

Name: _____

Nicht bestanden: ☐

Vorname: _____

Matrikelnummer: _____

Endnote: _____

**B.Sc. Landwirtschaft, B.Eng. Wirtschaftsingenieurwesen im Agri- und Hortibusiness,
B.Sc. Angewandte Pflanzenbiologie - Gartenbau, Pflanzentechnologie**

Klausur Angewandte Statistik und Versuchswesen

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

26. Juni 2024

Erlaubte Hilfsmittel für die Klausur

- Normaler Taschenrechner ohne Möglichkeit der Kommunikation mit anderen Geräten - also ausdrücklich kein Handy!
- Eine DIN A4-Seite als beidseitig, selbstgeschriebene, handschriftliche Formelsammlung - keine digitalen Ausdrucke.
- **You can answer the questions in English without any consequences.**

Ergebnis der Klausur

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

_____ von 63 Punkten sind aus dem Rechen- und Textteil erreicht.

_____ von 83 Punkten in Summe.

Es wird folgender Notenschlüssel angewendet.

Punkte	Note
79.5 - 83.0	1,0
75.5 - 79.0	1,3
71.0 - 75.0	1,7
67.0 - 70.5	2,0
63.0 - 66.5	2,3
59.0 - 62.5	2,7
55.0 - 58.5	3,0
50.5 - 54.5	3,3
46.5 - 50.0	3,7
41.5 - 46.0	4,0

Es ergibt sich eine Endnote von _____.

Multiple Choice Aufgaben

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

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

- Es sind ____ von 20 Punkten erreicht worden.

Rechen- und Textaufgaben

- Die Tabelle wird vom Dozenten ausgefüllt.

Aufgabe	11	12	13	14	15	16	17
Punkte	9	8	10	8	12	7	9

- Es sind ____ von 63 Punkten erreicht worden.

1 Aufgabe

(2 Punkte)

Sie führen ein Feldexperiment durch um das Gewicht von Lauch zu steigern. Die Pflanzen wachsen unter einer Kontrolle und zwei verschiedenen Behandlungsbedingungen. Nach der Berechnung einer einfaktoriellen ANOVA ergibt sich ein $\eta^2 = 0.3$. Welche Aussage ist richtig?

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

2 Aufgabe

(2 Punkte)

Die ANOVA ist ein statistisches Verfahren welches häufig in den Auswertungen von Experimenten in den Agrarwissenschaften angewendet wird. Dabei wird die ANOVA als ein erstes statistischen Werkzeug für die Übersicht über die Daten benutzt. Eine ANOVA testet dabei ...

- A** ☐ ... den Unterschied zwischen der globalen Varianz und der Varianz aus verschiedenen Behandlungsguppen. Wenn die ANOVA signifikant ist, ist nicht bekannt welcher Vergleich konkret unterschiedlich ist.
- B** ☐ ... den Unterschied zwischen der F-Statistik anhand der Varianz der Gruppen. Wenn die F-Statistik exakt 0 ist, kann die Nullhypothese abgelehnt werden.
- C** ☐ ... den Unterschied zwischen zwei paarweisen Mittelwerten aus verschiedenen Behandlungsguppen. Wenn die signifikant ist, ist daher bekannt welcher Vergleich konkret unterschiedlich ist.
- D** ☐ ... den Unterschied zwischen mehreren Varianzen aus verschiedenen Behandlungsguppen. Wenn die ANOVA signifikant ist, ist nicht bekannt welcher Vergleich konkret unterschiedlich ist.
- E** ☐ ... den Unterschied zwischen der Mittelwerte und der Varianz aus verschiedenen Behandlungsguppen. Wenn die ANOVA signifikant ist, ist bekannt welcher Vergleich konkret unterschiedlich ist.

3 Aufgabe

(2 Punkte)

Der Barplot stellt folgende statistische Maßzahlen in einer Abbildung dar. Damit gehört der Barplot zu einem der am meisten genutzten statistischen Verfahren zur Visualisierung von Daten.

