Handleiding Kansrekenen en Statistiek met R

dr. Wim P. Krijnen
Rijksuniversiteit Groningen
Faculteit Wiskunde en Natuurwetenschappen
Statistiek en Stochastiek

Oktober 2013

Inhoudsopgave

1 Inleiding

Er bestaan veel computer programma's om statistisch mee te rekenen. We geven enkele overwegingen om over te gaan tot het programmeren met R. Het programma VUSTAT houdt op te bestaan. Programma's als SPSS en Excell geven aanschafkosten en zijn minder geschikt voor het rekenen op grote data sets of het berekenen van kansen van verdelingen. Excell geeft relatief weinig uitleg bij het uitvoeren van statistische toetsen waardoor gemakkelijk verwarring kan ontstaan. Zowel Excell als SPSS geven de mogelijkheid om binnen deze programma's met R te rekenen.

Gelukkig vormt de statistische programmeertaal R een goed alternatief. De aanschaf is gratis en wereldwijd de standaard op het gebied van kansrekening en statistiek. Het wordt op universiteiten zeer veel gebruikt in de beta/gamma wetenschappen en in iets mindere mate in de alfa wetenschappen. Verdieping via R in kansrekening en statistiek vormt een uitstekende voorbereiding op het wetenschappelijk onderwijs.

Om R aan te leren vormt het menu systeem geboden door de R-commander een zeer handig hulpmiddel bij het inlezen van data, het berekenen van descriptief statistische grootheden, en het uitvoeren van statistische toetsen. Aangezien de ervaring leert dat het menu systeem van R-commander "in de weg gaat staan"bij herhaald gebruik, is ervoor gekozen om functies direct te verduidelijken. Het gebruik van functies kan in de vorm van kleine scripts worden opgeslagen voor verdere verbetering of later gebruik.

Met R komen veel functies beschikbaar om mee te kunnen programmeren. Het is relatief eenvoudig om simulaties uit te voeren, waardoor een dieper begrip van concepten uit de kansrekening mogelijk wordt. Door de vele data sets die via R toegankelijk worden is het zeer eenvoudig statistiek te combineren met elk ander vak op de middelbare school. Het is zeer geschikt voor een opdracht van 10 studielasturen of een profiel-werkstuk aangezien het eenvoudig is om zelf verzamelde data via een tekst of Excel file in te lezen en statistisch te verwerken.

Doel van de huidige tekst is het geven van een inleiding op het gebruik van statistiek en kansrekening met R. De tekst is geschreven onder de aanname dat statistische concepten zoals toetsen elders zijn toegelicht en met de bedoeling dat lezers de gegeven functies actief in R gebruiken. Pas dan komt de tekst tot leven. Er is niet naar gestreefd om de Engelse taal compleet te vermijden.

Er zal gebruik worden gemaakt van enkele bibliotheken (packages) met functies en data. De noodzaak een bibliotheek te installeren zal niet steeds worden herhaald. Letterlijke tekst voor de programmeertaal R zal weergegeven worden in de `verbatim` stijl. De prompt `>` geeft code en de verwerking daarvan. Een □ betekend het einde van een voorbeeld. Door met de linker muis knop naar een URL te verwijzen stuurt men de standaard browser naar het WWW adres. De tekst word aangevuld met een ASCI file met R-code.

2 Starten met R

2.1 R downloaden

Om R te downloaden ga je naar de URL `http://cran.r-project.org`, kiest het gewenste Operating System: Windows, Linux, of Mac OS, vervolgens `base` en `html help`. Volg de verdere instructies.

R kan ook geïnstalleerd worden op de Ipad, maar we hebben hier weinig ervaring mee.

Aanbevolen tekstverwerkers zijn Notepad ++, Kate, emacs, en Word. Voor laatst genoemde dient het automatisch gebruik van hoofdletters uitgezet te worden aangezien R een verschil maakt tussen kleine en hoofdletters. Het gebruikt van syntax highlighting {[(wordt aanbevolen.

2.2 Library installeren

Veel extra mogelijkheden komen beschikbaar door een bibliotheek te installeren via `Packages` bovenaan het scherm of direct door

```
install.packages("Rcmdr",repo="http://cran.r-project.org",dep=TRUE)
```

Dit installeert de library van Het menu systeem "R commander". Dit systeem wordt aanbevolen in een stadium van beginnend gebruik aangezien het de gegenereerde code en de output toont. R commander wordt opgestart door het laden de bibliotheek.

```
library(Rcmdr)
```

Hoewel het systeem zich door de eenvoud zelf wijst, verwijzen we voor extra informatie naar:

- `http://cran.r-project.org/doc/contrib/Karp-Rcommander-intro.pdf`
- `http://www.jstatsoft.org/v14/i09/paper`

3 Vinden van hulp

De site `http://cran.r-project.org` heeft

- een handige `Search` knop waarmee men deze kan doorzoeken;
- onder de `Contributed` knop, de tutorials: "R for Beginners", "The R Guide"
- onder de `Contributed` knop, de zeer handige "R reference card"(Tom Short) met een uitgebreid en beknopt overzicht van veel gebruikte functies

Hulp binnen het R-systeem:

- `help.search()` de manual "An Introduction to R"(pdf or HTML).
- indien de naam van de functie bekend is kan de manual geopend worden door `help(t.test)` of, nog eenvoudiger, `?t.test`

Opmerking: Het bespaart tijd eerst de correcte Engelse term te achterhalen en pas dan te zoeken.

