Praktikum zur Stochastik: Ausarbeitung

Chiying Zoe Lai

May 23, 2023

Inhaltsverzeichnis

1 Einführung

In diesem Paper werden zwei wichtige Werkzeuge in der statistischen Analyse jeweils mit einem in R vorhandenen Datensatz dargestellt: Unabhängigkeitstest bei kategoriellen Daten und ANOVA. Die können dazu beitragen, signifikante Unterschiede zwischen Gruppen zu identifizieren und Hypothesen zu testen.

Im Abschnitt 2 werden die Unabhängigkeit von zwei Merkmalen sowie die Gleichheit auf Verteilung jeweiliger Merkmale bei kategoriellen Daten mithilfe vom χ^2-Test geprüft. Im Abschnitt 3 werden der Test auf Varianzhomogenität, die einfaktorielle ANOVA und noch die Tukey-Methode durchgeführt. Am Ende im Abschnitt 4 wird die gesamte Ergebnisse kurz zusammengefasst.

2 Unabhängigkeits-Test bei kategoriellen Daten (Aufgabe 1)

Unabhängigkeit ist ein zentrales Konzept der Stochastik. Laut Engel (2023) Kapitel 4.6 und Ugarte, Militino, and Arnholt (2015) Kapitel 5.2 heißen zwei diskrete Zufallsvariablen X und Y (stochastisch) unabhängig, falls gilt:

$$p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y). \tag{1}$$

Falls es ein Datenpaar von (x, y) gibt, sodass $p_{X,Y}(x,y) \neq p_X(x) \cdot p_Y(y)$ ist, sind X und Y dann nicht unabhängig (Ugarte, Militino, and Arnholt (2015) Kapitel 5.2).

Um Unabhängigkeit zu testen, gibt es verschiedenen Möglichkeiten (Holzmann et al. (2023) Folien 448-468). Falls X und Y normalverteilt sind, kann man die Unabhängigkeit mit dem Pearson-Unabhängigkeitstest testen. Falls die nicht normalverteilt sind, dann ist die Spearman Rang-Korrelationskoeffizienten verwendbar. In Kontingenztafeln ist der χ^2-Test nützlich, und in 2×2-Kontingenztafeln ist der Fischers exakter Test auch geeignet.

Der χ^2-Test ist eine Art des Homogenitätstests, der in diesem Kapitel demonstriert wird. Laut Engel (2023) Kapitel 8.3 und Ugarte, Militino, and Arnholt (2015) Kapitel 10.8.1 definiert man die Nullhypothese und die Alternative wie folgt.

$$H_0 : X, Y \text{ unabhängig}$$

$$H_1 : X, Y \text{ nicht unabhängig}$$

Und die Prüfgröße ist durch

$$T = \sum_{k=1}^{r} \sum_{l=1}^{s} \frac{(n_{k,l} - \frac{1}{n} \cdot n_{k+} \cdot n_{+l})^2}{\frac{1}{n} \cdot n_{k+} \cdot n_{+l}} \tag{2}$$

gegeben, wobei r und s die Größe der $r \times s$-Kontingenztafel entsprechen. Falls $T \geq \chi^2_{(r-1)(s-1);(1-\alpha)}$ ist, wird H_0 abgelehnt, sonst gibt es kein Widerspruch und die Nullhypothese wird angenommen.

Zunächst wird die Unabhängigkeit von Merkmalen eines in R vorhandenen Datensatzes `HairEyeColor` untersucht, und dann mit dem χ^2-Test getestet.

2.1 Unabhängigkeitstests und Homogenitätstests (Teil a)

Unabhängigkeitstests und Homogenitätstests sind statistische Methoden, um den Zusammenhang zwischen zwei oder mehr Variablen zu untersuchen.

Ein Unabhängigkeitstest wird verwendet, um die Hypothese der Unabhängigkeit von zwei Zufallsvariablen zu prüfen (Engel (2023) Kapitel 8.3; Ugarte, Militino, and Arnholt (2015) Kapitel 10.8.1). Im Allgemeinen testet man die Nullhypothese, dass die beiden Variablen unabhängig voneinander sind. Wenn der Test signifikant ist, wird die Nullhypothese abgelehnt und man kann schließen, dass die Zufallsvariablen nicht unabhängig sind.

Ein Homogenitätstest wird verwendet, um zu bestimmen, ob die Verteilungen von zwei oder mehr Variablen gleich sind. (Holzmann et al. (2023) Folien 470-471)Der Test prüft, ob die Beziehung zwischen den Variablen in verschiedenen Gruppen gleich ist oder nicht. Wenn der Test signifikant ist, wird die Nullhypothese abgelehnt und man kann schließen, dass die Verteilungen unterschiedlich sind.

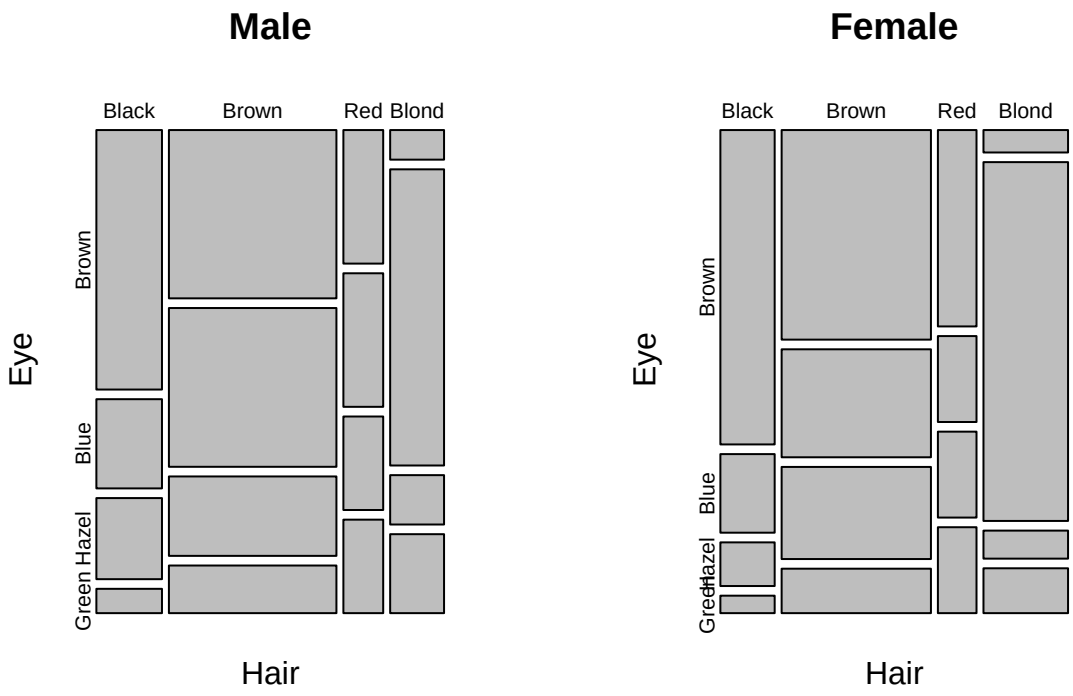
Es gibt einen engen Zusammenhang zwischen Unabhängigkeits- und Homogenitätstests. Ein Unabhängigkeitstest kann als spezieller Fall des Homogenitätstests betrachtet werden, bei dem nur zwei Gruppen verglichen werden. Wenn die Verteilungen der Variablen in den Gruppen gleich sind, sind die Variablen unabhängig voneinander. (Holzmann et al. (2023) Folien 471) Wenn die Verteilungen jedoch unterschiedlich sind, sind die Variablen abhängig voneinander.

