Værktøjshjælp for DataMeter
Struktur for appendiks:

Til hvert af de gennemgåede værktøjer findes der 5 afsnit. De enkelte afsnit kan læses uafhængigt af hinanden. Der forudsættes et elementært kendskab til det pågældende værktøj. Der er mange forskellige måder man kan benytte værktøjerne på – det følgende er kun et forslag – i forbindelse med den faktiske udførelse af undervisningen kan andre metoder sagtens vise sig mere hensigtsmæssige. Af samme grund er det heller ikke nødvendigt at gennemarbejde samtlige afsnit.

Det er valgfrit til såvel den skriftlige eksamen som den mundtlige eksamen om man vil benytte sig af teoretiske metoder eller eksperimentelle metoder. Til den skriftlige eksamen er de indbyggede fordelinger og rutiner et godt udgangspunkt (afsnit 4 og 5). Til den mundtlige eksamen er eksperimentel hypotesetest i forbindelse med et statistisk projekt et godt udgangspunkt (dele af afsnit 2).

Indholdsfortegnelse

Følgende **TI-Nspire CAS**-filer følger med:
Afsnit 1: Eksempler på grafiske fremstillinger.
Afsnit 2: Eksperimentel hypotesetest
Afsnit 3: De indbyggede fordelingsfunktioner
Afsnit 4: Teoretiske udregninger
Afsnit 5: De indbyggede test

1) Eksempler på grafisk fremstilling af data:
1a: Uafhængighed

Eksempel 1: (side 4 i kursusmaterialet)

Køn\Forbrug	<1500 kr./måned	≥1500 kr./måned	i alt
kvinder	98	102	200
mænd	60	100	160
i alt	158	202	360

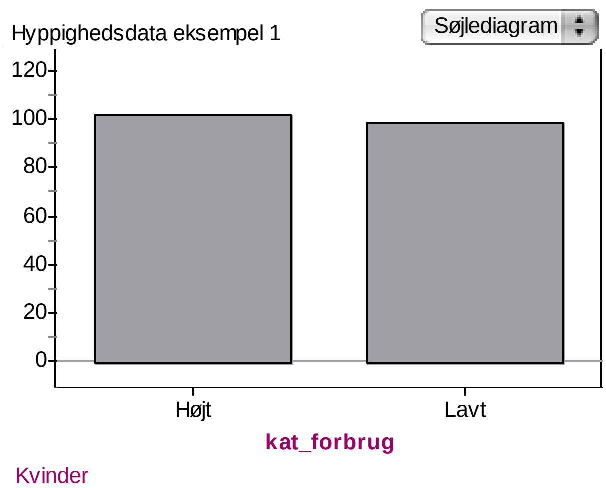
Når vi skal vurdere om der er samme fordeling af forbruget hos henholdsvis mænd og kvinder er det formentligt nemmest at taste data ind i lister og regneark og overføre dem til grafer som frekvensplot, her illustreret som søjlediagrammer. Det ses da tydeligt at mænds forbrugsmønster i den pågældende stikprøve ser helt anderledes ud end kvinders forbrugsmønster.

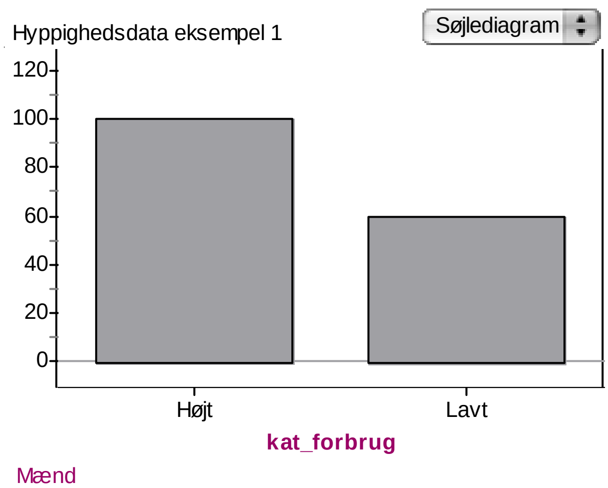
Bemærkning: Frekvensplot frembringes ved at trække den kategoriske variabel, her **kat_forbrug**, ind i grafrummets vandrette akse og dernæst erstatte **tæl**()-funktionen med den variabel, der rummer hyppighederne.

Hyppighedsdata eksempel 1

Hyppighedsdata eksempel 1

	kat_forbrug	Kvinder	Mænd
1	Lavt	98	60
2	Højt	102	100





Forskellen mellem de to køns forbrug fremstår endnu tydeligere i et blokdiagram. Men det kræver adgang til de rå data, som derfor først må opbygges på grundlag af hyppighedstabellen. Hvis forbruget var uafhængigt af køn skulle delelinjen ligge lige højt for de to køn.

Rådata eksempel 1

Køn =

$$\text{ombyt}(\text{indeks})\begin{cases}(\text{indeks} \le 200) & : \text{"Kvinde"} \\ \text{ellers} & : \text{"Mand"}\end{cases}$$

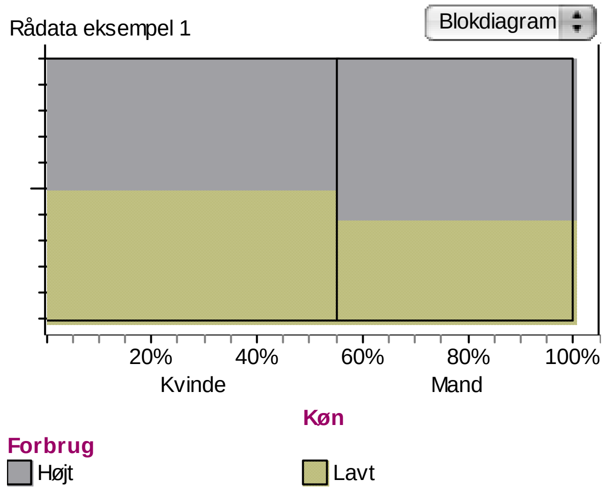
Forbrug =

$$\text{ombyt}(\text{indeks})\begin{cases}(\text{indeks} \le 98) & : \text{"Lavt"} \\ (\text{indeks} \le 200) & : \text{"Højt"} \\ (\text{indeks} \le 260) & : \text{"Lavt"} \\ \text{ellers} & : \text{"Højt"}\end{cases}$$

Bemærkning: Blokdiagrammer frembringes ved at trække den kategoriske variabel **køn** ind i grafrummets vandrette akse og dernæst trække den kategoriske variabel **forbrug** ind i selve grafrummet, så den splitter søjlediagrammet for **køn**, der derefter kan omdannes til et blokdiagram.

Rådata eksempel 1

	Køn	Forbrug
1	Kvinde	Lavt
2	Kvinde	Lavt
3	Kvinde	Lavt
4	Kvinde	Lavt
5	Kvinde	Lavt
6	Kvinde	Lavt
7	Kvinde	Lavt
8	Kvinde	Lavt
9	Kvinde	Lavt
10	Kvinde	Lavt



1b Goodness of fit

Eksempel 2 (side 24 i kursusmaterialet)

Indkomstfordelingen i stikprøven var: *I = Indkomst i 1000 kr.*

Observeret antal

I<50	50<=I<100	100<=I<150	150<=I<200	200<=I<300	300<=I<400	400<=I<500	500<=I
98	88 1	99	136	210	179	52	38

Den forventede fordeling i stikprøven baseret på de ovenstående procenter er tilsvarende givet ved:

Forventet antal

I<50	50<=I<100	100<=I<150	150<=I<200	200<=I<300	300<=I<400	400<=I<500	500<=I
64	93	178	123	243	180	66	53

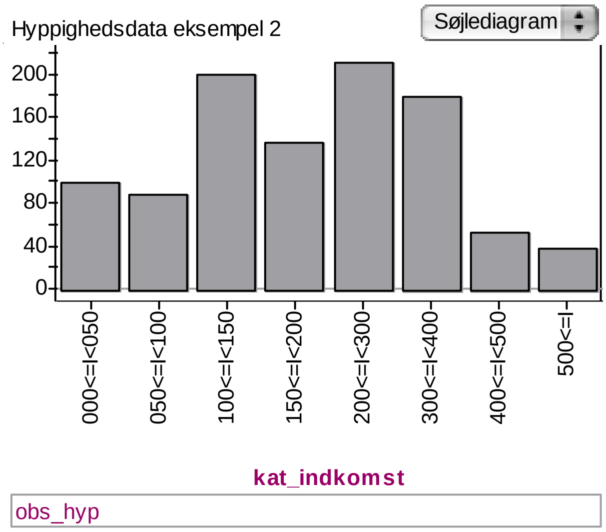
Sammenholder vi de observerede hyppigheder med de forventede følges de så nogenlunde ad. Men man kunne måske være bekymret for, om de laveste indkomster er overrepræsenteret i stikprøven. Her ligger den observerede hyppighed et godt stykke over den forventede.