In de Nederlandse taal bestaan weinig informatiebronnen over R gebruik. Een goede uitzondering hierop is het boek geschreven door Erik van Zwet: `http://www.math.leidenuniv.nl/~evanzwet/boek.pdf`.

Er bestaan twee gebruikersgroepen in Nederland:

- Amsterdam: `http://www.meetup.com/amst-R-dam/`
- Enschede: `http://twenterug.wordpress.com/about/` (Alleen intern UT gebruik.)

4 Enkele handige functies

- `getwd()` get working directory
- `setwd()` set working directory
- `choose.dir()` set working directory via clicking the mouse
- `dir()` lijst met files/mappen op working directory
- `str()` structure van een object
- `with()` geeft beschikking over namen van kolommen
- `apply()` past functie toe op elke rij of kolom van een matrix
- `lm(y ~x, data=dat` voer lineaire regressie uit op response variable `y` en verklarende variabele `x`, welke zich als kolommen in de dataset `dat` bevinden
- `plot()` maakt afbeelding

5 Data inladen

Een adres met veel data is: `http://www.statsci.org/datasets.html`
Bibliotheken met data sets zijn onder andere:

```
library(datasets)
library(MASS)
library(BSDA)
```

Via `?datasets` en `Index` krijg je toegang tot de inhoud van de bibliotheek `datasets`.

Voorbeelden van data sets zijn:

```
library(MASS)
?nlschools   # (N=2287) eighth-grade pupils
?SP500       # (n=2780)  Returns of Standard and Poors 500 Index

library(survey)
?election   # (n=4600)    8

library(alr3)
?heights  # (n=1375) lengte moeder/dochter Pearson (1903)
?UN2      #  193
```

Een dataset in de vorm van een tabel bestaande uit rijen en kolommen kan direct van het WWW worden laden:

```
d <- read.table("http://www.stat.washington.edu/hoff/Book/Data/hwdata/azdiabetes.dat",
  header=TRUE)
```

Met deze functie kunnen ook data in `.txt` formaat worden geladen, zie `?read.table()`.

De bibliotheek `foreign` heeft functies om data van verschillende talen, zoals SAS, Stata, Epi, Minitab, Octave, en SPSS, te laden.

Hoewel, de library `gdata` een `read.xls` functie heeft, wordt afgeraden om Excell data in `.xls` format in te lezen aangezien de typisch Nederlandse gewoonte met comma getallen te werken parten speelt. Het is eenvoudiger om alleen de eerste regel te reserveren voor namen van kolommen, de resterende regels voor data, vervolgens de file weg te schrijven in `CSV (MS-DOS)` formaat en deze in R te laden met de functie `read.csv2`.

6 Beschrijvende Statistiek

Enkele functies om beschrijvend statistische grootheden mee te berekenen op een steekproef van gegevens:

- gemiddelde `mean`
- mediaan `median`
- standaard deviatie: `sd`
- variantie: `var`
- inter quartile range: `IQR`

Opmerking 1: De standaard deviatie is de tweedemachtswortel van de variantie.
Opmerking 2: Het gemiddelde en de mediaan geven inzicht in de locatie van de verdeling, de standaard deviatie, variantie en inter quartile range in de variatie van de gegevens.
Opmerking 3: Het gemiddelde en de standaard deviatie zijn niet robuust tegen extreme uitbijters, de mediaan en de inter quartile range wel.
Opmerking 4: Als de data normaal verdeeld zijn, dan geeft `IQR(chem)/1.349` een schatting van de standaard deviatie.

Voorbeeld 1: Van 24 bepalingen van koper in volkorenmeel, delen per miljoen, berekenen we de beschrijvend statistische grootheden. Uit een grotere studie bleek dat het populatie gemiddelde $\mu = 3.68$. De data zijn door het laboratorium beoordeeld op transcriptie fouten en in orde bevonden (Venables and Ripley, 2002, p. 125).

```
> library(MASS)
> data(chem)
> median(chem)
[1] 3.385
> mean(chem)
[1] 4.280417
> sd(chem)
[1] 5.297396
> IQR(chem) / 1.349
[1] 0.6856931
```

De mediaan en het gemiddelde verschillen sterk. De standaard deviatie is veel groter dan de schatting daarvan met behulp van de IQR. Deze verschillen worden met name veroorzaakt door één extreme uitbijter. □

