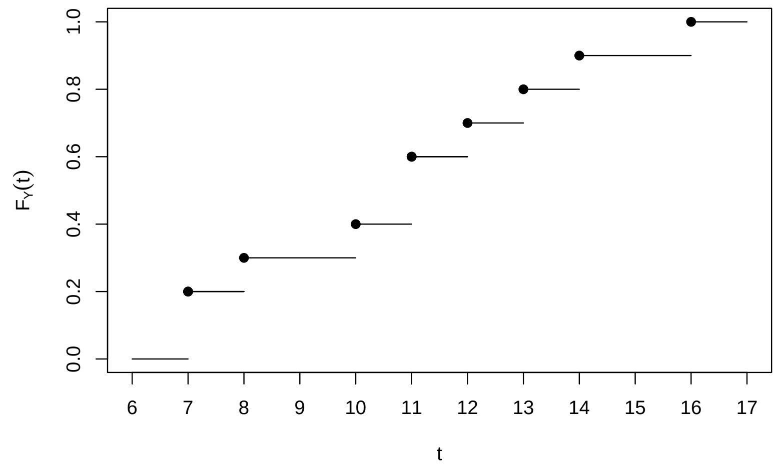
Lösungen zu Tutoriumsblatt 9

Aufgabe T1 (Empirische Verteilungsfunktion)

Folgendes Schaubild zeigt die empirische Verteilungsfunktion $F_Y(t)$ der Stichprobe $y = (y_1, \dots, y_{10})$.



a) Bestimmen Sie basierend auf dem Schaubild den minimalen Wert $y_{(1)}$, den maximalen Wert $y_{(10)}$, den Median $\widetilde{y}$ und das untere Quartil $\widetilde{y}_{0.25}$ von $(y_1, ..., y_{10})$.

b) Geben Sie die geordnete Stichprobe $y_{()}$ an.

Lösungsvorschlag:

a) Die gesuchten Größen können direkt aus dem Schaubild abgelesen werden: $y_{(1)}$ und $y_{(10)}$ sind die Werte auf der x-Achse bei der ersten und letzten Sprungstelle von F_Y, also bei 7 und 16. Für den Median und das 0.25-Quantil suchen wir den Wert t, zu dem der Wert von F_Y erstmals größer gleich 0.5 bzw. 0.25 ist. Wir erkennen damit, dass der Median bei $\widetilde{y} = 11$ und das 0.25-Quantil bei $\widetilde{y}_{0.25} = 8$ liegen. Alternativ hätte man zuerst die geordnete Stichprobe bestimmen und anschließend die Formeln aus der Vorlesung verwenden können.

b) Anhand der Sprünge der empirischen Verteilungsfunktion erkennen wir, dass die Werte $7, 8, 10, 11, 12, 13, 14$ und 16 in der Stichprobe y vorkommen. Da zwei der Sprünge Höhe 0.2 haben während die anderen Sprünge der Höhe 0.1 sind, erkennen wir, dass die Werte 7 und 11 genau zweimal in der Stichprobe y auftauchen, alle anderen Werte nur einmal. Daher ist die geordnete Stichprobe gegeben durch

$$y_{()} = (7, 7, 8, 10, 11, 11, 12, 13, 14, 16).$$

Aufgabe T2 (Statistik im Karlsruher Zoo)

Im Karlsruher Zoo wurden an einem Biber tägliche Messungen durchgeführt, um den Holzkonsum zu analysieren. Hierzu wurde ein Gehege konstruiert, sodass an jedem Tag für den Biber identische Bedingungen (Temperatur, Baumsorten, etc.) vorherrschten. Jeden Abend wurde die konsumierte Holzmenge gemessen, welche die folgenden Messwerte (in Gramm) lieferte:

j	1	2	3	4	5	6	7	8	9	10
x_j	950	921	925	925	960	980	960	930	950	925

(a) Berechnen Sie folgende Größen und geben Sie jeweils an, was diese Größen aussagen:

i) das empirische 70%–Quantil (0,7–Quantil),
ii) das arithmetische Mittel,
iii) den Median,
iv) die empirische Varianz,
v) die empirische Standardabweichung,
vi) den Quartilsabstand.

