

## Checkbox für die Version vom 23. Dezember 2022

Die gesamte Klausur beinhaltet aktuell in Summe **108** Fragen. Davon sind **50** Multiple Choice Fragen sowie **58** 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 23. Dezember 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.

## Multiple Choice Aufgaben

- Es wird nie mehr als fünfzig Multiple Choice Fragen geben.
- Im Laufe der Zeit werden einzelne Fragen durch andere Fragen *ersetzt*, bitte beachten Sie diesen Sachstand, wenn Sie eine *Wiederholungsklausur* im nächsten Semester schreiben.

### 1 Aufgabe

(2 Punkte)

Welche Aussage über die *confounder* Adjustierung ist richtig?

- A ☐ Die *confounder* Adjustierung wird meist ignoriert. Wenn die Annahmen an den statistischen Test richtig sind, kann auf eine Adjustierung verzichtet werden.
- B ☐ Die *confounder* 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.
- C ☐ Die *confounder* 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.
- D ☐ Die *confounder* Adjustierung wird durchgeführt um bei multiplen Vergleichen den Fehler 1. Art zu kontrollieren. Es wird die Irrtumswahrscheinlichkeit adjustiert, daher das  $\alpha$ -Niveau.
- E ☐ Die *confounder* 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: 0.21, 0.89, 0.42, 0.03 und 0.001. Sie adjustieren die p-Werte nach Bonferroni. Welche Aussage ist richtig?

- A ☐ Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 0.042, 0.178, 0.084, 0.006 und 0.0002. Die adjustierten p-Werte werden zu einem  $\alpha$ -Niveau von 1% verglichen.
- B ☐ Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 1, 1, 1, 0.15 und 0.005. Die adjustierten p-Werte werden zu einem  $\alpha$ -Niveau von 5% verglichen.
- C ☐ Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 1.05, 4.45, 2.1, 0.15 und 0.005. Die adjustierten p-Werte werden zu einem  $\alpha$ -Niveau von 5% verglichen.
- D ☐ Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 1, 1, 1, 0.15 und 0.005. Die adjustierten p-Werte werden zu einem  $\alpha$ -Niveau von 1% verglichen.
- E ☐ Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 0.042, 0.178, 0.084, 0.006 und 0.0002. Die adjustierten p-Werte werden zu einem  $\alpha$ -Niveau von 5% verglichen.

### 3 Aufgabe

(2 Punkte)

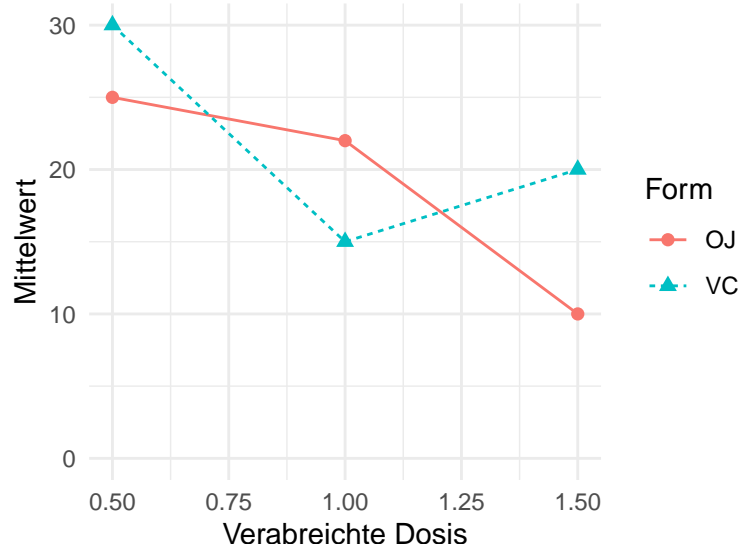
Der Datensatz PlantGrowth enthält das Gewicht von Pflanzen, die unter einer Kontrolle und zwei verschiedenen Behandlungsbedingungen erzielt wurden. Nach der Berechnung einer einfaktoriellen ANOVA ergibt sich ein  $\eta^2 = 0.24$ . Welche Aussage ist richtig?

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

## 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 1.5 mg/Tag) über eine von zwei Verabreichungsmethoden mit Orangensaft (OJ) oder Ascorbinsäure (VC) erhielt.



Welche Aussage ist richtig im Bezug auf eine zweifaktorielle ANOVA?

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

## 5 Aufgabe

(2 Punkte)

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

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

## 6 Aufgabe

(2 Punkte)

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

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

## 7 Aufgabe

(2 Punkte)

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

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

## 8 Aufgabe

(2 Punkte)

Welche Aussage über den Korrelationskoeffizienten nach Pearson ist richtig?

- A** ☐ Der Korrelationskoeffizienten nach Pearson wird genutzt, wenn das Outcome  $Y$  normalverteilt ist. Der Korrelationskoeffizienten liegt zwischen -1 und 1.
- B** ☐ Der Korrelationskoeffizienten nach Pearson wird genutzt, wenn der Korrelationskoeffizienten zwischen -1 und 1 liegt. Dann sind die Residuen normalverteilt.
- C** ☐ Der Korrelationskoeffizienten nach Pearson wird genutzt, wenn das Outcome  $Y$  nicht normalverteilt ist. Der Korrelationskoeffizienten liegt zwischen -1 und 1.

- D** ☐ Der Korrelationskoeffizienten nach Pearson wird genutzt, wenn das Outcome Y normalverteilt ist. Der Korrelationskoeffizienten liegt zwischen 0 und 1.
- E** ☐ Der Korrelationskoeffizienten nach Pearson wird genutzt, wenn das Outcome Y nicht normalverteilt ist. Der Korrelationskoeffizienten liegt zwischen 0 und 1.

## 9 Aufgabe

(2 Punkte)

Berechnen Sie den Mittelwert und Standardabweichung von y mit 10, 3, 6, 9 und 19.

- A** ☐ Es ergibt sich 10.4 +/- 3.01
- B** ☐ Es ergibt sich 8.4 +/- 18.15
- C** ☐ Es ergibt sich 9.4 +/- 36.3
- D** ☐ Es ergibt sich 9.4 +/- 6.02
- E** ☐ Es ergibt sich 9.4 +/- 3.01

## 10 Aufgabe

(2 Punkte)

Berechnen Sie den Median und das IQR von x mit 23, 15, 26, 13, 14, 14 und 42.

- A** ☐ Es ergibt sich 21 [14, 26]
- B** ☐ Es ergibt sich 15 [14, 26]
- C** ☐ Es ergibt sich 21 +/- 14
- D** ☐ Es ergibt sich 15 +/- 14
- E** ☐ Es ergibt sich 15 +/- 26

## 11 Aufgabe

(2 Punkte)

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

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

## 12 Aufgabe

(2 Punkte)

Welche Aussage über Cook's d und Cohen's d ist richtig?

- A** ☐ Wir nutzen Cook's d um Outlier in den Daten zu finden und Cohen's d um standardisierte Outlier für Gruppenvergeliche zu erhalten.
- B** ☐ Wir nutzen Cook's d um Outlier in den Daten zu finden. Cohen's d findet auch Outlier, ist aber ein veraltetes Konzept in der Statistik.



- C** ☐ Wir nutzen Cohen's d um Outlier in den Daten zu finden und Cook's d um einen standardisierten Effektschätzer für Gruppenvergeliche zu erhalten.
- D** ☐ Wir nutzen Cook's d um Outlier in den Daten zu finden und Cohen's d um einen standardisierten Effektschätzer für Gruppenvergeliche zu erhalten.
- E** ☐ Wir nutzen Cook's d um Outlier in den Daten zu finden und Cohen's d um einen nicht standardisierten Effektschätzer für Gruppenvergeliche zu erhalten.

### 13 Aufgabe

(2 Punkte)

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

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

### 14 Aufgabe

(2 Punkte)

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

- A** ☐ Die Verteilung von Daten. Meistens dem Outcome oder auch  $y$  genannt. Hierbei ist eine hohe Fallzahl notwendig mit mehr als 20 Beobachtungen.
- B** ☐ Die Eigenschaften von Daten aufgeteilt nach zwei Gruppen eines Faktors  $x$ .
- C** ☐ Die Verteilung von Daten. Meistens dem Outcome oder auch  $y$  genannt. Die Darstellung ist auch mit einer kleinen Fallzahl von Minimum 5 Beobachtungen möglich.
- D** ☐ Die Eigenschaften von Daten anhand der fehlenden Werte und den Leveln eines Faktors  $x$ .
- E** ☐ Die Verteilung von Daten. Auch 2 bis 5 Beobachtungen können noch dargestellt werden.

### 15 Aufgabe

(2 Punkte)

Nachdem Sie in einem Experiment die Daten  $D$  erhoben haben, berechnen Sie den Mittelwert und den Median. Der Mittelwert  $\bar{y}$  und der Median  $\tilde{y}$  unterscheiden sich nicht. Welche Aussage ist richtig?

- A** ☐ Da sich der Mittelwert und der Median nicht unterscheiden, liegen vermutlich keine Outlier in den Daten vor. Wir verwenden den Datensatz so wie er ist.
- B** ☐ Da sich der Mittelwert und der Median unterscheiden, ist der Datensatz nicht zu verwenden. Mittelwert und Median müssen gleich sein.
- C** ☐ Da sich der Mittelwert und der Median nicht unterscheiden, liegen vermutlich Outlier in den Daten vor.
- D** ☐ Da sich der Mittelwert und der Median unterscheiden, liegen vermutlich Outlier in den Daten vor. Wir untersuchen den Datensatz nach auffälligen Beobachtungen.
- E** ☐ Da sich der Mittelwert und der Median unterscheiden, liegen vermutlich keine Outlier in den Daten vor.

## 16 Aufgabe

(2 Punkte)

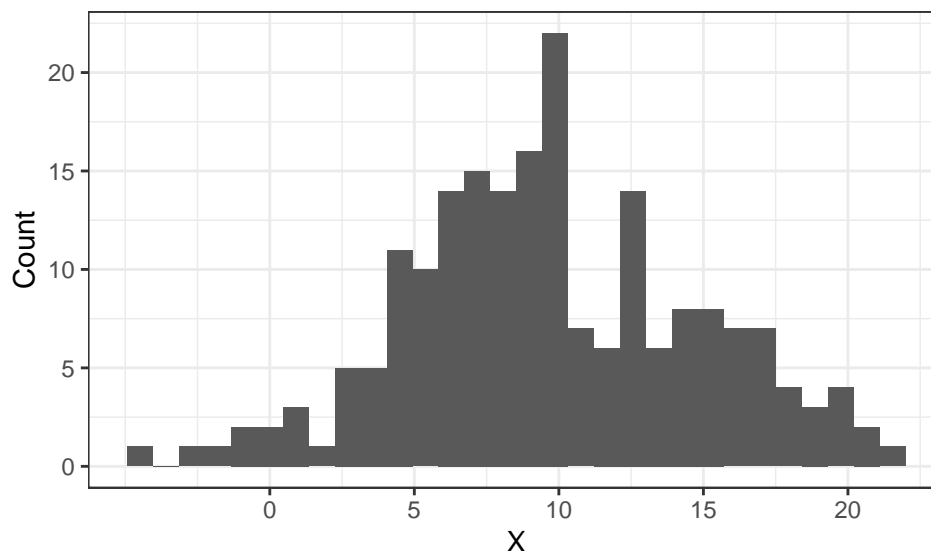
Price et al. (2016) untersuchte die Auswirkungen des Bergbaus und der Talauffüllung auf den Bestand und die Häufigkeit von Bachsalamandern. Um den Effekt zu Berechnen nutze Price et al. (2016) eine Poisson-Regression auf die Anzahl an aufgefundenen Bachsalamandern an den jeweiligen Suchorten. Welche 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 3.13 liegt Overdispersion vor. Die Lösung ist die Nutzung einer anderen Verteilungsfamilie wie die Quasipoisson Verteilung.
- B** ☐ Die Poisson-Regression schätzt drei Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzten Overdispersion von 3.13 liegt Overdispersion vor. Die Lösung ist die Nutzung von nur zwei der drei Verteilungsparameter:  $\gamma_1$  und  $\gamma_3$ .
- C** ☐ Die Poisson-Regression schätzt nur einen Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzten Overdispersion von 3.13 liegt Overdispersion vor. Damit kann keine Poisson-Regression gerechnet werden. Die Lösung ist eine Gaussian Regression mit Nullanpassung.
- D** ☐ Die Poisson-Regression schätzt zwei Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzten Overdispersion von 3.13 liegt keine Overdispersion vor.
- E** ☐ Die Poisson-Regression schätzt nur einen Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzten Overdispersion von 3.13 liegt keine Overdispersion vor. Overdispersion liegt vor, wenn die geschätzte Overdispersion unter 1 liegt.

