Einleitung

Im folgenden wird eine Untersuchung vorgestellt, deren Ziel es ist, die Anfälligkeit für Krankheiten in der Bevölkerung genauer zu klären. Zu diesem Zweck wurde in mehreren Gemeinden der Verbrauch von Taschentüchern dokumentiert. Mithilfe einer Varianzanalyse soll getestet werden, ob

1. der Taschentuchverbrauch davon beeinflusst wird, in welcher Region in Deutschland die Probanden leben: H_0^A: $\alpha_i = 0$ für alle i=1…, I vs. H_1^A: $\alpha_i \neq 0$ für mind. ein α_i,

2. der Taschentuchverbrauch davon beeinflusst wird, welche Haarfarbe die Probanden haben: H_0^B: $\beta_j = 0$ für alle j=1…, J vs. H_1^B: $\beta_j \neq 0$ für mind. ein β_j,

3. es eine Interaktion der Effekte beider Faktoren auf den Taschentuchverbrauch gibt, also ob der Effekt eines Faktors von der Ausprägung des anderen Faktors abhängt:

   H_0^{AxB}: $(\alpha\beta)_{ij} = 0$ für alle i, j vs. H_1^{AxB}: $(\alpha\beta)_{ij} \neq 0$ für mind. ein $(\alpha\beta)_{ij}$.

Die Daten

Es liegt ein Datensatz vor mit insgesamt n=66 gemittelten Beobachtungen aus den Gemeinden. Es gibt eine abhängige Variable, die den durchschnittlichen Taschentuchverbrauch pro Jahr in jeweils einer beobachteten Gemeinde darstellt. Die Hälfte der Beobachtungen wurden in Gemeinden in Norddeutschland und die andere Hälfte in Gemeinden in Süddeutschland vorgenommen. Es gibt also eine unabhängige Variable namens Region. Außerdem gibt es eine zweite unabhängige Variable, die die Haarfarbe der für die Untersuchung herangezogenen Probanden dokumentiert. Dieser Faktor hat drei Faktorstufen: dunkelhaarig, rothaarig und blond. Jede dieser Haarfarben umfasst genau ein Drittel der Stichprobe.

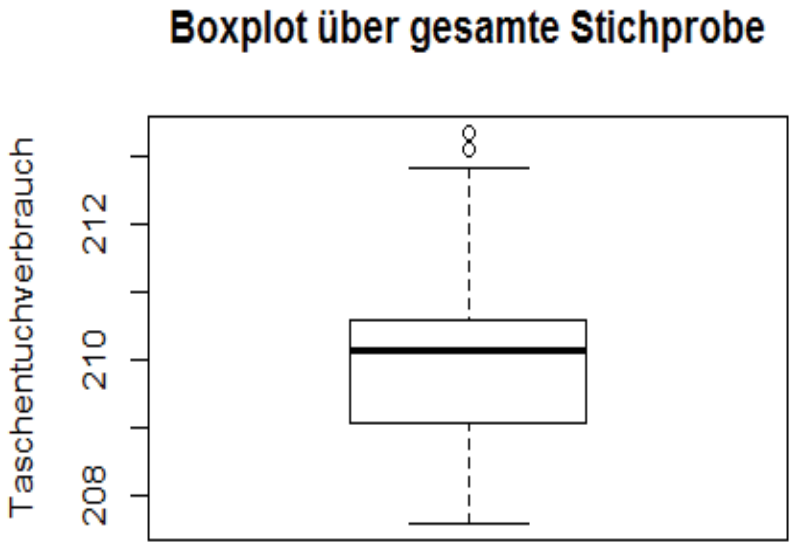
	Dunkelhaarig	Rothaarig	Blond	$\Sigma_{Haarfarbe}$
Norddeutschland	11	11	11	33
Süddeutschland	11	11	11	33
Σ_{Region}	22	22	22	66

Es handelt sich also um eine balancierte Stichprobe in einem 2x3 between-subject Design.

Der kleinste Wert, den die abhängige Variable annimmt, ist 207,6 Taschentuchpackungen durchschnittlich pro Jahr. Der größte Wert beträgt 213,4 Taschentuchpackungen durchschnittlich pro Jahr.

