

# Testen von Hypothesen

## eine Anwendung der Binomialverteilung

### I. Einseitiger Test einer Hypothese

Von einem Würfel wird vermutet, daß er öfters die Sechsen liefert, als es bei einem Laplace-Würfel zu erwarten ist. Es soll ein **Test** entworfen werden, um die **Hypothese**, es handle sich um einen Laplace-Würfel, zu untersuchen.

Dazu wird geplant, den Würfel  $n=100$  mal zu werfen und dabei die Zufallsvariable  $X$ =Anzahl der aufgetretenen Sechsen zu betrachten.

Sei  $H_0$ : "Es handelt sich um einen Laplace-Würfel." ( $p(\{6\}) = 1/6$ ) die **Nullhypothese**.

Sei  $H_1$ : "Die Sechsen erscheint zu häufig." ( $p(\{6\}) > 1/6$ ) die **Gegenhypothese**.

Mit einer zunächst willkürlich festgelegten Zahl  $k$ , etwa  $k=25$ , wird die folgende **Entscheidungsregel** festgelegt:

$X \leq k \rightarrow H_0$  wird akzeptiert.

$X > k \rightarrow H_1$  wird akzeptiert.

Das so gebildete Urteil kann natürlich falsch sein:

**Fehler 1. Art:** Es handelt sich in Wirklichkeit um einen Laplace-Würfel, aber  $X > k$ , und  $H_1$  wird also fälschlicherweise akzeptiert.

**Fehler 2. Art:** Es handelt sich in Wirklichkeit um keinen Laplace-Würfel, aber  $X \leq k$  und  $H_0$  wird also fälschlicherweise akzeptiert.

Es ist klar, daß die Größe dieser Fehler durch die Wahl von  $k$  beeinflusst wird, deshalb ist es wichtig, diese Fehler zu berechnen, um sie durch eine geeignete Wahl von  $k$  klein zu halten.

Bezeichne  $\alpha'$  den Fehler 1. Art, dann gilt:

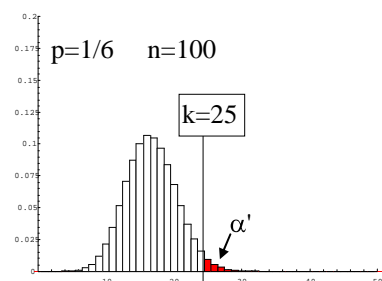
$$\alpha' = P_{H_0}(X > k)$$

$$= 1 - F_{\frac{1}{6}}^{100}(k) \quad \left( F_p^n(k) := \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i} \right)$$

$$= 1 - 0.9881$$

$$= 0.0119$$

$$= 1.19\%$$

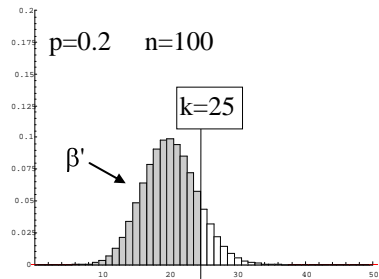


Das heißt: Mit einer Wahrscheinlichkeit von 1.19% wird ein Laplace-Würfel irrtümlicherweise für einen gefälschten Würfel gehalten. Diese Fehlerwahrscheinlichkeit ist also vertretbar klein.

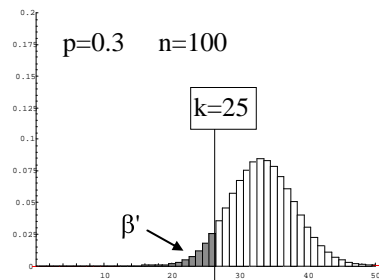
Den Fehler 2. Art zu berechnen ist schwierig, weil man die Wahrscheinlichkeit  $p$  für eine Sechsen nicht kennt. Nehmen wir an, der Würfel sei gefälscht und es gelte  $p(\{6\}) = 0.2$ . Dann gilt:

$$\text{Fehler 2. Art} =: \beta' = P_{H_1}(X \leq k) = F_{0.2}^{100}(k) = 0.9125 = 91.25\%.$$

Das bedeutet, daß auch ein gefälschter Würfel mit der Wahrscheinlichkeit von 91.25% noch irrtümlicherweise für einen echten Laplace-Würfel gehalten wird. Wenn ein Urteil mit einem solch großen Fehler behaftet ist, ist es natürlich fast wertlos.