(b) Ersetzt man den 6. Wert 980 durch den Wert 1098 (Schreibfehler), so ändern sich von den in (ii) bestimmten Größen **genau 3 nicht**. Welche sind dies?

Lösungsvorschlag:

(a) Da diverse Größen, die hier auftreten, mit Hilfe der geordneten Stichprobe berechnet werden müssen, schreiben wir diese zunächst auf:

j	1	2	3	4	5	6	7	8	9	10
$x_{(j)}$	921	925	925	925	930	950	950	960	960	980

i) Wir berechnen das α-Quantil mit $n = 10$, $\alpha = 0{,}7$. Damit ist $n \cdot \alpha = 7$, eine natürliche Zahl und wir erhalten $\tilde{x}_{0,7} = \frac{1}{2}\left(x_{(7)} + x_{(8)}\right) = 955$. Interpretation: Mindestens 70% der gemessenen Werte sind kleiner oder gleich $x_{0,7}$, die restlichen 30% größer oder gleich.

ii) Es gilt $\bar{x} = \frac{1}{10}\sum_{j=1}^{10} x_j = 942{,}6$. Das arithmetische Mittel entspricht dem, was man allgemein als Durchschnittswert bezeichnet.

iii) Der Median entspricht dem 50%-Quantil, d.h. mindestens die Hälfte der gemessenen Werte ist größer oder gleich, die andere Hälfte kleiner oder gleich. Hier ist $n = 10$ und $\alpha = 0{,}5$ also $n \cdot \alpha = 5$ und es gilt $\tilde{x} = \frac{1}{2}\left(x_{(5)} + x_{(6)}\right) = 940$.

iv) Es gilt $s_x^2 = \frac{1}{9}\sum_{j=1}^{10}\left(x_j - \overline{x}\right)^2 = \frac{1}{9}\left(\sum_{j=1}^{10} x_j^2 - 10\overline{x}^2\right) = 407{,}6$. Beachte hierbei den Vorfaktor $\frac{1}{9}$ statt $\frac{1}{10}$ (allgemein $n-1$ statt n im Nenner)! Die Varianz ist ein Maß für die durchschnittliche Streuung der Daten, je höher sie ist, desto mehr Messwerte sind vom errechneten arithmetischen Mittel weit entfernt (im Extremfall $s_x^2 = 0$ würden also alle gemessenen Daten gleich sein und mit dem arithmetischen Mittel übereinstimmen).

v) Es gilt $s_x = +\sqrt{s_x^2} = 20{,}19$. Die Standardabweichung ist die Wurzel der Varianz, im Gegensatz zu ihr besitzt sie aber die selbe Maßeinheit wie die Daten (hier also Gramm), was eine Interpretation erleichtert. In unserem Fall weichen die Messergebnisse also durchschnittlich um ca. 20 Gramm vom errechneten arithmetischen Mittel ab.

vi) Der Quartilsabstand berechnet sich als empirisches 75%-Quantil minus empirisches 25%-Quantil. Wir berechnen diese beiden Größen zuerst: Mit $n = 10$ und $\alpha = \frac{1}{4}$ gilt $n \cdot \alpha = \frac{5}{2} \notin \mathbb{N}$ und $\lfloor n\alpha \rfloor + 1 = 3$. Somit folgt $\tilde{x}_{0.25} = x_{(3)}$. Analog errechnet man mit für $\alpha = \frac{3}{4}$ das Quantil $x_{0.75} = x_{(8)}$ und somit den Quartilsabstand als $x_{0.75} - x_{0.25} = 35$.

(b) Da der größte Messwert verfälscht wurde, hat sich die Reihenfolge der Messwerte nicht verändert. Damit bleiben die Größen aus i), iii), vi) unverändert, denn in ihrer Berechnung kommt $x_{(10)}$ nicht vor.

Aufgabe T3 (Optimierung bzgl. der absoluten Abweichung)

Es sei $x = (x_1, \ldots, x_n)$ die Stichprobe eines beliebigen (reellwertigen) Zufallsexperiments.

