Nachklausur: Statistik

Jürgen Meisel
FH Ludwigshafen

Zugelassene Hilfsmittel: Nicht progr. Taschenrechner
Bearbeitungszeit: **120 Minuten**

Aufgabe 1:

Mittelwerte, Streuungsmaße und graphische Darstellungsformen
Für die 20 Industriebetriebe in einer Gemeinde liegen folgende Daten vor:

Umsatz (in Mio)	50	100	30	200	20
Anzahl der Betriebe	7	2	5	1	5

a) Bestimmen Sie arithm. Mittel, Median und Modalwert des Umsatzes eines Betriebs.

Lösung:

Ordnen nach der Größe:

Umsatz (in Mio)	20	30	50	100	200
Anzahl der Betriebe	5	5	7	2	1

$$\mu_{arith.} = \frac{20\cdot 5+30\cdot 5+50\cdot 7+100\cdot 2+200\cdot 1}{5+5+7+2+1} = \frac{1.000}{20} = 50$$

$$\bar{x}_{Median} = \frac{1}{2}\left(x_{10}+x_{11}\right) = \frac{30+50}{2} = 40$$

Modalwert: $\bar{x}_{Modal} = 50$

b) Berechnen Sie die Standardabweichung und die 1. und 3. Quartile.

Lösung: Varianz und Standardabweichung:

$$\sigma^2 = \frac{(20-50)^2\cdot 5+(30-50)^2\cdot 5+(50-50)^2\cdot 7+(100-50)^2\cdot 2+(200-50)^2\cdot 1}{5+5+7+2+1}$$

$$\sigma^2 = \frac{900\cdot 5+400\cdot 5+0\cdot 7+2.500\cdot 2+22.500\cdot 1}{20} = \frac{34.000}{20} = 1.700$$

$$\sigma = \sqrt{1.700} = 41,23$$

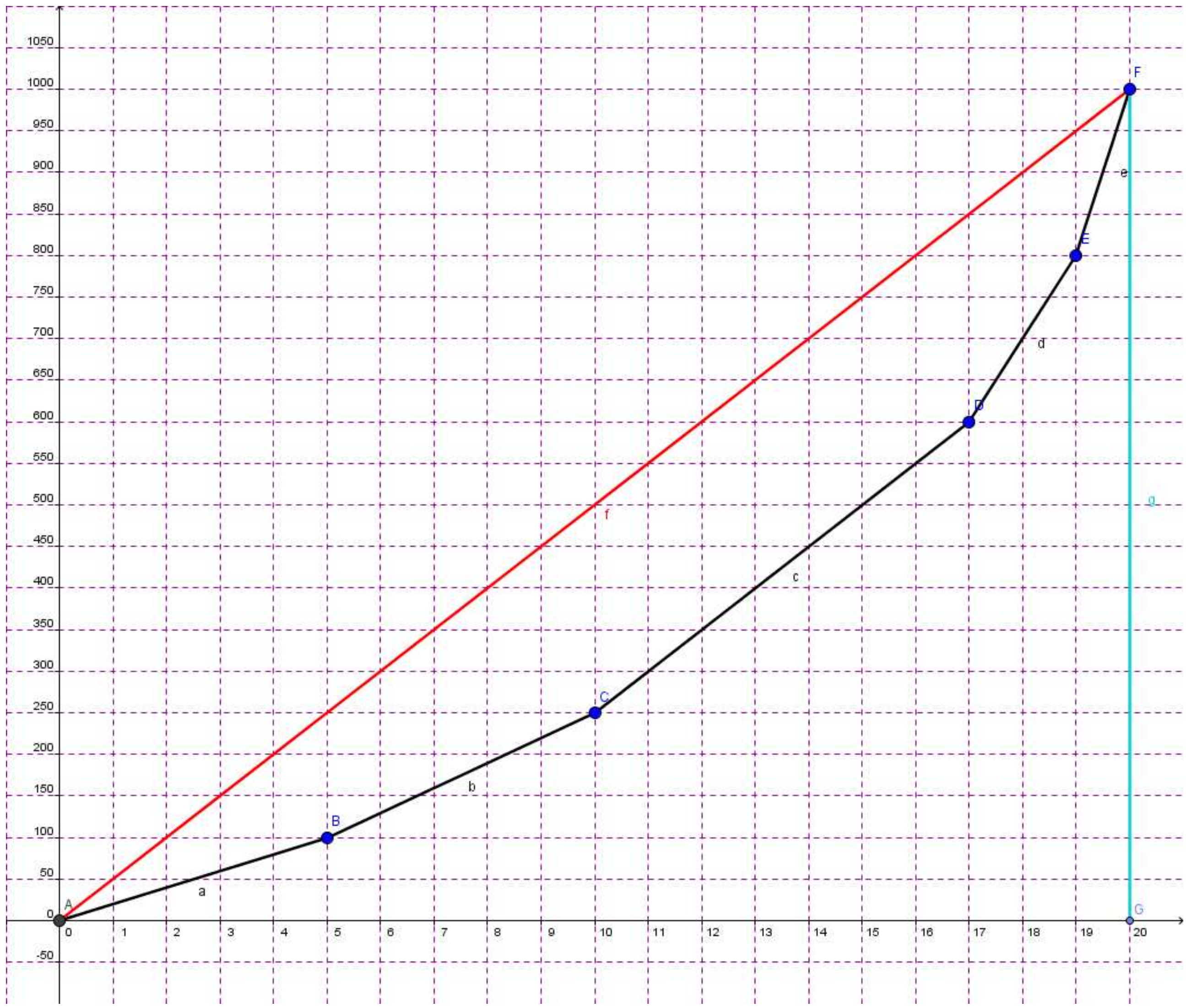
Quartil 1: $\overline{x_{0,25}} = \frac{1}{2}\left(x_{20\cdot 0,25} + x_{20\cdot 0,25+1}\right) = \frac{1}{2}\left(x_5 + x_6\right) = \frac{1}{2}(20+30) = 25$