- A** ☐ Den Median und die Quartile.
- B** ☐ Den Mittelwert und die Standardabweichung.
- C** ☐ Den Mittelwert sowie den Median und die Streuung.
- D** ☐ Den Mittelwert und die Varianz.
- E** ☐ Den Median und die Standardabweichung.

4 Aufgabe

(2 Punkte)

Betrachten wir die Teststatistik aus einem abstrakteren Blickwinkel. Beim statistischen Testen wird das „*signal*“ mit dem „*noise*“ zu einer Teststatistik T verrechnet. Welche der Formel berechnet korrekt die Teststatistik T?

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

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

5 Aufgabe

(2 Punkte)

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

- A** ☐ Wir vergleichen mit dem p -Wert und dem Signifikanzniveau α Wahrscheinlichkeiten und damit die absoluten Werte auf einem Zahlenstrahl, wenn die H_0 gilt.
- B** ☐ 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.
- 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 Signifikanzschwelle unter der Annahme der Nullhypothese.
- E** ☐ Wir machen eine Aussage über die individuelle Wahrscheinlichkeit des Eintretens der Nullhypothese H_0 .

6 Aufgabe

(2 Punkte)

Welche Aussage über den Effekt eines statistischen Tests ist richtig?

- A** ☐ Der Effekt eines statistischen Tests beschreibt den Output oder die Wiedergabe eines Tests in einem Computer.
- B** ☐ Der Effekt eines statistischen Tests beschreibt die mathematisch interpretierbare Ausgabe eines Tests. Damit ist der Effekt direkt mit dem Begriff der Signifikanz verbunden. Die Entscheidung über die Signifikanz trifft der Forschende unabhängig von der Relevanz eines statistischen Tests.
- C** ☐ Der Effekt eines statistischen Tests beschreibt die biologisch interpretierbare Ausgabe eines Tests. Modernen Algorithmen liefern keine Effekte mehr sondern nur noch bedingte Wahrscheinlichkeiten. Der Effekt spielt in der modernen Statistik keine Rollen mehr.
- D** ☐ Der Effekt eines statistischen Tests beschreibt die biologisch interpretierbare Ausgabe eines Tests. Zum Beispiel den mittleren Unterschied zwischen zwei Gruppen aus einem t-Test. Damit ist der Effekt direkt mit dem Begriff der Relevanz verbunden. Die Entscheidung über die Relevanz trifft der Forschende unabhängig von der Signifikanz eines statistischen Tests.
- E** ☐ Der Effekt eines statistischen Tests beschreibt die biologisch interpretierbare Ausgabe eines Tests. Damit ist der Effekt direkt mit dem Begriff der Signifikanz verbunden. Die Entscheidung über die Signifikanz trifft der Forschende unabhängig von der Relevanz eines statistischen Tests.

7 Aufgabe

(2 Punkte)

Die Testtheorie hat mehrere Säulen. Einer der Säulen ist das Falsifikationsprinzip. Das Falsifikationsprinzip besagt,

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

8 Aufgabe

(2 Punkte)

Die Abkürzung *CLD* steht für welches statistische Verfahren? Welche anschließende Beschreibung der Interpretation ist korrekt?

- A** ☐ Compact letter display. Gleichheit in den Behandlungen wird durch den gleichen Buchstaben oder Symbol dargestellt. Teilweise ist die Interpretation des CLD herausfordernd, da wir ja nach dem Unterschied suchen.
- B** ☐ Compact line display. Gleichheit in den Behandlungen wird durch den gleichen Buchstaben oder Symbol dargestellt. Früher wurden keine Buchstaben sondern eine durchgezogene Linie verwendet. Bei mehr als drei Gruppen funktioniert die Linie aber graphisch nicht mehr.
- C** ☐ Contrast letter display. Unterschiede in den Behandlungen werden durch den gleichen Buchstaben oder Symbol dargestellt. Die Interpretation des CLD führt häufig in die Irre.
- D** ☐ Compact letter detection. Gleichheit in den Behandlungen wird durch den gleichen Buchstaben oder Symbol dargestellt.
- E** ☐ Compound letter display. Gleichheit in dem Outcomes wird durch den gleichen Buchstaben oder Symbol dargestellt. Teilweise ist die Interpretation des Verbunds (eng. compound) herausfordernd, da wir ja nach dem Unterschied suchen.

