Musterloesung Aufgabe 2.2: einfaktorielle ANOVA

Download R-Skript

Download PDF

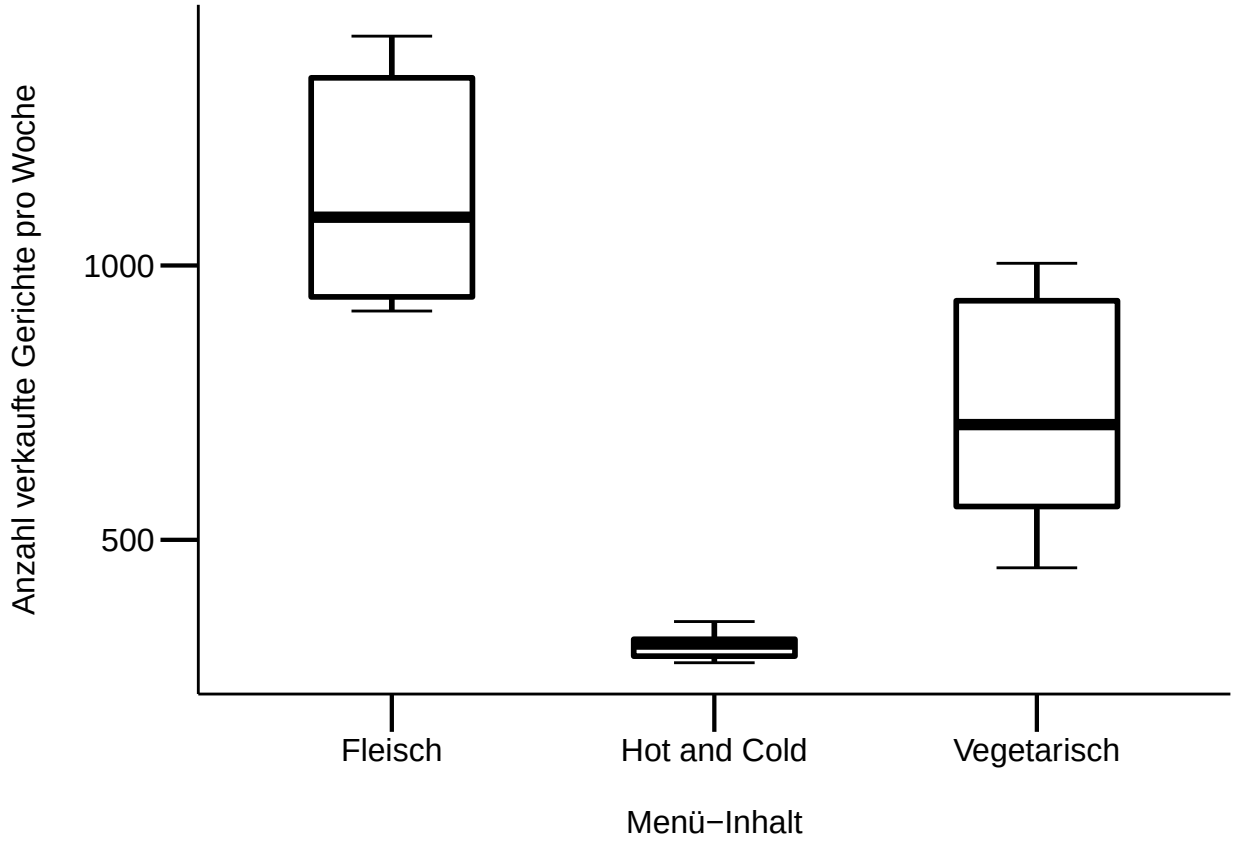
---

kommentierter Lösungsweg

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

# fasst die vier Inhalte der Gerichte zu drei Inhalten zusammen.
df %<>%
  # Geflügel & Fisch zu fleischgerichte zählen
  mutate(label_content = str_replace(label_content, "Geflügel|Fisch", "Fleisch")) %>%
  # achtung reihenfolge spielt eine rolle, wegen des + (plus)
  mutate(label_content = str_replace(label_content, "Pflanzlich[+]|Pflanzlich", "Vegetarisch"))