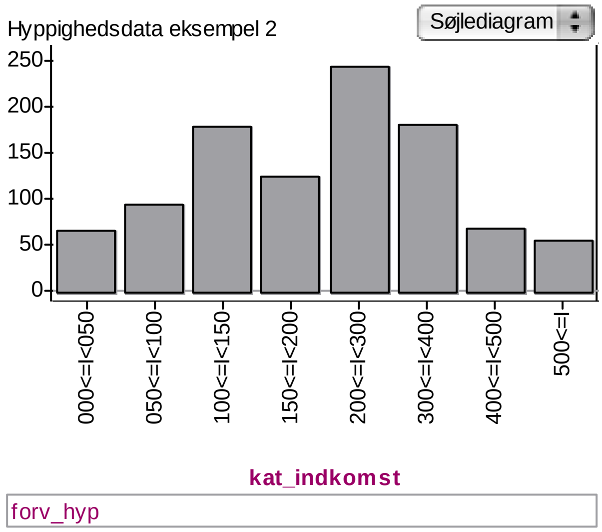
Når vi skal vurdere om der er samme fordeling af indkomster i interviewundersøgelsen (stikprøven) og landsgennemsnittet (populationen) er det formentligt nemmest at taste data ind i lister og regneark og overføre dem til grafer som frekvensplot, her illustreret som søjlediagrammer. Det ses da tydeligt at fx fordelingen for de to laveste indkomstgrupper er vendt om i stikprøven i forhold til populationen. Der synes altså at være grund til bekymringen!

Hyppighedsdata eksempel 2

	kat_indkomst	obs_hyp	forv_hyp
1	000<=I<050	98	64
2	050<=I<100	88	93
3	100<=I<150	199	178
4	150<=I<200	136	123
5	200<=I<300	210	243
6	300<=I<400	179	180
7	400<=I<500	52	66
8	500<=I	38	53

Hyppighedsdata eksempel 2





2) **Eksperimentel hypotesetest**

2a: Uafhængighed

Eksempel 1: (side 4 i kursusmaterialet)

Køn\Forbrug	<1500 kr./måned	≥ 1500 kr./måned	i alt
kvinder	98	102	200
mænd	60	100	160
i alt	158	202	360

Simulering af nulhypotesen

For at simulere nulhypotesen, der påstår at forbruget er uafhængigt af kønnet, må vi først fastlægge en fortolkning af hvad vi mener med uafhængighed. Det kan gøres på flere måder.

Metode 1: Vi diskuterer først **omrøring**.

Vi konstruerer først de rå data for **køn** og **forbrug** der er i overensstemmelse med de oplyste hyppigheder. Vi får også brug for de forventede hyppigheder til udregning af teststørrelsen. De udregnes ud fra en krydstabel frembragt ved at trække de to variable **Køn** og **Forbrug** ind i en beregningsboks.

Endelig kan vi udregne teststørrelsen ved at højreklikke i krydstabellen og overføre celler til nyt datasæt. Den viser sig at have værdien 4,7735305

Rå data

	Køn	Forbrug	<ny>
=	ombyt (indeks) $\begin{cases} (? \le 200) & : \text{"Kvinde"} \\ \text{ellers} & : \text{"Mand"} \end{cases}$	ombyt (indeks) $\begin{cases} (? \le 98) & : \text{"Lavt"} \\ (? \le 200) & : \text{"Højt"} \\ (? \le 260) & : \text{"Lavt"} \\ \text{ellers} & : \text{"Højt"} \end{cases}$	
1	Kvinde	Lavt	
2	Kvinde	Lavt	
3	Kvinde	Lavt	
4	Kvinde	Lavt	
5	Kvinde	Lavt	
6	Kvinde	Lavt	
7	Kvinde	Lavt	
8	Kvinde	Lavt	

Rå data

Rå data

		Forbrug: Højt	Forbrug: Lavt	Række total
Køn	Kvinde	102 112,22222	98 87,777778	200 200
	Mand	100 89,777778	60 70,222222	160 160
	Søjle total	202 202	158 158	360 360

R1 = tæl ( )

R2 = Forventet

Celler fra oversigtstabel for Rå data

Celler fra oversigtstabel for Rå data

	Køn	Forbrug	R1	R2
1	Kvinde	Højt	102	112,222
2	Kvinde	Lavt	98	87,7778
3	Mand	Højt	100	89,7778
4	Mand	Lavt	60	70,2222

Celler fra oversigtstabel for Rå data

	4,7735305

$$R1 = \text{sum}\left(\frac{(R1 - R2)^2}{R2}\right)$$

Så er vi klar til omrøringen! I DataMeter udføres omrøringen ved at højreklikke på datasættet for de rå data og vælge **Rør rundt i en variabel**. Som udgangspunkt røres der rundt i den første variabel, dvs. her **Køn** og det er jo fint nok! Der udføres en tilfældig permutation af elementerne i listen for **køn** og derved brydes en eventuel sammenhæng med listen for **forbrug**.

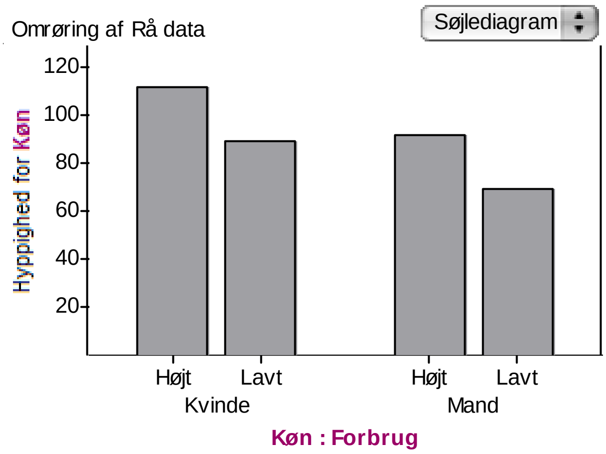
Omrøring af Rå data

	Køn	Forbrug
1	Kvinde	Lavt
2	Kvinde	Lavt
3	Mand	Lavt
4	Mand	Lavt
5	Kvinde	Lavt
6	Kvinde	Lavt
7	Kvinde	Lavt
8	Mand	Lavt
9	Mand	Lavt
10	Kvinde	Lavt

Omrøring af Rå data

tæl ( )

Omrøring af Rå data



Omrøring af Rå data

		Forbrug: Højt	Forbrug: Lavt	Række total
Køn	Kvinde	111 112,22222	89 87,777778	200 200
	Mand	91 89,777778	69 70,222222	160 160
	Søjle total	202 202	158 158	360 360

R1 = tæl ( )
R2 = Forventet

Celler fra oversigtstabel for Omrøring af Rå data

Celler fra oversigtstabel for Omrøring af Rå data

	Køn	Forbrug	R1	R2
1	Kvinde	Højt	111	112,222
2	Kvinde	Lavt	89	87,7778
3	Mand	Højt	91	89,7778
4	Mand	Lavt	69	70,2222

Celler fra oversigtstabel for Omrøring af Rå data

	0,068241634

$$R1 = \text{sum}\left(\frac{(R1 - R2)^2}{R2}\right)$$

Vi kan nu simulere nulhypotesen ved at markere datasættet for **Omrøring af rå data** og taste **CTRL-U** gentagne gange!
For at kunne opbygge fordelingen af teststørrelsen må vi først gemme teststørrelsen som en måling i et datasæt, her det ovenstående **Celler fra oversigtstabel for Omrøring af rå data**. Herefter kan den trækkes ud som gentagne målinger ved at højreklikke på datasættet.

