Störche, Geburten, Korrelationen, Kausalzusammenhänge

Hansruedi Künsch, Seminar für Statistik, ETH Zürich, kuensch@stat.math.ethz.ch

1 Einführung

Der empirische Korrelationskoeffizient, oder kurz die Korrelation, zweier Merkmale in einer Stichprobe bestehend aus n Wertepaaren $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ ist definiert als

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Dabei bezeichnen $\bar{x}$ und $\bar{y}$ die arithmetischen Mittelwerte der beiden Merkmale. Die Korrelation liegt immer im Intervall $[-1, 1]$, und je grösser der Absolutbetrag $|r_{xy}|$, desto näher liegen die Punkte (x_i, y_i) im Streudiagramm bei einer Geraden, deren Steigung das gleiche Vorzeichen hat wie r_{xy}. Für $n = 3$ ist das leicht einzusehen, denn r_{xy} ist dann einfach der Kosinus des Zwschenwinkels der beiden Vektoren mit Elementen $x_i - \bar{x}$ und $y_i - \bar{y}$. Die Korrelation misst also die Stärke und Richtung des linearen Zusammenhangs der beiden Merkmale.

Wenn im Unterricht die Korrelation besprochen wird, dann wird meist auch darauf hingewiesen, dass ein starker statistischer Zusammenhang, d.h. ein grosser Wert von $|r_{xy}|$, nur bedeutet, dass die beiden Merkmale assoziiert sind, und nicht, dass ein Kausalzusammenhang besteht im Sinne, dass eine Änderung des einen Merkmals direkt eine Auswirkung auf das andere Merkmal hat. Als Beispiel wird dabei oft erwähnt, dass man zwar eine starke Korrelation zwischen der Geburtenzahl und der Anzahl brütender Storchenpaare beobachtet, dass daraus aber nicht folgt, dass der Storch Babys bringt.

Dieses Beispiel wirkt im Unterricht viel überzeugender, wenn es mit echten Daten untermauert werden kann. Im Kontext von echten Daten kann man auch besser mögliche Gründe für die Assoziation diskutieren. Mir sind nur zwei solche Datensätze aus der Literatur bekannt: Die Korrelation der Anzahl Störche und der Einwohnerzahl von Oldenburg in den Jahren 1930-1936 im Streudiagramm auf S. 8 von [1] ist etwa 0.95, und [5] findet eine Korrelation von 0.62 zwischen der Einwohnerzahl und der Anzahl Brutpaare von Störchen in 17 europäischen Ländern im Jahr 1990. Ich zeige hier ein weiteres Beispiel mit Daten aus der Schweiz und gebe dann einen kurzen Einblick, wie der Kausalzusammenhang von Rauchen und Lungenkrebs nachgewiesen wurde.

2 Störche und Geburten

Abbildung 1 zeigt links das Streudiagramm der Geburten und Storchenbrutpaaren in der Schweiz von 2000 bis 2019. Die Daten der Storchenbrutpaare wurden mir freundlicherweise von der Vogelwarte Sempach zur Verfügung gestellt, die Anzahl der Lebendgeburten habe ich von der Webseite des Bundesamts für Statistik heruntergeladen. Der Zusammenhang ist genähert linear und die Korrelation beträgt 0.91, so dass diese Daten als Beispiel benutzt werden können. Ebenfalls eingezeichnet ist