## 17 Aufgabe

(2 Punkte)

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



- A** ☐ Eine rechtsschiefe, multivariate Normalverteilung.
- B** ☐ Es handelt sich um eine Poisson-Verteilung mit  $\text{Pois}(10)$ .

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

## 18 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 3.35 ist richtig?

- A** ☐ 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.
- 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 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
- D** ☐ 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.
- E** ☐ 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.

## 19 Aufgabe

(2 Punkte)

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

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

## 20 Aufgabe

(2 Punkte)

Sie haben das abstrakte Modell  $y \sim x_1 + x_2$  vorliegen. Welche Aussage über X ist richtig?

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

## 21 Aufgabe

(2 Punkte)

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

- A ☐ Ein kausales Modell basiert auf einem Trainingsdatensatz und einem Testdatensatz. Auf dem Trainingsdatensatz wird das Modell trainiert und auf dem Testdatensatz validiert.
- B ☐ Ein kausales Modell benötigt mindestens eine Fallzahl von über 100 Beobachtungen und darf keine fehlenden Werte beinhalten. Die Varianzkomponenten müssen homogen sein.
- C ☐ Ein kausales 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  $\beta$ .
- D ☐ Ein kausales Modell schliesst grundsätzlich lineare Modell aus. Es muss ein Graph gefunden werden, der alle Punkte beinhaltet. Erst dann kann das  $R^2$  berechnet werden.
- E ☐ Ein kausales Modell wird auf einem Trainingsdatensatz trainiert und anschliessend über eine explorative Datenanalyse validiert. Signifikanzen über  $\beta_i$  können hier nicht festgestellt werden.

## 22 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, erlauben es nur eine ANOVA zu rechnen.
- B ☐ Die Effektschätzer aus einem Modell, in diesem Fall ein polynomies Modell, erlauben es sowohl eine ANOVA zurechnen sowie auch eine Zusammenfassung der Mittelwerte zu betrachten.

- 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 einen Mittelwertsvergleich 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.

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

## 24 Aufgabe

(2 Punkte)

Sie führen ein Experiment zur Behandlung von Klaueninfektionen bei Kühen durch. Bei 5 Tieren finden Sie eine Erkrankung der Klauen vor und 8 Tiere sind gesund. Welche Aussage über den Odds ratio Effektschätzer ist richtig?

- A** ☐ Es ergibt sich ein Odds ratio von 0.38, da es sich um eine Chancenverhältnis handelt.
- B** ☐ Es ergibt sich ein Odds ratio von 1.6, da es sich um ein Anteil handelt.
- C** ☐ Es ergibt sich ein Odds ratio von 0.62, da es sich um eine Chancenverhältnis handelt.
- D** ☐ Es ergibt sich ein Odds ratio von 0.62, da es sich um ein Anteil handelt.
- E** ☐ Es ergibt sich ein Odds ratio von 0.38, da es sich um ein Anteil handelt.

## 25 Aufgabe

(2 Punkte)

Welche Aussage über die nicht-parametrische Statistik ist richtig?

- A** ☐ Die nicht-parametrische Statistik basiert auf dem Schätzen von Parametern aus einer a priori festgelegten Verteilung. Daher gibt es auch direkt zu interpretierenden Effektschätzer.
- B** ☐ Die nicht-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.
- C** ☐ Die nicht-parametrische Statistik basiert auf Rängen. 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 nicht-parametrische Statistik basiert auf dem Schätzen von Parametern aus einer festgelegten Verteilung. Daher gibt es auch direkt zu interpretierenden Effektschätzer.

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

## 27 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 ☐ Alle Funktionen und auch Anwendungen sind in  in englischer Sprache. Die Nutzung von deutschen Wörtern ist nicht schick und das ist zu vermeiden.
- B ☐ 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.
- C ☐ Die Spracherkennung von  ist nicht in der Lage Deutsch zu verstehen.
- D ☐ 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.
- E ☐ Es gibt keinen Grund nicht auch deutsche Wörter zu verwenden. Es ist ein Stilmittel.

## 28 Aufgabe

(2 Punkte)


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

- A ☐ Wir lesen als erstes die Daten über `read_excel()` ein, transformieren die Spalten über `mutate()` in die richtige Form und können dann über `ggplot()` uns die Abbildungen erstellen lassen. Wichtig ist, dass wir keine Faktoren sondern nur numerische Variablen vorliegen haben.
- B ☐ Wir lesen die Daten ein und mutieren die Daten. Dabei ist wichtig, dass wir nicht das Paket `tidyverse` nutzen, da dieses Paket veraltet ist. Über die Funktion `library(tidyverse)` entfernen wir das Paket von der Analyse.
- C ☐ Wir transformieren die Spalten über `mutate()` in ein `tibble` und können dann über `ggplot()` uns die Abbildungen erstellen lassen. Dabei beachten wir das wir keine Faktoren in den Daten haben.

- D** ☐ Wir lesen die Daten über eine generische Funktion `read()` ein und müssen dann die Funktion `ggplot()` nur noch installieren. Dann haben wir die Abbildungen als `*.png` vorliegen.
- E** ☐ Wir lesen als erstes die Daten über `read_excel()` ein, transformieren die Spalten über `mutate()` in die richtige Form und können dann über `ggplot()` uns die Abbildungen erstellen lassen.

## 29 Aufgabe

(2 Punkte)

In einem Stallexperiment mit  $n = 95$  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 *logistischen* linearen Regression. Welche Aussage über den **Effekt** ist richtig?

term	estimate	std.error
(Intercept)	0.04	0.28
light_binhigh	-3.80	1.05

- A** ☐ In einer logistischen Regression berechnet man das RR. Daher muss der Schätzer des Effektes  $\beta_1$  noch quadriert werden. Somit liegt das RR bei 14.44
- B** ☐ In einer logistischen Regression wird die Mittelwertsdifferenz betrachtet. Daher ist der Effekt zwischen den beiden Lichtregimen eine Gewichtsänderung von -3.8
- C** ☐ Eine logistischen Regression basiert auf dem maximum Likelihood Prinzip. Hierbei kann kein Effekt beschrieben werden. Im Zweifel hilft aber eine Quadrierung der Fehlerquadratrate  $\epsilon$ .
- D** ☐ In einer logistischen Regression muss für die Interpretation des Effektes das  $\beta_1$  quadriert werden. Somit liegt das OR bei 14.44
- E** ☐ In einer logistischen Regression kann kein Effekt roh interpretiert werden. Es muss erst eine Confounderadjustierung durchgeführt werden.

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

## 31 Aufgabe

(2 Punkte)

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

- A** ☐ Das GLM erlaubt auch nicht normalverteilte Residuen in der Schätzung der Regressionsgrade.
- B** ☐ Das GLM ist eine allgemeine Erweiterung der linearen Regression auf die Normalverteilung.

- C** ☐ Das GLM ist eine Vereinfachung des LM in R. Mit dem GLM lassen polygonale Regressionen rechnen.
- D** ☐ Das GLM ist ein faktisch maschineller Lernalgorithmus, der selbstständig die Verteilungsfamilie für Y wählt.
- E** ☐ Das GLM erlaubt auch weitere Verteilungsfamilien für das Y bzw. das Outcome in einer linearen Regression zu wählen.

## 32 Aufgabe

(2 Punkte)

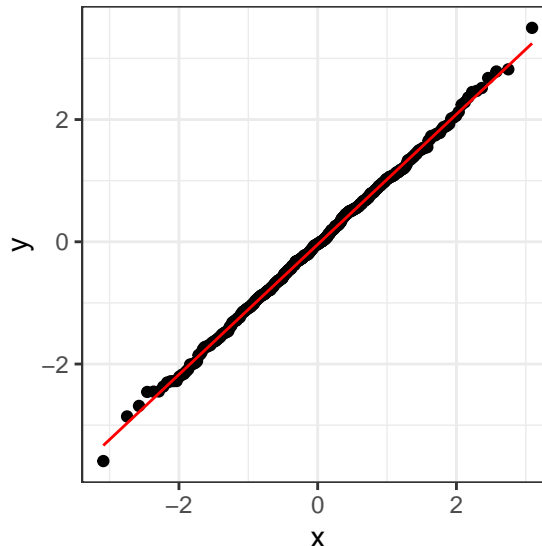
Sie rechnen in einer linearen Regression das Modell A und das Modell B. Nun stellt sich die Frage, welches der beiden Modelle das bessere Modell ist. Um die Modelle bewerten zu können berechnen Sie dafür das  $AIC_A$  für Modell A mit 653 und für das Modell B das  $AIC_B$  von 287. Welche Aussage über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

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

## 33 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 erfüllt. Die Punkte liegen zum überwiegenden Teil nicht auf der Geraden.
- B** ☐ Die Annahme der normalverteilten Residuen ist nicht erfüllt. Die Punkte liegen zum überwiegenden Teil auf der Geraden.
- C** ☐ Die Annahme der normalverteilten Residuen ist nicht erfüllt. Die Punkte liegen zum überwiegenden Teil nicht auf der Geraden.

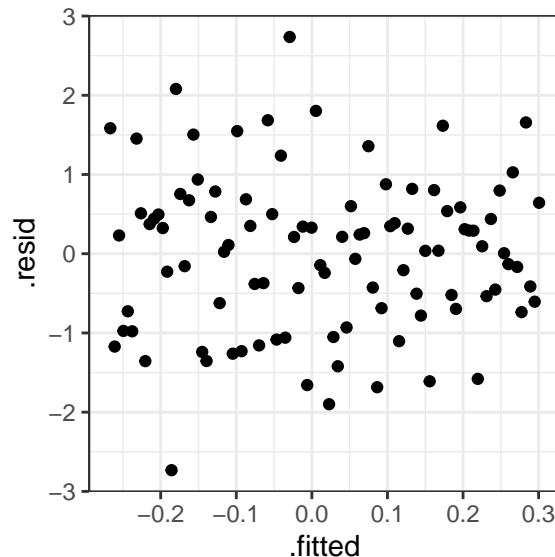


- D** ☐ Die Annahme der normalverteilten Residuen ist erfüllt. Die Punkte liegen zum überwiegenden Teil auf der Geraden.
- E** ☐ Die Annahme der normalverteilten Residuen ist erfüllt. Die Punkte liegen zum überwiegenden Teil nicht auf der Geraden und Korrelation ist negativ.

### 34 Aufgabe

(2 Punkte)

Sie rechnen eine lineare 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 nicht erfüllt. Vereinzelt Punkte liegen oberhalb bzw. unterhalb der Geraden um die 0 Linie weiter entfernt. Ein klares Muster ist zu erkennen.
- D** ☐ Die Annahme der normalverteilten Residuen ist erfüllt. Kein Muster ist zu erkennen und keine Outlier zu beobachten.
- E** ☐ Die Annahme der normalverteilten Residuen ist nicht erfüllt. Es ist kein Muster zu erkennen.

