Musterloesung Aufgabe 2.3S: ANOVA mit Interaktion

Lese-Empfehlung Kapitel 7 von Manny Gimond

Download R-Skript

Download PDF

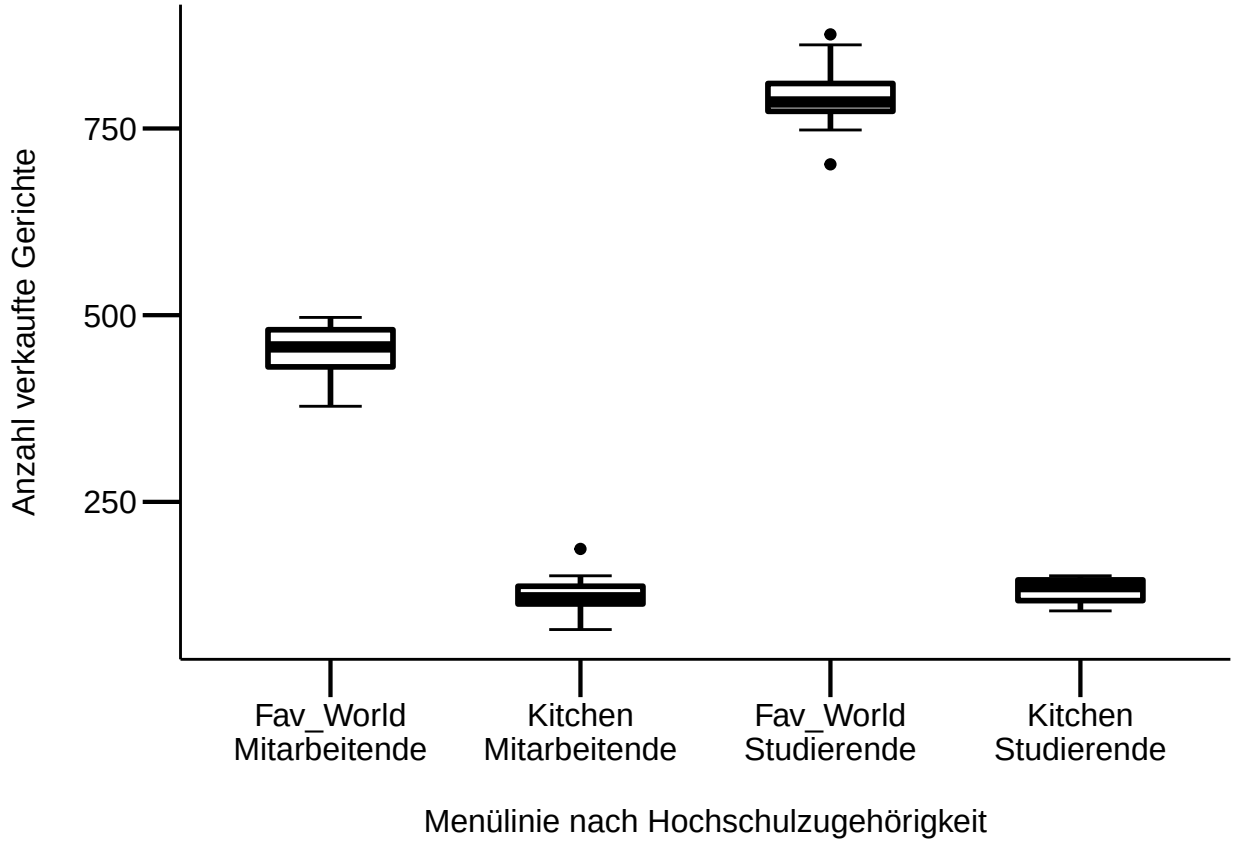
---

kommentierter Lösungsweg

```
# klone den originaler Datensatz
df <- nova

# Daten vorbereiten
df %<>% # schaut euch das Package "magrittr" an
  # ersetze Local mit einem leeren String
  mutate(article_description = str_replace(article_description, "Local ", "")) %>%
  filter(article_description != "Hot and Cold") %>% # lasse Buffet Gerichte weg
  filter(member != "Spezialkarten") %>% # Spezialkarten können vernachlässigt werden
  #  fasse die zwei Menülinien "World & Favorite" zusammen
  mutate(article_description = str_replace_all(article_description, "Favorite|World",
                                               "Fav_World"))

# gruppiere Daten nach Menülinie, Geschlecht und Hochschulzugehörigkeit
df %<>%
    group_by(article_description, member, week) %>%
    summarise(tot_sold = n()) %>%
    ungroup() %>%
    drop_na()  # lasst die unbekannten Menü-Inhalte weg

# überprüft die Voraussetzungen für eine ANOVA
# Schaut euch die Verteilungen der Mittelwerte der Responsevariable an
# Sind Mittelwerte nahe bei Null? Gäbe uns einen weiteren Hinweis auf
# eine spezielle Binomail-Verteilung (vgl. Statistik 4)
df %>%
  split(.$article_description) %>% # teilt den Datensatz in 3 verschiedene Datensätze auf