Målinger fra Celler fra oversigtstabel for Omrøring af Rå data

	teststørrelse	<ny>
1	0,00225592	
2	2,38282	
3	0,814388	
4	0,474308	
5	0,474308	
6	0,352488	
7	0,027635	
8	0,027635	
9	0,144379	
10	0,651961	

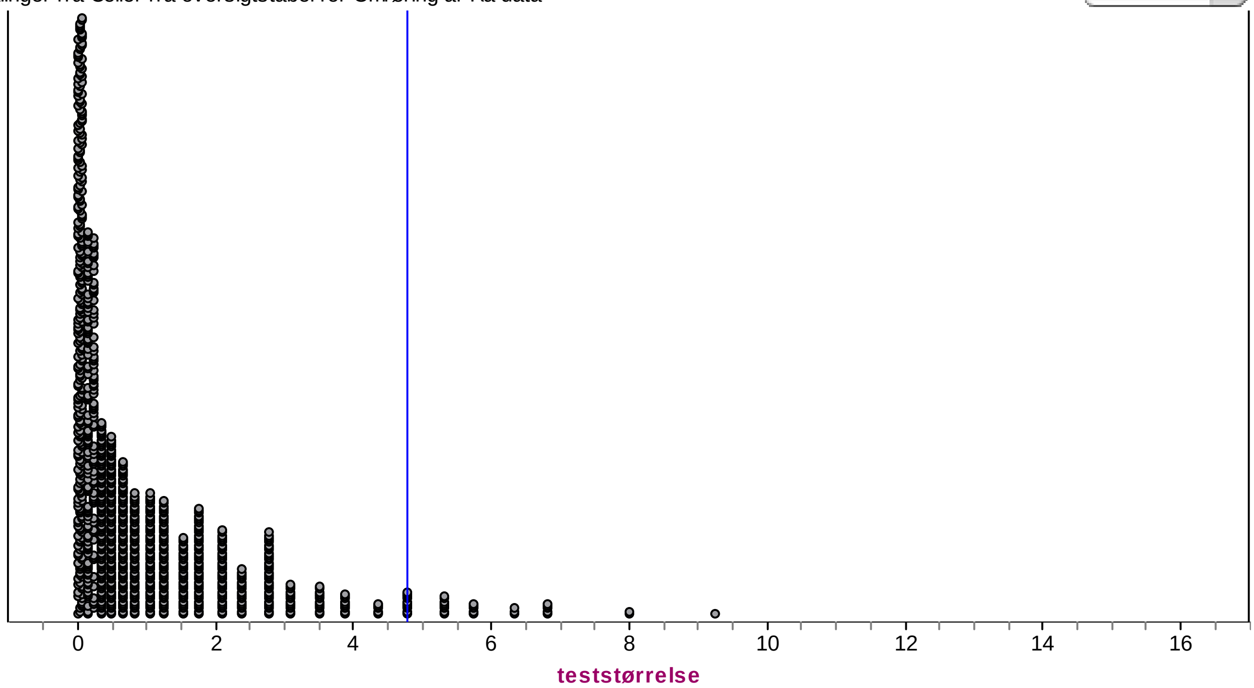
Målinger fra Celler fra oversigtstabel for Omrøring af Rå data

teststørrelse	34

R1 = tæl (teststørrelse ≥ 4,7735)

Målinger fra Celler fra oversigtstabel for Omrøring af Rå data

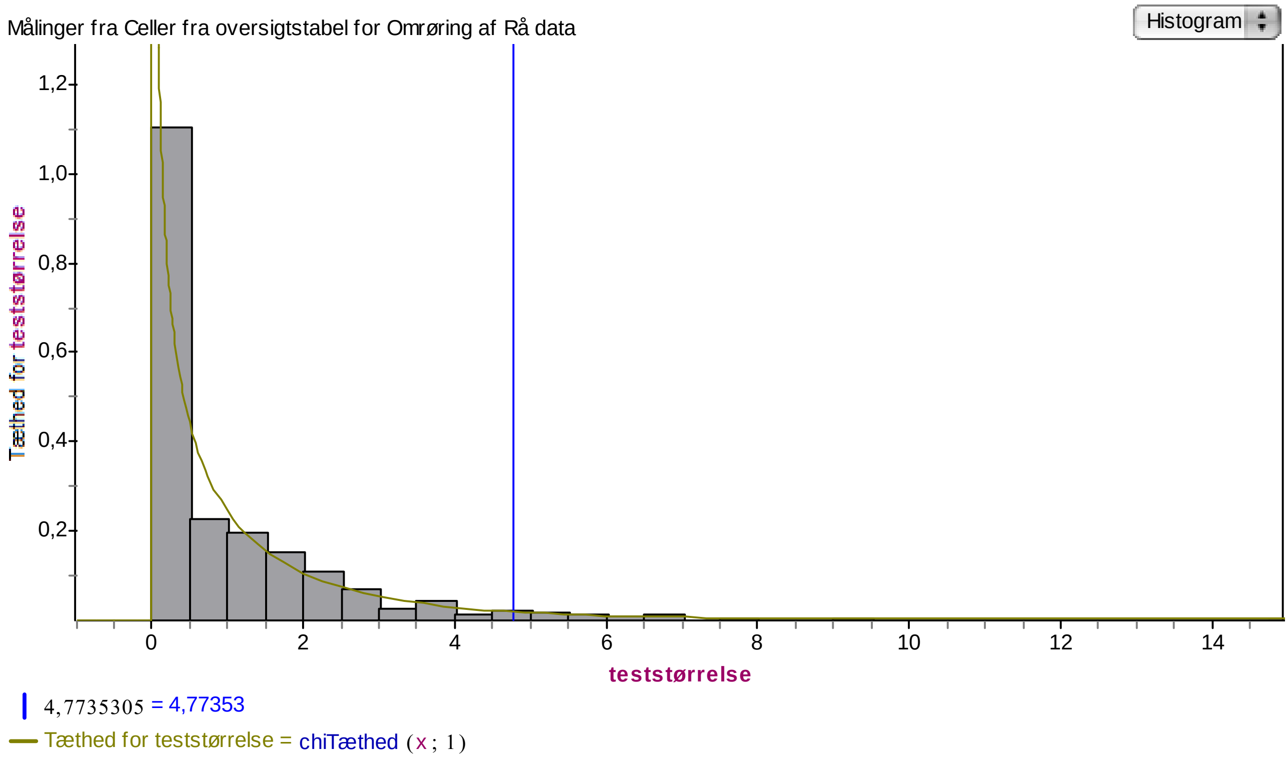
Prikdiagram



4,7735305 = 4,77353

Så skal vi blot have udført simuleringen systematisk rigtigt mange gange. Her har vi udført simuleringen 1000 gange. Prikdiagrammet understreger den grynede natur af chi2-testet. Vi kan da umiddelbart tælle, at der er 34 skæve målinger og dermed estimere p-værdien til ca. 3.4%. Den observerede fordeling er derfor forskellig fra den forventede fordeling på 5%-niveauet.

Vi kan også illustrere testfordelingen med et histogram overlejret med den teoretiske fordeling. Som det ses stemmer den empiriske simulerede fordeling og den teoretiske fordeling fint overens!



Metode 2: Denne gang lægger vi os tættere op af sandsynlighedsregningen og udnytter at sandsynlighedsfordelingen for et mix af to uafhængige variable er givet ved produktfordelingen, dvs. vi ganger de respektive sandsynligheder sammen.
Da vi ikke har fået oplyst sandsynlighedsfordelingen for de enkelte variable **køn** og **forbrug**, estimerer vi dem ud fra den observerede stikprøve:

Køn\Forbrug	<1500 kr./måned	≥ 1500 kr./måned	i alt
kvinder	98	102	200
mænd	60	100	160
i alt	158	202	360

Vi får da igen brug for at frembringe de rå data ud fra den ovenstående tabel, så vi kan trække dem ind i beregningsbokse. Det sker ligesom i det foregående eksempel ved hjælp af **ombyt**-kommandoen. Herefter kan vi få opstillet en krydstabel og få udregnet ikke blot de forventede værdier, der ligger til grund for teststørrelsen, men også de forventede sandsynligheder (ved at dividere med 360)

Rå data_1

	Køn	Forbrug
=	$\text{ombyt}(\text{indeks})\begin{cases}(? \le 200) & : \text{"Kvinde"} \\ \text{ellers} & : \text{"Mand"}\end{cases}$	$\text{ombyt}(\text{indeks})\begin{cases}(? \le 98) & : \text{"Lavt"} \\ (? \le 200) & : \text{"Højt"} \\ (? \le 260) & : \text{"Lavt"} \\ \text{ellers} & : \text{"Højt"}\end{cases}$
1	Kvinde	Lavt
2	Kvinde	Lavt
3	Kvinde	Lavt
4	Kvinde	Lavt
5	Kvinde	Lavt
6	Kvinde	Lavt
7	Kvinde	Lavt

Rå data_1

Rå data_1

		Forbrug: Højt	Forbrug: Lavt	Række total
Køn	Kvinde	102 112,22222 0,3117284	98 87,777778 0,24382716	200 200 0,55555556
	Mand	100 89,777778 0,24938272	60 70,222222 0,19506173	160 160 0,44444444
	Søjle total	202 202 0,56111111	158 158 0,43888889	360 360 1

R1 = tæl ( )
R2 = Forventet
$R3 = \dfrac{\text{Forventet}}{360}$

Celler fra oversigtstabel for Rå data_1

Celler fra oversigtstabel for Rå data_1

	4,7735305

$R1 = \text{sum}\left(\dfrac{(R1 - R2)^2}{R2}\right)$

Her har vi udregnet den observerede teststørrelse!

