Mehrfaches Testen[1]

QIUJIE ZHENG UND YONGGANG LU, ANCHORAGE, ALASKA

[1] Original: Do you catch undersized fish? Let's Go fishing to learn some important concepts in multiple testing. In *Teaching Statistics* 38 (2016) 3, 91–97. Kürzung, Bearbeitung und Übersetzung: JÖRG MEYER

Zusammenfassung: *Im Zeitalter von Big Data, wo die Datensammlung und -speicherung immer billiger wird, können eine Reihe von statistischen Tests alle gleichzeitig ausgeführt werden, um neue spannende Erkenntnisse zu bekommen. In diesem Artikel wird ein Zugang vorgestellt, um in das mehrfache Testen einzuführen, und zwar am Beispiel des Fischens mit einem Netz. Der Zugang macht Freude und erfordert keine mathematischen Berechnungen.*

1 Einleitung

Das Testen von Hypothesen gehört zur schließenden Statistik. Um Lernende mit Hypothesentests vertraut zu machen, wurden eine Reihe von innovativen Zugängen entwickelt, in letzter Zeit etwa Dambolena et al. (2009), Holland (2007) und Nordmoe (2004).

In diesem Artikel geht es um das Problem des mehrfachen Testens. Die Nullhypothese sei „Münze ist fair", die Alternativhypothese sei „Münze ist nicht fair". Aber anstatt den Test mit nur einer einzigen Münze auszuführen, sollen 100.000 Leute den Test ausführen, und alle Durchführungen sollen unabhängig voneinander sein. Jede Person bekommt eine faire Münze (zusammen müssen wir also 100.000 garantiert faire Münzen haben). Jede Person soll seine Münze 10-mal werfen und notieren, wie oft „Kopf" kam.

Wenn wir das Ablehnungskriterium auf $\alpha = 0{,}05$ festlegen, sind 5 % der Testergebnisse statistisch signifikant und lassen uns die Nullhypothese zurückweisen, obwohl wir wissen, dass sie bei allen 100.000 Münzen zutrifft. Das 5 %-Signifikanzniveau erscheint uns vernünftig für jeden individuellen Test. Wenn wir jedoch dies Kriterium auf alle 100.000 Tests anwenden, bekommen wir 5.000 falsche Zurückweisungen der (wahren) Nullhypothese. Insofern kann man nicht von einem seltenen Ereignis reden.

Zwar ist diese große Zahl falscher Zurückweisungen beim Münzwurf nicht unbedingt eine große Sache, eine solch große Zahl kann aber kaum in der wissenschaftlichen Forschung toleriert werden, in der statistische signifikante Ergebnisse zu wichtigen Erkenntnissen führen können. So ist es zum Beispiel für die US Food and Drug Administration und das Gesundheitsministerium von Großbritannien völlig inakzeptabel, mit 5.000 falschen Schlüssen bei 100.000 klinischen Tests von neu entwickelten Arzneien rechnen zu müssen.

Aufgrund der mittlerweile sehr billigen Möglichkeit, an sehr, sehr viele Daten schnell heranzukommen, können Lernende viele statistische Tests zusammen ausführen lassen. So kann man etwa in sozialen Netzwerken viele Leute auffordern, an dem oben beschriebenen Münzwurfexperiment teilzunehmen und die Ergebnisse mitzuteilen. Kennt man die oben beschriebene große Zahl falscher Zurückweisungen, wird man daraus dann keine falschen Schlüsse mehr ziehen. Anderson (2008) hat geschrieben, dass Big Data wissenschaftliche Methoden überflüssig mache und dass diese durch ein massive Suchen nach signifikanten Korrelationen ersetzt werden können, was keinen großen Aufwand erfordere.

Da andererseits die mathematische Komplexität beim mehrfachen Testen größer ist als beim einfachen Hypothesentest, muss man sich effektive Möglichkeiten überlegen, um Lernende an die Logik und an die Begriffe des mehrfachen Testens heranzuführen. White (2015) hat dazu zwei Aktivitäten mit Excel beschrieben. Andererseits ist es allgemein anerkannt, dass lebensnahe Beispiele Lernende besser dazu bringen können, abstrakte statistische Konzepte zu verstehen. Gelman/Nolan (2002) liefern mehrere solcher Beispiele.

In diesem Artikel beschreiben wir eine analoge Vorgehensweise, um wichtige Konzepte des mehrfachen Testens anhand eines lebensnahen Beispiels zu vermitteln. Mathematische Berechnungen sind dazu nicht erforderlich.

2 Wichtige Konzepte beim mehrfachen Testen

Wir stellen kurz einige wichtige Konzepte vor, die die Lernenden bei unserem Zugang kennenlernen.

2.1 Gleichzeitige statistische Inferenz

In einem Hypothesentest zieht man einen Inferenz-Schluss, indem man die Nullhypothese ablehnt oder nicht ablehnt. In dem oben beschriebenen Fall werden 100.000 solcher Inferenz-Schlüsse gezogen. Beim mehrfachen Testen werden alle Einzeltests als gleichzeitig behandelt.

