Prüfung aus

EDV-geschützte Statistik

107.378

O. Univ. Prof. Dr. R. Dutter

Claus Reichel*
9300654

14. Mai 1997

*http://stud1.tuwien.ac.at/~e9300654

1 Einleitung

Zu untersuchen sind Daten über drei historische Gattungen der grauen Känguruhs. Es wurden an insgesamt 148 Tieren jeweils 18 Messungen an deren Schädeln vorgenommen. Die Daten wurden bereits von Fehlmessungen befreit. Das Merkmal der Kammbreite sollte im Speziellen untersucht werden.

2 Vorbereitungen

2.1 Verwendete Objekte in S-PLUS

Die zu analysierende Variable *crest*.

```
> crest_kangaroos[,16]
```

Dataframe der gesamten Beobachtungen.

```
> kangaroos.df_data.frame(kangaroos)
```

Speciesvektor (Gruppierungsvektor) *species*.

```
> species_kangaroos[,3]
```

Typvektor für die Varianzanalyse.

```
> crest.typ_aovvector(species,"species")
```

Dataframe für die Varianzanalyse.

```
> crest.df_data.frame(crest.typ,crest)
```

Dataframe für die zufälligen Effekte.

```
> crest.df.rnd_crest.df
```

Zweite Variable für die Regressionsanalyse *nasal length*.

```
> nasall_kangaroos[,8]
```

Dataframe ohne *crest* erstellen.

```
> kangaroos.no.crest.df_data.frame(kangaroos[,-c(16)])
```

Dataframe für die Faktoranalyse.

```
> kangaroos.factor.df_data.frame(kangaroos[,10:15])
```

2.2 Funktionen in S-PLUS

2.2.1 Konfidenzintervall für die Standardabweichung

Definition einer neuen Funktion.

```
> ki.si_function(){}
```

Editiern der soeben angelegten Funktion.

```
> ki.si_vi(ki.si)
```

Die Funktion stellt die Implementierung der folgenden Formel für ein

$\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2};n-1}}$

Konfidenzintervall dar. Siehe auch [Josef Bleymüller, 1996] Kapitel 14 Abschnitt 4.

```
function(vector, niveau = 0.95, dig = 3)
{
    options(digits = dig)
    df <- length(vector) - 1
    alpha2 <- (1 - niveau)/2
    chi.lo <- qchisq(alpha2, df)
    chi.hi <- qchisq(1 - alpha2, df)
    s <- var(vector)
    sigma.lo <- sqrt((df * s)/chi.lo)
    sigma.hi <- sqrt((df * s)/chi.hi)
    cat("Confidence Interval for Standard Deviation:\n")
    cat(" niveau: ", niveau, "\n")
    cat(" [", format(sigma.lo), " / ", format(sigma.hi), "]\n\n")
}
```

2.2.2 Aufspalten eines Datenvektors

Diese Funktion dient zum aufspalten eines Datenvektors anhand eines Gruppierungsvektors in eine Matrix. Die beiden Vektoren müssen die gleiche Länge aufweisen. Die Werte im Gruppierungsvektor müssen in durchgehender Reihenfolge vorhanden sein, um Spalten ohne Werte zu vermeiden. Da die Funktion F.test nur mit Vektoren gleicher Länge arbeitet werden die restlichen Elemente mit Not Available (NA) aufgefüllt.
Eine neue Funktion anlegen und editieren.

```
> splitvector_function(){}
> splitvector_vi(splitvector)
function(vector, condvector)
{
    lo <- min(condvector)
    hi <- max(condvector)
    col <- hi + 1 - lo
    m <- matrix(0, length(vector), col)
    for(i in 1:length(vector)) {
        for(j in 1:col) {
            if((j == (condvector[i] + 1 - lo)) == T) {
                m[i, j] <- vector[i]
            }
            else {
                m[i, j] <- NA
            }
        }
    }
    as.matrix(m)
}
```

Ein kurzes Beispiel zur Veranschaulichung. Die beiden folgenden Vektoren

```
Datenvektor    Gruppierungsvektor
       141      0
       116      2
       149      1
       128      1
       103      0
        86      0
       107      2
```

liefern als Ergebnis die folgende Matrix.

```
 141    NA    NA
  NA    NA   116
  NA   149    NA
  NA   128    NA
 103    NA    NA
  86    NA    NA
  NA    NA   107
```

Auf die so gruppierten Daten können nun leicht diverse Testverfahren angewendet werden.

2.2.3 Erstellen eines Gruppenvektors

Diese Funktion dient zum Erstellen eines Gruppenvektors für die Varianzanalyse.

```
> aovvector_function(){}
> aovvector_vi(aovvector)

function(vector, name)
{
    v <- NULL
    for(i in 1:length(vector)) {
        v <- cbind(v, paste(name, vector[i], sep = ""))
    }
    cbind(v[1:length(v)])
}
```

2.3 BMDP

Datenfile aus S-PLUS Objekt erzeugen.

```
> write(crest,file= "crest.bmdp" ,ncolumns=1)
```

Kommandfile für BMDP [Dutter, 1996].

```
/BMDP2D
/PROBLEM        TITLE IS 'VO'.
/INPUT          VARIABLES ARE 1.
                FORMAT IS FREE.
                FILE='crest.bmdp'.
/VARIABLE       NAMES ARE Crest.
                USE IS 1.
/PRINT          ESTIMATE.
                STEM.
                WSTAT.
                LINE=80.
                PAGE=60.
/END
```

3 Lösung der Aufgaben

3.1 Aufgabe 1

3.1.1 Aufgabenstellung

Untersuchen Sie die Variable *crest width* mittels univariater Statistiken und Graphiken.