Quartil 3: $\overline{x_{0,75}} = \frac{1}{2}\left(x_{20\cdot 0,75} + x_{20\cdot 0,75+1}\right) = \frac{1}{2}\left(x_{15} + x_{16}\right) = \frac{1}{2}(50+50) = 50$

c) Zeichnen Sie die Lorenzkurve.

Lösung:

Umsatz (in Mio)	20	30	50	100	200
Summe Umsatz	100	250	600	800	1.000
Anzahl der Betriebe	5	5	7	2	1
Summe Anzahl Betriebe	5	10	17	19	20



Aufgabe 2: Preisindizes

Ein Landwirt stellt fest, dass sich seine Ausgaben für Futtermittel von 2004 bis 2008 erhöht haben.

Futtermittel	Ausgaben in €		$\frac{\text{Preis 2008}}{\text{Preis 2004}}$
	2004	**2008**	
Mais	1000	1500	1,25
Hafer	1250	1440	1,2
Heu	250	900	1,5

Berechnen Sie den Preisindex nach Laspeyres und den Preisindex nach Paasche.

Lösung:

$$L_P = \frac{\sum p_{1i} \cdot q_{0i}}{\sum p_{0i} \cdot q_{0i}} = \frac{1.000 \cdot 1,25 + 1.250 \cdot 1,2 + 250 \cdot 1,5}{1.000 + 1.250 + 250}$$

$$L_P = \frac{3.125}{2.500} = 1,25 \Rightarrow 25[\%]$$

$$P_P = \frac{\sum p_{1i} \cdot q_{1i}}{\sum p_{0i} \cdot q_{1i}} = \frac{1.500 + 1.440 + 900}{\frac{1.500}{1,25} + \frac{1.440}{1,2} + \frac{900}{1,5}} = \frac{3.840}{3.000} = 1,28 \Rightarrow 28[\%]$$

Aufgabe 3: Trendfunktionen – Lineare Regression und Korrelation

In einem bestimmten Bereich hängt der Ernteertrag eines landwirtschaftlichen Gutes von der Menge eines eingesetzten Düngemittels je Hektar ab.

Auf 6 Versuchsfeldern wird der Düngemitteleinsatz getestet.