Weitere deskriptive Kennwerte:

Arithmetisches Mittel	210
Standardabweichung	1,31
1. Quartil	209,1
Median	210,1
3. Quartil	210,6
Interquartilsabstand	1,49



Man kann sehen, dass die Werte sich in einem relativ kleinen Bereich verteilen (Spannweite = 5,8). Auch die Streuung scheint mit einer Standardabweichung von 1,31 um den Mittelwert 210 nicht groß zu sein. Insbesondere der Interquartilsabstand macht deutlich, dass sich tatsächlich die Hälfte der Werte in dem kleinen Bereich von 1,49 Punkten verteilen.

Der Boxplot zeigt, dass sich die zentralen 50% der Werte relativ mittig in der Verteilung befinden; allerdings nicht ganz mittig. Es sieht so aus, als wäre das Zentrum etwas zum unteren Ende der Verteilung versetzt. Es ist also zu vermuten, dass sie ein wenig linkssteil ist. Eine genauere Überprüfung, ob man eine Normalverteilung annehmen kann, kommt weiter unten.

Dadurch dass der Median und das Arithmetische Mittel fast identisch sind, liegt die Vermutung

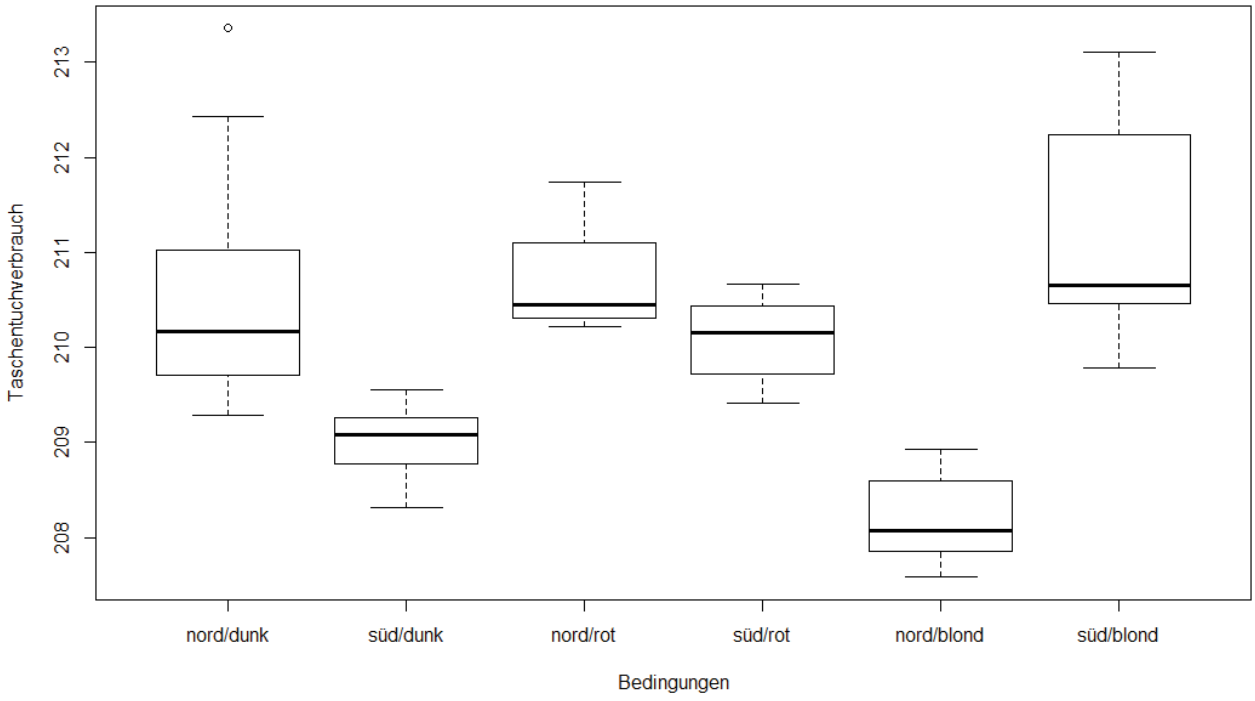
nahe, dass Letzterer nicht besonders durch Ausreißer verzerrt wird. Im Boxplot kann man zwei Punkte erkennen, die außerhalb der Whisker liegen. Eine Untersuchung auf Ausreißer (hier definiert als Werte, die mehr als den anderthalbfachen Interquartilsabstand unter dem ersten oder über dem dritten Quartil liegen) ergibt, dass es tatsächlich drei Werte sind. Alle drei befinden sich am oberen Ende der Spannweite (212,85; 213,11; 213,36).