3.1.2 S-PLUS

Erstellen eines Boxplots [siehe Abbildung 1, Seite 6].

```
> boxplot(crest)
```

Erstellen eines Histogramms [siehe Abbildung 2, Seite 6].

```
> hist(crest)
```

Erstellen eines QQ-Plots [siehe Abbildung 3, Seite 7].

```
> qqnorm(scale(crest))
> abline(0,1)
```

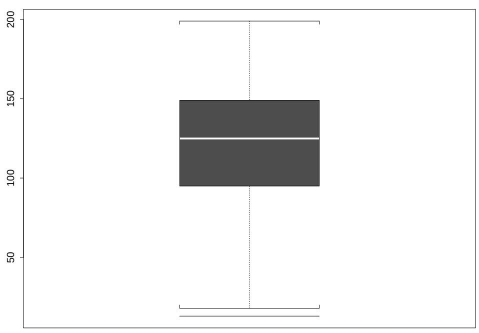

Abbildung 1: Boxplot der Variablen *crest width*

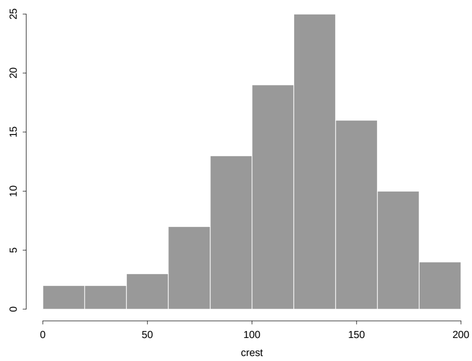

Abbildung 2: Histogramm der Variablen *crest width*

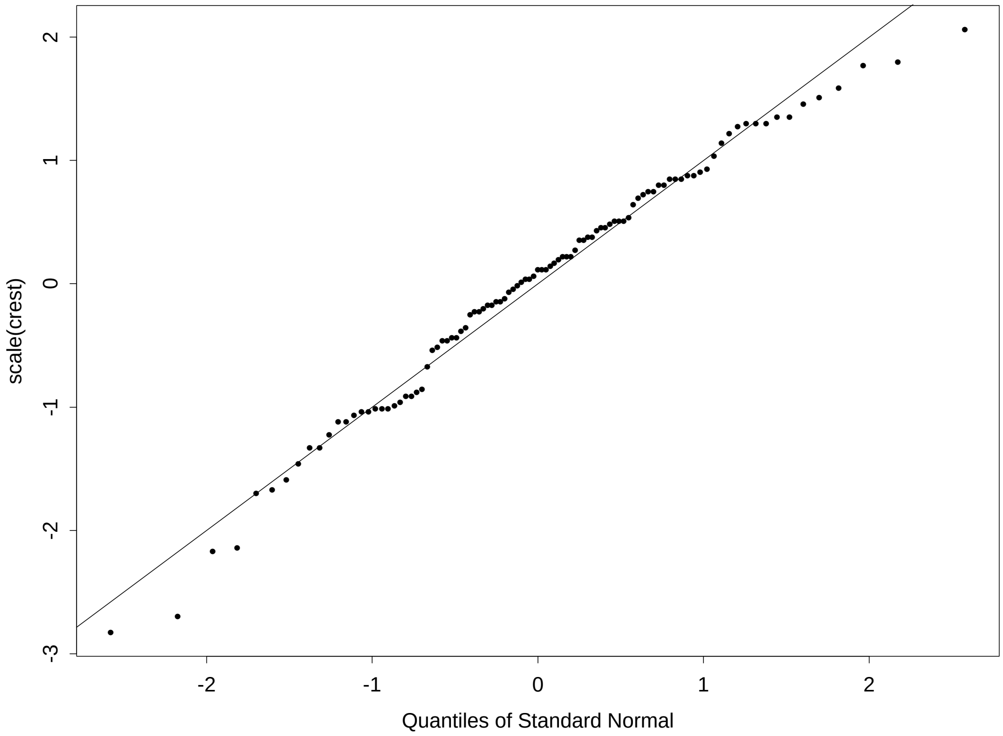

Abbildung 3: QQPlot der standardisierten Variablen *crest width*

3.1.3 BMDP

Aufruf von BMDP.

```
> 2d command
```

```
 VARIABLES TO BE USED
      1 Crest

 DATA FORMAT:  FREE

 THE LONGEST RECORD MAY HAVE UP TO   80 CHARACTERS.

 NUMBER OF CASES READ. . . . . . . . . . . . . .     101

   ************
   * Crest    *
   ************                                   VALUE    ZSCORE  CASE #
                                          MAX   199.0000000    2.061      83
 VARIABLE NUMBER . . . . . .        1     MIN    13.0000000   -2.826      47
```

```
NUMBER OF DISTINCT VALUES .       72
NUMBER OF VALUES COUNTED. .      101                  VALUE    VALUE/S.E.
NUMBER OF VALUES NOT COUNTED       0     SKEWNESS      -.43    -1.748
                                         KURTOSIS      -.09     -.181

                      EACH 'H' REPRESENTS     2 COUNTS
           H          EACH '-' REPRESENTS  10.0000 UNITS
        H  HHH
        H  HHH H      L =  10.0000
        H HHHH H      U =  210.000
        H HHHHHHH
     HHHH HHHHHHHHH
 H H HHHHHHHHHHHHHHH
L--------------------U

                  ESTIMATE      ST.ERROR                         ESTIMATE
MEAN           120.5544586     3.7873809          ST.DEV.       38.0627060
MEDIAN         125.0000000     4.6188045          VARIANCE    1448.7695313
MODE            NOT UNIQUE                        RANGE        186.0000000
                                                  (Q3-Q1)/2     28.7500000
TRIM(.15)      122.1845827
HAMPEL         122.3139013            LOWER 95% C.L. OF MEAN   113.0404053
BWEIGHT        122.5760498            UPPER 95% C.L. OF MEAN   128.0685120
                                                         Q1     91.5000000
TEST OF NORMALITY                                        Q3    149.0000000
W STATISTIC          .9726                               S-     82.4917526
SIGNIFICANCE LEVEL   .2057                               S+    158.6171570

              EACH '.' BELOW =       2.5000
                             S   Q                      Q   S
 M                           -   1          MHM         3   +            M
 I..........................................EAE..........................A
 N                                          AMD                          X
                                            NPI
                      PERCENTS                        PERCENTS
     VALUE    COUNT  CELL   CUM       VALUE    COUNT  CELL   CUM
        13.       1   1.0   1.0         122.       2   2.0  48.5
        18.       1   1.0   2.0         123.       1   1.0  49.5
        38.       1   1.0   3.0         125.       3   3.0  52.5
        39.       1   1.0   4.0         126.       1   1.0  53.5
        56.       1   1.0   5.0         127.       1   1.0  54.5
        57.       1   1.0   5.9         128.       1   1.0  55.4
        60.       1   1.0   6.9         129.       3   3.0  58.4
        65.       1   1.0   7.9         131.       1   1.0  59.4
        70.       2   2.0   9.9         134.       2   2.0  61.4
        74.       1   1.0  10.9         135.       2   2.0  63.4
        78.       2   2.0  12.9         137.       1   1.0  64.4
        80.       1   1.0  13.9         138.       2   2.0  66.3
        81.       2   2.0  15.8         139.       1   1.0  67.3
```

