Derefter taster vi **Vis tabel**. Data hentes nu fra Danmarks statistik og vises i en sammentalt tabel. Denne tabel kan derefter eksporteres til Excel, hvorfra vi henter den ind i **TI-Nspire CAS**. Dette er langt den nemmeste platform hvorfra man kan hente data over i **TI-Nspire CAS**.



STRAF5: Ofre for straffelovsforbrydelser efter overtrædelsens art, alder og køn

Abn / gem som... | Rediger tabel | Grafisk præsentation

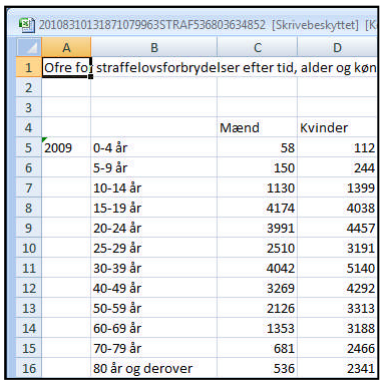
Excel (*.xls) | Pivot: Drej med uret | Kurvediagram | Sorter tabellen | Udskriv

Inkl. koder i sep. kolonner

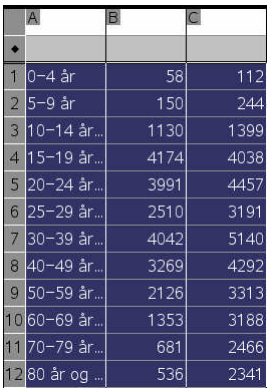
Inkl. fodnoter mv.

Ofre for straffelovsforbrydelser efter tid, alder og køn

	Mænd	Kvinder
2009		
0-4 år	58	112
5-9 år	150	244
10-14 år	1 130	1 399
15-19 år	4 174	4 038
20-24 år	3 991	4 457
25-29 år	2 510	3 191
30-39 år	4 042	5 140
40-49 år	3 269	4 292
50-59 år	2 126	3 313
60-69 år	1 353	3 188
70-79 år	681	2 466
80 år og derover	536	2 341



	A	B	C	D
1	Ofre fo	straffelovsforbrydelser efter tid, alder og køn		
2				
3				
4			Mænd	Kvinder
5	2009	0-4 år	58	112
6		5-9 år	150	244
7		10-14 år	1130	1399
8		15-19 år	4174	4038
9		20-24 år	3991	4457
10		25-29 år	2510	3191
11		30-39 år	4042	5140
12		40-49 år	3269	4292
13		50-59 år	2126	3313
14		60-69 år	1353	3188
15		70-79 år	681	2466
16		80 år og derover	536	2341



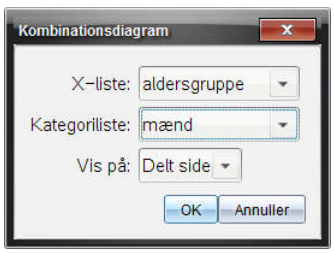
	A	B	C
1	0–4 år	58	112
2	5–9 år	150	244
3	10–14 år...	1130	1399
4	15–19 år...	4174	4038
5	20–24 år...	3991	4457
6	25–29 år...	2510	3191
7	30–39 år...	4042	5140
8	40–49 år...	3269	4292
9	50–59 år...	2126	3313
10	60–69 år...	1353	3188
11	70–79 år...	681	2466
12	80 år og ...	536	2341

Som vist er det kun tabeldataene, der overføres til **TI-Nspire CAS**. Variabelnavnene tilføjes efterfølgende. Men nu har vi fået en hyppighedstabel i kassen og kan begynde at se på data!

5.2 Kombinationsdiagram: Søjlediagrammer og cirkeldiagrammer

Første pointe er nu at da der er tale om en hyppighedstabel skal vi bruge et kombinationsdiagram til at afbilde data. De almindelige grafer virker kun på rå data, men her skal vi jo kombinere en *kategoriske variabel*, **aldersgruppe**, med en *numerisk variabel*, hyppighederne for **mænd** hhv. **kvinder**. Da vi i dette tilfælde har et meget stort antal ofre for kriminalitet, der langt overstiger det maksimale antal rækker på 2500 er der ingen mulighed for at rekonstruere de rå data som lister. Anden pointe er at da udgangspunktet er en kategorisk variabel, kan vi i første omgang kun frembringe et søjlediagram eller et cirkeldiagram (prikdiagrammer er ikke til rådighed for kombinationsdiagrammer, da antallet af datapunkter kan være vilkårligt stort). Vi kan enten markere de to første kolonner (ved at klikke i etiketterne $\boxed{A}$ og $\boxed{B}$) og derefter højreklikke i kolonnerne eller vi kan vælge menupunktet **Data>Kombinationsdiagram**.

	A aldersgruppe	B mænd	C kvinder
1	0–4 år	58	112
2	5–9 år	150	244
3	10–14 år	1130	1399
4	15–19 år	4174	4038
5	20–24 år	3991	4457
6	25–29 år	2510	3191
7	30–39 år	4042	5140
8	40–49 år	3269	4292
9	50–59 år	2126	3313
10	60–69 år	1353	3188
11	70–79 år	681	2466
12	80 år og derover	536	2341



Kombinationsdiagram

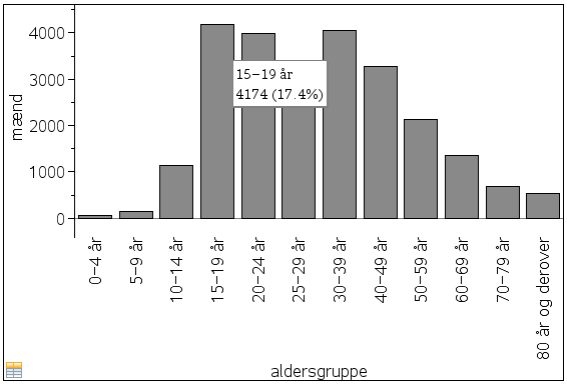
X-liste: aldersgruppe

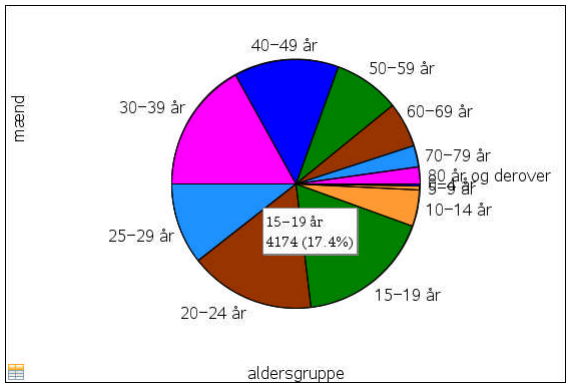
Kategoriliste: mænd

Vis på: Delt side

OK | Annuller

I begge tilfælde åbnes der for en dialogboks, der giver mulighed for at vælge den kategoriske variabel, her **aldersgruppe**, der skal afsættes ud af førsteaksen, samt den tilhørende værdiliste med antallene (hyppighederne, frekvenserne), her **mænd**, samt om kombinationsdiagrammet skal afsættes på samme side eller på en ny side.

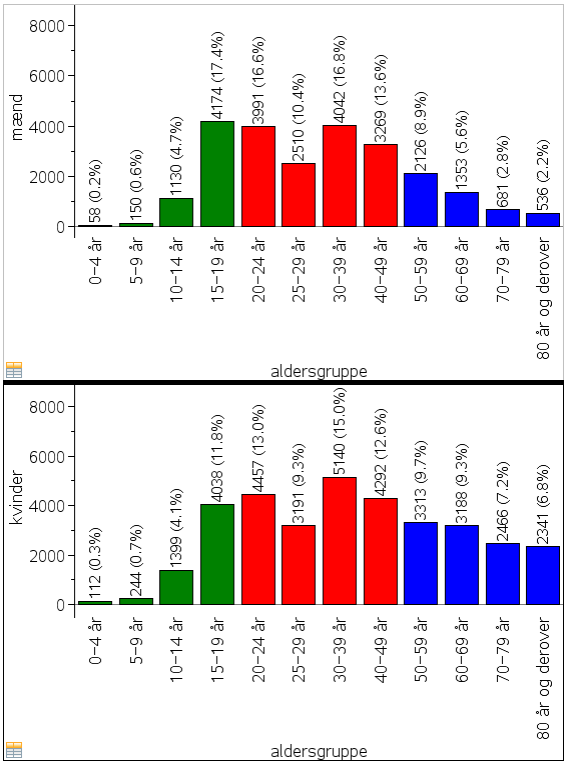




Som det ses kan vi få oplyst såvel antal (hyppigheder) som frekvenser (procenter) for de enkelte kategorier, der afsættes i samme rækkefølge som i tabellen (læst oppe fra og ned). I cirkeldiagrammet følger vi urets omløbsretning med start fra vandret!

Vi kan selvfølgelig slå **Vis alle etiketter til**. Ligeledes kan vi ændre farvekodningen, så vi fx skelner mellem unge (grøn), voksne (rød) og gamle (blå), ligesom det er nemt at sammenligne fordelingerne for de to køn. Farverne ændres ved at klikke i de kolonner man ønsker at markere, hvorefter man kan vælge farven fra menubjælkens farvepalet.

Du kan også oprette et kombinationsdiagram direkte fra **Diagrammer og Statistik**-værkstedet. Det kræver blot at du højreklikker på akseetiketten (Klik for at tilføje en variabel), så du får adgang til at tilføje en X-variabel med værdiliste. Men meget mere er der heller ikke at gøre godt med for søjlediagrammerne, der fungerer som en erstatning for histogrammer. I dette tilfælde er aldersgrupperne jo egentlig knyttet til en numerisk variabel, nemlig ofrets alder, så det ville være rart om vi også kunne konstruere et rigtigt histogram og tilsvarende frembringe rimelige skøn for de sædvanlige statistiske estimatorer: middelværdien, medianen osv.



5.3 Kombinationsdiagrammer: Histogrammer for grupperede data

Det første problem er da at intervallerne ikke har samme bredde. I det ovenstående datasæt er der tre intervaller med tre forskellige længder: 5-årsintervaller (op til 30 år), 10-årsintervaller op til 80 år og et uspecificeret interval fra 80 år og derover! Det er noget rod, som vi først må gøre op med.

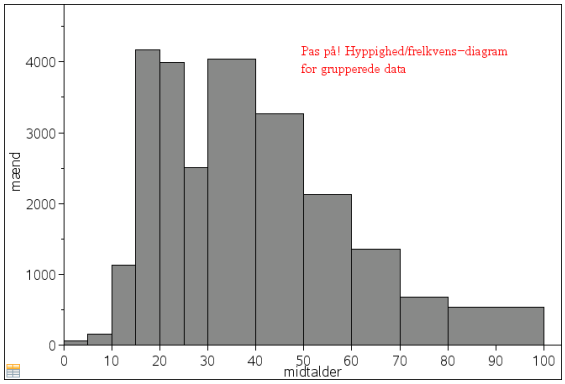
Hvert af intervaller lader vi nu repræsentere af intervallets midtpunkt, dvs. en variabel **midt_alder** der løber fra 5 til 90, da antagelsen om at dataene ligger jævnt fordelt i de enkelte intervaller, betyder at vi kan erstatte hver enkelt alder med intervalmidtpunktet uden at ændre det samlede gennemsnit! Tilsvarende indfører vi intervalendpunkterne i listen **interval**, der derfor kommer til at rumme en celle mere end de øvrige ☺. Derefter er det nemmest at indføre tre nye intervalvariable **grundlinje** (der er fælles) og **højde_mænd** henholdsvis **højde_kvinder**. Grundlinjerne indføres manuelt, men højderne af søjlerne skal beregnes individuelt efter formlerne

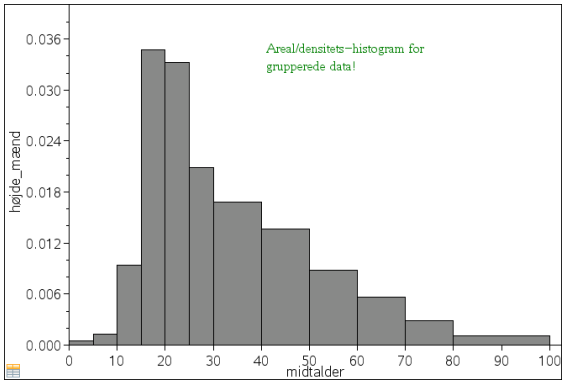
$$\textbf{højde\_mænd} := \frac{\textbf{mænd} / \text{sum}(\textbf{mænd})}{\textbf{grundlinje}} \qquad \textbf{højde\_kvinder} := \frac{\textbf{kvinder} / \text{sum}(\textbf{kvinder})}{\textbf{grundlinje}}$$

	A kategorier	B mænd	C kvinder	D midt...	E inter...	F grund...	G højde_m...	H højde_k...
◆							=mænd/(sum	=kvinder/(s
1	0–4 år	58	112	2.5	0	5	0.000483	0.000655
2	5–9 år	150	244	7.5	5	5	0.001249	0.001428
3	10–14 år...	1130	1399	12.5	10	5	0.009409	0.008186
4	15–19 år...	4174	4038	17.5	15	5	0.034754	0.023627
5	20–24 år...	3991	4457	22.5	20	5	0.033231	0.026079
6	25–29 år...	2510	3191	27.5	25	5	0.020899	0.018671
7	30–39 år...	4042	5140	35	30	10	0.016828	0.015038
8	40–49 år...	3269	4292	45	40	10	0.013609	0.012557
9	50–59 år...	2126	3313	55	50	10	0.008851	0.009693
10	60–69 år...	1353	3188	65	60	10	0.005633	0.009327
11	70–79 år...	681	2466	75	70	10	0.002835	0.007215
12	80 år og ...	536	2341	90	80	20	0.001116	0.003424
13					100			
14								

H13

Histogrammet tager nu udgangspunkt i disse reviderede data. Først oprettes et kombinationsdiagram over antallet af mænd som funktion af midtalderen, dvs. variablen **midtalder** på førsteaksen kombineres med værdilisten **mænd** på andenaksen. Vi kan nu skifte til et histogram som et kombinationsdiagram for mændenes hyppigheder. Vi skal da efterfølgende gå ind i søjleindstillinger (højreklik) og skifte til ulige store intervaller baseret på listen over intervalendepunkter:

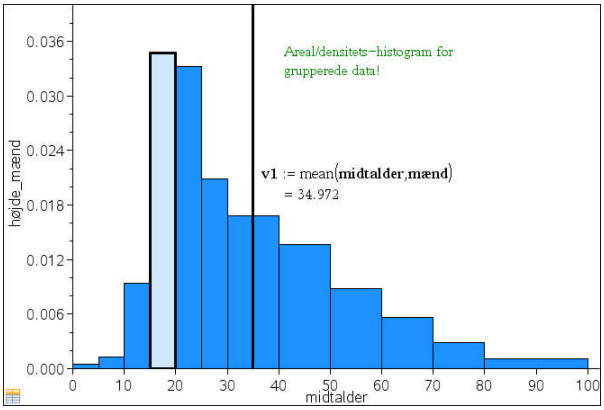


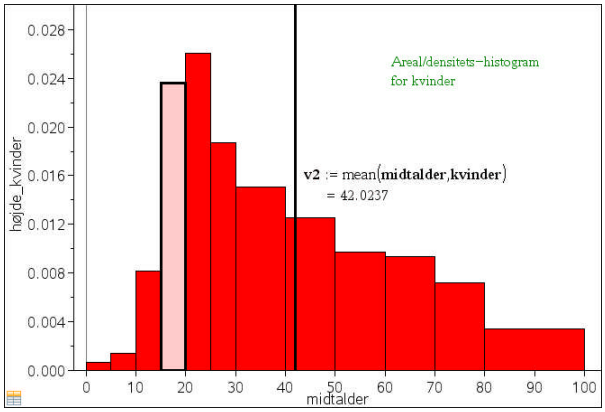


Men i dette histogram er *højden* af den enkelte søjle netop proportional med antallet (et frekvenshistogram). Vi vil normalt foretrække et areal/densitetshistogram, hvor det er *arealet* af søjlen, der er proportional med antallet. Hvis vi havde arbejdet med rå data kunne vi nu bare skifte til densitet på andenaksen ved at højreklikke og vælge **Histogramskala**. Men denne mulighed er ikke til stede for grupperede data! I stedet skifter vi manuelt til variablen **højde_mænd** på andenaksen og får endelig et traditionelt histogram med samlet areal 1.

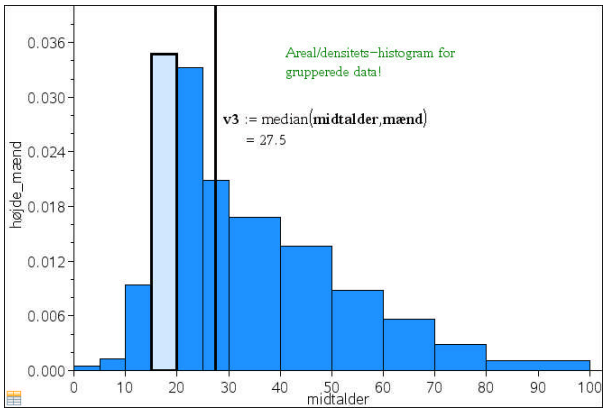
Vi får derved et meget klarere indtryk af fordelingen. Specielt giver det nu mening at snakke om **typeintervallet**, dvs. det interval, der har det største antal observationer, dvs. i dette tilfælde aldersgruppen fra 15-20. Med udgangspunkt i de reviderede data kan vi nu også skønne over middelalderen for ofrene for kriminalitet. Vi kan fx vælge **Plot værdi** i **Undersøg data**-menuen og indskrive kommandoen **mean**(**midt_alder**, **mænd**), hvor det selvfølgelig er afgørende at vi oplyser både alderen og hyppigheden. Det samme kan vi gøre for kvindernes hyppighedsfordeling og dermed sammenligne mænds og kvinders fordeling. Vi ser da bl.a. at typeintervallet for de to fordelinger er det samme, men at de kvindelige ofres middelalder ligger over de

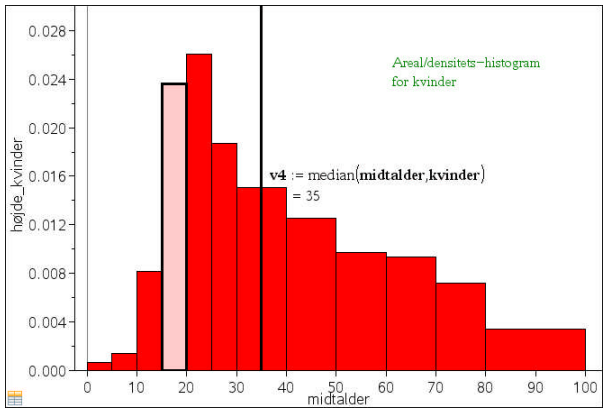
mandlige ofres middelalder, i overensstemmelse med at kvindernes fordeling er længere om at falde ned.





Når vi nu har fundet middelværdien kunne man tænke sig at vi kunne finde medianen på samme måde. Men det giver kun et groft skøn over medianen, fordi alle datapunkterne i et givet interval er samlet i midtpunktet. Hvis vi forsøger os med en beregning af medianen med udgangspunkt i histogrammet giver det derfor kun en oplysning om hvilket interval, der indeholder den midterste observation! Vi finder da fx de følgende skøn over medianerne for mænd henholdsvis kvinder





I praksis anvender man derfor en anden teknik til at skønne over medianer og kvartiler!

5.4 Sumkurver for grupperede data

Sumkurven bygger altid på en tabel over de kumulerede (summerede) frekvenser, så hvis udgangspunktet er en tabel over antal/hyppigheder, må denne først omdannes til en tabel over frekvenser/procenter. Til hvert intervalendepunkt, knytter vi den procentdel af observationerne, der går *forud* for endepunktet. Det er denne procentdel, der er den **kumulerede frekvens** (kumulere = opsamle). Ofte betegnes den dog også den **summerede frekvens**, fordi den fremkommer ved at lægge de foregående frekvenser sammen.

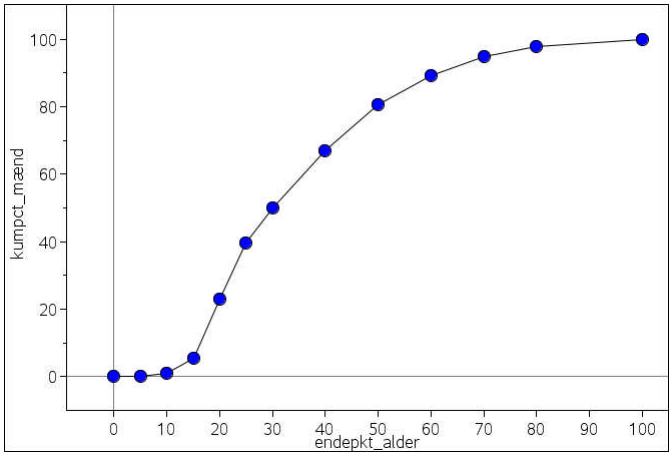
Vi starter derfor med at opbygge en tabel ud fra *samtlige* intervalendepunkter: 0, 5, 10, 15, …, 70, 80, 100. Det giver et ekstra intervalendepunkt til at begynde med, der tildeles antallet 0, idet der *ikke* ligger nogen observationer forud for det allerførste intervalendepunkt.

Derefter lægges antallene løbende sammen og der divideres med det samlede antal og ganges med hundrede for at få

	A endepkt_alder	B antal_mænd	C kumpct_mænd	D
◆			=cumulativesum(	
1	0	0	0.	
2	5	58	0.241465	
3	10	150	0.865945	
4	15	1130	5.57036	
5	20	4174	22.9475	
6	25	3991	39.5629	
7	30	2510	50.0125	
8	40	4042	66.8401	
9	50	3269	80.4496	
10	60	2126	89.3006	
11	70	1353	94.9334	
12	80	681	97.7685	
13	100	536	100.	

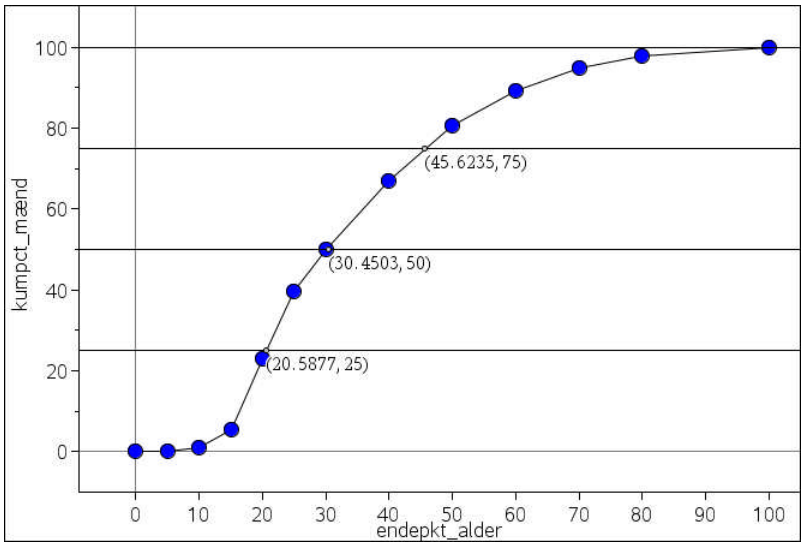
$$C \quad \text{kumpct\_mænd} := \frac{\text{cumulativesum}(\text{antal\_mænd})}{\text{sum}(\text{antal\_mænd})} \cdot 100.$$

resultatet ud i procent. Læg mærke til at den kumulerede frekvens altid starter med 0% og slutter med 100%. Med undtagelse af muligheden for en mindre afrundingsfejl (som højst må være 0.1-0.2%) er det afgørende, for det viser netop, at vi har fået *alle* observationerne talt med.



Vi kan nu tegne et **XY-plot** med **endepkt_alder** som uafhængig variabel og **Kumpct_mænd** som afhængig variabel. Derefter højreklikker vi og vælger **Forbind datapunkter**. Det giver os den ønskede sumkurve.

På sumkurven kan vi bl.a. aflæse det stejleste linjestykke hørende til typeintervallet, der i dette tilfælde viser sig at være aldersgruppen fra 15-20 år. Med udgangspunkt i sumkurven kan vi nu også aflæse kvartilerne og medianen. Hertil tilføjer vi vandrette linjer svarende til y = 25, y = 50 og y=75. Det sker ved at vælge **Plot funktion** i **Undersøg data**-menuen. Derefter vælger vi **Grafsporing** (højreklik på graf) og flytter grafpunktet så tæt på skæringspunkterne med sumkurven som muligt. Derefter tastes **ENTER** for at fastfryse koordinaterne. Af hensyn til aflæsningsnøjagtigheden kan det betale sig at sprede grafen ud så meget som muligt på skærmen (benyt evt. **CTRL 6** for at sprede ruderne på siden.



Nøjagtigheden ligger omkring heltalsværdien, men da der er tale om skøn er der ingen grund til at forsøge at beregne skæringspunkterne med større nøjagtighed. I dette tilfælde finder vi derfor

Mændenes nøgletal: Min = 0, Q1 = 21, Med = 30, Q3 = 46, Max = 100

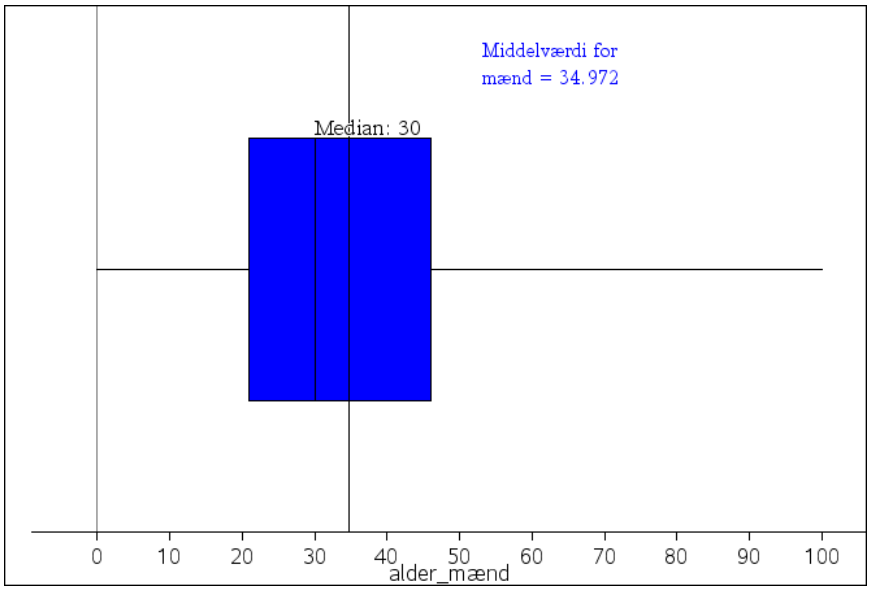
(Her er den maksimale alder sat til 100 på grund af tilskæringen af datasættet!)
Det kan naturligvis også gøres for kvinderne, hvor vi finder

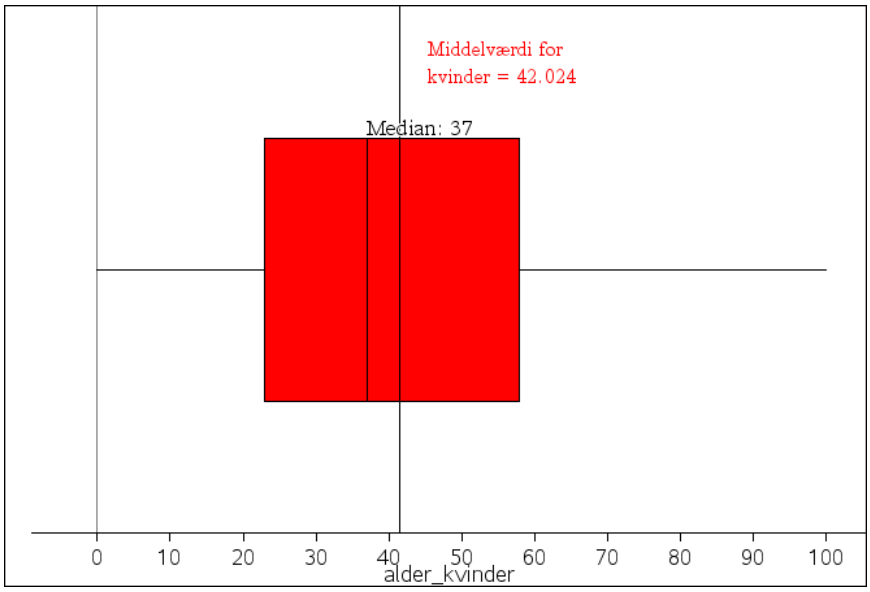
Kvindernes nøgletal: Min = 0, Q1 = 23, Med = 37, Q3 = 58, max = 100

5.5 Boksplot for grupperede data

Med udgangspunkt i de statistiske nøgletal kan vi nu tegne et boksplot. Vi indskriver da de statistiske nøgletal, **idet vi gentager medianen**! Derved fås et stærkt reduceret datasæt, der har de *samme* statistiske nøgletal, netop fordi vi har gentaget medianen! På grund af det stærkt reducerede datasæt kan det være nødvendigt at vælge **Udvid boksplotgrænser** for at forbinde min/max-punkterne med den centrale boks. Det ses da at kvindernes fordeling er bredere og forskudt mod højre i forhold til mændenes fordeling. Af boksplottene ses også at højre hale er længere end venstre hale og at højre del af kvartilboksen er bredere end venstre del. De to fordelinger er altså højreskæve. Det bekræftes af at middelværdierne ligger til højre for medianerne. I begge tilfælde ligger middelværdien ca. 5 år over medianen.

	A deskriptor	B alder_mænd	C alder_kvinder
◆			
1	min	0	0
2	Q1	21	23
3	Med	30	37
4	Med	30	37
5	Q3	46	58
6	max	100	100





Bekræftende statistik

6. Introduktion til bekræftende statistik

Den **beskrivende statistik** (*exploratory data analysis*) tager udgangspunkt i et konkret datamateriale og forsøger at afdække de strukturer, der findes i det konkrete materiale. Hvis det fx drejer sig om højderne i en klasse, udtaler den beskrivende statistik sig om forholdene i netop den klasse: Hvad er middelhøjden, den mindste højde, den største højde osv. Vi kan også inddrage flere variable, fx køn, og sammenligne højderne mellem de to køn osv.

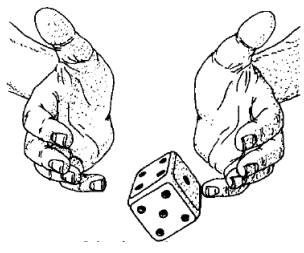
Ofte vil man nu være interesseret i at **generalisere** de fundne strukturer, så man kan drage konklusioner, der rækker *ud over* det pågældende datasæt. Fx kan man ønske at generalisere de fundne kønsforskelle i højderne til ikke bare at gælde den pågældende klasse, men alle unge i tilsvarende klasser. Man kunne også være interesseret i at skønne over middelhøjden for alle unge på basis af den fundne middelhøjde for klassen osv. Her til benyttes metoder fra den **bekræftende statistik** (*confirmatory data analysis*).

> Hovedformålet med bekræftende statistik er at kunne skelne mellem **systematiske variationer** og **tilfældige variationer** i et observeret datasæt.

I de ovenstående eksempler ønsker man altså at drage konklusioner om opførslen af en hel population på basis af opførslen af en passende repræsentativ stikprøve, fx en stikprøve, der er udvalgt rent tilfældigt fra populationen. Vi forventer da at stikprøvens opførsel afspejler hele populationens opførsel, ikke nøjagtigt, men med en rimelig tilnærmelse. Populationen kan fx bestå af alle vælgerne i Danmark og stikprøven af et tilfældigt udsnit af vælgere bestående af 1600 vælgere. Hvis fx en tredjedel af alle vælgere stemmer venstre, så forventer vi også at ca. en tredjedel af de adspurgte vælgere i stikprøven vil lægge deres stemme hos Venstre. Vi regner det fx for fuldstændigt usandsynligt at alle vælgerne i stikprøven er venstrestemmer, selv om det på ingen måde er umuligt at finde et udsnit på 1600 venstrestemmer blandt hele befolkningen.

