Checkbox für die Version vom 29. November 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 Aufgaben*fragen* 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`

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 29. November 2022

Erlaubte Hilfsmittel für die Klausur

- Normaler Taschenrechner ohne Möglichkeit der Kommunikation mit anderen Geräten - also ausdrücklich kein Handy!
- Eine DIN A4-Seite als beidseitig, selbstgeschriebene, handschriftliche Formelsammlung - keine digitalen Ausdrucke.

Ergebnis der Klausur

______ von 20 Punkten sind aus dem Multiple Choice Teil erreicht.

______ von 60 Punkten sind aus dem Rechen- und Textteil erreicht.

______ von 80 Punkten in Summe.

Es wird folgender Notenschlüssel angewendet.

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

	A	**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 α Adjustierung ist richtig?

A ☐ Die α 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.

B ☐ Die α Adjustierung wird durchgeführt um den Effekt von Interesse, meist die Behandlung, von anderen Effekten zu trennen. Daher eine Adjustierung auf den β-Werten einer Regression.

C ☐ Die α Adjustierung wird durchgeführt um bei multiplen Vergleichen den Fehler 1. Art zu kontrollieren. Es wird die Irrtumswahrscheinlichkeit adjustiert, daher das α-Niveau.

D ☐ Die α 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.

E ☐ Die α Adjustierung wird meist ignoriert. Wenn die Annahmen an den statistischen Test richtig sind, kann auf eine Adjustierung verzichtet werden.

2 Aufgabe (2 Punkte)

Sie haben folgende unadjustierten p-Werte gegeben: 0.03, 0.21, 0.42, 0.001 und 0.02. Sie adjustieren die p-Werte nach Bonferroni. Welche Aussage ist richtig?

A ☐ Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 0.15, 1.05, 2.1, 0.005 und 0.1. Die adjustierten p-Werte werden zu einem α-Niveau von 5% verglichen.

B ☐ Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 0.006, 0.042, 0.084, 0.0002 und 0.004. Die adjustierten p-Werte werden zu einem α-Niveau von 1% verglichen.

C ☐ Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 0.006, 0.042, 0.084, 0.0002 und 0.004. Die adjustierten p-Werte werden zu einem α-Niveau von 5% verglichen.

D ☐ Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 0.15, 1, 1, 0.005 und 0.1. Die adjustierten p-Werte werden zu einem α-Niveau von 1% verglichen.

E ☐ Nach der Bonferroni-Adjustierung ergeben sich die adjustierten p-Werte von 0.15, 1, 1, 0.005 und 0.1. Die adjustierten p-Werte werden zu einem α-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 $\eta^2 = 0.22$. Welche Aussage ist richtig?

A ☐ Das η^2 beschreibt den Anteil der Varianz, der von den Behandlungsbedingungen erklärt wird. Das η^2 ist damit mit dem R^2 aus der linearen Regression zu vergleichen.

B ☐ Das η^2 ist die Korrelation der ANOVA. Mit der Ausnahme, dass 0 der beste Wert ist.

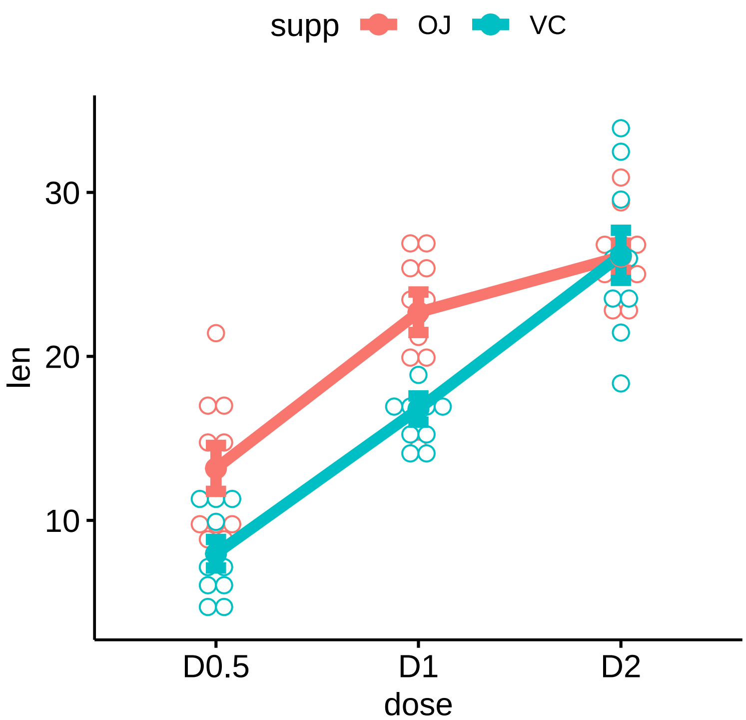
C ☐ Die Berechnung von η^2 ist ein Wert für die Interaktion.

D ☐ Das η^2 beschreibt den Anteil der Varianz, der von den Behandlungsbedingungen nicht erklärt wird. Somit der Rest an nicht erklärbarer Varianz.

E ☐ Das η^2 ist ein Wert für die Güte der ANOVA. Je kleiner desto besser. Ein η^2 von 0 bedeutet ein perfektes Modell mit keiner Abweichung. Die Varianz ist null.

4 Aufgabe (2 Punkte)

Die folgende Abbildung enthält die Daten aus einer Studie zur Bewertung der Wirkung von Vitamin C auf das Zahnwachstum bei Meerschweinchen. Der Versuch wurde an 60 Schweinen durchgeführt, wobei jedes Tier eine von drei Vitamin-C-Dosen (0.5, 1 und 2 mg/Tag) über eine von zwei Verabreichungsmethoden erhielt (Orangensaft oder Ascorbinsäure (eine Form von Vitamin C und als VC codiert). Die Zahnlänge wurde gemessen, und eine Auswahl der Daten ist unten dargestellt.



Welche Aussage ist richtig im Bezug auf eine zweifaktorielle ANOVA?

A ☐ Ein signifikanter Effekt ist nicht zu erwarten. Durch das Kreuzen der OJ und VC Geraden an der dose Faktorstufe D2 ist eine Interaktion vorhanden.

B ☐ Ein signifikanter Effekt ist zwischen allen dose Faktorstufen zu erwarten. Durch das Kreuzen der OJ und VC Geraden an der dose Faktorstufe D2 ist keine Interaktion vorhanden.

C ☐ Der η^2-Wert sollte liegt bei 1, da eine Gerade zu beobachten ist. Dadurch ist die Interaktion zwischen den dose Faktorstufen nicht mehr signifikant. Der Faktor supp hat keinen Einfluss.

D ☐ Ein signifikanter Effekt ist zwischen allen dose Faktorstufen nicht zu erwarten. Eine Interaktion liegt nicht vor. Der η^2-Wert sollte bei ca. 0 liegen.

E ☐ Ein signifikanter Effekt ist zwischen mindestens einer dose Faktorstufen zu erwarten. Durch das Kreuzen der OJ und VC Geraden an der dose Faktorstufe D2 ist mit einem signifkanten Interaktionsterm zu rechnen.

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 T-Statistik indem den Mittelwertsunterschied der Gruppen simultan durch die Standardabweichung der Gruppen teilt. Wenn die T-Statistik größer als 1.96 ist, kann die Nullhypothese abgelehnt werden.

B ☐ Die ANOVA berechnet die F-Statistik indem die MS der Behandlung durch die MS des Fehlers geteilt werden. Wenn die F-Statistik sich der 0 annähert kann die Nullhypothese nicht abgelehnt werden.

C ☐ Die ANOVA berechnet die 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.

D ☐ 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.

E ☐ Die ANOVA berechnt die F-Statistik aus den SS Behandlung geteilt durch die SS Fehler.

6 Aufgabe (2 Punkte)

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

A ☐ Es liegt Varianzhetrogenität vor.

B ☐ Es liegt Varianzhomogenität vor.

C ☐ Es handelt sich um ein balanciertes Design.

D ☐ Es handelt sich um abhängige Beobachtungen.

E ☐ Es handelt sich um ein unbalanciertes Design

7 Aufgabe (2 Punkte)

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

A ☐ 5 Beobachtungen

B ☐ 2 Beobachtungen

C ☐ 1 Beobachtung

D ☐ 10 Beobachtungen

E ☐ 0 Beobachtungen

8 Aufgabe (2 Punkte)

Berechnen Sie die Korrelation r zwischen x mit 12, 11, 6, 15 und 13 sowie y mit 10, 7, 7, 9 und 9. Nutzen Sie folgende Formel:

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

A ☐ Es ergibt sich eine Korrelation von 0.68

B ☐ Es ergibt sich eine Korrelation von 1.68

C ☐ Es ergibt sich eine Korrelation von 0.46

D ☐ Es ergibt sich eine Korrelation von 0.31

E ☐ Es ergibt sich eine Korrelation von -0.68

9 Aufgabe (2 Punkte)

Berechnen Sie die Kovarianz $s_{x,y}$ und die Korrelation r zwischen x mit 3, 13, 11, 5 und 15 sowie y mit 17, 14, 19, 19 und 18.

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

A ☐ Es ergibt sich eine Kovarianz von 2.7 sowie eine Korrelation von 0.06

B ☐ Es ergibt sich eine Kovarianz von 7.29 sowie eine Korrelation von 0.06

C ☐ Es ergibt sich eine Kovarianz von -2.7 sowie eine Korrelation von -0.25

D ☐ Es ergibt sich eine Kovarianz von -19.68 sowie eine Korrelation von 0.25

E ☐ Es ergibt sich eine Kovarianz von -2.7 sowie eine Korrelation von 0.25

10 Aufgabe (2 Punkte)

Berechnen Sie den Mittelwert und Standardabweichung von y mit 9, 8, 7, 9 und 5.

A ☐ Es ergibt sich 7.6 +/- 1.67

B ☐ Es ergibt sich 7.6 +/- 2.8

C ☐ Es ergibt sich 8.6 +/- 0.835

D ☐ Es ergibt sich 6.6 +/- 1.4

E ☐ Es ergibt sich 7.6 +/- 0.835

11 Aufgabe (2 Punkte)

Berechnen Sie den Median und das IQR von x mit 13, 9, 20, 17, 34, 30, 22, 12, 16, 33 und 51.

A ☐ Es ergibt sich 20 +/- 13

B ☐ Es ergibt sich 23 [13, 22]

C ☐ Es ergibt sich 20 [13, 22]

D ☐ Es ergibt sich 20 +/- 22

E ☐ Es ergibt sich 23 +/- 13

12 Aufgabe (2 Punkte)

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

A ☐ ... die Methode des absoluten, quadrierten Abstands.

B ☐ ... das Produkt der kleinsten Quadrate.

C ☐ ... die Methode des absoluten Abstands.

D ☐ ... die Methode der aufaddierten, absoluten Abstände.

E ☐ ... die kleinste Quadrate Methode oder auch least square method genannt.

13 Aufgabe (2 Punkte)

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

A ☐ Der Begriff *ordinary mean* beschreibt gewöhnliche Mittelwerte und wohingegen *marginal mean* kleine Mittelwerte beschreibt.

B ☐ Der Begriff *ordinary mean* beschreibt die einzelnen Mittelwerte der Level eines Faktors wohingegen der *marginal mean* den Mittelwert über alle Level des Faktors wiedergibt.

C ☐ Der Begriff *ordinary mean* beschreibt einen unbedeutenden Mittelwert und *marginal mean* einen Randmittelwert.

D ☐ Der Begriff *ordinary mean* beschreibt einen einzelnen Mittelwert eines Levels eines Faktors wohingegen der *marginal mean* die unbedeutenden Mittelwerte beschreibt.

E ☐ Der Begriff *ordinary mean* beschreibt die Mittelwerte eines Faktors wohingegen der *marginal mean* die einzelnen Level betrachtet.

14 Aufgabe (2 Punkte)

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

A ☐ 1 Beobachtung.

B ☐ 2-5 Beobachtungen.

C ☐ 10 Beobachtungen.

D ☐ mindestens 20 Beobachtungen.

E ☐ 5 und mehr Beobachtungen.

15 Aufgabe (2 Punkte)

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

A ☐ Die Verteilung von Daten. Meistens dem Outcome oder auch y genannt. Hierbei ist eine große Fallzahl notwendig.

B ☐ Die Verteilung von Daten. Meistens dem Outcome oder auch y genannt. Die Darstellung ist auch mit einer kleinen Fallzahl möglich.

C ☐ Die Eigenschaften von Daten aufgeteilt nach zwei Gruppen eines Faktors x.

D ☐ Die Verteilung von Daten. Auch 2 bis 5 Beonachtungen können noch dargestellt werden. Auch wenn es sich streng genommen um keine Darstellugn einer Verteilung handelt.

E ☐ Die Eigenschaften von Daten anhand der fehlenden Werte und den Leveln eines Faktors x.

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 Possion-Regression auf die Anzahl an aufgefundenen Bachsalamandern an den jeweiligen Suchorten. Welche Aussage zur Possion-Regression auf Zähldaten ist richtig?

A ☐ Die Possion-Regression schätzt drei Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzen Overdispersion von 3.06 liegt Overdispersion vor. Die Lösung ist die Nutzung von nur zwei der drei Verteilungsparameter: γ_1 und γ_3.

B ☐ Die Possion-Regression schätzt nur einen Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzen Overdispersion von 3.06 liegt Overdispersion vor. Die Lösung ist die Nutzung einer anderen Verteilungsfamilie wie die Quasipossion Verteilung.