Voorbeeld 2: De lichaamslengte van 1375 moeders en dochters is gemeten in inches (Pearson and Lee, 1903). We verduidelijken het berekenen van descriptief statistische grootheden op twee kolommen van een matrix (data frame) met behulp van `apply`. De aanpassing van de IQR functie laat zien dat we eenvoudig elke gewenste functie te gebruiken.

```
> library(alr3)
> apply(heights, 2, mean)
 Mheight  Dheight
62.45280 63.75105
> apply(heights, 2, median)
Mheight Dheight
   62.4    63.6
> apply(heights, 2, sd)
 Mheight  Dheight
2.355103 2.600053
> apply(heights, 2, function(x) IQR(x)/ 1.349)
 Mheight  Dheight
2.297999 2.668643
```

Opmerking: De twee soorten schattingen van locatie en standaarddeviatie komen sterk overeen. □

7 Afbeeldingen van data

Enkele afbeeldingen om normaliteit van data visueel te beoordelen zijn:

- histogram: `hist`
- box-and-wiskers-plot: `boxplot`
- Quantile-Quantile `qqnorm` gevolgd door `qqline`

Voorbeeld 1: Box-and-wiskers-plot van lengte moeders en dochters, histogram van moeders, en QQ-afbeelding van moeders.

```
library(alr3)
boxplot(heights)  #geeft twee plots naast elkaar

with(heights, hist(Mheight,freq=FALSE))
dens <- function(x){dnorm(x, mean=62.45, sd=2.35)}
x <- seq(55,75,0.1)
lines(x,dens(x),col='blue',lwd=3)  #toevoegen van dichtheid functie

with(heights, qqnorm(Mheight))
with(heights, qqline(Mheight))
```

Opmerking 1: De box-en-wiskers afbeeldingen suggereren dat de dochters iets langer zijn.
Opmerking 2: De afbeeldingen laten zien dat lengte normaal verdeeld is. □

In R kunnen afbeeldingen worden gemaakt in elk gewenst bestandsformaat. Het `.jpg` formaat is handig om een afbeelding in Word op te nemen.

Voorbeeld 2: Het opslaan van een afbeelding in `.jpg` formaat naar de harde schijf.

```
library(alr3)
setwd("C:\\work\\RuGOnderwijs\\VernieuwingWisVWO")
tiff(filename = "LengteMoedersDochters.tiff", width = 30, height = 15,
  units = "cm", pointsize = 12, bg = "white", res = 600, restoreConsole = TRUE)
with(heights, hist(Mheight,freq=FALSE))
dens <- function(x){dnorm(x, mean=62.45, sd=2.35)}
x <- seq(55,75,0.1)
lines(x,dens(x),col='blue',lwd=3)  #toevoegen van dichtheid functie
dev.off()
```

Opmerking: Met `getwd()` vind je de map met de file `LengteMoedersDochters.jpg`. □

8 Kansrekening

Het aantal permutaties van de letters a, b, c, d is $4 \cdot 3 \cdot 2 \cdot 1 =$ `prod(1:4)`.
Het aantal permutaties van 4 letters uit een totaal van 6 verschillende is

$$\frac{6!}{(6-4)!} = \texttt{prod(1:6)/prod(1:2)} = 360$$

Het aantal combinaties van 4 letters uit een totaal van 6 verschillende is

$$\frac{6!}{4!(6-4)!} = \texttt{prod(1:6)/(prod(1:4) * prod(1:2))} = \texttt{choose(6,4)} = 15$$

Voorbeeld 1: Het aantal manieren waarop 10 juiste antwoorden uit 40 vragen kunnen worden gegeven is

```
> choose(40,10)
[1] 847660528
```

□

Voorbeeld 2: Wat is de kans dat er in een klas van 23 leerlingen twee op dezelfde dag jarig zijn? We gaan ervan uit dat de verjaardagen volkomen onafhankelijk zijn. We bereken eerst de kans dat er geen twee leerlingen bestaan die op dezelfde dag jarig zijn.

```
> prod(365:343)/365^23
[1] 0.4927028
> 1 - prod(365:343)/365^23
[1] 0.5072972
```

Daarna gebruiken de complement-regel. □

9 Veel gebruikte kansverdelingen

Enkele veel gebruikte kansverdelingen zijn

- binomiaal (`binom`)
- Poisson (`pois`)
- uniform (`unif`)
- normaal (`norm`)

Van elk zijn vier functie beschikbaar beginnend met de letter:

- `d`: "density"(dichtheid). Voor discrete random variabele X is dit $P(X = x)$.
- `p`: "probability". De cumulatieve kans functie $P(X \leq x)$.
- `q`: "quantile"; Geeft waarde x_α zodat $P(X \leq x_\alpha) = \alpha$.
- `r`: "random deviate". Geeft $x_1, x_2, \cdots, x_n$ uit verdeling van random variable X.

Voorbeeld 1: De binomiale verdeling.

