7. Übung: Einfache statistische Schätzverfahren

Aufgabe 1

Mögliche Parameter der Grundgesamtheit sind

a) s ☐
b) σ ■
c) N ■
d) n ☐
e) μ ■

Erläuterung:
Die Grundgesamtheit bezeichnet die Menge aller statistischen Einheiten (Merkmalsträger), beispielsweise die Menge aller möglichen Ereignisse. Somit ist die Anzahl aller Einheiten oder Ereignisse N ein Parameter der Grundgesamtheit (→ Antwort c) ist richtig). Die Grundgesamtheit kann einen Mittelwert und eine Streuung (und somit auch eine Standardabweichung) aufweisen (→ Antworten b) und e) sind ebenfalls richtig).

Aufgabe 2

Eigenschaften guter Schätzfunktionen sind:

a) Konsistenz ■
b) Konfidenz ☐
c) Effizienz ■
d) Variabilität ☐
e) Erwartungstreue ■

Erläuterung:
Schätzungen werden benutzt, um für unbekannte Parameter θ einer Wahrscheinlichkeitsverteilung oder für Funktionen $g(\theta)$ geeignete Näherungswerte zu erhalten. Eine Stichprobenfunktion $T_n = T_n(X) = \hat{\theta}(X)$ der zugehörigen mathematischen Stichprobe X heißt Punktschätzung oder Schätzfunktion. Sie hat folgende Eigenschaften (vgl. [1], S. 197):

- T_n heißt erwartungstreu (unbiased) für $g(\theta)$, falls $E(T_n) = g(\theta)$.
- $(T_n)_{n=1,2\ldots}$ ist asymptotisch erwartungstreu für $g(\theta)$, falls $\lim_{n\to\infty} E(T_n) = g(\theta)$.
- $(T_n)_{n=1,2\ldots}$ heißt (schwach) konsistent für $g(\theta)$, falls für beliebiges $\varepsilon > 0$ die Beziehung $\lim_{n\to\infty} P(|T_n - g(\theta)| < \varepsilon) = 1$ gilt.
- Sind T_n und s_n erwartungstreu für $g(\theta)$, so heißt T_n wirksamer (oder effizienter) als s_n genau dann, wenn $Var(T_n) < Var(s_n)$ gilt. Das heißt, die erwartungstreue Schätzfunktion T_n weist eine geringere Varianz auf als eine andere erwartungstreue Schätzfunktion s_n, die zum Vergleich herangezogen wird.

Aufgabe 3

Unter Erwartungstreue versteht man,

a) dass eine erwartungstreue Schätzfunktion im Vergleich zu anderen erwartungstreuen Schätzfunktionen die kleinste Varianz aufweist. ☐

b) dass die Wahrscheinlichkeit, dass der Schätzwert nahe dem wahren Parameter der Grundgesamtheit ist, mit wachsendem Stichprobenumfang gegen 1 strebt. ☐

c) dass der Erwartungswert der Schätzfunktion gleich dem wahren Parameter der Grundgesamtheit ist. ■

d) dass der Erwartungswert der Schätzfunktion mit wachsendem n gegen den wahren Parameter der Grundgesamtheit strebt. ☐

e) dass der Bias (Verzerrung) der Schätzfunktion gleich 0 ist. ■

Erläuterung:
Eine Schätzung $\hat{\theta}$ gilt als erwartungstreu oder unverzerrt, wenn der Erwartungswert (Mittelwert) von $\hat{\theta}$ gleich dem zu schätzenden Parameter θ ist, d.h. $E(\hat{\theta}) = \theta$ → Antwort c) ist korrekt.

Die Verzerrung ist die Abweichung vom Erwartungswert der Schätzfunktion $E(\hat{\theta})$ zum wahren Parameter θ. Wenn die Schätzung erwartungstreu ist, gilt:

$$E(\hat{\theta}) = \theta \Leftrightarrow E(\hat{\theta}) - \theta = 0$$

Also ist die Verzerrung (Bias) der Schätzfunktion gleich 0 → Antwort e) ist ebenfalls korrekt.

Aufgabe 4

Unter Konsistenz versteht man,

a) dass eine erwartungstreue Schätzfunktion im Vergleich zu anderen erwartungstreuen Schätzfunktionen die kleinste Varianz aufweist. ☐

b) dass die Wahrscheinlichkeit, dass der Schätzwert nahe dem wahren Parameter der Grundgesamtheit ist, mit wachsendem Stichprobenumfang gegen 1 strebt. ■

c) dass der Erwartungswert der Schätzfunktion gleich dem wahren Parameter der Grundgesamtheit ist. ☐

d) dass der Erwartungswert der Schätzfunktion mit wachsendem n gegen den wahren Parameter der Grundgesamtheit strebt. ☐

e) dass der Bias (Verzerrung) der Schätzfunktion gleich 0 ist. ☐

Erläuterung:
Es wird angenommen, dass eine Schätzung $\hat{\theta}$ (schwach) konsistent sein soll. Das bedeutet, dass für $\hat{\theta}$ bei einem geringen Stichprobenumfang die Wahrscheinlichkeitsverteilung noch einen großen Abstand zum wahren, aber unbekannten Parameter θ aufweist. Mit wachsendem Stichprobenumfang n konvergiert die Wahrscheinlichkeitsverteilung gegen den Parameter θ. Die Wahrscheinlichkeit dafür, dass der Schätzwert nahe dem wahren Parameter der Grundgesamtheit ist, strebt somit mit wachsendem Stichprobenumfang gegen 1 → Antwort b) ist korrekt.

Aufgabe 5

Unter Effizienz versteht man,

a) dass eine erwartungstreue Schätzfunktion im Vergleich zu anderen erwartungstreuen Schätzfunktionen die kleinste Varianz aufweist. ■

b) dass die Wahrscheinlichkeit dafür, dass der Schätzwert nahe dem wahren Parameter der Grundgesamtheit liegt, mit wachsendem Stichprobenumfang gegen 1 strebt. ☐

c) dass der Erwartungswert der Schätzfunktion gleich dem wahren Parameter der Grundgesamtheit ist. ☐

d) dass der Erwartungswert der Schätzfunktion mit wachsendem n gegen den wahren Parameter der Grundgesamtheit strebt. ☐

e) dass der Bias (Verzerrung) der Schätzfunktion gleich 0 ist. ☐

Erläuterung:
Bei der Betrachtung der Effizienz betrachtet man die Varianz zweier (asymptotischer) Schätzungen $\hat{\theta}_1$ und $\hat{\theta}_2$. Ist die Varianz der Schätzung $\hat{\theta}_1$ kleiner als die der Schätzung $\hat{\theta}_2$ ($Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$), so gilt $\hat{\theta}_1$ als wirksamer (oder effizienter) gegenüber der Schätzung $\hat{\theta}_2$. Das Verhältnis

$$\eta = \frac{Var(\hat{\theta}_1)}{Var(\hat{\theta}_2)}$$

heißt Wirkungsgrad von $\hat{\theta}_2$ in Bezug auf $\hat{\theta}_1$ → Antwort a) ist richtig.

