Statistik I für WInf und WI

Prof. Dr. Wilhelm Stannat

Inhalt:

Das vorliegende Skript ist eine Zusammenfassung des dritten Teils der Vorlesung Statistik I für WInf und WI, die im Wintersemester 2007/08 an der TU Darmstadt gehalten wurde. Die Lektüre des Skriptes ist kein gleichwertiger Ersatz für den Besuch der Vorlesung.

Korrekturen bitte per Email an: `stannat@mathematik.tu-darmstadt.de`

III Induktive Statistik

Im Gegensatz zur deskriptiven Statistik, die sich auf die Beschreibung von Daten anhand von Kennzahlen und Grafiken beschränkt, versucht die induktive (d.h. die schließende) Statistik von beobachteten Daten auf deren Verteilungen (oder Eigenschaften ihrer Verteilungen) zu schließen. Dies kann zum Beispiel dann notwendig sein, wenn eine vollständige Datenerhebung unmöglich, zu zeitaufwendig oder zu kostspielig ist, wie es etwa bei Umfragen der Fall ist.

In der schließenden Statistik gibt es im Wesentlichen drei zu bearbeitende Problemstellungen:

1. Konstruktion eines **Schätzers** für einen Parameter der unbekannten Verteilung
2. Berechnung von **Konfidenzintervallen**, d.h. von Schranken, die einen unbekannten Parameter mit vorgegebener Wahrscheinlichkeit einfangen.
3. Entwicklung **statistischer Tests**, mit denen vorgegebene Parameter auf Verträglichkeit mit Beobachtungen überprüft werden können.

1. Schätzen

Ausgangspunkt ist wieder eine Grundgesamtheit G von Merkmalsträgern. Unter einer **Stichprobenerhebung** versteht man eine zufällige Entnahme von endlich vielen Objekten aus G. Dabei bedeutet zufällig, dass für jedes Objekt die Wahrscheinlichkeit der Entnahme gleich ist.
In der Sprache der Wahrscheinlichkeitstheorie handelt es sich bei der Stichprobenerhebung um Zufallsexperimente, deren Ausgang man durch Zufallsvariablen

$$X_1, X_2, \ldots, X_n$$

beschreiben kann. In diesem Zusammenhang nennt man die X_i auch **Stichprobenvariablen**. In der Regel betrachtet man nur unabhängige Wiederholungen desselben Zufallsexperiments, d.h. also, dass die Stichprobenvariablen $X_1, \ldots, X_n$ stochastisch unabhängig und identisch verteilt sind.

Unter dem **Stichprobenergebnis** oder der **Stichprobenrealisation** versteht man dann das n-Tupel $(x_1, \ldots, x_n)$ der Realisierung von $X_1, \ldots, X_n$.
Eine **Punktschätzung** ist eine Funktion $g : \mathbb{R}^n \to \mathbb{R}$. Sie ordnet der Stichprobe $X_1, \ldots, X_n$ den Schätzwert $g(x_1, \ldots, x_n)$ zu.
Die zugehörige **Schätzfunktion** (oder auch **Statistik**) $g(X_1, \ldots, X_n)$ ist diejenige Zufallsvariable, die man durch Einsetzen der Stichprobenvariablen X_i für x_i in die Funktion g erhält.

Beispiele $X_1, \ldots, X_n$ mit Mittel m und Varianz σ^2

Schätzfunktion	Bezeichnung	Erwartungswert	Varianz
$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	Stichprobenmittel	m	$\frac{\sigma^2}{n}$
$\sqrt{n}\frac{\overline{X}-m}{\sigma}$	Gauß-Statistik	0	1
$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$	mittlere quadratische Abweichung	$\frac{n-1}{n}\sigma^2$	
$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$	Stichprobenvarianz	σ^2	
$S = \sqrt{S^2}$	Stichprobenstandard - abweichung		
$\sqrt{n}\frac{\overline{X}-m}{S}$	t-Statistik		

Im Folgenden wollen wir annehmen, dass die (unbekannte) Verteilung der **Stichprobenvariablen** aus einer Menge möglicher Verteilungen stammt, die über einen Parameter $\theta \in \Theta$ parametrisiert sind.

Beispiel X_i seien $N(m, \sigma^2)$-verteilt mit unbekanntem Mittel m und unbekannter Varianz σ^2. In diesem Falle ist also $\theta = (m, \sigma^2)$ aus $\Theta = \mathbb{R} \times]0, \infty[$ eine (mögliche) Parametrisierung der zugrundeliegenden Verteilungen.

Ist nun $T = g(X_1, \ldots X_n)$ ein Schätzer, so wird der Erwartungswert $E(T)$ abhängen von der Verteilung der Zufallsvariablen X_i. Um diese Abhängigkeit im folgenden kenntlich zu machen, schreiben wir $E_\theta(T)$ für $E(T)$, wenn die zu θ gehörende Verteilung die tatsächliche Verteilung der X_i ist.

Ein zu schätzender Parameter aus der Menge der zugrundeliegenden Verteilungen kann nun realisiert werden als Abbildung

$$\tau : \Theta \to \mathbb{R}.$$

Eigenschaften von Schätzern

Erwartungstreue

Ein Schätzer $T = g(X_1, \ldots, X_n)$ heißt **erwartungstreu** für den Parameter τ, falls

$$E_\theta(T) = E_\theta\left(g(X_1, \ldots, X_n)\right) = \tau(\theta)$$

für jedes $\theta \in \Theta$.