# gruppiert Daten nach Menü-Inhalt und Woche
df %<>%
    group_by(label_content, week) %>%
    summarise(tot_sold = n()) %>%
    drop_na() # lasst die unbekannten Menü-Inhalte weg

# überprüft die Voraussetzungen für eine ANOVA
# Schaut euch die Verteilungen der Mittelwerte an (plus Standardabweichungen)
# Sind Mittelwerte nahe bei Null?
# Gäbe uns einen weiteren Hinweis auf eine spezielle Binomail-Verteilung
df %>%
  split(.$label_content) %>% # teilt den Datensatz in 3 verschiedene Datensätze auf
  purrr::map(~ psych::describe(.$tot_sold)) # mit map können andere Funktionen
```

```
## $Fleisch
##    vars  n    mean     sd median trimmed    mad min  max range skew kurtosis
## X1    1 12 1135.58 200.03   1088  1129.2 223.13 917 1418   501 0.19    -1.89
##       se
## X1 57.74
##
## $`Hot and Cold`
##    vars  n   mean    sd median trimmed   mad min max range skew kurtosis   se
## X1    1 12 308.33 23.53    310   307.3 30.39 276 351    75 0.32    -1.25 6.79
##
## $Vegetarisch
##    vars  n   mean     sd median trimmed    mad min  max range  skew kurtosis
## X1    1 12 739.25 213.54    710   741.8 323.95 449 1004   555 -0.01    -1.85
##       se
## X1 61.64
```

```r
# auf den Datensatz angewendet werden (alternative Funktionen sind aggregate oder apply)


# Boxplot
ggplot(df, aes(x = label_content, 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 = "\nMenü-Inhalt", y = "Anzahl verkaufte Gerichte pro Woche\n") +
  # achtung erster Hinweis einer Varianzheterogenität, wegen den Hot&Cold Gerichten
  mytheme
```

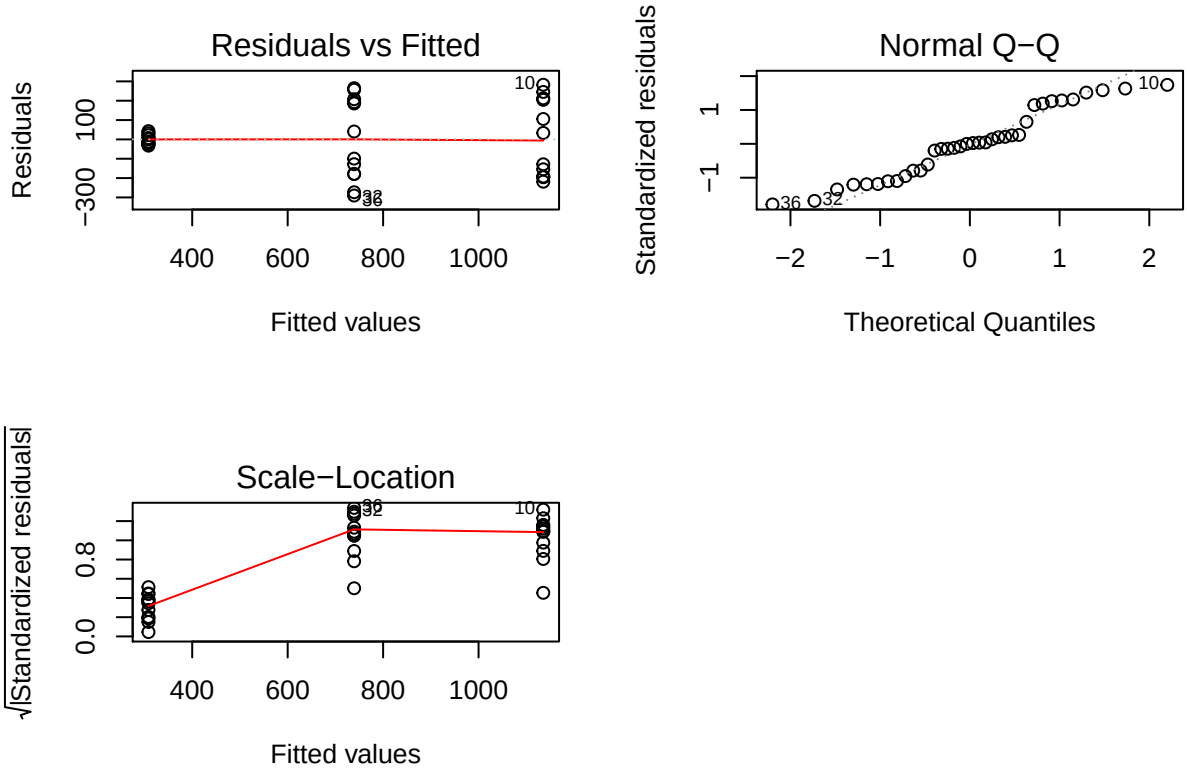



