Name:	Nicht bestanden: □
Vorname:	
Matrikelnummer:	Endnote:

M.Sc. Angewandte Nutztier- und Pflanzenwissenschaften

# **Klausur Biostatistik**

#### Hochschule Osnabrück

Prüfer: Prof. Dr. Jochen Kruppa Fakultät für Agrarwissenschaften und Landschaftsarchitektur j.kruppa@hs-osnabrueck.de

29. Juni 2023

#### Erlaubte Hilfsmittel für die Klausur

- Normaler Taschenrechner ohne Möglichkeit der Kommunikation mit anderen Geräten also ausdrücklich kein Handy!
- Eine DIN A4-Seite als beidseitig, selbstgeschriebene, handschriftliche Formelsammlung keine digitalen Ausdrucke.
- You can answer the questions in English without any consequences.

## **Ergebnis der Klausur**

\_\_\_\_\_ von 20 Punkten sind aus dem Multiple Choice Teil erreicht.

von 66 Punkten sind aus dem Rechen- und Textteil erreicht.

\_\_\_\_\_ von 86 Punkten in Summe.

Es wird folgender Notenschlüssel angewendet.

Punkte	Note
82.0 - 86.0	1,0
78.0 - 81.5	1,3
73.5 - 77.5	1,7
69.5 - 73.0	2,0
65.0 - 69.0	2,3
60.5 - 64.5	2,7
56.5 - 60.0	3,0
52.0 - 56.0	3,3
48.0 - 51.5	3,7
43.0 - 47.5	4,0

Es ergibt sich eine Endnote von \_\_\_\_\_.

#### **Multiple Choice Aufgaben**

- Pro Multipe Choice Frage ist *genau* eine Antwort richtig.
- Übertragen Sie Ihre Kreuze in die Tabelle auf dieser Seite.
- Es werden nur Antworten berücksichtigt, die in dieser Tabelle angekreuzt sind!

	A	В	С	D	E	✓
1 Aufgabe						
2 Aufgabe						
3 Aufgabe						
4 Aufgabe						
5 Aufgabe						
6 Aufgabe						
7 Aufgabe						
8 Aufgabe						
9 Aufgabe						
10 Aufgabe						

• Es sind \_\_\_\_ von 20 Punkten erreicht worden.

## **Rechen- und Textaufgaben**

• Die Tabelle wird vom Dozenten ausgefüllt.

Aufgabe	11	12	13	14	15	16	17
Punkte	9	10	8	10	9	11	9

• Es sind \_\_\_\_ von 66 Punkten erreicht worden.

1 Aufgabe (2 Punkte)

Der Datensatz PlantGrowth enthält das Gewicht von Pflanzen, die unter einer Kontrolle und zwei verschiedenen Behandlungsbedingungen erzielt wurden. Nach der Berechnung einer einfaktoriellen ANOVA ergibt sich ein $\eta^2 = 0.27$ . Welche Aussage ist richtig?
<b>A</b> $\square$ Die Berechnung von $\eta^2$ ist ein Wert für die Interaktion.
<b>B</b> $\square$ Das $\eta^2$ beschreibt den Anteil der Varianz, der von den Behandlungsbedingungen erklärt wird. Das $\eta^2$ ist damit mit dem $R^2$ aus der linearen Regression zu vergleichen.
${f C}$ Das $\eta^2$ beschreibt den Anteil der Varianz, der von den Behandlungsbedingungen nicht erklärt wird. Somit der Rest an nicht erklärbarer Varianz.
<b>D</b> $\square$ Das $\eta^2$ ist die Korrelation der ANOVA. Mit der Ausnahme, dass 0 der beste Wert ist.
<b>E</b> $\square$ Das $\eta^2$ ist ein Wert für die Güte der ANOVA. Je kleiner desto besser. Ein $\eta^2$ von 0 bedeutet ein perfektes Modell mit keiner Abweichung. Die Varianz ist null.
2 Aufgabe (2 Punkte)
In der Bio Data Science wird häufig mit sehr großen Datensätzen gerechnet. Historisch ergibt sich nun ein Problem bei der Auswertung der Daten und deren Bewertung hinsichtlich der Signifikanz. Welche Aussage ist richtig?
<b>A</b> □ Aktuell werden zu grosse Datensätze für die gänigige Statistik gemessen. Daher wendet man maschinelle Lernverfahren für kausale Modelle an. Hier ist die Relevanz gleich Signifikanz.
<b>B</b> □ Aktuell werden immer grössere Datensätze erhoben. Eine erhöhte Fallzahl führt automatisch auch zu mehr signifikanten Ergebnissen, selbst wenn die eigentlichen Effekte nicht relevant sind.
$\mathbf{C} \ \square$ Aktuell werden immer grössere Datensätze erhoben. Dadurch wird auch die Varianz immer höher was automatisch zu mehr signifikanten Ergebnissen führt.
<b>D</b> ☐ Big Data ist ein Problem der parametrischen Statistik. Parameter lassen sich nur auf kleinen Datensätzen berechnen, da es sich sonst nicht mehr um eine Stichprobe im engen Sinne der Statistik handelt.
<b>E</b> $\square$ Relevanz und Signifikanz haben nichts miteinander zu tun. Daher gibt es auch keinen Zusammenhang zwischen hoher Fahlzahl (n > 10000) und einem signifikanten Test. Ein Effekt ist immer relevant und somit signifikant.
3 Aufgabe (2 Punkte)
In der Theorie zur statistischen Testentscheidung kann " $H_0$ beibehalten obwohl die $H_0$ falsch ist" in welche richtige Analogie gesetzt werden?

