

# oefeningen hoofdstuk 7 - De $\chi^2$ kwadraat toets

*Tijs Martens*

*12 april 2019*

## oefening 7.1.

### opgave

Hoe komen we hier aan de noemer? Waar komt dit mee overeen? Hoe bepaal je de variantie van een binomiale verdeling?

$$r_i = \frac{O_i - n\pi_i}{\sqrt{n\pi_i(1-\pi_i)}}$$

formule voor het aanduiden welke klasse de grootste bedrijge levert

### oplossing

$$n \times \pi(1 - \pi)$$

## Oefening 7.2.

Inlezen van de file (afkomstig uit de library)

```
library(MASS)
```

```
View(survey)
```

```
attach(survey)
```

### sporten en rookgedrag

a

uit te voeren onderzoeken

- sporten & roken (onderzoek 1)
- schrijfhand & bovenste hand (onderzoek 2)
- geslacht & roken (onderzoek 3)
- geslacht & schrijfhand (onderzoek 4)

### b. kruistabellen maken

onderzoek 1:

```
kruistabel01 <- table(Smoke, Exer)
kruistabel01
```

```
##           Exer
## Smoke   Freq None Some
##  Heavy     7     1     3
##   Never    87    18    84
##   Occas    12     3     4
##   Regul     9     1     7
```

### onderzoek 2:

```
kruistabel02 <- table(Fold, W.Hnd)
kruistabel02
```

```
##           W.Hnd
## Fold      Left Right
## L on R     10    88
## Neither     1    17
## R on L      7   113
```

### onderzoek 3:

```
kruistabel03 <- table(Sex, Smoke)
kruistabel03
```

```
##           Smoke
## Sex      Heavy Never Occas Regul
## Female     5    99     9     5
## Male       6    89    10    12
```

### onderzoek 4:

```
kruistabel04 <- table(Sex, W.Hnd)
kruistabel04
```

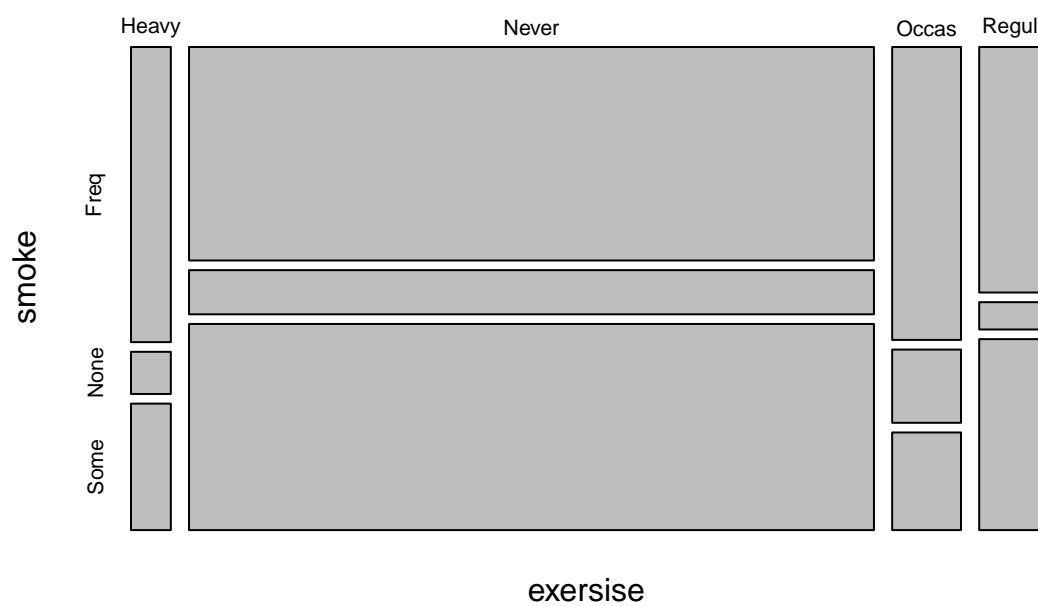
```
##           W.Hnd
## Sex      Left Right
## Female     7    110
## Male      10    108
```

