

# Værktøjshjælp for TI-Interactive

## 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

1) <b>Eksempler på grafisk fremstilling af data</b>	<b>side 1</b>
(Beskrivende statistik – Explorative Data Analysis)	
1a: Uafhængighed	side 1
1b: Goodness of fit	side 1
2) <b>Eksperimentel hypotesetest</b>	<b>side 3</b>
2a: Uafhængighed	side 3
Metode 1: Simulering ud fra omrøring	side 3
Metode 2: Simulering ud fra produktfordeling	side 6
2b: Goodness of Fit	side 11
3) <b>Teori: De indbyggede fordelingsfunktioner</b>	<b>side 15</b>
4) <b>Teoretiske udregninger hørende til hypotesetest</b>	<b>side 18</b>
4a: Uafhængighed	side 18
4b: Goodness of Fit	side 19
5) <b>Indbyggede testrutiner</b>	<b>side 21</b>
5a: Uafhængighed	side 21
5b: Goodness of Fit	side 22

Følgende **TI-Interactive**-filer følger med:

**Afsnit 1:** Eksempler på grafiske fremstillinger.

**Afsnit 2:** Eksperimentel hypotesetest Ia, Ib, II (skal åbnes med tålmodighed!)

**Afsnit 3:** De indbyggede fordelingsfunktioner

**Afsnit 4:** Teoretiske udregninger

**Afsnit 5:** De indbyggede test

**Supplerende note:** Nogle nyttige kommandoer i **TI-Interactive**.

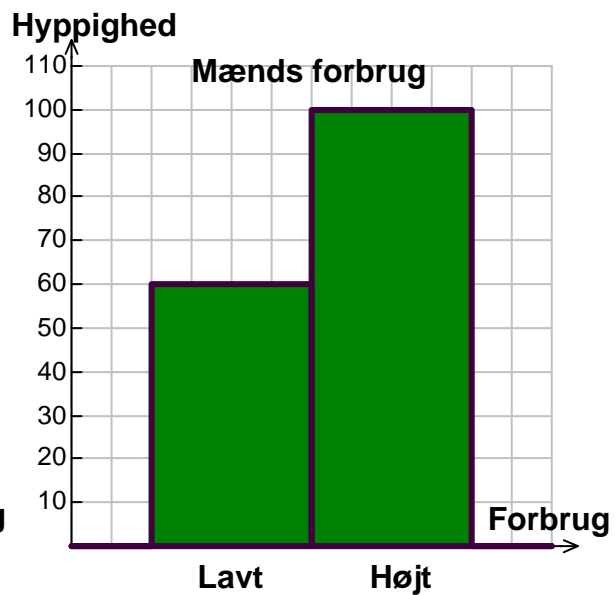
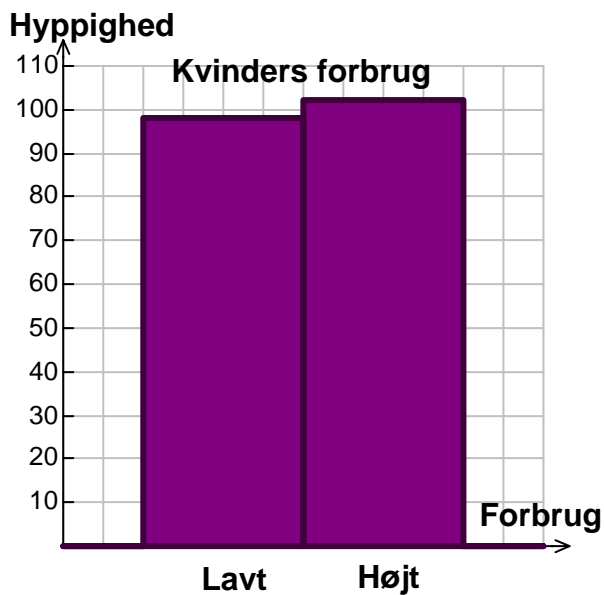
# 1) Eksempler på grafisk fremstilling af data (til brug for den beskrivende statistik – Explorative Data Analysis)

## 1a: Uafhængighed Eksempel 1: (side 4 i kursusmaterialet)

Koen\Toejforbrug	<1500kr./maaned	≥ 1500kr./maaned	I alt
Kvinde	98	102	200
Mand	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 liste-editoren og overføre dem til grafer som histogrammer med tilhørende hyppigheder. Det ses da tydeligt at mænds forbrugsmønster i den pågældende stikprøve (hvor de to blokke har meget forskellig højde) ser helt anderledes ud end kvinders forbrugsmønster (hvor de to blokke er næsten lige høje).

kat forbrug	Kvinder	Mænd
1	98	60
2	102	100



## 1b Goodness of fit Eksempel 2 (side 24 i kursusmaterialet)

Indkomstfordelingen i stikprøven var:  $I = \text{Indkomst i 1000 kr.}$

Observeret antal

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

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

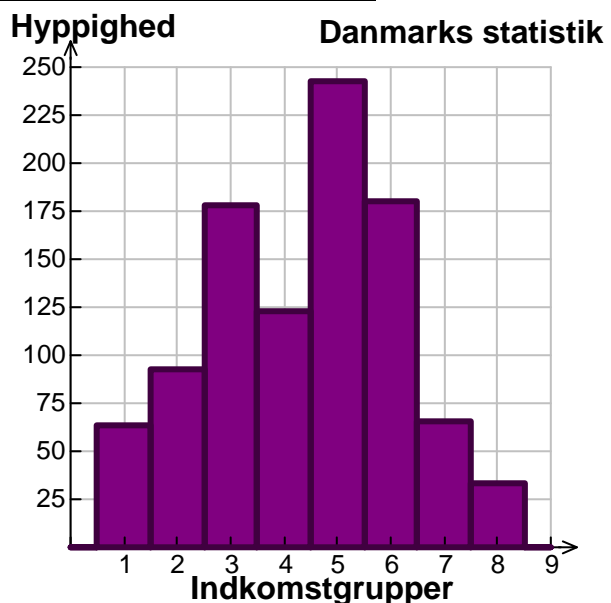
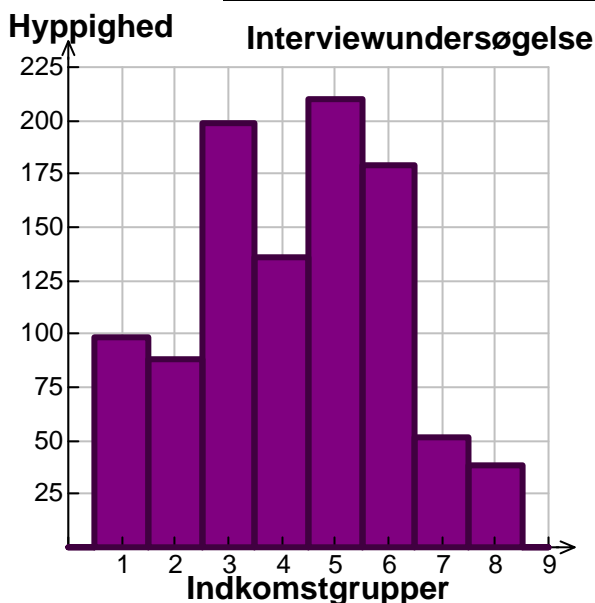
Forventet antal