2.2 Überblick auf den Datensatz `HairEyeColor` (Teile b-c)

Der Datensatz `HairEyeColor` ist von R vorhanden, in dem die Verteilung von Haare- und Augenfarben von 592 Studierende der Statistik in 1974 an der Universität in Delaware nach Geschlecht aufgenommen wurde. Der sieht wie folgt aus:

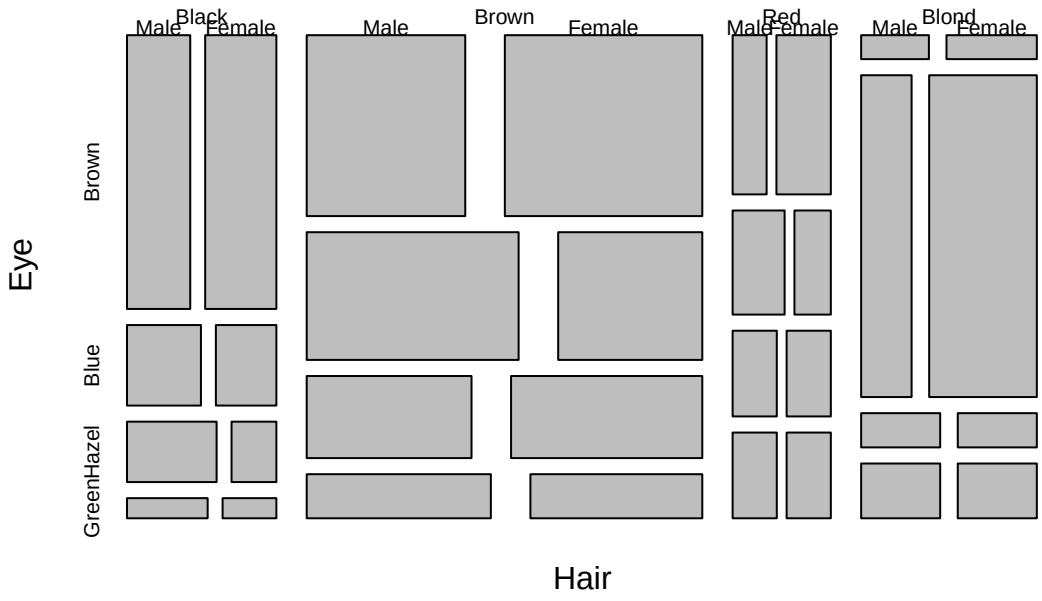
```
## , , Sex = Male
##
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    32   11    10     3
##   Brown    53   50    25    15
##   Red      10   10     7     7
##   Blond     3   30     5     8
##
## , , Sex = Female
##
##        Eye
## Hair    Brown Blue Hazel Green
##   Black    36    9     5     2
##   Brown    66   34    29    14
##   Red      16    7     7     7
##   Blond     4   64     5     8
```

Nun wird der mithilfe von Mosaik-Plots visualisiert, sodass wir leichter ein Überblick bekommen können. Zuerst werden die Daten nach Geschlechter getrennt untersucht, und danach zusammen.



Grafik 1: Verteilung der Augen- und Haarfarbe für männliche (links) bzw. weibliche (rechts) Studierende

Aus Grafik 1 kann man durch die Fläche sehen, dass die Kombinationen von braunen Haare mit braunen oder blauen Augen bei Männern am meistens auftreten, und dass die Kombinationen von blonden Haare mit blauen Augen und braunen Haare mit braunen Augen bei Frauen am meistens auftreten. Andererseits treten die Kombinationen von schwarzen Haare mit grünen Augen und blonden Haare mit braunen Augen bei Männern am seltenstens auf. Die von schwarzen Haare mit grünen Augen tritt auch bei Frauen am seltenstens auf.



Grafik 2: Verteilung der Augen- und Haarfarbe für beide männliche und weibliche Studierende

Aus Grafik 2 kann man analog sehen, dass die Fläche von der Kombination mit braunen Haare und braunen Augen am größten ist, so ist diese Kombination bei beiden Geschlechter am häufigsten aufgetroffen. Andererseits ist tritt die Kombination von schwarzen Haare mit grünen Augen am seltestens auf.

Wir können statt plotten auch die Wahrscheinlichkeiten berechnen. Der ursprüngliche Datensatz zeigt die absolute Häufigkeiten. Wenn wir die auf relative Häufigkeiten umwandeln, kann man die Wahrscheinlichkeiten für die einzelnen Haar- und Augenfarbekombination genau sehen und die Randhäufigkeiten dann berechnen (Engel (2023) Kapitel 8.1.1; Holzmann et al. (2023) Folien 423-424). Die relevante Kontingenztafel zu dem Datensatz `HairEyeColor` werden wie folgt gebildet.

