

Opgaven werkcollege Statistisch Toetsen

9 April 2019

dr. Wim P. Krijnen
Rijksuniversiteit Groningen
Faculteit Science and Engineering
Statistiek en Stochastiek

29 maart 2019

Inhoudsopgave

1	Inleiding	1
2	Starten met R	1
2.1	R downloaden	1
2.2	Library installeren	2
3	Vinden van hulp	2
4	Enkele handige functies	3
5	Data inladen	3
6	Statistische Toetsen	3
7	Opgaven	5
8	Antwoorden	7

1 Inleiding

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.

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 {[(is handig!

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 eenvoudig is, 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

Hulp binnen het R-systeem:

- `help.start()` opent HTML pagina om te browsen naar verdere info.
- `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.

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

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
- `plot()` maakt afbeelding
- `boxplot()` maakt box-and-wiskers-plot

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

6 Statistische Toetsen

- `t.test` toetst gelijkheid van gemiddelden $H_0 : \mu_1 = \mu_2$ tegen verschil $H_0 : \mu_1 \neq \mu_2$.
- `wilcox.test` toetst bij één steekproef de null-hypothese dat de data symmetrisch rond μ_0 zijn verdeeld.
- `SIGN.test` in library `BSDA` toetst verschil in twee random variabelen $H_0 : p = 0.50$ tegen $H_0 : p \neq 0.50$, waarbij $p = Pr(X > Y)$.
- `prop.test` toetst gelijkheid van proporties $H_0 : p_1 = p_2$ tegen verschil $H_0 : p_1 \neq p_2$.

Genoemde toetsen kunnen zowel op één als op twee steekproeven worden uitgevoerd. We nemen overall significantie niveau $\alpha = 0.05$, 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 vindt dat het productie proces 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-proces 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*. □

Opmerkingen:

- Het wordt in het algemeen aanbevolen om twee-zijdig te toetsen om de mogelijkheid van een verrassende uitkomst 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`).

7 Opgaven

1. Lichaamslengte van dochter en moeders.

Karl Pearson onderzocht een steekproef van meer dan 1100 families in Engeland in de periode 1893 tot 1898. De dataset geeft de lengte in inches van moeders en hun dochters, met maximaal twee dochters per moeder. Alle dochters zijn minstens 18 jaar oud en alle moeders zijn jonger dan 65.

K. Pearson and A. Lee (1903), On the laws of inheritance in man, *Biometrika*, 2, 357463, Table 31.

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

(a) Toets de nul-hypothese $H_0 : \mu_D = \mu_M$.

(b) Welke conclusie kun je trekken.

2. Proces controle. De onzuiverheid in in ppm van een tussen product in een chemisch proces wordt gemeten. Dit levert de volgende waarden in ppm,

```
y <- c(2.7, 2.5, 1.7, 1.6, 1.9, 2.6, 1.3, 1.9, 2.0, 2.5, 2.6, 2.3, 2.0,  
      1.8, 1.3, 1.7, 2.0, 1.9, 2.3, 1.9, 2.4, 1.5)
```

Het is van belang dat het gemiddelde lager blijft dan 2.5 ppm. Dergelijke process control komt vaak voor binnen chemische bedrijven¹.

(a) Toets de nul-hypothese $H_0 : \mu = 2.5$ ppm. Deel de data even door 10.

(b) Welke conclusie kun je trekken.

3. Halfgeleiders. Bij de vervaardiging van halfgeleiders wordt vloeibaar chemisch etsen vaak gebruikt om silicium te verwijderen van de ruggen van wafers voorafgaand aan metallisatie (aanbrengen dun metalen laagje). Twee verschillende vloeistoffen worden vergeleken op de ets snelheid (mm per minuut)².

¹Montgomery, D.C. & Runger, G.C. (2014). *Applied Statistics and Probability for Engineers*. 6th Edition.

²Montgomery, D.C. & Runger, G.C. (2014). *Applied Statistics and Probability for Engineers*. 6th Edition.

```
Vloeistof1 <- c(.251, .269, .239, .262, .236, .254, .244, .262, .259, .257)
Vloeistof2 <- c(.257, .254, .269, .259, .272, .272, .265, .265, .267, .262)
```

- (a) Toets de nul-hypothese $H_0 : \mu_1 = \mu_2$ tegen $H_A : \mu_1 \neq \mu_2$ (mm/min.).
- (b) Welke conclusie kun je trekken.

8 Antwoorden

1. Lichaamslengte. We toetsen de nul-hypothese $H_0 : \mu_D = \mu_M$ met de t-toets waarbij we opmerken dat de waarnemingen gepaard zijn.

(a)

```
> 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
```

De p -waarde $< 2.2 \cdot 10^{-16} < \alpha = 0.05$ geeft aanleiding de nul-hypothese te verwerpen.

- (b) De conclusie is dat de dochters gemiddelde genomen 1.3 inch langer zijn dan hun moeders.

2. Proces controle.

(a) We gebruiken de t-toets voor de nul-hypothese $H_0 : \mu = 2.5$ ppm als volgt

```
> y <- c(2.7, 2.5, 1.7, 1.6, 1.9, 2.6, 1.3, 1.9, 2.0, 2.5, 2.6, 2.3, 2.0,
+       1.8, 1.3, 1.7, 2.0, 1.9, 2.3, 1.9, 2.4, 1.5)
> t.test(y, mu=2.5)
```

One Sample t-test

```
data: y
t = -5.3893, df = 21, p-value = 2.405e-05
alternative hypothesis: true mean is not equal to 2.5
95 percent confidence interval:
 1.832258 2.204106
```

```
sample estimates:
mean of x
 2.018182
```

De p -waarde $2.405 \cdot 10^{-5} < \alpha = 0.05$ geeft aanleiding de nul hypothese te verwerpen.

- (b) De conclusie is dat de het gemiddelde ppm lager is dan 2.5.

3. Halfgeleiders.

(a) The metingen zijn onafhankelijk, nemen aan dat ze normaal verdeeld zijn en passen de t-toets voor twee groepen toe.

```
> v1 <- c(.251, .269, .239, .262, .236, .254, .244, .262, .259, .257)
> v2 <- c(.257, .254, .269, .259, .272, .272, .265, .265, .267, .262)
> t.test(v1,v2)
```

Welch Two Sample t-test

```
data: v1 and v2
t = -2.7819, df = 14.336, p-value = 0.01443
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.019285236 -0.002514764
sample estimates:
mean of x mean of y
 0.2533    0.2642
```

De p -waarde 0.01443 geeft aanleiding de nul hypothese te verwerpen.

(b) De conclusie is dat vloeistof 2 een grotere snelheid heeft.