

## Checkbox für die Version vom 5. April 2022

Die gesamte Klausur beinhaltet aktuell in Summe **65** Fragen.

Davon sind **33** Multiple Choice Fragen sowie **32** Rechen- und Textaufgaben.

## Frequently asked questions (FAQ)

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

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

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

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

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

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

**Solange kann ich nicht warten...** Dann einfach eine Mail an mich schreiben:

[j.kruppa@hs-onsabrueck.de](mailto:j.kruppa@hs-onsabrueck.de)

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

**B.Sc. Bioverfahrenstechnik in Agrar- und Lebensmittelwirtschaft, B.Sc. Landwirtschaft, B.Eng. Wirtschaftsingenieurwesen im Agri- und Hortibusiness, B.Sc. Angewandte Pflanzenbiologie - Gartenbau, Pflanzentechnologie, B.Eng. Wirtschaftsingenieurwesen im Agri- und Hortibusiness, M.Sc. Angewandte Nutztier- und Pflanzenwissenschaften**

# **Klausurfragen der (Bio) Data Science**

**Hochschule Osnabrück**

Prüfer: Prof. Dr. Jochen Kruppa  
Fakultät für Agrarwissenschaften und Landschaftsarchitektur  
[j.kruppa@hs-osnabrueck.de](mailto:j.kruppa@hs-osnabrueck.de)

Version vom 5. April 2022

## Ergebnis der Klausur

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

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

\_\_\_\_\_ von 80 Punkten in Summe.

Es wird folgender Notenschlüssel angewendet.

<b>Punkte</b>	<b>Note</b>
78 - 80	1,0
75 - 77	1,3
70 - 74	1,7
65 - 69	2,0
59 - 64	2,3
54 - 58	2,7
49 - 53	3,0
44 - 48	3,3
41 - 43	3,7
40	4,0

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

## Multiple Choice Aufgaben

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

	A	B	C	D	E
1 Aufgabe					
2 Aufgabe					
3 Aufgabe					
4 Aufgabe					
5 Aufgabe					
6 Aufgabe					
7 Aufgabe					
8 Aufgabe					
9 Aufgabe					
10 Aufgabe					

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

## 1 Aufgabe

(2 Punkte)

Price et al. (2016) untersuchte die Auswirkungen des Bergbaus und der Talauffüllung auf den Bestand und die Häufigkeit von Bachsalamandern. Um den Effekt zu Berechnen nutze Price et al. (2016) eine Poisson-Regression auf die Anzahl an aufgefundenen Bachsalamandern an den jeweiligen Suchorten. Nach einer statistischen Beratung wurde ihm nahegelegt auf Overdispersion zu achten, wenn er statistische Aussagen zur Signifikanz treffen will. Price et al. (2016) schätzt zwei Modelle. Modell 1 mit einer Poisson Verteilung und Modell 2 mit einer Quasi-Poisson Verteilung. Welche Aussage zu einer geschätzten Overdispersion von 2.62 ist richtig?

### Summary Output der Funktion tidy() von Modell 1 (Poisson):

term	estimate	std.error	statistic	p.value
sppPR	-1.4152	0.2925	-4.8387	0.0000
sppDM	0.2132	0.1832	1.1634	0.2447
sppDF	0.3104	0.1864	1.6653	0.0959

### Summary Output der Funktion tidy() von Modell 2 (Quasi-Poisson):

term	estimate	std.error	statistic	p.value
sppPR	-1.4152	0.4736	-2.9878	0.0030
sppDM	0.2132	0.2967	0.7184	0.4730
sppDF	0.3104	0.3019	1.0283	0.3045

- A** ☐ Bei einer geschätzten Overdispersion höher als 1.5 ist von Overdispersion in den Daten auszugehen. Daher wird die Varianz systematisch unterschätzt, was zu höheren p-Werten führt. Daher gibt es weniger signifikante Ergebnisse als es in Wirklichkeit gibt. Daher ist das Modell 1 die bessere Wahl.
- 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 geschätzten Overdispersion höher als 1.5 ist von keiner Overdispersion in den Daten auszugehen. Dennoch sind die p-Werte zu klein, dass diese p-Werte natürlich entstanden sein könnten. Die p-Werte müssen adjustiert werden.
- D** ☐ Bei einer geschätzten Overdispersion höher als 1.5 ist von Overdispersion in den Daten auszugehen. Daher wird die Varianz systematisch unterschätzt, was zu kleineren p-Werten führt. Daher gibt es mehr signifikante Ergebnisse als es in Wirklichkeit gibt. Daher ist das Modell 2 die bessere Wahl.
- E** ☐ Bei einer geschätzten Overdispersion höher als 1.5 ist von Overdispersion in den Daten auszugehen. Daher wird die Varianz systematisch überschätzt, was zu höheren p-Werten führt. Daher gibt es mehr signifikante Ergebnisse als es in Wirklichkeit gibt. Daher ist das Modell 1 die bessere Wahl

## 2 Aufgabe

(2 Punkte)

Das Falsifikationsprinzip besagt...

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

## 3 Aufgabe

(2 Punkte)

Welche Aussage über die *confounder* Adjustierung ist richtig?

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

## 4 Aufgabe

(2 Punkte)

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

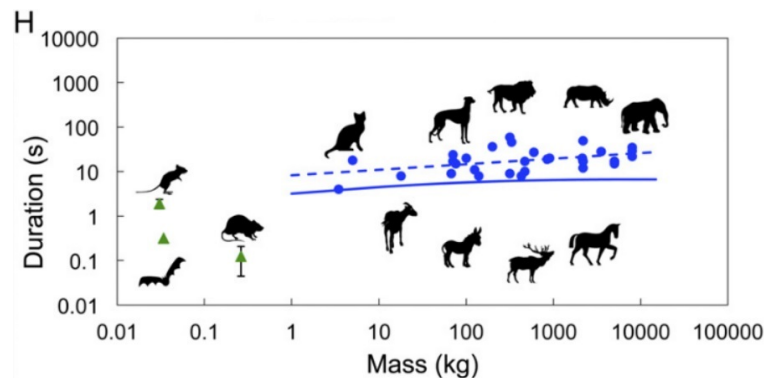
- A** ☐ Randomisierung war bis 1952 bedeutend, wurde dann aber in Folge besserer Rechnerleistung nicht mehr verwendet. Aktuelle Statistik nutzt keine Randomisierung mehr.
- B** ☐ Randomisierung erlaubt erst die Varianzen zu schätzen. Ohne eine Randomisierung ist die Berechnung von Mittelwerten und Varianzen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

- D** ☐ Die summary() Funktion in R kann nur auf randomisierten Daten angewendet werden.
- E** ☐ Randomisierung sorgt für Strukturgleichheit und erlaubt erst von der Stichprobe auf die Grundgesamtheit zurückzuschliessen.

## 5 Aufgabe

(2 Punkte)

Folgende Abbildung ist Yang et al. (2014) *Duration of urination does not change with body size* entnommen.



Welche der folgenden Aussagen ist richtig?

- A** ☐ Es ist kein Zusammenhang zwischen der Körpergrösse und der Dauer des Urinierens zu erkennen. Ohne einen Signifikanztest lassen sich aber keine Aussagen treffen.
- B** ☐ Es ist kein Zusammenhang zwischen der Körpergrösse und der Dauer des Urinierens zu erkennen. Dieser Zusammenhang liegt aber hauptsächlich an der log-Transformation.
- C** ☐ Es ist ein Zusammenhang zu erkennen, dieser wird aber durch die falsche Wahl der Tiere verzerrt.
- D** ☐ Es ist ein Zusammenhang zu erkennen, der aber durch die log-Transformation verzerrt wird.
- E** ☐ Es ist kein Zusammenhang zwischen der Körpergrösse und der Dauer des Urinierens zu erkennen. Die Gerade verläuft parallel zur x-Achse.

## 6 Aufgabe

(2 Punkte)

Berechnen Sie den Mittelwert und Standardabweichung von x mit -1, 8, 12, 8 und 11.