$I < 50$	$50 \leq I < 100$	$100 \leq I < 150$	$150 \leq I < 200$	$200 \leq I < 300$	$300 \leq I < 400$	$400 \leq I < 500$	$500 \leq I$
64	93	178	123	243	180	66	33

Sammenholder vi de observerede hyppigheder med de forventede følger 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 liste-editoren overføre dem til histogrammer med tilhørende hyppigheder. 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!

kat indkomst	obs hyp	forv hyp
1	98	64
2	88	93
3	199	178
4	136	123
5	210	243
6	179	180
7	52	66
8	38	33



## 2) Eksperimentel hypotesetest

### 2a: Uafhængighed Eksempel 1: (side 4 i kursusmaterialet)

Koen\Toejforbrug	<1500kr./maaned	≥ 1500kr./maaned	I alt
Kvinde	98	102	200
Mand	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 indsætter først et hjælpebibliotek, der rummer de nødvendige funktioner til at gennemføre de ønskede udregninger med krydstabeller/matricer:

$$\text{rowtotal}(Matrix) := \text{seq} \left( \sum_{j=1}^{\text{colDim}(Matrix)} (Matrix_{[i, j]}), i, 1, \text{rowDim}(Matrix) \right)$$

$$\text{colTotal}(Matrix) := \text{seq} \left( \sum_{i=1}^{\text{rowDim}(Matrix)} (Matrix_{[i, j]}), j, 1, \text{colDim}(Matrix) \right)$$

$$\text{grandTotal}(Matrix) := \sum_{j=1}^{\text{colDim}(Matrix)} \left( \sum_{i=1}^{\text{rowDim}(Matrix)} (Matrix_{[i, j]}) \right)$$

#### Expected(obs):

→ "Done"

$$\text{chiInd}(obsMatrix, expMatrix) :=$$

$$\sum_{j=1}^{\text{colDim}(obsMatrix)} \left( \sum_{i=1}^{\text{rowDim}(obsMatrix)} \left( \frac{(obsMatrix_{[i, j]} - expMatrix_{[i, j]})^2}{ExpMatrix_{[i, j]}} \right) \right)$$

#### FreqToList(list):

→ "Done"

→

#### Scramble(list):

"Done"

#### NewMatrix(rowList, colList):

→ "Done"

#### Simul(Matrix):

→ "Done"

Vi konstruerer først en krydstabel for kombinationen af køn og forbrug der er i overensstemmelse med de oplyste hyppigheder.

$$obs := \begin{bmatrix} 98 & 102 \\ 60 & 100 \end{bmatrix} \rightarrow \begin{bmatrix} 98 & 102 \\ 60 & 100 \end{bmatrix}$$

$$forv := \text{expected}(obs) \rightarrow \begin{bmatrix} 87.7778 & 112.222 \\ 70.2222 & 89.7778 \end{bmatrix}$$





$$\text{Simul}(obs) \rightarrow \begin{bmatrix} 92 & 108 \\ 66 & 94 \end{bmatrix}$$

$$\text{chiInd}(\text{Simul}(obs), \text{expected}(obs)) \rightarrow .027635$$

Så skal vi blot have udført simuleringen systematisk rigtigt mange gange. Det sker ved hjælp af en seq-kommando, som vi først udfører 10 gange for at se den fungerer som den skal!

$$\text{seq}(\text{chiInd}(\text{simul}(obs), \text{expected}(obs)), i, 1, 10)$$

$$\rightarrow \{.027635, .027635, 1.0428, .027635, .474308, .068242, .474308, 1.525, .474308, .068242\}$$

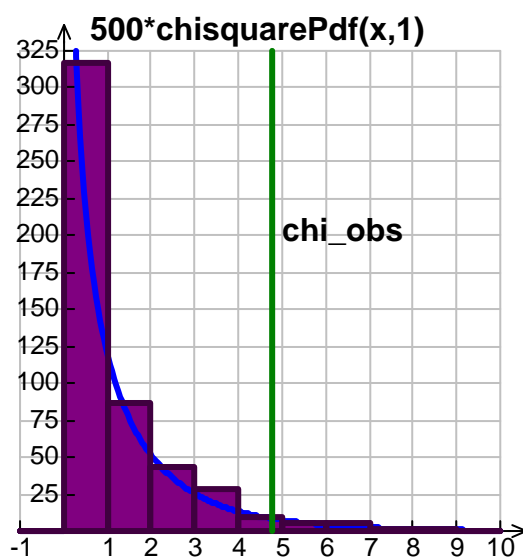
Herefter har vi med stor tålmodighed udført simuleringen 500 gange:

$$\text{test} := \text{seq}(\text{chiInd}(\text{simul}(obs), \text{expected}(obs)), i, 1, 500)$$

$$\text{crit} := \sum_{i=1}^{\dim(\text{test})} \begin{cases} 1 & \text{test}_{[i]} \geq \text{chi\_obs} \\ 0 & \text{else} \end{cases} \rightarrow 18$$