### 35 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  $s = 3$  und  $\delta_0 = 1$ . Es ergibt sich somit folgende Fallzahl.

- A** ☐ Es ergibt sich eine Fallzahl von 310
- B** ☐ Es ergibt sich eine Fallzahl von 308
- C** ☐ Es ergibt sich eine Fallzahl von 155
- D** ☐ Es ergibt sich eine Fallzahl von 154
- E** ☐ Es ergibt sich eine Fallzahl von 78

### 36 Aufgabe

(2 Punkte)

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

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

### 37 Aufgabe

(2 Punkte)

Das Falsifikationsprinzip besagt...

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

### 38 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 ☐ Auf einer Statistikkonferenz in Genf im Jahre 1942 wurde dieser Cut-Off nach langen Diskussionen festgelegt. Bis heute ist der Cut Off aber umstritten, da wegen dem 2. Weltkrieg viele Wissenschaftler nicht teilnehmen konnten.
- C ☐ Der Begründer der modernen Statistik, R. Fischer, hat die Grenze simuliert und berechnet. Dadurch ergibt sich dieser optimale Cut-Off.
- D ☐ Im Rahmen eines langen Disputes zwischen Neyman und Fischer wurde  $\alpha = 5\%$  festgelegt. Leider werden die Randbedingungen und Voraussetzungen an statistische Modelle heute immer wieder ignoriert.
- E ☐ 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.

### 39 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  $1 - \beta$  wird auf 80% gesetzt. Damit liegt die Wahrscheinlichkeit für die  $H_0$  bei 20%.
- D** ☐ Es gilt  $\alpha + \beta = 1$  und somit liegt  $\beta$  meist bei 95%.
- E** ☐ Die Power beschreibt die Wahrscheinlichkeit die  $H_A$  abzulehnen. Wir testen die Power jedoch nicht.

### 40 Aufgabe

(2 Punkte)

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

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

### 41 Aufgabe

(2 Punkte)

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

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

## 42 Aufgabe

(2 Punkte)

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

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

## 43 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 zu grosse Datensätze für die gängige Statistik gemessen. Daher wendet man maschinelle Lernverfahren für kausale Modelle an. Hier ist die Relevanz gleich Signifikanz.
- B ☐ Big Data ist ein Problem der parametrischen Statistik. Parameter lassen sich nur auf kleinen Datensätzen berechnen, da es sich sonst nicht mehr um eine Stichprobe im engen Sinne der Statistik handelt.
- C ☐ Aktuell werden immer grössere Datensätze erhoben. Dadurch wird auch die Varianz immer höher was automatisch zu mehr signifikanten Ergebnissen führt.
- D ☐ Aktuell werden immer grössere Datensätze erhoben. Eine erhöhte Fallzahl führt automatisch auch zu mehr signifikanten Ergebnissen, selbst wenn die eigentlichen Effekte nicht relevant sind.
- E ☐ 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.

## 44 Aufgabe

(2 Punkte)

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

- A ☐ Das Konfidenzintervall. Durch die Visualisierung des Konfidenzintervalls kann eine Relevanzschwelle vom Anwender definiert werden. Zusätzlich erlaubt das Konfidenzintervall auch eine Entscheidung über die Signifikanz.
- B ☐ Das OR. Als Chancenverhältnis gibt es das Verhältnis von Relevanz und Signifikanz wieder.
- C ☐ Der p-Wert. Durch den Vergleich mit  $\alpha$  lässt sich über die Signifikanz entscheiden und der  $\beta$ -Fehler erlaubt über die Power eine Einschätzung der Relevanz.
- D ☐ 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 ☐ 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.

## 45 Aufgabe

(2 Punkte)

Welche Aussage über die frequentistischen Testtheorie ist richtig?

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

## 46 Aufgabe

(2 Punkte)

Welche Aussage über den  $p$ -Wert und dem Signifikanzniveau  $\alpha$  gleich 5% ist richtig?

- A ☐ Wir vergleichen mit dem  $p$ -Wert und dem Signifikanzniveau  $\alpha$  absolute Werte auf einem Zahlenstrahl und damit den Unterschied der Teststatistiken, wenn die  $H_0$  gilt.
- B ☐ Wir vergleichen mit dem  $p$ -Wert und dem Signifikanzniveau  $\alpha$  Wahrscheinlichkeiten und damit die Flächen unter der Kurve der Teststatistik, wenn die  $H_0$  gilt.
- C ☐ Wir machen eine Aussage über die individuelle Wahrscheinlichkeit des Eintretens der Nullhypothese  $H_0$ .
- D ☐ Wir vergleichen die Effekte des  $p$ -Wertes mit den Effekten der Signifikanzschwelle unter der Annahme der Nullhypothese.
- E ☐ Wir vergleichen mit dem  $p$ -Wert und dem Signifikanzniveau  $\alpha$  Wahrscheinlichkeiten und damit die absoluten Werte auf einem Zahlenstrahl, wenn die  $H_0$  gilt.

## 47 Aufgabe

(2 Punkte)

Welche Aussage über den t-Test ist richtig?

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

## 48 Aufgabe

(2 Punkte)

Welche Aussage über den Welch t-Test ist richtig?

- A ☐ Der Welch t-Test ist ein Post-hoc Test der ANOVA und basiert daher auf dem Vergleich der Varianz.
- B ☐ Der Welch t-Test ist die veraltete Form des Student t-Test und wird somit nicht mehr verwendet.

- C** ☐ Der Welch t-Test vergleicht die Mittelwerte von zwei Gruppen unter der strikten Annahme von Varianzhomogenität.
- D** ☐ Der Welch t-Test vergleicht die Varianz von zwei Gruppen.
- E** ☐ Der Welch t-Test wird angewendet, wenn Varianzheterogenität zwischen den beiden zu vergleichenden Gruppen vorliegt.

## 49 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** ☐ Der Fehler liegt in den t-Tests. Wenn eine ANOVA signifikant ist, dann muss zwangsweise auch ein t-Test signifikant sein.
- B** ☐ Die ANOVA testet auf der gesamten Fallzahl. Es wäre besser die ANOVA auf der gleichen Fallzahl wie die einzelnen t-Tests zu rechnen.
- C** ☐ Es gibt einen Fehler in der Varianzstruktur. Daher kann die ANOVA nicht richtig sein und paarweise t-Tests liefern das richtige Ergebnis.
- 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.

## 50 Aufgabe

(2 Punkte)

Welche Aussage über den gepaarten t-Test für verbundene Stichproben ist richtig?

- A** ☐ Beim gepaarten t-Test kombinieren wir die Vorteile des Student t-Test für Varianzhomogenität mit den Vorteilen des Welch t-Test für Varianzheterogenität. Wir bilden dafür die Differenz der Einzelbeobachtungen.
- B** ☐ Der gepaarte t-Test nutzt die Varianz der Beobachtungen jeweils paarweise und bildet dafür eine verbundene Stichprobe. Dieser Datensatz  $d$  dient dann zur Differenzbildung.
- C** ☐ Der gepaarte t-Test wird genutzt, wenn die Differenzen der Beobachtungen verbunden sind und wir dadurch die Unabhängigkeit nicht mehr vorliegen haben.
- D** ☐ Der gepaarte t-Test wird gerechnet, wenn die Beobachtungen nicht unabhängig voneinander sind. Wir messen wiederholt an dem gleichen Probanden oder Tier oder Pflanze. Wir bilden die Differenzen um den gepaarten t-Test rechnen zu können.
- E** ☐ Der gepaarte t-Test wird gerechnet, wenn die Beobachtungen abhängig voneinander sind. Wir messen jede Beobachtung nur einmal und berechnen dann die Differenz zu dem Mittel der anderen Beobachtungen.

## Deskriptive Statistik & Explorative Datenanalyse

Mehr Informationen zu den Aufgaben in den folgenden Kapiteln aus dem Skript Bio Data Science.

- Kapitel 15 - Deskriptive Statistik
- Kapitel 16 - Visualisierung von Daten
- Kapitel 18 - Verteilung von Daten

### 51 Aufgabe

**(10 Punkte)**

Sie haben folgende Zahlenreihe  $y$  vorliegen  $y = \{17, 15, 12, 14, 18, 24, 16, 18, 17, 17, 17\}$ . Berechnen Sie folgende deskriptive Maßzahlen. Geben Sie Formeln und Rechenwege mit an!

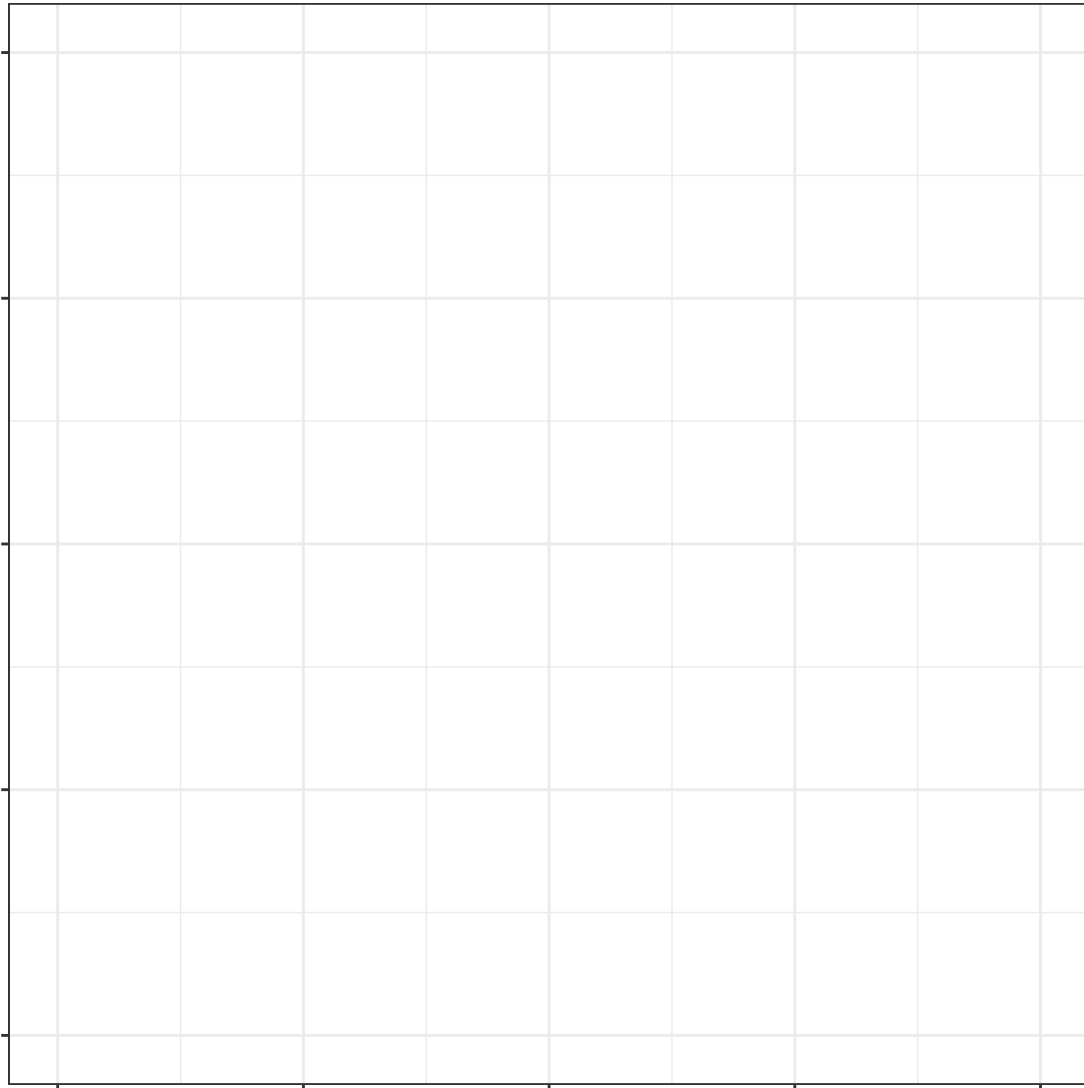
1. Den Interquartileabstand **(2 Punkte)**
2. Das 3rd Quartile **(2 Punkte)**
3. Den Mittelwert **(2 Punkte)**
4. Das 1st Quartile **(2 Punkte)**
5. Die Varianz **(2 Punkte)**

## 52 Aufgabe

(8 Punkte)

Sie haben folgende Zahlenreihe  $y$  vorliegen  $y = \{21, 17, 16, 19, 16\}$ .