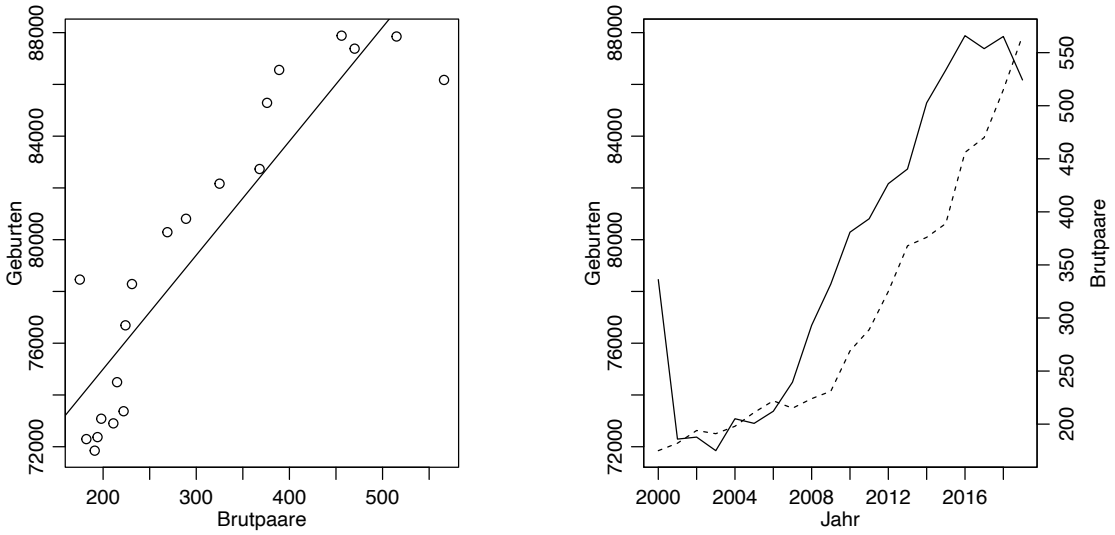

Abbildung 1 – Links: Lebendgeburten gegen die Anzahl Storchenbrutpaare in der Schweiz im Zeitraum 2000-2019 mit der Kleinste-Quadrate-Gerade. Rechts: Zeitreihe der Lebendgeburten (durchgezogen, Skala links) und der Storchenbrutpaare (gestrichelt, Skala rechts). Datenquellen: Vogelwarte Sempach und Bundesamt für Statistik.

die Regressionsgerade (Kleinste-Quadrate-Gerade). Sie hat die Steigung 44 und den Achsenabschnitt 66'200. Die Störche können also sicher nicht allein für den Nachwuchs beim Menschen sorgen.

Wie kann man diese Korrelation erklären? In der Abbildung 1 rechts ist die zeitliche Entwicklung der Geburten und der brütenden Stochenpaare dargestellt. Beide nehmen ungefähr linear zu, was sich natürlich auch im linearen Zusammenhang im Streudiagramm widerspiegelt. Der zeitliche Trend der beiden Merkmale, welcher zur hohen Korrelation führt, kann wie folgt erklärt werden: Es werden grosse Anstrengungen unternommen, um die Störche in der Schweiz wieder heimisch zu machen, während für die wachsenden Geburtenzahlen in erster Linie das Wachstum der Wohnbevölkerung durch Einwanderung verantwortlich ist. Die Geburtenrate, d.h. die Anzahl Geburten pro 1000 Einwohner, schwankt im ganzen Zeitraum um den Wert 10 und die Steigung der Kleinste-Quadrate-Geraden ist nur 0.013. Die Korrelation zwischen der Geburtenrate und der Anzahl Brutpaare beträgt noch 0.31. Das scheint zwar immer noch beachtlich, aber ein solcher Wert kann auch zufällig auftreten bei 20 Paaren von unabhängigen Merkmalen.

Dass die Korrelation von 0.91 zwischen der Anzahl Geburten und der Anzahl Brutpaare keine Bedeutung hat, wird völlig klar, wenn man nicht nur die Periode 2000-2019, sondern den ganzen Zeitraum von 1960-2019 betrachtet, in dem die Anzahl der Storchenbrutpaare lückenlos vorliegt. Abbildung 2 stellt diese Daten in der gleichen Form wie bei Abbildung 1 dar. Die Entwicklung der beiden Merkmale verläuft völlig unterschiedlich, und die Korrelation ist mit -0.30 sogar negativ. Die Anzahl Störche nimmt monoton zu und reflektiert wie gesagt die Massnahmen, um den Storch wieder heimisch zu machen. Bei den Geburten fällt in erster Linie der grosse Abfall in den sechziger Jahren auf, der mit dem Aufkommen der Anti-Baby-Pille zusammenhängt. Ferner sieht man einen genähert periodischen Effekt von 25-30 Jahren, der vermutlich einen Generationeneffekt darstellt.

Dieser Datensatz ergibt also nicht nur ein Beispiel für eine hohe Korrelation zwischen Brutpaaren von Störchen und Geburten, sondern auch ein warnendes Beispiel, wie man durch gezielte Auswahl eines Teils der Daten oft die gewünschten Resultate erhalten kann.