Kontingenztafel für männlichen Studierende:

```
male_relative <- prop.table(xtabs(Freq ~ Hair + Eye, data = HairEyeColor[,,1]))
male_table <- addmargins(male_relative,  margin = 1:2)
male_table
```

```
##        Eye
## Hair          Brown       Blue      Hazel      Green        Sum
##   Black 0.11469534 0.03942652 0.03584229 0.01075269 0.20071685
##   Brown 0.18996416 0.17921147 0.08960573 0.05376344 0.51254480
##   Red   0.03584229 0.03584229 0.02508961 0.02508961 0.12186380
##   Blond 0.01075269 0.10752688 0.01792115 0.02867384 0.16487455
##   Sum   0.35125448 0.36200717 0.16845878 0.11827957 1.00000000
```

Von der Tafel können wir sehen, dass die Wahrscheinlichkeit, dass Männer braune Haare mit braunen Augen haben, am höchstens ist (18.996416%). Dann folgen die Kombination von braune Haare mit blauen Augen (17.921147%) und dann die von schwarzen Haare mit braunen Augen (11.469534%).

Die Kombinationen von schwarzen Haare mit grünen Augen (1.075269%), die von blonden Haare mit braunen Augen (auch 1.075269%) und die von blonden Haare mit rehbraunen Augen (1.792115%) kommen jedoch am seltestens vor.

Analog können wir die andere Kontingenztafel auch so bilden und davon untersuchen.

Kontingenztafel für weiblichen Studierende:

```
##        Eye
## Hair          Brown        Blue       Hazel       Green         Sum
##   Black 0.115015974 0.028753994 0.015974441 0.006389776 0.166134185
##   Brown 0.210862620 0.108626198 0.092651757 0.044728435 0.456869010
##   Red   0.051118211 0.022364217 0.022364217 0.022364217 0.118210863
##   Blond 0.012779553 0.204472843 0.015974441 0.025559105 0.258785942
##   Sum   0.389776358 0.364217252 0.146964856 0.099041534 1.000000000
```

Bei Frauen liegt die Wahrscheinlichkeit von braunen Haare mit braunen Augen bei 21.0862620%, die noch höcher als bei Männern ist. Und die von blonden Haare mit blauen Augen hat auch eine deutlich höche Wahrscheinlichkeit bei 20.4472843%. Die dritte häufigste Kombination ist dann schwarze Haare mit braunen Augen (11.5015974%).

Am wenigsten Frauen haben andererseits schwarzen Haare mit grünen Augen (0.6389776%). Blonde Haare mit braunen Augen und rehbraunen Augen mit schwarzen/blonden Haare treten dann auch selten auf (1.2779553% bzw. 1.5974441%)

Nun werden beide Geschlechter zusammen untersucht.

Kontingenztafel für beide männliche und weibliche Studierende zusammen:

```
##        Eye
## Hair          Brown        Blue       Hazel       Green         Sum
##   Black 0.114864865 0.033783784 0.025337838 0.008445946 0.182432432
##   Brown 0.201013514 0.141891892 0.091216216 0.048986486 0.483108108
##   Red   0.043918919 0.028716216 0.023648649 0.023648649 0.119932432
##   Blond 0.011824324 0.158783784 0.016891892 0.027027027 0.214527027
##   Sum   0.371621622 0.363175676 0.157094595 0.108108108 1.000000000
```

Zusammengestellt kommt die Kombination von braunen Haare mit braunen Augen mit 20.1013514% am meistens vor. Danach kommen blonde Haare mit blauen Augen (15.8783784%) und braune Haare mit blauen Augen (14.1891892%) dann vor.

Schwarze Haare mit grünen Augen tritt andererseits am seltestens auf (0.8445946%). Und blonde Haare mit braunen bzw. rehbraunen Augen haben die zweitniedrigste bzw. drittniedrigste Wahrscheinlichkeiten (1.1824324% bzw. 1.6891892%).

Von diesen Kontingenztafeln kann man auch leicht auf die "`Sum`" Spalten und Reihen die relevante Randhäufigkeiten sehen. Zum Beispiel haben fast die Hälfte (48.3108108%) der Studierende braunen Haare.

2.3 Annahme der Unabhängigkeit (Teile d-e)

Wenn die Merkmale Haar- und Augenfarbe unabhängig wären, würde die Formel (1) für alle Datenpaare gelten. Dann könnten wir zum Beispiel die erwartete Wahrscheinlichkeit der männlichen Studenten, die schwarze Haare und braunen Augen haben, wie folgt bestimmen:

$p_{Haarfarbe,Augenfarbe}(schwarz, braun)$

$= p_{Haarfarbe}(schwarz) \cdot p_{Augenfarbe}(braun)$

$= 0.20071685 \cdot 0.35125448$

$= 0.070502692$

Dann gibt $0.070502692 * 279$ männliche Studenten die erwartete Anzahl von $19.67025128 \approx 20$ aus. Also 20 statt 32 Studenten, die schwarze Haare und braunen Augen haben, sind erwartet, falls beide Merkmale unabhängig voneinander wären. Ebenso können wir die erwartete Anzahlen für alle Kombinatiionsmöglichkeiten und auch für weibliche bzw. alle Studenten bestimmen.

Erwartete Wahrscheinlichkeiten für männlichen Studierende:

```
#Die Summen der Möglichkeiten von jeweilige Haarfarbe und jeweilige Augenfarbe
#in einen Vektor abspeichern
male_hair_marginals = male_table[1:4,"Sum"]
male_eye_marginals = male_table["Sum",1:4]

#Ergebnistabelle initialisieren
male_expected_prob <- xtabs(rep(0, times = 16) ~ Hair + Eye, data = HairEyeColor[,,1])

#Erwartete Wahrscheinlichkeiten bestimmen
for (i in 1:4) {
  for (j in 1:4) {
    male_expected_prob[i, j] <- male_hair_marginals[i] * male_eye_marginals[j]
  }
}
male_expected_prob
```

```
##        Eye
## Hair          Brown       Blue      Hazel      Green
##   Black 0.07050269 0.07266094 0.03381252 0.02374070
##   Brown 0.18003366 0.18554489 0.08634267 0.06062358
##   Red   0.04280521 0.04411557 0.02052903 0.01441400
##   Blond 0.05791293 0.05968577 0.02777457 0.01950129
```

Erwartete Anzahlen für männlichen Studierende:

```
#Erwartete Anzahlen bestimmen
male_expected_count <- sum(HairEyeColor[,,1])*male_expected_prob
male_expected_count
```

```
##        Eye
## Hair         Brown      Blue     Hazel     Green
##   Black 19.670251 20.272401  9.433692  6.623656
##   Brown 50.229391 51.767025 24.089606 16.913978
##   Red   11.942652 12.308244  5.727599  4.021505
##   Blond 16.157706 16.652330  7.749104  5.440860
```

Analog können die Anzahlen für weiblichen bzw. alle Studierenden berechnet werden.