### c. grafisch voorstellen

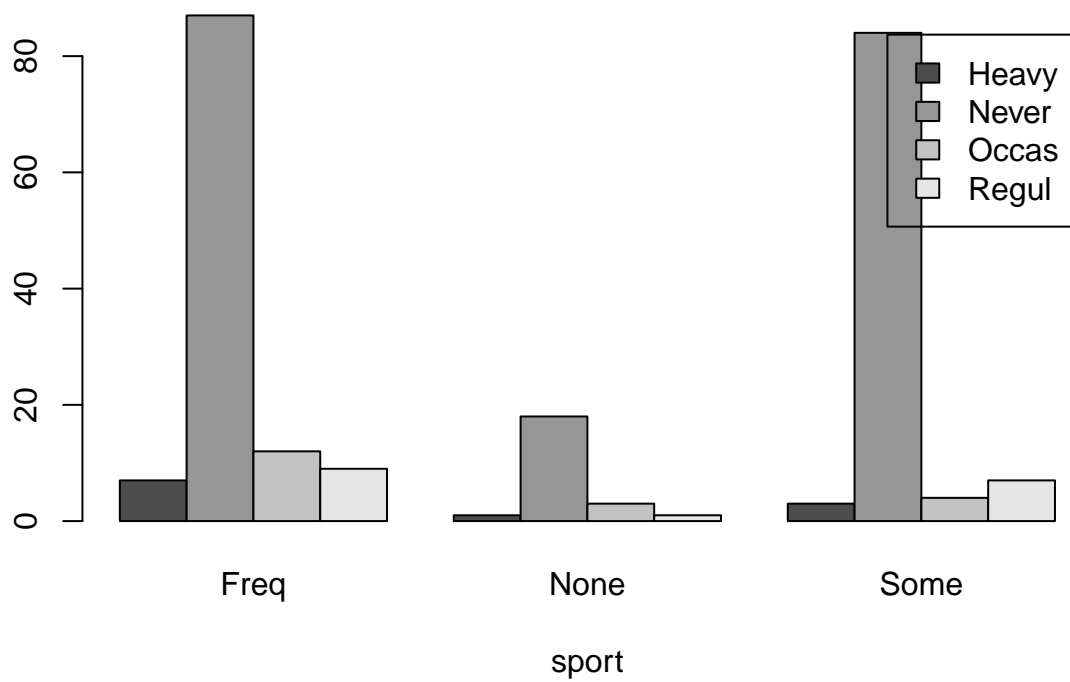
#### onderzoek 1:

```
plot(kruistabel01, ylab = 'smoke', xlab = 'exersise')
```

## kruistabel01

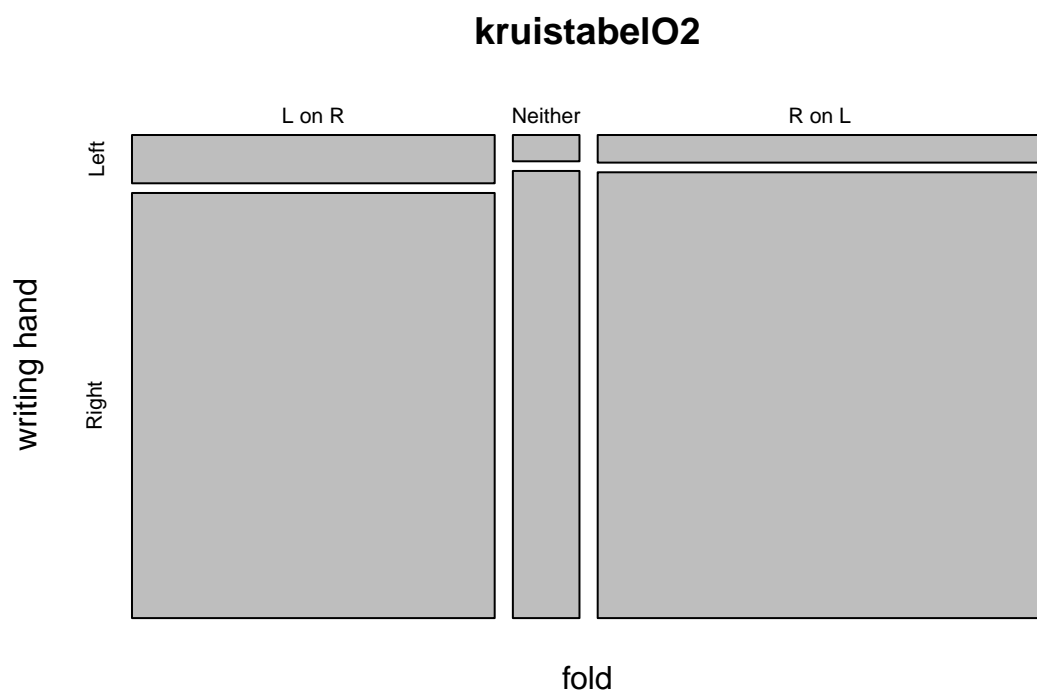


```
barplot(kruistabel01, xlab = 'sport', beside = TRUE, legend=rownames(kruistabel01))
```

