Vertiefende statistische Verfahren

3. Übungsblatt SS 2024

Allgemeine Information

Alle Aufgaben sind mit R zu lössen, wenn nicht explizit anders angegeben. Die Berechnungen sollen nachvollziehbar und dokumentiert sein. Um die vollständige Punktezahl zu erreichen, müssen alle Ergebnisse und Fragen entsprechend interpretiert bzw. beantwortet werden. Code alleine ist nicht ausreichend! Die Abgabe erfolgt über Moodle entsprechend der Abgaberichtlinien als pdf und Rmd File. Bitte inkludieren Sie namentlich alle beteiligten Gruppenmitglieder sowohl im Bericht als auch im Source Code. Die jeweiligen Datensätze die für diese Übung relevant sind finden Sie ebenfalls in Moodle.

1 Einfaktorielle ANOVA (händisch) [2P]

Ein Sportwissenschafter möchte sehen, ob es einen Unterschied in der Gewichtszunahme von Sportlern gibt, die einer von drei speziellen Diäten folgen. Die Athleten werden nach dem Zufallsprinzip drei Gruppen zugewiesen und unterziehen sich für 6 Wochen der jeweiligen Diät. Die Gewichtszunahmen (in Pfund) sind angegeben. Nehmen Sie an, dass die Gewichtszunahmen normalverteilt sind und die Varianzen gleich sind. Kann der Sportwissenschafter bei einem Signifikanzniveau von 0,05 schlussfolgern, dass es einen Unterschied zwischen den Diäten gibt? Führen Sie die ANOVA händisch durch und überprüfen Sie das Ergebnis mit `summary(aov(...))`.