Vi kan da umiddelbart tælle, at der er 18 skæve målinger og dermed estimere  $p$ -værdien til  $18/500$ , dvs. ca. 3.6%. 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. Først indsætter vi dog hjælpebiblioteket med matrix-kommandoer, der skal holde styr på krydstabellerne:

$$\text{rowTotal}(Matrix) := \text{seq} \left( \sum_{j=1}^{\text{colDim}(Matrix)} (Matrix_{[i, j]}), i, 1, \text{rowDim}(Matrix) \right)$$

$$\text{colTotal}(Matrix) := \text{seq} \left( \sum_{i=1}^{\text{rowDim}(Matrix)} (Matrix_{[i, j]}), j, 1, \text{colDim}(Matrix) \right)$$

$$\text{grandTotal}(Matrix) := \sum_{j=1}^{\text{colDim}(Matrix)} \left( \sum_{i=1}^{\text{rowDim}(Matrix)} (Matrix_{[i, j]}) \right)$$

**expected(obsMatrix):** → "Done"

$$\text{chiInd}(obsMatrix, expMatrix) := \sum_{j=1}^{\text{colDim}(obsMatrix)} \left( \sum_{i=1}^{\text{rowDim}(obsMatrix)} \left( \frac{(obsMatrix_{[i, j]} - expMatrix_{[i, j]})^2}{ExpMatrix_{[i, j]}} \right) \right)$$

**ListToFreq(List):** → "Done"

**FreqToList(FreqList):** → "Done"

Koen\Toejforbrug	<1500kr./maaned	≥ 1500kr./maaned	I alt
Kvinde	98	102	200
Mand	60	100	160
I alt	158	202	360

De observerede hyppigheder skrives ind i en krydstabel og de observerede sandsynligheder for simultanfordelingen udregnes som vist:

$$obs := \begin{bmatrix} 98 & 102 \\ 60 & 100 \end{bmatrix} \rightarrow \begin{bmatrix} 98 & 102 \\ 60 & 100 \end{bmatrix}$$

$$chi\_obs := \text{chiInd}(obs, \text{expected}(obs)) \rightarrow 4.77353$$

$$prob := \frac{\text{expected}(obs)}{\text{grandTotal}(obs)} \rightarrow \begin{bmatrix} .243827 & .311728 \\ .195062 & .249383 \end{bmatrix}$$

$$probList := \text{MatToList}(prob) \rightarrow \{.243827, .311728, .195062, .249383\}$$

$$kumList := \text{cumSum}(probList) \rightarrow \{.243827, .555556, .750617, 1.\}$$

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

$$\begin{bmatrix} \text{KvindeLavt} & \text{KvindeHoejt} \\ \text{MandLavt} & \text{MandHoejt} \end{bmatrix} = \begin{bmatrix} .243827 & .311728 \\ .195062 & .249383 \end{bmatrix}$$

Her skal vi nu denne gang bruge den kumulerede fordeling 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).

*Roulette* := rand(grandTotal(*obs*))

$$\text{Simulering} := \text{seq} \left( \begin{array}{l} 1 \quad \text{Roulette}_{[i]} \leq \text{kumList}_{[1]} \\ 2 \quad \text{Roulette}_{[i]} \leq \text{kumList}_{[2]}, i, 1, \dim(\text{Roulette}) \\ 3 \quad \text{Roulette}_{[i]} \leq \text{kumList}_{[3]} \\ 4 \quad \text{else} \end{array} \right)$$

→ {3, 1, 3, 4, 3, 2, 3, 4, 1, 1, 1, 4, 2, 2, 3, 1, 3, 4, 3, 1, 3, 4, 4, 2, 2, 4, 2, 3, 3, 3, 1, 2, 3, 2, 4, 3, 4, 2, 1, 1, 1, 4, 1, 4, 4, 4, 2, 4, 3, 4, 3, 2, 1, 2, 1, 2, 1, 3, 1, 1, 2, 2, 2, 3, 1, 3, 2, 2, 4, 1, 3, 2, 3, 2, 1, 1, 3, 2, 4, 4, 3, 1, 3, 2, 4, 1, 4, 4, 1, 4, 1, 3, 1, 1, 1, 2, 4, 1, 3, 1, 1, 3, 1, 2, 3, 2, 4, 4, 2, 2, 1, 4, 1, 1, 2, 2, 3, 2, 4, 2, 4, 2, 2, 1, 4, 4, 4, 4, 2, 3, 4, 2, 1, 2, 2, 2, 2, 3, 1, 1, 2, 3, 3, 3, 4, 2, 4, 1, 4, 2, 1, 4, 2, 4, 4, 4, 1, 2, 1, 3, 2, 4, 3, 3, 2, 2, 2, 2, 1, 2, 4, 4, 4, 4, 2, 1, 4, 3, 4, 3, 4, 1, 4, 3, 1, 2, 4, 3, 2, 1, 2, 4, 2, 3, 2, 1, 4, 4, 3, 4, 1, 1, 3, 2, 3, 1, 2, 3, 1, 1, 1, 3, 1, 2, 3, 1, 3, 4, 4, 1, 1, 1, 1, 2, 1, 1, 1, 3, 2, 3, 3, 2, 4, 3, 2, 1, 4, 2, 2, 4, 4, 2, 1, 1, 4, 4, 4, 2, 1, 1, 2, 2, 1, 1, 3, 1, 4, 1, 1, 3, 1, 1, 4, 1, 2, 1, 4, 4, 2, 2, 2, 2, 4, 2, 1, 4, 3, 1, 1, 4, 2, 1, 2, 2, 1, 1, 3, 3, 2, 3, 1, 1, 2, 3, 1, 3, 2, 1, 2, 2, 2, 3, 4, 3, 2, 1, 4, 2, 3, 1, 4, 2, 4, 4, 2, 2, 4, 1, 4, 1, 3, 4, 2, 3, 3, 2, 1, 1, 1, 1, 1, 2, 2, 4, 3, 2, 1, 2, 2, 2, 4, 2, 1, 3, 3, 3, 4, 4, 2, 2, 3, 1, 2, 3, 2, 3, 3, 3, 1, 1}

