Korrelation durch gemeinsame Elemente

von Robert M. Lynch, Greeley, übersetzt von Karl Röttel, Buxheim

Zusammenfassung

Diese Arbeit behandelt die Korrelation durch gemeinsame Elemente (KGE), ihre Erweiterung auf negative Korrelation und die Erzeugung bivariat normalverteilter Stichproben für eine festgelegte Korrelationsmatrix.

Einleitung

Die Pearsonschen Produkt-Moment-Korrelationen und ihre Darstellung in Matrizenform sind Themen, die im 1. oder 2. Jahr des Statistikunterrichtes abgedeckt werden. Vor kurzem hat der Autor in den Unterricht Material zum Ziehen von Stichproben aus bivariat normalverteilten Grundgesamtheiten mit festgelegten Korrelationen einbezogen.

Dies erlaubte ihm, die Beziehung zwischen mehreren Variablen, die Stichprobenverteilung von r, die Mehrfachkollinearität und die Erzeugung von Zufallszahlen zu besprechen. Zur Entwicklung des Materials erweiterte der Autor ein Verfahren, das sich ursprünglich auf die KGE bezog und von Carl Fischer (1933) vorgelegt wurde.

Die Methode mit den gemeinsamen Elementen ist ein einfaches und geeignetes Verfahren, Stichproben aus einer bivariat normalverteilten Grundgesamtheit mit einem festgelegten Korrelationskoeffizienten zu ziehen. Sie kann mit der meisten Standard-Statistik-Software einschließlich PC-Statistik-Software (z.B. Minitab, NCSS) oder Spreadsheets (z.B. Lotus 1-2-3, Quattro) durchgeführt werden.

Tiefere Kenntnisse der Computerbedienung oder Statistik sind nicht erforderlich. Die Schüler werden zudem mit einer andersartigen und weniger üblichen Interpretation der Korrelation bekanntgemacht.

Dieser Bericht zeigt die Korrelation gemeinsamer Elemente, ihre Erweiterung auf negative Korrelation und die Erzeugung bivariat normalverteilter Stichproben für eine festgelegte Korrelationsmatrix.

Die KGE

Die KGE verwendet (X_1, X_2)-Paare, die auf der Addition von Elementen beruhen, wobei einige Elemente dem Paar gemeinsam sind, einige nicht. Das Ausmaß der Gemeinsamkeit ist bestimmt durch die erwünschte Korrelation.

Stochastik in der Schule14(1994), Nr. 3, S. 39-43

Der KGE-Koeffizient ρ, wie ihn Fischer (1933) vorstellte, wird bestimmt durch $\rho = N_{12}/\sqrt{N_{11}N_{22}}$, (1)
worin N_{12} die Anzahl der gemeinsamen Elemente, N_{11} die Anzahl der Elemente, die summiert X_1 erzeugen, und N_{22} die Anzahl der Elemente, die summiert X_2 erzeugen, bezeichnen.
Zur Veranschaulichung wurden in Tabelle 1 drei (X_1, X_2)-Paare erzeugt, wobei jedes Paar aus der Summierung von 5 Elementen entsteht. Man sieht, daß jedes Paar 3 gemeinsame und 2 nicht-gemeinsame Elemente hat. Die Summierung von 5 Elementen für jedes X_1 und jedes X_2 erzeugt ein (X_1, X_2)-Paar.
Der Korrelationskoeffizient ρ der Grundgesamtheit zwischen (X_1, X_2)-Punkten, beruhend auf (1), ist 0,60 ($\rho = 3/\sqrt{5 \cdot 5}$).
Stichproben, die durch das KGE-Verfahren erzeugt sind, liefern die Pearsonschen Produkt-Moment-Korrelationen, die ρ schätzen.

Die Formel (1), wenn die Anzahl der für X_1 und X_2 verwendeten Elemente dieselbe ist, vereinfacht sich zu $\rho = N_{12}/N$, (2)
worin N_{12} die Anzahl der gemeinsamen Elemente und N die Anzahl der Elemente, die summiert werden, um jedes X_1 und X_2 zu erzeugen, bedeuten.

Tabelle 1: (X_1, X_2)-*Paare mit gemeinsamen Elementen*

		X_1	X_2
Paar 1		0,19	0,19
		0,57	0,57
		0,72	0,72
		0,10	0,14
		0,48	0,03
	Punkt	2,15	1,65
Paar 2		0,54	0,54
		0,53	0,53
		0,55	0,55
		0,25	0,06
		0,69	0,88
	Punkt	2,56	2,56
Paar 3		0,19	0,19
		0,96	0,96
		0,82	0,82
		0,05	0,78
		0,28	0,24
	Punkt	2,30	2,99