Mit anderen Worten: Bestimmt man den Erwartungswert von T unter der Voraussetzung, dass der Parameter θ zugrundeliegt, ergibt sich $\tau(\theta)$ als Erwartungswert.

Beispiele

(i) Das Stichprobenmittel $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ist ein erwartungstreuer Schätzer für das Mittel $m = E_\theta(X)$, denn

$$E_\theta(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E_\theta(X_i) = m\,.$$

(ii) Die mittlere quadratische Abweichung

$$T = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

ist kein erwartungstreuer Schätzer für die Varianz $\sigma^2 = E_\theta\left((X - E_\theta(X))^2\right)$, denn

$$\begin{aligned} E_\theta(T) &= \frac{1}{n}\sum_{i=1}^{n} E_\theta(X_i^2) - 2E_\theta(X_i\overline{X}) + E_\theta(\overline{X}^2) \\ &= E_\theta(X^2) - E_\theta\left(\overline{X}^2\right) = \frac{n-1}{n}\sigma^2\,, \end{aligned}$$

denn

$$E_\theta\left(\overline{X}^2\right) = \frac{1}{n^2}\sum_{i,j=1}^{n} E_\theta(X_iX_j) = \frac{n-1}{n}E_\theta(X)^2 + \frac{1}{n}E_\theta(X^2)\,.$$

Im Gegensatz hierzu ist die Stichprobenvarianz $S^2 = \frac{n}{n-1}T$ ein erwartungstreuer Schätzer für σ^2, denn

$$E_\theta\left(S^2\right) = \frac{n}{n-1}E_\theta(T) = \sigma^2\,.$$

Als Abschwächung der Erwartungstreue betrachtet man **asymptotische Erwartungstreue** bei wachsender Stichprobenlänge. Dazu nimmt man an, dass zu jeder Stichprobenlänge n ein Schätzer $T_n = g_n(x_1, \ldots, x_n)$ für $\tau(\theta)$ gegeben ist. Die Folge $T_1, T_2 \ldots$ heißt **asymptotisch erwartungstreu** (für τ), falls

$$\lim_{n\to\infty} E_\theta(T_n) = \tau(\theta)$$

für jedes $\theta \in \Theta$.

Beispiel Die mittlere quadratische Abweichung

$$T_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

ist asymptotisch erwartungstreu für die Varianz, denn

$$E_\theta(T_n) = \frac{n-1}{n}\,\mathsf{Var}_\theta(X) \to_{n\to\infty} \mathsf{Var}_\theta(X)\,.$$

Für einen nicht erwartungstreuen Schätzer T bezeichnet man die Abweichung

$$\mathsf{Bias}_\theta(T) := E_\theta(T) - \tau(\theta)$$

als **Bias** (oder **Verzerrung**).

Die **mittlere quadratische Abweichung**

$$MSE(T) := E_\theta\left((T-\tau(\theta))^2\right)$$

ist ein Maß für die Schätzgüte. MSE steht dabei für **mean squared error**.

Struktur des mittleren quadratischen Fehlers

$$MSE(T) = E_\theta((T-\tau(\theta))^2) = \mathsf{Var}_\theta(T) + \mathsf{Bias}\,(T)^2.$$

Beweis

$$\begin{aligned} E_\theta((T-\tau(\theta))^2) &= E_\theta(T^2) - 2E_\theta(T)\tau(\theta) + \tau(\theta)^2 \\ &= E_\theta(T^2) - (E_\theta(T))^2 + (E_\theta(T))^2 - 2E_\theta(T)\tau(\theta) + \tau(\theta)^2 \\ &= \mathsf{Var}_\theta(T) + (E_\theta(T) - \tau(\theta))^2\,. \end{aligned}$$

Es sei $T_1, T_2, \ldots$ wieder eine Folge von Schätzern für $\tau(\theta)$. Dann heißt diese Folge

- **konsistent im quadratischen Mittel**, falls

$$\lim_{n\to\infty} E_\theta\left((T_n-\tau(\theta))^2\right) = 0$$

- **schwach konsistent**, falls

$$\lim_{n\to\infty} P_\theta\left(|T_n-\tau(\theta)| \geq \varepsilon\right) = 0 \qquad \forall\, \varepsilon > 0\,.$$

Aufgrund der Ungleichung von Tschebychev ist klar, dass aus Konsistenz im quadratischen Mittel immer schwache Konsistenz folgt, denn

$$P_\theta\left(|T_n-\tau(\theta)| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2} E_\theta\left((T_n-\tau(\theta))^2\right) \to 0 \qquad \text{für } n\to\infty\,.$$

Beispiel Das Stichprobenmittel $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ ist konsistent im quadratischen Mittel für das Mittel $m = E_\theta(X)$ (und damit auch schwach konsistent), denn

$$E_\theta((\overline{X}-m)^2) = \frac{1}{n}\,\mathsf{Var}_\theta(X) \to 0 \text{ für } n\to\infty\,.$$

Effizienz von Schätzern

Der mittlere quadratische Fehler eines Schätzers liefert ein Vergleichskriterium zwischen den verschiedenen Schätzern für τ. Offensichtlich ist von zwei Schätzern T_1, T_2 mit

$$MSE\,(T_1) \leq MSE(T_2)$$

der Schätzer T_1 mit der kleineren mittleren quadratischen Abweichung **wirksamer** für die Schätzung von $\tau(\theta)$.