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)**





## 53 Aufgabe

(7 Punkte)

Nach einem Gewächshausexperiment mit drei Bewässerungstypen (*low*, *mid* und *high*) ergibt sich die folgende Datentabelle mit dem gemessenen Frischgewicht (*freshmatter*).

water_type	freshmatter
mid	30
mid	43
mid	28
high	12
low	15
low	13
high	16
low	10
high	16
high	14
low	19

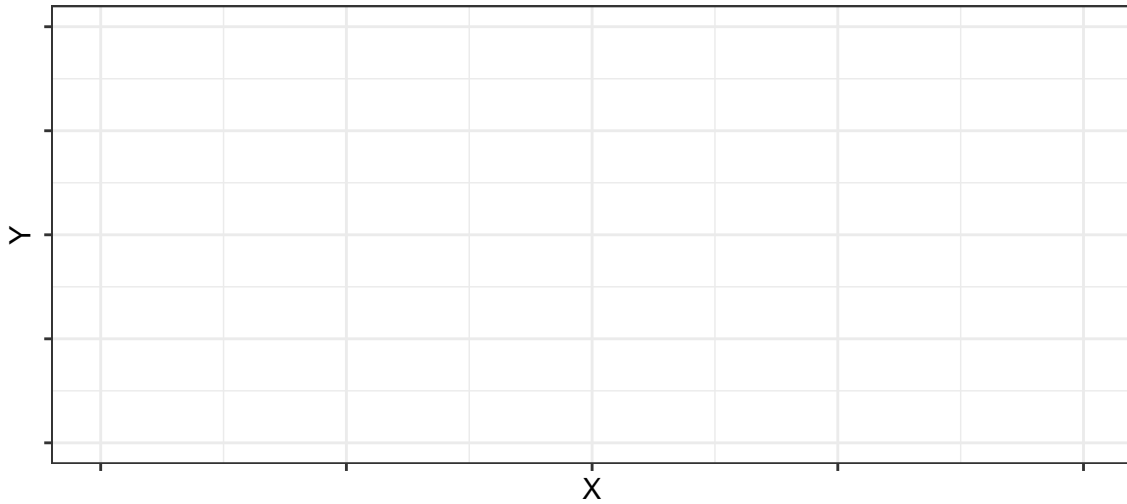
1. Zeichnen Sie in *einer* Abbildung die Barplots für die Bewässerungstypen! Beschriften Sie die Achsen entsprechend! **(4 Punkte)**
2. Beschriften Sie *einen* Barplot mit den gängigen statistischen Maßzahlen! **(2 Punkte)**
3. Wenn Sie *keinen Effekt* zwischen der Bewässerungstypen erwarten würden, wie sehen dann die Barplots aus? **(1 Punkt)**

## 54 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)**

$N(2, 2)$  und  $N(4, 1)$



$\text{Pois}(15)$



## 55 Aufgabe

**(6 Punkte)**

1. Skizzieren Sie 2 Normalverteilungen *in einer Abbildung* mit  $\mu_1 \neq \mu_2$  und  $s_1 \neq s_2$ ! **(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)**

## 56 Aufgabe

**(6 Punkte)**

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

3, 4, 5, 3, 1, 5, 3, 4, 8, 2, 4, 5, 2, 3, 5

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

## 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	14.8
B	15.6
A	16.9
B	17.9
A	16.9
B	14.5
A	20.7
A	20.5
A	20.1
B	11.6
B	11.6
A	16.6
A	21.4
B	14.1

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	19
low	23
low	24
low	18
low	25
low	23
low	17
low	21
low	25
low	19
low	16

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

(6 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	4
zeroX	1
roundUp	7
roundUp	4
roundUp	6
ctrl	8
zeroX	3
pestKill	4
roundUp	6
pestKill	2
ctrl	7
zeroX	1
ctrl	5

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



## 61 Aufgabe

(6 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	no
OutEx	no
RoundUp	yes
OutEx	yes
RoundUp	no
OutEx	yes
RoundUp	yes
RoundUp	no
OutEx	yes
OutEx	no
OutEx	yes
OutEx	yes
RoundUp	no
RoundUp	yes
RoundUp	no
RoundUp	yes
OutEx	no
RoundUp	no
RoundUp	yes
OutEx	yes
RoundUp	yes

1. Stellen Sie in einer 2x2 Tafel den Zusammenhang zwischen dem Pestizid und dem Infektionsstatus dar! **(2 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 sähe dann der Mosaic-Plot aus? **(2 Punkte)**

## 62 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
16	36
17	24
13	25
14	26
15	24
16	29
15	36
12	21
13	22

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)**

## Statistisches Testen

Mehr Informationen zu den Aufgaben in den folgenden Kapiteln aus dem Skript Bio Data Science.

- Kapitel 3 - Falsifikationsprinzip
- Kapitel 19 - Die Testentscheidung
- Kapitel 20 - Die Testtheorie

### 63 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)**

## 64 Aufgabe

(6 Punkte)

Geben ist folgende 2x2 Kreuztabelle.


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

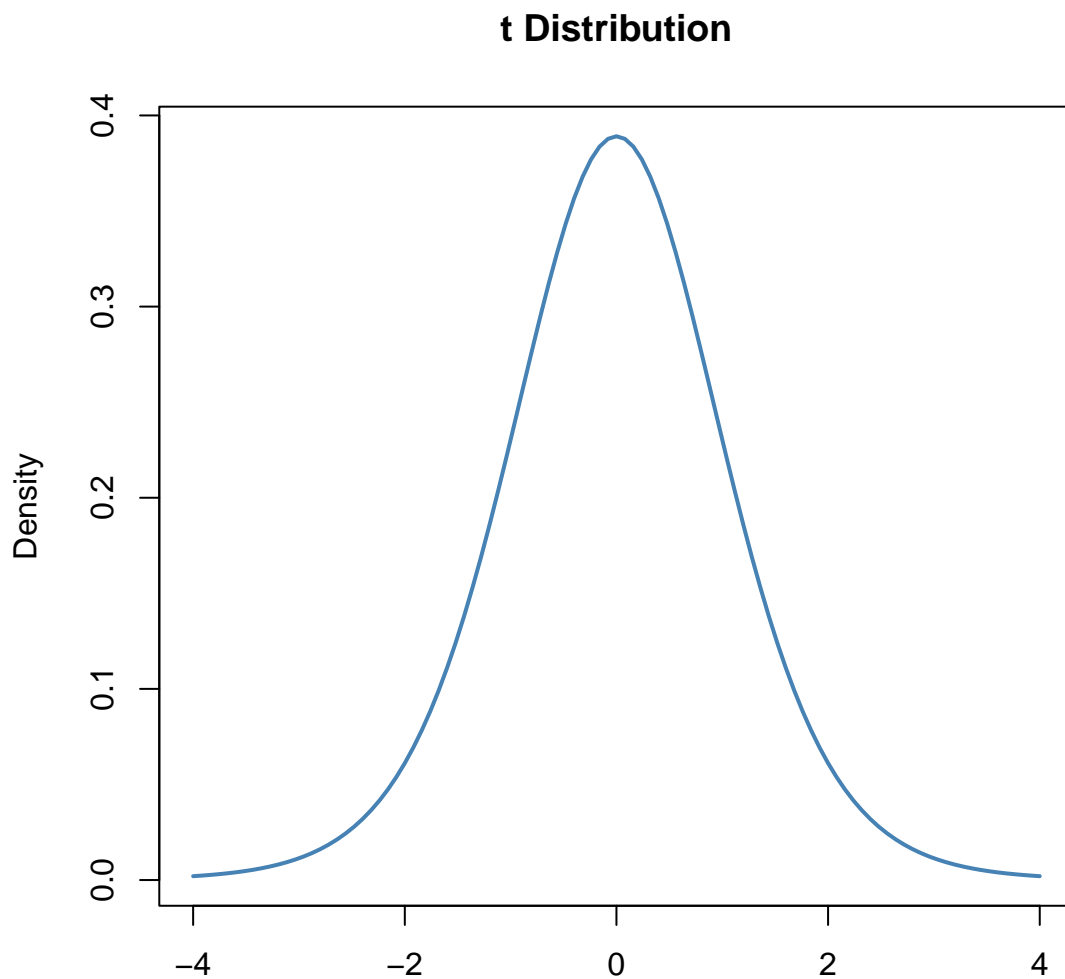
- (Unbekannte) Wahrheit
- $H_0$  wahr
- $H_0$  falsch
- $H_0$  abgelehnt
- $H_0$  beibehalten
- Testentscheidung
- $\alpha$ -Fehler
- $\beta$ -Fehler
- Richtige Entscheidung
- 5%
- 20%

## 65 Aufgabe

(4 Punkte)

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

1. Zeichnen Sie das Signifikanzniveau  $\alpha$  in die Abbildung! **(1 Punkte)**
2. Zeichnen Sie einen nicht signifikant p-Wert in die Abbildung! **(1 Punkt)**
3. Ergänzen Sie „ $\bar{y}_1 = \bar{y}_2$ “! **(1 Punkt)**
4. Ergänzen Sie „ $A = 0.95$ “! **(1 Punkt)**



## 66 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



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

## 68 Aufgabe

(8 Punkte)

Sie haben folgende Aussage gegeben.

### **Bin ich im Sommer?**

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)**



## 69 Aufgabe

**(8 Punkte)**

»Ohne ausreichende Zufuhr von Vitamin C stellen sich nach 68 Tagen die ersten Symptome ein; die ersten Toten sind nach 98 Tagen zu beklagen; nach 123 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)**

## 70 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)**

## Der t-Test

Mehr Informationen zu den Aufgaben in den folgenden Kapiteln aus dem Skript Bio Data Science.