Erwartete Anzahlen für weiblichen Studierende:

```
##        Eye
## Hair        Brown      Blue     Hazel     Green
##   Black 20.268371 18.939297  7.642173  5.150160
##   Brown 55.738019 52.083067 21.015974 14.162939
##   Red   14.421725 13.476038  5.437700  3.664537
##   Blond 31.571885 29.501597 11.904153  8.022364
```

Erwartete Anzahlen für alle Studierende zusammen:

```
##        Eye
## Hair         Brown       Blue      Hazel      Green
##   Black  40.135135  39.222973  16.966216  11.675676
##   Brown 106.283784 103.868243  44.929054  30.918919
##   Red    26.385135  25.785473  11.153716   7.675676
##   Blond  47.195946  46.123311  19.951014  13.729730
```

Nun werden die jeweiligen quadrierten Abweichungen der beobachteten Werte des Datensatzes von diesen berechneten erwarteten Werten bestimmen, um zu schauen, wie groß die Differenzen zwischen den erwarteten und den tatsächlichen Werten sind.

Quadrierte Abweichungen für männliche Studierende:

```
male_squared_deviations <-
  (xtabs(Freq ~ Hair + Eye, data = HairEyeColor[,,1]) - male_expected_count)^2
male_squared_deviations
```

```
##        Eye
## Hair          Brown        Blue       Hazel       Green
##   Black 152.0227130  85.9774283   0.3207050  13.1308822
##   Brown   7.6762760   3.1223777   0.8288177   3.6633137
##   Red     3.7738981   5.3279891   1.6190054   8.8714302
##   Blond 173.1252296 178.1603011   7.5575725   6.5491964
```

Quadrierte Abweichungen für weibliche Studierende:

```
##        Eye
## Hair           Brown         Blue        Hazel        Green
##   Black  247.4841634   98.7896273    6.9810756    9.9235064
##   Brown  105.3082506  326.9973155   63.7446641    0.0265492
##   Red      2.4909512   41.9390726    2.4407823   11.1253152
##   Blond  760.2088416 1190.1397789   47.6673335    0.0005002
```

Quadrierte Abweichungen für alle Studierende zusammen:

```
##        Eye
## Hair           Brown         Blue        Hazel        Green
##   Black  776.4506939  369.5226899    3.8660062   44.5646457
##   Brown  161.7021549  394.7470896   82.2820604    3.6822498
##   Red      0.1483291   77.1845354    8.1013314   39.9970782
##   Blond 1615.7140705 2292.1773677   99.0226699    5.1541271
```

Mit diesen Abweichungen können wir dann die Prüfgröße T des χ^2-Tests berechnen, indem wir die Formel (2) verwenden.

2.4 Unabhängigkeitstest: der χ^2-Test (Teile f-h)

Die χ^2-Teststatistik ist in der Formel (2) gegeben. Der Teil $\frac{1}{n} \cdot n_{k+} \cdot n_{+l}$ kann man auch umwandelt, wie wir in dem letzten Abschnitt gerechnet haben. $\frac{1}{n} \cdot n_{k+} \cdot n_{+l} = \frac{n^2}{n} \cdot \frac{n_{k+}}{n} \cdot \frac{n_{+l}}{n} = n \cdot \frac{n_{k+}}{n} \cdot \frac{n_{+l}}{n} = n \cdot p_X(x) \cdot p_Y(y)$ gibt genau die erwartete Anzahlen under Annahme der Unabhängigkeit aus. Deshalb kann man die Prüfgröße auch wie folgt definieren (Engel (2023) Kapitel 8.3):

$$T = \sum_k \sum_l \frac{(\text{beobachtet}_{k,l} - \text{erwartet}_{k,l})^2}{\text{erwartet}_{k,l}}. \tag{3}$$

Nun werden die Prüfgrößen jeweils für beide Geschlechter getrennt und nochmal alle zusammen berechnet.

```
#Prüfgröße für männliche Studierende:
T_male <- sum(male_squared_deviations / male_expected_count)
T_male
```

```
## [1] 41.28029
```

```
#Prüfgröße für weibliche Studierende:
T_female <- sum(female_squared_deviations / female_expected_count)
T_female
```

```
## [1] 106.6637
```

```
#Prüfgröße für alle Studierende zusammen:
T_maleandfemale <- sum(maleandfemale_squared_deviations / maleandfemale_expected_count)
T_maleandfemale
```

```
## [1] 138.2898
```

Und die 0.95-Quantil der entsprechenden χ^2-Verteilungen können mithilfe der "`qchisq`" wie folgt berechnet werden.

```
#0.95-Quantil der Chi-Quadrat-Verteilung:
#degree of freedom = (Anzahl der Haarfarbe - 1) * (Anzahl der Augenfarbe - 1) = 3*3 = 9
qchisq(p = 0.95, df = 9)
```

```
## [1] 16.91898
```

Die Ausgabe von 16.91898 ist viel kleiner als unsere berechnete Prüfgrößen, insbesondere für weibliche Studierende und für alle Studierende zusammen. Deshalb wird die Nullhypothese signifikant zum Niveau 0.05 abgelehnt (Holzmann et al. (2023) Folien 338), sodass die Merkmale als nicht unabhängig angenommen werden.

