# Klausur Datenanalyse und Statistik (SS 2016)

Matrikelnummer: Fachrichtung:

| Aufgabe: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | $\sum$ | ZP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Pkt. mgl. | 4 | 4 | 5 | 5 | 8 | 3 | 5 | 7 | 3 | 44 | |
| Pkt erreicht: | | | | | | | | | | | |

Unter der folgenden Nummer finden Sie Ihr Ergebnis später im Internet: | D | S | 1 | 6 | 3 | | | |

**Schreiben Sie sich die Nummer bitte jetzt auf!**

Diese Klausur wird nur dann als Prüfung gewertet, wenn Sie im Prüfungsamt angemeldet sind. Ansonsten werden die Ergebnisse nur für einen Schein gewertet.
Lesen Sie die Aufgaben genau durch. Nehmen Sie für diese Klausur grundsätzlich ein $\alpha$-Niveau von 5% an.

# 1 Daten und Analyseaufgaben

Diese Klausur steht ganz im Zeichen der $CO_2$ Bilanz. Dazu werden wir zwei Datensätze analysieren.

## 1.1 Datensatz "Kirsche"

Der bei dem ersten Datensatz ("Kirsche") geht es darum mit einfachen Messungen die in Wäldern gebundene CO2 Menge zu quantifizieren. Dazu soll ermittelt werden wie die in einem Baum gebundene $CO_2$ Menge (`C2O` in Tonnen [t]) von seiner Höhe (`hoehe` in Meter [m]) und seinem Umfang in 1,27 m Hoehe (`umfang` in [cm]) abhängt. In diesem Datensatz betrachten wir das für die Schwarzkirsche. Dafür wurden 31 Exemplare dieser Art vermessen und dann zur Bestimmung des gebundenen CO2 gefällt.

```
> head(Kirsche)

    hoehe umfang        CO2
1 21.3360 21.082 0.08749906
2 19.8120 21.844 0.08749906
3 19.2024 22.352 0.08664955
4 21.9456 26.670 0.13931889
```

```
5 24.6888 27.178 0.15970701
6 25.2984 27.432 0.16735256
```

⋮

## 1.2 Datensatz "uptake"

Im zweiten Datensatz "uptake" geht es darum zu verstehen, wie sich die $CO_2$ aufnahme von Pflanzen bei geänderten Klimabedingungen verhalten könnte. Gemessen wurde die $CO_2$-Aufnahme (`uptake` in $\frac{\mu mol}{m^2 s}$ von 12 Individuen (ID in Variable `Plant`) der Grassart "Echinochloa crus-galli". Wenn die Pflanzen unter gewissen Bedingungen mehr $CO_2$ abgaben als aufnahmen, war dies als negativer `uptake` zu berichten. Die Gräser kamen aus zwei Herkunftsregionen mit unterschiedlichem Klima ("Quebec" und "Mississippi" (Variable `Type`). Es wurde der Einfluss der Temperatur auf zwei Stufen (Variable `Treatment` mit den Werten "kaltünd "warm") und des $CO_2$-Partialdrucks (Variable `conc` für "concentration" in $\frac{ml}{l}$ in der Luft untersucht.

```
> head(uptake)
```

```
  Plant   Type Treatment conc uptake
1   Qn1 Quebec      warm   95   16.0
2   Qn1 Quebec      warm  175   30.4
3   Qn1 Quebec      warm  250   34.8
4   Qn1 Quebec      warm  350   37.2
5   Qn1 Quebec      warm  500   35.3
6   Qn1 Quebec      warm  675   39.2
```

⋮

Ziel ist festzustellen ob und welchen Einfluss Klima, Temperatur und $CO_2$ Partialdruck auf die $CO_2$ Aufnahme haben, um verstehen zu können, wie sich $CO_2$ Aufnahme im Falle einer $CO_2$-bedingten Klimaveränderung verhalten wird.
Im weiteren Verlauf der Klausur werden wir auch Zusammenfassungen und Teildatensätze verwenden.

