6 – Analyse van 2 variabelen

Als casus lezen we resultaten in van een enquête uitgevoerd onder de klanten van een restaurant aan een hogeschool. Vergeet in R niet de werkdirectory aan te passen met “Session > Set Working Directory > To Source File Location”, anders zal het bestand niet teruggevonden worden.

```
library(foreign)
resto <- read.spss("../cursus/data/catering_hogeschool.sav",
                   to.data.frame = TRUE)
```

```
## re-encoding from CP1252
```

Chi-kwadraat en Cramér’s V

Chi-kwadraat en Cramér’s V zijn maten om te bepalen of er een verband bestaat tussen twee kwalitatieve variabelen (in R: factors).

Stel, we wensen te bepalen of vrouwen en mannen (variabele `Geslacht`) een andere mening hebben over de keuzemogelijkheden in het basisassortiment (variabele `Keuze_basis`). Als dat zo is, dan zeggen we dat er een verband is tussen de variabelen `Geslacht` en `Keuze_basis`. Als de antwoorden tussen beide groepen (ongeveer) gelijk zijn, dan zeggen we dat er *geen* verband is.

De resultaten van de enquête zijn:

```
# Bereken een frequentietabel. Eerst de afhankelijke, vervolgens
# de onafhankelijke variabele.
observed <- table(resto$Keuze_basis,
                  resto$Geslacht)
# Voeg rij- en kolomtotalen toe aan de tabel
addmargins(observed)
```

```
##
##               vrouw man Sum
##   Goed            9   8  17
##   Voldoende       8  10  18
##   Onvoldoende     5   5  10
##   Slecht          0   4   4
##   Sum            22  27  49
```

Uitgewerkte berekening

Een eerste stap om na te gaan of er verschillen zijn is om eerst na te gaan wat de waarden in deze frequentietabel zouden moeten zijn als vrouwen en mannen op een gelijkaardige manier antwoorden op de vraag. In dat geval zou je een tabel moeten krijgen met dezelfde rij- en kolomtotalen, maar gelijkmatig verdeeld. Deze aantallen verkrijg je door in elke cel het product te nemen van het rij- en kolomtotaal en te delen door het totaal aantal respondenten. Bijvoorbeeld voor de cel linksboven (vrouwen die “Goed” antwoordden) krijg je `22 * 17 / 49`. De gehele tabel bereken je dan zo:

```
row_sums <- rowSums(observed)              # rijtotalen
col_sums <- colSums(observed)              # kolomtotalen
n <- sum(observed)                         # totaal hele tabel
expected <- outer(row_sums, col_sums) / n  # verwachte waarden
addmargins(expected)                       # voeg totalen toe
```

```
##                vrouw       man Sum
## Goed         7.632653  9.367347  17
## Voldoende    8.081633  9.918367  18
## Onvoldoende  4.489796  5.510204  10
## Slecht       1.795918  2.204082   4
## Sum         22.000000 27.000000  49
```

Zoals je kan zien zijn de rij- en kolomtotalen inderdaad gelijk aan deze van de geobserveerde waarden. Wat is nu het verschil tussen beide?

```
expected - observed
```

```
##                  vrouw         man
## Goed        -1.36734694  1.36734694
## Voldoende    0.08163265 -0.08163265
## Onvoldoende -0.51020408  0.51020408
## Slecht       1.79591837 -1.79591837
```

Sommige waarden lijken sterker af te wijken (bv. “Goed”), anderen veel minder (bv. “Voldoende”). Een maat om de totale afwijking in een frequentietabel te bepalen, bestaat er uit om de verschillen tussen verwachte en geobserveerde waarden te kwadrateren (net zoals men bij variantie/standaardafwijking doet) en te delen door de verwachte waarde:

```
diffs <- (expected - observed)^2 / expected
diffs
```

```
##                   vrouw          man
## Goed        0.2449525265 0.1995909475
## Voldoende   0.0008245723 0.0006718737
## Onvoldoende 0.0579777365 0.0472411187
## Slecht      1.7959183673 1.4633408919
```

De som van al deze waarden wordt χ^2 (“chi-kwadraat”) genoemd.

```
chi_squared <- sum(diffs)
chi_squared
```

```
## [1] 3.810518
```

Nu zegt deze waarde op zich nog steeds niet zo veel. Onder welke voorwaarden zeggen we dat er al dan niet een verband is tussen beide variabelen? Een en ander zal ook afhangen van de grootte van de tabel en het totaal aantal observaties. In een kruistabel met meer rijen/kolommen, zal je een grotere χ^2 moeten hebben om te besluiten dat er een verband is.

Cramér’s V is een formule waarmee de χ^2 kan genormaliseerd worden tot een waarde tussen 0 en 1 die onafhankelijk is van de tabelgrootte:

```
k <- min(nrow(observed), ncol(observed))
cramers_v <- sqrt(chi_squared / ((k - 1) * n))
cramers_v
```

```
## [1] 0.278865
```

Om een besluit te trekken uit dit getal, vergelijk je het met de waarden in onderstaande tabel:

Cramér's V	Besluit
0	Geen verband
0.1	Zwak verband
0.25	Redelijk sterk verband
0.50	Sterk verband
0.75	Zeer sterk verband
1	Volledig verband

We kunnen dus besluiten dat er een redelijk sterk verband bestaat tussen de variabelen `Geslacht` en `Keuze_basis`.

R-functies

In R zijn er al functies geschreven voor de berekening van χ^2 en Cramér's V. Je hoeft dus niet telkens de berekeningen van hierboven te herhalen, maar kan meteen de juiste functie gebruiken, meer bepaald `assocstats` uit de library `vcd`:

```
library(vcd)
```

```
## Loading required package: grid
```

```
assocstats(observed)
```

```
##                     X^2 df P(> X^2)
## Likelihood Ratio 5.3156  3  0.15009
## Pearson          3.8105  3  0.28267
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.269
## Cramer's V        : 0.279
```

Deze functie geeft zowel de waarde van "Pearson's χ^2" (derde regel, eerste cijfer) als Cramér's V (laatste regel). Beide komen overeen met de waarden die we hierboven berekend hebben.