- De kans op k succes uit n bij succeskans p is `dbinom(x, n, p)` $= \binom{n}{x} p^x (1-p)^{n-x}$.
- De kans op 10 of minder goed bij willekeurig invullen van 40 4-keuze vragen is `pbinom(10, 40, 0.25)`.
- Aantal succes bij 40 4-keuze vragen passend bij cumulatieve kans 90% is `qbinom(0.90, 40, 0.25)`
- Aantal juist van 50 studenten die willekeurig 40 4-keuze vragen beantwoorden: `rbinom(50, 40, 0.25)` □

Voorbeeld 2: De kans op 14 of minder goed bij het willekeurig beantwoorden van 40 vragen met 4-keuze mogelijkheden:

```
> pbinom(14, 40, 0.25)
[1] 0.9455628
```

Als men er minstens 95% zeker wil zijn van aanwezigheid van enige kennis, dan kan er gekozen worden voor een omreken tabel vanaf 15 juiste antwoorden. □

Voorbeeld 3: We berekenen de verwachting en de variantie van een binomiale verdeling met parameters $n = 40$ en $p = 1/4$.

```
n <- 40; p <- 1/4
x <- 0:n
(EX <- sum(x*dbinom(x,n,p)))
(VarX <-  sum((x-EX)^2*dbinom(x,n,p)))
```

Opmerking: Door de parameters n, p te veranderen kan eenvoudig de verwachting en variantie van een andere binomiale verdeling worden berekend. □

10 Statistische Toetsen

- `prop.test` toetst gelijkheid van proporties $H_0 : p_1 = p_2$ tegen verschil $H_0 : p_1 \neq p_2$.
- `t.test` toetst gelijkheid van gemiddelden $H_0 : \mu_1 = \mu_2$ tegen verschil $H_0 : \mu_1 \neq \mu_2$.
- `z.test` in library `BSDA` toetst gelijkheid van gemiddelde $H_0 : \mu = \mu_0$ tegen $H_0 : \mu \neq \mu_0$ onder aanname dat standaarddeviatie σ bekend is.
- `SIGN.test` in library `BSDA` toetst verschil in twee random variabelen $H_0 : p = 0.50$ tegen $H_0 : p \neq 0.50$, waarbij $p = Pr(X > Y)$.

Genoemde toetsen kunnen ook op één steekproef worden uitgevoerd. We nemen overal significantie niveau 0.05, dat wil zeggen dat we de nul hypothese verwerpen als de p-waarde kleiner is dan 0.05. Een equivalente manier is het verwerpen af te laten hangen van de gebeurtenis dat het betrouwbaarheidsinterval nul bevat.

Voorbeeld 1: Een machine fabriek test geproduceerde beeldschermen op kwaliteit voordat ze worden afgeleverd aan de klant. De engineer vind dat het productie process in orde is als er slechts 2% van de beeldschermen worden afgekeurd. Elke dag worden er 250 schermen getest. Op een dag worden er 11 afgekeurd. Is het productie-process in orde?

```
> prop.test(11, 250, p=0.02)

        1-sample proportions test with continuity correction

data:  11 out of 250, null probability 0.02
X-squared = 6.1735, df = 1, p-value = 0.01297
alternative hypothesis: true p is not equal to 0.02
95 percent confidence interval:
 0.02330015 0.07954098
sample estimates:
    p
0.044
```

Beslissing: De null-hypothese $H_0 : p_1 = 0.02$ wordt verworpen.
Conclusie: De proportie afgekeurde schermen is groter dan 2%.
Implicatie: Ingrijpen lijkt verstandig. □

Voorbeeld 2: Is de vleugel-lengte van de mug-soorten Amerohelea fasciata en Amerohelea pseudofasciata verschillend?

```
> library(Flury)
> data(midge)
> with(midge, t.test(Wing.Length ~ Species))

        Welch Two Sample t-test

data:  Wing.Length by Species
```

```
t = -2.1697, df = 12.967, p-value = 0.0492
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2439471978 -0.0004972466
sample estimates:
 mean in group Af mean in group Apf
         1.804444          1.926667
```

Beslissing: De null-hypothese wordt verworpen.
Conclusie: De vleugel lengte van Amerohelea pseudofasciata is groter dan van Amerohelea fasciata. □

Voorbeeld 3: Is de gemiddelde lengte van dochters groter dan die van hun moeders? Had Karl Pearson al in (1903) experimenteel bewijs voor deze conclusie?

```
> library(alr3)
> with(heights, t.test(Dheight, Mheight, paired=TRUE))

        Paired t-test

data:  Dheight and Mheight
t = 19.184, df = 1374, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.165499 1.431010
sample estimates:
mean of the differences
               1.298255
```

Beslissing: De null-hypothese wordt verworpen.
Conclusie: De lengte van dochters is groter dan die van moeders. □

Opmerkingen:

- De gepaarde t-toets is hier het meest geschikt omdat moeders en dochters paren vormen.
- Hier is $n = 1375$ en dus zo groot dat er geen waarneembaar verschil is met de z-toets.
- Het wordt in het algemeen aanbevolen om twee-zijdig te toetsen om de mogelijkheid van een verrassende uitkomt open te houden.
- Het wordt afgeraden om de t-toets voor twee steekproeven onder de expliciete aanname van gelijke populatie variantie te gebruiken. Deze kan eenvoudig worden vervangen door de meer algemene t-toets zonder genoemde aanname (`var.equal = FALSE`).
- Bij onderzoek naar gen-expressie in de bio-informatica komt het voor dat de onderzoeker de statistische toets 10^4 keer wil toepassen. Het is dan handig om alleen de

p-waarde op te slaan en deze verder te analyseren. Dit geldt ook voor simulaties. De output van `t.test()` is een lijst waaruit met `$p.value` de p-waarde kan worden gehaald en weggeschreven.