Dabei wurden die folgenden Erträge je Hektar erzielt:

Versuchsfeld	**1**	**2**	**3**	**4**	**5**	**6**
Düngemittel [100 kg]	**6**	**3**	**8**	**2**	**7**	**2**
Ernteertrag [t]	**30**	**10**	**22**	**14**	**36**	**24**

a) Stellen Sie den Ernteertrag in Abhängigkeit von dem Düngemitteleinsatz durch eine Funktion $y = b_0 + b_1 x$ dar.

Lösung:

Versuchsfeld	1	2	3	4	5	6	Σ
Düngemittel [100 kg]	6	3	8	2	7	2	28
Ernteertrag [t]	30	10	22	14	36	24	136
$x_i \cdot y_i$	180	30	176	28	252	48	714
$(x_i)^2$	36	9	64	4	49	4	166
$(y_i)^2$	900	100	484	196	1.296	576	3.552

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{28}{6} = 4,67 \qquad \mu_y = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{136}{6} = 22,67$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i y_i) - n\cdot\mu_x\cdot\mu_y}{\sum_{i=1}^{n}\left(x_i^2\right) - n\cdot\mu_x^2} = \frac{714 - 6\cdot\frac{14}{3}\cdot\frac{68}{3}}{166 - 6\cdot\left(\frac{28}{6}\right)^2} = \frac{238\cdot 3}{3\cdot 106} = 2,245$$

$$b_0 = \mu_y - b_1\cdot\mu_x = \frac{136}{6} - 2,245\cdot\frac{28}{6} = 12,19$$

Lineare Regression: $y = 2,245x + 12,19$

b) Ermitteln Sie den Korrelationskoeffizienten und beurteilen Sie Ihr Ergebnis.

Lösung:

$$r = \frac{\sum_{i=1}^{n}(x_i y_i) - n\cdot\mu_x\cdot\mu_y}{\sqrt{\sum_{i=1}^{n}\left(x_i^2\right) - n\cdot\mu_x^2}\cdot\sqrt{\sum_{i=1}^{n}\left(y_i^2\right) - n\cdot\mu_y^2}}$$

$$r = \frac{714 - 6\cdot\frac{28}{6}\cdot\frac{136}{6}}{\sqrt{166 - 6\cdot\left(\frac{28}{6}\right)^2}\cdot\sqrt{3.552 - 6\cdot\left(\frac{136}{6}\right)^2}} = \frac{238}{3\cdot\sqrt{\frac{106}{3}}\cdot\sqrt{\frac{1.408}{3}}} = 0,616$$

Aufgabe 4: Normalverteilung (mit bedingter Wahrscheinlichkeit)

Erfahrungsgemäß ist in einem Restaurant die Füllhöhe der Weingläser normalverteilt mit dem Mittelwert 0,26 l und s = 0,02 l.

Berechnen Sie die Wahrscheinlichkeit, dass die Füllhöhe

a) mindestens 0,27 l

b) höchstens 0,25 l

c) zwischen 0,24 l und 0,26 l beträgt,

d) über 0,27 l ist, wenn sie über dem „Eichstrich" (= 0,25 l) liegt.

<u>Lösung:</u>

a)

$$P(X \geq 0{,}27) = 1 - \Phi(0{,}5) = 1 - 0{,}691 = 0{,}309 \approx 30{,}9[\%]$$

$$Umrechnung: \quad x = \frac{0{,}27 - 0{,}26}{0{,}02} = 0{,}5$$

b)

$$P(X \leq 0{,}25) = \Phi(-0{,}5) = 1 - \Phi(0{,}5) = 1 - 0{,}691 = 0{,}309 \approx 30{,}9[\%]$$

$$Umrechnung: \quad x = \frac{0{,}25 - 0{,}26}{0{,}02} = -0{,}5$$

c)

$$P(0{,}24 \leq X \leq 0{,}26) = \Phi(0) - \Phi(-1) = 0{,}5 - (1 - 0{,}841) = 0{,}341 \approx 34{,}1[\%]$$

$$Umrechnung: \quad x_1 = \frac{0{,}24 - 0{,}26}{0{,}02} = -1 \qquad und \qquad x_2 = \frac{0{,}26 - 0{,}26}{0{,}02} = 0$$

d)

$$P_{(X \geq 0{,}25)}(X \geq 0{,}27) = \frac{P(X \geq 0{,}27)}{P(X \geq 0{,}25)} = \frac{0{,}309}{0{,}691} = 0{,}4472 \approx 44{,}72[\%]$$

Aufgabe 5: Normalverteilung mit Stetigkeitskorrektur

Bei der Produktion eines Massenartikels sind erfahrungsgemäß 20 % aller gefertigten Erzeugnisse unbrauchbar. Es wird eine Stichprobe vom Umfang n = 1000 entnommen.