Vi skal så have omdannet denne liste over kategorier til en krydstabel over hyppighederne:

$$\text{sim\_hyp} := \text{ListToFreq}(\text{Simulering}) \quad \rightarrow \quad \{101, 102, 75, 82\}$$

$$\text{sim} := \text{ListToMat}(\text{sim\_hyp}, \text{rowDim}(\text{obs})) \quad \rightarrow \quad \begin{bmatrix} 101 & 102 \\ 75 & 82 \end{bmatrix}$$

$$\text{sim\_exp} := \text{expected}(\text{sim}) \quad \rightarrow \quad \begin{bmatrix} 99.2444 & 103.756 \\ 76.7556 & 80.2444 \end{bmatrix}$$

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 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!

$$\text{chi\_sim} := \text{chiInd}(\text{sim}, \text{expected}(\text{sim})) \quad \rightarrow \quad .139319$$

Det er klart at vi kan gentage simuleringen ved at pakke de sidste kommandoer sammen i en matematik-boks:

```
Roulette := rand(grandTotal(obs)) :: Simulering := seq
(
  1  Roulette[i] ≤ kumList[1]
  2  Roulette[i] ≤ kumList[2], i, 1, dim(Roulette)
  3  Roulette[i] ≤ kumList[3]
  4  else
) :: sim_hyp := ListToFreq(Simulering) :: sim
:= ListToMat(sim_hyp, rowDim(obs)) :: sim_exp := expected(sim) :: chi_sim := chiInd
(sim, expected(sim))
→ .490624
```

```
Roulette := rand(grandTotal(obs)) :: Simulering := seq
(
  1  Roulette[i] ≤ kumList[1]
  2  Roulette[i] ≤ kumList[2], i, 1, dim(Roulette)
  3  Roulette[i] ≤ kumList[3]
  4  else
) :: sim_hyp := ListToFreq(Simulering) :: sim
:= ListToMat(sim_hyp, rowDim(obs)) :: sim_exp := expected(sim) :: chi_sim := chiInd
(sim, expected(sim))
→ .39828
```

Men skal det virkelig batte noget, så skal vi have udført simuleringen systematisk rigtigt mange gange. Det kræver konstruktion af en brugerdefineret kommando, der pakker de ovenstående kommandoer:

```
Define simul(obsMatrix, KumListe) = Func
:: Local RouletteLoc, SimulLoc, i, k, simHypLoc, simLoc
:: RouletteLoc := rand(grandTotal(obsMatrix))
:: SimulLoc := newList(dim(KumListe))
:: For i, 1, dim(RouletteLoc)
:: k := 1
:: SimulLoc[i] := 1
:: While k ≤ dim(KumListe) and RouletteLoc[i] ≥ KumListe[k]
:: k := k + 1
:: SimulLoc[i] := SimulLoc[i] + 1
:: EndWhile
:: EndFor
:: simHypLoc := ListToFreq(SimulLoc)
:: simLoc := ListToMat(simHypLoc, rowDim(obsMatrix))
:: chiInd(simLoc, expected(simLoc))
:: EndFunc
→ "Done"
```

```
simul(obs, kumList) → .296534
```

$$\text{seq}(\text{simul}(\text{obs}, \text{kumList}), i, 1, 10) \rightarrow \{.575299, .008181, 1.29854, .300151, .018313, 3.33209, 6.19936, .208765, .119659, 2.98803\}$$

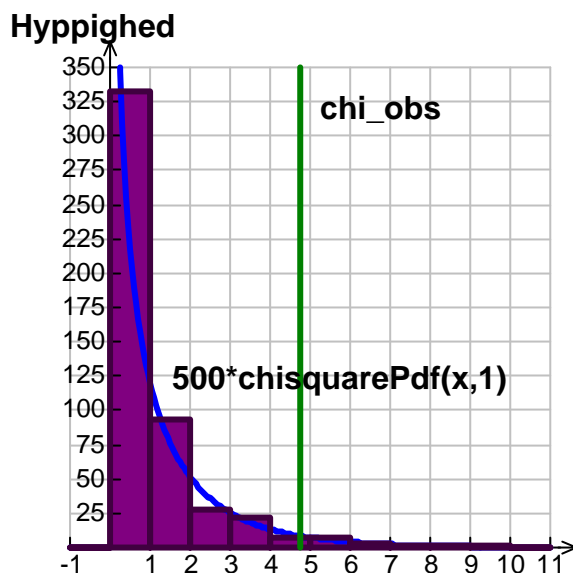
$$\text{test} := \text{seq}(\text{simul}(\text{obs}, \text{kumList}), i, 1, 500)$$

Her har vi med stor tålmodighed udført simuleringen 500 gange.

$$\text{crit} := \sum_{i=1}^{\dim(\text{test})} \begin{cases} 1 & \text{test}_{[i]} \geq \text{chi\_obs} \\ 0 & \text{else} \end{cases} \rightarrow 18$$

Vi kan da umiddelbart tælle, at der er 18 skæve målinger og dermed estimere  $p$ -værdien til  $18/500$ , dvs. 3.6%. 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 empirisk simulerede fordeling og den teoretiske fordeling fint overens!



## 2b: Goodness of Fit

**Eksempel 2:** (side 24 i kursusmaterialet)

Indkomstfordelingen i stikprøven var:  $I = \text{Indkomst i 1000 kr.}$

Observeret antal

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

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

Forventet antal

$I < 50$	$50 \leq I < 100$	$100 \leq I < 150$	$150 \leq I < 200$	$200 \leq I < 300$	$300 \leq I < 400$	$400 \leq I < 500$	$500 \leq I$
64	93	178	123	243	180	66	33

Sammenholder vi de observerede hyppigheder med de forventede følger 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 udtrækker derfor en stikprøve fra en ideel population, der repræsenterer den forventede fordeling, sådan som den fremgår af tallene fra Danmarks statistik. Vi indskrives derfor lister med indkomst-kategorier, de observerede hyppigheder og de forventede hyppigheder. Derefter benytter vi den følgende kommando til at omdanne hyppighedslisterne til lister over rådata, dvs. vi sætter  $\text{obs} = \text{FreqToList}(\text{obs\_hyp})$  og  $\text{ideel} = \text{FreqToList}(\text{forv\_hyp})$  - læg mærke til at de to lister ikke længere er koordinerede, de skal kun bruges til simuleringen!:

Define  $\text{FreqToList}(\text{FreqList}) = \text{Func} :: \text{Local } i, \text{list} :: \text{list} := \{ \} :: \text{For } i, 1, \text{Dim}(\text{FreqList}) :: \text{list} := \text{augment}(\text{list}, \text{seq}(i, n, 1, \text{FreqList}[i])) :: \text{EndFor} :: \text{list} :: \text{EndFunc}$

kat_ind	obs_hyp	forv_hyp	obs	ideel
1	98	64	1	1
2	88	93	1	1
3	199	178	1	1
4	136	123	1	1
5	210	243	1	1
6	179	180	1	1
7	52	66	1	1
8	38	53	1	1
			1	1
			1	1
			1	1
			1	1
			1	1
			1	1
			1	1
			1	1

$$\text{chi\_obs} := \text{approx} \left( \text{sumList} \left( \frac{(\text{obs\_hyp} - \text{forv\_hyp})^2}{\text{forv\_hyp}} \right) \right) \rightarrow 33.8848$$

Da stikprøven også består af 1000 individer skal vi nu have trukket 1000 individer fra populationen MED tilbagelægning, så hver indkomstkategori hver gang har samme sandsynlighed for at blive udtrukket!

```
stikproeve := seq(ideel[randInt(1, 1000)], i, 1, 1000)
```

Vi skal så have rekonstrueret hyppighederne hørende til denne liste. Det sker ved hjælp af følgende kommando:

```
Define ListToFreq(List)=Func
:: Local i,k, FreqList
:: FreqList:=NewList(max(List))
:: For k, 1, dim(List)
:: i:=List[k]
:: FreqList[i]:=FreqList[i]+1
:: EndFor
:: FreqList
:: EndFunc
```

```
Stik_hyp := ListToFreq(stikproeve) → {61, 76, 190, 130, 272, 173, 61, 37}
```

Herefter er vejen banet for at udregne testværdien:

```
chi_sim := approx(sumList((Stik_hyp - forv_hyp)2 / forv_hyp)) → 13.3976
```

Vi kan gentage udregningen af testværdien ved at pakke de forestående kommandoer sammen i en math-boks:

```
stikproeve := seq(ideel[randInt(1, 1000)], i, 1, 1000) :: Stik_hyp := ListToFreq(stikproeve) :: chi_sim
:= approx(sumList((Stik_hyp - forv_hyp)2 / forv_hyp))
→ 5.14849
```

```
stikproeve := seq(ideel[randInt(1, 1000)], i, 1, 1000) :: Stik_hyp := ListToFreq(stikproeve) :: chi_sim
:= approx(sumList((Stik_hyp - forv_hyp)2 / forv_hyp))
→ 8.79261
```

```
stikproeve := seq(ideel[randInt(1, 1000)], i, 1, 1000) :: Stik_hyp := ListToFreq(stikproeve) :: chi_sim
:= approx(sumList((Stik_hyp - forv_hyp)2 / forv_hyp))
→ 5.94811
```

Vi har nu gentaget simuleringen 3 gange, men skal det virkelig batte noget skal vi konstruere en brugerdefineret funktion, der pakker de foregående kommandoer sammen i én lang sekvens!

```
Define Simul(ideel, antal) = Func :: Local forventet :: forventet := ListToFreq(ideel):: stikproeve := se
q(ideel[randInt(1, dim(ideel))], i, 1, antal):: Stik_hyp := ListToFreq(stikproeve) :: approx(sumList(
  (Stik_hyp - forventet)^2))
  forventet)) :: EndFunc
```

simul(ideel, 1000) → 8.83794

Derefter kan vi udføre simuleringen mange gange ved hjælp af en seq-kommando, som vi først udfører 10 gange for at se den fungerer som den skal!

seq(simul(ideel, 1000), i, 1, 10) → {6.71565, 8.07661, 6.18424, 7.98584, 16.2214, 12.7585, 13.3885, 5.82388, 5.2944, 6.08142}

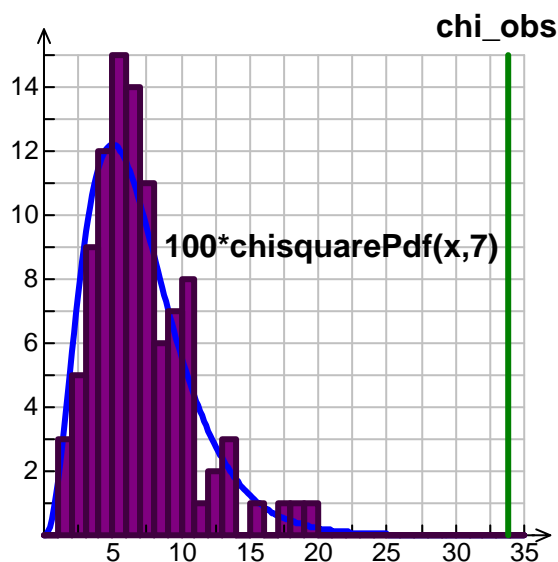
Herefter har vi med stor tålmodighed udført simuleringen 100 gange:

test := seq(simul(ideel,1000),i,1,100)

crit := sum(when(test[i] >= chi\_obs,1,0) ,i,1,dim(test)) → 0

Vi kan da umiddelbart se at der slet ikke er nogen skæve målinger, hvorfor resultatet synes at være statistisk signifikant på 1%-niveauet.

Vi kan også illustrere testfordelingen med et histogram overlejret med den teoretiske fordeling. Som det ses stemmer den empirisk simulerede fordeling rimeligt overens med den teoretiske fordeling.