82.	3	3.0	18.8
83.	1	1.0	19.8
84.	1	1.0	20.8
86.	2	2.0	22.8
87.	1	1.0	23.8
88.	1	1.0	24.8
95.	1	1.0	25.7
100.	1	1.0	26.7
101.	1	1.0	27.7
103.	2	2.0	29.7
104.	2	2.0	31.7
106.	1	1.0	32.7
107.	1	1.0	33.7
111.	1	1.0	34.7
112.	2	2.0	36.6
113.	1	1.0	37.6
114.	2	2.0	39.6
115.	2	2.0	41.6
116.	1	1.0	42.6
118.	1	1.0	43.6
119.	1	1.0	44.6
120.	1	1.0	45.5
121.	1	1.0	46.5
140.	3	3.0	70.3
141.	1	1.0	71.3
145.	1	1.0	72.3
147.	1	1.0	73.3
148.	1	1.0	74.3
149.	2	2.0	76.2
151.	2	2.0	78.2
153.	3	3.0	81.2
154.	2	2.0	83.2
155.	1	1.0	84.2
156.	1	1.0	85.1
160.	1	1.0	86.1
164.	1	1.0	87.1
167.	1	1.0	88.1
169.	1	1.0	89.1
170.	3	3.0	92.1
172.	2	2.0	94.1
176.	1	1.0	95.0
178.	1	1.0	96.0
181.	1	1.0	97.0
188.	1	1.0	98.0
189.	1	1.0	99.0
199.	1	1.0	100.0

```
              *
DEPTH    STEM * LEAVES
              *
   2        1 * 38
   2        2 *
   4        3 * 89
   4        4 *
   6        5 * 67
   8        6 * 05
  13        7 E 00488
  25        8 * 011222346678
  26        9 Q 5
  34       10 * 01334467
  45       11 * 12234455689
+ 14       12 M 01223555678999
  42       13 * 144557889
  33       14 Q 000157899
  24       15 * 113334456
  15       16 E 0479
  11       17 * 0002268
   4       18 * 189
   1       19 * 9
              *
DEPTH    STEM * LEAVES
              *
```

3.1.4 Aussagen

Die gemessenen Werte liegen in einem Bereich von 13 und 199 und ergeben somit eine Range von 186. Die beiden Zscores von -2.826 und 2.061 für das Minimum/Maximum in Verbindung mit dem Histogramm lassen auf eine Ausreisergruppe auf der linken Seite schließen. Bei einer Normalverteilung würden im Bereich von ± 1.96 95 % und in im Bereich von $\pm 2{,}575$ 99 % der Daten liegen.
Der Wert der Skewness von -0.43 zeigt, daß die Daten linksschief verteilt sind. Bei einem Wert von 0 wären die Daten symmetrisch verteilt.
Der Wert der Kurtosis von -0.09 gibt eine leichte Flachgipfligkeit im Vergleich zur Normalverteilung an.
Die Form eines Histogramms hängt stark von der Wahl der Klassenbreite ab, wie man leicht sieht, wenn man die Histogramme von BMDP [siehe Seite 8] und S-PLUS [siehe Abbildung 2, Seite 6] vergleicht.
Der Mean, hier 120.55, stellt das arithmetische Mittel ($\frac{1}{n}\sum_{i=1}^{n} x_i$) der beobachteten Wert dar. Der Median, hier 125, stellt den mittleren Wert dar ($\tilde{x} = x_{\frac{n+1}{2}}$ wenn n ungerade und $\tilde{x} = \frac{x_{\frac{n}{2}} + x_{\frac{n+2}{2}}}{2}$ wenn n gerade). Der Mode, hier nicht eindeutig, würde den häufigsten Wert angeben. Für den trimmed Mean wurden 15 % der Daten links und rechts abgeschnitten, um eventuelle Ausreißer zu eliminieren.
Die Varianz der Variablen beträgt 1448.77 woraus sich σ mit 38.06 ergibt. BMDP liefert desweiteren noch einen kleinen Boxplot, an dem diverse Werte gekennzeichnet sind. Den nächsten Punkt stellt ein detailierter Überblick über die Daten dar. Zum Abschluß liefert BMDP noch eine Stamm-und-Blatt-Darstellung.

3.2 Aufgabe 2

3.2.1 Aufgabenstellung

Testen Sie auf dem 5 %-Signifikanzniveau auf Normalverteilung. Transformieren Sie nötigenfalls die Daten (z.B. logarithmieren).

3.2.2 S-PLUS

χ^2-Anpassungstest durchführen.

```
> chianp(crest)

Chisquare-Test of Normal Distribution
*************************************

          interval            observed  predicted expected  T
-inf      -         76.8      11        0.1252    12.64     0.2132
76.8      -         94.9      14        0.125     12.62     0.1501
94.9      -         108       9         0.1206    12.18     0.8309
108       -         121       13        0.1339    13.52     0.02036
121       -         133       13        0.1235    12.47     0.02234
133       -         146       13        0.1199    12.11     0.06479
146       -         164       15        0.1251    12.63     0.4445
164       -         inf       13        0.1268    12.81     0.00277
sums                8         101       1         101       1.749

  T:  1.749    df:  5    p-value:  0.8827

Chi:  11.07   alpha: 0.05
```

3.2.3 Aussagen

Die Voraussetzungen, damit dieser Test überhaupt durchgeführt werden kann sind, daß die erwarteten absoluten Häufigkeiten nicht zu klein sind. Als Faustregel gilt, daß sie ≥ 5 sind, was in diesem Fall erfüllt ist. Der p-Wert mit 0.8827 sagt aus, daß
H_0 : Die Daten entsprechen einer Normalverteilung
nicht verworfen werden kann. Auch im Output von BMDP findet sich ein Test auf Normalverteilung. Der Wert der W-Statistik mit 0.9726 gibt an, wie gut die Daten an die Normalverteilung angepaßt sind. Der Signifikanzlevel entspricht dem p-Wert, und die
H_0 : Die Daten entsprechen einer Normalverteilung
kann ebenso nicht verworfen werden. Auch anhand des QQ-Plots [siehe Abbildung 3, Seite 7] sieht man, daß die Daten approximativ normalverteilt sind.

