Was ist eigentlich ein *p*-Wert?[1]

PATRICIA HUMPHREY, STATESBORO, GA, USA

[1] Original ‚What is a *p*-value?' in *Teaching Statistics* 34 (2012) 1, 18–20.
Übersetzung und Bearbeitung:
MANFRED BOROVCNIK, KLAGENFURT

Zusammenfassung: *Es wird eine Klassenaktivität beschrieben, die nicht nur zum Ziel hat, Hypothesentests zu motivieren, sondern auch den p-Wert und die Macht von statistischen Tests verständlicher zu machen. Schlagworte: Unterricht, Hypothesentesten, Simulation und Macht.*

1 Einleitung

Der *p*-Wert beim statistischen Testen wird von vielen Studierenden falsch verstanden. Sie glauben oft, dass es sich dabei um die Wahrscheinlichkeit handelt, dass die Nullhypothese wahr ist und die Gegenhypothese falsch oder sie verdrehen andere Dinge. „Die Wahrscheinlichkeit, berechnet unter der Annahme, dass H_0 wahr ist, dass die Teststatistik einen Wert annehmen würde, der ebenso extrem oder noch extremer ist als derjenige, der tatsächlich beobachtet wurde", so definieren Moore & McCabe (2003). Die Studierenden haben eine [einzige] Stichprobe mit einem konkreten Mittelwert oder einem Anteil vor sich. Warum dann der Fokus auf die Teststatistik [als Zufallsvariable] und warum der „oder noch extremer"-Teil der Aussage?

Es ist Jahre her, dass ich bei Einführung ins Hypothesentesten versuchte, mich an der Idee „wie viel Information (Evidenz) gebraucht wird, um die Nullhypothese zu verwerfen?" zu orientieren und dazu einen Stapel von Spielkarten mitbrachte. Ich versprach jedem, der eine schwarze Karte vom Stapel zieht, die Hausübung zu erlassen. Wie zu erwarten, vermuteten die Studierenden sofort, dass die Karten manipuliert sind – manche vermuteten sogar, dass überhaupt keine schwarzen Karten im Stapel seien.

Ich dachte hin und her und entschied mich, die Frage etwas anders zu stellen. Später modifizierte ich den Ablauf und bezog die Simulation der erwarteten Verteilung der Kartenfarben unter der Nullhypothese mit ein, damit die Studierenden ein Gefühl für den Zufall bekommen, was passieren kann, wenn das Kartenspiel fair ist. Wenn es die Zeit erlaubt (oder in einer späteren Einheit), kehren wir zum Problem zurück und simulieren die Macht des Tests für unseren besonders zusammengesetzten Stapel von Karten.

2 Das Experiment in der Klasse

Man ändere die Zusammensetzung eines Kartenstapels (oder noch besser, eine andere Person macht das), indem man zwei vollständige Stapel mischt und teilt. Danach präsentiert man den so präparierten Stapel der Klasse und gibt die zusätzliche Information, dass nun die roten und schwarzen Karten verändert sein könnten, sodass die Farben nicht mehr im gleichen Anteil repräsentiert sind. Dies führt in natürlicher Weise auf die Hypothesen

$$H_0: p = 0{,}5 \text{ gegen } H_A: p \neq 0{,}5;$$

dabei betrachten wir p als Anteil an roten Karten.

Weil Hypothesentests immer unter der Annahme, dass die Nullhypothese wahr ist, durchgeführt werden, bestimmen wir die Stichprobenverteilung des Stichprobenanteils $\hat{p}$ unter der „Null". Diese hängt auch von der Anzahl der Studierenden in der Klasse ab; sind 25 da und jeder zieht (mit Zurücklegen und neu Mischen des Stapels) eine Karte, so wird die Anzahl der Erfolge (rote Karten) einer Binomialverteilung mit $n = 25$ und $p = 0{,}5$ folgen. Wir können diese Verteilung durch eine Normalverteilung mit $\mu = 12{,}5$ und $\sigma = \sqrt{25 \cdot 0{,}5 \cdot 0{,}5} = 2{,}5$ approximieren (siehe Abb. 1). Der Anteil $\hat{p}$ wird daher approximativ $N(\mu = 0{,}5, \sigma = 0{,}1)$-verteilt sein.