(a) Sei $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$ die geordnete Stichprobe und $\tilde{x}$ der empirische Median von x. Zeigen Sie für $a \in \mathbb{R}$:

$$\sum_{j=1}^{n} |x_j - \tilde{x}| = -\sum_{j=1}^{\lfloor \frac{n}{2} \rfloor} \left(x_{(j)} - a\right) + \sum_{j=\lfloor \frac{n+1}{2} \rfloor + 1}^{n} \left(x_{(j)} - a\right).$$

(b) Beweisen Sie mit Hilfe von (a) die Minimalitätseigenschaft

$$\sum_{j=1}^{n} |x_j - \tilde{x}| = \min_{a \in \mathbb{R}} \sum_{j=1}^{n} |x_j - a|.$$

Lösungsvorschlag:

(a) Wir machen eine Fallunterscheidung danach, ob n gerade ist oder nicht.

(1) Zunächst sei n ungerade, $\tilde{x} = x_{(\frac{n+1}{2})}$. Dann gilt:

$$\begin{aligned}
\sum_{j=1}^{n} |x_j - \tilde{x}| &= \sum_{j=1}^{\frac{n-1}{2}} (-1) \left(x_{(j)} - \tilde{x}\right) + \sum_{j=\frac{n+1}{2}}^{n} \left(x_{(j)} - \tilde{x}\right) \\
&= -\sum_{j=1}^{\frac{n-1}{2}} \left(x_{(j)} - a\right) - \frac{n-1}{2}(a - \tilde{x}) + \sum_{j=\frac{n+1}{2}+1}^{n} \left(x_{(j)} - a\right) + \frac{n-1}{2}(a - \tilde{x}) \\
&= -\sum_{j=1}^{\lfloor \frac{n}{2} \rfloor} \left(x_{(j)} - a\right) + \sum_{j=\lfloor \frac{n+1}{2} \rfloor + 1}^{n} \left(x_{(j)} - a\right).
\end{aligned}$$

(2) Für gerades n gilt $\tilde{x} = \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right)$ und somit:

$$\begin{aligned}
\sum_{j=1}^{n} |x_j - \tilde{x}| &= -\sum_{j=1}^{\frac{n}{2}} \left(x_{(j)} - \tilde{x}\right) + \sum_{j=\frac{n}{2}+1}^{n} \left(x_{(j)} - \tilde{x}\right) \\
&= -\sum_{j=1}^{\lfloor \frac{n}{2} \rfloor} \left(x_{(j)} - a\right) + \sum_{j=\lfloor \frac{n+1}{2} \rfloor + 1}^{n} \left(x_{(j)} - a\right).
\end{aligned}$$

(b) Wir zeigen die Aussage für ungerades n. Der Fall „n gerade“ geht analog.

$$\begin{aligned}
\sum_{j=1}^{n} |x_j - \tilde{x}| &\overset{\text{(i)}}{=} -\sum_{j=1}^{\frac{n-1}{2}} \left(x_{(j)} - a\right) + \sum_{j=\frac{n+1}{2}+1}^{n} \left(x_{(j)} - a\right) \\
&\leqslant -\sum_{\substack{j=1,\\ x_{(j)} \leqslant a}}^{n} \left(x_{(j)} - a\right) + \sum_{\substack{j=1,\\ x_{(j)} > a}}^{n} \left(x_{(j)} - a\right) = \sum_{j=1}^{n} \left|x_{(j)} - a\right|.
\end{aligned}$$

Aufgabe T4 (Empirische Verteilung)

Es sei $(x_j, y_j) \in \mathbb{R}^2$, $j \in [n]$, mit $n \geqslant 2$ eine Stichprobe paarweise verschiedener zweidimensionaler Daten. Wir definieren über dem Grundraum $\Omega = \{(x_j, y_j) \,|\, j \in [n]\}$ die gemeinsame Verteilung der Zufallsvariablen $X(x, y) = x$ und $Y(x, y) = y$ durch

$$\mathbb{P}(X = x_j, Y = y_j) := \frac{1}{n}, \qquad j \in [n],$$

die man auch die empirische Verteilung der Stichprobe (x_j, y_j), $j \in [n]$, nennt. Zeigen Sie:

(a) $\mathbb{E}[X] = \bar{x}$, $\mathbb{E}[Y] = \bar{y}$

(b) $\mathbb{V}(X) = \frac{n-1}{n} s_x^2$, $\mathbb{V}(Y) = \frac{n-1}{n} s_y^2$

(c) $\mathbb{C}(X, Y) = \frac{n-1}{n} \widetilde{s}_{xy}$ mit $\widetilde{s}_{xy} := \frac{1}{n-1} \sum_{j=1} (x_j - \bar{x}) \cdot (y_j - \bar{y})$