- Kapitel 22 - Der t-Test

### 71 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
GoneEx	19
RoundUp	15
GoneEx	16
RoundUp	17
RoundUp	19
RoundUp	23
GoneEx	17
RoundUp	14
GoneEx	16
RoundUp	15
GoneEx	17

1. Formulieren Sie die wissenschaftliche 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)**

## 72 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	17	11
2	16	8
3	13	12
4	15	12
5	15	10
6	12	8
7	14	11

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)**

## 73 Aufgabe

(10 Punkte)


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

```
##  
## Two Sample t-test  
##  
## data:  freshmatter by N  
## t = -2.36993, df = 14, p-value = 0.032696  
## alternative hypothesis: true difference in means between group high and group low is not equal to 0  
## 95 percent confidence interval:  
## -8.61786237 -0.42975667  
## sample estimates:  
## mean in group ctrl mean in group trt1  
## 16.333333 20.857143
```

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 eine Abbildung in der Sie  $T_{calc}$ ,  $Pr(D|H_0)$ ,  $A = 0.95$ , sowie  $T_{\alpha=5\%} = |2.14|$  einzeichnen! **(4 Punkte)**
4. Beschriften Sie die Abbildung entsprechend! **(2 Punkte)**

## 74 Aufgabe

(8 Punkte)


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

```
##  
## Two Sample t-test  
##  
## data: waterintake by infusion  
## t = -0.0687208, df = 16, p-value = 0.94606  
## alternative hypothesis: true difference in means between group high and group low is not equal to 0  
## 95 percent confidence interval:  
## -3.5386758 3.3164536  
## sample estimates:  
## mean in group high mean in group trt1  
## 14.333333 14.444444
```

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 95% Konfidenzintervall! **(2 Punkte)**
4. Beschriften Sie die Abbildung und das 95% Konfidenzintervall entsprechend! **(2 Punkte)**

## 75 Aufgabe

(8 Punkte)

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

```
##
##  Welch t-Test
##
## data:  weight by group
## t = -1.02132, df = 8, p-value = 0.33699
## alternative hypothesis: true difference in means between group high and group low is not equal to 0
## 95 percent confidence interval:
##  -7.1672849  2.7672849
## sample estimates:
## mean in group ctrl mean in group trt1
##                17.0                19.2
```

1. Formulieren Sie die wissenschaftliche Fragestellung! **(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)**

## 76 Aufgabe

(8 Punkte)

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

```
##  
## Paired t-test  
##  
## data: drymatter by Fe  
## t = 1.1227, df = 8, p-value = 0.29413  
## alternative hypothesis: true mean difference is not equal to 0  
## 95 percent confidence interval:  
## -1.6395320 4.7506432  
## sample estimates:  
## mean difference  
## 1.5555556
```

1. Formulieren Sie die wissenschaftliche Fragestellung! **(2 Punkte)**
2. Liegt ein signifikanter Unterschied zwischen den Gruppen vor? Begründen Sie Ihre Antwort! **(2 Punkte)**
3. Skizzieren Sie den sich ergebenden Datensatz mit  $n = 4$ ! Beobachtungen! **(2 Punkte)**
4. Skizzieren Sie die sich ergebenden Boxplots! Welche Annahmen an die Daten haben Sie getroffen? Begründen Sie Ihre Antwort! **(2 Punkte)**



## Die ANOVA

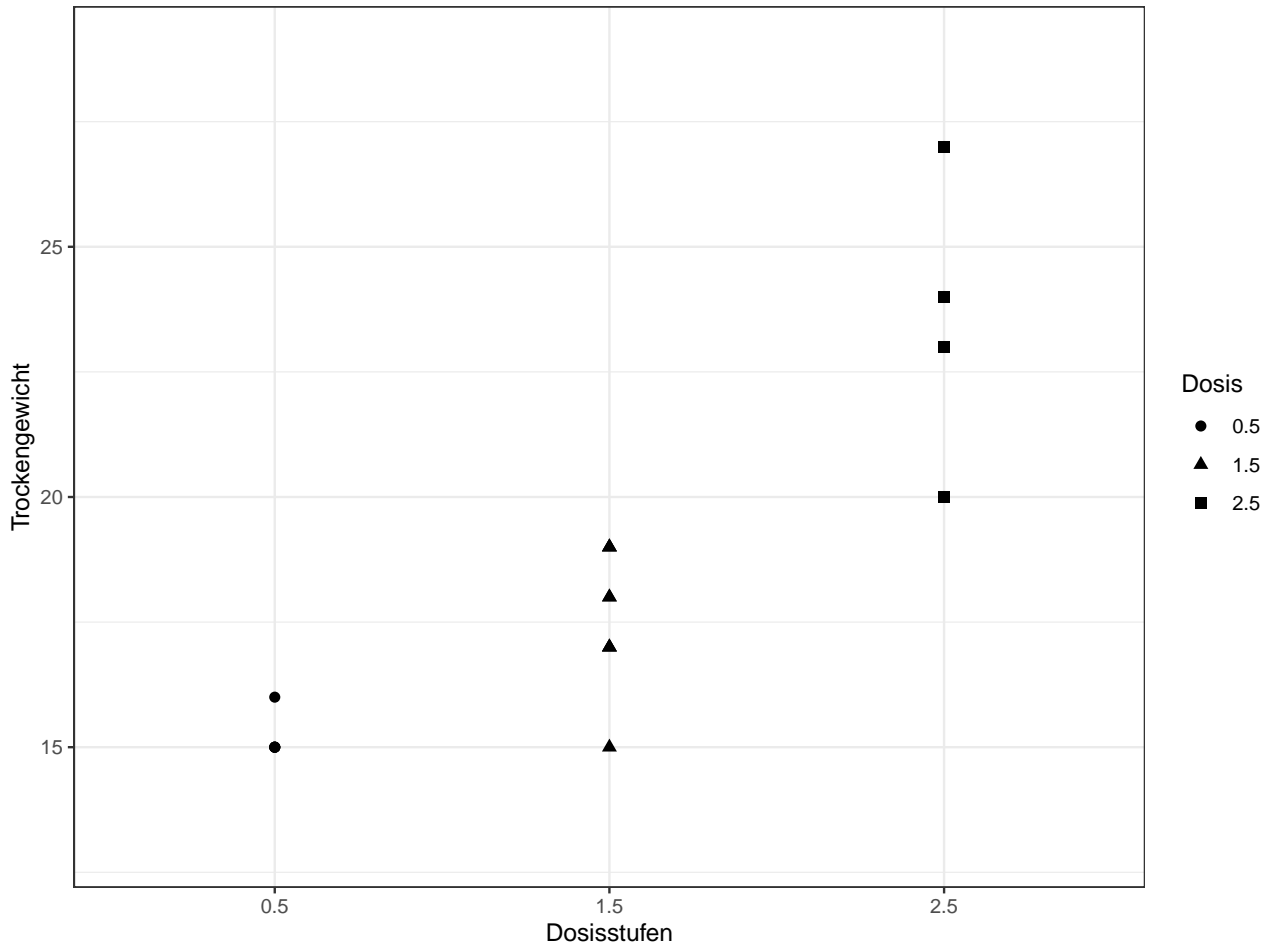
Mehr Informationen zu den Aufgaben in den folgenden Kapiteln aus dem Skript Bio Data Science.

- Kapitel 23 - Die ANOVA

### 77 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 stehend 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)

## 78 Aufgabe

(13 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)
<b>group</b>	2	3.69			
<b>Residuals</b>	27	10.09			

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 für den *vermutlich* signifikantesten Gruppenvergleich anhand der untenstehenden Tabelle mit  $T_{\alpha=5\%} = 2.03$ . Begründen Sie Ihre Auswahl! **(3 Punkte)**

group	n	mean	sd
ctrl	10	5.03	0.60
trt1	10	4.66	0.73
trt2	10	5.52	0.47

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

## 79 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  3.28070  1.640352  5.11761 0.013048 *
## Residuals 27  8.65433  0.320531
## ---
## 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)**

## 80 Aufgabe

(12 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! **(4 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	211.66			
<b>dose</b>	2	2432.32			
<b>supp:dose</b>	2	105.8			
<b>Residuals</b>	54	713.59			

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*! **(2 Punkte)**
4. Was sagt der Term *supp:dose* aus? Interpretieren Sie das Ergebnis des abgeschätzten p-Wertes! **(2 Punkte)**