3.3 Aufgabe 3

3.3.1 Aufgabenstellung

Ermitteln Sie ein 95 % Konfidenzintervall für Mittelwert und Standardabweichung.

3.3.2 S-PLUS

Das Konfidenzintervall für die Standardabweichung ermitteln.

```
> ki.si(crest)
```

Als Ergebnis erhält man [33.4 / 44.2].

3.3.3 BMDP

Für den Mittelwert lieferte BMDP das Ergebnis von [113.04 / 128.07].

3.3.4 Aussagen

Das heiß mit 95 %iger Wahrscheinlichkeit liegt der wahre Mittelwert bzw. die unbekannte Standardabweichung innerhalb des angegebenen Intervalls.

3.4 Aufgabe 4

3.4.1 Aufgabenstellung

Testen Sie die Hypothese $\mu > \overline{x}$.

3.4.2 S-PLUS

```
> t.test(crest,alternative="greater" ,mu=mean(crest))
         One-sample t-Test

data:  crest
t = 0, df = 100, p-value = 0.5
alternative hypothesis: true mean is greater than 120.554455445545
95 percent confidence interval:
 114  NA
sample estimates:
 mean of x
       121
```

3.4.3 Aussagen

Das heißt
$H_0 : \mu = \overline{x}$
kann mit einem p-Wert von 0.5 nicht verworfen werden. Die Alternativhypotese ist in diesem Fall
$H_A : \mu > \overline{x}$. Weiters ist das wahre Mittel der unbekannte Grundgesamtheit mit 95 %iger Wahrscheinlichkeit größer als 114 ist.

3.5 Aufgabe 5

3.5.1 Aufgabenstellung

Testen Sie auf Gleichheit der Mittelwerte der Variablen *crest width* bezüglich der Gruppierung durch die Variable *species* paarweise. Welche Voraussetzungen müssen gelten? Sind die Voraussetzungen erfüllt?

3.5.2 S-PLUS

Die gegebenen Daten gruppieren.

```
> crest.grouped_splitvector(crest, species)
```

Auf Gleichheit der Varianzen testen.

```
> F.test(crest.grouped[,1],crest.grouped[,2])

$f:
[1] 0.4130893
$df:
[1] 36 27
$prob:
[1] 0.9931567

> F.test(crest.grouped[,1],crest.grouped[,3])

$f:
[1] 0.7425409
$df:
[1] 36 35
$prob:
[1] 0.8108641

> F.test(crest.grouped[,2],crest.grouped[,3])

$f:
[1] 1.797531
$df:
[1] 27 35
$prob:
[1] 0.05151813
```

Auf Normalverteilung testen.

```
> chianp(crest.grouped[,1])

Chisquare-Test of Normal Distribution
************************************

          interval            observed  predicted expected  T
-inf      -         76.9      4         0.1428    5.283     0.3115
76.9      -         90.9      8         0.1428    5.285     1.395
90.9      -         102       3         0.1475    5.457     1.106
102       -         112       5         0.1422    5.261     0.01293
112       -         122       5         0.1329    4.918     0.001354
122       -         136       7         0.1449    5.361     0.5011
136       -         inf       5         0.1469    5.435     0.03489
sums                7         37        1         37        3.363

  T:  3.363    df:  4    p-value:  0.499

Chi:  9.488   alpha: 0.05

> chianp(crest.grouped[,2])

Chisquare-Test of Normal Distribution
************************************

          interval            observed  predicted expected  T
-inf      -         67.5      5         0.1665    4.662     0.02452
67.5      -         90.8      5         0.1666    4.664     0.02426
90.8      -         110       0         0.1712    4.794     4.794
110       -         128       6         0.1604    4.491     0.5068
128       -         152       7         0.1713    4.796     1.013
152       -         inf       5         0.164     4.593     0.03602
sums                6         28        1         28        6.398

  T:  6.398    df:  3    p-value:  0.09376

Chi:  7.815   alpha: 0.05

> chianp(crest.grouped[,3])
```

```
Chisquare-Test of Normal Distribution
************************************

         interval           observed  predicted expected  T
-inf     -        109       4         0.1443    5.197     0.2755
109      -        125       6         0.141     5.077     0.1678
125      -        138       4         0.1489    5.36      0.345
138      -        149       4         0.1349    4.855     0.1505
149      -        162       7         0.1484    5.343     0.5138
162      -        178       7         0.1401    5.042     0.7602
178      -        inf       4         0.1424    5.127     0.2476
sums                7       36        1         36        2.46

  T:  2.46    df:  4    p-value:  0.6517

Chi:  9.488   alpha: 0.05
```

Test auf Gleichheit der Mittelwerte durchführen.
$H_0 : \overline{x} = \overline{y}$
$H_A : \overline{x} \neq \overline{y}$

```
> t.test(crest.grouped[,1],crest.grouped[,2])

         Standard Two-Sample t-Test

data:  crest.grouped[, 1] and crest.grouped[, 2]
t = -0.3195, df = 63, p-value = 0.7504
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -20.55172  14.88569
sample estimates:
 mean of x mean of y
  106.7027  109.5357

> t.test(crest.grouped[,1],crest.grouped[,3])

         Standard Two-Sample t-Test

data:  crest.grouped[, 1] and crest.grouped[, 3]
t = -5.1854, df = 71, p-value = 0
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -50.75474 -22.56207
sample estimates:
 mean of x mean of y
  106.7027  143.3611
```

```
> t.test(crest.grouped[,2],crest.grouped[,3])

         Standard Two-Sample t-Test

data:  crest.grouped[, 2] and crest.grouped[, 3]
t = -3.571, df = 62, p-value = 7e-04
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -52.76018 -14.89061
sample estimates:
 mean of x mean of y
  109.5357  143.3611
```

