4. Aufgabenblatt vom Montag, den 22.Mai 2017 zur Vorlesung

Deskriptive Statistik für Soziologinnen und Soziologen (Mariana Nold)

Thema: Bivariate Exploration von quantitativen und qualitativen Merkmalen: Korrelation

Abgabe: bis Freitag, den 2. Juni 2017, Briefkasten des Instituts für Soziologie, in der Nähe der Cafeteria der Carls-Zeiss-Straße (In einem Umschlag, an mich adressiert) **oder** Mittwoch, Donnerstag und Freitag von 13-15 Uhr im Sekretariat von Frau Prof. Leuze, CZ-Straße 2, R286

15. **Wortschatz von Kindern** (20 Punkte) **Aufgabe abgeben!**
(in Anlehnung an: Fahrmeir et al, Statistik Der Weg zur Datenanalyse, S .151)
Bei fünf zufällig ausgewählten Kindern wurden der Körpergröße X in cm und die Wortschatz Y gemessen. Dabei erfolgte die Messung des Wortschatzes über die Anzahl der verschiedenen Wörter, die die Kinder in einem Aufsatz über die Ergebnisse in ihren Sommerferien benutzten. Nehmen wir an, wir hätten folgende Daten erhalten:

Kind i	1	2	3	4	5
Körpergröße x_i	130	112	108	114	136
Wortschatz y_i	37	30	20	28	35

Tabelle 1: Die Körpergröße X und der Wortschatz Y in cm gemessen von 5 zufällig ausgewählten Kindern.

(a) Zeichnen Sie ein Streudiagramm.
Lösung:
Siehe Abbildung **??**.

(b) Erklären Sie an Hand dieses Beispiels was eine

- positive bzw. negative lineare Korrelation
- positive bzw. negative monotone Korrelation

inhaltlich bedeuten.
Lösung:
Eine Korrelation beschreibt die Beziehung zwischen zwei Merkmalen. Man kann als Synonym für Korrelation den Begriff Wechselbeziehung verwenden. Es bedeutet, dass einen Einfluss von einem Merkmal X auf eine anderes Merkmal Y gibt, wenigstens in eine Richtung. Ist der Einfluss nur in eine Richtung, dann liegt eine Spezialfall einer Wechselbeziehung vor. Dann sind die Rollen des abhängigen und des unabhängigen Merkmals klar verteilt. Das abhängige Merkmal wird vom unabhängigen Merkmal beeinflusst.
Die Korrelation mach keine Aussage darüber, ob die Rollen klar verteilt sind. Ein ***positive*** *Korrelation der Merkmale X und Y bedeutet, einfach gesagt, „je mehr X, desto mehr Y."Diese Aussage beschreibt eine mittlere Tendenz. Eine positive Korrelation zwischen Körpergröße und Wortschatz bedeutet, dass man im Mittel beobachtet, dass größere Kinder einen höheren Wortschatz haben, oder umgekehrt,*

Kinder mit größerem Wortschatz größer sind. Die Korrelation macht keine Aussage über die Richtung des Zusammenhangs.
*Analog beschreibt eine **negative** Korrelation zwischen zwei Merkmalen X und Y eine negative mittlere Tendenz. Wenn z. B. mit zunehmender Zeit, seit Beginn der Bearbeitung einer Aufgabe die Konzentration sinkt, dann sind die Bearbeitungsdauer X und die Konzentrationsleistung Y negativ korreliert. In diesem Beispiel ist inhaltlich klar, dass die Bearbeitungsdauer ursächlich ist für die fallende Konzentration. Die Korrelation selbst beschreibt eine ungerichtete Wechselbeziehung.*
In folgendem Artikel wir Korrelation am Beispiel gemessener Intelligenz und der Schulleistung diskutiert und erklärt:
`http://www.zeit.de/1974/44/was-ist-eine-korrelation`

(c) Ist die folgende Aussage falsch oder richtig: Es ist im Allgemeinen möglich, dass ein positiver linearer Zusammenhang vorliegt, aber kein positiver monotoner Zusammenhang.
Lösung: *Diese Aussage ist falsch. Wenn ein positiver linearer Zusammenhang vorliegt, dann ist dieser Zusammenhang auch monoton. Jede Gerade mit positiver Steigung ist eine monoton wachsende Funktion. Umgekehrt gilt allerdings, ein positiver monotoner Zusammenhang braucht nicht linear zu sein.*

