KORRELATION ZWISCHEN DEN AUGENZAHLEN VON ZWEI WÜRFELN

nach L.A. Morgan

Orginaltitel in "Teaching Statistics" vol. 9(1987) Nr. 2:

Correlation from Throws of the Dice

Bearbeitung: Hans-Joachim Bentz, Osnabrück, und Manfred Borovcnik, Klagenfurt

Zusammenfassung: Anhand von wiederholten Wurfserien mit zwei Würfeln wird klar gemacht, welchen Schwankungen der Korrelationskoeffizient von Stichprobe zu Stichprobe unterworfen ist, und, daß er speziell bei kleinen Stichproben erheblich vom tatsächlichen Wert abweichen kann. Zur Demonstration der Problematik bietet es sich an, die Daten des Würfelns vom PC erzeugen zu lassen.

1. Schätzung von Parametern aus einer Stichprobe

Erwartungswert

Der Erwartungswert $\mu=E(X)$ einer Zufallsvariablen X wird üblicherweise durch das arithmetische Mittel $\bar{x}$ von n Daten von X : x_1, x_2, ..., x_n (einer Stichprobe von X) geschätzt. Klar, daß dieses $\bar{x}$ erheblich von E(X) abweichen kann.

Beispiel: Die Zufallsvariable X sei die Augenzahl eines (idealen) Würfels. Für den Erwartungswert gilt: $\mu=3,5$. Fünfmaliges Werfen des Würfels ergibt z.B. folgende Datenliste (die Stichprobe):

$$x_1=3,\ x_2=4,\ x_3=2,\ x_4=3,\ x_5=1.$$

Der Mittelwert $\bar{x}$ der Stichprobe ist daher

$$\bar{x} = \frac{3 + 4 + 2 + 3 + 1}{5} = 2,6$$

und weicht vom Erwartungswert μ "erheblich" ab.

Korrelationskoeffizient

Aus den Daten einer Stichprobe schätzt man den Erwartungswert, einen speziellen Parameter der Wahrscheinlichkeitsverteilung von X, mehr oder weniger zuverlässig. Was für den Erwartungswert μ leicht überschaubar ist, trifft im Prinzip auf den Korrelationskoeffizienten ρ, einem speziellen Parameter der Wahrscheinlichkeitsverteilung von zwei Variablen X und Y zu.

Bemerkung: ρ wird auch als Erwartungswert festgelegt, und zwar so:

$$\rho = \frac{E(X - \mu)(Y - \nu)}{\sigma_X \ \sigma_Y}$$

wobei μ und ν die Erwartungswerte von X bzw. Y, σ_X und σ_Y die entsprechenden Standardabweichungen sind. Die Berechung von ρ ist nur in einfachen Fällen elementar. Für die folgenden Überlegungen ist es nicht unbedingt notwendig, die obige Festlegung von ρ wirklich zu verstehen.

Dieser (gewöhnliche) Korrelationskoeffizient ρ mißt indirekt den Grad des gemeinsamen Variierens von zwei Zufallsvariablen X und Y. Je "enger" die zweidimensionale Wahrscheinlichkeitsverteilung längs einer Geraden gruppiert ist, desto stärker der Zusammenhang zwischen den Variablen, umso größer auch der numerische Wert des Korrelationskoeffizienten $\rho = \rho(X,Y)$. In aller Regel verfügt man nicht über die Kenntnis der gemeinsamen Wahrscheinlichkeitsverteilung der Zufallsvariablen X und Y sondern lediglich über Stichprobeninformation, das ist ein Satz von n Datenpaaren:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

Daraus berechnet man den Korrelationskoeffizienten nach der üblichen Formel:

$$r = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2 \; \Sigma(y_i-\bar{y})^2}} = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}}$$

Bemerkung: In der englischsprachigen Literatur, aber auch in der weiterführenden statistischen Literatur (zur Varianzanalyse z.B.) sind die folgenden Bezeichnungen SS_{xx}, SS_{xy} üblich:

SS_{xx} = "Sum of Squares of x"
= "Summe der quadratischen Abweichungen der x-Daten von ihrem Mittelwert $\bar{x}$"
= $\Sigma(x_i-\bar{x})^2$,

SS_{xy} = "Sum of Squares of xy"
= "Summe der Produkte der Abweichungen der x-Daten mit den Abweichungen der y-Daten vom jeweiligen Mittelwert $\bar{x}$ bzw. $\bar{y}$".
= $\Sigma(x_i-\bar{x})(y_i-\bar{y})$

Die Summation erstreckt sich jeweils über die Indices i=1,2,...,n.

