

## Checkbox für die Version vom 22. November 2022

Die gesamte Klausur beinhaltet aktuell in Summe **109** Fragen. Davon sind **47** Multiple Choice Fragen sowie **62** Rechen- und Textaufgaben.

## Frequently asked questions (FAQ)

**Was ist das hier?** Im Folgenden findet sich die Sammlung *aller* Klausurfragen der Bio Data Science über *alle* Veranstaltungen, die ich an der Fakultät für Agrarwissenschaften und Landschaftsarchitektur anbiete.

**Sind aber ein bisschen viele Fragen...** Ja, das stimmt. Die Sortierung und Überlegung welche Fragen zur Veranstaltung passen obliegt dem Studierenden. Gerne stehe ich für Rückfragen bereit. Teilweise sind Fragen auch ähnlich.

**Sind die Fragen fix?** Ein klares Jein. Die Zahlen und die *Reihenfolge* der Aufgaben - auch im Multiple Choice Teil - werden sich ändern, da die Klausurfragen zufällig erstellt werden. Die *Aufgabenfragen* hindoch werden die gleichen Fragen bleiben.

**Okay, aber woher weiß ich jetzt welche Fragen zu meiner Veranstaltung gehören?** Das ist der Trick. Durch das Durchlesen und das selbstständige Sortieren der Fragen zu Themen und Inhalten merkt man ziemlich schnell, welche Inhalte zu der Veranstaltung gehören und welche nicht. Ist also alles Teil des Lernprozesses. *Und* wenn Unsicherheiten da sind, gerne in der Wiederholungsveranstaltung - letzte Vorlesung - einfach mich fragen.

**Wie sieht denn die finale Klausur aus?** Die Klausur hat am Ende 10 Multiple Choice Fragen mit jeweils 2 Punkten sowie Rechen- und Textaufgaben mit in Summe ca. 60 Punkten. Ich peile daher eine Klausur mit 80 Punkten an, wobei 40 Punkte zum Bestehen der Klausur notwendig sind. **Bei geteilten Veranstaltungen mit mehr als einem Dozenten ändert sich die Zusammensetzung der endgültigen Punkteanzahl!**

**Sind aber mehr als zehn Multiple Choice Fragen...** Ja, aber es werden in der finalen Klausur nur zehn Multiple Choice Fragen sein. Ich wähle die Fragen dann zufällig aus. Ich berücksichtige natürlich die Veranstaltung und das Lernniveau.

**Solange kann ich nicht warten...** Dann einfach eine Mail an mich schreiben:  
[j.kruppa@hs-onsabrueck.de](mailto:j.kruppa@hs-onsabrueck.de)

Ich versuche dann die Frage kurzfristig zu beantworten oder aber in der Vorlesung nochmal (anonym) aufzugreifen.

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

*Nicht bestanden:* ☐

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

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

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

# Klausurfragen der Bio Data Science

**Hochschule Osnabrück**

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

Version vom 22. November 2022

## 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.

## Ergebnis der Klausur

\_\_\_\_\_ von 20 Punkten sind aus dem Multiple Choice Teil erreicht.  
\_\_\_\_\_ von 60 Punkten sind aus dem Rechen- und Textteil erreicht.  
\_\_\_\_\_ von 80 Punkten in Summe.

Es wird folgender Notenschlüssel angewendet.

<b>Punkte</b>	<b>Note</b>
78 - 80	1,0
75 - 77	1,3
70 - 74	1,7
65 - 69	2,0
59 - 64	2,3
54 - 58	2,7
49 - 53	3,0
44 - 48	3,3
41 - 43	3,7
40	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							

- Es sind \_\_\_\_ von 60 Punkten erreicht worden.

## 1 Aufgabe

(2 Punkte)

Welche Aussage über die  $\alpha$  Adjustierung ist richtig?

- A ☐ Die  $\alpha$  Adjustierung wird meist ignoriert. Wenn die Annahmen an den statistischen Test richtig sind, kann auf eine Adjustierung verzichtet werden.
- B ☐ Die  $\alpha$  Adjustierung wird durchgeführt um den Fehler 2. Art zu kontrollieren. Ohne diese Adjustierung würde der Fehler 2. Art nicht bei 80% liegen sondern sehr schnell gegen 0 laufen.
- C ☐ Die  $\alpha$  ist notwendig um Effekte gegeneinander aufzurechnen. Ohne diese Adjustierung würde der eigentliche Effekt nicht richtig geschätzt. Daher handelt es sich um eine Adjustierung der Fehlerwahrscheinlichkeiten.
- D ☐ Die  $\alpha$  Adjustierung wird durchgeführt um bei multiplen Vergleichen den Fehler 1. Art zu kontrollieren. Es wird die Irrtumswahrscheinlichkeit adjustiert, daher das  $\alpha$ -Niveau.
- E ☐ Die  $\alpha$  Adjustierung wird durchgeführt um den Effekt von Interesse, meist die Behandlung, von anderen Effekten zu trennen. Daher eine Adjustierung auf den  $\beta$ -Werten einer Regression.

## 2 Aufgabe

(2 Punkte)

Sie haben folgende unadjustierten p-Werte gegeben: 0.03, 0.42, 0.001, 0.02, 0.34 und 0.89. Sie adjustieren die p-Werte nach Bonferroni. Welche Aussage ist richtig?

- A ☐ Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 0.005, 0.07, 0.0002, 0.0033, 0.0567 und 0.1483. Die adjustierten p-Werte werden zu einem  $\alpha$ -Niveau von 0.83% verglichen.
- B ☐ Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 0.18, 1, 0.006, 0.12, 1 und 1. Die adjustierten p-Werte werden zu einem  $\alpha$ -Niveau von 5% verglichen.
- C ☐ Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 0.18, 1, 0.006, 0.12, 1 und 1. Die adjustierten p-Werte werden zu einem  $\alpha$ -Niveau von 0.83% verglichen.
- D ☐ Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 0.005, 0.07, 0.0002, 0.0033, 0.0567 und 0.1483. Die adjustierten p-Werte werden zu einem  $\alpha$ -Niveau von 5% verglichen.
- E ☐ Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 0.18, 2.52, 0.006, 0.12, 2.04 und 5.34. Die adjustierten p-Werte werden zu einem  $\alpha$ -Niveau von 5% verglichen.

## 3 Aufgabe

(2 Punkte)

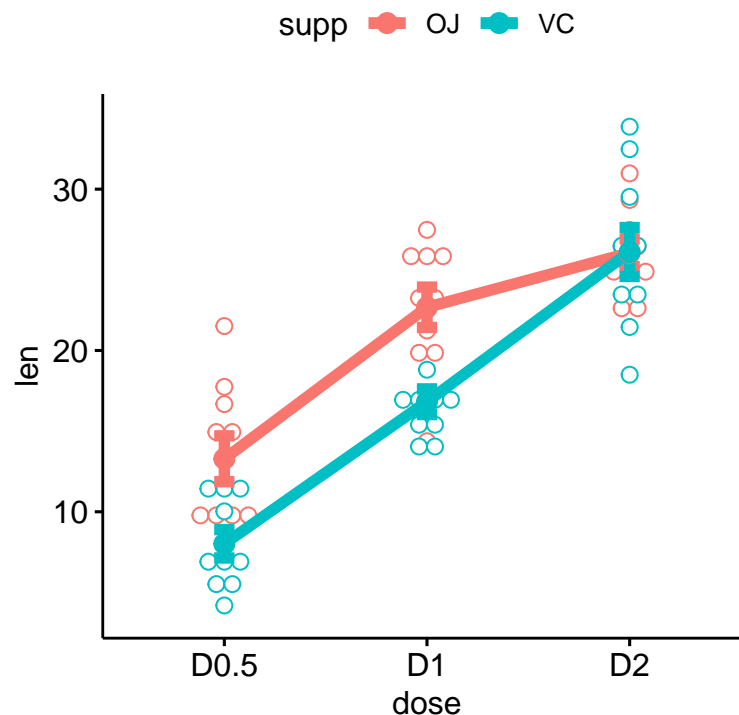
Der Datensatz PlantGrowth enthält das Gewicht von Pflanzen, die unter einer Kontrolle und zwei verschiedenen Behandlungsbedingungen erzielt wurden. Nach der Berechnung einer einfaktoriellen ANOVA ergibt sich ein  $\eta^2 = 0.2$ . Welche Aussage ist richtig?

- A ☐ Die Berechnung von  $\eta^2$  ist ein Wert für die Interaktion.
- B ☐ 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.
- C ☐ Das  $\eta^2$  ist die Korrelation der ANOVA. Mit der Ausnahme, dass 0 der beste Wert ist.
- D ☐ 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.
- E ☐ Das  $\eta^2$  beschreibt den Anteil der Varianz, der von den Behandlungsbedingungen nicht erklärt wird. Somit der Rest an nicht erklärbarer Varianz.

## 4 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 2 mg/Tag) über eine von zwei Verabreichungsmethoden erhielt (Orangensaft oder Ascorbinsäure (eine Form von Vitamin C und als VC codiert)). Die Zahnlänge wurde gemessen, und eine Auswahl der Daten ist unten dargestellt.



Welche Aussage ist richtig im Bezug auf eine zweifaktorielle ANOVA?

- A** ☐ Ein signifikanter Effekt ist nicht zu erwarten. Durch das Kreuzen der OJ und VC Geraden an der dose Faktorstufe D2 ist eine Interaktion vorhanden.
- B** ☐ Ein signifikanter Effekt ist zwischen allen dose Faktorstufen zu erwarten. Durch das Kreuzen der OJ und VC Geraden an der dose Faktorstufe D2 ist keine Interaktion vorhanden.
- C** ☐ Der  $\eta^2$ -Wert sollte liegt bei 1, da eine Gerade zu beobachten ist. Dadurch ist die Interaktion zwischen den dose Faktorstufen nicht mehr signifikant. Der Faktor supp hat keinen Einfluss.
- D** ☐ Ein signifikanter Effekt ist zwischen allen dose Faktorstufen nicht zu erwarten. Eine Interaktion liegt nicht vor. Der  $\eta^2$ -Wert sollte bei ca. 0 liegen.
- E** ☐ Ein signifikanter Effekt ist zwischen mindestens einer dose Faktorstufen zu erwarten. Durch das Kreuzen der OJ und VC Geraden an der dose Faktorstufe D2 ist mit einem signifikanten Interaktionsterm zu rechnen.

## 5 Aufgabe

(2 Punkte)

Sie haben das abstrakte Modell  $Y \sim X$  mit  $X$  als Faktor mit zwei Leveln vorliegen. Welche Aussage über  $n_1 < n_2$  ist richtig?

- A** ☐ Es liegt Varianzheterogenität vor.
- B** ☐ Es handelt sich um ein balanciertes Design.

- C** ☐ Es handelt sich um ein unbalanciertes Design
- D** ☐ Es handelt sich um abhängige Beobachtungen.
- E** ☐ Es liegt Varianzhomogenität vor.

## 6 Aufgabe

(2 Punkte)

Die Mindestanzahl an Beobachtungen für eine Zelle der Vierfeldertafel bei der Nutzung eines Chi-Quadrat-Testes ist...

- A** ☐ 0 Beobachtungen
- B** ☐ 10 Beobachtungen
- C** ☐ 5 Beobachtungen
- D** ☐ 2 Beobachtungen
- E** ☐ 1 Beobachtung

## 7 Aufgabe

(2 Punkte)

Berechnen Sie die Korrelation  $r$  zwischen  $x$  mit 15, 11, 12, 10 und 8 sowie  $y$  mit 17, 12, 2, 13 und 8. Nutzen Sie folgende Formel:

$$r_{x,y} = \frac{s_{x,y}}{s_x \cdot s_y}$$

- A** ☐ Es ergibt sich eine Korrelation von 0.14
- B** ☐ Es ergibt sich eine Korrelation von 0.38
- C** ☐ Es ergibt sich eine Korrelation von -0.38
- D** ☐ Es ergibt sich eine Korrelation von 0.05
- E** ☐ Es ergibt sich eine Korrelation von 1.38

## 8 Aufgabe

(2 Punkte)

Berechnen Sie die Kovarianz  $s_{x,y}$  und die Korrelation  $r$  zwischen  $x$  mit 10, 14, 9, 5 und 7 sowie  $y$  mit 18, 14, 10, 13 und 14.

$$r_{x,y} = \frac{s_{x,y}}{s_x \cdot s_y}$$

- A** ☐ Es ergibt sich eine Kovarianz von 2 sowie eine Korrelation von 0.21
- B** ☐ Es ergibt sich eine Kovarianz von 4 sowie eine Korrelation von 0.04
- C** ☐ Es ergibt sich eine Kovarianz von -2 sowie eine Korrelation von 0.04
- D** ☐ Es ergibt sich eine Kovarianz von 2 sowie eine Korrelation von -0.21
- E** ☐ Es ergibt sich eine Kovarianz von 8 sowie eine Korrelation von -0.21

## 9 Aufgabe

(2 Punkte)

Berechnen Sie den Mittelwert und Standardabweichung von  $y$  mit 12, 10, 9, 8 und 7.

- A ☐ Es ergibt sich  $9.2 \pm 0.96$
- B ☐ Es ergibt sich  $10.2 \pm 0.96$
- C ☐ Es ergibt sich  $9.2 \pm 3.7$
- D ☐ Es ergibt sich  $8.2 \pm 1.85$
- E ☐ Es ergibt sich  $9.2 \pm 1.92$

## 10 Aufgabe

(2 Punkte)

Berechnen Sie den Median und das IQR von  $x$  mit 30, 11, 17, 16, 26, 19 und 42.

- A ☐ Es ergibt sich  $19 \pm 30$
- B ☐ Es ergibt sich 23 [16, 30]
- C ☐ Es ergibt sich 19 [16, 30]
- D ☐ Es ergibt sich  $23 \pm 16$
- E ☐ Es ergibt sich  $19 \pm 16$

## 11 Aufgabe

(2 Punkte)

Eine der gängigsten Methode der Statistik um einen Fehler zu bestimmen ist...

- A ☐ ... die kleinste Quadrate Methode oder auch least square method genannt.
- B ☐ ... die Methode des absoluten, quadrierten Abstands.
- C ☐ ... die Methode der aufaddierten, absoluten Abstände.
- D ☐ ... die Methode des absoluten Abstands.
- E ☐ ... das Produkt der kleinsten Quadrate.

## 12 Aufgabe

(2 Punkte)

Welche Aussage über *ordinary mean* und *marginal mean* ist richtig?

- A ☐ Der Begriff *ordinary mean* beschreibt die Mittelwerte eines Faktors wohingegen der *marginal mean* die einzelnen Level betrachtet.
- B ☐ Der Begriff *ordinary mean* beschreibt einen unbedeutenden Mittelwert und *marginal mean* einen Randmittelwert.
- C ☐ Der Begriff *ordinary mean* beschreibt die einzelnen Mittelwerte der Level eines Faktors wohingegen der *marginal mean* den Mittelwert über alle Level des Faktors wiedergibt.
- D ☐ Der Begriff *ordinary mean* beschreibt einen einzelnen Mittelwert eines Levels eines Faktors wohingegen der *marginal mean* die unbedeutenden Mittelwerte beschreibt.
- E ☐ Der Begriff *ordinary mean* beschreibt gewöhnliche Mittelwerte und wohingegen *marginal mean* kleine Mittelwerte beschreibt.



### 13 Aufgabe

(2 Punkte)

Die empfohlene Mindestanzahl an Beobachtungen für einen Dotplot sind...

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

### 14 Aufgabe

(2 Punkte)

Mit einem Histogramm wird Folgendes in der Statistik und der explorativen Datenanalyse hauptsächlich dargestellt.

- A ☐ Die Verteilung von Daten. Meistens dem Outcome oder auch  $y$  genannt. Die Darstellung ist auch mit einer kleinen Fallzahl möglich.
- B ☐ Die Verteilung von Daten. Auch 2 bis 5 Beobachtungen können noch dargestellt werden. Auch wenn es sich streng genommen um keine Darstellung einer Verteilung handelt.
- C ☐ Die Verteilung von Daten. Meistens dem Outcome oder auch  $y$  genannt. Hierbei ist eine große Fallzahl notwendig.
- D ☐ Die Eigenschaften von Daten aufgeteilt nach zwei Gruppen eines Faktors  $x$ .
- E ☐ Die Eigenschaften von Daten anhand der fehlenden Werte und den Levels eines Faktors  $x$ .

### 15 Aufgabe

(2 Punkte)

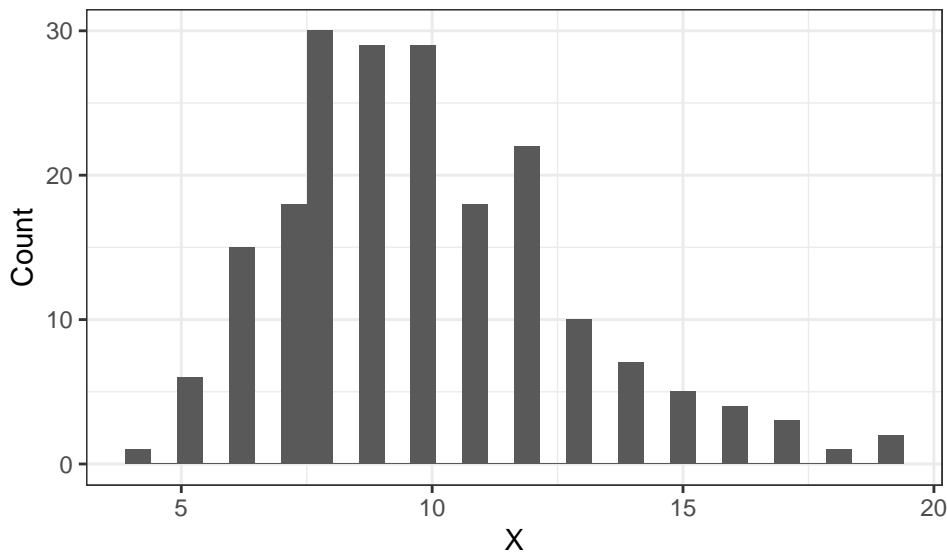
Price et al. (2016) untersuchte die Auswirkungen des Bergbaus und der Talauffüllung auf den Bestand und die Häufigkeit von Bachsalamandern. Um den Effekt zu Berechnen nutze Price et al. (2016) eine Poisson-Regression auf die Anzahl an aufgefundenen Bachsalamandern an den jeweiligen Suchorten. Welche Aussage zur Poisson-Regression auf Zähldaten ist richtig?