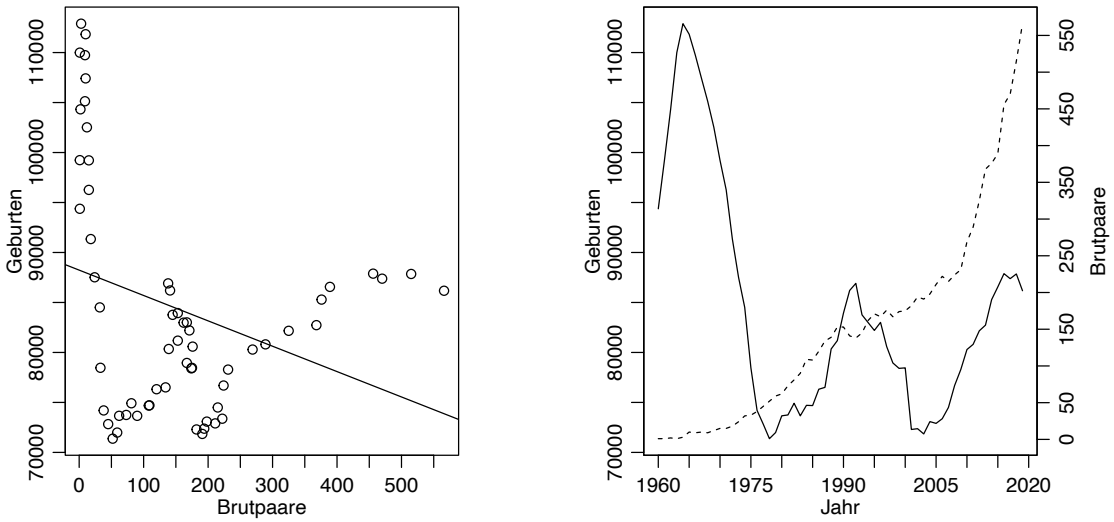

Abbildung 2 – Gleiche Darstellung wie in Abbildung 1, jedoch für den Zeitraum 1960-2019.

3 Von Assoziation zu einem Kausalzusammenhang

Die oben untersuchten Daten kommen von einer sogenannte Beobachtungsstudie: Für ein Land und eine Stichprobe von n Jahren beobachtet man die Anzahl Störche und die Anzahl Geburten. Beobachtungsstudien sind auch bei ernsthaften und wichtigen Fragestellungen weit verbreitet, obwohl es sehr schwierig ist, damit einen kausalen Effekt des einen Merkmals auf das andere nachzuweisen. Das bekannteste Beispiel ist wohl der Zusammenhang zwischen Rauchen und Lungenkrebs. Seit etwa 1900 hat der Tabakkonsum vor allem in der Form von Zigaretten in vielen Ländern stark zugenommen, und mit einer Verzögerung von ungefähr 20 Jahren auch die Todesrate von Lungenkrebs. Dies impliziert eine starke Korrelation dieser beiden Merkmale (siehe z.B. Fig. 2 in [2]), aber wie bei den Störchen und den Geburten könnte dieser gemeinsame Trend völlig unterschiedliche Ursachen haben. Der Zigarettenkonsum könnte z.B. wegen Werbung, Veränderung der Arbeitswelt oder Wachstum der Städte zugenommen haben, die Todesrate von Lungenkrebs wegen Luftverschmutzung, verbesserter Diagnosemöglichkeiten oder wegen des Rückgangs von anderen Krankheiten.

Informativer sind Studien, welche direkt das Auftreten von Lungenkrebs bei Rauchern und Nichtrauchern vergleichen. Zwei frühe solche Studien sind [2] und [3], die den Raucheranteil in zwei Patientengruppen von mehreren englischen Spitälern vergleichen: Die erste Gruppe bestand aus Patienten, die wegen Lungenkrebs behandelt wurden, die zweite aus gleich vielen Patienten mit anderen Krankheiten, jeweils aus dem gleichen Spital und mit der gleichen Altersverteilung. Ausserdem wurde der Tabakkonsum vor dem Auftreten der Krankheit erhoben, bzw. bei denen, die das Rauchen schon früher aufgegeben hatten, der Konsum im letzten Jahr davor. Die beobachteten Zahlen für die Männer aus [3] sind in Tabelle 1 gegeben, und in Abbildung 3 links ist der Anteil der Lungenkrebspatienten in der Studie in Abhängigkeit vom Tabakkonsum dargestellt.

Krankheit	Zigaretten pro Tag						Total
	0	1-4	5-14	15-24	25-49	50+	
Lungenkrebs	7	49	516	445	299	41	1’357
Andere	61	91	615	408	162	20	1’357
Total	68	140	1’131	853	461	61	2’714

Tabelle 1 – Daten von [3] für Männer. Tabak in anderer Form wurde in Zigaretten umgerechnet.