3.5.3 Aussagen

Da die Vorraussetzungen Gleichheit der Varianzen sowie Normalverteilung für alle drei Gruppen gelten kann der Test auf Gleichheit der Mittelwerte durchgeführt werden. Die
$H_0 : \mu_x = \mu_y$
kann nur für die ersten beiden Gruppen angenommen werden, alle anderen Kombination sind zu verwerfen, wie man auch leicht mit einem Blick auf die Boxplotdarstellung [siehe Abbildung 4, Seite 17] der einzelnen Gruppen ersehen kann.

3.6 Aufgabe 6

3.6.1 Aufgabenstellung

Vergleichen Sie die Mittelwerte Ihrer Variablen bezüglich obiger Gruppeneinteilung mittels Varianzanalyse. Welche Voraussetzungen müssen gelten? Sind die Voraussetzungen erfüllt? Vergleichen Sie die Ergebnisse mit denen im vorhergehenden Punkt und erklären Sie die Unterschiede.

3.6.2 S-PLUS

Durchführen der Varianzanalyse

```
> crest.aov_aov(crest~crest.typ,crest.df)
> summary(crest.aov)
          Df Sum of Sq  Mean Sq  F Value        Pr(F)
crest.typ  2     29224 14611.98 12.38164 1.605915e-05
Residuals 98    115653  1180.13
```

Boxplot der über die Variablen *species* gruppierten Variablen *crest*.

```
> plot.factor(crest.df)
```

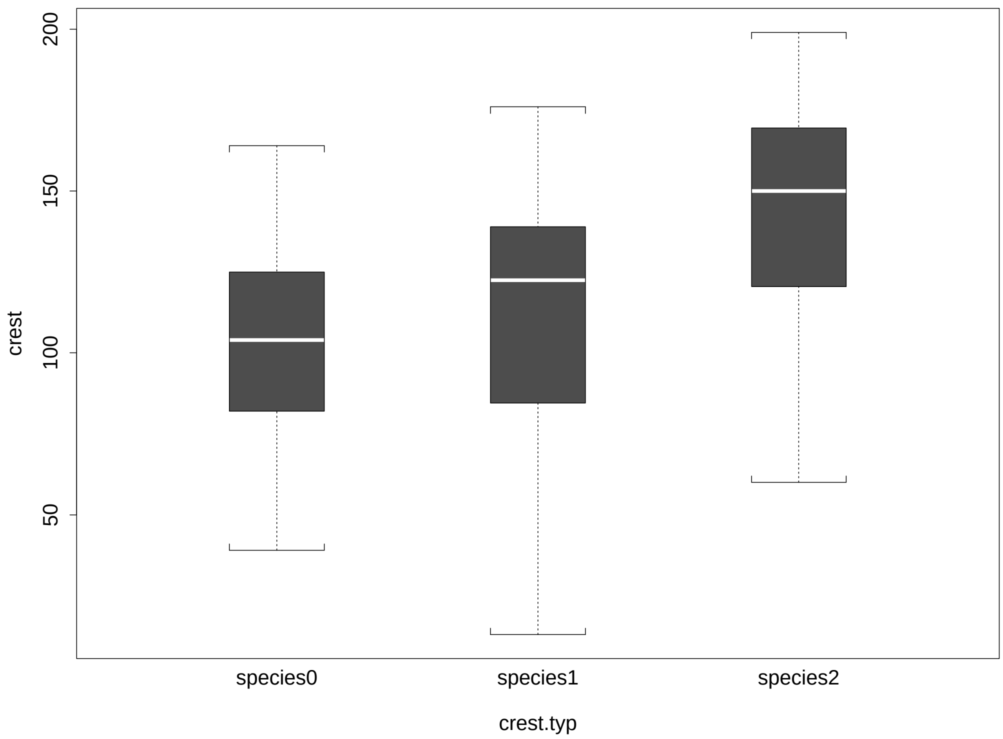

Abbildung 4: Boxplot der über *species* gruppierten Variablen *crest*

3.6.3 Aussagen

Die Daten der einzelnen Gruppen müssen normalverteilt mit Mittel μ_i und der endlichen Varianz σ_i sein. Desweiteren muß Homoskedastizität ($\sigma_1 = \ldots = \sigma_n$) gelten. Diese Voraussetzungen wurden bereits bei der vorigen Aufgabe getestet und sind erfüllt.
In diesem Fall handelt es sich um ein Modell I (feste Effekte), da der Faktor *species* fix vorgegeben ist. Bei der Varianzanalyse wird getestet, ob der Faktor *species* einen Einfluß auf die abhängige Größe *crest* hat. In diesem Fall kann die
H_0 : *species* hat keinen signifikanten Einfluß auf *crest*
nicht angenommen werden. Daraus folgt, daß *species* einen Einfluß auf die

Mittelwerte der Variablen *crest* hat. Die Bestätigung des mittels Varianzanalyse ermittelten Ergenisses erhält man bei Betrachtung der Boxplots [siehe Abbildung 4, Seite 17].

3.7 Aufgabe 7

3.7.1 Aufgabenstellung

Wie vorher, nur fassen Sie die Effekte der Gruppierung von *species* als zufällig auf und lassen Sie die graphischen Untersuchungen weg.

3.7.2 S-PLUS

Zufällige Effekte festlegen.

```
> crest.df.rnd_crest.df
```

Durchführen der Varianzanalyse

```
> crest.vc_varcomp(crest~crest.typ,crest.df.rnd)
> summary(crest.vc)
Call:
varcomp(formula = crest ~ crest.typ, data = crest.df.rnd)
Variance Estimates:
          Variance
crest.typ  423.3528
Residuals 1165.7519
Method:  minque0

Coefficients:
 (Intercept)
    119.9269
Approximate Covariance Matrix of Coefficients:
[1] 152.8272
```

3.7.3 Aussagen

In diesem Fall, übt der betrachtete Faktor einen signifikanten Einfluß auf *crest* aus. S-PLUS liefert in diesem Fall die erwartungstreuen Schätzer für $Var(\widehat{\sigma}_e^2)$, $Var(\widehat{\sigma}_A^2)$ sowie $Cov(\widehat{\sigma}_A^2, \widehat{\sigma}_e^2)$ [Dutter, 1996].