- A ☐ Die Poisson-Regression schätzt nur einen Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzten Overdispersion von 3.12 liegt Overdispersion vor. Damit kann keine Poisson-Regression gerechnet werden. Die Lösung ist eine Gaussian Regression mit Nullanpassung.
- B ☐ Die Poisson-Regression schätzt nur einen Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzten Overdispersion von 3.12 liegt keine Overdispersion vor. Overdispersion liegt vor, wenn die geschätzte Overdispersion unter 1 liegt.
- C ☐ Die Poisson-Regression schätzt drei Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzten Overdispersion von 3.12 liegt Overdispersion vor. Die Lösung ist die Nutzung von nur zwei der drei Verteilungsparameter:  $\gamma_1$  und  $\gamma_3$ .
- D ☐ Die Poisson-Regression schätzt zwei Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzten Overdispersion von 3.12 liegt keine Overdispersion vor.
- E ☐ Die Poisson-Regression schätzt nur einen Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzten Overdispersion von 3.12 liegt Overdispersion vor. Die Lösung ist die Nutzung einer anderen Verteilungsfamilie wie die Quasipoisson Verteilung.

## 16 Aufgabe

(2 Punkte)

In dem folgenden Histogramm von  $n = 200$  Pflanzen ist welche Verteilung mit welchen korrekten Verteilungsparametern dargestellt?



- ☐ A Es handelt sich um eine Poisson-Verteilung mit  $\text{Pois}(10)$ .
- ☐ B Eine Standardnormalverteilung mit  $N(0,1)$ .
- ☐ C Eine rechtsschiefe, multivariate Normalverteilung.
- ☐ D Es handelt sich um eine Binomial-Verteilung mit  $\text{Binom}(10)$ .
- ☐ E Es handelt sich um eine Normalverteilung mit  $N(10, 5)$ .

## 17 Aufgabe

(2 Punkte)

Price et al. (2016) untersuchte die Auswirkungen des Bergbaus und der Talauffüllung auf den Bestand und die Häufigkeit von Bachsalamandern. Um den Effekt zu Berechnen nutze Price et al. (2016) eine Poisson-Regression auf die Anzahl an aufgefundenen Bachsalamandern an den jeweiligen Suchorten. Nach einer statistischen Beratung wurde ihm nahegelegt auf Overdispersion zu achten, wenn er statistische Aussagen zur Signifikanz treffen will. Price et al. (2016) schätzt zwei Modelle. Modell 1 mit einer Poisson Verteilung und Modell 2 mit einer Quasi-Poisson Verteilung. Welche Aussage zu einer geschätzten Overdispersion von 3.37 ist richtig?

- ☐ A Das Vergleichen von verschiedenen Modellen muss erst über ein AIC Kriterium erfolgen. Die Abschätzung über die Overdispersion ist nicht notwendig. Die Varianzen werden später in einer ANOVA adjustiert. Die Confounder Adjustierung.
- ☐ B Bei einer geschätzten Overdispersion höher als 1.5 ist von Overdispersion in den Daten auszugehen. Daher wird die Varianz systematisch unterschätzt, was zu höheren p-Werten führt. Daher gibt es weniger signifikante Ergebnisse als es in Wirklichkeit gibt. Daher ist das Modell 1 die bessere Wahl.
- ☐ C Bei einer geschätzten Overdispersion höher als 1.5 ist von Overdispersion in den Daten auszugehen. Daher wird die Varianz systematisch unterschätzt, was zu kleineren p-Werten führt. Daher gibt es mehr signifikante Ergebnisse als es in Wirklichkeit gibt. Daher ist das Modell 2 die bessere Wahl.
- ☐ D Bei einer geschätzten Overdispersion höher als 1.5 ist von keiner Overdispersion in den Daten auszugehen. Dennoch sind die p-Werte zu klein, dass diese p-Werte natürlich entstanden sein könnten. Die p-Werte müssen adjustiert werden.

- E** ☐ Bei einer geschätzten Overdispersion höher als 1.5 ist von Overdispersion in den Daten auszugehen. Daher wird die Varianz systematisch überschätzt, was zu höheren p-Werten führt. Daher gibt es mehr signifikante Ergebnisse als es in Wirklichkeit gibt. Daher ist das Modell 1 die bessere Wahl

## 18 Aufgabe

(2 Punkte)

In einem Zuchtexperiment messen wir die Ferkel verschiedener Sauen. Die Ferkel einer Muttersau sind daher im statistischen Sinne...

- A** ☐ Untereinander abhängig. Die Ferkel stammen von einem Muttertier und haben vermutlich eine ähnliche Varianzstruktur.
- B** ☐ Untereinander unabhängig. Sollten die Mütter verwandt sein, so ist die Varianzstruktur ähnlich und muss modelliert werden.
- C** ☐ Untereinander abhängig, wenn die Mütter ebenfalls miteinander verwandt sind. Erst die Abhängigkeit 2. Grades wird in der Statistik modelliert.
- D** ☐ Untereinander stark korreliert. Die Ferkel sind von einer Mutter und somit miteinander korreliert. Dies wird in der Statistik jedoch meist nicht modelliert.
- E** ☐ Untereinander unabhängig. Die Ferkel sind eigenständig und benötigen keine zusätzliche Behandlung.

## 19 Aufgabe

(2 Punkte)

Sie haben das abstrakte Modell  $Y \sim X$  vorliegen. Welche Aussage über  $Y$  ist richtig?

- A** ☐  $Y$  beinhaltet eine Spalte. Die Spalte gibt die Verteilungsfamilie vor.
- B** ☐  $Y$  beinhaltet die Zeilen. Die Zeilen geben die Verteilungsfamilie vor.
- C** ☐  $Y$  beinhaltet eine Spalte. Die Spalte gibt nicht die Verteilungsfamilie vor.
- D** ☐  $Y$  beinhaltet mehrere Spalten. Die Spalten enthalten die Behandlung und weitere potenzielle Einflussvariablen
- E** ☐  $Y$  beinhaltet mehrere Spalten. Die Spalten geben die Verteilungsfamilie vor.

## 20 Aufgabe

(2 Punkte)

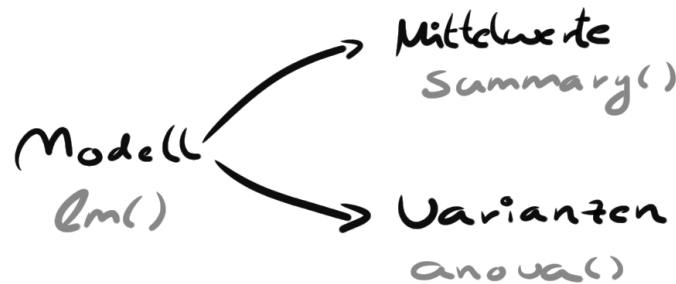
Sie haben das Modell  $Y \sim X$  vorliegen und wollen nun ein prädiktives Modell rechnen. Welche Aussage ist richtig?

- A** ☐ Ein prädiktives Modell benötigt mindestens eine Fallzahl von über 100 Beobachtungen und darf keine fehlenden Werte beinhalten. Die Varianzkomponenten müssen homogen sein.
- B** ☐ Ein prädiktives Modell wird auf einem Trainingsdatensatz trainiert und anschliessend über eine explorative Datenanalyse validiert. Signifikanzen über  $\beta_i$  können hier nicht festgestellt werden.
- C** ☐ Ein prädiktives Modell schliesst grundsätzlich lineare Modell aus. Es muss ein Graph gefunden werden, der alle Punkte beinhaltet. Erst dann kann das  $R^2$  berechnet werden.
- D** ☐ Ein prädiktives Modell basiert auf einem Trainingsdatensatz und einem Testdatensatz. Auf dem Trainingsdatensatz wird das Modell trainiert und auf dem Testdatensatz validiert.
- E** ☐ Ein prädiktives Modell möchte die Zusammenhänge von  $X$  auf  $Y$  modellieren. Hierbei geht es um die Effekte von  $X$  auf  $Y$ . Man sagt, wenn  $X$  um 1 ansteigt ändert sich  $Y$  um einen Betrag  $\beta$ .

## 21 Aufgabe

(2 Punkte)

In der folgenden Abbildung ist der Zusammenhang vom Modell zu der linearen Regression und der ANOVA skizziert.



Welche der folgenden Aussagen ist richtig?

- ☐ A Die Effektschätzer aus einem Modell, in diesem Fall ein polynomes Modell, erlauben es sowohl eine ANOVA zurechnen sowie auch eine Zusammenfassung der Mittelwerte zu betrachten.
- ☐ B Die Effektschätzer aus einem Modell, erlauben es nur einen Mittelwertsvergleich zu rechnen.
- ☐ C Die Effektschätzer aus einem Modell, in diesem Fall ein lineares Modell, erlauben es sowohl eine ANOVA zurechnen sowie auch eine Zusammenfassung der Mittelwerte zu betrachten.
- ☐ D Die Effektschätzer aus einem Modell, erlauben es nur eine ANOVA zu rechnen.
- ☐ E Die Effektschätzer aus einem Modell, in diesem Fall ein lineares Modell, erlauben es nur eine ANOVA zurechnen oder eine Zusammenfassung der Mittelwerte zu betrachten. Beides ist nicht möglich.

## 22 Aufgabe

(2 Punkte)

In der Statistik werden die Daten  $D$  modelliert in dem ein Modell der Form  $Y \sim X$  aufgestellt wird. Welche statistische Kenngrösse wird modelliert?

- ☐ A Die Varianzstruktur wird modelliert.
- ☐ B Die  $X$  werden modelliert.
- ☐ C Die Mittelwerte werden modelliert.
- ☐ D Die Varianz der  $X$  unabhängig vom  $Y$  wird modelliert.
- ☐ E Die  $Y$  werden modelliert.

## 23 Aufgabe

(2 Punkte)

Sie führen ein Experiment zur Behandlung von Klaueninfektionen bei Kühen durch. Bei 3 Tieren finden Sie eine Erkrankung der Klauen vor und 7 Tiere sind gesund. Welche Aussage über den Odds ratio Effektschätzer ist richtig?

- ☐ A Es ergibt sich ein Odds ratio von 2.33, da es sich um ein Anteil handelt.
- ☐ B Es ergibt sich ein Odds ratio von 0.3, da es sich um eine Chancenverhältnis handelt.
- ☐ C Es ergibt sich ein Odds ratio von 0.43, da es sich um eine Chancenverhältnis handelt.
- ☐ D Es ergibt sich ein Odds ratio von 0.43, da es sich um ein Anteil handelt.
- ☐ E Es ergibt sich ein Odds ratio von 0.3, da es sich um ein Anteil handelt.

## 24 Aufgabe

(2 Punkte)

Welche Aussage über die parametrische Statistik ist richtig?

- A ☐ Die parametrische Statistik basiert auf dem Schätzen von Parametern aus einer a priori festgelegten Verteilung. Daher gibt es auch direkt zu interpretierenden Effektschätzer.
- B ☐ Die parametrische Statistik basiert auf Rängen. Daher gibt es auch direkt zu interpretierenden Effektschätzer.
- C ☐ Die parametrische Statistik basiert auf dem Schätzen von Parametern aus einer festgelegten Verteilung. Daher gibt es auch direkt zu interpretierenden Effektschätzer.
- D ☐ Die nicht-parametrische Statistik ist ein Vorgänger der parametrischen Statistik und wurde wegen dem Mangel an Effektschätzern nicht mehr ab 1960 genutzt.
- E ☐ Die parametrische Statistik basiert auf Rängen. Daher wird jeder Zahl ein Rang zugeteilt. Nur auf den Rängen wird die Auswertung gerechnet. Daher gibt es auch keinen direkt zu interpretierenden Effektschätzer.

## 25 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 Varianzen zu schätzen. Ohne eine Randomisierung ist die Berechnung von Mittelwerten und Varianzen nicht möglich.
- C ☐ Randomisierung sorgt für Strukturgleichheit und erlaubt erst von der Stichprobe auf die Grundgesamtheit zurückzuschliessen.
- D ☐ Randomisierung erlaubt erst die Mittelwerte zu schätzen. Ohne Randomisierung keine Mittelwerte.
- E ☐ Randomisierung bringt starke Unstrukturiertheit in das Experiment und erlaubt erst von der Stichprobe auf die Grundgesamtheit zurückzuschliessen.

## 26 Aufgabe


(2 Punkte)


Wenn Sie einen Datensatz erstellen, dann ist es ratsam die Spalten und die Einträge in englischer Sprache zu verfassen, wenn Sie später die Daten in  auswerten wollen. Welcher folgende Grund ist richtig?

- A ☐ Es gibt keinen Grund nicht auch deutsche Wörter zu verwenden. Es ist ein Stilmittel.
- B ☐ Im Allgemeinen haben Programmiersprachen Probleme mit Umlauten und Sonderzeichen, die in der deutschen Sprache vorkommen. Eine Nutzung der englischen Sprache umgeht dieses Problem auf einfache Art.
- C ☐ Programmiersprachen können nur englische Begriffe verarbeiten. Zusätzliche Pakete können zwar geladen werden, aber meist funktionieren diese Pakete nicht richtig. Deutsch ist International nicht bedeutend genug.
- D ☐ Alle Funktionen und auch Anwendungen sind in  in englischer Sprache. Die Nutzung von deutschen Wörtern ist nicht schick und das ist zu vermeiden.
- E ☐ Die Spracherkennung von  ist nicht in der Lage Deutsch zu verstehen.

## 27 Aufgabe


(2 Punkte)

Das  Package `gtsummary` erlaubt viele Funktionalitäten auf einem Datensatz und deren Analyse. Welche Aussage zu der Funktion `tbl_summary()` ist richtig?

- ☐ A Die Funktion `tbl_summary()` erlaubt das Ergebnis einer Regressionsanalyse schnell in eine Übersichtstabelle umzuwandeln. Insbesondere die Zusammenfassung der Schätzwerte in eine Zeile macht die Analyse sehr praktisch.
- ☐ B Die Funktion `tbl_summary()` erlaubt das Ergebnis einer Regressionsanalyse schnell in eine Übersichtstabelle umzuwandeln. Dies funktioniert jedoch nur für simple lineare Regressionen. Multiple Regression sind nicht möglich.
- ☐ C Die Funktion `tbl_summary()` erlaubt einen Datensatz schnell in eine Übersichtstabelle umzuwandeln. Insbesondere die Einteilung der Spalten nach dem Strata der Behandlung macht die Funktion sehr nützlich.
- ☐ D Die Funktion `tbl_summary()` erlaubt in einem Setting in dem kein signifikantes Ergebnis vorliegt Daten in Form einer explorativen Datenanalyse darzustellen. Da dies ein komplizierter Schritt ist, wird von der Verwendung abgeraten.
- ☐ E Die Funktion `tbl_summary()` erlaubt einen Datensatz in eine Tabelle zusammenzufassen. Leider ist eine Stratifizierung nach der Behandlung nicht möglich. Deshalb nutzt man dann das  Package `tableone`.

## 28 Aufgabe

(2 Punkte)

In einem Stallexperiment mit  $n = 104$  Ferkeln wurden verschiedene Outcomes gemessen: der Gewichtszuwachs, Überleben nach 21 Tagen sowie Anzahl Verletzungen pro 7 Tagen. Zwei Lichtregime wurden als Einflussfaktor gemessen. Sie erhalten den  Output der Funktion `tidy()` einer simplen *poission* linearen Regression. Welche Aussage über den **Effekt** ist richtig?

term	estimate	std.error
(Intercept)	0.58	0.10
light_binhigh	0.16	0.14

- ☐ A In einer *poission* Regression muss für die Interpretation des Effektes das  $\beta_1$  quadriert werden. Somit liegt das OR bei 0.03
- ☐ B In einer *poission* Regression berechnet man das RR. Daher muss der Schätzer des Effektes  $\beta_1$  noch quadriert werden. Somit liegt das RR bei 0.03
- ☐ C Eine *poission* Regression basiert auf dem maximum Likelihood Prinzip. Hierbei kann kein Effekt beschrieben werden. Im Zweifel hilft aber eine Quadrierung der Fehlerquadrate  $\epsilon$ .
- ☐ D In einer *poission* Regression wird die Mittelwertsdifferenz betrachtet. Daher ist der Effekt zwischen den beiden Lichtregimen eine Gewichtsänderung von 0.16
- ☐ E In einer *poission* Regression kann kein Effekt roh interpretiert werden. Es muss erst eine Confounderadjustierung durchgeführt werden.

## 29 Aufgabe

(2 Punkte)

In einer linearen Regression werden die  $\epsilon$  oder Residuen geschätzt. Welcher Verteilung folgen die Residuen bei einer optimalen Modellierung?

- A ☐ Die Residuen sind normalverteilt mit  $\mathcal{N}(0, s^2)$ .
- B ☐ Die Residuen sind normalverteilt mit  $\mathcal{N}(0, 1)$ .
- C ☐ Die Residuen sind binomialverteilt.
- D ☐ Die Residuen sind normalverteilt mit  $\mathcal{N}(\bar{y}, s^2)$ .
- E ☐ Die Residuen folgen einer Poissonverteilung mit  $\text{Pois}(0)$ .

### 30 Aufgabe

(2 Punkte)

Welche Aussage über das *generalisierte lineare Modell (GLM)* ist richtig?

- A ☐ Das GLM ist ein faktisch maschineller Lernalgorithmus, der selbstständig die Verteilungsfamilie für Y wählt.
- B ☐ Das GLM erlaubt auch nicht normalverteilte Residuen in der Schätzung der Regressionsgrade.
- C ☐ Das GLM erlaubt auch weitere Verteilungsfamilien für das Y bzw. das Outcome in einer linearen Regression zu wählen.
- D ☐ Das GLM ist eine allgemeine Erweiterung der linearen Regression auf die Normalverteilung.
- E ☐ Das GLM ist eine Vereinfachung des LM in R. Mit dem GLM lassen polygonale Regressionen rechnen.

### 31 Aufgabe

(2 Punkte)

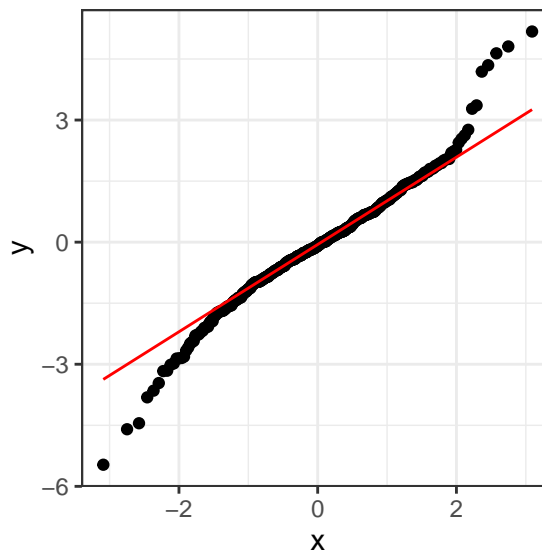
Sie rechnen in einer linearen Regression das Modell A und das Modell B. Nun stellt sich die Frage, welches der beiden Modelle das bessere Modell ist. Um die Modelle bewerten zu können berechnen Sie dafür das  $AIC_A$  für Modell A mit 653 und für das Modell B das  $AIC_B$  von 287. Welche Aussage über die beiden Modelle ist richtig?

