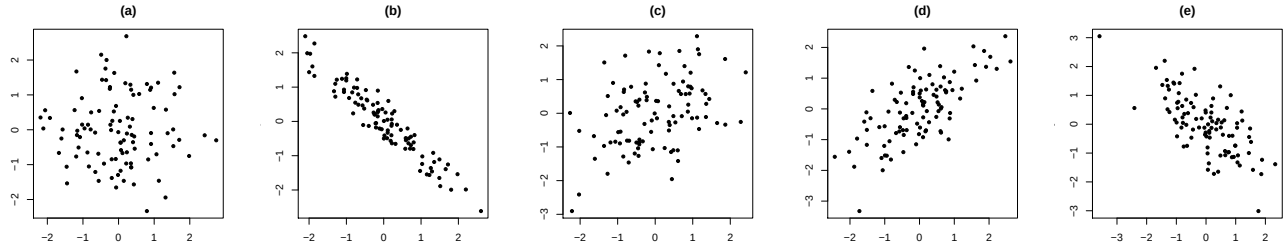
Prof. D. Metzler SS 2021

ÜBUNGEN ZUR WAHRSCHEINLICHKEITSRECHNUNG UND STATISTIK FÜR BIOLOGEN

Blatt 11

**Aufgabe 1**

(a) Die unten stehenden Scatterplots (a)–(e) zeigen Datensätze mit verschiedenen Korrelationskoeffizienten. Ordnen Sie die Bilder und Korrelationkoeffizienten einander zu.




| Korrelationkoffizient | -0.95 | -0.7 | 0 | 0.4 | 0.7 |
|---|---|---|---|---|---|
| Bild | | | | | |

(b) Berechnen Sie für folgenden Mini-Datensatz jeweils Mittelwert und Stichprobenvarianz der $x$- und der $y$-Werte sowie die Kovarianz von $x$ und $y$ und den Korrelationskoeffizienten:

| $x$ | 1 | 3 | 2 |
|---|---|---|---|
| $y$ | 0 | 2 | -2 |

Sowohl Skalierung mit $n-1$ als auch mit $n$ sind möglich; diskutieren Sie, welche Skalierung sinnvoller ist und wovon das abhängt.

(c) Berechnen Sie nur mit Papier, Stift, Taschenrechner und Quantiltafeln für den Mini-Datensatz aus (b) die Regressionsgrade für die Vorhersage von $y$ aus $x$ und prüfen Sie, ob der Zusammenhang zwischen $x$ und $y$ statistisch signifikant ist.

(d) Visualisieren Sie die Daten aus (b) zusammen mit Ihrer Regressionsgeraden aus (c).

**Aufgabe 2** Folgende Tabelle zeigt Dauer des Studiums (in Semestern) und Einstiegsgehalt (in Tausend €) der Absolventen eines Jahres am Fachbereich Mathematik und Informatik der Yule-Simpson-Universität:

| Semester | 12 | 14 | 16 | 12 | 15 | 14 | 13 | 14 | 11 | 13 | 10 | 12 | 14 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gehalt | 39.4 | 38.2 | 37.4 | 39.5 | 32.8 | 35.3 | 39.1 | 35.2 | 37.9 | 35.7 | 41 | 40.9 | 34.2 | 38.4 | 36.2 | 38.4 |
| Semester | 9 | 11 | 9 | 9 | 12 | 13 | 11 | 10 | 10 | 10 | 9 | 10 | 12 | 10 | | |
| Gehalt | 33.7 | 35.9 | 36.1 | 34.2 | 29.9 | 31.9 | 33.3 | 36.2 | 33.8 | 32.9 | 33.3 | 35.1 | 34.2 | 35.3 | | |

(a) Schlägt sich (für diese Absolventen) ein längeres Studium in einem höheren Anfangsgehalt nieder? Bestimmen Sie die Regressiongerade für Studiendauer gegen Anfangsgehalt.

(b) Ändert sich Ihr Befund, wenn Sie zusätzlich erfahren, dass die oberen beiden Zeilen der Tabelle sich auf die Absolventen des Fachs Informatik, die unteren beiden sich auf die Absolventen des Fachs Mathematik beziehen, und Sie dieselbe Regression jeweils innerhalb dieser beiden Gruppen durchführen?

(c) Führen Sie auch eine Regressionsanalyse durch, in der Sie sowohl dis Studiendauer als auch das Studienfach als erklärende Variablen für das Einstiegsgehalt berücksichtigen. Überprüfen Sie durch eine Varianzanalyse und weitere Kriterien, ob Sie zusätzlich einen Interaktionsterm zwischen Studienfach und Studiendauer im Modell haben sollten.

