Übung 6

Explorative Datenanalyse und Visualisierung

Wintersemester 2019

S. Döhler, S. Döhler (FBMN, h_da)

Name: Valentina Cisternas Seeger & Roman Kessler

Punkte:

Aufgabe 16. *In dieser Aufgabe soll untersucht werden, ob Text-Messaging die Rechtschreibung verschlechtert. Dazu wurde folgendes Experiment durchgeführt: Ein Gruppe von 25 Schülern wurde sechs Monate lang ermuntert über ihre Smartphones Textnachrichten zu versenden. Einer zweite Gruppe von 25 Schülern wurde hingegen sechs Monate lang verboten über ihre Smartphones Textnachrichten zu versenden. Am Anfang und Ende der sechs Monate wurde die Rechtschreibung der Schüler durch einen Test gemessen (Details – auch zur Durchführung des Verbots! – finden sich in "Discovering Statistics Using R" von Andy Field). Der Datensatz* `TextMessages.dat`*, den Sie in Moodle finden enthält folgende Variablen:*

- `Group`*: Beschreibt, ob die Person zur ersten oder zweiten Gruppe gehörte.*
- `Baseline`*: Ergebnis des Rechtschreibungstests (in % Richtige) zu Beginn der sechs Monate*
- `SixMonths`*: Ergebnis des Rechtschreibungstests (in % Richtige) am Ende der sechs Monate*

Sie sollen die Ergebnisse des Experiments mit den Methoden aus der LV explorativ analysieren. Einige Hinweise:

- *Starten Sie zunächst mit den Rohdaten. Nähern Sie sich dann der Fragestellung indem Sie neue Variablen einführen, mit denen Sie dann weiterarbeiten.*
- *Begründen Sie, welche Methoden zu welchen Variablentypen passen könnten.*
- *Achten Sie auf Titel, Legende, Achsenbeschriftung Ihrer plots.*

Lösung

Zunächst importieren wir den Datensatz.

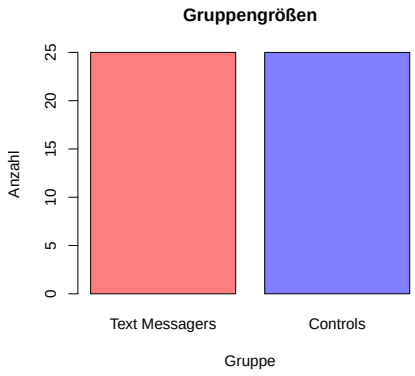
```
# import the data (delete the white spaces before importing.
# note: because the import failed, we modified the text data
# by deleting white spaces between "Text" and "Messagers")
```

```
df <- read.table("TextMessages.dat",
                 sep = "\t",
                 col.names = c("group", "pre", "post"),
                 colClasses = c("factor", "integer", "integer"),
                 skip = 1)
```

Wir überprüfen kurz, dass unsere Stichproben auch gleich groß sind.

```
barplot(height = c(length(df$group[df$group=="TextMessagers"] == TRUE),
                   length(df$group[df$group=="Controls"] == TRUE)),
        main = "Gruppengrößen",
        ylab = "Anzahl",
        xlab = "Gruppe",
        names = c("Text Messagers","Controls"),
        col = c(rgb(1,0,0,0.5),rgb(0,0,1,0.5))

        )
```



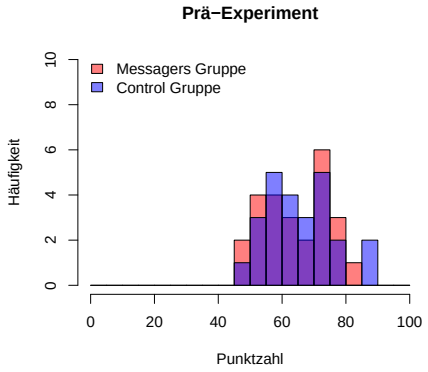
Dies ist die einzige Kategorische Variable in unserem Datensatz. Im folgenden benutzen wir nur Plots, die für metrisch skalierte Daten verwendung finden (z.B. Boxplots, QQ-Plots, etc).

