Manfred BOROVCNIK, Klagenfurt

# Statistische Zusammenhänge

*Hängt der Erfolg einer Therapie vom eingesetzten Medikament ab? Ist der Grad der Schädigung von Gefäßwänden aus Blutfettwerten und Gamma GT–Werten der Leber abzulesen? Beide Fragestellungen führen zum Test von Hypothesen. Ist in der ersten Fragestellung die Größe des Therapieerfolgs wenigstens dem medizinischen Experten einsichtig, so braucht man in der zweiten auch den so genannten Korrelationskoeffizienten als abstraktes Maß für die Zusammenhänge. Hier treffen sich Feinheiten des Testens mit dem Fehlen einer einfachen Deutung der Skala, auf der man die Zusammenhänge misst. Man erleichtert das Verständnis der Vorgangsweise und der damit erzielbaren Ergebnisse durch systematische Analyse von konkreten, auch schematischen Daten.*

## Einleitung

Einige Beispiele illustrieren, wie wesentlich es ist, zu verstehen, was man sich eigentlich unter der Aussage vorzustellen hat, dass ein Zusammenhang statistisch gesichert ist. Danach werden Möglichkeiten vorgestellt, die damit zusammenhängenden mathematischen Begriffe besser zu verstehen.

Unterstellt sei, dass Punktwolken den Zusammenhang zwischen zwei Merkmalen beschreiben und dass der Korrelationskoeffizient die Stärke eines solchen (linearen) Zusammenhangs erfasst, ohne dass man das Konzept Korrelationskoeffizient auch nur annähernd definiert. Dabei stellt man sich auf den Standpunkt, dass man durch den Gebrauch des Begriffes auch etwas über dessen Eigenschaften lernen kann.

Variiert man systematisch Punktwolken und betrachtet die Auswirkungen auf die Trendgerade, welche diese Punktwolke tendenziell beschreibt, sowie auf den Korrelationskoeffizienten, so sieht man rasch, dass Punkte, die relativ isoliert von den anderen liegen, einen übermäßigen Einfluss erhalten. Solche Studien kann man bei Borovcnik (2006) nachspielen.

Desgleichen kann man die Auswirkung der Änderung etwa eines Punktes in einer Punktwolke auf die Vorhersagen studieren: Interpretiert man die eine Variable $x$ als unabhängige, die andere – $y$ – als abhängige, so stellt sich das Problem, wie man eine Vorhersage der abhängigen Variablen $y$ durchführen und verbessern kann. Zwei einfache Ansätze zur Vorhersage sind: (i) Naive Vorhersage des Mittelwerts von $y$, was die Kenntnis der unabhängigen Variablen $x$ und allfällige Zusammenhänge zwischen $x$ und $y$ außer Acht lässt. (ii) Lineare Vorhersage von $y$ in Abhängigkeit vom Wert von $x$ und dem gewählten linearen Modell.

Auch hier sieht man durch Visualisieren der Fehler rasch, dass sich für die beiden Vorhersagen bei Korrelationskoeffizienten nahe bei 0 kaum Größenunterschiede in den Fehlern ergeben; dass aber die Größe der Fehler bei der linearen Vorhersage erheblich verringert wird, wenn der Korrelationskoeffizient in der Nähe von ±1 liegt. Rein durch systematisches Explorieren mit schematischen Daten erhält man tiefere Einblicke in das Wesen der Begriffe von Korrelation, Regression und Güte der Vorhersagen. Bleibt dann noch die Frage, wie man durch systematisches Probieren erhellen kann, was es bedeutet, dass ein Korrelationskoeffizient signifikant von Null verschieden ist. Alle Studien sind in EXCEL implementiert und von Borovcnik (2006) herunterladbar.

