

Einseitige Tests, gesunder Menschenverstand, KI-Chatbots und die Bayes'sche Regel

Max konnte die Hypothese $H_0: p \leq 0,3$ gegen die Alternative $H_1: p > 0,3$ verwerfen, weil eine beobachtete Häufigkeit so groß war, dass sie im Ablehnungsbereich von H_0 lag. Eine Antwort auf die Frage, was diese Aussage bedeutet, ist nicht einfach.

In einer Analyse von 800 wissenschaftlichen Studien zeigte sich, dass in mehr als der Hälfte der Fälle die Ergebnisse der Signifikanztests von den Forschenden falsch interpretiert wurden.

Viele Anwender, Erklärvideos und Chatbots sind davon überzeugt, dass durch Verwerfen der Nullhypothese auf einem Signifikanzniveau von 5% bewiesen wurde:

„Die Alternative $H_1: p > 0,3$ gilt mit einer Sicherheit von (mehr als) 95%, weil man bei (weniger als) 5% Irrtumswahrscheinlichkeit in (mehr als) 95% richtig liegt.“

Das ist deswegen falsch, weil das 5%-Signifikanzniveau nur abschätzt, wie oft man sich – *unter der Voraussetzung, dass H_0 gilt* – fälschlich gegen H_0 entscheidet. Und das ist weder eine Aussage über die Gültigkeit von H_0 noch von H_1 . Die Irrtumswahrscheinlichkeit prognostiziert nur, wie oft man ein Ergebnis im Ablehnungsbereich erwarten würde, wenn H_0 gelten würde. Die Aussage ist auch deswegen falsch, weil Statistik Wahrscheinlichkeitsaussagen nur über Zufallsgrößen macht. Hypothesen sind aber keine Zufallsgrößen. Sie sind Modelle, die die Wirklichkeit besser oder schlechter beschreiben.

Das zu akzeptieren, fällt den meisten Menschen extrem schwer, weil wir gewohnt sind, auch Vertrauen in Hypothesen durch Sicherheitswahrscheinlichkeiten zu quantifizieren, die im Gegensatz zu den Wahrscheinlichkeiten der Statistiker keine relativen Häufigkeiten vorhersagen. Beispielsweise sind sich Graphologen und auch Chatbots „zu 90% sicher“, dass die beiden Zeilen von der gleichen Person geschrieben wurden. Hier geht es aber um ein einzelnes, nicht wiederholbares Ereignis, bei dem es keine relative Häufigkeit gibt. Um Sicherheitsaussagen über Hypothesen formulieren zu können, nutzen Chatbots und Graphologen die Idee der Bayes'schen Regel. Sie revidieren mithilfe von beobachteten Indizien schrittweise A-priori-Vertrauensverteilungen, die selten objektiv, meist am Anfang nach „Bauchgefühl“ festgelegt wurden. Bei Chatbots entstehen Sicherheitswahrscheinlichkeiten sukzessive im Rahmen des Trainings von Sprachmodellen.

*Freu Dich nicht zu spät
Freu Dich nicht zu spät*

Das sagt ein KI-Chatbot:
Die 90% sind eine Bayes'sche posterior probability, also ein Maß für den aktuellen Wissensstand – nicht eine empirisch messbare Häufigkeit.

1 Sockenlabor

Um eine Antwort auf die Eingangsfrage zu einseitigen Hypothesentests zu erhalten, gehen wir wieder ins Sockenlabor (vgl. Erkundung auf Seite 270): Vor Hannah liegen 10 äußerlich gleich aussehende Socken, von denen man nur weiß, dass sie irgendwie mit roten und weißen Kugeln gefüllt sind. Rot gilt als Treffer. Den Anteil roter Kugeln deuten wir als Trefferwahrscheinlichkeit.



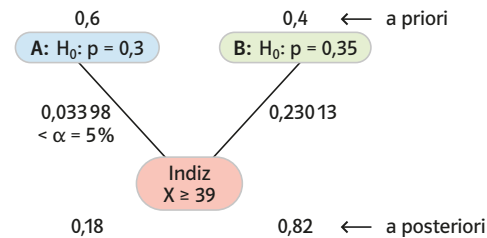
Hannah wählt eine Socke und übergibt sie an Max. Max teilt mit, dass er nach 100-maligem Ziehen mit Zurücklegen die Hypothese $H_0: p = 0,3$ gegen die Alternative $H_1: p > 0,3$ ablehnen konnte, weil er bei einem rechtsseitigen Test eine Trefferzahl k im Ablehnungsbereich von H_0 erhielt.