  # mit map können andere Funktionen auf den Datensatz angewendet werden
  # (alternative Funktionen sind aggregate oder apply)
  purrr::map(~ psych::describe(.$tot_sold))
```

```
## $Fav_World
##    vars  n   mean     sd median trimmed    mad min max range skew kurtosis   se
## X1    1 24 622.67 178.79  599.5   620.8 253.52 378 876   498 0.04    -1.88 36.5
##
## $Kitchen
##    vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
## X1    1 24 128.5 22.21  124.5   128.2 23.72  79 187   108 0.27     0.43 4.53
```

```r
# visualisiere dir dein Model, was siehst du?
# sind möglicherweise gewisse Voraussetzungen verletzt?
# Boxplot
ggplot(df, aes(x = interaction(article_description, member), y= tot_sold)) +
   # Achtung: Reihenfolge spielt hier eine Rolle!
  stat_boxplot(geom = "errorbar", width = 0.25) +
  geom_boxplot(fill="white", color = "black", size = 1, width = .5) +
  labs(x = "\nMenülinie nach Hochschulzugehörigkeit", y = "Anzahl verkaufte Gerichte\n") +
  scale_x_discrete(limits = c("Fav_World.Mitarbeitende", "Kitchen.Mitarbeitende",
                              "Fav_World.Studierende", "Kitchen.Studierende"),
                   breaks = c("Fav_World.Mitarbeitende", "Fav_World.Studierende",
                              "Kitchen.Mitarbeitende",  "Kitchen.Studierende"),
                   labels = c("Fav_World\nMitarbeitende", "Fav_World\nStudierende",
                              "Kitchen\nMitarbeitende",  "Kitchen\nStudierende")) +
  mytheme # wie sind die Voraussetzungen erfüllt?
```