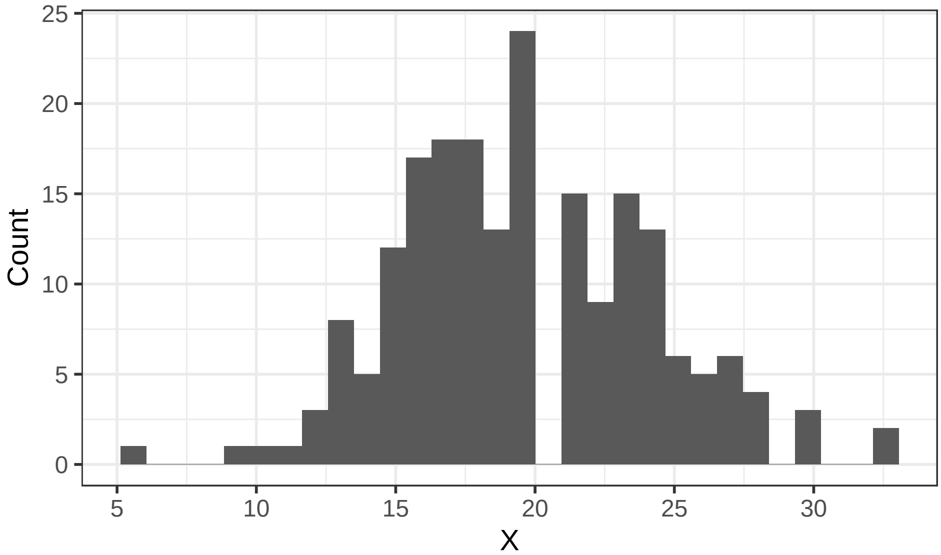
C ☐ Die Possion-Regression schätzt zwei Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzen Overdispersion von 3.06 liegt keine Overdispersion vor.

D ☐ Die Possion-Regression schätzt nur einen Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzen Overdispersion von 3.06 liegt keine Overdispersion vor. Overdispersion liegt vor, wenn die geschätzte Overdispersion unter 1 liegt.

E ☐ Die Possion-Regression schätzt nur einen Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzen Overdispersion von 3.06 liegt Overdispersion vor. Damit kann keine Possion-Regression gerechnet werden. Die Lösung ist eine Gaussian Regression mit Nullanpassung.

17 Aufgabe (2 Punkte)

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



A ☐ Eine Standardnormalverteilung mit N(0,1).

B ☐ Eine rechtsschiefe, multivariate Normalverteilung.

C ☐ Es handelt sich um eine Binomial-Verteilung mit Binom(10).

D ☐ Es handelt sich um eine Poisson-Verteilung mit Pois(20).

E ☐ Es handelt sich um eine Normalverteilung mit N(20, 5).

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 Possion-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 Possion Verteilung und Modell 2 mit einer Quasi-Poisson Verteilung. Welche Aussage zu einer geschätzen Overdispersion von 3.23 ist richtig?

A ☐ Bei einer geschätzen 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.

B ☐ 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.

C ☐ Bei einer geschätzen 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.

D ☐ Bei einer geschätzen 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

E ☐ Bei einer geschätzen 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.

19 Aufgabe (2 Punkte)

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

A ☐ Untereinander unabhängig. Die Ferkel sind eigenständig und benötigen keine zusätzliche Behandlung.

B ☐ Untereinander abhängig, wenn die Mütter ebenfalls miteinander verwandt sind. Erst die Abhängigkeit 2. Grades wird in der Statistik modelliert.

C ☐ Untereinander stark korreliert. Die Ferkel sind von einer Mutter und sommit miteinander korreliert. Dies wird in der Statistik jedoch meist nicht modelliert.

D ☐ Untereinander unabhängig. Sollten die Mütter verwandt sein, so ist die Varianzstruktur ähnlich und muss modelliert werden.

E ☐ Untereinander abhängig. Die Ferkel stammen von einem Muttertier und haben vermutliche eine ähnliche Varianzstruktur.

20 Aufgabe (2 Punkte)

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

A ☐ Y beinhaltet eine Spalte. Die Spalte gibt die Verteilungsfamilie vor.

B ☐ Y beinhaltet die Zeilen. Die Zeilen geben die Verteilungsfamilie vor.

C ☐ Y beinhaltet eine Spalte. Die Spalte gibt nicht die Verteilungsfamilie vor.

D ☐ Y beinhaltet mehrere Spalten. Die Spalten enthalten die Behandlung und weitere potenzielle Einflussvariablen

E ☐ Y beinhaltet mehrere Spalten. Die Spalten geben die Verteilungsfamilie vor.

21 Aufgabe (2 Punkte)

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

A ☐ Ein prädiktives Modell basiert auf einem Traingsdatensatz und einem Testdatensatz. Auf dem Trainingsdatensatz wird das Modell trainiert und auf dem Testdatensatz validiert.

B ☐ Ein prädiktives Modell benötigt mindestens eine Fallzahl von über 100 Beobachtungen und darf keine fehlenden Werte beinhalten. Die Varianzkomponenten müssen homogen sein.

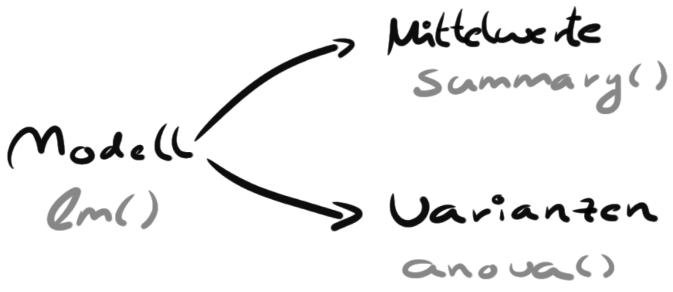
C ☐ Ein prädiktives Modell wird auf einem Trainingsdatensatz trainiert und anschliessend über eine explorative Datenanalyse validiert. Signifikanzen über β_i können hier nicht festgestellt werden.

D ☐ Ein prädiktives Modell schliesst grundsätzlich lineare Modell aus. Es muss ein Graph gefunden werden, der alle Punkte beinhaltet. Erst dann kann das R^2 berechnet werden.

E ☐ Ein prädiktives Modell möchte die Zusammenhänge von X auf Y modellieren. Hierbei geht es um die Effekte von X auf Y. Man sagt, wenn X um 1 ansteigt ändert sich Y um einen Betrag β.

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, in diesem Fall ein lineares Modell, erlauben es sowohl eine ANOVA zurechnen sowie auch eine Zusammenfassung der Mittelwerte zu betrachten.

B ☐ Die Effektschätzer aus einem Modell, in diesem Fall ein polynomes Modell, erlauben es sowohl eine ANOVA zurechnen sowie auch eine Zusammenfassung der Mittelwerte zu betrachten.

C ☐ 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.

D ☐ Die Effektschätzer aus einem Modell, erlauben es nur einen Mittelwertsvergleich zu rechnen.

E ☐ Die Effektschätzer aus einem Modell, erlauben es nur eine ANOVA zu rechnen.

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 X werden modelliert.

B ☐ Die Varianzstruktur wird modelliert.

C ☐ Die Varianz der X unabhängig vom Y wird modelliert.

D ☐ Die Y werden modelliert.

E ☐ Die Mittelwerte werden 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 12 Tiere sind gesund. Welche Aussage über den Odds ratio Effektschätzer ist richtig?

A ☐ Es ergibt sich ein Odds ratio von 0.42, da es sich um eine Chancenverhältnis handelt.

B ☐ Es ergibt sich ein Odds ratio von 0.29, da es sich um eine Chancenverhältnis handelt.

C ☐ Es ergibt sich ein Odds ratio von 0.29, da es sich um ein Anteil handelt.

D ☐ Es ergibt sich ein Odds ratio von 0.42, da es sich um ein Anteil handelt.

E ☐ Es ergibt sich ein Odds ratio von 2.4, 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 Rängen. Daher gibt es auch direkt zu interpretierenden Effektschätzer.

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

C ☐ Die nicht-parametrische Statistik ist ein Vorgänger der parametrischen Statistik und wurde wegen dem Mangel an Effektschätzern nicht mehr ab 1960 genutzt.

D ☐ Die nicht-parametrische Statistik basiert auf dem Schätzen von Parametern aus einer festgelegten Verteilung. Daher gibt es auch direkt zu interpretierenden Effektschätzer.

E ☐ 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.

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 war bis 1952 bedeutend, wurde dann aber in Folge besserer Rechnerleistung nicht mehr verwendet. Aktuelle Statistik nutzt keine Randomisierung mehr.

B ☐ Randomisierung sorgt für Strukturgleichheit und erlaubt erst von der Stichprobe auf die Grundgesamtheit zurückzuschliessen.

C ☐ Randomisierung bringt starke Unstrukturiertheit in das Experiment und erlaubt erst von der Stichprobe auf die Grundgesamtheit zurückzuschliessen.

D ☐ Randomisierung erlaubt erst die Varianzen zu schätzen. Ohne eine Randomisierung ist die Berechnung von Mittelwerten und Varianzen nicht möglich.

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 R auswerten wollen. Welcher folgende Grund ist richtig?

A ☐ 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.

B ☐ Im Allgemeinen haben Programmiersprachen Probleme mit Umlauten und Sonderzeichen, die in der deutschen Sprache vorkommen. Eine Nutzung der englischen Sprache umgeht dieses Problem auf einfache Art.

C ☐ Die Spracherkennung von R ist nicht in der Lage Deutsch zu verstehen.

D ☐ Es gibt keinen Grund nicht auch deutsche Wörter zu verwenden. Es ist ein Stilmittel.

E ☐ Alle Funktionen und auch Anwendungen sind in R in englischer Sprache. Die Nutzung von deutschen Wörtern ist nicht schick und das ist zu vermeiden.

28 Aufgabe (2 Punkte)

Bei der explorativen Datenanalyse (EDA) in R 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 abschließend über ggplot() uns die Abbildungen erstellen lassen. Wichtig ist, dass wir keine Faktoren sondern nur numerische Variablen vorliegen haben.

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

C ☐ Wir lesen 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)` schließen wir das Paket von der Analyse aus.

D ☐ Wir lesen als erstes die Daten über `read_excel()` ein, transformieren die Spalten über `mutate()` in die richtige Form und können dann abschließend über ggplot() uns die Abbildungen erstellen lassen.

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

29 Aufgabe (2 Punkte)

In einem Stallexperiment mit $n = 106$ 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 R Output der Funktion `tidy()` einer simplen *possion* linearen Regression. Welche Aussage über den **Effekt** ist richtig?

term	estimate	std.error
(Intercept)	0.63	0.10
light_binhigh	0.09	0.14

A ☐ Eine possion Regression basiert auf dem maximum Likelihood Prinzip. Hierbei kann kein Effekt beschrieben werden. Im Zweifel hilft aber eine Quadrierung der Fehlerqudrate ϵ.

B ☐ In einer possion Regression kann kein Effekt roh interpretiert werden. Es muss erst eine Confounderadjustierung durchgeführt werden.

C ☐ In einer possion Regression wird die Mittelwertsdifferenz betrachtet. Daher ist der Effekt zwischen den beiden Lichtregimen eine Gewichtsänderung von 0.09

D ☐ In einer possion Regression muss für die Interpretation des Effektes das β_1 quadriert werden. Somit liegt das OR bei 0.01

E ☐ In einer possion Regression berechnet man das RR. Daher muss der Schätzer des Effektes β_1 noch quadriert werden. Somit liegt das RR bei 0.01

30 Aufgabe (2 Punkte)

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

A ☐ Die Residuen sind binomialverteilt.

B ☐ Die Residuen folgen einer Poissonverteilung mit Pois(0).

C ☐ Die Residuen sind normalverteilt mit $\mathcal{N}(\bar{y}, s^2)$.

D ☐ Die Residuen sind normalverteilt mit $\mathcal{N}(0, 1)$.

E ☐ Die Residuen sind normalverteilt mit $\mathcal{N}(0, s^2)$.

31 Aufgabe (2 Punkte)

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

A ☐ Das GLM ist eine Vereinfachung des LM in R. Mit dem GLM lassen polygonale Regressionen rechnen.

B ☐ Das GLM ist eine allgemeine Erweiterung der linearen Regression auf die Normalverteilung.

C ☐ Das GLM erlaubt auch nicht normalverteilte Residuen in der Schätzung der Regressionsgrade.

D ☐ Das GLM erlaubt auch weitere Verteilungsfamilien für das Y bzw. das Outcome in einer linearen Regression zu wählen.

E ☐ Das GLM ist ein faktisch maschineller Lernalgorithmus, der selstständig die Verteilungsfamilie für Y wählt.

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 254 und für das Modell B das AIC_B von 435. Welche Aussage über die beiden Modelle ist richtig?

A ☐ Da $AIC_A < AIC_B$ ist das Modell B das bessere Modell.

B ☐ Da $AIC_A > 0$ ist das Modell A das bessere Modell. Der AIC Wert für B wird verworfen.

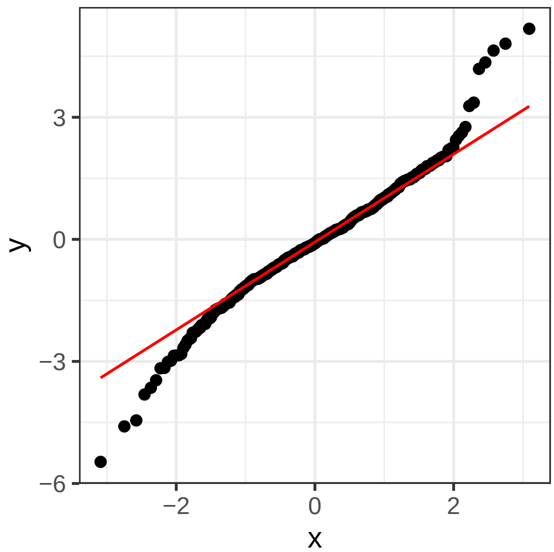
C ☐ Da $AIC_A < AIC_B$ ist das Modell A das bessere Modell.

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 auf der Geraden.

