# Statistische Tests

## 1 Begriffe und Vorgehensweise

| Allgemein | Beispiel |
|---|---|
| 1. Problem formulieren: *Nullhypothese* $H_0$ festlegen | |
| 2. Alternativen bestimmen: *Alternativhypothese* $H_A$ | |
| 3. zu beobachtende Grösse festlegen: *Teststatistik* $T$ | |
| 4. Wahrscheinlichkeitsverteilung von $T$ unter $H_0$ bestimmen | |
| 5. Menge aller "extremen" Beobachtungen definieren: *Verwerfungsbereich* $K$ mit *Signifikanzniveau* $\alpha$ | |
| 6. Daten erheben, Wert von $T$ berechnen: $T = t$ | |
| 7. *P-Wert* berechnen | |
| 8. Entscheidung: $t$ im Verwerfungbereich: verwerfe $H_0$<br>$t$ im Annahmebereich: behalte $H_0$ bei | |

**Erläuterungen:**

1. Wir nehmen an, dass kein Effekt oder Unterschied vorhanden ist und versuchen Evidenz gegen diese Annahme, die *Nullhypothese* $H_0$, zu finden. Die Nullhypothese ist das zu überprüfende Modell und besteht meistens aus einer Verteilungsannahme und einer Aussage über einen Parameter.

2. Weil man Evidenz **gegen** und nicht für etwas sucht, entspricht die Alternativhypothese der *Arbeitshypothese*. Die Nullhypothese möchte man möglichst widerlegen.

3. Die Teststatistik basiert meistens auf einer Schätzung des Parameters, der in Null- und Alternativhypothese auftritt.

5. Das Signifikanzniveau $\alpha$ (significance level) ist gleich der Wahrscheinlichkeit, ein „extremes“ Resultat zu erhalten, unter der Annahme, dass $H_0$ stimmt. Je grösser also $\alpha$ gewählt wird, desto grösser ist der Verwerfungsbereich und umgekehrt. Üblich sind $\alpha = 5\%$ oder $\alpha = 1\%$. Das Komplement zum Verwerfungsbereich heisst *Annahmebereich.*

7. Der $P$-Wert (p-value) ist die Wahrscheinlichkeit, dass in einem neuen Versuch ein mindestens so extremes Resultat herauskommt, unter der Annahme, dass $H_0$ richtig ist. Der $P$-Wert ist **nicht** die Wahrscheinlichkeit, dass $H_0$ richtig ist. Einer Hypothese kann gar keine Wahrscheinlichkeit zugeordnet werden, sie ist entweder richtig oder falsch.

8. Der Wert von $T$ ist genau dann im Verwerfungsbereich, wenn der $P$-Wert$\leq \alpha$. In diesem Fall wird die Nullhypothese abgelehnt oder verworfen. $H_0$ wird als statistisch widerlegt betrachtet. Der Test ist (statistisch) signifikant.

   Wenn der $P$-Wert$> \alpha$ ist, dann liegt der Wert von $T$ im Annnahmebereich und die Daten sprechen zu wenig gegen $H_0$. $H_0$ kann nicht verworfen werden. Der Test ist nicht signifikant.

**Fehler 1. Art:** $H_0$ wird verworfen, obschon $H_0$ richtig wäre. Die Wahrscheinlichkeit eines Fehlers 1. Art ist $\alpha$ und wird auf 5% oder 1% festgelegt.

**Fehler 2. Art:** $H_0$ wird beibehalten, obschon $H_A$ stimmt. Die Wahrscheinlichkeit eines Fehlers 2. Art wird meist mit $\beta$ bezeichnet.

**Macht:** $1 - \beta$ ist die Wahrscheinlichkeit, eine wahre Alternativhypothese zu erkennen und heisst die *Macht* des Tests (power).

# 2 Tests für Lageparameter

## 2.1 z-Test

Seien $X_1, \ldots, X_n$ unabhängig normalverteilt mit Erwartungswert $\mu$ und bekannter Varianz $\sigma^2$. Betrachte die folgenden Hypothesen:

$H_0$: $\mu = \mu_0$
$H_A$: $\mu \neq \mu_0$.

