

Statistische Methoden

Dr. C.J. Luchsinger

8 Lehren für's Management & das tägliche Leben

Wir präsentieren in diesem kleinen Kapitel kleine Resultate ohne Zusammenhang.

8.1 $\lambda \pm 2\sqrt{\lambda}$ revisited (prozentuale Angaben vs absolute Zahlen)

In der Vorlesung Angewandte Stochastik (Kapitel 9) sind wir auf obige Formel gestossen. Wie ging das damals? Wozu wird die Formel benutzt?

Die Problematik, dass wir für statistische Signifikanz die absoluten Zahlen und nicht relative (prozentuale) Zahlen oder relative Anteile brauchen, tritt in vielen Situationen der Statistik auf. Neben obiger Situation sind weitere bekannte Situationen etwa die χ^2 -Tests. In dem Zusammenhang noch eine Bitte: Wenn schon prozentuale Angaben über Änderungen (z.B. Wähleranteile von Parteien), dann bitte unterscheiden zwischen Prozenten und Prozentpunkten. Beispiel Vierfeldertafel:

8.2 Unnötigen Verlust von Information vermeiden - ein Beispiel

Angenommen, Sie messen von n Personen ein Merkmal (zum Beispiel Gewicht) vor und nach einer Behandlung. Damit haben Sie Daten x_1, \dots, x_n und y_1, \dots, y_n , welche wir mal als iid $\mathcal{N}(\mu_1, \sigma^2)$ und iid $\mathcal{N}(\mu_2, \sigma^2)$ modellieren wollen. Wir interessieren uns jetzt für die Frage, ob im Schnitt eine Gewichtsabnahme stattfindet. Wie können wir da vorgehen? Worauf ist dabei zu achten?

8.3 Randomisierung

Mit wenigen Änderungen einem Skript von Prof. Dr. A.D. Barbour und Dr. M. Roos entnommen:

Wie verteilt man zum Beispiel $2n$ Menschen in zwei gleich grosse Gruppen, um eine statistische Untersuchung zu machen? Die Antwort, welche erst auf die 1930-er Jahre und R.A. Fisher zurückgeht, ist klar: *streng nach dem Zufallsprinzip*. Nichts anderes gilt; es genügt nicht, die Menschen alternierend zu den Gruppen zuzuteilen, oder die ersten n Menschen zur ersten Gruppe zuzuteilen, oder eine Zuteilung, welche aus anderen Gründen bequem zu sein scheint, einfach hinzunehmen. Weshalb?

Die Begründung geht darauf zurück, wie wir uns den Begriff Zufallsfehler eigentlich vorstellen. In Tat und Wahrheit bestehen ‘Fehler’ hauptsächlich aus Unterschieden zwischen den Menschen, welche wir nicht imstande sind, wahrzunehmen. Sie sind keineswegs im mathematischen Sinne zufällig; nur können wir sie nicht besser beschreiben. Also müssen wir uns dagegen schützen, dass bei irgendeiner systematischen oder vorgegebenen Verteilungsregel, die Fehler ungleich zu den Gruppen zugeteilt werden - z.B. mehrheitlich positive Fehler zur Gruppe 1, mehrheitlich negative Fehler zur Gruppe 2 - sodass die Fehler einen erheblichen Beitrag zur Differenz $\bar{Y}_1 - \bar{Y}_2$ liefern. Die einzige Art und Weise, wie wir davon sicher sein können, ist durch eine Einteilung *streng nach dem Zufallsprinzip*. Dadurch ist eine ungleiche Verteilung der Fehler zu den Gruppen zwar nicht völlig ausgeschlossen; jedoch erhalten wir über diese sogenannte *Randomisierung* ein genau definiertes Zufallsmodell, nach welchem wir Signifikanz und Macht ehrlich berechnen können, indem wir die Wahrscheinlichkeiten jeder ungleichen Verteilung der Fehler genau wissen. Die Fehler, wie immer sie aussehen möchten, werden zu den Gruppen rein zufällig verteilt; jede von den möglichen $\binom{2n}{n}$ Zuteilungen ist gleich wahrscheinlich. Also beträgt zum Beispiel die Wahrscheinlichkeit, dass die grössten n Fehler zur ersten Gruppe und die restlichen Fehler zur zweiten Gruppe verteilt werden, genau $1/\binom{2n}{n}$, und ähnliche aber kompliziertere kombinatorische Rechnungen ergeben die Wahrscheinlichkeit von anderen ungleichen Verteilungen der Fehler.