Vi forventer altid en vis tilfældig variation i en repræsentativ stikprøve, men der er trods alt grænser for hvor stor denne tilfældige variation kan blive, hvis vi fortsat skal tro på at stikprøven er repræsentativ. Vi deler derfor alle de mulige stikprøver fra populationen i to kategorier: De typiske/repræsentative stikprøver, der opfører sig som forventet af en tilfældig stikprøve, henholdsvis de atypiske /skæve stikprøver, hvis opførsel ligger overraskende langt fra den forventede opførsel af en tilfældig stikprøve. Det er denne opdeling af stikprøver, der er hjørnestenen i den statistiske metode, idet den kan bruges til at afgøre hvilke populationer, der er i overensstemmelse med den udtrukne stikprøve.

6.1 Retssagsmetaforen for hypotesetest



Vi vil nu illustrere metoden med et typisk eksempel. Vi ser på et stokastisk fænomen, fx kast med en terning i et prøvespil, før det går løs for alvor med et spil om penge baseret på antallet af seksere i 24 kast. Vi observerer udfaldene for en bestemt serie bestående af 24 kast med terningen, hvor der viser sig at være 10 seksere. Skal vi være bekymrede over udfaldet? Vi ville forvente de seks sider på terningen var lige sandsynlige, så hvis gambleren kastede en **ideel terning** 24 gange ville vi forvente 4 seksere.

Men hvis vi kun ser på en enkelt udførelse af eksperimentet forventer vi selvfølgelig ikke lige præcis 4 seksere – nogle gange vil der være flere seksere, nogle gange færre. Men hvis eksperimentet med de 24 kast med en ideel terning blev gentaget rigtigt mange gange (uendeligt mange gange) så ville vi i gennemsnit forvente 4 seksere i de mange stikprøver på 24 kast.
I det konkrete tilfælde fandt vi 10 seksere. Vi skal nu træffe et valg: Vil vi stadigvæk tro på at der er tale om ideel terning, dvs. at de overskydende seksere tilskrives tilfældige variationer, og dermed deltage i spillet, eller vil vi i stedet afvise terningen som ideel, og i stedet tro på at der

er en systematisk variation involveret som favoriserer seksere – fx at terningen er skæv eller at der er boret en lille blyklump ind i terningens ene side, og dermed afvise spillet. Det er dette valg mellem en tilfældig variation og en systematisk variation vi skal forholde os rationelt til. Vi skal altså udvikle en objektiv metode til at træffe valget, der alene bygger på den faktiske observation, dvs. i dette tilfælde de 10 seksere.

Hvis udfaldet i stedet havde været 24 seksere ville vi ikke have været i tvivl: Der ville helt klart være blevet manipuleret med terningen, idet sandsynligheden for at en ideel terning giver 24 seksere i træk kan udregnes som $(1/6)^{24} \approx 2.11 \cdot 10^{-19}$. Men det er en svimlende lille sandsynlighed. Hvis vi kaster 24 ideelle terninger i minuttet, skal vi vente i størrelsesordenen $5 \cdot 10^{18}$ minutter, svarende til 900 milliarder år, før vi kan forvente at få en enkelt serie på 24 seksere i træk. Selv om vi forestillede os hele Jordens fem milliarder store befolkning blev sat til at kaste terninger på denne måde dag og nat, skulle vi vente 150 år, før vi kan forvente en enkelt serie på 24 seksere i træk. Så det ville være absurd at tro på at terningen var ideel, hvis man i første og eneste forsøg fik 24 seksere i træk.

Men hvad med 10 seksere i 24 forsøg? Det er 6 mere end forventet. Er det også helt urimeligt at tro på en ideel terning? Er det virkeligt helt usandsynligt? Det er problemstillingen!

Vi indfører derfor en **nulhypotese** H_0, som hævder at afvigelsen fra det forventede ikke er større end at den med rimelighed kan tilskrives tilfældige variationer. Tilsvarende indfører vi en **alternativ hypotese** H_a, som hævder at afvigelsen er blevet så stor, at den ikke længere med rimelighed alene kan tilskrives tilfældige variationer. Støtter vi den alternative hypotese vælger vi altså at tro på, at den store afvigelse også skyldes indflydelsen fra en systematisk variation.

Men hvordan træffer vi valget mellem de to hypoteser? Her kan vi illustrere strategien med den såkaldte retssagsmetafor, hvor vi altså skal fælde dom i en sag om to påstande:

1. **Nulhypotesen** H_0 hævder at den observerede afvigelse ligeså godt kan tilskrives **tilfældigheder**, dvs. der er ikke nogen grund til at tro på en sammensværgelse.
2. **Den alternative hypotese** H_a hævder modsat at den observerede afvigelse er resultatet af en **systematisk variation**, dvs. der er grund til at tro at der er fusket med terningen.

Tilhængerne af den alternative hypotese H_a forsøger at få nulhypotesen forkastet, dvs. de optræder som *anklagere* i retssagen, mens tilhængerne af nulhypotessen H_0 tilsvarende forsøger at få den bekræftet, dvs. de optræder som forsvarere i *retssagen*. Der er ikke noget fældende bevis i sagen (vi mangler 'den rygende pistol'), så afgørelsen skal alene træffes på grundlag af **indicier**, dvs. vi skal foretage en vurdering af hvor overbevisende forskellen mellem de **observerede antal** og de **forventede antal** egentlig er. Begge parter kan i princippet have ret, og det eneste vi kan gøre er at forsøge at **sandsynliggøre** den ene hypotese frem for den anden. Det gør vi ved at vurdere **troværdigheden af nulhypotesen**: Den anklagede erklærer sig uskyldig og afviser al snak om fusk – han var bare på et uheldigt sted på et uheldigt tidspunkt. Hvis den anklagede under retssagen synes at have en høj troværdighed, fastholder vi nulhypotesen og frikender terningen for fusk. Ellers forkaster vi den og dømmer altså terningen for fusk på grundlag af de foreliggende indicier. Fokus for retssagen, dvs. den statistiske metode, er altså nulhypotesen. Der findes andre mere avancerede statistiske test-metoder, hvor den alternative hypotese udbygges og kommer mere i spil, men dem vil vi ikke beskæftige os yderligere med her.

Vi skal da på forhånd blive enige om tre forhold:

1. Hvilket **signifikansniveau** vil vi lægge til grund for domfældelsen, dvs. hvor stærke skal indicierne være?
2. Hvilken **teststørrelse** vil vi benytte til at afgøre sagen, dvs. hvordan vil vi måle afvigelsen mellem det observerede og det forventede og dermed sætte tal på den. Det er størrelsen af denne afvigelse, der skal sætte os i stand til at træffe et fornuftigt valg mellem at tro på tilfældigheder kontra systematiske variationer.
3. Hvilket mål vil vi bruge for **nulhypotesens troværdighed**, den såkaldte **p-værdi**. Hvis troværdigheden er høj er indicierne for sammensværgelsen svage og omvendt.

Signifikansniveauet: Her anvender vi det typiske niveau på 5%. Det er ikke en objektiv given størrelse, men beror alene på erfaringen: Den har vist sig at virke i praksis i mange sammenhænge. Men i konkrete sammenhænge kan det sagtens forekomme at man i stedet aftaler et andet signifikansniveau på fx 1% eller 10%. Det kan fx dreje sig om en industriproduktion, der skal opfylde et vist aftalt kvalitetsniveau. Signifikansniveauet er da afgørende for hvornår vi kan fastslå at afvigelsen fra den forventede kvalitet er blevet så høj, at der er tale om et kontraktbrud og der derfor skal udbetales erstatning. I sådanne erhvervskontrakter vil den konkrete værdi for signifikansniveauet være aftalt som en del af kontrakten.

Teststørrelsen: Den afhænger af hvilken type problemstilling vi undersøger. Først og fremmest skal det afgøres om der er tale om kategoriske eller numeriske variable. I terningeeksemplet med de seks udfald vil vi typisk opfatte en sekser som et kategorisk udfald (en succes), og vi vil tælle antallet af succeser. I andre tilfælde (fx højderne for en klasse) er der tale om udpræget numeriske variable. Her vil vi udelukkende kigge på test med kategoriske variable. For sådanne test har man siden 1900 brugt en teststørrelse udviklet af Pearson, hvor man ikke blot kigger på afvigelsen mellem det observerede og det forventede antal, men man kvadrerer denne afvigelse og dividerer med det forventede antal. I det ovenstående eksempel med terningen går man derfor frem på følgende måde: Antallet af observerede succeser er 10, men vi forventede kun 4, dvs. der er en afvigelse på 6. Succeserne bidrager derfor med $6^2/4 = 36/4 = 9$. Tilsvarende er antallet af fiaskoer 14 (eftersom vi havde 24 forsøg i alt), hvor vi forventede 20 fiaskoer. Fiaskoerne bidrager derfor med $(-6)^2/20 = 36/20 = 1.8$. Den samlede teststørrelse er ifølge Pearson derfor givet ved $9 + 1.8 = 10.8$. Det kan forekomme indviklet, men styrken ved Pearsons teststørrelse er dels, at den kan bruges i mange forskelligartede situationer, dels at den gør det nemt i praksis at vurdere nulhypotesens troværdighed. Pearsons teststørrelse for kategoriske data fører til det såkaldte chi2-test (udtales chi-i-anden eller chi-kvadrat), som er en af de mest udbredte statistiske test, og som er den test vi vil sætte fokus på i de følgende kapitler.

Nulhypotesens troværdighed, dvs. p-værdien: Resten er avanceret mekanik. Man vælger i første omgang at tro på nulhypotesen, dvs. kaster en ideel terning 24 gange og tæller antallet af seksere. Ved hjælp af teoretiske udregninger, konkrete terningkast eller simulerede terningekast, finder man nu ud af, hvor ofte 24 kast med en ideel terning giver anledning til en teststørrelse, der er mindst lige så stor som den observerede, dvs. der skal være mindst 10 seksere – eller Pearsons teststørrelse skal være mindst lige så stor som de observerede 10.8. De 24 kast betegnes i den forbindelse ofte som stikprøver fra en ideel population, der afspejler nulhypotesen, dvs. sekserne udgør netop en sjettedel af hele populationen. p-værdien angiver dermed netop hvor stor en brøkdel af de gentagne stikprøver, der viser sig at være skæve, dvs. rammer mindst lige så skævt som den observerede stikprøve – der måske, måske ikke, også stammer fra den ideelle population. Det er jo det, hele retssagen handler om.

Domfældelsen: Hvordan fældes der man nu dom i retssagen? Det gør man så snart man har fundet p-værdien, dvs. nulhypotesens troværdighed.

1) Hvis p-værdien ligger under signifikansniveauet, der her er på 5%, er troværdigheden blevet så lille, at det ikke længere er rimeligt at tro på at terningen er ideel. Hvis vi fx finder en p-værdi på 0.2% betyder det jo at den observerede forskel er så sjælden at den for en ideel terning kun forekommer for hver femhundrede af de gentagne stikprøver foretaget med en ideel terning. Det gør det ikke rimeligt at tro på at vi skulle have fremkaldt en sådan sjælden begivenhed i det ene eksperiment med den anklagede terning, vi har til rådighed. Vi dømmer derfor at der har været fusk i eksperimentet. Indicierne er så stærke at vi vælger at tro på at terningen er forfalsket.

2) Hvis p-værdien omvendt ligger over signifikansniveauet, der her er på 5%, er troværdigheden derimod så stor, at vi ikke kan afvise nulhypotesen på det foreliggende grundlag. Det betyder ikke at terningen ikke godt kunne være forfalsket. Det betyder alene at indicierne ikke er stærke nok til at begrunde en domfældelse.

Ligesom en retssag er en statistisk test aldrig helt sikker: Vi kan ikke ved hjælp af en statistisk test bevise, hvilken af de to hypoteser, der i virkeligheden er den rigtige. Den statistiske test handler dybest set om forholdene i en stor population, hvor man drager slutninger om hele populationen på basis af en mindre, tilfældigt valgt stikprøve. Man satser da på at stikprøven er *repræsentativ*, dvs. at den på korrekt vis afspejler forholdene i hele populationen. Men der er altid en risiko for at stikprøven på trods af vores omhyggelige tilrettelægning falder helt skævt ud og overhovedet ikke afspejler populationen. Den statistiske test kan altså ikke bevise eller modvise en hypotese, men kun udtale sig som om hvor troværdig hypotesen er. Uanset, hvor skævt resultatet forekommer at være i forhold til det resultat man forventer fra nulhypotesen, er der altid en teoretisk mulighed for at skævheden rent faktisk skyldes tilfældigheder, dvs. at det i virkeligheden er i overensstemmelse med nulhypotesen. Når man bekræfter henholdsvis forkaster en nulhypotese risikerer man derfor at begå to typer af fejl:

1) Man kan komme til at *forkaste en korrekt nulhypotese* (dvs. dømme en uskyldig, altså begå et justitsmord). Dette kaldes en **type I fejl**.

2) Man kan komme til at *bekræfte en forkert nulhypotese* (dvs. frikende en skyldig). Dette kaldes en **type II fejl**.

I den mere avancerede testteori må man derfor indgående kigge på disse to muligheder. Hele ideen bag testen er nu at vi har vurderet/udregnet sandsynligheden for at det forventede resultat er mindst lige så skævt som det observerede. Hvis den ligger under 5% opfatter vi det som ret usandsynligt at resultatet er fremkommet som resultatet af en tilfældighed ud fra nulhypotesen og vi forkaster nulhypotesen. Men hvis nulhypotesen rent faktisk er korrekt vil vi jo i 5% af tilfældene alligevel få et resultat, der er så skævt at vi alligevel forkaster den, dvs. vi kommer til at begå et justitsmord. Signifikansniveauet kan derfor også tolkes som risikoen for at begå et justitsmord, og den bør selvfølgelig være ret lille.

I den ovenstående fremstilling har vi fokuseret på den statistiske test som et beslutningsværktøj, der tillader os at vælge mellem to hypoteser. I traditionel statistik er man ofte meget mere forsigtig i sine formuleringer og opfatter udelukkende den statistiske test som et redskab, der kan benyttes til at vurdere troværdigheden af en hypotese. Først når der er indsamlet mange data fra forskellige uafhængige undersøgelser vil man ud fra en samlet vurdering af de forskellige undersøgelser kunne træffe et kvalificeret valg.

Fx har det vist sig at spørgsmålet om hvorvidt rygning er farlig ikke kan afgøres med en enkelt undersøgelse. Det er kun det samlede resultat af et meget stort antal forskelligartede undersøgelser, der peger på at rygningen er sundhedsfarlig.

Prototest: Hvad så med terningen, der gav 10 seksere i 24 forsøg? Bliver den dømt for fusk eller frikendes den? Ja, det kræver jo, at vi kan få styr på, hvordan en ideel terning opfører sig, når vi kaster den 24 gange. Prøv nu selv i klassen at kaste en terning 24 gange og tæl op hvor mange seksere I fik. Hvis p-værdien er over 5% burde der i en almindelig klasse med mindst 20 elever dukke mindst en serie op med mindst 10 seksere. Gjorde der det? Er det så et tilfredsstillende grundlag for at afgøre retssagen? Hvor mange gange bør I kaste en sådan serie, før I kan træffe en afgørelse?
Der er noget at spekulere over ☺

Vi slutter med en mere traditionel opstilling af den statistiske metode (hentet fra '**Introduction to the Practice of Statistics**' af David S. Moore and George P. McCabe (W.H. Freeman and Company):

En **signifikanstest** vurderer i hvor høj grad de fremlagte data svækker en **nulhypotese** H_0 til fordel for en **alternativ hypotese** H_a.

Hypoteserne er angivet i form af populationsparametre, her fx sandsynligheden for at få en sekser ved kast med terningen. Normalt udsiger nulhypotesen H_0 at der ingen effekt er til stede (dvs. at sandsynligheden for at få en sekser er netop 1/6), mens den alternative hypotese H_a, udsiger, at en parameter afviger fra sin nulværdi i en bestemt retning (**et-sidet alternativ**, her at sandsynligheden for at få en sekser er større end 1/6) eller i begge retninger (**to-sidet alternativ** her at sandsynligheden for at få en sekser med terningen er forskellig fra 1/6).

Testen er baseret på en **teststørrelse**. **p-værdien** er sandsynligheden for at den statistiske teststørrelse vil antage en værdi, der er mindst lige så ekstrem som den faktisk observerede, beregnet under forudsætning af, at nulhypotesen H_0 er sand. En lille p-værdi tolkes som et stærkt argument mod H_0. Beregningen af p-værdien kræver kendskab til stikprøvefordelingen for teststørrelsen under forudsætning af at nulhypotesen H_0 er sand.

Hvis p-værdien er lige så lille eller mindre end en aftalt værdi α, siger man at data er **statistisk signifikant** på **signifikansniveauet** α.

Eksempel: Når man undersøger indlæring hos dyr benytter man sig fx af et eksperiment hvor en mus skal finde vej gennem en labyrint. For en bestemt labyrint har det gennem tidligere forsøg vist sig at gennemsnitstiden er 18 sekunder for at finde vej gennem labyrinten. En forsker tror på at en høj lyd vil få musen til at finde hurtigere vej gennem labyrinten. Hun gennemfører derfor et forsøg med 10 mus, der skal finde vej gennem labyrinten samtidigt med at de udsættes for en høj lyd. Det viser sig, at gennemsnittet for de 10 mus er 16 sekunder.

I dette tilfælde er nulhypotesen H_0 altså at gennemsnitstiden for en mus, der udsat for en høj lyd løber gennem labyrinten, er 18 sekunder. Den alternative hypotese H_a er tilsvarende at gennemsnitstiden for en mus, der udsat for en høj lyd løber gennem labyrinten er mindre end 18 sekunder. Nulhypotesen udsiger netop at den høje lyd er uden effekt. Den alternative hypotese udsiger tilsvarende, at den høje lyd har en effekt - nemlig at gennemsnitstiden falder.

Resultatet af eksperimentet er en gennemsnitstid for 10 mus på 16 sekunder. Vi skal vurdere om denne afvigelse er statistisk signifikant på 5%-niveauet. Som teststørrelse kunne vi da fx anvende forskellen i middelværdi: Den forventede middelværdi er 18 sekunder, den observerede middelværdi er 16 sekunder. Selv om det er lavere end det forventede kunne forskellen godt skyldes en tilfældighed. Vi er derfor nødt til at vurdere p-værdien, dvs. sandsynligheden for at de samme 10 mus ville kunne opnår et gennemsnit på 16 sekunder, hvis de *ikke* var udsat for en høj lyd.

Det er beregningen af p-værdien, der er kernen i den statistiske test. I det ovenstående eksempel har vi *ikke* oplysninger nok til at kunne gennemføre beregningen. Men hvis vi fx vidste fra de tidligere forsøg at gennemløbstiden for en mus er normalfordelt med en middelværdi på 18 sekunder og en spredning på fx 1.5 sekund, så ville vi kunne gennemføre beregningen teoretisk. Hvis vi i stedet for gennemsnitstiden for de 10 mus havde alle gennemløbstiderne til rådighed for de 10 mus, der løb gennem labyrinten under påvirkning af en høj lyd, ville vi stadigvæk kunne udføre testen eksperimentelt.

Vi vender tilbage til eksemplet!

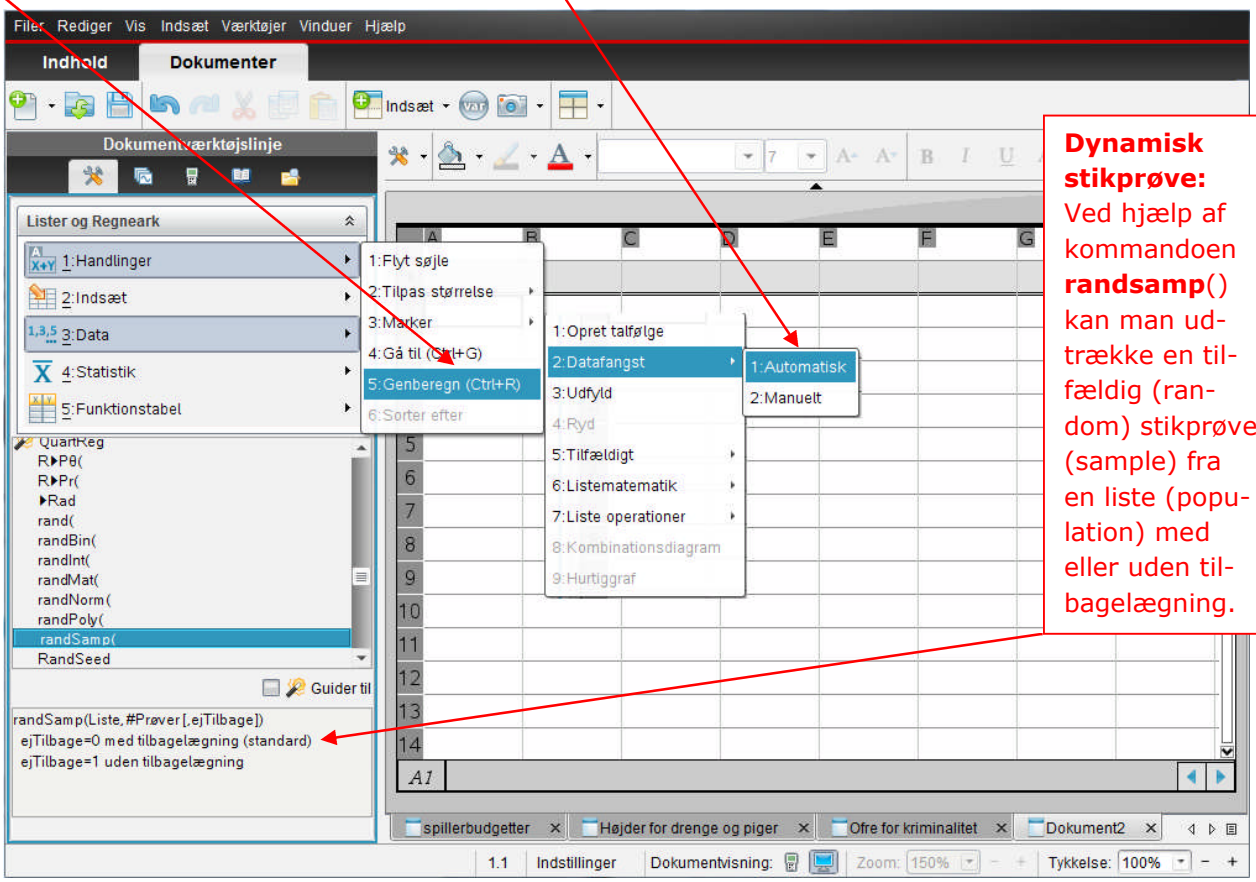
6.2 Tilfældig variation: Eksperimentelle metoder

Historisk set er mange metoder til at kunne håndtere tilfældige variationer først udviklet eksperimentelt, hvorefter der er udviklet en avanceret statistisk teori med formler til at automatisere udregningen af den tilfældige variation. Denne udvikling skyldes ikke blot ønsket om en bedre forståelse af metoderne, men også at de eksperimentelle metoder, så længe de måtte udføres med håndkraft, var besværlige og tidsrøvende. Med indførslen af computerne har det sidste aspekt imidlertid ændret sig radikalt, og i vore dage er de eksperimentelle metoder lige så tilgængelige som de teoretiske.

Begge typer af tilgange understøttes af **TI-Nspire CAS**, idet de teoretiske metoder typisk ligger i menuerne for test og konfidensintervaller, mens de eksperimentelle metoder især bygger på tilfældige stikprøver, dvs. kommandoen **RandSamp()**, såvel som muligheden for at udføre gentagne målinger ved at kombinere automatisk genberegning (**CTRL R**) og automatisk datafangst, dvs. regnearksproceduren **Capture()**.

Genberegn: **CTRL R**
Regnearket opdateres løbende, men de tilfældige rutiner **rand**() osv. genberegnes kun hvis man vælger **Genberegn**!

Datafangst: Automatisk eller manuel
Hvis man har oprettet en navngivet variabel (en **måling**) kan man fange dens værdier automatisk, når de ændres, og manuelt (**CTRL .**), når som helst. De indfangne værdier gemmes i en liste.



Dynamisk stikprøve:
Ved hjælp af kommandoen **randsamp**() kan man udtrække en tilfældig (random) stikprøve (sample) fra en liste (population) med eller uden tilbagelægning.

Begge typer af tilgange har fordele og ulemper. Tilegnelsen af de teoretiske metoder kræver en indføring i vanskeligt tilgængelige teoretiske begreber, der nemt kan komme til at skygge for de mere principielle og grundlæggende træk ved den statistiske metode. Ydermere er mange grundlæggende problemstillinger principielt umulige at håndtere ved eksakte metoder, fordi der ikke findes færdige formler. Endelig er de teoretiske metoder ikke nødvendigvis mere præcise end de eksperimentelle. Kigger vi igen på test med kategoriske variable findes der godt nok eksakte test til specielt simple problemstillinger. Den mest kendte hedder binomialtesten. Men den er *ikke* implementeret i **TI-Nspire CAS**, hvor den er erstattet af dels den såkaldte z-test, der bygger på den såkaldte normalfordelingsapproksimation og derfor kun er en approksimativ test,

dels af chi2-testen, der også kun er approksimativ– alle formlerne knyttet til binomialfordelingen er dog tilgængelige, så hvis man ønsker det kan man selv udføre binomialtesten eksakt.

I modsætning hertil er de eksperimentelle metoder mere fleksible. De kan ikke blot altid anvendes – alene eller som supplement til de teoretiske metoder; De kan også anvendes på en langt større gruppe af problemstillinger end de enkelte teoretiske metoder. De teoretiske metoder er nemlig ofte skræddersyede til kun at kunne anvendes på en snæver gruppe af problemstillinger, der oven i købet ofte kræver specielle forudsætninger opfyldt før man fuldt ud kan stole på de opnåede resultater. Til gengæld er de eksperimentelle metoder ikke så præcise som de eksakte teoretiske metoder, og de giver typisk kun resultater, der fortæller noget om det helt konkrete datasæt. Ændres der i datasættet må man derfor begynde helt forfra med undersøgelsen.

I det følgende vil vi nu redegøre for de vigtigste eksperimentelle metoder til at simulere en tilfældig variation. Dermed bliver vi selv i stand til at løse simple opgaver indenfor den bekræftende statistik.

1. **Tilfældighedsgeneratorer**. Dette er den klassiske metode til at simulere tilfældighed. Ved hjælp af en tilfældighedsgenerator kan vi frembringe serier af tilfældige tal, der igen kan benyttes som udgangspunkt for tilfældige valg i et datasæt. Denne teknik har altid været til rådighed, idet man tidligere benyttede officielle tabeller over tilfældige tal til at frembringe de ønskede serier af tilfældige tal. Men i vore dage er de sådanne tabeller erstattet af simple funktioner på lommeregnere og i regneark.
2. **Omrøring af variable**. Dette er en ny og elegant metode til at simulere uafhængighed mellem to grupper af data, dvs. til at sikre at enhver forskel mellem de to grupper netop *kun* kan skyldes tilfældige variationer. Man fjerner altså ganske enkelt eventuelle systematiske variationer ved en omrøring.
3. **Bootstrap**. Dette er endnu en ny og elegant metode til at simulere den naturlige variation i et datasæt. I modsætning til omrøringen bevarer bootstrappet stadigvæk de systematiske variationer i datasættet. Men da man nu får overlejret de naturlige variationer, kan man direkte se hvor stor indflydelse de naturlige tilfældige variationer har på den systematiske variation.

I de sidste 20 år har den eksperimentelle metode vundet betydelig accept i såvel industrien som undervisningsverdenen. Det skyldes dels den fleksibilitet, den tilbyder, dels at den er langt mere konkret end den teoretiske metode, som forudsætter et indgående kendskab til sandsynlighedsregning. Der er altså store begrebsmæssige fordele ved at anvende den eksperimentelle metode i undervisningen.

Selv om den eksperimentelle metode i princippet kan udføres i hånden, så kan den i praksis kun udføres med brug af en computer (hvilket igen forklarer, hvorfor den først er slået igennem de sidste 20 år). Ydermere kræver den specielle programmer, der er designet til at udføre eksperimentelle tests. Sådanne programmer kan hentes på nettet (søg på 'resampling' på Google!), men i undervisningen kan det være en fordel at benytte et program, der er designet til undervisningsbrug.

6.3 Hvad så med gambleren med terningen? Blev han 'dømt' for fusk?

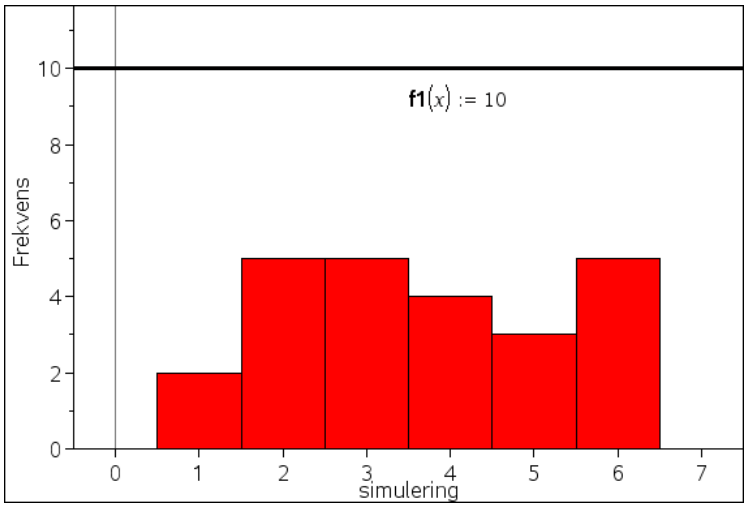
Her var udgangspunktet et prøvespil, hvor gambleren kastede en terning 24 gange og opnåede 10 seksere. Hvis vi klassificerer seksere som succes, opnåede han altså 10 succeser, hvor vi kun forventede 4 succeser. Er det så usædvanligt at vi må forkaste nulhypotesen:

H_0: Terningen opfører sig ideelt og giver i gennemsnit kun seksere i en sjettedel af tilfældene.

For at afgøre det bliver vi nødt til at vide noget mere om hvordan en ideel terning opfører sig: Hvor tit sker det at man kaster en ideel terning 24 gange og får mindst 10 seksere? Samtidigt skal vi på forhånd have aftalt et signifikansniveau, dvs. hvor sjælden en begivenhed skal der være tale om før vi forkaster nulhypotesen. Vi kan da holde p-værdien (andelen af serier med et antal succeser på mindst 10) op mod signifikansniveauet (hvor utroværdig skal nulhypotesen være, for at vi opgiver den). Vi vælger det traditionelle signifikansniveau på 5%.