### 3) Teori: De indbyggede fordelingsfunktioner

#### De indbyggede fordelingsfunktioner:

Chi-kvadrat ( $\chi^2$ ) fordelingen hedder chisquare. Når man skal arbejde med chi-kvadratfordelingen kan man benytte de følgende to operatører:

$$\begin{aligned} y &= \text{chisquarePdf}(x, df): && \text{(Point Distribution Function)} \\ p &= \text{chisquareCdf}(x_{lav}, x_{høj}, df) && \text{(Cumulative Distribution Function)} \end{aligned}$$

Vi ser først på tæthedsfunktionen:

$$\text{chisquarePdf}(x, df) \rightarrow \text{chisquarepdf}(x, df)$$

Som det ses kan vi ikke uden videre få oplyst forskriften. Vil man arbejde med forskriften skal den derfor indføres som en brugerdefineret funktion:

$$\text{chi2density}(x, df) := \frac{x^{(df/2-1)} \cdot e^{-x/2}}{\int_0^{\infty} (x^{(df/2-1)} \cdot e^{-x/2}) dx} \mid x \geq 0 \rightarrow \text{"Done"}$$

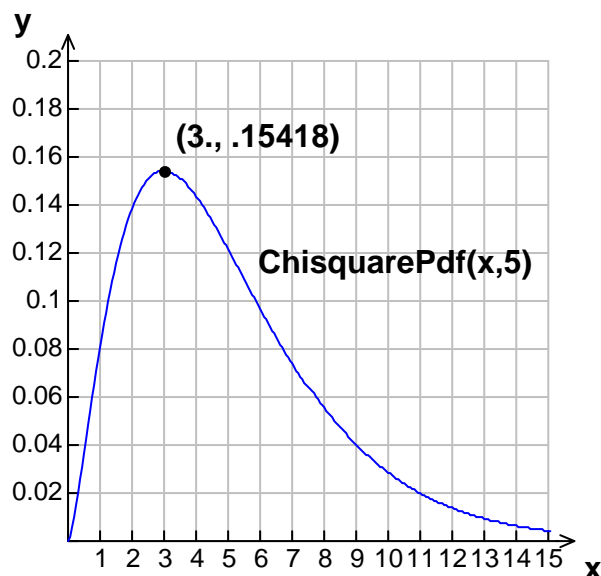
Det kan godt se lidt uoverskueligt ud, men for konkrete værdier af antallet af frihedsgrader forenkles udtrykket - ikke mindst for de lige frihedsgrader:

$$\text{chi2density}(x, 4) \rightarrow \frac{x \cdot e^{-x/2}}{4}$$

$$\text{chi2density}(x, 5) \rightarrow .132981 \cdot x^{3/2} \cdot e^{-x/2}$$

For de ulige frihedsgrader fører integrationen kun til en numerisk approksimation.

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



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

$$\text{chisquareCdf}(x_{low}, x_{high}, df) \rightarrow \text{chisquarecdf}(x_{low}, x_{high}, df)$$

Som det ses kan vi igen ikke få oplyst forskriften. Vil man arbejde med forskriften skal den derfor indføres som en brugerdefineret funktion:

$$\text{chikumuleret}(x_{low}, x_{high}, df) := \frac{\int_{x_{low}}^{x_{high}} (x^{(df/2-1)} \cdot e^{-x/2}) dx}{\int_0^{\infty} (x^{(df/2-1)} \cdot e^{-x/2}) dx} \rightarrow \text{"Done"}$$

Igen forenkles det betydeligt for et konkret antal frihedsgrader, ikke mindst for de lige frihedsgrader:

$$\text{chikumuleret}(x_{low}, x_{high}, 4)$$

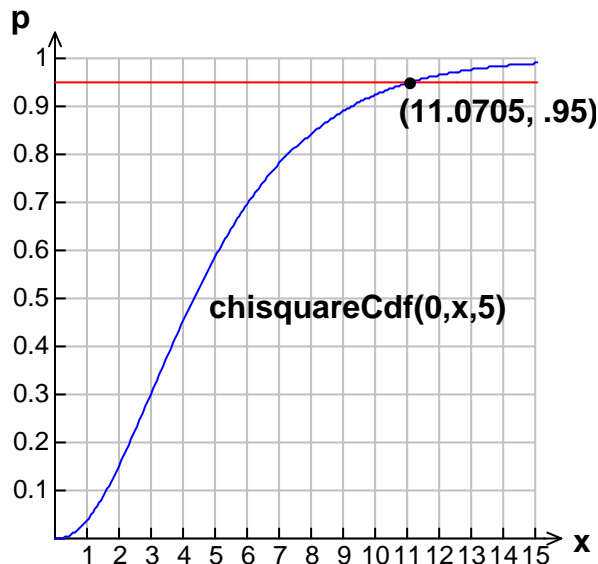
$$\rightarrow \frac{(e^{x_{high}/2} \cdot (x_{low} + 2) - (x_{high} + 2) \cdot e^{x_{low}/2}) \cdot e^{-(x_{high}/2 - x_{low}/2)}}{2}$$

$$\text{chikumuleret}(0, x_{high}, 4) \rightarrow \frac{e^{-x_{high}/2} \cdot (2 \cdot e^{x_{high}/2} - x_{high} - 2)}{2}$$

$$\text{expand}(ans) \rightarrow \frac{-x_{high}}{2 \cdot \sqrt{e^{x_{high}}}} - \frac{1}{\sqrt{e^{x_{high}}}} + 1$$

(hvor det sidste udtryk er fremkommet ved at anvende en expand-kommando)

$$\text{chikumuleret}(x_{low}, x_{high}, 5) \rightarrow .132981 \cdot \int_{x_{low}}^{x_{high}} (x^{3/2} \cdot e^{-x/2}) dx$$



Endelig kan man finde fraktilerne (den inverse kumulerede fordeling). Men da den inverse chi2-funktion ikke er indbygget må vi her ty til en **solve**-kommando:

```
solve(chisquareCdf(0, x, 5) = 0.95, x)
```

```
→ x = 11.0705
```

```
Warning: More solutions may exist
```

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

```
chisquareCdf(0, 11.0705, 5) → .95
```