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

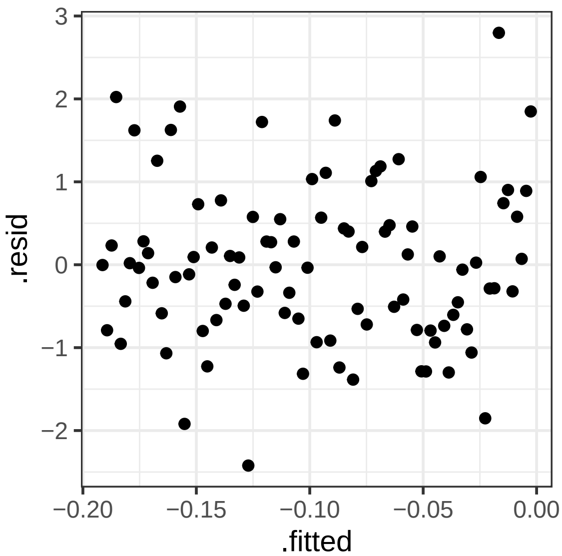
C ☐ Die Annahme der normalverteilten Residuen ist erfüllt. Die Punkte liegen zum überwiegenden Teil nicht auf der Geraden.

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

E ☐ Die Annahme der normalverteilten Residuen ist nicht erfüllt. Die Punkte liegen zum überwiegenden Teil nicht auf der Geraden.

34 Aufgabe (2 Punkte)

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



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

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

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

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

E ☐ Die Annahme der normalverteilten Residuen ist erfüllt. Es ist ein 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 $s = 4$ und $\delta_0 = 1$. Es ergibt sich somit folgende Fallzahl.

A ☐ Es ergibt sich eine Fallzahl von 274

B ☐ Es ergibt sich eine Fallzahl von 548

C ☐ Es ergibt sich eine Fallzahl von 275

D ☐ Es ergibt sich eine Fallzahl von 138

E ☐ Es ergibt sich eine Fallzahl von 550

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 in der Wissenschaft immer etwas falsch sein muss. Sonst gebe es keinen Fortschritt.

B ☐ ... dass Modelle meist falsch sind und selten richtig.

C ☐ ... dass Fehlerterme in statistischen Modellen nicht verifiziert werden können.

D ☐ ... dass Annahmen an statistische Modelle meist falsch sind.

E ☐ ... dass ein schlechtes Modell durch ein weniger schlechtes Modell ersetzt wird. Die Wissenschaft lehnt ab und verifiziert nicht.

38 Aufgabe (2 Punkte)

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

A ☐ 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.

B ☐ Im Rahmen eines langen Disputs zwischen Neyman und Fischer wurde α = 5% festgelegt. Leider werden die Randbedingungen und Voraussetzungen an statistsiche Modelle heute immer wieder ignoriert.

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

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

E ☐ 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.

39 Aufgabe (2 Punkte)

Welche Aussage über die Power ist richtig?

A ☐ Die Power $1-\beta$ wird auf 80% gesetzt. Alle statistischen Tests sind so konstruiert, dass die H_A mit 80% "bewiesen wird".

B ☐ Die Power beschreibt die Wahrscheinlichkeit die H_A abzulehnen. Wir testen die Power jedoch nicht.

C ☐ Es gilt $\alpha + \beta = 1$ und somit liegt β meist bei 95%.

D ☐ Die Power ist nicht in der aktuellen Testthorie mehr vertreten. Wir rechnen nur noch mit dem Fehler 1. Art.

E ☐ Die Power $1-\beta$ wird auf 80% gesetzt. Damit liegt die Wahrscheinlichkeit für die H_0 bei 20%.

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 $T = \dfrac{signal}{noise}$

B ☐ Es gilt $T = (signal \cdot noise)^2$

C ☐ Es gilt $T = signal \cdot noise$

D ☐ Es gilt $T = \dfrac{signal}{noise^2}$

E ☐ Es gilt $T = \dfrac{noise}{signal}$

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 brennenden Hauses ohne Rauchmelder: *House without noise*.

C ☐ In die Analogie eines Rauchmelders: *Alarm without fire*, dem α-Fehler.

D ☐ In die Analogie eines Rauchmelders: *Fire without alarm*, dem β-Fehler.

E ☐ In die Analogie eines Rauchmelders: *Alarm with fire*.

42 Aufgabe (2 Punkte)

Sie rechnen eine simple Poisson Regression. Welche Aussage bestreffend der Konfidenzintervalle ist für die Poisson Regression richtig?

A ☐ Wenn die 0 im Konfidenzinterval enthalten ist, kann die Nullhypothese abgelehnt werden.

B ☐ Wenn die Konfidenzintervalle den p-Wert der Regression enthalten, kann die Nullhypothese abgelehnt werden.

C ☐ Wenn die 1 im Konfidenzinterval enthalten ist, kann die Nullhypothese nicht abgelehnt werden.

D ☐ Wenn die 0 im Konfidenzinterval enthalten ist, kann die Nullhypothese nicht abgelehnt werden.

E ☐ Wenn die Relevanzschwelle mit enthalten ist, 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 immer grössere Datensätze erhoben. Eine erhöhte Fallzahl führt automatisch auch zu mehr signifikanten Ergebnissen, selbst wenn die eigentlichen Effekte nicht relevant sind.

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 ☐ Relevanz und Signifikanz haben nichts miteinander zu tun. Daher gibt es auch keinen Zusammenhang zwischen hoher Fahlzahl (n > 10000) und einem signifikanten Test. Ein Effekt ist immer relevant und somit signifikant.

D ☐ Aktuell werden immer grössere Datensätze erhoben. Dadurch wird auch die Varianz immer höher was automatisch zu mehr signifikanten Ergebnissen führt.

E ☐ Aktuell werden zu grosse Datensätze für die gänigige Statistik gemessen. Daher wendet man maschinelle Lernverfahren für kausale Modelle an. Hier ist die Relevanz gleich Signifikanz.

44 Aufgabe (2 Punkte)

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

A ☐ Das OR. Als Chancenverhältnis gibt es das Verhältnis von Relevanz und Signifikanz wieder.

B ☐ Das Δ. Durch die Effektstärke haben wir einen Wert für die Relevanz, die vom Anwender bewertet werden muss. Da Δ antiproportional zum p-Wert ist, bedeutet auch ein hohes Δ ein sehr kleinen p-Wert.

C ☐ 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.

D ☐ Das Konfidenzintervall. Durch die Visualizierung des Konfidenzintervals kann eine Relevanzschwelle vom Anwender definiert werden. Zusätzlich erlaubt das Konfidenzinterval auch eine Entscheidung über die Signifikanz.

E ☐ Der p-Wert. Durch den Vergleich mit α lässt sich über die Signifikanz entscheiden und der β-Fehler erlaubt über die Power eine Einschätzung der Relevanz.

45 Aufgabe (2 Punkte)

Welche Aussage über die frequentistischen Testtheorie ist richtig?

A ☐ Wir machen Aussagen über den Effekt!

B ☐ Wir machen keine Aussagen über Wahrscheinlichkeiten!

C ☐ Wir machen Aussagen über den Nicht-Zustand/Keinen Effekt!

D ☐ Wir machen Aussagen über die individuelle Wahrscheinlichkeit eines Effektes!

E ☐ Wir machen Aussagen über Individuen!

46 Aufgabe (2 Punkte)

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

A ☐ Wir machen eine Aussage über die indivduelle Wahrscheinlichkeit des Eintretens der Nullhypothese H_0.

B ☐ Wir vergleichen mit dem p-Wert und dem Signifikanzniveau α Wahrscheinlichkeiten und damit die absoluten Werte auf einem Zahlenstrahl, wenn die H_0 gilt.

C ☐ Wir vergleichen mit dem p-Wert und dem Signifikanzniveau α Wahrscheinlichkeiten und damit die Flächen unter der Kurve der Teststatistik, wenn die H_0 gilt.

D ☐ Wir vergleichen die Effekte des p-Wertes mit den Effekten der Signifiaknzschwelle unter der Annahme der Nullhypothese.

E ☐ Wir vergleichen mit dem p-Wert und dem Signifikanzniveau α absolute Werte auf einem Zahlenstrahl und damit den Unterschied der Teststatistiken, 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.

B ☐ 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.

C ☐ Der t-Test testet generell zu einem erhöhten α-Niveau von 20%.

D ☐ Der t-Test ist ein Vortest der ANOVA und basiert daher auf dem Vergleich von Streuungsparametern

E ☐ Der t-Test vergleicht die Varianzen von mindestens zwei oder mehr Gruppen

48 Aufgabe (2 Punkte)

Welche Aussage über den Welch t-Test ist richtig?

A ☐ Der Welch t-Test wird angewendet, wenn Varianzheterogenität zwischen den beiden zu vergleichenden Gruppen vorliegt.

B ☐ Der Welch t-Test ist ein Post-hoc Test der ANOVA und basiert daher auf dem Vergleich der Varianz.

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 ist die veraltete Form des Student t-Test und wird somit nicht mehr verwendet.

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 α-Schwelle. Sie schauen sich auch die rohen, unadjustierten p-Werte an und finden hier als niedrigsten p-Wert $p_{3-2} = 0.053$. Welche Aussage ist richtig?

A ☐ Die ANOVA testet auf der gesamten Fallzahl. Es wäre besser die ANOVA auf der gleichen Fallzahl wie die einzelnen t-Tests zu rechnen.

B ☐ 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.

C ☐ Der Fehler liegt in den t-Tests. Wenn eine ANOVA signifikant ist, dann muss zwangsweise auch ein t-Test signifikant sein.

D ☐ Die ANOVA testet auf der gesamten Fallzahl. Die einzelnen t-Tests immer nur auf einer kleineren Subgruppe. Da mit weniger Fallzahl weniger signifikante Ergebnisse zu erwarten sind, kann eine Diskrepenz zwischen der ANOVA und den paarweisen t-Tests auftreten.

E ☐ Es gibt einen Fehler in der Varianzstruktur. Daher kann die ANOVA nicht richtig sein und paarweise t-Tests liefern das richtige Ergebnis.

50 Aufgabe (2 Punkte)

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

A ☐ Der gepaarte t-Test nutzt die Varianz der Beobachtungen jeweils paarweise und bildet dafür eine verbundene Stichprobe. Dieser Datensatz d dient dann zur Differenzbildung.

B ☐ Der gepaarte t-Test wird genutzt, wenn die Differenzen der Beobachtungen verbunden sind und wir dadurch die Unabhängigkeit nicht mehr vorliegen haben.

C ☐ 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.

D ☐ Der gepaarte t-Test wird gerehnet, wenn die Beobachtungen abhängig voneinander sind. Wir messen jede Beobachtung nur einmal und berechnen dann die Differenz zu dem Mittel der anderen Beobachtungen.

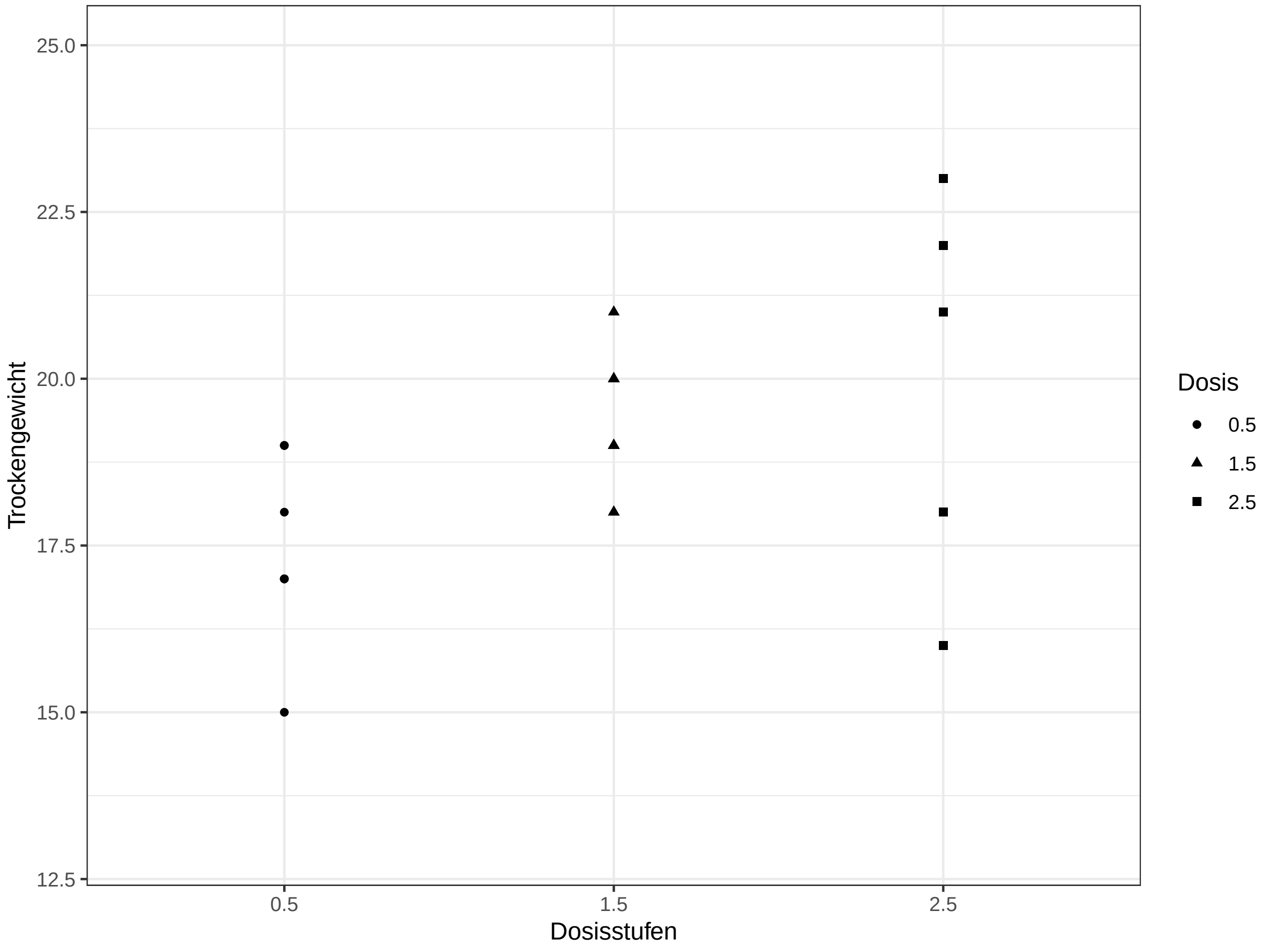
E ☐ Der gepaarte t-Test wird gerechnet, wenn die Beobachtungen nicht unabhängig voneinander sind. Wir messen wiederholt an dem gleichen Probanden oder Tier oder Pflanze. Wir bilden die Differenzen um den gepaarten t-Test rechnen zu können.