Nun werden noch die p-Werte bestimmt, gegeben sind unsere jeweilige berechnete Prüfgröße:

```r
#p-Wert für männliche Studierende:
pchisq(q = T_male, df = 9, lower.tail = FALSE)
```

```
## [1] 4.447279e-06
```

```r
#p-Wert für weibliche Studierende:
pchisq(q = T_female, df = 9, lower.tail = FALSE)
```

```
## [1] 7.014013e-19
```

```r
#p-Wert für alle Studierende zusammen:
pchisq(q = T_maleandfemale, df = 9, lower.tail = FALSE)
```

```
## [1] 2.325287e-25
```

Alle p-Werte sind eher insignifikant, sodass die Daten eher stark gegen die Nullhypothese sprechen (Holzmann et al. (2023) Folien 348). Hier sind die Daten wieder bei weiblichen Studierende und bei allen Studierende zusammen noch signifikanter als bei männlichen Studierende, sodass es noch sicherer ist, dass die Merkmale hier nicht unabhängig sind.

Im Folgenden zeigt eine Alternative, die die quadrierte Abweichungen mit entsprechendem p-Wert mithilfe einer in R vorhandenen Funktion "`chisq.test`"ausgibt.

```r
#Für männliche Studierende:
chisq.test(HairEyeColor[,,1])
```

```
##
##  Pearson's Chi-squared test
##
## data:  HairEyeColor[, , 1]
## X-squared = 41.28, df = 9, p-value = 4.447e-06
```

```r
#Für weibliche Studierende:
chisq.test(HairEyeColor[,,2])
```

```
##
##  Pearson's Chi-squared test
##
## data:  HairEyeColor[, , 2]
## X-squared = 106.66, df = 9, p-value < 2.2e-16
```

```r
#Für alle Studierende zusammen:
chisq.test(HairEyeColor[,,1] + HairEyeColor[,,2])
```

```
##
##  Pearson's Chi-squared test
##
## data:  HairEyeColor[, , 1] + HairEyeColor[, , 2]
## X-squared = 138.29, df = 9, p-value < 2.2e-16
```

Und die Werten stimmen mit den oben gerechneten Werten überein.

2.5 Homogenitätstest: Auch der χ^2-Test (Teil i)

Zuletzt wird noch getestet, ob die Augenfarbe bzw. Haarfarbe bei Frauen und Männern identisch verteilt sind. Und der χ^2-Test ist hier auch verwendbar (Holzmann et al. (2023) Folien 471, Wollschläger (2010) Kapitel 6.2). Die Nullhypothese entspricht die identische Verteilung, und die Alternative keine identische Verteilung.

```
eye_gender <- xtabs(Freq ~ Eye + Sex, data = HairEyeColor)
chisq.test(eye_gender)
```

```
##
##  Pearson's Chi-squared test
##
## data:  eye_gender
## X-squared = 1.5298, df = 3, p-value = 0.6754
```

```
hair_gender <- xtabs(Freq ~ Hair + Sex, data = HairEyeColor)
chisq.test(hair_gender)
```

```
##
##  Pearson's Chi-squared test
##
## data:  hair_gender
## X-squared = 7.9942, df = 3, p-value = 0.04613
```

Bei Augenfarbe ist der p-Wert des Tests 0.6754, die größer als 0.05 ist. Deshalb gibt es kein Widerspruch zur Nullhypothese. Also die Augenfarbe ist bei beiden Geschlechter identisch verteilt. Aber der p-Wert des Tests bei Haarfarbe ist 0.04613, die etwas kleiner als 0.05 ist. Die Nullhypothese wird abgelehnt, also die Haarfarbe ist bei beiden Geschlechter nicht identisch verteilt.

3 Varianzanalyse - ANOVA (Aufgabe 2)

Die Varianzanalyse ist eine wichtige Anwendungsgebiet des linearen Modells. Ein lineare Modell untersucht einen Zusammenhang zwischen einer quantitativen Zielvariablen und einer oder mehreren meist quantitativen erklärenden Variablen, und die Varianzanalyse ist mit dem Spezialfall von qualitativen erklärenden Variablen beschäftigt (Engel (2023) Kapitel 9). Der Name "ANOVA" stammt aus der englischen "Analysis of Variance".

Um Varianzhomogenität zu testen, gibt es verschiedenen Möglichkeiten: der F-Test, der Fligner-Killeen-Test, der Bartlett-Test und der Levene-Test (Wollschläger (2010) Kapitel 8.1). Die ersten 3 Tests sind aber sensibel auf Verletzungen der Voraussetzung von Normalverteiltheit. Die Nullhypothese und die Alternative werden im Allgemeinen wie folgt definiert:

$$H_0 : \sigma_1^2 = \cdots = \sigma_k^2 \text{ (d.h. gleiche Varianzen in den } k \text{ Gruppen).}$$

$$H_1 : \exists i, j \in \{1, ..., k\} : \sigma_i^2 \neq \sigma_j^2 \text{ (d.h. nicht alle Varianzen sind gleich).}$$

Laut Engel (2023) (Kapitel 10.2) nennen wir die ANOVA einfaktoriell, wenn es nur eine qualitative Variable im Modell gibt. Mit mehreren qualitativen Variablen nennt man dann die zweifactorielle ANOVA, aber hier wird nur die einfaktorielle ANOVA untersucht. Wenn wir einen Faktor mit k verschiedenen Gruppen/Klassen/Levels haben, modellieren wir die Zielvariable y mit μ als den globalen Erwartungswert der Zeilgröße und $\alpha_i = \mu_i - \mu$ als die Einflüsse jeder Faktorstufe wie folgt:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),\ i = 1, ..., k,\ j = 1, ..., n_i.$$

Vorausgesetzt ist, dass die Zeilvariable normalverteilt ist und die Fehler ε_{ij} für alle Beobachtungen die gleiche Varianz σ^2 haben. Die zugehörige Nullhypothese und Alternative sind wie folgt:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k.$$

$$H_1 : \exists 1 \leq i, j \leq k : \mu_i \neq \mu_j.$$

Die Tukey-Methode testet noch genauer, sodass man wissen kann, welche Paare unterschiedliche Mittelwerte haben. Es gilt unter der Nullhypothese auch $H_0 : \mu_1 = \mu_2 = ... = \mu_k$, und die Hypothese $\mu_{i_1} = \mu_{i_2}$ für $1 \leq i_1, i_2 \leq k$ auf dem Niveau α für diejenigen Paare (i_1, i_2), bei denen die Null nicht im Konfidenzintervall liegt. (Engel (2023) Kapitel 10.2)

In diesem Kapitel wird der Test auf Varianzhomogenität, die ANOVA und die Tukey-Methode durch einen Datensatz `ewr` aus dem Paket `UsingR` dargestellt.