Aufgabe 6

Wie lautet die Alternativhypothese H_1 eines zweiseitigen Mittelwerts?

a) $H_1 : \mu = \mu_0$ ☐

b) $H_1 : \mu \neq \mu_0$ ■

c) $H_1 : \mu > \mu_0$ ☐

d) $H_1 : \mu \leq \mu_0$ ☐

e) $H_1 : \mu < \mu_0$ ☐

Erläuterung:
Die Nullhypothese H_0 eines zweiseitigen Mittelwerts lautet (vgl. [1], S. 202):

$$H_0 : \mu = \mu_0$$

Damit lautet die Alternativhypothese H_1 (vgl. [1], S. 202):

$$H_1 : \mu \neq \mu_0$$

Aufgabe 7

Wie lautet die Alternativhypothese H_1 eines linksseitigen Mittelwerts?

a) $H_1 : \mu = \mu_0$ ☐
b) $H_1 : \mu \neq \mu_0$ ☐
c) $H_1 : \mu > \mu_0$ ☐
d) $H_1 : \mu \leq \mu_0$ ☐
e) $H_1 : \mu < \mu_0$ ■

Erläuterung:
Die Nullhypothese H_0 eines linksseitigen Mittelwerts lautet (vgl. [1], S. 202):

$$H_0 : \mu \geq \mu_0$$

Damit lautet die Alternativhypothese H_1 (vgl. [1], S. 202):

$$H_1 : \mu < \mu_0$$

Aufgabe 8

Wie lautet die Alternativhypothese H_1 eines rechtsseitigen Mittelwerts?

a) $H_1 : \mu = \mu_0$ ☐
b) $H_1 : \mu \neq \mu_0$ ☐
c) $H_1 : \mu > \mu_0$ ■
d) $H_1 : \mu \leq \mu_0$ ☐
e) $H_1 : \mu < \mu_0$ ☐

Erläuterung:
Die Nullhypothese H_0 eines rechtsseitigen Mittelwerts lautet (vgl. [1], S. 202):

$$H_0 : \mu \leq \mu_0$$

Damit lautet die Alternativhypothese H_1 (vgl. [1], S. 202):

$$H_1 : \mu > \mu_0$$

Aufgabe 9

Der Fehler 1. Art α

a) besteht darin, die Nullhypothese fälschlicherweise abzulehnen. ■

b) besteht darin, die Nullhypothese fälschlicherweise beizubehalten. □

c) kann nur gemacht werden, wenn die Nullhypothese durch den Test abgelehnt wurde. ■

d) kann nur gemacht werden, wenn die Nullhypothese durch den Test beibehalten wurde. □

e) entspricht dem Signifikanzniveau. ■

Erläuterung:
Zur Lösung und zum besseren Verständnis verwende man folgende Entscheidungsstruktur:

Entscheidung	Tatsächliche (unbekannte) Situation	
	H_0 ist wahr	H_0 ist nicht wahr
H_0 wird abgelehnt	Fehler erster Art	richtige Entscheidung
H_0 wird nicht abgelehnt	richtige Entscheidung	Fehler zweiter Art

Daraus ist abzulesen, dass der Fehler erster Art (α) nur dann vorliegen kann, wenn die Nullhypothese fälschlicherweise abgelehnt wird. Um das zu überprüfen, muss folglich der Test gemacht werden, welcher dann zur Ablehnung der Nullhypothese führt (das wiederum bedeutet, dass die Nullhypothese entgegen der Entscheidung doch richtig ist).

Aufgabe 10

Der Fehler 2. Art β

a) besteht darin, die Nullhypothese fälschlicherweise abzulehnen. □

b) besteht darin, die Nullhypothese fälschlicherweise beizubehalten. ■

c) kann nur gemacht werden, wenn die Nullhypothese durch den Test abgelehnt wurde. □

d) kann nur gemacht werden, wenn die Nullhypothese durch den Test beibehalten wurde. ■

e) entspricht dem Signifikanzniveau. □

Erläuterung:
Zur Lösung und zum besseren Verständnis verwende man folgende Entscheidungsstruktur:

Entscheidung	Tatsächliche (unbekannte) Situation	
	H_0 ist wahr	H_0 ist nicht wahr
H_0 wird abgelehnt	Fehler erster Art	richtige Entscheidung
H_0 wird nicht abgelehnt	richtige Entscheidung	Fehler zweiter Art

Daraus ist abzulesen, dass der Fehler zweiter Art (β) nur dann vorliegen kann, wenn die Nullhypothese fälschlicherweise nicht abgelehnt wird. Um das zu überprüfen, muss folglich der Test gemacht werden. Ergibt der Test dann, dass die Nullhypothese entgegen der Entscheidung doch falsch ist, liegt ein Fehler zweiter Art vor.

Aufgabe 11

Sei X in seiner Grundgesamtheit normalverteilt und seine Varianz bekannt, dann gilt für die Verteilung des Stichprobenmittelwertes:

a) $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0;1)$ ■

b) $T = \frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t(n-1)$ ☐

c) $Z = \frac{\bar{X} - \mu}{\sigma} \sim N(0;1)$ ☐

d) $T = \frac{\bar{X} - \mu}{s} \sim t(n-1)$ ☐

e) $T = \frac{\bar{X} - \mu}{s / \sqrt{n}} \sim N(0;1)$ ☐

Erläuterung:
Schritt 1) Um zu entscheiden, ob man von einer t- oder Normalverteilung ausgehen muss, muss man betrachten, ob die zur Standardisierung des Mittelwerts benötigte Varianz des Merkmals bekannt oder unbekannt ist. Im vorliegenden Fall ist die Varianz bekannt, also muss die Verteilung des Stichprobenmittelwerts einer Normalverteilung folgen.

Schritt 2) Da es sich um die Verteilung des Stichprobenmittelwerts handelt, muss bei der durchgeführten Transformation $\bar{X} - \mu$ durch die (bekannte) Varianz des Stichprobenmittelwerts geteilt werden. Für diese gilt:

$$\sigma_{\bar{x}} = \sigma / \sqrt{n}$$

Somit bleibt nur noch Lösungsmöglichkeit a) übrig.

Aufgabe 12

Sei X in seiner Grundgesamtheit normalverteilt, seine Varianz aber unbekannt, dann gilt für die Verteilung des Stichprobenmittelwertes:

a) $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0;1)$ ☐

b) $T = \frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t(n-1)$ ■

c) $Z = \frac{\bar{X} - \mu}{\sigma} \sim N(0;1)$ ☐

d) $T = \frac{\bar{X} - \mu}{s} \sim t(n-1)$ ☐

e) $T = \frac{\bar{X} - \mu}{s / \sqrt{n}} \sim N(0;1)$ ☐

Erläuterung:
Schritt 1) Um zu entscheiden, ob man von einer t- oder Normalverteilung ausgehen muss, muss man betrachten, ob die zur Standardisierung des Mittelwerts benötigte Varianz des Merkmals bekannt oder unbekannt ist. Im vorliegenden Fall ist die Varianz unbekannt, also muss die Verteilung des Stichprobenmittelwerts einer t-Verteilung folgen.