In den einzelnen Bedingungen sehen die Kennwerte folgendermaßen aus:

Mittelwerte	Dunkelhaarig	Rothaarig	Blond
Norddeutschland	210,61	210,72	208,19
Süddeutschland	209	210,11	211,23

Standardabweichungen	Dunkelhaarig	Rothaarig	Blond
Norddeutschland	1,29	0,52	0,48
Süddeutschland	0,39	0,45	1,16

Interquartilsabstände	Dunkelhaarig	Rothaarig	Blond
Norddeutschland	1,32	0,8	0,75
Süddeutschland	0,5	0,71	1,78

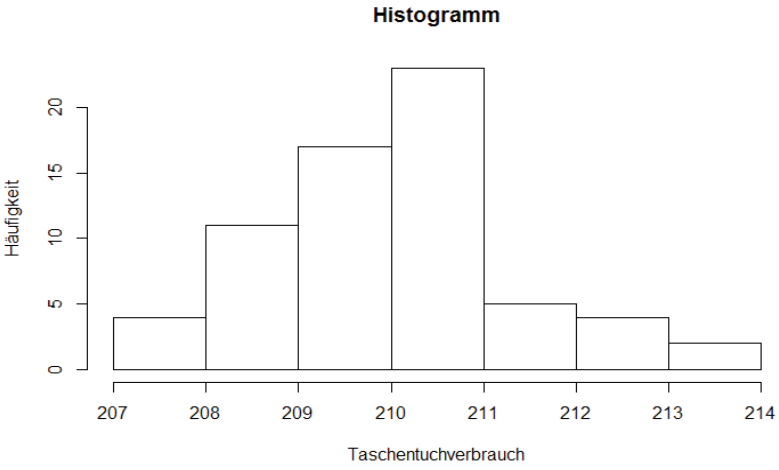


Während sich die Mittelwerte über die meisten Bedingungen nur gering unterscheiden, lassen sich bei der Streuung teilweise erhebliche Unterschiede beobachten. Sowohl in der Standardabweichung wie auch in den Interquartilsabständen sind die Werte einzelner Bedingungen mehr als doppelt so groß wie unter anderen Bedingungen. Besonders die Bedingungen Norddeutschland/dunkelhaarig und Süddeutschland/blond fallen deutlich aus dem Muster.