Beschränkt man sich beim Vergleich zweier Schätzer auf erwartungstreue Schätzer, also Schätzer mit

$$\text{Bias}\,(T_i) = 0 \text{ und damit } MSE(T_i) = \text{Var}(T_i)\,,$$

so reduziert sich der Vergleich der mittleren quadratischen Abweichung auf den Vergleich der Varianzen.

Sind T_1, T_2 zwei erwartungstreue Schätzer für $\tau(\theta)$, so heißt T_1 **effizienter** (bzw. **wirksamer**), falls

$$\text{Var}(T_1) \leq \text{Var}(T_2).$$

Bemerkung (Cramèr-Rao Schranke) Die Varianz eines erwartungstreuen Schätzers kann nicht beliebig klein werden, sondern wird nach unten beschränkt durch die "Cramér-Rao Schranke". Wir wollen diese Schranke hier nicht angeben, sondern nur bemerken, dass sie von der Variation der Verteilungen in Abhängigkeit von θ abhängt. Ein erwartungstreuer Schätzer, der diese untere Schranke annimmt, heißt **effizient** (oder **wirksamst**).

Beispiele für effiziente Schätzer

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ für den Erwartungswert, wenn man

- alle Verteilungen mit endlicher Varianz zulässt
- alle Normalverteilungen zulässt.

Prinzipien zur Konstruktion von Schätzern

(a) Maximum Likelihood Schätzer

Es seien $X_1, \ldots, X_n$ zunächst diskret verteilt und

$$f(x_1, \ldots, x_n \mid \theta) = P_\theta(X_1 = x_1, \ldots, X_n = x_n)$$

die Wahrscheinlichkeitsfunktion zur gemeinsamen Verteilung der Stichprobenvariablen bei zugrundeliegender Verteilung zum Parameter θ. Zu gegebener Stichprobe $x_1, \ldots, x_n$ heißt die Funktion

$$L \,:\, \theta \longmapsto f(x_1, \ldots, x_n \mid \theta)\,,\ \theta \in \Theta,$$

die **Likelihoodfunktion**, denn sie gibt an, wie wahrscheinlich die gewonnene Stichprobe $x_1, \ldots, x_n$ bei angenommener zugrundeliegender Verteilung zum Parameter θ ist.

Die Grundidee der **Maximum Likelihood Schätzung** besteht darin, als Schätzer für θ gerade denjenigen Parameter $\widehat{\theta}$ zu wählen, für den die gewonnene Stichprobe am **wahrscheinlichsten** ist, also $\widehat{\theta}$ mit

$$L(\widehat{\theta}) = \max_{\theta\in\Theta} L(\theta) = \max_{\theta\in\Theta} f(x_1, \ldots, x_n \,|\, \theta)\,.$$

Der Einfachheit halber betrachten wir im folgenden nur unabhängig und identisch verteilte Stichprobenvariablen. Dann bekommt die Likelihoodfunktion die Produktgestalt

$$L(\theta) = f(x_1, \ldots, x_n \,|\, \theta) = f(x_1 \,|\, \theta) \cdot \ldots \cdot f(x_n \,|\, \theta) \tag{3.1}$$

mit $f(x \,|\, \theta) = P_\theta(X = x)$.

Sind die Stichprobenvariablen zu $\theta \in \Theta$ stetig verteilt mit Dichte $f(x \mid \theta)$, so ersetzt man in der Likelihoodfunktion (3.1) die Wahrscheinlichkeitsfunktion der Verteilung durch die entsprechende Dichte.

Bemerkung Über Existenz (und Eindeutigkeit) des Maximums der Likelihoodfunktion wird hier keine Aussage gemacht! Insbesondere muss i.a. der Maximum-Likelihood Schätzer nicht existieren, oder er muss nicht eindeutig bestimmt sein.

Die Bestimmung der Maximum-Likelihood Schätzung erfolgt in der Regel durch Nullsetzen der Ableitung der Likelihoodfunktion L. Wegen der Produktgestalt von L in (3.1) ist es zweckmäßig, L zunächst zu logarithmieren:

$$\ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i \mid \theta) \tag{3.2}$$

und dann zu maximieren. $\ln L$ heißt **Log-Likelihood Funktion**.

Beispiele

(a) **Bernoulli-Experiment**

$$S_n = X_1 + \ldots + X_n$$

sei die Anzahl der Erfolge in einem Bernoulli-Experiment der Länge n bei unbekanntem Erfolgsparameter p. Die Likelihoodfunktion hat die Form

$$L(p) = \binom{n}{S_n} p^{S_n}(1-p)^{n-S_n}, p \in [0,1],$$

wobei S_n die beobachtete Anzahl der Erfolge ist. In diesem Falle ist $\widehat{p} = \frac{S_n}{n}$ das eindeutig bestimmte Maximum, also

$$\widehat{p} = \frac{S_n}{n}$$

die (eindeutig bestimmte) Maximum-Likelihood Schätzung für p.

Insbesondere Die Maximum-Likelihood Schätzung $\widehat{p} = \frac{S_n}{n} = \frac{1}{n}(X_1 + \ldots + X_n)$ ist gerade das Stichprobenmittel!