Klausur Biostatistik 4

**D** □ In die Analogie eines brennenden Hauses ohne Rauchmelder: *House without noise*.

**A**  $\square$  In die Analogie eines Rauchmelders: *Fire without alarm*, dem  $\beta$ -Fehler.

**E**  $\square$  In die Analogie eines Rauchmelders: *Alarm without fire*, dem  $\alpha$ -Fehler.

**B** □ In die Analogie eines Rauchmelders: *Alarm with fire*.

**C** □ In die Analogie eines Feuerwehrautos: *Car without noise*.

4 Aufgabe (2 Punkte)

Nachdem Sie in einem Experiment die Daten $D$ erhoben haben, berechnen Sie den Mittelweund den Median. Der Mittelwert $ar{y}$ und der Median $ar{y}$ unterscheiden sich nicht. Welche Aussage irichtig?	
A □ Da sich der Mittelwert und der Median unterscheiden, ist der Datensatz nicht zu verwende Mittelwert und Median müssen gleich sein.	ŀn.

C □ Da sich der Mittelwert und der Median nicht unterscheiden, liegen vermutlich Outlier in den Daten vor.

**B** Da sich der Mittelwert und der Median nicht unterscheiden, liegen vermutlich keine Outlier

in den Daten vor. Wir verweden den Datensatz so wie er ist.

**D** □ Da sich der Mittelwert und der Median unterscheiden, liegen vermutlich keine Outlier in den Daten vor.

**E** □ Da sich der Mittelwert und der Median unterscheiden, liegen vermutlich Outlier in den Daten vor. Wir untersuchen den Datensatz nach auffälligen Beobachtungen.

5 Aufgabe (2 Punkte)

Sie haben folgende unadjustierten p-Werte gegeben: 0.001, 0.89, 0.02, 0.34 und 0.21. Sie adjustieren die p-Werte nach Bonferroni. Welche Aussage ist richtig?

**A**  $\square$  Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 0.005, 1, 0.1, 1 und 1. Die adjustierten p-Werte werden zu einem  $\alpha$ -Niveau von 5% verglichen.

**B**  $\square$  Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 0.005, 4.45, 0.1, 1.7 und 1.05. Die adjustierten p-Werte werden zu einem  $\alpha$ -Niveau von 5% verglichen.

**C**  $\square$  Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 2e-04, 0.178, 0.004, 0.068 und 0.042. Die adjustierten p-Werte werden zu einem  $\alpha$ -Niveau von 1% verglichen.

**D**  $\square$  Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 2e-04, 0.178, 0.004, 0.068 und 0.042. Die adjustierten p-Werte werden zu einem  $\alpha$ -Niveau von 5% verglichen.

