Data-Frames, Faktoren, Deskriptive Statistik

Jonathan Harrington

0. Vorbereitung von Daten

Data-Frame
besteht aus Reihen und Spalten.

Reihen
sind Beobachtungen (jedes Mal, dass eine Versuchsperson etwas getan hat, gibt es eine Reihe).

Spalten
können aus **kontinuierlichen Variablen** und **Faktoren** (**kategorialen Daten**) bestehen.

Typische kontinuierliche Variablen in der Phonetik
z.B. Dauermessungen, Formantwerte, Reaktionszeiten usw.

Typische Faktoren (kategoriale Daten) in der Phonetik (und Linguistik)
- Sprecherkürzel (wer hat was gesprochen?)
- Sprechereigenschaften (Geschlecht, Dialekt, Alter...)
- Spracheigenschaften (Vokale, Wörter, Silben...)

Ein Faktor besteht aus Stufen
z.B. Geschlecht hat 2 Stufen (m, w); wenn drei Altersgruppen aufgenommen wurden, dann hat der Faktor Alter 3 Stufen (jung, mittel, alt) usw.

Beispiel 1
20 Versuchspersonen (10 männlich, 10 weiblich) produzierten /i, a, u/ Vokale jeweils 10 Mal. Die Dauerwerte dieser Vokale wurden erhoben.

Kontinuierliche Variablen
Dauer

Faktoren
- Versuchspersonen (20 Stufen)
- Geschlecht (2 Stufen: m, w)
- Vokal (3 Stufen, i, u, a)
- Wiederholung (10 Stufen: 1, 2, ...10)

Data-Frame Struktur
20 (Versuchspersonen) × 3 Vokale × 10 Wiederholungen = 600 Reihen
5 Spalten: Dauer, Versuchsperson, Geschlecht, Vokal, Wiederholung.

Eine 600 × 5 Matrix (3000 Werte)

Beispiel 2
F1 und F2 Werte wurden erhoben für lange und kurze Vokale, die sowohl in Funktionswörtern also auch in Inhaltswörtern vorkamen. Die Daten wurden von 40 Personen

produziert, 20 aus Bayern, 20 aus Hessen; 12 der Versuchspersonen waren männlich, die anderen 28 weiblich.

Kontinuierliche Variablen
F1, F2

Faktoren
Vokallänge (2 Stufen: lang, kurz)
Kategorie (2 Stufen: Funktion, Inhalt)
Versuchsperson (40 Stufen)
Dialekt (2 Stufen)
Geschlecht (2 Stufen)

Data-Frame Struktur
40 (Versuchspersonen) × 2 Vokallängen × 2 Kategorien = 160 Reihen
(also jede Versuchsperson produzierte 4 Werte: langer Vokal im Inhaltswort, kurzer Vokal im Inhaltswort, langer Vokal im Funktionswort, kurzer Vokal im Funktionswort)

7 Spalten: F1, F2, Vokallänge, Kategorie, Versuchsperson, Dialekt, Geschlecht

Beispiel 3
Die Reaktionszeiten wurden in 40 Versuchspersonen gemessen. Den Versuchspersonen wurden 50 Stimuli präsentiert, einmal orthographisch auf einem Bildschirm, einmal auditiv über Kopfhörer. Die Versuchspersonen mussten entscheiden, ob der präsentierte Stimulus ein reales Deutsches Wort war, oder nicht. Die Stimuli begannen mit einem /s/, /st/, oder /str/. 20 der Vpn. waren L1-Deutsch, 20 L1-Spanisch. Das Experiment wurde zwei Mal pro Versuchsperson durchgeführt: Einmal mit den auditiven Stimuli zuerst, einmal mit den orthographischen Stimuli zuerst.

1. Data-Frame in R erzeugen

```
# Dauermessungen
d = c(100, 56, 190, 80, 110, 45, 89, 90, 120)

# für diese Vokale
lab = c("i", "i", "e", "e", "i", "a", "a", "e", "a")

# für diese Versuchspersonen
vpn = c("AB", "AB", "EJ", "AB", "EJ", "MP", "DN", "WN", "DN")

# Geschlecht der Versuchspersonen
g = c("m", "m", "w", "m", "w", "m", "w", "m", "w")

mein.df = data.frame(dauer = d, Vok = factor(lab), Vpn = factor(vpn), G = factor(g))

class(mein.df)

dim(mein.df)

# um die ersten paar Zeilen zu sehen
head(mein.df)
```

```
# um die Spaltennamen zu sehen
names(mein.df)
```

Hier erzeugen wir einen Data-Frame aus den Dauerwerten, Konsonant /t, d, St/, Silbenanzahl, Wort der Sprachdatenbank `sabinevot`. Zur Erinnerung:

```
#
library(emu)
vot.s = emu.query("sabinevot", "*", "Phonetik = h")
# Die Wörter
wort.s = emu.requery(vot.s, "Phonetik", "Wort")
wort.l = label(wort.s)

# wir können die obigen zwei Befehle so kombinieren
wort.l = emu.requery(vot.s, "Phonetik", "Wort", j = T)

# Die Silbenzahl
zahl.l = emu.requery(vot.s, "Phonetik", "Zahl", j = T)
# Der Konsonant
kons.l = emu.requery(vot.s, "Phonetik", "Phonetik", seq = -1, j = T)

# Data-frame genannt vot.df bauen mit diesen Daten.
```