Schritt 2) Da es sich um die Verteilung des Stichprobenmittelwerts handelt, muss bei der durchgeführten Transformation $\bar{X} - \mu$ durch die (unbekannte) Varianz des Stichprobenmittelwerts geteilt werden. Für diese gilt:

$$s_{\bar{x}} = s / \sqrt{n}$$

Somit bleibt nur noch Lösungsmöglichkeit b) übrig.

Aufgabe 13

Welche Formel für die Stichprobenvarianz ist erwartungstreu?

a) $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad falls\ \mu\ unbekannt$ ☐

b) $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad falls\ \mu\ unbekannt$ ■

c) $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \quad falls\ \mu\ bekannt$ ■

d) $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2 \quad falls\ \mu\ bekannt$ ☐

e) $s^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 \quad falls\ \mu\ unbekannt$ ☐

Erläuterung:
Ist der Erwartungswert μ bekannt, werden die Abweichungsquadrate unter Zugrundelegung des Erwartungswerts und nicht auf Basis des arithmetischen Mittels (empirischer Mittelwert) $\bar{X}$ ermittelt. Außerdem ist bei der Berechnung der (erwartungstreuen) Stichprobenvarianz die Summe der Abweichungsquadrate auf die gesamte Stichprobenmenge n zu normieren und nicht auf $n-1$.

➔ Antwort c) erfüllt diese Bedingungen.

Ist der Erwartungswert μ hingegen unbekannt, werden die Abweichungsquadrate unter Zugrundelegung des arithmetischen Mittels (empirischer Mittelwert) $\bar{X}$ ermittelt. Außerdem ist bei der Berechnung der (erwartungstreuen) Stichprobenvarianz die Summe der Abweichungsquadrate nicht auf die gesamte Stichprobenmenge n zu normieren, sondern auf $n-1$. Dies hängt damit zusammen, dass durch die Schätzung des Stichprobenmittelwerts die Zahl der Freiheitsgrade um 1 reduziert wird.

➔ Antwort b) erfüllt diese Bedingungen.

Aufgabe 14

Der Tonerverbrauch eines Laserdruckers soll kontrolliert werden. Dazu wird die Anzahl der Seiten (X), die mit einem neuen Toner gedruckt werden kann, protokolliert. Mit den ersten sechs Tonern erziele man folgende Seitenzahlen:

3.450 2.900 3.100 2.850 3.100 3.200

Die Anzahl der gedruckten Seiten kann als näherungsweise normalverteilt angesehen werden.

a) Kann die Anzahl der gedruckten Seiten exakt normalverteilt sein?

b) Führen Sie auf der Grundlage der gegebenen Daten eine erwartungstreue Punktschätzung für $E(X) = \mu$ bzw. $Var(X) = \sigma^2$ durch.

c) Geben Sie explizit das Konfidenzintervall für den tatsächlichen Durchschnittswert μ zum Konfidenzintervall $1 - \alpha$ an.

d) Bestimmen Sie das 95%-Konfidenzintervall für die durchschnittliche Anzahl der mit einem Toner bedruckbaren Seiten.

e) Verbalisieren Sie das Ergebnis von d).

f) Welche Möglichkeiten haben Sie, bei der Planung einer solchen Untersuchung Einfluss auf die Länge des Schätzintervalls zu nehmen?

g) Wie würde sich das Schätzintervall verändern, wenn σ^2 als bekannt vorausgesetzt werden kann?

h) Wie viele Toner muss man verbrauchen, damit bei 90% aller Stichproben die durchschnittlich erzielte Seitenzahl um höchstens 50 von der wahren durchschnittlichen Seitenzahl abweicht, wenn die wahre Standardabweichung 250 [Seiten] beträgt?

i) An einem Lehrstuhl sollen in Kürze vier Diplomarbeiten abgegeben werden. Aus Erfahrung weiß man, dass die Anzahl der Seiten einer Diplomarbeit näherungsweise normalverteilt ist mit $\mu = 80$ und $\sigma^2 = 225$.
   i. Wie ist die Gesamtzahl der Seiten von vier Diplomarbeiten verteilt?
   ii. Mit welcher Wahrscheinlichkeit müssen bei den vier Diplomarbeiten zusammen mindestens 350 Seiten gelesen werden?

Lösung:

a) Der Wertebereich einer Normalverteilung geht von $-\infty$ bis $+\infty$, auch wenn die Wahrscheinlichkeiten an den Rändern der Verteilung sehr gering sind. Die Anzahl der gedruckten Seiten kann daher niemals exakt normalverteilt sein, weil es keine negativen Seitenzahlen gibt.

b) Schritt 1) Aus der Stichprobentheorie folgt, dass das Stichprobenmittel (vgl. [1], S. 197)

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

eine erwartungstreue Schätzung von $E(X) = \mu$ ist, denn es gilt: $E(\bar{X}) = \mu$.

Schritt 2) Ebenso gilt für die Stichprobenvarianz s^2 mit (vgl. [1], S. 197)

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

dass s^2 eine erwartungstreue Schätzung von σ^2 ist (wegen $E(s^2) = \sigma^2$).

Schritt 3) Zur Lösung der Aufgabe werden diese beiden Schätzfunktionen benutzt und alle gegebenen Werte X_i eingesetzt:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{1}{6}\sum_{i=1}^{6} X_i$$
$$= \frac{1}{6}(3450 + 2900 + 3100 + 2850 + 3100 + 3200) = \frac{18600}{6} = 3100$$

Der Schätzwert des Erwartungswerts beträgt 3100.