- A ☐ Da  $AIC_A > AIC_B$  ist das Modell B das bessere Modell.
- B ☐ Da  $AIC_A < AIC_B$  ist das Modell B das bessere Modell.
- C ☐ Da  $AIC_A < AIC_B$  ist das Modell A das bessere Modell.
- D ☐ Da  $AIC_A > 0$  ist das Modell A das bessere Modell. Der AIC Wert für B wird verworfen.
- E ☐ Da  $AIC_A > AIC_B$  ist das Modell A das bessere Modell.

### 32 Aufgabe

(2 Punkte)

Sie rechnen in eine linearen Regression und erhalten folgenden QQ Plot. Welche Aussage ist richtig?

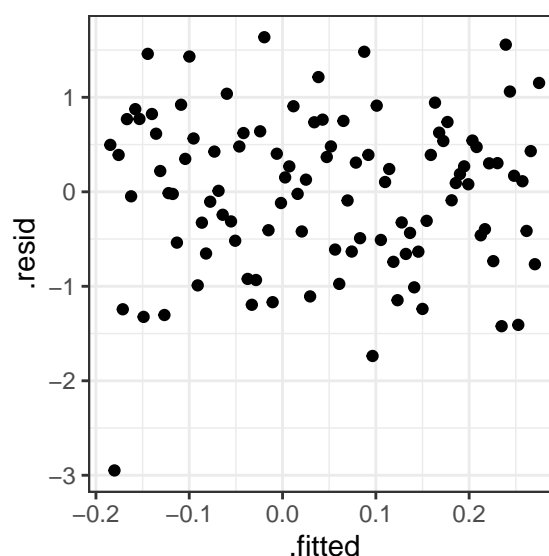


- A** ☐ Die Annahme der normalverteilten Residuen ist nicht erfüllt. Die Punkte liegen zum überwiegenden Teil auf der Geraden.
- B** ☐ Die Annahme der normalverteilten Residuen ist erfüllt. Die Punkte liegen zum überwiegenden Teil nicht auf der Geraden.
- C** ☐ Die Annahme der normalverteilten Residuen ist nicht erfüllt. Die Punkte liegen zum überwiegenden Teil nicht auf der Geraden.
- D** ☐ Die Annahme der normalverteilten Residuen ist erfüllt. Die Punkte liegen zum überwiegenden Teil auf der Geraden.
- E** ☐ Die Annahme der normalverteilten Residuen ist erfüllt. Die Punkte liegen zum überwiegenden Teil nicht auf der Geraden und Korrelation ist negativ.

### 33 Aufgabe

(2 Punkte)

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



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



- B** ☐ Die Annahme der normalverteilten Residuen ist erfüllt. Die Punkte liegen zum überwiegenden Teil auf der Diagonalen.
- C** ☐ Die Annahme der normalverteilten Residuen ist erfüllt. Kein Muster ist zu erkennen und keine Outlier zu beobachten.
- D** ☐ Die Annahme der normalverteilten Residuen ist nicht erfüllt. Es ist kein Muster zu erkennen.
- 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.

### 34 Aufgabe

(2 Punkte)

Sie wollen ein Feldexperiment mit zwei Düngestufen durchführen. Berechnen Sie die benötigte Fallzahl mit  $t_{1-\alpha/2} = 1.645$  und  $t_{1-\beta} = 1.282$  sowie  $s = 1$  und  $\delta_0 = 1$ . Es ergibt sich somit folgende Fallzahl.

- A** ☐ Es ergibt sich eine Fallzahl von 17
- B** ☐ Es ergibt sich eine Fallzahl von 9
- C** ☐ Es ergibt sich eine Fallzahl von 34
- D** ☐ Es ergibt sich eine Fallzahl von 36
- E** ☐ Es ergibt sich eine Fallzahl von 18

### 35 Aufgabe

(2 Punkte)

Welche Aussage zum mathematische Ausdruck  $Pr(D|H_0)$  ist richtig?

- A** ☐  $Pr(D|H_0)$  ist die Wahrscheinlichkeit der Alternativhypothese und somit  $1 - Pr(H_A)$
- B** ☐  $Pr(D|H_0)$  ist die Wahrscheinlichkeit die Daten D zu beobachten wenn die Nullhypothese wahr ist.
- C** ☐ Die Wahrscheinlichkeit für die Nullhypothese, wenn die Daten wahr sind.
- D** ☐ Die Inverse der Wahrscheinlichkeit unter der die Nullhypothese nicht mehr die Alternativhypothese überdeckt.
- E** ☐ Die Wahrscheinlichkeit der Daten unter der Nullhypothese in der Grundgesamtheit.

### 36 Aufgabe

(2 Punkte)

Das Falsifikationsprinzip besagt...

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

### 37 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 ☐ 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.
- B ☐ Der Begründer der modernen Statistik, R. Fischer, hat die Grenze simuliert und berechnet. Dadurch ergibt sich dieser optimale Cut-Off.
- C ☐ 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.
- 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 ☐ 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.

### 38 Aufgabe

(2 Punkte)

Welche Aussage über die Power ist richtig?

- A ☐ Die Power ist nicht in der aktuellen Testtheorie mehr vertreten. Wir rechnen nur noch mit dem Fehler 1. Art.
- B ☐ Die Power  $1 - \beta$  wird auf 80% gesetzt. Alle statistischen Tests sind so konstruiert, dass die  $H_A$  mit 80% "bewiesen wird".
- C ☐ Die Power beschreibt die Wahrscheinlichkeit die  $H_A$  abzulehnen. Wir testen die Power jedoch nicht.
- D ☐ Die Power  $1 - \beta$  wird auf 80% gesetzt. Damit liegt die Wahrscheinlichkeit für die  $H_0$  bei 20%.
- E ☐ Es gilt  $\alpha + \beta = 1$  und somit liegt  $\beta$  meist bei 95%.

### 39 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  $T = \frac{\text{signal}}{\text{noise}}$
- B ☐ Es gilt  $T = \frac{\text{noise}}{\text{signal}}$
- C ☐ Es gilt  $T = \text{signal} \cdot \text{noise}$
- D ☐ Es gilt  $T = (\text{signal} \cdot \text{noise})^2$
- E ☐ Es gilt  $T = \frac{\text{signal}}{\text{noise}^2}$

## 40 Aufgabe

(2 Punkte)

In der Theorie zur statistischen Testentscheidung kann „ $H_0$  ablehnen obwohl die  $H_0$  gilt“ in welche richtige Analogie gesetzt werden?

- A ☐ In die Analogie eines Rauchmelders: *Alarm with fire*.
- B ☐ In die Analogie eines Feuerwehrautos: *Car without noise*.
- C ☐ In die Analogie eines brennenden Hauses ohne Rauchmelder: *House without noise*.
- D ☐ In die Analogie eines Rauchmelders: *Alarm without fire*, dem  $\alpha$ -Fehler.
- E ☐ In die Analogie eines Rauchmelders: *Fire without alarm*, dem  $\beta$ -Fehler.

## 41 Aufgabe

(2 Punkte)

Sie rechnen eine simple Gaussian Regression. Welche Aussage betreffend der Konfidenzintervalle ist für die Gaussian Regression richtig?

- A ☐ Wenn die Relevanzschwelle mit enthalten ist, kann die Nullhypothese abgelehnt werden.
- B ☐ Wenn die Konfidenzintervalle den p-Wert der Regression enthalten, kann die Nullhypothese abgelehnt werden.
- C ☐ Wenn die 1 im Konfidenzintervall enthalten ist, kann die Nullhypothese nicht abgelehnt werden.
- D ☐ Wenn die 0 im Konfidenzintervall enthalten ist, kann die Nullhypothese abgelehnt werden.
- E ☐ Wenn die 0 im Konfidenzintervall enthalten ist, kann die Nullhypothese nicht abgelehnt werden.

## 42 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?

- A ☐ Aktuell werden immer grössere Datensätze erhoben. Eine erhöhte Fallzahl führt automatisch auch zu mehr signifikanten Ergebnissen, selbst wenn die eigentlichen Effekte nicht relevant sind.
- B ☐ Relevanz und Signifikanz haben nichts miteinander zu tun. Daher gibt es auch keinen Zusammenhang zwischen hoher Fallzahl ( $n > 10000$ ) und einem signifikanten Test. Ein Effekt ist immer relevant und somit signifikant.
- C ☐ 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.
- D ☐ Aktuell werden immer grössere Datensätze erhoben. Dadurch wird auch die Varianz immer höher was automatisch zu mehr signifikanten Ergebnissen führt.
- E ☐ Aktuell werden zu grosse Datensätze für die gängige Statistik gemessen. Daher wendet man maschinelle Lernverfahren für kausale Modelle an. Hier ist die Relevanz gleich Signifikanz.

### 43 Aufgabe

(2 Punkte)

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

- A ☐ 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.
- B ☐ 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.
- C ☐ Das OR. Als Chancenverhältnis gibt es das Verhältnis von Relevanz und Signifikanz wieder.
- D ☐ 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.
- E ☐ Das Konfidenzintervall. Durch die Visualisierung des Konfidenzintervalls kann eine Relevanzschwelle vom Anwender definiert werden. Zusätzlich erlaubt das Konfidenzintervall auch eine Entscheidung über die Signifikanz.

### 44 Aufgabe

(2 Punkte)

Welche Aussage über die frequentistischen Testtheorie ist richtig?

- A ☐ Wir machen Aussagen über Wahrscheinlichkeiten!
- B ☐ Wir machen Aussagen über die individuelle Wahrscheinlichkeit eines Effektes!
- C ☐ Wir machen Aussagen über den Effekt!
- D ☐ Wir machen Aussagen über Individuen!
- E ☐ Wir machen keine Aussagen über Wahrscheinlichkeiten!

### 45 Aufgabe

(2 Punkte)

Welche Aussage über den t-Test ist richtig?

- A ☐ Der t-Test vergleicht die Varianzen von mindestens zwei oder mehr Gruppen
- B ☐ Der t-Test ist ein Vortest der ANOVA und basiert daher auf dem Vergleich von Streuungsparametern
- C ☐ Der t-Test testet generell zu einem erhöhten  $\alpha$ -Niveau von 20%.
- D ☐ 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.
- E ☐ Der t-Test vergleicht die Mittelwerte von zwei Gruppen.

## 46 Aufgabe

(2 Punkte)

Welche Aussage über den Welch t-Test ist richtig?

- A** ☐ Der Welch t-Test vergleicht die Mittelwerte von zwei Gruppen unter der strikten Annahme von Varianzhomogenität.
- B** ☐ Der Welch t-Test vergleicht die Varianz von zwei Gruppen.
- C** ☐ Der Welch t-Test ist die veraltete Form des Student t-Test und wird somit nicht mehr verwendet.
- D** ☐ Der Welch t-Test wird angewendet, wenn Varianzheterogenität zwischen den beiden zu vergleichenden Gruppen vorliegt.
- E** ☐ Der Welch t-Test ist ein Post-hoc Test der ANOVA und basiert daher auf dem Vergleich der Varianz.

## 47 Aufgabe

(2 Punkte)

Nach einem Experiment mit fünf Weizensorten ergibt eine ANOVA ( $p = 0.041$ ) einen signifikanten Unterschied für den Ertrag. Sie führen anschließend die paarweisen t-Tests für alle Vergleiche der verschiedenen Weizensorten durch. Nach der Adjustierung für multiples Testen ist kein p-Wert unter der  $\alpha$ -Schwelle. Sie schauen sich auch die rohen, unadjustierten p-Werte an und finden hier als niedrigsten p-Wert  $p_{3-2} = 0.053$ . Welche Aussage ist richtig?

- A** ☐ Die ANOVA testet auf der gesamten Fallzahl. Es wäre besser die ANOVA auf der gleichen Fallzahl wie die einzelnen t-Tests zu rechnen.
- B** ☐ Es gibt einen Fehler in der Varianzstruktur. Daher kann die ANOVA nicht richtig sein und paarweise t-Tests liefern das richtige Ergebnis.
- C** ☐ Der Fehler liegt in den t-Tests. Wenn eine ANOVA signifikant ist, dann muss zwangsweise auch ein t-Test signifikant sein.
- D** ☐ Die ANOVA testet auf der gesamten Fallzahl. Die einzelnen t-Tests immer nur auf einer kleineren Subgruppe. Da mit weniger Fallzahl weniger signifikante Ergebnisse zu erwarten sind, kann eine Diskrepanz zwischen der ANOVA und den paarweisen t-Tests auftreten.
- E** ☐ Die adjustierten p-Werte deuten in die richtige Richtung. Zusammen mit den nicht signifikanten rohen p-Werten ist von einem Fehler in der ANOVA auszugehen.

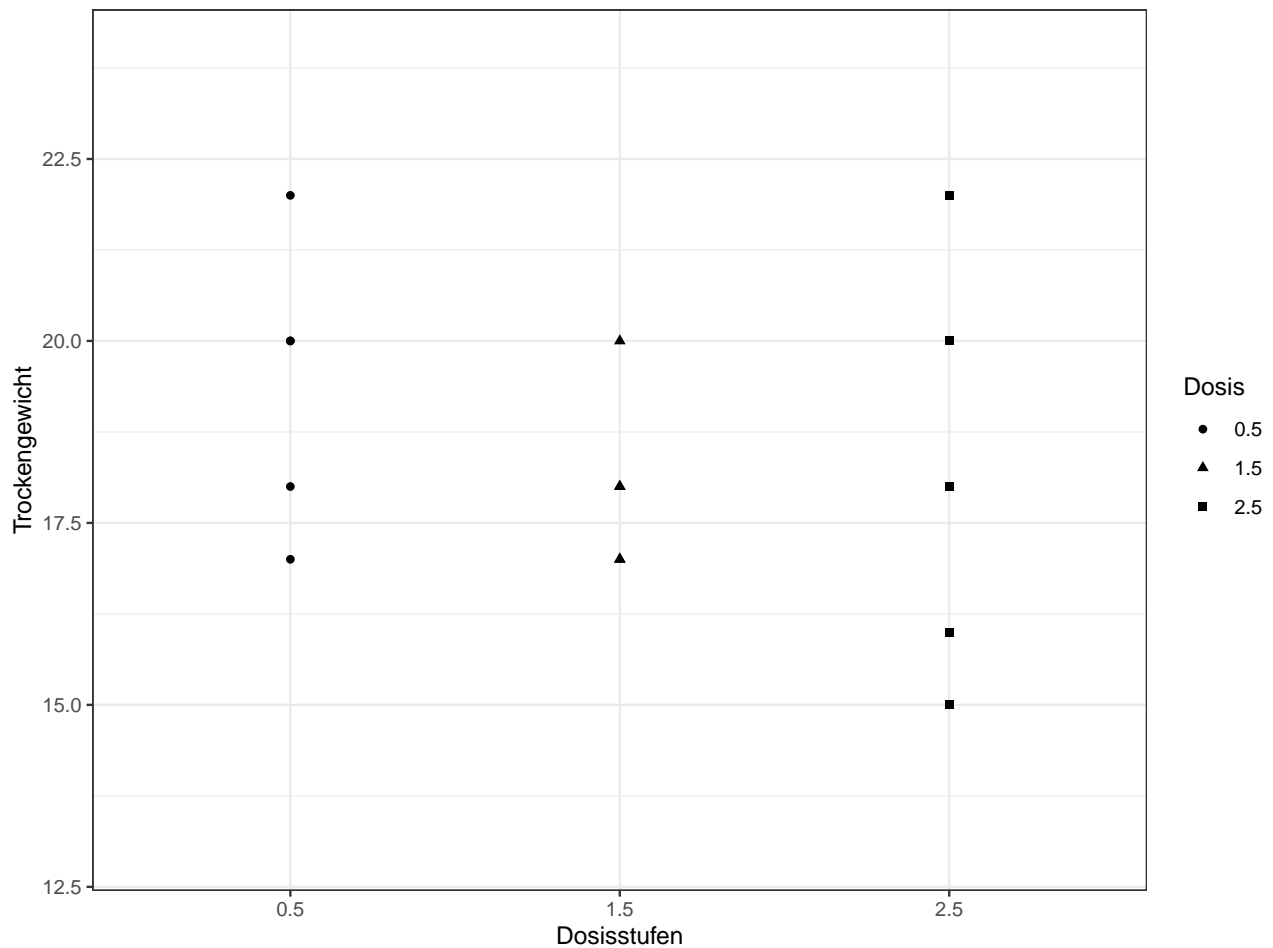
## Rechen- und Textaufgaben

- Die Zahlen und Abbildungen werden in *jeder* Version dieses Dokuments neu erstellt.
- Es kann daher sein, dass *seltsame* Ergebnisse oder Abbildungen entstehen. Im Falle der Klausur werde ich das nochmal korrigieren — hier lasse ich es so stehen.

## 48 Aufgabe

(8 Punkte)

In einem Experiment wurde der Ertrag von Erbsen unter drei verschiedenen Pestizid-Dosen 0.5 g/l, 1.5 g/l und 2.5 g/l gemessen. Unten stehenden sehen Sie die Visualisierung des Datensatzes.



1. Zeichnen Sie folgende statistischen Masszahlen in die Abbildung ein! (6 Punkte)

- Total (grand) mean:  $\beta_0$
- Mittelwerte der Dosen:  $\bar{y}_{0.5}$ ,  $\bar{y}_{1.5}$  und  $\bar{y}_{2.5}$
- Effekt der einzelnen Level der Dosen:  $\beta_{0.5}$ ,  $\beta_{1.5}$ , und  $\beta_{2.5}$
- Residuen oder Fehler:  $\epsilon$

2. Schätzen Sie den p-Wert einer einfaktoriellen ANOVA ab. Liegt ein *vermutlicher* signifikanter Unterschied zwischen den Dosisstufen vor? Begründen Sie Ihre Antwort! (2 Punkte)

## 49 Aufgabe

(14 Punkte)

Der Datensatz PlantGrowth enthält das Gewicht der Pflanzen (*weight*), die unter einer Kontrolle und zwei verschiedenen Behandlungsbedingungen erzielt wurden – dem Faktor *group* mit den Faktorstufen *ctrl*, *trt1*, *trt2*.