Wenn wir nun den Anteil $\hat{p}$ in der Stichprobe vorhersagen wollen und dabei die Bedingungen der Nullhypothese zutreffen (gleich viele schwarze wie rote Karten), so wenden wir die σ-Regeln (68-95-99,7) an [bei der Normalverteilung liegt im Bereich von 1σ-Einheit um den Erwartungswert μ ca. 0,68 Wahrscheinlichkeit, das erhöht sich auf ca. 0,95 bzw. 0,997 für den zentralen 2- bzw. 3-σ-Bereich]; wir können daher sagen, dass ungefähr 95 % aller Stichproben einen Anteil für die roten Karten zwischen 30 und 70 % haben werden.

Hochgerechnet auf die 25 Studierenden werden in 95 % der Stichproben (der wiederholten Experimente) zwischen 7,5 und 17,5 (also von 8 bis 17) rote Karten in der Klasse gezogen werden. Zieht man weniger als 8 oder mehr als 17, so deutet das an, dass der Stapel tatsächlich verändert worden ist.

Die Ergebnisse der einzelnen Ziehungen werden an der Tafel notiert; die gezogene Karte wird zurückgelegt und der Stapel erneut gemischt. Es ergeben sich

bald Lieblingskarten und eine sportliche Sicht darauf, wer nun „gewinnt“.

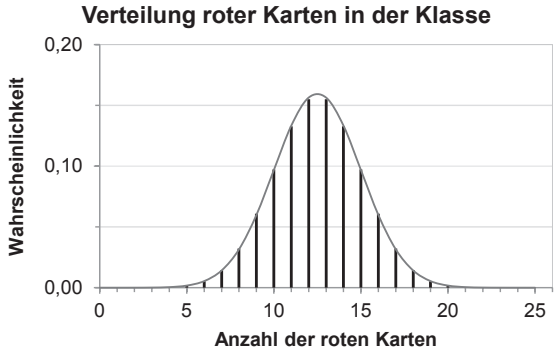

Abb. 1: Anzahl der roten Karten – Stichprobenverteilung unter der Nullhypothese

3 Den *p*-Wert berechnen

Im letzten Semester hatten wir 19 rote Karten bei 25 in der Klasse. Aus den vorhergehenden Überlegungen wissen wir, dass wir die Nullhypothese verwerfen und den Schluss ziehen, dass der Stapel durch die eingehende Bearbeitung zugunsten der roten Karten verändert worden ist. Aber wie wahrscheinlich ist denn unser beobachtetes Ergebnis?

[*Likely*, nicht *probable*; im Englischen ist das nicht immer synonym zueinander; die Frage nach der Wahrscheinlichkeit stellt sich ja nicht mehr, denn das hat sich ja schon ereignet; man muss also virtuell nachfragen „wie wahrscheinlich wäre das denn, wenn …?“]

Vom Graphen in Abb. 1 sehen wir, dass jeder beliebige Wert für die Anzahl von roten Karten wenig wahrscheinlich ist. Wie wahrscheinlich ist es nun, 19 oder mehr bzw. 6 oder weniger rote Karten (6 ist gleich weit vom Erwartungswert 12,5 wie 19) zu erhalten, wenn der Stapel ausgeglichen ist (die Nullhypothese also zutrifft)? Dies ist der *p*-Wert des Tests und dies können die Studierenden in der Regel verstehen. [Das *p* von *p*-Wert und die Bezeichnung der Erfolgswahrscheinlichkeit mit *p* ist einer der lästigen Bezeichnungskonflikte; die beiden haben nichts gemeinsam.]

Im ersten Schritt (kann auch weggelassen werden) lässt man die Studierenden erneut vom ganzen Stapel (mit gleich vielen roten und schwarzen Karten) ziehen. In derselben Klasse erhielten wir 12 rote Karten, was enger mit dem zusammenpasst, was wir erwarten, aber es ist ja nur *eine* Wiederholung des Experiments. Wie wir wissen, hat Wahrscheinlichkeit etwas mit relativen Häufigkeiten auf lange Sicht zu tun. Wie können wir den *p*-Wert schätzen? Einfach durch Simulation.