Für die Schätzung der Stichprobenvarianz gilt damit:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2 = \frac{1}{6-1}\sum_{i=1}^{6}(X_i - \bar{X})^2$$

$$= \frac{1}{5}[(3450 - 3100)^2 + (2900 - 3100)^2 + \cdots + (3200 - 3100)^2]$$

$$= \frac{235000}{5} = 47000$$

Der Schätzwert der Varianz beträgt 47000.

c) Für das zweiseitige Konfidenzintervall des Erwartungswerts μ gilt bei unbekannter Varianz der Grundgesamtheit (diese musste ja gerade geschätzt werden) (vgl. [1], S. 200):

$$\left[\bar{X} - t_{df;1-\frac{\alpha}{2}} * \frac{s}{\sqrt{N}}; \bar{X} + t_{df;1-\frac{\alpha}{2}} * \frac{s}{\sqrt{N}}\right]$$

Als Formel ausgedrückt:

$$P\left(\bar{X} - t_{df;1-\frac{\alpha}{2}} * \frac{s}{\sqrt{N}} \leq \mu \leq \bar{X} + t_{df;1-\frac{\alpha}{2}} * \frac{s}{\sqrt{N}}\right) = 1 - \alpha$$

Das bedeutet, dass der wahre Wert von μ mit einer Wahrscheinlichkeit von $1 - \alpha$ in dem gegebenen Konfidenzintervall liegt. Dabei ist $t_{df;1-\frac{\alpha}{2}}$ ein Ausdruck für das $\left(1 - {}^{\alpha}/_{2}\right)$-Quantil der t_{N-1}-Verteilung mit $df = N - 1$ Freiheitsgraden. α wird als die Irrtumswahrscheinlichkeit bezeichnet und ist ein Ausdruck des Risikos, das bei der jeweiligen Schätzung eingegangen wird.

d) Schritt 1) Ermittlung aller unbekannten Parameter. Bei einem Konfidenzniveau von 95 % gilt:

$$Konfidenzniveau = 1 - \alpha = 0{,}95 \rightarrow \alpha = 1 - 0{,}95 = 0{,}05$$

$$1 - \frac{\alpha}{2} = 1 - \frac{0{,}05}{2} = 1 - 0{,}025 = 0{,}975$$

Für die Standardabweichung gilt:

$$s = \sqrt{s^2} = \sqrt{47000} \approx 216{,}79$$

Für die Freiheitsgrade gilt:

$$df = N - 1 = 6 - 1 = 5$$

Schritt 2) Bestimmen des Quantils $t_{N-1;1-\frac{\alpha}{2}}$ der t-Verteilung aus der Verteilungstabelle:

$$t_{df;1-\frac{\alpha}{2}} = t_{5;0{,}975}$$

n	...	0,95	0,975	0,99	...
⋮	⋮	⋮	⋮	⋮	...
4	⋮	2,1318	2,7764	3,7469	...
5	⋮	2,0150	2,5706	3,3649	...
6	⋮	1,9432	2,4469	3,1427	...
⋮	⋮	⋮	⋮	⋮	

Damit gilt:

$$t_{df;1-\frac{\alpha}{2}} = t_{5;0{,}975} = 2{,}5706$$

Schritt 3) Berechnen des gesuchten Konfidenzintervalls:

$$\left[\bar{X} - t_{df;1-\frac{\alpha}{2}} * \frac{s}{\sqrt{N}}; \bar{X} + t_{df;1-\frac{\alpha}{2}} * \frac{s}{\sqrt{N}}\right]$$
$$= \left[\bar{X} - t_{5;0{,}0975} * \frac{s}{\sqrt{N}}; \bar{X} + t_{5;0{,}0975} * \frac{s}{\sqrt{N}}\right]$$

$$= \left[3100 - 2{,}5706 * \frac{216{,}79}{\sqrt{6}}; 3100 + 2{,}5706 * \frac{216{,}79}{\sqrt{6}}\right]$$

$$\approx [3100 - 227{,}51; 3100 + 227{,}51] = [2872{,}49; 3327{,}51]$$

Das gesuchte 95 % Konfidenzintervall lautet [2872,49; 3327,51].

e) Mit einer Sicherheit von 95 % überdeckt das berechnete Konfidenzintervall zwischen $g_u = 2872{,}49$ und $g_o = 3327{,}51$ Seiten den wahren Parameter μ für die Anzahl der mit einem neuen Toner bedruckbaren Seiten.

f) Für die Länge l des Schätzintervalls gilt folgende Formel:

$$l = g_o - g_u = \bar{X} + t_{df;1-\frac{\alpha}{2}} * \frac{s}{\sqrt{N}} - \left(\bar{X} - t_{df;1-\frac{\alpha}{2}} * \frac{s}{\sqrt{N}}\right) = 2 * t_{df;1-\frac{\alpha}{2}} * \frac{s}{\sqrt{N}}$$

Darin finden sich also 5 Parameter, die vor der Untersuchung nicht bestimmt sind:

- Standardabweichung s: Ist die Realisierung einer Zufallsvariablen aufgrund der Stichprobenziehung (also zufällig).
- Quantil $t_{df;1-\frac{\alpha}{2}}$: Ergibt sich aus der Anzahl der Freiheitsgrade df und dem Wert $1-\frac{\alpha}{2}$.
- Anzahl Freiheitsgrade $df = N - 1$: Ist abhängig von der Größe der Stichprobe.
- $1-\frac{\alpha}{2}$: Ergibt sich aus dem festgelegten Konfidenzniveau.
- Größe der Stichprobe N

Die Länge des Schätzintervalls kann also nur durch die Größe der Stichprobe N und durch das vorgegebene Konfidenzniveau α beeinflusst werden.

g) Bei bekannter Varianz der Grundgesamtheit gilt für das zweiseitige Konfidenzintervall des Erwartungswerts μ (vgl. [1], S. 200):

$$\left[\bar{X} - z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{N}}; \bar{X} + z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{N}}\right]$$

Es ändert sich also lediglich der Faktor $t_{N-1;1-\frac{\alpha}{2}}$ zu $z_{1-\frac{\alpha}{2}} = z_{0,975}$, d.h. es muss anstelle der Tabelle der t-Verteilung die der Standardnormalverteilung benutzt werden. Das gesuchte Quantil $z_{0,975}$ erhält man durch Ablesen des z-Wertes mit $\Phi(z) = 0{,}975$, also für $z = 1{,}96$:

z	…	0,05	0,06	0,07	…
⋮	⋮	⋮	⋮	⋮	…
1,8	⋮	0,9678	0,9686	0,9693	…
1,9	⋮	0,9744	0,9750	0,9756	…
2,0	⋮	0,9798	0,9803	0,9808	…
⋮	⋮	⋮	⋮	⋮	

Dann erhält man unter der Annahme, dass die bekannte Varianz σ^2 der Schätzung der Varianz s^2 entspricht, für das gesuchte Konfidenzintervall:

$$\left[\bar{X} - z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{N}}; \bar{X} + z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{N}}\right] = \left[\bar{X} - z_{0,975} * \frac{s}{\sqrt{N}}; \bar{X} + z_{0,975} * \frac{s}{\sqrt{N}}\right]$$

$$= \left[3100 - 1{,}96 * \frac{216{,}79}{\sqrt{6}}; 3100 + 1{,}96 * \frac{216{,}79}{\sqrt{6}}\right]$$

$$\approx [3100 - 173{,}47; 3100 + 173{,}47] = [2926{,}53; 3273{,}47]$$

Man sieht, dass das Konfidenzintervall aufgrund des kleineren Faktors $z_{1-\frac{\alpha}{2}}$ kleiner wird.