Rechen- und Textaufgaben

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

51 Aufgabe (8 Punkte)

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



1. Zeichnen Sie folgende statistischen Masszahlen in die Abbildung ein! **(6 Punkte)**
   - Total (grand) mean: β_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: ϵ
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)**

52 Aufgabe (14 Punkte)

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

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

	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
group	2	3.82			
Residuals	27	10.51			

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

group	n	mean	sd
ctrl	10	5.00	0.56
trt1	10	4.65	0.80
trt2	10	5.52	0.46

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

53 Aufgabe (16 Punkte)

Der Datensatz *ToothGrowth* enthält Daten aus einer Studie zur Bewertung der Wirkung von Vitamin C auf das Zahnwachstum bei Meerschweinchen. Der Versuch wurde an 60 Schweinen durchgeführt, wobei jedes Tier eine von drei Vitamin-C-Dosen *dose* (0.5 mg/Tag, 1 mg/Tag und 2 mg/Tag) über eine von zwei Verabreichungsmethoden *supp* erhielt (Orangensaft oder Ascorbinsäure). Die Zahnlänge wurde als normalverteiltes Outcome gemessen.

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

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
supp	1	203.12			
dose	2	2440.86			
supp:dose	2	110.82			
Residuals	54	724.31			

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

54 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 R.

```
## Analysis of Variance Table
##
## Response: drymatter
##           Df   Sum Sq  Mean Sq F value    Pr(>F)
## trt        2  4.75229 2.376144 5.62487 0.0090767 **
## Residuals 27 11.40576 0.422436
## ---
## 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. Berechen Sie den Effektschätzer η^2. Was sagt Ihnen der Wert von η^2 aus? **(2 Punkte)**
4. Skizieren Sie eine Abbildung, der dem obigen Ergebnis der einfaktoriellen ANOVA näherungsweise entspricht! **(2 Punkte)**

55 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 75 Mastschweinen durchgeführt, wobei jedes Tier eine von drei Selen-Dosen (0.5 ng/Tag, 1 ng/Tag und 5 ng/Tag) über eine von zwei Verabreichungsmethoden erhielt (Wasser oder Festnahrung). Das Gewicht wurde als normalverteiltes Outcome gemessen. Sie erhalten folgenden Output in .

```
## Analysis of Variance Table
##
## Response: len
##           Df  Sum Sq  Mean Sq   F value                 Pr(>F)
## supp       1  133.19  133.186  10.14720              0.0021705 **
## dose       2 3433.82 1716.908 130.80772 < 0.0000000000000000222 ***
## supp:dose  2  103.38   51.688   3.93799              0.0240168 *
## Residuals 69  905.65   13.125
## ---
## 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 dose auf! **(2 Punkte)**
2. Interpretieren Sie das Ergebnis der zweifaktoriellen ANOVA. Gehen Sie im besonderen auf den Term $supp : dose$ ein! **(3 Punkte)**
3. Zeichnen Sie eine Abbildung, der dem obigen Ergebnis der zweifaktoriellen ANOVA näherungsweise entspricht! **(3 Punkte)**

56 Aufgabe (8 Punkte)

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

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

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

58 Aufgabe (8 Punkte)

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

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

59 Aufgabe (9 Punkte)

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

water_type	freshmatter
A	15
A	19
C	11
A	13
B	33
A	16
B	30
C	8
C	12
A	15
C	8
B	27

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

60 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	16.4
B	13.3
A	11.4
B	8.1
A	14.5
A	13.2
B	11.6
A	8.7
B	13.9
A	12.7
B	10.2
A	9.9
A	12.6
A	16.9
A	13.2
A	14.9
A	18.0
B	11.1

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)**

61 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	23
low	20
low	26
low	17
low	22
low	24
low	21
low	20
low	20
low	27
low	23

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 Normalveteilung vorliegt. Begründen Sie Ihre Antwort. **(2 Punkte)**

62 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)**

63 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 $n = 142$ Pflanzen untersucht.

	Erkrankt (ja)	**Erkrankt (nein)**	
Pestizid (ja)	56	21	
Pestizid (nein)	13	52	

1. Ergänzen Sie die Tabelle um die Randsummen! **(1 Punkt)**
2. Formulieren Sie die Fragestellung! **(1 Punkt)**
3. Formulieren Sie das Hypothesenpaar! **(2 Punkte)**
4. Berechnen Sie die Teststatistik eines Chi-Quadrat-Test auf der 2x2 Tafel. Geben Sie Formeln und Rechenweg mit an! **(4 Punkte)**
5. Treffen Sie eine Entscheidung im Bezug zu der Nullhypothese gegeben einem $T_k = 3.841$! **(1 Punkt)**
6. Skizzieren Sie eine 2x2 Tabelle mit $n = 30$ Pflanzen in dem *vermutlich* die Nullhypothese nicht abgelehnt werden kann! **(1 Punkt)**

64 Aufgabe (7 Punkte)

Gegeben sind folgende Randsummen in einer 2x2 Kreuztabelle aus einem Experiment mit $n =$ 138 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)			63
Medikament (nein)			75
	87	51	138

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

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

   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)**

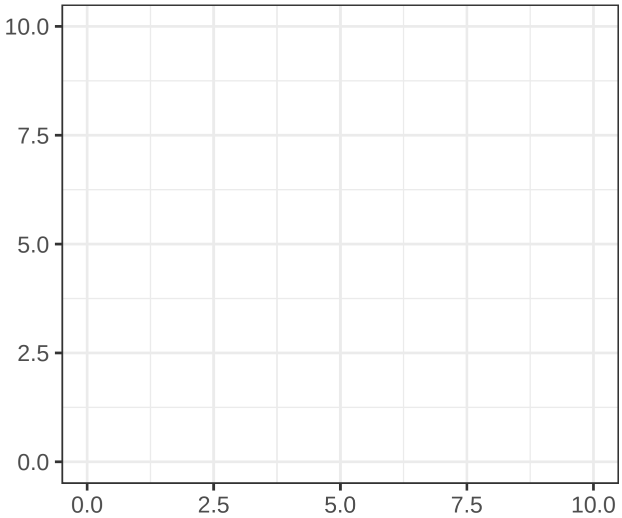
65 Aufgabe (9 Punkte)