Celler fra oversigtstabel for Rå data_1

	Køn	Forbrug	R1	R2	R3	kumSand
=						forrige (kumSand; 0) + R3
1	Kvinde	Højt	102	112,222	0,311728	0,311728
2	Kvinde	Lavt	98	87,7778	0,243827	0,555556
3	Mand	Højt	100	89,7778	0,249383	0,804938
4	Mand	Lavt	60	70,2222	0,195062	1

Her har vi suppleret med de forventede sandsynligheder!

Hvis nulhypotesen er korrekt er de to variable **Køn** og **Forbrug** uafhængige og de tilhørende sandsynligheder for den simultane fordeling er derfor netop givet ved produktfordelingen:

{"KvindeHøjt","KvindeLavt","MandHøjt","MandLavt"}

{0.311728, 0.243827, 0.249383, 0.195062}

Her skal vi nu denne gang bruge den kumulerede sandsynlighedsfordeling til at konstruere stikprøven! Vi trækker derfor 360 tilfældige tal mellem 0 og 1 (**Roulette**) og afgør i hvert enkelt tilfælde, hvor det tilfældige tal falder indenfor den kumulerede fordeling. Derved simulerer vi netop produktfordelingen for de to uafhængige variable (dvs. i det væsentlige nulhypotesen).

Rå data_1

	Køn	Forbrug	Mix	Ro
=	ombyt (indeks) $\begin{cases} (? \le 200) & : \text{"Kvinde"} \\ \text{ellers} & : \text{"Mand"} \end{cases}$	ombyt (indeks) $\begin{cases} (? \le 98) & : \text{"Lavt"} \\ (? \le 200) & : \text{"Højt"} \\ (? \le 260) & : \text{"Lavt"} \\ \text{ellers} & : \text{"Højt"} \end{cases}$	sammenkædning (Køn; Forbrug)	tilfæl
1	Kvinde	Lavt	KvindeLavt	0
2	Kvinde	Lavt	KvindeLavt	0
3	Kvinde	Lavt	KvindeLavt	0
4	Kvinde	Lavt	KvindeLavt	0
5	Kvinde	Lavt	KvindeLavt	0
6	Kvinde	Lavt	KvindeLavt	0
7	Kvinde	Lavt	KvindeLavt	0

Vi skal nu have udregnet teststørrelsen hørende til simuleringen. Vi skal derfor have skilt variablen **SimMix** ad i sine to bestanddele: **SimKøn** og **SimForbrug**:

Rå data_1

	Køn	Forbrug	Mix
=	ombyt (indeks) $\begin{cases} (? \le 200) & : \text{"Kvinde"} \\ \text{ellers} & : \text{"Mand"} \end{cases}$	ombyt (indeks) $\begin{cases} (? \le 98) & : \text{"Lavt"} \\ (? \le 200) & : \text{"Højt"} \\ (? \le 260) & : \text{"Lavt"} \\ \text{ellers} & : \text{"Højt"} \end{cases}$	sammenkædning (Køn
1	Kvinde	Lavt	KvindeLavt
2	Kvinde	Lavt	KvindeLavt
3	Kvinde	Lavt	KvindeLavt
4	Kvinde	Lavt	KvindeLavt
5	Kvinde	Lavt	KvindeLavt
6	Kvinde	Lavt	KvindeLavt
7	Kvinde	Lavt	KvindeLavt
8	Kvinde	Lavt	KvindeLavt

Tilsvarende skal vi have udregnet teststørrelsen for simuleringen, men den er mere subtil: Vi kan ikke bare bruge de forventede værdier hørende til produktfordelingen for de observerede hyppigheder, for så er vi jo reelt i gang med at teste om de observerede hyppigheder passer med produktfordelingen, hvilket er en goodness-of-fit test med 3 frihedsgrader for en kendt fordeling. I stedet må vi til hver af de simulerede hyppigheder udregne de tilhørende forventede hyppigheder - ud fra antagelsen om uafhængighed (nulhypotesen). Det var ikke noget problem ved omrøringen, for der holder vi jo fast i marginalhyppighederne. Men denne gang ændres antallet af kvinder, osv. sig i hver simulering. Så nu er der forskel - OG FORSKELLEN ER AFGØRENDE! Vi må derfor oprette en krydstabel for det **simulerede køn** og det **simulerede forbrug** og så på ny udregne teststørrelsen:

Rå data_1

		Sim Forbrug: Højt	Sim Forbrug: Lavt	Række total
Sim Køn	Kvinde	99 105	81 75	180 180
	Mand	111 105	69 75	180 180
	Søjle total	210 210	150 150	360 360

R1 = tæl ( )
R2 = Forventet

Celler fra oversigtstabel for Rå data_1

Celler fra oversigtstabel for Rå data_1

	0,036743442

R1 = $\text{sum}\left(\frac{(R1 - R2)^2}{R2}\right)$

Så er modellen for simulering af nulhypotesen færdigkonstrueret og du kan checke at den fungerer ved at markere datasættet **Rådata_1** og så taste **CTRL U** et passende antal gange. For at kunne opbygge fordelingen af den simulerede teststørrelse må vi først gemme teststørrelsen som en måling i et datasæt, her det ovenstående **Celler fra oversigtstabel for Omrøring af rå data_1**. Herefter kan den trækkes ud som gentagne målinger ved at højreklikke på datasættet.

Målinger fra Celler fra oversigtstabel for Rå data_1

Målinger fra Celler fra oversigtstabel for Rå data_1

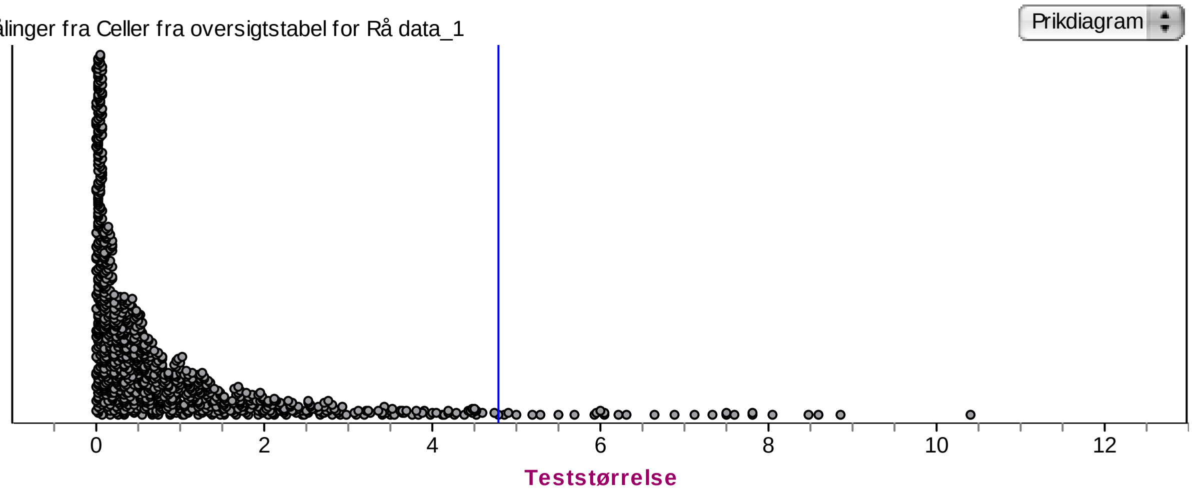
	Teststørrelse	<ny>
1	0,248381	
2	4,85569	
3	0,651443	
4	0,165602	
5	0,320043	

Målinger fra Celler fra oversigtstabel for Rå data_1

Teststørrelse	29

R1 = tæl (Teststørrelse ≥ 4,7735)