Nun schauen wir uns die Verteilung der anderen Variablen in Histogrammen an.

Zunächst, wie sehen die Punktzahlen der beiden Gruppen vor dem Experiment aus?

```
hist(df$pre[df$group == "TextMessagers"],
     col=rgb(1,0,0,0.5),
     xlim=c(0,100),
     ylim=c(0,10),
     main="Prä-Experiment",
     xlab="Punktzahl",
     ylab = "Häufigkeit",
```

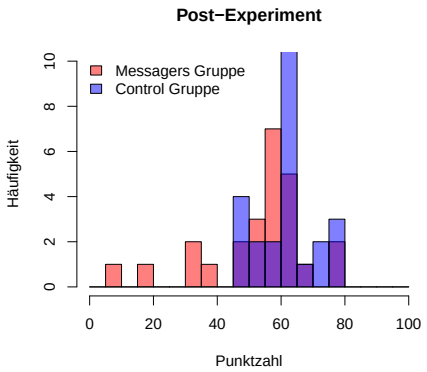
```
     breaks = seq(0,100,5))
hist(df$pre[df$group == "Controls"],
     col=rgb(0,0,1,0.5),
     breaks = seq(0,100,5),
     add=T)
legend("topleft",
       legend=c("Messagers Gruppe", "Control Gruppe"),
       fill = c(rgb(1,0,0,0.5),rgb(0,0,1,0.5)),
       cex=1, bty = "n")
```



Vor dem Experiment sehen die Verteilungen der Punktzahlen der beiden Gruppen erstmal sehr ähnlich aus.

Anschließend, wie sehen die Punktzahlen der beiden Gruppen nach dem Experiment aus?

```
hist(df$post[df$group == "TextMessagers"],
     col=rgb(1,0,0,0.5),
     xlim=c(0,100),
     ylim=c(0,10),
     main="Post-Experiment",
     xlab="Punktzahl",
     ylab = "Häufigkeit",
     breaks = seq(0,100,5))
hist(df$post[df$group == "Controls"],
     col=rgb(0,0,1,0.5),
     breaks = seq(0,100,5),
     add=T)
legend("topleft",
       legend=c("Messagers Gruppe", "Control Gruppe"),
       fill = c(rgb(1,0,0,0.5),rgb(0,0,1,0.5)),
       cex=1,
       bty = "n")
```

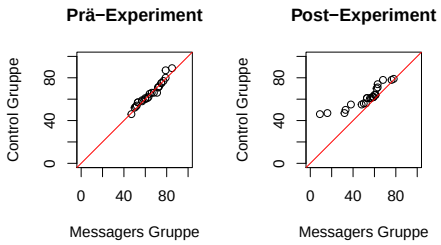


Wir sehen hier, dass die “Text Messagers” Gruppe nach dem Experiment einige Werte im unteren Bereich hat, also niedrige Punktzahlen. Es scheint, als hätten sich einzelne Probanden der “Text Messagers” Gruppe durch das Experiment verschlechtert.

Nun vergleichen wir mithilfe eines QQ-Plots noch einmal die Verteilung der beiden Gruppen.

```
{
par(mfrow=c(1,2))
par(pty="s")
qqplot(df$pre[df$group == "TextMessagers"],
       df$pre[df$group == "Controls"],
       ylim = c(0,100),
       xlim = c(0,100),asp=1,
       xlab = "Messagers Gruppe",
       ylab = "Control Gruppe",
       main = "Prä-Experiment")
abline(0, 1, col = 'red')

par(pty="s")
qqplot(df$post[df$group == "TextMessagers"],
       df$post[df$group == "Controls"],
       ylim = c(0,100), xlim = c(0,100),asp=1,
       xlab = "Messagers Gruppe",
       ylab = "Control Gruppe",
       main = "Post-Experiment")
abline(0, 1, col = 'red')
}
```