**E**  $\square$  Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 0.005, 1, 0.1, 1 und 1. Die adjustierten p-Werte werden zu einem  $\alpha$ -Niveau von 1% verglichen.

6 Aufgabe (2 Punkte)

Bei der explorativen Datenanalyse (EDA) in  $\P$  gibt es eine richtige Abfolge von Prozessschritten, auch *Circle of life* genannt. Wie lautet die richtige Reihenfolge für die Erstellung einer EDA?

**A** ☐ Wir lesen als erstes die Daten über read\_excel() ein, transformieren die Spalten über mutate() in die richtige Form und können dann über ggplot() uns die Abbildungen erstellen lassen.

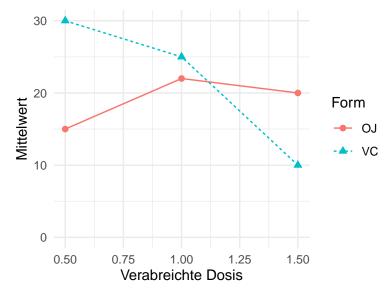
**B** ☐ Wir transformieren die Spalten über mutate() in ein tibble und können dann über ggplot() uns die Abbildungen erstellen lassen. Dabei beachten wir das wir keine Faktoren in den Daten haben.

C ☐ Wir lesen die Daten über eine generische Funktion read() ein und müssen dann die Funktion ggplot() nur noch installieren. Dann haben wir die Abbildungen als \*.png vorliegen.

D 🗆	Wir lesen als erstes die Daten über read_excel() ein, transformieren die Spalten über mutate() in die richtige Form und können dann über ggplot() uns die Abbildungen erstellen lassen. Wichtig ist, dass wir keine Faktoren sondern nur numerische Variablen vorliegen haben.
E	Wir lesen die Daten ein und mutieren die Daten. Dabei ist wichtig, dass wir nicht das Paket tidyverse nutzen, da dieses Paket veraltet ist. Über die Funktion library(tidyverse) entfernen wir das Paket von der Analyse.
7 /	Aufgabe (2 Punkte)
	Randomisierung von Beobachtungen bzw. Samples zu den Versuchseinheiten ist bedeutend er Versuchsplanung. Welche der folgenden Aussagen ist richtig?
<b>A</b> 🗆	Randomisierung erlaubt erst die Mittelwerte zu schätzen. Ohne Randomisierung keine Mittelwerte.
В□	Randomisierung bringt starke Unstrukturiertheit in das Experiment und erlaubt erst von der Stichprobe auf die Grundgesamtheit zurückzuschliessen.
<b>C</b> 🗆	Randomisierung sorgt für Strukturgleichheit und erlaubt erst von der Stichprobe auf die Grundgesamtheit zurückzuschliessen.
D 🗆	Randomisierung war bis 1952 bedeutend, wurde dann aber in Folge besserer Rechnerleistung nicht mehr verwendet. Aktuelle Statistik nutzt keine Randomisierung mehr.
E	Randomisierung erlaubt erst die Varianzen zu schätzen. Ohne eine Randomisierung ist die Berechnung von Mittelwerten und Varianzen nicht möglich.
8 4	Aufgabe (2 Punkte)
	nem Zuchtexperiment messen wir die Ferkel verschiedener Sauen. Die Ferkel einer Muttersau daher im statistischen Sinne
<b>A</b> 🗆	Untereinander unabhängig. Die Ferkel sind eigenständig und benötigen keine zusätzliche Behandlung.
В□	Untereinander abhängig, wenn die Mütter ebenfalls miteinander verwandt sind. Erst die Abhängigkeit 2. Grades wird in der Statistik modelliert.
<b>C</b> 🗆	Untereinander stark korreliert. Die Ferkel sind von einer Mutter und sommit miteinander korreliert. Dies wird in der Statistik jedoch meist nicht modelliert.
D 🗆	Untereinander abhängig. Die Ferkel stammen von einem Muttertier und haben vermutliche eine ähnliche Varianzstruktur.
E	Untereinander unabhängig. Sollten die Mütter verwandt sein, so ist die Varianzstruktur ähnlich und muss modelliert werden.