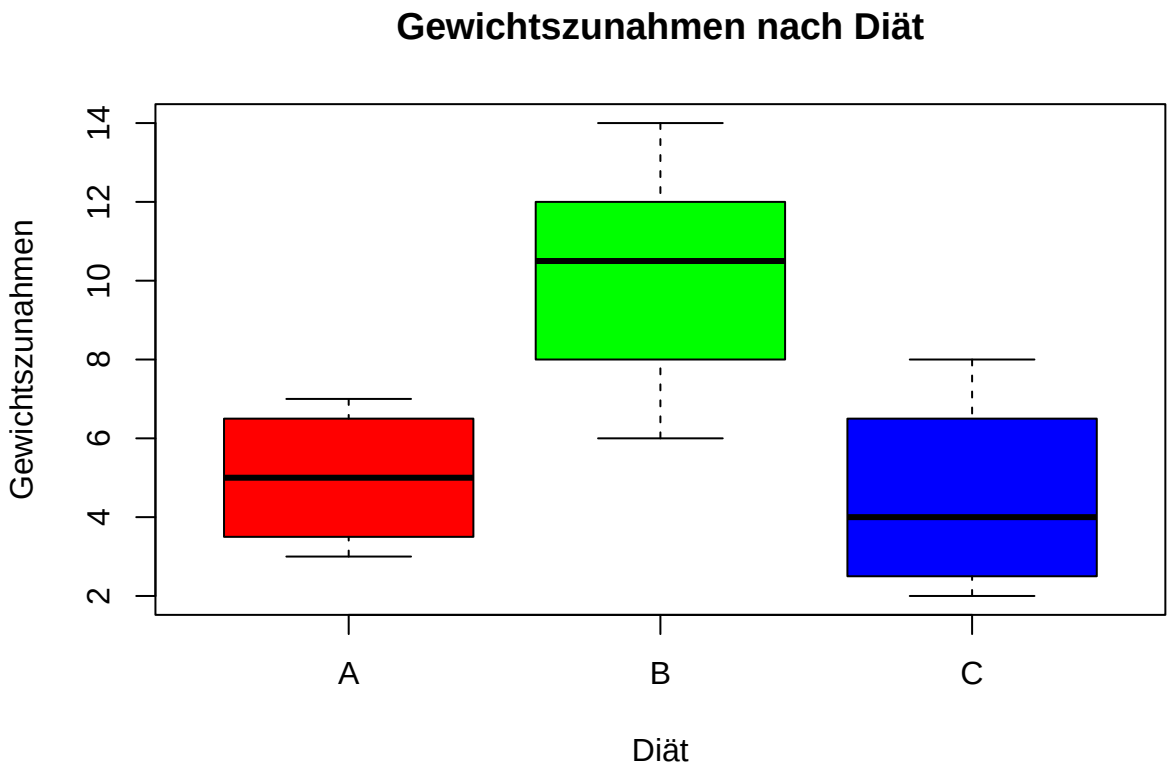
Diät	Messwerte
A	3, 6, 7, 4
B	10, 12, 11, 14, 8, 6
C	8, 3, 2, 5

```r
# ANOVA händische durchführung
gains <- list(
  A = c(3, 6, 7, 4),
  B = c(10, 12, 11, 14, 8, 6),
  C = c(8, 3, 2, 5))

# Überblick über die Daten / Farben nach Diet
str(gains)
```

```
## List of 3
##  $ A: num [1:4] 3 6 7 4
##  $ B: num [1:6] 10 12 11 14 8 6
##  $ C: num [1:4] 8 3 2 5
```

```r
boxplot(gains, col = c("red", "green", "blue"),
        ylab = "Gewichtszunahmen",
        xlab = "Diät",
        main = "Gewichtszunahmen nach Diät")
```




```r
# Hypothesen
# H0: mu_a = mu_b = mu_c
# H1: mindestens ein mu_i ist ungleich

# Berechnung der Mittelwerte
mean_total = mean(unlist(gains))

# Berechnung der Quadratsummen
qs_between2 <- 0
for (group in gains) {
  group_mean <- mean(group)
  n <- length(group)
  qs_between2 <- qs_between2 + n * (group_mean - mean_total)^2
}

qs_within2 <- 0
for (group in gains) {
  group_mean <- mean(group)
  for (value in group) {
    qs_within2 <- qs_within2 + (value - group_mean)^2
```

```
  }
}

qs_total <- sum((unlist(gains) - mean_total)^2)

# Überprüfen der Vorraussetzungen

# Normalverteilung
for (group in gains) {
  print(shapiro.test(group))
}
```

```
##
##  Shapiro-Wilk normality test
##
## data:  group
## W = 0.94971, p-value = 0.7143
##
##
##  Shapiro-Wilk normality test
##
## data:  group
## W = 0.98901, p-value = 0.9866
##
##
##  Shapiro-Wilk normality test
##
## data:  group
## W = 0.94563, p-value = 0.6889
```

```
# Varianzhomogenität (Barlet-Test)
bartlett.test(gains)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  gains
## Bartlett's K-squared = 0.61192, df = 2, p-value = 0.7364
```

```
# Freiheitsgrade
df_between <- length(gains) - 1
df_within <- length(unlist(gains)) - length(gains)
df_total <- length(unlist(gains)) - 1

# Berechnung der Varianzen (MS)
ms_total <- qs_total / df_total
ms_between <- qs_between2 / df_between
ms_within <- qs_within2 / df_within

# Berechnung des F-Wertes
f_value <- ms_between / ms_within
```

```r
# kritischer F-Wert aus Tabelle
alpha <- 0.05
f_critical <- qf(1 - alpha, df_between, df_within)

# Ergebnis
if (f_value > f_critical) {
  print("H0 wird verworfen")
} else {
  print("H0 wird nicht verworfen")
}
```

```
## [1] "H0 wird verworfen"
```

```r
# Überprüfung mit R
diet <- factor(rep(names(gains), times = sapply(gains, length)))
all_gains <- unlist(gains)

aov_result <- aov(all_gains ~ diet)
summary(aov_result)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## diet         2 101.10   50.55    7.74 0.00797 **
## Residuals   11  71.83    6.53
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation: Es gibt einen signifikanten Unterschied zwischen den Diäten ($p < 0.05$). Der F-Wert beträgt 7.74 und der kritische F-Wert 3.98. Die Wahl der Diät hat somit einen signifikanten Einfluss auf die Gewichtszunahme. Der Unterschied in den Diäten ist mit einem Signifikanzniveau von 0.05 signifikant.

2 Einfaktorielle ANOVA [2P]

Analysieren Sie die Daten zu niedrigem Geburtsgewicht von Neugeborenen in `birthwt_aov.xlsx`. Untersuchen Sie ob die ethnische Zugehörigkeit der Mutter (Variable `ethnic`, wobei `1 = "white"`, `2 = "black"`, `3 = "other"`) einen Einfluss auf das Geburtsgewicht von Neugeborenen hat (Variable `bwt`). Beachten Sie dabei folgende Punkte:

i) Überprüfen Sie die Vorraussetzungen.
ii) Verwenden Sie eine geeignete grafische Darstellung der Daten / Ergebnisse.
iii) Gibt es signifikante Unterschiede zwischen den Gruppen, wenn ja zwischen welchen Gruppen?
iv) Achten Sie auf eine "statistisch korrekte" Formulierung des Ergebnisses.