- A** ☐ Es ergibt sich 8.6 +/- 2.565
- B** ☐ Es ergibt sich 6.6 +/- 13.15
- C** ☐ Es ergibt sich 7.6 +/- 5.13
- D** ☐ Es ergibt sich 7.6 +/- 2.565
- E** ☐ Es ergibt sich 7.6 +/- 26.3

## 7 Aufgabe

(2 Punkte)

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

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

## 8 Aufgabe

(2 Punkte)

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

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

## 9 Aufgabe

(2 Punkte)

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

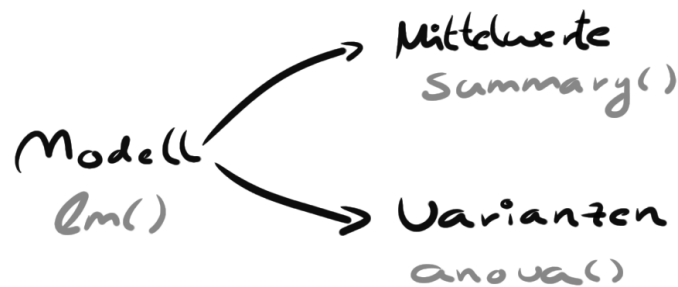
- A ☐ Der p-Wert. Durch den Vergleich mit  $\alpha$  lässt sich über die Signifikanz entscheiden und der  $\beta$ -Fehler erlaubt über die Power eine Einschätzung der Relevanz.
- B ☐ Das  $\Delta$ . Durch die Effektstärke haben wir einen Wert für die Relevanz, die vom Anwender bewertet werden muss. Da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



## 10 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 `summary()` Funktion ist veraltet und wurde durch die ANOVA ersetzt. Grundsätzlich sind die Varianzen mehr aussagekräftig, da die Varianzen die Mittelwerte schon mit beinhaltet. Hier spart man einen Analyseschritt.
- B** ☐ Ein Modell in **R** ist in der Lage parallel die ANOVA sowie auch eine lineare Regression zu rechnen.
- C** ☐ Die Effektschätzer aus einem Modell, in diesem Fall ein lineares Modell unter der Annahme der Poissonverteilung, erlauben es sowohl eine ANOVA zurechnen sowie auch eine Zusammenfassung der Mittelwerte zu betrachten.
- D** ☐ Die Effektschätzer eines `lm()` sind als einziges in der Lage die ANOVA als auch eine lineare Regression zu rechnen. Die Funktion `glm()` erlaubt dies nicht.
- E** ☐ Die Effektschätzer aus einem Modell, in diesem Fall ein lineares Modell unter der Annahme der Normalverteilung, erlauben es sowohl eine ANOVA zurechnen sowie auch eine Zusammenfassung der Mittelwerte zu betrachten.

## 11 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** ☐ Alle Funktionen und auch Anwendungen sind in **R** in englischer Sprache. Die Nutzung von deutschen Wörtern ist nicht schick und das ist zu vermeiden.
- B** ☐ Es gibt keinen Grund nicht auch deutsche Wörter zu verwenden. Es ist ein Stilmittel.
- C** ☐ Programmiersprachen können nur englische Begriffe verarbeiten. Zusätzliche Pakete können zwar geladen werden, aber meist funktionieren diese Pakete nicht richtig. Deutsch ist International nicht bedeutend genug.
- D** ☐ Die Spracherkennung von **R** ist nicht in der Lage Deutsch zu verstehen.
- E** ☐ 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.

## 12 Aufgabe

(2 Punkte)

Berechnen Sie den Median und das IQR von  $x$  mit 0, 14, 28, 25, 28, 23 und 63.

- A** ☐ Es ergibt sich 25 [18.5, 28]
- B** ☐ Es ergibt sich 25 +/- 28
- C** ☐ Es ergibt sich 26 +/- 18.5
- D** ☐ Es ergibt sich 25 +/- 18.5
- E** ☐ Es ergibt sich 26 [18.5, 28]

## 13 Aufgabe

(2 Punkte)

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

- A** ☐ Der Begründer der modernen Statistik, R. Fischer, hat die Grenze simuliert und berechnet. Dadurch ergibt sich dieser optimale Cut-Off.
- B** ☐ Im Rahmen eines langen Disputs zwischen Neyman und Fischer wurde  $\alpha = 5\%$  festgelegt. Leider werden die Randbedingungen und Voraussetzungen an statistische Modelle heute immer wieder ignoriert.
- C** ☐ Auf einer Statistikkonferenz in Genf im Jahre 1942 wurde dieser Cut-Off nach langen Diskussionen festgelegt. Bis heute ist der Cut Off aber umstritten, da wegen dem 2. Weltkrieg viele Wissenschaftler nicht teilnehmen konnten.
- D** ☐ Der Wert ergab sich aus einer Auswertung von 1042 wissenschaftlichen Veröffentlichungen zwischen 1914 und 1948. Der Wert 5% wurde in 28% der Veröffentlichungen genutzt. Daher legte man sich auf diese Zahl fest.
- E** ☐ Die Festlegung von  $\alpha = 5\%$  ist eine Kulturkonstante. Wissenschaftler benötigt eine Schwelle für eine statistische Testentscheidung, der Wert von  $\alpha$  wurde aber historisch mehr zufällig gewählt.

## 14 Aufgabe

(2 Punkte)

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

- A** ☐ Die Poisson-Regression schätzt zwei Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzten Overdispersion von 2.86 liegt keine Overdispersion vor.

- B** ☐ Die Poisson-Regression schätzt drei Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzten Overdispersion von 2.86 liegt Overdispersion vor. Die Lösung ist die Nutzung von nur zwei der drei Verteilungsparameter:  $\gamma_1$  und  $\gamma_3$ .
- C** ☐ Die Poisson-Regression schätzt nur einen Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzten Overdispersion von 2.86 liegt keine Overdispersion vor. Overdispersion liegt vor, wenn die geschätzte Overdispersion unter 1 liegt.
- D** ☐ Die Poisson-Regression schätzt nur einen Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzten Overdispersion von 2.86 liegt Overdispersion vor. Damit kann keine Poisson-Regression gerechnet werden. Die Lösung ist eine Gaussian Regression mit Nullanpassung.
- E** ☐ Die Poisson-Regression schätzt nur einen Verteilungsparameter. Deshalb muss überprüft werden, ob Overdispersion vorliegt. Mit einer geschätzten Overdispersion von 2.86 liegt Overdispersion vor. Die Lösung ist die Nutzung einer anderen Verteilungsfamilie wie die Quasipoisson Verteilung.

## 15 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?

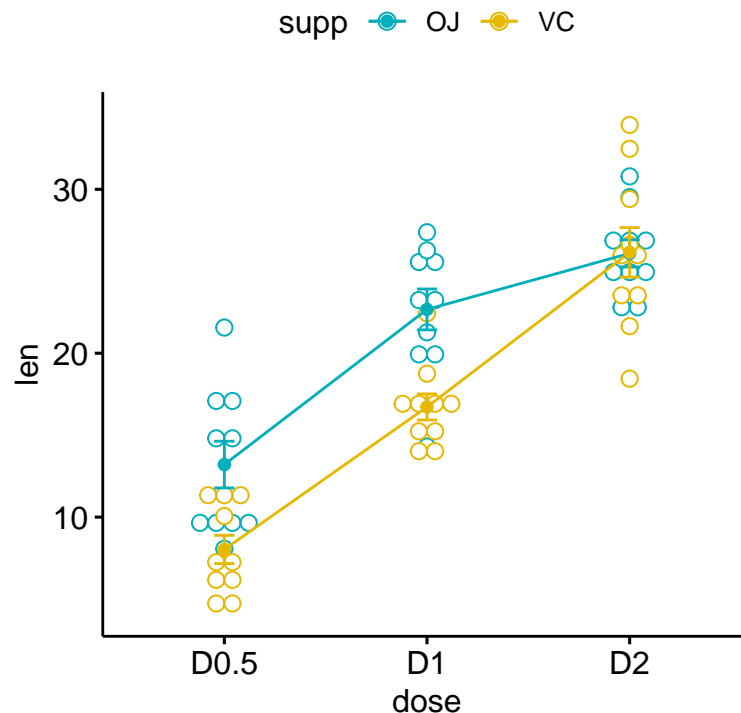
- A** ☐ Aktuell werden immer grössere Datensätze erhoben. Dadurch wird auch die Varianz immer höher was automatisch zu mehr signifikanten Ergebnissen führt.
- B** ☐ Relevanz und Signifikanz haben nichts miteinander zu tun. Daher gibt es auch keinen Zusammenhang zwischen hoher Fallzahl ( $n > 10000$ )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

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

## 17 Aufgabe

(2 Punkte)

Welche Aussage über den Welch t-Test ist richtig?

- A** ☐ Der Welch t-Test vergleicht die Varianz von zwei Gruppen.

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

## 18 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 7 Tiere sind gesund. Welche Aussage über den Risk ratio Effektschätzer ist richtig?

- A** ☐ Es ergibt sich ein Risk ratio von 0.71, da es sich um ein Anteil handelt.
- B** ☐ Es ergibt sich ein Risk ratio von 1.4, da es sich um ein Anteil handelt.
- C** ☐ Es ergibt sich ein Risk ratio von 0.71, da es sich um eine Chancenverhältnis handelt.
- D** ☐ Es ergibt sich ein Risk ratio von 0.42, da es sich um eine Chancenverhältnis handelt.
- E** ☐ Es ergibt sich ein Risk ratio von 0.42, da es sich um ein Anteil handelt.

## 19 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 wird auf einem Trainingsdatensatz trainiert und anschließend über eine explorative Datenanalyse validiert. Signifikanzen über  $\beta_i$  können hier nicht festgestellt werden.
- 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 möchte die Zusammenhänge von X auf Y modellieren. Hierbei geht es um die Effekte von X auf Y. Man sagt, wenn X um 1 ansteigt ändert sich Y um einen Betrag  $\beta$ .
- D** ☐ Ein prädiktives Modell basiert auf einem Trainingsdatensatz und einem Testdatensatz. Auf dem Trainingsdatensatz wird das Modell trainiert und auf dem Testdatensatz validiert.
- E** ☐ Ein prädiktives Modell schließt grundsätzlich lineare Modell aus. Es muss ein Graph gefunden werden, der alle Punkte beinhaltet. Erst dann kann das  $R^2$  berechnet werden.