Inferenzstatistische Analysen

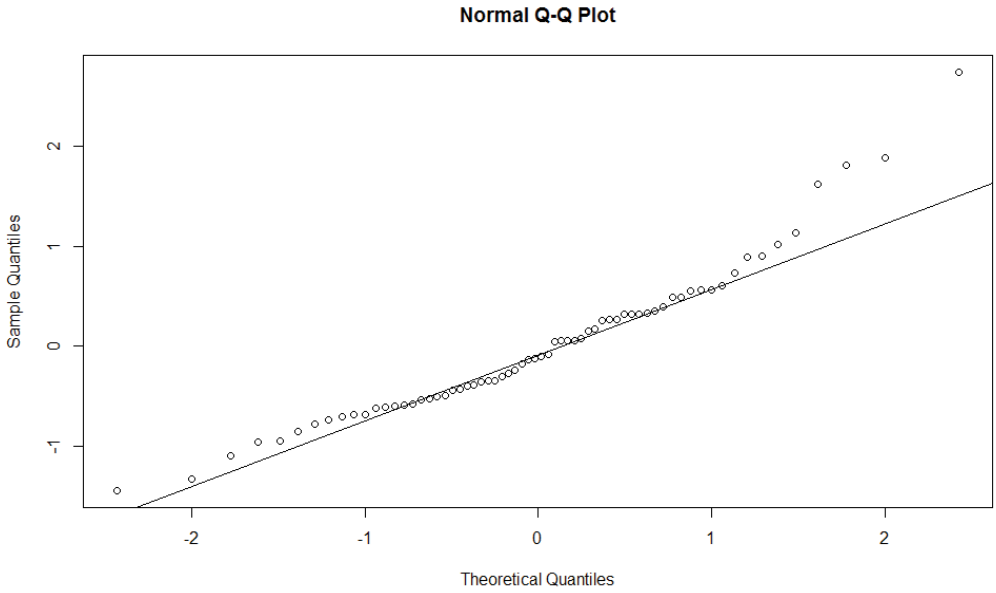
Vor der Durchführung einer Varianzanalyse müssen erst einige Annahmen überprüft werden. Zu den Voraussetzungen der Varianzanalyse gehören die Normalverteilung der Stichprobenvariablen, die Normalverteilung der Residuen, Unabhängigkeit der Beobachtungen und Homoskedastizität.

Weiter oben wurden bereits einige Vermutungen zur Art der Verteilung genannt. Der Boxplot zur abhängigen Variablen ließ darauf schließen, dass sie eventuell linkssteil sein könnte. Vielleicht verdeutlicht die Darstellung in einem Histogramm die Verteilungssituation.



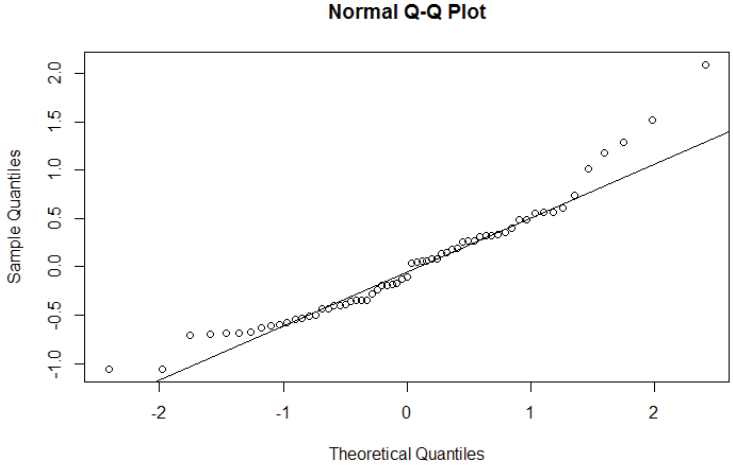
Auch in dieser Grafik ist der Fall nicht ganz klar. Die Verteilung nimmt grob die Form einer Normalverteilung an, jedoch wird auch hier deutlich, dass sie es nicht perfekt tut. Der Verdacht auf Linkssteilheit erhärtet sich. Eine weitere Möglichkeit zur Überprüfung der Normalverteilungsannahme ist ein Hypothesentest. Der Shapiro-Wilk-Test ergibt einen p-Wert von 0,1525. Das heißt, dass bei einem Signifikanzniveau von $\alpha = 0.05$ die Annahme auf Normalverteilung nicht abgelehnt werden kann.

Um zu untersuchen, ob auch die Residuen normalverteilt sind, eignet sich ein Q-Q-Plot der Residuen, in dem die beobachteten Residuen mit den unter Normalverteilung erwarteten Residuen verglichen werden:

