Testen von Hypothesen

Ein Text zum Studium

1) Der Spielwürfel

Mr X gibt uns einen Würfel, von dem er behauptet, es sei ein ganz normaler, symmetrischer Würfel. Wir werfen ihn 30 Mal und es erscheint nur eine einzige "6".

Normalerweise sollten in 30 Würfen etwa 5 Sechser erscheinen, denn im Schnitt ist zu erwarten, dass etwa jede 6. Zahl eine Sechs ist. Das eingetretene Ergebnis (nur eine einzige "6") ist ziemlich extrem. *Wir hegen wegen dieses extremen Ergebnisses den Verdacht, der betrachtete Würfel sei nicht symmetrisch* und die Sechs komme zu selten vor. Wir hegen also den Verdacht, der Würfel von Mr X sei gezinkt.

Etwas mathematischer formuliert lautet unser Verdacht: die Wahrscheinlichkeit für eine Sechs ist bei dem betrachteten Würfel nicht $p_6 = \frac{1}{6}$, sondern kleiner.

Der vorliegende Text beschreibt eine Möglichkeit, wie man solche extremen Ergebnisse deutet und wie man verdächtige Vermutungen bestätigt oder widerlegt.

a) Vorgehen bei der Lösung

Es stehen sich zwei Hypothesen gegenüber: Mr X behauptet, der Würfel sei symmetrisch, es sei also kein besonderer Effekt vorhanden. Dieser Annahme sagt man **Nullhypothese** und bezeichnet sie mit $\mathbf{H_0}$.

Wir hegen hingegen die Vermutung, der Würfel sei nicht symmetrisch. Dieser Annahme sagt man **Gegenhypothese** und bezeichnet sie mit $\mathbf{H_1}$.

Die beiden Hypothesen werden einander gegenübergestellt und wir werden uns für eine der beiden Möglichkeiten entscheiden müssen.

Nochmals, in anderen Worten formuliert: die Nullhypothese H_0 bezeichnet den "Normalfall" und der lautet in unserem Beispiel: $p_6 = \frac{1}{6}$; die Gegenhypothese H_1 bezeichnet den "Verdachtsfall". In unserem Beispiel lautet die Gegenhypothese: $p_6 < \frac{1}{6}$.

b) Konkrete Berechnung

Nun müssen wir etwas berechnen, damit wir die beiden Hypothesen gegeneinander abwägen können. Da wir nicht wissen, ob der Würfel *tatsächlich* asymmetrisch ist, gehen wir vom Normalfall aus, d.h. wir nehmen vorerst an, der Würfel sei symmetrisch, es sei kein besonderer Effekt vorhanden, oder juristisch ausgedrückt, wir starten bei der "Unschuldsvermutung".

Die nachfolgenden Berechnungen werden also mit der Annahme $p_6 = \frac{1}{6}$ durchgeführt.

Das eingetretene Ergebnis (nur eine einzige "6") ist ziemlich extrem. Wir betrachten nun das eingetretene Ergebnis und alle noch extremeren Fälle (warum man die noch extremeren Fälle dazunimmt, wird später begründet). Wir betrachten also als **Extrembereich** diejenigen Fälle, wo in 30 Würfen mit einem Würfel höchstens eine "6" erscheint.

Wir berechnen nun die **Wahrscheinlichkeit** dafür, dass ein Ergebnis aus dem Extrembereich eintritt, also die Wahrscheinlichkeit, mit einem symmetrischen Würfel in 30 Würfen höchstens eine "6" zu werfen.

Diese Wahrscheinlichkeit beträgt 2.95% (bitte selber nachrechnen).

Die Wahrscheinlichkeit, mit einem symmetrischen Würfel auf ein so extremes Ergebnis zu kommen, ist sehr klein (das war ja zu erwarten). Die Frage ist nun, ob diese Wahrscheinlichkeit

genügend klein ist, damit wir berechtigterweise sagen können, der Würfel sei als asymmetrisch anzusehen.

Aus verschiedenen Gründen (die hier nicht näher erläutert werden) wählt man häufig eine Grenze von 5%. Unsere erhaltene Wahrscheinlichkeit ist kleiner. Also werden wir die **Nullhypothese verwerfen** und zur Gegenhypothese übergehen. **Wir sehen durch die Berechnung den Verdacht erhärtet, dass der Würfel von Mr X *nicht* symmetrisch sei**.

c) Zusammenfassung

Durch das Experiment haben wir ein sehr extremes Ergebnis erhalten, nämlich nur eine "6" in 30 Würfen. Wir haben somit den Verdacht, der Würfel sei nicht symmetrisch.

Die Wahrscheinlichkeit, mit einem symmetrischen Würfel auf ein so extremes oder noch extremeres Ergebnis zu kommen, ist ziemlich klein; insbesondere kleiner als 5%.

Als Folge davon werden wir mit einer gewissen Berechtigung annehmen dürfen, der Würfel sei nicht symmetrisch. **Wir verwerfen die Nullhypothese**.

