

**Name:** \_\_\_\_\_

*Nicht bestanden:* ☐

**Vorname:** \_\_\_\_\_

**Matrikelnummer:** \_\_\_\_\_

**Endnote:** \_\_\_\_\_

**B.Sc. Bioverfahrenstechnik in Agrar- und Lebensmittelwirtschaft**

## **Klausur Angewandte Statistik für Bioverfahrenstechnik**

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

29. Januar 2024

### 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 64 Punkten sind aus dem Rechen- und Textteil erreicht.

\_\_\_\_\_ von 84 Punkten in Summe.

Es wird folgender Notenschlüssel angewendet.

Punkte	Note
80.5 - 84.0	1,0
76.0 - 80.0	1,3
72.0 - 75.5	1,7
67.5 - 71.5	2,0
63.5 - 67.0	2,3
59.5 - 63.0	2,7
55.0 - 59.0	3,0
51.0 - 54.5	3,3
46.5 - 50.5	3,7
42.0 - 46.0	4,0

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

## Multiple Choice Aufgaben

- Pro Multiple 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	B	C	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	6	6	12	10	11	10	9

- Es sind \_\_\_\_ von 64 Punkten erreicht worden.

## 1 Aufgabe


(2 Punkte)

Welche Aussage über den t-Test ist richtig?

- A** ☐ Der t-Test ist ein Vortest der ANOVA und basiert daher auf dem Vergleich von Streuungsparametern
- B** ☐ Der t-Test vergleicht die Mittelwerte von zwei Gruppen.
- C** ☐ Der t-Test testet generell zu einem erhöhten  $\alpha$ -Niveau von 20%.
- D** ☐ Der t-Test vergleicht die Varianzen von mindestens zwei oder mehr Gruppen
- E** ☐ Der t-Test vergleicht die Mittelwerte von zwei Gruppen unter der strikten Annahme von Varianzhomogenität. Sollte keine Varianzhomogenität vorliegen, so gibt es keine Möglichkeit den t-Test in einer Variante anzuwenden.

## 2 Aufgabe

(2 Punkte)

Bei der explorativen Datenanalyse (EDA) in  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 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.
- B** ☐ 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.
- C** ☐ 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.
- 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.
- E** ☐ 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.

## 3 Aufgabe

(2 Punkte)

Eine einfaktorielle ANOVA berechnet eine Teststatistik um zu die Nullhypothese abzulehnen. Welche Aussage über die Teststatistik der ANOVA ist richtig?

- A** ☐ Die ANOVA berechnet die T-Statistik aus der Multiplikation der MS Behandlung mit der MS der Fehler. Wenn die F-Statistik genau 0 ist, kann die Nullhypothese abgelehnt werden.
- B** ☐ Die ANOVA berechnet die F-Statistik indem die MS der Behandlung durch die MS des Fehlers geteilt werden. Wenn die F-Statistik sich der 0 annähert kann die Nullhypothese nicht abgelehnt werden.
- C** ☐ Die ANOVA berechnet die T-Statistik indem den Mittelwertsunterschied der Gruppen simultan durch die Standardabweichung der Gruppen teilt. Wenn die T-Statistik höher als 1.96 ist, kann die Nullhypothese abgelehnt werden.
- D** ☐ Die ANOVA berechnet die F-Statistik aus den SS Behandlung geteilt durch die SS Fehler.
- E** ☐ Die ANOVA berechnet die F-Statistik indem die MS des Fehlers durch die MS der Behandlung geteilt werden. Wenn die F-Statistik sich der 1 annähert kann die Nullhypothese nicht abgelehnt werden.

#### 4 Aufgabe

(2 Punkte)

Die Randomisierung von Beobachtungen bzw. Samples zu den Versuchseinheiten ist bedeutend in der Versuchsplanung. Welche der folgenden Aussagen ist richtig?

