PROBEKLAUSUR STATISTIK I (mit Lösungen)

Berufsbegleitender Studiengang Betriebswirtschaftslehre
Berufsbegleitender Studiengang Wirtschaftsingenieurwesen

Sommersemester 2016
Christian Reinboth

Aufgabenteil I: Theorie (10 Punkte)

Sind die nachfolgenden Aussagen richtig oder falsch? (1 Punkt pro korrekter Beantwortung)

1) Der Bravais-Pearson-Korrelationskoeffizient kann nur Werte zwischen -1 und 1 annehmen.

X richtig O falsch

2) In Stem-and-Leaf-Diagrammen repräsentiert ein Blatt stets einen Fall.

O richtig **X falsch**

3) Ist ein Merkmal häufbar, kann es mehrere Ausprägungen gleichzeitig annehmen.

X richtig O falsch

4) Die Berechnung der Varianz setzt mindestens metrisch skalierte Daten voraus.

X richtig O falsch

5) Die Zäune im Box-Plot umfassen stets eine Strecke mit Länge des 1,5-fachen IQR.

O richtig **X falsch**

6) 2D-Streudiagramme stellen die gemeinsame Verteilung der Werte zweier Variablen dar.

X richtig O falsch

7) Der Quartilskoeffizient der Schiefe ist unempfindlich gegenüber Ausreißern.

X richtig O falsch

8) Das arithmetische Mittel kann nur für ordinalskalierte Daten gebildet werden.

O richtig **X falsch**

9) Variablen mit nur zwei möglichen Ausprägungen werden als bivariat bezeichnet.

O richtig **X falsch**

10) Eine per Zufallsauswahl zusammengestellte Stichprobe ist stets repräsentativ.

O richtig **X falsch**

Aufgabenteil II: Grafische Darstellungsformen (20 Punkte)

Bei einer Mathematikarbeit ergibt sich die nachfolgend tabellierte Notenverteilung.

Note	**1**	**2**	**3**	**4**	**5**	**6**
Schüler/innen	3	7	11	8	2	1

Konstruieren Sie einen (erweiterten) Box-Plot für diese Verteilung (10 Punkte)
und benennen Sie alle für die Konstruktion benötigten Größen (10 Punkte).

<u>Geordnete Verteilung</u>

1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 6 n = 32

<u>Erforderliche Größen</u>

Oberes Quartil:

$n * p = 0{,}75 * 32 = 24$ -> ganzzahlig, daher gilt (bei k = 24):

$$x_p = \frac{1}{2}(x_{(k)} + x_{(k+1)}) = x_{0,75} = \frac{1}{2}(x_{(24)} + x_{(25)}) = \frac{1}{2}(4+4) = 4$$

Median:

$n * p = 0{,}50 * 32 = 16$ -> ganzzahlig, daher gilt (bei k = 16):

$$x_p = \frac{1}{2}(x_{(k)} + x_{(k+1)}) = x_{0,75} = \frac{1}{2}(x_{(16)} + x_{(17)}) = \frac{1}{2}(3+3) = 3$$

Unteres Quartil:

$n * p = 0{,}25 * 32 = 8$ -> ganzzahlig, daher gilt (bei k = 8):

$$x_p = \frac{1}{2}(x_{(k)} + x_{(k+1)}) = x_{0,75} = \frac{1}{2}(x_{(8)} + x_{(9)}) = \frac{1}{2}(2+2) = 2$$

Interquartilsabstand:

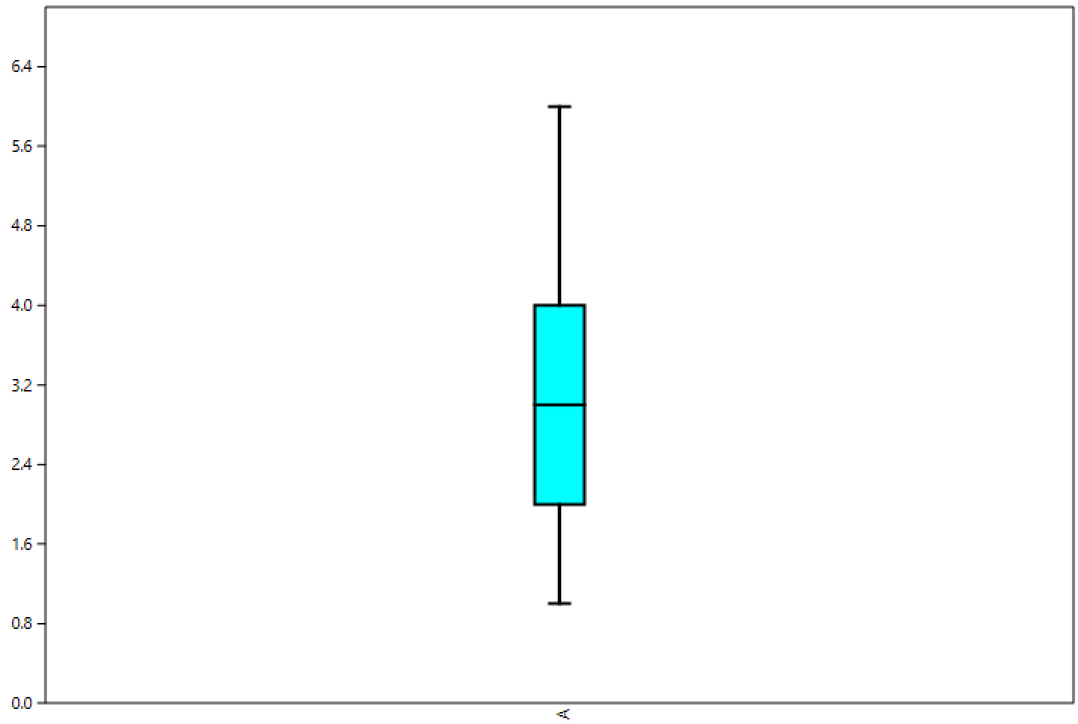
$$IQR = x_{(0,75)} - x_{(0,25)} = 4 - 2 = 2$$

Aus diesen Größen lässt sich der Box-Plot nun wie folgt konstruieren:

- Das obere Ende der Box liegt bei 4.
- Das untere Ende der Box liegt bei 2.
- In die Box wird der Median bei 3 eingezeichnet.

- Der größte Nicht-Ausreißer könnte maximal bei 4 + 1,5 * IQR = 7 liegen. Der größte noch zwischen 4 und 7 auftretende Wert ist 6. Der obere Zaun verläuft also bis zur 6.
- Der kleinste Nicht-Ausreißer könnte minimal bei 2 – 1,5 * IQR = -1 liegen. Der kleinste noch zwischen 2 und -1 auftretende Wert ist 1. Der untere Zaun verläuft also bis zur 1.

- Da es keine Werte mehr gibt, die außerhalb der Zäune liegen, muss im finalen Schritt nun nicht mehr nach Ausreißern oder Extremwerten gesucht werden.



Aufgabenteil III: Verteilungsparameter (30 Punkte)

Die Bestimmung der Körpergröße von 10 Personen erbringt folgende Ergebnisse:

1,62 m	1,77 m
1,71 m	1,82 m
1,61 m	1,55 m
1,82 m	1,74 m
1,75 m	1,63 m

Maße der zentralen Tendenz:

- Bestimmen Sie das arithmetische Mittel. (5 Punkte)
- Bestimmen Sie den Median. (5 Punkte)
- Bestimmen Sie den Modus. (5 Punkte)

Streuungsmaße / Dispersionsparameter:

- Bestimmen Sie die Varianz. (5 Punkte)
- Bestimmen Sie die Standardabweichung. (5 Punkte)

Verteilungsmaße / Schiefe und Wölbung:

- Bestimmen Sie den Quartilskoeffizienten der Schiefe. (5 Punkte)

Bestimmung des arithmetischen Mittels

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= \frac{1}{10}(1{,}62 + 1{,}71 + 1{,}61 + 1{,}82 + 1{,}75 + 1{,}77 + 1{,}82 + 1{,}55 + 1{,}75 + 1{,}63)$$

$$= \frac{1}{10}(17{,}02) = 1{,}702 \approx 1{,}70$$

Bestimmung des Median

Schritt 1: Ordnung der Werte der Verteilung.