> Quelle: Potvin, C., Lechowicz, M. J. and Tardif, S. (1990) "The statistical analysis of ecophysiological response curves obtained from experiments involving repeated measures", *Ecology*, **71**, 1389-1400.

# 2 Aufgaben

**Aufgabe 1: Skalen**
Welche statistische Skala haben die folgenden Variablen (4):

**(1)** `Treatment`

---

**(2)** `conc`

---

**(3)** `uptake`

---

**(4)** `Plant`

---

**Aufgabe 2: Graphiken auswählen**

**(1)** Welche statistische Graphik eignet sich am besten, um die Variable `Treatment` darzustellen?(1)

---

**(2)** Welche statistische Graphik eignet sich am besten zur Darstellung der Abhängigkeit $CO_2$-Partialdruck und $CO_2$-Aufnahme (1)

---

**(3)** Welche statistische Graphik eignet sich am besten zur Darstellung der Abhängigkeiten von `Plant`, `Type` und `Treamtment`.(1)

---

**(4)** Warum sollten Sie kein Histogram wählen, um die Variable `Plant` darzustellen? (1)

---

**Aufgabe 3: Gültigkeitsbereich der Analyse**

**(1)** Wie muss die Stichprobe im Datensatz "Kirsche" erhoben werden, damit die Ergebnisse der Studie später für die Berechung des in deutschen Kirschbäumen gebunden $CO_2$ verwendet werden kann?(2)

---

---

**(2)** Welche Eigenschaft der Stichprobe haben Sie damit versucht herzustellen? (1)

---

**(3)** Was ist in diesem Fall die Grundgesamtheit? (1)

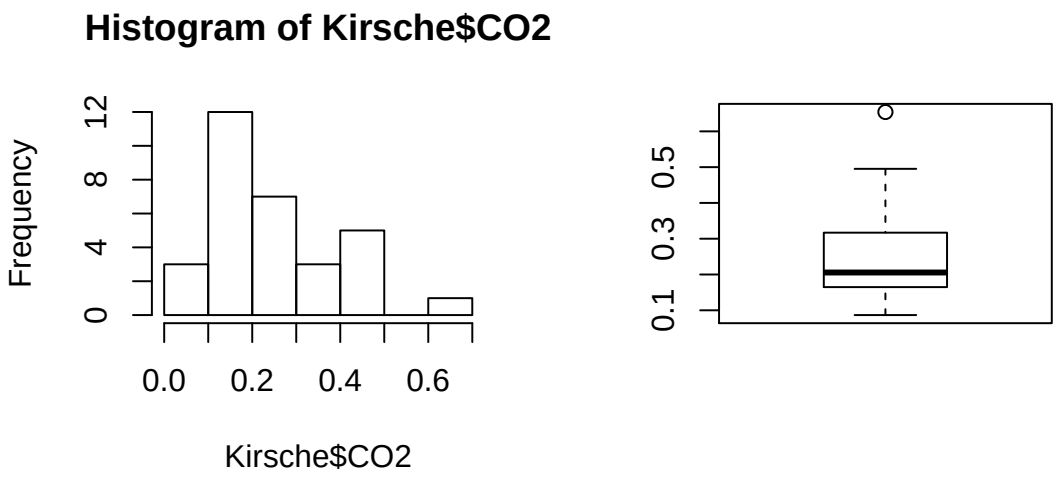
---

**(4)** Was ist in diesem Fall die Stichprobe? (1)

---

**Aufgabe 4: Deskriptive Statistik**

```
> par(mfrow=c(1,2))
> hist(Kirsche$CO2)
> boxplot(Kirsche$CO2)
```




**(1)** Geben Sie zwei Eigenschaften der Verteilung für in Kirschbäumen gebundes $CO_2$ an (2).

---

**(2)** Welche Kenngröße sollte man verwenden, wenn den Leser der Studie interessiert wieviel $CO_2$ wohl in den 100Millionen deutschen Kirschbäumen gebunden ist?(1)