Wie groß ist die Wahrscheinlichkeit, dass sich

a) höchstens 100

Lösung:

$$\mu \;=\; 1.000 \cdot 0,2 \;=\; 200 \qquad und \qquad \sigma^2 \;=\; 1.000 \cdot 0,2 \cdot 0,8 \;=\; 160 \xrightarrow{\sqrt{}} \sigma \;=\; \sqrt{160}$$

$$P(X \leq 100) \;=\; \Phi(-7,87) \;=\; 1-\Phi(7,87) \;=\; 1-0,999 \;=\; 0,001 \;\approx\; 0,1[\%]$$

$$Umrechnung: \quad x \;=\; \frac{100+0,5-200}{\sqrt{160}} \;=\; -7,87$$

b) mehr als 150 schlechte Stücke in der Stichprobe befinden

Lösung:

$$P(X > 150) \;=\; 1-P(X \leq 150) \;=\; 1-\Phi(-3,9) \;=\; 1-\left[1-\Phi(-3,9)\right]$$

$$P(X > 150) \;=\; \Phi(3,9) \;=\; 0,999 \;\approx\; 99,9[\%]$$

$$Umrechnung: \quad x \;=\; \frac{150+0,5-200}{\sqrt{160}} \;=\; -3,9$$

Aufgabe 6: Binomialverteilung

Polizeilichen Statistiken zufolge beträgt der Anteil der Autolenker, die während der Fahrt keinen Sicherheitsgurt tragen, 15 %. Diese Fahrer werden ab jetzt „Gurtmuffel" genannt. Man darf annehmen, dass die Autofahrer unabhängig voneinander den Gurt anlegen oder nicht.

a) Wie viele Autos muss man überprüfen, um mit mind. 99 %iger Wahrscheinlichkeit mind. einen Gurtmuffel zu finden?

Lösung:

$$B(X \geq 1) \;=\; 1-B(X=0) \;\geq\; 0,99 \;\Rightarrow\; \binom{n}{0} \cdot 0,15^0 \cdot 0,85^n \;\leq\; 0,01$$

$$\xrightarrow{\ln} \; n \cdot \ln(0,85) \;\leq\; \ln(0,01) \;\Rightarrow\; n \;\geq\; \frac{\ln(0,01)}{\ln(0,85)} \;=\; 28,34 \;\Rightarrow\; n \;\geq\; 29$$

b) Wie groß wäre der Anteil p der Gurtmuffel mind., wenn von 25 vorbeifahrenden Autos mit 99 %iger Wahrscheinlichkeit mind. eines von einem Gurtmuffel gelenkt würde.

Lösung:

$$B(X \geq 1) \;=\; 1-B(X=0) \;\geq\; 0,99 \xrightarrow{\text{vereinfachen}} \binom{25}{0} p^0 \cdot (1-p)^{25} \;\leq\; 0,01$$

$$\xrightarrow{\text{vereinfachen und } \sqrt[25]{(1\text{-p})}} (1-p) \;\leq\; \sqrt[25]{0,01} \xrightarrow{\text{vereinfachen}} p \;\geq\; 1-\sqrt[25]{0,01} \;=\; 0,168 \;=\; 16,8\,\%$$

Auswahlfragen:

Kreuzen Sie bei den folgenden Fragestellungen jeweils den korrekten Antwortbuchstaben eindeutig an!