## Der Zusammenhang zwischen XX und YY ist statistisch gesichert

*Risikofaktoren für cardiovaskuläre Krankheiten (CVK)*

Der Zusammenhang von γ-Glutamyltransferase (GGT) zum Sterberisiko aus CVK wurde prospektiv untersucht bei über 160 000 Patienten; die Probanden wurden über 17 Jahre lang beobachtet. Als mathematisches Modell wurden Cox proportional hazards verwendet; die Daten wurden korrigiert nach schon bekannten Risikofaktoren. Für Männer und Frauen ergab sich: ein hoher GGT ist signifikant (P<0,001) assoziiert mit totaler Mortalität aus CVK: Dabei steigen die Hazard ratios um 1,66 (95% KI: 1,40-1,98) pro Einheit log GGT Zuwachs bei Männern …Hohe Werte von GGT sind positiv assoziiert mit

- tödlichen Ereignissen aus chronischen Koronarherzerkrankungen (P=0,009)
- Herzversagen (P<0,001)
- ischemischem und hämorrhagischen Schlaganfall (P<0,001)
- kein signifikanter Zusammenhang mit akutem Herzinfarkt (P=0,16)
- bei Frauen …
- bei Jüngeren: ein stärkerer Einfluss des GGT

Die Studie zeigt, dass GGT ein eigenständiger Faktor ist, unabhängig von bekannten Risikofaktoren, der mit cardiovaskulärer Mortalität verknüpft ist. Als unabhängige Variable spielt der Logarithmus von GGT mit, als abhängige Variable dient das proportionale Risiko – die sogenannten odds (Verhältnis $p : 1–p$) bzw. auch wieder der Logarithmus davon. Das wahre Ausmaß der Einflüsse verbirgt sich hinter abstrakten Merkmalen; man muss sich erst über Szenarien ein Bild davon machen. Wichtig aber ist, dass der Korrelationskoeffizient bzw. der Regressionskoeffizient signifikant von Null verschieden sind. Das bedeutet, die unabhängige Variable ist ein – statistisch gesicherter – Prädiktor für das Auftreten der Krankheit. Man kann die Messung von GGT etwa in allgemeine Gesundenuntersuchungsprogramme einbauen, oder bei Vorhandensein anderer Risiken

messen, damit man eine Handhabe hat, bei Vorliegen dieses Risikofaktors frühzeitig gezielte Gegenmaßnahmen zu empfehlen.

*Optimieren von chemischen Prozessen*

Sie untersuchen die Ausbeute einer Reaktion bei verschiedenen Temperaturen, um herauszufinden, ob es sich lohnt, die optimale Temperatur zu suchen – http://www.chemie.unibas.ch/%7Ehuber/Statistik/LinReg/LRBeispiele/beispiele.html#B4:

| | | | | | | |
|---|---|---|---|---|---|---|
| Temperatur T/K | 300 | 307 | 310 | 319 | 322 | 328 |
| Ausbeute A/% | 80 | 79 | 83 | 82 | 84 | 84 |

a. Ist der lineare Zusammenhang vom Sachzusammenhang her gesichert ? – Ein linearer Zusammenhang ist von der Sache her nicht gesichert. Wir kennen keinerlei Gesetz zu diesem Vorgang.

b. Lohnt es sich gemäß der Varianzanalyse bei einem Signifikanzniveau von 95 (99)% ? – Die Varianzanalyse ergibt F = 7.77

$$\frac{\text{Summe Quadrate } \textit{erklärte Variation}/\text{Freiheitsgrade}}{\text{Summe Quadrate } \textit{nicht-erklärte Variation}/\text{ Freiheitsgrade}}$$

Schwellenwert von F für 95 % Signifikanz = 7.71 <7.77. Folglich lohnt, die optimale Temperatur zu suchen sich [auf Niveau 95%]. Dieselbe Aussage erhält man auch so:

Achsenabschnitt = 30,70 Streuung = 18,46 Steigung = 0,163
Streuung = 0,059 R = 0,81

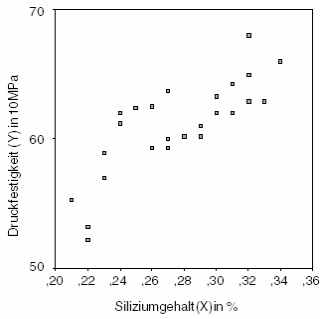
Wenn wir nun die untere Vertrauensgrenze der Steigung auf diesem Niveau berechnen erhalten wir 0,0006; der Wert ist also positiv wie die Steigung selbst. Steigung **0** – keine Abhängigkeit von Temperatur – liegt *nicht* dazwischen. Für 99%: Schwellenwert von F = 21.20: Mit 99% Sicherheit können wir nicht sagen: Ausbeute = temperaturabhängig.