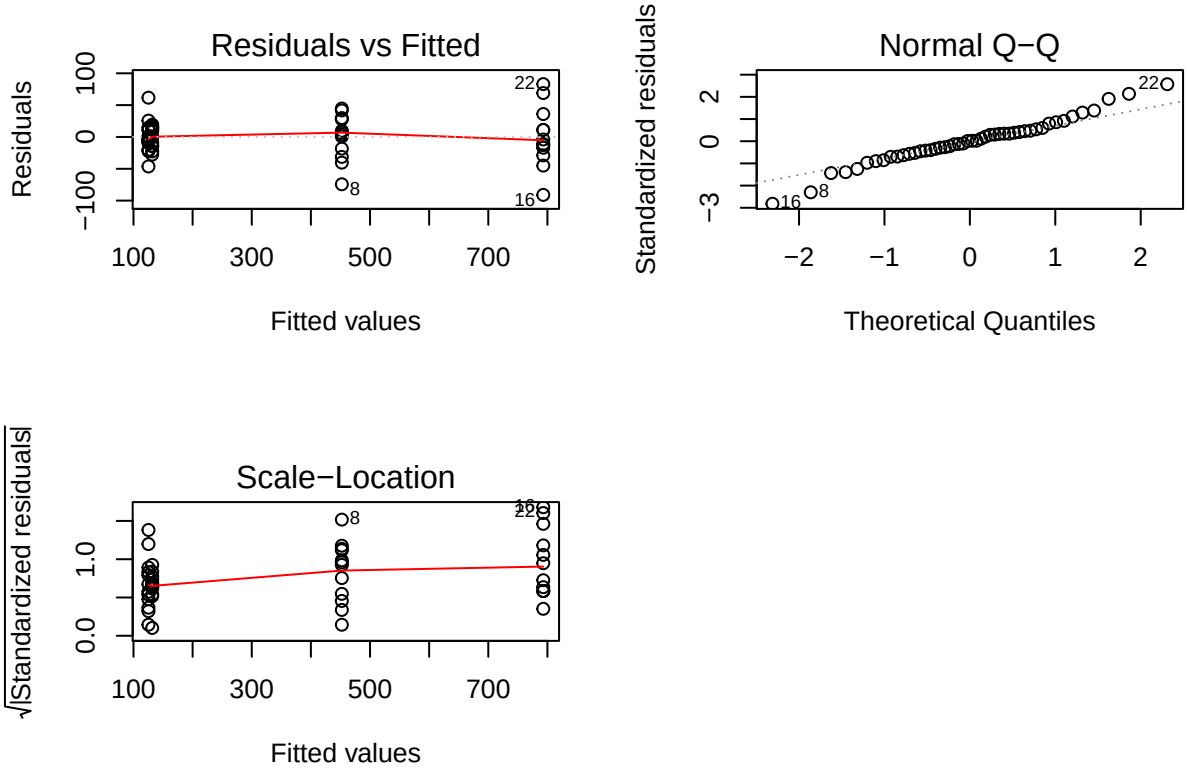



```r
# definiert das Modell (Skript Statistik 2)
model <- aov(tot_sold ~ article_description * member, data = df)

summary.lm(model)
```

```
##
## Call:
## aov(formula = tot_sold ~ article_description * member, data = df)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.00 -17.33   0.50  14.83  83.00
##
## Coefficients:
##                                             Estimate Std. Error t value
## (Intercept)                                  452.333      9.734   46.47
## article_descriptionKitchen                  -327.000     13.766  -23.75
## memberStudierende                            340.667     13.766   24.75
## article_descriptionKitchen:memberStudierende -334.333     19.469  -17.17
##                                             Pr(>|t|)
## (Intercept)                                   <2e-16 ***
## article_descriptionKitchen                    <2e-16 ***
## memberStudierende                             <2e-16 ***
## article_descriptionKitchen:memberStudierende  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.72 on 44 degrees of freedom
## Multiple R-squared:  0.9864, Adjusted R-squared:  0.9855
## F-statistic:  1063 on 3 and 44 DF,  p-value: < 2.2e-16
```

```r
# überprüft die Modelvoraussetzungen (Statistik 2)
par(mfrow = c(2,2)) # alternativ gäbe es die ggfortify::autoplot(model) funktion
plot(model)
```



Fazit: Die Inspektion des Modells zeigt kleinere Verletzungen bei der Normalverteilung der Residuen (Q-Q Plot). Aufgrund keiner starken Verbesserung durch eine Transformation der Responsevariable, entscheide ich mich für eine ANOVA ohne log-tranformierten Responsevariablen (AV).