9 Aufgabe (2 Punkte)

Die folgende Abbildung enthält die Daten aus einer Studie zur Bewertung der Wirkung von Vitamin C auf das Zahnwachstum bei Meerschweinchen. Der Versuch wurde an 60 Schweinen durchgeführt, wobei jedes Tier eine von drei Vitamin-C-Dosen (0.5, 1 und 1.5 mg/Tag) über eine von zwei Verabreichungsmethoden mit Orangensaft (OJ) oder Ascorbinsäure (VC) erhielt.



Welche Aussage ist richtig im Bezug auf eine zweifaktorielle ANOVA?

- **A** ☐ Keine Interaktion ist zu erwatzen. Die Geraden der Verabreichungsmethode laufen parallel und mit ähnlichen Abständen.
- **B** □ Keine Interaktion liegt vor. Die Geraden scheiden sich und laufen nicht parallel.
- **C** □ Eine starke Interaktion liegt vor. Die Geraden laufen parallel und schneiden sich nicht.
- f D  $\Box$  Eine starke Interaktion ist zu erwarten. Die Geraden schneiden sich und die Abstände sind nicht gleichbleibend.
- **E** ☐ Eine leichte Interaktion ist zu erwarten. Die Geraen schneiden sich noch nicht, aber die Abstände unterscheiden sich stark.

10 Aufgabe (2 Punkte)

Das Falsifikationsprinzip besagt...

- **A** □ ... dass Fehlerterme in statistischen Modellen nicht verifiziert werden können.
- **B** □ ... dass in der Wissenschaft immer etwas falsch sein muss. Sonst gebe es keinen Fortschritt.
- **C** □ ... dass Annahmen an statistische Modelle meist falsch sind.
- **D** □ ... dass Modelle meist falsch sind und selten richtig.
- **E** □ ... dass ein schlechtes Modell durch ein weniger schlechtes Modell ersetzt wird. Die Wissenschaft lehnt ab und verifiziert nicht.

11 Aufgabe (9 Punkte)

Geben Sie grundsätzlich Formeln und Rechenweg zur Lösung der Teilaufgaben mit an!



Nach einem Feldexperiment mit zwei Düngestufen (A und B) ergibt sich die folgende Datentabelle mit dem gemessenen Trockengewicht (*drymatter*).

trt	drymatter
Α	23.0
Α	16.4
Α	18.5
Α	18.1
Α	15.7
В	20.0
В	21.2
В	8.2
Α	21.7
В	13.6
Α	18.7
В	12.9
В	12.9
В	13.0

- 1. Zeichnen Sie in *einer* Abbildung die beiden Boxplots für die zwei Düngestufen A und B! Beschriften Sie die Achsen entsprechend! **(6 Punkte)**
- 2. Beschriften Sie *einen* der beiden Boxplots mit den gängigen statistischen Maßzahlen! **(2 Punkte)**
- 3. Wenn Sie *keinen Effekt* zwischen den Düngestufen erwarten würden, wie sehen dann die beiden Boxplots aus? (1 Punkt)

12 Aufgabe (10 Punkte)



Sie rechnen einen t-Test für Gruppenvergleiche. Sie schätzen den Unterschied zwischen dem mittleren Befall mit Parasiten zu einer unbehandelten Kontrolle.

- 1. Beschriften Sie die untenstehende Abbildung mit der Signifikanzschwelle! Begründen Sie Ihre Antwort! (2 Punkte)
- 2. Ergänzen Sie eine *in den Kontext passende* Relevanzschwelle! Begründen Sie Ihre Antwort! **(2 Punkte)**
- 3. Skizieren Sie in die untenstehende Abbildung sechs einzelne Konfidenzintervalle (a-f) mit den jeweiligen Eigenschaften! (6 Punkte)
  - (a) Ein signifikantes, relevantes 95%-Konfidenzintervall
  - (b) Ein nicht signifikantes, nicht relevantes 95%-Konfidenzintervall
  - (c) Ein 95%-Konfidenzintervall mit höherer Varianz  $s_p$  in der Stichprobe als der Rest der 95%-Konfidenzintervalle
  - (d) Ein signifikantes, relevantes 90%-Konfidenzintervall.
  - (e) Ein 95%-Konfidenzintervall mit niedriger Varianz  $s_p$  in der Stichprobe als der Rest 95%-der Konfidenzintervalle
  - (f) Ein signifikantes, nicht relevantes 95%-Konfidenzintervall