Auch dieser Stichprobenkorrelationskoeffizient r weicht mitunter erheblich vom "wirklichen" Korrelationskoeffizienten ρ für die gemeinsame Wahrscheinlichkeitsverteilung ab, insbesondere wenn die Stichproben klein sind. Dies wird, im Gegensatz zum Sachverhalt mit dem Erwartungswert μ nicht nur von Lernenden sondern auch von Anwendern oft "übersehen".

2. Fluktuation des Korrelationskoeffizienten in Stichproben

Die Variablen X und Y sind unkorreliert: $\rho=0$

Zwei unterscheidbare Würfel werden geworfen (rot und blau z.B.). Das Versuchsergebnis ist jeweils ein Paar (x,y), fünf Würfe ergeben z.B. folgende Daten:

(3,3), (1,4), (5,2), (4,3), (6,1).

Der Stichprobenkorrelationskoeffizient r ist nach obiger Formel gleich -0,96. Falls das letzte Paar statt (6,1) nun (6,5) lautet, so würde sich der Stichprobenkorrelationskoeffizient drastisch zu r=+0,05 ändern.

Sind X und Y die Zufallsvariablen Augenzahl des "roten" bzw. "blauen" Würfels, so ist die gemeinsame Wahrscheinlichkeitsverteilung von X und Y eine Gleichverteilung auf dem Gitter $\{1,2,...,6\}x\{1,2,...,6\}$. Es ist intuitiv einsichtig, daß Y und X "minimal" gemeinsam variieren (z.B. je größer X desto größer im "Durchschnitt" Y). Eine formale Berechnung des Korrelationskoeffizienten ρ ergibt tatsächlich $\rho=0$. Die folgende Figurenfolge gibt die Daten der beiden fingierten Stichproben von oben sowie die Gleichverteilung der Zufallsvariablen (X,Y) wieder:

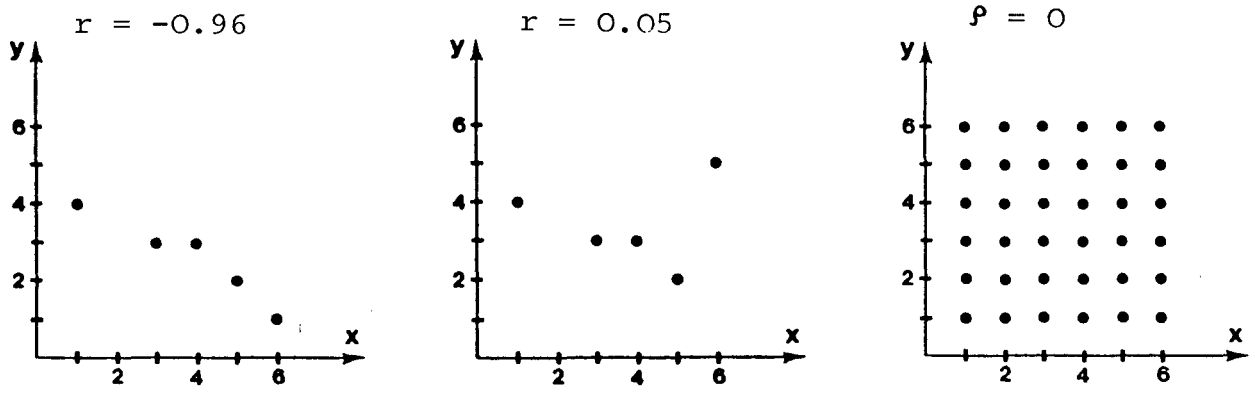

Fig. 1: Stichprobendaten und Wahrscheinlichkeitsverteilung zweier Augenzahlen X und Y.

Aus dem Beispiel wird deutlich, daß für den Wert von r erhebliche Schwankungen möglich sind, falls der Stichprobenumfang n=5 beträgt. Bei einem Korrelationskoeffizient von $\rho=0$ für die Zufallsvariablen X und Y kann man durch die kleine Stichprobe derart in die Irre geführt werden, daß man glaubt, daß eine hohe Korrelation vorliegt

(r=-0,96 bei den ersten Daten).