- A** ☐ Randomisierung war bis 1952 bedeutend, wurde dann aber in Folge besserer Rechnerleistung nicht mehr verwendet. Aktuelle Statistik nutzt keine Randomisierung mehr.
- B** ☐ Randomisierung erlaubt erst die Mittelwerte zu schätzen. Ohne Randomisierung keine Mittelwerte.
- C** ☐ Randomisierung sorgt für Strukturgleichheit und erlaubt erst von der Stichprobe auf die Grundgesamtheit zurückzuschliessen.
- D** ☐ Randomisierung bringt starke Unstrukturiertheit in das Experiment und erlaubt erst von der Stichprobe auf die Grundgesamtheit zurückzuschliessen.
- E** ☐ Randomisierung erlaubt erst die Varianzen zu schätzen. Ohne eine Randomisierung ist die Berechnung von Mittelwerten und Varianzen nicht möglich.

#### 5 Aufgabe

(2 Punkte)

Welche statistische Masszahl erlaubt es *Relevanz* mit *Signifikanz* zuverbinden? Welche Aussage ist richtig?

- A** ☐ Das OR. Als Chancenverhältnis gibt es das Verhältnis von Relevanz und Signifikanz wieder.
- B** ☐ Das Konfidenzintervall. Durch die Visualisierung des Konfidenzintervalls kann eine Relevanzschwelle vom Anwender definiert werden. Zusätzlich erlaubt das Konfidenzintervall auch eine Entscheidung über die Signifikanz.
- C** ☐ Der p-Wert. Durch den Vergleich mit  $\alpha$  lässt sich über die Signifikanz entscheiden und der  $\beta$ -Fehler erlaubt über die Power eine Einschätzung der Relevanz.
- D** ☐ Die Teststatistik. Durch den Vergleich von  $T_c$  zu  $T_k$  ist es möglich die  $H_0$  abzulehnen. Die Relevanz ergibt sich aus der Fläche rechts vom dem  $T_c$ -Wert.
- E** ☐ Das  $\Delta$ . Durch die Effektstärke haben wir einen Wert für die Relevanz, die vom Anwender bewertet werden muss. Da  $\Delta$  antiproportional zum p-Wert ist, bedeutet auch ein hohes  $\Delta$  ein sehr kleinen p-Wert.

#### 6 Aufgabe

(2 Punkte)

Berechnen Sie den Mittelwert und Standardabweichung von y mit 12, 12, 16, 10 und 7.

- A** ☐ Es ergibt sich 12.4 +/- 1.645
- B** ☐ Es ergibt sich 10.4 +/- 5.4
- C** ☐ Es ergibt sich 11.4 +/- 1.645
- D** ☐ Es ergibt sich 11.4 +/- 10.8
- E** ☐ Es ergibt sich 11.4 +/- 3.29

#### 7 Aufgabe

(2 Punkte)

Beim statistischen Testen wird signal mit noise zur Teststatistik T verrechnet. Welche der Formel berechnet korrekt die Teststatistik T?

- A** ☐ Es gilt  $T = \text{signal} \cdot \text{noise}$
- B** ☐ Es gilt  $T = \frac{\text{signal}}{\text{noise}}$
- C** ☐ Es gilt  $T = \frac{\text{noise}}{\text{signal}}$

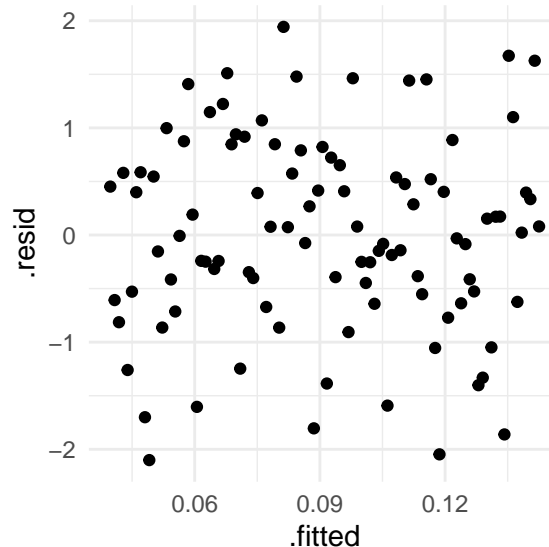
**D** ☐ Es gilt  $T = (\text{signal} \cdot \text{noise})^2$