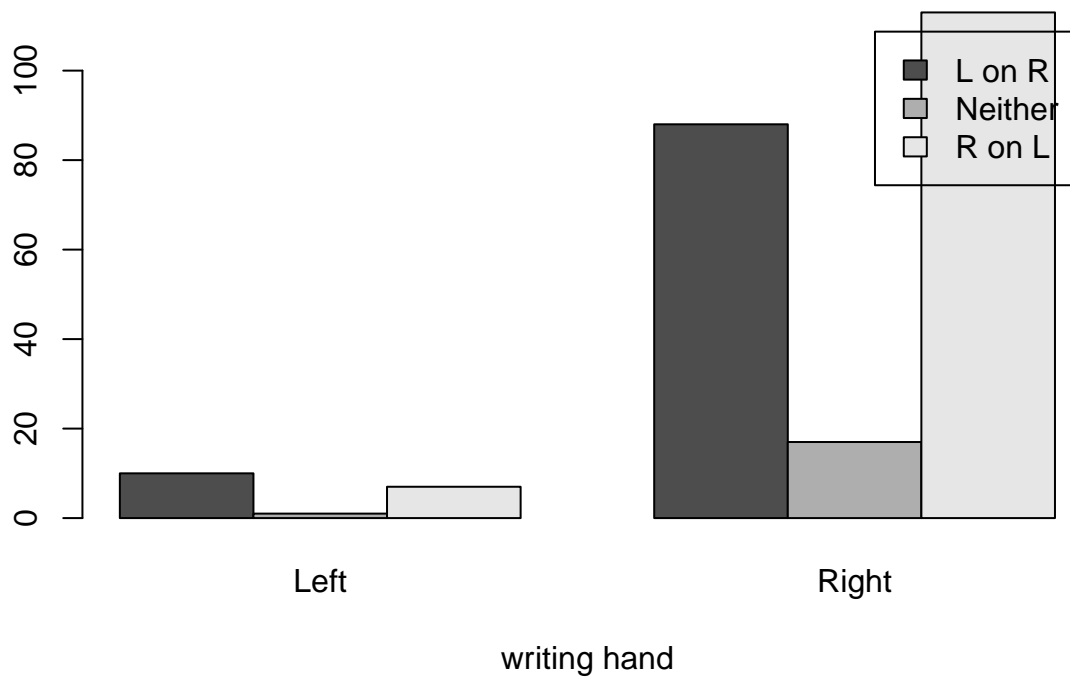


onderzoek 2:

```
plot(kruistabel02, ylab = 'writing hand', xlab = 'fold')
```

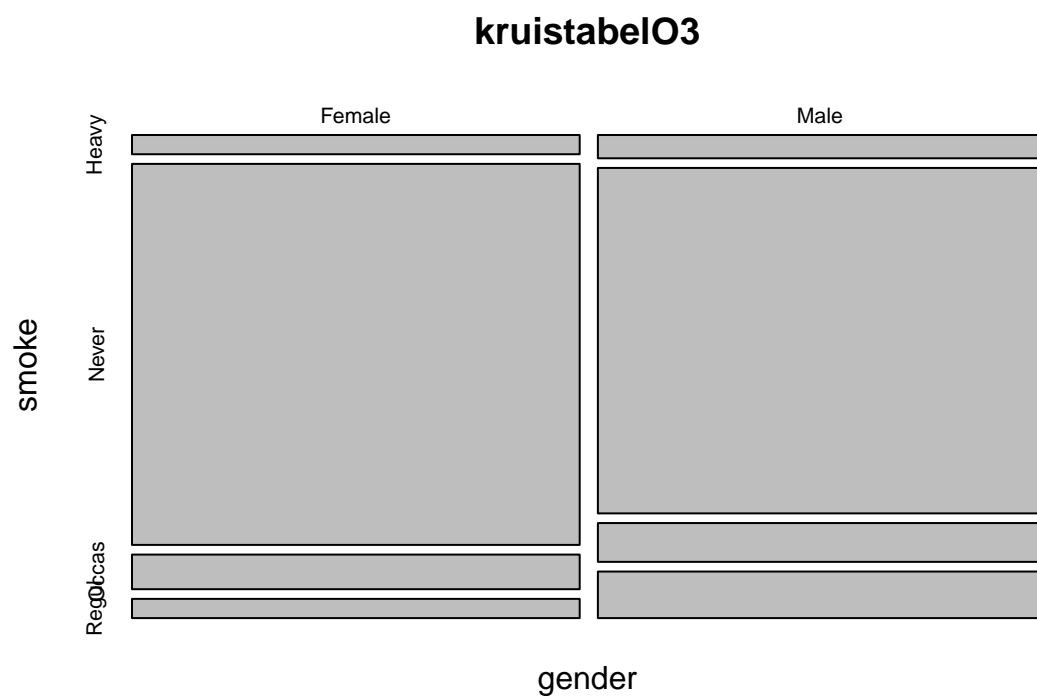


```
barplot(kruistabel02, xlab = 'writing hand', beside = TRUE, legend=rownames(kruistabel02))
```