h) Gesucht ist der notwendige Stichprobenumfang (also die Anzahl der Toner N), für die mit einer Sicherheitswahrscheinlichkeit von 90 % der wahre Mittelwert der gedruckten Seiten um maximal 50 Seiten (Schätzfehler $e \leq 50$) von der wahren Standardabweichung $\sigma = 250$ abweicht. Als Formel ausgedrückt:

$$P(|\bar{X} - \mu| \leq e) = 1 - \alpha = 0{,}9 \rightarrow \alpha = 0{,}1$$

Für das zweiseitige Konfidenzintervall gilt bei bekannter Varianz der Grundgesamtheit (vgl. [1], S. 200):

$$P\left(\bar{X} - z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{X} + z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{N}}\right) = 1 - \alpha$$

Die Länge l entspricht der doppelten Länge des Schätzfehlers. Analog zu Aufgabenteil f) gilt für die Länge:

$$2 * e = l = 2 * z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{N}}$$

Diese Gleichung kann dann (mit $\alpha = 0{,}1$) nach N umgeformt werden:

$$\Leftrightarrow z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{N}} = e$$

$$\Leftrightarrow N = \frac{\sigma^2 z^2_{1-\frac{0{,}1}{2}}}{e^2} = \frac{\sigma^2 z^2_{1-0{,}05}}{e^2} = \frac{\sigma^2 z^2_{0{,}95}}{e^2}$$

Das gesuchte Quantil $z_{0{,}95}$ erhält man durch Ablesen des z-Wertes mit $\Phi(z) = 0{,}95$ aus der Standardnormalverteilungstabelle, also für $z = 1{,}64$:

z	…	0,03	0,04	0,05	…
⋮	⋮	⋮	⋮	⋮	…
1,5	⋮	0,9370	0,9382	0,9394	…
1,6	⋮	0,9484	0,9495	0,9505	…
1,7	⋮	0,9582	0,9591	0,9599	…
⋮	⋮	⋮	⋮	⋮	

Zusammen mit $e \leq 50$ folgt dann für den mindestens notwendigen Stichprobenumfang:

$$N = \frac{\sigma^2 z^2_{0{,}95}}{e^2} \geq \frac{250^2 * 1{,}64^2}{50^2} = 67{,}24$$

Da es nur ganzzahlige Stichprobenumfänge gibt, muss die Zahl auf die nächste ganze Zahl aufgerundet werden. Es müssen also $N = 68$ Toner verbraucht werden.

i)

(i) Es sei Y die Gesamtzahl der Seiten aller vier Diplomarbeiten. Jede einzelne Diplomarbeit ist näherungsweise normalverteilt mit:

$$X_i \sim N(\mu; \sigma) = N(80; \sqrt{225})$$

Da Y eine Linearkombination normalverteilter Zufallsgrößen ist, gilt für ihren Erwartungswert und ihre Varianz:

$$\mu_y = \sum_{i=1}^{n} \mu_i = 4 * \mu = 4 * 80 = 320$$

$$\sigma_y^2 = \sum_{i=1}^{n} \sigma_i^2 = 4 * \sigma^2 = 4 * 225 = 900$$

Y ist also näherungsweise normalverteilt mit:

$$Y \sim N\left(320; \sqrt{900}\right) = N(320; 30)$$

(ii) Gesucht ist die Wahrscheinlichkeit, dass bei vier Diplomarbeiten insgesamt über 350 Seiten gelesen werden müssen. Dazu muss Y über eine Z-Transformation $Z = \frac{Y-\mu}{\sigma}$ in die Standardnormalverteilung überführt werden. Anschließend kann der gesuchte Wahrscheinlichkeitswert aus der Tabelle der Standardnormalverteilung abgelesen werden:

$$P(Y > 350) = 1 - P(Y \leq 350) = 1 - P\left(Z = \frac{Y-\mu}{\sigma} \leq \frac{350-320}{30}\right)$$

$$= 1 - P\left(Z \leq \frac{30}{30}\right) = 1 - P(Z \leq 1) = 1 - \Phi(1)$$

Für $\Phi(1)$ gilt:

z	0,00	0,01	…
⋮	⋮	⋮	…
0,9	0,8159	0,8186	…
1,0	0,8413	0,8438	…
1,1	0,8643	0,8665	…
⋮	⋮	⋮	

Also gilt für die gesuchte Wahrscheinlichkeit:

$$P(Y > 350) = 1 - \Phi(1) = 1 - 0{,}8413 = 0{,}1587$$

Es müssen also mit einer Wahrscheinlichkeit von 15,87 % mehr als 350 Seiten gelesen werden.

Aufgabe 17

Der Wasserverbrauch einer alten Waschmaschine soll untersucht werden. Man weiß, dass der Wasserverbrauch näherungsweise normalverteilt ist. Bei einer Versuchsreihe mit $N = 10$ Durchgängen wurde folgendes Ergebnis [in Liter] erzielt:

55 69 40 58 69 62 48 65 45 59

Weiterhin wird für diese Aufgabe eine Irrtumswahrscheinlichkeit von 5% angenommen.

a) Wie lauten die Hypothesen für den Test auf einen mittleren Wasserverbrauch von 50 Litern?

b) Geben Sie die zugrunde liegende Stichprobenfunktion formal und verbal an.

c) Welche Verteilung hat die Stichprobenfunktion unter H_0?

d) Wie lautet die Testfunktion, und wie ist diese unter H_0 verteilt?

e) Bestimmen Sie den Ablehnungsbereich für diesen Test.

f) Wie lautet Ihre Testentscheidung (Interpretation)?

g) Welcher Fehler kann bei dieser Entscheidung (Aufgabenteil f) unterlaufen sein?

h) Wiederholen Sie den Test an derselben Stichprobe, wenn man in der Gebrauchsanweisung den Hinweis gefunden hätte, dass die Waschmaschine höchstens 50 [Liter] verbraucht.

i) Welcher Fehler kann bei dieser Entscheidung (Aufgabenteil h) unterlaufen sein?

j) Interpretieren Sie eine Irrtumswahrscheinlichkeit von $\alpha = 5\%$ für diese Aufgabe am Beispiel.

Lösung:

a) Im vorliegenden Fall muss eine Hypothese bzgl. des Erwartungswerts aufgestellt werden. Die entsprechende Nullhypothese sagt aus, dass der mittlere Wasserverbrauch 50 Litern entspricht (vgl. [1], S. 201):

$$H_0 : \mu = 50$$

Die Alternativhypothese (auch Gegenhypothese) steht der Nullhypothese entgegen. Da keine spezifischen Zusatzinformationen darüber vorliegen, welche Richtung die Alternativhypothese vorgibt (mehr oder weniger als 50 Liter), muss ein zweiseitiger Test durchgeführt werden. Die Alternativhypothese lautet daher (vgl. [1], S. 201):

$$H_1 : \mu \neq 50$$

b) Die zugrunde liegende Stichprobenfunktion des Erwartungswerts μ ist das arithmetische Mittel $\bar{X}$. Es bedeutet: „Der mittlere Wasserverbrauch der alten Waschmaschine [in Liter] bei einer Zufallsstichprobe von $N = 10$ Durchgängen“. Dabei ist X_i der „Wasserverbrauch [in Liter] im i-ten Durchgang“. Für $\bar{X}$ gilt (vgl. [1], S. 174):