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

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

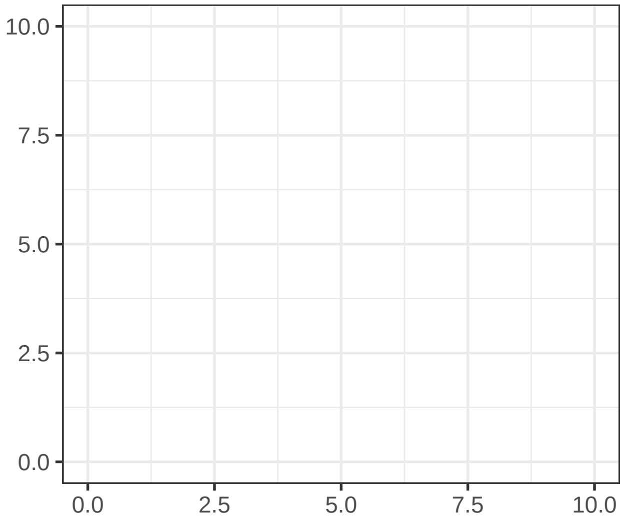
Pearsons $\rho = 1$

$R^2 = 0.75$



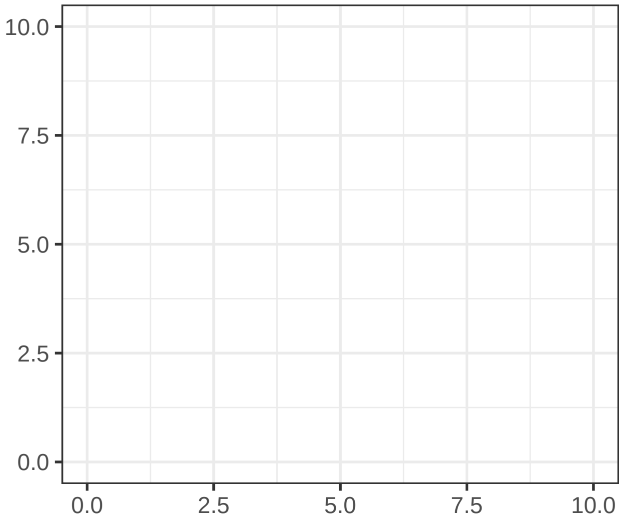
Pearsons $\rho = 0.5$

$R^2 = 1$



Pearsons $\rho = -0.5$

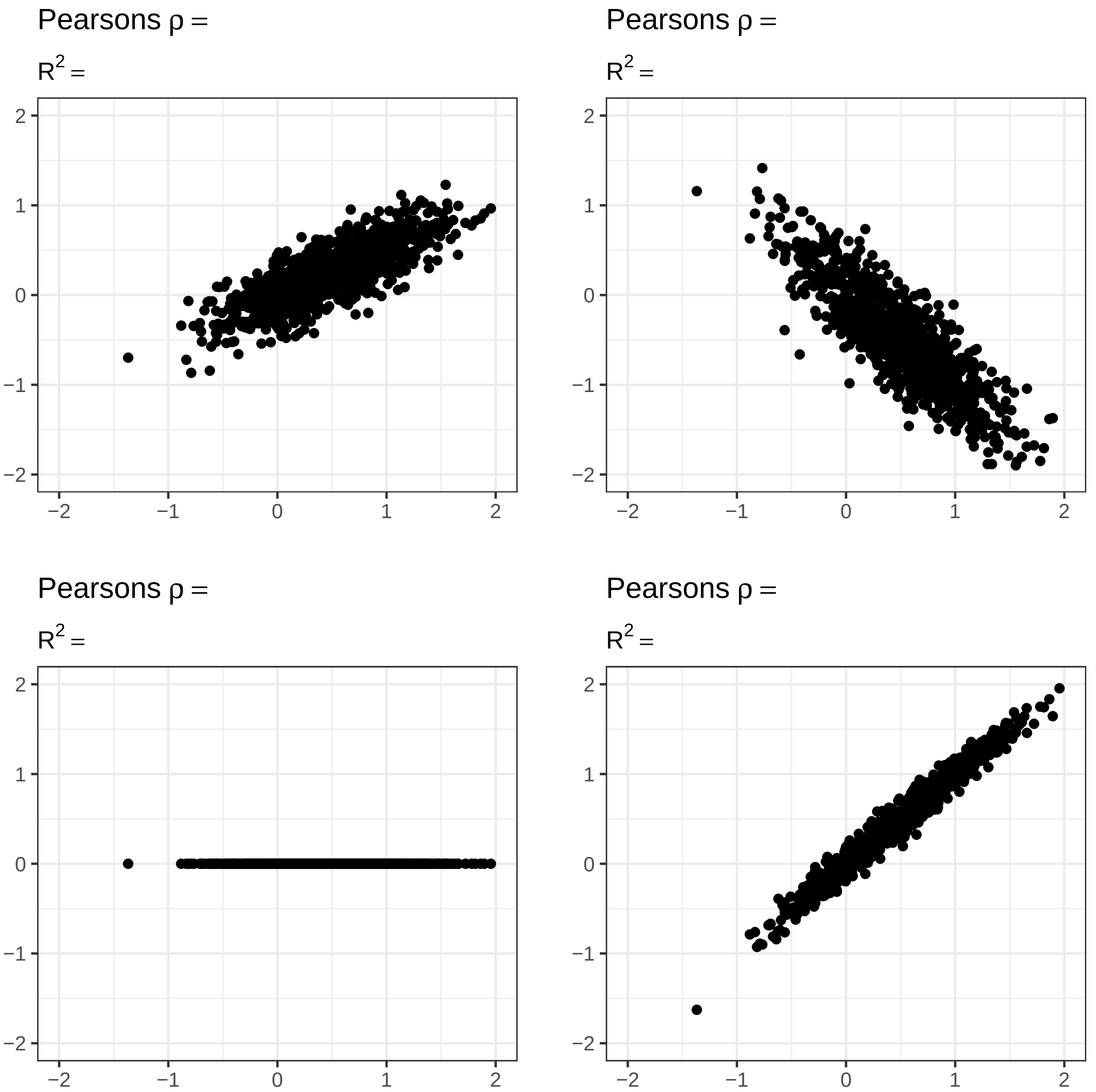
$R^2 = 0.5$



66 Aufgabe (10 Punkte)

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

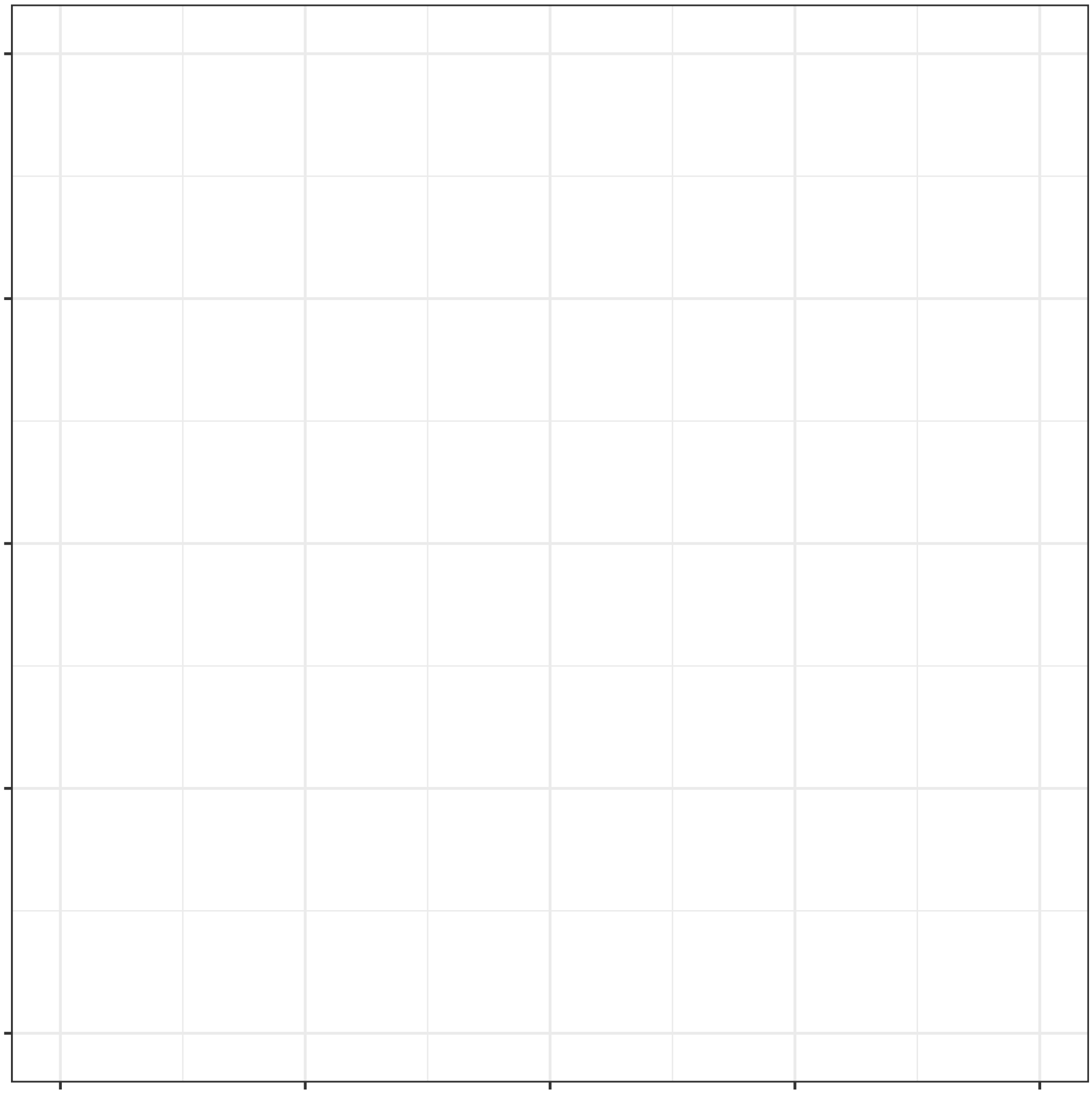
1. Schätzen Sie die ρ-Werte in der entsprechenden Abbildung! **(4 Punkte)**
2. Schätzen Sie die R^2-Werte in der entsprechenden Punktewolke um die Gerade! **(4 Punkte)**
3. Sie rechnen ein statistisches Modell. Was sagen Ihnen die R^2-Werte über das jeweilige Modell? **(2 Punkte)**



67 Aufgabe (8 Punkte)

Sie haben folgende Zahlenreihe y vorliegen $y = \{9, 8, 16, 11, 8, 15\}$.

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)**



68 Aufgabe (10 Punkte)

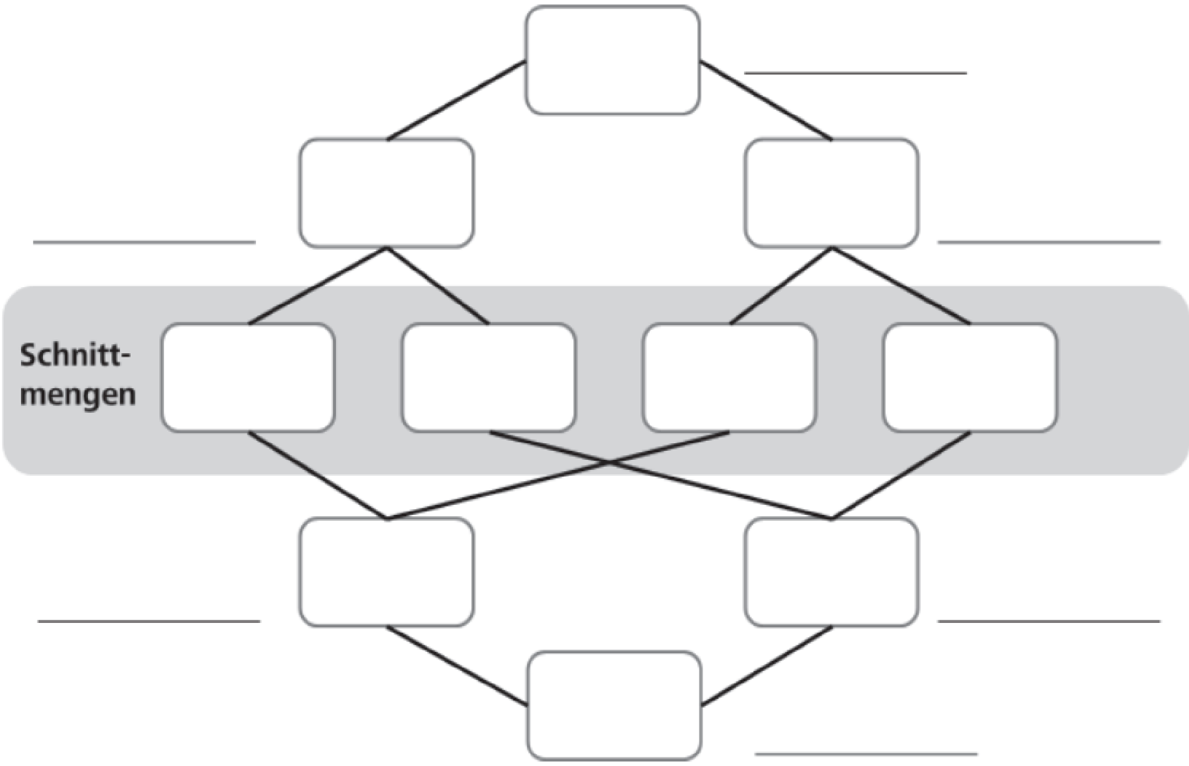
Sie haben folgende Zahlenreihe y vorliegen $y = \{17, 15, 16, 18, 18, 20, 15, 21, 17, 17, 17\}$. Berechnen Sie folgende deskriptive Maßzahlen. Geben Sie Formeln und Rechenwege mit an!

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

69 Aufgabe (11 Punkte)

Die Prävalenz von Klauenseuche bei Wollschweinen wird mit 3% angenommen. In 85% 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 2000 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)**



70 Aufgabe (9 Punkte)

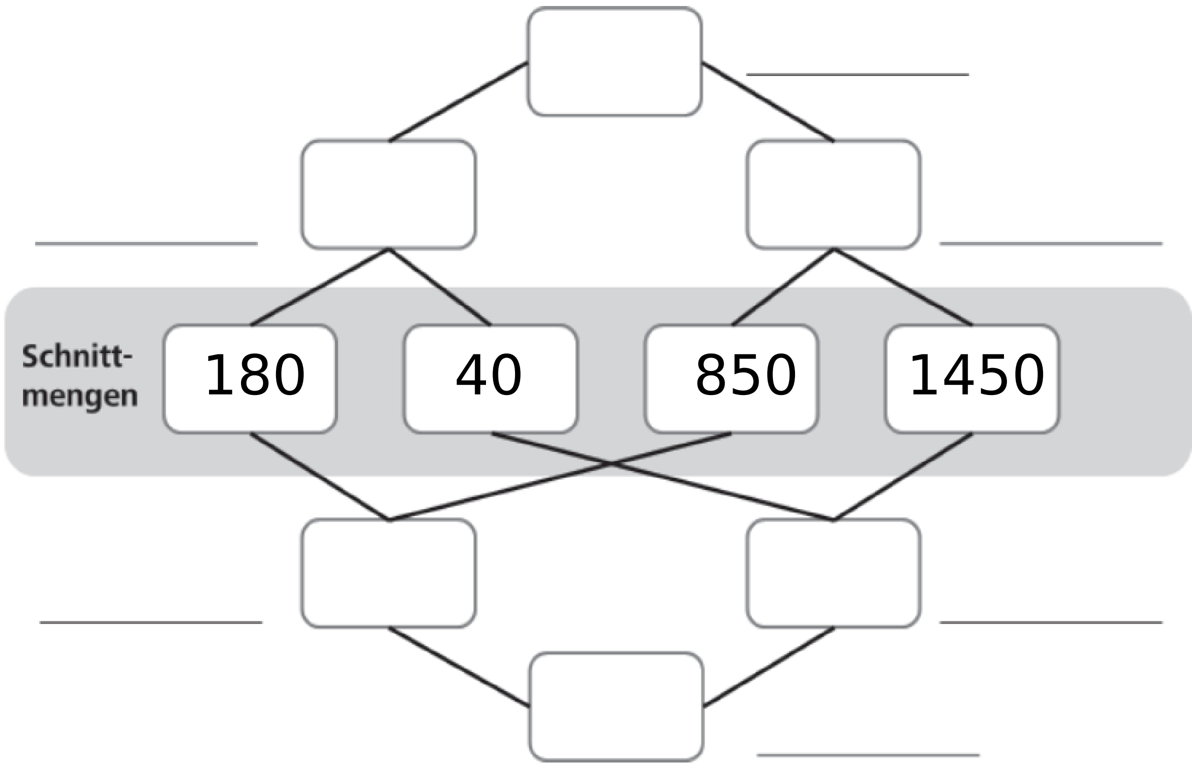
Beim diagnostischen Testen erhalten Sie *True Positives (TP)*, *True Negatives (TN)*, *False Positives (FP)* und *False Negatives (FN)*. Erklären Sie den Zusammenhang wir folgt.

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 Kreuztablle ein. Beschriften Sie die Tabelle entsprechend! **(3 Punkte)**

71 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 Sensifität und Spezifität des diagnostischen Tests für Klauenseuche! Erstellen Sie dafür zunächst eine 2x2 Kreuztabelle aus dem ausgefüllten Doppelbaum! **(4 Punkte)**



72 Aufgabe (7 Punkte)

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

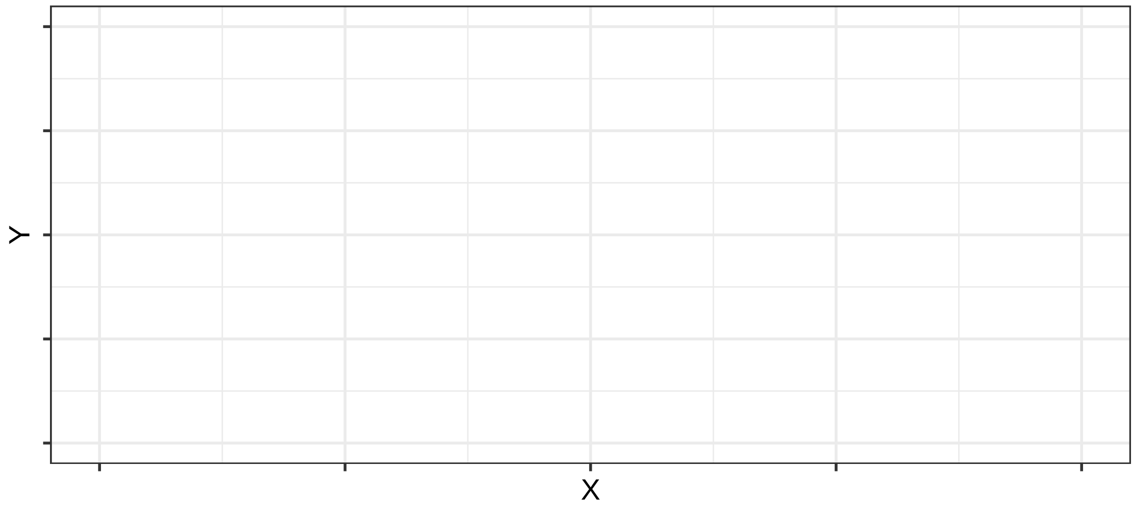
pesticide	grade
ctrl	5
zeroX	1
ctrl	8
roundUp	5
roundUp	4
zeroX	0
pestKill	4
pestKill	2
roundUp	7
zeroX	1
pestKill	3
zeroX	-1
pestKill	5
ctrl	4