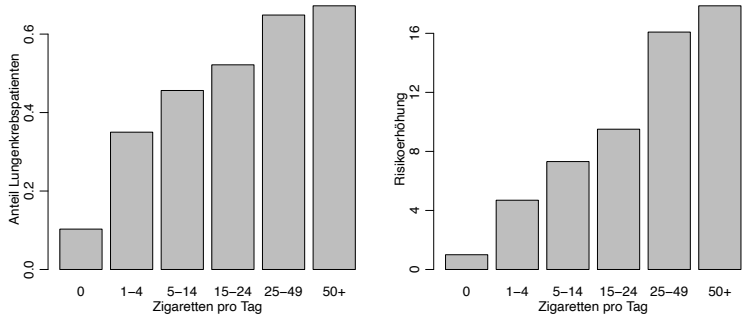

Abbildung 3 – Links: Relative Häufigkeit von Lungenkrebspatienten in der Studie [3] in Abhängigkeit vom Tabakkonsum. Rechts: Geschätzte Risikozunahme für Lungenkrebs bei Tabakkonsum; siehe Text.

Wenn kein Zusammenhang zwischen Rauchen und Lungkrebs bestehen würde, müsste der Anteil der Lungenkrebspatienten in jeder Kategorie um 50% schwanken. Offensichtlich nimmt aber der Anteil Lungenkrebspatienten mit wachsendem Tabakkonsum stark zu. Zur Berechnung der Korrelation von Tabakkonsum und Anteil Lungenkrebspatienten habe ich angenommen, dass der Konsum in jeder Kategorie gleich dem mittleren Wert dieser Kategorie, bzw. gleich 50 in der letzten Kategorie ist. Dann ergibt sich eine Korrelation von 0.88.

Trotzdem kann daraus noch nicht geschlossen werden, dass Rauchen das Auftreten von Lungenkrebs erhöht: Es könnte z.B. der Tabakkonsum bei den Lungenkrebspatienten zu hoch eingeschätzt worden sein, oder es könnten andere Faktoren wie Beruf, soziale Klasse oder Wohnort sowohl den Tabakkonsum als auch das Auftreten von Lungenkrebs begünstigen. Es ist beeindruckend, wie in [3] solche Unsicherheiten erwähnt und ihre Plausibilität diskutiert wird. In den Jahren nach dieser Publikation wurden dann auch noch genetische Faktoren in die Diskussion eingebracht, welche sowohl das Risiko von Lungenkrebs erhöhen als auch schneller zu Nikotinabhängigkeit führen könnten.

Die theoretisch sauberste Lösung, um solche anderen Faktoren auszuschliessen und einen kausalen Zusammenhang zwischen zwei Merkmalen nachzuweisen, ist ein randomisiertes Experiment. “Experiment” bedeutet, dass man nicht einfach zwei Merkmale in einem vorliegenden Datensatz erfasst und den Zusammenhang mit Hilfe der Korrelation oder einem anderen Mass quantifiziert, sondern die Werte desjenigen Merkmals, das als Ursache vermutet wird, für jedes Individuum in der Studie (oder – im Beispiel der Störche – für jedes Jahr der untersuchten Periode) festlegt und dann beobachtet, welche Werte des zweiten Merkmals herauskommen. “Randomisiert” bedeutet, dass der Wert des ersten Merkmals zufällig festgelegt wird. Im Beispiel von Rauchen und Lungenkrebs würde man also für jeden Studienteilnehmer durch das Los entscheiden, wieviele Zigaretten er täglich in den nächsten 5 oder 10 Jahren rauchen soll. Dies ist natürlich aus praktischen und ethischen Gründen nicht durchführbar. Es würde es aber erlauben, den Einwand mit den genetischen Faktoren auszuschliessen, denn wegen der Randomisierung wäre garantiert, dass die Höhe des Tabakkonsums unabhängig ist von solchen Faktoren.

Da ein randomisiertes Experiment nicht möglich war, basierte die Entscheidung, ob Rauchen eine Ursache von Lungenkrebs ist, auf einer Kombination von Argumenten. Erstens gab es plausible medizinische Erklärungen, wie Rauchen die Lunge schädigen kann. Zweitens gab es eine Vielzahl von Studien unter unterschiedlichen Bedingungen, die alle zweifelsfrei einen Zusammenhang nicht nur zwischen Rauchen und Lungenkrebs sondern auch zwischen Rauchen und anderen Krankheiten nachwiesen. Die Studien [2] und [3] sind retrospektiv (es wird der Tabakkonsum nach der Erkrankung erfasst), aber die gleichen Autoren haben anschliessend auch eine prospektive Studie durchgeführt, in der die Rauchgewohnheit von gesunden britischen Ärzten erfasst und danach während Jahren deren Gesundheitszustand verfolgt wurde. Drittens nahm im gleichen Zeitraum in vielen Ländern bei Männern der Zigarettenkonsum und die Todesrate von Lungenkrebs ab, während bei Frauen diese beiden Merkmale zunahmen. Daher kam eine Studie von vier medizinischen Gesellschaften und Bundesäm-

tern in den USA 1957 zu folgendem Schluss: “The sum total of scientific evidence establishes beyond reasonable doubt that cigarette smoking is a causative factor in the rapidly increasing incidence of human epidermoid carcinoma of the lung.”

