R - data wrangling

J.J. van Nijnatten

Contents

Generate data

```
nrofsubs  = 4
nrofconds = 3
subj = rep(1:nrofsubs,nrofconds)
time = rep(LETTERS[1:nrofconds],each=nrofsubs)
score = as.vector( replicate(
          nrofconds , rnorm(n = nrofsubs, mean = sample(8,1)+10 , sd = sample(5,1) )
        ) )
data.long = data.frame(subj, time, score);
rm(list=setdiff(ls(), c("data.long", "nrofsubs","nrofconds")))
```

```
##      subj     time     score
##         1        A      8.83
##         2        A      8.65
##         3        A     15.39
##         4        A     15.12
##         1        B     20.03
##         2        B     20.77
##         3        B     11.30
##         4        B     15.71
##         1        C      5.31
##         2        C     14.14
##         3        C      9.46
##         4        C     10.15
```

```
##      subj         A        B         C
##         1      8.83     20.0      5.31
##         2      8.65     20.8     14.14
##         3     15.39     11.3      9.46
##         4     15.12     15.7     10.15
```

Data inspecteren

```
##      subj     time     score
##         1        A      8.83
##         2        A      8.65
##         3        A        NA
##         4        A     15.12
##         1     <NA>        NA
##         2        B     20.77
##         3        B     11.30
##         4        B     15.71
##         1        C     35.90
##         2        C     14.14
##         3        C      9.46
##         4        C     10.15
##         4        C     10.15
```

Foutieve data opsporen (missende data, extreme waardes, verkeerde data types)

Missende data

De functie *complete.cases()* geeft voor iedere rij uit de dataset aan of deze compleet is of niet (TRUE/FALSE).

```
(complRows = complete.cases(badData))
```

```
##  [1]  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [12]  TRUE  TRUE
```

In combinatie met de functie *which()* weet je welke rijnummers compleet zijn:

```
which(complete.cases(badData))
```

```
##  [1]  1  2  4  6  7  8  9 10 11 12 13
```

Andersom kun je met het symbool *!* de selectie omdraaien en krijg je terug welke rijen incompleet zijn:

```
(incomRows = which(!complete.cases(badData)))
```

```
## [1] 3 5
```

Een andere optie is om specifiek te zoeken naar missende waardes. In R worden die weergegeven als *NaN, NA* of *NULL*. Hiervoor kun je de functies *is.na(), is.nan(), is.null()* gebruiken. In dit geval zijn er missende waardes die als NA zijn weergegeven:

```
(naRows = which(is.na(badData$score)))
```

```
## [1] 3 5
```

Dubbele data

De functie *duplicated()* controleert of er data dubbel voorkomt in het dataframe:
(Deze functie vergelijkt hele rijen met elkaar, dus niet of een waarde meerdere keren voorkomt binnen een kolom.)

```
(dupRows = which(duplicated(badData)))
```

```
## [1] 13
```

Extreme of foutieve waardes

Soms komt het voor dat er foutieve waardes in je dataset terecht komen, of je wilt een subset van waardes selecteren om mee verder te werken. Stel dat in deze dataset alleen scores van 1 t/m 20 mogelijk zijn kunnen alsvolgt controleren of er waardes zijn die daar buiten vallen en welke dat zijn:

```
minScore = 1
maxScore = 20
(badData$score < minScore | badData$score > maxScore)
```

```
##  [1] FALSE FALSE    NA FALSE    NA  TRUE FALSE FALSE  TRUE FALSE FALSE
## [12] FALSE FALSE
```

```
(extremeRows = which((badData$score < minScore | badData$score > maxScore)))
```

```
## [1] 6 9
```

Data opschonen

Wanneer je foutieve waardes in je dataset hebt opgespoord kun je twee dingen doen: 1) De proefobjecten met die foutieve waardes verwijderen uit je dataset, of 2) alleen de correcte data opslaan in een nieuwe dataset.

Verwijderen van foutieve datapunten

Als je weet welke rijen in je dataset waardes bevatten die je wilt verwijderen kan dat alsvolgt:

```
naRows = which(is.na(badData$score)) # rij-nummers met incomplete data
badData = badData[-naRows,] # selecteer alles behalve incomplete rijen
```

Let op! wanneer je de rijnummers bewaart met slechte data, zorg dan dat je alles in een keer verwijdert. Stel dat je eerst de incomplete rijen verwijdert, en daarna de extreme waardes zullen na de eerste keer verwijderen de rijnummers opschuiven en gooi je mogelijk de vereerde data weg.

Dus niet zo

```
# BAD CODE !!
badData = badData[-dupRows]
badData = badData[-extremeRows,]
badData = badData[-incomRows,]
badData = badData[-naRows,]
# Bad CODE !!
```

```
##      subj     time     score
##         1        A      8.83
##         2        A      8.65
##         3        B     11.30
##         2        C     14.14
##         3        C      9.46
##         4        C     10.15
##         4        C     10.15
```

maar zo

```
# maak verzameling met unieke rij-nummers met foute data.
badRows = unique(c(incomRows, naRows, extremeRows, dupRows))
# verwijder alle rijen in een keer om te voorkomen dat rijen opschuiven
badData = badData[-badRows,]
```

```
##      subj     time     score
##         1        A      8.83
##         2        A      8.65
##         4        A     15.12
##         3        B     11.30
##         4        B     15.71
##         2        C     14.14
##         3        C      9.46
##         4        C     10.15
```

Let hierbij op het min-teken (-) wat aangeeft dat je alles behalve die data selecteert.