Regressie

Als voorbeeld voor het berekenen van de regressierechte nemen we de dataset die ook in de slides gebruikt wordt:

```
weight_gain <- read.csv("../cursus/data/santa.txt",
                        sep = "")
```

Kleinste kwadratenmethode: uitgewerkte berekening

We proberen een verzameling van punten (x_i, y_i) (voor $i : 1, \ldots, n$) zo goed mogelijk te benaderen met een rechte $\hat{y} = \beta_0 + \beta_1 x$. Het symbool $\hat{y}$ betekent “een schatting voor y”. De parameters β_0 en β_1 worden als volgt berekend:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - x)^2}$$

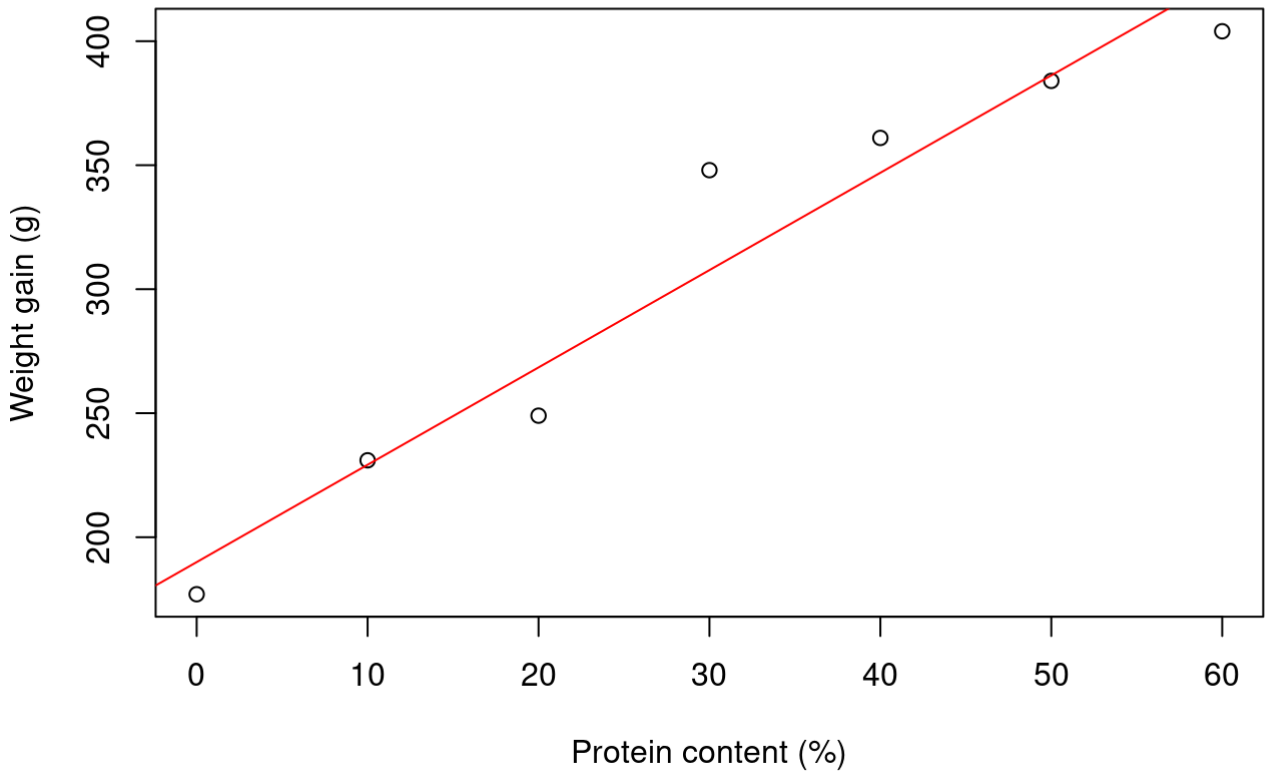
$$\beta_0 = \overline{y} - \beta_1 \overline{x}$$

In R kan je dat als volgt uitrekenen:

```
mx <- mean(weight_gain$x)  # gemiddelde van x
my <- mean(weight_gain$y)  # gemiddelde van y
xx <- weight_gain$x - mx   # x - mx
yy <- weight_gain$y - my   # y - my
beta_1 <- sum(xx * yy) / sum(xx^2)
beta_0 <- my - beta_1 * mx
```

Een plot toont dat dit een goede benadering is:

```
plot(x = weight_gain$x, y = weight_gain$y,
     xlab = "Protein content (%)",
     ylab = "Weight gain (g)")
abline(a = beta_0,  # snijpunt y-as
       b = beta_1,  # richtingscoëfficiënt
       col = 'red')
```