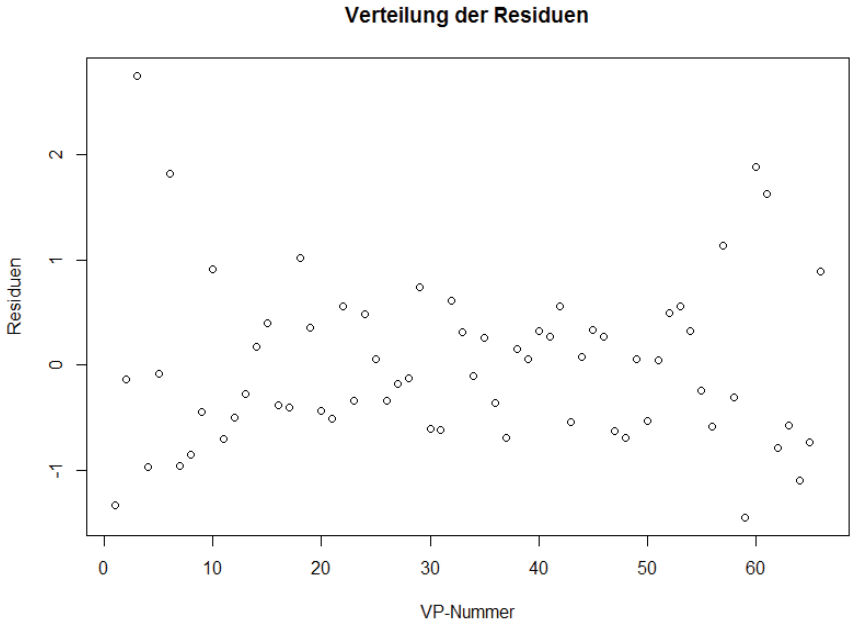


Wenn die Residuen normalverteilt sind, müssten die Quantile der beobachteten Residuen den theoretisch erwarteten Quantilen entsprechen, also auf der eingezeichneten Linie liegen. Auf die meisten Quantile trifft das sehr gut zu. Es gibt aber leider einige sehr starke Ausreißer, die die Normalverteilungsannahme gefährden. Und tatsächlich ergibt ein erneuter Shapiro-Wilk-Test für die Residuen einen p-Wert von 0,003823. Die Normalverteilungsannahme wird also abgelehnt.

Nun ist der Shapiro-Wilk-Test sehr anfällig für Ausreißer. Da die im deskriptiven Teil durchgeführte Ausreißeranalyse ergeben hat, dass es drei Ausreißerwerte gibt, wiederhole ich das Vorgehen noch einmal mit der von den Ausreißern bereinigten Stichprobe. Der p-Wert wird tatsächlich etwas größer, liegt mit 0,005141 aber immer noch deutlich unter dem Signifikanzniveau. Ein weiterer Q-Q-Plot zeigt, dass die Ausreißer in den Residuen offensichtlich nicht direkt mit den gefundenen Ausreißern zusammenhängen. (Alle folgenden Berechnungen werden also weiterhin mit der vollständigen Stichprobe durchgeführt.)



Um ein Urteil über die Unabhängigkeit der Messwerte zu machen, lohnt es sich erneut einen Blick auf die Residuen zu werfen:



In diesem Streudiagramm kann man sehr gut sehen, dass die Residuen zufällig streuen, abgesehen von einzelnen Ausreißern verteilen sich die Werte in einer Punktwolke. Obwohl die Versuchspersonen nach den Bedingungen sortiert sind, lässt sich keine bestimmte Struktur in der Verteilung der Residuen erkennen. Mit einem Hypothesentest kann diese Vermutung eventuell bestätigt werden. Der Durbin-Watson-Test berechnet einen Autokorrelationswert von -0,113. Das heißt es scheint eine negative Korrelation zu geben, die jedoch sehr schwach ist. Außerdem wird ein p-Wert von 0,968 berechnet, womit die Nullhypothese, dass es keine Autokorrelation gibt, nicht abgelehnt werden kann. Man scheint also Unabhängigkeit der Messwerte vermuten zu können.