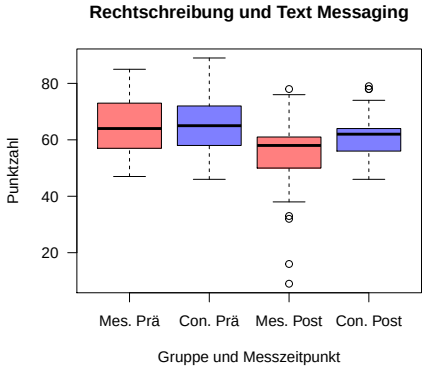
Wir sehen dass die beiden Gruppen vor dem Experiment ähnlich verteilt sind (Punkte liegen an der Winkelhalbierenden des QQ-Plots). Nach dem Experiment sehen wir klare Abweichungen der Punkte von der Winkelhalbierenden. Wenn wir die Richtung der Abweichung anschauen, bestätigt dies nochmal eine mögliche Verschlechterung in der Messagers Gruppe.

Wir formulieren uns hier schwammig (z.B. “mögliche Verschlechterung”), da wir ohne eine Statistik erstmal keine endgültige Aussage treffen wollen.

Nun Vergleichen wir mit Boxplots die Gruppen prä und post gegeneinander:

```
#boxplot(df$pre[df$group == "TextMessagers"], col="darkred")

boxplot(
        df$pre[df$group == "TextMessagers"],
        df$pre[df$group == "Controls"],
        df$post[df$group == "TextMessagers"],
        df$post[df$group == "Controls"],
        main = "Rechtschreibung und Text Messaging",
        ylab = "Punktzahl",
        xlab = "Gruppe und Messzeitpunkt",
        at = c(1,2,3,4),
        names = c("Mes. Prä","Con. Prä",
                  "Mes. Post",  "Con. Post"),
        las = 1,
        col = c(rgb(1,0,0,0.5),rgb(0,0,1,0.5)),
        border = "black",
        notch = FALSE
)
```



Durch die Boxpolts bestätigt sich nochmal, dass sich die Gruppen vor dem Experiment nicht grob zu unterscheiden scheinen.

Nach dem Experiment sehen wir eine leichte verschlechterung in der Messagers Gruppe. Es gibt einige sehr niedrige Werte (siehe auch Histogram). Weiterhin sehen wir, dass der Median in der Gruppe im Vergleich zur Kontrollgruppe ein kleines wenig niedriger ist. Die Streuung der Daten ist auf jeden Fall größer, als in der Kontrollgruppe.

Nun berechnen wir die Differenzen der einzelnen Probanden der beiden Gruppen. Dadurch können wir hoffentlich besser sehen, wie die individuelle Entwicklung der Probanden der beiden Gruppen durch das Experiment aussieht.

Wir definieren die Differenz als Post minus Prä - Experiment, somit bedeutet eine positive Differenz eine Verbesserung, und eine negative Differenz eine Verschlechterung.

Wir plotten die Differenzen erstmal als stem-and-leave Plot, um uns die Verteilung anzuschauen.

Für die Kontrollgruppe:

```
x = seq(1,100)
df$imp <- df$post - df$pre
stem(df$imp[df$group == "Controls"])
#>
#>   The decimal point is 1 digit(s) to the right of the |
#>
#>   -1 | 63311000
#>   -0 | 976555444431
#>    0 | 03
#>    1 | 4
#>    2 | 00
```

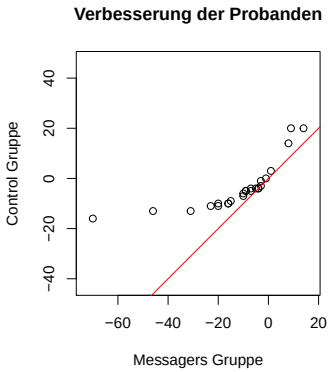
Und für die Experimentalgruppe:

```r
stem(df$imp[df$group == "TextMessagers"])
#>
#>   The decimal point is 1 digit(s) to the right of the |
#>
#>   -6 | 0
#>   -4 | 6
#>   -2 | 1300
#>   -0 | 665009977544331
#>    0 | 1894
```

Wir sehen, dass es in beiden Gruppen hauptsächlich negative Werte auftreten, somit eine Verschlechterung. In der Text Messagers Gruppe sind jedoch mehr Werte im Negativen als in der Control Gruppe.