2) Münzwurf

(In diesem Beispiel ist der Text auf die entscheidenden Schritte zusammengekürzt.)

Wir werfen eine Münze 24 Mal und es erscheinen 16 "Kopf".

Da die Anzahl "Kopf" unerwartet gross ist (es wären ja etwa 12 "Kopf" zu erwarten), hegen wir den Verdacht, die Münze sei nicht symmetrisch.

Die Nullhypothese H_0 lautet jetzt $p = ½$, die Gegenhypothese H_1 lautet $p > ½$.

Der "Extrembereich" (das eingetretene Ergebnis und alle noch extremeren Fälle) lautet in diesem Beispiel: mindestens 16 "Kopf"-Würfe in 24 Würfen.

Unter der Annahme der Nullhypothese ist die Wahrscheinlichkeit dafür 7.58% (bitte selber nachrechnen). Diese Wahrscheinlichkeit ist aber zu gross, als dass wir die Nullhypothese verwerfen könnten. Wir werden also die **Nullhypothese beibehalten** und die Münze weiterhin als symmetrisch betrachten.

Weil es wichtig ist, nochmals: die Wahrscheinlichkeit, mit einer symmetrischen Münze in 24 Würfen 16 oder mehr "Kopf" zu erhalten, beträgt 7.58%, ist also nicht genügend klein. **Wir bleiben bei der Nullhypothese und betrachten die Münze als symmetrisch**.

3) Testen von Hypothesen (dasselbe nochmals, diesmal theoretisch)

Bei einem Hypothesentest stellt man eine **Nullhypothese H_0** (diese bezeichnet den Normalfall) einer **Gegenhypothese H_1** gegenüber.

Wenn man das Experiment durchgeführt hat und ein extremes Ergebnis eingetreten ist, dann bildet *dieses und alle noch extremeren Ergebnisse* den Extrembereich E. Wir berechnen dann unter Annahme der Nullhypothese die Wahrscheinlichkeit s dafür, in den Extrembereich zu gelangen.

s heisst das erreichte oder beobachtete **Signifikanzniveau** und stellt also die Wahrscheinlichkeit dar, auf ein so extremes Ergebnis zu kommen, wenn man zunächst annimmt, dass kein besonderer Effekt vorhanden sei.

Nun stellt sich die Frage, was mit der Zahl s anzufangen ist. Ist s sehr klein (sagen wir, kleiner als eine Zahl α), dann wird man einen Effekt vermuten und die Nullhypothese **verwerfen**; ist s nicht besonders klein (aus praktischen Gründen: grösser oder gleich α), wird man die Nullhypothese **beibehalten**.

Die vorerst unbestimmte Zahl α wird je nach Wichtigkeit des Tests und je nach allfälligen zu ziehenden Konsequenzen verschieden gross gewählt werden. Üblich sind Werte von $\alpha = 5\%$ (man spricht dann von einem **signifikanten Ergebnis**) oder $\alpha = 1\%$ (man spricht dann von einem **hochsignifikanten Ergebnis**).

Da uns das so beschriebene Testverfahren nur gestattet, irgendwelche Abweichungen (vom Normalfall) als signifikant oder nichtsignifikant zu erklären, heisst ein solcher Test auch **Signifikanztest**.

4) Eine Bemerkung zum Extrembereich

Warum nimmt man eigentlich für den Extrembereich das eingetretene Ergebnis *und alle noch extremeren Fälle* und nicht nur das eingetretene Ergebnis allein?

Bei sehr vielen möglichen Ergebnissen werden die Wahrscheinlichkeiten jedes einzelnen Ergebnisses sehr klein. Folgendes Beispiel möge das illustrieren: wir werfen eine Münze 1000 Mal und es erscheinen 497 "Kopf". Kaum jemand wird auf Grund dieses Resultates die Vermutung haben, die Münze sei nicht symmetrisch, denn ein Ergebnis von 497 "Kopf"-Würfe von 1000 ist nun wirklich völlig normal.

Aber die Wahrscheinlichkeit, *genau* 497 "Kopf"-Würfe zu haben, beträgt nur 2.48% (und das wäre ein signifikantes Ergebnis). Hingegen ist die Wahrscheinlichkeit für höchstens 497 "Kopf"-Würfe 43.72%, was den absolut normalen Versuchsausgang bestätigt.

5) Noch eine allgemeine Bemerkung

Wenn wir nach dem Durchführen des Zufallsexperimentes zum Schluss kommen, H_0 sei zu verwerfen, dann wird *nicht* gesagt, H_0 sei falsch oder H_1 sei wahr. Es besteht nur ein **Grund zur Annahme**, die Gegenhypothese als die bessere anzusehen.

Analog gilt: wenn H_0 beibehalten wird, dann wird auch *nicht* gesagt, H_1 sei falsch.

Man wird also stets nur die beiden Hypothesen gegeneinander abwägen und sich für die aufgrund des Ergebnisses günstigere der beiden Alternativen entscheiden.