3.8 Aufgabe 8

3.8.1 Aufgabenstellung

Suchen Sie sich eine zweite Variable, die einen linearen Zusammenhang mit Ihrer vermuten läßt. Führen Sie eine Regressionsanalyse graphisch und numerisch durch. Verwenden Sie zusätzlich einen robusten Schätzer Ihrer Wahl. Gelangt man durch Einsatz des robusten Schätzers zu einer Verbesserung? Erklären Sie die Einflüsse der Gewichtsfunktionen und wann robuste Schätzer zum Einsatz kommen.

3.8.2 S-PLUS

Korrelation der Variablen *crest* mit den anderen Variablen.

```
> cor(kangaroos,crest)
                         [,1]
           number -0.5130400
              sex  0.2685671
          species  0.4100899
   basilar length -0.5823773
    occipitonasal -0.6647996
        palatilar -0.6344833
           palate -0.5452242
     nasal length -0.7122812
      nasal width -0.5578438
        squamosal -0.3440413
   inter lacrymal -0.4711398
        zygometic -0.3205799
     post orbital  0.1282284
          rostral -0.5280582
  supra-occipital -0.5341114
            crest  1.0000000
incisive foramina -0.2018792
  mandible length -0.5804219
   mandible width -0.2016026
   mandible depth -0.3037763
              arh -0.4110995
```

Numerischen Regressionsanalyse (*crest* hängt von *nasall* ab).

```
> summary(lm(crest~nasall))

Call: lm(formula = crest ~ nasall)
Residuals:
    Min     1Q  Median    3Q   Max
 -86.55 -17.85 -0.2831 15.68 78.28

Coefficients:
                Value Std. Error   t value  Pr(>|t|)
(Intercept)  337.2259   21.6244    15.5947    0.0000
     nasall   -0.3245    0.0321   -10.0971    0.0000

Residual standard error: 26.85 on 99 degrees of freedom
Multiple R-Squared: 0.5073
F-statistic: 102 on 1 and 99 degrees of freedom, the p-value is 1.11e-16

Correlation of Coefficients:
       (Intercept)
nasall -0.9923
```

L1-Schätzung

```
> l1fit(nasall, crest)$coef
 Intercept          X
  341.0255 -0.3316326
```

LS-Schätzung

```
> lsfit(nasall, crest)$coef
 Intercept          X
  337.2259 -0.3245076
```

Robuste Schätzung (Bisquare)

```
> rreg(nasall,crest,method=wt.bisquare)$coefficients
 (Intercept)          x
    337.2081 -0.3246565
```

Robuste Schätzung (Hampel)

```
> rreg(nasall,crest,method=wt.hampel)$coefficients
 (Intercept)          x
    336.1291 -0.3230822
```

Graphische Regressionsanalyse

```
> plot(nasall,crest)
> abline(l1fit(nasall, crest)$coef)
> abline(rreg(nasall,crest,method=wt.hampel)$coefficients)
> text(880,45,"l1fit")
> text(880,58,"Hampel")
```

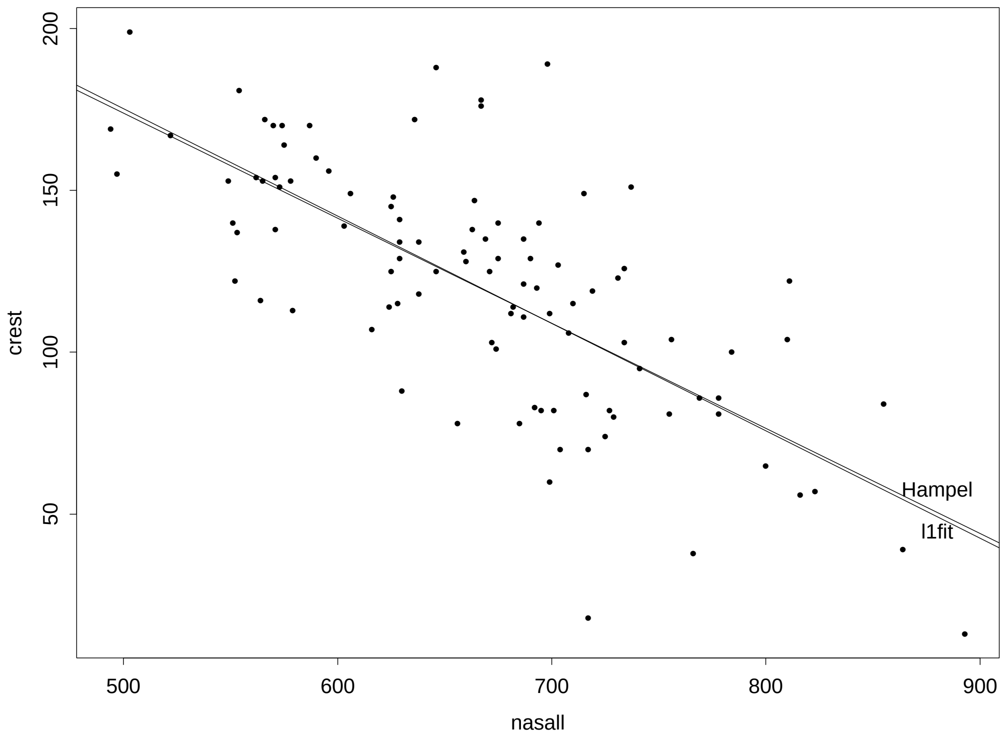

Abbildung 5: Graphische Regressionsanalyse

3.8.3 Aussagen

Die Variable *nasal length* mit einem Korrelationskoeffizienten von -0.71 wurde als Variable ausgewählt. Die Variable *crest* ist negativ von *nasal length* abhängig. Wie man bereits bei der numerischen Lösung sieht liegen alle vier ausgewählten Methoden sehr knapp beieinander. Aus diesem Grund wurden in den Plot [siehe Abbildung 5, Seite 21] auch nur die beiden äußeren Regressionsgeraden eingezeichnet.
Gründe für die minimalen Unterschiede sind: bereinigte Daten, es gibt kaum

Ausreißer - der Ausreißer vom Boxplot [siehe Abbildung 1, Seite 6] findet sich leicht wieder, eben dieser Ausreißer bewirkt eine minimale Drehung bei der Verwendung des Hampel Schätzers (Gewichtsfunktion zum Ausgleich).