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

c) Unter der Annahme der Richtigkeit der Nullhypothese H_0 ist $\bar{x}$ normalverteilt mit $\mu = 50$. Für den Standardfehler des arithmetischen Mittels gilt:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{N} = \frac{\sigma^2}{10}$$

Da die wahre Varianz σ^2 unbekannt ist, muss diese über die Stichprobenvarianz s^2 geschätzt werden.

d) Die vorliegende Grundgesamtheit ist annähernd normalverteilt und es wurde eine Stichprobe mit dem Umfang N durchgeführt. Da die Varianz σ^2 unbekannt ist, kann der einfache t-Test durchgeführt werden (vgl. [1], S. 202):

$$T = \frac{\bar{X} - \mu_0}{s}\sqrt{N}$$

Dieser ist unter H_0 t-verteilt mit $df = N - 1$ Freiheitsgraden.

e) Der Ablehnungsbereich eines Tests wird über den kritischen Wert des Tests bestimmt (vgl. [1], S. 202):

$$|t| \geq t_{df;1-\frac{\alpha}{2}} = t_{krit}$$

$$\Leftrightarrow -t_{df;1-\frac{\alpha}{2}} \leq t \leq t_{df;1-\frac{\alpha}{2}}$$

Dabei ist $t_{krit} = t_{df;1-\frac{\alpha}{2}}$ der kritische Wert des Tests. Er ist abhängig von der Irrtumswahrscheinlichkeit und dem Stichprobenumfang (da damit die Anzahl der Freiheitsgrade berechnet wird).

Schritt 1) Bestimmen der noch unbekannten Parameter. Mit einer Irrtumswahrscheinlichkeit von 5% und einer Stichprobe von $N = 10$ gilt:

$$\alpha = 0{,}05 \rightarrow 1 - \frac{\alpha}{2} = 0{,}975$$

$$N = 10 \rightarrow df = N - 1 = 9$$

Damit liegt der kritische Testwert bei:

$$t_{df;1-\frac{\alpha}{2}} = t_{9;0,975}$$

Schritt 2) Bestimmen des entsprechenden Quantils $t_{9;0,975}$ der t-Verteilung aus der Verteilungstabelle:

n	...	0,95	0,975	0,99	...
⋮	⋮	⋮	⋮	⋮	...
8	⋮	1,8595	2,3060	2,8965	...
9	⋮	1,8331	2,2622	2,8214	...
10	⋮	1,8125	2,2281	2,7638	...
⋮	⋮	⋮	⋮	⋮	

Damit gilt:

$$t_{krit} = t_{9;0,975} = 2,2622$$

Schritt 3) Damit ergibt sich der Ablehnungsbereich dieses Tests zu:

$$(-\infty; -2,2622] \cup [2,2622; \infty)$$

Für den Nichtablehnungsbereich gilt entsprechend:

$$(-2,2622; 2,2622)$$

f) Um eine Testentscheidung treffen zu können, muss die Testgröße für die vorliegende Stichprobe berechnet werden. Dann muss geprüft werden, ob der Testwert in den Ablehnungs- oder Nichtablehnungsbereich fällt.

Schritt 1) Berechnen der Testgröße. Um die Formel aus Aufgabenteil d) nutzen zu können, müssen noch das arithmetische Mittel $\bar{X}$ der Stichprobe berechnet sowie die Stichprobenvarianz abgeschätzt werden:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{1}{10}(55 + 69 + 40 + \cdots + 59) = \frac{570}{10} = 57$$

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2} = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(X_i - \bar{X})^2}$$

$$= \sqrt{\frac{1}{9}[(55-57)^2 + (69-57)^2 + \cdots + (59-57)^2]}$$

$$= \sqrt{\frac{(-2)^2 + 12^2 + \cdots + 2^2}{9}} = \sqrt{\frac{900}{9}} = \sqrt{100} = 10$$

Dann kann der Testwert berechnet werden:

$$T = \frac{\bar{X} - \mu_0}{s}\sqrt{N} = \frac{57 - 50}{10}\sqrt{10} \approx 2{,}214$$

Schritt 2) Prüfen, ob der Testwert im Ablehnungsbereich liegt. Es gilt:

$$|T| = |2{,}214| = 2{,}214 < t_{krit} = 2{,}2622$$

T liegt also im Nichtablehnungsbereich der Nullhypothese $H_0 : \mu = 50$. Daher kann diese bei einer Irrtumswahrscheinlichkeit von $\alpha = 0{,}05$ nicht abgelehnt werden.

g) Bei dieser Entscheidung (Nichtablehnung der Nullhypothese) kann der Fehler 2. Art begangen werden, d.h. die Nullhypothese wird nicht abgelehnt, obwohl der wahre Parameter nicht $\mu_0 = 50$ beträgt.

h) Schritt 1) Aufstellen der Testhypothesen. Es müssen einseitige Hypothesen aufgestellt werden, da hier nicht angenommen wird, dass die Waschmaschine genau 50 Liter Wasser verbraucht, sondern höchstens 50 Liter. So ergibt sich (vgl. [1], S. 202):

$$H_0 : \mu \leq \mu_0 = 50 \qquad H_1 : \mu > \mu_0 = 50$$

Dies entspricht einem rechtsseitigen Test und einem einseitigen Testproblem.

Schritt 2) Bestimmen des kritischen Testwerts t_{krit}. Für den Ablehnungsbereich gilt (vgl. [1], S. 202):

$$t \geq t_{df;1-\alpha} = t_{krit}$$

Mit $N = 10$ und $1 - \alpha = 1 - 0{,}05 = 0{,}95$ gilt:

$$t_{krit} = t_{9;0{,}95}$$

Dieser Wert kann aus der Tabelle der t-Verteilung abgelesen werden:

n	…	0,9	0,95	0,975	…
⋮	⋮	⋮	⋮	⋮	…
8	⋮	1,3968	1,8595	2,3060	…
9	⋮	1,3830	1,8331	2,2622	…
10	⋮	1,3722	1,8125	2,2281	…
⋮	⋮	⋮	⋮	⋮	

Damit gilt:

$$t_{krit} = 1{,}8331$$

Schritt 3) Damit ergibt sich der Ablehnungsbereich dieses Tests zu:

$$[1{,}8331; \infty)$$

Für den Nichtablehnungsbereich gilt entsprechend:

$$(-\infty; 1{,}8331)$$

Schritt 4) Bestimmen des Testwerts. Für den Testwert gilt mit den gleichen Parametern wie in Aufgabenteil f) (vgl. [1], S. 202):

$$T = \frac{\bar{X} - \mu_0}{s}\sqrt{N} = \frac{57 - 50}{10}\sqrt{10} \approx 2{,}214$$