2.2 Mehrfache Fehler erster Art

Bei einem Hypothesentest begeht man einen Fehler erster Art, wenn man die wahre Nullhypothese fälschlicherweise zurückweist. Beim 100.000-fachen Testen kann man beim Signifikanzniveau von 5 % erwarten, dass 5.000-mal ein solcher Fehler eintritt unter der Voraussetzung, dass für jeden Einzeltest die Nullhypothese zutrifft. Wie oben beschrieben, ist das in der wissenschaftlichen Forschung inakzeptabel. Daher sollte man beim mehrfachen Testen den Fehler 1. Art so definieren, dass man eine oder mehr fälschliche Zurückweisungen macht unter der Voraussetzung, dass für jeden Einzeltest die Nullhypothese zutrifft. Wenn für den Individualtest das Signifikanzniveau nicht höher als $\alpha = 0{,}05$ sein soll, sollte bei n voneinander unabhängigen Tests die Fehlerrate nicht höher als

$$\alpha_m = 1 - (1 - \alpha)^n = 0{,}05 \qquad (1)$$

sein.

2.3 Festlegung der Grenze beim mehrfachen Fehler 1. Art

Aus (1) folgt, dass bei $\alpha = 0{,}05$ und $n = 10$ der Wert $\alpha_m \approx 0{,}4$ beträgt. Daher muss α viel kleiner gewählt werden. Die einfachste Festlegung heißt *Bonferroni-Korrektur* und definiert den angepassten α-Wert als

$$\alpha_{\text{korr}} = \alpha / n\,. \qquad (2)$$

Bei 10 voneinander unabhängigen Test ist dann

$$\alpha_m = 1 - \left(1 - \frac{0{,}05}{10}\right)^{10} \approx 0{,}0408$$

(man beachte die Erörterungen in White (2015)).

Als nächstes wird der Zugang geschildert, der ohne mathematische Berechnungen auskommt.

3 Warum kann man Fische in Untergröße fangen?

Fischen mit Netzen ist sehr effektiv und deshalb sehr weit verbreitet sowohl im kommerziellen als auch im privaten Bereich. Da die Beute i. a. groß ist, wird die Netzfischerei von Behörden vieler Länder stark kontrolliert. So wird die maximale Netz- und die minimale Maschengröße häufig vorgeschrieben.

Einer der Gründe für diese Regulierungen ist es, das Abfischen von Fischen in Untergröße zu verhindern. Fische in Untergröße sind i. a. junge Fische.

Eine naheliegende Frage ist nun: Wie gut greifen solche Restriktionen?

Wenn man nur einen einzigen Fisch mit einem Netz fangen möchte, so wäre das Problem einfach: Wenn der Fisch kleiner ist als die Maschengröße, könnte er mit einer Wahrscheinlichkeit von vielleicht 95 % durch das Netz hindurchschwimmen. Ein zugehöriges Video auf YouTube ist unter https://www.youtube.com/watch?v=c_G_W2-wvfI abrufbar; Abb. 1 zeigt ein Screenshot.

Abb. 1: Versuch, nur einen Fisch zu fangen

Aber wenn man zwei Fische mit Untergröße fangen will, wie groß ist dann die Wahrscheinlichkeit, dass einem beide entkommen? Wie steht es mit drei, vier, fünf Fischen? Durch solche Fragen wird eine „indirekte“ Behandlung von Gleichung (1) induziert. Nun kann man das 2. Video zeigen (Abb. 2 zeigt das Screenshot).

Abb. 2: Versuch, mehrere Fische zu fangen

Bei zunehmender Anzahl der Fische steigt die Wahrscheinlichkeit, einen der Fische in Untergröße zu fangen.

Eine andere Strategie, um das Fangen von Fischen in Untergröße zu vermeiden, besteht darin, dass man immer nur einen Fisch fangen will, aber dies mehrere Male nacheinander. Was ist der Unterschied zwischen den Ergebnissen beider Strategien? Ein einfaches Beispiel macht das deutlich. Zwei Fischer wollen jeweils 100 Fische fischen. Ein Fischer will nur einen Fisch pro Netzwurf und wirft sein Netz 100-mal aus. Der andere Fischer lässt sein großes Netz so lange im Wasser, bis er 100 Fische hat. Natürlich ist die 2. Me-

thode sinnvoller. Aber vom Standpunkt, junge Fische schützen zu wollen, ist die Wahrscheinlichkeit beim 2. Fischer, viele Fische in Untergröße zu haben, viel größer als beim 1. Fischer. Die Abb. 3a und 3b zeigen den Unterschied zwischen beiden Methoden.