3.9 Aufgabe 9

3.9.1 Aufgabenstellung

Lassen Sie Ihre Variable von allen anderen Elementen linear abhängen. Wie lautet die Regressionsgleichung? Welche Elemente tragen dazu nichts bei?

3.9.2 S-PLUS

Durchführen der Regressionsanalyse.

```
> summary(lm(crest ~ ., kangaroos.no.crest.df))
Call: lm(formula = crest ~ ., data = kangaroos.no.crest.df)
Residuals:
    Min     1Q Median    3Q   Max
 -48.85 -12.82 -2.699 12.65 56.21

Coefficients:        Value Std. Error  t value Pr(>|t|)
      (Intercept) 33.6408 59.2015     0.5682   0.5715
           number -0.5235  0.6052    -0.8650   0.3896
              sex 12.7597  5.7106     2.2344   0.0282
          species 10.3015  5.3418     1.9285   0.0573
   basilar.length  0.1330  0.1365     0.9739   0.3330
    occipitonasal -0.1482  0.1187    -1.2491   0.2153
        palatilar -0.2838  0.1387    -2.0468   0.0440
           palate -0.2400  0.1462    -1.6414   0.1046
     nasal.length  0.0882  0.1400     0.6297   0.5307
      nasal.width  0.1540  0.2326     0.6620   0.5098
        squamosal  0.1626  0.2101     0.7741   0.4412
   inter.lacrymal  0.0303  0.2111     0.1437   0.8861
        zygometic  0.5284  0.1294     4.0838   0.0001
     post.orbital  0.5087  0.1743     2.9182   0.0046
          rostral -0.2303  0.1555    -1.4808   0.1426
  supra.occipital  0.0974  0.1398     0.6965   0.4881
incisive.foramina  0.0566  0.1736     0.3261   0.7452
  mandible.length -0.1026  0.1319    -0.7776   0.4391
   mandible.width  0.1830  0.4075     0.4490   0.6546
   mandible.depth  0.1446  0.2521     0.5735   0.5679
              arh -0.2932  0.1396    -2.1010   0.0388

Residual standard error: 21.19 on 80 degrees of freedom
Multiple R-Squared: 0.752
F-statistic: 12.13 on 20 and 80 degrees of freedom, the p-value is 1.11e-16
```

3.9.3 Aussagen

Je geringer der p-Wert der Variable ist, desto mehr trägt die betreffende Variable zu dem Modell bei. Zum Beispiel würde man das *Geschlecht* mit einem p-Wert von 0.0282 in das Modell aufnehmen, den *Tränenapparat* mit 0.8861 aber nicht.

3.10 Aufgabe 10

3.10.1 Aufgabenstellung

Führen Sie eine schrittweise Regression durch. Wieviele Elemente bleiben übrig?

3.10.2 S-PLUS

Durchführen der schrittweisen Regression

```
> stepwise(kangaroos.no.crest.df,crest,method="forward")
$rss:
 [1] 71374.43 60649.33 55328.79 53564.61 44851.90 42889.69 41586.12 40457.36
 [9] 39150.02 38407.16 37679.20 37414.43 37156.81 36920.62 36601.79 36323.97
[17] 36096.20 35998.55 35943.28 35934.01

$which:
 1(+ 8) nasal.length
 2(+13) post.orbital
 3(+ 2) sex
 4(+12) zygometic
 5(+17) mandible.length
 6(+ 7) palate
 7(+14) rostral
 8(+ 6) palatilar
 9(+20) arh
10(+ 3) species
11(+19) mandible.depth
12(+ 9) nasal.width
13(+ 5) occipitonasal
14(+15) supra.occipital
15(+ 4) basilar.length
16(+ 1) number
17(+10) squamosal
18(+18) mandible.width
19(+16) incisive.foramina
20(+11) inter.lacrymal
```

```
$f.stat:
 [1] 101.9517803  17.3301035   9.3277380   3.1618215  18.4542327   4.3005109
 [7]   2.9152060   2.5668041   3.0387730   1.7407457   1.7194674   0.6227594
[13]   0.6031858   0.5501631   0.7404307   0.6424567   0.5237365   0.2224271
[19]   0.1245514   0.0206496
```

3.10.3 Aussagen

Es werden schrittweise Variablen zu dem Modell hinzugefügt → die Residuenquadratsummen werden kleiner. Es wird jeweils die Variable mit dem größten F-to-enter Wert in das Modell aufgenommen, für *nasal.length* wäre dies zum Beispiel 101.95. Bei BMDP wäre die Default-Schranke bei dem der Algorithmus abbricht zum Beispiel bei 4.0. Für diese Beispiel würde dies bedeuten, daß die ersten drei Variablen in das Modell aufgenommen werden.

3.11 Aufgabe 11

3.11.1 Aufgabenstellung

Untersuchen Sie den Datensatz mittels Faktorenanalyse. Betrachten Sie nur die Merkmale 10 bis 15. Lassen sich die drei Tierarten in getrennten Clustern darstellen?