1. Füllen Sie die unterstehende einfaktorielle ANOVA Ergebnistabelle aus mit den gegebenen Informationen von Df und Sum Sq! (4 Punkte)
2. Schätzen Sie den p-Wert der Tabelle mit der Information von  $F_{\alpha=5\%} = 3.35$  ab. Begründen Sie Ihre Antwort! (2 Punkte)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	4.07			
Residuals	27	9.25			

3. Was bedeutet ein signifikantes Ergebnis in einer einfaktoriellen ANOVA im Bezug auf die möglichen Unterschiede zwischen den Gruppen? Beziehen Sie sich auf den obigen Fragetext bei Ihrer Antwort! (2 Punkte)
4. Berechnen Sie *einen* Student t-Test mit  $T = \frac{\bar{x}_1 - \bar{x}_2}{s_p \cdot \sqrt{2/n_g}}$  für den *vermutlich* signifikantesten Gruppenvergleich anhand der untenstehenden Tabelle mit  $T_k = 2.03$ . Begründen Sie Ihre Auswahl! (4 Punkte)

group	n	mean	sd
ctrl	10	5.05	0.54
trt1	10	4.63	0.78
trt2	10	5.53	0.36

5. Gegebenen der ANOVA Tabelle war das Ergebnis des t-Tests zu erwarten? Begründen Sie Ihre Antwort! (2 Punkte)



## 50 Aufgabe

(16 Punkte)

Der Datensatz *ToothGrowth* enthält 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 *dose* (0.5 mg/Tag, 1 mg/Tag und 2 mg/Tag) über eine von zwei Verabreichungsmethoden *supp* erhielt (Orangensaft oder Ascorbinsäure). Die Zahnlänge wurde als normalverteiltes Outcome gemessen.


1. Füllen Sie die unterstehende zweifaktorielle ANOVA Ergebnistabelle aus mit den gegebenen Informationen von Df und Sum Sq! **(6 Punkte)**
2. Schätzen Sie den p-Wert der Tabelle mit der Information von den kritischen F-Werten mit  $F_{supp} = 4.02$  und  $F_{dose} = 3.17$  sowie  $F_{supp:dose} = 3.17$  ab. Begründen Sie Ihre Antwort! **(4 Punkte)**

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<b>supp</b>	1	205.38			
<b>dose</b>	2	2430.35			
<b>supp:dose</b>	2	107.36			
<b>Residuals</b>	54	713.49			

3. Was bedeutet ein signifikantes Ergebnis in einer zweifaktoriellen ANOVA im Bezug auf die möglichen Unterschiede zwischen den Gruppen? Beziehen Sie sich dabei einmal auf den Faktor *supp* und einmal auf den Faktor *dose*! **(4 Punkte)**
4. Was sagt der Term *supp:dose* aus? Interpretieren Sie das Ergebnis des abgeschätzten p-Wertes! **(2 Punkte)**

## 51 Aufgabe

(8 Punkte)


Der Datensatz *Crop* enthält das Trockengewicht der Maispflanzen (*drymatter*), die unter drei verschiedenen Düngerbedingungen erzielt wurden – dem Faktor *trt* mit den Faktorstufen *low*, *mid*, *high*. Sie erhalten folgenden Output in .

```
## Analysis of Variance Table
##
## Response: drymatter
##          Df    Sum Sq  Mean Sq F value    Pr(>F)
## trt        2   3.65161   1.825804  4.65503 0.018325 *
## Residuals 27  10.59000   0.392222
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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! **(2 Punkte)**

## 52 Aufgabe

(8 Punkte)

Der Datensatz *PigGain* enthält Daten aus einer Studie zur Bewertung der Wirkung vom Vitamin Selen auf das Wachstum bei Mastschweinen. Der Versuch wurde an 63 Mastschweinen durchgeführt, wobei jedes Tier eine von drei Selen-Dosen (0.5 ng/Tag, 1 ng/Tag und 5 ng/Tag) über eine von zwei Verabreichungsmethoden erhielt (Wasser oder Festnahrung). Das Gewicht wurde als normalverteiltes Outcome gemessen. Sie erhalten folgenden Output in .

```
## Analysis of Variance Table
##
## Response: len
##      Df    Sum Sq Mean Sq  F value    Pr(>F)
## supp     1 1330.563 1330.563 117.66550 0.00000000000000017535 ***
## dose      2 1939.644  969.822  85.76416 < 0.0000000000000000222 ***
## supp:dose  2    8.777    4.388   0.38807    0.68014
## Residuals 57  644.557   11.308
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

## 53 Aufgabe

(8 Punkte)

In einer ANOVA wird die F Statistik nach der Formel  $F_{calc} = \frac{MS_{treatment}}{MS_{error}}$  berechnet. In einem t-test berechnen wir die T Statistik nach der Formel  $T_{calc} = \frac{\bar{y}_1 - \bar{y}_2}{s_p \cdot \sqrt{2/n_g}}$ .

1. Erklären Sie den konzeptionellen Zusammenhang zwischen der  $F_{calc}$  Statistik und  $T_{calc}$  Statistik! **(2 Punkte)**
2. Visualisieren Sie eine nicht signifikante  $F_{calc}$  Statistik sowie eine signifikante  $F_{calc}$  Statistik! Beschriften Sie die Abbildung! **(2 Punkte)**
3. Erklären Sie an der Formel des F-Tests sowie an der Abbildung warum das Minimum der F-Statistik 0 ist! **(2 Punkte)**
4. Wenn die F-Statistik 0 ist, spricht dies eher für oder gegen die Nullhypothese? Begünden Sie Ihre Antwort! **(2 Punkte)**

## 54 Aufgabe

**(6 Punkte)**

Sie rechnen eine zweifaktorielle ANOVA und erhalten einen signifikanten Interaktionseffekt zwischen den beiden Faktoren  $f_1$  und  $f_2$ . Der Faktor  $f_1$  hat drei Level. Der Faktor  $f_2$  hat dagegen nur zwei Level.

1. Visualisieren Sie in zwei getrennten Abbildungen eine mittlere und eine starke Interaktion zwischen den Faktoren  $f_1$  und  $f_2$ ! **(2 Punkte)**
2. Erklären Sie den Unterschied zwischen Ihren beiden Zeichnungen! **(2 Punkte)**
3. Wenn Sie eine signifikante Interaktion in Ihren Daten vorliegen haben, wie ist dann Ihr weiteres Vorgehen bei dem Posthoc-Tests? **(2 Punkte)**

## 55 Aufgabe

**(8 Punkte)**

Sie rechnen eine einfaktorielle ANOVA mit einem Faktor  $f_1$  mit fünf Leveln. Nachdem Sie die einfaktorielle ANOVA gerechnet haben, erhalten Sie einen p-Wert von 0.078 und eine F Statistik mit  $F_{calc} = 1.2$ . Als Sie sich die Boxplots der Behandlungen anschauen, stellen Sie fest, dass es eigentlich einen Mittelwertsunterschied zwischen dem dritten und ersten Level geben müsste. Die *IQR*-Bereiche überlappen sich nicht und die Mediane liegen auch weit vom globalen Mittel entfernt.

1. Erklären Sie die Annahme der Normalverteilung und die Annahme der Varianzhomogenität an einer passenden Abbildung! **(2 Punkte)**
2. Visualisieren Sie die Berechnung von  $F_{calc}$  am obigen Beispiel! **(2 Punkte)**
3. Erklären Sie das Ergebnis der obigen einfaktoriellen ANOVA unter der Berücksichtigung der Annahmen an eine ANOVA! Geben Sie ein numerisches Beispiel! **(4 Punkte)**

## 56 Aufgabe

**(9 Punkte)**

Nach einem Gewächshausexperiment mit drei Bewässerungstypen (A, B und C) ergibt sich die folgende Datentabelle mit dem gemessenen Frischgewicht (*freshmatter*).

water_type	freshmatter
A	17
A	20
B	34
B	30
C	12
A	17
C	7
C	11
A	12
C	12
B	30
A	18
B	32

1. Zeichnen Sie in *einer* Abbildung die Barplots für die Bewässerungstypen! Beschriften Sie die Achsen entsprechend! **(6 Punkte)**
2. Beschriften Sie *einen* Barplot mit den gängigen statistischen Maßzahlen! **(2 Punkte)**
3. Wenn Sie *keinen Effekt* zwischen der Bewässerungstypen erwarten würden, wie sehen dann die Barplots aus? **(1 Punkt)**

## 57 Aufgabe

(9 Punkte)

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

trt	drymatter
B	12.7
B	10.1
B	13.9
A	18.5
B	12.3
B	13.6
A	18.5
B	15.4
B	9.7
B	15.5
A	21.0
A	18.0
B	12.6
B	16.8
B	12.2
A	19.6
A	19.5
A	20.8

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)**



## 58 Aufgabe

**(9 Punkte)**

Nach einem Feldexperiment mit mehreren Düngestufen stellt sich die Frage, ob die Düngestufe *low* im Bezug auf das Trockengewicht normalverteilt sei. Sie erhalten folgende Datentabelle.

fertilizer	drymatter
low	12
low	18
low	18
low	16
low	16
low	16
low	15
low	16
low	12
low	18
low	15

1. Zeichnen Sie eine passende Abbildung in der Sie visuell überprüfen können, ob eine Normalverteilung des Trockengewichts vorliegt! **(4 Punkte)**
2. Beschriften Sie die Achsen und ergänzen Sie die statistischen Maßzahlen. **(3 Punkte)**
3. Entscheiden Sie, ob eine Normalverteilung vorliegt. Begründen Sie Ihre Antwort. **(2 Punkte)**

## 59 Aufgabe

(4 Punkte)

1. Zeichnen Sie über den untenstehenden Boxplot die entsprechende zugehörige Verteilung! (2 Punkte)
2. Zeichnen Sie unter den untenstehenden Boxplot die entsprechende zugehörige Beobachtungen! (2 Punkte)



## 60 Aufgabe

**(10 Punkte)**

Nach einem Experiment ergibt sich die folgende 2x2 Datentabelle mit einem Pestizid (ja/nein), dargestellt in den Zeilen. Im Weiteren mit dem infizierten Pflanzenstatus (ja/nein) in den Spalten. Insgesamt wurden  $n = 152$  Pflanzen untersucht.

	Erkrankt (ja)	Erkrankt (nein)	
Pestizid (ja)	56	21	
Pestizid (nein)	23	52	

1. Ergänzen Sie die Tabelle um die Randsummen! **(1 Punkt)**
2. Formulieren Sie die Fragestellung! **(1 Punkt)**
3. Formulieren Sie das Hypothesenpaar! **(2 Punkte)**
4. Berechnen Sie die Teststatistik eines Chi-Quadrat-Test auf der 2x2 Tafel. Geben Sie Formeln und Rechenweg mit an! **(4 Punkte)**
5. Treffen Sie eine Entscheidung im Bezug zu der Nullhypothese gegeben einem  $T_k = 3.841$ ! **(1 Punkt)**
6. Skizzieren Sie eine 2x2 Tabelle mit  $n = 26$  Pflanzen in dem *vermutlich* die Nullhypothese nicht abgelehnt werden kann! **(1 Punkt)**

## 61 Aufgabe

(7 Punkte)

Gegeben sind folgende Randsummen in einer 2x2 Kreuztabelle aus einem Experiment mit  $n = 120$  Sauen. In dem Experiment wurde gemessen, ob eine Sau nach einer Behandlung mit einem Medikament (ja/nein) mehr als 30 Ferkel pro Jahr bekommen konnte (ja/nein).

	>30 Ferkel (ja)	≤30 Ferkel (nein)	
Medikament (ja)			62
Medikament (nein)			58
	57	63	120

1. Ergänzen Sie die Felder innerhalb der 2x2 Kreuztabelle in dem Sinne, dass *kein* signifikanter Effekt zu erwarten wäre! **(2 Punkte)**
2. Erklären und Begründen Sie Ihr Vorgehen an der Formel des Chi-Quadrat-Tests mit

$$\chi^2 = \sum \frac{(O-E)^2}{E}.$$

Sie können dies an einem Beispiel erklären! **(2 Punkte)**

3. Was ist die Mindestanzahl an Beobachtungen je Zelle? Wenn in einer der Zellen weniger Beobachtungen auftreten, welchen Test können Sie anstatt des „normalen“ Chi-Quadrat-Tests anwenden? **(2 Punkte)**
4. Warum hat die obige Vierfeldertafel einen Freiheitsgrad von  $df = 1$ ? **(1 Punkt)**

## 62 Aufgabe

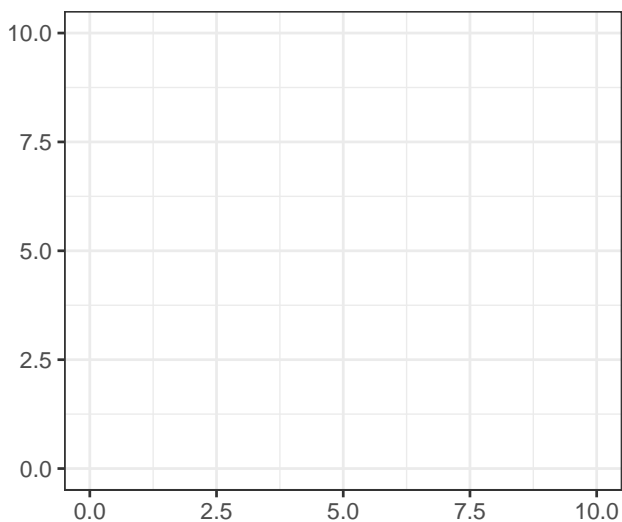
(9 Punkte)

Im folgenden sehen Sie drei leere Scatterplots. Füllen Sie diese Scatterplots nach folgenden Anweisungen.

1. Zeichnen Sie für die angegebene  $\rho$ -Werte eine Gerade in die entsprechende Abbildung! **(3 Punkte)**
2. Zeichnen Sie für die angegebenen  $R^2$ -Werte die entsprechende Punktwolke um die Gerade. **(3 Punkte)**
3. Sie rechnen ein statistisches Modell. Was sagen Ihnen die  $R^2$ -Werte über das jeweilige Modell? **(3 Punkte)**

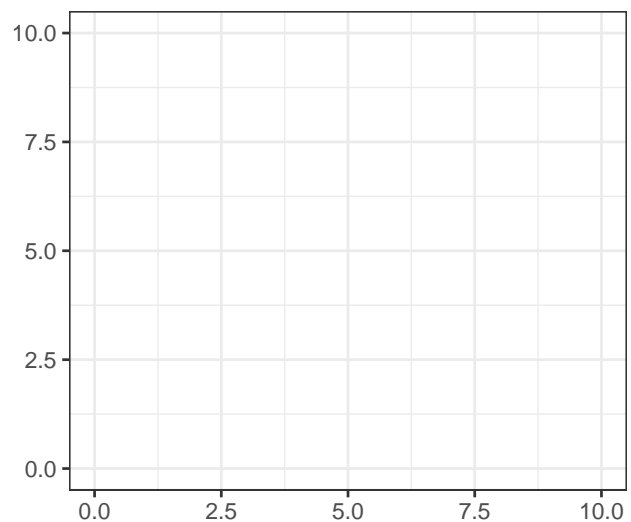
Pearsons  $\rho = -0.25$

$R^2 = 0.5$



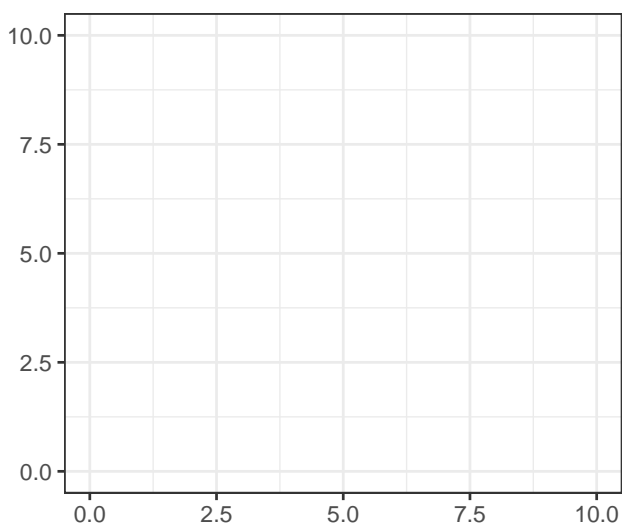
Pearsons  $\rho = 0.25$

$R^2 = 0.25$



Pearsons  $\rho = 0.75$

$R^2 = 1$



## 63 Aufgabe

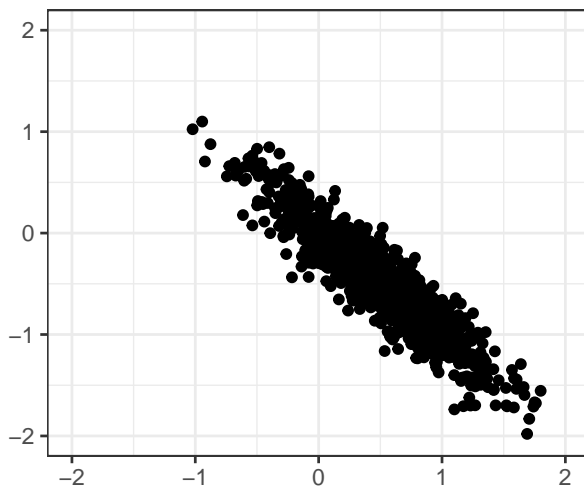
(10 Punkte)

Im folgenden sehen Sie vier Scatterplots. Ergänzen Sie die Überschriften der jeweiligen Scatterplots.