(d) Schreiben Sie die Tabelle der Ränge von X und Y, berechnen die den Rangkorrelationskoeffizient r_{XY}^{SP} und interpretieren Sie diesen Wert.
Lösung:
Siehe Tabelle **??**. *Die Formel für den Rangkorrelationskoeffizient ist:*

$$r_{XY}^{SP} = \frac{\sum_{i=1}^{n}(rg(x_i) - \bar{rg}_X) \cdot (rg(y_i) - \bar{rg}_Y)}{\sqrt{\sum_{i=1}^{n}(rg(x_i) - \bar{rg}_X)^2} \cdot \sqrt{\sum_{i=1}^{n}(rg(y_i) - \bar{rg}_Y)^2}}, \tag{1}$$

dabei gilt:

$$\bar{rg}_X = \frac{1}{n}\sum_{i=1}^{n} rg(x_i) = \frac{1}{n}\sum_{i=1}^{n} i = (n+1)/2$$

und

$$\bar{rg}_Y = \frac{1}{n}\sum_{i=1}^{n} rg(y_i) = \frac{1}{n}\sum_{i=1}^{n} i = (n+1)/2$$

Für die mittleren Ränge $\bar{rg}_X$ und $\bar{rg}_Y$ ergibt sich also in unserem Beispiel: $\bar{rg}_X = \bar{rg}_Y = (5+1)/2 = 3$
*Mit Hilfe der Tabelle kann man die Formel (**??**) in drei Teile zerlegen. Im Zähler steht die Summe über die Produkte die in Zeile 6 der Tabelle berechnet sind. Summiert man diese Summe auf, so erhält man den Wert 8. Der Nenner besteht aus zwei Teilen, die miteinander multipliziert werden. Um den ersten Teil zu berechnen, summiert man die 7. Zeile auf und zieht dann die Wurzel. Es ergibt sich der Wert $\sqrt{10} \approx 3.162$. Entsprechend erhält man den zweiten Teil aus Zeile 8. Der Werte ist $\sqrt{10} \approx 3.162$. Insgesamt ergibt sich also*

$$r_{X,Y}^{SP} = \frac{8}{\sqrt{10} \cdot \sqrt{10}} = 0.8.$$

Dieser Wert spricht für einen ziemlich starken monotonen Zusammenhang. Er macht keine Aussage darüber, ob dieser Zusammenhang linear ist, oder nicht.

Kind i	1	2	3	4	5
$rg(x_i)$	4	2	1	3	5
$rg(y_i)$	5	3	1	2	4
$rg(x_i) - \bar{rg}_X$	1	-1	-2	0	2
$rg(y_i) - \bar{rg}_Y$	2	0	-2	-1	1
$(rg(x_i) - \bar{rg}_X)(rg(y_i) - \bar{rg}_Y)$	2	0	4	0	2
$(rg(x_i) - \bar{rg}_X)^2$	1	1	4	0	4
$(rg(y_i) - \bar{rg}_Y)^2$	4	0	4	1	1

Tabelle 2: Die Ränge der Körpergröße X und des Wortschatzes Y von fünf zufällig ausgewählten Kindern. Erweiterte Tabelle zur Berechnung des Rangkorrelationskoeffizient.

(e) Berechnen Sie nun den Korrelationskoeffizienten nach Pearson $r_{X,Y}$ und interpretieren Sie diesen Wert.
Lösung:
Die Formel für den Korrelationskoeffizienten nach Pearson $r_{X,Y}$ ist:

$$r_{X,Y} = \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_X \cdot \hat{\sigma}_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

Man erkennt, dass der Aufbau der gleiche ist, wie im Rangkorrelationskoeffizienten. Der Unterschied ist: Man rechnet hier nicht mit den Rängen der Beobachtungen, sondern mit den Beobachtungen selbst. Man berechnet $\bar{x} = 120$ und $\bar{y} = 30$. Entsprechend bildet man die Tabelle: Wie oben berechnet man die Summer der 6. Zeile

Kind i	1	2	3	4	5
Körpergröße x_i	130	112	108	114	136
Wortschatz y_i	37	30	20	28	35
$x_i - \bar{x}$	10	-8	-12	-6	16
$y_i - \bar{y}$	7	0	-10	-2	5
$(x_i - \bar{x})(y_i - \bar{y})$	70	0	120	12	80
$(x_i - \bar{x})^2$	100	64	144	36	256
$(y_i - \bar{y})^2$	49	0	100	4	25