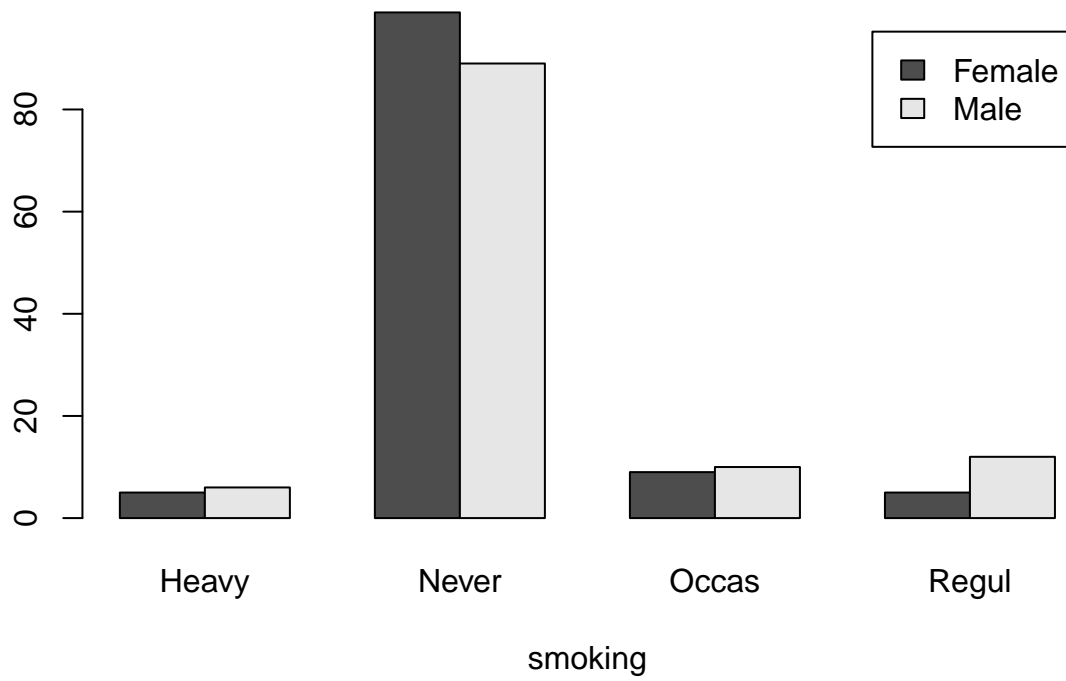


#### onderzoek 3:

```
plot(kruistabel03, ylab = 'smoke', xlab = 'gender')
```



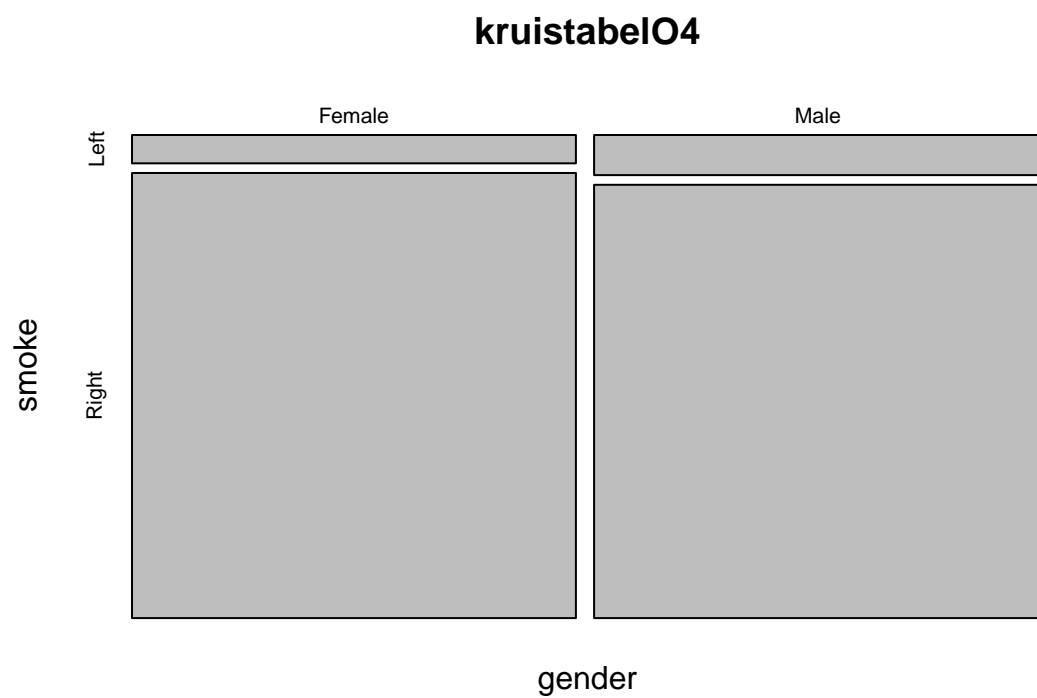
```
barplot(kruistabel03, xlab = "smoking" , beside = TRUE, legend=rownames(kruistabel03))
```



#### onderzoek 4:

```
plot(kruistabel04, ylab = 'smoke', xlab = 'gender')
```





```
barplot(kruistabel04, xlab = 'writing hand', beside = TRUE, legend=rownames(kruistabel04))
```



d

onderzoek 1:

hoog

onderzoek 2:

laag

onderzoek 3:

laag

onderzoek 4:

laag

e / f

onderzoek 1:

```
chisq.test(kruistabel01)
```

```
## Warning in chisq.test(kruistabel01): Chi-squared approximation may be
## incorrect
```

```
##
```

```
## Pearson's Chi-squared test
```

```
##
## data:  kruistabel01
## X-squared = 5.4885, df = 6, p-value = 0.4828
df01 <- (nrow(kruistabel01)-1) * (ncol(kruistabel01)-1)
df01

## [1] 6
qchisq(0.95, df01)

## [1] 12.59159
chi2 = 5.48 vrijheidsgraden = 6 grenswaarde = 12.59 p-value = 0.48
```