## 20 Aufgabe

(2 Punkte)

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

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

## 21 Aufgabe

(2 Punkte)

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

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

## 22 Aufgabe

(2 Punkte)


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


- A** ☐  $X$  beinhaltet eine Spalte. Die Spalte gibt dann auch die Verteilungsfamilie vor und ist somit wichtig für die Erstellung von explorativen Abbildungen (EDA).
- B** ☐  $X$  beinhaltet meist eine Matrix. Das heisst, es wird nicht nur eine Spalte berücksichtigt sondern mehrere Spalten aus den Daten  $D$ .
- C** ☐  $X$  ist grundsätzlich immer in einem kausales Modell. Ein prädiktives Modell ist nicht möglich.

- D** ☐ X beschreibt die Prädiktion. Das Modell muss aber noch die Matrixform beinhalten sonst ist eine Unabhängigkeit nicht gewährleistet.
- E** ☐ Die explorative Datenanalyse basiert auf der Transformation von Y auf X durch Summary-Tabellen. Hierbei spielt die Variable eine bedeutende Rolle zur Abschätzung der Varianzkomponenten.

## 23 Aufgabe

(2 Punkte)

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

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

## 24 Aufgabe

(2 Punkte)

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

- A** ☐ Das  $\eta^2$  beschreibt den Anteil der Varianz, der von den Behandlungsbedingungen erklärt wird. Das  $\eta^2$  ist damit mit dem  $R^2$  aus der linearen Regression zu vergleichen.
- B** ☐ Das  $\eta^2$  beschreibt den Anteil der Varianz, der von den Behandlungsbedingungen nicht erklärt wird. Somit der Rest an nicht erklärbarer Varianz.
- C** ☐ Das  $\eta^2$  ist die Korrelation der ANOVA. Mit der Ausnahme, dass 0 der beste Wert ist.
- D** ☐ Das  $\eta^2$  ist ein Wert für die Güte der ANOVA. Je kleiner desto besser. Ein  $\eta^2$  von 0 bedeutet ein perfektes Modell mit keiner Abweichung. Die Varianz ist null.
- E** ☐ Die Berechnung von  $\eta^2$  ist ein Wert für die Interaktion.

## 25 Aufgabe

(2 Punkte)

Welche Aussage über die nicht-parametrische Statistik ist richtig?

- A** ☐ Die nicht-parametrische Statistik ist ein Vorgänger der parametrischen Statistik und wurde wegen dem Mangel an Effektschätzern nicht mehr ab 1960 genutzt.
- 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 basiert auf Rängen. Daher gibt es auch direkt zu interpretierenden Effektschätzer.
- D** ☐ 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.
- E** ☐ Die nicht-parametrische Statistik basiert auf dem Schätzen von Parametern aus einer a priori festgelegten Verteilung. Daher gibt es auch direkt zu interpretierenden Effektschätzer.

## 26 Aufgabe

(2 Punkte)

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

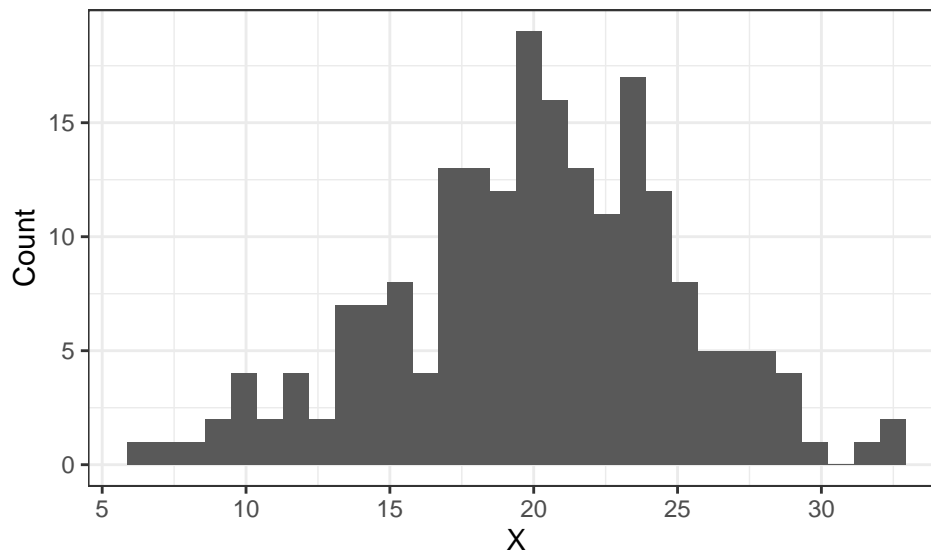
- A** ☐ Die Mittelwerte werden modelliert.
- B** ☐ Die  $X$  werden modelliert.
- C** ☐ Die Varianzstruktur wird modelliert.
- D** ☐ Die  $Y$  werden modelliert.
- E** ☐ Die Varianz der  $X$  unabhängig vom  $Y$  wird modelliert.

## 27 Aufgabe

(2 Punkte)

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





- A** ☐ Es handelt sich um eine Poisson-Verteilung mit  $\text{Pois}(20)$ .
- B** ☐ Es handelt sich um eine Binomial-Verteilung mit  $\text{Binom}(10)$ .
- C** ☐ Es handelt sich um eine Normalverteilung mit  $N(20, 5)$ .
- D** ☐ Eine Standardnormalverteilung mit  $N(0,1)$ .
- E** ☐ Eine rechtsschiefe, multivariate Normalverteilung.

## 28 Aufgabe

(2 Punkte)

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

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

## 29 Aufgabe

(2 Punkte)

Welche Aussage über den t-Test ist richtig?

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

### 30 Aufgabe


(2 Punkte)

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

- A** ☐ Wenn die 1 im Konfidenzintervall enthalten ist, kann die Nullhypothese abgelehnt werden.
- B** ☐ Wenn die Relevanzschwelle mit enthalten ist, dann kann der p-Wert nur signifikant sein. Dies gilt nicht bei multiplen linearen Regressionen.
- C** ☐ Wenn die 0 im Konfidenzintervall enthalten ist, kann die Nullhypothese abgelehnt werden.
- D** ☐ Wenn die 0 im Konfidenzintervall enthalten ist, kann die Nullhypothese abgelehnt werden, aber nur wenn der p-Wert höher als 5% ist. Die Inverse der Null gilt.
- E** ☐ Wenn die Konfidenzintervalle den p-Wert der Regression enthalten.

### 31 Aufgabe

(2 Punkte)

In einem Stallexperiment mit  $n = 101$  Ferkeln wurden verschiedene Outcomes gemessen: der Gewichtszuwachs, Überleben nach 21 Tagen sowie Anzahl Verletzungen pro 7 Tagen. Zwei Lichtregime wurden als Einflussfaktor gemessen. Sie erhalten den  Output der Funktion `tidy()` einer simplen *poisson* linearen Regression. Welche Aussage über den **Effekt** ist richtig?

term	estimate	std.error
(Intercept)	0.75	0.09
light_binhigh	0.15	0.13

- A** ☐ In einer poisson Regression wird die Mittelwertsdifferenz betrachtet. Daher ist der Effekt zwischen den beiden Lichtregimen eine Gewichtsänderung von 0.15
- B** ☐ In einer poisson Regression berechnet man das RR. Daher muss der Schätzer des Effektes  $\beta_1$  noch quadriert werden. Somit liegt das RR bei 0.02

- C** ☐ Eine poisson Regression basiert auf dem maximum Likelihood Prinzip. Hierbei kann kein Effekt beschrieben werden. Im Zweifel hilft aber eine Quadrierung der Fehlerquadratrate  $\epsilon$ .
- D** ☐ In einer poisson Regression muss für die Interpretation des Effektes das  $\beta_1$  quadriert werden. Somit liegt das OR bei 0.02
- E** ☐ In einer poisson Regression kann kein Effekt roh interpretiert werden. Es muss erst eine Confounderadjustierung durchgeführt werden.

### 32 Aufgabe

(2 Punkte)

Der Fehler 2. Art wird auch Power  $\beta$  genannt. Welche Aussage über die Power ist richtig?

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

### 33 Aufgabe

(2 Punkte)

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

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

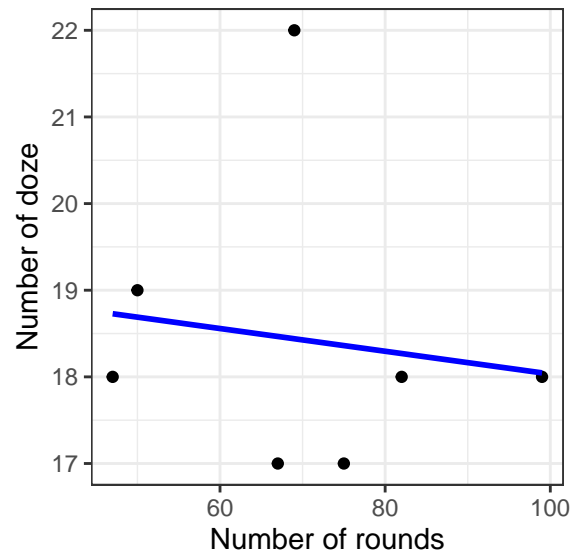
## Rechen- und Textaufgaben

- 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.
- Die Punkte pro Aufgabe dienen als Orientierung. Die Punkte können in der finalen Klausur anders sein.