9 Aufgabe


(2 Punkte)

Um die Varianz zu berechnen müssen wir folgende Rechenoperationen durchführen.

- A** ☐ Den Mittelwert berechnen, dann die quadratischen Abstände zum Mittelwert aufsummieren und durch die Fallzahl teilen.
- B** ☐ Den Mittelwert berechnen und die Abstände quadrieren. Die Summe mit der Fallzahl multiplizieren.
- C** ☐ Den Mittelwert berechnen, dann die absoluten Abstände zum Mittelwert aufsummieren
- D** ☐ Den Mittelwert berechnen, dann die quadratischen Abstände zum Mittelwert aufsummieren und durch die Fallzahl teilen, dann die Wurzel ziehen.
- E** ☐ Den Median berechnen, dann die quadratischen Abstände zum Median aufsummieren, dann die Wurzel ziehen.

10 Aufgabe

(2 Punkte)

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


- A** ☐ Wir transformieren die Spalten über `mutate()` in ein `tibble` und können dann über `ggplot()` uns die Abbildungen erstellen lassen. Dabei beachten wir das wir keine Faktoren in den Daten haben.
- B** ☐ Wir lesen als erstes die Daten über `read_excel()` ein, transformieren die Spalten über `mutate()` in die richtige Form und können dann über `ggplot()` uns die Abbildungen erstellen lassen. Wichtig ist, dass wir keine Faktoren sondern nur numerische Variablen vorliegen haben.
- 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)` entfernen wir das Paket von der Analyse.
- D** ☐ Wir lesen als erstes die Daten über `read_excel()` ein, transformieren die Spalten über `mutate()` in die richtige Form und können dann über `ggplot()` uns die Abbildungen erstellen lassen.
- E** ☐ Wir 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.

11 Aufgabe

(9 Punkte)

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



In einem Experiment für den Proteingehalt von Wasserlinsen in g/l mit vier Dosisstufen (ctrl, low, mid und high) erhalten Sie folgende Matrix als  Ausgabe mit den rohen, unadjustierten p -Werten.

```
##          ctrl      high      low      mid
## ctrl 1.0000000 0.0808074 0.3967099 0.0231227
## high 0.0808074 1.0000000 0.0120476 0.5640457
## low  0.3967099 0.0120476 1.0000000 0.0027516
## mid  0.0231227 0.5640457 0.0027516 1.0000000
```

Im Weiteren erhalten Sie folgende Informationen über die Fallzahl n , den Mittelwert $mean$ und die Standardabweichung sd in den jeweiligen Dosisstufen.


trt	n	mean	sd
ctrl	9	11.79	1.19
high	9	8.54	4.64
low	9	13.33	5.10
mid	9	7.48	3.08

1. Zeichnen Sie in eine Abbildung, die sich ergebenden Barplots! **(2 Punkte)**
2. Adjustieren Sie die rohen p -Werte nach Bonferroni. Begründen Sie Ihre Antwort! **(3 Punkte)**
3. Ergänzen Sie das *Compact letter display (CLD)* zu der Abbildung. Nutzen Sie dazu die rohen p -Werte! **(2 Punkte)**
4. Interpretieren Sie das *Compact letter display (CLD)*! **(2 Punkte)**

12 Aufgabe

(8 Punkte)



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

```
##  
## Two Sample t-test  
##  
## data:  freshmatter by N  
## t = 1.1095, df = 12, p-value = 0.2889  
## alternative hypothesis: true  is not equal to [condensed]  
## 95 percent confidence interval:  
##  -2.398618  7.376396  
## sample estimates:  
## mean in group ctrl mean in group trt2  
##      20.60000      18.11111
```


1. Formulieren Sie das statistische Hypothesenpaar! **(2 Punkte)**
2. Liegt ein signifikanter Unterschied zwischen den Gruppen vor? Begründen Sie Ihre Antwort! **(2 Punkte)**
3. Skizzieren Sie das sich ergebende 95% Konfidenzintervall! **(2 Punkte)**
4. Beschriften Sie die Abbildung und das 95% Konfidenzintervall entsprechend! **(2 Punkte)**