Die Teststatistik des $z$-Tests ist:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}, \quad \text{unter } H_0 \text{ standardnormalverteilt.}$$

## 2.2 t-Test für eine Stichprobe

Seien $X_1, \ldots, X_n$ unabhängig normalverteilt mit Erwartungswert $\mu$ und unbekannter Varianz. Betrachte die folgenden Hypothesen:

$H_0$: $\mu = \mu_0$
$H_A$: $\mu \neq \mu_0$.

Die Teststatistik des $t$-Tests ist:

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}, \quad \text{unter } H_0 \text{ t-verteilt mit } n-1 \text{ Freiheitsgraden (degrees of freedom).}$$

## 2.3 Vorzeichentest (sign test)

Seien $X_1, \ldots, X_n$ iid mit Median $m$.

$H_0$: $m = \mu_0$
$H_A$: $m \neq \mu_0$.

Zur Berechnung der Teststatistik $T$ wird von jedem Wert $\mu_0$ subtrahiert. Dann ist $T$ die Anzahl positiver (oder negativer) Beobachtungen; $T \sim \mathcal{B}(n, 0.5)$.

### 2.4 Wilcoxontest (Rangsummentest, signed rank test)

Seien $X_1, \dots, X_n$ iid stetige, symmetrisch verteilte Zufallsvariablen mit Median $m$.

$H_0$: $m = \mu_0$
$H_A$: $m \neq \mu_0$.

Zuerst wird von allen Beobachtungen $\mu_0$ subtrahiert. Die Werte werden dann dem Absolutbetrag nach geordnet und die zugehörigen Ränge bestimmt. Sind mehrere Werte gleich gross, werden die Ränge gemittelt.
Die Teststatistik ist die Rangsumme aller positiven (oder negativen) Werte $T^+$ (oder $T^-$).
Es gibt Tabellen mit den kritischen Werten; für $n > 30$ Normalapproximation.

### 2.5 t-Test für zwei unabhängige Stichproben

Seien $X_1, \dots, X_n$ unabhängig normalverteilt mit Erwartungswert $\mu_X$ und Varianz $\sigma^2$ und $Y_1, \dots, Y_m$ unabhängig normalverteilt mit Erwartungswert $\mu_Y$ und Varianz $\sigma^2$. Die $Y_i$ seien unabhängig von den $X_i$.

$H_0$: $\mu_X = \mu_Y$
$H_A$: $\mu_X \neq \mu_Y$.

Die Teststatistik des $t$-Tests ist:

$$t = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \qquad \text{mit } S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

Unter $H_0$ ist t t-verteilt mit $n + m - 2$ Freiheitsgraden.

### 2.6 Mann-Whitney-Test (Rangsummentest, rank sum test)

Seien $X_1, \dots, X_n$ iid stetige Zufallsvariablen mit Erwartungswert $\mu_X$ und $Y_1, \dots, Y_m$ iid stetige Zufallsvariablen mit Erwartungswert $\mu_Y$. Die $Y_i$ seien unabhängig von den $X_i$.

$H_0$: $\mu_X = \mu_Y$
$H_A$: $\mu_X \neq \mu_Y$.

Bestimme die Rangsummen $T^{(1)}$ der $X_i$ und $T^{(2)}$ der $Y_i$ in der „gemeinsamen“ Stichprobe.
Setze

$$U^{(1)} = T^{(1)} - \frac{n(n+1)}{2}, \quad U^{(2)} = T^{(2)} - \frac{m(m+1)}{2} \qquad \text{und } U = \min(U^{(1)}, U^{(2)}).$$

Die kritischen Werte für $U$ sind tabelliert.

## 3 Dualität zwischen Tests und Vertrauensintervallen

Bei einem Test lautet die Frage: „Welche Beobachtungen sind vereinbar mit $H_0$, bzw. einem Parameterwert $\mu_0$?“ Der Annahmebereich liefert die Antwort.
Umgekehrt wird bei einem Vertrauensintervall gefragt: „Welche Parameter sind vereinbar mit den Beobachtungen?“ Alle diese Parameterwerte bilden dann das Vertrauensintervall.

**Satz 3.1 (Dualitätssatz)**
*Ein Test mit Signifikanzniveau $\alpha$ verwirft $H_0 : \mu = \mu_0$ genau dann nicht, wenn $\mu_0$ innerhalb des $(1-\alpha)100\%$-Vertrauensintervalls liegt.*