### 34 Aufgabe

(12 Punkte)

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



1. Beschriften Sie die Gerade mit den gängigen statistischen Maßzahlen.
2. Liegt ein Zusammenhang zwischen der Anzahl an gefahrenen Runden und der Müdigkeit vor? Begründen Sie Ihre Antwort mit statistischen Maßzahlen und der Abbildung.
3. Wenn kein Zusammenhang zu beobachten wäre, wie würde die Gerade aussehen?

### 35 Aufgabe

(8 Punkte)

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

Vergleich	Raw p-val	Adjusted p-val	Reject $H_0$
dose 10 - ctrl	0.760		
dose 15 - ctrl	0.012		
dose 20 - ctrl	0.030		
dose 40 - ctrl	0.001		

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

### 36 Aufgabe

(5 Punkte)

Geben ist folgende 2x2 Kreuztabelle.


1. Tragen Sie folgende Fachbegriffe korrekt in die 2x2 Kreuztabelle ein.

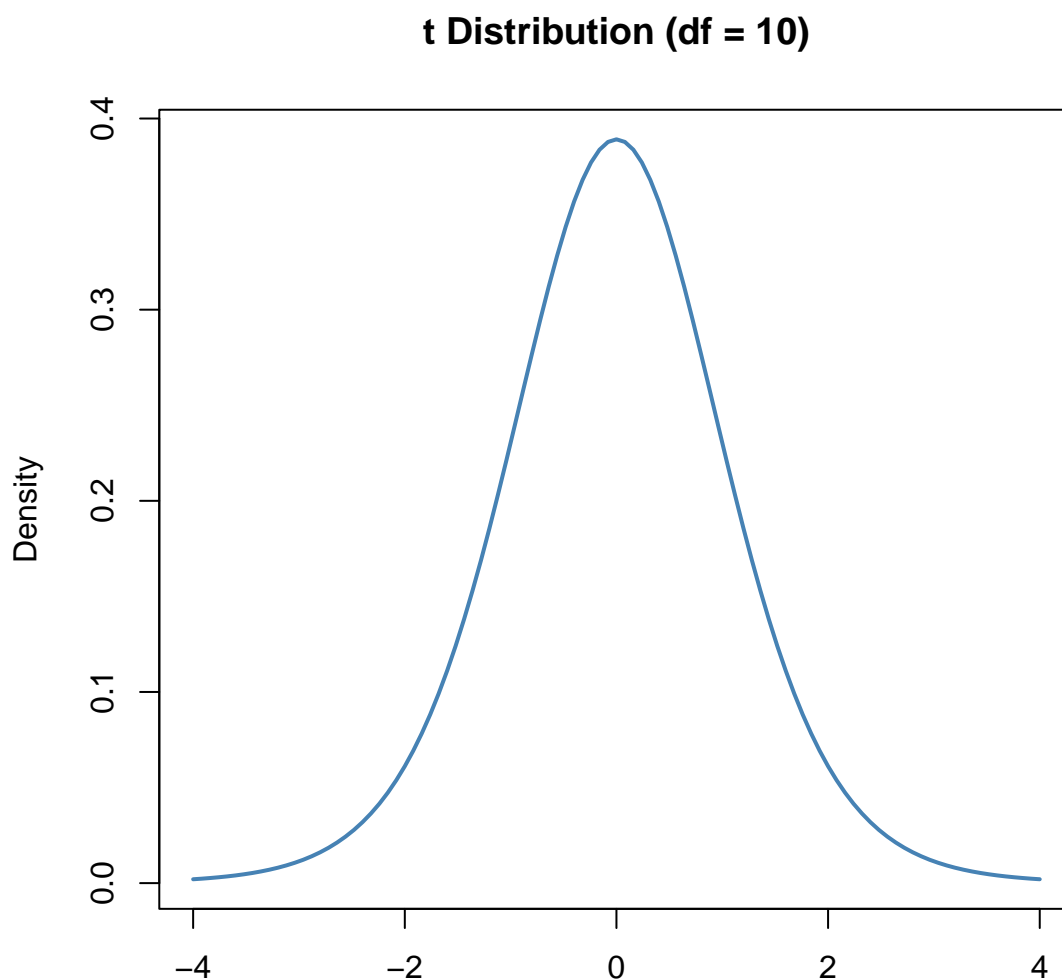
- (Unbekannte) Wahrheit
- $H_0$  wahr
- $H_0$  falsch
- $H_0$  abgelehnt
- $H_0$  beibehalten
- Testentscheidung
- $\alpha$ -Fehler
- $\beta$ -Fehler
- Richtige Entscheidung

### 37 Aufgabe

(8 Punkte)

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

1. Zeichnen Sie das zweiseitige  $\alpha$  Niveau in die Abbildung.
2. Zeichnen Sie einen nicht signifikant p-Wert in die Abbildung.
3. Ergänzen Sie „ $\bar{x}_1 = \bar{x}_2$ “.
4. Ergänzen Sie „ $H_0$  ist wahr“.
5. Ergänzen Sie eine t-Verteilung mit 100 Freiheitsgraden. Zeichnen Sie im Zweifel über den Rand der Abbildung.





### 38 Aufgabe

(10 Punkte)

Sie rechnen eine simple Gaussian Regression.

1. Beschriften Sie die **untenstehende Abbildung** mit der Konfidenzschwelle.
2. Ergänzen Sie eine Relevanzschwelle.
3. Skizzieren Sie in die **untenstehende Abbildung** fünf einzelne Konfidenzintervalle (a-e) mit den jeweiligen Eigenschaften.
  - (a) Ein signifikantes, relevantes Konfidenzintervall
  - (b) Ein nicht signifikantes, relevantes Konfidenzintervall
  - (c) Ein signifikantes, nicht relevantes Konfidenzintervall
  - (d) Ein Konfidenzintervall mit höher Varianz  $s_p$  in der Stichprobe als der Rest der Konfidenzintervalle
  - (e) Ein Konfidenzintervall mit niedriger Varianz  $s_p$  in der Stichprobe als der Rest der Konfidenzintervalle




### 39 Aufgabe

(5 Punkte)

Sie erhalten folgende Ausgabe eines Daten tibbles.

```
## # A tibble: 5 x 4
##   weight number_lesion severe shed
##   <dbl>         <int> <fct> <chr>
## 1   78.4             10 high  Berlin
## 2   54.9             12 high  Kiel
## 3   64.3              9 high  Berlin
## 4   79.2             13 mid   Kiel
## 5   82.1              7 high  Berlin
```

1. Nennen Sie drei Unterschiede zwischen einem `tibble()` und einem `data.frame()` in .
2. Was ist die Bedeutung von `dbl`, `int`, `fct` und `chr`?
3. Die Spalte `severe` ist auch numerisch codiert. Wie lauten die Zahlen hinter den Einträgen in der Spalte?

## 40 Aufgabe

(7 Punkte)

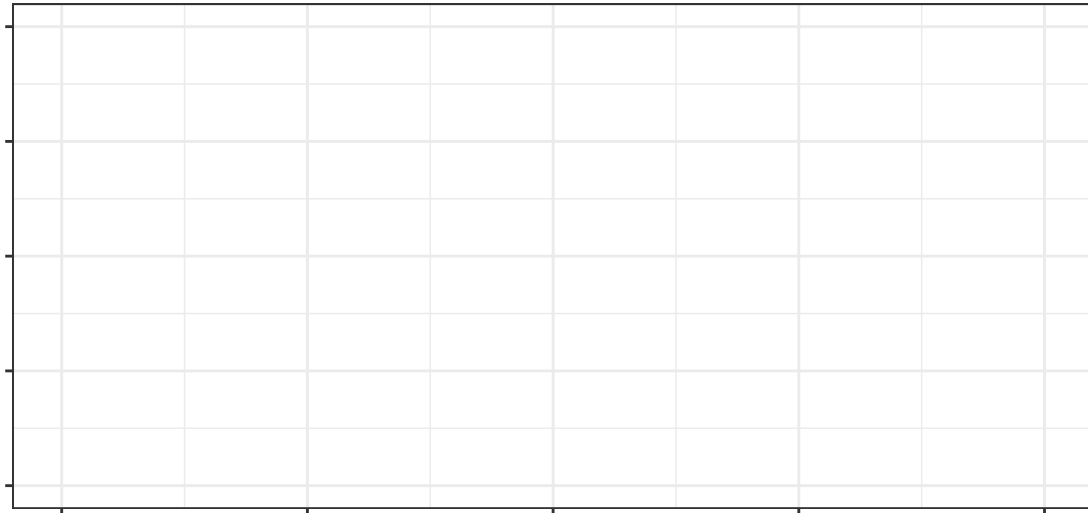
1. Ergänzen Sie vor den jeweiligen Wörtern in der runden Klammer ein **Y** oder ein **X**, je nachdem welchen der beiden Teile eines statistischen Modells  $Y \sim X$  das Wort zugeordnet ist.
  - ( ) simple
  - ( ) Verteilungsfamilie
  - ( ) endpoint
  - ( ) variable
  - ( ) multiple
  - ( ) mehrere Spalten
  - ( ) Effektschätzer
  - ( ) outcome
  - ( ) eine Spalte
  - ( ) response
2. Skizzieren Sie ein simples und ein multiples lineares Regressions Modell.
3. Nenne Sie den Unterschied in modellschreibweise zwischen einer simplen und multiplen linearen Regression.