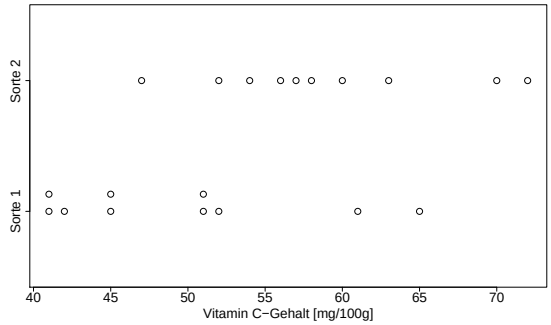
(d) Visualisieren Sie in geeigneter Weise die Daten und die Regressionsmodelle ohne und mit Berücksichtigung des Studienfachs.

**Aufgabe 3** In einer Studie wurde die Wirksamkeit eines Schlafmittels geprüft und für zehn Patienten die zusätzliche Anzahl Stunden Schlaf pro Tag (gemittelt über den Zeitraum der Medikamentengabe) im Vergleich zu einem Referenzzeitraum ohne Medikament bestimmt:

1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4

Der Stichproben-Mittelwert der zusätzlichen Stunden Schlaf ist $\mu = 2.33$ mit Stichproben-Varianz $s^2 = 4.01$. Testen Sie (unter einer Normalverteilungsannahme an die Daten) die Wirksamkeit des Schlafmittels. Hinweise: Wie immer zuerst die Nullhypothese formulieren. Geben Sie auch ein 95%-Konfidenzintervall für die mittlere zusätzliche Anzahl Stunden Schlaf mit Medikament an.

**Aufgabe 4** Der Vitamin C-Gehalt zweier Kohlsorten wurde in jeweils 10 Stichproben pro Sorte untersucht:




Dabei wurde für die Proben von Sorte 1 ein mittlerer Gehalt $\mu_1 = 49.4$ mg pro 100g mit Standardabweichung $s_1 = 8.33$ gefunden, für Sorte 2 waren die Werte $\mu_2 = 58.9$ und $s_2 = 7.74$. Testen Sie die Hypothese, dass der mittlere Vitamin C-Gehalt für beide Sorten gleich ist, mittels eines $t$-Tests zum Irrtumsniveau 5% und formulieren Sie Ihr Ergebnis in einem Satz. Nehmen Sie dazu an, dass die wahren Varianzen gleich sind.

**Aufgabe 5** Phenylketonurie ist eine autosomal-rezessiv vererbte menschliche Stoffwechselkrankheit (d.h. nur homozygote Träger des „kranken“ Allels $a$ erkranken), an der (in Deutschland) ca. eines von 8000 Neugeborenen leidet. Nehmen wir an, die Population befinde sich (bezüglich dieses Gens) im Hardy-Weinberg-Gleichgewicht. Welchen Anteil der Population haben dann die Genotypen $AA$, $aA$ und $aa$?
Ein gesundes Paar (gebildet gemäß der „Hardy-Weinberg-Regeln“) habe ein gesundes Kind. Wie wahrscheinlich ist es, gegeben diese Information, dass beide Eltern Genotyp $AA$ haben?

**Aufgabe 6** ($t$-Statistik und Permutationstests) Anstatt zur Bestimmung des $p$-Werts (approximative) Aussagen über die Verteilung der $t$-Statistik zu benutzen, kann man die Hypothese „die beiden Stichproben stammen aus derselben Population“ auch mittels eines Permutationstests prüfen, beispielsweise folgendermaßen: Verteile die Gruppenbezeichnungen rein zufällig (Hinweis: der R-Befehl `sample(x)` erzeugt eine zufällige Permutation des Vektors `x`), berechne die $t$-Statistik für die so permutierten Daten (Hinweis: `t.test()$statistic`) und bestimme durch wiederholte Simulation, mit welcher Wahrscheinlichkeit sich dabei ein betragsmäßig mindestens so großer Wert der $t$-Statistik wie der tatsächlich beobachtete ergibt. Dieses Vorgehen ist zwar aufwändiger als der klassische $t$-Test, dafür aber „immun“ gegen Verletzungen der Normalverteilungsannahme. Führen Sie den Permutationstest wie gerade beschrieben für die Milben-Daten aus `milben.csv` durch, gemäß dem Beispiel aus der Vorlesung über den zwei-Stichproben-t-test. Vergleichen Sie dann das Ergebnis mit dem p-Wert des t-Tests.