Neben der Frage nach der Kausalität möchte man auch herausfinden, um wieviel sich die Wahrscheinlichkeit für Lungenkrebs bei einem bestimmten Tabakkonsum erhöht. Da bei [3] die Anzahl Patienten mit Lungenkrebs, bzw. anderen Krankheiten fixiert war, entsprechen die relativen Häufigkeiten von Abbildung 3 nicht den relativen Häufigkeiten von Lungenkrebs in der männlichen Population. Wenn wir annehmen, dass sich die Verteilung des Merkmals Tabakkonsum in der Population nicht von der Verteilung bei den Patienten mit anderen Krankheiten unterscheidet, lässt sich die Risikoerhöhung für Lungenkrebs durch Tabakkonsum aus den Daten von Tabelle 1 abschätzen. Um dies zu erklären, bezeichne ich die Ereignisse “Lungenkrebs”, bzw. “Kein Lungenkrebs” mit L, bzw. $\bar{L}$, und die Ereignisse Tabakkonsum in Kategorie j für $j = 0, 1, \ldots, 5$ mit Z_j. Die bedingten Wahrscheinlichkeiten $P(Z_j \mid L)$ und $P(Z_j \mid \bar{L})$ können mit den relativen Häufigkeiten von Tabelle 1 geschätzt werden, z.B.

$$P(Z_1 \mid L) \approx \frac{49}{1357} = 0.036, \quad P(Z_1 \mid \bar{L}) \approx \frac{91}{1357} = 0.067.$$

Wir interessieren uns jedoch für die bedingten Wahrscheinlichkeiten $P(L \mid Z_j)$, und bekanntlich sind diese nicht symmetrisch in den beiden Argumenten. Aus der Definition der bedingten Wahrscheinlichkeit folgt aber, dass gewisse Verhältnisse von bedingten Wahrscheinlichkeiten symmetrisch sind:

$$\frac{P(Z_j \mid L)}{P(Z_0 \mid L)} : \frac{P(Z_j \mid \bar{L})}{P(Z_0 \mid \bar{L})} = \frac{P(Z_j \cap L)P(Z_0 \cap \bar{L})}{P(Z_0 \cap L)P(Z_j \cap \bar{L})} = \frac{P(L \mid Z_j)}{P(\bar{L} \mid Z_j)} : \frac{P(L \mid Z_0)}{P(\bar{L} \mid Z_0)}.$$

Weil Lungenkrebs auch bei Rauchern eine seltene Krankheit ist, gilt in erster Näherung für jedes j $P(\bar{L} \mid Z_j) \approx 1$ und somit

$$P(L \mid Z_j) \approx \frac{P(Z_j \mid L)}{P(Z_0 \mid L)} : \frac{P(Z_j \mid \bar{L})}{P(Z_0 \mid \bar{L})} \cdot P(L \mid Z_0).$$

Mit den Zahlen von Tabelle 1 erhält man z.B. $P(L \mid Z_1) \approx \frac{49}{7} \cdot \frac{61}{91} \cdot P(L \mid Z_0) = 4.7 \cdot P(L \mid Z_0)$. Die Resultate für alle Kategorien sind in Abb. 3 rechts dargestellt.

Etwas ausführlichere Darstellungen der Arbeiten [2] und [3] und weitere Literaturhinweise findet man in Abschnitt 3 von [4] und auf den Seiten 115-121 von [6].

Links

[1] Box, George E. P, Hunter, J. Stuart, and Hunter, William G.: *Statistics for Experimenters: Design, Innovation, and Discovery, 2nd Edition* (2005), Wiley, New York.

[2] Doll, Richard and Hill, A. Bradford: *Smoking and Carcinoma of the Lung; Preliminary report*, British Medical Journal, 2 (1950), 739-748. `https://pubmed.ncbi.nlm.nih.gov/14772469/`

[3] Doll, Richard and Hill, A. Bradford: *A Study of the Aetiology of Carcinoma of the Lung*, British Medical Journal, 2 (1952), 1271-1786.`https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2022425/`

[4] Gail, Mitchell H.: *Statistics in Action*, J. Amer. Statist. Assoc. 91 (1996), 1-13.

[5] Matthews, Robert: *Storks Deliver Babies ($p = 0.008$)*, Teaching Statistics 22, 2 (2000), 36-38.

[6] Senn, Stephen: *Dicing with Death: Chance, Risk and Health* (2003), Cambridge University Press, Cambridge.