13 Aufgabe

(10 Punkte)

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



In einem Stallexperiment mit $n = 120$ Ferkeln wurde der Gewichtszuwachs in kg unter ansteigender Lichteinstrahlung in nm gemessen. Sie erhalten den  Output einer simplen Gaussian linearen Regression sieben Wochen nach der ersten Messung.

term	estimate	std.error	t statistic	p-value
(Intercept)	2.334751	1.3311480		
light	1.939039	0.1316322		

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

14 Aufgabe

(8 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 keine Interaktion zwischen den Faktoren f_1 und f_2 ! **(4 Punkte)**
2. Erklären Sie den Unterschied zwischen den beiden Stärken der Interaktion! **(2 Punkte)**
3. Wenn eine signifikante Interaktion in den Daten vorliegt, wie ist dann das weitere Vorgehen bei einem Posthoc-Test? **(2 Punkte)**

15 Aufgabe

(12 Punkte)

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



Der Datensatz `fertilizer_growth_tbl` enthält den Ertrag pro Hektar der Erbsenschooten, die unter einer Kontrolle und zwei verschiedenen Behandlungsbedingungen erzielt wurden. Dabei wurden die Erbsenschooten unter verschiedenen Konzentrationen von einem alternativen Dünger angebaut. Als Behandlung haben Sie daher den Faktor *group* mit den Faktorstufen *ctrl*, *pos* und *extreme* vorliegen.

1. Füllen Sie die unterstehende einfaktorielle ANOVA Ergebnistabelle mit den gegebenen Informationen von **Df** und **Sum Sq** aus! (3 Punkte)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	94.17			
error	20	253.75			

2. Schätzen Sie den p-Wert der Tabelle mit der Information von $F_{\alpha=5\%} = 3.49$ ab. Begründen Sie Ihre Antwort! (2 Punkte)
3. Was bedeutet ein signifikantes Ergebnis in einer einfaktoriellen ANOVA im Bezug auf die möglichen Unterschiede zwischen den Gruppen? Beziehen Sie sich auf den obigen Fragetext bei Ihrer Antwort! (2 Punkte)
4. Berechnen Sie *einen* Student t-Test mit für den *vermutlich* signifikantesten Gruppenvergleich anhand der untenstehenden Tabelle mit $T_{\alpha=5\%} = 2.03$. Begründen Sie Ihre Auswahl! (3 Punkte)

group	n	mean	sd
ctrl	7	15.29	0.49
pos	7	17.29	0.95
extreme	9	20.11	5.56

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

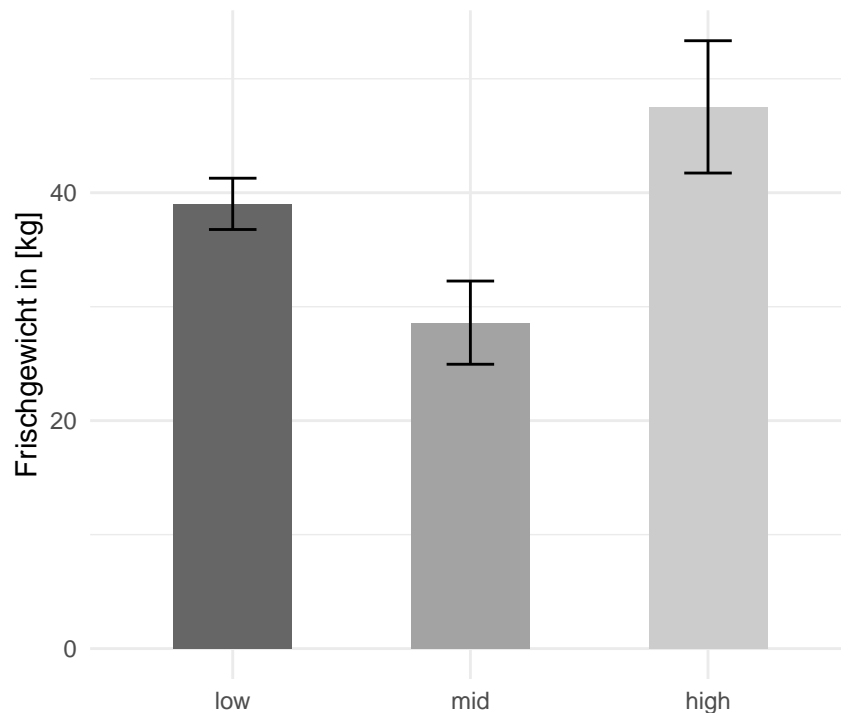
16 Aufgabe


(7 Punkte)