1,55; 1,61; 1,62; 1,63; 1,71; 1,74; 1,75; 1,77; 1,82; 1,82

Schritt 2: Bestimmung des Median mit Hilfe der Formel (bei n = gerade).

$$x_{med} = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) = \frac{1}{2}(x_{(5)} + x_{(6)})$$

$$= \frac{1}{2}(1{,}71 + 1{,}74) = \frac{1}{2}(3{,}45) = 1{,}725 \approx 1{,}73$$

Bestimmung des Modus

Der Modus kann nicht bestimmt werden, da keine eindeutige unimodale Verteilung vorliegt.

Bestimmung der Varianz

Für die Bestimmung der (empirischen) Varianz werden das arithmetische Mittel (dieses ist ja bereits bekannt: 1,70) sowie die folgende Hilfstabelle benötigt:

x	$(x_i - \bar{x})^2$
1,62	(1,62 – 1,70) ² = 0,0064
1,71	(1,71 – 1,70) ² = 0,0001
1,61	(1,61 – 1,70) ² = 0,0081
1,82	(1,82 – 1,70) ² = 0,0144
1,75	(1,75 – 1,70) ² = 0,0025
1,77	(1,77 – 1,70) ² = 0,0049
1,82	(1,82 – 1,70) ² = 0,0144
1,55	(1,55 – 1,70) ² = 0,0225
1,74	(1,74 – 1,70) ² = 0,0016
1,63	(1,63 – 1,70) ² = 0,0049
	Summe: 0,0798

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{10}(0{,}0798) = 0{,}00798 \approx 0{,}01$$

Bestimmung der Standardabweichung

Die Standardabweichung berechnet sich als positive Wurzel aus der Varianz:

$$s = +\sqrt{s^2} = \sqrt{0{,}01} = 0{,}1$$

Bestimmung des Quartilskoeffizienten der Schiefe

Oberes Quartil:

$n * p = 0{,}75 * 10 = 7{,}5$ -> nicht ganzzahlig, daher gilt (bei k = 8):

$$x_p = x_{(k)} = x_{(8)} = 1{,}77$$

Median: Der Median ist bereits bekannt und liegt bei 1,73.

Unteres Quartil:

$n * p = 0{,}25 * 10 = 2{,}5$ -> nicht ganzzahlig, daher gilt (bei k = 3):

$$x_p = x_{(k)} = x_{(3)} = 1{,}62$$

Der Quartilskoeffizient der Schiefe berechnet sich somit wie folgt:

$$g_{0,25} = \frac{(x_{0,75} - x_{med}) - (x_{med} - x_{0,25})}{x_{0,75} - x_{0,25}} = \frac{(1{,}77 - 1{,}73) - (1{,}73 - 1{,}62)}{(1{,}77 - 1{,}62)} = \frac{0{,}04 - 0{,}11}{0{,}15} = -0{,}4666 \approx -0{,}47$$

Aufgabenteil IV: Zusammenhangsmaße (20 Punkte)

Eine Befragung unter 10 Mitarbeitern/innen eines Unternehmens nach der Dauer ihrer Unternehmenszugehörigkeit (in Jahren) sowie ihrem Gehalt (gerundet) ergab folgende Werte:

Zugehörigkeit (x)	Gehalt (y)	Zugehörigkeit (x)	Gehalt (y)
10	3.200	23	2.300
8	2.400	3	1.400
11	2.500	2	2.000
27	3.100	9	2.600
13	2.800	12	2.500

Berechnen und interpretieren Sie den Konkordanzkoeffizienten nach Kendall.