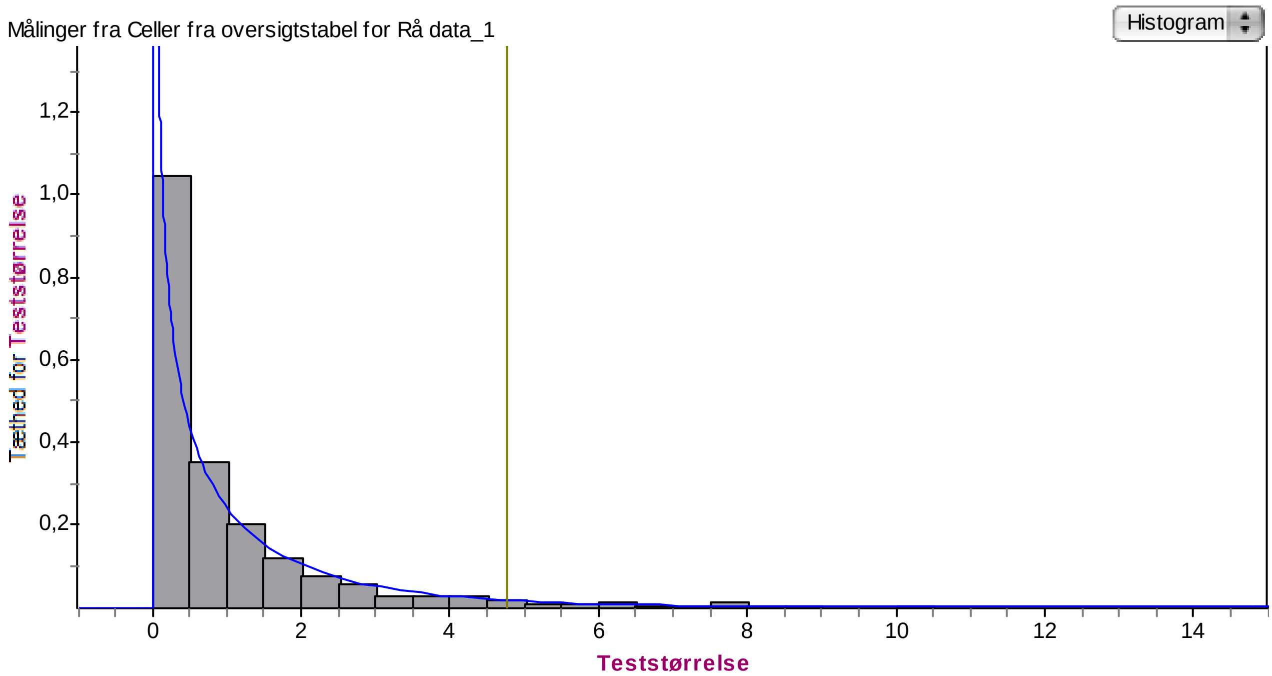
Målinger fra Celler fra oversigtstabel for Rå data_1



4,7735305 = 4,77353

Så skal vi blot have udført simuleringen systematisk rigtigt mange gange. Her har vi udført simuleringen 1000 gange. Vi kan da umiddelbart tælle, at der er 29 skæve målinger og dermed estimere p-værdien til ca. 2.9%. Den observerede fordeling er derfor forskellig fra den forventede fordeling på 5%-niveauet.
Vi kan også illustrere testfordelingen med et histogram overlejret med den teoretiske fordeling. Som det ses stemmer den empiriske simulerede fordeling og den teoretiske fordeling fint overens!

Målinger fra Celler fra oversigtstabel for Rå data_1



Tæthed for Teststørrelse = chiTæthed (x ; 1)
4,7735305 = 4,77353

2b: Goodness of fit test

Eksempel 2: (side 24 i kursusmaterialet)

Indkomstfordelingen i stikprøven var: *I = Indkomst i 1000 kr.*

Observeret antal

I<50	50≤I<100	100≤I<150	150≤I<200	200≤I<300	300≤I<400	400≤I<500	500≤I
98	88	199	136	210	179	52	38

Den forventede fordeling i stikprøven baseret på de ovenstående procenter er tilsvarende givet ved:

Forventet antal

I<50	50≤I<100	100≤I<150	150≤I<200	200≤I<300	300≤I<400	400≤I<500	500≤I
64	93	178	123	243	180	66	53

Sammenholder vi de observerede hyppigheder med de forventede følges de så nogenlunde ad. Men man kunne måske være bekymret for, om de laveste indkomster er overrepræsenteret i stikprøven. Her ligger den observerede hyppighed et godt stykke over den forventede.

Løsning: Vi skal have simuleret nulhypotesen og benytter derfor **Udtag Stikprøve**-kommandoen til at udtrække en stikprøve fra en ideel population, der repræsenterer den forventede fordeling, sådan som den fremgår af tallene fra Danmarks statistik. Vi konstruerer derfor en liste med de forventede indkomstkategorier, dvs. den ideelle population. Vi får da glæde af først at konstruere et datasæt for såvel de observerede hyppigheder som de observerede som de forventede hyppigheder samt de kumulerede forventede hyppigheder. Undervejs udregner vi også den observerede teststørrelse!

Hyppighedsdata

Hyppighedsdata

	indkomst	obs_hyp	forv_hyp	chi2_test	kum_forv
=				$\text{sum}\left(\frac{(\text{obs\_hyp} - \text{forv\_hyp})^2}{\text{forv\_hyp}}\right)$	forv_hyp + forrige (kum_forv; 0)
1	I<=50	98	64	33,8848	64
2	50<=I<100	88	93	33,8848	157
3	100<=I<150	199	178	33,8848	335
4	150<=I<200	136	123	33,8848	458
5	200<=I<300	210	243	33,8848	701
6	300<=I<400	179	180	33,8848	881
7	400<=I<500	52	66	33,8848	947
8	500<=I	38	53	33,8848	1000

Ud fra de kumulerede forventede hyppigheder kan vi nu konstruere en ideel population, der netop afspejler den forventede fordeling.

Rådata forventet

	ideel
=	ombyt (indeks) { (? ≤ 64) : "I<50" (? ≤ 157) : "50<=I<100" (? ≤ 335) : "100<=I<150" (? ≤ 458) : "150<=I<200" (? ≤ 701) : "200<=I<300" (? ≤ 881) : "300<=I<400" (? ≤ 947) : "400<=I<500" ellers : "500<=I"
1	I<50
2	I<50
3	I<50
4	I<50
5	I<50
6	I<50
7	I<50
8	I<50

Rådata forventet

Rådata forventet

ideel	I<50	64
	50<=I<100	93
	100<=I<150	178
	150<=I<200	123
	200<=I<300	243
	300<=I<400	180
	400<=I<500	66
	500<=I	53
	Søjle total	1000

R1 = tæl ( )

Da stikprøven også består af 1000 individer skal vi nu have trukket 1000 individer fra den ideelle population MED tilbagelægning, så hver indkomstkategori hver gang har samme sandsynlighed for at blive udtrukket! Vi kan gentage stikprøven ved at taste **CTRL U** (mens vi har markeret datasættet **Stikprøve fra Rådata forventet**).

Stikprøve fra Rådata forventet

	ideel	<ny
1	200<=I<300	
2	150<=I<200	
3	300<=I<400	
4	100<=I<150	
5	300<=I<400	
6	200<=I<300	
7	400<=I<500	
8	100<=I<150	
9	400<=I<500	
10	200<=I<300	

Stikprøve fra Rådata forventet

Stikprøve fra Rådata forventet

ideel	I<50	59
	50<=I<100	97
	100<=I<150	180
	150<=I<200	122
	200<=I<300	256
	300<=I<400	161
	400<=I<500	69
	500<=I	56
	Søjle total	1000

R1 = tæl ( )

For at kunne udregne teststørrelsen overfører vi først cellerne fra beregningstabellen til et nyt datasæt og tilføjer en beregning af de forventede hyppigheder:

Celler fra oversigtstabel for Stikprøve fra Rådata forventet

Celler fra oversigtstabel for Stikprøve fra Rådata forventet

	ideel	R1	R2
=			ombyt (ideel) { ("I<50") : 64; ("50<=I<100") : 93; ("100<=I<150") : 178; ("150<=I<200") : 123; ("200<=I<300") : 243; ("300<=I<400") : 180; ("400<=I<500") : 66; ellers : 53
1	I<50	59	64
2	50<=I<100	97	93
3	100<=I<150	180	178
4	150<=I<200	122	123
5	200<=I<300	256	243
6	300<=I<400	161	180
7	400<=I<500	69	66
8	500<=I	56	53

Celler fra oversigtstabel for Stikprøve fra Rådat...