---

**(3)** Welche Kenngröße sollten wir verwenden, wenn wir wissen wollen, wie stark das in Kirschbäumen gebundene $CO_2$ variiert? (2)

---

**Aufgabe 5: Regression**

Hier und in Graphik **??** wurde eine detailierte Regressionsanalyse für den Datensatz "Kirsche" durchgeführt.

```
> R2<- function(mod) var(predict(mod))/var(resid(mod)+predict(mod))
> mod1 <- lm(CO2~hoehe,data=Kirsche)
> mod1
```

```
Call:
lm(formula = CO2 ~ hoehe, data = Kirsche)

Coefficients:
(Intercept)        hoehe
   -0.74012      0.04301
```

```
> R2(mod1)
```

```
[1] 0.3579026
```

```
> anova(mod1)
```

```
Analysis of Variance Table

Response: CO2
          Df  Sum Sq  Mean Sq F value    Pr(>F)
hoehe      1 0.20937 0.209367  16.165 0.0003784
Residuals 29 0.37562 0.012952
```

```
> mod2 <- lm(CO2~umfang,data=Kirsche)
> mod2
```

```
Call:
lm(formula = CO2 ~ umfang, data = Kirsche)

Coefficients:
(Intercept)       umfang
   -0.31384      0.01694
```

```
> R2(mod2)
```

```
[1] 0.9353199
```

```
> anova(mod2)
```

```
Analysis of Variance Table

Response: CO2
          Df  Sum Sq Mean Sq F value    Pr(>F)
umfang     1 0.54715 0.54715  419.36 < 2.2e-16
Residuals 29 0.03784 0.00130
```

```
> mod3 <- lm(CO2~umfang+hoehe,data=Kirsche)
> mod3
```

```
Call:
lm(formula = CO2 ~ umfang + hoehe, data = Kirsche)

Coefficients:
(Intercept)       umfang        hoehe
  -0.492608     0.015746     0.009455
```

```
> R2(mod3)

[1] 0.94795

> anova(mod3)

Analysis of Variance Table

Response: CO2
          Df  Sum Sq Mean Sq  F value  Pr(>F)
umfang     1 0.54715 0.54715 503.1503 < 2e-16
hoehe      1 0.00739 0.00739   6.7943 0.01449
Residuals 28 0.03045 0.00109

> mod4 <- lm(log(CO2)~log(umfang),data=Kirsche)
> mod4

Call:
lm(formula = log(CO2) ~ log(umfang), data = Kirsche)

Coefficients:
(Intercept)  log(umfang)
     -9.172        2.200

> R2(mod4)

[1] 0.9538743

> anova(mod4)

Analysis of Variance Table

Response: log(CO2)
            Df Sum Sq Mean Sq F value    Pr(>F)
log(umfang)  1 7.9254  7.9254  599.72 < 2.2e-16
Residuals   29 0.3832  0.0132

> mod5 <- lm(log(CO2)~log(umfang)+log(hoehe),data=Kirsche)
> mod5

Call:
lm(formula = log(CO2) ~ log(umfang) + log(hoehe), data = Kirsche)

Coefficients:
(Intercept)  log(umfang)   log(hoehe)
    -11.921        1.983        1.117

> R2(mod5)

[1] 0.9776784

> anova(mod5)
```

```
Analysis of Variance Table

Response: log(CO2)
            Df Sum Sq Mean Sq F value    Pr(>F)
log(umfang)  1 7.9254  7.9254 1196.53 < 2.2e-16
log(hoehe)   1 0.1978  0.1978   29.86 7.805e-06
Residuals   28 0.1855  0.0066
```

```
> mod6 <- lm(log(CO2)~log(umfang)*log(hoehe),data=Kirsche)
> mod6

Call:
lm(formula = log(CO2) ~ log(umfang) * log(hoehe), data = Kirsche)

Coefficients:
           (Intercept)             log(umfang)              log(hoehe)
               -8.9789                  1.1197                  0.1823
log(umfang):log(hoehe)
                0.2740

> R2(mod6)

[1] 0.9778

> anova(mod6)

Analysis of Variance Table

Response: log(CO2)
                      Df Sum Sq Mean Sq   F value    Pr(>F)
log(umfang)            1 7.9254  7.9254 1160.1177 < 2.2e-16
log(hoehe)             1 0.1978  0.1978   28.9509 1.097e-05
log(umfang):log(hoehe) 1 0.0010  0.0010    0.1479    0.7035
Residuals             27 0.1845  0.0068
```

```
> par(mfrow=c(2,3))
> plot(exp(predict(mod5)),Kirsche$CO2,main="Vorhersage mod 5")
> abline(0,1)
> plot(predict(mod5),resid(mod5),main="mod 5")
> infl <-influence.measures(mod5)$infmat
> plot(infl[,c("hat","cook.d")],xlab="Hebelwirkung",ylab="Cook's distance",main="m
> infl <-influence.measures(mod3)$infmat
> plot(predict(mod3),Kirsche$CO2,main="Vorhersage mod 3")
> abline(0,1)
> plot(predict(mod3),resid(mod3),main="mod 3")
> plot(infl[,c("hat","cook.d")],xlab="Hebelwirkung",ylab="Cook's distance",main="m
```