3.11.2 S-PLUS

Faktoranalyse durchführen

```
> kangaroos.factor_factanal(kangaroos.factor.df,factors=2)
```

Die Ladungsmatrix

```
> kangaroos.factor$loadings
                Factor1 Factor2
      squamosal 0.881
 inter.lacrymal 0.853   0.475
      zygometic 0.922   0.181
   post.orbital 0.106   0.591
        rostral 0.887   0.181
supra.occipital 0.900   0.264
```

Erstellen der Plots

```
> biplot(kangaroos.factor,olab=species)
```

```
> plot(kangaroos.factor$loadings)
```

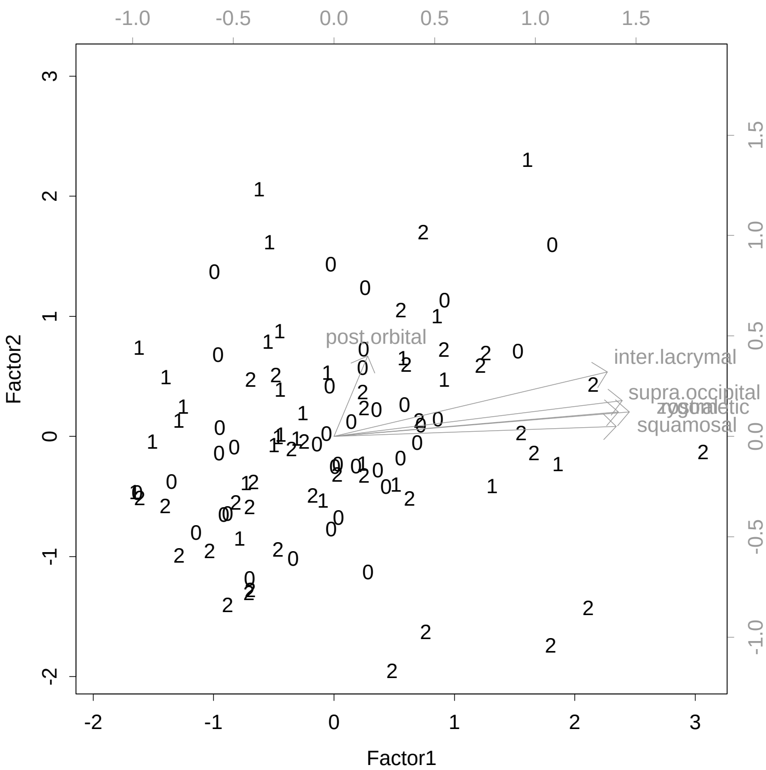

Abbildung 6: Faktorwerte

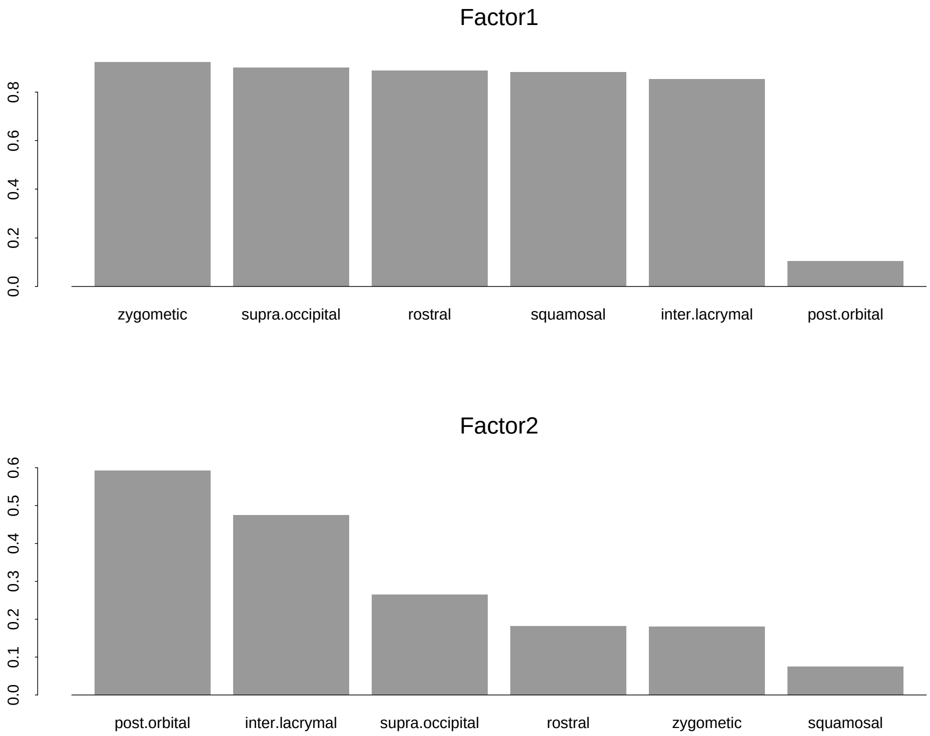

Abbildung 7: Graphische Darstellung der Ladungsmatrix

3.11.3 Aussagen

Auf dem Plot der Faktorwerte [siehe Abbildung 6, Seite 25] kann man erkennen, daß alle Variablen, bis auf post.orbital, zum Großteil einen Beitrag zum Faktor 1 leisten. Die einzelnen Tierarten bilden keine Cluster. Anhand der Ladungsmatrix [siehe Abbildung 7, Seite 26] erkennt man den Einfluß der einzelnen Variablen auf die beiden Faktoren.

4 Anhang

Dieses Dokument wurde mit LaTeX [Michel Goossens, 1994] unter Linux (Version: 5.0.1 Kernel: Linux stargate 2.0.11 #3 Thu Aug 8 23:20:52 MET DST 1996 i586) erstellt. Die folgenden Pakete wurden verwendet: german, isolatin1, psfig, float und fancyheadings. Das LaTeX Dokument umfaßt 37410 Zeichen in 1058 Zeilen.

Literatur

[Dutter, 1996] Dutter, R. (1996). *EDV-gestützte Statistik*. Institut für Statistik und Wahrscheinlichkeitstheorie, Wien, nach WS 1994/95 überarbeitete Auflage.

[Josef Bleymüller, 1996] Josef Bleymüller, G. G. u. H. G. (1996). *Statistik für Wirtschaftswissenschaftler*. Verlag Vahlen, München, zehnte überarbeitete Auflage.

[Michel Goossens, 1994] Michel Goossens, F. M. u. A. S. (1994). *Der LaTeXBegleiter*. Addison-Wesley, Bonn; Paris, erste Auflage / erster korrigierter Nachdruck 1996.