Es ist offensichtlich, daß der Fehler 1. Art klein wird, wenn  $k$  größer gewählt wird. Der Fehler 2. Art jedoch kann prinzipiell nicht durch  $k$  kontrolliert werden, da die Wahrscheinlichkeit für die Sechsen bei einem gefälschten Würfel nicht bekannt ist.



Man muß deshalb die Entscheidungsregel abändern:

$$\begin{aligned} X \leq k &\rightarrow H_0 \text{ wird nicht abgelehnt.} \\ X > k &\rightarrow H_0 \text{ wird abgelehnt} \quad (= H_1 \text{ wird akzeptiert.}) \end{aligned}$$

Nur wenn die Versuchsreihe mehr als  $k$  Sechsen ergeben hat, (Man sagt dann: "Der Test zeigt ein **signifikantes Ergebnis**.") kann man also eine praktisch brauchbare Schlußfolgerung aus dem Test ziehen: Es handelt sich mit einem möglichen Fehler von 1.19% um einen gefälschten Würfel. Im anderen Fall ist keine Aussage möglich (Häufig findet man jedoch auch die irrije Meinung, der Test habe gezeigt, daß der Würfel nicht gefälscht sei.).

Bei der praktischen Planung eines Tests gibt man häufig eine obere Schranke  $\alpha$ , etwa  $\alpha=5\%$  für den Fehler 1. Art vor, und bestimmt dann die kleinste Zahl  $k$ , für die der Fehler 1. Art höchstens gleich  $\alpha$  ist:

$$\alpha \geq \alpha' = P_{H_0}(X > k) = 1 - F_{\frac{1}{6}}^{100}(k)$$

$$\Rightarrow F_{\frac{1}{6}}^{100}(k) \geq 1 - \alpha' = 95\%$$

$$\Rightarrow k_{\min} = 23$$

Ergeben sich bei dem Versuch also mehr als 23 Sechsen, so kann man auf dem Signifikanzniveau 5% (mit einer Sicherheit von mindestens 95%) sagen, daß der Würfel gefälscht ist. Andere Versuchsergebnisse bezeichnet man als nicht signifikant (auf dem Niveau von 5%) und es ist keine Schlußfolgerung möglich.

Dieser Test heißt **einseitig**, weil der **Ablehnungsbereich**  $[k+1, k+2, \dots, 100]$  nur auf einer Seite des Erwartungswertes von  $X$  liegt. Man wählte hier diesen Test deshalb, weil von vornherein vermutet wurde, daß die Sechsen zu häufig auftrat. Hätte man nur vermutet, daß die Wahrscheinlichkeit für eine Sechsen von  $1/6$  verschieden ist, so hätte man einen

Ablehnungsbereich wählen müssen, der auf beiden Seiten des Erwartungswertes von  $X$  gelegen ist.

## II. Zweiseitiger Test einer Hypothese

Von einem Würfel wird vermutet, daß er die Sechsen mit einer Wahrscheinlichkeit liefert, die nicht gleich  $1/6$  ist, wie es bei einem Laplace-Würfel zu erwarten wäre. Es soll ein **Test** entworfen werden, um die **Hypothese**, es handle sich um einen Laplace-Würfel, zu untersuchen.

Es wird wieder geplant, den Würfel  $n=100$  mal zu werfen und dabei die Zufallsvariable  $X$ =Anzahl der aufgetretenen Sechsen zu betrachten.

Sei  $H_0$ : "Es handelt sich um einen Laplace-Würfel." ( $p(\{6\})=1/6$ ) die **Nullhypothese**.

Sei  $H_1$ : "Es handelt sich um **keinen** Laplace-Würfel." ( $p(\{6\})\neq 1/6$ ) die **Gegenhypothese**.

Da hier, anders als im vorangegangenen Beispiel, auch bedacht werden muß, daß der Würfel vielleicht zu selten eine Sechsen produziert, muß der Ablehnungsbereich der Nullhypothese auf beiden Seiten des Erwartungswertes für  $X$  eines Laplace-Würfels gelegen sein (**zweiseitiger Test**). Das heißt, wenn entweder sehr wenige oder sehr viele Sechsen auftreten, werden wir die Nullhypothese verwerfen.