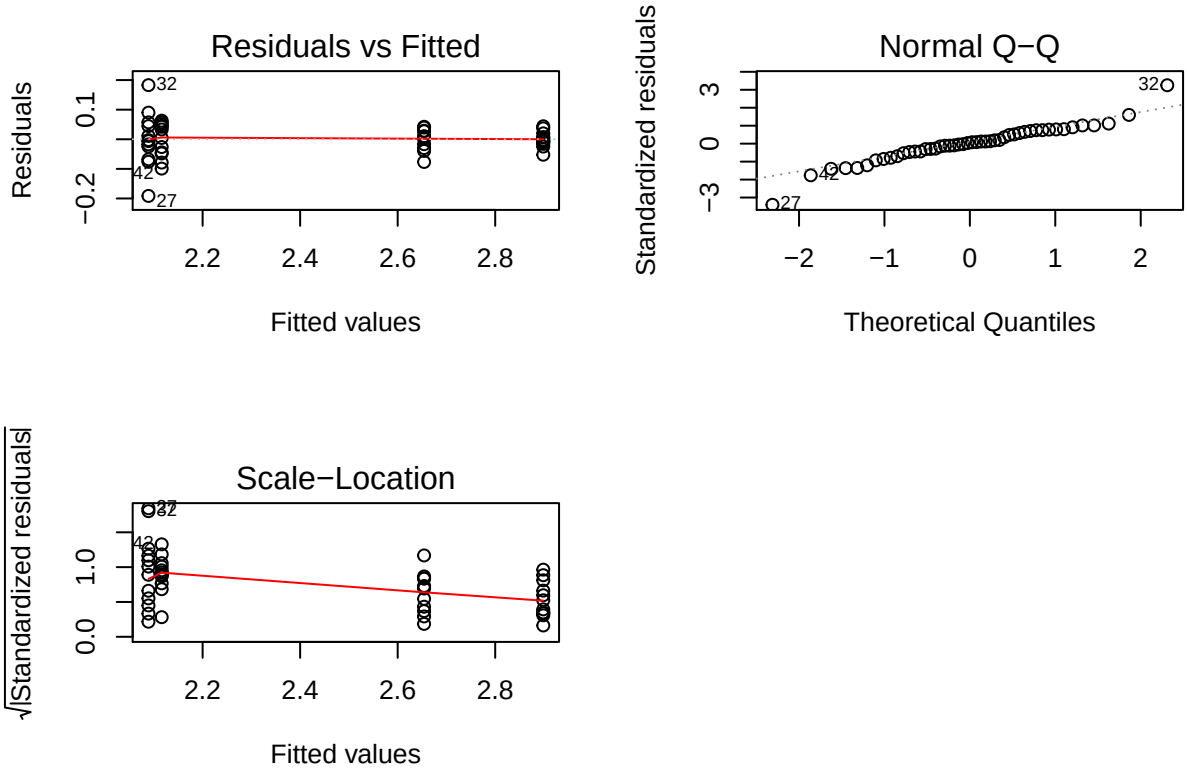
```
# sieht aus, als ob die Voraussetzungen für eine Anova nur geringfügig verletzt sind
# mögliche alternativen:
# 1. log-transformation um die grossen werte zu minimieren (nur möglich, wenn
# keine 0 enthalten sind und die Mittelwerte weit von 0 entfernt sind
# => bei Zähldaten ist dies leider nicht immer gegeben)
# 2. nicht parametrische Test z.B. Welch-Test, da hohe Varianzheterogenität
# zwischen den Residuen


#1) log-transformation
model_log <- aov(log10(tot_sold) ~ article_description * member, data = df)

summary.lm(model_log) # interaktion ist nun nicht mehr signifikant: vgl.
```

```
##
## Call:
## aov(formula = log10(tot_sold) ~ article_description * member,
##     data = df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.191372 -0.025043  0.003191  0.037604  0.182842
##
## Coefficients:
##                                            Estimate Std. Error t value
## (Intercept)                                 2.65417    0.01696 156.533
## article_descriptionKitchen                 -0.56517    0.02398 -23.569
## memberStudierende                           0.24438    0.02398  10.191
## article_descriptionKitchen:memberStudierende -0.21726    0.03391  -6.407
##                                            Pr(>|t|)
## (Intercept)                                 < 2e-16 ***
## article_descriptionKitchen                  < 2e-16 ***
## memberStudierende                          3.71e-13 ***
## article_descriptionKitchen:memberStudierende 8.51e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05874 on 44 degrees of freedom
## Multiple R-squared:  0.9745, Adjusted R-squared:  0.9728
## F-statistic: 561.4 on 3 and 44 DF,  p-value: < 2.2e-16
```

```
# nochmals euren Boxplot zu beginn, machen diese Koeffizienten sinn?

# überprüft die Modelvoraussetzungen (vgl. Skript Statistik 2)
# bringt aber keine wesentliche Verbesserung, daher bleibe ich bei den
# untranfromierten Daten
par(mfrow = c(2,2))
plot(model_log)
```

```
## hat values (leverages) are all = 0.08333333
```

```
##  and there are no factor predictors; no plot no. 5
```




```
# post-hov Vergleiche
TukeyHSD(model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = tot_sold ~ article_description * member, data = df)
##
## $article_description
##                         diff      lwr       upr p adj
## Kitchen-Fav_World -494.1667 -513.785 -474.5484     0
##
## $member
##                              diff      lwr      upr p adj
## Studierende-Mitarbeitende 173.5 153.8817 193.1183     0
##
## $`article_description:member`
##                                                          diff        lwr        upr
## Kitchen:Mitarbeitende-Fav_World:Mitarbeitende -327.000000 -363.75650 -290.24350
## Fav_World:Studierende-Fav_World:Mitarbeitende  340.666667  303.91017  377.42317
## Kitchen:Studierende-Fav_World:Mitarbeitende   -320.666667 -357.42317 -283.91017
## Fav_World:Studierende-Kitchen:Mitarbeitende    667.666667  630.91017  704.42317
## Kitchen:Studierende-Kitchen:Mitarbeitende        6.333333  -30.42317   43.08983
```

```
## Kitchen:Studierende-Fav_World:Studierende     -661.333333 -698.08983 -624.57683
##                                                    p adj
## Kitchen:Mitarbeitende-Fav_World:Mitarbeitende 0.0000000
## Fav_World:Studierende-Fav_World:Mitarbeitende 0.0000000
## Kitchen:Studierende-Fav_World:Mitarbeitende   0.0000000
## Fav_World:Studierende-Kitchen:Mitarbeitende   0.0000000
## Kitchen:Studierende-Kitchen:Mitarbeitende     0.9672944
## Kitchen:Studierende-Fav_World:Studierende     0.0000000
```