## 4) Teoretiske udregninger hørende til hypotesetest

### 4a: Uafhængighed (Eksempel 1, side 4 i kursusmaterialet)

Koen\Tøjforbrug	<1500kr./maaned	≥ 1500kr./maaned	I alt
Kvinde	98	102	200
Mand	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. Vi bruger da det følgende hjælpebibliotek til at styre gennem udregningerne med krydstabellerne:

$$\text{rowTotal}(Matrix) := \text{seq} \left( \sum_{j=1}^{\text{colDim}(Matrix)} (Matrix_{[i, j]}), i, 1, \text{rowDim}(Matrix) \right)$$

$$\text{colTotal}(Matrix) := \text{seq} \left( \sum_{i=1}^{\text{rowDim}(Matrix)} (Matrix_{[i, j]}), j, 1, \text{colDim}(Matrix) \right)$$

$$\text{grandTotal}(Matrix) := \sum_{j=1}^{\text{colDim}(Matrix)} \left( \sum_{i=1}^{\text{rowDim}(Matrix)} (Matrix_{[i, j]}) \right)$$

Define  $\text{expected}(obs) = \text{Func} :: \text{Local } i, j, \text{expMatrix} :: \text{expMatrix} := \text{NewMat}(\text{rowDim}(obs), \text{colDim}(obs)) :: \text{For } i, 1, \text{rowDim}(obs) :: \text{For } j, 1, \text{colDim}(obs) :: \text{ExpMatrix}_{[i, j]} := \text{approx} \left( \frac{(\text{rowTotal}(obs))_{[i]} \cdot (\text{colTotal}(obs))_{[j]}}{\text{GrandTotal}(obs)} \right) :: \text{EndFor} :: \text{EndFor} :: \text{ExpMatrix} :: \text{EndFunc}$

$$\text{chiInd}(obsMatrix, \text{expMatrix}) := \sum_{j=1}^{\text{colDim}(obsMatrix)} \left( \sum_{i=1}^{\text{rowDim}(obsMatrix)} \left( \frac{(\text{obsMatrix}_{[i, j]} - \text{expMatrix}_{[i, j]})^2}{\text{ExpMatrix}_{[i, j]}} \right) \right)$$

Det er faktisk kun de to sidste kommandoer **expected** og **chiInd**, der bruges i det følgende, men de bygger på de tre første kommandoer, der derfor også skal med ind i dokumentet!

Vi indtaster da først krydstabellen for de observerede hyppigheder:

$$obs := \begin{bmatrix} 98 & 102 \\ 60 & 100 \end{bmatrix} \rightarrow \begin{bmatrix} 98 & 102 \\ 60 & 100 \end{bmatrix}$$

Dernæst udregnes krydstabellen for de forventede værdier:

$$forv := \text{expected}(obs) \rightarrow \begin{bmatrix} 87.7778 & 112.222 \\ 70.2222 & 89.7778 \end{bmatrix}$$

Og endeligt finder vi testværdien:

$$chi\_obs := \text{chiInd}(obs, forv) \rightarrow 4.77353$$

Herefter er vejen banet for udregning af  $p$ -værdien, dvs. sandsynligheden for at vi rammer mindst lige så skævt som det observerede:

$$p\_val := \text{chisquareCdf}(chi\_obs, \infty, 1) \rightarrow .028901$$

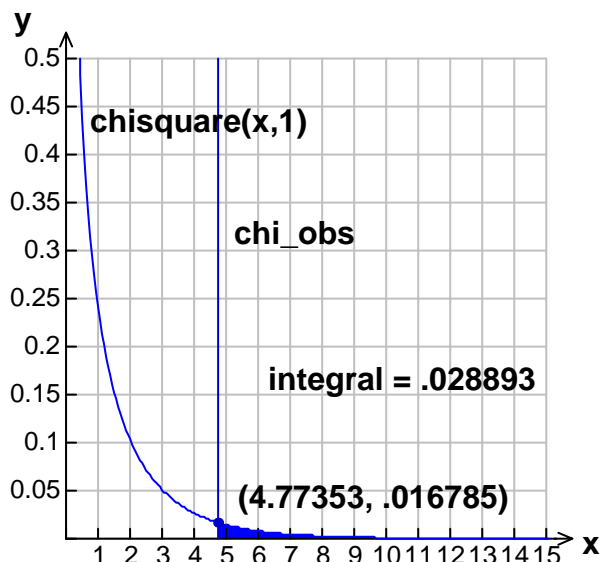
Den kritiske sandsynlighed er altså 2.89%, hvorfor afvigelsen er signifikant på

5%-niveau, dvs. vi forkaster nulhypotesen om uafhængighed på 5%-niveauet. Vi kunne også have fundet den kritiske grænse for teststørrelsen:

$$\text{solve}(\text{chisquareCdf}(0, x, 1) = 0.95, x) \rightarrow x = 3.84146$$

Warning: More solutions may exist

Hvis teststørrelsen ligger over 3.84 er afvigelsen altså kritisk, dvs. vi må forkaste nulhypotesen (på 5%-niveauet). Endelig kan vi illustrere testen grafisk (hvor vi nøjes med at integrere fra chi\_obs til 100) :



## 4b Goodness of fit

### Eksempel 2 (side 24 i kursusmaterialet)

Indkomstfordelingen i stikprøven var:  $I = \text{Indkomst i 1000 kr.}$

Observeret antal

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

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

Forventet antal

$I < 50$	$50 \leq I < 100$	$100 \leq I < 150$	$150 \leq I < 200$	$200 \leq I < 300$	$300 \leq I < 400$	$400 \leq I < 500$	$500 \leq I$
64	93	178	123	243	180	66	33

Sammenholder vi de observerede hyppigheder med de forventede følger 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:

kat_indkomst	obs_hyp	forv_hyp
1	98	64
2	88	93
3	199	178
4	136	123
5	210	243
6	179	180
7	52	66
8	38	53

$$chi\_obs := \text{approx}\left(\text{sumList}\left(\frac{(obs\_hyp - forv\_hyp)^2}{forv\_hyp}\right)\right) \rightarrow 33.8848$$

Herefter er vejen banet for udregning af  $p$ -værdien, dvs. sandsynligheden for at vi rammer mindst lige så skævt som det observerede:

$$p\_val := \text{chisquareCdf}(chi\_obs, \infty, 7) \rightarrow .000018$$

Den kritiske sandsynlighed er altså 0.0018%, hvorfor afvigelsen er signifikant på 1%-niveau, dvs. vi forkaster nulhypotesen om at stikprøven er repræsentativ for landsfordelingen på 1%-niveauet.