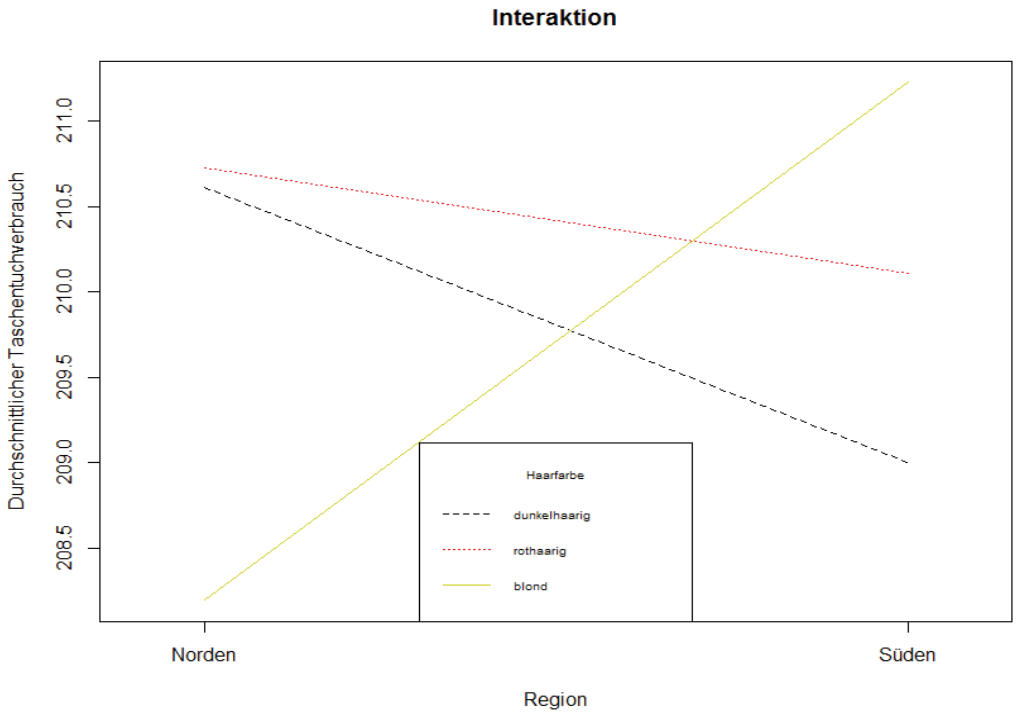
Zuletzt bleibt noch die Gleichheit der Varianzen zu prüfen. Im deskriptiven Teil wurde bereits festgestellt, dass die Streuungsparameter teilweise deutlich voneinander abweichen. Es würde sich der Bartlett-Test an dieser Stelle anbieten. Dieser ist allerdings empfindlich auf Verletzungen der Normalverteilungsannahme. Aufgrund der unsicheren Situation in Bezug auf die Normalverteilung, weiche ich auf den Fligner-Killeen-Test aus. Mit einem p-Wert von 0,1456 kommt der erstaunlicherweise zu dem Ergebnis, dass die Annahme von Varianzhomogenität nicht abgelehnt werden kann.

Die Überprüfung der Voraussetzungen für die Varianzanalyse ergibt somit, dass die Unabhängigkeit

der Messwerte gegeben ist, Varianzhomogenität (unerwarteterweise) angenommen werden kann, die Stichprobenvariable normalverteilt zu sein scheint, aber die Residuen offensichtlich nicht. Obwohl eine Annahme verletzt ist, werde ich eine Varianzanalyse rechnen, denn auch wenn die Normalverteilung eine Voraussetzung ist, zeigt sich die Varianzanalyse sehr robust gegen Verstöße gegen die Normalverteilungsannahme.

Die ANOVA ergibt einen signifikanten Haupteffekt des Faktors Haarfarbe: $F(2,60) = 4{,}987$; $p<0{,}01$ und einen signifikanten Interaktionseffekt: $F(2,60)= 50{,}939$; $p < 0{,}001$. Der Haupteffekt für den Faktor Region bleibt mit $F(1,60) = 1{,}832$; $p = 0{,}18093$ nicht signifikant.

Im Interaktionsplot wird deutlich, wie sich der Interaktionseffekt auswirkt. Während der durchschnittliche Taschentuchverbrauch bei den Faktorstufen „dunkelhaarig“ und „rothaarig“ des Faktors Haarfarbe geringer ist unter der Bedingung „Süddeutschland“ als unter der Bedingung „Norddeutschland“, ist es bei der Faktorstufe „blond“ genau umgekehrt.