**E** ☐ Es gilt  $T = \frac{\text{signal}}{\text{noise}^2}$

## 8 Aufgabe

(2 Punkte)

Sie rechnen eine lineare Regression und erhalten folgenden Residual Plot. Welche Aussage ist richtig?



**A** ☐ Die Annahme der normalverteilten Residuen ist nicht erfüllt. Es ist kein Muster zu erkennen.

**B** ☐ Die Annahme der normalverteilten Residuen ist erfüllt. Kein Muster ist zu erkennen und keine Outlier zu beobachten.

**C** ☐ Die Annahme der normalverteilten Residuen ist erfüllt. Es ist ein Muster zu erkennen.

**D** ☐ Die Annahme der normalverteilten Residuen ist erfüllt. Die Punkte liegen zum überwiegenden Teil auf der Diagonalen.

**E** ☐ Die Annahme der normalverteilten Residuen ist nicht erfüllt. Vereinzelt Punkte liegen oberhalb bzw. unterhalb der Geraden um die 0 Linie weiter entfernt. Ein klares Muster ist zu erkennen.

## 9 Aufgabe

(2 Punkte)

Der Fehler 1. Art oder auch Signifikanzniveau  $\alpha$  genannt, liegt bei 5%. Welcher der folgenden Gründe für diese Festlegung auf 5% ist richtig?

**A** ☐ Im Rahmen eines langen Disputs zwischen Neyman und Fischer wurde  $\alpha = 5\%$  festgelegt. Leider werden die Randbedingungen und Voraussetzungen an statistische Modelle heute immer wieder ignoriert.

**B** ☐ Auf einer Statistikkonferenz in Genf im Jahre 1942 wurde dieser Cut-Off nach langen Diskussionen festgelegt. Bis heute ist der Cut Off aber umstritten, da wegen dem 2. Weltkrieg viele Wissenschaftler nicht teilnehmen konnten.

**C** ☐ Der Begründer der modernen Statistik, R. Fischer, hat die Grenze simuliert und berechnet. Dadurch ergibt sich dieser optimale Cut-Off.

**D** ☐ Der Wert ergab sich aus einer Auswertung von 1042 wissenschaftlichen Veröffentlichungen zwischen 1914 und 1948. Der Wert 5% wurde in 28% der Veröffentlichungen genutzt. Daher legte man sich auf diese Zahl fest.

**E** ☐ Die Festlegung von  $\alpha = 5\%$  ist eine Kulturkonstante. Wissenschaftler benötigt eine Schwelle für eine statistische Testentscheidung, der Wert von  $\alpha$  wurde aber historisch mehr zufällig gewählt.

## 10 Aufgabe

(2 Punkte)

Die empfohlene Mindestanzahl an Beobachtungen für ein Histogramm sind...

- A** ☐ 1 Beobachtung.
- B** ☐ mindestens 20 Beobachtungen.
- C** ☐ 2-5 Beobachtungen.
- D** ☐ 10 Beobachtungen.
- E** ☐ 5 und mehr Beobachtungen.

## 11 Aufgabe

(6 Punkte)

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



Nach einem Gewächshausexperiment zählen Sie folgende Anzahl an Läsionen auf den Blättern von Brokoli nach einer durchgestandenen Infektion der Pflanze.