In unserer Klasse benutzen wir den TI 83/84. Damit kann man Binomialexperimente wie dieses simulieren. Im Menü MATH, PRB gibt es den Befehl *randBin*. Wir wählen diesen, bringen ihn auf das Anzeigefeld und tragen unsere Parameterwerte für die Simulation ein: *n* die Anzahl der Versuche (Studierenden), *p* die Wahrscheinlichkeit einer roten Karte unter der Nullhypothese (0,5) und die Anzahl der Wiederholungen der Simulation (wir wählten 25). Wir speichern das Ergebnis der Simulation in Listen von L1 beginnend bis L25. Um die Suche zu erleichtern, ob sich ein extremes Ergebnis (6 oder weniger bzw. 19 oder mehr) ereignet hat, ordnen wir die Listen mit dem Befehl *SortA* oder *SortD* aus dem STAT-Menü. In dieser Simulation hatte keine der Wiederholungen weniger als 7 bzw. mehr als 18 Erfolge.

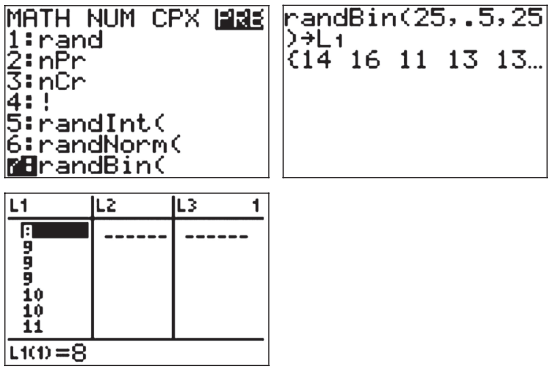

Abb. 2: Das erste Bild zeigt das Anzeigefenster mit dem *randBin*-Befehl zur Erzeugung der binomialverteilten Zufallszahlen (im Menü MATH, PRB); das zweite zeigt den eigentlichen Befehl, 25 Zufallszahlen zu erzeugen; das dritte zeigt einen Teil der sortierten Listen, die damit erzeugt wurden.

Wir haben dann die Simulationsläufe aller in der Klasse aggregiert. Das illustriert auch, wie unterschiedlich die Ergebnisse der einzelnen Studierenden sind (d. h., wie stark die Stichprobenergebnisse fluktuieren). Wir dividieren die Anzahl „extremer Ereignisse“ wie oben definiert durch die Anzahl aller Experimente (25 mal 25) und haben damit einen approximativen *p*-Wert für die ausgangs erhaltene Beobachtung. Wir erhielten im ganzen Szenario 12 extreme Ereignisse und schätzten damit den *p*-Wert durch 12/625 = 0,0192.

Den exakten Wert erhält man durch die Anwendung der Binomialverteilung. Dazu verwenden wir den Befehl *binomialcdf*(*n*, 0.5, *k*), der die Verteilungsfunktion an der Stelle *k* auswertet (d. h., die Wahrscheinlichkeit, ein Ergebnis von 0 bis zu *k* zu erhalten). Für den exakten *p*-Wert erhalten wir:

$$binomialcdf(25, 0.5, 6) + 1 - binomialcdf(25, 0.5, 18),$$

was 0,0146 liefert. Man kann auch thematisieren, warum unsere Simulation einen davon abweichenden Wert liefert. Wir hatten ja eine relative kleine Zahl von Simulationsläufen. Wenn wir mehr Genauigkeit anstreben, müssen wir mehr Daten simulieren.

Ist es möglich, im Unterricht auf PCs zurückzugreifen, so kann die Simulation etwa in Minitab oder in einer anderen Software [Fathom, Excel] durchgeführt werden. Mit mehr simulierten Daten (mehrere Tausend) erhält man dann auch bessere Schätzungen.