## 41 Aufgabe

(10 Punkte)

Sie haben folgende Zahlenreihe  $x$  vorliegen  $x = \{20, 20, 18, 19, 18, 15, 19, 22, 16\}$ .

1. Visualisieren Sie den Mittelwert von  $x$  in der untenstehenden Abbildung.
2. Beschriften Sie die Y und X-Achse entsprechend.
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.



## 42 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 (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.

1. Füllen Sie die unterstehende zweifaktorielle ANOVA Ergebnistabelle aus mit den gegebenen Informationen von Df und Sum Sq.
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.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
<b>supp</b>	1	204.65			
<b>dose</b>	2	2425.02			
<b>supp:dose</b>	2	107.33			
<b>Residuals</b>	54	701.05			

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

### 43 Aufgabe

(5 Punkte)

1. Zeichnen Sie über den untenstehenden Boxplot die entsprechende zugehörige Verteilung.
2. Zeichnen Sie den entsprechenden Lageparameter in die Abbildung.
3. Handelt es sich eher um eine Poisson- oder Gaussianverteilung? Begründen Sie Ihre Antwort.



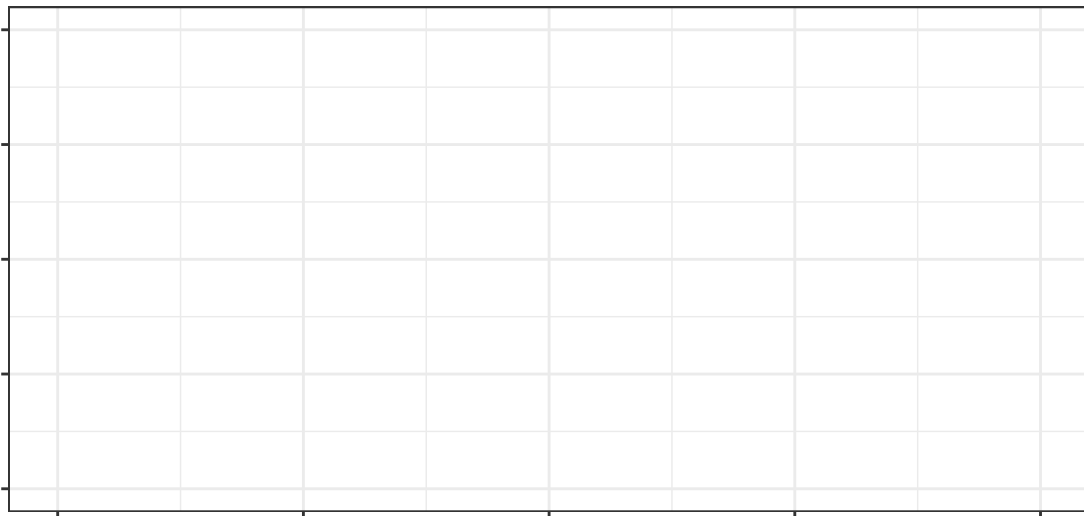
## 44 Aufgabe

(10 Punkte)

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

```
##
## Two Sample t-test
##
## data: activity by timepoint
## t = -0.96421, df = 12, p-value = 0.35397
## alternative hypothesis: true difference in means between group A and group B is
## 95 percent confidence interval:
## -8.9822515 3.4711404
## sample estimates:
## mean in group A mean in group B
## 15.444444 18.200000
```

1. Formulieren Sie die Hypothesenpaare für den t-Test.
2. Liegt ein signifikanter Unterschied zwischen den Gruppen vor? Begründen Sie Ihre Antwort.
3. Skizzieren Sie eine Abbildung in der Sie  $T_c$ ,  $Pr(D|H_0)$  sowie  $T_k = |2.18|$  einzeichnen.
4. Beschriften Sie die Abbildung und den Graphen entsprechend.



## 45 Aufgabe

(14 Punkte)

Das Gewicht von Küken wurde vor der Behandlung mit STARTex gemessen und 1 Woche nach der Behandlung. Es ergab sich die Fragestellung, ob es einen Effekt von STARTex auf das Gewicht von Küken nach einer Woche gibt.

animal_id	before	after
1	19	11
2	12	13
3	23	17
4	7	13
5	12	12
6	8	14
7	10	12

1. Bestimmen Sie die Teststatistik  $T_c$  eines Paired t-Tests für den Vergleich der beiden Zeitpunkte.
2. Formulieren Sie die Fragestellung.
3. Formulieren Sie die **statistischen** Hypothesenpaare.
4. Treffen Sie mit  $T_k = 2.04$  eine Aussage zur Signifikanz der Hypothesenpaare.
5. Wenn Sie keinen Unterschied zwischen den beiden Zeitpunkten erwarten würden, wie groß wäre dann die Teststatistik  $T_c$ ?
6. Schätzen Sie  $Pr(D|H_0)$  ab.



## 46 Aufgabe

(8 Punkte)

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

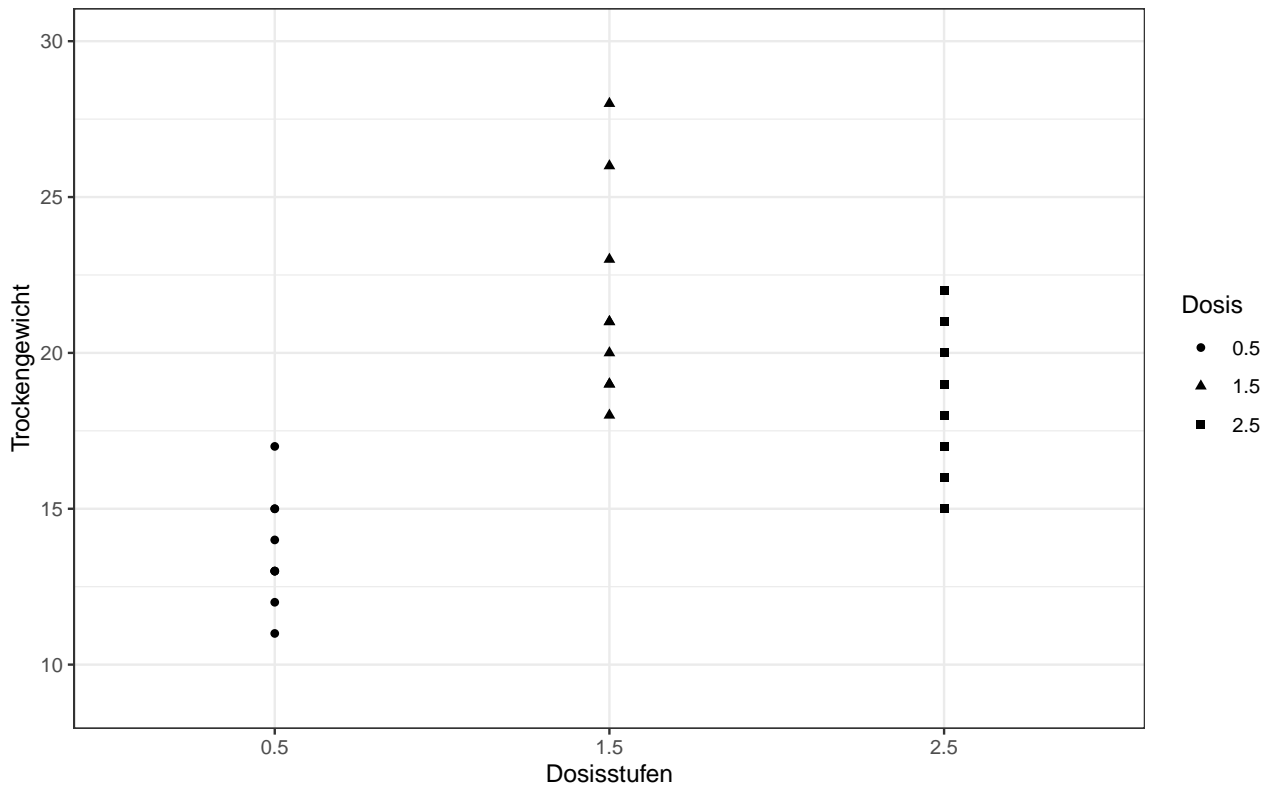
trt	rsp
A	15
A	16
A	17
A	17
A	15
A	15
A	12
A	16
A	12

1. Zeichnen Sie eine passende Abbildung in der Sie visuell überprüfen können, ob eine Normalverteilung des Trockengewichts vorliegt.
2. Entscheiden Sie, ob eine Normalverteilung vorliegt. Begründen Sie Ihre Antwort.

## 47 Aufgabe

(6 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 Abbildung des Modells *Trockengewicht* ~ *Dosisstufen*.