Tabelle 3: Die Körpergröße X und des Wortschatzes Y von fünf zufällig ausgewählten Kindern. Erweiterte Tabelle zur Berechnung des Korrelationskoeffizient.

um den Zähler zu bestimmen. Man erhält 282. Für den Zähler erhält man beruhend auf der 7. Zeile. Es ergibt sich $\sqrt{600}$. Den zweiten Teil des Produkts im Zähler erhält man mit Zeile 9. Es ergibt sich $\sqrt{178}$. Zusammenfassend erhält man

$$r_{X,Y} = \frac{282}{\sqrt{600} \cdot \sqrt{178}} = 0.863.$$

Diese Zahl spricht für einen starken linearen Zusammenhang.

(f) Die Tabelle **??** enthält zusätzlich das Alter Z der Kinder, berechnen Sie jeweils r_{YZ} und r_{XZ} und interpretieren Sie auch diese Werte.
Lösung:

Kind i	1	2	3	4	5
Körpergröße x_i	130	112	108	114	136
Wortschatz y_i	37	30	20	28	35
Alter z_i	12	7	6	7	13

Tabelle 4: Die Körpergröße X in cm gemessen, der Wortschatz Y und das Alter Z von fünf zufällig ausgewählten Kindern.

Wie oben in Teilaufgabe e, berechnet man $r_{X,Z} = 0.995$ *und* $r_{Y,Z} = 0.867$. *Man erkennt, dass der Körpergröße* X *und dem Alter* Z *ein sehr deutlicher linearer Zusammenhang besteht. Auch zwischen dem Alter* Z *und dem Wortschatz* Y *besteht ein deutlicher Zusammenhang.*

16. **Kreuztabellen interpretieren: Habilitationsdichte** (11 Punkte) **Aufgabe abgeben!**

Die Habilitation ist die höchstrangige Hochschulprüfung in Deutschland durch Anfertigung einer wissenschaftlichen Arbeit. In einer Untersuchung zur Habilitationsdichte an deutschen Hochschulen wurden u. a. die Merkmale Geschlecht und Habilitationsfach erhoben. In Tabelle **??** ist - nach Fächern aufgeschlüsselt- zusammengefasst, wieviele Habilitationen im Jahre 2015 erfolgreich abgeschlossen wurden (Quelle: Statistisches Bundesamt) Hier stellt sich die Frage, ob die Habilitationsdichte in den einzelnen Fächern im Jahr 2015 geschlechtsspezifisch ist, d. h. man interessiert sich dafür, ob zwischen den Merkmalen Geschlecht (=:Y) und Habilitationsfach (=:X) ein Zusammenhang besteht.

$$\text{Ausprägungen } X \triangleq \begin{cases} a_1, & \text{Geisteswissenschaften} \\ a_2, & \text{Rechts-,Wirtschafts-,Sozialwiss.} \\ a_3, & \text{Mathe u. Naturwiss.} \\ a_4, & \text{Human-,Gesundheitswiss.} \\ a_5, & \text{übrige Fächer} \end{cases}$$

$$\text{Ausprägungen } Y \triangleq \begin{cases} b_1, & \text{Frauen} \\ b_2, & \text{Männer} \end{cases}$$

		X					$\sum$
		a_1	a_2	a_3	a_4	a_5	
Y	b_1	77	62	66	225	32	
	b_2	159	139	181	571	115	
	$\sum$						

Tabelle 5: Habilitationen im Jahre 2015 erfolgreich abgeschlossen wurden nach Fächern und Geschlecht aufgeschlüsselt.

(a) Ergänzen sie die fehlenden Randhäufigkeiten.
Lösung:
Siehe Tabelle **??**

		X					$\sum$
		a_1	a_2	a_3	a_4	a_5	
Y	b_1	77	62	66	225	32	462
	b_2	159	139	181	571	115	1165
	$\sum$	236	201	247	796	147	1627

Tabelle 6: Habilitationen im Jahre 2015 erfolgreich abgeschlossen wurden nach Fächern und Geschlecht aufgeschlüsselt mit Randhäufigkeiten.