(b) **Normalverteilung**

$X_1, \ldots, X_n$ unabhängig $N(m, \sigma^2)$-verteilt, also $\theta = (m, \sigma)$, und die zugehörige Likelihoodfunktion hat die Gestalt.

$$L(m,\sigma) = \Pi_{i=1}^{n} f_{m,\sigma^2}(x_i) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-m)^2}{\sigma^2}\right)$$

Logarithmieren ergibt

$$\ln L(m,\sigma) = -n\ln\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-m)^2}{\sigma^2}$$

mit partiellen Ableitungen

$$\frac{\partial \ln L}{\partial m}(m,\sigma) = \sum_{i=1}^{n} \frac{x_i - m}{\sigma^2}$$
$$\frac{\partial \ln L}{\partial \sigma}(m,\sigma) = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i - m)^2}{\sigma^3}\,.$$

Nullsetzen der partiellen Ableitungen liefert die Maximum-Likelihood Schätzung

$$\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{und} \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{m})^2\,. \tag{3.3}$$

Hier hat man schließlich noch zu überprüfen, dass (3.3) tatsächlich (eindeutig bestimmtes) Maximum der Likelihoodfunktion ist.

Insbesondere Die Maximum-Likelihood Schätzung m für $\widehat{m}$ entspricht dem Stichprobenmittel, diejenige für σ^2 der mittleren quadratischen Abweichung S^2.

(b) **Kleinste Quadrate Schätzung**

Ein weiteres Prinzip der Parameterschätzung besteht in der Minimierung der Summe der quadratischen Abweichungen zwischen Beobachtungswert und geschätzem Wert. Dies haben wir bereits bei der Regression kennengelernt.

Beispiel Arithmetisches Mittel

$$\min_{m\in\mathbb{R}} \sum_{i=1}^{n}(x_i - m)^2$$

führt wieder auf das Stichprobenmittel

$$\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Intervallschätzungen

Die bisher konstruierten Schätzer liefern zu gegebenen Beobachtungen $x_1, \ldots, x_n$ eine Schätzung $g(x_1, \ldots, x_n)$ für den unbekannten Parameter $\tau(\theta)$. Daher spricht man auch von Punktschätzungen. In den seltensten Fällen wird die Schätzung exakt mit $\tau(\theta)$ übereinstimmen, sondern bestenfalls "in der Nähe" liegen.

Daher ist es zweckmäßiger, zu gegebener Beobachtung ein Intervall

$$I(x_1, \ldots, x_n) = [U(x_1, \ldots, x_n),\, O(x_1, \ldots, x_n)]$$

anzugeben, in dem der wahre Parameter $\tau(\theta)$ mit einer vorgegebenen Wahrscheinlichkeit $1-\alpha$ liegt, also:

$$P_\theta\left(\tau(\theta) \in [U(x_1, \ldots, x_n),\, O(x_1, \ldots, x_n)]\right) \geq 1-\alpha \text{ für alle } \theta \in \Theta\,.$$

$1-\alpha$ heißt **Konfidenzniveau**, das Intervall $I(x_1, \ldots, x_n)$ **Konfidenzintervall** für $\tau(\theta)$ (zum Niveau $1-\alpha$).

Konfidenzintervalle für unabhängige normalverteilte Stichprobenvariablen

Es seien $(X_1, \ldots, X_n)$ unabhängig $N(m, \sigma^2)$-verteilt.

(i) **Konfidenzintervall für m bei bekannter Varianz $\sigma^2 = \sigma_0^2$**

$$\Theta = \{(m, \sigma_0) : m \in \mathbb{R}\},\ \tau(m, \sigma_0) = m\,.$$

Eine Punktschätzung für τ ist das Stichprobenmittel

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,.$$

Die zugehörige Schätzfunktion

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

ist bei zugrundeliegendem Parameter $\theta = (m, \sigma_0)$ $N(m, \frac{\sigma_0^2}{n})$-verteilt und damit ist die zugehörige Gauß-Statistik

$$\overline{Y} = \sqrt{n}\frac{\overline{X} - m}{\sigma_0} \qquad N(0,1) - \text{verteilt.} \tag{3.4}$$

Zu gegebenem Konfidenzniveau $1 - \alpha$ ist also

$$P\left(-z_{1-\frac{\alpha}{2}} \le \overline{Y} \le z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha\,,$$

wobei z_p das p-Quantil der Standard-Normalverteilung bezeichnet, denn

$$P\left(-z_{1-\frac{\alpha}{2}} \le \overline{Y} \le z_{1-\frac{\alpha}{2}}\right) = \Phi\left(z_{1-\frac{\alpha}{2}}\right) - \Phi(-z_{1-\frac{\alpha}{2}}) = 2\underbrace{\Phi\left(z_{1-\frac{\alpha}{2}}\right)}_{=1-\frac{\alpha}{2}} - 1 = 1 - \alpha\,.$$

Das zugehörige Konfidenzintervall hat also die Form

$$I(x_1, \ldots, x_n) = \left[\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\,,\ \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\right]\,.$$

(ii) **Konfidenzintervall für m bei unbekannter Varianz σ^2**

$$\Theta = \{(m, \sigma) : m \in \mathbb{R}, \sigma > 0\}\,, \tau(m, \sigma) = m\,.$$

Bei unbekannter Varianz σ^2 muss diese erst anhand der Stichprobe $x_1, \ldots, x_n$ geschätzt werden. Dafür bietet sich die Stichprobenvarianz an:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

mit zugehöriger Schätzfunktion

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\,.$$

Einsetzen in (3.4) liefert als Schätzfunktion

$$\sqrt{n}\frac{\overline{X}-m}{S}$$

und diese ist gerade t_{n-1}-verteilt. Zu gegebenem Konfidenzniveau $1-\alpha$ ist also

$$P\left(-t_{n-1,1-\frac{\alpha}{2}} \leq \sqrt{n}\frac{\overline{X}-m}{S} \leq t_{n-1,1-\frac{\alpha}{2}}\right) = 1-\alpha\,,$$

wobei $t_{n-1,1-\frac{\alpha}{2}}$ das $1-\frac{\alpha}{2}$-Quantil der t-Verteilung mit $n-1$-Freiheitsgraden bezeichnet. Das zugehörige Konfidenzintervall hat also die Form

$$I(X_1,\ldots,X_n) = \left[\overline{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\,,\ \overline{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right]\,.$$