## 81 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 66 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). Sie erhalten folgenden Output in .

```
## Analysis of Variance Table
##
## Response: len
##      Df    Sum Sq  Mean Sq    F value    Pr(>F)
## supp     1   333.955   333.955   33.96055 0.00000023801 ***
## dose      2  3005.935  1502.968  152.83997 < 0.000000000000000222 ***
## supp:dose  2    20.293    10.147    1.03183    0.36259
## Residuals 60   590.016     9.834
## ---
## 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! **(2 Punkte)**
3. Zeichnen Sie eine Abbildung, der dem obigen Ergebnis der zweifaktoriellen ANOVA näherungsweise entspricht! **(4 Punkte)**

## 82 Aufgabe

(8 Punkte)

In der untenstehenden Tabelle ist die Formel für den F-Test aus der ANOVA und die Formel für den t-Test dargestellt. In der ANOVA berechnen Sie die F-Statistik  $F_{calc}$  und in dem t-Test die T-Statistik  $T_{calc}$ .

$$F_{calc} = \frac{MS_{treatment}}{MS_{error}} \quad T_{calc} = \frac{\bar{y}_1 - \bar{y}_2}{s_p \cdot \sqrt{2/n_g}}$$

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 anhand von  $MS_{treatment}$  und  $MS_{error}$ ! 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)**

## 83 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 starke und eine schwache Interaktion zwischen den Faktoren  $f_1$  und  $f_2$ ! **(2 Punkte)**
2. Erklären Sie den Unterschied zwischen den beiden Stärken der Interaktion! **(2 Punkte)**
3. Wenn eine signifikante Interaktion in den Daten vorliegt, wie ist dann das weitere Vorgehen bei einem Posthoc-Test? **(2 Punkte)**

## 84 Aufgabe

(7 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 für eine ANOVA 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! **(3 Punkte)**



## Der $\chi^2$ -Test & Der diagnostische Test

Mehr Informationen zu den Aufgaben in den folgenden Kapiteln aus dem Skript Bio Data Science.

- Kapitel 28 - Der  $\chi^2$ -Test
- Kapitel 29 - Der diagnostische Test

### 85 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  $n = 114$  Pflanzen untersucht.

	Erkrankt (ja)	Erkrankt (nein)	
Pestizid (ja)	38	11	
Pestizid (nein)	27	38	

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)**

## 86 Aufgabe

(7 Punkte)

Gegeben sind folgende Randsummen in einer 2x2 Kreuztabelle aus einem Experiment mit  $n = 178$  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)			87
Medikament (nein)			91
	107	71	178

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)**

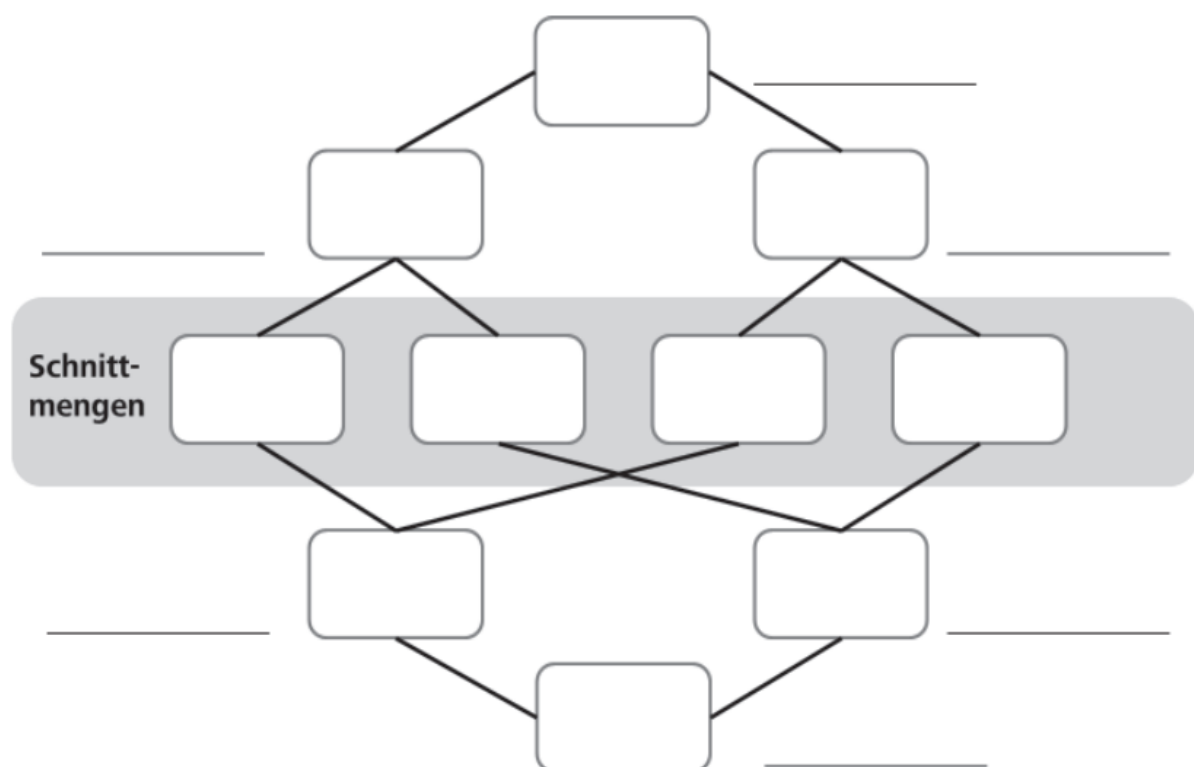
3. Was ist die Mindestanzahl an Beobachtungen je Zelle? Wenn in einer der Zellen weniger Beobachtungen auftreten, welchen Test können Sie anstatt des „normalen“ Chi-Quadrat-Tests anwenden? **(2 Punkte)**
4. Warum hat die obige Vierfeldertafel einen Freiheitsgrad von  $df = 1$ ? **(1 Punkt)**

## 87 Aufgabe

(11 Punkte)

Die Prävalenz von Klauenseuche bei Wollschweinen wird mit 4% angenommen. In 75% der Fälle ist ein Test positiv, wenn das Wollschwein erkrankt ist. In 7.5% der Fälle ist ein Test positiv, wenn das Wollschwein *nicht* erkrankt ist und somit gesund ist. Sie werten 4000 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)**



## 88 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.

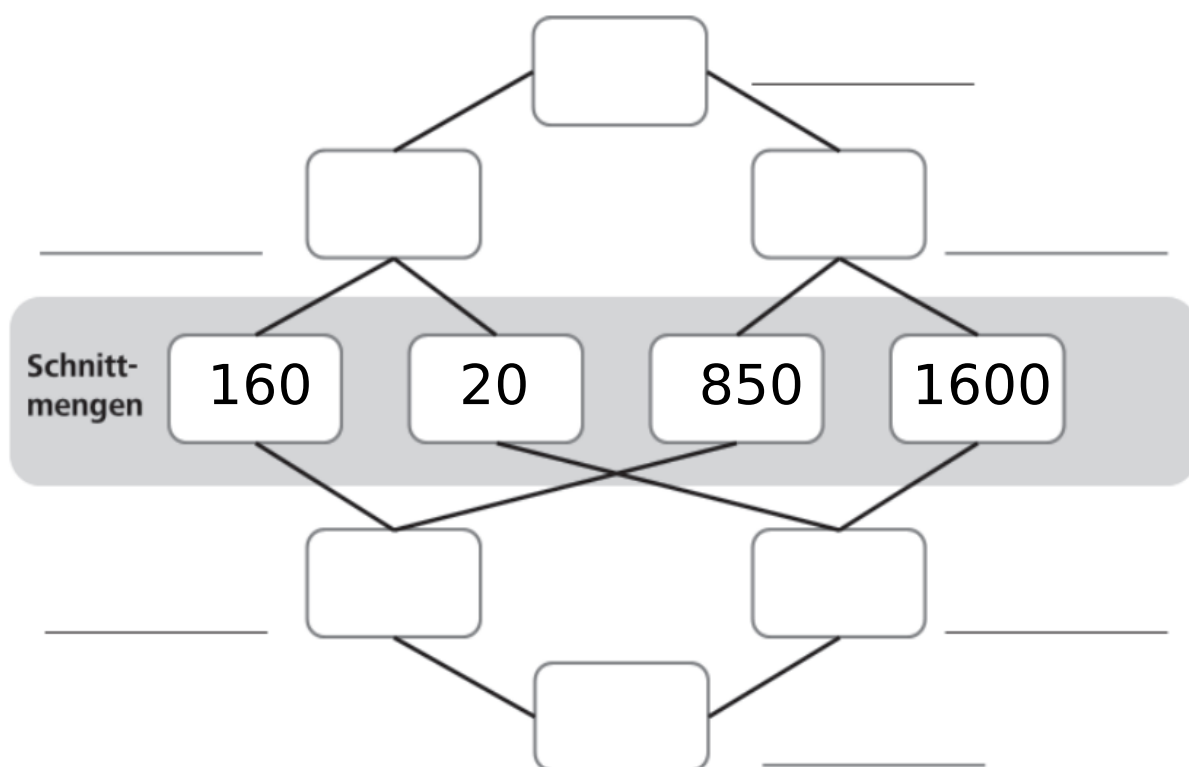
1. Visualisieren Sie *TP*, *TN*, *FP* und *FN* in einer Abbildung. Beschriften Sie die Abbildung und die Achsen entsprechend! **(6 Punkte)**
2. Tragen Sie *TP*, *TN*, *FP* und *FN* in eine 2x2 Kreuztabelle ein. Beschriften Sie die Tabelle entsprechend! **(3 Punkte)**

## 89 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)**



## Simple linear Regression

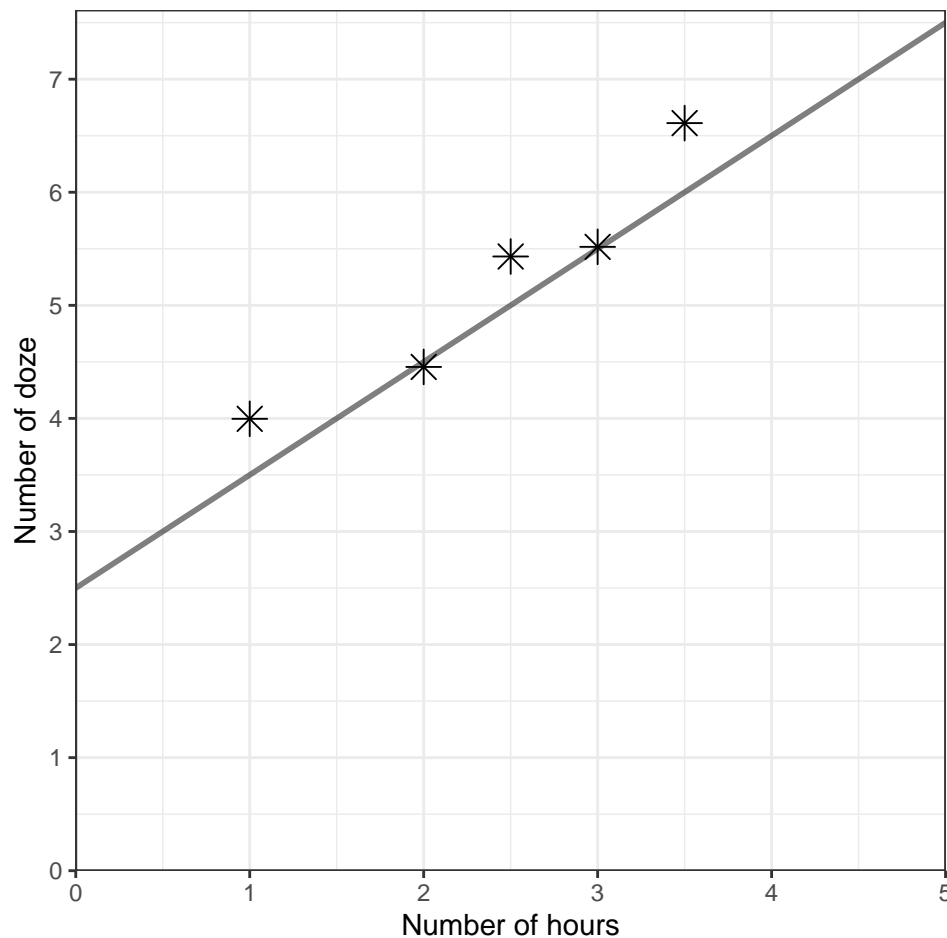
Mehr Informationen zu den Aufgaben in den folgenden Kapiteln aus dem Skript Bio Data Science.

- [Kapitel 32 - Simple lineare Regression](#)
- [Kapitel 33 - Maßzahlen der Modelgüte](#)
- [Kapitel 34 - Korrelation](#)

## 90 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)**

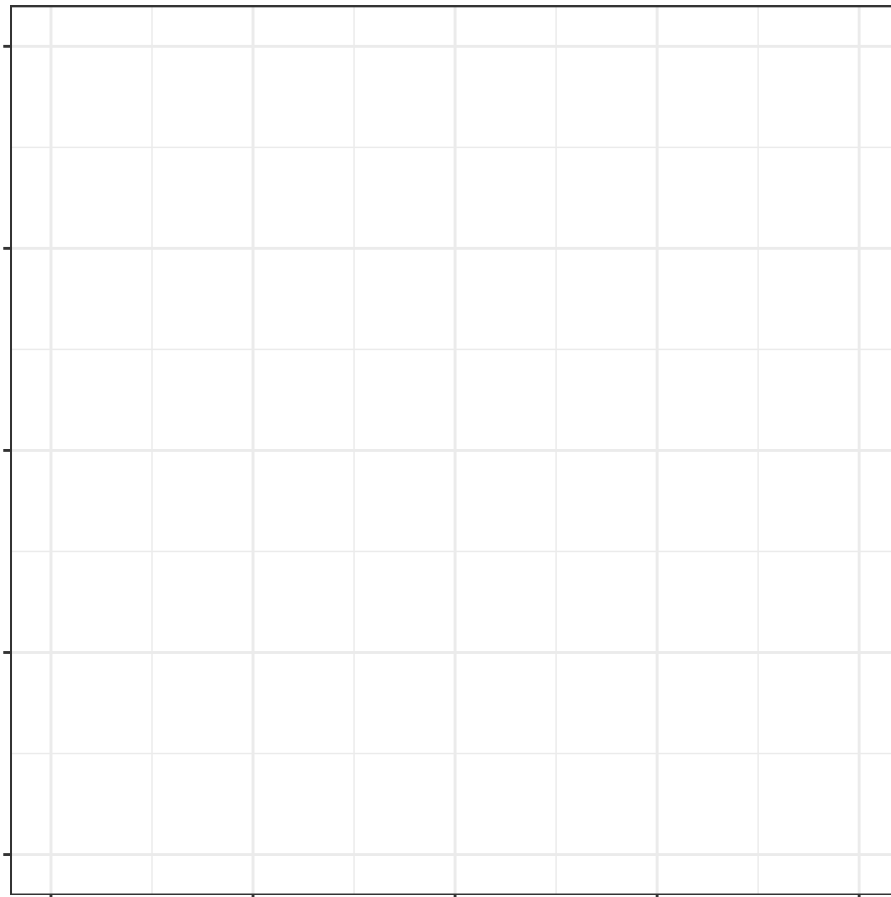
## 91 Aufgabe

(10 Punkte)

In einem Stallexperiment mit  $n = 97$  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)	24.73	0.98		
light	1.52	0.10		

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)**






## 92 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
## -3.20000 -1.42857 -0.81429  1.24286  5.80000
##
## Coefficients:
##              Estimate Std. Error t value    Pr(>|t|)
## (Intercept)   19.4286     1.0037  19.3576 0.000000002952 ***
## trtB           7.7714     1.5549   4.9981  0.0005388 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.6555 on 10 degrees of freedom
## Multiple R-squared:  0.71413, Adjusted R-squared:  0.68554
## F-statistic: 24.981 on 1 and 10 DF,  p-value: 0.00053883
```