Formel (1) erhält man leicht (Snedecor (1967[6])) aus $\rho = \mathrm{Cov}(X_1, X_2)/(\sigma_1 \cdot \sigma_2)$.
Wenn X_1 die Summe von n_{11} unabhängigen Zügen aus einer Grundgesamtheit mit der Standardabweichung σ ist, dann ist $\sigma_1 = \sigma\sqrt{n_{11}}$ und entsprechend $\sigma_2 = \sigma\sqrt{n_{22}}$.
Um $\mathrm{Cov}(X_1, X_2)$ zu finden, setze man $X_1 = c + u_1$, $X_2 = c + u_2$, worin c, der gemeinsame Anteil, die Summe der gleichen Menge von n_{12} Zügen ist.
Wenn wir Züge aus einer Grundgesamtheit mit dem Mittelwert 0 annehmen, haben X_1 und X_2 den Mittelwert 0, und die Kovarianz $\mathrm{Cov}(X_1, X_2)$ ist gleich dem Mittelwert von (X_1, X_2) und gleich dem Mittelwert von $(c + u_1)(c + u_2)$. Dies ist der Mittelwert von c^2 oder die Varianz von c, weil die Terme cu_2, cu_1 und u_1u_2 alle den Mittelwert 0 haben. (Man beachte, daß c, u_1 und u_2 aus unabhängigen Zügen resultieren.)

Die Varianz von c ist schließlich $\sigma^2 n_{12}$, was $\rho = \sigma^2 n_{12}/(\sigma\sqrt{n_{11}} \cdot \sigma\sqrt{n_{22}}) = n_{12}/\sqrt{n_{11}n_{22}}$ ergibt.
Bei der Arbeit im Klassenzimmer verwenden die Schüler einen das Intervall gleichmäßig erfasssenden Zufallszahlengenerator, um die gemeinsamen und nicht-gemeinsamen Elemente zu erzeugen. Dies verhindert eine Verzerrung der Größe und der resultierenden Varianz zwischen den gemeinsamen und den nicht-gemeinsamen Elementen. Die Schüler können leicht den Einfluß sehen, den die gemeinsamen Elemente beim Erzeugen des Paares haben, und daraus erkennen, weshalb eine Korrelation auftritt und wie groß sie zu erwarten ist. Die Schüler müßten daran erinnert werden, daß die Summen vieler gleichmäßiger [0,1]-Zufallszahlen normalverteilt sind.
Dies verhindert eine Verzerrung der Größe und der resultierenden Varianz zwischen den gemeinsamen und den nicht-gemeinsamen Elementen.

Negative Korrelationen

Um eine Stichprobe mit festgelegter negativer Korrelation aus einer bivariat normalverteilten Grundgesamtheit zu ziehen, würde ein Schüler so vorgehen:
1. Eine 2variable Stichprobe erstellen, wie sie für den positiven Wert der gewünschten Korrelation dargestellt ist.
2. Alle Werte der einen Variablen mit -1 multiplizieren.
3. Die Pearsonsche Produkt-Moment-Korrelation zwischen den beiden Variablen berechnen.

Die Korrelation zwischen diesen zwei Variablen schätzt die festgelegte negative Korrelation. Eine Veranschaulichung ist im nächsten Abschnitt gegeben. Die Gewichtung mit -1 spiegelt die Punkte des 1. Quadranten in den 2. oder 4. Quadranten. Die Lage des Quadranten hängt von der Variablen ab, die gewichtet wurde.

Das Erstellen von Stichproben zu einer hypothetischen Korrelationsmatrix

Um Stichproben aus hypothetischen bivariat normalverteilten Grundgesamtheiten zu bilden, muß vom Schüler verlangt werden:
1. Festsetzen der gewünschten Korrelationsmatrix.
2. Die Anzahl der gemeinsamen und nicht-gemeinsamen Elemente bestimmen, die genommen werden, um die festgesetzten Korrelationen zu erzeugen.

3. Die bivariat normalverteilten Stichproben erzeugen durch Benutzen des Schemas mit den gemeinsamen Elementen.
4. Die geeignete(n) Variable(n) mit -1 gewichten für Variablen, die mit anderen Variablen negativ korrelieren soll(en).
5. Die Pearsonsche Produkt-Moment-Korrelation auf alle Paare der Variablen anwenden.

Beispiel für den Ablauf: In Tabelle 2 ist die hypothesierte (vorausgesetzte) Korrelationsmatrix festgesetzt als die Korrelationsmatrix der Grundgesamtheit. Der Tabelle 2 entnimmt der Schüler, daß die drei Variablen und alle Korrelationen aus 5 Elementen erzeugt werden können durch Werte mit entweder zwei ($\rho = \pm 0,40$) oder drei ($\rho = 0,60$) gemeinsamen Elementen.