Zu diesem Zeitpunkt wollen die Studierenden dann endlich die wirkliche Zusammensetzung des Kartenstapels wissen. Wir deckten die Karten auf und fanden 32 rote (bei insgesamt 52) Karten; das ergibt $P(\text{rot}) = 0{,}615$.

4 Die Macht des Tests simulieren

Man kann die Lerneinheit ausbauen, um auch den Begriff der Macht eines statistischen Tests zu erklären. Macht ist die Wahrscheinlichkeit, die Nullhypothese zu Recht zu verwerfen, das ist $P(\text{verwerfe } H_0 \mid H_0 \text{ falsch})$.

Weil diese Terminologie die Studierenden sehr verwirren kann, übersetzen wir sie in die Frage „Welche Chance haben wir bei unserem Test, zu erkennen, dass ein Stapel in bestimmter Weise verändert worden ist?“ Diese Frage kam mir selber auf, als es gerade passiert war, dass wir genau gleich viele rote und schwarze Karten gezogen hatten, obwohl der Stapel (das wusste ich) *verändert* worden war.

An der Stelle muss man in der Klasse daran erinnern, wie wir die Entscheidungsregel formuliert haben: wir haben jedes Ergebnis mit mehr als 17 bzw. weniger als 8 roten Karten als inkompatibel mit der Nullhypothese angesehen und diese entsprechend verworfen. Wenn wir nun die tatsächliche Zusammensetzung des Kartenstapels kennen, wie wahrscheinlich wird das zutreffen?

Wir könnten mehrere Male 25 Karten (mit Zurücklegen) aus dem veränderten Stapel ziehen, jedes Mal feststellen, ob wir die Nullhypothese verwerfen, und dann diese Wahrscheinlichkeit schätzen. Das wäre sehr zeitaufwändig.

Da wir ja die tatsächliche Zusammensetzung mit $P(\text{rot}) = 0{,}615$ kennen, simulieren wir auf dem Taschenrechner gleich mit diesem Wert. Wir spielen die Ziehung mehrfach durch, konsolidieren die Ergebnisse und erhalten damit eine Schätzung der Macht unseres Tests (siehe Abb. 3). Wir erhielten dabei 124 Simulationen mit mindestens 18 roten sowie überhaupt keine mit höchstens 7 bei insgesamt 625 Durchläufen. Das ergibt eine Schätzung der Macht von $124/625 \approx 19{,}8\ \%$.

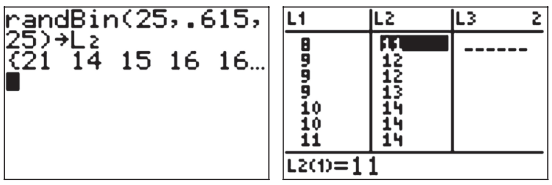

Abb. 3: Ergebnis eines Simulationslaufs zur Schätzung der Macht des Tests, falls $p = 32/52 = 0{,}615$.

Man kann das Simulationsszenario unter verschiedenen Bedingungen wiederholen. Ändert man die Stichprobengröße, so erkennt man, dass die Macht mit zunehmendem Stichprobenumfang größer wird. Ändert man die tatsächliche Zusammensetzung des Kartenstapels, so erkennt man, dass die Macht mit zunehmender Entfernung von der Nullhypothese größer wird.

5 Schlussfolgerungen

Diese Lerneinheit hat sich schon durch mehrere Semester hindurch bewährt. Die Studierenden haben durch die Demonstration ein gutes Verständnis dafür bekommen, was der *p*-Wert bedeutet. Sie erhalten langsam auch ein Gefühl für den Begriff Macht eines Tests. Der zeitliche Aufwand ist gering (eine Stunde bis etwas mehr). Der Zugang wirkt außerordentlich motivierend.

Literatur

Moore, D. S.; McCabe, G. P. (2003): Introduction to the Practice of Statistics. 4th edition. New York: W. H. Freeman.

Anschrift der Verfasserin

Patricia Humphrey
Georgia Southern University
Georgia Ave 2030 P.O.Box 8093
Statesboro, GA 30460, USA.
phumphre@georgiasouthern.edu