Methode

Ziel war es die Unterschiede zwischen den preisgünstigeren und teureren Menülinien und der Hochschulzugehörigkeit herauszufinden: Hierfür wurde eine ANOVA mit Interaktion gerechnet, da wir eine (quasi)-metrische Responsevariable und zwei Prädiktorvariablen (Menülinie und Hochschulzugehörigkeit) haben. Die Voraussetzungen für eine ANOVA waren im ersten Model nicht stark verletzt, lediglich die Normalverteilung der Residuen: Deshalb habe wurde auf eine log-Transformation der Responsevariable verzichtet. Anschliessend wurden noch post-hoc Einzelvergleiche nach Tukey durchgeführt.

Ergebnisse

Die wöchentlichen Verkaufszahlen der Menülinien unterscheiden sich nach Hochschulzugehörigkeit signifikant ($F(3,44) = 561.42$, $p < .001$). Inhaltich bedeutet dies, dass Studierende signifikant häufiger die preisgünstigere Menülinie "Favorite & World" als Mitarbeitende kaufen. Entgegen der Annahme gibt es aber keine signifikanten Unterschiede zwischen Studierende und Mitarbeitende bei dem Kauf der teureren Menülinie "Kitchen". Über die möglichen Gründe können nur spekuliert werden, hierfür bedarf es weiteren Analysen z.B. mit dem Prädiktor "Menüinhalt".