13 Aufgabe (9 Punkte)



In einem Experiment für den Proteingehalt von Wasserlinsen in g/l mit vier Dosisstufen (ctrl, low, mid und high) erhalten Sie folgende Matrix als  $\mathbb{R}$  Ausgabe mit den rohen, unadjustierten p-Werten.

```
## ctrl high low mid

## ctrl 1.0000000 0.1143358 0.0576407 0.5420519

## high 0.1143358 1.0000000 0.7316889 0.3214892

## low 0.0576407 0.7316889 1.0000000 0.1855728

## mid 0.5420519 0.3214892 0.1855728 1.0000000
```

Im Weiteren erhalten Sie folgende Informationen über die Fallzahl n, den Mittelwert mean und die Standardabweichung sd in den jeweiligen Dosisstufen.

trt	n	mean	sd
ctrl	9	5.50	2.37
high	9	7.23	2.88
low	9	7.60	1.77
mid	9	6.16	1.84

- 1. Zeichnen Sie in eine Abbildung, die sich ergebenden Barplots! (2 Punkte)
- 2. Adjustieren Sie die rohen p-Werte nach Bonferroni. Begründen Sie Ihre Antwort! (3 Punkte)
- 3. Ergänzen Sie das Compact letter display (CLD) zu der Abbildung! (2 Punkte)
- 4. Interpretieren Sie das Compact letter display (CLD)! (2 Punkte)

14 Aufgabe (11 Punkte)



In einem Experiment zur Steigerung der Milchleistung (gain) in dl/h von Kühen wurden zwei Arten von Musik in den Ställen gespielt. Zum einen ruhige Musik (calm) und eher flotte Musik (pop). Die Messungen wurden an jeder Kuh (subject) wiederholt durchgeführt. Darüber hinaus wurden verschiedene Ställe (barn) mit der Musik bespielt.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: gain ~ attitude + (1 | subject) + (1 | barn)
     Data: data_tbl
##
## REML criterion at convergence: 795.5
##
## Scaled residuals:
      Min 10 Median
                               30
                                      Max
## -2.3081 -0.6191 -0.0156 0.5324 3.1841
## Random effects:
## Groups Name
                        Variance Std.Dev.
## barn
             (Intercept) 195.1
                                 13.97
## subject (Intercept) 3908.3
                                 62.52
                                 25.91
## Residual
                         671.3
## Number of obs: 83, groups: barn, 7; subject, 6
## Fixed effects:
              Estimate Std. Error t value
##
## (Intercept) 202.016
                           26.367
                                    7.662
## attitudepop -19.756
                            5.693 -3.470
## Correlation of Fixed Effects:
               (Intr)
## attitudepop -0.106
```

- 1. Ist die Annahme der Normalverteilung an das Outcome *gain* erfüllt? Begründen Sie Ihre Antwort! (2 Punkte)
- 2. Wie groß ist der Effekt der Musikart attitude? Liegt ein signifikanter Effekt vor? Schätzen Sie den p-Wert mit einem kritischen t-Wert von  $T_k = 1.96$  ab. Begründen und visualisieren Sie Ihre Antwort und Entscheidung! (3 Punkte)
- 3. Was ist der Unterschied zwischen einem "random" und "fixed" Effekt. Gehen Sie in der Begründung Ihrer Antwort auf dieses konkrete Beispiel ein! (3 Punkte)
- 4. Wie groß ist die Varianz, die durch die zufälligen Effekte erklärt wird? (1 Punkt)
- 5. Schreiben Sie das Ergebnis der R Ausgabe in einen Satz nieder, der die Information zum Effekt und der Signifikanz enthält! (2 Punkte)

15 Aufgabe (9 Punkte)



Maispflanzen sollen auf die ertragssteigerende Wirkung von verschiedenen Einflussfaktoren untersucht werden. Gemessen wurde als Outcome die Trockenmasse in kg/m². Dafür wurde für jede Maispflanze gemessen wieviel Wasser (I/m²) die Pflanze erhalten hat oder ob die Pflanze ein neuartiges Lichtregime (0 =alt, 1 =neu) erhalten hatte. Zusätzlich wurde die Anzahl an Nematoden im Boden bestimmt sowie der Eisen- und Phosphorgehalt ( $\mu$ g/kg) des Bodens. Es ergibt sich folgender Auszug aus den Daten.

water	light	Р	Fe	drymatter	nematodes
6.93	0	9.78	99.69	68.01	10
10.11	0	6.72	101.05	74.12	4
9.09	0	9.16	97.98	67.17	5
11.36	1	9.71	96.78	67.18	11

Sie rechnen nun eine Gaussian lineare Regression auf den Daten und erhalten folgenden 😱 Output.