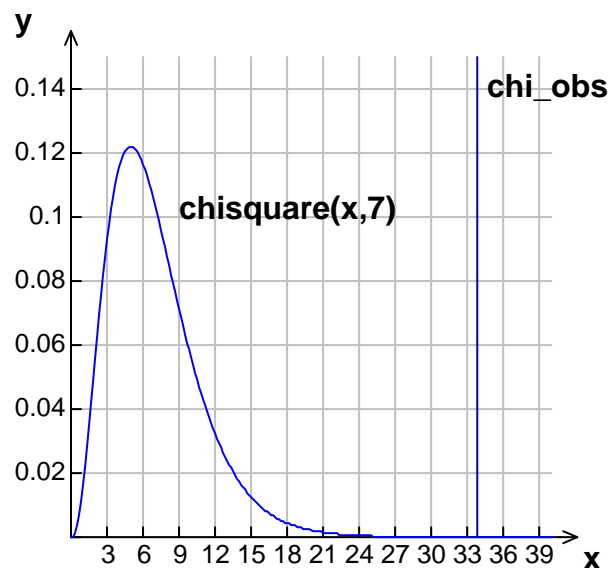
Vi kunne også have fundet den kritiske grænse for teststørrelsen:

$$\text{solve}(\text{chisquareCdf}(0, x, 7) = 0.99, x) \rightarrow x = 18.4753$$

Warning: More solutions may exist

Hvis teststørrelsen ligger over 18.47 er afvigelsen altså kritisk, dvs. vi må forkaste nulhypotesen (på 1%-niveauet).

Endelig kan vi illustrere testen grafisk. I dette tilfælde er det dog meget svært at illustrere det kritiske område uden at grafen for fordelingen bliver meget uoverskuelig.



## 5) Indbyggede testrutiner

### 5a: Uafhængighed Eksempel 1: (side 4 i kursusmaterialet)

Koen\Toejforbrug	<1500kr./maaned	≥ 1500kr./maaned	I alt
Kvinde	98	102	200
Mand	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 (chi-square test), der forventer at få oplyst såvel matricen for de observerede hyppigheder som matricen for de forventede værdier. Vi starter derfor med at indsætte biblioteket for hjælpefunktioner til chi-kvadrat testet:

$$\text{rowTotal}(Matrix) := \text{seq} \left( \sum_{j=1}^{\text{colDim}(Matrix)} (Matrix_{[i, j]}), i, 1, \text{rowDim}(Matrix) \right)$$

$$\text{colTotal}(Matrix) := \text{seq} \left( \sum_{i=1}^{\text{rowDim}(Matrix)} (Matrix_{[i, j]}), j, 1, \text{colDim}(Matrix) \right)$$

$$\text{grandTotal}(Matrix) := \sum_{j=1}^{\text{colDim}(Matrix)} \left( \sum_{i=1}^{\text{rowDim}(Matrix)} (Matrix_{[i, j]}) \right)$$

Define  $\text{expected}(obs) = \text{Func} :: \text{Local } i, j, \text{expMatrix} :: \text{expMatrix} := \text{NewMat}(\text{rowDim}(obs), \text{colDim}(obs)) :: \text{For } i, 1, \text{rowDim}(obs) :: \text{For } j, 1, \text{colDim}(obs) :: \text{ExpMatrix}_{[i, j]} \\ := \text{approx} \left( \frac{(\text{rowTotal}(obs))_{[i]} \cdot (\text{colTotal}(obs))_{[j]}}{\text{GrandTotal}(obs)} \right) :: \text{EndFor} :: \text{EndFor} :: \text{ExpMatrix} :: \text{EndFunc}$

$$obs := \begin{bmatrix} 98 & 102 \\ 60 & 100 \end{bmatrix} \rightarrow \begin{bmatrix} 98 & 102 \\ 60 & 100 \end{bmatrix}$$

$$forv := \text{expected}(obs) \rightarrow \begin{bmatrix} 87.7778 & 112.222 \\ 70.2222 & 89.7778 \end{bmatrix}$$

Herefter kan testen gennemføres inklusive en grafisk illustration (Brug **Stat**

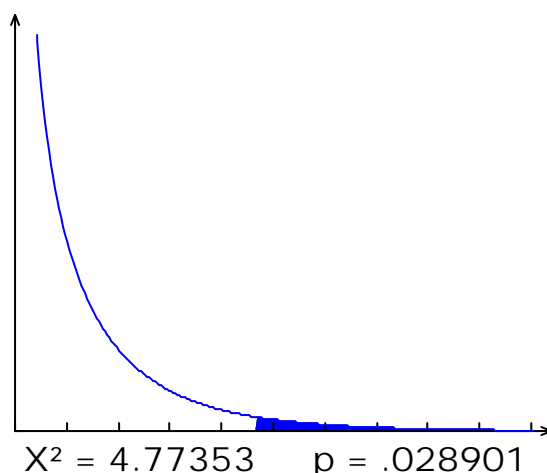
**Tests & Interval Tools**-værktøjet  på værktøjsbjælken):

Chi-square test

$p = .028901$

$X^2 = 4.77353$

$df = 1.$



Vi får da som vist en række koncentrerede oplysninger:

1)  $p$ -værdien er 2.89%, 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!

2) Teststørrelsen har værdien 4.77

3) Teststørrelsen er  $\chi^2$  fordelt med 1 frihedsgrad.

Vi har allerede udregnet matricen for de forventede værdier. Vi kan derfor supplere med matricen for de enkelte bidrag til teststørrelsen (læg mærke til punktummerne foran regneoperationerne. De sikrer at vi udregner udtrykket element for element):

$$comp := (obs .- forv) .^ 2 ./ forv \rightarrow \begin{bmatrix} 1.19044 & .931133 \\ 1.48805 & 1.16392 \end{bmatrix}$$

Summen af matricelementerne i compare-matricen er da netop teststørrelsen. I dette tilfælde bidrager de dog alle ca. det samme!

### 5b: Goodness of fit test:

Der er **ikke** indbygget en Goodness-of-fit test i TI-Interactive, så den må man selv gennemføre, jfr. afsnit 4!