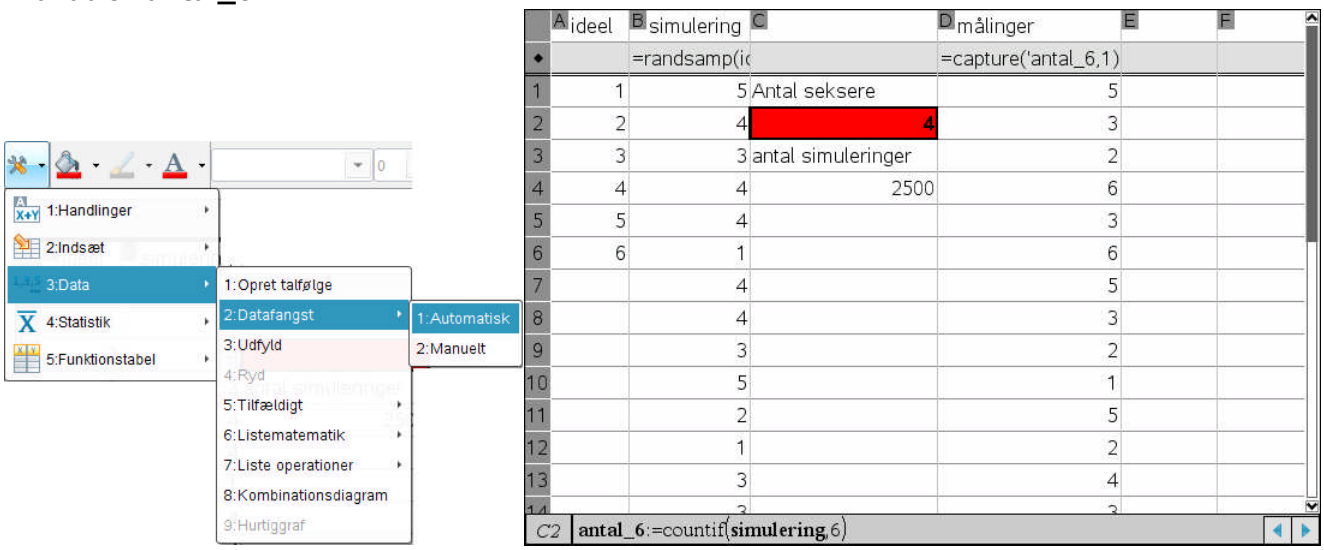
Den nemmeste måde at få hold på det er at simulere terningekastet. Vi opretter derfor en liste, **ideel**, med seks elementer, nemlig udfaldene 1, 2, 3, 4, 5 og 6. Trækker vi tilfældigt fra denne liste *med tilbagelægning*, vil vi netop forvente at få en sekser i en sjettedel af tilfældene. Vi trækker altså stikprøver på 24 elementer *med* tilbagelægning fra listen **ideel**, og tæller op grafisk eller med en cellekommando, hvor mange seksere vi opnår:

	A ideel	B simulering
◆		=randsamp(ideel,24)
1	1	6
2	2	3
3	3	2
4	4	3
5	5	6
6	6	3
7		4
8		3
9		5
10		1
11		1
12		5
13		4
14		6
15		6
16		5
17		2
18		4
19		2
20		6
21		2
22		2
23		3
24		4
B	simulering:=randsamp(ideel,24)	

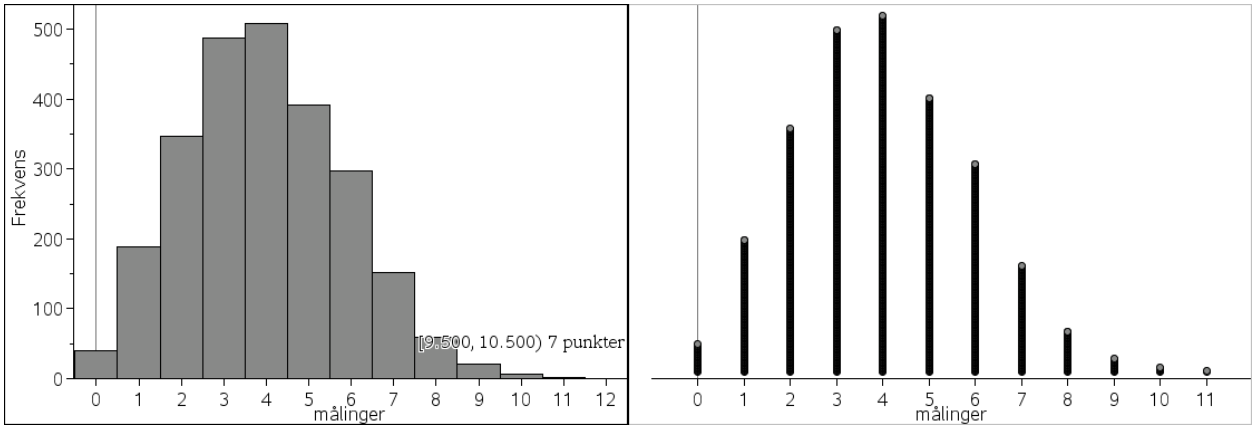


I dette tilfælde opnåede vi altså fem succeser (næsten som forventet), men taster vi **CTRL-R** inde i **Lister og Regneark**-værkstedet, vil vi få gentaget simuleringen, og søjlerne vil da blafre op og ned. Vi kan da hurtigt se at det fx er ret let at kaste 24 terninger uden at få succes en eneste gang, men til gengæld synes det ret svært at få mindst 10 succeser – vi har indlagt en overligger, y = 10, for at gøre det nemmere at se, hvornår vi får et ekstremt udfald. Denne simple observation svækker altså nulhypotesen: Hvis gambleren ikke har fusket med terningen skal man være ret så heldig for i et enkelt forsøg at få 10 seksere.

Men vi er nødt til at få et mere præcist hold over hvor svært det er at få en skæv serie med mindst 10 seksere. Vi tæller derfor antallet af succeser med en Countif-kommando i cellen C2. Det sker via cellekommandoen antal_6:= countif(**simulering**,6) som gemmer værdien af C2 i variablen **antal_6** :



Det åbner netop muligheden for at vi som vist kan fange værdien af variablen antal og gemme den i listen **målinger** ved at bruge capture-proceduren. Læg godt mærke til capture-proceduren: Det er *ikke* en normal matematisk kommando, som kan bruges i alle værksteder. Den kan *kun* anvendes i **Lister og regneark**-værkstedet, hvor den kaldes fra **Data**-menuen ved at markere formelfeltet (men ikke gå ind i det!). Den ekstra parameter 1, fortæller at der er tale om en automatisk datafangst, dvs. **antal_6** registreres automatisk, hver gang den ændres. Taster vi nu **CTRL R** rigtigt mange gange vil vi få fyldt listen med målinger – der er i alt plads til 2500 målinger. Vi kan da afbilde målingerne i deres eget histogram henholdsvis prikdiagram:



Vi ser da, at når man får lov at forsøge sig 2500 gange med 24 kast med en terning, kan det faktisk godt lade sig gøre at få 10 seksere. Det skete i dette tilfælde i 7 af serierne. Tilsvarende har vi fået 11 succeser i 2 af serierne, mens vi aldrig fik fat i 12 succeser. Altså har vi ramt skævt 9 gange i løbet af de 2500 forsøg. Men det er ikke nemt at ramme skævt, for andelen af skæve ligger under 1‰. Med en p-værdi på under 1‰ er nulhypotesens troværdighed så lav sammenlignet med signifikansniveauet, at vi ikke kan opretholde den.

Så svaret er JA! Gambleren blev dømt for fusk. Men vi har faktisk også fusket lidt ☺ For vi har anvendt automatisk datafangst og den virker kun når værdien af **antal_6** skifter. Og det sker faktisk – om end det er sjældent – at vi får det samme antal seksere to gange i træk. Det er dog mest de typiske værdier som 3, 4 og 5, der på denne måde falder ud af statistikken. Så i virkeligheden har vi brugt mere end 2500 forsøg på at frembringe mindst 10 skæve. Men det gør det jo heldigvis bare endnu sværere at opretholde nul-hypotesen.

6.4 Hvad så med musene – var der belæg for at de løb hurtigere?*

Dette afsluttende eksempel kan roligt overspringes i en første gennemlæsning. Det er et sidespor i forhold til chi2-testen, der er i fokus i resten af hæftet og er kun medtaget for at man også kan få et indtryk af bootstrap-metoden, som er meget anvendelig i projektsammenhænge.

For at kunne træffe en afgørelse vedrørende laboratoriemusene, skal vi først og fremmest have nogle flere oplysninger: Som et minimum skal vi kende de individuelle gennemløbstider for de ti mus. De fremgår af følgende tabel, hvor vi også har afbildet dem som et boksplot:

Som det ses ligger medianen på 16 sekunder (som også var gennemsnittet for den tid, de brugte på at løbe gennem labyrinten under påvirkning af en høj lyd). Tilsvarende ligger tredje kvartil på 18 sekunder (som er det typiske gennemsnit, for den tid det tager en mus at løbe gennem labyrinten *uden* påvirkning af en høj lyd). Tre af musene bruger faktisk en tid, der er længere end de 18 sekunder. Kunne det så ikke godt tyde på at løbetiderne for disse ti mus bare er tilfældige variationer over løbetiderne for en typisk mus, der ikke påvirkes af en høj lyd, altså at den høje lyd slet ikke har haft nogen påviselig indflydelse? Det fører til nulhypotesen

> H_0: Lyden har ingen indflydelse – dvs. der er blot tale om tilfældige variationer i gennemløbstiderne, hvor den typiske gennemløbstid stadigvæk er 18 sekunder.

Det er nu vi skal anvende metoder fra den bekræftende statistik. Hvis vi ikke har flere oplysninger end de ovenstående, vender vi nu problemstillingen og spørger: På basis af de ti gennemløbstider, er det så rimeligt at tro på at den typiske gennemløbstid for en mus, der løber gennem labyrinten under påvirkning af en høj lyd, kan være 18 sekunder (nulhypotesen)? For hvis det er tilfældet, har vi ikke jo påvist nogen sikker indflydelse af lyden.

Vi antager da at de 10 gennemløbstider er repræsentative for mus, der løber gennem labyrinten under påvirkning af en høj lyd. Vi kan derfor bruge disse 10 gennemløbstider som en ideel fordeling af gennemløbstiden for sådanne mus. Hvis vi trækker en stikprøve fra denne fordeling *med* tilbagelægning, så finder vi altså typiske stikprøver fra mus, der påvirkes af en høj lyd. Specielt kan vi tage en stikprøve på 10 mus *med* tilbagelægning. Det svarer til en virtuel gentagelse af selve eksperimentet. En sådan stikprøve kaldes også en bootstrap, fordi vi så at sige trækker stikprøven uden at bruge anden information, end den der ligger i selve eksperimentet. Metoden kaldes derfor **bootstrap**-metoden. Men det afgørende for metoden er altså antagelsen om at den oprindelige stikprøve er repræsentativ og derfor afspejler den generelle fordeling hørende til gennemløbstider for mus, der løber gennem labyrinten under påvirkning af en høj lyd. Hvis denne antagelse viser sig ikke at holde er metoden meningsløs!

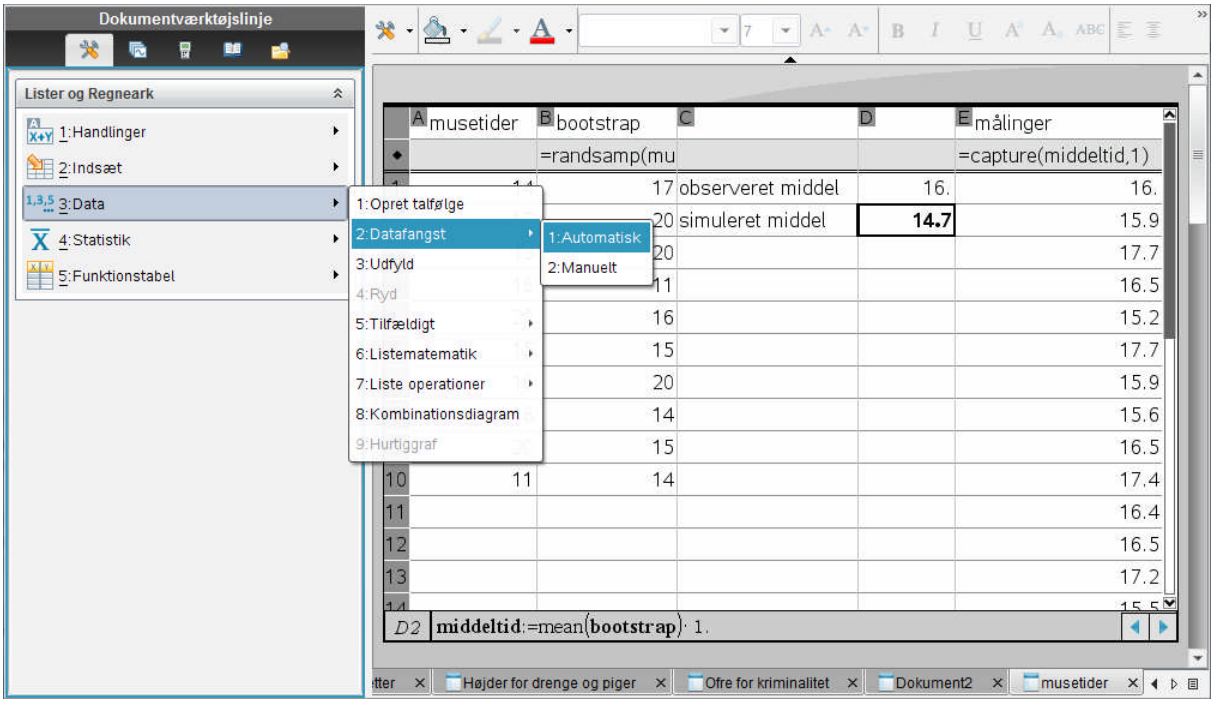
Selve bootstrappet udføres nu med kommandoen RandSamp(**musetider**,10), idet det er underforstået at randsamp-kommandoen som standard trækker *med* tilbagelægning, hvis vi ikke tilføjer en ekstra parameter. Derefter udregner vi gennemsnitstiden for bootstrappet. Hvis dette gennemsnit ligger under 18 sekunder favoriserer det antagelsen om at lyden har en indflydelse (dvs. den alternative hypotese). Men ligger det over 18 sekunder favoriserer det antagelsen om at lyden ikke har nogen indflydelse (dvs. nulhypotesen).

	A musetider	B bootstrap	C	D
◆		=randsamp(musetider,1		
1	14	11	Observeret middel	16
2	17	11	Simuleret middel	**14.8**
3	13	20		
4	16	14		
5	20	18		
6	15	20		
7	16	14		
8	18	14		
9	20	11		
10	11	15		
11				

D2 **middeltid**:=mean(**bootstrap**)·1.

Læg mærke til at vi har gemt det simulerede gennemsnit i variablen **middeltid** ved at indsætte **middeltid**:= foran formlen mean(**bootstrap**), og at vi har ganget med 1. for at tvinge **TI-Nspire CAS** til at regne med decimaltal.

Vi kan nu gentage simuleringen mange gange ved at taste **CTRL R** i **Lister og regneark**-værkstedet og på den måde se hvor mange gentagelser, der favoriserer antagelsen om at lyden har en indflydelse og hvor mange der ikke gør det. Det vil da være bekvemt at opsamle de simulerede gennemsnit i en ny variabel, kaldet **målinger**:

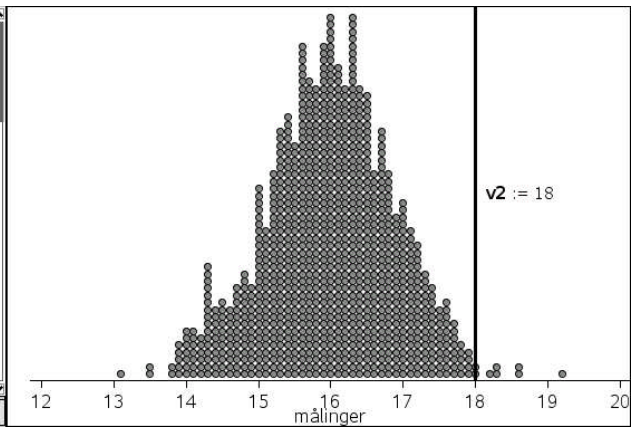


Læg mærke til capture-kommandoen. Det er *ikke* en normal matematisk kommando, som kan bruges i alle værkstederne. Den kan *kun* anvendes i **Lister og regneark**-værkstedet, hvor den kaldes fra **Data**-menuen. Den ekstra parameter 1, fortæller at der er tale om en automatisk datafangst, dvs. **middeltid** registreres automatisk, hver gang den ændres. Men så kan vi jo nemt samle tusinde målinger, dvs. gentage eksperimentet tusinde gange, og se hvor mange af gentagelserne der favoriserer antagelsen om at lyden ikke har en indflydelse (nulhypotesen). Hvis der er mange sådanne gentagelser er det nemt at forsvare nulhypotesen, men hvis der er meget få sådanne gentagelser, gør det det meget svært at opretholde nulhypotesen.

Som sædvanlig skal vi da først aftale et signifikansniveau, dvs. hvilken andel skiller mellem at vi vælger fortsat at tro på nulhypotesen henholdsvis at vi opgiver at tro på nulhypotesen. Vi vælger her det traditionelle signifikansniveau på 5%.

	B bootstrap	C	D	E målinger
◆	=randsamp(m			=capture('middeltid,1)
1	11	Observeret middel	16	15.4
2	11	Simuleret middel	**14.8**	16.2
3	20	Antal målinger	1000	15.3
4	14	Antal skæve	8	15.6
5	18	p-værdi	0.8	17.
6	20			15.3
7	14			15.
8	14			16.6
9	11			16.4
10	15			16.5
11				17.

D2 **middeltid**:=mean(**bootstrap**) 1.



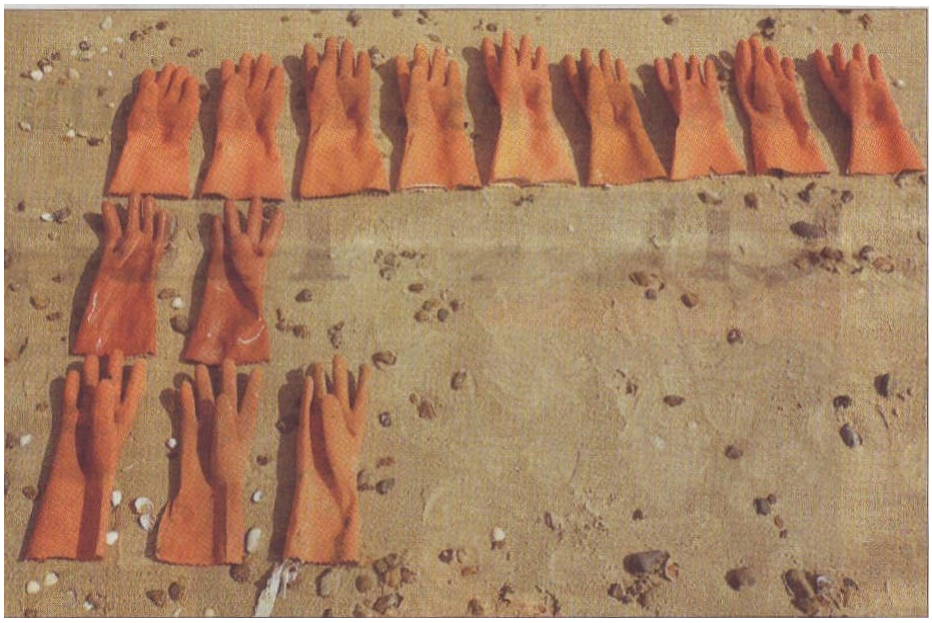
Vi ser da, at der kun er i 8 af gentagelserne at vi frembringer et gennemsnit på 18 sekunder eller derover. Det er så sjældent, idet det kun svarer til 0.8%, at vi på signifikansniveauet 5% må afvise nulhypotesen. Det er ikke rimeligt at tro at den observerede gennemsnitstid for de ti laboratoriemus i virkeligheden blot skyldes tilfældige variationer i gennemløbstiderne for mus med en typisk gennemløbstid på 18 sekunder. Eksperimentet svækker altså nulhypotesen betydeligt til fordel for den alternative hypotese: En høj lyd nedsætter gennemløbstiden i labyrinten.

Så svaret er Ja! Der er rent faktisk belæg for at de løb hurtigere.

Bemærkning: Ordet **bootstrap** (dvs. støvlestrop) stammer fra Baron von Münchausens eventyr, hvor den løgnagtige baron fortæller, hvordan han reddede sig op af en sump ved at trække i støvlestropperne. I den klassiske illustration af Doré trækker han sig i stedet op af sumpen ved at hive sig i håret. I statistik benyttes det som kælenavn for en metode, hvor man genskaber den oprindelige population ud fra tilfældige udtrækninger i en repræsentativ stikprøve. Når stikprøven afspejler de væsentligste træk fra hele populationen kan man ved at trække fra stikprøven opnå tilnærmelsesvis de samme resultater, som hvis man trak fra hele populationen.

7 Handskerne på stranden: χ^2-test for Goodness-of-fit

For nylig kunne man læse følgende lille notits på bagsiden af Politiken:



Takket være en utrættelig indsats af Thyge Steffensen og hans frue kan vi i dag præsentere dette års imponerende høst af gummihandsker fra parrets strandture ved Jammerbugten. Som det fremgår, er det blevet til 9 venstrehandsker, 1 intakt par - og kun 3 højrehandsker! Er det, fordi de formentlig primært højrehåndede kystfiskere ofte smider venstre handske væk? Eller er der en anden forklaring? »Det bemærkelsesværdige forhold må interessere mange«, skriver det årvågne strandvandrerpar, vi er ganske enige. (bs)

Det er tydeligt at asymmetrien mellem venstrehandsker og højrehandsker har overrasket ægteparret. Men er det nu virkelig så usædvanligt, som det umiddelbart ser ud til? Umiddelbart ville vi forvente at der blev tabt lige mange venstre og højrehandsker, så der burde vel også drive lige mange af hver slags ind på stranden. Men handskerne driver jo for vind og vejr, så det er ret tilfældigt, hvor de havner på stranden. Vi forventer derfor heller ikke lige præcis lige mange hver gang – men sådan ca. lige mange, hvor der somme tider vil være flest venstre handsker og somme tider flest højrehandsker. Vi kan derfor opstille en nulhypotese H_0, der siger at sandsynligheden for at finde en venstrehandske er lige så stor som sandsynligheden for at finde en højrehandske, og at vi derfor i gennemsnit forventer at finde lige mange. Samtidig har vi en observation, hvor vi rent faktisk har fundet 10 venstrehandsker og 4 højrehandsker. Er det nu set rimeligt ud fra nulhypotesen? Det er hvad vi skal forsøge at finde et kvalificeret svar på!

7.1 Simulering af nulhypotesen

Hvor ekstrem er denne observation nu? For at undersøge det må vi først opbygge en **ideel** fordeling i overensstemmelse med nulhypotesen. Da der skal være lige mange venstrehandsker og højrehandsker behøver den ideelle fordeling kun at bestå af en enkelt venstrehandske og en enkelt højrehandske. Læg mærke til at venstre og højre skal skrives ind med gåseøjne for at markere dem som tekstvariable (kategoriske variable).

	A ideel	B
◆		
1	Venstre	
2	Højre	

A1	"Venstre"

A1	"venstre"

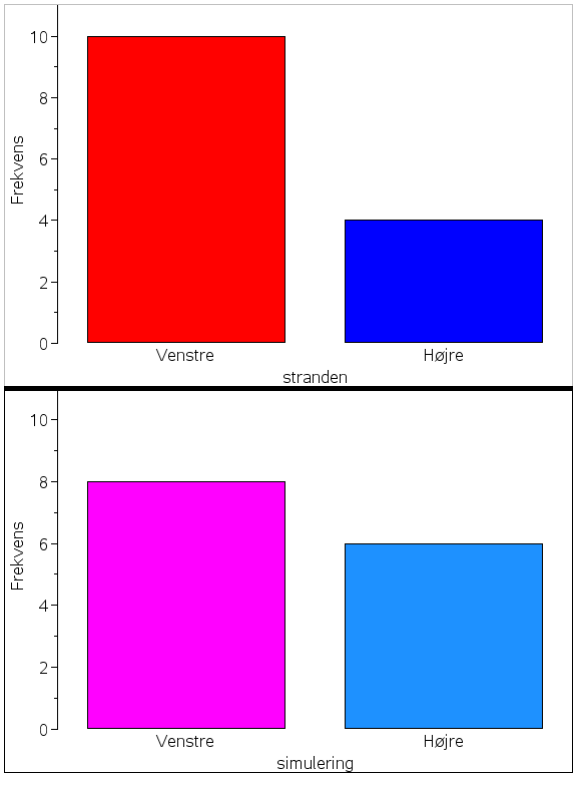
Vi vil også opbygge en liste over de observationer, som hr. og fru Steffensen har gjort. Det kan vi fx gøre ved hjælp af kommandoen **freqtable@>list**({"Venstre","Højre"},{10,4}), hvor kategorilisten {"Venstre","Højre"} fortæller os hvad slags handsker der skal stå i listen, mens hyppighedslisten angiver hvor mange gange de enkelte kategorier skal optræde, dvs. først 10 venstrehandsker og derefter 4 højrehandsker.

Endelig skal vi have opbygget en *simulering* af nulhypotesen, dvs. vi skal have trukket 14 tilfældige handsker fra den ideelle fordeling. Det sker ved hjælp af listekommandoen **simulering**

= randsamp(**ideel**,14). Den opbygger netop en tilfældig stikprøve med 14 elementer trukket fra listen **ideel**, hvor der er tale om stikprøve *med* tilbagelægning, dvs. vi trækker igen og igen fra den samme ideelle fordeling. Endelig kan vi afbilde såvel de originale observationer fra strandturen i Jammerbugten som de simulerede observationer i søjlediagrammer, der grafisk viser antallet af venstre- og højrehandsker. Ved at sammenholde resultaterne fra simuleringen med resultaterne fra den oprindelige strandtur kan vi nu vurdere hvor usædvanlig en strandtur, der egentlig har været tale om.

	A ideel	B stranden	C simulering
◆		=freqtable▸list({"Venstre",	=randsamp(id
1	Venstre	Venstre	Venstre
2	Højre	Venstre	Højre
3		Venstre	Venstre
4		Venstre	Venstre
5		Venstre	Venstre
6		Venstre	Højre
7		Venstre	Højre
8		Venstre	Venstre
9		Venstre	Venstre
10		Venstre	Højre
11		Højre	Venstre
12		Højre	Venstre
13		Højre	Højre
14		Højre	Højre
C	**simulering**:=randsamp(**ideel**,14)		

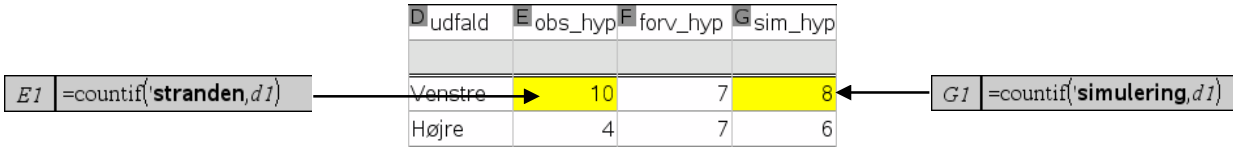
Vi ser da, at i dette tilfælde ligger antallet af venstre og højrehandsker i simuleringen noget tættere på hinanden end de gjorde på strandturen. Det er helt som forventet. Men for at kunne vurdere hvor sjældent resultatet fra strandturen egentlig er gentager vi nu simuleringen ved indefra **Lister og regneark**-værkstedet at taste **CTRL R** (ReCalculate). Vi ser da søjlerne blafre i takt med at antallene går op og ned. Jeg prøvede fx 20 gange og i løbet af disse tyve gange oplevede jeg nogle få gange at ramme en simulering, der var lige så skæv som strandturen, så mere sjælden er den altså heller ikke. Jeg oplevede også, at simuleringen gav den samme fordeling af venstre og højrehandsker to gange i træk, selv om rækkefølgen i de to tilfælde var forskellig. På denne måde kan man begynde at danne sig en ret god formodning om hvorvidt observationen er i rimelig overensstemmelse med nulhypotesen hvorfor vi ikke kan forkaste nulhypotesen, eller om den i den grad strider mod nulhypotesen, at vi bør forkaste den.



I langt de fleste tilfælde vil man på baggrund af de første tyve gentagelser kunne se, hvorvidt det er værd at gå videre med undersøgelsen. Hvis vi allerede ser flere ekstreme tilfælde blandt de 20 betyder det jo at p-værdien er noget højere end 5% (hvilket svarer til at vi i gennemsnit finder mere end et ekstremt tilfælde for hver gang vi udfører 20 simuleringer). Og så kan vi jo ikke forkaste nulhypotesen på signifikansniveauet 5%.

7.2 Hvor stor er afvigelsen: Opbygningen af teststørrelsen

For nu yderligere at underbygge denne fornemmelse og for at få en mere præcis vurdering af p-værdien, dvs. sandsynligheden for at få et udfald, der er mindst lige så ekstremt som det observerede vil vi indføre en teststørrelse, der kan benyttes til at vurdere hvor ekstremt det observerede egentligt er. Først må vi da have talt handskerne, hvilket gøres ved hjælp af den betingede kommando **CountIf**(kategori, betingelse). I vores tilfælde er der to kategorier: Venstre- og Højre-handsker. Vi indfører derfor en kategoriliste med disse to muligheder (i dette tilfælde falder den sammen med den ideelle liste, men i almindelighed vil dette ikke være tilfældet, fordi udfaldene fra den ideelle liste ikke behøver være lige sandsynlige). Derefter er vi klar til at tælle såvel de observerede som de simulerede hyppigheder. Endelig tilføjer vi en liste over de forventede hyppigheder: Ifølge nulhypotesen er der lige mange venstre og højrehandsker og det er kun tilfældigheder, der afgør om det er den ene eller anden handske fra et givet par, der lander på stranden.

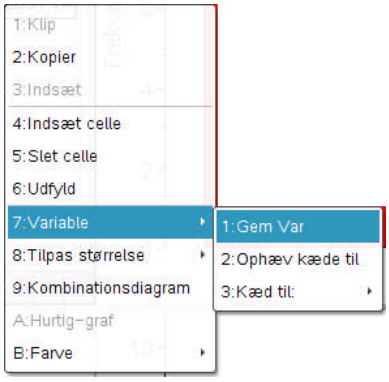


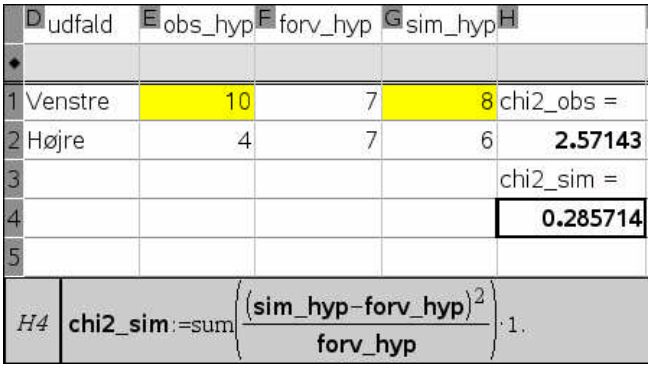
Der er flere muligheder for valget af en sådan teststørrelse, men et godt bud på en samlet forskel kunne bestå i at *kvadrere* forskellene mellem de observerede og de forventede værdier og lægge kvadraterne på disse forskelle sammen $sum((observeret - forventet)^2)$. Men som påvist af statistikeren Pearson i 1900 kan det betale sig yderligere at dividere med de forventede værdier

$$sum\left(\frac{(observeret - forventet)^2}{forventet}\right)$$

Læg mærke til, at jo længere væk de observerede værdier ligger fra de forventede, dvs. jo mere skæv den observerede fordeling af venstre- og højrehandsker er, jo større er teststørrelsen. Den fungerer derfor som et godt mål for hvor skæv den observerede fordeling af handskerne egentlig er. Vi udregner derfor denne teststørrelse først for de observerede værdier på stranden og dernæst for de simulerede værdier, der jo følger nulhypotesen om at der smides lige mange venstre og højrehandsker i havet og kun rene tilfældigheder afgør, hvem der når stranden. Taster vi **CTRL R** i **Lister og Regneark**-værkstedet kan vi nu igen nemt se, hvornår den simulerede teststørrelse er lige så ekstrem, dvs. lige så stor som den observerede.