Lineaire regressie in R

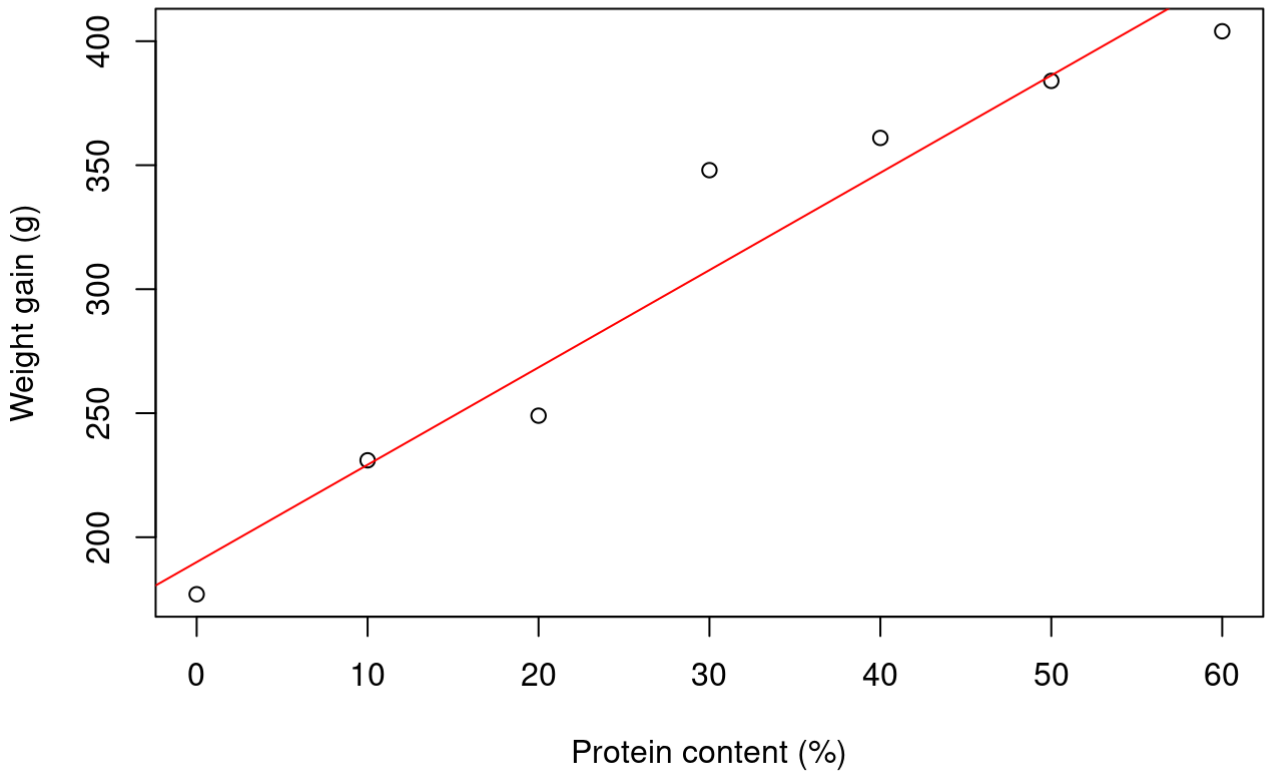
Uiteraard bestaat ook hiervoor een manier om dit in R makkelijk te berekenen, meer bepaald met de functie `lm()`, afkorting voor *linear model*.

```
lm(weight_gain$y ~ weight_gain$x)
```

```
##
## Call:
## lm(formula = weight_gain$y ~ weight_gain$x)
##
## Coefficients:
##   (Intercept)  weight_gain$x
##       189.964          3.925
```

Merk op dat hier de "group by" operator `~` gebruikt wordt. De onafhankelijke variabele komt **rechts** van de tilde te staan. Het resultaat van `lm()` kan meteen meegegeven worden aan de functie `abline()`:

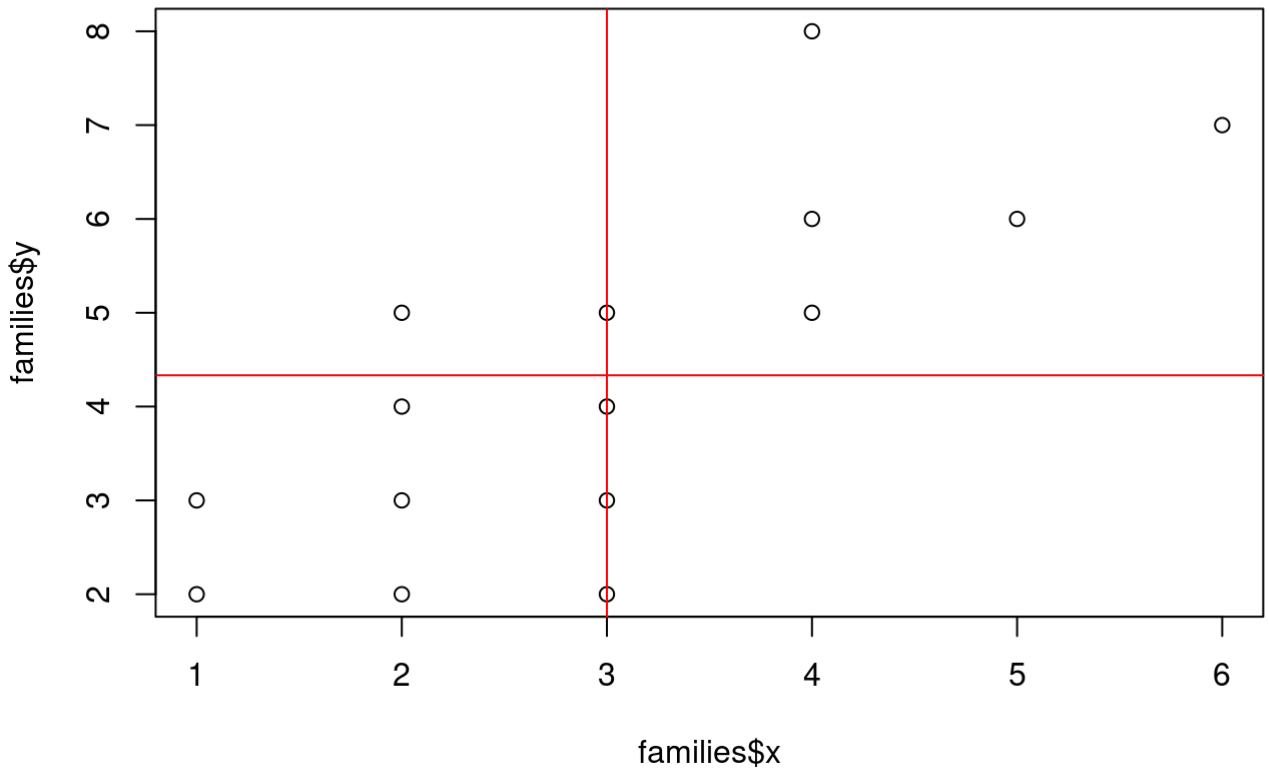
```
plot(x = weight_gain$x, xlab = "Protein content (%)",
     y = weight_gain$y, ylab = "Weight gain (g)")
regression <- lm(weight_gain$y ~ weight_gain$x)
abline(regression,
       col = 'red')
```