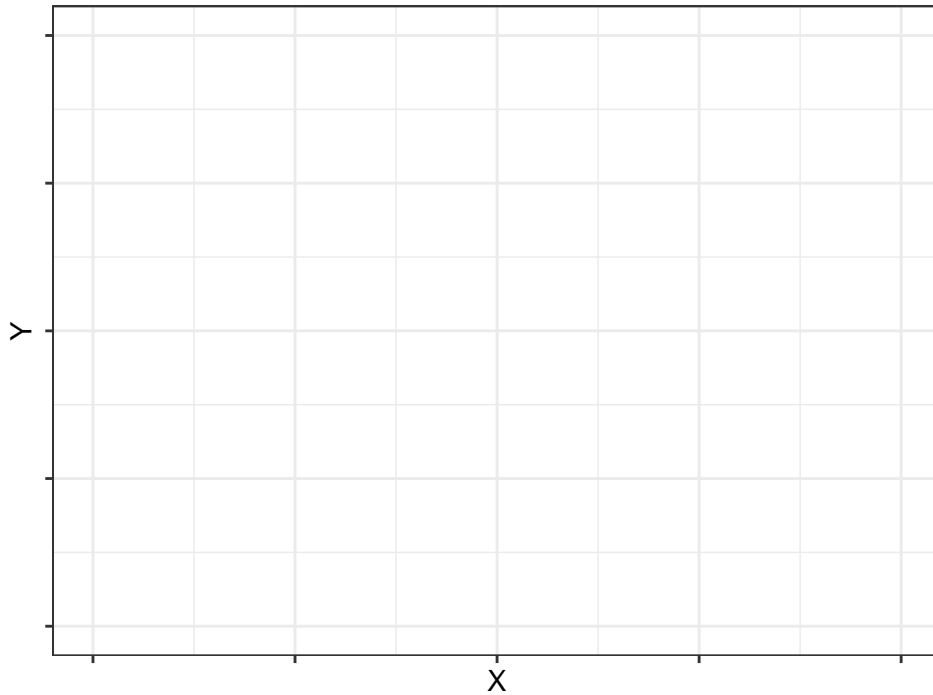
1. Ist die Annahme der Normalverteilung an das Outcome *rsp* erfüllt? Begründen Sie die Antwort! **(2 Punkte)**
2. Wie groß ist der Effekt des *Trt*? Liegt ein signifikanter Effekt vor? Geben Sie die Formel und den Rechenweg mit an! **(2 Punkte)**
3. Schreiben Sie das Ergebnis der  Ausgabe in einen Satz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

### 93 Aufgabe

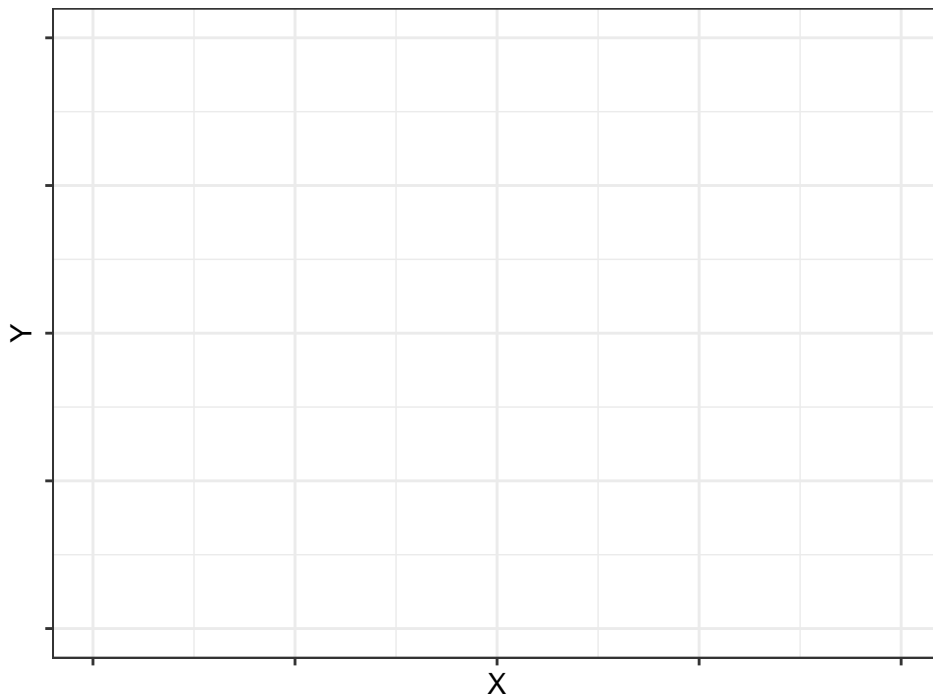
(6 Punkte)

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 effect of X



Predictive model with effect of X



## 94 Aufgabe

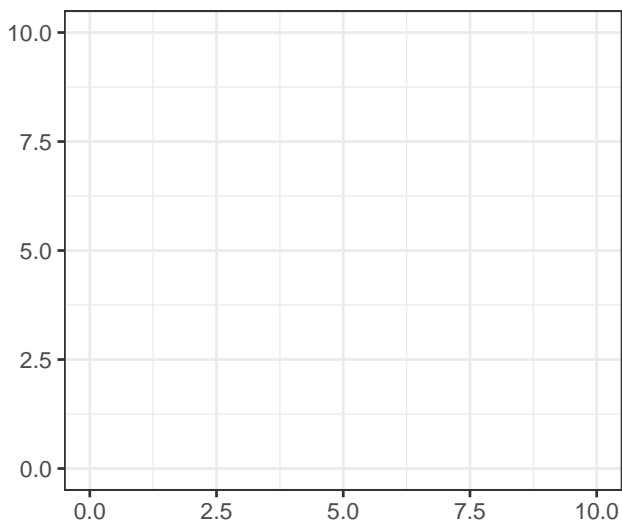
(9 Punkte)

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)**

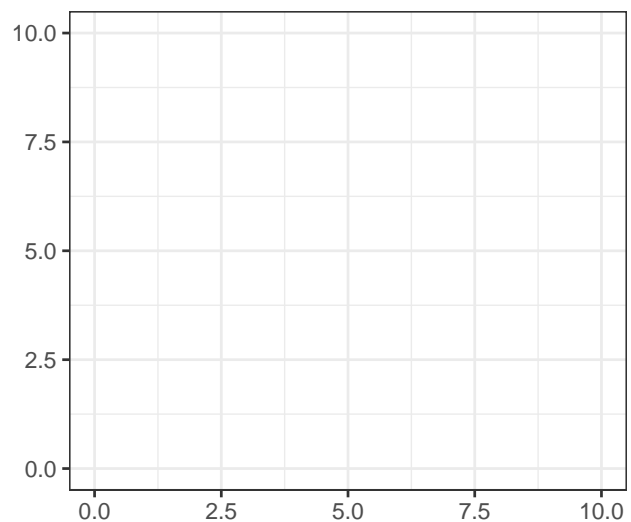
Pearsons  $\rho = -0.5$

$R^2 = 1$



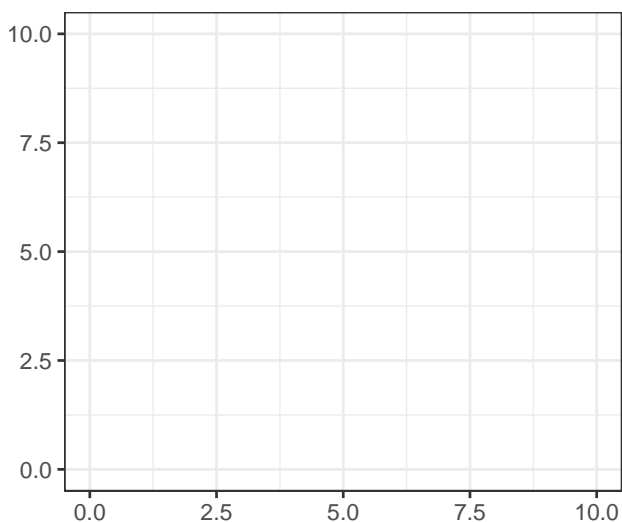
Pearsons  $\rho = -0.25$

$R^2 = 0.5$



Pearsons  $\rho = 1$

$R^2 = 0.75$



## 95 Aufgabe

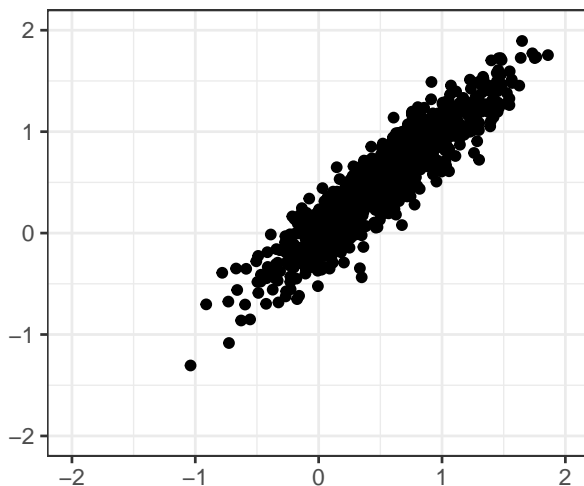
(10 Punkte)

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)