Schauen wir uns die beiden Verteilungen noch einmal mittels eines QQ-Plots an:

```r
par(pty="s")
Xmin = -3 + min(df$imp[df$group == "TextMessagers"])
Xmax = +3 + max(df$imp[df$group == "TextMessagers"])
Ymin = -3 + min(df$imp[df$group == "Controls"])
Ymax = +3 + max(df$imp[df$group == "Controls"])
qqplot(df$imp[df$group == "TextMessagers"],
       df$imp[df$group == "Controls"],
       ylim = c(Ymin,Ymax),
       xlim = c(Xmin,Xmax),asp=1,
       xlab = "Messagers Gruppe",
       ylab = "Control Gruppe",
       main = "Verbesserung der Probanden")
abline(0, 1, col = 'red')
```



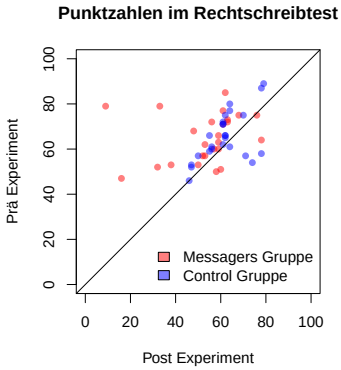
Dieser Plot bestätigt noch einmal alle schon oben getätigten aussagen.

In einem finalen Schritt (das könnten wir aber schon viel früher machen), schauen wir uns mal die einzelnen Datenpunkte in einem Streudiagramm an.

Wir tragen sowohl die Punktzahl eines jeden Probanden vor und nach dem Experiment auf, für alle Probanden der beiden Gruppen (in unterschiedlichen Farben). Wir erwarten zunächst mal einen groben Linearen Zusammenhang zwischen Vorher und Nacher, da wir davon ausgehen würden, dass die Probanden, die vorher *sehr gut* waren, später nicht *sehr schlecht* sein werden, und wenn doch, dann eher als Ausnahme.

Wir können jedoch noch etwas viel interessanteres sehen: Die Probanden, die auf der einen Seite der Winkelhalbierenden liegen, haben sich verschlechtert, und die auf der anderen Seite, haben sich verbessert. Der Abstand zu der Winkelhalbierenden impliziert gleichzeitig das Ausmaß der Verbesserung/Verschlechterung.

```
{
par(pty="s")
plot(x = df$post, y = df$pre,
     xlim = c(0,100), ylim = c(0,100),
     col = ifelse(df$group=="Controls",
                  rgb(0,0,1,0.5),rgb(1,0,0,0.5)),
     ylab = "Prä Experiment",
     xlab = "Post Experiment",
     main = "Punktzahlen im Rechtschreibtest",
     pch=16)
abline(0, 1, col = 'black')
legend("bottomright",
       legend=c("Messagers Gruppe", "Control Gruppe"),
       fill = c(rgb(1,0,0,0.5),rgb(0,0,1,0.5)),
       cex=1, bty = "n")
}
```



Dieser Plot gibt uns jetzt eigentlich keine neue Information, ist aber eine schöne Zusammenfassung der einzelnen Datenpunkte. Wir sehen auch hieran, dass es eine Tendenz zur Verschlechterung in der “Messagers” Gruppe vorhanden ist, da viele Datenpunkte nach oben links abweichen.

Anmerkungen/Korrektur

"Wir formulieren uns hier schwammig (z.B."mögliche Verschlechterung"), da wir ohne eine Statistik erstmal keine endgültige Aussage treffen wollen."
Was für eine Statistik bräuchtet ihr denn um eine Aussage zu treffen?
Boxplot und Säulendiagramm: Wenn ihr Farben im Plot benutzt, fügt bitte eine Legende hinzu!
Ansonsten eine gute Abgabe mit interessanten Plots und Analysen!
✓