1. Schätzen Sie die  $\rho$ -Werte in der entsprechenden Abbildung! (4 Punkte)
2. Schätzen Sie die  $R^2$ -Werte in der entsprechenden Punktwolke um die Gerade! (4 Punkte)
3. Sie rechnen ein statistisches Modell. Was sagen Ihnen die  $R^2$ -Werte über das jeweilige Modell? (2 Punkte)

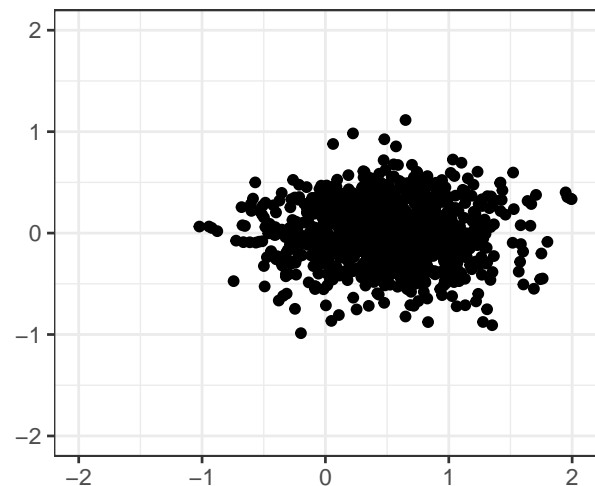
Pearsons  $\rho =$

$R^2 =$



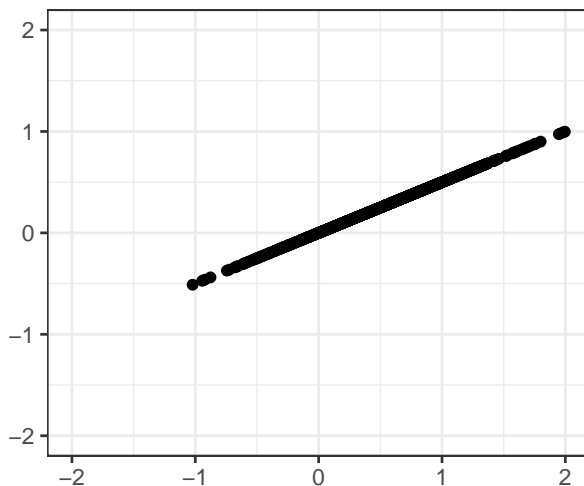
Pearsons  $\rho =$

$R^2 =$



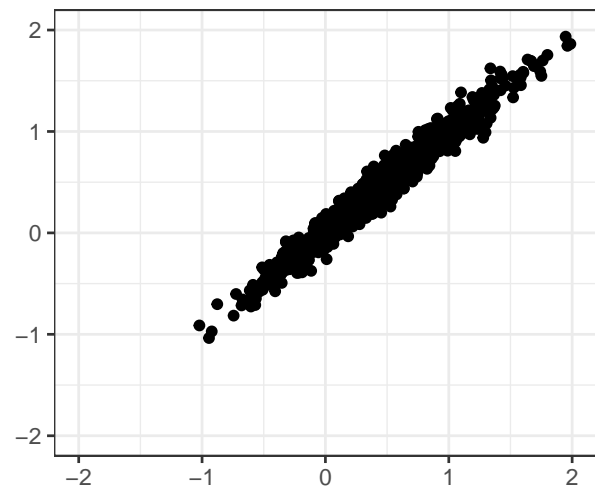
Pearsons  $\rho =$

$R^2 =$



Pearsons  $\rho =$

$R^2 =$

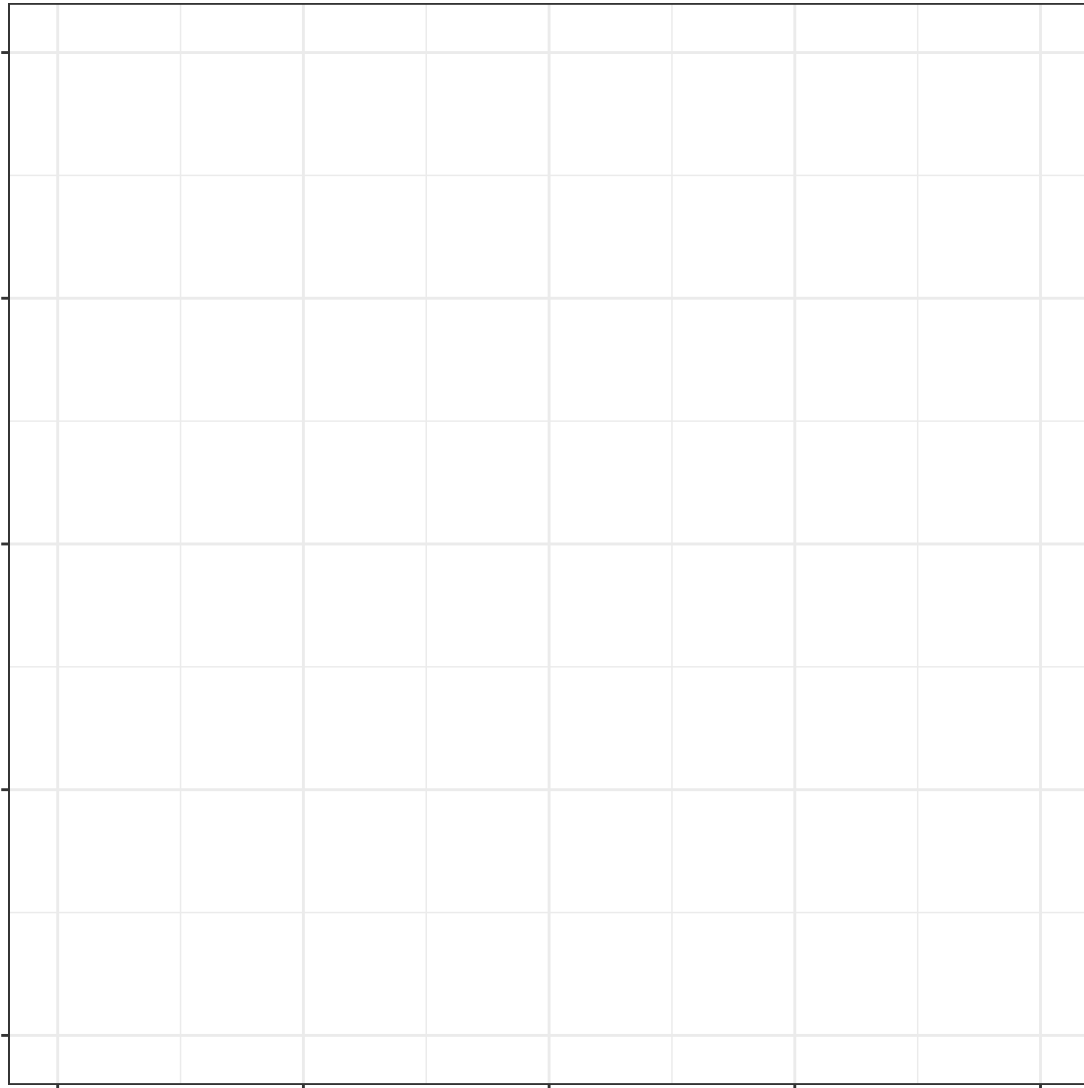


## 64 Aufgabe

(8 Punkte)

Sie haben folgende Zahlenreihe  $y$  vorliegen  $y = \{18, 11, 15, 18, 17, 16\}$ .

1. Visualisieren Sie den Mittelwert von  $y$  in der untenstehenden Abbildung! **(4 Punkte)**
2. Beschriften Sie die  $Y$  und  $X$ -Achse entsprechend! **(2 Punkte)**
3. Für die Berechnung der Varianz wird der Abstand der einzelnen Werte  $x_i$  zum Mittelwert  $\bar{x}$  quadriert. Warum muss der Abstand,  $x_i - \bar{x}$ , in der Varianzformel quadriert werden? Erklären Sie den Zusammenhang unter Berücksichtigung der Abbildung! **(2 Punkte)**



## 65 Aufgabe

**(10 Punkte)**

Sie haben folgende Zahlenreihe  $y$  vorliegen  $y = \{17, 19, 17, 22, 19, 25, 21, 16, 22, 19, 13\}$ . Berechnen Sie folgende deskriptive Maßzahlen. Geben Sie Formeln und Rechenwege mit an!

1. Den Mittelwert **(2 Punkte)**
2. Die Standardabweichung **(2 Punkte)**
3. Den Interquartileabstand **(2 Punkte)**
4. Das 3rd Quartile **(2 Punkte)**
5. Das 1st Quartile **(2 Punkte)**

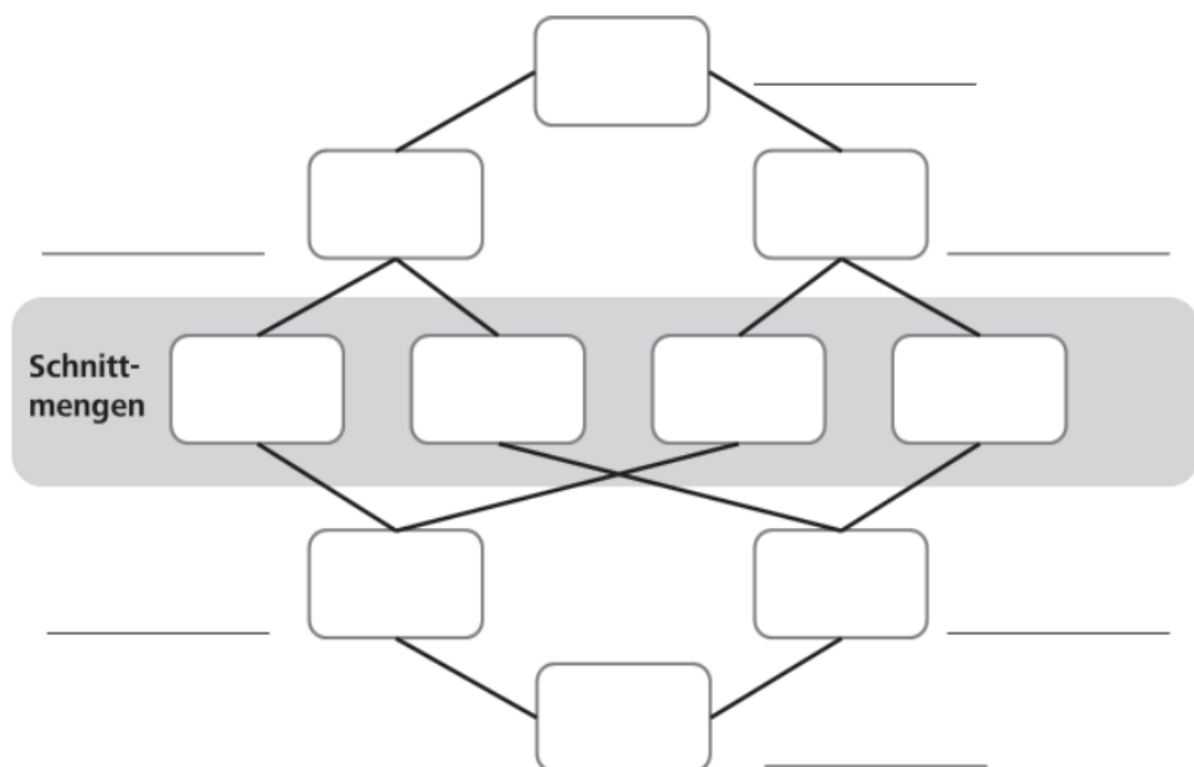


## 66 Aufgabe

(11 Punkte)

Die Prävalenz von Klauenseuche bei Wollschweinen wird mit 2% angenommen. In 80% der Fälle ist ein Test positiv, wenn das Wollschwein erkrankt ist. In 8% der Fälle ist ein Test positiv, wenn das Wollschwein *nicht* erkrankt ist und somit gesund ist. Sie werten 1000 Wollschweine mit einem diagnostischen Test auf Klauenseuche aus.

1. Füllen und beschriften Sie den untenstehenden Doppelbaum! Beschriften Sie auch die Äste des Doppelbaumes, mit denen Ihnen bekannten Informationen! **(8 Punkte)**
2. Berechnen Sie die Wahrscheinlichkeit  $Pr(K^+|T^+)$ ! **(2 Punkte)**
3. Was sagt Ihnen die Wahrscheinlichkeit  $Pr(K^+|T^+)$  aus? **(1 Punkt)**



## 67 Aufgabe

**(9 Punkte)**

Beim diagnostischen Testen erhalten Sie *True Positives (TP)*, *True Negatives (TN)*, *False Positives (FP)* und *False Negatives (FN)*. Erklären Sie den Zusammenhang wie folgt.

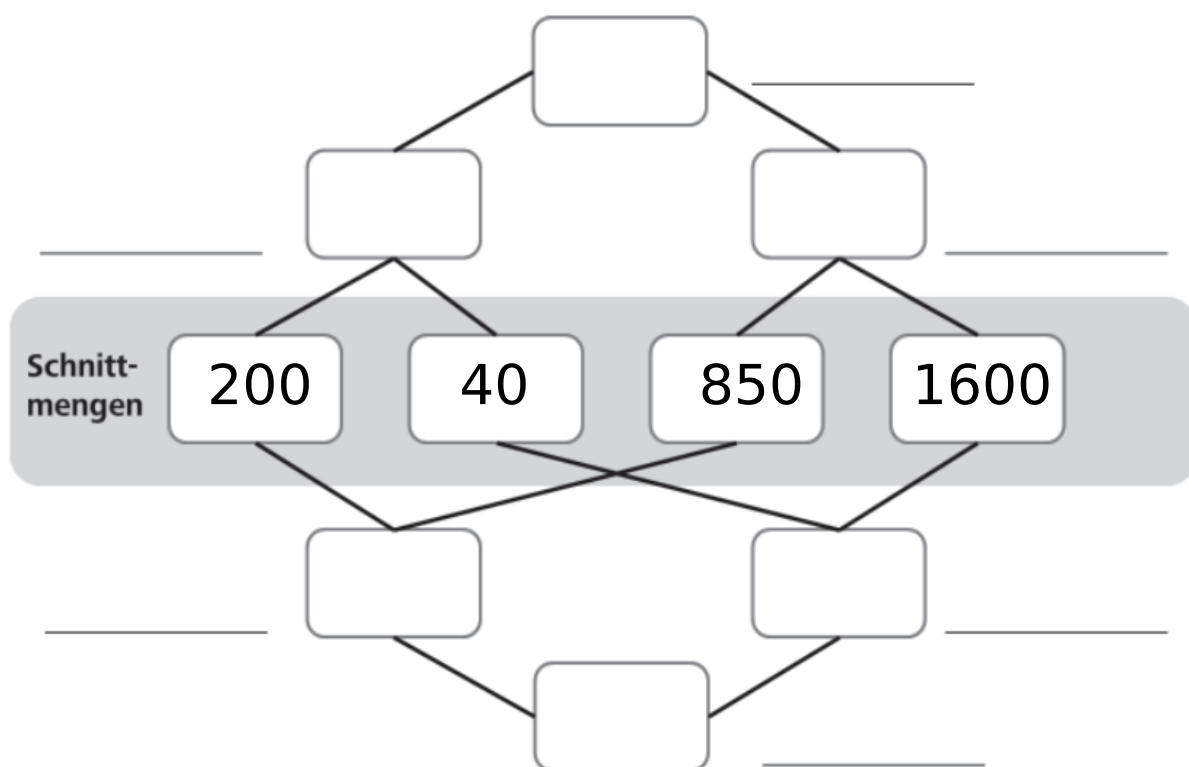
1. Visualisieren Sie *TP*, *TN*, *FP* und *FN* in einer Abbildung. Beschriften Sie die Abbildung und die Achsen entsprechend! **(6 Punkte)**
2. Tragen Sie *TP*, *TN*, *FP* und *FN* in eine 2x2 Kreuztabelle ein. Beschriften Sie die Tabelle entsprechend! **(3 Punkte)**

## 68 Aufgabe

(12 Punkte)

Folgender diagnostischer Doppelbaum nach der Testung auf Klauenseuche bei Fleckvieh ist gegeben.

1. Füllen und beschriften Sie den untenstehenden Doppelbaum! **(4 Punkte)**
2. Berechnen Sie die Wahrscheinlichkeit  $Pr(K^+|T^+)$ ! **(2 Punkte)**
3. Berechnen Sie die Prävalenz für Klauenseuche! **(2 Punkte)**
4. Berechnen Sie die Sensitivität und Spezifität des diagnostischen Tests für Klauenseuche! Erstellen Sie dafür zunächst eine 2x2 Kreuztabelle aus dem ausgefüllten Doppelbaum! **(4 Punkte)**



## 69 Aufgabe

(7 Punkte)

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

pesticide	grade
pestKill	2
pestKill	3
pestKill	2
ctrl	4
ctrl	6
roundUp	5
zeroX	1
zeroX	2
zeroX	4
roundUp	0
roundUp	9
roundUp	3
ctrl	5
zeroX	0
pestKill	0
ctrl	4

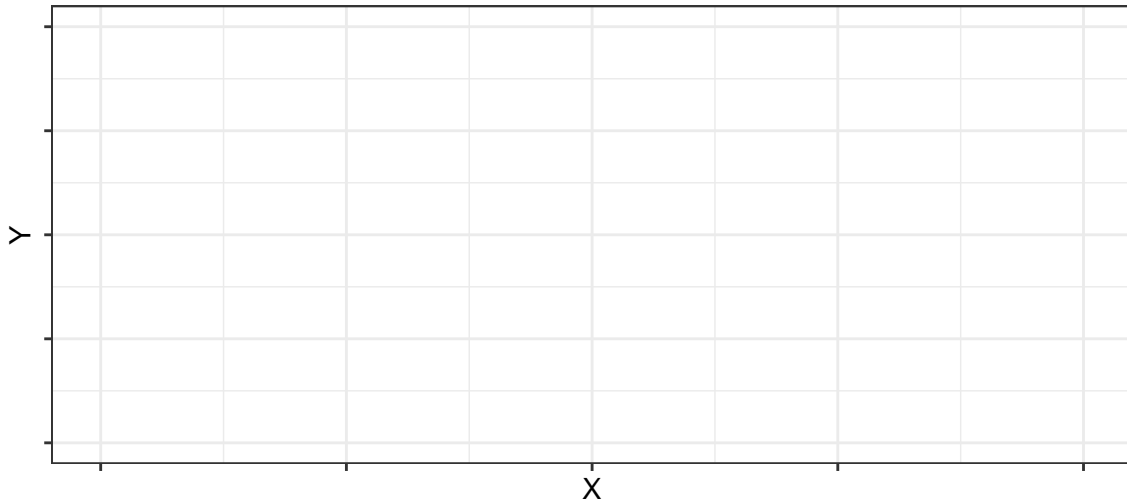
1. Zeichnen Sie in *einer* Abbildung die Dotplots für die vier Pestizidlevel! Beschriften Sie die Achsen entsprechend! **(4 Punkte)**
2. Ergänzen Sie die Dotplots mit einer gängigen statistischen Maßzahl. **(2 Punkte)**
3. Wenn Sie *keinen Effekt* zwischen den Pestizidlevel erwarten würden, wie sehen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

## 70 Aufgabe

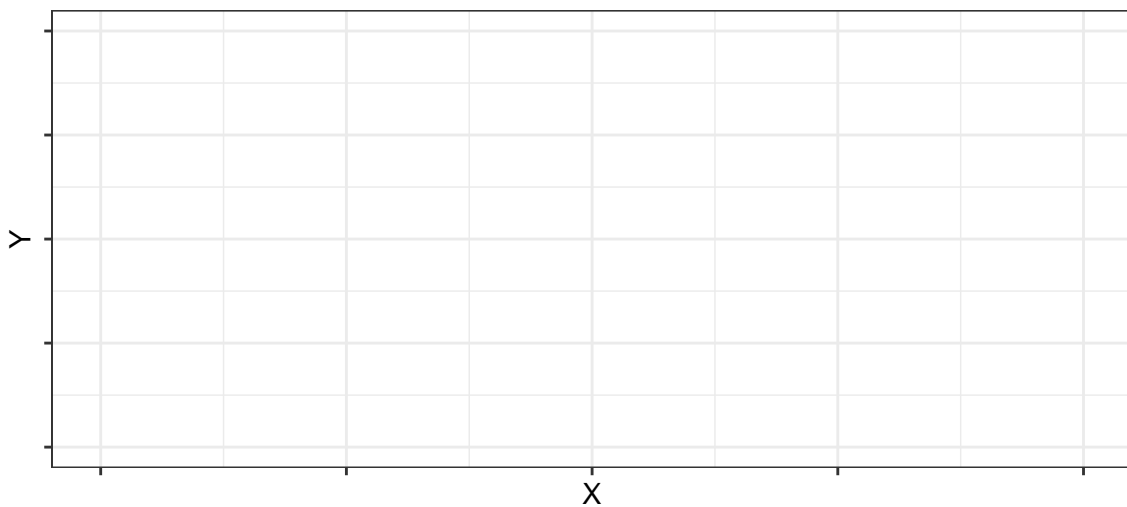
(6 Punkte)

1. Skizzieren Sie in die unten stehenden, freien Abbildungen die Verteilungen, die sich nach der Abbildungsüberschrift ergeben! **(4 Punkte)**
2. Achten Sie auf die entsprechende Skalierung der beiden Verteilungen in der ersten Abbildung! **(2 Punkte)**

$N(0, 2)$  und  $N(2, 2)$



$\text{Pois}(15)$



## 71 Aufgabe

(6 Punkte)

1. Skizzieren Sie 4 Normalverteilungen *in einer Abbildung* mit  $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$  und  $s_1 = s_2 = s_3 = s_4$ ! **(2 Punkte)**
2. Beschriften Sie die Normalverteilungen mit den entsprechenden Parametern! **(2 Punkte)**
3. Liegt Varianzhomogenität oder Varianzheterogenität vor? Begründen Sie Ihre Antwort! **(2 Punkte)**

## 72 Aufgabe

**(6 Punkte)**

Nach einem Experiment zählen Sie folgende Anzahl an Läsionen auf den Beobachtungen nach einer durchgestandenen Infektion.