Den ovenstående teststørrelse kaldes χ^2-teststørrelsen (udtales chi-2) og skrives altså ofte med et græsk c, dvs. med symbolet χ, men i **TI-Nspire CAS** kan dette overalt erstattes med chi. Vi lagrer derefter såvel den observerede teststørrelse som den simulerede teststørrelse fx ved hjælp af kommandoen **CTRL L** (for Link), men vi kan også højreklikke i cellen og vælge menupunktet **Variable...** > **Gem Var** ligesom vi simpelthen selv kan tilføje variabelnavnet og kolonet : med kolontegnet lige foran lighedstegnet i celleformlen



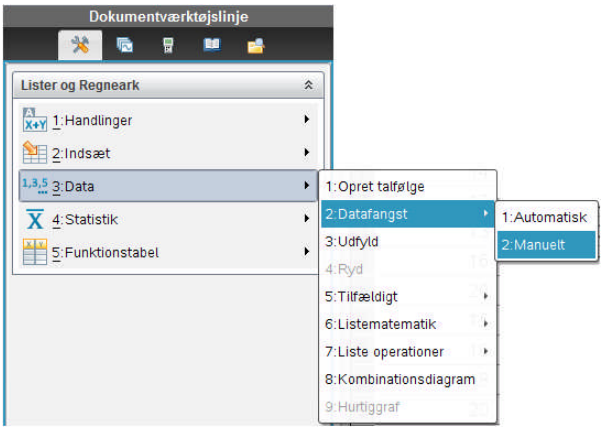


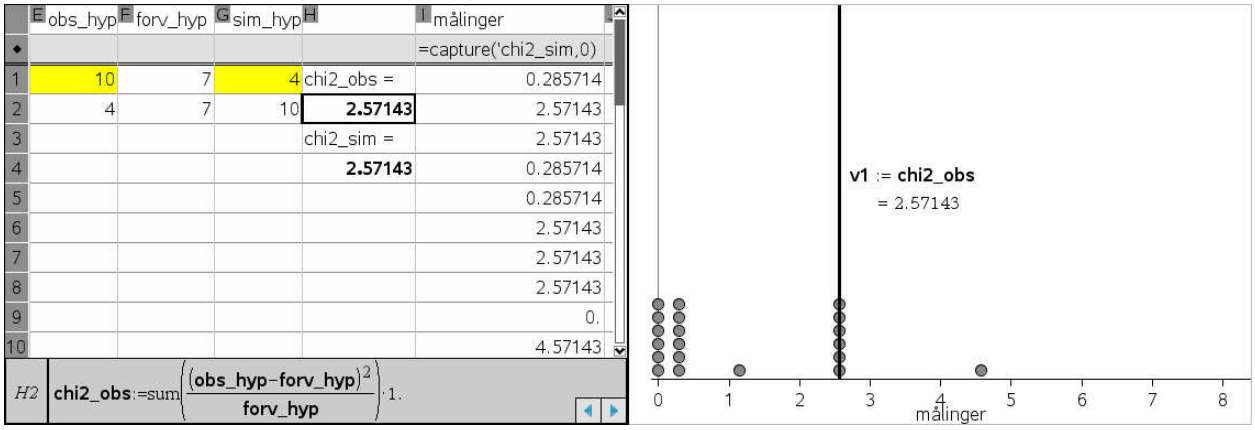
Læg mærke til, at cellen nu fremstår med fed skrift som tegn på at den er lagret som en variabel og vi derfor kan referere til cellens indhold i alle andre værksteder, der måtte være åbne.

7.3 Datafangst: Den eksperimentelle stikprøvefordeling

Det er den variable teststørrelse **chi2_sim** vi nu vil *måle* på en masse gange for at finde ud af hvordan den fordeler sig, dvs. vi vil undersøge *stikprøvefordelingen* af teststørrelsen rent eksperimentelt. Specielt vil vi holde øje med hvordan den ligger i forhold til den observerede teststørrelse **chi2_obs**, der er lagret i celle H2. Vi ved nemlig at den simulerede teststørrelse i kraft af sin konstruktion afspejler nulhypotesen, så ved at kigge på, hvor den observerede teststørrelse ligger i stikprøvefordelingen, kan vi få en god fornemmelse for hvor sjælden den er ifølge nulhypotesen – og dermed også hvor troværdigt det er, at den skulle stamme fra nulhypotesen.

Vi indfører derfor en liste kaldet **målinger**. Ved at gå ned i formelfeltet og vælge **Datafangst > Manuel Datafangst** på **Data** menuen kan vi nu oprette en liste over alle vores målinger af teststørrelsen





Parameteren 0 i Capture-værktøjet angiver at der er tale om en manuel datafangst. Manuel datafangst foregår ved at taste **CTRL .** (dvs. Kontrol punktum). Af hensyn til den gentagne simulering skifter vi nu tilbage til det **Lister og regneark**-værksted, hvor simuleringen udføres. Derefter taster vi først **CTRL R** for at gentage simuleringen og dernæst **CTRL .** for at få fanget målingen – og sådan fortsætter vi. Det sker i praksis ved at holde **CTRL**-tasten nede og først taste **R** og dernæst **.** osv. Vi ser da målingerne dukke op i **Diagrammer og statistik**-værkstedet. Her har vi også plottet værdien **chi2_obs**, dvs. den observerede teststørrelse. Vi ser da at vi i de første 20 målinger har ramt lige så skævt i 7 af tilfældene. Dermed er vores skøn over p-værdien baseret på de førte 20 målinger givet ved 7/20 = 35%.

Det ville nu være rart at kunne holde automatisk styr på antallet af målinger, ligesom det ville være rart at få udregnet vores bud på p-værdien helt automatisk. Vi finder antallet af målinger ved hjælp af celleformlen

= count(**målinger**).

Tilsvarende kan vi tælle antallet af skæve målinger ved hjælp af celleformlen

= countIf(**målinger,?≥chi2_obs**).

Endelig finder vi et bud på p-værdien, dvs. sandsynligheden for at det simulerede udfald er mindst lige så ekstremt som det observerede udfald, ved at dividere antallet af skæve med det

samlede antal målinger. Denne gang refererer vi direkte til de foregående målinger i formlen, dvs. vi peger på cellen med værdien for antallet af skæve og dividerer derefter med værdien for det samlede antal målinger ved at pege på cellen. Til sidst ganger vi med 100. (hvor decimal-punktummet sikrer, at det skrives som decimaltal).

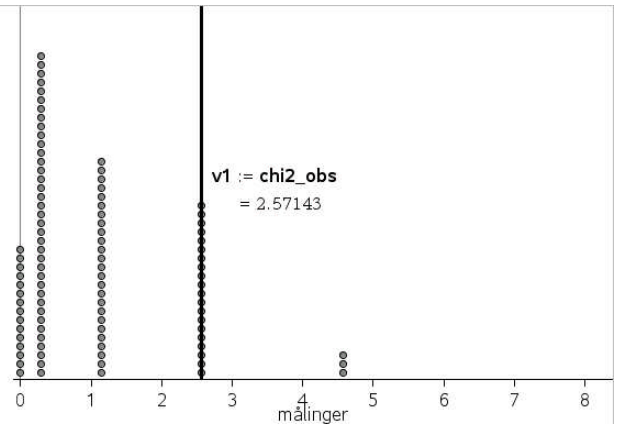
	s_hyp	forv_hyp	sim_hyp		målinger
◆					=capture('chi2_sim,0)
1	10	7	4	chi2_obs =	0.285714
2	4	7	10	**2.57143**	2.57143
3				chi2_sim =	2.57143
4				**2.57143**	0.285714
5				Antal målinger =	0.285714
6				20	2.57143
7				Antal skæve =	2.57143
8				7	2.57143
9				p-værdi =	0.
10				35.	4.57143

H10 $= \frac{h8}{h6} \cdot 100.$

Nu er 20 målinger ikke ret mange, så vi kan ikke regne med at p-værdien har stabiliseret sig endnu. Industristandarden er 500 målinger, så vi fortsætter ufortrødent. Vi ser da at p-værdien lige så stille falder til et lavere niveau. Efter 100 målinger ser det således ud:

	s_hyp	forv_hyp	sim_hyp		målinger
◆					=capture('chi2_sim,0)
1	10	7	8	chi2_obs =	0.285714
2	4	7	6	**2.57143**	2.57143
3				chi2_sim =	2.57143
4				**0.285714**	0.285714
5				Antal målinger =	0.285714
6				100	2.57143
7				Antal skæve =	2.57143
8				23	2.57143
9				p-værdi =	0.
10				23.	4.57143
11					0.285714

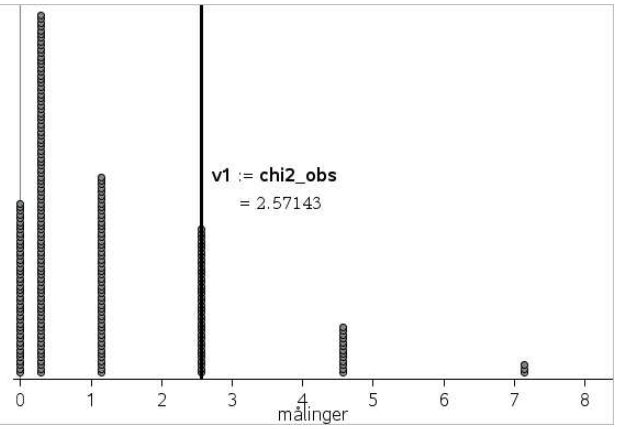
F4



Efter 250 målinger ser det således ud (men så orker man næsten heller ikke at lave flere)

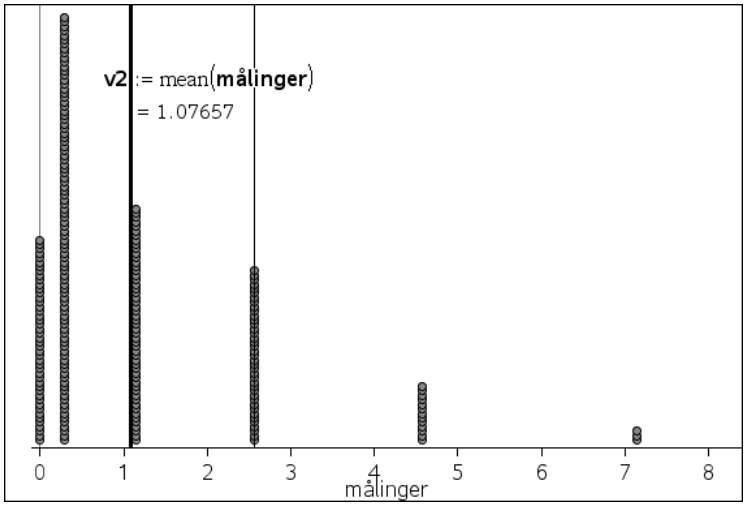
	s_hyp	forv_hyp	sim_hyp		målinger
◆					=capture('chi2_sim,0)
1	10	7	6	chi2_obs =	0.285714
2	4	7	8	**2.57143**	2.57143
3				chi2_sim =	2.57143
4				**0.285714**	0.285714
5				Antal målinger =	0.285714
6				250	2.57143
7				Antal skæve =	2.57143
8				55	2.57143
9				p-værdi =	0.
10				22.	4.57143
11					0.285714

F4

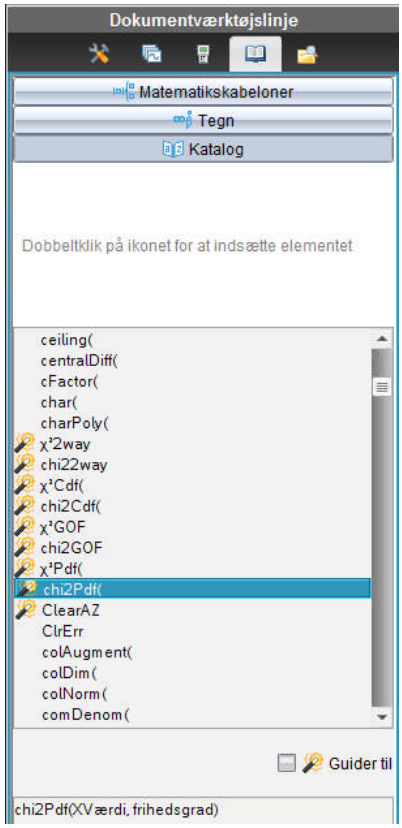


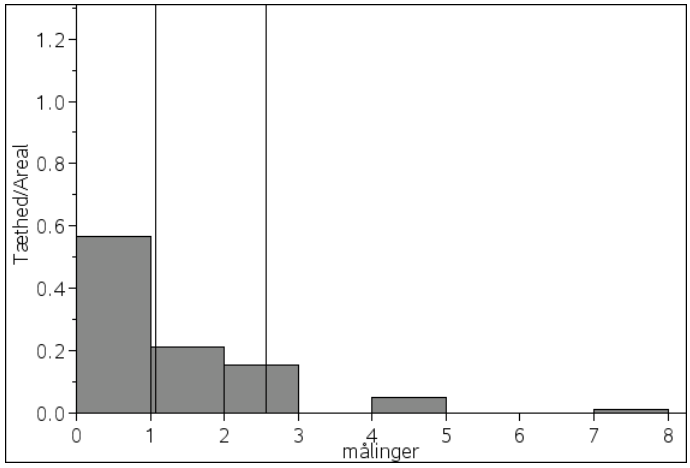
P-værdien ligger til sidst på ca. 22%, så når man simulerer nulhypotesen er sandsynligheden for at få et udfald, der er mindst lige så ekstremt som det observerede givet ved ca. 22%. Det er ikke særligt sjældent, nulhypotesen er altså rimeligt troværdig. Med et signifikansniveau på 5% er resultatet derfor *ikke* statistisk signifikant.

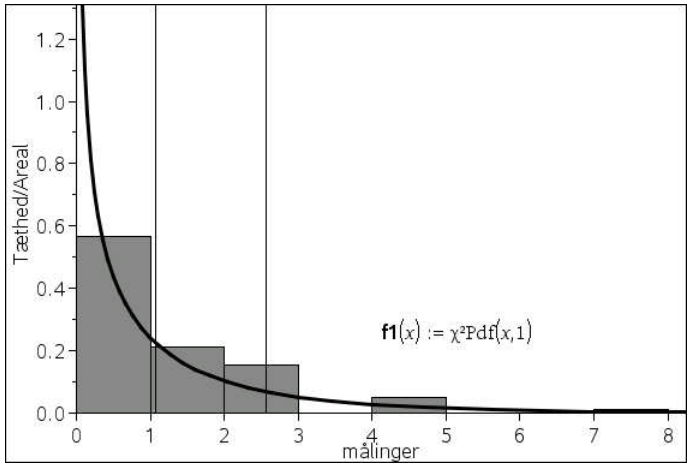
7.4 Den teoretiske stikprøvefordeling

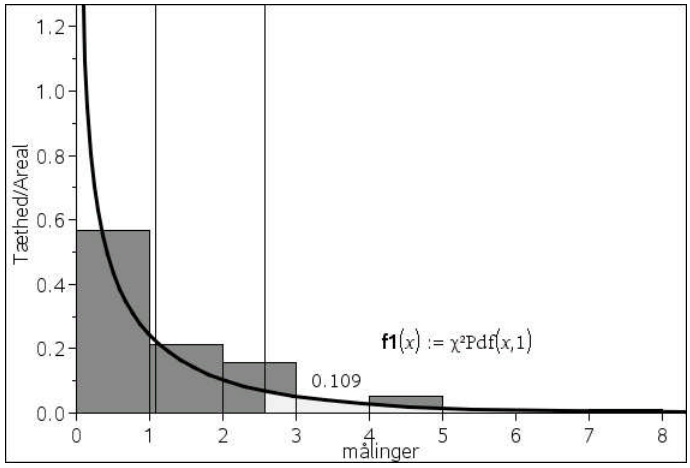


For at kunne sammenligne de fundne eksperimentelle resultater med en mere teoretisk tilgang får vi i det følgende brug for begrebet **frihedsgrad**. Der er **to kategorier**: venstre- og højre-handsker, men når vi skal fordele de 14 handsker på de to kategorier er der kun **én frihedsgrad**: Så snart vi har valgt et antal venstrehandsker, kender vi jo antallet af højrehandsker. Plotter vi middelværdien for teststørrelsen ser vi netop at middelværdien ligger meget tæt på 1. Dette er en af de væsentlige begrundelser for at dividere med den forventede værdi i teststørrelsen. Derved sikres det netop at den forventede middelværdi svarer til antallet af frihedsgrader. Hvis vi nu den forventede værdi i teststørrelsen. Derved sikres det netop at den forventede middelværdi svarer til antallet af frihedsgrader. Hvis vi nu ønsker at kigge nærmere på stikprøvefordelingen for teststørrelsen kan vi omdanne histogrammet til et *tæthedshistogram* (hvor det samlede areal er 1) ved dels at højreklikke og ændre **skalaen** til **tæthed/areal**, dels gå ind i **søjleindstillinger** og sætte **bredden** til 1, så vi får frembragt et sammenhængende histogram. Men det kan vi nu sammenligne med den teoretiske tæthedsfunktion for chi2-fordelingen, der hedder chi2Pdf(x,1), fordi der er 1 frihedsgrad (idet Pdf er en forkortelse for PointDistributionFunction). Som det ses, følger histogrammet pænt den teoretiske fordeling. Når man udfører en teoretisk beregning af chi2-fordelingen vil man typisk støtte sig til tæthedsfordelingen chi2Pdf(x,df), hvor df er antallet af frihedsgrader (Degrees of Freedom)









Vi kan endda bruge den teoretiske fordeling til at skønne over p-værdien. Vi kan fx højreklikke på grafen og vælge **Skraver under funktion**. Vi skal da først klikke på den observerede teststørrelse og derefter på boksen med +∞. Resultatet ligger noget lavere, 10.9%, hvilket skyldes at den teoretiske stikprøvefordeling ikke er så god en tilnærmelse til den faktiske stikprøvefordeling i dette tilfælde (se evt.næste afsnit). Men resultatet er selvfølgelig stadigvæk ikke statistisk signifikant, så konklusionen er uændret.
Man kan godt korrigere for det teoretiske skøn, så det bliver lidt mere præcist, men det er omstændeligt! Resultatet var altså slet ikke så overraskende som vi i første omgang troede. Det samme resultat kom Politiken også frem til idet de næste dag bragte følgende kommentar til historien:

Slet ikke så sært endda

BJARNE SCHILLING

Vi beskrev i går, hvordan Jammerbugt-strandvandrerparret Steffensen blev slået med forbløffelse, da de gjorde årets arbejdshandskehøst op: 9 venstre- og kun 3 højrehandsker. Tager man de to handsker med, som udgør et egentligt par, er fordelingen 10 venstre og 4 højre.
»Men det er faktisk ikke så bemærkelsesværdigt«, skriver Bendix Carstensen, der som seniorstatistiker på Steno Diabetes Center så afgjort må siges at have forstand på tal.
»Hvis der på stranden ligger præcis lige mange højre- og venstrehandsker, og vi sender 100 indsamlingshold ud og beder dem om at indsamle 12 enkelthandsker hver, så er der ca. 14 af holdene, der vil få 3 eller færre af den ene slags«, skriver han.
»Hvis vi i stedet sender holdene ud for at samle 14 handsker hver, vil omkring 18 af de 100 hold finde 4 eller færre af den ene slags. Formelt set skal man formulere det, som at der er 17,96 procent sandsynlighed for at finde 4 eller færre af den ene slags handsker, hvis der indsamles 14 fra en strand, hvor der er tilfældigt fordelt lige mange af begge slags (og der i øvrigt er uendeligt mange handsker til rådighed ...) «, påpeger Bendix Carstensen og går videre med noget, der hedder »binomialfordelingen«, og så er det, at vi står af. Her er heller ikke mere plads.
bjarne.schilling@pol.dk

Antal målinger =
2500
Antal skæve =
460
p−værdi =
18.4

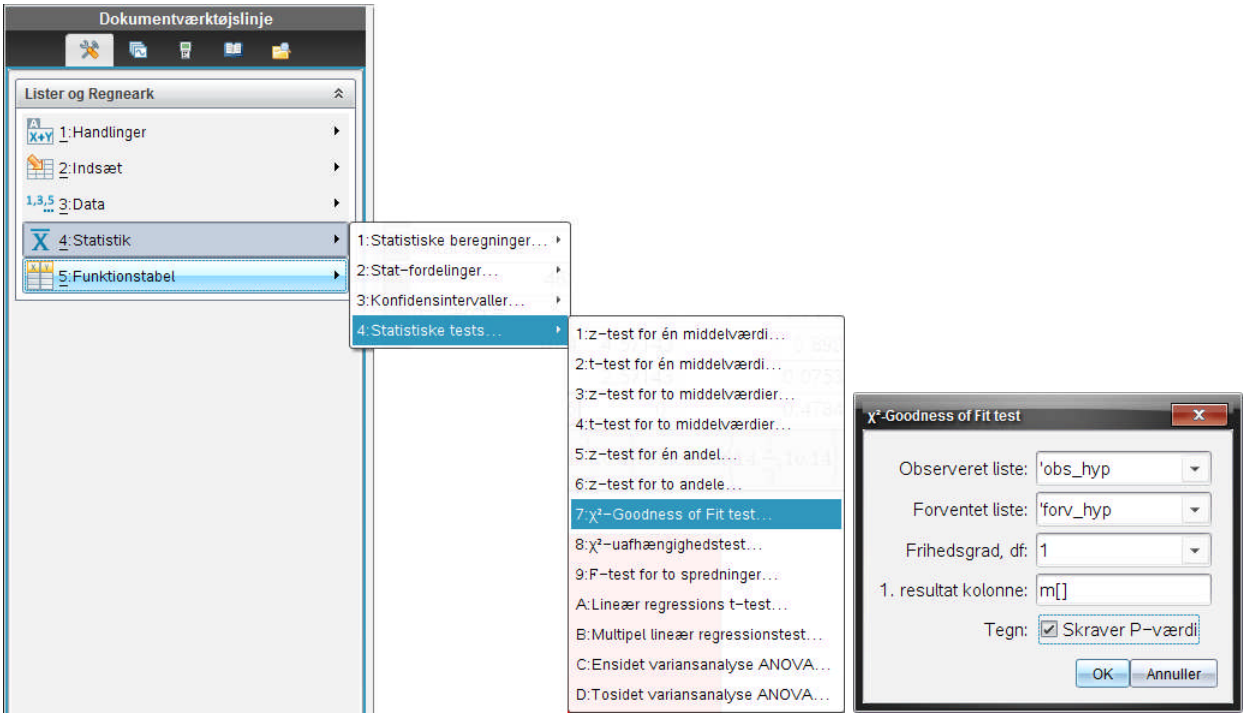
I artiklen kommer seniorstatistikeren Bendix Carstensen frem til at der er sandsynligheden 17,96% for at finde et resultat, der er mindst lige så ekstremt, hvis man simulerer nulhypotesen ved at sende 100 hold ud for at finde 14 tilfældige handsker på stranden. Det afviger lidt fra vores hidtidige resultat for p-værdien, men det skyldes i det væsentlige at vi ikke har simuleret tilstrækkeligt mange gange. Gennemfører vi i stedet 2500 enkeltmålinger, der er det maksimalt mulige antal målinger i **Lister og Regneark**-værkstedet når vi samler målingerne i en enkelt liste, fås i stedet en p-værdi på 17.6%. Denne gang ligger vi altså meget tættere på artiklens værdi. Man kan godt regne sig frem til den eksakte p-værdi, men som antydet i artiklen kræver det et godt kendskab til binomialfordelingen.

11	Binomialtest
12	0.179565
H12	=binomcdf(14,1/2,0,4)+binomcdf(14,1/2,10,14)

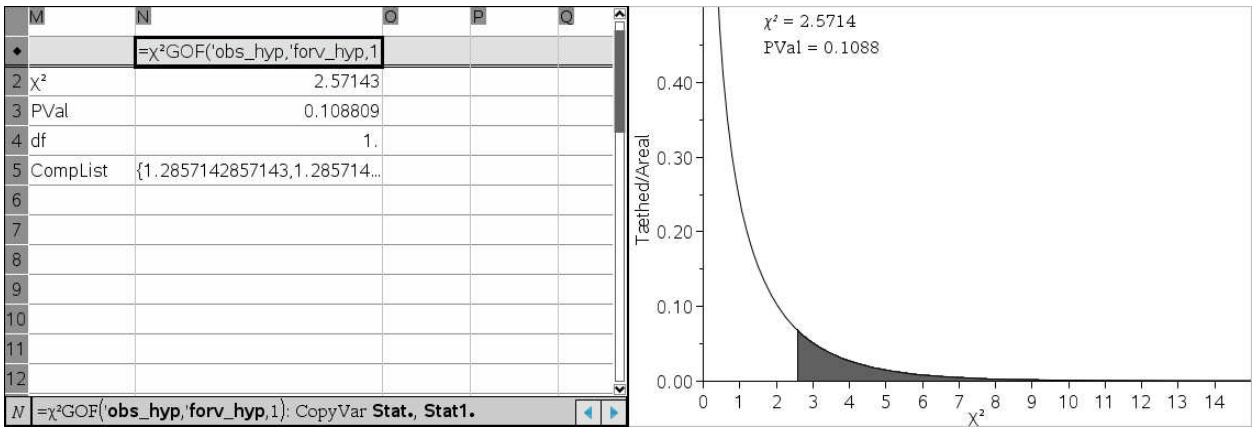
I modsætning til binomialtestet er chi2-testet kun et approksimativt test og det fører derfor kun til et – forhåbentligt – rimeligt skøn over p-værdien.

7.5 Chi2-testet for Goodness-of-Fit som et kanonisk test

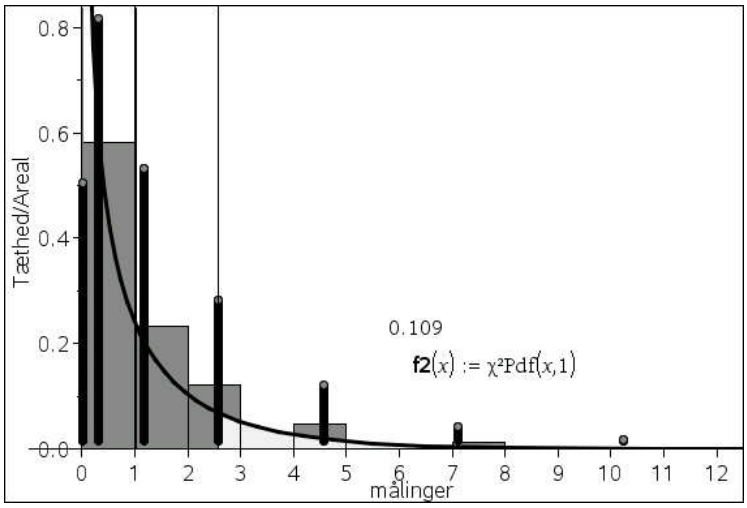
Ti-Nspire CAS har også indbygget chi2-testet som et kanonisk test kaldet **Goodness-of-Fit**-testet χ^2 GOF, fordi det tester hvorvidt et givet kategorisk datasæt ser ud til at følge den forventede fordeling. I **Lister og Regneark**-værkstedet hentes det fra **Statistik**-menuen:



Der åbnes da som vist for en dialogboks, hvor man skal angive listen over de observerede hyppigheder, listen over de forventede hyppigheder, antallet af frihedsgrader, samt hvor man vil have anbragt resultaterne i **Lister og Regneark**-værkstedet. Endelig har man som en ekstra bonus mulighed for at sætte kryds ved **Skraver P-værdi**. Der opbygges da automatisk en grafisk illustration af testen med stikprøvefordelingen og en skravering af området for ekstreme udfald, dvs. udfald, der ligger mindst lige så langt ude som den observerede teststørrelse. I vores tilfælde har vi allerede opbygget lister over såvel de observerede hyppigheder som de forventede hyppigheder inde i selve regnearket. Vi kan derfor referere direkte til disse lister i dialogboksen, hvor vi vælger dem på rullegardin-menuerne. Men vi kan også indsætte dem direkte ved hjælp af Tuborg-parenteser, dvs. den observerede liste som {10,4} og den forventede liste som {7,7}.



Vi får da oplyst teststørrelsen 2.57143, p-værdien 10.88%, antallet af frihedsgrader df = 1 (som vi selv har indtastet), samt Comp-listen over de enkelte kategoriers bidrag til teststørrelsen. Dertil kommer den grafiske illustration, hvor vi også får gentaget teststørrelsen og p-værdien.



I dette tilfælde er p-værdien på 10.88% tydeligvis for lav. Det skyldes chi2-fordelings-approksimationen, der ikke tager hensyn til den meget grynede natur af målingerne. Prøver vi i stedet at finde p-værdien svarende til 9 venstrehandsker finder vi i stedet:

	M	N	O	P	Q
◆		=χ²GOF({9,5},{7,7},1) : CopyV			
1	Titel	χ²–Goodness of Fit test			
2	χ²	1.14286			
3	PVal	0.285049			
4	df	1.			
5	CompList	{0.57142857142857,0.57142...			
6					
7					
8					
9					
10					
11					

N =χ²GOF({9,5},{7,7},1) : CopyVar **Stat., Stat1.**

Denne gang rammer vi for højt med en p-værdi på 28.50%. Den 'rigtige p-værdi' svarende til at antallet af succeser er givet ved 9.5 (som ikke accepteres af programmet!) er da ca. givet ved gennemsnittet (28.50%+10.88%)/2 = 19.69%, hvilket er rimeligt tæt på det korrekte resultat!

7.6 Automatisk datafangst på den forsigtige måde*

Det følgende afsnit er lidt teknisk og kan roligt overspringes i en første gennemlæsning. Hvis vi gerne vil udføre rigtigt mange målinger skal vi slå automatisk datafangst til i stedet for den manuelle datafangst. Her skal vi imidlertid være forsigtige: Den manuelle datafangst udfører en datafangst, hver gang vi taster **CTRL .**, men den automatiske datafangst kræver ikke at vi taster noget bestemt. Den overvåger variablen og registrerer hver gang den ændrer værdi. Den opfører sig altså ligesom min gamle jagthund, der var helt vild med at løbe efter vildt. Men hvis vildtet kunne holde sig i ro skete der ingen ting. Det var kun når vildtet løb, dvs. når der skete ændringer, at hunden satte efter vildtet. Når vi udfører en automatisk datafangst gentager vi simuleringen ved at holde **CTRL R** nede, men en ny simulering fører ikke nødvendigvis til en ny værdi for teststørrelsen. Vi mister altså nogle målinger svarende til gentagelserne af teststørrelsen, hvis vi ikke tænker os om! Da det er sværere at gentage de ekstreme målinger, fordi de forekommer mere sjældent, overvurderer vi derfor størrelsen af p-værdien, hvis vi ikke korrigerer for gentagelserne. Tilsvarende opfører manuel datafangst lidt ligesom en velopdragen jagthund. Den rør sig kun når den får lov, dvs. den skal have en apport-kommando, der altså i **TI-Nspire CAS** hedder **CTRL punktum** .

Når vi udfører en automatisk datafangst og vil gardere os mod at miste målinger, fordi teststørrelsen tilfældigvis giver det samme resultat, går vi derfor frem på følgende forsigtige måde: Sammen med teststørrelsen fanger vi en dummyvariabel givet ved celleformlen

dummy:= rand(),

som vi kan være sikker på skifter værdi hver gang vi genberegner regnearket. Endelig fanger vi også *summen* af vores teststørrelse **chi2_sim** og dummyvariablen. Den kaldes **dummysum** og skal bruges til at fange gentagelserne af værdien for teststørrelsen! Det ser sådan ud:

	H	I målinger	J	K test	L testsum
◆		='testsum−'test		=capture('dummy,1)	=capture('dummysum,1)
1	chi2_obs =	0.285714	dummy =	0.837441	1.12316
2	**2.57143**	2.57143	**0.280419**	0.041443	2.61287
3	chi2_sim =	0.285714	dummysum =	0.868792	1.15451
4	**0.**	0.285714	**0.280419**	0.166584	0.452298
5	Antal målinger =	2.57143		0.587199	3.15863
6	2500	0.285714		0.743856	1.02957
7	Antal skæve =	0.		0.725335	0.725335
8	460	0.		0.582025	0.582025
9	p−værdi =	0.		0.747263	0.747263
10	18.4	4.57143		0.89979	5.47122
11	Binomialtest	2.57143		0.075351	2.64678
J4	**dummysum**:=*h4+j2*				

Holder vi **CTRL R** nede starter målingerne for fuld udblæsning, men efter ca. 500 målinger taber maskinen pusten og det begynder at gå noget langsommere. Man kan stoppe op undervejs og taste enkeltvis for at bringe antallet af kontroller op på fx 1000 kontrollerede målinger. Men man kan også køre igennem til den bitre ende. Efter 2500 målinger stopper målingerne da, da der kun er plads til 2500 rækker i regnearket. Hvis man vil have flere målinger må man da oprette en ny søjle og fortsætte med denne. I dette tilfælde fører 2500 målinger til en p-værdi på 17.56%, hvilket er rimeligt tæt på seniorstatistikerens værdi på 17.96%. Vi kan altså styre den automatiske datafangst på den ovenfor beskrevne måde, hvor vi fanger tre variable: **chi2_sim**, **dummy** (:= rand()) og **dummysum** (= **chi2_sim** + **dummy**)! Værdierne for de to sidste variable gemmes i listerne **test** og **test_sum**. Endelig får vi vores kontrollerede testværdi tilbage ved at trække dummyvariablen fra igen, dvs. ved at udregne listen

målinger = **test_sum** - **test**

Det er lidt besværligt, og heldigvis findes der mange statistiske test, hvor teststørrelsen varierer så bredt at vi kan ignorere problemet med gentagelserne. Er vi i tvivl kan vi altid tilføje dummyvariablen og se hvor mangelfuld den simple ukontrollerede måling er. Når vi både måler på den simulerede teststørrelse **chi2_sim** og dummy-variablen rand(), så kan vi jo umiddelbart tælle dem og forskellen i de to antal viser da netop hvor mange gentagelser, der må have været.

