

Statistisches Testen (Einführung)

M. Gruber

SS 2009

– Typeset by Foil \LaTeX –

Testen der Hypothese “die Münze ist fair”

Test der Hypothese H_0 “die Münze ist fair”: Wir werfen die Münze fünfmal. Wenn dabei das Ereignis AAAAA oder ZZZZZ eintritt, **verwerfen** wir H_0 , ansonsten **nehmen** wir H_0 an.

Formal: Über die $\mathcal{B}(1, p)$ -verteilten ZVn X_1, \dots, X_5 wird die Hypothese $H_0: p = 0.5$ aufgestellt. Zur **Testvariablen**

$$U(X_1, \dots, X_5) = \sum_{1 \leq i \leq 5} X_i$$

geben wir uns die **Entscheidungsvorschrift**: Verwirf H_0 , wenn die Realisierung von $U(X_1, \dots, X_5)$ den Wert 0 oder 5 ergibt.

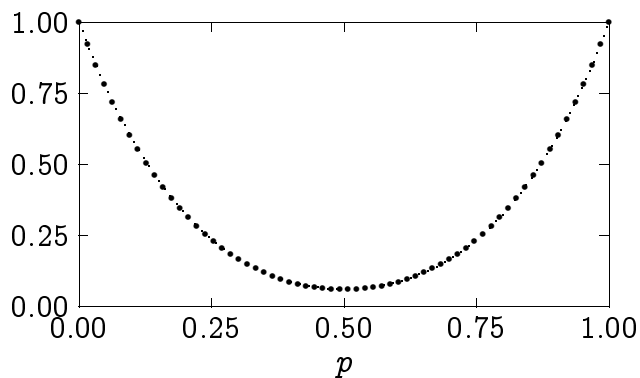
α -Fehler und β -Fehler

Mögliche Entscheidungsverläufe dieses Tests:

1. (Berechtigte) Ablehnung von H_0 , wenn H_0 nicht zutrifft.
2. (Unberechtigte) Ablehnung von H_0 , wenn H_0 zutrifft (**α -Fehler** oder **Fehler 1. Art**); Wahrscheinlichkeit hierfür: $2 \cdot \left(\frac{1}{2}\right)^5 = 0.0625$.
3. (Berechtigte) Annahme von H_0 , wenn H_0 zutrifft.
4. (Unberechtigte) Annahme von H_0 , wenn H_0 nicht zutrifft (**β -Fehler** oder **Fehler 2. Art**); Wahrscheinlichkeit hierfür: $1 - p^5 - (1 - p)^5$ mit dem (wahren) Parameter $0 \leq p < \frac{1}{2}$ bzw. $\frac{1}{2} < p \leq 1$.

Gütefunktion des Tests

Die **Gütefunktion** des Tests $G(p) = p^5 + (1 - p)^5$ zeigt die Ablehnwahrscheinlichkeit von H_0 als Funktion von p . An ihr kann man die Wahrscheinlichkeit für den α -Fehler ($G(p)$, $p = \frac{1}{2}$) und die β -Fehler ($1 - G(p)$, $p \neq \frac{1}{2}$) ablesen.



Man erkennt: Die Ablehnwahrscheinlichkeit von H_0 wird erst groß, wenn p stark von $\frac{1}{2}$ abweicht.

Kritische Region und Annahmereich

Der Test unterscheidet 2 Teilmengen des Wertebereichs der Testfunktion U :

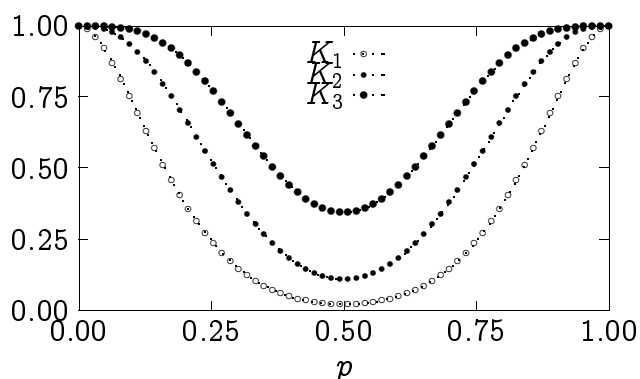
1. Führt die Realisierung von U zu den Werten 0 oder 5, so verwerfen wir H_0 :

Die Menge $\{0, 5\}$ heißt **kritischer Bereich** des Tests.