```r
# definiert das Modell (vgl. Skript Statistik 2)
model <- aov(tot_sold ~ label_content, data = df)

summary.lm(model)
```

```
##
## Call:
## aov(formula = tot_sold ~ label_content, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -290.250 -135.083    1.667  125.500  282.417
##
```

```
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1135.58      48.92  23.211  < 2e-16 ***
## label_contentHot and Cold    -827.25      69.19 -11.956 1.54e-13 ***
## label_contentVegetarisch     -396.33      69.19  -5.728 2.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 169.5 on 33 degrees of freedom
## Multiple R-squared:  0.8125, Adjusted R-squared:  0.8012
## F-statistic: 71.52 on 2 and 33 DF,  p-value: 1.007e-12
```

```
# überprüft die Modelvoraussetzungen
par(mfrow = c(2,2))
plot(model)
```

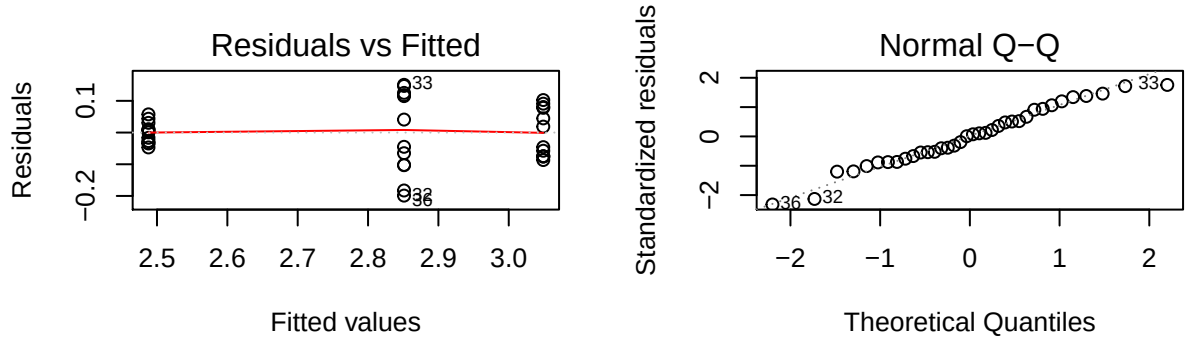


Fazit: Inspektion der Modellvoraussetzung zeigt klare Verletzungen des Residuelplots (zeigt einen "Trichter", siehe Skript Statistik 2), somit Voraussetzung der Homoskedastizität verletzt. Mögliche nächste Schritte:

- Menüinhalt "Buffet" aus der Analyse ausschliessen, da sowieso kein richtiger Menüinhalt (aber Informationsverlust)
- Datentransformation z.B. log-Transformation
- nicht-parametrischer Test (Achtung, auch dieser setzt Voraussetzungen voraus)
- ein glm Model (general linear model) mit einer poisson/quasipoisson link Funktion (vgl. Skript Statistik 4), weitere Infos dazu Link