Schritt 5) Prüfen, ob der Testwert im Ablehnungsbereich liegt. Es gilt:

$$T = 2{,}214 > t_{krit} = 1{,}8331$$

T liegt also im Ablehnungsbereich der Nullhypothese $H_0 : \mu \leq 50$. Daher muss diese bei einer Irrtumswahrscheinlichkeit von $\alpha = 0{,}05$ abgelehnt werden.

i) Bei dieser Entscheidung (Ablehnung der Nullhypothese) kann der Fehler 1. Art begangen werden, d.h. die Nullhypothese wird abgelehnt, obwohl sie richtig ist.

j) Die Irrtumswahrscheinlichkeit von 5% gibt an, wie groß die Wahrscheinlichkeit ist, den Fehler 1. Art zu begehen. Mit dieser gegebenen Irrtumswahrscheinlichkeit kann es passieren, dass die Nullhypothese „Der Durchschnittsverbrauch der alten Waschmaschine liegt bei höchstens 50 Litern“ abgelehnt wird, obwohl sie in der Grundgesamtheit (Wasserverbrauch der alten Waschmaschine bei unabhängigen Durchgängen) wahr ist, d.h. der Durchschnittsverbrauch wirklich höchstens 50 Liter beträgt. Dann würde die vorliegende Stichprobe, die zum Ablehnen der Nullhypothese führte, ein seltenes (unwahrscheinliches) Ereignis darstellen.

Aufgabe 18

In der folgenden Tabelle ist der Bestand an Gütermotorschiffen der Binnenschifffahrt eines westeuropäischen Landes an einem bestimmten Stichtag nach dem Alter (Y) und der Tragfähigkeit (X) aufgegliedert.

Alter (in Jahren) / Trag-fähigkeit (t)	unter 20	20 bis unter 40	40 bis unter 60	60 bis unter 80	80 und mehr	ins-ges.
Unter 400	60	60	170	200	40	530
400 b.u. 1000	80	280	370	380	170	1280
1000 b.u. 3000	370	280	130	50	20	850
insgesamt	510	620	670	630	230	2660

a) Bestimmen Sie die ausgleichende Regressionsgerade $\bar{X}(Y) = b_0 + b_1 Y$!
Berechnen und interpretieren Sie den Wert $\bar{X}(36)$! Gehen Sie bei Ihren Berechnungen davon aus, dass kein Güterschiff älter als 100 Jahre ist.

b) Geben Sie, ohne die Parameter der ausgleichenden Regressionsgeraden $\bar{Y}(X) = a_0 + a_1 X$ zu berechnen, das Vorzeichen des Parameters a_1 an!
c) Berechnen Sie die Parameter a_0 und a_1!
d) Berechnen Sie den Pearsonschen Korrelationskoeffizienten und interpretieren Sie diesen!

Lösung:
Merkmale: Alter (metrisch), Tragfähigkeit (metrisch)
Merkmalsträger: Gütermotorschiffe

Vorbemerkung: Da die beiden metrischen Merkmale Alter (Y) und Tragfähigkeit (X) klassiert sind, können die zu berechnenden Maßzahlen nur approximativ bestimmt werden.

a) Für die Regressionsgerade gilt in allgemeiner Form:

$$\bar{X}(Y) = b_0 + b_1 Y$$

Für die beiden Parameter b_0 (y-Achsenabschnitt) und b_1 (Steigung) gilt:

$$b_0 = \bar{x} - b_1 \bar{y}$$

$$b_1 = \frac{s_{xy}}{s_y^2}$$

Für die Kovarianz s_{xy}, die Varianz s_y^2 und die arithmetischen Mittel gilt (klassierte Daten):

$$s_{xy} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{l} x_i y_j \, N_{i,j} - \bar{x}\bar{y}$$

$$s_y^2 = \frac{1}{N} \sum_{j=1}^{l} y_j^2 N_j^y$$

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{k} x_i \, N_i^x \qquad \bar{y} = \frac{1}{N} \sum_{j=1}^{l} y_i N_j^y \qquad \overline{y^2} = \frac{1}{N} \sum_{j=1}^{l} y_j^2 \, N_j^y$$

Dabei steht k für die Anzahl an Klassen des Merkmals Tragfähigkeit und l für die Anzahl an Klassen des Merkmals Alter. N_i^x und N_j^y stehen für die Anzahl in der jeweiligen Klasse, x_i und y_j für die jeweiligen Klassenmitten. Diese berechnen sich zu:

$$Klassenmitte = \frac{Klassenobergrenze - Klassenuntergrenze}{2}$$

Unter Annahme eines Maximalalters von 100 Jahren kann die Tabelle wie folgt umgeschrieben werden:

x_i \ y_j	10	30	50	70	90	$\sum (= N_i^x)$
200	60	60	170	200	40	530
700	80	280	370	380	170	1280
2000	370	280	130	50	20	850
$\sum (= N_j^y)$	510	620	670	630	230	2660

Damit ergeben sich mit $N = 2660$:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} x_i\, N_i^x = \frac{1}{2660}\sum_{i=1}^{3} x_i\, N_i^x$$

$$= \frac{1}{2660}(200 * 530 + 700 * 1280 + 2000 * 850) = \frac{2702000}{2660} \approx 1015{,}79$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{l} y_i N_j^y = \frac{1}{2660}\sum_{i=1}^{5} y_i N_j^y$$

$$= \frac{1}{2660}(10 * 510 + 30 * 620 + 50 * 670 + 70 * 630 + 90 * 230) \approx 45{,}86$$

Um die Kovarianz zu berechnen, müssen die einzelnen Produkte aus den beiden Klassenmitten x_i und y_j sowie der entsprechenden Anzahl an Kombinationen $N_{i,j}$ berechnet werden. Diese sind in folgender Tabelle zusammengefasst:

x_i \ y_j	10	30	50	70	90
200	120000	360000	1700000	2800000	720000
700	560000	5880000	12950000	18620000	10710000
2000	7400000	16800000	13000000	7000000	3600000

Dann gilt für die Kovarianz s_{xy}:

$$s_{xy} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} x_i y_j\, N_{i,j} - \bar{x}\bar{y} = \frac{1}{2660}\sum_{i=1}^{3}\sum_{j=1}^{5} x_i y_j\, N_{i,j} - \bar{x}\bar{y}$$

$$= \frac{120000 + 360000 + 1700000 + \cdots + 3600000}{2660} - 1015{,}79 * 45{,}86$$

$$= \frac{102220000}{2660} - 1015{,}79 * 45{,}86 \approx 38428{,}57 - 46584{,}13 = -8155{,}56$$