(iii) **Konfidenzintervall für σ^2**

$$\Theta = \{(m,\sigma) : \sigma > 0\}\,,\ \tau(m,\sigma) = \sigma^2.$$

Eine Punktschätzung für die Varianz σ^2 ist die Stichprobenvarianz

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

mit zugehöriger Schätzfunktion

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\,.$$

Da $X_1,\ldots,X_n$ unabhängig $N(m,\sigma^2)$-verteilt, also $Y_i = \frac{X_i-m}{\sigma}$ unabhängig $N(0,1)$-verteilt, folgt, dass

$$\frac{n-1}{\sigma^2}S^2 = \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 = \sum_{i=1}^{n}(Y_i-\overline{Y})^2$$

χ^2_{n-1}-verteilt ist. Zu gegebenem Niveau $1-\alpha$ ist also

$$P\left(\chi^2_{n-1,\frac{\alpha}{2}} \leq \frac{n-1}{\sigma^2}S^2 \leq \chi^2_{n-1,1-\frac{\alpha}{2}}\right) = 1-\alpha\,,$$

wobei $\chi^2_{n-1,p}$ das p-Quantil der χ^2_{n-1}-Verteilung bezeichnet, denn

$$P\left(\chi^2_{n-1,1-\frac{\alpha}{2}} \leq \frac{n-1}{\sigma^2}S^2 \leq \chi^2_{n-1,\frac{\alpha}{2}}\right) = F_{\chi^2_{n-1}}\left(1-\frac{\alpha}{2}\right) - F_{\chi^2_{n-1}}\left(\frac{\alpha}{2}\right) = 1-\alpha\,.$$

Es ist also

$$I(X_1,\ldots,X_n) = \left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\,,\ \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right]$$

Konfidenzintervall zum Niveau $1-\alpha$.

Einschub (zur χ^2-Verteilung)

Bemerkung Ihre Bedeutung erhält die χ^2-Verteilung in der induktiven Statistik durch folgende Beobachtung: Sind $X_1, \ldots, X_n$ unabhängig $N(0,1)$-verteilt, so gilt für die Stichprobenvarianz

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

(mit $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$)), dass $(n-1)S^2$ χ^2_{n-1}-verteilt ist.

Bemerkung Ist die Normalverteilungsannahme an die Stichprobenvariablen $X_1, \ldots, X_n$ nicht gerechtfertigt, so kann man unter Ausnutzung des zentralen Grenzwertsatzes eine Normalapproximation für die standardisierte Summe $\sqrt{n}\frac{\overline{X}-m}{\sigma}$ betrachten (siehe Bemerkung zu Konfidenzintervallen in Abschnitt 2.4).

Zum Abschluss noch der wichtige Spezialfall von unabhängig Bernoulli-verteilten Stichprobenvariablen

$$X_1, \ldots, X_n \,.$$

In diesem Falle ist die Summe

$$S_n = X_1 + \ldots + X_n$$

Bin (n,p)-verteilt, bei unbekannter Erfolgswahrscheinlichkeit p. Nach dem zentralen Grenzwertsatz ist

$$S_n^* := \frac{S_n - np}{\sqrt{np(1-p)}} = \sqrt{n}\frac{\overline{X} - p}{\sqrt{p(1-p)}}$$

näherungsweise $N(0,1)$-verteilt, also

$$P\left(-z_{1-\frac{\alpha}{2}} \leq S_n^* \leq z_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha \,.$$

Auflösen der Ungleichungen

$$-z_{1-\frac{\alpha}{2}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq z_{1-\frac{\alpha}{2}}$$

nach p liefert

$$\overline{X} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\overline{X}(1-\overline{X})}{n}} \leq p \leq \overline{X} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\overline{X}(1-\overline{X})}{n}} \,,$$

also ist

$$I(X_1, \ldots, X_n) = \left[\overline{X} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\overline{X}(1-\overline{X})}{n}} \,,\, \overline{X} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right]$$

ein (approximatives) Konfidenzintervall für p zum Niveau $1-\alpha$.

Beispiel zur Illustration In einem Warenposten aus DVD-Scheiben soll der Anteil der defekten Scheiben geschätzt werden. Dazu wird eine Stichprobe von 200 DVD-Scheiben überprüft. Angenommen, es werden dabei 6 defekte Scheiben gefunden, so ergibt sich für die Ausschusswahrscheinlichkeit zum Niveau 0.95, also $\alpha = 0.05$, das approximative Konfidenzintervall

$$[0.0063, 0.0537] .$$

Mit einer Wahrscheinlichkeit von 95% liegt also der tatsächliche Anteil der defekten DVD-Scheiben im getesteten Warenposten zwischen 0.6 Prozent und 5.37 Prozent.