Frage 1

Bei einer Untersuchung an 200 Ratten wurden Durchmesser der Nervenzellkerne in µ gemessen. Die Daten wurden klassiert. In die Klasse (8.9, 9.1] fielen 40 Daten.
Die relative Häufigkeit für die Klasse ist

(A) 0.2/40 (B) 0.2/40 µ (C) 9/40 µ (D) 9/200 µ (E) 40/200

Frage 2

Bei einer Untersuchung wurde das Körpergewicht von 300 weißen Mäusen gemessen. Die Daten wurden klassiert. Für die Klasse (50, 70] wurde eine relative Häufigkeit von 0.3 festgestellt.
Die absolute Häufigkeit für diese Klasse ist

(A) 60 * 0.3 = 20 (B) 0.5 * (50 * 0.3 + 70 * 0.3) = 18

(C) 50 * 0.3 = 15 (D) 70 * 0.3 = 21 (E) 300 x 0.3 = 90

Frage 3

Bei der Auswertung einer Klausur mit 10 Fragen ergab sich, dass die achte Frage von 50 % der Klausurteilnehmer falsch beantwortet wurde.
Die absolute Häufigkeit, mit der diese Frage falsch beantwortet wurde, ist

(A) 0.5 (B) 5 (C) 4 (D) 50

(E) Die absolute Häufigkeit lässt sich aus diesen Angaben nicht berechnen

Frage 4

In einer Stadt mit 800 000 Einwohnern sind von den 308 000 männlichen Einwohner 5 000 farbenblind und von den 492 000 weiblichen Einwohnern 600 farbenblind.
Die relative Häufigkeit der Männer unter den Farbenblinden ist

(A) $\frac{5.000}{308.000}$ (B) $\frac{5.000}{800.000}$ (C) $\frac{5.600}{800.000}$ (D) $\frac{5.000}{5.600}$ (E) $\frac{600}{5.600}$

Frage 5

Für die relative Häufigkeit h_i einer Merkmalsausprägung a_i gilt stets

(A) $0 < h_i < 1$ (B) $h_i \leq 0$ (C) $h_i \geq 1$ (D) $0 \leq h_i \leq 1$

(E) Der Wertebereich von h_i hängt von der Gesamtanzahl der Beobachtungen ab.

Frage 6

s^2 soll die empirische Varianz der Daten $x_1, x_2, \dots, x_n$ sein.
Die Daten $x_1, x_2, \dots, x_n$ können beliebig positiv, negativ oder null sein.
Es ist stets

(A) $s^2 \geq 0$ (B) $s^2 \neq 0$ (C) $s^2 > 0$ (D) $s^2 \leq 0$ (E) $s^2 = 0$.

Frage 7

Die empirische Varianz ist stets ein

(A) Lagemaß (B) Streuungsmaß (C) Maß für die Anzahl der Daten

(D) Quantil (E) Lageparameter

Frage 8

Die Vierfeldertafel ist eine spezielle Form der Kontingenztafel, die benutzt wird, wenn,

(A) jedes beobachtete Merkmal 4-fach klassiert ist
(B) landwirtschaftliche Versuche durchgeführt werden
(C) pro Beobachtungseinheit 4 Daten vorliegen
(D) an jeder Beobachtungseinheit 2 Merkmale mit jeweils 2 Ausprägungen beobachtet werden
(E) an jeder Beobachtungseinheit die Ausprägung von 4 Merkmalen beobachtet werden

Frage 9

Die geeignete Darstellungsform der Häufigkeiten eines stetigen Merkmals mit klassierten Daten ist

(A) die Kontingenztafel (B) die Punktwolke (C) das Stabdiagramm

(D) das Histogramm (E) das Venn-Diagramm

Frage 10

Bei einem Regressionsproblem hat man die Gleichung $y = 5 - 2x$ für die Regressionsgerade berechnet. Hieraus folgt, dass

(A) die Maßzahl x im Mittel zwei Einheiten abnimmt, wenn die Maßzahl y eine Einheit zunimmt
(B) die Maßzahl y im Mittel zwei Einheiten zunimmt, wenn die Maßzahl x eine Einheit zunimmt
(C) die Maßzahl y im Mittel zwei Einheiten abnimmt, wenn die Maßzahl x eine Einheit zunimmt
(D) die Maßzahl y im Mittel stets 5 Einheiten größer ist als die Maßzahl x
(E) Keine der Aussagen A - D ist richtig