3.1 Überblick auf den Datensatz `ewr` (Teile a-b)

Der Datensatz `ewr` enthält monatliche Taxi-In- und Taxi-Out-Time zwischen Januar 1999 und November 2000 von 8 verschiedenen Fluglinien am EWR Newark Flughafen. Die ersten 6 Zeilen sehen wie folgt aus.

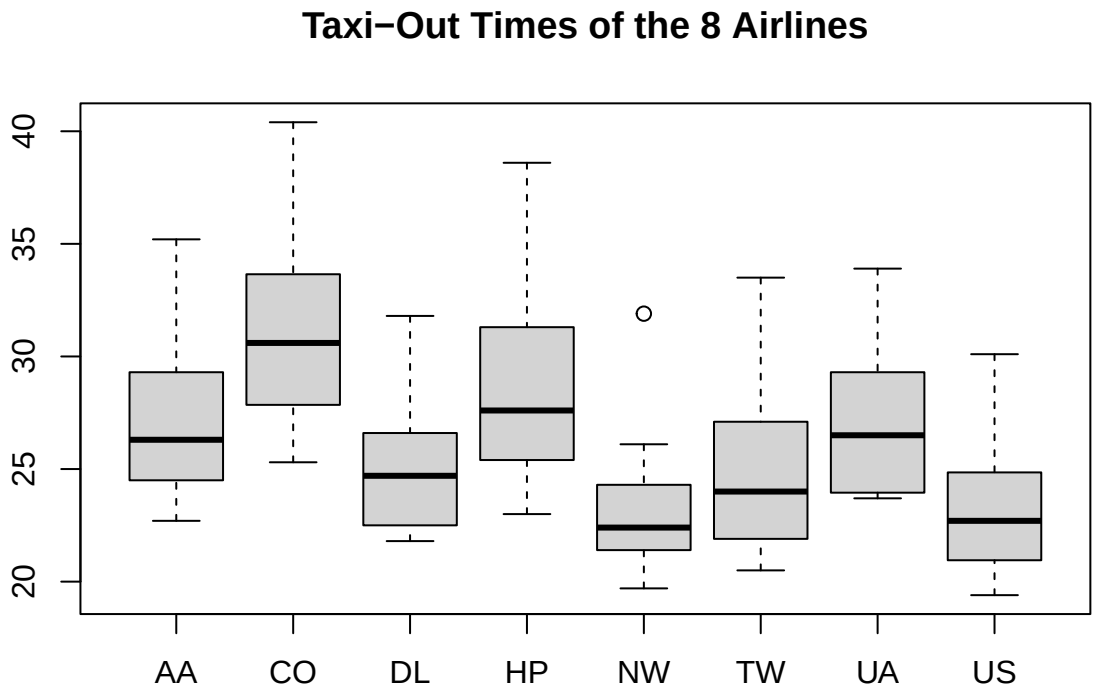
```
##   Year Month  AA  CO  DL   HP  NW   TW  UA  US inorout
## 1 2000   Nov 8.6 8.3 8.6 10.4 8.1  9.1 8.4 7.6      in
## 2 2000   Oct 8.5 8.0 8.4 11.2 8.2  8.5 8.5 7.8      in
## 3 2000   Sep 8.1 8.5 8.4 10.2 8.3  8.6 8.2 7.6      in
## 4 2000   Aug 8.9 9.1 9.2 14.5 9.0 10.3 9.2 8.7      in
## 5 2000   Jul 8.3 8.9 8.2 11.5 8.8  9.1 9.2 8.2      in
## 6 2000   Jun 8.8 9.0 8.8 14.9 8.4 10.8 8.9 8.3      in
```

Hier wird nur die Taxi-Out Time untersucht. Dafür werden 2 Variablen vom Datensatz extrahiert:

```
taxi_out <- ewr[ewr$inorout == "out", 3:10] # speichert die Taxi-Out Zeiten bzgl. Fluglinien
airlines <- colnames(taxi_out) # speichert die 8 Namen der Fluglinien
```

Um einen Eindruck über die Verteilungen von diesen Taxi-Out Time (in Minuten) verschiedener Fluglinien zu bekommen, werden zunächst die entsprechenden Boxplots erstellt.

```
boxplot(taxi_out, main = "Taxi-Out Times of the 8 Airlines")
```



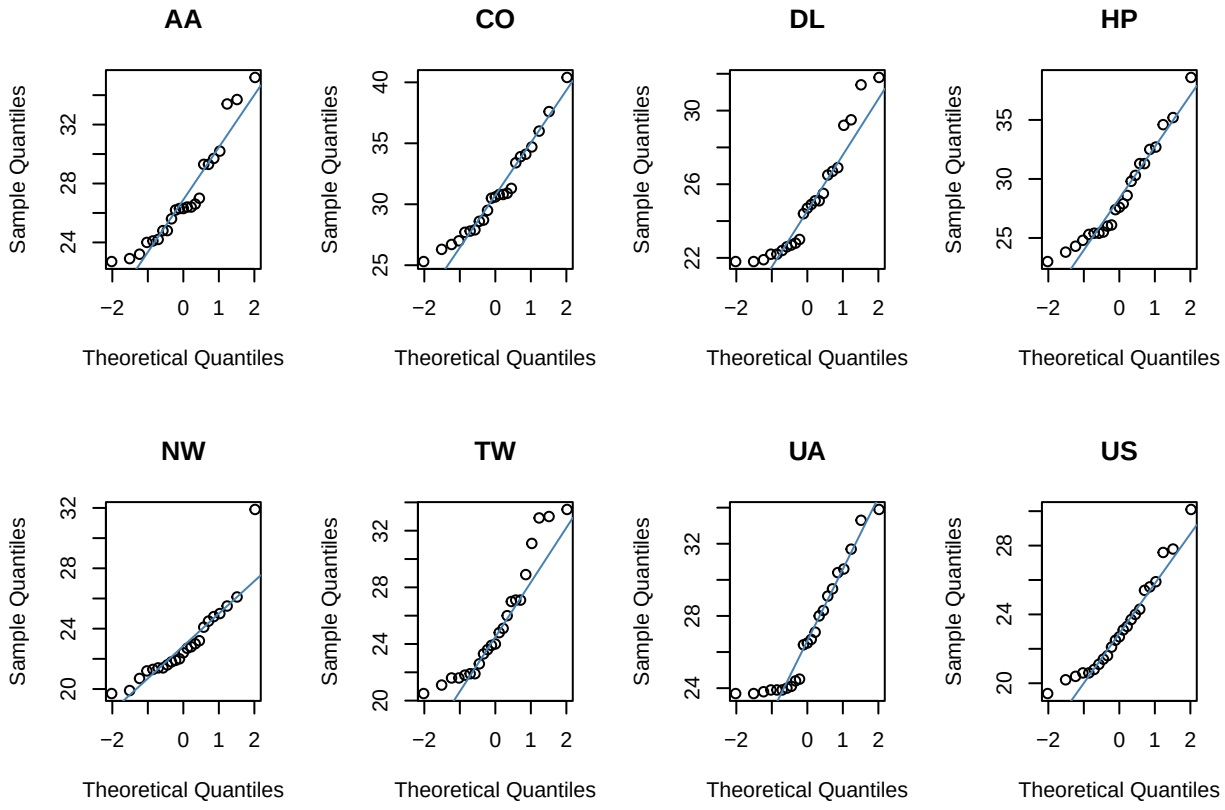
Grafik 3: Boxplots der Taxi-Out Time von den 8 Fluglinien