Tabelle 2:

Korrelationen der Grundgesamtheit

	X_1	X_2	X_3
X_1	1,00	0,40	$-0,40$
X_2		1,00	$-0,60$
X_3			1,00

Tabelle 3: *Gleichförmig verteilte Zufallszahlen und X_1-, X_2- und X_3-Werte.*

X_1 abcde	X_2 abfgh	X_3 (defgh)	a	b	c	d	e	f	g	h
2.15	1.04	-0.95	0.19	0.57	0.72	0.19	0.48	0.14	0.03	0.11
2.56	2.14	-2.01	0.54	0.53	0.55	0.25	0.69	0.06	0.88	0.13
2.30	2.40	-1.58	0.19	0.96	0.82	0.05	0.28	0.78	0.24	0.23
3.04	2.24	-2.41	0.42	0.79	0.47	0.79	0.59	0.69	0.12	0.24
2.56	1.79	-2.17	0.33	0.34	0.85	0.40	0.65	0.38	0.73	0.01
2.47	2.92	-1.56	0.65	0.90	0.74	0.18	0.00	0.49	0.09	0.79
3.17	2.00	-1.73	0.95	0.69	0.17	0.63	0.73	0.20	0.08	0.09
3.56	2.65	-2.67	0.88	0.47	0.83	0.39	0.99	0.43	0.75	0.11
2.14	2.22	-1.92	0.18	0.74	0.59	0.50	0.13	0.33	0.47	0.50
2.18	1.65	-1.16	0.26	0.76	0.64	0.04	0.48	0.29	0.06	0.29
3.34	3.25	-3.39	0.66	0.92	0.04	0.93	0.79	0.33	0.84	0.50
2.16	2.11	-1.93	0.45	0.52	0.41	0.20	0.58	0.59	0.54	0.01
3.83	2.35	-2.64	0.52	0.98	0.54	0.92	0.88	0.03	0.77	0.05
2.51	3.01	-2.42	0.65	0.82	0.16	0.20	0.68	0.29	0.29	0.96
2.16	3.05	-3.03	0.78	0.09	0.43	0.67	0.19	0.63	0.99	0.55
2.02	2.09	-1.71	0.60	0.24	0.72	0.19	0.28	0.29	0.85	0.11
1.65	1.52	-1.30	0.09	0.64	0.41	0.25	0.26	0.17	0.15	0.47
2.98	2.56	-3.41	0.34	0.40	0.66	0.65	0.93	0.26	0.83	0.73
2.83	3.30	-2.53	0.93	0.53	0.69	0.38	0.30	0.29	0.98	0.58
2.66	3.37	-2.23	0.85	0.83	0.43	0.28	0.27	0.95	0.51	0.23
2.23	2.72	-2.41	0.06	0.78	0.85	0.30	0.23	0.56	0.82	0.50
0.92	1.58	-1.40	0.48	0.05	0.04	0.19	0.17	0.04	0.78	0.23
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...

In Tabelle 3 wurden Beobachtungen für jede der drei Variablen erzeugt durch Verwenden der 9 Spalten, die mit a bis h bezeichnet sind. Dieses Beispiel kann mit 9 Spalten durchgeführt werden. Einige Probleme erfordern mehr, andere hingegen weniger als 9 Spalten. Jedes Element in jeder Spalte ist eine gleichförmig verteilte Zufallszahl zwischen 0 und 1.

1. Um die X_1 zu erzeugen, werden die Elemente a, b, c, d, e in jeder Zeile addiert.
2. Um die X_2 zu erzeugen, müssen den Werten X_1 und X_2 2 Elemente gemeinsam und 3 nicht-gemeinsam sein ($\rho = 0,40$). Zwei gemeinsame Elemente (a, b) und 3 nicht-gemeinsame (f, g, h) werden addiert.
3. Um die X_3 zu erzeugen, müssen 2 Elemente mit X_1 und 3 mit X_2 gemeinsam sein. Die Elemente d, e werden für X_1, die Elemente f, g, h für X_2 verwendet. Die Elemente d, e, f, g, h werden addiert und mit -1 multipliziert, um die X_3 zu erzeugen. Dies schätzt die angesetzte negative Korrelation zwischen (X_1, X_2) und (X_2, X_3).

50 Beobachtungen der 3 Variablen aus Tabelle 3 erzeugten die Korrelationsmatrix in Tabelle 4. Alle Korrelationen sind einsichtig und sicher innerhalb einer Standardabweichung der Korrelationen der Grundgesamtheit.

Tabelle 4:

Korrelationen der Stichprobe

	X_1	X_2	X_3
X_1	1,00	0,51	$-0,46$
X_2		1,00	$-0,64$
X_3			1,00

Die oben beschriebenen Techniken können zusammen mit passender Software den Anfängern in den Statistikkursen einen guten Einblick in Korrelationen und die Stichprobenverteilung von r vermitteln. Tabellenkalkulationsprogramme (spreadsheets) haben sich bei diesen Techniken gut bewährt, da sie den Schülern gestatten, die Verfahren schrittweise durchzuführen, und das Verstehen bei jedem Schritt visuell bekräftigen.

Literatur:

Fischer Carl H.: Common Elements Correlation. In: The Annals of Mathematical Statistics 4(1933), 103 - 126.
Snedecor George W. und Cochran William G.: Statistical Methods. Ames, Iowa 1967[6].

Dr. Karl Röttel,
Am Aschweg 57, 85114 Buxheim