```r
# überprüft die Voraussetzungen des Welch-Tests:
# Gibt es eine hohe Varianzheterogenität und ist die relative Verteilung der
# Residuen gegeben? (siehe Statistik 2)
# Ja Varianzheterogenität ist gegeben, aber die Verteilung der Residuen folgt
# einem "Trichter", also keiner "normalen/symmetrischen" Verteilung um 0
# Daher ziehe ich eine Transformation der AV einem nicht-parametrischen Test vor
# für weitere Infos:
# https://data.library.virginia.edu/interpreting-log-transformations-in-a-linear-model/

# achtung hier log10, bei Rücktransformation achten
model_log <- aov(log10(tot_sold) ~ label_content, data = df)

par(mfrow = c(2,2))
plot(model_log) # scheint ok zu sein
```

```
## hat values (leverages) are all = 0.08333333
##  and there are no factor predictors; no plot no. 5
```




```r
summary.lm(model_log) # Referenzkategorie ist der Buffet-Inhalt
```

```
##
## Call:
## aov(formula = log10(tot_sold) ~ label_content, data = df)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.198920 -0.059343  0.003477  0.062579  0.150567
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.04908    0.02585 117.942  < 2e-16 ***
## label_contentHot and Cold -0.56121    0.03656 -15.350  < 2e-16 ***
## label_contentVegetarisch  -0.19792    0.03656  -5.413 5.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08956 on 33 degrees of freedom
## Multiple R-squared:  0.8802, Adjusted R-squared:  0.8729
## F-statistic: 121.2 on 2 and 33 DF,  p-value: 6.238e-16
```

```
TukeyHSD(model_log) # (Statistik 2)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = log10(tot_sold) ~ label_content, data = df)
##
## $label_content
##                                diff        lwr        upr   p adj
## Hot and Cold-Fleisch     -0.5612085 -0.6509215 -0.4714955 0.0e+00
## Vegetarisch-Fleisch      -0.1979175 -0.2876305 -0.1082044 1.6e-05
## Vegetarisch-Hot and Cold  0.3632910  0.2735780  0.4530041 0.0e+00
```

```
# Achtung Beta-Werte resp. Koeffinzienten sind nicht direkt interpretierbar
# sie müssten zuerst wieder zurück transformiert werden, hier ein Beispiel dafür:
# für Buffet
10^model_log$coefficients[1]
```