```
##
## Call:
## lm(formula = reformulate(response = "drymatter", termlabels = wanted_vec),
      data = data_tbl)
##
##
## Residuals:
##
     Min
               10 Median
                               30
                                      Max
## -6.0093 -1.4168 0.0395 1.6640 5.9925
## Coefficients:
##
               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.37948 7.50489 -1.383 0.169706
## P
               -0.58640
                           0.16581 -3.537 0.000614
## Fe
                0.85242
                           0.07442 11.455 < 2e-16
## nematodes
                0.04536
                           0.07512
                                    0.604 0.547240
## Residual standard error: 2.347 on 101 degrees of freedom
## Multiple R-squared: 0.6116, Adjusted R-squared:
## F-statistic: 53 on 3 and 101 DF, p-value: < 2.2e-16
```

- 1. Sind die Residuals approximativ Normalverteilt? Begründen Sie Ihre Antwort! (3 Punkte)
- 2. Welche der Einflussfaktoren sind signifikant? Begründen Sie Ihre Antwort! (3 Punkte)
- 3. Interpretieren Sie die Spalte *estimate* im Bezug auf den Ertrag in Trockenmasse der Maispflanzen! (3 Punkte)

16 Aufgabe (8 Punkte)



Sie erhalten folgende R Ausgabe der Funktion t.test().

```
##
## Two Sample t-test
##
## data: freshmatter by N
## t = -0.64984, df = 16, p-value = 0.525
## alternative hypothesis: true difference in means between group high and group low is no
## 95 percent confidence interval:
## -2.841476 1.508142
## sample estimates:
## mean in group high mean in group trt2
## 18.55556 19.22222
```

- 1. Formulieren Sie die wissenschaftliche Fragestellung! (2 Punkte)
- 2. Liegt ein signifikanter Unterschied zwischen den Gruppen vor? Begründen Sie Ihre Antwort! (2 Punkte)
- 3. Skizieren Sie das sich ergebende 95% Konifidenzintervall! (2 Punkte)
- 4. Beschriften Sie die Abbildung und das 95% Konfidenzintervall entsprechend! (2 Punkte)

17 Aufgabe (10 Punkte)



Der Datensatz <code>pig\_gain\_weight\_tbl</code> enthält Daten aus einer Studie zur Bewertung der Wirkung vom Vitamin Selen auf das Wachstum bei Mastschweinen. Der Versuch wurde an 18 Mastschweinen durchgeführt, wobei jedes Tier eine von drei Selen-Dosen <code>dose</code> (0.5 ng/Tag, 3 ng/Tag und 10 ng/Tag) über eine von zwei Verabreichungsmethoden <code>form</code> erhielt (Wasser oder Festnahrung). Sie erhalten folgende Ausgabe in  $\square$ .

- 1. Stellen Sie die statistische  $H_0$  und  $H_A$  Hypothese für die obige zweifaktorielle ANOVA für den Faktor *form* auf! **(2 Punkte)**
- 2. Interpretieren Sie das Ergebnis der zweifaktoriellen ANOVA. Gehen Sie im besonderen auf den Term dose : form ein! (2 Punkte)
- 3. Zeichnen Sie eine Abbildung, der dem obigen Ergebnis der zweifaktoriellen ANOVA näherungsweise entspricht! (4 Punkte)
- 4. Beschriften Sie die Abbildung entsprechend der R Ausgabe! (2 Punkte)