[In der Regel ist es in der Mathematik so, dass von zwei gegenteiligen Ergebnissen das eine richtig und das andere falsch ist. In der Statistik stellt man *nur Vermutungen* auf, die sich aus durchgeführten Versuchen ergeben.]

6) Der Zöllner

Beim Zoll passieren 9 Personen, davon sind 4 Schmuggler und 5 ehrliche Leute. Der Zöllner wählt drei Personen zur Kontrolle aus. Jetzt stellt sich heraus, dass er lauter Schmuggler erwischt hat. Handelt es sich somit um einen guten Zöllner?

In diesem Beispiel besagt die Nullhypothese: der Zöllner wählt zufällig; die Gegenhypothese lautet: der Zöllner wählt nicht zufällig.

Weil der Zöllner sicher 3 *verschiedene* Personen zur Kontrolle auswählt, müssen wir die Aufgabe umformulieren und die Wahrscheinlichkeit s wie folgt berechnen: In einem Behälter befinden sich 9 Kugeln, davon sind 4 rote. Wie gross ist die Wahrscheinlichkeit, lauter rote Kugeln zu ziehen, wenn man 3 Mal *zufällig (H_0!)* ohne Zurücklegen zieht?

Die Wahrscheinlichkeit s beträgt 4.76%, liegt also unter 5%. Somit wird die **Nullhypothese verworfen** und der Zöllner als gut angesehen.

7) Medikamentenprüfung

(Dieses angewandte Beispiel ist so weit vereinfacht, dass wir die nötigen Berechnungen sinnvoll anstellen können.)

Zur Heilung einer bestimmten Krankheit gab es bisher nur das bewährte Mittel A, das mit Wahrscheinlichkeit ¾ heilte. Nun wurde von der Pharmaindustrie ein Mittel B entwickelt und an 25 Patienten getestet. Dabei wurden 22 geheilt. Ist jetzt B als das bessere Medikament anzusehen?

In diesem Beispiel ist die Nullhypothese die vom Medikament A bekannte Wahrscheinlichkeit $p = ¾$. Die Pharmaindustrie stellt für das Medikament B natürlich die Gegenhypothese $p > ¾$ auf. Die zu berechnende Wahrscheinlichkeit können wir nun so formulieren: Wir setzen vorerst $p = ¾$ für "Erfolg" im einzelnen Versuch, d.h. am einzelnen Patienten. Das beobachtete Signifikanzniveau s ist dann die Wahrscheinlichkeit, in mindestens 22 von 25 Fällen "Erfolg" zu haben. Berechne selber.

Das Resultat ist $s = 9.62\%$ und liegt somit weit über $\alpha = 5\%$. Folglich wird man die **Nullhypothese beibehalten**, d.h. beim Medikament A bleiben. Das Medikament B muss noch verbessert und neu getestet werden.

8) Fehler 1. Art und Fehler 2. Art

Zum Schluss des Textes eine Bemerkung zu Fehlern. Wenn man zwei Hypothesen gegeneinander abwägt, dann kann es sein, dass man sich für die falsche der beiden Hypothesen entscheidet.

Beim Durchführen eines Tests könnte es sein, dass man ein sehr extremes Ergebnis erhält, obschon H_0 korrekt ist. Die Wahrscheinlichkeit dafür beträgt genau s. In diesem Fall begeht man einen Fehler, indem man H_0 **fälschlicherweise verwirft**. Man sagt diesem Fehler ein **Fehler 1. Art**.

Andererseits kann es sein, dass die Gegenhypothese richtig ist, dass man dies aber im Test gar nicht merkt (weil kein extremes Ergebnis eintritt). Dann begeht man einen sogenannten **Fehler 2. Art**, indem man H_0 **fälschlicherweise beibehält**. Da man beim Durchführen eines Tests meist über H_1 keine genauen Angaben machen kann, kann man die Wahrscheinlichkeit eines Fehlers 2. Art in der Regel nicht berechnen.

Die Wahrscheinlichkeit eines Fehlers 1. Art muss gering gehalten werden, da man H_0 nur verwerfen sollte, wenn genügend Grund dazu besteht. Beispielsweise wird man ein bewährtes Medikament (= Nullhypothese!) nur dann absetzen (= H_0 verwerfen) und durch ein neues Medikament (= H_1) ersetzen, wenn man sehr sicher ist (= Signifikanzniveau), dass das neue Medikament wirklich besser ist (= extrem gutes Testergebnis, z.B. das neue Medikament heilt viel besser).

Ein Fehler 2. Art ist in der Regel nicht so "schlimm", da man das Bewährte, nämlich die Nullhypothese, beibehält. Dazu ein Beispiel aus der Wirtschaft: eine bewährte Maschine für einen Massenartikel (z.B. Bolzen) wird man nicht gerade auf den Schrotthaufen werfen, wenn eine zweite, neue Maschine, *etwas* bessere Resultate (d.h. etwas weniger Ausschuss) produziert. Man wird die bewährte Maschine (= Nullhypothese) beibehalten.