Die Grafik 3 zeigt, dass die Fluglinie `CO` einen deutlich längeren Taxi-Out Time mit einem Mittelwert (50% Quantil) von ungefähr 31 Minuten hat. Aber 6 anderen Fluglinien haben die 75% Quantile unter 30 Minuten, und die restliche Fluglinie hat auch eine 75% Quantil von ungefähr 32 Minuten. Die Verteilung von `CO` ist insgesamt höcher als allen anderen Fluglinien.

Die Fluglinie `NW` hat gegenseitig den niedrigsten Mittelwert von ungefähr 22.5 Minuten mit einer 75%-Quantil von unter 25 Minuten. Die hat jedoch einen Außreiser von rund 32.5 Minuten, bleibt die 90%-Quantil aber auf 27 Minuten.

Mithilfe von QQ-Plots kann man zusätzlich noch schauen, ob die Variablen normalverteilt sind. Grafik 4 zeigt auch, dass die Verteilung von `CO` insbesonderes hoch mit einem Mittelwert von rund 31 Minuten ist, und dass die `NW` meistens unter rund 26 Minuten ist aber einen Außreiser von rund 32 Minuten hat. Außerdem sehen die Punkte jeweils von den Fluglinien `CO`, `HP` und `US` am besten auf der Gerade aus, sodass diese Werte möglich normalverteilt sind.

```
par(mfrow = c(2,4))
for (i in 1:8) {
  qqnorm(taxi_out[,i], main = airlines[i])
  qqline(taxi_out[,i], col = "steelblue")
}
```



Grafik 4: QQ-Plots der Taxi-Out Time von den 8 Fluglinien

3.2 Varianzhomogenität (Teil c)

Nun werden die Standardabweichungen für die einzelnen Beobachtungen bestimmt. Und die Airlines, die die größte Unterschied in deren Standardabweichung haben, werden dann auf Varianzhomogenität noch getestet. Für viele Tests auf Varianzhomogenitätist die Normalverteiltheit vorausgesetzt (Wollschläger (2010) Kapitel 8).

Wir können zusätzlich zu den oben gestellten QQ-Plots den Shapiro-Test durchführen, um zu schauen, ob die Werte normalverteilt sind. Die Nullhypothese spricht, dass eine Normalverteilung vorliegt, und die Alternative keine (Engel (2023) Kapitel 7.3.2.). Wenn der p-Wert größer als 0.05 ist, nehmen wir an, dass die Beobachtungen normalverteilt sind.

Die Ergebnisse vom Shapiro-Test und den Standardabweichungen stehen im Folgenden:

```
# Vectoren initialisieren
shapiro.test.pvalues <- numeric(8)
normalverteilt <- logical(8)

# shapiro.test() verwenden, um die Normalverteilung zu testen
for (i in 1:8) {
  # speichere den p-wert
  shapiro.test.pvalues[i] <- shapiro.test(taxi_out[,i])$p

  # Nullhypothese nicht verwerfen, falls den p-Wert > 0.05 ist
  normalverteilt[i] <- shapiro.test.pvalues[i] > .05
}
```

```r
# in eine neue Dataframe speichern und die Zeilennamen auf die Fluglinien setzen
airlines.data <- data.frame(shapiro.test.pvalues, normalverteilt)
rownames(airlines.data) <- airlines

# Standardabweichungen bestimmen und auf Dataframe hinzufügen
sd.dev <- apply(taxi_out, MARGIN = 2, FUN = sd)
airlines.data <- cbind(airlines.data, sd.dev)
airlines.data
```

```
##    shapiro.test.pvalues normalverteilt   sd.dev
## AA          0.023119229          FALSE 3.504909
## CO          0.230858830           TRUE 3.860159
## DL          0.008399171          FALSE 3.067347
## HP          0.120117222           TRUE 4.127144
## NW          0.002226341          FALSE 2.602006
## TW          0.012938358          FALSE 4.056698
## UA          0.007928876          FALSE 3.297610
## US          0.132234267           TRUE 2.786692
```

Die Fluglinien `HP` hat die größte Standardabweichung von rund 4.13, und die `NW` hat die niedrigste Standardabweichung von rund 2.60. Da nicht beide normalverteilt sind, ist der Levene-Test zum Testen der Varianzhomogenität geeignet (Wollschläger (2010) Kapitel 8.1). Davor muss man das Paket `cars` laden.

```r
# leveneTest() verwenden, da die Taxi-Out-Zeiten bei NW nicht normalverteilt ist.
# alle Zeiten von beiden Fluglinien eingeben, und dann als Gruppe 1 bzw. 2 aufteilen.
leveneTest(y = c(taxi_out[,"HP"], taxi_out[,"NW"]), group = factor(c(rep(1, 23), rep(2, 23))))
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  1  5.0747 0.02931 *
##       44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Der p-Wert ist 0.02931, deshalb würde man nur auf 1% Niveau die Nullhypothese, dass die Varianzen homogen sind, nicht verwerfen. Wenn die allgemeine 5% Niveau erwartet ist, verwirft man die Nullhypothese.