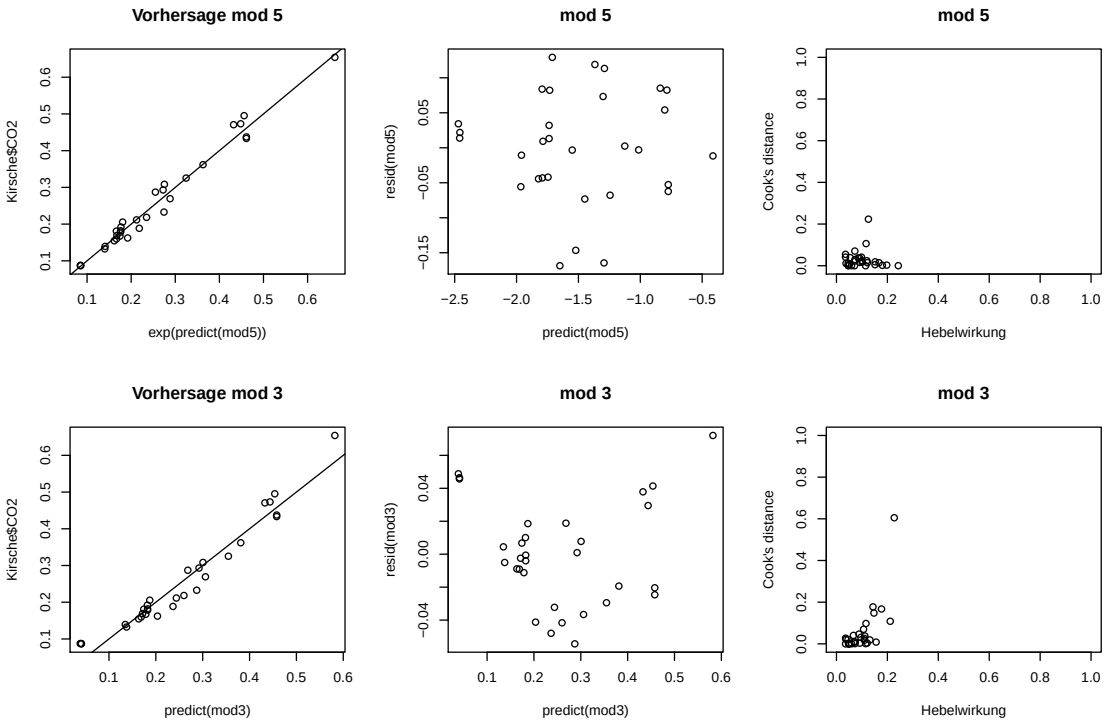



Abbildung 1: Regressionsdiagnostik des Modells mod5.