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



In einer Klimakammer mit drei Bewässerungstypen (*low*, *mid* und *high*) als Behandlung (*treatment*) ergeben sich die folgenden Barplots mit dem gemessenen Frischgewicht (*freshmatter*) von Kartoffeln.



1. Erstellen Sie eine Tabelle mit den statistischen Maßzahlen aus der obigen Abbildung der drei Barplots! Beachten Sie die korrekte Darstellungsform der statistischen Maßzahlen! **(3 Punkte)**
2. Erstellen Sie einen beispielhaften Datensatz, aus dem die drei Barplots *möglicherweise* erstellt wurden, im  üblichen Format! **(2 Punkte)**
3. Erwarten Sie einen Unterschied zwischen den Behandlungen? Begründen Sie Ihre Antwort! **(2 Punkte)**

17 Aufgabe

(9 Punkte)

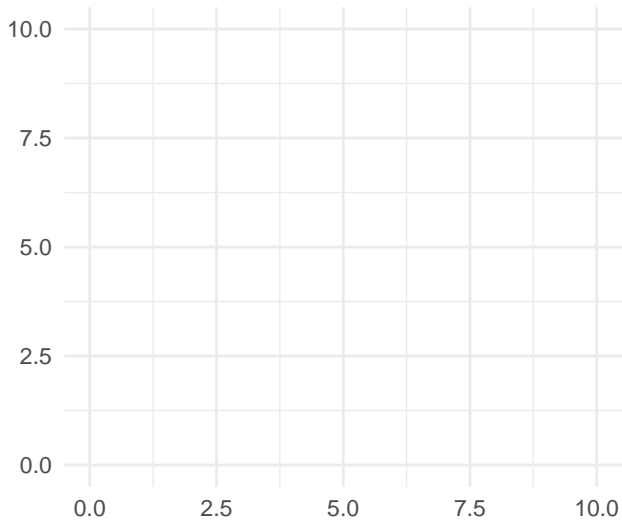


Im folgenden sehen Sie drei leere Scatterplots. Füllen Sie diese Scatterplots nach folgenden Anweisungen.

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 Punktwolke um die Gerade. **(3 Punkte)**
3. Sie rechnen ein statistisches Modell. Was sagen Ihnen die R^2 -Werte über das jeweilige Modell? **(3 Punkte)**

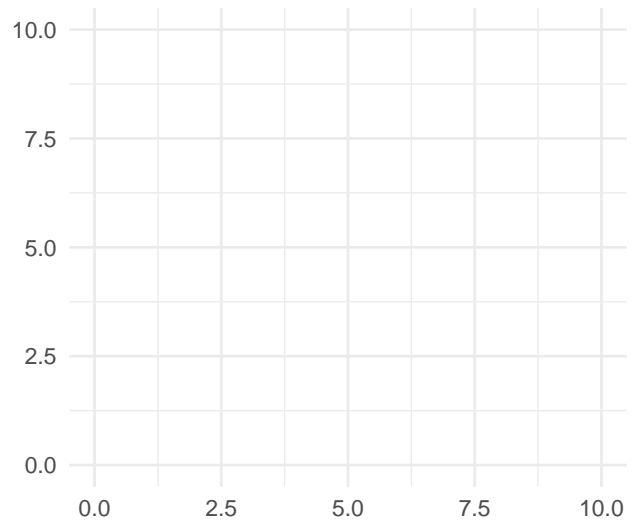
Pearsons $\rho = -0.5$

$R^2 = 1$



Pearsons $\rho = -0.25$

$R^2 = 0.25$



Pearsons $\rho = 0.75$

$R^2 = 0.75$