Covariantie en correlatie

```
families <- read.csv("../cursus/data/families.txt", sep ="")
mx <- mean(families$x)
my <- mean(families$y)

plot(families$x,families$y)
abline(h = my, col = 'red')
abline(v = mx, col = 'red')
```



Covariantie: $Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$

```
# Covariantie manueel berekend
covar <- sum((families$x - mx) * (families$y - my)) / (length(families$x) - 1)
covar
```

```
## [1] 2
```

```
# R-functie
cov(families$x, families$y)
```

```
## [1] 2
```

Correlatie (Pearson's product-momentcorrelatiecoëfficiënt):

$$R = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 (y_i - \overline{y})^2}}$$

```
# Berekenen vanuit covariantie
covar / (sd(families$x) * sd(families$y))
```

```
## [1] 0.7533708
```

```
# Uitgewerkte formule
sum((families$x - mx) * (families$y - my)) /
  sqrt(sum((families$x- mx)^2 * sum((families$y - my)^2)))
```

```
## [1] 0.7533708
```

```
# R-functie
cor(families$x, families$y)
```

```
## [1] 0.7533708
```

Determinatiecoëfficiënt:

- Definieer $SS_{tot} = \sum_{i=1}^{n}(y_i - \overline{y})^2$, de totale variantie van de steekproef (SS is een afkorting voor *squared sum*)
- Definieer $SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y})^2$, de residuën t.o.v. de regressierechte, of de niet-verklaarde variantie van de steekproef
- De determinatiecoëfficiënt is dan $R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}}$

De determinatiecoëfficiënt heeft als eigenschap dat die het kwadraat is van de correlatiecoëfficiënt (wat de notatie R^2 verklaart). Dit is een getal tussen 0 en 1 dat je kan interpreteren als het percentage van de variantie in y die kan verklaard worden door x. Dit is een maat die aangeeft hoe goed de regressielijn de echte datapunten benadert. Hoe dichter bij 1, hoe beter de benadering is, dus hoe dichter de geobserveerde datapunten bij de regressierechte liggen.

```
# gemiddelde van y
my <- mean(weight_gain$y)
# som van kwadraten tov gemiddelde
ss_tot <- sum((weight_gain$y - my)^2)
# som van kwadraten "residuën", i.e. verschil tussen geobserveerde
# en voorspelde waarde op basis van regressie
regression <- lm(weight_gain$y ~ weight_gain$x) # regressiemodel
yy <- predict.lm(regression, weight_gain)       # voorspelde waarden
ss_res <- sum((weight_gain$y - yy)^2)           # som v kwadraten
# R^2 aan de hand van de definite
(ss_tot - ss_res) / ss_tot
```

```
## [1] 0.9383165
```

```
# R^2 aan de hand van correlatie:
correlation <- cor(x = weight_gain$x, y = weight_gain$y)
correlation^2
```

```
## [1] 0.9383165
```

Visualiseren van verbanden tussen twee variabelen

Wanneer je het verband tussen twee variabelen wil visualiseren, dan hangt het meest geschikte grafiektype af van het meetniveau van de variabelen. Deze vind je de tabel hieronder voor verschillende combinaties van meetniveaus van enerzijds de onafhankelijke en anderzijds de afhankelijke variabele.

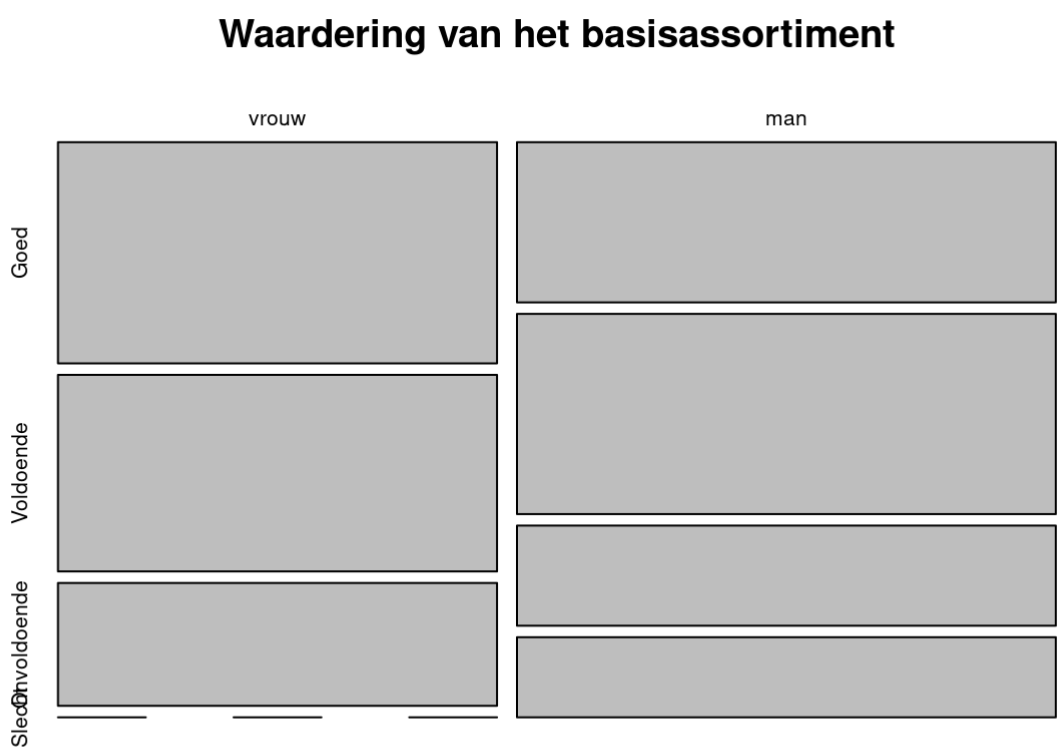
Onafhankelijke	Afhankelijke	Grafiektype
Kwalitatief	Kwalitatief	mozaïekdiagram
		geclusterde staafgrafiek
		rependiagram
Kwalitatief	Kwantitatief	boxplot
		staafgrafiek gemiddelde
Kwantitatief	Kwantitatief	spreidings/XY-grafiek