**(1)** Geben Sie zwei Gründe, warum `mod3` dem Modell `mod2` vorzuziehen ist (2).

**(2)** Geben Sie drei Aspekte in denen `mod5` dem Modell `mod3` überlegen ist Geben Sie mit an, woran Sie das sehen (4).

**(3)** Welches Modell sollten wir für die Vorhersage wählen? (1)

Warum?(1)

**Aufgabe 6: Vorhersage**
Schreiben Sie in einer Formel wieviel Tonnen $CO_2$ in einm Kirschbaum der Hoehe 3m und einem Umfang von 10cm gebunden ist.

**(1)** Gemäß Modell mod1 (1):

---

**(2)** Gemäß Modell mod5 (2):

---

**Aufgabe 7: Korrelationsanalyse**
Jemand hat eine kurze Korrelationsanalyse für den Teildatensatz mit nur einer Pflanze durchgeführt:

```
> plant1 <- uptake[uptake$Plant=="Qn1",]
> shapiro.test(plant1$conc)

        Shapiro-Wilk normality test

data:  plant1$conc
W = 0.9304, p-value = 0.5545

> shapiro.test(plant1$uptake)

        Shapiro-Wilk normality test

data:  plant1$uptake
W = 0.7801, p-value = 0.02592

> with(plant1,cor.test(conc,uptake,method="pearson"))

        Pearson's product-moment correlation

data:  conc and uptake
t = 2.2835, df = 5, p-value = 0.07123
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.08350523  0.95415861
sample estimates:
      cor
0.7144826

> with(plant1,cor.test(conc,uptake,method="spearman"))

        Spearman's rank correlation rho

data:  conc and uptake
S = 2, p-value = 0.002778
```

```
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9642857
```

**(1) Bonferronie Korrektur**
Wie und warum haben Sie die Bonferronie Korrektur zur Beantwortung für die Interpretation der Shapiro-Wilk-Tests verwendet? (2)

---

---

**(2) Schlussfolgerungen** Welcher der folgenden Aussagen lassen sich aus den Ergebnissen der beiden Korrelationstests ableiten? (3)

- **O** Die $CO_2$-Aufnahme ist vom $CO_2$-Partialdruck in der Luft unabhängig.
- **O** Die $CO_2$-Aufnahme ist vom $CO_2$-Partialdruck in der Luft abhängig.
- **O** Eine höhere $CO_2$-Aufnahme steigert den $CO_2$-Partialdruck in der Luft.
- **O** Bei mehr $CO_2$ in der Luft wird tendentiell weniger $CO_2$ aufgenommen (Sättigungseffekt)
- **O** Es besteht ein linearer Zusammenhang zwischen $CO_2$-Aufnahme und $CO_2$-Partialdruck.
- **O** Statistisch gesehen nimmt die Pflanze bei mehr $CO_2$ in der Luft auch mehr davon auf.
- **O** Es gibt klare Hinweise, dass der Zusammenhang zwischen $CO_2$-Partialdruck und $CO_2$-Aufnahme nichtlinear ist.