*Verbesserung von Materialeigenschaften*

Gibt es einen linearen Zusammenhang zwischen dem Siliziumgehalt ($x$ %) und der Druckfestigkeit ($y$ 10MPa) einer bestimmten Stahlsorte? Dazu wurden beide Merkmale an n = 25 Stahlproben erfasst; siehe: www.elsevier.de/sixcms/media.php/795/Einfache%20Korrelationsanalyse.pdf

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 0,34 | 0,27 | 0,26 | 0,33 | 0,29 | 0,3 | 0,32 | 0,21 | 0,24 | 0,22 | 0,26 | 0,27 | 0,3 |
| $y_i$ | 66 | 59,3 | 59,3 | 62,9 | 60,2 | 63,3 | 62,9 | 55,3 | 61,2 | 53,2 | 62,5 | 60 | 60 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 0,29 | 0,3 | 0,31 | 0,32 | 0,31 | 0,27 | 0,32 | 0,22 | 0,23 | 0,23 | 0,24 | 0,25 |
| $y_i$ | 61 | 62 | 62 | 65 | 64,2 | 63,7 | 68 | 52,2 | 58,9 | 57 | 62 | 62,4 |



Der Korrelationskoeffizient r = 0, 796 ist signifikant verschieden von 0. Der Zusammenhang ist *statistisch gesichert*. Es *lohnt sich*, den Siliziumgehalt je nach wirtschaftlichen Gegebenheiten hoch anzusetzen.

## Zusammenhänge modellieren und absichern

Was bedeutet es, Zusammenhänge zu modellieren? Man sucht im Rahmen einer Systemanalyse nach jenen Variablen, die eine Zielgröße maßgeblich beeinflussen. Im Fall der Koronarerkrankungen also neuerdings auch GGT, bei der Prozessoptimierung nach der Temperatur, bei der Verbesserung von Druckfestigkeit bei Stahl etwa nach dem Siliziumgehalt. In aller Regel hat man mehrere Kandidaten für Einflussgrößen. Sodann werden in einer experimentellen Phase Daten „erzeugt“, welche die vermuteten Zusammenhänge prüfen lassen sollen. Man trennt dabei die Daten in Signal – so hängen die Daten für die abhängige Variable über ein mathematisches Modell von den Werten der unabhängigen Variablen ab plus Rauschen, das ist noch nicht erklärtes Variieren der Daten – die so genannten Residuen. Damit erhält man eine Strukturgleichung für das „Entstehen“ der Daten:


Hat man einmal Zusammenhänge in einem Modell erfasst, so kann man die Werte der unabhängigen Variablen als Prädiktoren für die Vorhersage der abhängigen Variablen nutzen. Die Genauigkeit der Vorhersagen kann man mittels Beschreibender Statistik untersuchen. Hier gelangt man direkt und einfach zu tieferen Zusammenhängen, wie das Bestimmtheitsmaß – das Quadrat des Korrelationskoeffizienten – benutzt werden kann, um die Verbesserung der Vorhersagen durch das benutzte Modell im Vergleich zur naiven Vorhersage (durch den Mittelwert der *y*-Daten) zu quantifizieren. Der Vorteil von EXCEL liegt auch darin, dass man durch Simulation von Daten (*x*, *y*) den Spielraum für den Korrelationskoeffizienten einschätzen kann, das ist dessen natürliche Fluktuation unter der Bedingung, dass *kein* Zusammenhang zwischen *x* und *y* besteht. Ist der festgestellte Korrelationskoeffizient außerhalb dieser durch die Studie ermittelten Schwellenwerte, so kann man sagen, der Zusammenhang ist *statistisch gesichert.*

## Literatur


Borovcnik, M.: EXCEL-Files für den Unterricht in Stochastik 2006: http://www.uni-klu.ac.at/stochastik.schule/ unter Links.

Christie, D.: Resampling mit Excel. Stochastik in der Schule 24 (2004), Heft 3, 22-27.

Neuwirth, E., Arganbright, D.: The Active Modeler: Mathematical Modeling with Microsoft Excel. Brooks/Cole 2004.

Reckelkamm, B.: Der Tanz der Residuen - Erarbeitung statistischer Grundbegriffe mit Hilfe von EXCEL. Stochastik in der Schule 24 (2004), Heft 3, 14-21.