Ein Tukey-Test zeigt, welche Bedingungen sich genau voneinander unterscheiden. Die Werte der Faktorstufe „rothaarig“ sind signifikant größer als die von „dunkelhaarig“ ($p=0{,}039$) und als die von „blond“ ($p=0{,}014$). „Dunkelhaarig“ und „blond“ unterscheiden sich nicht signifikant voneinander.

signifikante Bedingung	Nord/dunk vs. Süd/dunk	Nord/dunk vs. Nord/blond	Süd/dunk vs. Nord/rot	Süd/dunk vs. Süd/rot	Süd/dunk vs. Süd/blond	Nord/rot vs. Nord/blond	Süd/rot vs. Nord/blond	Süd/rot vs. Süd/blond	Nord/blond vs. Süd/blond
Richtung d. Differenz	N/d > S/d	N/d > N/b	S/d < N/r	S/d < S/r	S/d < S/b	N/r > N/b	S/r > N/b	S/r < S/b	N/b < S/b

Die Ergebnisse des Tukey-Tests zeigen, dass auch wenn der Unterschied zwischen den Faktorstufen Norddeutschland und Süddeutschland nicht signifikant ist, es über die Bedingungen gerechnet doch Unterschiede gibt. So ist die Differenz zwischen der Bedingung „Norddeutschland/dunkelhaarig“ und „Süddeutschland/dunkelhaarig“ ebenso signifikant wie die Differenz zwischen „Norddeutschland/blond“ und „Süddeutschland/blond“. Vermutlich liegt die fehlende Signifikanz des Haupteffekts Region also daran, dass die sich durch den Interaktionseffekt sehr stark voneinander unterscheidenden Werte der verschiedenen Bedingungen (wie z. B. „Norddeutschland/dunkelhaarig“ und „Norddeutschland/blond“) in den Faktorstufen gemittelt werden.

Interpretation

Die Varianzanalyse hat die Hypothese, dass es einen Effekt der Haarfarbe auf die Krankheitsanfälligkeit gibt, bestätigt. Es zeigt sich, dass die Rothaarigen durchschnittlich den größten Taschentuchverbrauch haben, bei ihnen also die größte Krankheitsanfälligkeit angenommen werden kann. Dies bestätigt alltagspsychologische Theorien über die Krankheitsanfälligkeit bei Rothaarigen. Dahingegen gibt es kein so eindeutiges Bild bei den Blonden und den Dunkelhaarigen. Interessant ist hier die signifikante Interaktion zwischen dem Effekt der Haarfarbe und des Lebensortes. Während Dunkelhaarige in süddeutschen Regionen einen geringen Taschentuchverbrauch haben, ist er in norddeutschen Regionen deutlich größer. Bei den Blonden zeigt sich das Gegenteil: In Norddeutschland haben sie den geringsten Taschentuchverbrauch, in Süddeutschland den größten. Der Grund hierfür liegt wahrscheinlich in der evolutionstechnischen Anpassung an die Umwelt. Während blonde Menschen besonders gut an die Umweltbedingungen

in nördlichen Regionen angepasst sind, sind Dunkelhaarige besonders gut an südlicheres Klima angepasst. In Regionen, für die sie keine Anpassung entwickelt haben, scheinen sie dann erheblich anfälliger für Krankheiten zu werden. Bei Rothaarigen wirkt sich der Anpassungseffekt weniger stark aus, weil die Krankheitsanfälligkeit grundsätzlich auf einem höheren Niveau liegt, als bei den anderen Haarfarben.

Interessant wäre es interessant, weiter zu untersuchen, wie sich die Regionseffekte auswirken, wenn die untersuchten Regionen weiter voneinander entfernt sind und sich stärker in ihren klimatischen Bedingungen unterscheiden. Dies ist eine Aufgabe für künftige Forschung. Für dieses Vorhaben müssen allerdings erst Wege entwickelt werden, wie die kulturellen und somit eventuell hygienischen Umstände vergleichbar gehalten werden können.