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

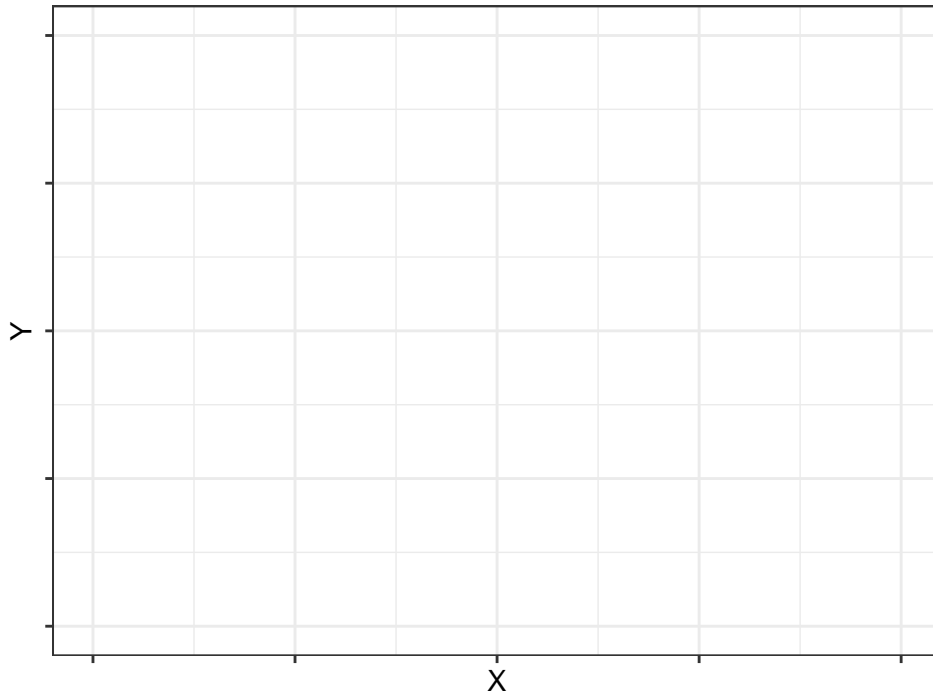
1. Zeichnen Sie ein Histogramm um die Verteilung der Daten zu visualisieren. **(3 Punkte)**
2. Beschriften Sie die Achsen der Abbildung! **(2 Punkte)**
3. Wie unterscheidet sich eine Poissonverteilung von einer Normalverteilung dargestellt in einem Histogramm? **(1 Punkt)**

## 73 Aufgabe

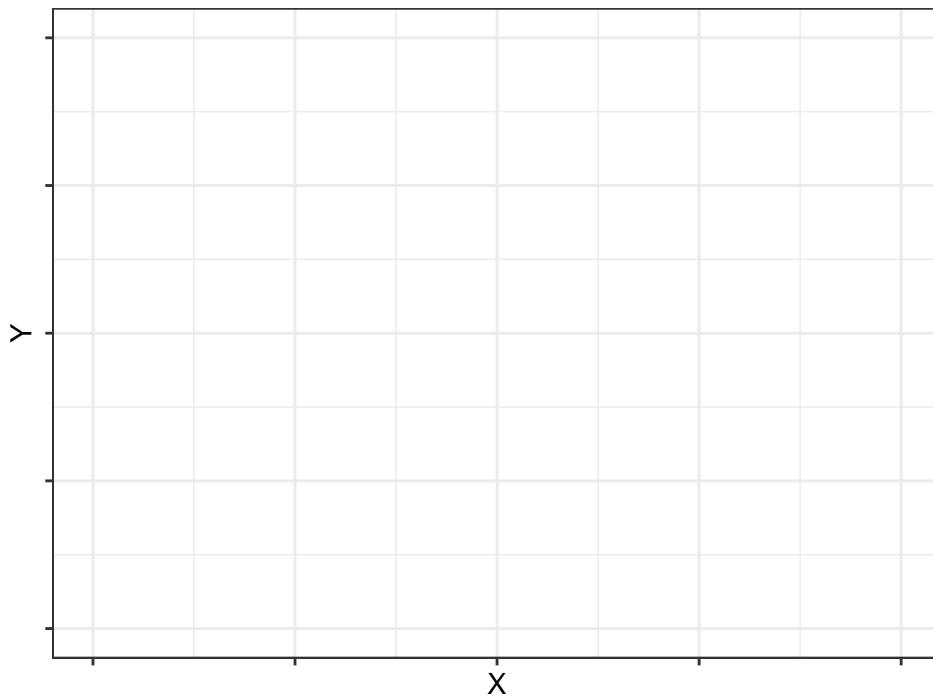
(6 Punkte)

1. Skizzieren Sie in die unten stehenden, freien Abbildungen ein kausales und ein prädiktives Modell mit  $n = 5$  Beobachtungen! **(4 Punkte)**
2. Beachten Sie bei der Erstellung der Skizze, ob ein Effekt von X vorliegt oder nicht! **(2 Punkte)**

Causal model with no effect of X



Predictive model with no effect of X





## 74 Aufgabe



(7 Punkte)

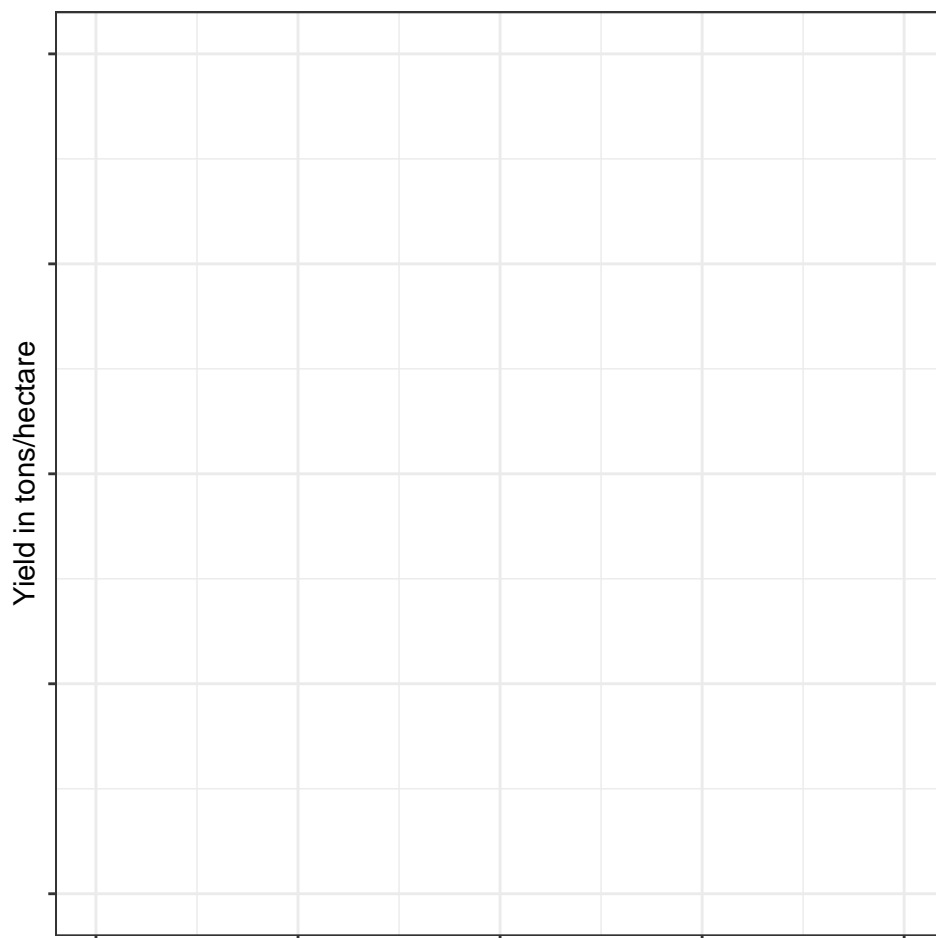
1. Ergänzen Sie vor den jeweiligen Wörtern in der runden Klammer ein  $Y$  oder ein  $X$ , je nachdem welchen der beiden Teile eines statistischen Modells  $Y \sim X$  das Wort zugeordnet ist! (5 Punkte)
  - ( ) simple
  - ( ) mehrere Spalten
  - ( ) Effektschätzer
  - ( ) multiple
  - ( ) eine Spalte
  - ( ) risk factor
  - ( ) outcome
  - ( ) endpoint
  - ( ) Verteilungsfamilie
  - ( ) variable
2. Skizzieren Sie ein simples und ein multiples lineares Regressions Modell! (2 Punkte)

## 75 Aufgabe

(9 Punkte)

Ein Feldexperiment wurde mit  $n = 200$  Pflanzen durchgeführt. Folgende Einflussvariablen ( $x$ ) wurden erhoben: S, dry und rainfall. Als mögliche Outcomevariablen stehen Ihnen nun folgende gemessene Endpunkte zu Verfügung: drymatter, yield, count, quality\_score und dead.

1. Wählen Sie ein Outcome was zu der Verteilungsfamilie *Gaussian* gehört! **(1 Punkt)**
2. Schreiben Sie das Modell in der Form  $y \sim x$  wie es in  üblich ist *ohne Interaktionsterm*! **(2 Punkte)**
3. Schreiben Sie das Modell in der Form  $y \sim x$  wie es in  üblich ist und ergänzen Sie *einen* Interaktionsterm nach Wahl! **(2 Punkte)**
4. Zeichnen Sie eine *schwache* Interaktion in die Abbildung unten für den Endpunkt *yield*. Ergänzen Sie eine aussagekräftige Legende. Wie erkennen Sie eine Interaktion? Begründen Sie Ihre Antwort! **(4 Punkte)**



## 76 Aufgabe

(5 Punkte)

Sie führen ein Feldexperiment zur Bestimmung verschiedener Schlangen in verschiedenen Habitaten durch. Sie messen verschiedene Masszahlen zu den Habitaten und Schlangen. Nachdem Sie sechs Schlangen gemessen haben, wollen Sie ein Modell mit *mass* als Outcome erstellen. Sie haben folgende Datentabelle gegeben:

mass	region	hab
6	1	1
8	1	2
5	1	3
7	1	1
9	2	2
11	2	3

1. Stellen Sie das statistische multiple lineare Regressionsmodell in  $\beta$ -schreibweise auf! **(1 Punkt)**
2. Schreiben Sie das statistische Modell in der Matrixform.
  - Schreiben Sie das Modell einmal als effect parametrization! **(2 Punkte)**
  - Schreiben Sie das Modell einmal als mean parametrization! **(2 Punkte)**

## 77 Aufgabe

(8 Punkte)

Nach einem Feldexperiment mit zwei Pestiziden (*RoundUp* und *OutEx*) ergibt sich die folgende Datentabelle mit dem jeweiligen beobachteten Infektionsstatus.

pesticide	infected
OutEx	no
OutEx	yes
OutEx	no
RoundUp	no
RoundUp	no
RoundUp	yes
OutEx	no
OutEx	no
RoundUp	no
RoundUp	yes
OutEx	no
OutEx	yes
RoundUp	no
RoundUp	yes
OutEx	no
OutEx	yes
OutEx	yes
RoundUp	yes
RoundUp	yes

1. Stellen Sie in einer 2x2 Tafel den Zusammenhang zwischen dem Pestizid und dem Infektionsstatus dar! **(4 Punkte)**
2. Zeichnen Sie den zugehörigen Mosaic-Plot. Berechnen Sie das Verhältnis pro Spalte! **(2 Punkte)**
3. Wenn das Pestizid keine Auswirkung auf den Infektionsstatus hätte, wie sehe dann der Mosaic-Plot aus? **(2 Punkte)**

## 78 Aufgabe

(6 Punkte)

In einem Experiment zur Dosiswirkung wurden verschiedene Dosisstufen mit einer Kontrollgruppe verglichen. Es wurden vier t-Test für den Mittelwertsvergleich gerechnet und es ergab sich folgende Tabelle mit den rohen p-Werten.

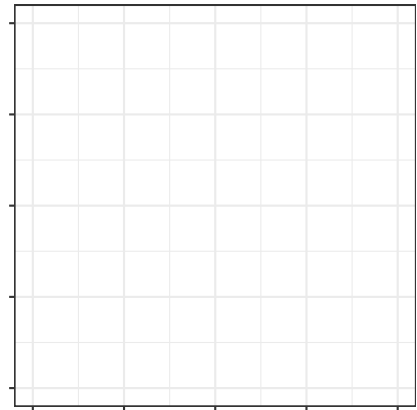
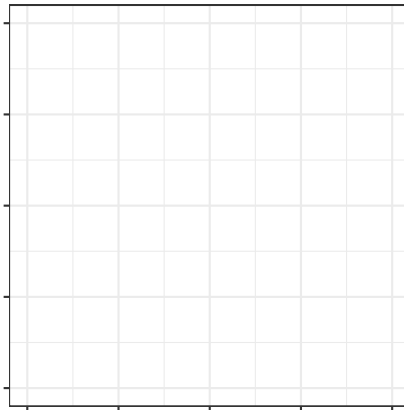
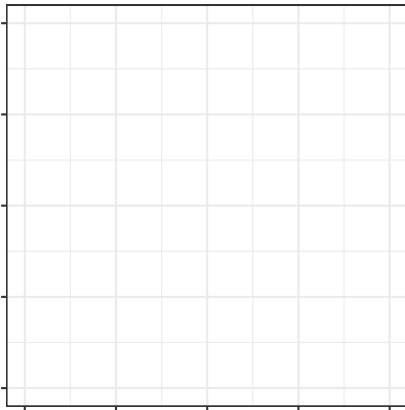
Vergleich	Raw p-val	Adjusted p-val	Reject $H_0$
dose 10 - ctrl	0.030		
dose 15 - ctrl	0.340		
dose 20 - ctrl	0.020		
dose 40 - ctrl	0.001		

1. Füllen Sie die Spalte „adjustierte p-Werte“ mit den adjustierten p-Werten nach Bonferroni aus! **(2 Punkte)**
2. Entscheiden Sie, ob nach der Adjustierung die Nullhypothese weiter abgelehnt werden kann. Tragen Sie Ihre Entscheidung in die obige Tabelle ein. Begründen Sie Ihre Antwort! **(4 Punkte)**

## 79 Aufgabe


(10 Punkte)

1. Zeichnen Sie in die drei untenstehenden, leeren Abbildungen die Zeile des Regressionskreuzes der Poissonverteilung. Wählen Sie die Beschriftung der y-Achse sowie der x-Achse entsprechend aus! **(6 Punkte)**
2. Welchen Effektschätzer erhalten Sie aus der entsprechend linearen Regression? Geben Sie ein Beispiel! **(2 Punkte)**
3. Wenn Sie keinen Effekt erwarten, welchen *Zahlenraum* nimmt dann der Effektschätzer ein? Geben Sie ein Beispiel! **(2 Punkte)**



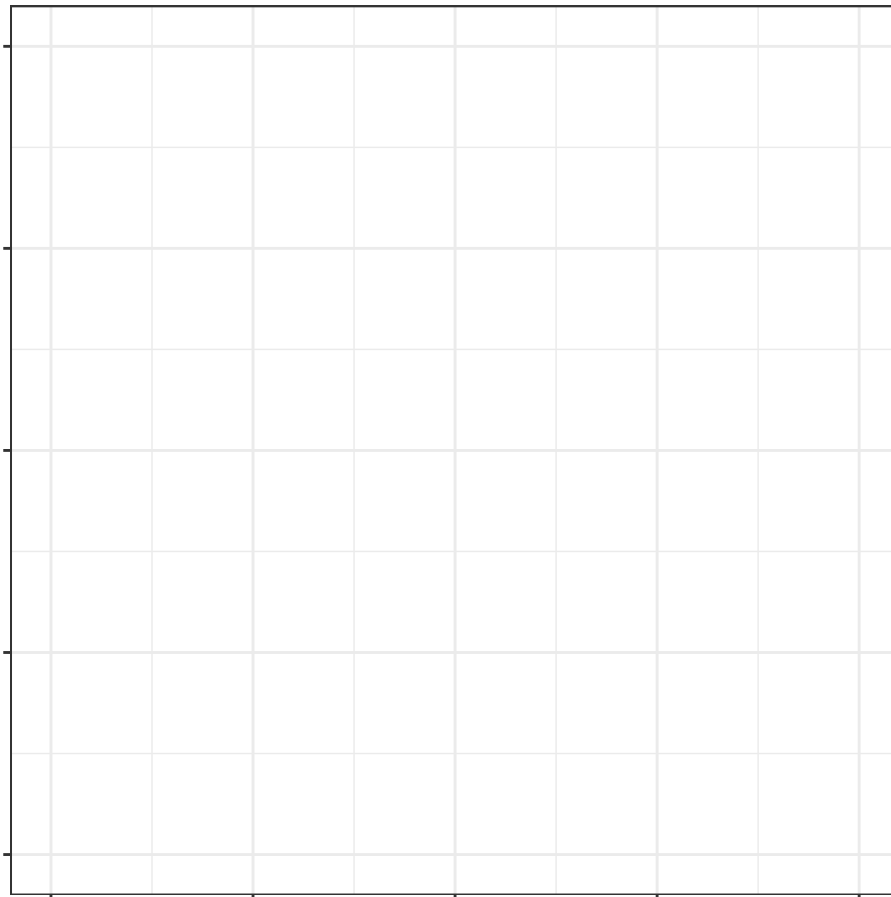
## 80 Aufgabe

(10 Punkte)

In einem Stallexperiment mit  $n = 94$  Ferkeln wurde der Gewichtszuwachs unter bestimmten Lichtverhältnissen gemessen. Sie erhalten den  Output der Funktion `tidy()` einer simplen Gaussian linearen Regression sieben Wochen nach der ersten Messung.

term	estimate	std.error	t statistic	p-value
(Intercept)	25.92	1.17		
light	1.40	0.12		

1. Berechnen Sie die t Statistik für *(Intercept)* und *light*! **(2 Punkte)**
2. Schätzen Sie den p-Wert für *(Intercept)* und *light* mit  $T_k = 1.96$  ab. Was sagt Ihnen der p-Wert aus? Begründen Sie Ihre Antwort! **(3 Punkte)**
3. Zeichnen Sie die Gerade aus der obigen Tabelle in die untenstehende Abbildung! **(1 Punkt)**
4. Beschriften Sie die Abbildung und die Gerade mit den statistischen Kenngrößen! **(2 Punkte)**
5. Formulieren Sie die Regressionsgleichung! **(2 Punkte)**



## 81 Aufgabe

(4 Punkte)

1. Skizzieren Sie in die unten stehende, freie Abbildung die Abbildung, die sich nach der Überschrift ergibt! **(2 Punkte)**
2. Beschriften Sie die Achsen entsprechend! **(2 Punkte)**

Residual plot with 2 outlier fulfilling the normality assumption.