Kwalitatief - kwalitatief

Mozaïekdiagram

Een mozaïekdiagram is een grafische weergave van een frequentietabel waarbij de oppervlakte van elke tegel proportioneel is met de frequentie in de overeenkomstige cel van de tabel.

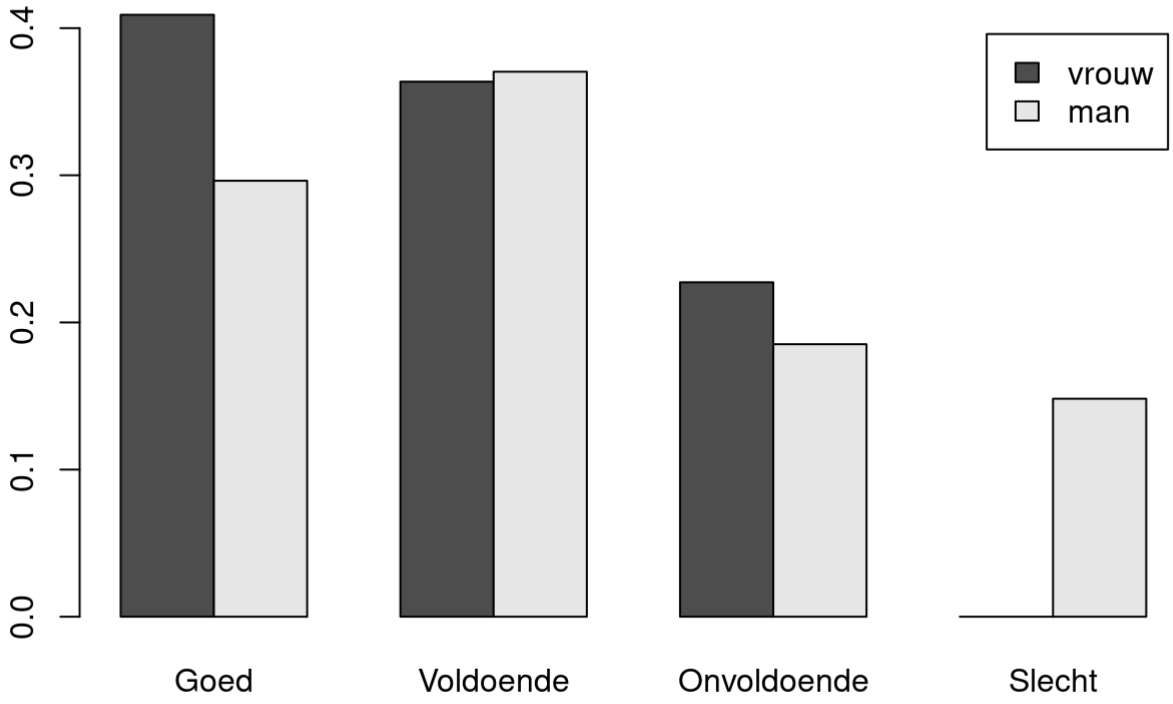
```
mosaicplot(t(observed),
           main = "Waardering van het basisassortiment")
```



In deze grafiek is de frequentietabel getransponeerd met de functie `t()`. Op die manier wordt de onafhankelijke variabele in de kolommen weergegeven en kan je duidelijk de verschillen in frequenties zien.

Geclusterde staafgrafiek

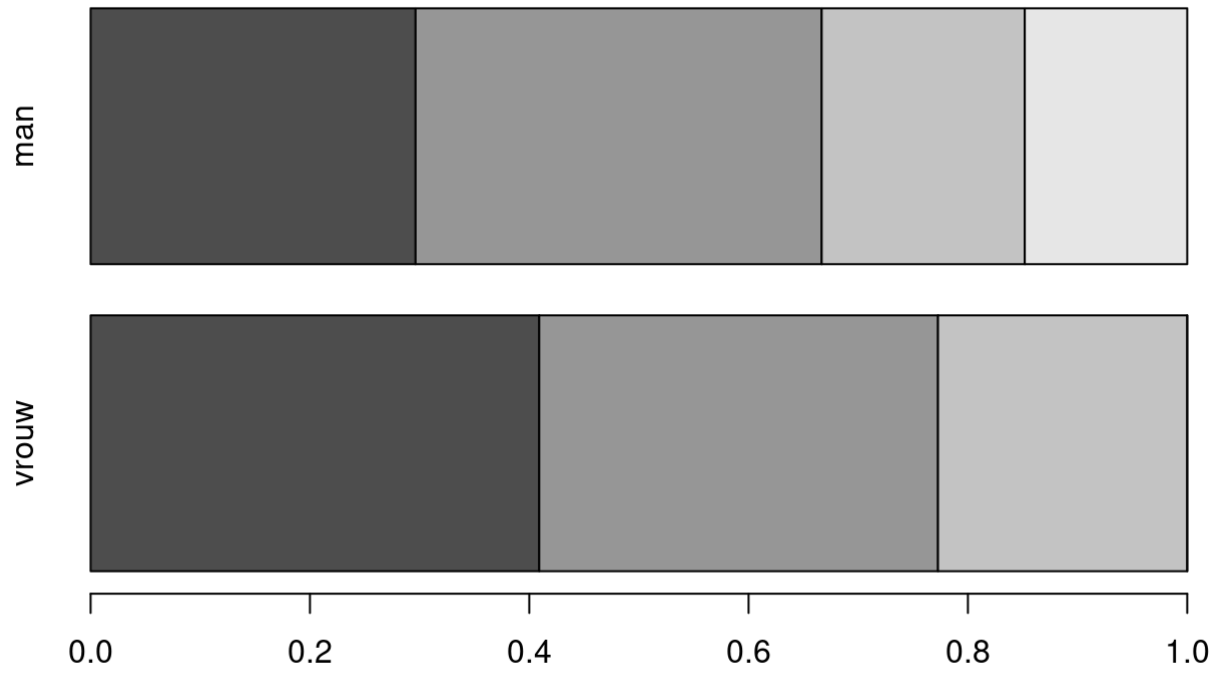
```
barplot(t(prop.table(observed, 2)),
        beside = TRUE,
        legend = TRUE)
```