8 Er piger venstreorienterede? χ^2-test for uafhængighed

Susanne Højte og Lars Donatzky fra Gladsakse Gymnasium har med hjælp fra deres elever fremstillet en både underholdende og instruktiv You-tube film om chi2-testen. Det anbefales kraftigt at se filmen ☺. Her vil vi analysere problemstillingen ved hjælp af **TI-Nspire CAS**. I filmen er eleverne Sidsel og Rasmus uenige om, hvorvidt piger er mere venstreorienterede end drenge. De bliver enige om at afgøre deres uenighed ved hjælp af en hypotesetest og aftaler derfor at spørge eleverne på skolen om hvad de ville stemme på, hvis der var folketingsvalg: Rød blok eller blå blok. De udspørger 60 elever, hvoraf 35 er piger og 25 er drenge. Blandt disse 60 elever ville 33 stemme rødt og 27 stemme blåt. Endvidere oplyses det, at 25 af de 33, der stemte rødt, var piger.

På basis af disse oplysninger kan vi nu udfylde en krydstabel for de observerede hyppigheder. Men da vi kender totalerne er det nu som vist nemt at finde de resterende tal:

Holdning\Køn	Pige	Dreng	I alt
Rød blok	25		33
Blå blok			27
I alt	35	25	60

Holdning\Køn	Pige	Dreng	I alt
Rød blok	25	**8** = 33-25	33
Blå blok	**10** = 35-25	**17** = 27-10 = 25-8	27
I alt	35	25	60

Det ser altså ud til pigerne foretrækker rød blok, mens drengene foretrækker blå blok. Men kan det nu virkeligt tages til indtægt for at piger er mere venstreorienterede end drenge? Kunne det ikke lige så godt forklares som et resultat af de tilfældige udsving i opinionsundersøgelser, der jo bygger på 'tilfældige stikprøver' i 'befolkningen' (hvor vi først må acceptere at klasserne på Gladsakse Gymnasium rent faktisk er repræsentative for den danske ungdom eller i det mindste den Københavnske gymnasieungdom ☺)? Vi kan derfor opstille en nulhypotese H_0, der alene tilskriver forskellen de uundgåelige fluktuationer i stemmetallene hos de tilfældigt udvalgte adspurgte

> H_0: Fordelingen hos de to køn afspejler den samme fælles fordeling af de forventede stemmer på Rød blok og Blå blok.

Sandsynligheden for at finde en vælger, der vil stemme på fx Rød blok er ifølge nulhypotesen derfor den samme for begge kønnene. Ikke desto mindre har vi en observation, hvor rød blok klart dominerer hos pigerne, samtidigt med at blå blok dominerer hos drengene! Er det nu rimeligt set ud fra nulhypotesen? Virker den troværdig? Det er dette spørgsmål som den statistiske test skal forsøge at kaste lys over.

Som før ønsker vi nu at simulere nulhypotesen. Men denne gang kender vi ikke den fælles fordeling af stemmerne, så vi kan ikke bare gentage Goodness-of-fit-testen fra det forrige eksempel. I stedet bygger vi på den følgende fundamentale ide: Hver af de afgivne stemmer i spørgeskemaundersøgelsen afspejler dels hvilket køn den stammer fra, dels hvilken politisk blok, der stemmes på. Men ifølge nulhypotesen spiller det nu ikke nogen rolle, hvilket køn, der er tale om, når vi kigger på fordelingen af stemmerne på de to blokke. Vi siger derfor også at ifølge nulhypotesen er udfaldet af variablen **blok** *uafhængig* af udfaldet af variablen **køn**. Denne gang kaldes testen derfor et **uafhængighedstest**, idet vi reelt tester om udfaldet af hvilken blok, der stemmes på, *ikke afhænger* af hvilket køn, der er tale om.

8.1 Krydstabeller: Observerede versus forventede hyppigheder

Når man skal undersøge en eventuel sammenhæng mellem to kategoriske variable, i dette tilfælde **køn** og **blok** opskrives de fundne hyppigheder typisk i en *krydstabel* i regnearket (i regneark som Excel kaldes krydstabeller for *pivottabeller*). Derefter benyttes celleformler til at til at udregne rækketotaler, søjletotaler og den samlede tabeltotal

B4	=sum(b2:b3)

D2	=sum(b2:c2)

Det er kun disse to formler, der indskrives. Resten fås da ved at 'trække' formlerne på langs eller tværs af regnearket ved at klikke i cellen med formlen og derefter trække i ankeret i nederste højre hjørne:

	A	B	C	D
1	obs_tabel	Pige	Dreng	I alt
2	Rød blok	25	8	
3	Blå blok	10	17	
4	I alt	35		
5				

	A	B	C	D
1	obs_tabel	Pige	Dreng	I alt
2	Rød blok	25	8	33
3	Blå blok	10	17	27
4	I alt	35	25	60
5				

Vi har nu frembragt en tabel over de *observerede hyppigheder*. Af denne tabel aflæses fx det følgende skøn over den fælles fordeling: Ifølge nulhypotesen er det ligegyldigt hvilket køn, der ligger til grund for tildelingen af stemmer til blokkene. Vi kan derfor lige så godt slå de to køn sammen og ser da at ud af de 60 afgivne stemmer er de 33 tildelt Rød blok og de resterende 27 er tildelt blå blok. Det er da ifølge nulhypotesen vores bedste bud på hvordan den underliggende fælles fordeling for tilslutningen til blokkene ser ud. Det giver nu anledning til at vi kan udregne de forventede stemmer ud fra nulhypotesen. Vi forventer nemlig at 33/60 af stemmerne går til rød blok og de resterende 27/60 af stemmerne går til blå blok. Fx forventer vi derfor at 33/60·35 piger stemmer på Rød blok, da der er 35 piger med i undersøgelsen osv. Læg mærke til at *de forventede værdier godt kan være decimaltal*. De repræsenterer ikke et faktisk antal vælgere, men det *gennemsnitlige* antal vælgere i en meget lang række opinionsundersøgelser baseret på den samme fordeling, og et gennemsnit kan godt være et decimaltal.

Vi indskriver derfor den forventede fordeling i en ny krydstabel, der fremkommer ved at kopiere den allerede fundne krydstabel for de observerede hyppigheder

	A	B	C	D
6	forv_tabel	Pige	Dreng	I alt
7	Rød blok	19.25	13.75	33.
8	Blå blok	15.75	11.25	27.
9	I alt	35.	25.	60.

Det er ikke svært at finde de enkelte forventede værdier, men vi kan også systematisere udregningen med en celleformel

	A	B	C	D
1	obs_tabel	Pige	Dreng	I alt
2	Rød blok	25	8	33
3	Blå blok	10	17	27
4	I alt	35	25	60
5				
6	forv_tabel	Pige	Dreng	I alt
7	Rød blok	19.25	13.75	33.
8	Blå blok	15.75	11.25	27.
9	I alt	35.	25.	60.

B7	$=\frac{\$d2}{\$d\$4}\cdot b\$4\cdot 1.$

$$19.25 = \frac{33}{60}\cdot 35$$

Læg mærke til dollartegnene, der binder cellereferencen til en absolut reference. Når vi flytter rundt på celle B7, skal tælleren hele tiden hentes dels i søjlen for søjletotaler (D-søjlen), dels i rækken for række-totaler (række 4). Der er derfor sat dollartegn foran D og 4 for at binde disse to celler til at blive i søjlen og rækken for totalerne. Tilsvarende skal nævneren hele tiden hentes i celle D4, hvor både D og 4 i nævneren er bundet gennem et foranstillet dollartegn.

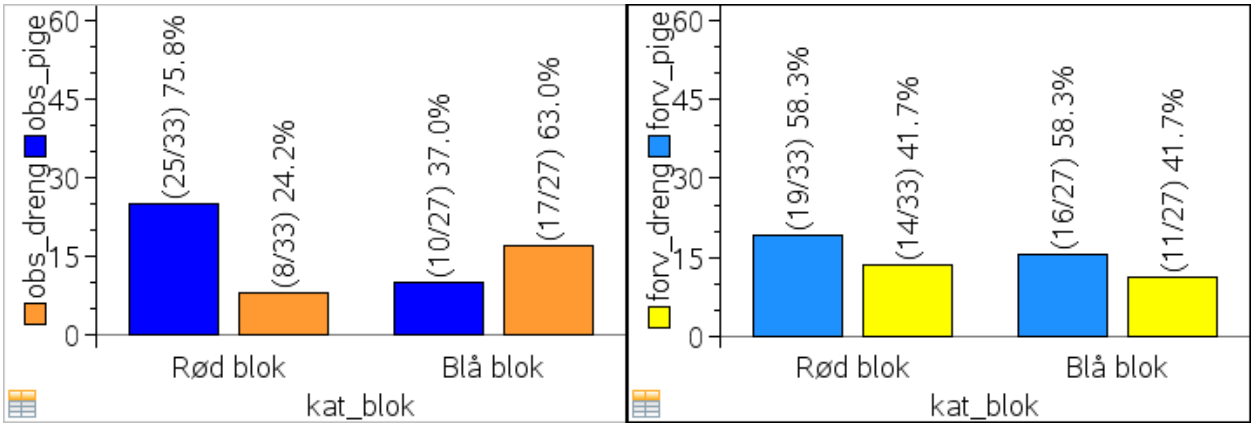
Hvis vi har gjort det rigtigt skal de forventede summer i række- og søjletotalerne nu netop give de samme værdier som i krydstabellen for de observerede hyppigheder.

Vi ønsker nu grafisk at illustrere de ovenstående krydstabeller. Dataene fra krydstabellerne skal da først overføres til navngivne *lister*, idet det kun er navngivne variable, der kan afbildes grafisk. Vi skal med andre ord have flyttet tabellen op over dobbeltstregen. Vi skal da oprette lister for kandidat-kategorierne og for de tilhørende hyppigheder, såvel de observerede som de forventede.

	A	B	C	D	E kat_blok	F obs_pi...	G obs_dr...	H forv_pi...	I forv_d...
◆						=b2:b3	=c2:c3	=b7:b8	=c7:c8
1	obs_tabel	Pige	Dreng	I alt	Rød blok...	25	8	19.25	13.75
2	Rød blok	25	8	33	Blå blok	10	17	15.75	11.25
3	Blå blok	10	17	27					
4	I alt	35	25	60					
5									
6	forv_tabel	Pige	Dreng	I alt					
7	Rød blok	19.25	13.75	33.					
8	Blå blok	15.75	11.25	27.					
9	I alt	35.	25.	60.					

Kategori-listen og hyppighedslisterne overføres som vist direkte fra krydstabellerne. Derefter kan vi netop frembringe kombinationsdiagrammer, hvor vi kobler blok-kategorierne med såvel de observerede hyppigheder som de forventede hyppigheder. Først afsættes **kat_blok** på førsteaksen. Derefter *højreklikkes* i aksefeltet på andenaksen og man vælger **Tilføj Y-værdiliste** for at oprette et søjlediagram baseret på **obs_pige**. Der højreklikkes endnu engang for at vælge **Tilføj Y-værdiliste** og denne gang tilføjes **obs_dreng**. Det samme gentages bagefter i et nyt **Diagrammer og statistik**-værksted, men denne gang med de forventede hyppigheder.

De forventede hyppigheder viser den forventede fordeling ifølge nulhypotesen, hvor der derfor ikke er forskel på de to køn, og hvor pigerne altså dominerer alene i kraft af deres overvægt i stikprøven. Dette er tydeligvis ikke tilfældet med de observerede hyppigheder, hvor pigerne dominerer rød blok, mens drengene tilsvarende dominerer blå blok. Men spørgsmålet er så om forskellen mellem den observerede fordeling og den forventede fordeling er stor nok til at den er statistisk signifikant? For at afgøre dette vender vi os derfor mod den bekræftende statistik. Vi skal da have en idé til hvordan man simulerer nulhypotesen om uafhængighed mellem opinionsundersøgelserne og stemmeafgivningen, dvs. vi skal have en ide til hvordan man kan simulere uafhængighed af to kategoriske variable.



8.2 Simulering af uafhængighed: Omrøring i data

Her kan vi nu tænke på dataindsamlingen som resulterende i en stak kartotekskort, hvor der på hvert enkelt kartotekskort dels er noteret hvilket køn, den stammer fra, dels hvilken blok, der er stemt på. I alt er der 60 sådanne kartotekskort. Hvis vi klipper dem over i to halvdele, vil venstre halvdel af kartotekskortene kun rumme oplysningen om kønnet, mens højre halvdel kun vil rumme oplysningen om blokken. I følge nulhypotesen er der nu ingen sammenhæng mellem opinionsundersøgelsen og stemmeafgivningen, dvs. kandidatfordelingen er uafhængig af opinionsundersøgelsen. Hvis vi derfor blander den venstre halvdel af kortene – den med kønnet – og derefter limer kartotekskortene sammen igen, får vi en fordeling, hvor der med sikkerhed ikke er nogen sammenhæng mellem kønnet og blokken. Vi simulerer med andre ord uafhængigheden mellem de to variable. Når man på denne måde adskiller to variable, blander den ene, og sætter dem sammen igen, siger man at man har foretaget en *omrøring*.

I praksis kan man udføre omrøringen som vist på filmen ved hjælp af flere sæt spillekort: Rasmus bestyrer blokkene, så han har et sæt spillekort med 33 røde kort for Rød blok og 27 sorte kort for Blå blok, dvs. i alt 60 spillekort. Sidsel bestyrer kønnene, så hun har et sæt spillekort med 35 røde kort for piger og 25 sorte kort for drenge, dvs. igen i alt 60 spillekort. Sidsel blander nu sine kort og de lægger derefter kortene på bordet ét for ét. Hver gang de begge lægger et rødt kort har de fundet en pige, der stemmer rødt osv. Faktisk er det nok at tælle hvor mange piger, der stemmer rødt, idet resten af tabellen udfylder sig selv. Det er en fin måde at få styr på simuleringen af nulhypotesen!

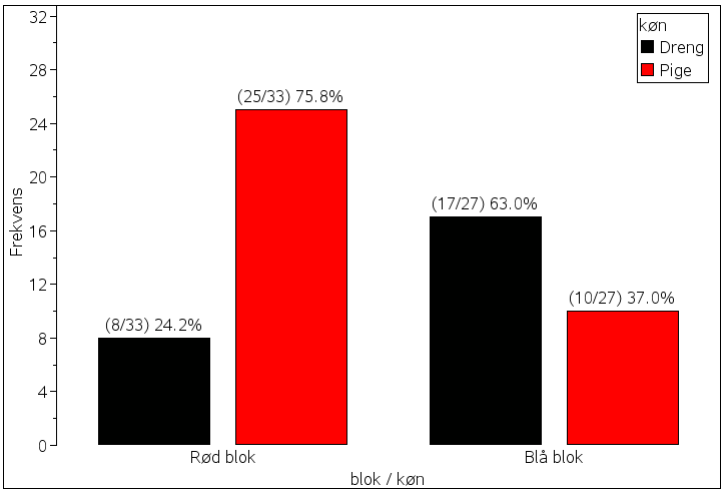
For nu at få foretaget omrøringen i **TI-Nspire CAS** skal vi først have frembragt "kartotekskortene", dvs. vi skal gå fra hyppighedstabeller til lister, der dels rummer kønnene, dels de blokke, der svarer til de afgivne stemmer. Vi skal med andre ord have rekonstrueret *rådata*. Det sker ved hjælp af listekommandoen **freqtable@>list**(), der skal have to argumenter, nemlig de to lister fra en hyppighedstabel: kategorierne og hyppighederne. Husk at hyppighed på engelsk hedder *frequency* og at hyppighedstabellen derfor hedder *frequency table*, heraf navnet på kommandoen. Husk også som vist at transformationssymbolet ▸ kan erstattes af @>. For at kunne anvende denne kommando effektivt skal vi først have opbygget to skyggetabeller, hvor tallene er erstattet af kategorierne for blokke henholdsvis køn, som kopieres ind i tabellen i stedet for tallene:

	A	B	C	D
◆				
11	obs_tabel	Pige	Dreng	I alt
12	Rød blok	Rød blok	Rød blok	2*"Rød b...
13	Blå blok	Blå blok	Blå blok	2*"Blå bl...
14	I alt	Rød blok"...	Rød blok...	2*"Rød b...
15				
16	obs_tabel	Pige	Dreng	I alt
17	Rød blok	Pige	Dreng	Pige"+"D...
18	Blå blok	Pige	Dreng	Pige"+"D...
19	I alt	2*"Pige"	2*"Dreng"	2*"Pige"...

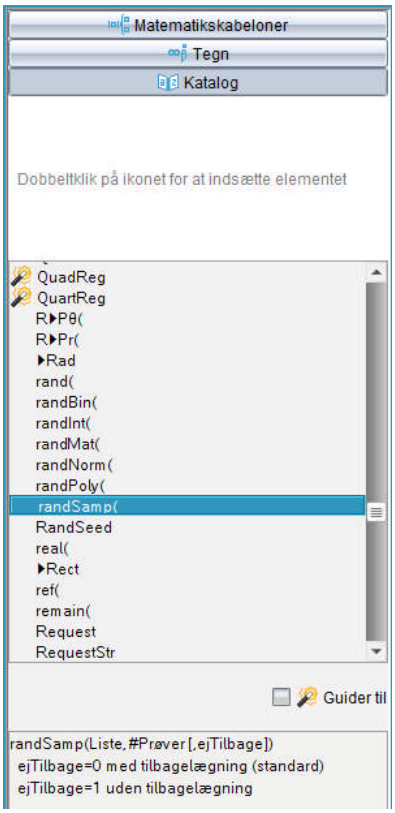
	J køn	K blok
◆	=freqtable▸list(b17:c18,b2:c3)	=freqtable▸list(b12:c13,b2:c3)
1	Pige	Rød blok
2	Pige	Rød blok
3	Pige	Rød blok
4	Pige	Rød blok
5	Pige	Rød blok
6	Pige	Rød blok
7	Pige	Rød blok
8	Pige	Rød blok
9	Pige	Rød blok

De to lister, vi på den måde får frembragt, indeholder da netop de to halvdele af vores 60 "kartotekskort": Listen for **køn** indeholder først 25 kartotekskort med Pige (der er koblet til Rød blok), derefter 10 kartotekskort med Pige, der er koblet til Blå blok osv. og tilsvarende for listen for **blok**. Vi kan afbilde kartotekskortene i et grupperet søjlediagram ved først at afbilde **blok** på førsteaksen og derefter højreklikke i aksefeltet for førsteaksen og vælge **Opdel kategorier ef-**

ter variabel, hvorved vi får splittet den kategoriske variabel **blok** i forhold til den kategoriske variabel **køn**:



Det fremkomne diagram minder meget om kombinationsdiagrammet, men denne gang er det frembragt direkte fra de rekonstruerede rå-data. Der kan derfor godt være forskel på rækkefølgen af kategorierne langs aksen. Rå-diagrammet er ordnet alfanumerisk (dvs. alfabetisk, hvor tallene går forud for bogstaverne), men man kan godt flytte rundt på søjlerne i den overordnede variabel.



Vi skal nu have simuleret **nulhypotesen**, dvs. rørt rundt i den første liste med **køn**, for at bryde en eventuel sammenhæng mellem kønnet og stemmefordelingerne. Det sker ved hjælp af kommandoen

RandSamp(population, antal stikprøver [,type]),

som beskrevet i kataloget i det venstre sidepanel. Typeparameteren kan undværes: den afgør om stikprøven sker med eller uden tilbagelægning. Som standard sker den **med** tilbagelægning, dvs. vi trækker igen og igen fra den samme fordeling! Men for at frembringe en omrøring skal vi jo i stedet have foretaget en permutation af listens elementer, dvs. have sat elementerne sammen i en ny tilfældig rækkefølge. Det kræver at vi trækker *uden tilbagelægning*, dvs. først trækker vi det element, der nu skal være første element, derefter trækker vi det element, der nu skal være andet element og sådan fortsætter vi til der ikke er flere elementer tilbage at trække af. Vi sætter derfor parameteren til 1. Samtidigt skal vi sikre os at vi trækker i bund, dvs. at der ikke er flere elementer tilbage. I dette tilfælde skal vi altså trække 60 gange. Men da dette antal varierer fra opgave til opgave er det bedre at skrive det som dim(**opinion**), hvor dim-kommandoen netop udregner listens længde, dvs. antallet af elementer i listen. Her er dim en forkortelse af ordet dimension. Se figuren næste side, hvor kommandoen også er skrevet helt ud.

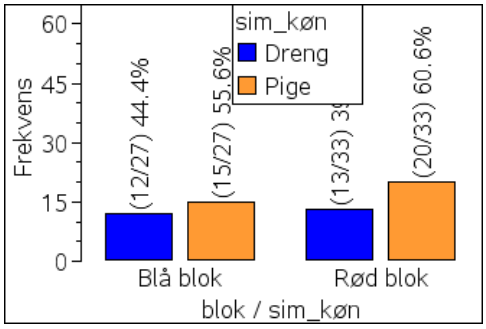
Taster vi **CTRL R** får vi nu netop gentaget den tilfældige omrøring, og dermed en ny version af en tilfældig kobling mellem opinionsundersøgelsen og stemmeafgivningen.

	at_blok	F obs_pi...	G obs_dr...	H forv_pi...	I forv_d...	J køn	K blok	L sim_køn
◆		=b2:b3	=c2:c3	forv_pige	=c7:c8	=freqtab	=freqtable	=randsam
1	d blok	25	8	19.25	13.75	Pige	Rød blok	Pige
2	blok	10	17	15.75	11.25	Pige	Rød blok	Pige
3						Pige	Rød blok	Pige
4						Pige	Rød blok	Pige
5						Pige	Rød blok	Pige
6						Pige	Rød blok	Dreng
7						Pige	Rød blok	Pige
8						Pige	Rød blok	Pige
9						Pige	Rød blok	Pige
10						Pige	Rød blok	Pige
11						Pige	Rød blok	Pige
12						Pige	Rød blok	Pige
13						Pige	Rød blok	Dreng

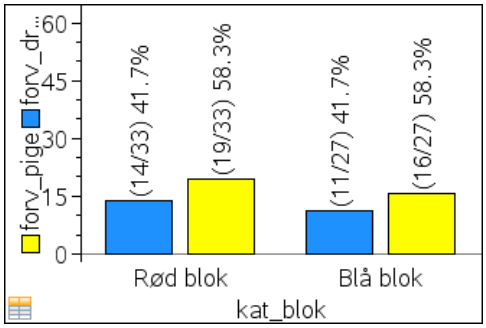
sim_køn:=randsamp(køn,dim(køn),1)

Men når vi har omrørt den kategoriske variable **køn**, så kan vi jo oprette en graf over fordelingen for det *omrørte køn* **sim_køn** koblet til fordelingen for **blok**. Denne graf anbringer som sagt kategorierne i alfanumerisk rækkefølge, så den kan afvige fra kombinationsdiagrammerne i rækkefølgen af kategorierne. Men da der er tale om et usammensat søjlediagram kan vi bare flytte rundt på søjlerne som det passer os, ligesom vi kan farvelægge søjlerne som det passer os!

Vi ser da at simuleringen typisk ligner den forventede fordeling (de to fordelinger har samme form og pigerne dominerer i begge kategorier i overensstemmelse med at der er flere piger end drenge i stikprøven). Dette er i modsætning til den observerede fordeling, hvor pigerne dominerer i den røde blok, mens drengene dominerer i den blå blok - på trods af at der er færre drenge i stikprøven. Gentages simuleringen mange gang kan man nu se hvor ofte det lykkes at frembringe en simulering, hvor der er mindst 25 piger der stemmer rødt.



Da jeg fx prøvede simuleringen igennem lykkedes det ikke for mig en eneste gang i tyve forsøg at få en fordeling, hvor 25 piger stemte på rød blok (men nok en enkelt gang, hvor 24 piger stemte rødt!). Noget tyder altså på at den observerede fordeling er relativ sjælden, dvs. ret svær at frembringe ud fra nulhypotesen og det svækker selvfølgelig tiltroen til denne! Men igen er det nødvendigt at præcisere, hvor sjældent det egentlig forekommer, dvs. udregne en p-værdi.



8.3 Hvor stor er afvigelsen: Opbygningen af teststørrelsen

Spørgsmålet er altså hvor meget forskel på de to køn man egentlig kan forvente i en sådan simulering. For at kunne svare på dette spørgsmål vil vi først oprette en automatisk optælling af de simulerede stemmer fordelt på de to **køn**. Det gøres ved hjælp af celleformler, hvor vi udnytter kommandoen Sum(iffn(betingelse på lister, hvis sand=1, hvis falsk=0)). Som betingelse på lister bruger vi da den boolske kombination **sim_køn** = "pige" and **blok** = "Rød blok", der netop rummer begge oplysninger. Hvis vi møder en position, hvor en simuleret pige stemmer rødt fås 1 og ellers 0. Til sidst lægger vi alle 1-tallerne sammen og finder netop antallet af gange, hvor piger stemmer rødt:

	A	B	C	D
◆				
20				
21	sim_tabel	Pige	Dreng	I alt
22	Rød blok	21	12	33
23	Blå blok	14	13	27
24	I alt	35	25	60

B22 =sum(iffn(**sim_køn**=*b*$21 and **blok**=$*a*22,1,0))

Læg mærke til dollartegnene, der sikrer at vi hele tiden mikser kategorier fra den øverste række $21 og den yderste venstre søjle $a.

Der er nu flere muligheder for at konstruere en passende teststørrelse, men i chi-i-anden-testen går man således frem: I første omgang udregner man forskellen på de observerede og forventede hyppigheder, derefter kvadrerer man forskellen og dividerer med de forventede værdier. Til sidst lægger man alle bidragene sammen:

$$\chi^2 = sum\left(\frac{(\text{observeret} - \text{forventet})^2}{\text{forventet}}\right)$$

$$= \frac{(21-19.25)^2}{19.25} + \frac{(14-15.75)^2}{15.75} + \frac{(12-13.75)^2}{13.75} + \frac{(13-11.25)^2}{11.25}$$

I begge tilfælde er det nu klart at jo mere de observerede antal afviger fra de forventede, jo større bliver teststørrelsen, dvs. den kan netop anvendes som passende mål for skævheden, hvor det er de store værdier, som er problematiske –men også små testværdier kan være problematiske! De optræder fx i tilfælde af videnskabeligt snyd, hvor en forsker har pyntet på resultaterne ved at lægge sine observationer alt for tæt på de forventede antal. Fx har der været rejst anklager om snyd med tvillingeundersøgelser, hvor det mest kendte og omdiskuterede eksempel handler om den engelske psykolog Cyril Burt.

Vi indskriver derfor Pearsons teststørrelse i en celleformel for sig selv, dvs. i en ikke-navngivet liste:

	J køn	K blok	L sim_køn	M
◆	=freqtab	=freqtable▸l	=randsam	
1	Pige	Rød blok	Pige	Chi2obs ...
2	Pige	Rød blok	Pige	**9.16017**
3	Pige	Rød blok	Pige	Chi2_sim..
4	Pige	Rød blok	Dreng	**0.848485**
5	Pige	Rød blok	Dreng	

M2 **chi2obs**:=sum$\left(\frac{(b2:c3-b7:c8)^2}{b7:c8}\right)$

Læg mærke til at vi henter de observerede og forventede hyppigheder direkte fra tabellerne.

Vi kan tilsvarende udregne teststørrelsen for de simulerede opinionsundersøgelser. Det giver mulighed for at gentage simuleringen ved at taste **CTRL-R** og se hvordan den simulerede teststørrelse varierer. Læg også mærke til at vi har lagret de to teststørrelser, så vi kan lade programmet holde udkig efter den. For at lagre dem markerer vi cellen **N2** henholdsvis **N4**, højre-

klikker i den og vælger menupunktet **Variable > Gem var** (eller taster **CTRL L** for Link til en variabel og vælger menupunktet **Gem VAR** eller går direkte ind i cellen og skriver selv **Chi2Obs :** med kolontegnet lige foran lighedstegnet). Læg mærke til at cellernes værdi skrives med fed, når de først er lagt på køl!

	J køn	K blok	L sim_køn	M
◆	=freqtab	=freqtable▸l	=randsam	
1	Pige	Rød blok	Pige	Chi2obs …
2	Pige	Rød blok	Pige	**9.16017**
3	Pige	Rød blok	Pige	Chi2_sim..
4	Pige	Rød blok	Dreng	**0.848485**
5	Pige	Rød blok	Dreng	

M4 **chi2_sim**:=sum$\left(\frac{(b22:c23-b7:c8)^2}{b7:c8}\right)$

Som det ses er den observerede teststørrelse **Chi2Obs** meget større end den vi får fra simuleringen; og selv om vi gentager simuleringen mange gange er det faktisk svært at finde en simuleret teststørrelse, der rammer lige så skævt.

8.4 Datafangst: Den eksperimentelle stikprøvefordeling

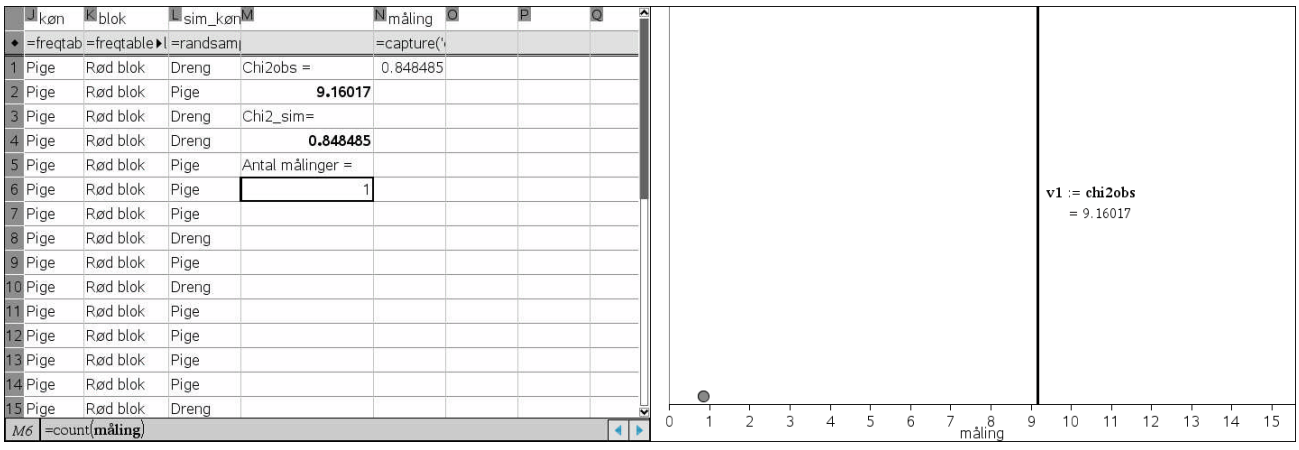
Dermed er jagten klar til at gå ind på den simulerede tesstørrelse. Vi opretter nu en ny kolonne med navnet **Måling** og markerer formelfeltet, hvorefter vi kan højreklikke i det eller blot vælge menupunktet **Datafangst > Automatisk** fra **Data**-menuen. Vi skal derefter navngive den variabel vi ønsker at fange og indskriver derfor variabelnavnet **chi2_sim** (eller henter det med **CTRL L**):

	J køn	K blok	L sim_køn	M	N måling
◆	=freqtab	=freqtable▸l	=randsam		=capture('
1	Pige	Rød blok	Dreng	Chi2obs =	0.848485
2	Pige	Rød blok	Pige	**9.16017**	
3	Pige	Rød blok	Dreng	Chi2_sim=	
4	Pige	Rød blok	Dreng	**0.848485**	
5	Pige	Rød blok	Pige		
6	Pige	Rød blok	Pige		

Vi vender så tilbage til kolonnen med teststørrelserne og indskriver endnu en celleformel

= count(**måling**)

til at finde antallet af målinger, så vi kan holde øje med målingerne. Tilsvarende opretter vi en graf for målingerne, hvor vi kan holde øje med hvorvidt de rammer indenfor eller udenfor den observerede teststørrelse **chi2_obs**, som indsættes ved at benytte kommandoen **Plot værdi** fra **Undersøg data**-menuen



Den første måling!