	3,6004738

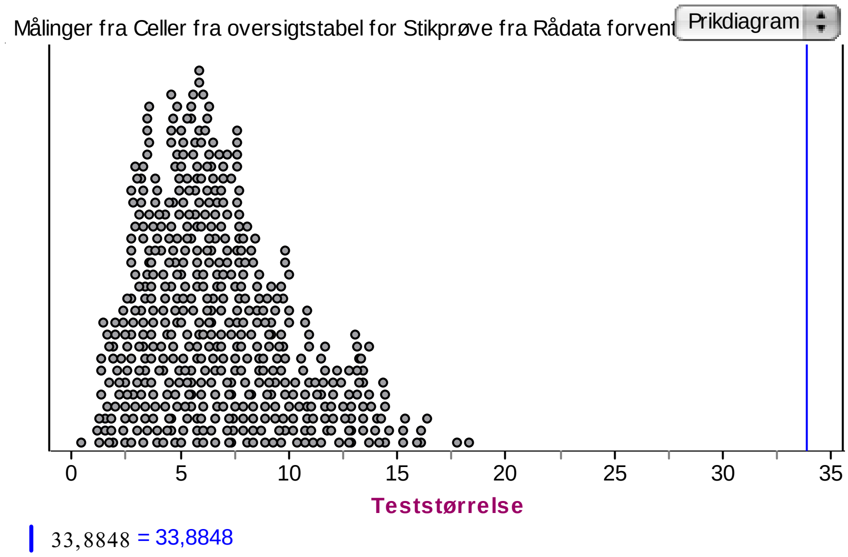
R1 = $\text{sum}\left(\frac{(R1 - R2)^2}{R2}\right)$

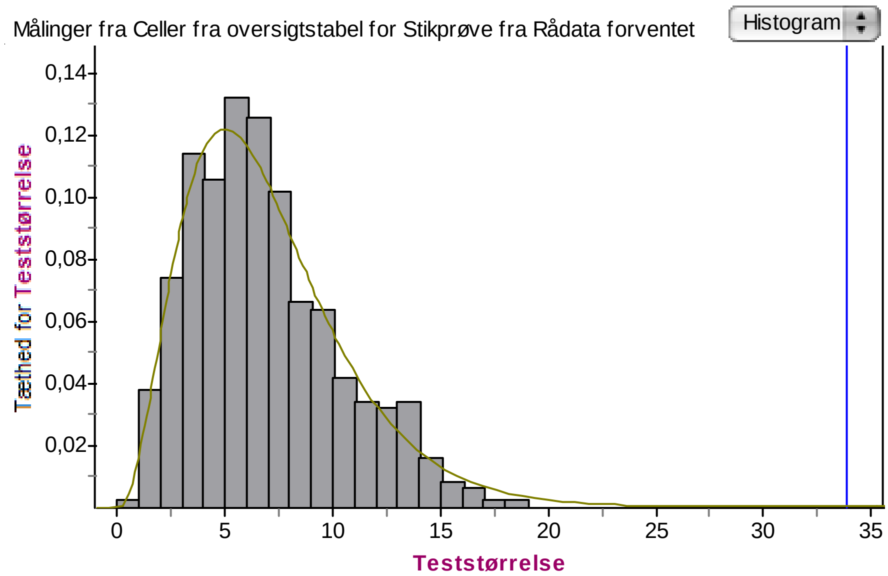
Vi kan så tilføje beregningen af teststørrelsen som en **måling** i datasættet **Celler fra oversigtstabel for Stikprøve fra Rådata forventet**. Herefter kan vi udføre gentagne målinger fra dette datasæt!
Så skal vi blot have udført simuleringen systematisk rigtigt mange gange. Her har vi udført simuleringen 500 gange . Vi kan da umiddelbart se, at der ikke er nogen skæve målinger (det er faktisk uhyre sjældent man fanger en skæv!) og dermed estimere p-værdien til 1/500, dvs. ca. 0.2%. Den observerede fordeling er derfor forskellig fra den forventede fordeling på 5%-niveauet.
Vi kan også illustrere testfordelingen med et histogram overlejret med den teoretiske fordeling. Som det ses stemmer den empiriske simulerede fordeling og den teoretiske fordeling rimeligt overens!

Målinger fra Celler fra oversigtstabel for Stikprøve fra Rådata forventet

Målinger fra Celler fra oversigtstabel for Stikprøve fra Rådata forventet

	Teststørrelse	<ny>
1	14,3937	
2	5,74974	
3	2,43922	
4	5,7328	
5	12,2758	
6	1,86781	
7	1,67923	
8	11,1171	


Målinger fra Celler fra oversigtstabel for Stikprøve fra Rådata forvent
Prikdiagram
0
5
10
15
20
25
30
35
Teststørrelse
33,8848 = 33,8848


Målinger fra Celler fra oversigtstabel for Stikprøve fra Rådata forventet
Histogram
0,14
0,12
0,10
0,08
0,06
0,04
0,02
Tæthed for Teststørrelse
0
5
10
15
20
25
30
35
Teststørrelse

3) Teori: De indbyggede fordelingsfunktioner

De indbyggede fordelingsfunktioner:

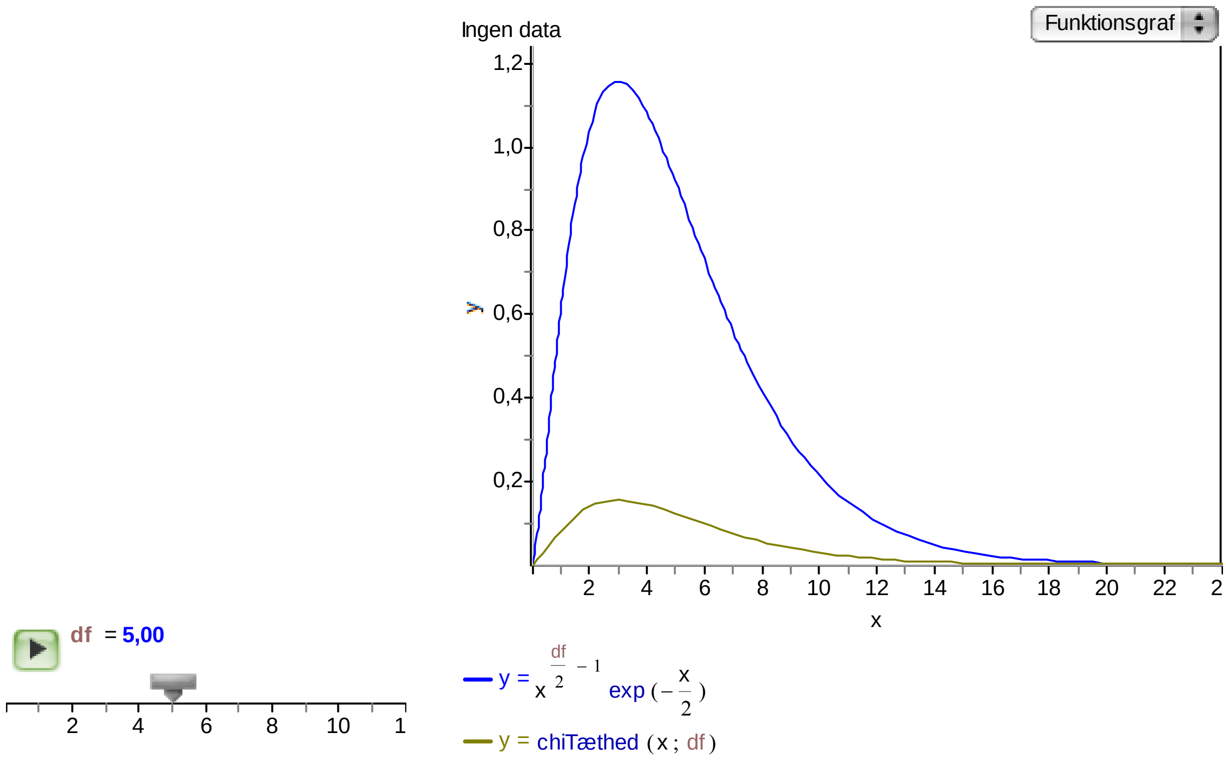
Chi-kvadrat (χ^2)) fordelingen hedder chi. Når man skal arbejde med chi-kvadratfordelingen kan man benytte de følgende operatorer:

y = chiTæthed(x,df):
p = chiSummeret(x,df)
x = chiInv(p,df)

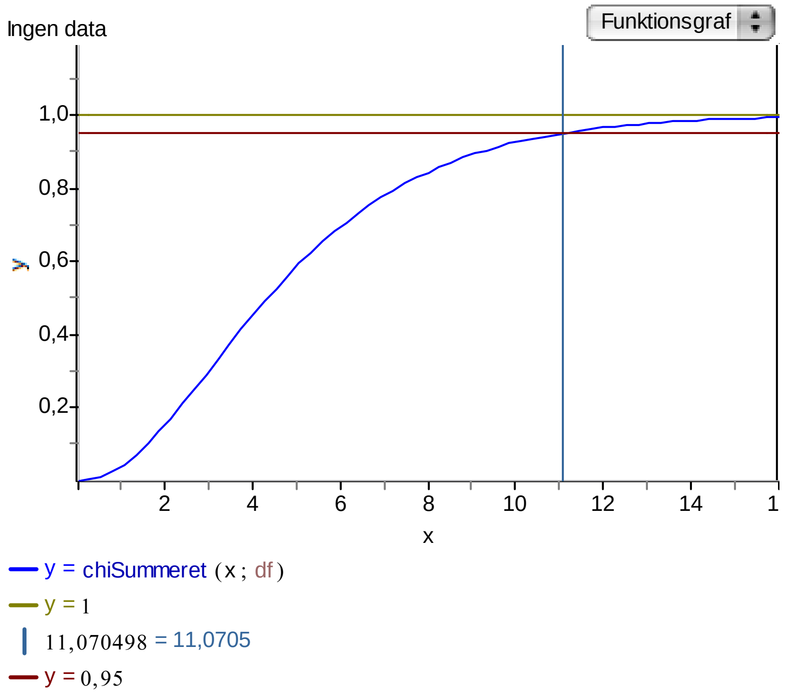
Da DataMeter ikke er et symbolsk CAS-program, kan det ikke for alvor betale sig at undersøge de symbolske forskrifter i DataMeter. Man kan dog nemt eftervise at forskriften for tæthedsfunktionen er proportional med

$(x^{df/2-1})\exp(-x/2)$

Det er også nemt at afbilde tæthedsfunktionen (som har maksimum i $x = df$ -2, dvs. i dette tilfælde 3):



Vi ser dernæst på grafen for den **kumulerede fordeling**:



Endelig kan man finde fraktilerne (den inverse kumulerede fordeling).