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

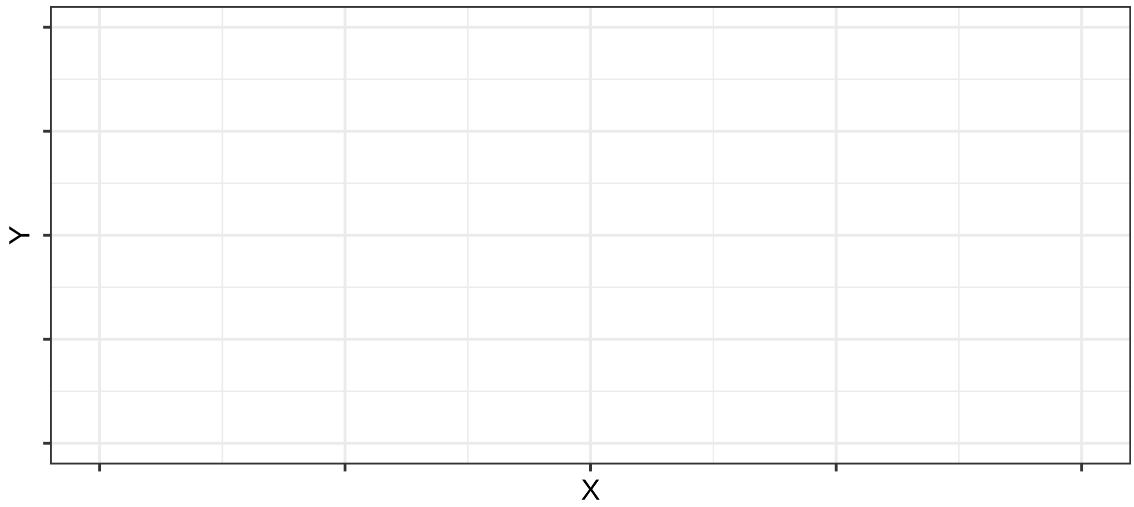
73 Aufgabe (6 Punkte)

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

N(3, 1) und N(4, 2)



Pois(5)



74 Aufgabe **(6 Punkte)**

1. Skizieren Sie 3 Normalverteilungen *in einer Abbildung* mit $\mu_1 \neq \mu_2 \neq \mu_3$ und $s_1 = s_2 = s_3$! **(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)**

75 Aufgabe (6 Punkte)

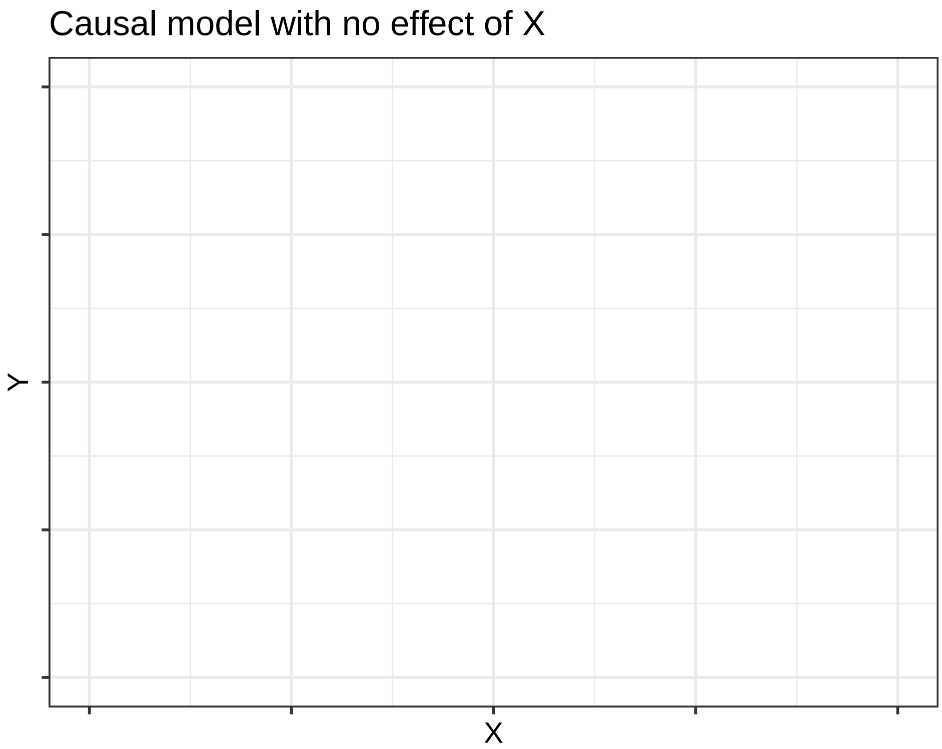
Nach einem Experiment zählen Sie folgende Anzahl an Läsionen auf den Beobachtungen nach einer durchgestandenen Infektion.

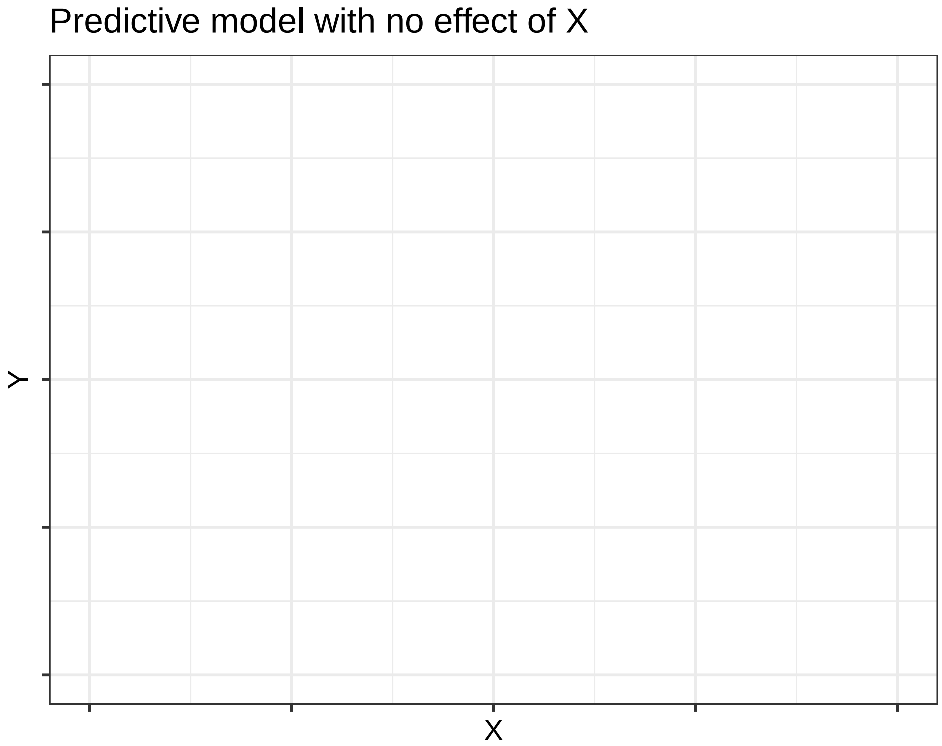
$$5, 5, 5, 6, 3, 6, 2, 2, 2, 2, 6, 3$$

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

76 Aufgabe (6 Punkte)

1. Skizieren Sie in die unten stehenden, freien Abbildungen ein kausales und ein prädiktives Modell mit $n = 9$ Beobachtungen! **(4 Punkte)**
2. Beachten Sie bei der Erstellung der Skizze, ob ein Effekt von X vorliegt oder nicht! **(2 Punkte)**





77 Aufgabe (9 Punkte)

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

1. Wählen Sie ein Outcome was zu der Verteilungsfamilie *Gaussian* gehört! **(1 Punkt)**

2. Schreiben Sie das Modell in der Form $y \sim x$ wie es in R üblich ist *ohne Interaktionsterm*! **(2 Punkte)**

3. Schreiben Sie das Modell in der Form $y \sim x$ wie es in R üblich ist und ergänzen Sie *einen* Interaktionsterm nach Wahl! **(2 Punkte)**

4. Zeichen 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)**

78 Aufgabe (8 Punkte)

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

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

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

79 Aufgabe (6 Punkte)

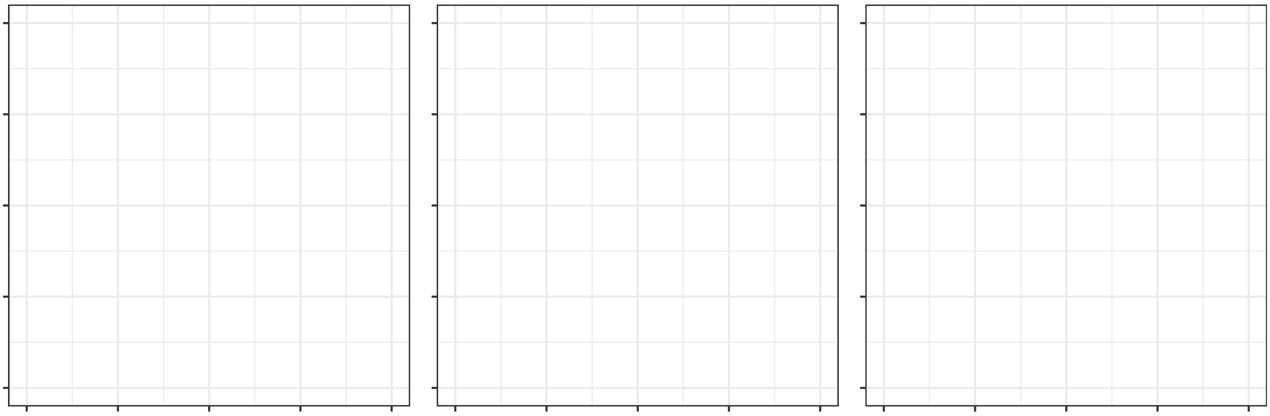
In einem Experiment zur Dosiswirkung wurden verschiedene Dosisstufen mit einer Kontrollgruppe vergleichen. Es wurden vier t-Test für den Mittelwertsvergleich gerechnet und es ergab sich folgende Tabelle mit den rohen p-Werten.

Vergleich	Raw p-val	Adjusted p-val	Reject H_0
dose 10 - ctrl	0.001		
dose 15 - ctrl	0.030		
dose 20 - ctrl	0.020		
dose 40 - ctrl	0.060		

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

80 Aufgabe (10 Punkte)

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

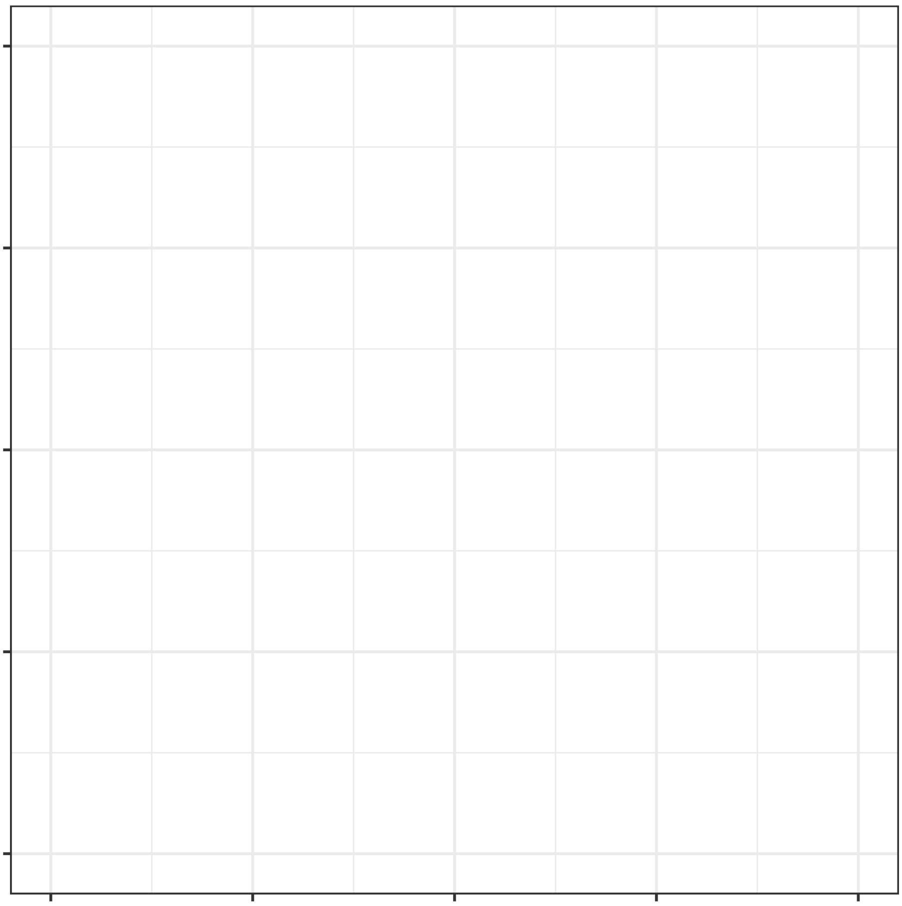


81 Aufgabe (10 Punkte)

In einem Stallexperiment mit $n = 98$ Ferkeln wurde der Gewichtszuwachs unter bestimmten Lichtverhältnissen gemessen. Sie erhalten den R Output der Funktion `tidy()` einer simplen Gaussian linearen Regression sieben Wochen nach der ersten Messung.

term	estimate	std.error	t statistic	p-value
(Intercept)	23.80	1.27		
light	1.61	0.13		