## 82 Aufgabe

(4 Punkte)

1. Skizzieren Sie in die unten stehende, freie Abbildung die Abbildung, die sich nach der Überschrift ergibt! **(2 Punkte)**
2. Beschriften Sie die Achsen entsprechend! **(2 Punkte)**

QQ plot with residuals fulfilling the normality assumption.



### 83 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 Spargel zu bestimmen. Sie erhalten folgende Datentabelle.

.id	drymatter	water	.fitted	.resid
1	29.3	11.3	27.2	
2	26.0	8.7	24.2	
3	30.5	14.3	30.7	
4	28.0	12.0	28.0	
5	31.3	16.7	33.4	
6	16.7	4.4	19.2	
7	27.9	12.4	28.4	
8	25.5	8.0	23.4	
9	27.3	10.9	26.7	
10	19.1	4.0	18.8	
11	19.3	5.8	20.9	


1. Ergänzen Sie die Werte in der Spalte `.resid` in der obigen Tabelle. Geben Sie den Rechenweg und Formel mit an! **(4 Punkte)**
2. Zeichnen Sie den sich aus der obigen Tabelle ergebenden Residualplot. Beschriften Sie die Abbildung! **(4 Punkte)**
3. Gibt es auffällige Werte anhand des Residualplots? Begründen Sie Ihre Antwort! **(2 Punkte)**

## 84 Aufgabe



(8 Punkte)

In verschiedenen Flüssen (*stream*) wurde die Anzahl an Knochenhechten (*longnose*) gezählt. Daneben wurden noch andere Eigenschaften der entsprechenden Flüsse gemessen. Es ergibt sich folgender Auszug aus den Daten.

stream	longnose	no3	temp	so4	do2
LITTLE_BEAR_CR	6	0.81	14.9	8.16	9.1
BEAVER_DAM_CR	19	5.44	17.0	16.53	9.2
PRETTYBOY_BR	19	6.81	19.2	9.20	9.8
NORTH_BR_CASSELMAN_R	16	0.29	18.0	2.50	7.4

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

```
##
## Call:
## glm(formula = reformulate(response = "longnose", termlabels = wanted_vec),
##      family = quasipoisson, data = data_tbl)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -7.72688  -3.95349  -1.99173   0.19015  22.82786
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.7655407  2.5784819   0.6847  0.4967
## no3          0.1498918  0.0929390   1.6128  0.1132
## temp         0.0451759  0.0703941   0.6418  0.5240
## so4         -0.0074486  0.0262792  -0.2834  0.7780
## do2          0.0587911  0.2070569   0.2839  0.7777
##
## (Dispersion parameter for quasipoisson family taken to be 47.41686)
##
## Null deviance: 1722.18  on 53  degrees of freedom
## Residual deviance: 1567.15  on 49  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```


1. Warum wurde hier eine Poisson bzw. Quasipoisson-Verteilung gewählt? Begründen Sie Ihre Antwort mit dem  Output! **(2 Punkte)**
2. Können Sie die *Estimate* der einzelnen Einflussvariablen direkt interpretieren? Begründen Sie Ihre Antwort! **(2 Punkte)**
3. Interpretieren Sie die *signifikanten* Effekte auf die Anzahl an Knochenhechten! **(2 Punkte)**
4. Erklären Sie am  Output wie sich die *t value* Spalte errechnet! **(2 Punkte)**

## 85 Aufgabe


(10 Punkte)

Auf einer Erdbeerplantage treten unerwartet häufig infizierte Erdbeerpflanzen auf. In einem Versuch sollen verschiedene Einflussfaktoren auf den Infektionsstatus betrachtet werden. Dafür wurde für jede Erdbeerpflanze gemessen, wieviel Wasser die Pflanze erhalten hat oder ob die Pflanze ein neuartiges Lichtregime erhalten hatte. Zusätzlich wurde die Anzahl an Nematoden im Boden bestimmt. Es ergibt sich folgender Auszug aus den Daten.

	infected	water	light	nematodes
1		11.69	1	7
1		9.84	1	2
0		9.11	0	4
1		8.32	1	1

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

```
## # A tibble: 3 x 4
##   term          std.error statistic p.value
##   <chr>         <dbl>      <dbl>   <dbl>
## 1 (Intercept)    1.51        1.18    0.237
## 2 nematodes      0.132       -0.498   0.618
## 3 water          0.143       -0.942   0.346
```


1. Die Spalte *estimate* wurde gelöscht. Berechnen Sie die Werte der Spalte *estimate* aus den  Output! **(2 Punkte)**
2. Welche Einflussfaktoren sind protektiv, welche ein Risiko? Berechnen Sie hierfür zunächst das OR aus der Spalte *estimate*! **(4 Punkte)**
3. Interpretieren Sie die Spalte *estimate* im Bezug auf den Infektionsstatus der Erdbeerpflanzen! **(2 Punkte)**
4. Was ist der Unterschied zwischen einem OR und einem RR? Geben Sie ein numerisches Beispiel! **(2 Punkte)**

## 86 Aufgabe

(7 Punkte)

Maispflanzen sollen auf die ertragssteigernde Wirkung von verschiedenen Einflussfaktoren untersucht werden. Gemessen wurde als Outcome die Trockenmasse in  $\text{kg/m}^2$ . Dafür wurde für jede Maispflanze gemessen wieviel Wasser ( $\text{l/m}^2$ ) 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 ( $\mu\text{g/kg}$ ) des Bodens. Es ergibt sich folgender Auszug aus den Daten.

water	light	P	Fe	drymatter	nematodes
10.75	0	10.03	99.97	69.80	11
7.95	1	8.24	97.44	69.68	9
11.77	0	8.93	101.46	73.32	7
11.23	0	8.46	99.96	69.19	9

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       1Q   Median       3Q      Max
## -6.79850 -1.23246 -0.07872  1.25436  5.09928
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  4.690538    9.234814   0.5079    0.61306
## P           -0.373458    0.183418  -2.0361    0.04542 *
## Fe            0.684522    0.093861   7.2929 0.0000000003176 ***
## light         1.415222    0.602241   2.3499    0.02152 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.3252 on 72 degrees of freedom
## Multiple R-squared:  0.47122, Adjusted R-squared:  0.44919
## F-statistic: 21.388 on 3 and 72 DF, p-value: 0.00000000052002
```


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

## 87 Aufgabe

(6 Punkte)

Sie erhalten folgende R Ausgabe der Funktion `lm()`.

```
##
## Call:
## lm(formula = rsp ~ trt, data = data_tbl)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.40  -2.55   1.40   2.35   4.60
##
## Coefficients:
##              Estimate Std. Error t value    Pr(>|t|)
## (Intercept)  18.6000     1.6763   11.096 0.000003886 ***
## trtB         -2.2000     2.3707   -0.928    0.3805
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.7483 on 8 degrees of freedom
## Multiple R-squared:  0.097189, Adjusted R-squared:  -0.015663
## F-statistic: 0.86121 on 1 and 8 DF,  p-value: 0.38055
```


1. Ist die Annahme der Normalverteilung an das Outcome *rsp* erfüllt? Begründen Sie die Antwort! **(2 Punkte)**
2. Wie groß ist der Effekt des *Trt*? Liegt ein signifikanter Effekt vor? Geben Sie die Formel und den Rechenweg mit an! **(2 Punkte)**
3. Schreiben Sie das Ergebnis der  Ausgabe in einen Satz nieder, der die Information zum Effekt und der Signifikanz enthält! **(2 Punkte)**

## 88 Aufgabe

(11 Punkte)

In einem Experiment zur Steigerung der Milchleistung (*gain*) 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: 793.2
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.92778 -0.60744 -0.07438  0.59675  3.34077
##
## Random effects:
## Groups   Name      Variance Std.Dev.
## barn     (Intercept) 225.19   15.006
## subject  (Intercept) 3956.81  62.903
## Residual                    642.72  25.352
## Number of obs: 83, groups:  barn, 7; subject, 6
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 202.7895    26.5884  7.6270
## attitudepop -21.0583     5.5704 -3.7804
##
## Correlation of Fixed Effects:
##              (Intr)
## attitudepop -0.103
```

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  Ausgabe in einen Satz nieder, der die Information zum Effekt und der Signifikanz enthält! **(2 Punkte)**

## 89 Aufgabe

(9 Punkte)

Nach einem Pilotversuch können Sie für vier Düngestufen die jeweiligen Erträge ermitteln. Für den Hauptversuch wollen Sie nun die benötigte Fallzahl abschätzen.

1. Skizzieren Sie den zugehörigen Boxplot für die vier Düngestufen mit folgenden Informationen.

treatment	mean	median	min	max	quartile_1st	quartile_3rd	var
A	11.29	11	9	15	9.5	12.5	4.90
B	15.71	17	10	17	16.0	17.0	6.90
C	11.86	12	8	18	8.0	14.5	16.14
D	11.71	12	9	14	11.0	12.5	2.57

Beschriften Sie die Achsen entsprechend! (4 Punkte)

2. Berechnen Sie die benötigten Fallzahlen für die Gruppenvergleiche mit

$$n_{Gruppe} = \frac{2 \cdot (1.96 + 0.77)^2}{\left( \frac{\Delta}{s_{pool}} \right)^2}$$

Nutzen Sie hierfür das gepoolte  $s_{pool}$  mit  $s_{pool} = (s_1 + s_2)/2$ ! (4 Punkte)

3. Was ist die Fallzahl, die Sie für den Versuch benötigen? (1 Punkte)



## 90 Aufgabe

**(10 Punkte)**

In einem Feldexperiment für die Bodendurchlässigkeit wurde der Niederschlag pro Parzelle sowie der durchschnittliche Ertrag gemessen. Es ergibt sich folgende Datentabelle.

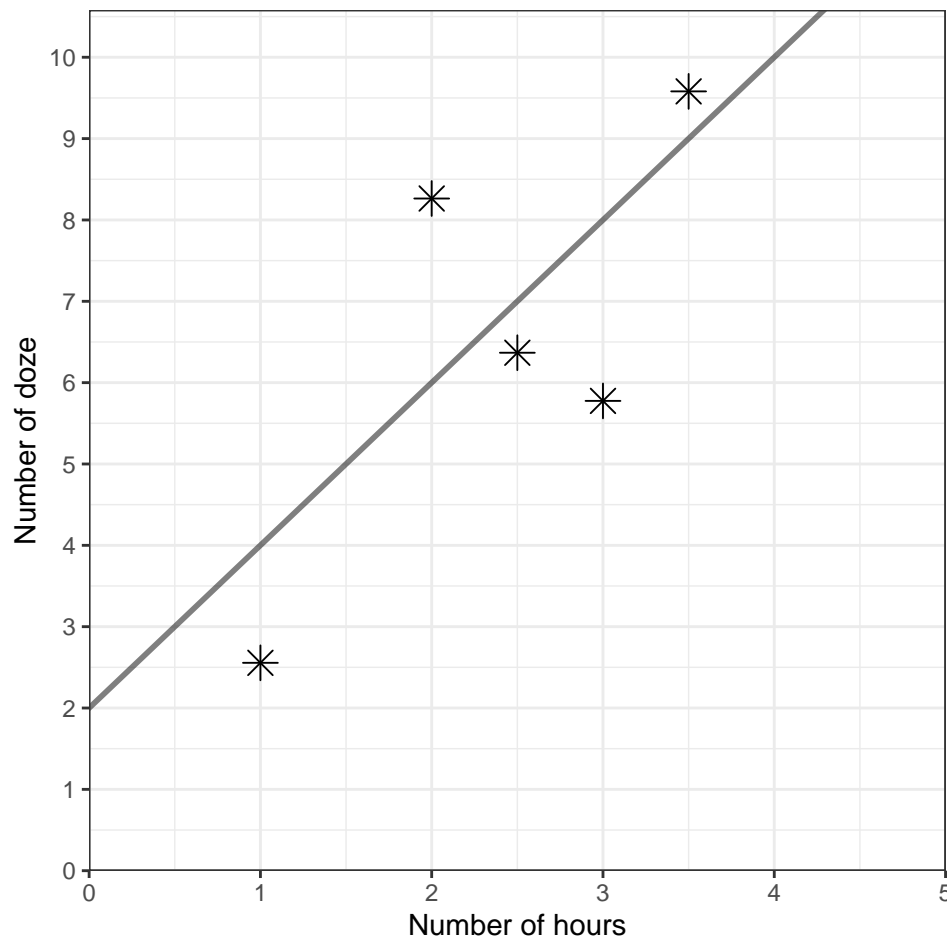
water	drymatter
19	23
22	24
21	26
22	17
16	26

1. Erstellen Sie den Scatter-Plot für die Datentabelle. Beschriften Sie die Achsen entsprechend! **(4 Punkte)**
2. Zeichnen Sie eine Gerade durch die Punkte! **(1 Punkt)**
3. Beschriften Sie die Gerade mit den gängigen statistischen Maßzahlen! **(3 Punkte)**
4. Wenn kein Effekt von dem Niederschlag auf das Trockengewicht vorhanden wäre, wie würde die Gerade verlaufen und welche Werte würden die statistischen Maßzahlen annehmen? **(2 Punkt)**

## 91 Aufgabe

(7 Punkte)

In einer Studie zur „Arbeitssicherheit auf dem Feld“ wurde gemessen wie viele Stunden auf einem Feld gefahren wurden und wie oft der Fahrer dabei drohte einzunicken. Es ergab sich folgende Abbildung.



1. Erstellen Sie die Regressionsgleichung aus der obigen Abbildung in der Form  $y \sim \beta_0 + \beta_1 \cdot x$ ! **(2 Punkte)**
2. Beschriften Sie die Gerade mit den Parametern der linearen Regressionsgleichung! **(2 Punkte)**
3. Liegt ein Zusammenhang zwischen der Anzahl an gefahrenen Runden und der Müdigkeit vor? Begründen Sie Ihre Antwort! **(2 Punkte)**
4. Wenn kein Zusammenhang zu beobachten wäre, wie würde die Gerade aussehen? **(1 Punkt)**

## 92 Aufgabe

**(6 Punkte)**

1. Erklären Sie den Zusammenhang zwischen Stichprobe und Grundgesamtheit an einem Schaubild! **(3 Punkte)**
2. Was ist der Unterschied zwischen  $\mu$  und  $\sigma$  und  $\bar{x}$  und  $s$  im Kontext der Stichprobe und Grundgesamtheit? **(2 Punkte)**
3. Warum müssen wir überhaupt zwischen einer Stichprobe und einer Grundgesamtheit unterscheiden? **(1 Punkt)**

### 93 Aufgabe

(6 Punkte)

Geben ist folgende 2x2 Kreuztabelle.


1. Tragen Sie folgende Fachbegriffe korrekt in die 2x2 Kreuztabelle ein! (6 Punkte)

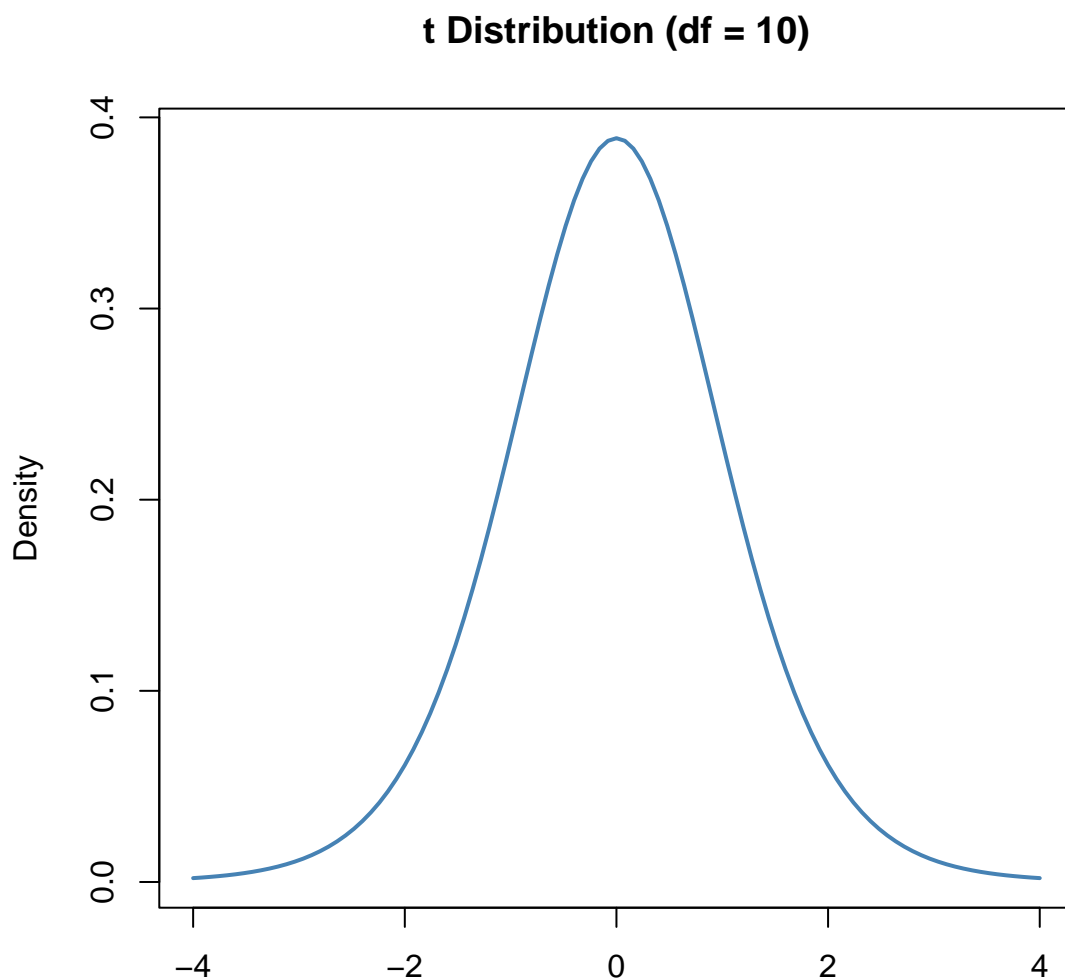
- (Unbekannte) Wahrheit
- $H_0$  wahr
- $H_0$  falsch
- $H_0$  abgelehnt
- $H_0$  beibehalten
- Testentscheidung
- $\alpha$ -Fehler
- $\beta$ -Fehler
- Richtige Entscheidung
- 5%
- 20%

## 94 Aufgabe

(6 Punkte)

Im folgenden ist eine t-Verteilung mit 10 Freiheitsgraden abgebildet. Ergänzen Sie die Abbildung wie folgt.

1. Zeichnen Sie das zweiseitige  $\alpha$  Niveau in die Abbildung! **(2 Punkte)**
2. Zeichnen Sie einen signifikant p-Wert in die Abbildung! **(1 Punkt)**
3. Ergänzen Sie „ $\bar{x}_1 = \bar{x}_2$ “! **(1 Punkt)**
4. Ergänzen Sie „ $H_0$  ist wahr“! **(1 Punkt)**
5. Ergänzen Sie eine t-Verteilung mit 5 Freiheitsgraden. Zeichnen Sie im Zweifel über den Rand der Abbildung! **(1 Punkte)**



## 95 Aufgabe

(8 Punkte)

Sie rechnen einen t-Test. Sie schätzen einen Mittelwertsunterschied.

1. Beschriften Sie die untenstehende Abbildung mit der Signifikanzschwelle! **(1 Punkt)**
2. Ergänzen Sie eine Relevanzschwelle! **(1 Punkt)**
3. Skizzieren Sie in die untenstehende Abbildung fünf einzelne Konfidenzintervalle (a-e) mit den jeweiligen Eigenschaften! **(6 Punkte)**
  - (a) Ein Konfidenzintervall mit niedriger Fallzahl  $n$  in der Stichprobe als der Rest der Konfidenzintervalle
  - (b) Ein nicht signifikantes, relevantes Konfidenzintervall
  - (c) Ein signifikantes, relevantes Konfidenzintervall
  - (d) Ein Konfidenzintervall mit höherer Fallzahl  $n$  in der Stichprobe als der Rest der Konfidenzintervalle
  - (e) Ein signifikantes, nicht relevantes Konfidenzintervall



## 96 Aufgabe

(6 Punkte)

Gegeben ist die vereinfachte Formel für den Zweistichproben t-Test mit der gepoolten Standardabweichung  $s_p$  und gleicher Gruppengrösse  $n_g$  der beiden Sample.

$$T = \frac{\bar{y}_1 - \bar{y}_2}{s_p \cdot \sqrt{\frac{2}{n_g}}}$$

Welche Auswirkung hat die Änderungen der jeweiligen statistischen Masszahl auf den T-Wert und damit auf die *vermutliche* Signifikanz? Füllen Sie hierzu die untenstehende Tabelle aus! **(6 Punkte)**

	T Statistik	$Pr(D H_0)$	$KI_{1-\alpha}$		T Statistik	$Pr(D H_0)$	$KI_{1-\alpha}$
$\Delta \uparrow$				$\Delta \downarrow$			
$s \uparrow$				$s \downarrow$			
$n \uparrow$				$n \downarrow$			

## 97 Aufgabe

(8 Punkte)

Sie haben folgende Aussage gegeben.

### **Bin ich im Herbst?**

1. Erklären Sie den Gedankengang der Testtheorie sowie des Falsifikationsprinzips an der Aussage! **(4 Punkte)**
2. Erklären Sie Ihre Entscheidung zu der Aussage! **(3 Punkte)**
3. Schätzen Sie den p-Wert zu der Aussage ab! **(1 Punkt)**



## 98 Aufgabe

**(8 Punkte)**

»Ohne ausreichende Zufuhr von Vitamin C stellen sich nach 73 Tagen die ersten Symptome ein; die ersten Toten sind nach 101 Tagen zu beklagen; nach 116 Tagen rafft die Skorbut eine ganze Schiffsbesatzung dahin.«

1. Stellen Sie den Verlauf der Skorbuterkrankung auf einem Schiff mit einer typischen Anzahl an Matrosen dar. Welche Annahmen haben Sie getroffen? Beschriften Sie die Achsen entsprechend! **(4 Punkte)**
2. Schifffarzt James Lind (1716-1794) ist dem Phänomen Skorbut systematisch nachgegangen. Wie würden Sie solch ein Experiment planen und welche (ethischen) Probleme sehen Sie? **(4 Punkte)**

## 99 Aufgabe

**(11 Punkte)**

Der Wissenschaftler Pettenkofer arbeitete streng naturwissenschaftlich-experimentell und gilt als Begründer der experimentellen Hygiene ("Konditionalhygiene"). Auch seine Untersuchungen zu Kleidung, Heizung, Lüftung, Kanalisation und Wasserversorgung trugen experimentelle Züge. Pettenkofer war ein Positivist, das heißt, er erkannte ausschließlich sichtbare, zum Beispiel in Experimenten gewonnene Tatsachen als Erkenntnisquelle an.

Pettenkofer unterlief ein Irrtum, der bis heute nachwirkt, indem viele Menschen glauben, es gebe eine "Atmende Wand": Er stellte bei frühen Luftwechsel-Messungen in einem Zimmer fest, dass sich nach dem vermeintlichen Abdichten sämtlicher Fugen die Luftwechselrate weniger als erwartet verminderte. Daraus schlussfolgerte er einen erheblichen Luftaustausch durch die Ziegelwände hindurch. Vermutlich kam er nicht darauf, den Kamin eines im Raum befindlichen Ofens abzudichten. Luftaustausch durch die Zimmerwände hindurch sei, so Pettenkofer, ein wesentlicher Beitrag zur Reinigung der Raumluft.

1. Bestimmen Sie das  $y$  bzw. den Endpunkt/Outcome für das Experiment "Atmende Wand". Wie messen Sie den Endpunkt? **(1 Punkt)**
2. Bestimmen Sie mögliche Einflussfaktoren auf den Endpunkt! Begründen Sie Ihre Entscheidung! **(2 Punkt)**
3. Beschreiben Sie das Modell in der Form  $y \sim x$ ! **(1 Punkt)**
4. Wie realisieren Sie Wiederholung in dem Experiment? Skizzieren Sie das Experiment! **(2 Punkte)**
5. Wie könnten Sie das Experiment graphisch darstellen? Was befindet sich auf der x-Achse und was auf der y-Achse? Beschriften Sie die Achsen entsprechend! **(2 Punkte)**
6. Basiert die heutige Wissenschaft auch auf dem Positivismus? Begründen Sie Ihre Antwort! **(2 Punkte)**

## 100 Aufgabe

**(12 Punkte)**

Nach einem Experiment mit zwei Pestiziden (*RoundUp* und *GoneEx*) ergibt sich die folgende Datentabelle mit dem gemessenen Trockengewicht (*drymatter*) von Weizen.

pesticide	drymatter
RoundUp	20
GoneEx	7
RoundUp	19
GoneEx	14
RoundUp	19
GoneEx	14
GoneEx	14
GoneEx	12
RoundUp	19
RoundUp	18
GoneEx	18
RoundUp	20
GoneEx	15

1. Formulieren Sie die Fragestellung! **(1 Punkt)**
2. Formulieren Sie das statistische Hypothesenpaar! **(2 Punkte)**
3. Bestimmen Sie die Teststatistik  $T_{calc}$  eines Student t-Tests für den Vergleich der beiden Pestizide. Geben Sie den Rechenweg und die Formeln mit an! **(5 Punkte)**
4. Treffen Sie mit  $T_{\alpha=5\%} = 2.04$  und dem berechneten  $T_{calc}$  eine Aussage zur Nullhypothese! **(2 Punkte)**
5. Wenn Sie keinen Unterschied zwischen den beiden Pestiziden erwarten würden, wie große wäre dann die Teststatistik  $T_{calc}$ ? Begründen Sie Ihre Antwort! **(2 Punkte)**

## 101 Aufgabe

**(13 Punkte)**


Das Gewicht von Küken wurde vor der Behandlung mit STARTex und 1 Woche nach der Behandlung gemessen. Es ergibt sich die folgende Datentabelle.

animal_id	before	after
1	12	22
2	14	22
3	15	26
4	6	32
5	11	23
6	13	38
7	7	25

1. Formulieren Sie die Fragestellung! **(1 Punkt)**
2. Formulieren Sie das statistische Hypothesenpaar! **(2 Punkte)**
3. Bestimmen Sie die Teststatistik  $T_{calc}$  eines gepaarten t-Tests für den Vergleich der beiden Zeitpunkte. Geben Sie den Rechenweg und die Formeln mit an! **(4 Punkte)**
4. Treffen Sie mit  $T_{\alpha=5\%} = 2.04$  und dem berechneten  $T_{calc}$  eine Aussage zur Nullhypothese! **(2 Punkte)**
5. Wenn Sie keinen Unterschied zwischen den beiden Zeitpunkten erwarten würden, wie große wäre dann die Teststatistik  $T_{calc}$ ? Begründen Sie Ihre Antwort! **(2 Punkte)**
6. Schätzen Sie  $Pr(D|H_0)$  ab. Begründen Sie Ihre Antwort! **(2 Punkte)**

## 102 Aufgabe

(10 Punkte)


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