Der Ablehnungsbereich ist also von der Form:  $[0, 1, \dots, k_1] \cup [k_r, k_r+1, \dots, 100]$ . Bei der Planung des Tests gibt man sich wieder eine obere Schranke  $\alpha$  (z.B.  $\alpha = 5\%$ ) für den Fehler 1. Art  $\alpha'$  an. Es soll also gelten:

$$5\% = \alpha \geq \alpha' = P_{H_0}(X \leq k_1 \vee k_r \leq X) = F_{\frac{1}{6}}^{100}(k_1) + 1 - F_{\frac{1}{6}}^{100}(k_r - 1) .$$

Es gibt nun viele Möglichkeiten,  $k_1$  und  $k_r$  so zu wählen, daß diese Bedingung erfüllt ist: Wenn die linke Teilmenge klein gehalten wird ( $k_1$  klein), dann kann man die rechte Teilmenge etwas größer wählen ( $k_r$  klein) oder umgekehrt. Man würde jedoch nur dann diese beiden Teilmengen unsymmetrisch wählen, wenn man a priori schon eine Vermutung über die Art der Fälschung des Würfels hat. Wenn man glaubt, daß der Würfel eher zu häufig als zu selten die Sechsen liefert, dann sollte man die rechte Teilmenge des Ablehnungsbereiches größer und die linke kleiner wählen. Das bedeutet, daß die linke Teilmenge leer sein sollte, wenn man annimmt, es komme nur in Frage, daß der Würfel entweder echt sei oder er zu viele Sechsen produziere. Dann handelt es sich wieder um den vorher diskutierten einseitigen Test.

Ist a priori keine Information über die mögliche Art der Fälschung des Würfels vorhanden, so wählt man  $k_1$  und  $k_r$  symmetrisch. Das heißt: Die Ungleichungen

$$2.5\% = \frac{\alpha}{2} \geq P_{H_0}(X \leq k_1) \quad \text{und} \quad 2.5\% = \frac{\alpha}{2} \geq P_{H_0}(k_r \leq X) \quad \text{sollten erfüllt sein.}$$

$$\Rightarrow 0.025 \geq F_{\frac{1}{6}}^{100}(k_1) \quad \text{und} \quad 0.025 \geq 1 - F_{\frac{1}{6}}^{100}(k_r - 1)$$

$$\Rightarrow 9 \geq k_1 \quad \text{und} \quad k_r - 1 \geq 24$$

Der Ablehnungsbereich für die Nullhypothese lautet jetzt  $[0,1,..9] \cup [25, 26,.. 100]$ . Erhält man also bei 100 Würfeln eine Anzahl von Sechsen, die in diese Menge fällt, so kann man bei einer Sicherheit von 95% behaupten, der Würfel sei gefälscht.

### III. Konstruktion eines Tests

Welchen Einfluß hat die Wahl der Fehlerschranke  $\alpha$  bzw. der Zahl  $k$  (beim oben beschriebenen einseitigen Test) auf die Aussagekraft eines Testergebnisses? Dazu stelle man sich vor, daß viele, unbekannte Würfel daraufhin getestet werden, ob sie zu häufig die Sechsen liefern.

**Je größer nun  $k$  gewählt wird**, desto kleiner ist der Fehler 1. Art; das heißt, daß man nur sehr selten einen echten Würfel irrtümlicherweise für einen gefälschten hält. Oder, positiv ausgedrückt: Fast jeder als gefälscht gehaltene Würfel ist tatsächlich gefälscht. Erkauft wird diese relative Sicherheit des Urteils durch eine hohe Rate von Würfeln, die nicht als gefälscht erkannt werden, obwohl sie es sind (großer Fehler 2. Art).

Es gibt durchaus reale Situationen, in denen ein solches Testverhalten sinnvoll ist: Betrachtet man ein Gerichtsverfahren als einen Test (Nullhypothese: "Der Angeklagte ist unschuldig."), so ist es gerade wünschenswert, daß eine etwaige Verurteilung des Angeklagten (Die Nullhypothese wird abgelehnt.) nur dann erfolgt, wenn das Gericht sich seiner Sache sehr sicher ist (Der Fehler 1. Art sollte sehr klein sein.). Der Grundsatz "in dubio pro reo" drückt gerade aus, daß wir bereit sind, große Fehler 2. Art hinzunehmen.