### onderzoek 2:

```
chisq.test(kruistabel02)

## Warning in chisq.test(kruistabel02): Chi-squared approximation may be
## incorrect

##
## Pearson's Chi-squared test
##
## data:  kruistabel02
## X-squared = 1.5814, df = 2, p-value = 0.4535
df02 <- (nrow(kruistabel02)-1) * (ncol(kruistabel02)-1)
df02

## [1] 2
qchisq(0.95, df02)

## [1] 5.991465
chi2 = 1.58 vrijheidsgraden = 2 grenswaarde = 5.99 p-value = 0.45
```

### onderzoek 3:

```
chisq.test(kruistabel03)

##
## Pearson's Chi-squared test
##
## data:  kruistabel03
## X-squared = 3.5536, df = 3, p-value = 0.3139
df03 <- (nrow(kruistabel03)-1) * (ncol(kruistabel03)-1)
df03

## [1] 3
qchisq(0.95, df03)

## [1] 7.814728
chi2 = 3.55 vrijheidsgraden = 3 grenswaarde = 7.81 p-value = 0.31
```

#### onderzoek 4:

```
chisq.test(kruistabel04)

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  kruistabel04
## X-squared = 0.23563, df = 1, p-value = 0.6274
df04 <- (nrow(kruistabel04)-1) * (ncol(kruistabel04)-1)
df04

## [1] 1
qchisq(0.95, df04)

## [1] 3.841459
```

$\chi^2 = 0.24$  vrijheidsgraden = 1 grenswaarde = 3.84 p-value = 0.62

g

#### onderzoek 1:

de nulhypothese wordt NIET verworpen, p (0.48) is groter dan alpha (0.05)

we hebben een representatieve steekproef,  $\chi^2(5.48)$  is kleiner dan de grenswaarde (12.59)

#### onderzoek 2:

de nulhypothese mag NIET verworpen worden, p (0.45) is groter dan alpha (0.05)

we hebben een representatieve steekproef,  $\chi^2(1.58)$  is kleiner dan de grenswaarde (5.99)

#### onderzoek 3:

de nulhypothese mag NIET verworpen worden, p (0.31) is groter dan alpha (0.05)

we hebben een representatieve steekproef,  $\chi^2(3.55)$  is kleiner dan de grenswaarde (7.81)

#### onderzoek 4:

$\chi^2 = 0.24$  vrijheidsgraden = 1 grenswaarde = 3.84 p-value = 0.62

de nulhypothese mag NIET verworpen worden, p (0.61) is groter dan alpha (0.05)

we hebben een representatieve steekproef,  $\chi^2(0.24)$  is kleiner dan de grenswaarde (3.84)

## oefening 7.3.

### opgave

Laad de dataset Aids2 uit package MASS (zie Oefening 7.2) die informatie bevat over 2843 patiënten die vóór 1991 in Australië met AIDS besmet werden. Deze dataset werd in detail besproken door Ripley en Solomon (2007). Onderzoek of er een relatie is tussen de variabele geslacht (Sex) en de manier van besmetting (T.categ).

1. Ga op de gebruikelijke manier te werk: visualiseren van de data,  $\chi^2$ , g en p-waarde berekenen (alpha = 0,05), en tenslotte een conclusie formuleren.
2. Bepaal de gestandaardiseerde residuën om te bepalen welke categorieën extreme waarden bevatten.