Selecteren van correcte data (indexing)

Correcte data opslaan in nieuwe dataset

Herstructureren van data

In de praktijk van het wetenschappelijk onderzoek worden experimenten vaak uitgevoerd met andere apparatuur en computers dan waar je de uiteindelijke statistische analyse gaat uitvoeren. De data zoals die uit het experiment komen rollen zijn meestal niet geschikt om direct een statistische toets op uit te voeren. Het kan zijn dat de data eerst moet worden opgeschoond. Dat wil zeggen, je moet controleren of de gemeten waardes wel binnen een plausibel bereik vallen, dus geen onmogelijke of onwaarschijnlijke waardes bevatten. Het kan ook gebeuren dat een proefdier of proefpersoon niet altijd de taak uitvoert zoals bedoeld en er een deel van de data mist. Dan is het belangrijk in beeld te brengen hoe vaak en wanneer dat gebeurt, en deze metingen weg te laten uit de analyse, of eventueel zelf alle metingen van betreffende proefdier / -persoon weg te laten.

Daarnaast werken sommige functies alleen correct wanneer de data van het juiste type zijn (bv. *numeric, character, factor*) en dat kun je niet altijd duidelijk zien op het eerste oog (b.v.: een "5" kan door R als *character* worden gezien, en niet als een nummer waar je mee kunt rekenen). Wat ook mis kan gaan is wanneer je condities of proefpersonen aanduid met nummers en R die getallen gebruikt om mee te rekenen i.p.v. deze te gebruiken om te weten bij welk proefobject / welke conditie een meetwaarde hoort. Het beste is om categorische data aan te duiden met letters, dus "S1","S2", .. i.p.v. "1","2",.. om proefobjecten aan te duiden, of "Conditie 1", "conditie 2" i.p.v. "1","2" (nog beter is om informatieve namen te gebruiken zoals: "Control", "LowDose","HighDose").

Datastructuren: wide vs. long format

Wanneer je alle data hebt opgeschoond en gecontroleerd kan het nog zijn dat de dataset niet de juiste *structuur* heeft. Met structuur bedoelen we hier hoe de onafhankelijk en afhankelijke variabelen en de subjectnummers etc. zijn ingedeeld. Er wordt een onderscheid gemaakt tussen het **LONG** en **WIDE** format.

WIDE FORMAT Wanneer een proefobject meermaals hebt gemeten en alle meetwaarden (afhankelijke variabelen) van de verschillende condities naast elkaar in aparte kolommen staan (dus 1 rij per proefobject) spreken we van een **WIDE** format zoals geillustreerd in Fig X.

LONG FORMAT Bij dit format staan alle meetwaarden van de verschillende condities in dezelfde kolom, met daarnaast een kolom die aanduid bij welke conditie iedere meetwaarde hoort, en een kolom die aanduid van welk proefobject de meetwaarde afkomstig is zoals geillustreerd in Fig X.

Van wide naar long

Stel je hebt 4 proefobjecten gemeten in 3 verschillende condities je data staan in het WIDE format:

```
##      subj         A        B         C
##         1      8.83     20.0      5.31
##         2      8.65     20.8     14.14
##         3     15.39     11.3      9.46
##         4     15.12     15.7     10.15
```

Om de data te herstructureren naar het LONG format kun je de *reshape()* functie alsvolgt gebruiken:

```
# transform data from wide to long format ----
data.long =
    reshape(data        = data.wide       # naam van de oude   dataset in het WIDE format
            ,direction  = 'long'          # richting van de data-transformatie
            ,varying    = c('A','B','C')  # kolomnamen die worden samengevoegd
            ,idvar      = 'subj'          # kolomnaam met de ppn-nummers
            ,v.names    = 'score'         # naam van nieuwe kolom met meetwaarden
            ,timevar    = 'time'          # naam van nieuwe kolom met condities
            ,times      = c('A','B','C')  # waardes
    )
```

```
##      subj     time     score
##         1        A      8.83
##         2        A      8.65
##         3        A     15.39
##         4        A     15.12
##         1        B     20.03
##         2        B     20.77
##         3        B     11.30
##         4        B     15.71
##         1        C      5.31
##         2        C     14.14
##         3        C      9.46
##         4        C     10.15
```

Van long naar wide

Stel je hebt 4 proefobjecten gemeten in 3 verschillende condities je data staan in het LONG format:

```
##      subj     time     score
##         1        A      8.83
##         2        A      8.65
##         3        A     15.39
##         4        A     15.12
##         1        B     20.03
##         2        B     20.77
##         3        B     11.30
##         4        B     15.71
##         1        C      5.31
##         2        C     14.14
##         3        C      9.46
##         4        C     10.15
```

Om de data te herstructureren naar het WIDE format kun je de *reshape()* functie alsvolgt gebruiken:

```
# transform data from long to wide format ----
data.wide =
reshape(data          = data.long      # naam van de dataset
          ,direction  = 'wide'         # richting van de data-transformatie
          ,v.names    = 'score'        # kolomnaam met de meetwaarden
          ,timevar    = 'time'         # kolomnaam met de conditienamen
          ,idvar      = 'subj'         # kolomnaam met de ppn-nummers
  )
names(data.wide)[2:4] = c('A','B','C') # pas de kolomnamen met scores aan
```

```
##      subj         A        B         C
##         1      8.83     20.0      5.31
##         2      8.65     20.8     14.14
##         3     15.39     11.3      9.46
##         4     15.12     15.7     10.15
```