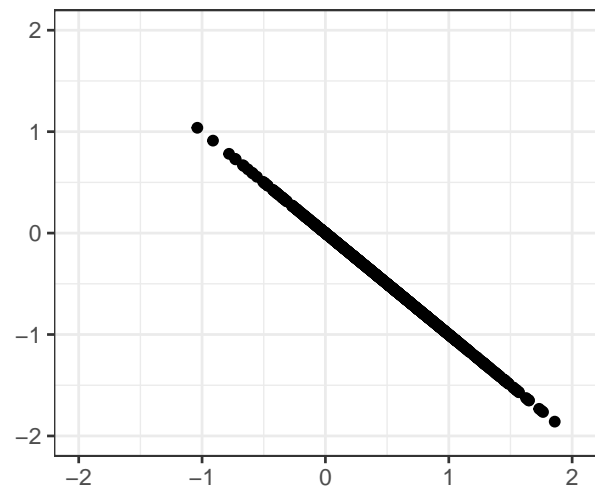
Pearsons  $\rho =$

$R^2 =$



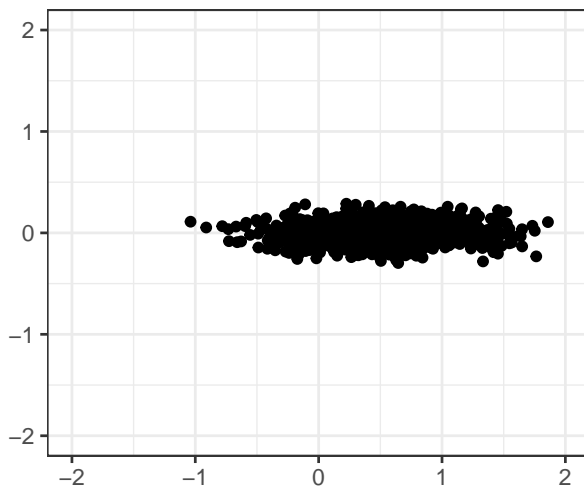
Pearsons  $\rho =$

$R^2 =$



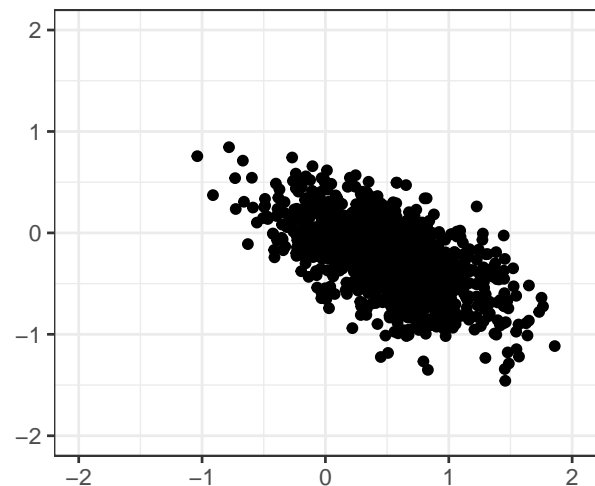
Pearsons  $\rho =$

$R^2 =$



Pearsons  $\rho =$

$R^2 =$

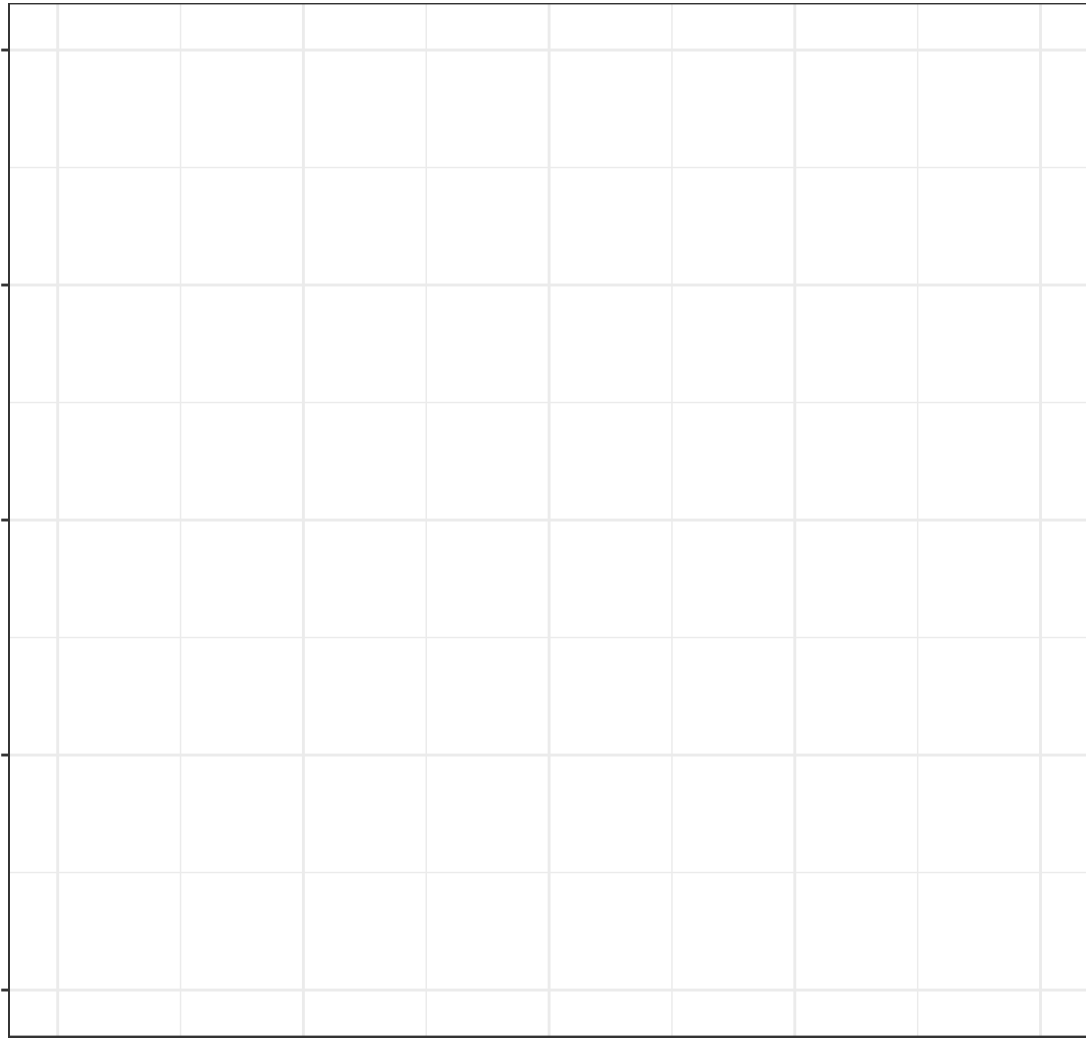


## 96 Aufgabe

(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 3 outlier fulfilling the normality assumption.



## 97 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.



## 98 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	20.5	7.9	22.1	
2	15.7	4.8	17.0	
3	34.2	15.7	35.0	
4	29.7	12.0	28.8	
5	26.6	9.4	24.7	
6	25.0	9.8	25.2	
7	24.0	6.9	20.6	
8	32.8	14.4	32.8	
9	21.0	6.9	20.5	
10	19.4	7.3	21.2	
11	24.8	9.9	25.5	
12	21.8	7.7	21.9	
13	23.2	8.8	23.6	

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)**

## Multiple linear Regression

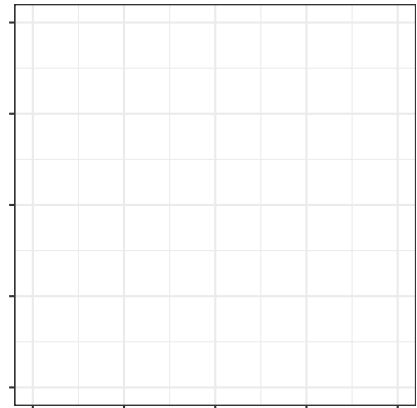
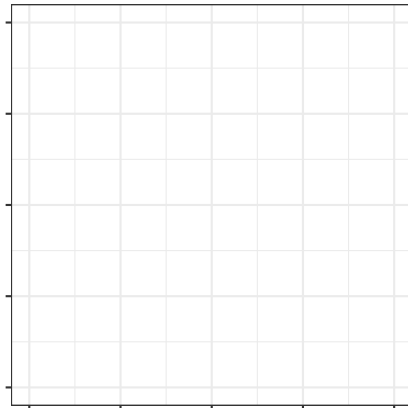
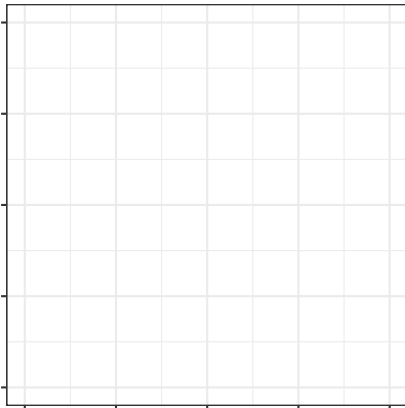
Mehr Informationen zu den Aufgaben in den folgenden Kapiteln aus dem Skript Bio Data Science.

- Kapitel 35 - Multiple lineare Regression
- Kapitel 40 - Gaussian Regression
- Kapitel 41 - Poisson Regression
- Kapitel 43 - Logistische Regression
- Kapitel 44 - Lineare gemischte Modelle

### 99 Aufgabe

(10 Punkte)

1. Zeichnen Sie in die drei untenstehenden, leeren Abbildungen die Zeile des Regressionskreuzes der Normalverteilung. 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)**







## 100 Aufgabe

(9 Punkte)

Ein Feldexperiment wurde mit  $n = 200$  Pflanzen durchgeführt. Folgende Einflussvariablen ( $x$ ) wurden erhoben: rainfall, dosage und fertilizier. 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 *Poisson* 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)**




## 101 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	maxdepth	acerage	so4
DORSEY_RUN	8	0.44	160	7814	20.24
HAY_MEADOW_BR	5	3.53	26	639	4.46
HOLLANDS_BR	38	2.34	85	1934	11.44
BEAVER_RUN	37	5.66	43	1722	5.91

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
## -9.9039  -4.0899  -2.5297   1.9196  16.1205
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.988203928  0.498386473   3.9893 0.0002209 ***
## no3          0.227800237  0.062952610   3.6186 0.0007000 ***
## maxdepth     0.011929391  0.003927903   3.0371 0.0038215 **
## acerage      0.000045321  0.000012397   3.6557 0.0006251 ***
## so4          -0.025268377  0.023723906  -1.0651 0.2920503
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 33.198023)
##
## Null deviance: 2463.84  on 53  degrees of freedom
## Residual deviance: 1411.14  on 49  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```


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)

## 102 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
	0	10.70	0	0
	0	10.88	0	1
	0	8.21	0	5
	0	7.83	0	4

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)    1.36        1.02    0.308
## 2 nematodes      0.124        0.603    0.547
## 3 water          0.136       -1.26    0.209
```


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)**

## 103 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
9.64	1	9.93	97.89	67.47	10
9.90	1	10.38	95.58	67.63	8
12.42	1	7.78	103.31	77.10	16
13.38	0	10.54	99.72	68.84	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
## -6.62687 -1.98724  0.13915  1.98437  5.92105
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  75.82186    2.52377  30.0431 < 0.0000000000000002 ***
## nematodes    -0.07466    0.12404  -0.6019    0.54886
## light         1.05810    0.70041   1.5107    0.13467
## P            -0.57924    0.23183  -2.4985    0.01444 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.994 on 83 degrees of freedom
## Multiple R-squared:  0.11879, Adjusted R-squared:  0.086935
## F-statistic: 3.7294 on 3 and 83 DF, p-value: 0.014372
```


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)**

## 104 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: 794.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.97515 -0.55750 -0.14600  0.59804  3.31704
##
## Random effects:
## Groups   Name      Variance Std.Dev.
## barn     (Intercept) 237.18   15.401
## subject  (Intercept) 4173.57   64.603
## Residual                    652.23   25.539
## Number of obs: 83, groups: barn, 7; subject, 6
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 203.0419    27.2948  7.4388
## attitudepop -19.4621     5.6115 -3.4683
##
## Correlation of Fixed Effects:
##              (Intr)
## attitudepop -0.101
```

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)**

## Nicht parametrische Tests

Mehr Informationen zu den Aufgaben in den folgenden Kapiteln aus dem Skript Bio Data Science.

- Kapitel 25 - Der Wilcoxon-Mann-Whitney-Test
- Kapitel 26 - Der Kruskal-Wallis-Test

### 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
8	11					
9	14					
10	13					
6	12					
10	16					
16	16					
9	12					
9	12					
12	11					
11	11					
15	14					
11	13					
7	14					
12	11					
7	14					

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  $z = \frac{W - \mu_W}{\sigma_W}$ ! **(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  $r = \left| \frac{z}{\sqrt{n}} \right|$  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	9.3
RootsGoneX	15.1
RootsGoneX	9.2
RootsGoneX	10.7
RootsGoneX	5.8
RootsGoneX	11.0
RootsGoneX	8.6
Kontrolle	16.1
Kontrolle	12.2
Kontrolle	14.1
Kontrolle	11.1
Kontrolle	10.4
Kontrolle	12.6
Kontrolle	11.6

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

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

- 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)**

- 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
Dosis 20	5.0			
Dosis 20	5.1			
Kontrolle	6.6			
Dosis 20	8.0			
Kontrolle	8.2			
Kontrolle	8.3			
Dosis 20	9.4			
Kontrolle	9.9			
Dosis 40	11.2			
Kontrolle	11.8			
Kontrolle	12.7			
Dosis 40	13.8			
Dosis 40	14.4			
Dosis 20	15.0			
Dosis 20	15.9			
Kontrolle	16.1			
Dosis 40	17.4			
Dosis 40	19.5			

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)



## Multiple Gruppenvergleiche

Mehr Informationen zu den Aufgaben in den folgenden Kapiteln aus dem Skript Bio Data Science.

- Kapitel 20.3 Adjustierung für multiple Vergleiche

### 108 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	0.060		
dose 15 - ctrl	0.760		
dose 20 - ctrl	0.030		
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! **(2 Punkte)**
3. Erklären Sie warum die p-Werte bei multiplen Vergleichen adjustiert werden müssen? **(2 Punkte)**