Für die Berechnung des Konkordanzkoeffizienten ist die folgende Hilfstabelle erforderlich.

x	Rang x	y	Rang y	K	D
2	1	2.000	2	8	1
3	2	1.400	1	8	0
8	3	2.400	4	6	1
9	4	2.600	7	3	3
10	5	3.200	10	0	5
11	6	2.500	5,5	2	1
12	7	2.500	5,5	2	1
13	8	2.800	8	1	1
23	9	2.300	3	1	0
27	10	3.100	9	0	0
				K = 31	D = 13

$$tau = \frac{2*(K-D)}{n*(n-1)} = \frac{2*(31-13)}{10*(10-1)} = \frac{36}{90} = 0,4$$

Der Konkordanzkoeffizient nach Kendall deutet auf eine eher schwache monotone Korrelation hin.

Aufgabenteil V: Lineare Regressionsanalyse (20 Punkte)

Berechnen Sie für die bereits bekannte Verteilung der Gehälter und Unternehmenszugehörigkeiten die lineare Regressionsfunktion und bewerten Sie deren Güte. Passen die gefundenen Ergebnisse mit den Ergebnissen der Berechnung des Konkordanzkoeffizienten nach Kendall zusammen?

Zugehörigkeit (x)	**Gehalt (y)**	**Zugehörigkeit (x)**	**Gehalt (y)**
10	3.200	23	2.300
8	2.400	3	1.400
11	2.500	2	2.000
27	3.100	9	2.600
13	2.800	12	2.500

Auch hier empfiehlt sich wieder die Anlage einer Hilfstabelle.

x	y	x^2	(x * y)
10	3.200	100	32.000
8	2.400	64	19.200
11	2.500	121	27.500
27	3.100	729	83.700
13	2.800	169	36.400
23	2.300	529	52.900
3	1.400	9	4.200
2	2.000	4	4.000
9	2.600	81	23.400
12	2.500	144	30.000
Summe = 118	Summe = 24.800	Summe = 1950	Summe = 313.300
Mittel = 11,8	Mittel = 2.480		

$$b = \frac{\sum_{i=1}^{n}(x_i * y_i) - n*\bar{x}*\bar{y}}{\sum_{i=1}^{n}(x_i^2) - n*\bar{x}^2} = \frac{313300 - 10*11,8*2480}{1950 - 10*11,8^2} = \frac{20660}{557,60} = 37,05$$

$$a = \bar{y} - b*\bar{x} = 2480 - 37,05*11,8 = 2042,81$$

$$y = a + b*x = 2042,81 + 37,05*x$$

Im zweiten Schritt wird das Gütemaß R^2 (ebenfalls mit einer Hilfstabelle) berechnet.

x	y	y^*	$(y^* - \bar{y})^2$	$(y - y^*)^2$
10	3.200	2.413,31	4.447,56	618881,16
8	2.400	2.339,21	19.821,82	3695,42
11	2.500	2.450,36	878,53	2464,13
27	3.100	3.043,16	317.149,19	3230,79
13	2.800	2.524,46	1.976,69	75922,29
23	2.300	2.894,96	172.191,80	353977,40
3	1.400	2.153,96	106.302,08	568455,68
2	2.000	2.116,91	131.834,35	13667,95
9	2.600	2.376,26	10.761,99	50059,59
12	2.500	2.487,41	54,91	158,51
Mittel = 11,8	Mittel = 2.480	//	ESS = 765.418,91	RSS = 1.690.512,91

$$TSS = ESS + RSS = 765418,91 + 1690512,91 = 2455931,83$$

$$R^2 = \frac{ESS}{TSS} = \frac{765418,91}{2455931,83} = 0,31$$

Das Gütemaß in Höhe von 0,31 spricht für einen eher geringen linearen Zusammenhang zwischen x und y, d.h. das lineare Regressionsmodell erklärt die Streuung der abhängigen Variablen eher schlecht. Dies deckt sich mit dem Ergebnis der Berechnung des Konkordanzkoeffizienten.

Der Blick auf das mit PSPP erstellte Streudiagramm stützt diese Interpretation ebenfalls.