(d) Verifizieren Sie mithilfe von Satz 12.10 die Formel von Folie 13.32 für a^* und b^*, d.h. zeigen Sie, dass das Minimum

$$\min_{a,b \in \mathbb{R}} \mathbb{E}\left[(Y - a - bX)^2\right]$$

für die Parameter

$$b^* := \frac{\sum_{j=1} (x_j - \bar{x}) \cdot (y_j - \bar{y})}{\sum_{j=1}^{n} (x_j - \bar{x})^2},$$
$$a^* := \bar{y} - b^* \cdot \bar{x}$$

angenommen wird.

Lösungsvorschlag:

Wir zeigen die ersten beiden Aussagen aus Symmetriegründen nur für X.

(a) Es gilt

$$\mathbb{E}[X] = \sum_{j=1}^{n} x_j \mathbb{P}(X = x_j, Y = y_j) = \frac{1}{n} \sum_{j=1}^{n} x_j.$$

(b) Mit Teil (a) gilt

$$\begin{aligned}\mathbb{V}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] &= \sum_{j=1}^{n} (x_j - \mathbb{E}[X])^2 \mathbb{P}(X = x_j, Y = y_j) \\ &= \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2 \\ &= \frac{n-1}{n} \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})^2 \\ &= \frac{n-1}{n} s_x^2.\end{aligned}$$

(c) Es gilt

$$\begin{aligned}\mathbb{C}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] &= \sum_{j=1}^{n} (x_j - \mathbb{E}[X])(y_j - \mathbb{E}[Y]) \mathbb{P}(X = x_j, Y = y_j) \\ &= \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x}) \cdot (y_j - \bar{y}) = \frac{n-1}{n} \widetilde{s}_{xy}.\end{aligned}$$

(d) Nach Satz 12.10 wird der Ausdruck für die Parameter

$$\begin{aligned}\tilde{b} &= \frac{\mathbb{C}(X,Y)}{\mathbb{V}(X)},\\ \tilde{a} &= \mathbb{E}[Y] - \tilde{b}\cdot\mathbb{E}[X].\end{aligned}$$

minimiert. Wir zeigen also $b^* = \tilde{b}$ und $a^* = \tilde{a}$. Mit den vorangegangenen Aufgabenteilen folgt

$$\tilde{b} = \frac{\frac{1}{n}\sum_{j=1}(x_j-\bar{x})\cdot(y_j-\bar{y})}{\frac{n-1}{n}s_x^2} = \frac{\sum_{j=1}(x_j-\bar{x})\cdot(y_j-\bar{y})}{(n-1)\frac{1}{n-1}\sum_{j=1}^{n}(x_j-\bar{x})^2} = b^*$$

und somit

$$\tilde{a} = \bar{y} - \tilde{b}\cdot\bar{x} = \bar{y} - b^*\cdot\bar{x} = a^*.$$

Aufgabe T5 (Lineare Regression)

Angenommen, wir beobachten in vier verschiedenen Experimenten jeweils $n = 50$ zweidimensionale Datenpaare (x_i, y_i), $i = 1, \ldots, n$. Wir möchten diese nun auf einen linearen Zusammenhang untersuchen. Die folgende Abbildung enthält die Punktewolken, die sich aus den vier Experimenten ergeben haben:

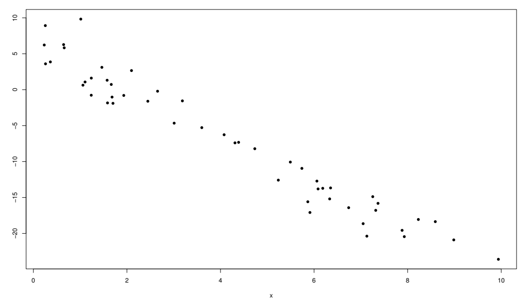


Abbildung 1: Daten Experiment 1

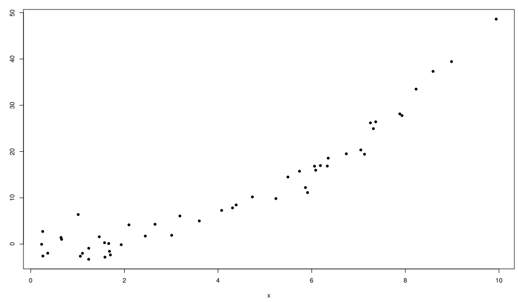


Abbildung 2: Daten Experiment 2

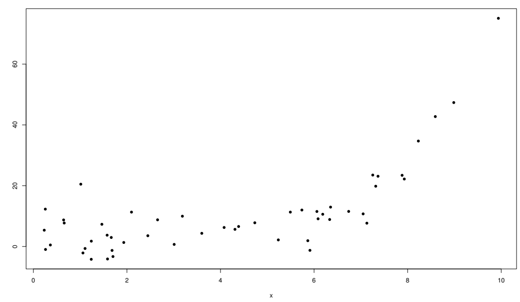


Abbildung 3: Daten Experiment 3

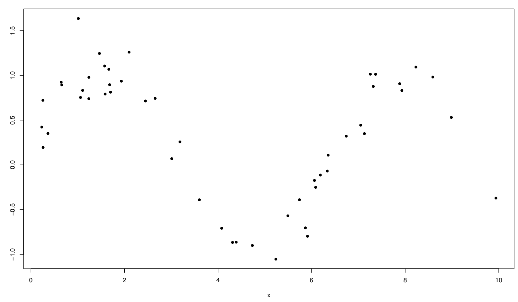


Abbildung 4: Daten Experiment 4

a) Skizzieren Sie in jeder der Abbildungen 1-4 eine Regressionsgerade. Bei welchen der vier Experimenten vermuten Sie einen linearen Zusammenhang?

b) Nun erhalten Sie als Zusatzinformation die Pearson-Korrelationskoeffizienten, die sich aus den vier Experimenten ergeben haben:

- Experiment 1: $r_{xy} = -0.97$
- Experiment 2: $r_{xy} = 0.94$
- Experiment 3: $r_{xy} = 0.68$
- Experiment 4: $r_{xy} = -0.3$

Passen diese Größen zu den Überlegungen aus Teil a) bzgl. eines möglichen linearen Zusammenhangs?

c) Um sicherzugehen, dass wirklich (k)ein linearer Zusammenhang vorliegt, betrachten Sie nun noch sogenannte Residuenplots. In diesen werden die angepassten Werte aus der Regressionsgeraden $y = \hat{a} + \hat{b}x$ mit den Residuen, also den Fehlern aus der Regression, in einem Schaubild aufgetragen. Genauer enthalten besagte Plots auf der x-Achse die Werte $\hat{a} + \hat{b}x_i$, $i = 1, \ldots, n$, und auf der y-Achse die Residuen $y_i - (\hat{a} + \hat{b}x_i)$, $i = 1, \ldots, n$. Die Plots sind nachfolgend in den Abbildungen 5-8 dargestellt. Bei welchen Experimenten vermuten Sie jetzt einen linearen Zusammenhang?