2. Führt die Realisierung von U zu einem Wert der Menge $\{1, 2, 3, 4\}$, so behalten wir H_0 (bis auf weiteres) bei.

Die Menge $\{1, 2, 3, 4\}$ heißt **Annahmereich** des Tests.

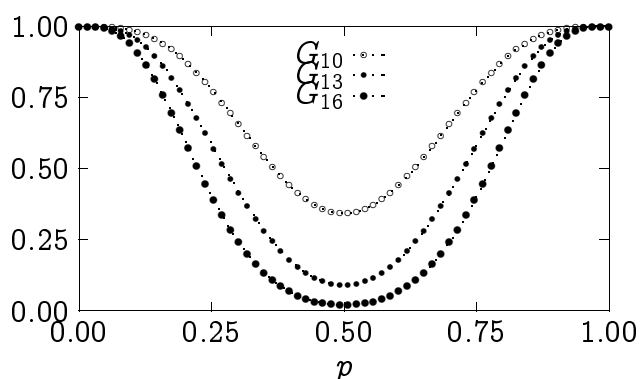
Vergrößerung des kritischen Bereichs



Durch Vergrößerung des kritischen Bereichs wird der β -Fehler kleiner, der α -Fehler dagegen größer.

Die Graphik zeigt die Gütefunktionen des Münzwurfttests für Stichprobenumfang $n = 10$ und die kritischen Bereiche $K_1 = \{0, 1, 9, 10\}$, $K_2 = \{0, 1, 2, 8, 9, 10\}$, $K_3 = \{0, 1, 2, 3, 7, 8, 9, 10\}$.

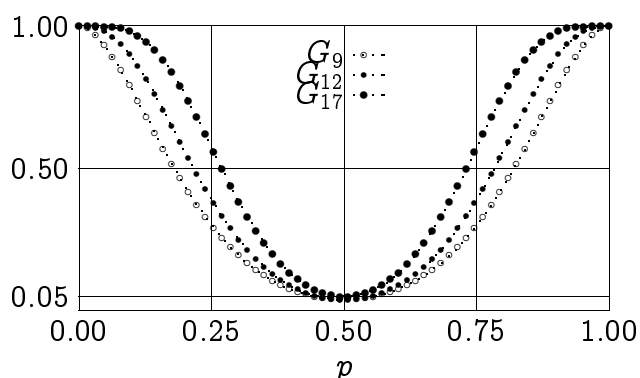
Erhöhung des Stichprobenumfangs



Durch Erhöhung des Stichprobenumfangs wird der α -Fehler kleiner, der β -Fehler dagegen größer.

Die Graphik zeigt die Gütefunktionen G_n des Münzwurftests zum kritischen Bereich $K = \{0, 1, 2, 3, n - 3, n - 2, n - 1, n\}$ und Stichprobenumfang $n = 10, 13, 16$.

Kontrolle des α -Fehlers



α -Fehler wiegen schwerer als β -Fehler. Deshalb gestaltet man Tests so, dass die Wahrscheinlichkeit für einen α -Fehler unterhalb eines vorgegebenen **Signifikanzniveaus** α bleibt (übliche Werte: $\alpha = 0.05, 0.01, 0.001$).

Die Graphik zeigt die Gütefunktionen G_n der Tests zum Niveau $\alpha = 0.05$ mit Stichprobenumfang $n = 9, 12, 17$ und kritischen Bereichen $K_9 = \{0, 1, 8, 9\}$, $K_{12} = \{0, 1, 2, 10, 11, 12\}$ bzw. $K_{17} = \{0, 1, 2, 3, 4, 13, 14, 15, 16, 17\}$.

Trennschärfe

Eine wünschenswerte Eigenschaft eines Tests ist die **Trennschärfe**. Darunter versteht man die Eigenschaft, dass ein Test (neben der Einhaltung der α -Schranke für den Fehler 1. Art) möglichst kleine Werte für die Wahrscheinlichkeit des Fehlers 2. Art aufweist. Damit “trennt” der Test die Hypothese H_0 möglichst “scharf” von den Alternativen zu H_0 .

Im Fall des Münzwurfs heißt dies, dass die Gütefunktion des Tests für $p \neq \frac{1}{2}$ möglichst große Werte annehmen soll.