11 Enkele toepassingen

11.1 Wet van de grote aantallen

De wet van de grote aantallen zegt dat het gemiddelde van een rij stochastische variabelen bij toenemende omvang van de rij convergeert. Stel we hebben een rij met onafhankelijk verdeelde stochastische variabelen $X_1, X_2, X_3, \cdots$ die elk dezelfde verdeling hebben met populatie gemiddelde μ en eindige variantie σ^2. Het gemiddelde van de eerste n variabelen is te schrijven als $\overline{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$. De afstand tussen μ en $\overline{X}_n$ schrijven we als $\left|\overline{X}_n - \mu\right|$. De gebeurtenis dat de afstand tussen het gemiddelde $\overline{X}_n$ en het populatie gemiddelde kleiner is dan een positive ϵ kan geschreven worden als $\left|\overline{X}_n - \mu\right| < \varepsilon$. Een dergelijke ongelijkheid is waar of onwaar. De geldigheid van de ongelijkheid is van toeval afhankelijk aangezien $\overline{X}_n$ een stochastische variabele is die waardes aanneemt met een bepaalde kans. Het is dus onmogelijk om met zekerheid iets te zeggen over de waarheid van de ongelijkheid. We kunnen wel iets zeggen over de kans dat de ongelijkheid waar is als de grootte van de steekproef n steeds blijft toenemen. Namelijk als n toeneemt dan volgt dat de kans dat de ongelijkheid waar is naar 1 gaat, ofwel

$$\lim_{n\to\infty} P\left(\left|\overline{X}_n - \mu\right| < \varepsilon\right) = 1.$$

Met andere woorden: Voor elke kleine maar vaste waarde van ε is de ongelijkheid waar bij toenemende n.

Voorbeeld: Meerkeuze toets. We simuleren 1000 leerlingen die elk 40 4-keuze vragen willekeurig beantwoorden. Vervolgens berekenen het gemiddelde voor toenemende steekproefomvang n en maken een afbeelding van het gemiddelde.

```
s <- 10^3; n <- 40; p <- 1/40
mu <- n*p
x <- rbinom(s, n, p)
cmean <- cumsum(x)/(1:s)  # cumulatief gemiddelde
plot(1:s,cmean, ylim=c(0,20), type="l",col='red')
abline(h=mu)
```

Opmerking 1: Het is leerzaam de afbeelding herhaald naar het scherm te sturen om het verschil in snelheid van de convergentie te observeren.
Opmerking 2: Een plot is een object waarop we kunnen programmeren bijvoorbeeld een rechte lijn toevoegen. □

Opmerking: De stelling zegt niets over de snelheid van de benadering. Door de parameters n, p te veranderen en het script een aantal keer te herhalen krijgt je inzicht in de snelheid van de benadering. De wet van de grote aantallen speelt een belangrijke rol in het benaderen van posterior verdelingen in de Bayesiaanse statistiek.

11.2 Centrale limiet stelling

De wet van de grote aantallen kan worden afgeleid uit de centrale limiet stelling. Deze stelling zegt dat als we een rij met onafhankelijk verdeelde stochastische variabelen $X_1, X_2, X_3, \cdots$ uit dezelfde verdeling komen met gemiddelde μ en variantie σ^2, dan geldt dat

$$\frac{\sum_{i=1}^{s} X_i - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} \mathrm{N}(0,1).$$

De verdeling van het linker deel nadert de standaard normale verdeling. Er bestaan verschillende uitdrukkingen voor deze benadering

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathrm{N}(0,1).$$

maar ook speciaal voor toepassing op de binomiale verdeling

$$\sum_{i=1}^{s} X_i \xrightarrow{d} \mathrm{N}(n\mu, n\sigma^2).$$

Voorbeeld: Meerkeuze toets. We simuleren 10^5 studenten die elk 40 4-keuze vragen willekeurig beantwoorden. Hier is $n = 40$, $p = 1/4$, zodat $\mu = np$ en $\sigma^2 = p(1-p)$. We maken een kansverdelings histogram en tekenen daar doorheen de dichtheidsfunctie van de normale verdeling.

```
s <- 10^5; n <- 40; p <- 1/4
y <- rbinom(s, n, p)
hist(y, xlim=c(0,2*n*p), ylim=c(0,0.18), freq=FALSE, nclass = 21)
dens <- function(x){dnorm(x, mean=n*p, sd=sqrt(n*p*(1-p)))}
x <- seq(0, 2*n*p, 0.01)
lines(x, dens(x), col='blue',lwd=3)
```

Opmerking: Door n op te hogen wordt de benadering beter. □

11.3 Lineaire regressie

In het lineaire model wordt de waarde van de response variable Y_i voortgebracht door een lineaire functie van de verklarende variable x_i plus een fout term ε_i, waarbij de index i loopt van 1 tot n. Het model wordt geschreven als

$$Y_i = \alpha + \beta x_i + \varepsilon_i,$$

hierbij is α het begin getal (intercept) en β de hellingscoëfficiënt (slope). Indien $n > 2$ dan past er geen rechte lijn door de punten $(x_1, y_1), \cdots, (x_1, y_1)$ (waarnemingen) en wordt de best passende lijn volgens het kleinste kwadraten criterium berekend door de functie `lm`. De input van deze functie bestaat uit de vectoren `x`, `y` en de model notatie `y ~ x`. De output bestaat uit een lange lijst met resultaten van de schatting van het model. Hiervan wordt met `summary` de essentie samengevat.