## oplossing

```
library(MASS)

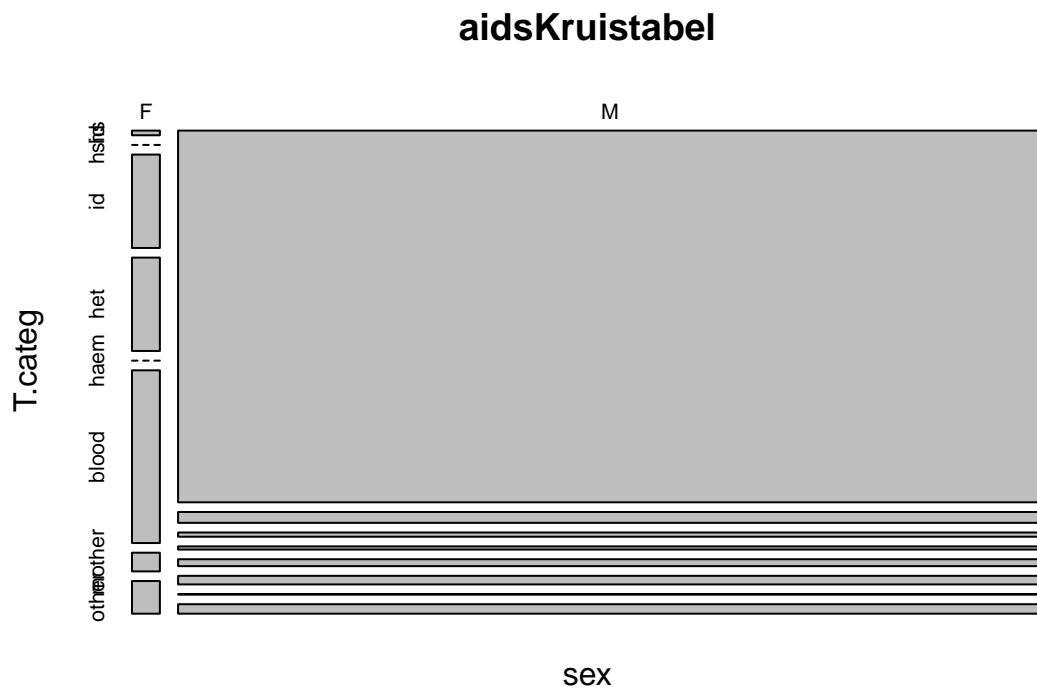
attach(Aids2)
```

### deel 1

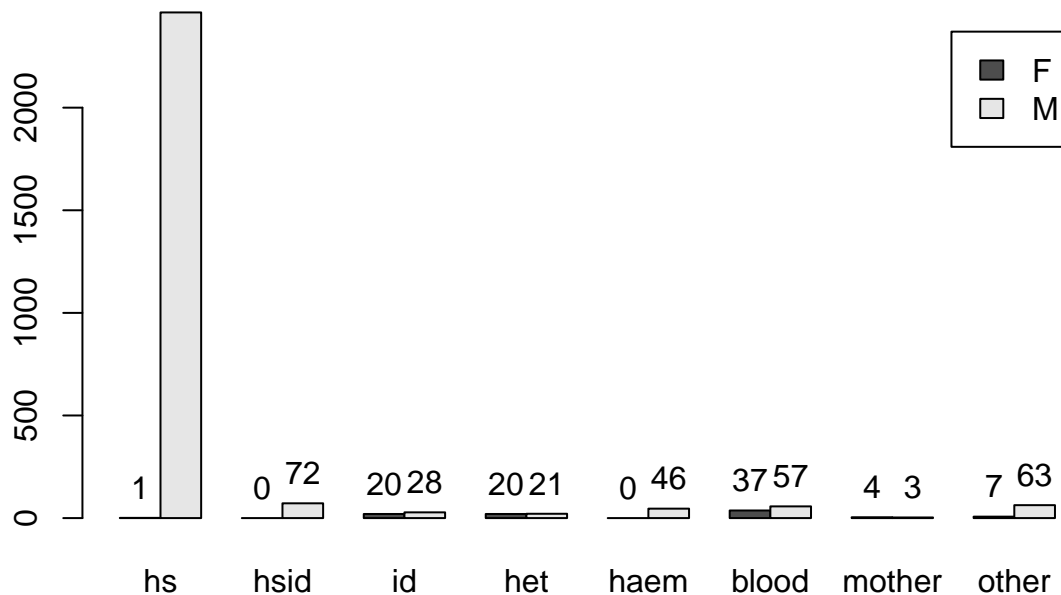
```
aidsKruistabel <- table(sex, T.categ)
aidsKruistabel
```

```
##      T.categ
## sex   hs hsid   id  het haem blood mother other
##  F     1   0   20   20   0   37     4     7
##  M 2464  72   28   21   46   57     3    63
```

```
plot(aidsKruistabel)
```



```
bp <- barplot(aidsKruistabel, beside = TRUE, legend=rownames(aidsKruistabel))
text(bp, aidsKruistabel, aidsKruistabel, pos = '3')
```



```
summary(aidsKruistabel)
```

```
## Number of cases in table: 2843
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 1083.4, df = 7, p-value = 1.157e-229
##  Chi-squared approximation may be incorrect
```

```
vrijheidsgraden <- (nrow(aidsKruistabel) - 1) * (ncol(aidsKruistabel) - 1)
vrijheidsgraden
```

```
## [1] 7
```

```
grens <- qchisq(0.95, vrijheidsgraden)
grens
```

```
## [1] 14.06714
```