1. Data-Frame als Text-Datei speichern

```
write.table(mein.df, file.path(pfad, "mein.txt"))
# oder
write.table(mein.df, file.path(pfad, "mein.txt"), quote=F)
```

2. Data-Frame einlesen

```
m = read.table(file.path(pfad, "mein.txt"))
```

3. Daten von einem Data-Frame manipulieren: die with() Funktion

Alle Befehle in R können als zweites Argument zu der with() Funktion eingegeben werden und sie funktionieren auf die selbe Weise. Das erste Argument ist irgendein Data-Frame, zB

```
4+10
with(mein.df, 4+10)

c("ja", "nein")
with(mein.df, c("ja", "nein"))
```

Aber zusätzlich sind die Spalten eines Data-Frames sichtbar. z.B.

```
colnames(mein.df)
Vok
Error: object 'Vok' not found

with(mein.df, Vok)
```

```
dauer
Error: object 'dauer' not found

with(mein.df, dauer)
```

Wie bekomme ich den Dauer-Mittelwert?

Wie bekomme ich eine Tabellierung der Vokale?

4. Deskriptive Statistik I: boxplot()
Um kontinuierliche Daten darzustellen, und um festzustellen, wie eine kontinuierliche Variable von einem oder mehreren Faktoren beeinflusst wird.

```
vot = dur(vot.s)
mean(vot)
median(vot)
[1] 14.5495
```

Bedeutung von Median: man sortiert die Werte. Der Median ist so nah wie möglich am der mittleren Wert in den sortierten Daten

```
sort(vot)
[1]   9.601  10.383  11.082  11.171  11.304  11.373  11.590
[8]  11.803  11.882  12.530  12.960  13.301  13.692  13.840
[15]  14.127  14.972  15.364  15.633  17.940  20.422  40.215
[22]  43.040  44.413  54.944  59.827  76.957  79.971  80.486
[29]  89.477 108.142
```

Man sagt auch für Median: der 50% Quantal

```
quantile(vot, .5)
```

Im Boxplot wird der Median als eine horizontale dicke Linie dargestellt; der 'Box' dehnt sich aus zwischen den 25% und 75% Quantalen, genannt den **Interquartalen Bereich**

```
quantile(vot, .25)
quantile(vot, .75)
# oder
quantile(vot, c(.25, .75))

boxplot(vot)
```

Meistens will man die Verteilung in Abhängigkeit von einem Faktor sehen:
```
boxplot(vot ~ zahl.l)

boxplot(vot ~ kons.l)
```

Beide zusammen (Die Faktoren werden 'gekreuzt')

```
boxplot(vot ~ kons.l * zahl.l)
```

Meistens erstellen wir solche Daten aus einem Data-Frame. Dafür gibt es zwei Möglichkeiten...

Fragen

Entpacken Sie die zip-Datei unter 4.1 in der Webseite, so dass die darin enthaltenen .txt-Dateien sich in Ihrem pfad-Verzeichnis befinden (Öffnen Sie eine Text-Datei mit einem Editor, um zu bestätigen, dass sie sich in dem pfad-Verzeichnis befinden).

1. `lok.txt`

Lesen Sie in R mit read.table() die .txt-Datei `lok.txt` ein:

```
lok = read.table(file.path(pfad, "lok.txt"))
```

Schreiben Sie R-Befehle um diese Informationen zu ermitteln.

1.1 Die Anzahl der Beobachtungen
1.2 Die Anzahl der Spalten
1.3. Die Stufen des Faktors `P`
1.4 Den Mittelwert der Variable `slopes`
1.5. Einen Boxplot der Variable `slopes` als Funktion der Faktoren `Kons` und `P`
1.6 Das gleiche wie 1.5 aber nur für die Männer (für die `M`-Stufe des Faktors `G`; hier müssen Sie einen logischen Vektor verwenden).

2. `rtdaten.txt`

Daten aus Johnson (2008), modifiziert. Die Daten zeigen Reaktionszeitmessungen (RT) für 58 Versuchspersonen (Listener) aufgeteilt in 4 Sprechergruppen (Gruppe). Für die Reaktionszeitmessungen bekamen die Versuchsperson eine Reihenfolge von 2 spanischen Sprachlauten. Der erste Laut (Cons) war entweder "d" oder "r". Das Urteil (Pair) der Versuchspersonen war entweder "same" oder "different" . z.B. Cons = "d" und Pair = "different" heißt: die Reaktionszeit wurde gemessen, wenn der erste Laut ein /d/ war, und wenn im nächsten Laut keinen /d/ wahrgenommen wurde. Die vier Sprechergruppen (Gruppe) sind: begin (L2-spanisch Anfänger), intermed (L2-spanisch Fortgeschrittene), nospan (L2-spanisch mit keinen Kenntnissen der spanischen Sprache), spannat (L1-Spanisch).

Inwiefern werden die Reaktionszeiten (a) von der Hörergruppe, (b) dem ersten Konsonant, und (c) Urteil beeinflusst? Verwenden Sie Boxplots um (a), (b), und (c) zu beantworten.