Frage 11
Die Regressionsgerade $y = b_0 + b_1 * x$ der Regression von y auf x minimiert die Summe der Quadrate der Abstände der Punkte (x_i, y_i) von der Regressionsgeraden. Gemessen werden dabei die Abstände

(A) parallel zur Regressionsgeraden
(B) senkrecht zur x-Achse
(C) senkrecht zur y-Achse
(D) senkrecht zur x-Achse und zur Regressionsgeraden
(E) senkrecht zur y-Achse und zur Regressionsgeraden

Frage 12
Für den empirischen Korrelationskoeffizienten r gilt stets

(A) $r \leq 0$ (B) $0 \leq r \leq 1$ (C) $-1 \leq r \leq 1$ (D) $-1 < r < 1$ (E) $r \leq -1$

Frage 13
Der Grad der linearen Abhängigkeit zweier Merkmale wird ausgedrückt durch

(A) den Regressionskoeffizient (B) die Regressionsgerade

(C) die beiden Mittelwerte (D) den Korrelationskoeffizienten

(E) die Kontingenztafel

Frage 14
Bei Daten eines stetigen Merkmals ist (sind)

(A) der Mittelwert gegenüber Ausreißern robust
(B) die Spannweite gegenüber Ausreißern robust
(C) der Median gegenüber Ausreißern robust
(D) Mittelwert und Median gegenüber Ausreißern robust
(E) der Mittelwert gegenüber Ausreißern robuster als der Median

Frage 15
Bei 200 Familien wurden folgende Anzahlen von Personen pro Familie festgestellt:

Personen pro Familie	2	3	4	5	6	7	8	9	10	11
Häufigkeit	15	35	50	30	20	16	14	10	5	5

Der Median dieser Daten ist

(A) 4 (B) 4.5 (C) 6.5 (D) 40 (E) 100

Auswahlfragen

Lösungen sind jeweils fett hervorgehoben!

Frage 1
Bei einer Untersuchung an 200 Ratten wurden Durchmesser der Nervenzellkerne in μ gemessen. Die Daten wurden klassiert. In die Klasse (8.9, 9.1] fielen 40 Daten.
Die relative Häufigkeit für die Klasse ist

(A) 0.2/40 (B) 0.2/40 μ (C) 9/40 μ (D) 9/200 μ **(E) 40/200**

Frage 2
Bei einer Untersuchung wurde das Körpergewicht von 300 weißen Mäusen gemessen. Die Daten wurden klassiert. Für die Klasse (50, 70] wurde eine relative Häufigkeit von 0.3 festgestellt.
Die absolute Häufigkeit für diese Klasse ist

(A) 60 * 0.3 = 20 (B) 0.5 * (50 * 0.3 + 70 * 0.3) = 18

(C) 50 * 0.3 = 15 (D) 70 * 0.3 = 21 **(E) 300 x 0.3 = 90**.

Frage 3
Bei der Auswertung einer Klausur mit 10 Fragen ergab sich, dass die achte Frage von 50 % der Klausurteilnehmer falsch beantwortet wurde.
Die absolute Häufigkeit, mit der diese Frage falsch beantwortet wurde, ist

(A) 0.5 (B) 5 (C) 4 (D) 50

(E) Die absolute Häufigkeit lässt sich aus diesen Angaben nicht berechnen

Frage 4
In einer Stadt mit 800 000 Einwohnern sind von den 308 000 männlichen Einwohner 5 000 farbenblind und von den 492 000 weiblichen Einwohnern 600 farbenblind.
Die relative Häufigkeit der Männer unter den Farbenblinden ist

(A) $\frac{5.000}{308.000}$ (B) $\frac{5.000}{800.000}$ (C) $\frac{5.600}{800.000}$ **(D)** $\frac{5.000}{5.600}$ (E) $\frac{600}{5.600}$

Frage 5
Für die relative Häufigkeit h_i einer Merkmalsausprägung a_i gilt stets

(A) $0 < h_i < 1$ (B) $h_i \leq 0$ (C) $h_i \geq 1$ **(D)** $\mathbf{0 \leq h_i \leq 1}$

(E) Der Wertebereich von h_i hängt von der Gesamtanzahl der Beobachtungen ab.