Sobald man Beobachtungen ausserhalb vom Labor erhebt, hat man schwierige Probleme, solche Randomisierungen zu gestalten. Man möchte z.B. untersuchen, ob Kinder von Rauchern mehr oder weniger unter Erkältungen leiden. Wie geht man vor, um Daten zu dieser Frage zu bekommen? Meist versucht man Elternpaare zu finden, welche bereit sind, über ein ganzes Jahr täglich zu notieren, wieviele Ihrer Kinder erkältet waren; vielleicht über Werbung, und mit einer angemessenen Belohnung. Einige Elternpaare sind schon Raucher, die anderen Nichtraucher; also kann man die ‘Behandlung’ Raucher bzw. Nichtraucher einfach nicht nach dem Zufallsprinzip zu den Elternpaaren zuteilen, und die schöne statistische Sicherheit, welche uns die Randomisierung liefert, ist sofort weg. Strenge Aussagen über Signifikanz sind danach meistens unmöglich. Es ist einfach sich vorzustellen, dass diejenigen Elternpaare, welche sich die Mühe geben, sich für eine solche Studie zu melden, nicht unvoreingenommen in Bezug auf den Inhalt von der Studie sein könnten - Raucher haben z.B. eventuell unbewusst die Tendenz, die Folgen vom Rauchen herunterzuspielen, und daher die Tendenz, den Zustand ‘erkältet’ weniger häufig als Nichtraucher zu merken. In solchen Fällen können die ϵ_i ’s gut über die Gruppen ungleich verteilt sein; mehrheitlich positiv in einer Gruppe, mehrheitlich negativ in der anderen wäre gut vorstellbar. Also muss man alle ‘statistischen’ Schlüsse, welche aus solchen Studien stammen, mit grosser Vorsicht betrachten.

Ähnlich sieht es aus, wenn man Tiere in der Natur beobachtet. Vielleicht will man wissen, ob die Populationsdichte einer bestimmten Art Schnecke grösser ist in der Westschweiz als in der Ostschweiz. Da untersucht man eine gewisse Fläche in der Westschweiz nach Schnecken, macht dasselbe in der Ostschweiz. Natürlich ist es unmöglich, die Behandlung ‘Ost- oder Westschweiz’ zu den Standorten zufällig zuzuteilen, sodass keine Randomisierung vorliegt. Nur sind die Bedingungen an den zwei Standorten nicht ganz gleich; wegen des Bodentyps, wegen der Exponierung, wegen des Wetters usw., sodass die gemessenen Populationsdichten auch ‘Fehler’ beinhalten, welche auf die unterschiedlichen Bedingungen (und nicht auf die Populationsdichte) zurückzuführen sind. Also darf man die Fehler keineswegs als gleichverteilt zwischen den Standorten annehmen. Auch mit mehreren Standorten in den beiden Gebieten hat man dasselbe Problem. Man kann versuchen, die Bedingungen

soweit möglich vergleichbar zu gestalten; nur kann es gut Unterschiede geben, welche aus der Sicht der Schnecke gesehen wichtig sind, von uns jedoch nicht wahrgenommen werden. Wiederum muss man 'statistische' Schlüsse mit grosser Vorsicht betrachten.

8.4 Eine politische Stellungnahme zu Statistik, Geheimhaltung und Political Correctness

Es gibt immer wieder Fälle, wo seriöse, politisch unbelastete Wissenschaftler massiv unter Beschuss kommen, wenn Sie Untersuchungen anstellen zu Tabu-Themen und Fragen wie

- * Intelligenz und Geschlecht/Rasse/Herkunft/Religion etc.
- * Kriminalitätsrate und Geschlecht/Rasse/Herkunft/Religion etc.
- * weitere Beispiele

Auch schon vorgekommen ist, dass eine Behörde einen Bericht nicht publizieren liess, weil die Resultate den Politiker/innen nicht genehm waren.

Hierzu sind ein paar Bemerkungen notwendig; man muss da kurz ausholen:

Seit etwa 1970 werden alle möglichen Untersuchungen gemacht und "statistisch" ausgewertet (vgl dazu auch Kapitel 11: 2 Artikel zur Qualität der Statistik in der angewandten Forschung). Der Grund liegt vor allem darin, dass in den Sozialwissenschaften (inkl Oekonomie) Seminar-/Diplom- und Doktorarbeiten in derart grossen Mengen verfasst werden müssen, dass sonst der Stoff ausginge. Über die Qualität und vor allem Relevanz solcher Arbeiten sprechen wir besser nicht. Man muss sich diesbezüglich vor allem fragen, ob dieser Missstand auch wäre, wenn studieren nicht (fast) kostenlos wäre und an höhere Anforderungen geknüpft würde. Wegen der oben erwähnten Masse von Untersuchungen kommt es dann sehr schnell vor, dass wegen des α 's massenhaft Fehler erster Art begangen werden.

Wenn aber tatsächlich wichtige Fragen untersucht werden wie "Ausländerkriminalität", so muss dies durch gute Statistiker/innen und ohne weitergehenden politischen Auftrag geschehen. Um bei diesem Beispiel zu bleiben: Wenn nicht seriöse Statistiker/innen diese Fragen untersuchen, werden es andere mit tendenziösem politischem Auftrag machen. Man wird in obigem Beispiel bei seriöser Datenanalyse übrigens unschwer feststellen, dass ein sehr grosser Teil der (erhöhten) sogenannten "Ausländerkriminalität" durch die

demographische Zusammensetzung der ausländischen Bevölkerung erklärt werden kann: Wenn es viele junge Personen in einer Bevölkerungsgruppe hat (15-40 Jahre), dann hat es automatisch in dieser Bevölkerungsgruppe mehr Kriminalität als bei den anderen - aber eben: weil es viele Junge hat und nicht (zumindest nicht in erster Linie) weil es Ausländer sind.

Dass eine Behörde - das Volk ist nicht nur in der Schweiz der Souverän! - einen Bericht nicht publiziert ist nicht akzeptabel. Die einzigen Ausnahmen sind Fragen der Geheimhaltung bei Polizei und Verteidigung sowie Datenschutz oder bei laufenden Verhandlungen des Staates mit anderen Staaten.