1. Zeichnen Sie folgende statistischen Masszahlen in die Abbildung ein.

- Total (grand) mean:  $\bar{y}_{..}$
- Level means:  $\bar{y}_{i.}$
- Effect of the level:  $\beta_i$
- Residual:  $\epsilon_{ij}$

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

## 48 Aufgabe

(14 Punkte)

Nach einem Experiment ergibt sich die folgende 2x2 Datentabelle mit einem Pestizid (yes/no), dargestellt in den Spalten. Im Weiteren mit dem infizierten Pflanzenstatus (yes/no) in den Zeilen. Insgesamt wurden  $n = 88$  Pflanzen untersucht.



	no	yes
no	14	12
yes	30	32

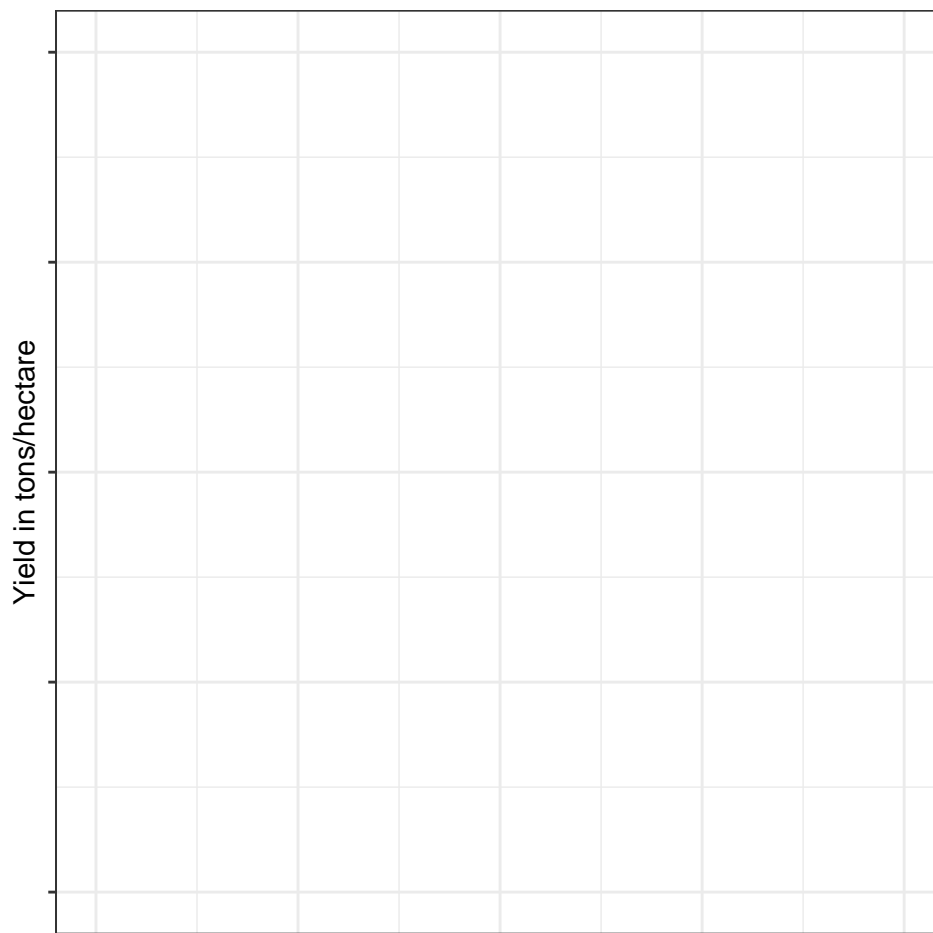
1. Ergänzen Sie die Tabelle um die Randsummen.
2. Berechnen Sie die Teststatistik eines Chi-Quadrat-Test auf der 2x2 Tafel.
3. Formulieren Sie die statistische Fragestellung.
4. Formulieren Sie das statistische Hypothesenpaar.
5. Treffen Sie eine Entscheidung im Bezug zu der Signifikanz gegeben einem  $T_k = 3.841$ .
6. Skizzieren Sie eine 2x2 Tabelle mit  $n = 20$  Pflanzen in dem *vermutlich* die Nullhypothese nicht abgelehnt werden kann.

## 49 Aufgabe

(10 Punkte)

Ein Feldexperiment wurde mit  $n = 200$  Pflanzen durchgeführt. Folgende Einflussvariablen ( $x$ ) wurden erhoben: S, P und dosage. 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 Binomial gehört. Begründen Sie Ihre Auswahl.
2. Schreiben Sie das Modell in der Form  $y \sim x$  wie es in  üblich ist (**Ohne Interaktionsterm**).
3. Schreiben Sie das Modell in der Form  $y \sim x$  wie es in  üblich ist und ergänzen Sie **einen** Interaktionsterm nach Wahl.
4. Zeichnen Sie eine **schwache** signifikante Interaktion in die Abbildung unten. Ergänzen Sie eine aussagekräftige Legende. Wie erkennen Sie eine Interaktion? Begründen Sie Ihre Antwort.



## 50 Aufgabe

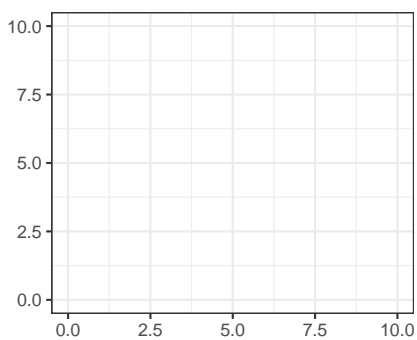
(9 Punkte)

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

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

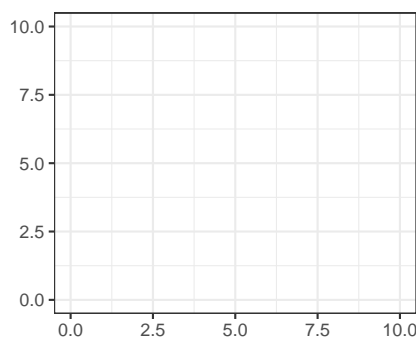
Spearman's  $\rho = -0.75$

$R^2 = 1$



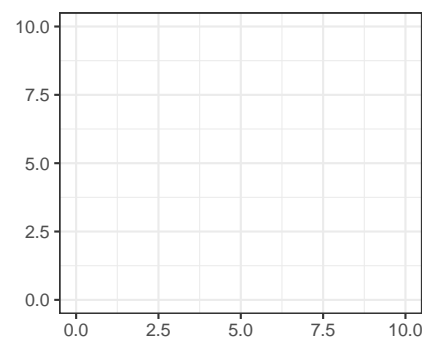
Spearman's  $\rho = 0.25$

$R^2 = 0.75$



Spearman's  $\rho = 1$

$R^2 = 0.25$



## 51 Aufgabe

(8 Punkte)

Gegeben sind folgende Randsummen in einer 2x2 Kreuztabelle aus einem Experiment mit  $n = 199$  Sauen. In dem Experiment wurde gemessen, ob eine Sau nach einer Behandlung mit einem Medikament (0 = nein / 1 = ja) mehr als 30 Ferkeln Jahr bekommen konnte (0 = nein / 1 = ja).

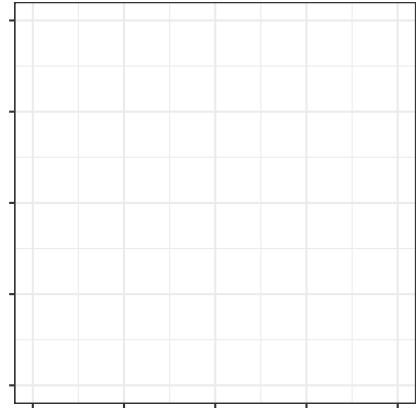
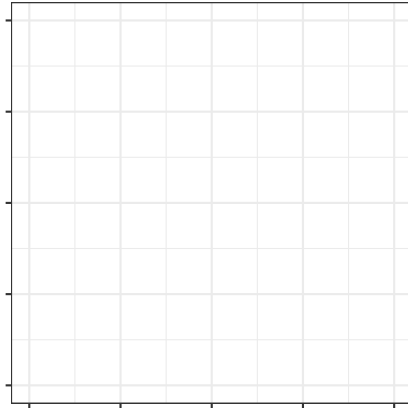
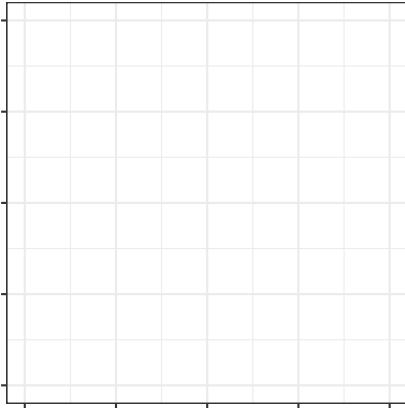
	1	0	
1			49
0			56
	36	58	199

1. Ergänzen Sie die Felder innerhalb der 2x2 Kreuztabelle in dem Sinne, dass **ein** signifikanter Effekt zu erwarten wäre.
2. Erklären und Begründen Sie Ihr Vorgehen an der Formel des Chi-Quadrat-Tests mit  $\chi^2 = \sum \frac{(O-E)^2}{E}$ .
3. Wenn in einer der Felder weniger als 5 Beobachtungen zu erwarten wären, welchen Test können Sie anstatt des „normalen“ Chi-Quadrat-Tests anwenden?