(b) Berechnen Sie die Randverteilungen (= marginale relative Häufigkeit).
Lösung:
Die Randverteilung des Merkmals Geschlecht (Y) ist $f_{1,\bullet} = 0.284, f_{2,\bullet} = 0.716$.
Die Randverteilung des Merkmals Habilitationsfach (X) ist:
$f_{\bullet,1} = 0.145, f_{\bullet,2} = 0.124, f_{\bullet,3} = 0.152, f_{\bullet,4} = 0.489, f_{\bullet,5} = 0.090$

(c) Wie hoch ist der Anteil der Frauen, die im Jahr 2015 eine Habilitation abgeschlossen haben.
Lösung:
Der Anteil beträgt $\frac{462}{1627} = 0.284$. *Von allen die* 2015 *eine Habilitation abgeschlossen haben, sind* 28.4% *Frauen.*

(d) Wie hoch ist der Anteil an Habilitationen aus dem Fachbereich „Mathematik und Naturwissenschaften“?
Lösung:
Der Anteil ist $f_{\bullet,3} = 0.152$. *Damit sind* 15.2% *der Habilitationen aus diesem Fachbereich.*

(e) Berechnen Sie die bedingten Verteilungen (= bedingte relative Häufigkeit) gegeben dem Fachbereich und interpretieren Sie das Ergebnis.
Lösung:
Siehe Tabelle **??**. *Die Tabelle zeigt wie sich in den einzelnen Fachbereichen die Anteile an Frauen und Männern darstellen. Ein Vergleich mit der Randverteilung zeigt, wo man mehr bzw. weniger Frauen als mit Bezug auf die Randverteilung erwartet, findet. In den Geisteswissenschaften ist der Frauenanteil am höchsten.*

		X					
		a_1	a_2	a_3	a_4	a_5	
Y	b_1	0.326	0.308	0.267	0.283	0.218	0.284
	b_2	0.674	0.692	0.733	0.717	0.782	0.716
		1.0	1.0	1.0	1.0	1.0	1.0

Tabelle 7: Bedingte Verteilung des Geschlechts (Y) gegeben dem Habilitationsfach (X) .

(f) Berechnen Sie die bedingten Verteilungen (= bedingte relative Häufigkeit) gegeben das Geschlecht und interpretieren Sie das Ergebnis.
Lösung:
Siehe Tabelle **??**. *Die Tabelle zeigt für die Gruppe der Männer und die Gruppe der Frauen jeweils, wie hoch der Anteil der Habilitationen in den einzelnen Fachbereichen ist. Unter den Frauen wurden die meisten Habilitationen in den Human- und Gesundheitswissenschaften abgeschlossen (*48.7%*). In der Gruppe der Männer liegt dieser Fachbereich ebenfalls ganz vorne mit* 49.0%

		X					
		a_1	a_2	a_3	a_4	a_5	
Y	b_1	0.167	0.134	0.143	0.487	0.069	1.0
	b_2	0.136	0.119	0.155	0.490	0.099	1.0
		0.145	0.124	0.152	0.489	0.090	1.0

Tabelle 8: Bedingte Verteilung des Habilitationsfach (X) gegeben dem Geschlechts (Y).

17. **Die Korrelation mit STATA berechnen**

Das Streudiagramm **??** zeigt die Spielbewertung eines neuen Spiels aufgetragen auf der Ordinate und die Mathe-Punkte (hier simuliert, nicht aus den PISA-Daten) von 300 Schülerinnen bzw. Schülern. Sie finden den entsprechenden Datensatz auf den Rechnern im Methoden-Labor im Ordner `Methoden/Statistik`.

(a) Öffnen Sie den Datensatz und geben Sie den Befehl `summarize` in das `command`-Fenster ein. Interpretieren Sie die von `STATA` erzeugte Tabelle.

(b) Geben Sie die Befehle `graph box x` und `graph box y` ein und interpretieren Sie die entsprechenden Boxplots.

(c) Erzeugen Sie mit dem Befehl `scatter y x` das Streudiagramm.

(d) Berechnen Sie mit `pwcorr x y` den Korrelationskoeffizient nach Pearson. Wie ändert sich das Ergebnis, wenn Sie den Befehl `pwcorr y x` eingeben. Interpretieren Sie diese Veränderung inhaltlich.

(e) Berechnen Sie mit Hilfe des Befehls `spearman x y` den Wert des Rangkorrelationskoeffizienten.