2. Testen

Ein zentrales Problem der Statistik ist die Frage, wie eine Vermutung über eine Eigenschaft der Verteilung einer Grundgesamtheit anhand einer Stichprobe überprüft werden kann.
Eine solche Vermutung bezeichnet man als **Nullhypothese** H_0. Ein **statistischer Test** ist dann zunächst einmal eine **Entscheidungsregel**

$$\varphi(x_1 \ldots, x_n) \in \{0, 1\}$$

die als Funktion der n Beobachtungen $x_1 \ldots, x_n$ die Nullhypothese H_0 annimmt $(\varphi(x_1 \ldots, x_n) = 0)$ oder verwirft $(\varphi(x_1 \ldots, x_n) = 1)$.
Demnach ist ein Test durch seinen **Verwerfungsbereich** (oder auch **kritischer Bereich**), also durch die Menge

$$K = \{(x_1 \ldots, x_n) : \varphi(x_1 \ldots, x_n) = 1\}$$

eindeutig bestimmt.

Beispiel Wir betrachten wieder das Beispiel der Warenposten aus DVD-Scheiben. Als Vermutung über den Anteil der defekten DVD-Scheiben soll die Nullhypothese

$$H_0\text{: Anteil der defekten DVD-Scheiben beträgt } 10\%$$

mit Hilfe eines statistischen Tests anhand einer Stichprobe von $n = 100$ DVD-Scheiben überprüft werden.
In diesem Fall wird man den Verwerfungsbereich mit Hilfe einer kritischen Schranke c definieren, ab der man sagt: Ist die beobachtete Anzahl $S_{100} = \sum_{i=1}^{100} X_i \geq c$, so wird die Nullhypothese verworfen.
Es kann nun allerdings vorkommen, dass die Hypothese in Wahrheit zutrifft, aber aufgrund der getroffenen Entscheidungsregel verworfen wird, da die beobachtete Anzahl s_n der defekten DVD-Scheiben die kritische Schranke übersteigt **(Fehler 1. Art)**. Die Wahrscheinlichkeit für eine solche fälschliche Ablehnung von H_0 soll möglichst klein sein. Dazu gibt man sich ein **Niveau** α vor (etwa $\alpha = 0.05$) und bestimmt die kritische Schranke c so, dass die Wahrscheinlichkeit für eine fälschliche Ablehnung der Hypothese maximal α ist.
Jetzt könnte man natürlich c so wählen, dass die Wahrscheinlichkeit eines Fehlers 1. Art Null ist (einfach: Hypothese immer annehmen!). Dann wird der statistische Test aber sinnlos, da nicht mehr zwischen "guter" und "schlechter" Warenprobe unterschieden wird. Deshalb wählt man c minimal, um damit die Wahrscheinlichkeit dafür, die Nullhypothese zu verwerfen, wenn sie tatsächlich nicht zutrifft, zu maximieren. Diese Wahrscheinlichkeit nennt man die **Macht** des statistischen Tests. Die Komplementärwahrscheinlichkeit hierzu, d.h. die Wahrscheinlichkeit dafür, die Hypothese anzunehmen, obwohl sie in Wahrheit nicht zutrifft, heißt **Fehler 2. Art**.

Die möglichen Ausgänge eines statistischen Tests im Überblick:

	Entscheidung	
	für H_0	gegen H_0
H_0 wahr	richtig	falsch Fehler 1.Art
H_0 falsch	falsch Fehler 2.Art	richtig

- **Niveau** = Wahrscheinlichkeit für einen Fehler 1. Art
- **Macht** = Komplementärwahrscheinlichkeit für einen Fehler 2. Art

Ein **Signifikanztest** zum **Signifikanzniveau** α, $0 < \alpha < 1$, ist ein statistischer Test zum Niveau α, d.h. ein Test mit

$$P(\text{ Fehler 1. Art }) \leq \alpha\,.$$

Im Beispiel geht man also wie folgt vor: Zu α wähle c minimal mit

$$P_{0.1}(S_{100} \geq c) \leq \alpha\,.$$

Hierbei deutet der Index 0.1 an, dass S_{100} unter P Bin $(100, 0.1)$-verteilt ist. Normalapproximation für

$$S^*_{100} = \frac{S_{100} - 100 \cdot 0.1}{\sqrt{100 \cdot 0.1(1-0.1)}} = \frac{S_{100} - 10}{3}$$

ergibt

$$P_{0.1}(S_{100} \geqslant c) = P_{0.1}\left(S^*_{100} \geqslant \frac{c-10}{3}\right) \approx 1 - \Phi\left(\frac{c-10}{3}\right)\,.$$

Also ist c minimal zu wählen mit

$$\Phi\left(\frac{c-10}{3}\right) = 1-\alpha \text{ und das liefert } c = 3z_{1-\alpha} + 10\,.$$

Allgemein: Approximativer Binomialtest ("Gut - Schlecht" Prüfung)

Gegeben sei die Summe

$$S_n = X_1 + \ldots + X_n$$

von n unabhängig Bernoulli-verteilten Zufallsvariablen X_i mit unbekanntem Parameter p und die Nullhypothese

$$H_0 : p = p_0$$

zu fest gewähltem Parameter $p_0 \in [0, 1]$.

Zu gegebenem Niveau α bestimme man dann die kritische Schranke

$$c = \sqrt{np_0(1-p_0)}z_{1-\alpha} + np_0\,.$$

Dann ist die Hypothese zu verwerfen, falls die Stichprobensumme $s_n = \sum_{i=1}^n x_i$ größer als c ist.