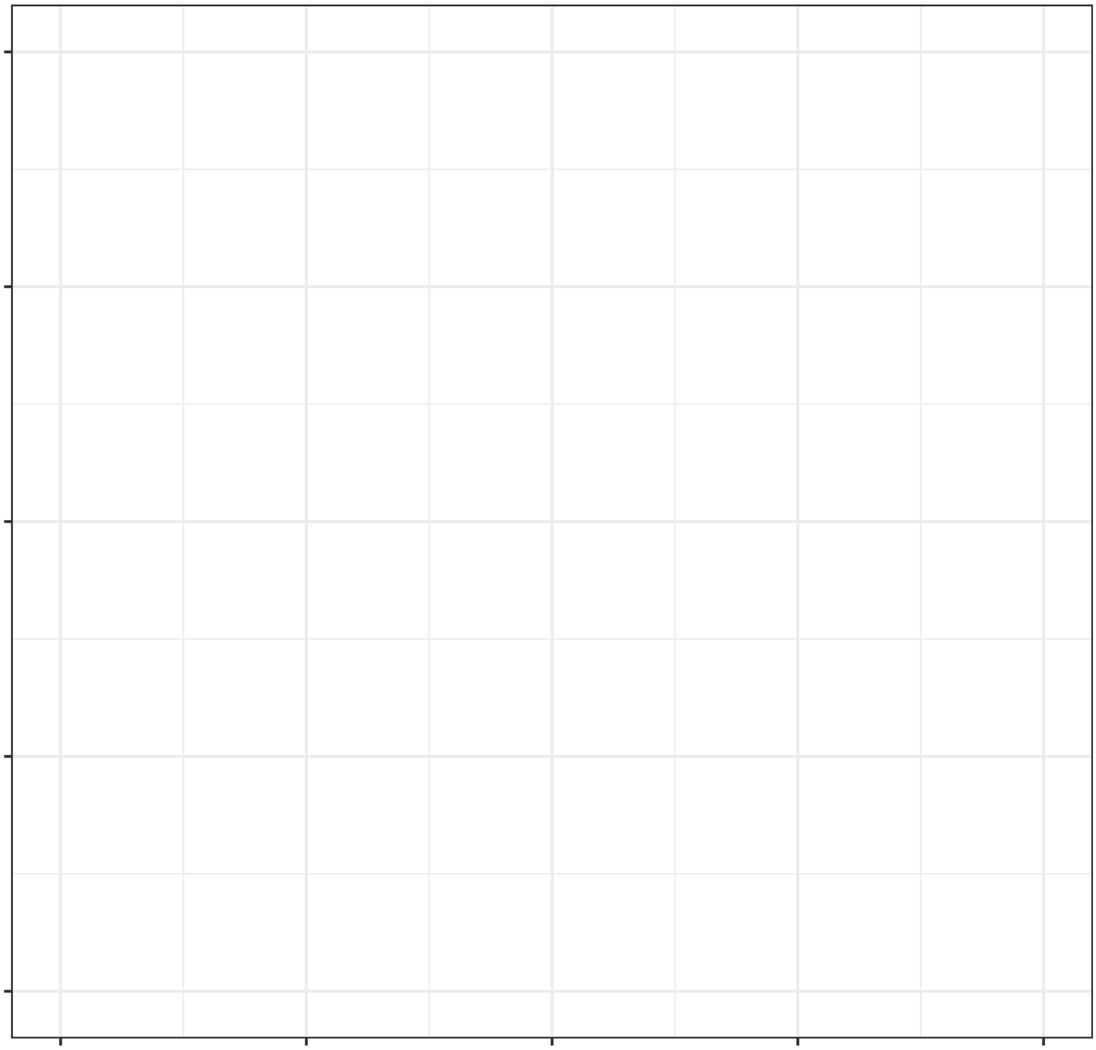
1. Berechnen Sie die t Statistik für *(Intercept)* und *light*! **(2 Punkte)**
2. Schätzen Sie den p-Wert für *(Intercept)* und *light* mit $T_k = 1.96$ ab. Was sagt Ihnen der p-Wert aus? Begründen Sie Ihre Antwort! **(3 Punkte)**
3. Zeichnen Sie die Gerade aus der obigen Tabelle in die untenstehende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



82 Aufgabe (4 Punkte)

1. Skizieren 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)**

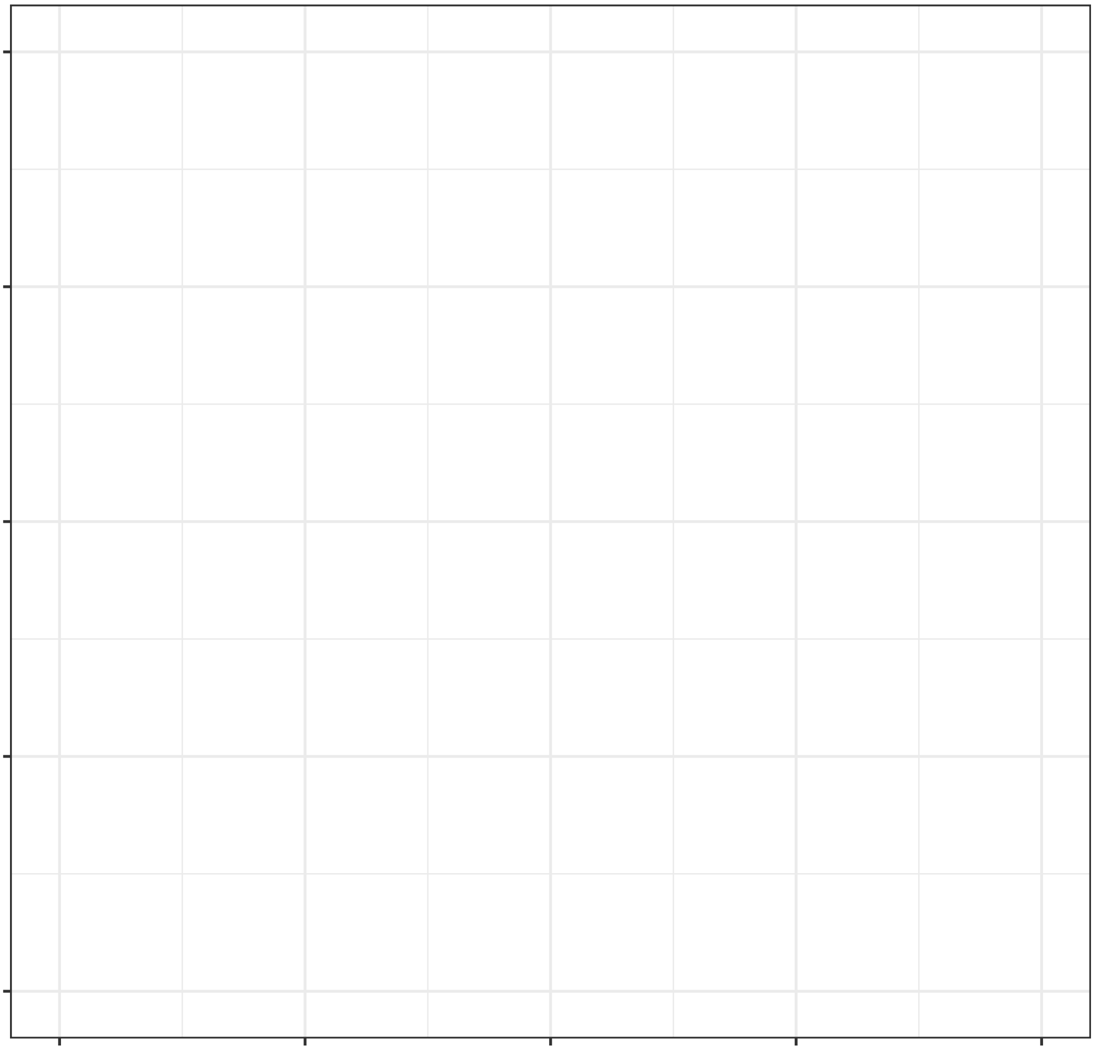
Residual plot with 1 outlier fullfiling the normality assumption.



83 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 fullfiling the normality assumption.



84 Aufgabe (10 Punkte)

Sie rechnen eine lineare Regression um nach einem Feldexperiment den Zusammenhang zwischen Trockengewicht kg/m^2 (*drymatter*) und Wassergabe l/m^2 (*water*) bei Spargel zu bestimmen. Sie erhalten folgende Datentabelle.

.id	drymatter	water	.fitted	.resid
1	31.8	14.8	35.0	
2	29.9	9.5	26.2	
3	20.3	7.5	22.7	
4	16.2	5.3	18.9	
5	30.9	11.3	29.2	
6	26.3	8.3	24.1	
7	31.7	12.4	31.1	
8	20.4	6.5	21.1	
9	24.7	8.8	24.9	
10	19.3	5.0	18.5	
11	30.6	12.2	30.7	

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)**

85 Aufgabe (8 Punkte)

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

stream	longnose	acerage	maxdepth	so4	no3
SAVAGE_R	106	29708	73	12.28	0.63
BEAR_BR	12	3333	83	7.74	5.34
BIG_BR	72	4790	91	5.65	4.10
DEER_CR	26	8297	60	6.36	5.26

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
## -13.8810   -3.4427   -1.6165    1.2162   18.0176
##
## Coefficients:
##                 Estimate   Std. Error t value  Pr(>|t|)
## (Intercept)  1.872650364  0.529073514  3.5395 0.0008893 ***
## acerage      0.000044470  0.000012828  3.4667 0.0011062 **
## maxdepth     0.012787949  0.004038208  3.1667 0.0026524 **
## so4         -0.032259022  0.026659580 -1.2100 0.2320707
## no3          0.214003046  0.071351623  2.9993 0.0042444 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 31.803522)
##
##     Null deviance: 2431.88  on 53  degrees of freedom
## Residual deviance: 1365.45  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)**

86 Aufgabe (10 Punkte)

Auf einer Erdbeerplantage treten unerwartet häufig infizierte Erdbeerpflanzen auf. In einem Versuch sollen verschiedende 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	9.14	0	1
1	10.93	0	1
1	10.73	0	3
1	8.54	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)     1.52       2.55  0.0107
## 2 water           0.146     -2.32  0.0206
## 3 nematodes       0.127     -1.31  0.191
```

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)**

87 Aufgabe (7 Punkte)

Maispflanzen sollen auf die ertragssteigerende Wirkung von verschiedenen Einflussfaktoren untersucht werden. Gemessen wurde als Outcome die Trockenmasse in kg/m^2. Dafür wurde für jede Maispflanze gemessen wieviel Wasser (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 (μg/kg) des Bodens. Es ergibt sich folgender Auszug aus den Daten.

water	light	P	Fe	drymatter	nematodes
8.36	1	8.87	102.20	72.95	8
8.91	0	10.00	96.98	68.19	8
9.65	0	8.79	104.34	74.77	5
7.84	0	10.69	99.87	71.16	11

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

```
##
## Call:
## lm(formula = reformulate(response = "drymatter", termlabels = wanted_vec),
##     data = data_tbl)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.42946 -1.54872 -0.04124  1.48202  4.50355
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 10.700641   6.358815  1.6828              0.09566 .
## water       -0.022390   0.136798 -0.1637              0.87034
## nematodes    0.034837   0.066903  0.5207              0.60377
## Fe           0.589830   0.063699  9.2596 0.0000000000000005798 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.9822 on 96 degrees of freedom
## Multiple R-squared:  0.48614,Adjusted R-squared:  0.47008
## F-statistic: 30.274 on 3 and 96 DF,  p-value: 0.000000000000073289
```

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)**

88 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
## -3.11111 -2.11111 -0.61111  1.13095  5.88889
##
## Coefficients:
##             Estimate Std. Error t value          Pr(>|t|)
## (Intercept) 20.11111    0.86256 23.3155 0.000000000001332 ***
## trtB        -5.96825    1.30407 -4.5766         0.0004312 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.5877 on 14 degrees of freedom
## Multiple R-squared:  0.59938,Adjusted R-squared:  0.57076
## F-statistic: 20.945 on 1 and 14 DF,  p-value: 0.00043124
```