Man kann zeigen, dass man “beliebig scharfe” Tests erhält, wenn der Stichprobenumfang entsprechend erhöht werden kann.

Einseitige Alternative H_1

Der soeben betrachtete Münzwurftest war symmetrisch in dem Sinn, daß als Alternative zur **Nullhypothese** $H_0: p = \frac{1}{2}$ (stillschweigend) die Hypothese $H_1: p \neq \frac{1}{2}$ betrachtet wurde.

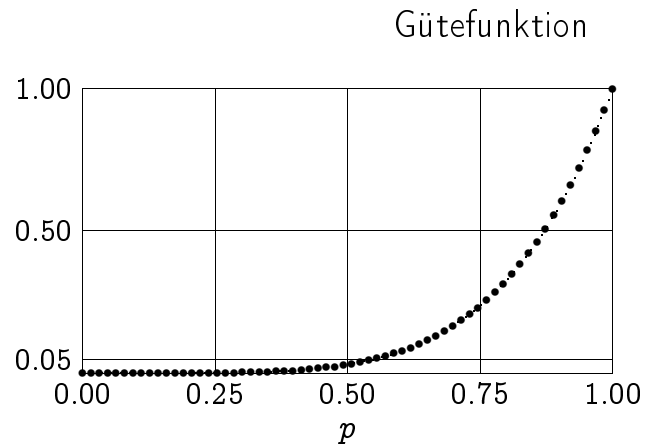
Es kann jedoch sinnvoll sein, die Alternative H_1 **einseitig** zu formulieren.

Beispiel 1. S_1 fordert S_2 zu einer Wette auf. S_1 setzt auf A einen gewissen Betrag, S_2 soll auf Z einen ebenso großen Betrag setzen. Die Münze gehört S_1 . S_2 hat Zweifel, ob die Münze fair ist und vermutet, dass A öfter fällt als Z .

S_2 erhält Gelegenheit für einen Test. S_2 nimmt sich vor, die Münze fünfmal zu werfen und sie zu beanstanden, wenn $AAAAA$ eintritt.

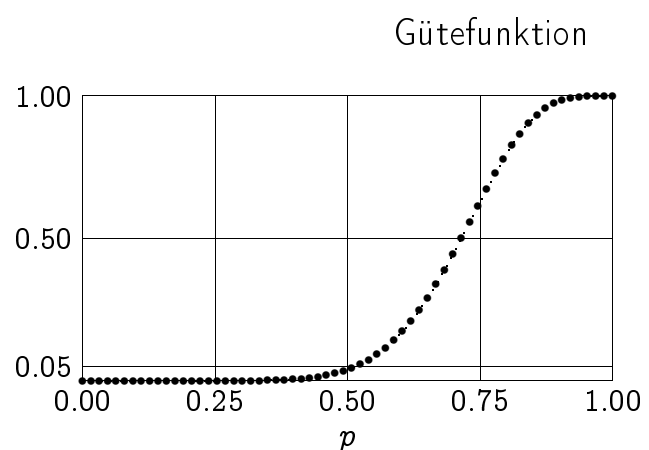
Formale Beschreibung des einseitigen Tests

- $H_0 : p \leq \frac{1}{2}, H_1 : p > \frac{1}{2}$,
- Testgröße
$$U(X_1, \dots, X_5) = \sum_{1 \leq i \leq 5} X_i,$$
- kritischer Bereich $K = \{5\}$.
- Signifikanzniveau $\alpha = 0.05$



Steigerung der Trennschärfe: $n = 16$

- $H_0 : p \leq \frac{1}{2}, H_1 : p > \frac{1}{2}$,
- Testvariable
$$U(X_1, \dots, X_{16}) = \sum_{i=1}^{16} X_i,$$
- kritischer Bereich
$$K = \{12, 13, 14, 15, 16\}.$$
- Signifikanzniveau $\alpha = 0.05$



Anwendung: Einseitiger Zeichentest

Beispiel 2. [Einseitiger Zeichentest] Zwei Lehrmethoden, A und B , sollen verglichen werden (angeblich ist B besser als A). Von 16 Zwillingspaaren soll dazu jeweils eine Person mit A , die andere mit B geschult und dann der Lehrerfolg anhand eines einheitlichen Tests festgestellt werden. Ist B besser als A , dann ist der Anteil p der Zwillingspaare, bei denen mit B ein besseres Resultat als mit A erzielt worden ist, größer als $\frac{1}{2}$.