Ein Computerprogramm zur Erzeugung der Daten

Aus den vorhergehenden Überlegungen wird klar, daß der Stichprobenkorrelationskoeffizient r bei kleinen Stichproben erheblich schwanken *kann*: 0,05 bzw. -0,96. Man mag dagegen einwenden, daß es zwar zu so "unrepräsentativen" Stichproben mit r=-0,96 kommen kann, die völlig von $\rho=0$ abweichen, daß dies aber *selten* vorkomme. Um festzustellen, wie selten bzw. wie häufig das eigentlich eintritt, kann man etwa 100 Fünferserien von Würfen mit zwei Würfeln durchführen und von jeder dieser Fünferserien den Stichprobenkorrelationskoeffizienten bestimmen. Dann hat man 100 Daten für r und ersieht aus der Häufigkeitsverteilung dieser Daten, daß z.B. 10 % der erhaltenen Korrelationskoeffizienten größer als 0,8 bzw. kleiner als -0,8, also sehr wesentlich verschieden von $\rho=0$, sind. Das ist mühselig! Sowohl das Werfen der Würfel als auch das anschließende Berechnen des jeweiligen Stichprobenkorrelationskoeffizienten.

An dieser Stelle kann man mit Vorteil einen Personal Computer einsetzen. Das folgende Programm in BASICA (z.B. auf IBM PC) ist imstande, die nötigen Stichprobendaten zu simulieren (jeweils fünf Datenpaare) und anschließend den Wert von r zu berechnen. Typische Ergebnisse des Durchlaufens des Programms sind z.B.:
(3,1)(1,4)(3,4)(1,4)(5,6) 0,30 (3,1)(2,5)(6,3)(4,5)(3,3) -0,12 usw.

```
10 DIM X(20), Y(20) : RANDOMIZE : N=5
20 FOR S=1 TO 100
30   X=0: Y=0: XY=0: XX=0: YY=0
40   FOR I=1 TO N
50     X(I)=INT(6*RND)+1: Y(I)=INT(6*RND)+1
60     PRINT"(";X(I);",";Y(I);")";
70     X=X+X(I): Y=Y+Y(I): XY=XY+X(I)*Y(I)
75     XX=XX+X(I)*X(I): YY=YY+Y(I)*Y(I)
80   NEXT I
90   SSXY=XY-X*Y/N: SSXX=XX-X*X/N: SSYY=YY-Y*Y/N
95   PRINT SSXY/SQR(SSXX*SSYY);
99 NEXT S
```

Ergebnisse einer Simulationsstudie

Im folgenden werden die Ergebnisse einer solchen Simulationsstudie dargestellt. Es wurden 100 Datenpaare von Fünferstichproben, dann von Zehnerstichproben sowie von Zwanzigerstichproben erzeugt. Die Anzahl der für eine Stichprobe erzeuten Datenpaare steuert man im Programm in Zeile 10 durch den Befehl N=5 bzw. N=10 oder N=20. Es ergaben sich für r folgende, bereits geordnete Werte:

Für n=5:

-0,999,-0,98,-0,93,-0,91,-0,85) -0,80 ... 0,79(0,80, 0,90, 0,91, 0,91, 0,96

Für n=10:

-0,75, -0,70, -0,65, -0,60, -0,58) -0,54 ... 0,52(0,56, 0,59, 0,66, 0,72, 0,75

Für n=20:

-0,59, -0,47, -0,45, -0,43, -0,42) -0,41 ... 0,45(0,45, 0,49, 0,53, 0,55, 0,55

Aus den folgenden Boxplots für die Daten von r geht deutlich hervor:
Die Datenpunkte ("Würfelergebnisse") "breiten sich besser aus", falls der Stichprobenumfang größer wird, das hat kleinere Werte von |r| zur Folge.
Die Datenpunkte liegen weniger häufig längs einer Geraden, falls der Stichprobenumfang größer wird, was Werte von |r| nahe bei 1 zur Folge hätte.