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

89 Aufgabe (11 Punkte)

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: gain ~ attitude + (1 | subject) + (1 | barn)
##    Data: data_tbl
##
## REML criterion at convergence: 793.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.44818 -0.60954 -0.08330  0.47204  3.44971
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  barn     (Intercept)  225.12  15.004
##  subject  (Intercept) 4020.40  63.407
##  Residual              645.42  25.405
## Number of obs: 83, groups:  barn, 7; subject, 6
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 202.5110    26.7879  7.5598
## attitudepop -20.0186     5.5821 -3.5862
##
## Correlation of Fixed Effects:
##             (Intr)
## attitudepop -0.103
```

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

90 Aufgabe (9 Punkte)

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

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

treatment	mean	median	min	max	quartile_1st	quartile_3rd	var
A	17.92	18.0	14	21	17.00	19.0	3.90
B	13.08	12.5	6	23	10.75	15.0	17.90
C	13.67	13.5	9	21	11.75	14.5	9.52
D	13.50	13.5	4	20	9.75	17.5	26.27

   Beschriften Sie die Achsen entsprechend! **(4 Punkte)**

2. Berechnen Sie die benötigten Fallzahlen für die Gruppenvergleiche mit

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

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

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

91 Aufgabe (10 Punkte)

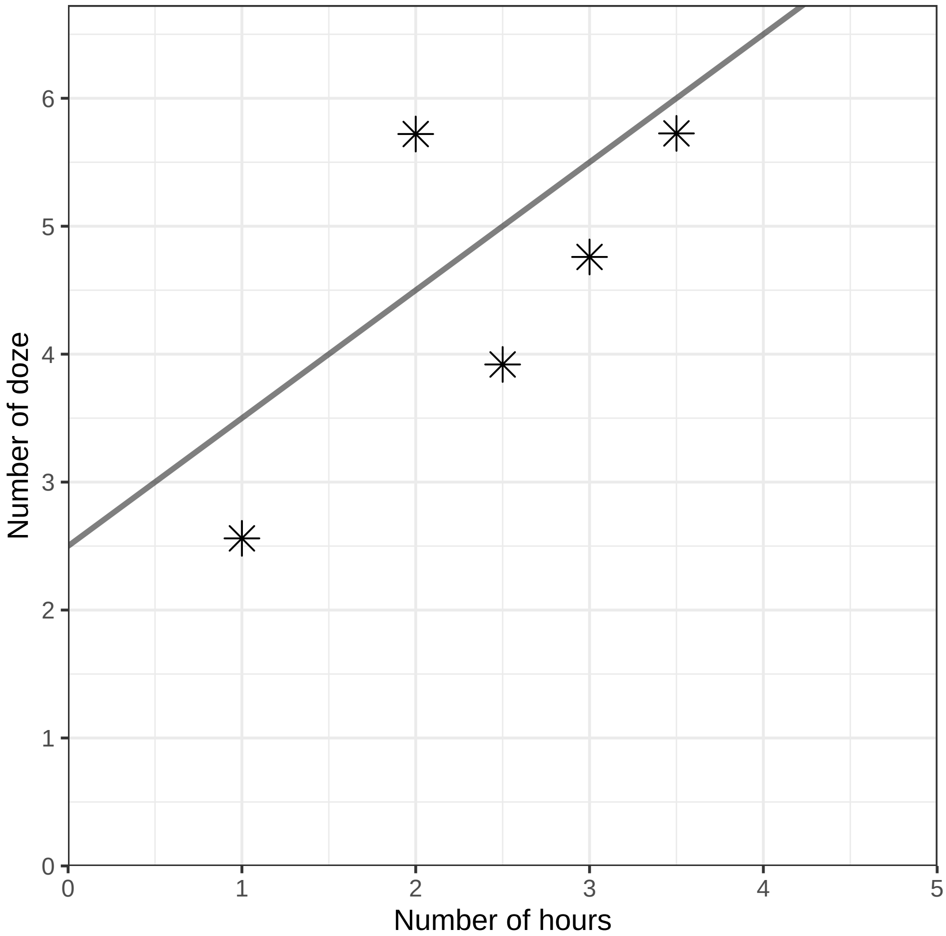
In einem Feldexperiment für die Bodendurchlässigkeit wurde der Niederschlag pro Parzelle sowie der durchschnittliche Ertrag gemessen. Es ergibt sich folgende Datentabelle.

water	drymatter
13	9
34	16
20	8
27	13
31	9

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)**

92 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.



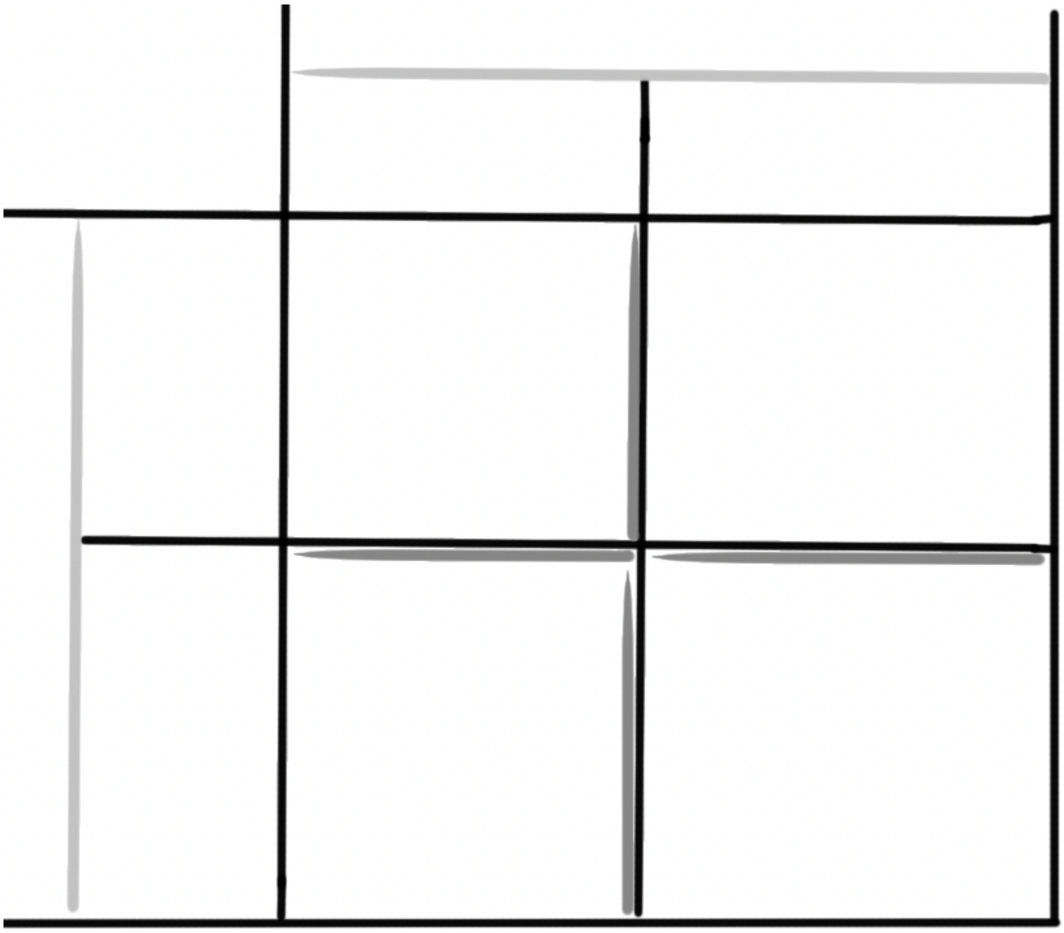
1. Erstellen Sie die Regressionsgleichung aus der obigen Abbildung in der Form $y \sim \beta_0 + \beta_1 \cdot x$! **(2 Punkte)**
2. Beschriften Sie die Gerade mit den Parametern der linearen Regressionsgleichung! **(2 Punkte)**
3. Liegt ein Zusammenhang zwischen der Anzahl an gefahrenen Runden und der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

93 Aufgabe (6 Punkte)

1. Erklären Sie den Zusammenhang zwischen Stichprobe und Grundgesamtheit an einem Schaubild! **(3 Punkte)**
2. Was ist der Unterschied zwischen μ und σ 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)**

94 Aufgabe (6 Punkte)

Geben ist folgende 2x2 Kreuztabelle.

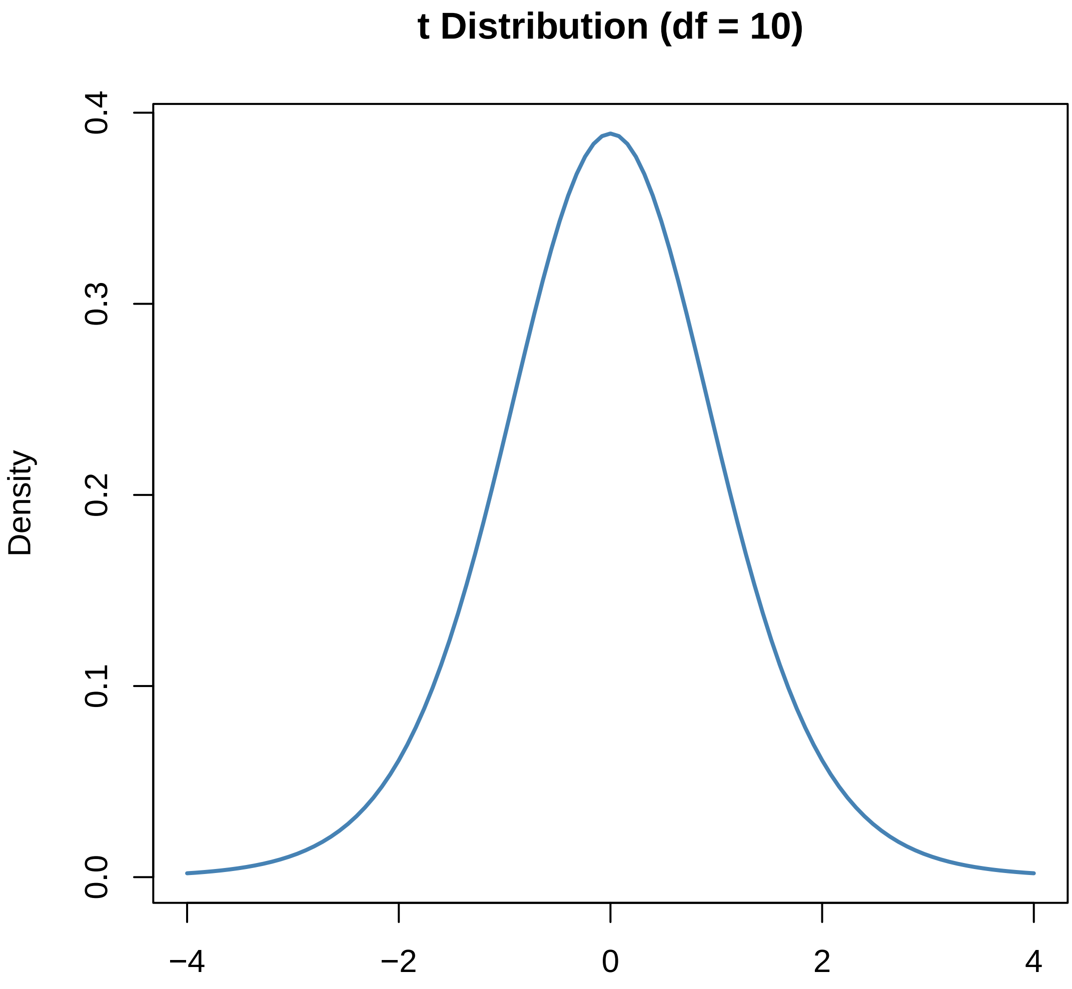


1. Tragen Sie folgende Fachbegriffe korrekt in die 2x2 Kreuztabelle ein! **(6 Punkte)**
   - (Unbekannte) Wahrheit
   - H_0 wahr
   - H_0 falsch
   - H_0 abgelehnt
   - H_0 beibehalten
   - Testentscheidung
   - α-Fehler
   - β-Fehler
   - Richtige Entscheidung
   - 5%
   - 20%

95 Aufgabe (6 Punkte)

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

1. Zeichnen Sie das einseitige α Niveau in die Abbildung! **(2 Punkte)**
2. Zeichnen Sie einen nicht signifikant p-Wert in die Abbildung! **(1 Punkt)**
3. Ergänzen Sie „$\bar{x}_1 = \bar{x}_2$"! **(1 Punkt)**
4. Ergänzen Sie „H_0 ist wahr"! **(1 Punkt)**
5. Ergänzen Sie eine t-Verteilung mit 100 Freiheitsgraden. Zeichnen Sie im Zweifel über den Rand der Abbildung! **(1 Punkte)**



96 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. Skizieren 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 signifikantes, nicht relevantes Konfidenzintervall
   (e) Ein Konfidenzintervall mit höhererFallzahl *n* in der Stichprobe als der Rest der Konfidenzintervalle

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

98 Aufgabe (8 Punkte)

Sie haben folgende Aussage gegeben.

Bin ich im Winter?

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)**

99 Aufgabe (8 Punkte)

»Ohne ausreichende Zufuhr von Vitamin C stellen sich nach 67 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. Schiffarzt 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)**

100 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? Skizieren 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)**

101 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	15
GoneEx	20
RoundUp	17
RoundUp	16
RoundUp	17
RoundUp	17
RoundUp	17
GoneEx	16
RoundUp	18
GoneEx	16
GoneEx	17
RoundUp	15
GoneEx	15
GoneEx	16

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)**

102 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	13	14
2	18	22
3	14	12
4	19	18
5	13	12
6	12	15
7	15	8
8	14	17
9	13	20

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)**

103 Aufgabe (10 Punkte)

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

```
##
##  Two Sample t-test
##
## data:  freshmatter by N
## t = 1.58756, df = 12, p-value = 0.13837
## alternative hypothesis: true difference in means between group high and group low is no
## 95 percent confidence interval:
##  -0.9044682  5.7616111
## sample estimates:
## mean in group mid mean in group low
##         17.142857         14.714286
```

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

104 Aufgabe (8 Punkte)

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

```
##
##  Two Sample t-test
##
## data:  freshmatter by N
## t = 0.461447, df = 10, p-value = 0.65435
## alternative hypothesis: true difference in means between group high and group low is no
## 95 percent confidence interval:
##  -1.9689872  2.9975587
## sample estimates:
## mean in group high mean in group trt1
##          18.800000          18.285714
```

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

105 Aufgabe (8 Punkte)

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

```
##
##  Welch t-Test
##
## data:  drymatter by Fe
## t = -3.61177, df = 12, p-value = 0.003568
## alternative hypothesis: true difference in means between group high and group low is no
## 95 percent confidence interval:
##  -12.1389176  -3.0039395
## sample estimates:
## mean in group ctrl mean in group trt2
##          11.571429          19.142857
```

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

106 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
13	13					
11	10					
10	10					
11	12					
8	14					
11	15					
9	10					
11	12					
10	12					
11	12					
9	10					
10	13					
9	8					
11	13					
14	9					

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

107 Aufgabe (8 Punkte)

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

Treatment	Count
RootsGoneX	9.0
RootsGoneX	9.6
RootsGoneX	8.0
RootsGoneX	8.3
RootsGoneX	11.4
RootsGoneX	13.9
RootsGoneX	10.1
Kontrolle	11.2
Kontrolle	11.0
Kontrolle	15.1
Kontrolle	11.6
Kontrolle	8.4
Kontrolle	13.1
Kontrolle	15.7

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

1. Bestimmen Sie hierfür U_c mit $U_c = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$! **(4 Punkte)**
2. Geben Sie eine Aussage über die Signifikanz von U_c durch $z = \frac{U_c - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$ und dem kritischen Wert von $z = 1.96$. Begründen Sie Ihre Antwort! **(2 Punkte)**
3. Berechnen Sie die Effektstärke mit $r = |\frac{z}{\sqrt{n}}|$ und interpretieren Sie die Effektstärke! **(2 Punkte)**

108 Aufgabe (9 Punkte)

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

Treatment	Count	Rang Kontrolle	Rang Dosis 20	Rang Dosis 40
Kontrolle	9.5			
Dosis 40	10.2			
Dosis 20	10.5			
Dosis 40	11.3			
Kontrolle	11.4			
Dosis 20	11.7			
Dosis 40	11.8			
Kontrolle	11.9			
Dosis 20	12.0			
Kontrolle	12.2			
Dosis 20	12.2			
Dosis 40	12.2			
Dosis 20	13.5			
Kontrolle	13.8			
Kontrolle	15.1			
Kontrolle	17.4			
Dosis 20	17.4			
Dosis 40	18.0			

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

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