Vi ser da at 95% af observationerne ligger under 11.0705, hvis vi har en stokastisk variabel, der er chi-kvadrat fordelt med 5 frihedsgrader.

Ingen data

Slip en variabel her	
	11,070498
	0,95000001

R1 = chiInv (0,95; 5)
R2 = chiSummeret (11,070498; 5)

4) **Teoretiske udregninger hørende til hypotesetest**

4a: Uafhængighed

Eksempel 1: (side 4 i kursusmaterialet)

Køn\Forbrug	<1500 kr./måned	≥ 1500 kr./måned	i alt
kvinder	98	102	200
mænd	60	100	160
i alt	158	202	360

Løsning: Vi skal nu undersøge om der er afhængighed mellem køn og tøjforbrug! Vi får da først og fremmest brug for at beregne de forventede værdier og teststørrelsen.
De forventede værdier udregnes nok nemmest ved at indskrive hyppighedstabellen som krydslister og så anvende de viste formler i en beregningsboks til at finde de forventede værdier.

Hyppighedsdata

	Køn	Forbrug	obs_hyp	forv_hyp	chi2_obs
=					$\text{sum}\left(\frac{(\text{obs\_hyp} - \text{forv\_hyp})^2}{\text{forv\_hyp}}\right)$
1	Kvinder	Lavt	98	87,777778	4,77353
2	Kvinder	Højt	102	112,222222	4,77353
3	Mænd	Lavt	60	70,222222	4,77353
4	Mænd	Højt	100	89,777778	4,77353

Hyppighedsdata

Hyppighedsdata

	87,777778 112,22222 70,222222 89,777778

$$R1 = \frac{\text{sum}(\text{obs\_hyp};\ \text{Køn} = \text{"Kvinder"}) \cdot \text{sum}(\text{obs\_hyp};\ \text{Forbrug} = \text{"Lavt"})}{\text{sum}(\text{obs\_hyp})}$$

$$R2 = \frac{\text{sum}(\text{obs\_hyp};\ \text{Køn} = \text{"Kvinder"}) \cdot \text{sum}(\text{obs\_hyp};\ \text{Forbrug} = \text{"Højt"})}{\text{sum}(\text{obs\_hyp})}$$

$$R3 = \frac{\text{sum}(\text{obs\_hyp};\ \text{Køn} = \text{"Mænd"}) \cdot \text{sum}(\text{obs\_hyp};\ \text{Forbrug} = \text{"Lavt"})}{\text{sum}(\text{obs\_hyp})}$$

$$R4 = \frac{\text{sum}(\text{obs\_hyp};\ \text{Køn} = \text{"Mænd"}) \cdot \text{sum}(\text{obs\_hyp};\ \text{Forbrug} = \text{"Højt"})}{\text{sum}(\text{obs\_hyp})}$$

Vi finder da:
forv_hyp = {87.77778,112.2222,70.22222,89.77778} **chi2_obs** = 4.77353
Herefter er vejen banet for udregning af *p*-værdien, dvs. sandsynligheden for at vi rammer mindst lige så skævt som det observerede, idet vi udnytter at antallet af frihedsgrader er 1:

Ingen data

Slip en variabel her	
	0,02890051

R1 = $1 - $ chiSummeret $(4{,}77353;\ 1)$

Den kritiske sandsynlighed *p* er altså 2.89%, hvorfor afvigelsen er signifikant på 5%-niveau, dvs. vi forkaster nulhypotesen om uafhængighed på 5%-niveauet.

Ingen data

Slip en variabel her	
	3,8414588

R1 = chiInv $(0{,}95;\ 1)$

Vi kunne også have fundet den **kritiske grænse** for teststørrelsen (med signifikansniveauet sat til 5%):

chiinv(0,95,1) = 3.84146

Hvis teststørrelsen ligger over 3,84 er afvigelsen altså kritisk, dvs. vi må forkaste nulhypotesen (på 5%-niveauet).

Hvis man vil have mere 'flydende' beregninger skal man have adgang til de rå data.

Køn\Forbrug	<1500 kr./måned	≥1500 kr./måned	i alt
kvinder	98	102	200
mænd	60	100	160
i alt	158	202	360

Vi opbygger da de rå data ud fra den oplyste tabel over hyppighederne:

Rå data

	Køn	**Forbrug**
=	ombyt (indeks) $\begin{cases} (? \le 200) & : \text{"Kvinde"} \\ \text{ellers} & : \text{"Mand"} \end{cases}$	ombyt (indeks) $\begin{cases} (? \le 98) & : \text{"Lavt"} \\ (? \le 200) & : \text{"Højt"} \\ (? \le 260) & : \text{"Lavt"} \\ \text{ellers} & : \text{"Højt"} \end{cases}$
1	Kvinde	Lavt
2	Kvinde	Lavt
3	Kvinde	Lavt
4	Kvinde	Lavt
5	Kvinde	Lavt
6	Kvinde	Lavt

Rå data

Herefter kan vi få trukket de to variable ind i en beregningsboks og få konstrueret en krydstabel med tilhørende forventede værdier

Celler fra oversigtstabel for Rå data

Rå data

		Forbrug		Række total
		Højt	Lavt	
Køn	Kvinde	102	98	200
		112,22222	87,777778	200
	Mand	100	60	160
		89,777778	70,222222	160
	Søjle total	202	158	360
		202	158	360

R1 = tæl ( )
R2 = Forventet

Celler fra oversigtstabel for Rå data

	Køn	Forbrug	R1	R2
1	Kvinde	Højt	102	112,222
2	Kvinde	Lavt	98	87,7778
3	Mand	Højt	100	89,7778
4	Mand	Lavt	60	70,2222

Krydstabellen overføres til et nyt datasæt ved at højreklikke på den og trække resultaterne ud (overfør celler til nyt datasæt). Herefter kan vi få udregnet teststørrelsen og alt er nu som før - en beregning af *p*-værdien viser umiddelbart nulhypotesen om uafhængighed må forkastes på det statistiske signifikansniveau 5%.

Celler fra oversigtstabel for Rå data

	4,7735305

R1 = $\text{sum}\left(\frac{(R1 - R2)^2}{R2}\right)$

Ingen data

Slip en variabel her	
	0,02890051

R1 = $1 - \text{chiSummeret}(4{,}77353;\ 1)$

4b Goodness of fit

Eksempel 2: (side 24 i kursusmaterialet)

Indkomstfordelingen i stikprøven var: *I = Indkomst i 1000 kr.*

Observeret antal

I<50	50≤I<100	100≤I<150	150≤I<200	200≤I<300	300≤I<400	400≤I<500	500≤I
98	88	199	136	210	179	52	38

Den forventede fordeling i stikprøven baseret på de ovenstående procenter er tilsvarende givet ved:

Forventet antal

I<50	50≤I<100	100≤I<150	150≤I<200	200≤I<300	300≤I<400	400≤I<500	500≤I
64	93	178	123	243	180	66	53

Sammenholder vi de observerede hyppigheder med de forventede følges de så nogenlunde ad. Men man kunne måske være bekymret for, om de laveste indkomster er overrepræsenteret i stikprøven. Her ligger den observerede hyppighed et godt stykke over den forventede.

Løsning: Vi skal undersøge om den observerede fordeling følger den forventede. Vi overfører derfor data til lister og udregner teststørrelsen. Vi finder da **chi2_obs**=33,8848.