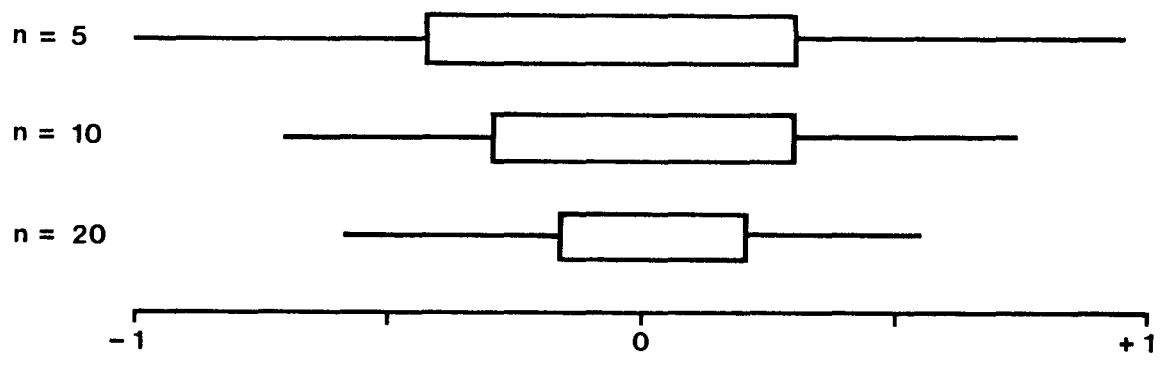

Fig. 2: Simulierte Verteilung von r, falls $\rho=0$.

Für die Praxis ist es nun wichtig, festzustellen, ob in einer gegebenen Stichprobe der Korrelationskoeffizient r erheblich vom Wert 0 abweicht, sodaß man die Hypothese $\rho=0$ ablehnen kann - ob also der Zusammenhang zwischen X und Y statistisch gesichert ist oder nicht. Dazu hat man folgendes Testproblem: Nullhypothese H_o: $\rho=0$ gegen Alternative H_1: $\rho\neq 0$, Signifikanzniveau α, z.B. $\alpha=0,10$.

Die kritischen Werte, bei deren Über- oder Unterschreiten man H_o ablehnen kann, bestimmt man aus der *Wahrscheinlichkeitsverteilung* des Stichprobenkorrelationskoeffizienten unter der Nullhypothese H_o. Diese zu bestimmen ist eine schwierige mathematische Aufgabe, wir verfügen jedoch über eine Näherung dieser Wahrscheinlichkeitsverteilung durch die Häufigkeitsverteilung der Werte von r aus den simulierten Daten. Man braucht nur jene Punkte zu bestimmen, die die 5 % kleinsten bzw. 5 % größten Werte von r von der übrigen Verteilung abtrennen, das ist dann der kritische Bereich für H_o, der Verwerfungsbereich von H_o. Denn: Werte nahe bei -1 sowie Werte nahe bei 1 sprechen am ehesten gegen H_o: $\rho=0$.

Man erhält für n=5:

Untere kritische Grenze: $\frac{-0,85 + (-0,80)}{2} = -0,825,$

obere kritische Grenze: $\frac{0,79 + 0,80}{2} = 0,795.$

Da aber ferner die Verteilung der Werte von r (theoretisch) symmetrisch ist, nimmt man als kritische Werte für r:

$$\pm \frac{0,825 + 0,795}{2} = \pm 0,81.$$

Die Nullhypothese H_o: $\rho=0$ ist demnach abzulehnen, falls in der konkreten Stichprobe vom Umfang n=5 für den Korrelationskoeffizienten $|r| \geq 0,81$ zutrifft ($\alpha=0,10$).

Es ist bemerkenswert, daß dieser kritische Wert 0,81 mit jenem gut übereinstimmt, den man erhält, falls X und Y normalverteilt sind. Im konkreten Beispiel hat X und Y ja eine diskrete Gleichverteilung. Die Entwicklung der kritischen Werte mit zunehmendem Stichprobenumfang ist in folgender Tabelle wiedergegeben:

Stichprobenumfang	n	5	10	20
Kritische Werte (Simulationsstudie)	α=0,10	0,81	0,55	0,43
Normalverteilungs-theorie	α=0,10	0,81	0,55	0,38

Die Variablen X und Y sind korreliert, z.B. $\rho=0,875$

Was passiert mit dem Stichprobenkorrelationskoeffizienten, falls die Zufallsvariablen X und Y korreliert sind? Auch dazu ein Beispiel: Nehmen wir von den Wurfergebnissen mit zwei Würfeln nur jene, wo sich die Augenzahlen um höchstens 1 unterscheiden, d.h. $|X-Y|\leq 1$, andere Wurfergebnisse ignorieren wir. Die Zufallsvariablen X und Y haben dann eine Gleichverteilung auf folgendem Gitter:

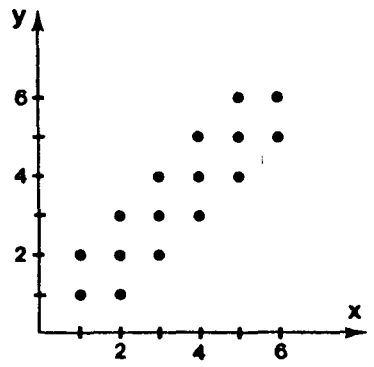

Fig.3: Wahrscheinlichkeitsverteilung von (X,Y), falls $|X-Y|\leq 1$, ρ=0,875

Eine formale Berechnung ergibt ρ=0,875, was deutlich widerspiegelt, daß nun die Variablen *gemeinsam* variieren: je höher X, desto höher im "Durchschnitt" Y. Die Wahrscheinlichkeitsverteilung gruppiert sich "eng" um eine Gerade.
Im Programm zur Erzeugung der Daten muß man nun eine neue Zeile einfügen:

```
55 IF ABS(X(I)-Y(I))>1 THEN 50
```

Diese Zeile garantiert, daß alle Datenpaare, die sich um mehr als 1 unterscheiden, ignoriert werden. Man geht zurück in die Schleife zu jenem Punkt, wo ein neues Datenpaar erzeugt wird. Ein typisches

Ergebnis wäre: (4,5) (3,2) (2,1) (5,6) (6,6) 0,94 (für r).
Eine analoge Simulationsstudie ergab, wie vorhin zusammengefaßt, folgende Daten für r:

Für n=5: 0,25, 0,29, 0,53, 0,56, 0,58)0,58...0,97(0,98, 0,98, 0,99, 0,999, 0,999

Für n=20: 0,72, 0,76, 0,76, 0,77, 0,78)0,80...0,92(0,92, 0,92, 0,93, 0,93, 0,94

Die folgenden Boxplots geben ein Bild dieser Verteilungen von r. Wiederum sieht man, daß der Wert von r bei einer kleinen Stichprobe (n=5) weit weg vom tatsächlichen Korrelationskoeffizienten ρ sein kann. Die Verteilung der Werte von r ist überdies sehr *schief*. Für größere Stichproben (n=20) ist es viel wahrscheinlicher, daß der Wert von r nahe beim richtigen Wert liegt: In ungefähr 90 % der Stichproben liegt r zwischen 0,80 und 0,92 verglichen mit dem tatsächlichen Wert ρ=0,875.

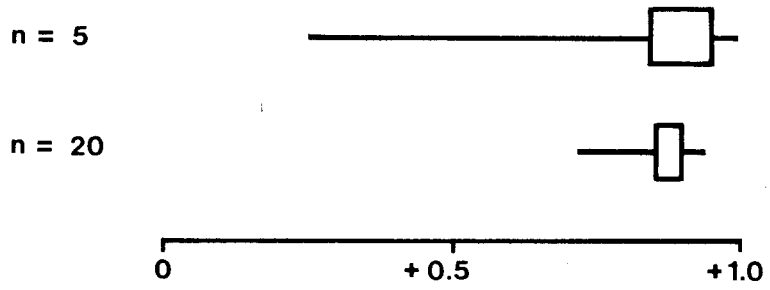

Fig.4: Simulierte Verteilung von r, falls ρ=0,875

Stichproben weisen eine unvermeidbare Fluktuation in den Daten auf. Daher schwanken Schätzungen aus Stichproben. Dies gilt auch für die Schätzung des Korrelationskoeffizienten.

Tatsächlicher Korrelationskoeffizient $\rho=\rho(X,Y)$	ρ=0	ρ=0,875	Umfang der Stichprobe
In 90 % der Stichproben schwankt der Korrelationskoeffizient innerhalb von	(-0,81, 0,81) (-0,43, 0,43)	(0,58, 0,98) (0,79, 0,92)	n=5 n=20

Es ist also ganz normal, daß man, speziell bei kleinen Stichproben zu voreiligen falschen Schlüssen etwa hinsichtlich hoher Korrelation kommen kann, obwohl tatsächlich *keine* Korrelation vorliegt. Aus der Simulationsstudie wird klar, was als normaler Schwankungsbereich anzusehen ist.