De waarden van de afhankelijke variabele worden op de X-as uitgezet, en elke cluster toont de *relatieve* frequenties van de waarden in de onafhankelijke variabele. Per waarde in de onafhankelijke variabele (hier: man/vrouw) werden de frequenties in de afhankelijke variabele (goed t/m slecht) herberekend tot percentages.

Rependiagram

```
proportions <- prop.table(observed, margin = 2)
barplot(proportions, horiz = TRUE)
```



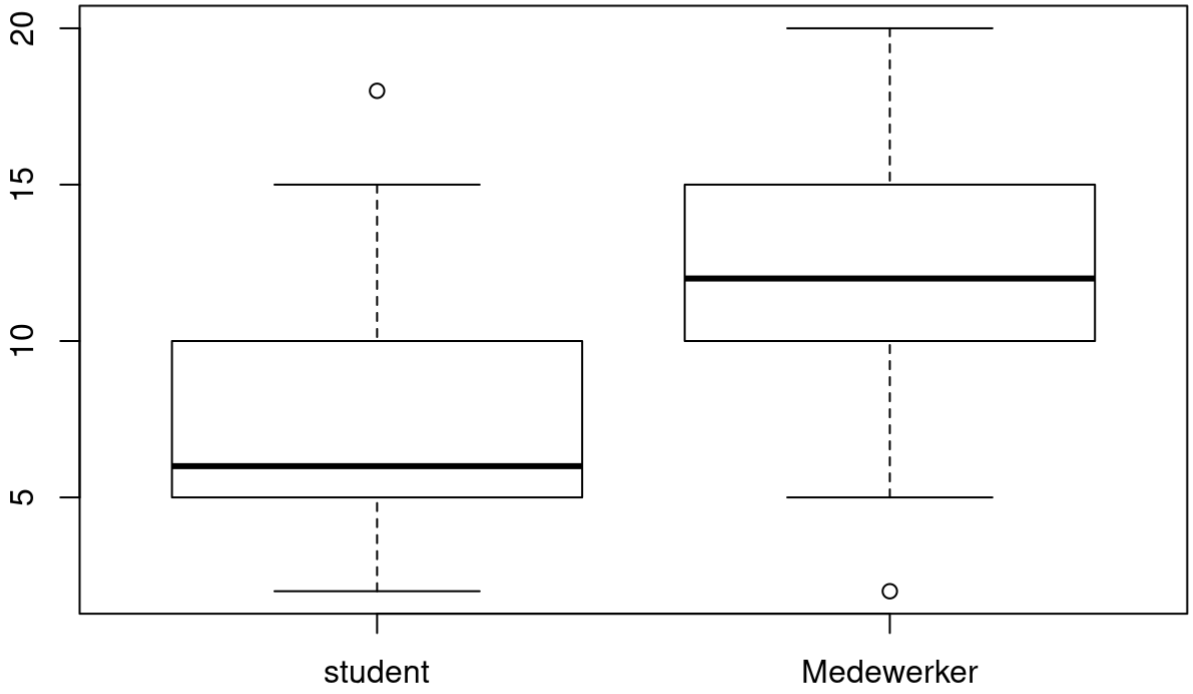
In deze grafiek worden eveneens *relatieve* frequenties getoond. Merk de gelijkenissen met het mozaïekdiagram op! Een mozaïekdiagram geeft echter nog iets meer informatie, want in het rependiagram is niet meer duidelijk of er een verschillend aantal mannen dan wel vrouwen is ondervraagd.

Kwalitatief - kwantitatief

Als voorbeeld van een verband tussen een kwalitatieve onafhankelijke en een kwantitatieve afhankelijke variabele nemen we “Besteden personeelsleden wekelijks een groter bedrag in het restaurant dan studenten?”