Man testet $H_0 : p \leq \frac{1}{2}$ einseitig gegen $H_1 : p > \frac{1}{2}$ zum Niveau $\alpha = 0.05$.

Die Durchführung ergibt: 13 mal hat B zu einem besseren Resultat geführt als A . Man kann damit H_0 auf dem 5%-Niveau ablehnen und sagen, dass B besser als A ist.

Haben statistische Tests Beweiskraft?

- Eine möglicherweise wahre Hypothese über eine statistische Grundgesamtheit wird als H_1 -Hypothese, die Gegenhypothese als H_0 -Hypothese formuliert.
- Ein statistischer Test kann die Hypothese H_1 , auch wenn sie zutrifft, nicht beweisen und die Hypothese H_0 , auch wenn sie falsch ist, nicht widerlegen.
- Ähnlich wie indirekte mathematische Beweise zielen statistische Tests darauf ab, die möglicherweise falsche Hypothese H_0 zugunsten der Hypothese H_1 zu verwerfen, und zwar auf der Grundlage von Testresultaten, die der Hypothese H_0 widersprechen.
- Anstatt auf absolute Beweiskraft kann man sich nur auf eine kontrollierbar kleine (aber positive) **Irrtumswahrscheinlichkeit (Signifikanzniveau)** berufen.

Asymmetrie zwischen α - und β -Fehler

Statistische Test können versagen, nämlich

1. wenn H_0 zutrifft und trotzdem abgelehnt wird (α -Fehler),
2. wenn H_0 falsch ist und H_0 dennoch nicht abgelehnt wird (β -Fehler).

Bei keinem Test kann man die Wahrscheinlichkeit für α -Fehler und β -Fehler gleichzeitig minimieren. Daher muss man sich damit begnügen,

- durch ein vorgegebenes Signifikanzniveau den α -Fehler zu kontrollieren und
- den β -Fehler durch Steigerung der Trennschärfe zu verringern.

Struktur eines statistischen Tests

1. Über die (teilweise) unbekannte Verteilung einer statistischen Grundgesamtheit sind alternative Hypothesen H_0 (möglicherweise falsch) und H_1 (möglicherweise wahr) formuliert.
2. Stichprobenvariablen X_1, \dots, X_n zur Grundgesamtheit bilden eine Testgröße $T_n(X_1, \dots, X_n)$, deren theoretische Verteilungseigenschaften in Abhängigkeit von den Hypothesen bekannt sind.
3. Stichprobenumfang und kritischer Bereich für die Ablehnung von H_0 sind so festgelegt, dass eine vorgegebene Irrtumswahrscheinlichkeit für eine fälschliche Ablehnung von H_0 nicht überschritten wird.
4. Liegt der Wert von T_n bei Realisierung der Stichprobe im kritischen Bereich, wird H_0 zugunsten H_1 verworfen, sonst angenommen.

Beispiel: Test für den Anteilswert p (zweiseitig)

- Hypothesen: $H_0 : p = p_0$, $H_1 : p \neq p_0$,

- Testgröße:

$$T_n(X_1, \dots, X_n) = \sum_{1 \leq i \leq n} X_i$$

(mit $\mathcal{B}(1, p)$ -verteilten ZVn X_i).

- Untere und obere Grenzen A_u, A_o des Annahmebereichs für die $\mathcal{B}(n, p)$ -verteilte Testgröße T_n bei Test auf dem Niveau α :

$$A_u = \min \left\{ k \in \mathbf{N} \mid P_{p_0}(T_n \leq k) > \frac{\alpha}{2} \right\},$$

$$A_o = \min \left\{ k \in \mathbf{N} \mid P_{p_0}(T_n \leq k) \geq 1 - \frac{\alpha}{2} \right\}.$$

Korrektes statistisches Testen

1. Testverfahren, Hypothesen, Signifikanzniveau und Stichprobenumfang sind **vorab** festzulegen.
2. Kann durch das Testverfahrens H_0 nicht abgelehnt werden, ist es **unzulässig**
 - (a) andere Tests auszuprobieren,
 - (b) α nachträglich zu erhöhen,
 - (c) so lange Stichproben zu ziehen, "bis H_0 schließlich doch noch abgelehnt wird".