```r
# trotz Verletzung der Normalverteilung var.test() verwenden.
var.test(x = taxi_out[,"HP"], y = taxi_out[,"NW"])
```

```
##
##  F test to compare two variances
##
## data:  taxi_out[, "HP"] and taxi_out[, "NW"]
## F = 2.5158, num df = 22, denom df = 22, p-value = 0.03543
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.066991 5.932048
## sample estimates:
## ratio of variances
##           2.515838
```

Trotz der Verletzung der Normalverteilung ist der p-Wert 0.03543, der ähnlich wie beim Levene-Test ist. Da der p-Wert auch weniger als 5% ist, wird die Nullhypothese abgelehnt, d.h. die beide Varianzen sind nicht homogen. Da die beide sind die Gruppen, die die größte Unterschied auf Standardabweichungen haben, folgt auch, dass die gesamte Varianzen von allen 8 Fluglinien verschieden sind.

3.3 ANOVA (Teil d)

Nun wird die ANOVA durchgeführt, in der die Taxi-Out-Time als Zielvariable verwendet wird und die Fluglinien als Gruppen. In R kann man "anova" auf dem "lm()" Modell anwenden, oder einfach "aov" verwenden (Engel (2023) Kapitel 10.2).

```
aovmodel = aov(values ~ ind, data = stack(taxi_out))
summary(aovmodel)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## ind           7   1155  165.01   13.82 3.27e-14 ***
## Residuals   176   2101   11.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Der Ergebnis zeigt, dass die F-Statistik einen p-Wert von 3.27e-14 hat, die signifikant kleiner als 1% ist. Deshalb wird die Nullhypothese ziemlich sicher verworfen. Es gibt dann mindestens ein Mitterlwert, der unterschiedlich von den anderen ist. Aber man weiß noch nicht, welche Gruppe von anderen anders ist.

3.4 Tukey-Methode (Teil e)

Zuletzt kann man noch die Tukey-Methode anwenden, um die Konfidenzintervalle jedes Paares von Gruppen zu bestimmen (Engel (2023) Kapitel 10.2).

```
# Tukey-Methode auf 99% Konfidenzintervalle verwenden und auf Dataframe setzen
tukey_result <- TukeyHSD(aovmodel, conf.level = .99)
tukey_table <- as.data.frame(tukey_result$ind)

# Finde Intervalle, die nicht die Null beinhaltet
extracted_tukey_table <- tukey_table[tukey_table$lwr > 0 | tukey_table$upr < 0, ]
extracted_tukey_table
```

```
##            diff         lwr        upr        p adj
## CO-AA  3.834783   0.1782378  7.4913274 5.479131e-03
## NW-AA -4.060870  -7.7174143 -0.4043248 2.452621e-03
## US-AA -3.830435  -7.4869796 -0.1738900 5.562045e-03
## DL-CO -5.886957  -9.5435013 -2.2304117 9.307550e-07
## NW-CO -7.895652 -11.5521969 -4.2391074 1.982225e-11
## TW-CO -5.486957  -9.1435013 -1.8304117 6.283010e-06
## UA-CO -3.873913  -7.5304578 -0.2173683 4.782751e-03
## US-CO -7.665217 -11.3217622 -4.0086726 7.379131e-11
## NW-HP -5.586957  -9.2435013 -1.9304117 3.931556e-06
## US-HP -5.356522  -9.0130665 -1.6999770 1.147758e-05
## UA-NW  4.021739   0.3651944  7.6782839 2.827573e-03
## US-UA -3.791304  -7.4478491 -0.1347596 6.362216e-03
```

```
# Bestimme, wieviele solche Intervalle es gibt
dim(extracted_tukey_table)[1]
```

```
## [1] 12
```

Es gibt insgesamt 12 von 28 Paare ein 99% Konfidenzintervall, das die Null nicht enthält. Alle diese Paare haben sogar einen p-Wert $< 1\%$, sodass die Nullhypothese abgelehnt wird, d.h. es gibt deutliche Unterschiede zwischen den Mittelwerte dieses Paares. Das stimmt mit den vorherige F-Statistik überein, dass mindestens eine Fluglinie ein unterschiedliche Mittelwert der Taxi-Out-Time von anderen hat.

4 Zusammenfassung

Insgesamt spielen verschiedene Tests im Bereich Stochastik eine große Rolle, sodass man sich leicht von dem relevanten p-Wert entscheiden kann, ob die zugehörige Nullhypothese abgelehnt oder nicht abgelehnt wird. Diese Tests geben Eindrucken aus, wie die Daten verhalten.

Mithilfe von χ^2-Test haben wir festgelegt, dass die Merkmale Augen- und Haarfarbe bei allen Gruppen vom Datensatz `HairEyeColor` zum Niveau 0.05 nicht unabhängig sind, und dass die Augenfarbe bei beiden Geschlechter identisch verteilt ist, aber die Haarfarbe nicht.

Der Datensatz `ewr` wurde andererseits mit verschiedenen Methode untersucht, wie die Verteilungen der Taxi-Out-Time bei den 8 Fluglinien aussehen. Es wurde vom Test auf Varianzhomogenität überzeugt, dass die Varianzen der Taxi-Out-Time von den Fluglinien verschieden sind. Und durch die ANOVA und die Tukey-Methode wurde es schlussfolgert, dass die Mittelwerte auch eher sicher unterschiedlich sind. Insgesamt sind sowohl die Varianzen als auch den Mittelwerte von Taxi-Out-Time der 8 Fluglinien versechieden.

References

Engel, Dirk. 2023. "Statistische Datenanalyse Und Modellierung Mit r: Kursunterlagen Für Das Praktikum Zur Stochastik."

Holzmann, Prof. Dr. Hajo, Max Berger, Kevin Wilk, and Lisa Drescher. 2023. "Praktikum Zur Stochastik: Statistische Datenanalyse Und Modellierung Mit r."

Ugarte, M. D., A. F. Militino, and A. T. Arnholt. 2015. *Probability and Statistics with r.* Chapman & Hall/CRC Press.

Wollschläger, D. 2010. *Grundlagen Der Datenanalyse Mit r.* Springer.