Derefter er det blot at holde **CTRL-R** tasten ned og vente tålmodigt! Til at begynde med vokser antallet af målinger jævnt, men efter et stykke tid begynder maskinen at hakke i det og den foretager nu målingerne i klumper. Med lidt held kan man stoppe målingerne ved ca. 1000 målinger. Men man kan også køre dem igennem til den bitre ende, dvs. det maksimale antal målinger, som er 2500, idet regnearket kun kan rumme 2500 rækker i en enkelt liste.

	blok	sim_køn	
	=freqtable▸l	=randsam(	
1	Rød blok	Dreng	Chi2obs =
2	Rød blok	Pige	9.16017
3	Rød blok	Dreng	Chi2_sim=
4	Rød blok	Pige	0.4329
5	Rød blok	Pige	Antal målinger =
6	Rød blok	Pige	20
7	Rød blok	Dreng	
8	Rød blok	Dreng	
9	Rød blok	Pige	
10	Rød blok	Dreng	
11	Rød blok	Pige	
12	Rød blok	Pige	
13	Rød blok	Dreng	
14	Rød blok	Pige	
15	Rød blok	Dreng	

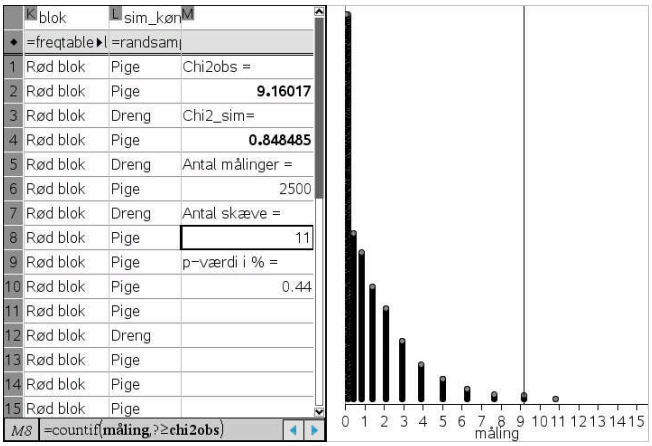
M6 =count(måling)

	blok	sim_køn	
	=freqtable▸l	=randsam(	
1	Rød blok	Pige	Chi2obs =
2	Rød blok	Pige	9.16017
3	Rød blok	Dreng	Chi2_sim=
4	Rød blok	Pige	0.848485
5	Rød blok	Dreng	Antal målinger =
6	Rød blok	Pige	2500
7	Rød blok	Dreng	
8	Rød blok	Pige	
9	Rød blok	Pige	
10	Rød blok	Pige	
11	Rød blok	Pige	
12	Rød blok	Dreng	
13	Rød blok	Pige	
14	Rød blok	Pige	
15	Rød blok	Pige	

M6 =count(måling)

De første tyve målinger *Resultaterne fra 2500 målinger*

Vi kan så også nemt udregne et skøn over **p-værdien**, dvs. hvor stor en procentdel, der har ramt mindst lige så skævt som opinionsundersøgelsen. På grafen ser det ud som om der er en 4-5 skæve målinger (den første prik dækker over flere prikker). En sikker optælling sker ”p-værdi i % =ved hjælp af **CountIf**-kommandoen. Der er faktisk 11 skæve målinger!



	blok	sim_køn	
	=freqtable▸l	=randsam(	
1	Rød blok	Pige	Chi2obs =
2	Rød blok	Pige	9.16017
3	Rød blok	Dreng	Chi2_sim=
4	Rød blok	Pige	0.848485
5	Rød blok	Dreng	Antal målinger =
6	Rød blok	Pige	2500
7	Rød blok	Dreng	Antal skæve =
8	Rød blok	Pige	11
9	Rød blok	Pige	p-værdi i % =
10	Rød blok	Pige	0.44
11	Rød blok	Pige	
12	Rød blok	Dreng	
13	Rød blok	Pige	
14	Rød blok	Pige	
15	Rød blok	Pige	

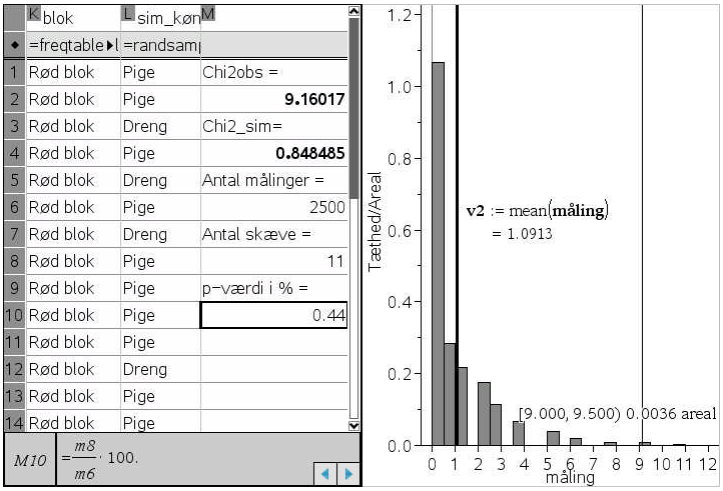
M8 =countif(måling,?≥chi2obs)

I dette tilfælde fås p-værdien 0.44%. Nulhypotesen har altså en meget lav troværdighed, langt under det almindeligt brugte signifikansniveau på 5%. Forskellen på de to opinionsundersøgelser er altså statistisk signifikant: Alt tyder på at pigerne i den Københavnske gymnasieungdom er venstreorienterede.

8.5 Den teoretiske stikprøvefordeling

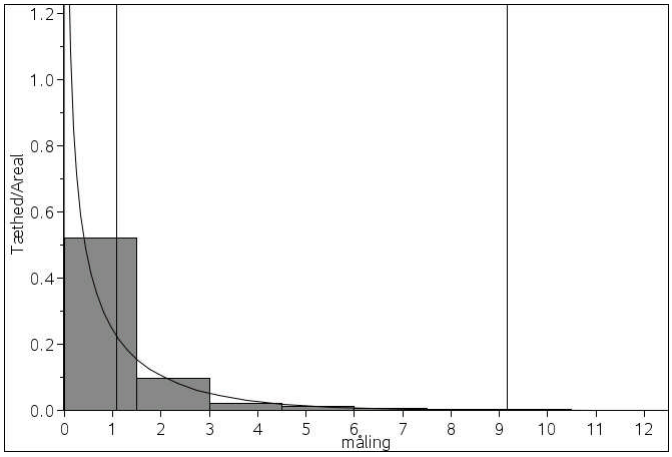
Vi kan få et endnu bedre indtryk af fordelingen, ved dels at konvertere den til et histogram, dels tilføje middelværdien for målingerne ved hjælp af menupunktet **Plot værdi** på **Undersøg**-menuen. Histogrammet omsættes til et tæthedshistogram ved at højreklikke og vælge menupunktet **Tæthed/Areal** på **Skala**-menuen. Vi lægger mærke til at middelværdien ligger meget tæt på 1. Det svarer netop til antallet af frihedsgrader i krydstabellen.

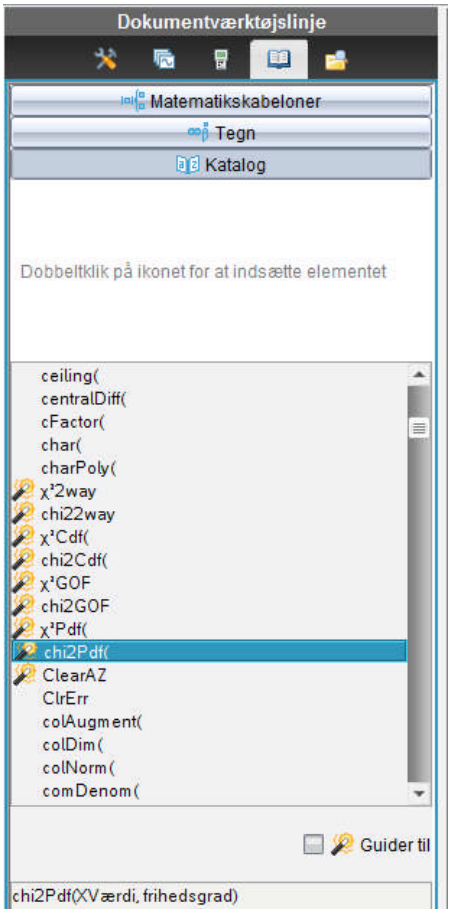
obs_tabel	Pige	Dreng	I alt
Rød blok	x	33−x	33
Blå blok	35−x	x−8	27
I alt	35	25	60



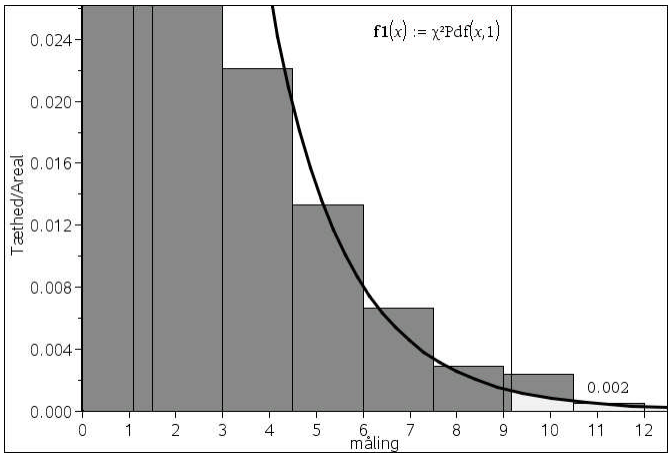
Krydstabellen har netop én frihedsgrad fordi vi kan vælge et af stemmetallene frit, her pigernes stemmer på rød blok angivet som x i den blå celle. Resten følger da som vist af søjletotalerne og rækketotalerne. Det svarer helt til vinklerne i en trekant, hvor der er to frihedsgrader, fordi den tredje vinkel følger af vinkelsummen på 180°.

Vi kan endda sammenligne stikprøvefordelingen med den teoretiske fordeling, idet vi vælger menupunktet **Plot funktion** på **Undersøg data**-menuen og vælger tæthedsfunktionen for chi-i-anden-fordelingen med 1 frihedsgrad fra kataloget (husk at vælge punktfordelingen, dvs. den der ender på Pdf = Point Distribution Function):





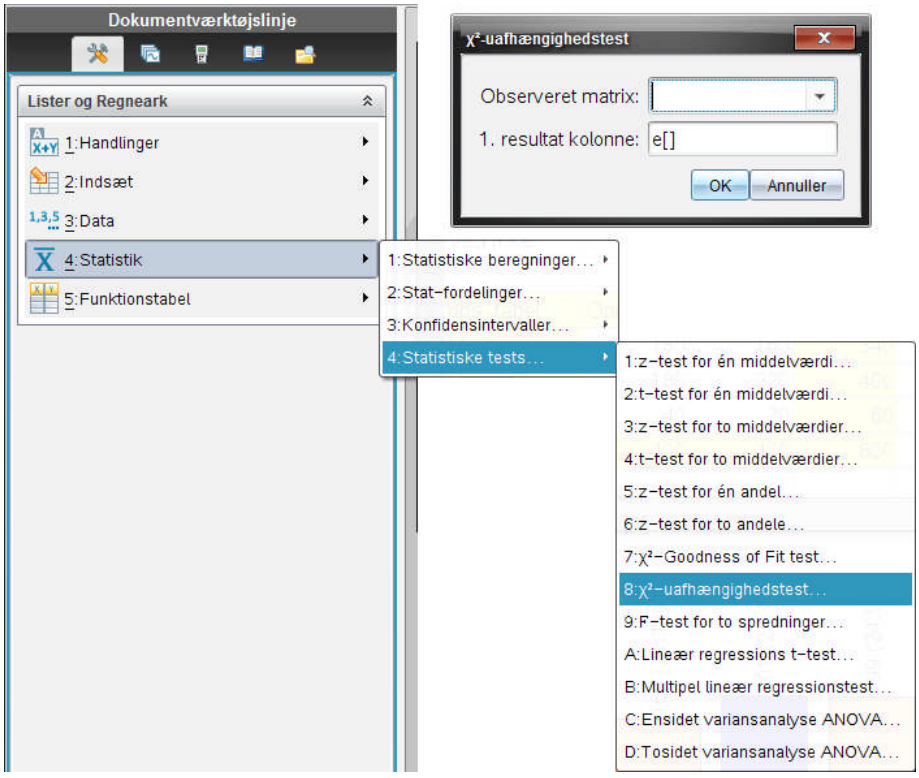
Som det ses følger den teoretiske fordeling meget fint tæthedshistogrammet for stikprøvefordelingen. Vi kan så også finde den teoretiske p-værdi. Vi skal da integrere, dvs. finde arealet under grafen for den teoretiske tæthedsfordeling, fra **chi2_obs**-værdien til uendelig. Det sker fx ved at højreklikke på grafen og vælge menupunktet **Skraver under funktion**. Derefter klikker vi først på den lodrette linje for chi2-obs-værdien, derefter på kassen med den øvre grænse givet ved $+\infty$. Hvis du ikke kan se resultatet skal du trække lidt i skalaen, fordi resultatet gemmer sig helt nede ved skalaen, da området er så snævert.



Vi ser altså at den teoretiske p-værdi er givet ved 0.003 = 0.3%, hvilket som lovet ligger et godt stykke under signifikansniveauet på 5 procent!

8.6 Chi2-testet for uafhængighed som en kanonisk test

Endelig vil vi til sidst se på hvordan vi kunne have udført testen som en teoretisk test. Denne gang skal vi vælge en χ^2-uafhængighedstest. Vi får da besked om at oplyse en *matrix* for de observerede værdier



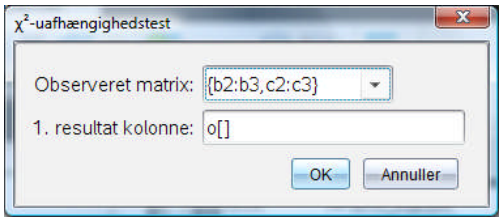
En matrix svarer nu i princippet direkte til en krydstabel (hvor vi kun fokuserer på tabellens værdier ordnet i rækker og søjler), men desværre kan **Lister og regneark**-værkstedet ikke arbejde direkte med matricer. Vi er derfor nødt til at gå omvejen over lister, og udnytte at *en matrix kan repræsenteres som en liste af lister*, hvor de inderste lister står for søjlerne i matricen. Fx vil den følgende liste af lister

	A	B	C	D
◆				
1	obs_tabel	Pige	Dreng	I alt
2	Rød blok	25	8	33
3	Blå blok	10	17	27
4	I alt	35	25	60

$$\{b2:b3,c2:c3\} = \{ \{25,10\},\{8,17\} \}$$

repræsentere matricen bestående af de observerede hyppigheder, dvs. netop kerneområdet i tabellen for de observerede hyppigheder. Men de tilsvarende lister er også oprettet som obs_pige og obs_dreng, hvorfor man også kan skrive matricen som

{obs_pige,obs_dreng}

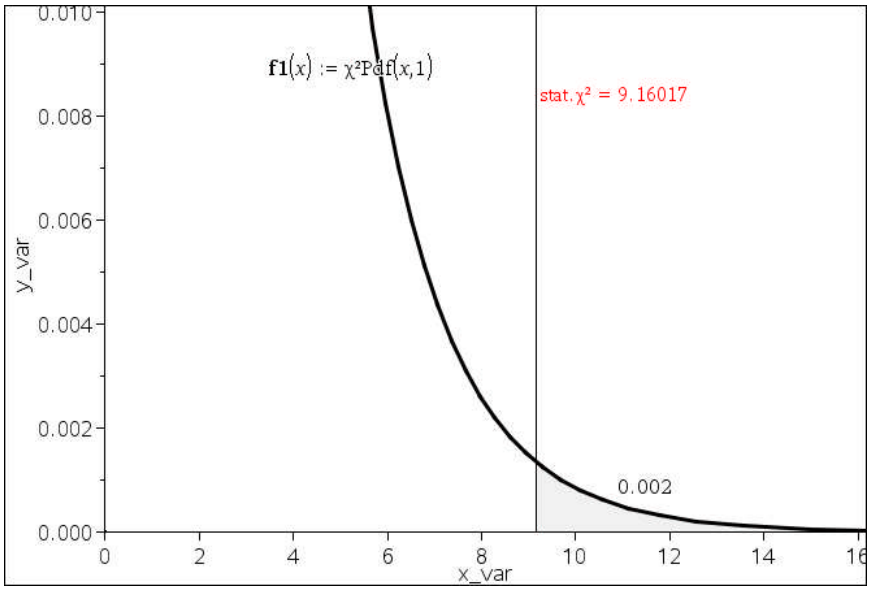


O	P
	=χ²2way({b2:b3,c2:c3}):
Titel	χ²-uafhængighedstest
χ²	9.16017
PVal	0.002473
df	1.
ExpMatrix	[[19.25,15.75][13.75,1...
CompMatrix	[[1.7175324675325,2.0...

Vi får nu oplyst χ^2-værdien, p-værdien, antallet af frihedsgrader (df = degrees of freedom) samt de komplette matricer for de forventede værdier såvel som de enkelte kategoriers bidrag til chi2-værdien (så vi kan se om der er kategorier, der i særlig grad bidrager til chi2-værdien og derfor pådrager sig særlig opmærksomhed, hvis vi som her finder at klart skævt resultat).

Denne gang får vi desværre ikke automatisk opbygget en graf over stikprøvefordelingen, dvs. tæthedsfordelingen for chi-i-anden-fordelingen med 1 frihedsgrad. Den må vi selv tegne i et **Diagrammer og statistik**-værksted. For at frembringe et koordinatsystem får vi da brug for at indføre to tomme lister: **x_var**:={} og **y_var** := {}. Den teoretiske chi2-værdi frembringes ved hjælp af **Plot værdi**- kommandoen fra **Undersøg data**-menuen, hvor vi plotter stat.χ^2. Endelig finder vi p-værdien grafisk ved at højreklikke på grafen og vælge kommandoen **Skraver under funktion**

S x_var	T y_var



Endelig kan vi se på de forventede værdier såvel som bidragene fra de enkelte kategorier: Da **Lister og regneark**-værkstedet som sagt ikke understøtter matricer direkte er vi nødt til først at konvertere dem tilbage til lister for at få lov til at se dem. Det sker ved hjælp af kommandoen **mat@>list**(matrix). Vi skal da huske at fjerne gåseøjnene omkring matricerne i testresultatet! Det sker nemmeste ved hjælp af expr()-kommandoen, der omdanner en tekst til et matematisk udtryk (expression). Vi må endvidere pege direkte på cellerne da **Lister og regneark**-værkstedet ikke accepterer stat.-variablen direkte!

forventet	bidrag
=mat▸list(expr(p5))	=mat▸list(expr(p6))
19.25	1.71753
15.75	2.09921
13.75	2.40455
11.25	2.93889

Her skal vi så huske på opbygningen af tabellen:

obs_tabel	Pige	Dreng
Rød blok	25	8
Blå blok	10	17

De to første elementer kommer fra den første søjle og de to næste fra den anden søjle i tabellen. Det er altså især de blå drenge, der skæpper i kassen, hvilket egentlig ikke er så overraskende, da der er flest piger i spørgeskemaundersøgelsen, og derfor forventer vi også flere piger end drenge i den røde blok. Men tallene er ikke meget markante, så man skal nok være varsom med at drage for vidtgående slutninger - der er jo kun én frihedsgrad, så det giver under alle omstændigheder ikke megen mening at udpege en synder.

Dermed kan vi altså med lidt øvelse hurtigt gennemføre en teoretisk chi-i-anden test for uafhængighed og drage de fornødne slutninger: Skal vi fortsat tro på nulhypotesen eller er resultatet så skævt, at vi må forkaste den og forlade os på en alternativ hypotese. Nøgletallet er da p-værdien som et mål for nulhypotesens troværdighed: Hvis den er meget lille, dvs. ligger under signifikansniveauet forkastes nulhypotesen og ellers accepteres den.

9 Spørgeskemaanalyser og χ^2-test

UGEBREVET A4 Arbejdsmarked | Politik | Velfærd | Værdier

10-04 06 NR. 14 | OM A4 | KONTAKT OS | ARKIV | ABONNEMENT | ABOUT A4 | SØG

DENNE UGE
Leder
Nyheder og aktuelle tal

Print
Tip en ven

Om samme emne:
Bøder skal forhindre sort byggeri

SORT ARBEJDE I EKSPLOSIV VÆKST BLANDT DE UNGE

Mens resten af samfundet er begyndt at arbejde mindre sort, har de unge for alvor smøget ærmerne op. Ifølge en ny undersøgelse fra Rockwoolfonden om sort arbejde har 46 procent af de helt unge arbejdet sort inden for det seneste år, og for dem i 20'erne gælder det 30 procent. De unge står dermed for en langt større stigning i det sorte arbejde end kvinderne, som ellers har fået det største fokus i kølvandet på den nye undersøgelse. Men hvad er der galt med de unge? Er de fuldstændig amoralske og skruppelløse, eller er det nød, der driver dem ud på det sorte arbejdsmarked?

Skat for Unge

Et frækt jobtilbud

Via et opslag i supermarkeder og på uddannelsessteder, blev man som en del af kampagnen opfordret til at ringe på et 'sort' jobtilbud, men når man gjorde det, ventede der ikke et jobtilbud, men en sjov telefonsvarer, der fortalte lidt om, hvad man som ansat mister ved at arbejde sort; feriepenge, arbejdsskadeforsikring, overarbejdsbetaling og så videre. Ingen af dem, der ringede op, blev dog registreret.

Vil du tjene lidt ekstra?

Så har vi et job med god løn.

Du kan tjene op til 150 kr. i timen (sort).

Ring til Peter K. på 80 60 18 92 og hør mere.

Ring 12 56 23 56 | Ring 80 60 18 92 | Ring 80 60 18 92 | Ring 80 60 18 92 | Ring 80 60 18 92 | Ring 80 60 18 92 | Ring 80 60 18 92 | Ring 80 60 18 92 | Ring 80 60 18 92 | Ring 80 60 18 92 | Ring 80 60 18 92 | Ring 80 60 18 92 | Ring 80 60 18 92 | Ring 12 56 23 56

Som et større eksempel på anvendelse af kategoriske data vil vi nu se nærmere på analysen af en spørgeskemaundersøgelse. Vi skal da have valgt et tema og en målgruppe. Her vil vi anvende et spørgeskema om **gråzonekriminalitet** udarbejdet til brug for et tværfagligt samarbejde mellem samfundsfag og matematik[1]

Spørgsmål til gråzonekriminalitet:

- Køn
- Alder

Har du uden tilladelse inden for det sidste år:

- Taget en cykel som ikke tilhørte dig selv
- Taget penge eller spiritus fra en jævnaldrende
- Taget penge eller spiritus fra forældre
- Taget noget i en forretning

- Har du indenfor det sidste år arbejdet sort
- Har du indenfor det sidste år ødelagt noget for sjov
- Har du inden for det sidste år ulovligt lavet graffiti

[1] Eksemplet er udarbejdet af Marianne Kesselhahn i forbindelse med det paradigmatiske eksempel 209: **Kriminaliteten i tal – et forløb i statistik**.

Dette spørgeskema kan nu sendes rundt til fx alle eleverne i 1g på skolen. Jo flere elever man kan inddrage i undersøgelsen jo bedre. På den måde får man en **stikprøve** af den danske gymnasieungdom. Man bør så overveje i hvilket omfang den er *repræsentativ*? Den er jo i hvert fald ikke fremkommet ved at foretage et tilfældigt udvalg af danske gymnasieelever - hvis fx undersøgelsen stammer fra et provinsgymnasium er det ikke sikkert den også dækker gymnasielever fra storbyerne. Stammer den fra et alment gymnasium er det ikke sikkert den også dækker elever fra handelsgymnasier (hhx) og tekniske gymnasier (stx) osv. Her vil blot notere os at undersøgelsen må formodes at være repræsentativ for en større population end de adspurgte.

For at kunne diskutere teknikker til at analysere spørgeskemaerne vil vi nu bruge resultaterne fra en autentisk undersøgelse i en 2g-klasse. Disse data kan også hentes på TI's danske hjemmeside **Besøg Education.ti.com** fra **Hjælp**-menuen. Det er en ret lille undersøgelse, der kun involverer 24 elever og derfor kun kan betragtes som en pilotundersøgelse. Men den vil kunne bruges til at illustrere metoderne, som så kan anvendes på den rigtige undersøgelse. De anførte svar på den anonyme undersøgelse indtastes i **TI-Nspire CAS**. Ved større undersøgelser gøres dette af flere omgange. Det er da afgørende at man først opretter en fælles skabelon, så alle variablene får nøjagtigt de samme navne i de enkelte delundersøgelser. Derefter kan man nemlig samle de enkelte delundersøgelser oprettet i hver sin **TI-Nspire CAS**-fil til én samlet undersøgelse ved at kopiere resultaterne og indsætte dem i ét samlet datasæt.

Man kan også gennemføre spørgeskemaundersøgelsen i fx Lectio. De samlede resultater kan da overføres til et Excel-regneark og derfra kopieres over i **TI-Nspire CAS**. Der er da tre ting man især skal være opmærksomme på:

1) Der er *ikke* grænser for hvor mange variable man kan håndtere i **TI-Nspire CAS**. Men det enkelte **Lister og regneark**-værksted kan kun håndtere op til 26 variable ad gangen, fordi der kun er 26 søjler i rengearket svarende til de 26 bogstaver i det engelske alfabet. Involverer undersøgelsen flere end 26 variable må man altså oprette flere **Lister og regneark**-værksteder. Da al kommunikation ud af regnearkene foregår via navngivne variable skal man blot sørge for at de enkelte variable i de forskellige regneark har forskellige navne.

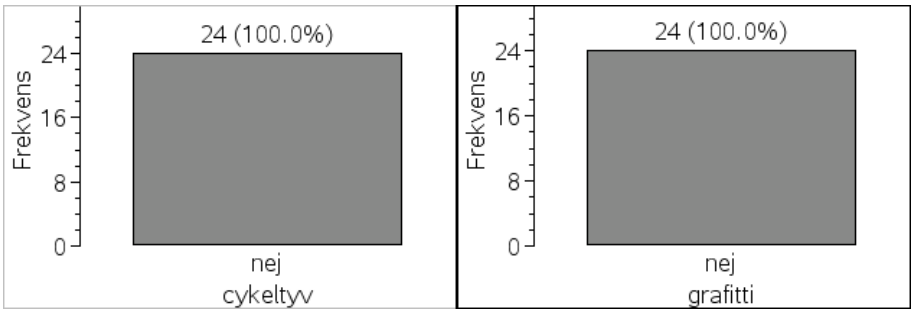
2) Der er *ikke* grænser for hvor mange data en variabel kan indeholde i **TI-Nspire CAS**. Men i **Lister og regneark**-værkstedet er der kun 2500 rækker. Hvis en liste indeholder mere end 2500 data må den derfor oprettes i et andet værksted, fx **Noter**-værkstedet. Man kan eventuelt kopiere listen ind af flere omgange i regnearket og så siden hen slå listerne sammen i **Noter**-værkstedet ved hjælp af sammenkædningskommandoen augment().

3) Der er *ikke* grænser for hvor store datamængder man kan operere med i **TI-Nspire CAS**, men der er grænser for hvor store datamængder man kan overføre via klippebordet. Når man kopierer et datasæt og vil sætte det ind i et **TI-Nspire CAS Lister og regneark**-værksted kan man derfor risikere at få en fejlmeddelelse om at datasættet fylder for meget. I så fald må man blot overføre det ad flere omgange.

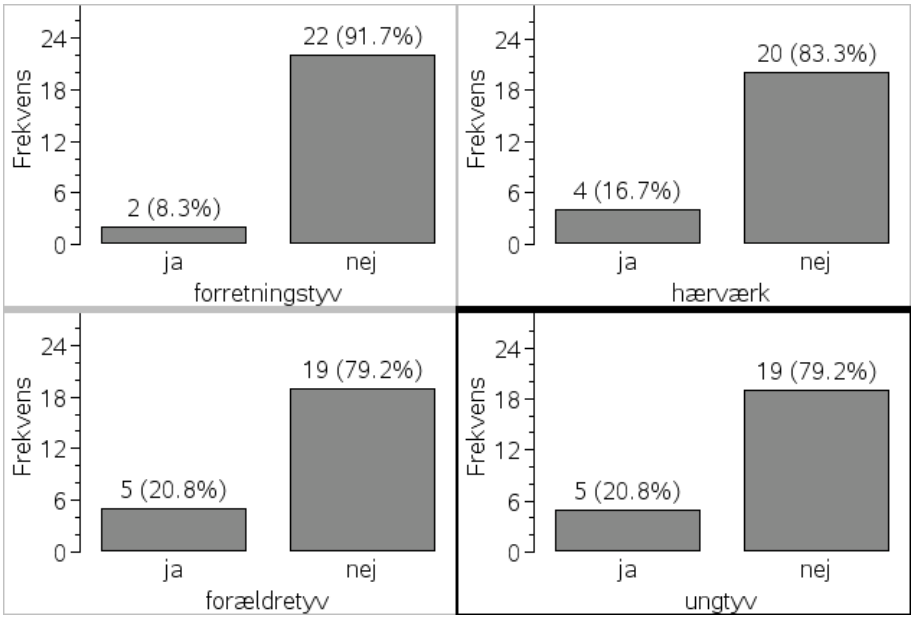
	A køn	B alder	C cykeltyv	D ungtyv	E forældretyv	F forretningstyv	G sortarbejde	H hærværk	I grafitti
◆									
1	dreng	17	nej	nej	ja	ja	nej	ja	nej
2	pige	17	nej	nej	nej	nej	nej	nej	nej
3	pige	17	nej	nej	nej	nej	ja	nej	nej
4	pige	17	nej	nej	nej	nej	ja	nej	nej
5	pige	17	nej	nej	nej	nej	nej	nej	nej
6	pige	17	nej	ja	nej	nej	nej	nej	nej
7	pige	17	nej	nej	nej	nej	nej	nej	nej
8	pige	17	nej	nej	nej	nej	nej	nej	nej
9	pige	17	nej	ja	ja	nej	ja	nej	nej
10	dreng	18	nej	nej	nej	nej	ja	nej	nej
11	dreng	18	nej	nej	ja	ja	ja	ja	nej
12	dreng	18	nej	ja	nej	nej	nej	ja	nej
13	dreng	18	nej	nej	nej	nej	ja	nej	nej
14	pige	18	nej	nej	ja	nej	ja	nej	nej
15	pige	18	nej	nej	nej	nej	ja	nej	nej
16	pige	18	nej	nej	nej	nej	ja	nej	nej
17	pige	18	nej	ja	nej	nej	nej	nej	nej
18	pige	18	nej	nej	nej	nej	ja	nej	nej
19	pige	18	nej	nej	nej	nej	ja	nej	nej
20	pige	18	nej	nej	nej	nej	nej	nej	nej
21	pige	18	nej	nej	nej	nej	ja	ja	nej
22	dreng	19	nej	ja	ja	nej	nej	nej	nej
23	dreng	19	nej	nej	nej	nej	ja	nej	nej
24	dreng	19	nej	nej	nej	nej	ja	nej	nej

9.1 På jagt efter sammenhænge: Søjlediagrammer

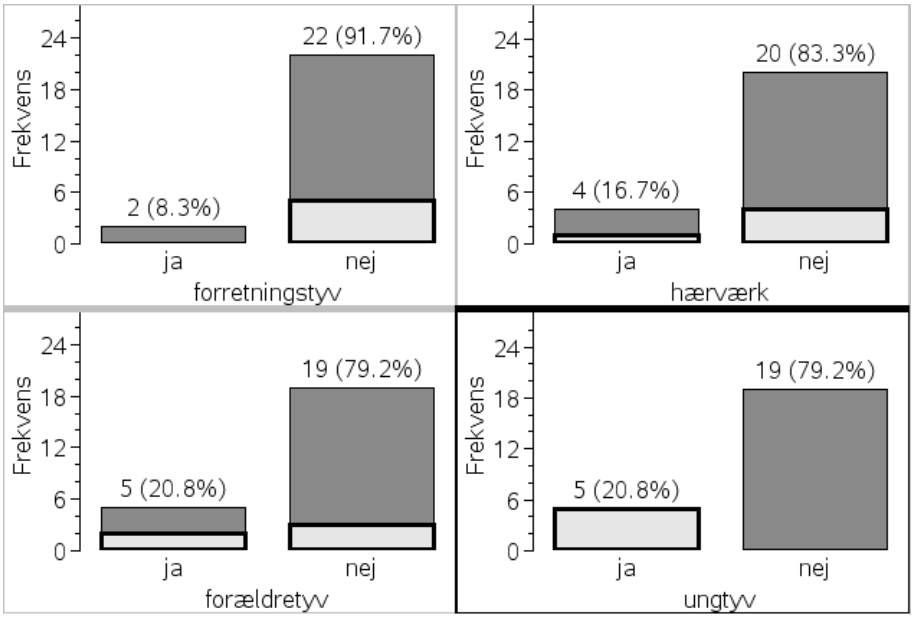
Ovenfor ses nu resultatet af undersøgelsen. De to første variable, **Køn** og **Alder**, vil vi opfatte som **uafhængige forklarende variable**. De syv sidste variable, **cykeltyv** ... **grafitti**, vil vi tilsvarende opfatte som **afhængige responsvariable**. Udover at kunne dokumentere omfanget af de forskellige former for gråzonekriminalitet er vi også interesserede i at undersøge om der kan konstateres en sammenhæng mellem køn og alder på den ene side og de forskellige typer gråzonekriminalitet på den anden side. Vi starter med de velkendte grafiske metoder fra den beskrivende statistik ('Vi skal tegne før vi kan regne'). Vi kan vælge mellem prikdiagrammer, søjlediagrammer og cirkeldiagrammer. Vi vælger søjlediagrammer. Det viser sig nu, at der ikke er nogen elever, der har *stjålet cykler* eller *tegnet graffiti*



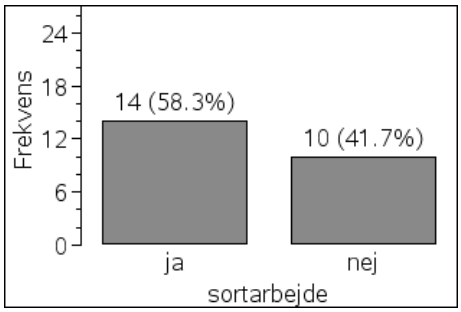
Tilsvarende er der kun et mindre, men stigende antal, der har *stjålet fra forretninger*, *begået hærværk* eller *stjålet fra forældre henholdsvis jævnaldrende*:



Vi kunne så umiddelbart tro at det er de samme fem elever, der både har stjålet fra jævnaldrende og fra forældrene, men hvis vi fx klikker i søjlen for dem, der har begået tyveri fra jævnaldrene (**ungtyv**), kan man umiddelbart se i alle de andre grafer, at det er ret tilfældigt om de også har begået andre typer af tyveri eller hærværk.