## 52 Aufgabe

(12 Punkte)

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




## 53 Aufgabe


(10 Punkte)

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


stream	longnose	maxdepth	do2	acerage	no3
MEADOW_BR	234	93	8.5	4803	5.01
PRETTYBOY_BR	19	39	9.8	904	6.81
MUDLICK_RUN	8	51	7.4	1507	0.84
LITTLE_GUNPOWDER_R	3	85	9.7	15305	2.60
HAINES_BR	98	50	8.6	1967	7.71

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

```
##
## Call:
## glm(formula = reformulate(response = "longnose", termlabels = wanted_vec),
##      family = quasipoisson, data = data_tbl)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -9.3468  -3.9477  -1.6980   1.5046  15.9072
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.936944297  1.299581183   1.4904 0.1425189
## maxdepth     0.010507579  0.004051315   2.5936 0.0124885 *
## do2          -0.029074299  0.149989402  -0.1938 0.8471007
## acerage       0.000052432  0.000014213   3.6891 0.0005643 ***
## no3           0.255614425  0.075645839   3.3791 0.0014339 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 33.39366)
##
##      Null deviance: 2355.66  on 53  degrees of freedom
## Residual deviance: 1405.86  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. Können Sie die Estimate der einzelnen Einflussvariablen direkt interpretieren? Begründen Sie Ihre Antwort.



3. Interpretieren Sie die **nicht signifikanten** Effekte auf die Anzahl an Knochenhechten.
4. Erklären Sie am  Output wie sich die t value Spalte errechnet.

## 54 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.

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

Welche Auswirkung hat die Änderungen der jeweiligen statistischen Masszahl auf den T-Wert und damit auf die *vermutlich* Signifikanz.

	T Statistik	Signifikanz		T Statistik	Signifikanz
$\Delta \uparrow$			$\Delta \downarrow$		
$s \uparrow$			$s \downarrow$		
$n \uparrow$			$n \downarrow$		

## 55 Aufgabe

(16 Punkte)

Nach einem Experiment mit zwei Pestiziden (A und B) ergibt sich die folgende Datentabelle mit dem gemessenen Trockengewicht (yield). Im Weiteren wurde noch bestimmt, ob das Zieltrockengewicht erreicht wurde und die tägliche mittlere, summierte UV-Einstrahlung protokolliert.

trt	yield	reach	uv
B	19	low	3.1
B	16	low	4.9
A	30	high	3.2
B	20	high	2.6
B	22	high	4.9
A	26	high	2.5
A	21	high	4.4
A	36	high	3.4
B	19	low	3.9
A	13	low	5.5
B	20	high	4.0
A	45	high	3.2
B	20	high	3.2
A	27	high	3.6
B	21	high	4.3
B	17	low	6.0
A	32	high	2.8
A	21	high	3.8

1. Bestimmen Sie die Teststatistik  $T_c$  eines Student t-Tests für den Vergleich der beiden Pestizide.
2. Formulieren Sie die Fragestellung.
3. Formulieren Sie die **statistischen** Hypothesenpaare.
4. Treffen Sie mit  $T_k = 2.04$  eine Aussage zur Signifikanz der Hypothesenpaare.
5. Wenn Sie keinen Unterschied zwischen den beiden Pestiziden erwarten würden, wie groß wäre dann die Teststatistik  $T_c$ ?
6. Schätzen Sie  $Pr(D|H_0)$ . Was ist der Vorteil von  $Pr(D|H_0)$  gegenüber der Teststatistik  $T_c$ ?
7. Was wäre der Unterschied zu einem Welch t-Test?

## 56 Aufgabe

(14 Punkte)

Nach einem Feldexperiment mit zwei Düngestufen (A und B) ergibt sich die folgende Datentabelle mit dem gemessenen Trockengewicht. Im Weiteren wurde noch bestimmt, ob das Zieltrockengewicht erreicht wurde und die tägliche mittlere Wassergabe protokolliert.

trt	rsp	gain	water
A	20	high	9.5
B	27	high	11.3
B	25	high	10.3
B	22	high	11.6
B	24	high	8.6
A	10	low	7.6
B	24	high	9.9
B	24	high	11.2
A	18	low	10.2
B	25	high	10.4
A	10	low	9.9
B	24	high	9.7
A	13	low	11.3
B	25	high	10.1
A	9	low	10.6
A	11	low	12.0

1. Zeichnen Sie in **einer** Abbildung die beiden Boxplots für die zwei Düngestufen A und B.
2. Beschriften Sie **einen** der beiden Boxplots mit den gängigen statistischen Maßzahlen.
3. Wenn Sie **keinen Effekt** zwischen de Düngestufen erwarten würden, wie sehen dann die beiden Boxplots aus?

## 57 Aufgabe

(8 Punkte)

Sie erhalten folgende R Ausgabe der Funktion lm().