Gegroepeerde boxplot

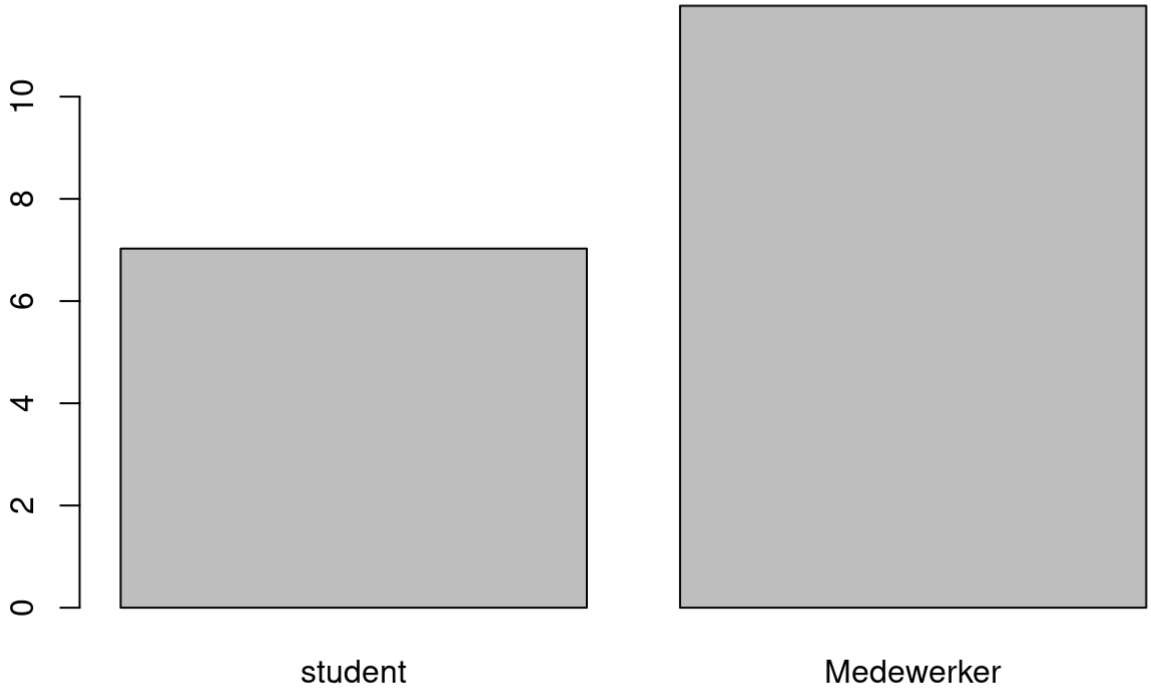
```
boxplot(resto$Bedrag ~ resto$Klanttype)
```



Staafdiagram van gemiddelden

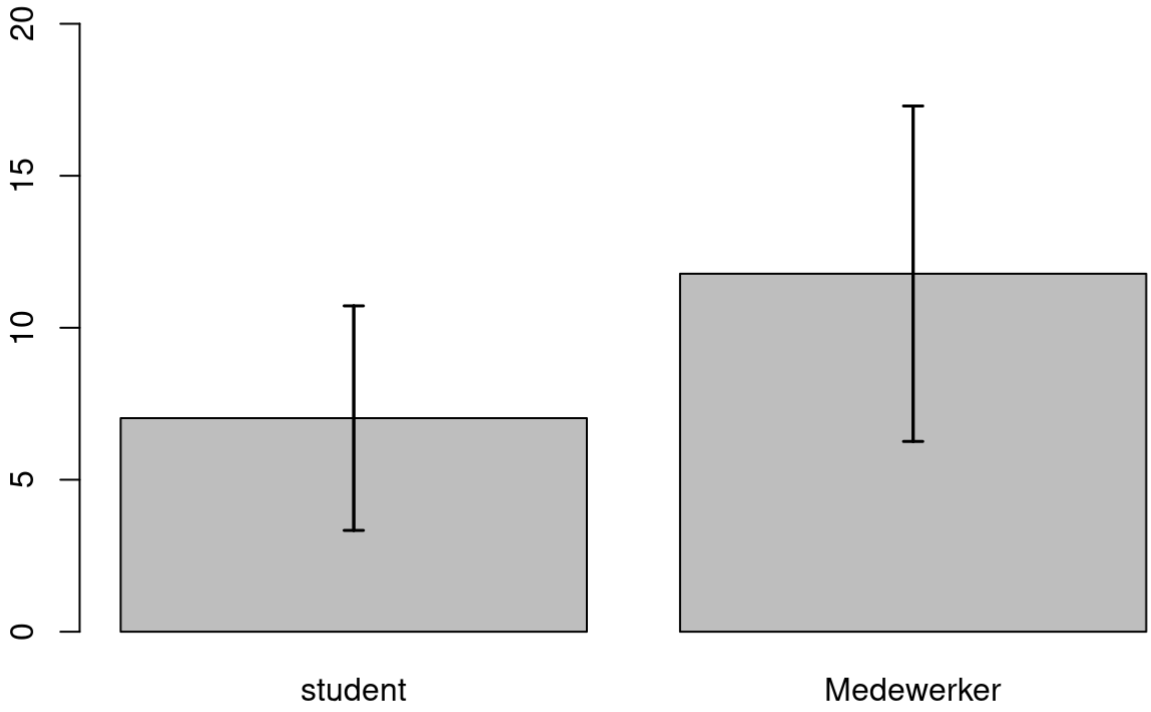
Een grafiektype dat we vaak tegenkomen voor dit soort onderzoeksvragen is een staafdiagram van het gemiddelde voor elke groep:

```
# Bereken eerst de gemiddelden voor elke groep
mean_amounts <- aggregate(resto$Bedrag ~ resto$Klanttype, FUN = mean)
# Plot het staafdiagram
barplot(mean_amounts$`resto$Bedrag`,
        names.arg = mean_amounts$`resto$Klanttype`)
```



In deze grafiek ontbreekt een uiterst belangrijk gegeven, en dat is de spreiding. Zoals de grafiek nu getekend is, geeft die **onvoldoende informatie** over het verschil tussen beide groepen. Die kan toegevoegd worden in de vorm van foutstaven (error bars) die de grootte van de standaardafwijking tonen.

```
# Bereken eerst de gemiddelden en standaardafwijkingen voor elke groep
mean_amounts <- aggregate(resto$Bedrag ~ resto$Klanttype, FUN = mean)
sd_amounts   <- aggregate(resto$Bedrag ~ resto$Klanttype, FUN = sd)
# Plot het staafdiagram
ymax <- max(resto$Bedrag, na.rm = TRUE) # zorg dat Y-as hoog genoeg is
mean_plot <- barplot(mean_amounts$`resto$Bedrag`,
                     names.arg = mean_amounts$`resto$Klanttype`,
                     ylim = c(0,ymax))
# Teken de vertikale lijnen van de error bars
segments(mean_plot,
         mean_amounts$`resto$Bedrag` - sd_amounts$`resto$Bedrag`,
         mean_plot,
         mean_amounts$`resto$Bedrag` + sd_amounts$`resto$Bedrag`,
         lwd = 1.5)
# Teken de einden van de error bars
arrows(mean_plot,
       mean_amounts$`resto$Bedrag` - sd_amounts$`resto$Bedrag`,
       mean_plot,
       mean_amounts$`resto$Bedrag` + sd_amounts$`resto$Bedrag`,
       lwd = 1.5, angle = 90, code = 3, length = 0.05)
```



Jammer genoeg is er geen eenvoudige manier om foutstaven toe te voegen aan een grafiek.