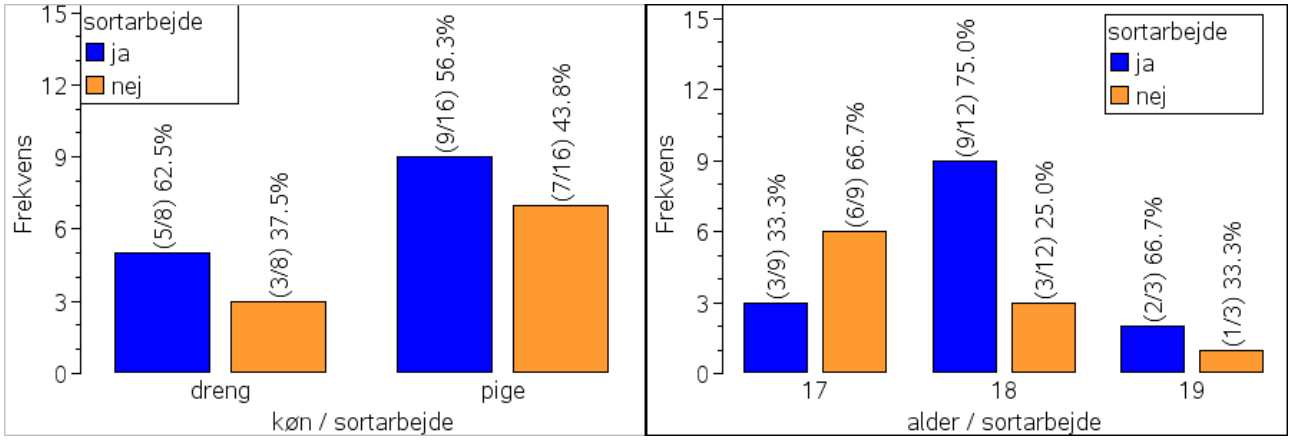
Endelig er der rigtigt mange, der har haft *sort arbejde*:



Bemærkning: Læg mærke til at vi i alle 7 grafer har brugt samme interval på hyppighedsaksen (der her betegnes frekvens i overensstemmelse med en udbredt sprogbrug i andre fag end matematik, hvor der ikke skelnes så nøje mellem frekvens og hyppighed);

For at undersøge mulige sammenhænge mellem en forklarende variabel og en responsvariabel skal vi have oprettet et grupperet søjlediagram ved at spille den forklarende variabel ved hjælp af responsvariablen. Vi afsætter derfor først den forklarende variabel og højreklikker derefter i aksefeltet og vælger **Opdel kategorier efter variabel**. Endelig skifter vi til søjlediagram og slår **Vis alle etiketter til**. Det fremkomne diagram svarer til oprettelsen af et pivot-diagram i Excel og rummer de samme informationer. Derimod understøtter **TI-Nspire CAS** ikke direkte pivottabeller, som må opbygges med håndkraft, hvad vi vender tilbage til.

Kigger vi nu først på sammenhængen mellem **sort arbejde** og **køn** ses en svag tendens til at drenge arbejder forholdsmæssigt mere sort end piger. Kigger vi derefter på sammenhængen mellem **alder** og **sort_arbejde** (hvor vi skal huske at højreklikke og tvinge alder til at være kategorisk!), tegner der sig der i mod en tydelig sammenhæng: Jo ældre eleverne bliver, jo flere laver sort arbejde: For de 17-årige er det kun ca. en tredjedel, der laver sort arbejde. For de 18- og 19-årige er det ca. to tredjedele.



Vi har nu ved hjælp af metoder fra den beskrivende statistik afdækket en mulig sammenhæng mellem alder og sort arbejde, der viser sig som en mulig tendens til at ældre elever laver mere sort arbejde end yngre. Men kunne denne forskel ikke lige så godt kunne forklares som et udslag af tilfældige variationer? Man kunne jo forestille sig at forskellen forsvandt, hvis vi udstrakte undersøgelsen til hele populationen af danske gymnasieelever. Vi har trods alt kun kigget på en mindre stikprøve, hvor forskellene måske kun skyldes de uundgåelige tilfældige variationer i små stikprøver. Vi indfører derfor den følgende nulhypotese

H_0: i virkeligheden afhænger det sorte arbejde slet ikke af alderen

Vi skal så i det følgende forsøge at vurdere hvor **troværdig** nulhypotesens påstand om en tilfældig variation er, dvs. om forskellen måske alligevel er så stor, at den må anses for at være statistisk signifikant og vi derfor må forkaste nulhypotesen.

9.2 Krydstabeller: Observerede versus forventede hyppigheder

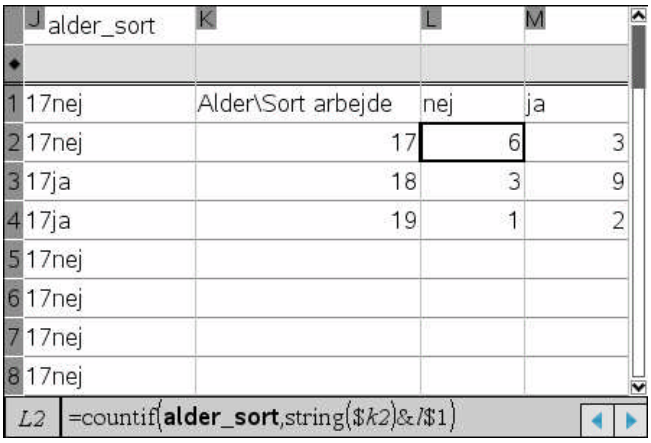
Vi starter da med at regne lidt på sammenhængen mellem alder og sort arbejde. Normalt sker det i form af en *krydstabel*, der viser hyppighederne for de to variable. Hyppighederne kan nemt aflæses fra det grupperede søjlediagram, men de kan også beregnes ved optælling. Her viser vi en teknik, hvor vi først kæder de to variable **alder** og **sort_arbejde** sammen til en enkelt mix-variabel **alder_sort**. Igen skal vi tage hensyn til at **alder** er indført som en numerisk variabel og vi bruger derfor string()-kommandoen til at konvertere den til en kategorisk variabel. Vi mixer derfor de to variable ved hjælp af cellekommandoen:

J1	=string(b1)&g1

Derefter foregår optællingen ved hjælp af cellekommandoen:

L2	=countif(**alder_sort**,string($k2)&l$1)

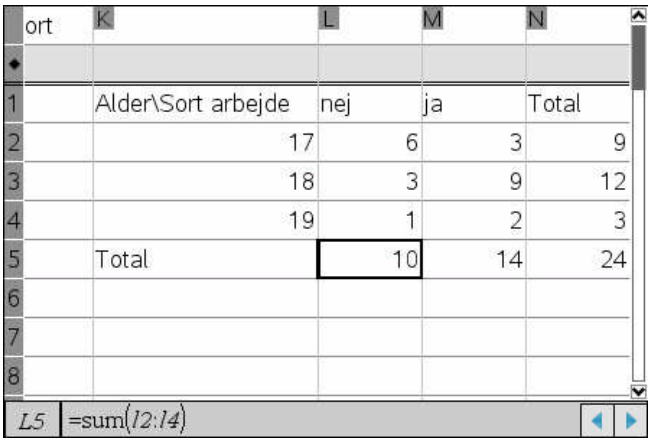
Læg mærke til dollartegnene $, der binder den første søjle **k** og den første række **1** i krydstabellen. De sikrer, at vi efterfølgende kan kopiere formlen rundt i krydstabellen:



	J alder_sort	K	L	M
1	17nej	Alder\Sort arbejde	nej	ja
2	17nej	17	6	3
3	17ja	18	3	9
4	17ja	19	1	2
5	17nej			
6	17nej			
7	17nej			
8	17nej			

L2 =countif(**alder_sort**,string($k2)&l$1)

Søjletotaler, Rækketotaler og den samlede Tabeltotal udregnes ved hjælp af **sum**-kommandoen anvendt på et celle-område:



	ort	K	L	M	N
1		Alder\Sort arbejde	nej	ja	Total
2		17	6	3	9
3		18	3	9	12
4		19	1	2	3
5		Total	10	14	24
6					
7					
8					

L5 =sum(l2:l4)

Vi får da dels optalt hyppighederne for de enkelte krydskategorier, fx at der er 3 elever, der er 17 år gamle og har sort arbejde, dels de såkaldte **marginaler** i form af **søjletotaler**, der angiver aldersfordelingen og **rækketotaler**, der angiver fordelingen for sort arbejde, og endelig den **samlede tabeltotal**, dvs. at der er 24 observationer i alt. Husk, at hvis du synes det er for besværligt at opbygge en formel til optælling af hyppighederne kan du bare aflæse dem direkte fra det grupperede søjlediagram og skrive dem ind i hånden – men nu ved du altså at det også kan lade sig gøre at oprette en pivottabel i **TI-Nspire CAS** ☺ (i afsnit 8 viste vi en anden teknik).

Ud fra krydstabellen skal vi derfor kunne aflæse om der synes at være en afhængighed mellem de to variable, dvs. om fordelingen af den ene variabel ændres afgørende, hvis vi indskrænker os til en bestemt værdi af den anden variabel. Fx ser vi, at der blandt de 17-årige er dobbelt så mange, der ikke har sort arbejde, mens der samlet set er flest, der har sort arbejde.

Vi kan også forholdsvis nemt udfylde en **forventet krydstabel** under forudsætning af at de to variable er uafhængige (nulhypotesen!). Hvis antallet af elever der har sort arbejde er uafhængigt af alderen, må antallet nemlig i hvert tilfælde udgøre 14/24 af det samlede antal. Tilsvarende må antallet af elever, der aldrig har udført sort arbejde, i hvert enkelt tilfælde udgøre 10/24 af det samlede antal. Da der fx er 9 elever med alderen 17 år, må de derfor forventes at fordele sig med 14/24·9, der har haft sort arbejde, og 10/24·9, der aldrig har haft sort arbejde. Det giver anledning til de følgende **forventede antal** i krydstabellen.

Det forventede antal – under forudsætning af at de to variable i en krydstabel er uafhængige – er givet ved formlen

$$\frac{\text{Rækketotal} \cdot \text{Søjletotal}}{\text{Samlet total}}$$

Vi kan nu nemt lave en krydstabel over de forventede værdier, ved at kopiere tabellen med de observerede hyppigheder og erstatte celleformlerne i krydstabellen med de relevante produkter:

	K	L	M	N
1	Alder\Sort arbejde	nej	ja	Total
2	17	6	3	9
3	18	3	9	12
4	19	1	2	3
5	Total	10	14	24
6	Forventede værdier			
7	Alder\Sort arbejde	nej	ja	Total
8	17	3.75	5.25	9.
9	18	5.	7.	12.
10	19	1.25	1.75	3.
11	Total	10.	14.	24.

L8 $=\frac{\$n2}{\$n\$5}\cdot/\$5\cdot 1.$

Læg mærke til dollartegnene, der 'fastfryser' udvalgte rækker og søjler.

9.3 Frihedsgrader i en krydstabel

Vi fortsætter undersøgelsen med en fundamental bemærkning: De to variable er karakteriseret ved deres fordelinger, der står i marginalerne, dvs. den yderste række og søjle. Disse marginalfordelinger må vi ikke røre ved. Men selve fordelingen inde i krydstabellen kan varieres på mangfoldige måder, idet der er stor frihed i hvordan vi udfylder de enkelte celler inde i krydstabellen. Her er det imidlertid vigtigt at være opmærksom på at det kun er nogle af cellerne, der giver stor frihed. Når vi fx har udfyldt den første celle i øverste venstre hjørne, så følger cellen lige til højre automatisk med, da summen af rækken jo skal give 9:

	A	B	C	D
1	Alder\Sort arbejde	nej	ja	Total
2	17	x	9−x	9
3	18			12
4	19			3
5	Total	10	14	24

Udfylder vi derefter den næste celle følger resten automatisk:

	A	B	C	D
1	Alder\Sort arbejde	nej	ja	Total
2	17	x	9−x	9
3	18	y	12−y	12
4	19	10−x−y	x+y−7	3
5	Total	10	14	24

Hvis vi fx udfylder den første celle med 7 og den anden med 2 (idet summen af de to første celler skal ligge mellem 7 og 10) fås den følgende krydstabel:

	A	B	C	D
◆				
1	Alder\Sort arbejde	nej	ja	Total
2	17	7	2	9
3	18	2	10	12
4	19	1	2	3
5	Total	10	14	24

Vi siger derfor at den ovenstående krydstabel har 2 **frihedsgrader**, fordi to af cellerne kan udfyldes frit, mens værdien af resten af cellerne derefter følger automatisk, idet summerne vandret og lodret er givet ved række og søjletotalerne.

I almindelighed gælder der at en krydstabel med **r** rækker og **s** søjler har netop (**r** – 1)·(**s** – 1) frihedsgrader.

9.4 Simulering af nulhypotesen – opbygning af teststørrelsen

Vi bryder en eventuel sammenhæng mellem de to variable, ved at foretage en tilfældig ombytning (permutation), dvs. **omrøring**, af værdierne i den ene variabel. Det sikrer, at der ikke længere er nogen sammenhæng mellem elevernes alder og deres erfaring med sort arbejde.

I dette tilfælde vælger vi at røre rundt i den forklarende variabel **alder** og opretter derfor en ny liste **sim_alder** og udfører en tilfældig permutation ved at trække en stikprøve af samme størrelse, men uden tilbagelægning ved hjælp af kommandoen

o	**sim_alder**:=randsamp(**alder**,dim(**alder**),1)

Vi mixer nu den simulerede alder med den oprindelige liste for sort arbejde.

P1	=string(*o1*)&*g1*

Dermed er vejen banet for at oprette en krydstabel for de simulerede data og finde de simulerede hyppigheder ved simpel optælling:

Simulerede værdier			
Alder\Sort arbejde	nej	ja	Total
17	3	6	9
18	5	7	12
19	2	1	3
Total	10	14	24

Hertil bruges cellekommandoen

L14	=countif(**sim_mix**,string($*k14*)&*l*$13)

Ikke overraskende ligner de observerede antal nu de forventede antal. Denne gang ved vi jo med sikkerhed at eventuelle afvigelser udelukkende skyldes de tilfældige variationer i krydstabellen.

B alder	G sortarbejde	O sim_alder	P sim_mix
		=randsamp(a	
17	nej	18	18nej
17	nej	18	18nej
17	ja	17	17ja
17	ja	18	18ja
17	nej	18	18nej
17	nej	19	19nej
17	nej	17	17nej
17	nej	17	17nej
17	ja	18	18ja
18	ja	18	18ja
18	ja	18	18ja
18	nej	18	18nej
18	ja	18	18ja
18	ja	17	17ja
18	ja	18	18ja
18	ja	17	17ja
18	nej	18	18nej
18	ja	18	18ja
18	ja	17	17ja
18	nej	19	19nej
18	ja	19	19ja
19	nej	17	17nej
19	ja	17	17ja
19	ja	17	17ja

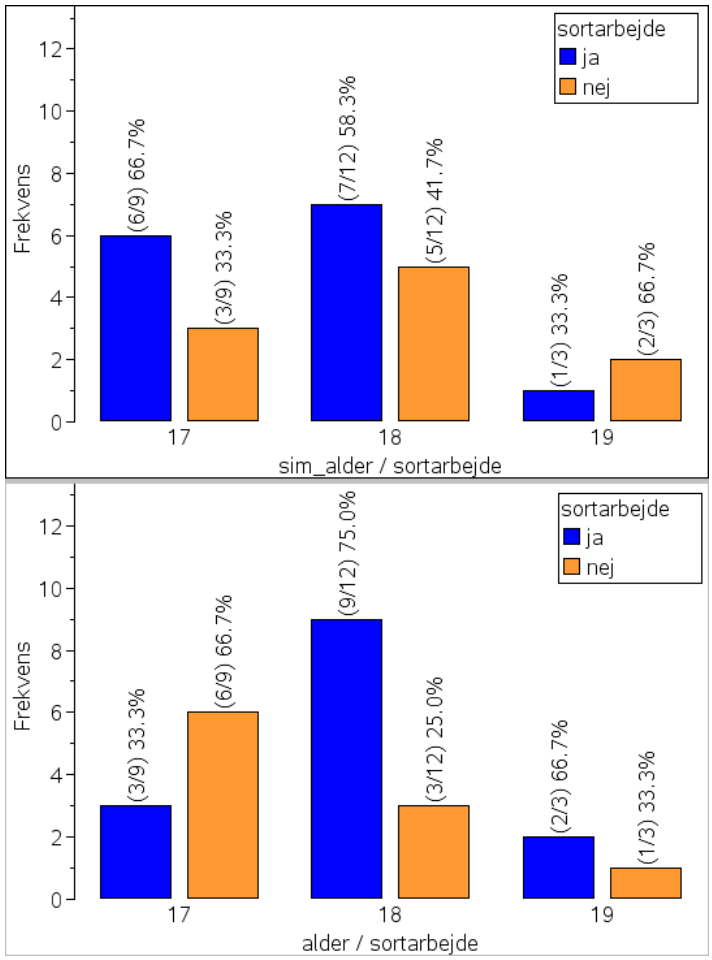
Men så er vejen jo banet for at udregne teststørrelsen, dvs. chi-kvadratet, for såvel det observerede datasæt som det omrørte datasæt:

O sim_alder	P sim_mix	Q
=randsamp(a		
18	18nej	Chi2_obs =
18	18nej	3.77143
17	17ja	Chi2_sim =
18	18ja	1.02857
18	18nej	
19	19nej	

Q2	$=\text{sum}\left(\frac{(l2:m4-l8:m10)^2}{l8:m10}\right)$

Q4	$=\text{sum}\left(\frac{(l14:m16-l8:m10)^2}{l8:m10}\right)$

Tilsvarende kan man tilføje det grupperede søjlediagram for de simulerede hyppigheder, så det direkte kan sammenlignes med det grupperede søjlediagram for de observerede hyppigheder:



Prøver man derefter at gentage omrøringen (klik i **Lister og regneark**-værkstedet efterfulgt af gentagne tast med **CTRL-R**) kan man hurtigt få en fornemmelse for hvor nemt/svært det er at frembringe en χ^2-testværdi, der er mindst lige så stor som den observerede 3.77. Det er hverken helt svært eller helt nemt. Jeg brugte for eksempel i første omgang 20 forsøg på bare at finde en enkelt krydstabel, der er præcis lige så slem.

9.5 Datafangst: Den eksperimentelle stikprøvefordeling

Det bliver altså nødvendigt med en mere præcis vurdering af p-værdien, dvs. sandsynligheden for at finde en testværdi, der er mindst lige så slem, for at få truffet en afgørelse. Her til skal vi udføre **gentagne målinger** på det omrørte datasæt. Vi starter da med at gemme såvel den observerede som den omrørte teststørrelse i variablene **chi2_obs** og **chi2_sim**

	O sim_alder	P sim_mix	Q
◆	=randsamp(a		
1	19	19nej	Chi2_obs =
2	18	18nej	**3.77143**
3	17	17ja	Chi2_sim =
4	17	17ja	**1.02857**
5	18	18nej	
6	17	17nej	

Q2 **chi2_obs**:=sum$\left(\frac{(l2:m4-l8:m10)^2}{l8:m10}\right)$

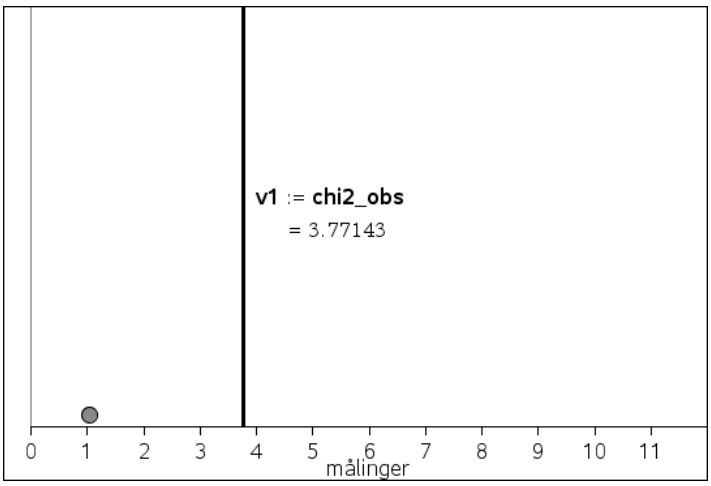
Q4 **chi2_sim**:=sum$\left(\frac{(l14:m16-l8:m10)^2}{l8:m10}\right)$

Læg mærke til at værdierne nu fremstår med fed, som tegn på, at de er tilknyttet en variabel.

Derefter oprettes en liste med navnet **målinger**, hvor vi udfører en **automatisk datafangst** af teststørrelsen **chi2_sim** via **data**-menuen. Tilsvarende tilføjer vi en celle **Q6** med formlen = count(**målinger**) så vi kan holde øje med antallet af målinger.

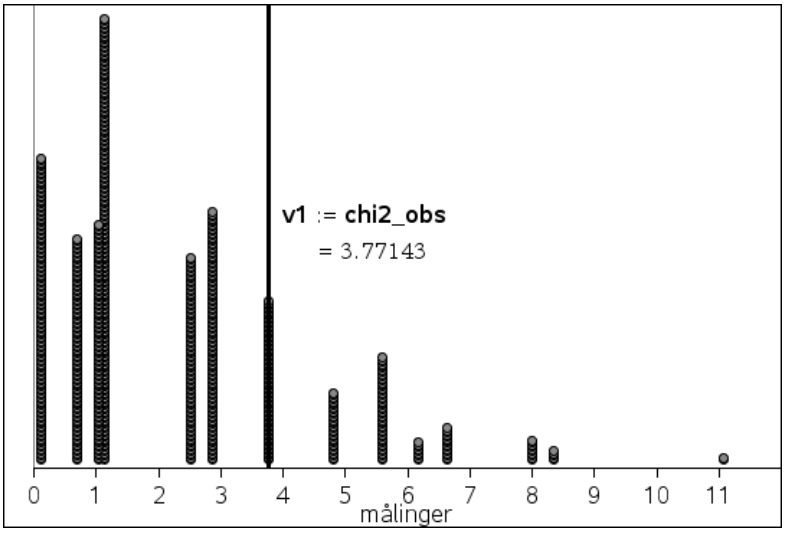
	O sim_alder	P sim_mix	Q	R målinger
◆	=randsamp(a			=capture(chi2_sim,1)
1	19	19nej	Chi2_obs =	1.02857
2	18	18nej	**3.77143**	
3	17	17ja	Chi2_sim =	
4	17	17ja	**1.02857**	
5	18	18nej	Antal målinger	
6	17	17nej	1	

Vi opretter nu et prikdiagram for variablen målinger og plotter værdien for den observerede teststørrelse **chi2_obs**, så vi kan holde øje med eventuelle skæve målinger:



Vi kan nu se fordelingen for teststørrelsen dukke langsomt frem på skærmen, selvom det selvfølgelig tager lidt tid at få udført de 1000 omrøringer med tilhørende målinger og dermed få udregnet de 1000 χ^2-testværdier for de tilhørende krydstabeller under antagelsen af at de to variable fra krydstabellen i virkeligheden er uafhængige (nulhypotesen).

Stikprøvefordelingen for de første 1000 målinger ser da således ud



Prikdiagrammet viser klart, hvor nemt det faktisk er at frembringe en testværdi, der er mindst lige så stor som den observerede værdi på 3.77143. Hvis vi tæller antallet af skæve målinger med kommandoen countif finder vi da også et skøn over p-værdien på ca. 20%.

Q	R målinger
	=capture(chi2_sim,1)
Chi2_obs =	1.02857
3.77143	0.114286
Chi2_sim =	1.14286
3.77143	2.85714
Antal målinger	1.14286
1000	6.17143
Antal skæve	1.02857
202	0.114286
p-værdi	2.85714
20.2	5.6

V8	=countif(**målinger**,?≥**chi2_obs**)

Det viser sig da at der i dette tilfælde er 202 ud af 1000 målinger, der ligger lige så langt eller længere ude end det rent faktisk observerede chi-kvadrat, så sandsynligheden for at den observerede forskel kan tilskrives rene tilfældigheder er altså helt oppe på ca. 20%. Når jeg således måtte bruge 20 forsøg på at ramme skævt bare en enkelt gang var det altså rent og skært uheld.

Nulhypotesen kan derfor ikke forkastes på det foreliggende grundlag. Selv om vi har observeret en tydelig tendens i stikprøven til at det sorte arbejde er mest udbredt blandt de elever, der allerede er fyldt 18 år, kan vi altså *ikke* generalisere vores antagelse om en systematisk sammenhæng mellem alder og sort arbejde, dvs. udvide den til hele populationen af danske gymnasieelever. Den observerede forskel er ganske enkelt ikke statistisk signifikant, dvs. den er ikke stor nok. Det betyder selvfølgelig ikke at påstanden nødvendigvis er forkert; men vores undersøgelse kan bare ikke bruges som dokumentation for påstanden, hvor rimelig den end måtte synes fra et samfundsfagligt synspunkt (der gælder fx helt forskellige lønningsregler for unge under 18 år og unge over 18 år, hvorfor mange unge gymnasieelever mister deres fritidsjob når de fylder 18 osv.).

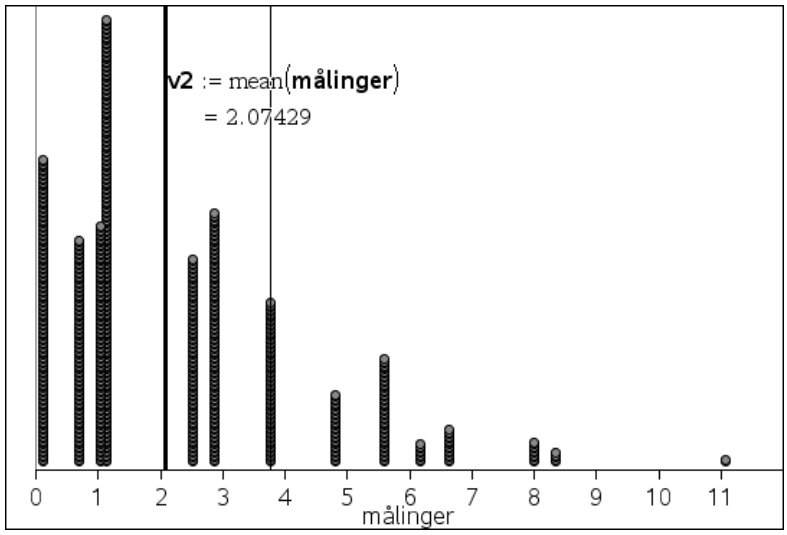
Hvad kunne der nu være forkert ved undersøgelsen? Det alvorligste problem er nok stikprøvens størrelse. Den er ikke særligt stor og det gør det svært at finde en signifikant forskel. Hvis vi bare kunne rykke én elev blandt de 17-årige fra gruppen med sort arbejde til gruppen uden sort arbejde (og tilsvarende omplacere én elev blandt de 19-årige for at bevare marginalforde-

lingerne) havde resultatet være signifikant. Hvad kunne vi så have gjort bedre? Først og fremmest bør vi skaffe os selv en større stikprøve ved at inddrage flere klasser. Hvis vi fx forestillede os at vi havde spurgt dobbelt så mange elever, dvs. 48, og at forholdene mellem dem der har sort arbejde, og dem der er uden sort arbejde i øvrigt var uændrede (dvs. vi fordobler alle hyppighederne), ville vi have frembragt en krydstabel med den dobbelte testværdi, dvs. $\chi^2 = 7.543$, men antallet af frihedsgrader ville stadig kun være 2. Resultatet ville derfor denne gang være klart signifikant.

> Hvis man ønsker at dokumentere en hypotese ved en spørgeskemaundersøgelse skal man altså sørge for at dimensionere den passende!

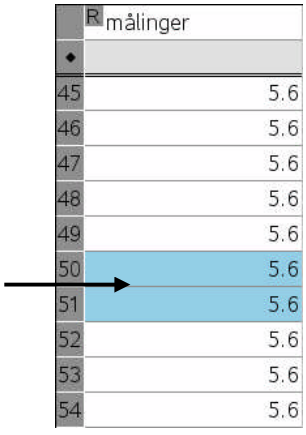
Vi slutter med nogle generelle bemærkninger om den χ^2-test vi her har udført.

1) Hvis vi tilføjer den eksperimentelle middelværdi til plottet over målinger



ser vi, at den eksperimentelle middelværdi viser sig at være 2.074. Den ligger netop rimeligt tæt på antallet af frihedsgrader, som er 2.

2) Vi kan også finde et skøn over den kritiske testværdi, dvs. den værdi som det observerede chi-kvadrat skal over for at afvigelsen er statistisk signifikant. Da signifikansniveauet i dette tilfælde er 5% og 5% udgør 50 målinger ud af 1000, skiller den kritiske testværdi de øverste 50 målinger fra de nederste 950 målinger. Vi skal derfor ordne målingerne aftagende efter størrelse. Da målinger er styret efter en 'formel' for automatisk datafangst, forsvinder denne formel, når vi ordner målingerne. Vi ser da at vores skøn over den kritiske testværdi er givet ved 5.6 i overensstemmelse med at vi skal have fordoblet den observerede teststørrelse for at frembringe et statistisk signifikant resultat.