**Je kleiner  $k$  gewählt wird**, desto größer wird der Fehler 1. Art, und der Fehler 2. Art wird kleiner. In einem solchen Fall zeigt der Test sehr häufig ein signifikantes Ergebnis: Viele Würfel werden, vielleicht auch irrtümlicherweise, als gefälscht erklärt. In den anderen Fällen aber, wenn der Test kein signifikantes Ergebnis zeigt, sind die Würfel echt oder nur schwach gefälscht ( $p(\{6\}) = 1/6 + \epsilon$ ).

Ein solches Testverhalten ist zum Beispiel bei einer Krebsvorsorgeuntersuchung (Nullhypothese: "Der Patient ist gesund.") erwünscht: Bei möglichst wenigen Menschen sollte die einfache Vorsorgeuntersuchung eine bereits vorhandene Erkrankung unerkannt lassen. Der hohe Fehler 1. Art (Relativ viele Menschen erhalten die zunächst beunruhigende Nachricht, erkrankt zu sein, obwohl sie es tatsächlich nicht sind.) ist in dieser Situation vertretbar, denn eine nachfolgende genauere Gewebeuntersuchung, die man aus Zeit- und Kostengründen nicht bei allen Testpersonen anwenden will, wird bald für Klarheit sorgen.

Die Fehlerschranke  $\alpha$  kann also nicht mathematisch **berechnet** werden, sondern entscheidend für ihre **Wahl** ist die Absicht, die mit dem Test verbunden ist.

Abschließend soll die Konstruktion eines Tests anhand eines Beispiels erläutert werden.

*In einem Spielkasino wird ein Spiel mit einem Würfel angeboten, das an zwanzig verschiedenen Tischen gleichzeitig gespielt wird. Nachdem einige Kunden der Polizei von spektakulären Spielverlusten berichtet haben, vermutet der Kommissar, dass einige der benutzten Würfel keine Laplacewürfel sind, sondern so gefälscht sind, dass sie*

- a) *die Sechsen nur mit einer Wahrscheinlichkeit erzeugen, die unter  $1/6$  liegt.*
- b) *die Sechsen mit einer Wahrscheinlichkeit erzeugen, die größer als  $1/6$  ist.*
- c) *die Sechsen mit einer Wahrscheinlichkeit erzeugen, die ungleich  $1/6$  ist.*

Anstatt nun alle Angestellte des Spielkasinos zu verhaften und die Würfel zu beschlagnahmen, um die Personen zu verhören und die Würfel auf mögliche Bleieinlagen zu untersuchen, dieses Vorgehen erscheint angesichts bloßer Verdächtigungen als unangemessen, erwägt der Kommissar, einen Test durchzuführen. Dazu sollen seine Mitarbeitern die Ergebnisse von 50 Würfeln eines jeden der zwanzig eingesetzten Würfel notieren, um in Abhängigkeit dieser Ergebnisse zu entscheiden, auf welchen Angestellten des Kasinos und auf welchen Würfeln er seine Untersuchungen konzentrieren sollte. In den Fällen a) oder b) (Es liegt ein Vorwissen über die Art der möglichen Fälschung vor.) wählt also der Kommissar eine natürliche Zahl  $k$  mit  $0 \leq k \leq 50$  und stellt dann die folgende Entscheidungsregel auf:

Für den Fall a)  $\mathbf{X} \leq k \rightarrow \mathbf{H}_0 \text{ wird abgelehnt}$  (=  $H_1$  wird akzeptiert.)  
 $\mathbf{X} > k \rightarrow \mathbf{H}_0 \text{ wird nicht abgelehnt.}$

$H_0$  ist wieder die Nullhypothese: "Es handelt sich um einen Laplace-Würfel." Nun muss der Kommissar entscheiden, welchen Wert er für  $k$  nehmen soll. Bei einem Laplace-Würfel wäre zu erwarten, dass etwa 8 Sechsen bei 50 Würfeln erscheinen. Wenn er also für  $k$  den Wert 4 einsetzt, so wird es bei einem ungefälschten Würfel nur selten passieren, dass so wenige Sechsen erscheinen und er deshalb fälschlicherweise für einen gefälschten Würfel gehalten wird. Der Fehler erster Art  $\alpha'$  ist also klein:  $\alpha' = F_{\frac{1}{6}}^{50}(4) = 6.43\%$ . Selbst wenn also alle