$\chi^2 = 1083.4$  p-waarde =  $1.15 \times 10^{-229}$  vrijheidsgraden = 7 grenswaarde = 14.07

geen representatieve steekproef aangezien dat  $\chi^2$  groter is dan de grenswaarde

## deel 2

wat is dit???

## oefening 7.4.

### opdracht

Elk jaar voert Imec (voorheen iMinds) een studie uit over het gebruik van digitale technologieën in Vlaanderen, de Digimeter (Vanhaelewyn & De Marez, 2016). In deze oefening zullen we nagaan of de steekproef van de Digimeter 2016 ( $n = 2164$ ) representatief is voor de bevolking wat betreft de leeftijdscategorieën van de deelnemers. In Tabel 7.2a worden de relatieve frequenties van de deelnemers weergegeven. De absolute frequenties voor de verschillende leeftijdscategorieën van de Vlaamse bevolking worden samengevat in Tabel 7.2b. Deze gegevens zijn ook te vinden in bijgevoegd CSV-bestand oefeningen/data/bestat-vl-ages.csv.

1. De tabel met leeftijdsgegevens van de Vlaamse bevolking als geheel heeft meer categorieën dan deze gebruikt in de Digimeter. Maak een samenvatting zodat je dezelfde categorieën overhoudt dan deze van de Digimeter. Tip: dit gaat misschien makkelijker in een rekenblad dan in R.
2. Om de goodness-of-fit test te kunnen toepassen hebben we de absolute frequenties nodig van de geobserveerde waarden in de steekproef. Bereken deze.
3. Bereken ook de verwachte percentages (??i) voor de populatie als geheel.
4. Voer de goodness-of-fit test uit over de verdeling van leeftijdscategorieën in de steekproef van de Digimeter. Is de steekproef in dit opzicht inderdaad representatief voor de Vlaamse bevolking?

### oplossing

```
technologiegebruikt <- read.csv("C:\\Users\\tijsm\\Google Drive\\HoGent 2018-2019\\2e semester\\Onderzo  
technologiegebruikt
```

```
##      age.group population  
## 1         0-5      352017  
## 2         5-9      330320  
## 3        10-14      341303  
## 4        15-19      366648  
## 5        20-24      375469  
## 6        25-29      387131  
## 7        30-34      401285  
## 8        35-39      409587  
## 9        40-44      458485  
## 10       45-49      493720  
## 11       50-54      463668  
## 12       55-59      413315  
## 13       60-64      379301  
## 14       65-69      299152  
## 15       70-74      279789  
## 16       75-79      249260  
## 17       80-84      182352  
## 18       85-89      104449  
## 19       90-94       29888  
## 20       95-99       7678  
## 21      100+        923
```

### interessante libraries

gplots graphics - interessant voor examen blijkbaar visualize - duidelijk tonen van toetsen