Für die Varianz s_y^2 ergibt sich:

$$s_y^2 = \frac{1}{N}\sum_{j=1}^{l} y_j^2 N_j^y - \bar{y}^2 = \frac{1}{2660}\sum_{j=1}^{5} y_j^2 N_j^y - \bar{y}^2$$

$$= \frac{1}{2660}(10^2 * 510 + 30^2 * 620 + 50^2 * 670 + 70^2 * 630 + 90^2 * 230) - 45{,}86^2$$

$$= \frac{7234000}{2660} - 45{,}86^2 \approx 2719{,}55 - 2103{,}14 = 616{,}41$$

Damit lassen sich nun die beiden Parameter der Regressionsgeraden berechnen:

$$b_1 = \frac{s_{xy}}{s_y^2} = \frac{-8155{,}56}{616{,}41} \approx -13{,}23$$

$$b_0 = \bar{x} - b_1\bar{y} = 1015{,}79 - (-13{,}23) * 45{,}86 \approx 1622{,}52$$

$$\rightarrow \bar{X}(Y) = 1622{,}52 - 13{,}23Y$$

An der Stelle $Y = 36$ nimmt die Regressionsgerade folgenden Wert an:

$$\bar{X}(Y = 36) = 1622{,}52 - 13{,}23 * 36 = 1146{,}24$$

Das bedeutet, dass 36-jährige Schiffe im Mittel eine Tragfähigkeit von 1146,24 t aufweisen.

b) Für die Steigung a_1 gilt:

$$a_1 = \frac{s_{xy}}{s_y^2}$$

Da der Nenner in jedem Fall positiv ist, wird das Vorzeichen der Steigung durch das Vorzeichen der Kovarianz bestimmt. Diese ist in beiden Fällen gleich, also ist auch $a_1 < 0$ und hat dementsprechend ein negatives Vorzeichen.

c) Um die Regressionsgerade $\bar{Y}(X) = a_0 + a_1 X$ berechnen zu können, muss nur noch ein Wert berechnet werden:

$$s_x^2 = \frac{1}{N}\sum_{i=1}^{k} x_i^2 N_i^x - \bar{x}^2 = \frac{1}{2660}\sum_{i=1}^{3} x_i^2 N_i^x - \bar{x}^2$$

$$= \frac{1}{2660}(200^2 * 530 + 700^2 * 1280 + 2000^2 * 850) - 1015{,}79^2$$
$$= \frac{1}{2660}(40000 * 530 + 490000 * 1280 + 4000000 * 850) - 1015{,}79^2$$

$$= \frac{4048400000}{2660} - 1015{,}79^2 \approx 1521954{,}89 - 1031829{,}32 = 490125{,}57$$

Damit lassen sich nun die beiden Parameter der Regressionsgeraden berechnen:

$$a_1 = \frac{s_{xy}}{s_x^2} = \frac{-8155{,}56}{490125{,}57} \approx -0{,}0166$$

$$a_0 = \bar{x} - a_1\bar{y} = 45{,}86 - (-0{,}0166) * 1015{,}79 \approx 62{,}72$$

Damit lautet die Regressionsgerade:

$$\bar{Y}(X) = 62{,}72 - 0{,}0166X$$

d) Der Korrelationskoeffizient ist definiert als (vgl. [1], S. 176):

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{-8155{,}56}{\sqrt{490125{,}57}\sqrt{616{,}41}} \approx \frac{-8155{,}56}{700{,}09 * 24{,}83} \approx -0{,}47$$

Er kann im Falle bekannter Steigungen der Regressionsgeraden auch anders berechnet werden:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \begin{cases} \sqrt{\frac{s_{xy}^2}{s_x^2 * s_y^2}} = \sqrt{\frac{s_{xy}}{s_x^2} * \frac{s_{xy}}{s_y^2}} = \sqrt{a_1 b_1} & falls\ s_{xy} > 0 \\ -\sqrt{\frac{s_{xy}^2}{s_x^2 * s_y^2}} = -\sqrt{\frac{s_{xy}}{s_x^2} * \frac{s_{xy}}{s_y^2}} = -\sqrt{a_1 b_1} & falls\ s_{xy} < 0 \end{cases}$$

Hier also mit $s_{xy} = -8155{,}56 < 0$:

$$r_{xy} = -\sqrt{a_1 b_1} = -\sqrt{-0{,}0166 * (-13{,}23)} = -\sqrt{0{,}219618} \approx -0{,}47$$

Dies bedeutet, dass zwischen den Merkmalen X (Tragfähigkeit) und Y (Alter) im Mittel ein negativer linearer statistischer Zusammenhang vorliegt, der nicht sehr ausgeprägt ist. Tendenziell gilt: Mit zunehmendem Alter nimmt die Tragfähigkeit der Schiffe ab.

Referenzen:

Die in der Übung aufgeführten Aufgaben wurden folgenden Lehr- und Arbeitsbüchern entnommen:

[1] *Luderer, B.; Nollau, V.; Vetters, K.:* Mathematische Formeln für Wirtschaftswissenschaftler, 8. Auflage, Wiesbaden: Springer Gabler 2015.

[2] *Beyer, O.; Hackel, H.; Pieper, V.; Tiedge, J.:* Mathematik für Ingenieure, Naturwissenschaftler, Ökonomen und Landwirte - Wahrscheinlichkeitsrechnung und mathematische Statistik, Bd. 17, Leipzig: B.G. Teubner1985.

[3] *Böhm, P.:* Induktive Statistik und Wahrscheinlichkeitsrechnung Arbeitsbuch II. Berlin: Studeo Verlag 2004.

[4] *Böhm, P.; Ringhut, S.; Engler, S.*: Deskriptive Statistik Arbeitsbuch II. Berlin: Studeo Verlag 2004.

[5] *Heller, W.-D.; Lindenberg, H.; Nuske, M.; Schriever, K.-H.*: Wahrscheinlichkeitsrechnung – Teil 2. Basel, Boston, Stuttgart: Birkhäuser Verlag 1979.

[6] *Gillert, H.; Nollau, V.; Pieper, V.; Tiedge, J.:* Mathematik für Ingenieure, Naturwissenschaftler, Ökonomen und Landwirte – Übungsaufgaben zur Wahrscheinlichkeitsrechnung und mathematische Statistik, Bd. Ü4, Leipzig: B.G. Teubner1989.

[7] *Kuchinke, L.; Schleusener, M.*: Induktive Statistik und Wahrscheinlichkeitsrechnung Arbeitsbuch I – Aufgaben mit Lösungen zur Vorbereitung auf Klausuren. Berlin: Studeo Verlag 2004.

[8] *Maibaum, G.:* Wahrscheinlichkeitsrechnung. Frankfurt (Main): Harri Deutsch, 1980.

[9] *Menges, G.:* Grundriß der Statistik – Teil1: Theorie. Opladen: Westdeutscher Verlag, 1972.

[10] *Nollau, V.; Patzsch, L.; Storm, R.; Lange, C.*: Wahrscheinlichkeitsrechnung und Statistik in Beispielen und Aufgaben. Stuttgart Leipzig: B.G. Teubner Verlagsgesellschaft 1997.

[11] *Vogel, F.*: Beschreibende und schließende Statistik Aufgaben und Beispiele. München Wien: Oldenbourg Wissenschaftsverlag 2001.