- a) Weisen Sie nach: Bei diesem einseitigen Test besteht der Ablehnungsbereich von H_0 aus dem Intervall $[39; 100]$ und die Irrtumswahrscheinlichkeit liegt mit 0,033 98 unter dem Signifikanzniveau von 5%.

Die hohe Trefferzahl $X \geq 39$ macht uns „recht sicher“, dass die ausgewählte Socke mehr als 30 % Treffer enthält. Um die Sicherheit quantifizieren zu können, braucht man aber weitere Informationen, die außerhalb des Sockenlabors selten präzise verfügbar sind, und im Rahmen von Datenerhebungen während der Trainingsphase von Chatbots gesammelt werden.

Wenn Max erfährt „Unter den 10 Socken gab es genau 6 mit einer Trefferwahrscheinlichkeit $p = 0,3$ und vier mit einer Trefferwahrscheinlichkeit $p = 0,35$ “, dann kann er a-priori $H_0: p = 0,3$ bzw. $H_1: p = 0,35$ mit den Sicherheitswahrscheinlichkeiten 0,6 bzw. 0,4 bewerten. Und Max kann sein Testergebnis im Rahmen der Bayes'schen Regel als Indiz $X \geq 39$ verwenden, das gegen H_0 spricht.

- b) Weisen Sie nach, dass $X \geq 39$ unter H_1 mit 0,230 13 deutlich wahrscheinlicher zu erwarten ist als unter H_0 (vgl. die Irrtumswahrscheinlichkeit aus Teilaufgabe a).
- c) Weisen Sie nach, dass man vor der Versuchsdurchführung das Indiz $X \geq 39$ mit einer totalen Wahrscheinlichkeit von 0,1124 erwartet hätte.



- d) Nutzen Sie die Bayes'sche Regel, um nachzuweisen: Dadurch, dass H_0 (wegen $X \geq 39$) auf einem Signifikanzniveau von 5% abgelehnt werden konnte, ist das Vertrauen in H_0 von 60 % auf 18 %, also um 42 Prozentpunkte, gesunken, und die Irrtumswahrscheinlichkeit, mit der H_0 verworfen wurde, und die Wahrscheinlichkeit der Alternative H_1 ergänzen sich nicht zu 100 %.
- e) Weisen Sie nach: Wenn man $H_0: p = 0,3$ auf einem Signifikanzniveau von 1% verwerfen kann, dann sinkt das Vertrauen in H_0 von 60 % auf 11 %, also um 49 Prozentpunkte.
- f) Halten Sie in einem „je-desto-Merksatz“ fest, wie sich in Abhängigkeit vom Signifikanzniveau das Vertrauen in H_0 dadurch verändert, dass man H_0 verwerfen konnte.

2 Einfluss von Parametern auf die Interpretation von Testergebnissen

Die Tabelle zeigt, wie sich das Vertrauen in die Hypothese $H_0: p = 0,3$ dadurch verändert, dass man H_0 auf dem Niveau α gegen $H_1: p > 0,3$ verwerfen konnte.

	A	B	C	D	E	F	G	H	I	J	K	L
1	H_0	H_1	priori (H_0)	posteriori (H_0)	Differenz	α	n	k	Irrtum	Pfad ₀	Pfad ₁	total
2	0,3	0,31	0,6	0,4833	0,1167	0,05	100	39	0,034	0,0204	0,0218	0,0422
3	0,3	0,32	0,6	0,3796	0,2204	0,05	100	39	0,034	0,0204	0,0333	0,0537
4	0,3	0,35	0,6	0,1813	0,4187	0,05	100	39	0,034	0,0204	0,0921	0,1124
5	0,3	0,4	0,6	0,0762	0,5328	0,05	100	39	0,034	0,0204	0,2471	0,2675
6	0,3	0,5	0,6	0,049	0,551	0,05	100	39	0,034	0,0204	0,3958	0,4162

- a) Stellen Sie die Beziehung der Tabelle zum obigen Baumdiagramm her.
- b) Erläutern Sie, welchen Einfluss die Alternative H_1 auf das Vertrauen hat, das man der Nullhypothese noch entgegenbringt, nachdem man sie auf dem Niveau α verwerfen konnte.
- c) Erforschen Sie mithilfe der interaktiven Anwendung „Bayes Deutung rechtsseitiger Tests“, welchen Einfluss der Stichprobenumfang n auf das Vertrauen hat, das man der Nullhypothese noch entgegenbringt, nachdem man sie auf dem Signifikanzniveau α verwerfen konnte.

138 **interaktive Anwendung**
Bayes Deutung
rechtsseitiger
Tests