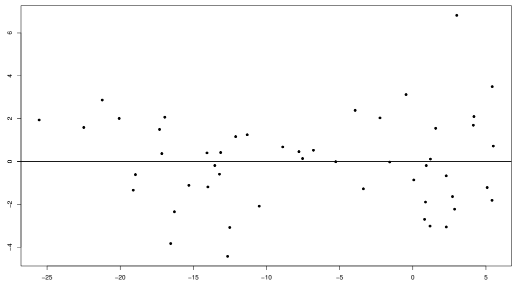


Abbildung 5: Residuen Experiment 1

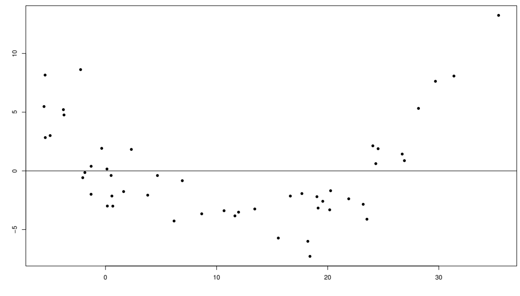


Abbildung 6: Residuen Experiment 2

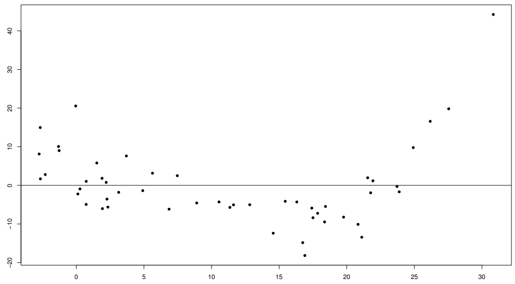


Abbildung 7: Residuen Experiment 3

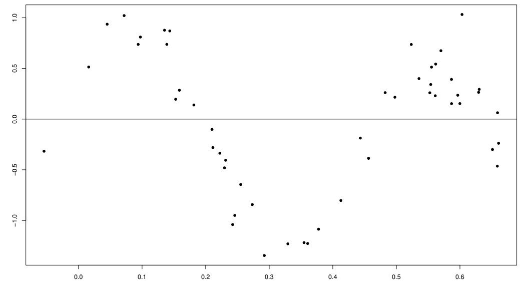


Abbildung 8: Residuen Experiment 4

Lösungsvorschlag:

a) Die Regressionsgeraden sind in den Abbildungen 9-12 dargestellt. Anhand der Bilder scheint das lineare Modell für Experiment 1 sehr gut zu passen. Für Experiment 4 passt es ganz klar nicht (Liegt hier vielleicht etwas Periodisches wie ein Sinus vor?). Bei Experiment 2 ist nicht klar ersichtlich, ob das lineare Modell passt oder nicht. In Experiment 3 scheint eher kein linearer Zusammenhang vorzuliegen.

b) In der Vorlesung haben wir gelernt, dass der Korrelationskoeffizient ein Maß für den affinen Zusammenhang zweier Stichproben darstellt. Für Experiment 1 und 2 ist r_{xy} nahe 1, was unsere Beobachtung eines passenden bzw. unklaren linearen Zusammenhangs unterstützt. Bei Experiment 3 und 4 haben wir die lineare Annahme nicht unterstützt, was auch durch die Korrelationskoeffizienten dargestellt wird. Wir sehen insbesondere unsere Beobachtung, dass das lineare Modell in Experiment 4 am schlechtesten passt.

c) Bei den Residuenplots sehen wir in Experiment 1 eine recht gleichmäßige Streuung um 0. Genau diese Beobachtung erwarten wir bei Vorliegen eines linearen Modells, sodass auch der Residuenplot unsere Vermutung eines linearen Zusammenhangs unterstützt. Bei Experiment 2 erkennen wir jetzt eine quadratische Form der Residuen, was klar gegen ein lineares Modell spricht. Anhand des Residuenplots können wir hier also davon ausgehen, dass kein lineares Modell vorliegt (obwohl r_{xy} nahe bei 1 liegt!). Bei Experiment 3 und 4 sehen wir unsere vorherige Vermutung, dass kein linearer Zusammenhang besteht, bestätigt.