11, 6, 3, 3, 1, 2, 4, 2, 3, 6, 3, 2, 2, 5, 5

1. Zeichnen Sie ein Histogramm um die Verteilung der Daten zu visualisieren! **(3 Punkte)**
2. Beschriften Sie die Achsen der Abbildung! **(2 Punkte)**
3. Ergänzen Sie die relativen Häufigkeiten in der Abbildung! **(1 Punkt)**



## 12 Aufgabe

(6 Punkte)



Nach einer Bonitur von Kartoffeln mit einer Kontrolle und drei Pestiziden (*ctrl*, *pestKill*, *roundUp* und *zeroX*) ergibt sich die folgende Datentabelle mit den Boniturnoten (*grade*).

pesticide	grade
pestKill	4
roundUp	6
roundUp	2
pestKill	2
pestKill	5
zeroX	3
zeroX	4
roundUp	4
pestKill	5
ctrl	5
ctrl	7
ctrl	4
zeroX	3
roundUp	7

1. Zeichnen Sie in *einer* Abbildung die Dotplots für die vier Pestizidlevel! Beschriften Sie die Achsen entsprechend! **(4 Punkte)**
2. Ergänzen Sie die Dotplots mit der gängigen statistischen Maßzahl! **(1 Punkt)**
3. Wenn Sie *keinen Effekt* zwischen den Pestizidlevel erwarten würden, wie sehen dann die Dotplots aus? **(1 Punkt)**

### 13 Aufgabe

(12 Punkte)

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



Nach einem Experiment mit zwei Futtermitteln (*FatDown* und *ProGain*) an Puten ergibt sich die folgende Datentabelle mit den gemessenen Gewichtszunahmen nach fünf Wochen Mast.


feed	weight
FatDown	15
ProGain	15
FatDown	17
ProGain	14
ProGain	15
FatDown	16
FatDown	18
FatDown	19
ProGain	16
FatDown	17
ProGain	14
ProGain	14
ProGain	15

1. Formulieren Sie die wissenschaftliche Fragestellung! **(1 Punkt)**
2. Formulieren Sie das statistische Hypothesenpaar! **(2 Punkte)**
3. Bestimmen Sie die Teststatistik  $T_{calc}$  eines Welch t-Tests für den Vergleich der beiden Futtermittel! **(4 Punkte)**
4. Treffen Sie mit  $T_{\alpha=5\%} = 1.77$  und dem berechneten  $T_{calc}$  eine Aussage zur Nullhypothese! **(1 Punkt)**
5. Berechnen Sie das 95% Konfidenzintervall unter der Verwendung von  $s_p$  und der gemittelten Fallzahl über die beiden Gruppen! **(3 Punkte)**
6. Nennen Sie den statistischen Grund, warum Sie sich zwischen einem Student t-Test und einem Welch t-Test entscheiden müssen! **(1 Punkt)**


## 14 Aufgabe

(10 Punkte)



Sie rechnen eine lineare Regression um nach einem Feldexperiment den Zusammenhang zwischen Trockengewicht  $\text{kg/m}^2$  (*drymatter*) und Wassergabe  $\text{l/m}^2$  (*water*) bei Erdbeerpflanzen zu bestimmen. Sie erhalten folgende  Ausgabe.

```
##
## Call:
## lm(formula = weight ~ water, data = data_tbl)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5714  -2.0556   0.4444   2.4286   6.4444
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   19.571      1.773   11.041  2.7e-08
## waterB         7.984      2.364    3.378  0.00451
##
## Residual standard error: 4.69 on 14 degrees of freedom
## Multiple R-squared:  0.4491, Adjusted R-squared:  0.4097
## F-statistic: 11.41 on 1 and 14 DF, p-value: 0.004505
```

1. Ist die Annahme der Normalverteilung an das Outcome *water* erfüllt? Begründen Sie die Antwort! **(2 Punkte)**
2. Wie groß ist der Effekt der Wassergabe? Begründen Sie Ihre Antwort! **(2 Punkte)**
3. Liegt ein signifikanter Effekt vor? Begründen Sie Ihre Antwort! **(2 Punkte)**
4. Erklären Sie *kurz* den Begriff R-squared! Was sagt Ihnen der Wert 0.4491 aus? **(2 Punkte)**
5. Schreiben Sie das Ergebnis der  Ausgabe in einen Satz nieder, der die Information zum Effekt und der Signifikanz enthält! **(2 Punkte)**