Voorbeeld: Galileo deed in 1609 een natuurkundig experiment waarbij een balletje van een baan met een bepaalde hoogte liet rollen. De horizontaal afgelegde afstand werd opgemeten in de eenheid punti (169/180 millimeter). In de dataset is de horizontaal afgelegde afstand (`h.d`) de response variabele en begin hoogte (`init.h`) de verklarende variabele. We schatten de best passende rechte lijn door de metingen en presenteren de punten en de lijn in een afbeelding.

```
library(UsingR)
lin.mod <-  lm(h.d ~ init.h, data = galileo)
plot(h.d ~ init.h, data = galileo)
abline(lin.mod)
summary(lin.mod)
```

Opmerking: Het lineaire model lijkt te eenvoudig. □

11.4 Driehoek van Pascal

De driehoek van Pascal is een rangschikking van de binomiaalcoëfficiënten $\binom{n}{k}$ in rijen voor toenemende n beginnend met $n = 0$ en op elke rij de $n + 1$ binomiaalcoëfficiënten voor de mogelijke waarden van k. In de driehoek komt de eigenschap tot uitdrukking dat elke binomiaalcoëfficiënt de som is van de twee bovenliggende. De getallen in de driehoek geven het aantal wegen aan vanaf de top naar de plaats van zo'n getal, waarmee ook de besproken eigenschap verklaard is. Omdat er steeds 2 mogelijkheden zijn om de weg naar onder te vervolgen is de som van de getallen op een rij de overeenkomstige macht van 2, zie `http://nl.wikipedia.org/wiki/Driehoek_van_Pascal`.

Voorbeeld 1: De driehoek kan als volgt worden berekend.

```
for (n in 0:10)  cat(choose(n,0:n),'\n')
```

Met de functie `cat` schrijven we regels tekst naar een file of scherm. □

Voorbeeld 2: We gaan na dat $\binom{n}{k} = \binom{n}{n-k}$ voor $n = 10$ en $k = 0, 1, 2, \cdots, 10$

```
n <- 10 ; k <- 0:n
choose(n,k) == choose(n,n-k)
```

Opmerking: `k` is hier een vector met waarden van 0 to 10. Het dubbele gelijkheidsteken wordt geëvalueerd op waarheid. □

Voorbeeld 3: We gaan na dat $\sum_{k=0}^{n} \binom{n}{k} = 2^k$ voor $n = 0, 1, 2, \cdots, 10$

```
for (n in 0:10) print(sum(choose(n, k = 0:n))==2^n )
```

Opmerking 1: R is zeer geschikt om samengestelde functies te gebruiken.
Opmerking 2: De output elf keer `TRUE` geeft aan dat het klopt. □

12 Literatuur

Partridge, L. and Farquhar, M. (1981). Sexual Activity and the Lifespan of Male Fruitflies. Nature, 580-581
Pearson, K. and Lee, A. (1903). On the laws of inheritance in man. Biometrika, 2, 357463, Table 31.
R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org/`.
Venables, W. N. and Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth edition. Springer.

13 Opgaven

1. Verjaardag. (Geschikt als opdracht?)

   (a) Maak een grafiek die de kans (verticaal) aangeeft dat er niemand tegelijk jarig is afhankelijk van de grootte van de klas (horizontaal).

   (b) Simuleer een klas en ga na of er dezelfde verjaardagen voorkomen.
   Hint: Gebruik `unique` om na te gaan welke verjaardagen uniek zijn en `length` om na te gaan hoeveel er dat zijn.

   (c) Herhaal (b) 100000 keer en benader de kansverdeling van het aantal verjaardagen in een klas met 23 leerlingen.

   (d) Herhaal (c) een aantal keer en vergelijk de antwoorden.

2. Grote steekproef. Genereer een steekproef van grootte 10^6 uit de binomiale verdeling met parameters $n = 5$ en $p = 1/3$. Tel het aantal keer dat je $X = 2$ hebt gekregen en deel dit door 10^6. Vergelijk dit met $P(X = 2)$ van de binomiale verdeling. Wat valt je op?

3. Verwachting en variantie Poisson verdeling. Bereken de verwachting en variantie van een Poisson verdeling met parameters $\lambda = 3$. Maak gebruik van de definitie van verwachting en variantie.

4. Normale benadering van Binomiale verdeling. De binomiale verdeling met parameters n, p kan benaderd worden door de normale verdeling met parameters np en $np(1-p)$.

   (a) Neem $n = 40$, $p = 1/2$ en maak één afbeelding van zowel kansverdeling als de benadering.