zwanzig eingesetzten Würfel echt sind, wird er nur etwa einen oder zwei davon (6.43% von 20) nach seinem Test für unecht halten. Die Gefahr sich in der Öffentlichkeit durch vorschnelle, letztlich ungerechtfertigte, vorläufige Festnahmen zu diskreditieren ist bei dieser Wahl von  $k$  für den Kommissar also vertretbar klein. Der Preis für diese Sicherheit ist jedoch ein großer Fehler zweiter Art: Nehmen wir an, 10 der im Spielkasino eingesetzten Würfel seien so gefälscht, dass sie die Sechsen nur mit einer Wahrscheinlichkeit  $1/10$  zeigen. Dann gilt  $\beta' = 1 - F_{\frac{1}{10}}^{50}(4) = 56.88\%$ . Der Kommissar muss damit rechnen, dass etwa 5 oder 6 (56.88% von

10) der tatsächlich gefälschten Würfel bei dieser Wahl  $k=4$  von seinem Test nicht entdeckt werden. Wären die Würfel nicht ganz so stark gefälscht (Wahrscheinlichkeit für eine Sechsen =  $1/8$ ), dann sähe die Bilanz für den Kommissar noch schlechter aus: Dann entkämen 76.54% also etwa 7 oder 8 der angenommenen 10 gefälschten Würfel unerkant.

Um den Wert von  $k$  festzulegen, muss der Kommissar also zuerst entscheiden, ob es ihm wichtiger ist, möglichst keinen Kasinoangestellten zu unrecht zu verdächtigen, dann muss er für  $k$  einen kleinen Wert wählen; viele Gauner werden ihm so jedoch entwischen. Oder möchte er möglichst viele Ganoven entlarven, dann wird er für  $k$  größere Werte einsetzen. Viele Unschuldige werden dann jedoch auch verdächtigt. Da der Fehler zweiter Art prinzipiell unkontrollierbar ist, **setzt** sich der Kommissar also eine obere Schranke für den Fehler erster Art **gemäß seiner Testabsichten** und bestimmt dann den dazu gehörigen größtmöglichen Wert für  $k$ : Als Beispiel gehen wir von 15% als obere Schranke für den Fehler erster Art aus.  $15\% \geq \alpha' = F_{\frac{1}{6}}^{50}(k) \Rightarrow k = 5$ .

Im Fall b)  $\mathbf{X} \geq k \rightarrow \mathbf{H}_0 \text{ wird abgelehnt}$  (=  $H_1$  wird akzeptiert.)

**$X < k \rightarrow H_0$  wird nicht abgelehnt,**

verläuft die Argumentation ähnlich zu der im Fall a). Zum Signifikanzniveau 15% findet der Kommissar den kleinstmöglichen Wert für  $k$  aus der Bedingung:

$$15\% \geq \alpha' = 1 - F_{\frac{1}{6}}^{50}(k-1) \Rightarrow k = 12.$$

Hat der Kommissar im Fall c) kein Vorwissen über die Art, wie die Würfel gefälscht sind, dann wird er, weiterhin zu dem Signifikanzniveau 15%, einen symmetrischen Ablehnungsbereich für die Nullhypothese wählen:

Die Entscheidungsregel lautet dann:

**$X \leq k_1 \vee k_r \leq X \rightarrow H_0$  wird abgelehnt (=  $H_1$  wird akzeptiert.)**

**$k_1 < X < k_r \rightarrow H_0$  wird nicht abgelehnt,**

$$7.5\% \geq \frac{\alpha'}{2} = F_{\frac{1}{6}}^{50}(k_1) \Rightarrow k_1 = 4, \text{ und}$$

$$7.5\% \geq \frac{\alpha'}{2} = 1 - F_{\frac{1}{6}}^{50}(k_r - 1) \Rightarrow k_r = 13.$$

Durch einen Vergleich mit dem Ergebnis im Fall a) bzw. im Fall b) erkennt man hier sogar mathematische Gründe dafür, dass die Ermittlungen in einer Spielhölle um so erfolgreicher sind, je mehr zutreffendes Vorwissen über die Art der Fälschung (Wahrscheinlichkeit der Sechsen ist erhöht oder erniedrigt) vorhanden ist.