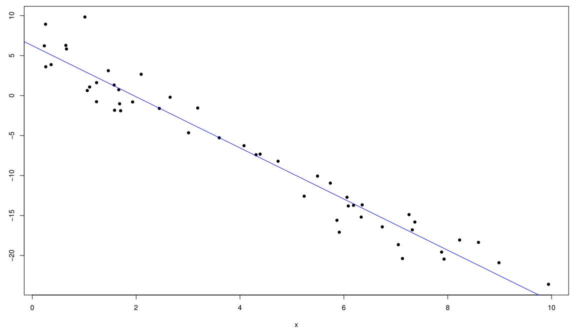


Abbildung 9: Experiment 1

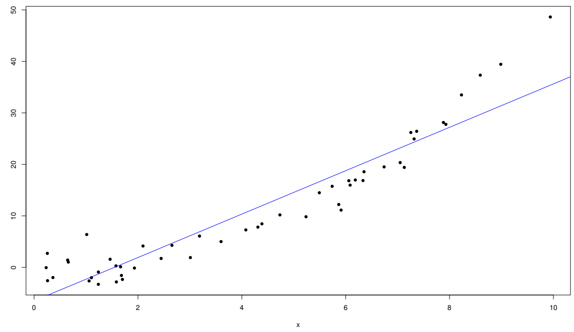


Abbildung 10: Experiment 2

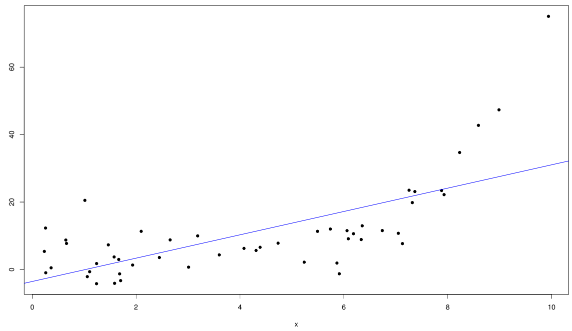


Abbildung 11: Experiment 3

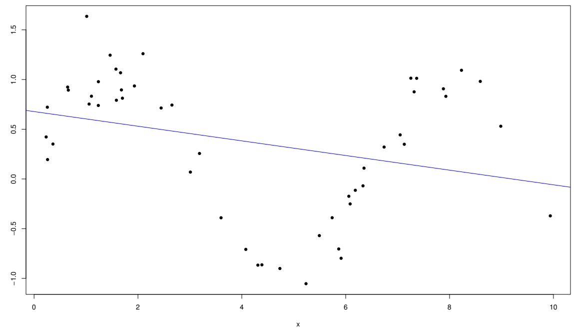


Abbildung 12: Experiment 4

Aufgabe T6 (Eindimensionale Randverteilungen eines multinomialverteilten Zufallsvektors)

Für $s \in \mathbb{N}_{\geqslant 2}$ und $p_j \in (0,1)$, $j \in [s]$, sei $(X_1, \ldots, X_s) \sim \text{Mult}(n, p_1, \ldots, p_s)$ ein multinomialverteilter Zufallsvektor. Zeigen Sie, dass die Zufallsvariablen $X_1, \ldots, X_s$ binomialverteilt sind und bestimmen Sie die zugehörigen Parameter.

Lösungsvorschlag:

Wir betrachten o.B.d.A. die Zufallsvariable X_1. Es sei $k_1 \in \{0, \ldots, n\}$. Dann gilt

$$\begin{aligned}
\mathbb{P}(X_1 = k_1) &= \sum_{\substack{k_2,\ldots,k_s \in \{0,\ldots,n\} \\ k_2+\ldots+k_s = n-k_1}} \mathbb{P}(X_1 = k_1, X_2 = k_2, \ldots, X_s = k_s) \\
&= \sum_{\substack{k_2,\ldots,k_s \in \{0,\ldots,n\} \\ k_2+\ldots+k_s = n-k_1}} p_1^{k_1} \cdot \ldots \cdot p_s^{k_s} \frac{n!}{k_1! \cdot \ldots \cdot k_s!} \\
&= \frac{p_1^{k_1}}{k_1!} \sum_{\substack{k_2,\ldots,k_s \in \{0,\ldots,n\} \\ k_2+\ldots+k_s = n-k_1}} p_2^{k_2} \cdot \ldots \cdot p_s^{k_s} \frac{n!}{k_2! \cdot \ldots \cdot k_s!} \\
&= \frac{n!}{k_1!(n-k_1)!} p_1^{k_1} \sum_{\substack{k_2,\ldots,k_s \in \{0,\ldots,n\} \\ k_2+\ldots+k_s = n-k_1}} p_2^{k_2} \cdot \ldots \cdot p_s^{k_s} \frac{(n-k_1)!}{k_2! \cdot \ldots \cdot k_s!} \\
&= \frac{n!}{k_1!(n-k_1)!} p_1^{k_1} (1-p_1)^{n-k_1} \underbrace{\sum_{\substack{k_2,\ldots,k_s \in \{0,\ldots,n\} \\ k_2+\ldots+k_s = n-k_1}} \left(\frac{p_2}{1-p_1}\right)^{k_2} \cdot \ldots \cdot \left(\frac{p_s}{1-p_1}\right)^{k_s} \frac{(n-k_1)!}{k_2! \cdot \ldots \cdot k_s!}}_{=1} \\
&= \frac{n!}{k_1!(n-k_1)!} p_1^{k_1} (1-p_1)^{n-k_1}.
\end{aligned}$$