Hyppighedsdata

	Indkomst	obs_hyp	forv_hyp	chi2_obs
=				$\text{sum}\left(\frac{(\text{obs\_hyp} - \text{forv\_hyp})^2}{\text{forv\_hyp}}\right)$
1	I<50	98	64	33,8848
2	50<=I<100	88	93	33,8848
3	100<=I<150	199	178	33,8848
4	150<=I<200	136	123	33,8848
5	200<=I<300	210	243	33,8848
6	300<=I<400	179	180	33,8848
7	400<=I<500	52	66	33,8848
8	500<I	38	53	33,8848

Hyppighedsdata

Herefter er vejen banet for udregning af *p*-værdien, dvs. sandsynligheden for at vi rammer mindst lige så skævt som det observerede, idet vi udnytter at antallet af frihedsgrader er 1:

Ingen data

Slip en variabel her	
	1,8100966e-05

R1 = $1 - \text{chiSummeret}(33{,}8848;\ 7)$

Den kritiske sandsynlighed *p* er altså 0,0018%, hvorfor afvigelsen er signifikant på 1%-niveau, dvs. vi forkaster nulhypotesen om stikprøvens repræsentativitet på 1%-niveauet.

Ingen data

Slip en variabel her	
	18,475307

R1 = $\text{chiInv}(0{,}99;\ 7)$

Vi kunne også have fundet den **kritiske grænse** for teststørrelsen (med signifikansniveauet sat til 1%):

chiinv(0,99,7) = 18;475307

Hvis teststørrelsen ligger over 18,48 er afvigelsen altså **kritisk**, dvs. vi må forkaste nulhypotesen (på 1%-niveauet).

5) De indbyggede testrutiner

5a: Uafhængighed

Eksempel 1: (side 4 i kursusmaterialet)

Køn\Forbrug	<1500 kr./måned	≥ 1500 kr./måned	i alt
kvinder	98	102	200
mænd	60	100	160
i alt	158	202	360

Løsning: Vi skal afgøre om de oplyste data er i rimelig overensstemmelse med nulhypotesen om uafhængighed mellem **Køn** og **Forbrug**. Vi benytter det indbyggede test for uafhængighed af to variable , der forventer at få oplyst krydstabellen for de observerede hyppigheder:

Stikprøveresultater | Test af uafhængighed

Første variabel (kategoriseret): Ikke tildelt

Anden variabel (kategoriseret): Ikke tildelt

		Forbrug		Række total
		Lavt	Højt	
Køn	Kvinder	98 (87,8)	102 (112,2)	200
	Mænd	60 (70,2)	100 (89,8)	160
	Søjle total	158	202	360

Første variabel: **Forbrug**
Antal kategorier: **2**
Anden variabel: **Køn**
Antal kategorier: **2**
Alternativ hypotese: Der er en sammenhæng mellem **Forbrug** og **Køn**

Teststørrelsen, chi-i-anden, er **4,774**. Der er **1** frihedsgrader (antallet af rækker minus én ganget med antallet af søjler minus én).

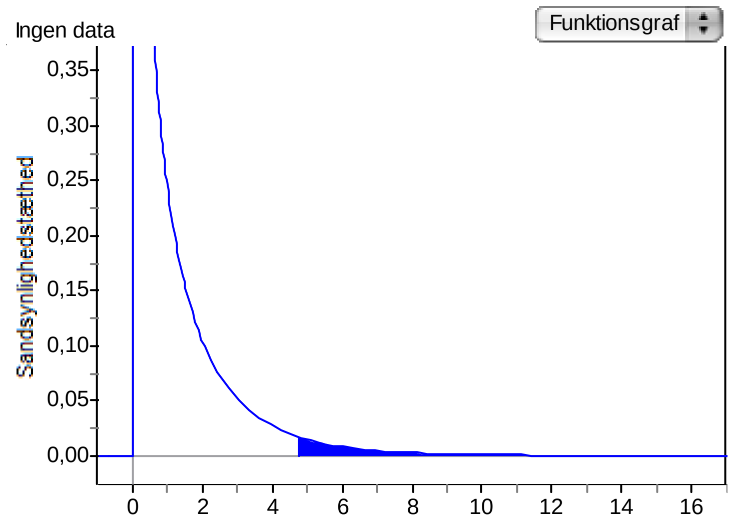
Hvis det var sandt at **Forbrug** var uafhængig af **Køn** (nulhypotesen), og stikprøven blev gentaget mange gange, så ville sandsynligheden for at få en værdi for chi-i-anden, der var mindst lige så stor være **0,029**.

Tallene i parentes i tabellen er de forventede antal.

Vi får da som vist en række koncentrerede oplysninger:
1) Teststørrelsen har værdien 4.774.
2) Teststørrelsen er chi2 fordelt med 1 frihedsgrad.
3) *p*-værdien er 2.9%, dvs. sandsynligheden for at finde en teststørrelse, der er mindst lige så skæv som den observerede er 2.89%. Nulhypotesen forkastes altså på signifikansniveauet 5%, men den forkastes ikke på 1% niveau!
4) Vi får adgang til krydstabellen for de forventede værdier:

Køn\Forbrug	<1500 kr./måned	≥ 1500 kr./måned	i alt
kvinder	87,8	112,2	200
mænd	70,2	89,8	160
i alt	158	202	360

5) Vi får også adgang til en grafisk fremstilling for chi2-testet (højreklik i testboksen)



Bemærkning: Den ovenstående tabel er sat til at være **ordrig**! Slår man det fra får man i stedet den kompakte version:

Første variabel: Forbrug
Antal kategorier: 2
Anden variabel: Køn
Antal kategorier: 2
Ho: Forbrug er uafhængig af Køn
Chi-i-anden: 4,774
Frihedsgr.: 1
p-værdi: 0,029

Tallene i parentes er de forventede antal.

5b: Goodness of fit

Eksempel 2: (side 24 i kursusmaterialet)
Indkomstfordelingen i stikprøven var: *I = Indkomst i 1000 kr.*
Observeret antal

I<50	50≤I<100	100≤I<150	150≤I<200	200≤I<300	300≤I<400	400≤I<500	500≤I
98	88	199	136	210	179	52	38

Den forventede fordeling i stikprøven baseret på de ovenstående procenter er tilsvarende givet ved:
Forventet antal

I<50	50≤I<100	100≤I<150	150≤I<200	200≤I<300	300≤I<400	400≤I<500	500≤I
64	93	178	123	243	180	66	53

Sammenholder vi de observerede hyppigheder med de forventede følges de så nogenlunde ad. Men man kunne måske være bekymret for, om de laveste indkomster er overrepræsenteret i stikprøven. Her ligger den observerede hyppighed et godt stykke over den forventede.

Løsning: Vi skal afgøre om de oplyste data er i rimelig overensstemmelse med nulhypotesen om samme fordeling for stikprøve og population. Vi benytter det indbyggede test for en fordeling, der forventer at få oplyst lister for dels de observerede hyppigheder, dels de forventede sandsynligheder.

Stikprøveresultater — Test af en fordeling

Variabel: (kategoriseret): Ikke tildelt

		Antal	Sandsynlighed
indkomst	I<50	98	0,064
	50<=I<100	88	0,093
	100<=I<150	199	0,178
	150<=I<200	136	0,123
	200<=I<300	210	0,243
	300<=I<400	179	0,180
	400<=I<500	52	0,066
	500<=I	38	0,053
	Søjle total	1000	1,000

Variabel: **indkomst**
Antal kategorier: **8**
Alternativ hypotese: Kategorierne for **indkomst har andre sandsynligheder end de ovenstående**

Teststørrelsen, chi-i-anden, er **33,88**. Der er **7** frihedsgrader (én mindre end antallet af kategorier).

Hvis det var sandt at kategorierne for **indkomst havde de ovenstående sandsynligheder** (nulhypotesen), og stikprøven blev gentaget mange gange, så ville sandsynligheden for at få en værdi for chi-i-anden som var mindst lige så stor være **< 0,0001**.

Vi får da som vist en række oplysninger:

1) Teststørrelsen har værdien 33.88
2) Teststørrelsen er chi2 fordelt med 7 frihedsgrader
3) p-værdien er 0.0018%, dvs. sandsynligheden for at finde en teststørrelse mindst lige så stor som den observerede er under 0.01%. Da den ligger under signifikansniveauet 1% forkastes nulhypotesen på signifikansniveauet 1%.
4) Vi får også adgang til en grafisk fremstilling for chi2-testet (højreklik i testboksen), men vi kan ikke se det kritiske område, fordi testdata er så ekstreme!

Den ovenstående tabel er sat til at være ordrig! Slår man det fra får man i stedet den kompakte version:

Ho: Kategorierne fra indkomst har de ovenstående sandsynligheder
Antal kategorier: 8
Chi-i-anden: 33,88
Frihedsgrader: 7
p-værdi: < 0,0001