## 15 Aufgabe

(11 Punkte)

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



Nach einem Feldexperiment ergibt sich die folgende 2x2 Datentabelle mit einem Pestizid (ja/nein) der Marke CleanX, dargestellt in den Zeilen, und dem infizierten Pflanzenstatus (ja/nein) von Kohl, dargestellt in den Spalten. Insgesamt wurden  $n = 122$  Pflanzen untersucht.

	Erkrankt (ja)	Erkrankt (nein)	
Pestizid (ja)	38	19	
Pestizid (nein)	27	38	

1. Formulieren Sie die wissenschaftliche Fragestellung! **(1 Punkt)**
2. Ergänzen Sie die Tabelle um die Randsummen! **(1 Punkt)**
3. Berechnen Sie die Teststatistik eines Chi-Quadrat-Test auf der 2x2 Tafel! **(3 Punkte)**
4. Treffen Sie eine Entscheidung im Bezug zu der Nullhypothese gegeben einem  $\chi^2_{\alpha=5\%} = 3.841$ ! Begründen Sie Ihre Antwort! **(2 Punkte)**
5. Skizzieren Sie die  $\chi^2$ -Verteilung, wenn die  $H_0$  wahr ist! Ergänzen Sie  $\chi^2_{\alpha=5\%}$  und  $\chi^2_{calc}$  in der Abbildung! **(2 Punkte)**
6. Berechnen Sie den Effektschätzer *Cramers V*! Interpretieren Sie den Effektschätzer! **(2 Punkte)**

## 16 Aufgabe

(10 Punkte)



Sie rechnen einen t-Test für Gruppenvergleiche der Mittelwerte. 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. Skizzieren Sie in die untenstehende Abbildung sechs einzelne Konfidenzintervalle (a-f) mit den jeweiligen Eigenschaften! **(6 Punkte)**
  - (a) Ein 95% Konfidenzintervall mit höherer Varianz  $s_p$  in der Stichprobe als der Rest der 95% Konfidenzintervalle
  - (b) Ein nicht signifikantes, nicht relevantes 95% Konfidenzintervall
  - (c) Ein 95% Konfidenzintervall mit niedriger Varianz  $s_p$  in der Stichprobe als der Rest 95% der Konfidenzintervalle
  - (d) Ein signifikantes, nicht relevantes 95% Konfidenzintervall
  - (e) Ein signifikantes, relevantes 99% Konfidenzintervall.
  - (f) Ein signifikantes, relevantes 95% Konfidenzintervall



## 17 Aufgabe

(9 Punkte)



Der Datensatz *crop\_tbl* enthält das Outcome *freshmatter* für ein Experiment mit Maispflanzen, welches unter drei verschiedenen Düngerbedingungen erzielt wurden. Die Düngerbedingungen sind in dem Faktor *trt* mit den Faktorstufen *A*, *B* und *C* codiert. Sie erhalten folgenden Output in .

```
## Analysis of Variance Table
##
## Response: freshmatter
##           Df Sum Sq Mean Sq F value    Pr(>F)
## trt         2  85.262   42.631   29.514 5.911e-07
## Residuals  22  31.778    1.444
```

1. Stellen Sie die statistische  $H_0$  und  $H_A$  Hypothese für die obige einfaktorielle ANOVA auf! **(2 Punkte)**
2. Interpretieren Sie das Ergebnis der einfaktoriellen ANOVA! **(2 Punkt)**
3. Berechnen Sie den Effektschätzer  $\eta^2$ . Was sagt Ihnen der Wert von  $\eta^2$  aus? **(2 Punkte)**
4. Skizzieren Sie eine Abbildung, der dem obigen Ergebnis der einfaktoriellen ANOVA näherungsweise entspricht! **(3 Punkte)**