   (b) Neem $n = 40$, $p = 1/20$ en maak één afbeelding van zowel kansverdeling als de benadering.

   (c) Start nu de animatie `vis.binom()` op uit library `TeachingDemos`.

5. IQR als schatting voor SD. Ga met een simulatie na dat de IQR gedeeld door 1.349 een schatting geeft van de standaard deviatie bij normaal verdeelde data. Hoe kun je een grote n gebruiken?

6. Lichaamslengte. De data set `heights` in de library `alr3` geeft lengte van moeders en dochters.

   (a) Bereken het aantal dochters dat langer is dan de moeders.

   (b) Bereken de mediaan van het verschil in length tussen de dochters en de moeders.

   (c) Gebruik de teken toets om na te gaan of het verschil in lengte significant is.

   (d) Geef commentaar op het gebruik van de teken toets in deze situatie.

7. Fruitvlieg. In een experiment werden 125 fruitvliegen willekeurig verdeeld in 5 groepen van 25 elk. We concentreren ons op de groep mannetjes die solitair werd gehouden en de groep die samen leefden met 8 maagdelijke vrouwtjes. De response was de levensduur van de fruitvlieg in dagen (Partridge and Farquhar, 1981). De data zijn beschikbaar onder de naam `fruitfly` in de library `faraway`.

   (a) Selecteer de data van beide groepen in een data frame. Hint: Gebruik de functie `subset`.

   (b) Maak een box-and-wiskers plot van de levensduur van de geïsoleerde groep en de groep met 8 vrouwtjes.

   (c) Toets of de groepsgemiddelden verschillen.

8. Klas acht. Snijders en Bosker (1999) verzamelden gegevens van 2287 leerlingen uit 132 groep acht klassen in 131 verschillende scholen in Nederland. Zie `nlschools` in `MASS`.

   (a) Maak een boxplot van het IQ voor leerlingen die wel en die niet in een combinatie klas zaten. Wat valt je op?

   (b) Toets de verschillen in IQ tussen leerlingen die wel en die niet in een combinatie klas zaten.

   (c) Doe hetzelfde voor sociaal economische status. Wat is jouw conclusie?

   (d) Doe hetzelfde voor de taal toets. Wat is jouw conclusie?

9. Power t-toets. De pH van bloed is verdeeld als $X \sim \text{normal}(7.4, 0.025)$.

   (a) Simuleer dat er 10^4 keer van een groep van 10 mensen een pH test wordt afgenomen en er 10^4 keer wordt getoetst of $H_0 : \mu = 7.4$ waar is. Neem aan dat de metingen normaal verdeeld zijn met gemiddelde $\mu = 7.4$ en variantie 0.025^2. Neem overschrijdingskans $\alpha = 0.05$. Bewaar de p-waarde van elke toets. Bereken over de 10^4 p-waarden de kans dat deze kleiner is dan 0.05. Welke uitkomst verwacht je? Herhaal de procedure enkele keren om te kijken of de uitkomt vergelijkbaar is.

   (b) Herhaal het onderdeel (a) waarbij de metingen normaal verdeeld zijn met gemiddelde $\mu = 7.375$ en variantie 0.025^2. Toets of $H_0 : \mu = 7.4$ waar is. De kans om een foutieve nulhypothese te verwerpen wordt de power van de statistische toets genoemd.

10. Formaldehyde. Een scheikundig experiment werd uitgevoerd om de calibratie lijn op te stellen voor de bepaling van de hoeveelheid formaldehyde door de toevoeging van zuur en het aflezen van de resulterende paarse kleur in foto-spectrometer. De data staan in de library `datasets` onder de naam `Formaldehyde`.

    (a) Neem `optden` als response variabele en `carb` als verklarende variable en schat het lineaire model.

    (b) Maak een afbeelding van de gegevens en de geschatte lijn.

    (c) Past de lijn goed bij de data?

11. UN. Enkele economische grootheden zijn bepaald voor 193 landen. Zie `UN3` in de library `alr3`.

    (a) Maak een afbeelding van percentage van de populatie dat stedelijk is (`Purban`) tegen het bruto binnenlands product per hoofd van de bevolking (`PPgdp`).

    (b) Doe hetzelfde als hiervoor maar gebruik nu de log van (`PPgdp`). Welk verband zien je?

    (c) Maak een afbeelding tussen vruchtbaarheid en jaarlijkse groei van de populatie. Welk verband zie je?

    (d) Schat het lineaire model voor jaarlijkse groei van de populatie als response en vruchtbaarheid als verklarende variabele. Wat valt je op?