Het is belangrijk in het bijschrift van de grafiek te vermelden wat de error bars precies voorstellen. Soms wordt ook 2x de standaardafwijking getoond.

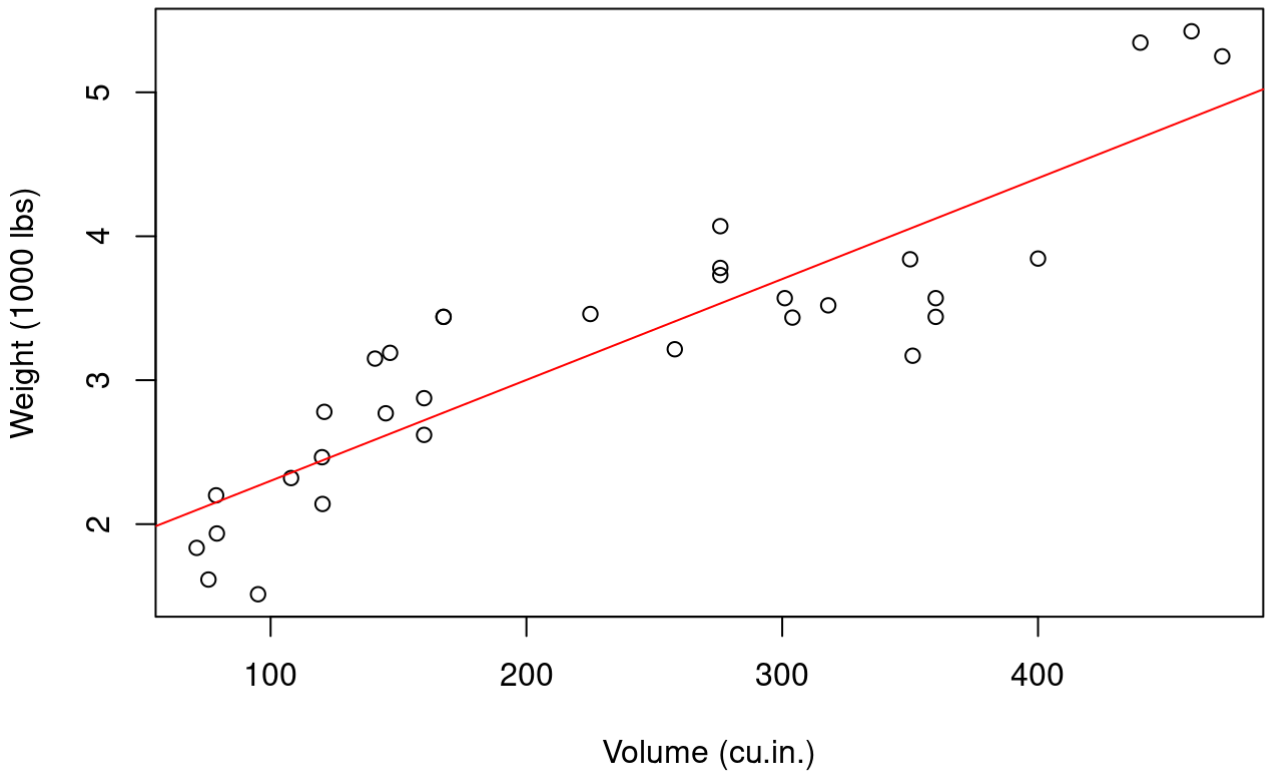
In de meeste gevallen is een boxplot een betere grafiek die meer informatie bevat. Bekijk het codevoorbeeld in de handleiding van `?boxplot` voor een vergelijking van beide grafiektypes. Vooral als er uitschieters in de data zitten, zal het gemiddelde en standaardafwijking een verkeerd beeld van de spreiding geven. Ook als de afhankelijke variabele niet normaal verdeeld is, heeft gemiddelde en standaardafwijking geen zin.

Kwantitatief - kwantitatief

Voor dit soort verbanden wordt typisch een XY-grafiek (scatter plot) gebruikt. De onafhankelijke variabele wordt typisch op de X-as uitgezet, de afhankelijke op de Y-as.

Laten we als voorbeeld bekijken of er in de dataset `mtcars` een verband is tussen het volume van een wagen (variabele `disp`, *displacement*, uitgedrukt in kubieke duim) en het gewicht (variabele `wt`, *weight*, in 1000 pond):

```
plot(mtcars$disp, mtcars$wt,
     xlab = "Volume (cu.in.)",
     ylab = "Weight (1000 lbs)")
regression <- lm(mtcars$wt ~ mtcars$disp) # bereken de regressierechte
abline(regression, col = 'red')           # teken die in het rood
```



De algemene `plot()`-functie zal voor elk datapunt een cirkel teken in een Cartesiaans assenstelsel. Je kan uiteraard het symbool veranderen (kruisje, punt, sterretje, …). Zie de handleiding van `?plot` voor meer info.

De regressierechte wordt hier ook getoond in het rood. Daarmee is wel duidelijk dat er een positief verband bestaat tussen volume en gewicht (wat je ook wel kan verwachten: grotere wagens zijn zwaarder).