R-Code

```
#Workspace leeren
rm(list=ls())

#Einlesen des Datensatzes
dat <- read.table("C:/Texte/Uni HH/Statistik/R -
Datenanalyse/Hausarbeit/DatVa23.txt",header=T)
colnames(dat) <- c("vp", "tasch",  "reg", "haar")

#Faktorendefinition
dat$reg <- factor(dat$reg, levels = c(1,2))
dat$haar <- factor(dat$haar, levels = c(1:3))
dat$vp <- factor(dat$vp, levels = c(1:length(dat$vp)))

###Deskriptive Statistiken
summary(dat$tasch)
sd(dat$tasch)
IQR(dat$tasch)
boxplot(dat$tasch, ylab="Taschentuchverbrauch", main="Boxplot über gesamte
Stichprobe")
```

```
ausr <- subset(dat, dat$tasch > quantile(dat$tasch, .75)+1.5*IQR(dat$tasch) |
dat$tasch < quantile(dat$tasch, .25)-1.5*IQR(dat$tasch))


mean_bed <- tapply(dat$tasch, list(dat$reg, dat$haar), mean)

iqr_bed <- tapply(dat$tasch, list(dat$reg, dat$haar), IQR)

sd_bed <- tapply(dat$tasch, list(dat$reg, dat$haar), sd)

boxplot(dat$tasch~dat$reg*dat$haar, xlab="Bedingungen",
ylab="Taschentuchverbrauch", main="Boxplots über alle Bedingungen", xaxt="n")

axis(side=1,at=c(1:6), labels=c("nord/dunk", "süd/dunk", "nord/rot", "süd/rot",
"nord/blond", "süd/blond"))


###Interferenzstatistik


##Überprüfen der Voraussetzungen


library("car")

linm <- lm(dat$tasch ~ dat$reg*dat$haar)


#Test auf Normalverteilung (Wenn p>0.05 Annahme auf NV nicht abgelehnt)

shapiro.test(dat$tasch)

hist(dat$tasch, xlab="Taschentuchverbrauch", ylab="Häufigkeit",
main="Histogramm")


#Test auf Normalverteilung der Residuen

shapiro.test(linm$res)

qqnorm(linm$res)

qqline(linm$res)


dat2 <- subset(dat, tasch < quantile(dat$tasch, .75)+1.5*IQR(dat$tasch) & tasch
> quantile(dat$tasch, .25)-1.5*IQR(dat$tasch))

linm2 <- lm(dat2$tasch ~ dat2$reg*dat2$haar)

shapiro.test(linm2$res)

qqnorm(linm2$res)

qqline(linm2$res)


#Test auf Unabhängigkeit (Wenn p>0.05 Annahme auf keine Autokorrelation nicht
abgelehnt)

durbinWatsonTest(linm)

plot(linm$res, xlab="VP-Nummer", ylab="Residuen", main="Verteilung der
Residuen")
```

```
#Test auf Varianzhomogenität (Wenn p>0.05 Annahme auf Homoskedastizität nicht
abgelehnt)
fligner.test(split(dat$tasch,list(dat$haar,dat$reg)))


##ANOVA
tasch.aov <- aov(dat$tasch ~ dat$reg * dat$haar)
summary(tasch.aov)
TukeyHSD(tasch.aov)


#Interaktionsplot
interaction.plot(dat$reg, dat$haar, dat$tasch,main="Interaktion", xlab="Region",
ylab="Durchschnittlicher Taschentuchverbrauch", lty=c(2,3,1), col=c("black",
"red", "yellow3"), legend=F, xaxt="n")
axis(side=1,at=c(1,2), labels=c("Norden", "Süden"))
legend(x="bottom", legend=c("dunkelhaarig", "rothaarig", "blond"),
col=c("black", "red", "yellow3"), lty=c(2,3,1), cex=0.6, title="Haarfarbe")


interaction.plot(dat$haar, dat$reg, dat$tasch, main="Interaktion",
xlab="Haarfarbe", ylab="Durchschnittlicher Taschentuchverbrauch", lty=c(2,3),
col=c("black", "red"), legend=F, xaxt="n")
axis(side=1,at=c(1,2,3), labels=c("Dunkel", "Rot", "Blond"))
legend(x="bottom", legend=c("Nord", "Süd"), col=c("black", "red"), lty=c(2,3),
cex=0.6, title="Region")
```