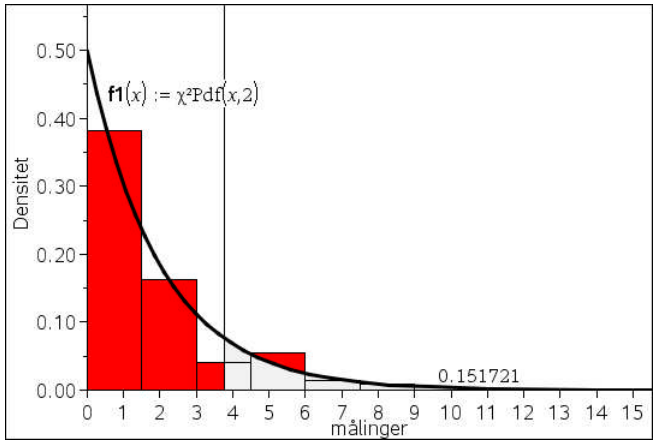


	R målinger
◆	
45	5.6
46	5.6
47	5.6
48	5.6
49	5.6
50	5.6
51	5.6
52	5.6
53	5.6
54	5.6

3) Når vi anvender automatisk dataopsamling begår vi faktisk en mindre men systematisk fejl, idet den automatiske dataopsamling kun registrerer *ændringer* i den målte værdi. Hvis to simuleringer fører til den samme chi2-testværdi registreres derfor kun den første af værdierne. Vi mangler derfor gentagelserne. Da gentagelser optræder hyppigst for de normale testværdier fører det til en overvurdering af antallet af skæve i forhold til antallet af normale målinger. Den fundne p-værdi er derfor overvurderet. Hvis man vil undgå denne fejlkilde skal man enten bruge manuel dataopsamling, hvilket er mere omstændeligt, eller man skal sno sig mere behændigt igennem den automatiske dataopsamling, hvilket igen er mere omstændeligt. Metoden er beskrevet i afsnit 7.6

9.6 Den teoretiske stikprøvefordeling

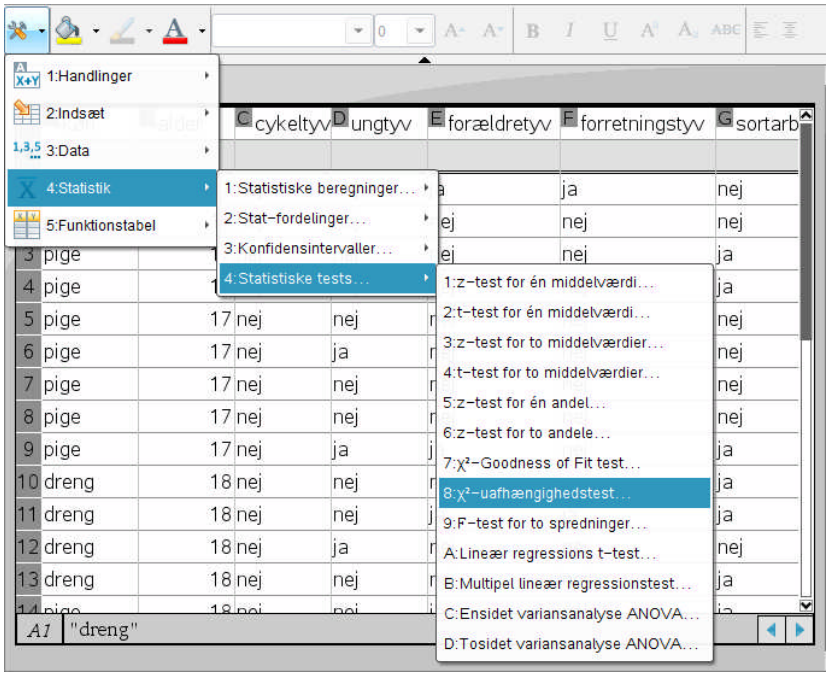
Vi har også set at prikplottet er rimeligt grynet. Der er jo kun et endeligt antal krydstabeller, fordi vi arbejder med heltallige værdier indenfor et snævert område. Men vi kan illustrere den asymptotiske fordeling ved at skifte over til et passende **histogram** (det er søjlebredden, der skal justeres for at tage højde for gryningen) og benytte skalaen **densitet**:



Her kan vi så tilføje den teoretiske chi-kvadratfordeling med 2 frihedsgrader og tilsvarende som vist udregne den teoretiske p-værdi ved at højre klikke på grafen og vælge **Skraver under funktion** og dernæst klikke på den lodrette linje for **chi2_obs** samt boksen med $\boxed{+\infty}$. Som det ses passer det rimeligt godt med vores eksperimentelle undersøgelse af teststørrelsens fordeling. Læg i øvrigt mærke til, at med så få frihedsgrader er fordelingen **ikke** klokkeformet! Faktisk er der tale om grafen for eksponentialfunktionen $\frac{1}{2} \cdot e^{-x/2}$.

9.7 Chi2-testet for uafhængighed som et kanonisk test

Til eksamen, eller når man er mere øvet i hypotesetest, benytter man ikke den tidskrævende eksperimentelle metode, men i stedet det indbyggede teoretiske test. Vi vælger da menupunktet **Statistiske tests... > χ^2 uafhængighedstest** fra **Statistik**-menuen:



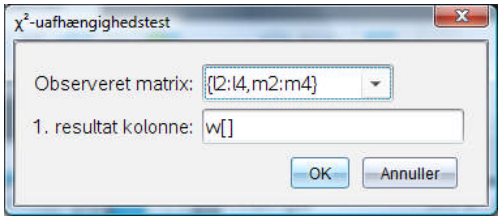
Vi får nu besked på at angive en krydstabel i form af en *matrix* bestående af de observerede hyppigheder. **Lister og regneark**-værkstedet understøtter *ikke* matricer. Men vi kan igen udnytte at en matrix kan repræsenteres af en liste af lister, hvor de inderste lister svarer til søjlerne i matricen:

Matricen for de observerede hyppigheder er derfor givet ved

{ L2:L4 , M2:M4 }

Resultatet udskrives da på samme måde som de andre testresultater, men denne gang rummer cellerne også matricer (indesluttet i gåseøjne, dvs. der er kun tale om tekst, idet **Lister og regneark**-værkstedet jo ikke understøtter matricer).

	K	L	M	N
◆				
1	Alder\Sort arbejde	nej	ja	Total
2	17	6	3	9
3	18	3	9	12
4	19	1	2	3
5	Total	10	14	24



W	X
	=χ²2way({l2:l4,m2:m4})
Titel	χ²–uafhængighedstest...
χ^2	3.77143
PVal	0.151721
df	2.
ExpMatrix	[[3.75,5.,1.25][5.25,7....
CompMatrix	[[1.35,0.8,0.05][0.964...

Vi får altså en vrimmel af oplysninger, men den vigtigste er den tredje **PVal** (Probability value, dvs. p-værdien), der netop fortæller os at sandsynligheden for at få en testværdi, der er mindst lige så stor som den faktisk observerede under forudsætning af at de to variable i virkeligheden er uafhængige (nulhypotesen, dvs. at den observerede forskel alene skyldes tilfældige variationer) er helt oppe på 15%. Da det ligger langt over signifikansniveauet på 5% er den observerede forskel altså *ikke* statistisk signifikant, i overensstemmelse med konklusionen i vores egen undersøgelse.

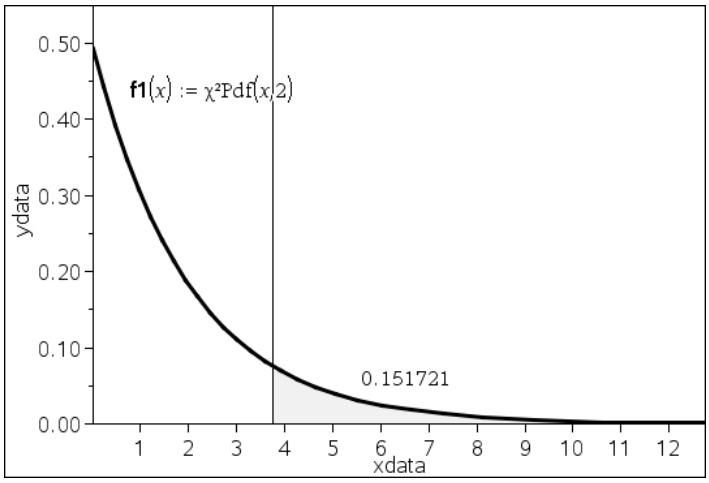
De øvrige oplysninger er dels teststørrelsen χ^2 = 3.77143, antallet af frihedsgrader **df** = 2 og den forventede matrix **ExpMatrix**. Endelig er der de enkelte matrixelementers bidrag til teststørrelsen. Disse matricer udskrives nemmest i **Beregninger**- eller **Noter**-værkstedet

stat.ExpMatrix	$\begin{bmatrix} 5.25 & 7. & 1.75 \\ 3.75 & 5. & 1.25 \end{bmatrix}$
stat.CompMatrix	$\begin{bmatrix} 0.964286 & 0.571429 & 0.035714 \\ 1.35 & 0.8 & 0.05 \end{bmatrix}$

Vi kan også få vist en graf med det kritiske område som støtte for fortolkningen. Men i uafhængighedstestet skal vi selv oprette grafen. Det sker ved at tegne funktionen for chi2-fordelingen med 2 frihedsgrader Chi2Pdf(x,2) samt plotte værdien for teststørrelsen stat.χ^2 = 3.77143. For at få oprettet et koordinatsystem i **Diagrammer og Statistik**-værkstedet er det ydermere nødvendigt at oprette to tomme lister **xdata** og **ydata**. Derefter højreklikkes på grafen og man vælger **Skraver under funktion** og angiver dels venstre grænse som den plottede værdi 3.77143, dels den højre grænse som $+\infty$. Det sværtede område ude til højre har da netop arealet 15%. Men for at forstå testværktøjet fuldt ud skal man altså dels have et indgående kendskab til den teoretiske fordeling af teststørrelsen, dels være opmærksom på at der skal være en række forudsætninger opfyldt før man kan drage troværdige konklusioner af testen, forudsætninger som vi ikke her har mulighed for at gøre fyldestgørende rede for.

xdata:={ } { }

ydata:={ } { }



10 Chi^2-test versus Gamma-test*

Vi har set at chi2-testen ikke altid er stærk nok til at se en sammenhæng og at én løsning på problemet kan være at øge stikprøvens størrelse for at gøre testen mere følsom. En anden mulighed foreligger, hvis der i virkeligheden er tale om ordinale variable, dvs. kategoriske variable med en naturlig rangordning. I så fald vil vi kunne udnytte denne rangordning og skifte til en gammatest, der er stærkere end chi2-testen når det drejer sig om ordinale variable. Problemstillingen forekommer hyppigt i forbindelse med spørgeskemaer, hvor man netop typisk involverer ordinale variable, fx ved at spørge om respondentens holdning til en påstand og bede dem anføre om de er "helt uenige", "lidt uenige", "neutrale", "lidt enige" eller "helt enige" i påstanden. I så fald foreligger der jo en klar rangordning. Gammatesten er ikke kernepensum og det følgende afsnit kan derfor roligt overspringes i en første gennemlæsning. Men da gammatesten forekommer ofte i undersøgelser af relevans for samfundsfag/sociologi eller biologi, og da det ikke er så nemt at finde materiale om den, har jeg valgt at inkludere den[2].

10.1 Rygevaner i Glostrup – den kanoniske χ^2-test *

Som eksempel vil vi kigge på et autentisk datamateriale der vedrører rygevaner og helbredstilstand. Helbredstilstanden stammer fra respondentens egen vurdering og kaldes derfor SRH (selv rapporteret helbredstilstand). Rygevaner går fra aldrig at have røget til at ryge mindst 25 cigaretter om dagen. Helbredstilstanden går fra dårlig til meget fin. Datasættet omfatter 417 respondenter fra Glostrup og kan hentes på TI's hjemmeside, hvis du vil lege med ☺:

	A	B	C	D	E	F
◆						
1	Total	6	39	317	55	417
2	25+	1	3	29	1	34
3	15–24	3	17	81	10	111
4	1–14	1	7	59	13	80
5	Stoppet	0	6	75	15	96
6	ikke–ryger	1	6	73	16	96
7	Rygning/Helbred	Dårligt	Rimeligt	Godt	Rigtigt godt	Total

I overensstemmelse med den tidligere diskussion af ordinale variable i afsnit 4.4 har vi anført rækkefølgen af kategorierne så den passer med et koordinatsystem: Rygningen vokser lodret opad og helbredet forbedres vandret mod højre. Vi kan indledningsvis foretage en kanonisk chi2-test for uafhængighed og finder da:

	W	X	Y
◆		=χ²2way({b2:b6,c2:c6,d2:d6,e	
1	Titel	χ²–uafhængighedstest	
2	χ²	16.1976	
3	PVal	0.182351	
4	df	12.	
5	ExpMatr...	[[0.48920863309353,1.5971...	
6	CompMa...	[[0.53332628015235,1.2322...	

X =χ²2way({b2:b6,c2:c6,d2:d6,e2:e6}): CopyVar **Stat.**, **Stat1.**

Der er 12 frihedsgrader, ch2-teststørrelsen er 16.20, hvilket er rimeligt tæt på den forventede værdi 12, og p-værdien er da også oppe på 18.2%. Det er altså *ikke* muligt at forkaste nulhypo-

[2] Du kan finde mere materiale om gamma-testen på følgende website: http://staff.pubhealth.ku.dk/~skm/fsvpage/Analysis%20of%20multivariate%20categorical%20data/CatData%20course.htm

tesen om uafhængighed på et signifikansniveau på 10%. Men kan være en smule bekymret for de små yderværdier. Men det skulle jo bare gøre det nemmere at forkaste nulhypotesen. De forventede værdier sammenholdt med bidragene til chi2-teststørrelsen ser heller ikke foruroligende ud. Konklusionen er altså at chi2-testen frikender nulhypotesen!

$$\text{stat1.ExpMatrix}^T \rightarrow \begin{bmatrix} 0.489209 & 3.17986 & 25.8465 & 4.48441 \\ 1.59712 & 10.3813 & 84.3813 & 14.6403 \\ 1.15108 & 7.48201 & 60.8153 & 10.5516 \\ 1.38129 & 8.97842 & 72.9784 & 12.6619 \\ 1.38129 & 8.97842 & 72.9784 & 12.6619 \end{bmatrix}$$

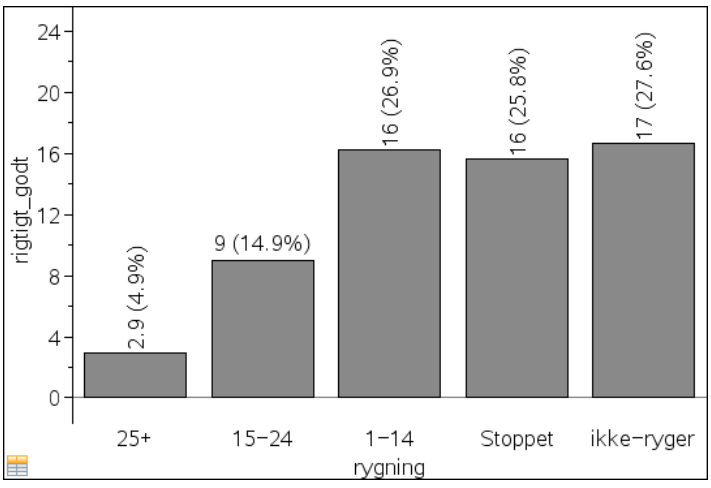
$$\text{stat1.CompMatrix}^T \rightarrow \begin{bmatrix} 0.533326 & 0.010173 & 0.384749 & 2.70741 \\ 1.23226 & 4.21983 & 0.135494 & 1.47075 \\ 0.019829 & 0.031053 & 0.054188 & 0.56815 \\ 1.38129 & 0.988033 & 0.056 & 0.431757 \\ 0.105253 & 0.988033 & 0.000006 & 0.880052 \end{bmatrix}$$

10.2 Rygevaner versus helbredstilstand undersøgt med gammatesten*

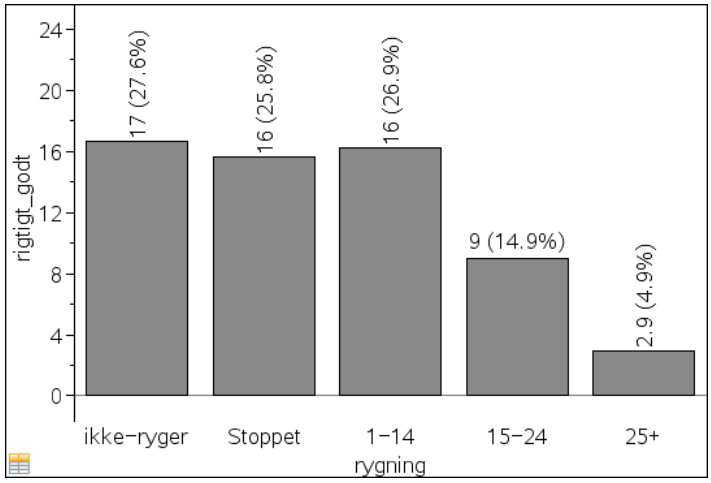
Men hvis vi i stedet håndterer de to variable som ordinale variable kan vi inddrage den skjulte information, der ligger gemt i rangordningen! Vi ser da et tydeligt mønster, hvis vi flytter data op over stregen, så vi kan afbilde dem i et passende diagram:

	A	B	C	D	E	F	G rygning	H dårligt	I godt	J rigtigt_godt
◆							=a10:a14	=b10:b14	=d10:d14	=e10:e14
1	Total	6	39	317	55	417	25+	2.94118	85.2941	2.94118
2	25+	1	3	29	1	34	15−24	2.7027	72.973	9.00901
3	15−24	3	17	81	10	111	1−14	1.25	73.75	16.25
4	1−14	1	7	59	13	80	Stoppet	0.	78.125	15.625
5	Stoppet	0	6	75	15	96	ikke−ryger	1.04167	76.0417	16.6667
6	ikke−ryger	1	6	73	16	96				
7	Rygning/Helbred	Dårligt	Rimeligt	Godt	Rigtigt godt	Total				
8										
9	Total	1.43885	9.35252	76.0192	13.1894	100.				
10	25+	2.94118	8.82353	85.2941	2.94118	100.				
11	15−24	2.7027	15.3153	72.973	9.00901	100.				
12	1−14	1.25	8.75	73.75	16.25	100.				
13	Stoppet	0.	6.25	78.125	15.625	100.				
14	ikke−ryger	1.04167	6.25	76.0417	16.6667	100.				

Vi ser fx på fordelingen af rygere henover de forskellige helbredstilstande. For de testpersoner der selv synes de har det rigtigt godt ser fordelingen således ud:



For de testpersoner, der kun synes de har det rimeligt godt, ser fordelingen derimod således ud:



Vi ser altså at fordelingen er den samme for de to helbredskategorier: I begge tilfælde falder andelen af rygere i takt med at de ryger mere og mere. Det kunne tyde på at aftagende sammenhæng: Megen rygning fører til dårlig helbredstilstand. For at undersøge en sådan sammenhæng gennemfører vi derfor et gammatest. Vi viser det først eksperimentelt:

Vi skal da have udregnet gammagraden, der viser hvilken grad af overensstemmelse eller uoverensstemmelse, der er mellem opførslen af de to variable. Det er en lille smule omstændeligt at udregne gamma-graden, idet der er lige så mange bidrag til gammagraden, som der er frihedsgrader:

	D	E	F	G	H
5	6	75	15	96 Voksende	10146
6	6	73	16	96 Aftagende	16618
7	imeligt	Godt	Rigtigt godt	Total Gamma	**-0.241817**
8					

H5 =b6·sum(c2:e5)+c6·sum(d2:e5)+d6·sum(e2:e5)+b5·sum(c2:e

voksende

$$=b6\cdot\text{sum}(c2:e5)+c6\cdot\text{sum}(d2:e5)+d6\cdot\text{sum}(e2:e5)+b5\cdot\text{sum}(c2:e4)$$
$$+c5\cdot\text{sum}(d2:e4)+d5\cdot\text{sum}(e2:e4)+b4\cdot\text{sum}(c2:e3)+c4\cdot\text{sum}(d2:e3)$$
$$+d4\cdot\text{sum}(e2:e3)+b3\cdot\text{sum}(c2:e2)+c3\cdot\text{sum}(d2:e2)+d3\cdot e2$$

aftagende

$$=b2\cdot\text{sum}(c3:e6)+b3\cdot\text{sum}(c4:e6)+b4\cdot\text{sum}(c5:e6)+b5\cdot\text{sum}(c6:e6)$$
$$+c2\cdot\text{sum}(d3:e6)+c3\cdot\text{sum}(d4:e6)+c4\cdot\text{sum}(d5:e6)+c5\cdot\text{sum}(d6:e6)$$
$$+d2\cdot\text{sum}(e3:e6)+d3\cdot\text{sum}(e4:e6)+d4\cdot\text{sum}(e5:e6)+d5\cdot e6$$

Man skal altså have tungen lige i munden for at få dem udregnet korrekt. Men der findes også en biblioteksfunktion, der udregner gammagraden, som vi vender tilbage til i det sidste afsnit.

Her og nu konstatere vi at gammagraden er -0.24 og at der derfor som forventet synes at være en negativ sammenhæng mellem rygevaner og helbredstilstand. Men er sammenhængen signifikant? Det kan vi afgøre med en uafhængighedstest: Hvis de to variable var uafhængige ville vi i middel forvente en gammagrad på 0, altså at der hverken var en voksende eller aftagende tendens til at rygevanerne hang sammen med helbredstilstanden. Og det kan vi jo afgøre eksperimentelt med de samme teknikker, som vi har anvendt til chi2-testen for uafhængighed. Først dublerer vi tabellen, idet vi erstatter hyppighederne med kategorierne:

	A	B	C	D	E
◆					
17	25+	25+	25+	25+	25+
18	15−24	15−24	15−24	15−24	15−24
19	1−14	1−14	1−14	1−14	1−14
20	Stoppet	Stoppet	Stoppet	Stoppet	Stoppet
21	ikke−ryger	ikke−ryg...	ikke−ryg...	ikke−ryg...	ikke−ryger
22	Rygning/Helbred	Dårligt	Rimeligt	Godt	Rigtigt godt
23					
24	25+	Dårligt	Rimeligt	Godt	Rigtigt godt
25	15−24	Dårligt	Rimeligt	Godt	Rigtigt godt
26	1−14	Dårligt	Rimeligt	Godt	Rigtigt godt
27	Stoppet	Dårligt	Rimeligt	Godt	Rigtigt godt
28	ikke−ryger	Dårligt	Rimeligt	Godt	Rigtigt godt
29	Rygning/Helbred	Dårligt	Rimeligt	Godt	Rigtigt godt

Det giver os mulighed for at rekonstruere de rå data ved hjælp af freqTable@>list()-kommandoen, idet vi nu både kan pege mod kategorierne og hyppighederne:

	I rygevaner	J helbred	K mix
◆	=freqtable▸lis	=freqtable	
1	25+	Dårligt	25+Dårligt
2	15−24	Dårligt	15−24Dårligt
3	15−24	Dårligt	15−24Dårligt
4	15−24	Dårligt	15−24Dårligt
5	1−14	Dårligt	1−14Dårligt
6	ikke−ryger	Dårligt	ikke−rygerDårligt
7	25+	Rimeligt	25+Rimeligt
8	25+	Rimeligt	25+Rimeligt
9	25+	Rimeligt	25+Rimeligt
10	15−24	Rimeligt	15−24Rimeligt
11	15−24	Rimeligt	15−24Rimeligt
12	15−24	Rimeligt	15−24Rimeligt
I	**rygevaner**:=freqtable▸list(*b17:e21,b2:e6*)		

	L sim	M sim_mix
◆	=randsamp(helbred,417,1)	
1	Rigtigt godt	25+Rigtigt godt
2	Godt	15−24Godt
3	Rigtigt godt	15−24Rigtigt godt
4	Godt	15−24Godt
5	Godt	1−14Godt
6	Godt	ikke−rygerGodt
7	Rigtigt godt	25+Rigtigt godt
8	Godt	25+Godt
9	Godt	25+Godt
10	Godt	15−24Godt
11	Godt	15−24Godt
12	Godt	15−24Godt
M1	*=i1&l1*	

Her er mix-variablen taget med for at give mulighed for at kontrollere grafisk at de rå data er i overensstemmelse med krydstabellen. Vi omrører nu som vist helbredstilstanden og mikser den med rygevanerne, så vi kan simulere uafhængigheden på sædvanlig vis! Vi kan nemlig nu trække de simulerede hyppigheder ud af **sim_mix** og dermed opbygge en tabel over de simulerede hyppigheder. På basis af denne tabel kan vi så genbruge formlerne til udregning af antallet af voksende og aftagende par i den simulerede tabel og dermed beregne den simulerede gamma-grad:

	N	O	P	Q	R	S	T	U
◆								
1	Total	6	39	317	55	417		
2	25+	1	3	26	4	34		
3	15−24	1	9	83	18	111		
4	1−14	2	9	57	12	80		
5	Stoppet	1	8	79	8	96	Voksende	14071
6	ikke−ryg...	1	10	72	13	96	Aftagende	12833
7	Rygning/...	Dårligt	Rimeligt	Godt	Rigtigt g...	Total	Gamma	**0.046015**

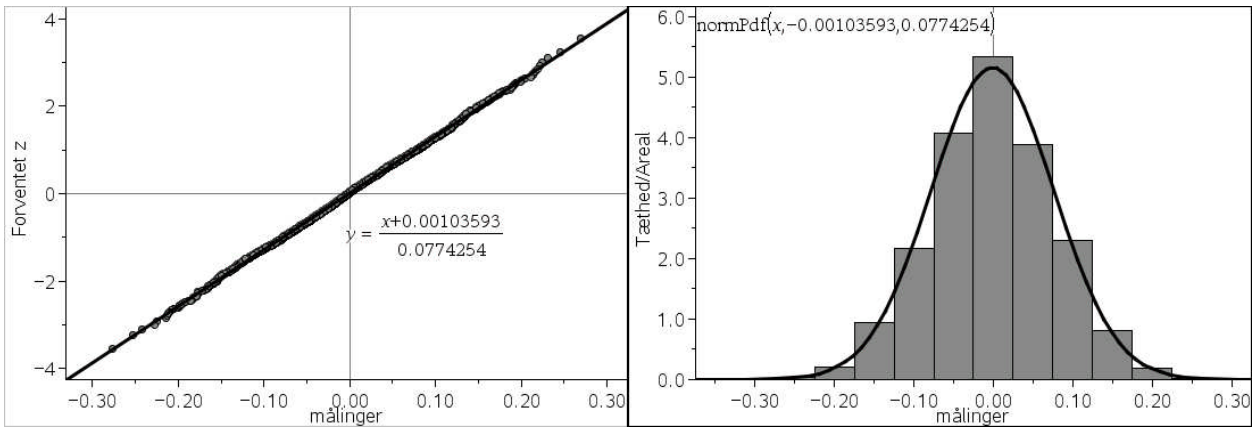
U7 **gamma**:=$\frac{u5-u6}{u5+u6}\cdot 1.$

Dermed er vejen banet for en automatisk måling af den simulerede gammagrad og ved at gentage simuleringen 2500 gange finder vi at der kun fem gange kommer en gammagrad, der er mindst lige så skæv som den observerede, hvad enten den er positiv eller negativ. Optællingen viser altså at p-værdien er nede på 0.2% og den aftagende sammenhæng mellem rygevaner og helbredsstilstand er derfor signifikant på signifikansniveauet 5% (og for den sags skyld også på 1%).

	T	U	V målinger	W
◆			=capture(gamma,1)	
5	Voksende	14071	0.058092	ExpMatrix
6	Aftagende	12833	-0.093078	CompMatrix
7	Gamma	**0.046015**	-0.088648	
8			-0.113969	
9	Antal målinger		-0.036968	
10	2500		-0.165441	
11	Antal skæve		0.013562	
12	5		-0.032035	
13	p−værdi		-0.003646	
14	0.2		-0.064048	

T12 =countif(**målinger**,?≤**gamma_obs**)+countif(**målinger**,?≥-**gamma_obs**)

Fordelingen af målingerne ligner en typisk klokkeformet fordeling. Den er endog meget tæt på at være normafordelt:



Asymptotisk kan man derfor udforme gammatestet som et normalfordelingstest.

Konklusion: I følge gammatesten er der en signifikant aftagende sammenhæng, som vi også med chi2-testen.

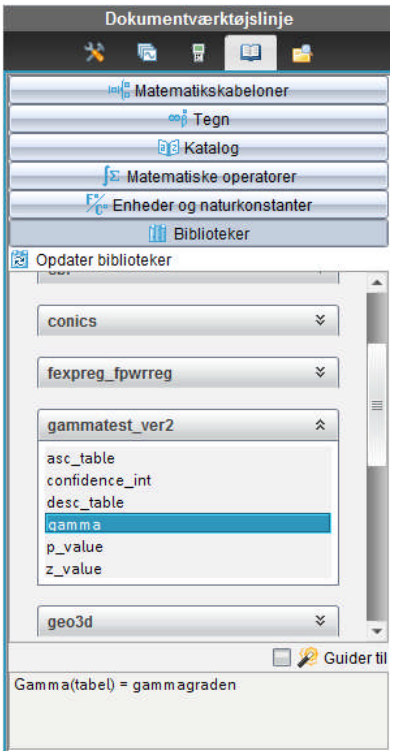
10.3 Gammatesten udført som en kanonisk test*

Gammatesten er *ikke* inkluderet blandt de statistiske tests som **TI-Nspire CAS** understøtter. Men man kan tilføje den via et *offentligt* biblioteksprogram, der indeholder rutiner for udregning af gammagraden, p-værdien og konfidensintervallerne, idet vi udfører testen som en asymptotisk normalfordelingstest. For at kunne anvende biblioteksfunktionerne skal filen **gammatest** gemmes i mappen

Dokumenter >TI-Nspire > MyLib

Mappen MyLib oprettes automatisk af **TI-Nspire CAS**, når man installerer programmet. Hvis du vil lege med skal du derfor hente filen gammatest på TI's hjemmeside og gemme den i MyLib for at få adgang til gammatesten.

For at anvende filen skal du nu åbne **Hjælpeprogrammer** og vælge **Biblioteker:**



Som det ses får du da adgang til en stribe biblioteksfunktioner til udregning af fx gammagraden, p-værdien, 95%-konfidensintervallet for gammagraden og z-værdien (dvs. den asymptotiske normalfordelings teststørrelse.

Her vil vi kalde gammagraden. Vi skal da indsætte en matrix, men ligesom i de andre tests kan den også anføres som en liste af lister. Vær opmærksom på, at biblioteksfunktionerne opfatter matricen som ordnet på sædvanlig vis, dvs. i læseretningen lodret nedad og vandret henad mod højre i modsætning til den måde vi læser koordinatsystemer! VI risikerer altså et fortegnsskift på gammagraden ☺

Vi finder da de følgende resultater i regnearket, hvor vi trækker tallene ud af krydstabellen:

	A	B	C	D	E
◆					
1	Total	6	39	317	55
2	25+	1	3	29	1
3	15–24	3	17	81	10
4	1–14	1	7	59	13
5	Stoppet	0	6	75	15
6	ikke–ryger	1	6	73	16
7	Rygning/Helbred	Dårligt	Rimeligt	Godt	Rigtigt godt
8	Gamma	p–værdi	z–værdi	Konfindens	
9	0.241817	0.000977	3.29703	[0.1033;0.3803]	
10					

A9 =gammatest\gamma({b2:b6,c2:c6,d2:d6,e2:e6},1.

Det skal sammenholdes med den eksperimentelle test, hvor vi fandt gammagraden-0.24817. Fortegnsskiftet skyldes alene den matrix-konvention vi allerede har omtalt. Til svarende fandt vi den eksperimentelle 2-sidede p-værdi til 0.2%, hvor den teoretiske er 0.1%. Den eksperimentelle teststørrelse er gammagraden divideret med spredningen, som vi fandt til 0.0774254, dvs. z_{eks} er givet ved

$$z = \frac{\gamma}{s_0} = \frac{-0.241817}{0.0774254} \approx 3.12$$

hvor den teoretiske z-værdi er givet ved 3.30. Alt i alt er der altså rimelig overensstemmelse mellem den eksperimentelle test og den kanoniske test. Endelig får vi også oplyst 95% konfidensintervallet [0.10;0.38], der *ikke* indeholder 0 i overensstemmelse med at nulhypotesen forkastes.