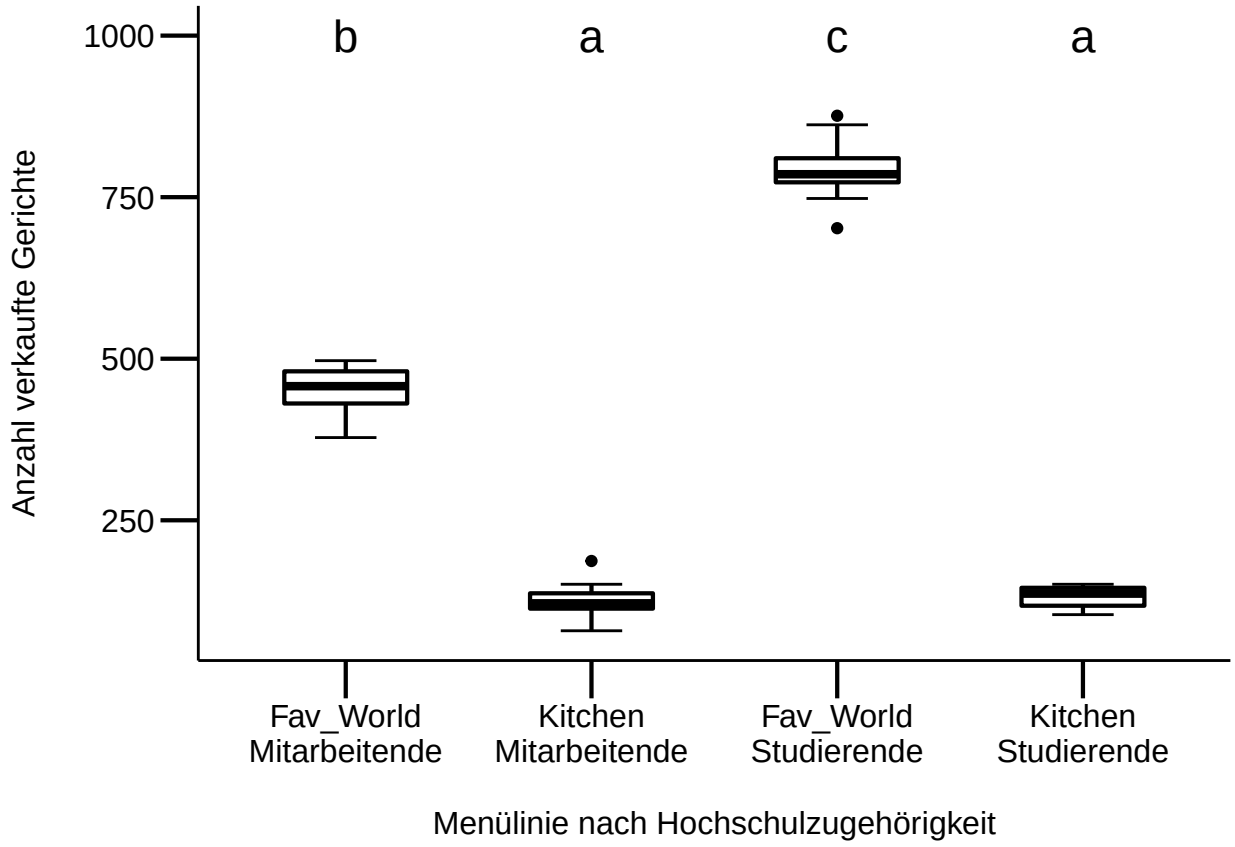


Figure 1: Box-Whisker-Plots der wöchentlichen Verkaufszahlen pro Menü-Inhalte. Kleinbuchstaben bezeichnen homogene Gruppen auf $p < .05$ nach Tukeys post-hoc-Test.