Frage 6
s^2 soll die empirische Varianz der Daten $x_1, x_2, \ldots, x_n$ sein.
Die Daten $x_1, x_2, \ldots, x_n$ können beliebig positiv, negativ oder null sein.
Es ist stets

(A) $s^2 \geq 0$ (B) $s^2 \neq 0$ (C) $s^2 > 0$ (D) $s^2 \leq 0$ (E) $s^2 = 0$.

Frage 7
Die empirische Varianz ist stets ein

(A) Lagemaß **(B) Streuungsmaß** (C) Maß für die Anzahl der Daten
(D) Quantil (E) Lageparameter

Frage 8
Die Vierfeldertafel ist eine spezielle Form der Kontingenztafel, die benutzt wird, wenn,

(A) jedes beobachtete Merkmal 4-fach klassiert ist
(B) landwirtschaftliche Versuche durchgeführt werden
(C) pro Beobachtungseinheit 4 Daten vorliegen
(D) an jeder Beobachtungseinheit 2 Merkmale mit jeweils 2 Ausprägungen beobachtet werden
(E) an jeder Beobachtungseinheit die Ausprägung von 4 Merkmalen beobachtet werden

Frage 9
Die geeignete Darstellungsform der Häufigkeiten eines stetigen Merkmals mit klassierten Daten ist

(A) die Kontingenztafel (B) die Punktwolke (C) das Stabdiagramm
(D) das Histogramm (E) das Venn- Diagramm

Frage 10
Bei einem Regressionsproblem hat man die Gleichung $y = 5 - 2x$ für die Regressionsgerade berechnet. Hieraus folgt, dass

(A) die Maßzahl x im Mittel zwei Einheiten abnimmt, wenn die Maßzahl y eine Einheit zunimmt
(B) die Maßzahl y im Mittel zwei Einheiten zunimmt, wenn die Maßzahl x eine Einheit zunimmt
(C) die Maßzahl y im Mittel zwei Einheiten abnimmt, wenn die Maßzahl x eine Einheit zunimmt
(D) die Maßzahl y im Mittel stets 5 Einheiten größer ist als die Maßzahl x
(E) Keine der Aussagen A - D ist richtig

Frage 11
Die Regressionsgerade $y = b_0 + b_1 * x$ der Regression von y auf x minimiert die Summe der Quadrate der Abstände der Punkte (x_i, y_i) von der Regressionsgeraden. Gemessen werden dabei die Abstände

(A) parallel zur Regressionsgeraden
(B) senkrecht zur x-Achse
(C) senkrecht zur y-Achse
(D) senkrecht zur x-Achse und zur Regressionsgeraden
(E) senkrecht zur y-Achse und zur Regressionsgeraden

Frage 12
Für den empirischen Korrelationskoeffizienten r gilt stets

(A) $r \leq 0$ (B) $0 \leq r \leq 1$ **(C) $-1 \leq r \leq 1$** (D) $-1 < r < 1$ (E) $r \leq -1$

Frage 13
Der Grad der linearen Abhängigkeit zweier Merkmale wird ausgedrückt durch

(A) den Regressionskoeffizient (B) die Regressionsgerade

(C) die beiden Mittelwerte **(D) den Korrelationskoeffizienten**

(E) die Kontingenztafel

Frage 14
Bei Daten eines stetigen Merkmals ist (sind)

(A) der Mittelwert gegenüber Ausreißern robust
(B) die Spannweite gegenüber Ausreißern robust
(C) der Median gegenüber Ausreißern robust
(D) Mittelwert und Median gegenüber Ausreißern robust
(E) der Mittelwert gegenüber Ausreißern robuster als der Median

Frage 15
Bei 200 Familien wurden folgende Anzahlen von Personen pro Familie festgestellt:

Personen pro Familie	2	3	4	5	6	7	8	9	10	11
Häufigkeit	15	35	50	30	20	16	14	10	5	5

Der Median dieser Daten ist

(A) 4 **(B) 4.5** (C) 6.5 (D) 40 (E) 100