Begründen Sie:

---

---

---

---

**Aufgabe 8: Varianzanalyse**
Wir wollen nun die Abh"ängigkeit der $CO_2$ Aufnahme von der Herkunftsregion der Pflane und der Temperatur bei einem festen Partialdruck von $350\frac{\mu l}{l}$ untersuchen:

```
> u350 <- uptake[uptake$conc==350,]
> u350
```

```
   Plant        Type Treatment conc uptake
4    Qn1      Quebec      warm  350   37.2
11   Qn2      Quebec      warm  350   41.8
18   Qn3      Quebec      warm  350   42.1
25   Qc1      Quebec      kalt  350   34.6
32   Qc2      Quebec      kalt  350   38.8
39   Qc3      Quebec      kalt  350   34.0
46   Mn1 Mississippi      warm  350   30.0
53   Mn2 Mississippi      warm  350   31.8
60   Mn3 Mississippi      warm  350   27.9
67   Mc1 Mississippi      kalt  350   18.9
74   Mc2 Mississippi      kalt  350   13.0
81   Mc3 Mississippi      kalt  350   17.9
```

```
> anova(lm(uptake~Treatment+Type,data=u350))
```

```
Analysis of Variance Table

Response: uptake
          Df Sum Sq Mean Sq F value    Pr(>F)
Treatment  1 239.41  239.41  18.979  0.001833
Type       1 660.08  660.08  52.328 4.901e-05
Residuals  9 113.53   12.61
```

```
> anova(lm(uptake~Type+Treatment,data=u350))
```

```
Analysis of Variance Table

Response: uptake
          Df Sum Sq Mean Sq F value    Pr(>F)
Type       1 660.08  660.08  52.328 4.901e-05
Treatment  1 239.41  239.41  18.979  0.001833
Residuals  9 113.53   12.61
```

```
> anova(lm(uptake~Type*Treatment,data=u350))
```

```
Analysis of Variance Table

Response: uptake
               Df Sum Sq Mean Sq F value    Pr(>F)
Type            1 660.08  660.08 93.7507 1.079e-05
Treatment       1 239.41  239.41 34.0036  0.000391
Type:Treatment  1  57.20   57.20  8.1245  0.021467
Residuals       8  56.33    7.04
```

**(1)** Warum darf die Varianzanalyse nur auf solche Teildatensätz, aber nicht auf den gesamten "uptake" Datensatz angewendet werden? (Tip: Voraussetzungen) (2)

---

**(2)** Warum ist es nicht notwendig noch die Modelle ohne einen der beiden Einflussfaktoren zu betrachten? (1)

---

**(3)** Angenommen die Voraussetzungen sind erfüllt: Was wurde dann mit dieser statistischen Auswertung nachgewiesen? (4)

- **O** Die $CO_2$-Aufnahme wird von Klima und Wetter nicht beeinflußt.
- **O** Die $CO_2$-Aufnahme ist von der Temperatur abhängig, aber nicht vom Klima an das die Pflanzen gewöhnt sind.
- **O** Es konnte nur die Abhängigkeit von der Tagestemperatur nachgewiesen Werden. Eine Abhängigkeit vom Klima bleibt unklar.
- **O** Die $CO_2$-Aufnahme dieser Pflanzen hängt von Klima und Wetter ab.
- **O** Es gibt starke Ausreißer im Datensatz, die auf eine Korrelation hindeuten.
- **O** $R^2$ ist negativ.
- **O** Der Einfluss von Temperatur und Herkunftsregion konnte auf dem 5%-Niveau signifikation nachgewiesen werden.
- **O** Diese Analyse war nicht hilfreich und trägt nichts zum Verständnis der $CO_2$-Aufnahme bei, da alle Tests signifikant waren und somit nichts nachgewiesen werden konnte.
- **O** Aus den Koeffizienten läßt sich die $CO_2$-Aufnahme aller Pflanzen für globale Klimamodelle exakt ableiten.
- **O** Die Interaktion der beiden Effekte konnte nachgewiesen werden.
- **O** Für diese Aufgabe wäre besser gewesen zwei t-Tests anzuwenden.

Begründen Sie:

---

**Aufgabe 9:** Formulieren Sie die wichtigsten Ergebnisse aus dieser Klausur für einen Klimawissenschaftler, der sich bisher nicht mit der Bilanzierung pflanzengebundenen $CO_2$ beschäftigt hat, dieses aber ab sofort tun möchte. (3)