Bemerkung (zweiseitiger approximativer Binomialtest) In Wahrheit haben wir bei obigem Test nur getestet, ob der Anteil der defekten DVD-Scheiben gleich 10% ist, wenn der unbekannte Parameter p aus der Menge $[0.1, 1]$ stammt. Ist allerdings auch $p < 0.1$ möglich, so setzt sich der Verwerfungsbereich aus einer unteren kritischen Schranke c_u und einer oberen kritischen Schranke c_o zusammen:

$$K = \{(x_1, \ldots, x_n) : s_u \leq c_u\} \cup \{(x_1, \ldots, x_n) : s_n \geq c_o\}$$

und man spricht von einem **zweiseitigen Ablehnungsbereich**.

Zweckmäßigerweise wählt man dann zu gegebenem Niveau α

$$c_u = -\sqrt{np_0(1-p_0)}z_{1-\frac{\alpha}{2}} + np_0$$
$$c_o = \sqrt{np_0(1-p_0)}z_{1-\frac{\alpha}{2}} + np_0\,,$$

d.h., die Nullhypothese wird verworfen, wenn die Stichprobensumme kleiner gleich c_u oder größer gleich c_o ist, oder in Größen von S^*_{100}, falls $|S^*_{100}| > z_{1-\frac{\alpha}{2}}$.

Der **approximative Binomialtest** im Überblick:

Test auf den Parameter p einer Binomialverteilung

Annahme $X_1, \ldots, X_n$ unabhängig Bernoulli-verteilt, also $S_n = X_1 + \ldots + X_n$ binomialverteilt.

Hypothese

(a) $H_0 : p = p_0$ (b) $H_0 : p \leq p_0$ (c) $H_0 : p \geq p_0$

Entscheidungsregel Betrachte als Testgröße

$$T(X_1, \ldots, X_n) = \frac{S_n - np_0}{\sqrt{np_0(1-p_0)}} = \sqrt{n}\frac{\overline{X} - p_0}{\sqrt{p_0(1-p_0)}} \qquad \text{(approx. N(0,1)-verteilt, falls } p = p_0)$$

Hierbei ist $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$ das Stichprobenmittel.
Ablehnung, falls

(a) $|T| > z_{1-\frac{\alpha}{2}}$ (b) $T > z_{1-\alpha}$ (c) $T < -z_{1-\alpha}$.

Will man die Annahme an die Verteilung der Stichprobenvariablen fallenlassen, muss man sich im allgemeinen auf das Testen einiger weniger Kennzahlen beschränken.

Gauß-Test

Test auf das Mittel m einer Verteilung bei bekannter Varianz

Annahme $X_1, \ldots, X_n$ unabhängig, identisch verteilt mit bekannter Varianz $\text{Var}(X_i) = \sigma_0^2$ und $X_i \sim N(m, \sigma_0^2)$ oder bei $n \geq 30$ X_i beliebig verteilt mit $E(X_i) = m$

Hypothese

(a) $H_0 : m = m_0$ (b) $H_0 : m \leq m_0$ (c) $H_0 : m \geq m_0$

Entscheidungsregel Betrachte als Testgröße

$$T(X_1, \ldots, X_n) = \sqrt{n}\frac{\overline{X} - m_0}{\sigma_0} \qquad \text{(approx. N(0,1)-verteilt, falls } m = m_0)$$

wobei

$$\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$$

das Stichprobenmittel bezeichnet.

Ablehnung, falls

(a) $|T| > z_{1-\frac{\alpha}{2}}$ (b) $T > z_{1-\alpha}$ (c) $T < -z_{1-\alpha}$.

Der Rest dieses Abschnittes dient der Übersicht einiger wichtiger Testprobleme. Dabei wird danach unterschieden, ob es sich um Ein-Stichproben oder Mehr-Stichproben Tests handelt.

Ein-Stichproben Tests

t-Test

Test auf das Mittel m einer Verteilung mit σ^2 unbekannt

Annahme $X_1, \ldots, X_n$ unabhängig, identisch verteilt mit
$X_i \sim N(m, \sigma^2)$ bzw. bei $n \geq 30$ beliebig verteilt mit $E(X_i) = m$ und Var$(X_i) = \sigma^2$

Hypothese
(a) $H_0 : m = m_0$ (b) $H_0 : m \leq m_0$ (c) $H_0 : m \geq m_0$

Entscheidungsregel Betrachte als Testgröße

$$T(X_1, \ldots, X_n) = \sqrt{n}\frac{\overline{X} - m_0}{S} \quad (\text{ approx. } t_{n-1} - \text{verteilt, falls } m = m_0)\,.$$

Hierbei ist

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ die Stichprobenvarianz.}$$

Ablehnung, falls
(a) $|T| > t_{n-1,1-\frac{\alpha}{2}}$ (b) $T > t_{n-1,1-\alpha}$ (c) $T < -t_{n-1,1-\alpha}$.

Für $n \geq 30$ kann man die Quantile der t-Verteilung durch die entsprechenden Quantile der Standardnormalverteilung ersetzen.

χ^2-Test für die Varianz

Annahme $X_1, \ldots, X_n$ unabhängig $N(m, \sigma^2)$-verteilt, m unbekannt

Hypothese
(a) $H_0 : \sigma^2 = \sigma_0^2$ (b) $H_0 : \sigma^2 \leq \sigma_0^2$ (c) $H_0 : \sigma^2 \geq \sigma_0^2$

Entscheidungsregel Betrachte als Testgröße

$$T(X_1, \ldots, X_n) = \frac{n-1}{\sigma_0^2}S^2 \quad (\,\chi^2_{n-1} - \text{verteilt, falls } \sigma^2 = \sigma_0^2).$$

Ablehnung, falls
(a) $T < \chi^2_{n-1,\frac{\alpha}{2}}$ oder $T > \chi^2_{n-1,1-\frac{\alpha}{2}}$ (b) $T > \chi^2_{n-1,1-\alpha}$ (c) $T < \chi^2_{n-1,\alpha}$.