```
##
## Two Sample t-test
##
## data:  freshmatter by N
## t = -1.71884, df = 12, p-value = 0.11131
## alternative hypothesis: true difference in means between group high and group low is not equal to 0
## 95 percent confidence interval:
##  -8.16338637  0.96338637
## sample estimates:
## mean in group high  mean in group low
##                14.4                18.0
```

1. Formulieren Sie das statistische Hypothesenpaar! **(2 Punkte)**
2. Liegt ein signifikanter Unterschied zwischen den Gruppen vor? Begründen Sie Ihre Antwort! **(2 Punkte)**
3. Skizzieren Sie eine Abbildung in der Sie  $T_{calc}$ ,  $Pr(D|H_0)$ ,  $A = 0.95$ , sowie  $T_{\alpha=5\%} = |2.18|$  einzeichnen! **(4 Punkte)**
4. Beschriften Sie die Abbildung und den Graphen entsprechend! **(2 Punkte)**

## 103 Aufgabe

(8 Punkte)


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

```
##  
## Two Sample t-test  
##  
## data:  freshmatter by N  
## t = -0.797347, df = 12, p-value = 0.44074  
## alternative hypothesis: true difference in means between group high and group low is not equal to 0  
## 95 percent confidence interval:  
## -4.2658050  1.9800907  
## sample estimates:  
## mean in group high  mean in group low  
##          13.714286          14.857143
```

1. Formulieren Sie das statistische Hypothesenpaar! **(2 Punkte)**
2. Liegt ein signifikanter Unterschied zwischen den Gruppen vor? Begründen Sie Ihre Antwort! **(2 Punkte)**
3. Skizzieren Sie das sich ergebende Konifidenzintervall! **(2 Punkte)**
4. Beschriften Sie die Abbildung und den Graphen entsprechend! **(2 Punkte)**

## 104 Aufgabe

(8 Punkte)

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

```
##  
## Two Sample t-test  
##  
## data:  freshmatter by N  
## t = -1.87356, df = 10, p-value = 0.090476  
## alternative hypothesis: true difference in means between group high and group low is not equal to 0  
## 95 percent confidence interval:  
## -6.69285555  0.57856984  
## sample estimates:  
## mean in group high  mean in group low  
##          14.142857          17.200000
```

1. Formulieren Sie das statistische Hypothesenpaar! **(2 Punkte)**
2. Liegt ein signifikanter Unterschied zwischen den Gruppen vor? Begründen Sie Ihre Antwort! **(2 Punkte)**
3. Skizzieren Sie die sich ergebenden Boxplots! Welche Annahmen an die Daten haben Sie getroffen? Begründen Sie Ihre Antwort! **(4 Punkte)**

## 105 Aufgabe

(12 Punkte)

Die Anzahl an Nematoden wurde vor und nach einer Behandlung mit einem bioaktiven Dünger gezählt. Es ergibt sich folgende Datentabelle.

Vorher	Nachher	Differenz	Vorzeichen	Rang	Positiv Rang	Negativ Rang
9	14					
7	13					
9	10					
11	12					
8	12					
11	9					
10	14					
9	12					
8	10					
12	13					
8	11					
10	10					
9	12					
10	10					
11	12					

1. Ergänzen Sie die obige Tabelle mit den notwendigen Informationen, die Sie benötigen um einen Wilcoxon-Vorzeichen-Rang-Test zu rechnen! **(4 Punkte)**
2. Bestimmen Sie die Teststatistik  $W$  mit  $W = \min(T_-; T_+)$  und berechnen Sie den erwarteten Wert  $\mu_W = \frac{n \cdot (n + 1)}{4}$ ! **(2 Punkte)**
3. Berechnen Sie anschließend den z-Wert mit  $z = \frac{W - \mu_W}{0}$ ! **(2 Punkte)**
4. Liegt mit einer Signifikanzschwelle von  $z = 1.96$  ein Unterschied zwischen den beiden Zeitpunkten vor? Begründen Sie Ihre Antwort! **(2 Punkte)**
5. Berechnen Sie die Effektstärke mit  $r = |\frac{z}{\sqrt{n}}|$  und interpretieren Sie die Effektstärke! **(2 Punkte)**



## 106 Aufgabe

(8 Punkte)

Nach einer Behandlung mit RootsGoneX wurde die mittlere Anzahl an Wurzeln an der invasiven Lupine (*Lupinus polyphyllus*) gezählt. Es ergab sich folgender Datensatz an mittleren Wurzelanzahl.

Treatment	Count
RootsGoneX	8.5
RootsGoneX	8.5
RootsGoneX	4.0
RootsGoneX	9.5
RootsGoneX	7.5
RootsGoneX	9.5
RootsGoneX	8.0
Kontrolle	19.3
Kontrolle	12.8
Kontrolle	8.9
Kontrolle	10.3
Kontrolle	10.4
Kontrolle	9.6
Kontrolle	6.2

Rechnen Sie einen Mann-Whitney-U-Test auf den obigen Daten.

1. Bestimmen Sie hierfür  $U_c$  mit  $U_c = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$ ! (4 Punkte)

2. Geben Sie eine Aussage über die Signifikanz von  $U_c$  durch  $z = \frac{U_c - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$  und dem kritischen Wert von  $z = 1.96$ . Begründen Sie Ihre Antwort! (2 Punkte)

3. Berechnen Sie die Effektstärke mit  $r = |\frac{z}{\sqrt{n}}|$  und interpretieren Sie die Effektstärke! (2 Punkte)

## 107 Aufgabe

(9 Punkte)

Die Anzahl an Blüten der Vanillepflanze pro Box wurde nach der Gabe von zusätzlichen Phosphorlösung (Kontrolle, Dosis 20 und Dosis 40) bestimmt. Es ergeben sich folgende nach der Anzahl der Blüten geordnete Daten.

Treatment	Count	Rang Kontrolle	Rang Dosis 20	Rang Dosis 40
Kontrolle	7.6			
Kontrolle	8.4			
Kontrolle	9.9			
Kontrolle	10.0			
Kontrolle	10.5			
Dosis 20	10.6			
Dosis 20	11.0			
Kontrolle	11.4			
Dosis 20	11.8			
Kontrolle	11.9			
Dosis 40	12.0			
Dosis 20	12.5			
Dosis 40	12.9			
Dosis 20	14.2			
Dosis 20	14.3			
Dosis 40	15.1			
Dosis 40	15.4			
Dosis 40	20.9			

Rechnen Sie einen Kruskal-Wallis-Test auf den obigen Daten.

- Bestimmen Sie hierfür  $H_c$  mit  $H_c = \frac{12}{n(n+1)} \left( \frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3} \right) - 3(n+1)!$  (6 Punkte)
- Geben Sie eine Aussage über die Signifikanz von  $H_c$  durch den kritischen Wert von  $H = 5.99!$  (1 Punkt)
- Wie lautet die Nullhypothese die Sie mit dem Kruskal-Wallis-Test überprüfen? (1 Punkt)
- Was sagt ein signifikantes Ergebnis des Kruskal-Wallis-Test in Bezug auf die einzelnen Gruppenvergleiche aus? (1 Punkt)

## 108 Aufgabe

(6 Punkte)

1. Folgende statistische Fachbegriffe bzw. Methoden sind gegeben:  
Kruskal-Wallis-Test, Simple Gaussian linear Regression, t-Test, Mann-Whitney-U-Test, Wilcoxon-Test, Paired t-Test, Pearson Korrelation, ANOVA und Spearman Korrelation.
2. Ordnen Sie die Begriffe den untenstehenden Boxen, *Parametrische Methoden* und *Nicht-parametrische Methoden*, zu! **(3 Punkte)**
3. Verbinden Sie – *wenn möglich* – ähnliche Methoden mit einer Linie! **(2 Punkte)**
4. Wie unterscheiden sich parametrische von nicht-parametrischen Methoden? **(1 Punkt)**

### Parametrische Methoden

### Nicht-parametrische Methoden

## 109 Aufgabe

(6 Punkte)

Ergänzen Sie folgende Worte auf die freien Linien.

*wahren Effekt  $\delta$ , Power  $1 - \beta$ , groß, klein, steigenden, sinkenden, Varianz  $\sigma^2$ , sinkende, statistischen Test, sinkende, Fehler 1. Art  $\alpha$*

**Die benötigte Fallzahl steigt** für \_\_\_\_\_ Fehler 1. Art  $\alpha$  **(0.5 Punkte)**  
für \_\_\_\_\_ wahren Effekt  $\delta$  **(0.5 Punkte)**  
für \_\_\_\_\_ Power  $1 - \beta$  **(0.5 Punkte)**  
für \_\_\_\_\_ Varianz  $\sigma^2$  **(0.5 Punkte)**

Die Information und Annahmen zu \_\_\_\_\_ und \_\_\_\_\_ erhalten Sie aus der Literatur. Die Information zu \_\_\_\_\_ und \_\_\_\_\_ können Sie statistischen Tabellen entnehmen. Die Fallzahl-Formel hängt vom genutzten \_\_\_\_\_ ab. Allgemein gesprochen, ist die Fallzahl so \_\_\_\_\_ wie nötig, aber so \_\_\_\_\_ wie möglich zu halten. **(4 Punkte)**