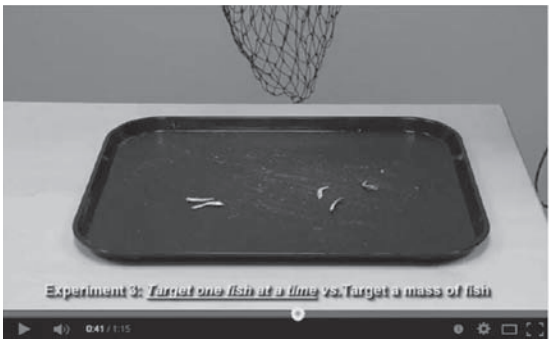

Abb. 3a: Ausbeute kleiner Fische beim 1. Fischer

Abb. 3b: Ausbeute kleiner Fische beim 2. Fischer

Dieses einfache Beispiel zeigt weiter, dass nicht die Maschen-, sondern auch die Netzgröße entscheidend ist. Wenn das Netz des 2. Fischers sehr klein ist, wird seine Strategie sich im Ergebnis nicht von der des 1. Fischers unterscheiden (Abb. 4).

Abb. 4: Der 2. Fischer hat nur ein kleines Netz

Ist andererseits das Netz des 2. Fischers so groß und so breit wie der Fluss, aus dem die Fische geholt werden sollen, wird seine Strategie den Fischbestand sehr schnell auch dann sehr dezimieren, wenn die Maschengröße den Anforderungen entspricht.

4 Wie kann man das Fischen von Fischen in Untergröße vermeiden?

Erstens muss man sowohl die Maschen- als auch die Netzgröße kontrollieren. Zweitens muss man beide Größen zusammen kontrollieren.

Es erscheint sinnvoll, Maschengröße M und Netzgröße N aufeinander zu beziehen, etwa durch

$$M_{\text{angepasst}} = M/\lambda_N. \quad (3)$$

Dabei ist $M_{\text{angepasst}}$ die angepasste Maschengröße, M die auf die maximal zu erhaltende Menge von Fisch bezogene Maschengröße und γ_N ein von der Netzgröße invers abhängiger Koeffizient (große Netze führen zu einem kleinen Wert von γ_N).

Entspricht ein Netz in seiner Größe der maximal zu erhaltenden Menge an Fisch, ist $\lambda_N = 1$. Ist das Netz 100-mal so groß, wird man etwa $\lambda_N = 0{,}8$ setzen.

Man muss (3) nicht anwenden, es reicht, wenn Lernende einsehen, dass Netzgröße und Maschengröße aufeinander bezogen sein sollten. Dann werden sie leichter die Bonferroni-Korrektur einsehen und verstehen, dass M und $M_{\text{angepasst}}$ analog sind zu α und α_{korr} und dass λ_N analog ist zu n in (2).

5 Analogien in der Statistik

Die inhaltlichen Analogien sind:

1. Wenn man ein Netz benutzt, um Fische zu fangen, so vollführt man eine simultane statistische Inferenz in Bezug auf viele statistische Tests.
2. Wenn man einen Fisch in Untergröße fängt, begeht man einen Fehler 1. Art und weist fälschlicherweise eine wahre Nullhypothese zurück.
3. Wenn die Maschengröße eingeschränkt wird, um Fische in Untergröße davor zu bewahren, gefischt zu werden, hat man ein Signifikanzniveau definiert, um die Wahrscheinlichkeit für einen Fehler 1. Art zu kontrollieren.
4. Wenn man gar keinen Fisch mit Untergröße angeln will, will man den Fehler 1. Art für Mehrfachtests vermeiden.
5. Wenn Maschen- und Netzgröße nicht aufeinander bezogen sind, wird der Fehler 1. Art für Mehrfachtests wahrscheinlicher, weil das Signifikanzniveau nicht mit der Anzahl der voneinander unabhängigen Tests verbunden ist.
6. Gleichung (3) bereitet die Idee der Bonferroni-Korrektur vor.

Literatur

Anderson, C. (2008): The end of theory: the data deluge makes the scientific method obsolete, In: *Wired Magazine*, 16.07. http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory/

Dambolena, I. G.; Eriksen, S. E.; Kopcso, D. P. (2009): An intuitive introduction to hypothesis testing. In: *INFORMS Transactions on Education*, 9, 53–62. DOI:10.1287/ited.1080.0019.

Gelman, A.; Nolan, D. (2002); Teaching statistics: A bag of tricks, Oxford University Press.

Holland, B. K. (2007); A classroom demonstration of hypothesis testing. In: *Teaching Statistics*, 29, 71–73. DOI:10.1111/j.1467-9639.2007.00269.x. Abdruck einer gekürzten Übersetzung in diesem Heft.

Nordmoe, E. D. (2004): Of Poohsticks and p-values: hypothesis testing in the hundred acre wood. In: *Teaching Statistics*, 26, 56–58. DOI:10.1111/j.1467-9639.2004.00163.x.

White, D. (2015): Active learning and threshold concepts in multiple testing that can further develop student critical statistical thinking. In: *Teaching Statistics*, 37, 48–53. DOI:10.1111/test.12069.

Anschrift der Verfasser

Qiujie Zheng
Department of Economics and Public Policy,
University of Alaska Anchorage,
Anchorage, AK, USA
qzheng3@uaa.alaska.edu

Yonggang Lu
Department of Information System and Decision Sciences,
University of Alaska Anchorage,
Anchorage, AK, USA
ylu4@uaa.alaska.edu