Im letzten Schritt wurde dabei verwendet, dass die Summanden

$$\left(\frac{p_2}{1-p_1}\right)^{k_2} \cdot \ldots \cdot \left(\frac{p_s}{1-p_1}\right)^{k_s} \frac{(n-k_1)!}{k_2! \cdot \ldots \cdot k_s!}$$

gerade die Zähldichte eines $(s-1)$-dimensionalen, Mult $\left(n-k_1, \frac{p_2}{1-p_1}, \ldots, \frac{p_s}{1-p_1}\right)$-verteilten Zufallsvektors beschreiben, da

$$\sum_{j=2}^{s} \frac{p_j}{1-p_1} = \frac{1-p_1}{1-p_1} = 1$$

aufgrund der Voraussetzung $p_1 + p_2 + \ldots + p_s = 1$ gilt. Somit ist $X_1 \sim \mathrm{Bin}(n, p_1)$. Analog gilt natürlich $X_j \sim \mathrm{Bin}(n, p_j)$.

Aufgabe T7 (Bedingte Multinomialverteilung)

Es sei $(X_1, \ldots, X_s) \sim \mathrm{Mult}(n, p_1, \ldots, p_s)$, wobei $p_s < 1$ gelte. Zeigen Sie:

$$\mathbb{P}(X_1 = k_1, \ldots, X_{s-1} = k_{s-1} \mid X_s = k) = \mathbb{P}(Y_1 = k_1, \ldots, Y_{s-1} = k_{s-1}), \qquad k \in \{0, 1, \ldots, n-1\}$$

mit

$$(Y_1, \ldots, Y_{s-1}) \sim \mathrm{Mult}\left(n-k, \frac{p_1}{1-p_s}, \ldots, \frac{p_{s-1}}{1-p_s}\right).$$

Lösungsvorschlag:

Seien $k_1, \ldots, k_{s-1}$ und k jeweils nichtnegativ und es gelte $\sum\limits_{j=1}^{s-1} k_j = n-k$. Dann gilt:

$$\begin{aligned}\mathbb{P}(X_1 = k_1, \ldots, X_{s-1} = k_{s-1} | X_s = k) &= \frac{\mathbb{P}(X_1 = k_1, \ldots, X_{s-1} = k_{s-1}, X_s = k)}{\mathbb{P}(X_s = k)} \\ &= \frac{\frac{n!}{k_1! \ldots k_{s-1}! k!}}{\frac{n!}{k!(n-k)!}} \cdot \frac{p_1^{k_1} \cdot \ldots \cdot p_{s-1}^{k_{s-1}} \cdot p_s^k}{p_s^k \cdot (1-p_s)^{n-k}} = \frac{(n-k)!}{k_1! \cdot \ldots \cdot k_{s-1}!} \left(\frac{p_1}{1-p_s}\right)^{k_1} \cdot \ldots \cdot \left(\frac{p_{s-1}}{1-p_s}\right)^{k_{s-1}}.\end{aligned}$$

Hierbei wurde verwendet, dass X_s eine Binomialverteilung mit Parametern n und p_s besitzt (siehe Aufgabe T6). Also folgt nach Definition der Multinomialverteilung, dass dies gerade gleich der Wahrscheinlichkeit

$$\mathbb{P}(Y_1 = k_1, \ldots, Y_{s-1} = k_{s-1})$$

ist.

Intuitive Erklärung:
$(X_1, \ldots, X_s) \sim \mathrm{Mult}(n, p_1, \ldots, p_s)$ gibt die Anzahl der Treffer erster, …, s-ter Art in einer Reihe von n unabhängigen Versuchen an, mit Wahrscheinlichkeit p_j für einen Treffer j-ter Art.

Steht nun die Randbedingung $X_s = k$ fest (k Treffer s-ter Art sind schon bekannt), so bleiben für die restlichen Treffer erster, …, $(s-1)$-ter Art noch $n-k$ unabhängige Versuche.

Für die Trefferwahrscheinlichkeiten gilt $p_1 + \ldots + p_{s-1} = 1 - p_s$ und Reskalieren auf Gesamtmasse 1 unter Beibehaltung der Verhältnisse innerhalb von $(p_1, \ldots, p_{s-1})$ ergibt die neuen Trefferwahrscheinlichkeiten

$$\left(\frac{p_1}{1-p_s}, \ldots, \frac{p_{s-1}}{1-p_s}\right).$$