```
## (Intercept)
##    1119.655
```

```
# für Fleisch
10^(model_log$coefficients[1] + model_log$coefficients[2])
```

```
## (Intercept)
##    307.5216
```

```
# für Vegi
10^(model_log$coefficients[1] + model_log$coefficients[3])
```

```
## (Intercept)
##    709.8501
```

Methoden

Ziel war es, die Unterschiede in den wöchentlichen Verkaufszahlen pro Menüinhalt aufzuzeigen. Da die Responsevariable (Verkaufszahlen) metrisch und die Prädiktorvariable kategorial sind, wurde eine einfaktorielle ANOVA gerechnet. Die visuelle Inspektion des Modells zeigte insbesondere schwere Verletzungen der Homoskedastizität. Der Boxplot bestätigt dieser Befund. Weil die Voraussetzungen schwer verletzt sind, wurde eine log-Transformation der Responsevariable vorgenommen. Anschliessend wurde erneut eine ANOVA gerechnet und die Modelvoraussetzungen visuell inspiziert: Homoskedastizität und Normalverteilung der Residuen sind gegeben.

Ergebnisse

Die Menüinhalte (Fleisch, Vegetarisch und Buffet) unterscheiden sich in den wöchentlichen Verkaufszahlen signifikant ($F(2,15) = 121.22$, $p < .001$). Die Abbildung 1 zeigt die wöchentlichen Verkaufszahlen pro Menüinhalt.

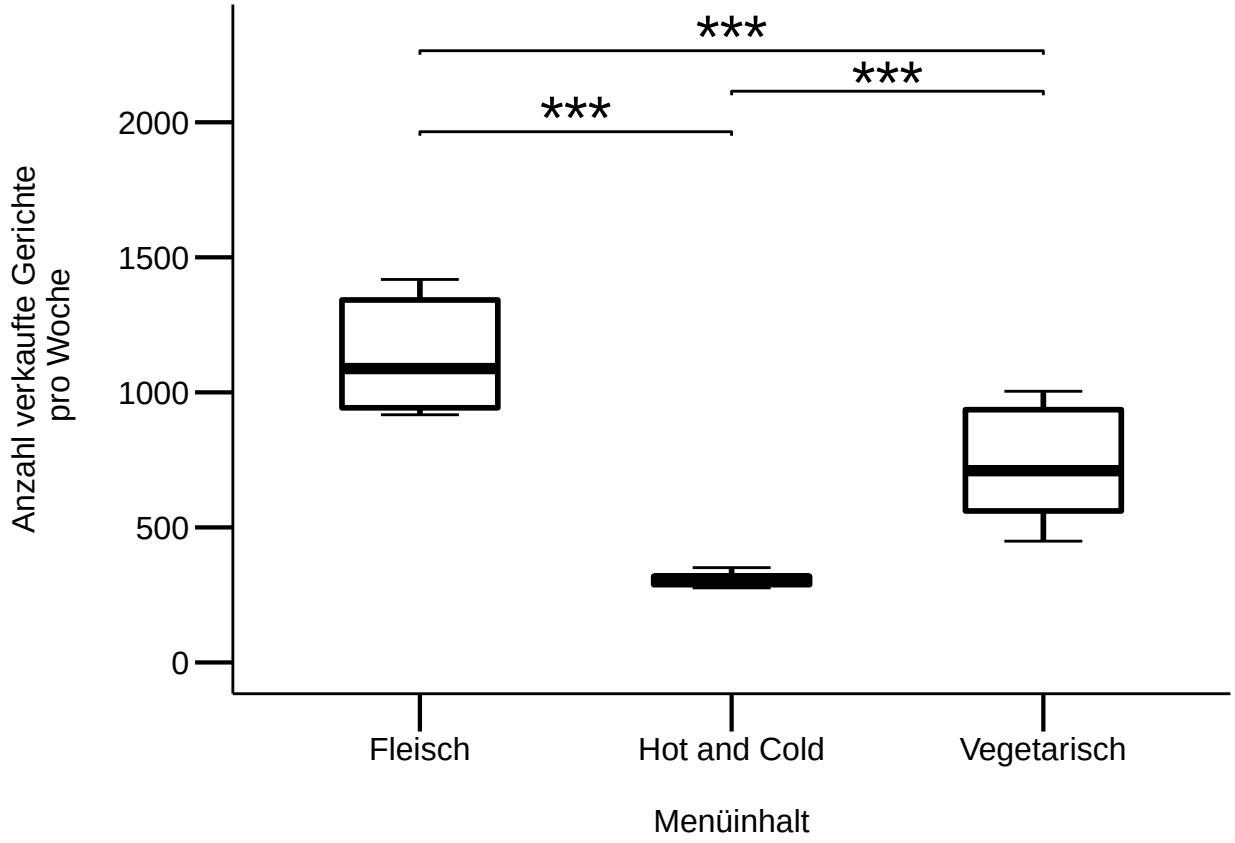


Figure 1: Die wöchentlichen Verkaufzahlen unterscheiden sich je nach Menüinhalt stark. Das Modell wurde mit den log-tranformierten Daten gerechnet.