14 Antwoorden

1. Verjaardag. Antwoord.

```
#(a)
n <- 1:70                          # vector van aan leerlingen
p <- numeric(70)                   # initializer van vector
for (i in n) {
  q = prod(1 - (0:(i-1))/365)   # P(Geen match) als i in de klas
  p[i] = 1 - q                  # complement
}
plot(n, p)                       # plot van p tegen n

#(b)
v = sample(1:365, n, replace=TRUE)    # 23 random verjaardagen in een klas van 23
23 - length(unique(v))                # no. of matches in i-de klas


#(c)
s <- 100000;  n <- 23                  # iterations; aantal leerlingen in klas
x <- numeric(s)                        # vector met aantal matches
for (i in 1:s){
  v <- sample(1:365, n, replace=TRUE)  # n random verjaardagen in de i-de klas
  x[i] = n - length(unique(v))         # aantal matches in i-de klas
```

```
}
table(x)/s                                  # of met mean(x == 0)
#(d) antwoorden komen zeer sterk overeen.
```

2. Grote steekproef. Antwoord.

```
dbinom(2,5,1/3)
mean(rbinom(10^6,5,1/3)==2)
```

   Het gemiddelde is in drie decimalen gelijk aan de kans.

3. Verwachting en variantie Poisson verdeling. Antwoord.

```
x <- 0:100
(EX <- sum(x*dpois(x,lambda=3)))
(VarX <- sum((x-EX)^2*dpois(x,lambda=3)))
```

4. Normale benadering van Binomiale verdeling.

```
n <- 40; p <- 1/20
x <- 0:n
y1 <- dbinom(x,n,p)
y2 <- dnorm(x,n*p,sqrt(n*p*(1-p)))
plot(x,y1,col='blue')
points(x,y2,col='green')

n <- 100
x <- 0:n
pp <- seq(0,1,1/200)
for (i in 1:length(pp)){
  p <- pp[i]
  y1 <- dbinom(x,n,p)
  y2 <- dnorm(x,n*p,sqrt(n*p*(1-p)))
  plot(x,y1,ylim=c(0,0.2),col='green')
  points(x,y2,col='blue')
  Sys.sleep(1)  #wacht 1 seconde en ga dan verder
}
```

5. IQR als schatting voor SD.

```
IQR(rnorm(10^7,7.40,0.025))/1.349
```

   Opmerking 10^8 werkt ook.

6. Lichaamslengte. Antwoorden

```
library(alr3); library(BSDA)
with(heights, sum(Dheight > Mheight))
with(heights, median(Dheight - Mheight))
with(heights,SIGN.test(Dheight,Mheight))
```

De teken-toets is hier niet optimaal aangezien de data normaal verdeeld zijn en deze informatie niet wordt meegenomen door de toets.

7. Fruitvlieg. Antwoord.

```
> library(faraway)
> dat <- subset(fruitfly, activity %in% c('isolated','high'))
> with(dat, boxplot(longevity ~ activity))
# data lijken normaal verdeeld met verschil in gemiddelde
> with(dat, t.test(longevity ~ activity))

        Welch Two Sample t-test

data:  longevity by activity
t = 6.0811, df = 44.09, p-value = 2.545e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 16.60817 33.07183
sample estimates:
mean in group isolated     mean in group high
                 63.56                  38.72
```

Beslissing: De nul hypothese wordt verworpen.
Conclusie: De fruitvliegmannetjes in de conditie met 8 vrouwtjes leven korter.

8. Klas acht.

```
library(MASS)
data(nlschools)
with(nlschools, boxplot(IQ ~ COMB))
with(nlschools, t.test(IQ ~ COMB))
with(nlschools, boxplot(SES ~ COMB))
with(nlschools, t.test(SES ~ COMB))
with(nlschools, boxplot(lang ~ COMB))
with(nlschools, t.test(lang ~ COMB))
```

9. Power t-toets.

```
pval <- NULL; n <- 10^4
for (i in 1:n){
  x <- rnorm(10,7.4,0.025)
  pval[i] <- t.test(x,mu=7.4)$p.value
```

```
}
sum(pval < 0.05)/n

for (i in 1:n){
  x <- rnorm(10,7.375,0.025)
  pval[i] <- t.test(x,mu=7.4)$p.value
}
sum(pval < 0.05)/n
```

10. Formaldehyde. Antwoord

```
> library(datasets)
> plot(optden ~ carb, data = Formaldehyde,
+   xlab = "Carbohydrate (ml)", ylab = "Optical Density",
+    main = "Formaldehyde data", col = 4, las = 1)
> abline(mod.lin <- lm(optden ~ carb, data = Formaldehyde))
> summary(mod.lin)

Call:
lm(formula = optden ~ carb, data = Formaldehyde)

Residuals:
        1         2         3         4         5         6
-0.006714  0.001029  0.002771  0.007143  0.007514 -0.011743

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.005086   0.007834   0.649    0.552
carb        0.876286   0.013535  64.744 3.41e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.008649 on 4 degrees of freedom
Multiple R-squared: 0.999,      Adjusted R-squared: 0.9988
F-statistic:  4192 on 1 and 4 DF,  p-value: 3.409e-07
```

De rechte lijn past zeer goed bij de data.

11. UN.

```
library(alr3)
data(UN3)
with(UN3, plot(Purban, PPgdp))
with(UN3, plot(Purban, log(PPgdp) ))
with(UN3, plot(Fertility, Change))
mod.lin <- lm(Change ~ Fertility, data = UN3)
abline(mod.lin)
summary(mod.lin)
```