```
##
## Call:
## lm(formula = rsp ~ trt, data = data_tbl)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.00000 -0.77778  0.22222  0.97222  2.00000
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  19.77778     0.48644  40.6581 0.0000000000000003171 ***
## trtB         -4.77778     0.81397  -5.8697 0.00007601610733342 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.4593 on 12 degrees of freedom
## Multiple R-squared:  0.74168, Adjusted R-squared:  0.72015
## F-statistic: 34.453 on 1 and 12 DF, p-value: 0.000076016
```

1. Ist die Annahme der Normalverteilung an das Outcome rsp erfüllt? Begründen Sie die Antwort.
2. Wie groß ist der Effekt des Trt? Liegt ein signifikanter Effekt vor?
3. Schreiben Sie das Ergebnis der R Ausgabe in einen Satz nieder, der die Information zum Effekt und der Signifikanz enthält.

## 58 Aufgabe

(16 Punkte)

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

water	drymatter
23	21
22	21
21	20
18	20
21	21

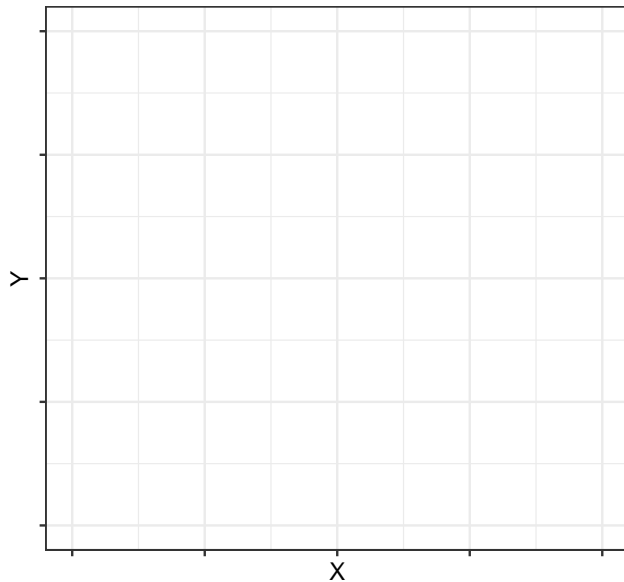
1. Erstellen Sie den Scatter-Plot für die Datentabelle.
2. Zeichnen Sie eine Gerade durch die Punkte.
3. Beschriften Sie die Gerade mit den gängigen statistischen Maßzahlen.
4. Wenn kein Effekt von dem Niederschlag auf das Trockengewicht vorhanden wäre, wie würde die Gerade verlaufen und welche Werte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

## 59 Aufgabe

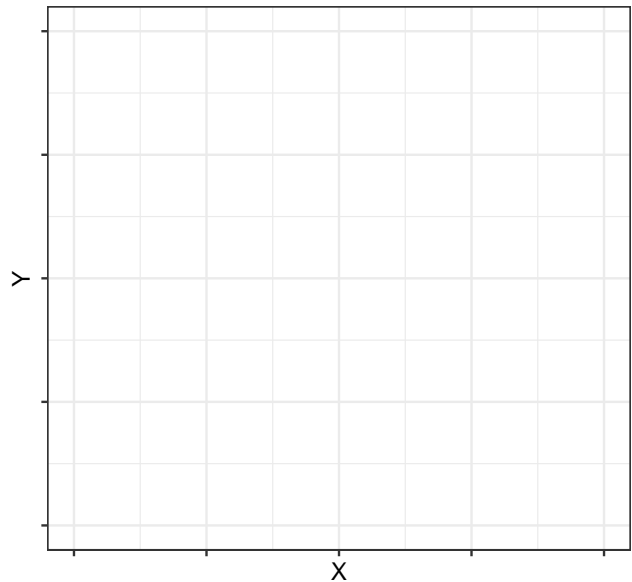
(6 Punkte)

1. Skizzieren Sie in die unten stehenden, freien Abbildungen ein kausales und ein prädiktives Modell mit  $n = 7$  Beobachtungen.
2. Beachten Sie bei der Erstellung der Skizze, ob ein Effekt von X vorliegt oder nicht.

Causal model with effect of X



Predictive model with effect of X



## 60 Aufgabe

(16 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.
2. Schätzen Sie den p-Wert der Tabelle mit der Information von  $F_k = 3.35$  ab.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	3.91			
Residuals	27	11.05			

3. Was bedeutet ein signifikantes Ergebnis in einer einfaktoriellen ANOVA im Bezug auf die möglichen Unterschiede zwischen den Gruppen?
4. Berechnen Sie **einen** t-Test für den *vielversprechendsten* Gruppenvergleich anhand der untenstehenden Tabelle mit  $T_k = 2.03$ . Begründen Sie Ihre Auswahl.

group	mean	sd
ctrl	4.95	0.57
trt1	4.71	0.82
trt2	5.57	0.47

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

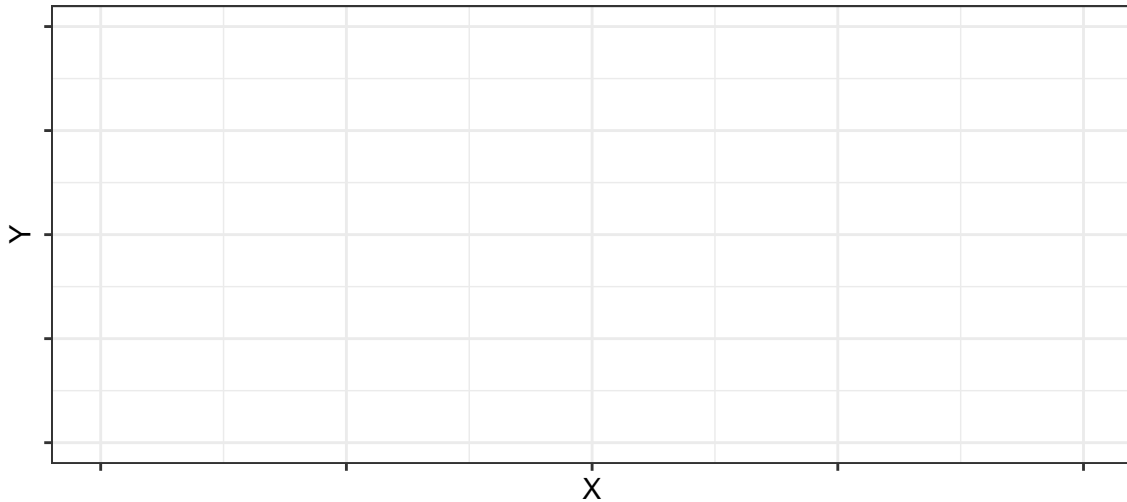


## 61 Aufgabe

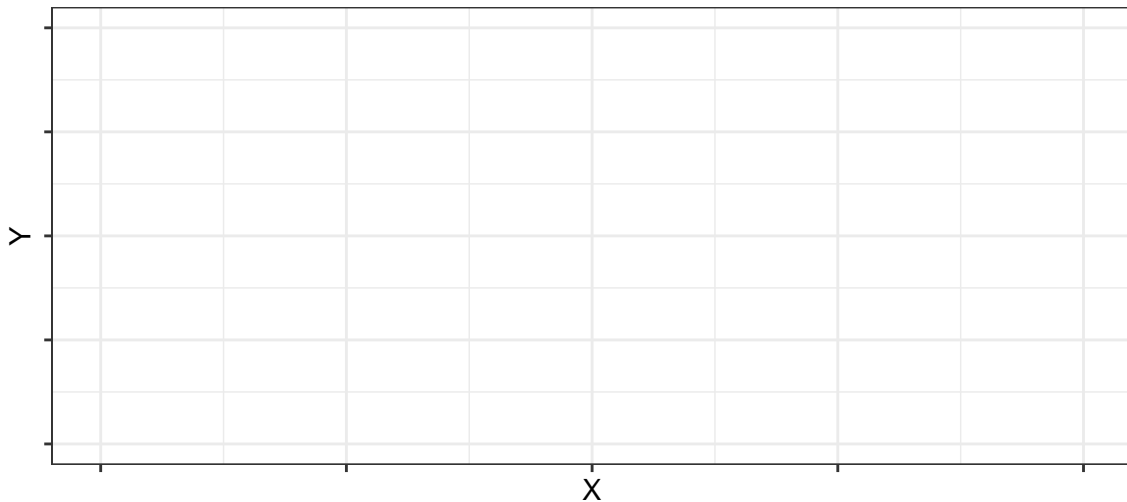
(8 Punkte)

1. Skizzieren Sie in die unten stehenden, freien Abbildungen die Verteilungen, die sich nach der Abbildungsüberschrift ergeben.
2. Beschriften Sie die Achsen entsprechend und achten Sie auf die entsprechende Skalierung der beiden Verteilungen in der ersten Abbildung.

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



$\text{Pois}(10)$



## 62 Aufgabe

(8 Punkte)


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

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

1. Stellen Sie das statistische Regressionsmodell in  $\beta$ -schreibweise auf.
2. Schreiben Sie das statistische Modell in der Matrixform.
  - Schreiben Sie das Modell einmal als effect parametrization.
  - Schreiben Sie das Modell einmal als mean parametrization.

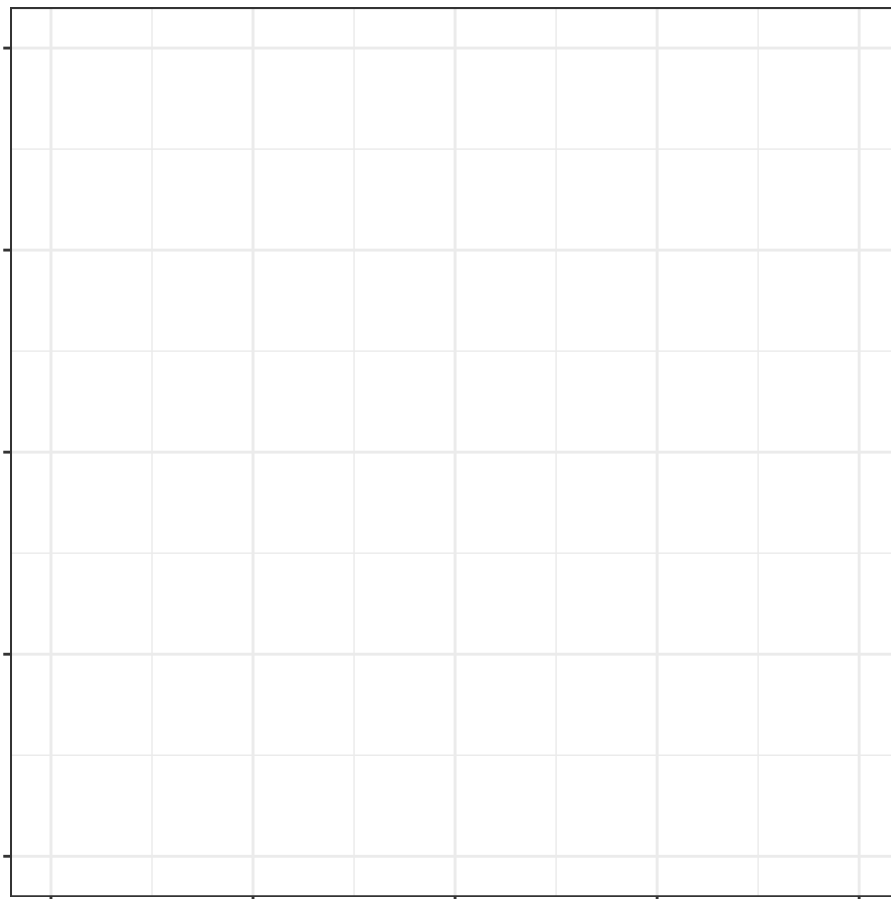
## 63 Aufgabe

(10 Punkte)

In einem Stallexperiment mit  $n = 106$  Ferkeln wurde der Gewichtszuwachs unter bestimmten Lichtverhältnissen gemessen. Sie erhalten den  Output der Funktion `tidy()` einer simplen linearen Regression zu einem Zeitpunkt  $t_2$ .

term	estimate	std.error
(Intercept)	25.22	1.23
light	1.49	0.12

1. Berechnen Sie die t Statistik für (Intercept) und light.
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. Zeichnen Sie die Gerade aus der obigen Tabelle in die untenstehende Abbildung.
4. Beschriften Sie die Abbildung und die Gerade mit den statistischen Kenngrößen.
5. Formulieren Sie die Regressionsgleichung.



## 64 Aufgabe

(14 Punkte)

Nach einem Feldexperiment mit zwei Pestiziden (A und B) ergibt sich die folgende Datentabelle mit dem jeweiligen beobachteten Infektionsstatus.

pest	infected	dead
A	yes	0
A	yes	1
A	no	1
A	yes	0
A	yes	0
B	no	0
B	no	0
B	no	0
A	yes	1
B	yes	0
B	no	0
A	yes	0
B	no	0
B	no	1

1. Stellen Sie in einer 2x2 Tafel den Zusammenhang zwischen dem Pestizid und dem Infektionsstatus dar.
2. Zeichnen Sie den zugehörigen Mosaic-Plot.
3. Welche Maßzahl nimmt jede Fläche des Mosaic-Plots ein?
4. Wenn das Pestizid keine Auswirkung auf den Infektionsstatus hätte, wie sehe dann der Mosaic-Plot aus?