```r
# Daten einlesen
library(readxl)
birthwt <- read_excel("UE3 Daten/birthwt_aov.xlsx")

# Überblick über die Daten
str(birthwt)
```

```
## tibble [189 x 10] (S3: tbl_df/tbl/data.frame)
##  $ low   : num [1:189] 0 0 0 0 0 0 0 0 0 0 ...
##  $ age   : num [1:189] 19 33 20 21 18 21 22 17 29 26 ...
##  $ lwt   : num [1:189] 182 155 105 108 107 124 118 103 123 113 ...
##  $ ethnic: num [1:189] 2 3 1 1 1 3 1 3 1 1 ...
##  $ smoke : num [1:189] 0 0 1 1 1 0 0 0 1 1 ...
##  $ ptl   : num [1:189] 0 0 0 0 0 0 0 0 0 0 ...
##  $ ht    : num [1:189] 0 0 0 0 0 0 0 0 0 0 ...
##  $ ui    : num [1:189] 1 0 0 1 1 0 0 0 0 0 ...
##  $ ftv   : num [1:189] 0 3 1 2 0 0 1 1 1 0 ...
##  $ bwt   : num [1:189] 2523 2551 2557 2594 2600 ...
```

```
summary(birthwt)
```

```
##       low              age              lwt             ethnic      
##  Min.   :0.0000   Min.   :14.00   Min.   : 80.0   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:19.00   1st Qu.:110.0   1st Qu.:1.000  
##  Median :0.0000   Median :23.00   Median :121.0   Median :1.000  
##  Mean   :0.3122   Mean   :23.24   Mean   :129.8   Mean   :1.847  
##  3rd Qu.:1.0000   3rd Qu.:26.00   3rd Qu.:140.0   3rd Qu.:3.000  
##  Max.   :1.0000   Max.   :45.00   Max.   :250.0   Max.   :3.000  
##      smoke            ptl               ht                ui        
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.0000  
##  Mean   :0.3915   Mean   :0.1958   Mean   :0.06349   Mean   :0.1481  
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000  
##  Max.   :1.0000   Max.   :3.0000   Max.   :1.00000   Max.   :1.0000  
##       ftv              bwt        
##  Min.   :0.0000   Min.   : 709  
##  1st Qu.:0.0000   1st Qu.:2414  
##  Median :0.0000   Median :2977  
##  Mean   :0.7937   Mean   :2945  
##  3rd Qu.:1.0000   3rd Qu.:3487  
##  Max.   :6.0000   Max.   :4990
```

3 Einfaktorielle ANOVA [2P]

Verwenden Sie den bereits bekannten `Framingham.sav` Datensatz. Analysieren Sie ob es Unterschiede im BMI in Abhängigkeit von der Schulbildung gibt. Achten Sie auf Ausreißer und fehlende Daten (`NaN, NA's`).

i) Überprüfen Sie die Vorraussetzungen.
ii) Verwenden Sie eine geeignete grafische Darstellung der Daten / Ergebnisse.
iii) Gibt es signifikante Unterschiede zwischen den Bildungsstufen, wenn ja zwischen welchen Stufen?
iv) Achten Sie auf eine "statistisch korrekte" Formulierung des Ergebnisses.

4 Mehrfaktorielle ANOVA [2P]

Sie führen eine Studie bezüglich des Einflusses unterschiedlicher Diäten und Aktivitätslevels auf den Erfolg bei der Gewichtsabnahme durch. Jedem Probanden wird eine Diät und ein Aktivitätslevel zugewiesen

und die Differenz zum Ausgangsgewicht nach 2 Monaten gemessen (in kg). Die Daten der Studie sind in `weightloss.sav` zusammengefasst. Analysieren Sie ob die beiden Faktoren einen Einfluss auf das Gewicht haben.

i) Überprüfen Sie die Vorraussetzungen.
ii) Verwenden Sie eine geeignete grafische Darstellung der Daten / Ergebnisse.
iii) Gibt es signifikante Haupteffekte sowie Interaktionseffekte (inkl. Interaktions-Plot)?
iv) Achten Sie auf eine "statistisch korrekte" Formulierung des Ergebnisses.

5 Mehrfaktorielle ANOVA [2P]

Verwenden Sie den erneut den `Framingham.sav` Datensatz. Analysieren Sie den systolischen Blutdruck `sysbp` abhängig von Geschlecht und Bildungsstufe. Achten Sie auf Ausreißer und fehlende Daten (`NaN, NA's`).

i) Überprüfen Sie die Vorraussetzungen.
ii) Verwenden Sie eine geeignete grafische Darstellung der Daten / Ergebnisse.
iii) Gibt es signifikante Haupteffekte sowie Interaktionseffekte (inkl. Interaktions-Plot)?
iv) Achten Sie auf eine "statistisch korrekte" Formulierung des Ergebnisses.