Mehr-Stichproben Tests

Bei mehr-Stichproben Tests sollen Zusammenhänge mehrerer unabhängiger Stichproben mit möglicherweise verschiedenen Längen

$$\begin{array}{c} X_1^{(1)}, \ldots, X_{n_1}^{(1)} \\ X_1^{(2)}, \ldots, X_{n_2}^{(2)} \\ \vdots \quad \vdots \quad \vdots \\ X_1^{(k)}, \ldots, X_{n_k}^{(k)} \end{array}$$

getestet werden. Die zentrale Frage in diesem Zusammenhang ist dann die nach der Gleichheit der zugrundeliegenden Verteilungen bzw. nach der Gleichheit gewisser Kennzahlen der zugrundeliegenden Verteilungen.

Zwei-Stichproben Gauß-Test

Test auf Gleichheit der Mittel m_X und m_Y zweier Verteilungen bei bekannten Varianzen

Annahme $X_1, \ldots, X_n$ unabhängig, identisch verteilt
$Y_1, \ldots, Y_m$ unabhängig, identisch verteilt mit
$X_i \sim N(m_X, \sigma_X^2)$, $Y_i \sim N(m_Y, \sigma_Y^2)$-verteilt
oder
X_i, Y_j mit beliebiger (stetiger) Verteilung

$$E(X_i) = m_X\,, \mathsf{Var}(X_i) = \sigma_X^2\,, E(Y_j) = m_Y\,, \mathsf{Var}(Y_j) = \sigma_Y^2 \qquad \text{und } n, m \geq 30\,.$$

In beiden Fällen seien σ_X^2 und σ_Y^2 bekannt.

(a) $H_0 : m_X = m_Y$ (b) $H_0 : m_X \leq m_Y$ (c) $H_0 : m_X \geq m_Y$

Entscheidungsregel Betrachte als Testgröße

$$T(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \qquad \text{approx. N(0,1)-verteilt, falls } m_X = m_Y\,.$$

Ablehnung, falls
(a) $|T| > z_{1-\frac{\alpha}{2}}$ (b) $T > z_{1-\alpha}$ (c) $T < -z_{1-\alpha}$.

Bei unbekannten Varianzen verwendet man

Zweistichproben t-Test

Test auf Gleichheit der Mittel m_X und m_Y zweier Verteilungen bei unbekannten Varianzen

Annahme $X_1, \ldots, X_n$ unabhängig $N(m_X, \sigma_X^2)$-verteilt
$Y_1, \ldots, Y_m$ unabhängig $N(m_Y, \sigma_Y^2)$-verteilt
$\sigma_X^2 = \sigma_Y^2$ unbekannt.

Hypothesen wie im zwei-Stichproben Gauß-Test

Entscheidungsregel Betrachte als Testgröße

$$T(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = \sqrt{\frac{nm(n+m-2)}{n+m}} \frac{\bar{X} - \bar{Y}}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}}$$

mit

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \text{ und } S_Y^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y_i - \bar{Y})^2\,.$$

(t_{n+m-2}-verteilt, falls $m_X = m_Y$.)

Ablehnung, falls

(a) $|T| > t_{n+m-2,1-\frac{\alpha}{2}}$ (b) $T > t_{n+m-2,1-\alpha}$ (c) $T < -t_{n+m-2,1-\alpha}$.

Für eine Erweiterung des zwei-Stichproben t-Tests auf den Fall ungleicher Varianzen siehe [2].

F-Test

Test auf Gleichheit der Varianzen zweier Normalverteilungen

Annahme $X_1, \ldots, X_n$ unabhängig $N(m_X, \sigma_X^2)$-verteilt
$Y_1, \ldots, Y_m$ unabhängig $N(m_Y, \sigma_Y^2)$-verteilt
Mittel unbekannt

Hypothese
(a) $H_0 : \sigma_X^2 = \sigma_Y^2$ (b) $H_0 : \sigma_X^2 \leq \sigma_Y^2$ (c) $H_0 : \sigma_X^2 \geq \sigma_Y^2$

Entscheidungsregel Betrachte als Testgröße

$$T(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = \frac{S_X^2}{S_Y^2} \quad (\, F_{n-1,m-1} - \text{verteilt, falls } \sigma_X^2 = \sigma_Y^2)$$

Ablehnung, falls

(a) $T < F_{n-1,m-1,\frac{\alpha}{2}}$ oder $T > F_{n-1,m-1,1-\frac{\alpha}{2}}$
(b) $T > F_{n-1,m-1,1-\alpha}$ (c) $T < F_{n-1,m-1,\alpha}$.

Statt Mittel und Varianzen auf Gleichheit zu überprüfen kann man schließlich auch zwei (oder mehr) Verteilungen auf Gleichheit überprüfen.

χ^2-Homogenitätstest

Annahme

$X_1^{(1)}, \ldots, X_{n_1}^{(1)}$	unabhängig und identisch verteilt mit Verteilungsfunktion F_1
$X_1^{(2)}, \ldots, X_{n_2}^{(2)}$	unabhängig und identisch verteilt mit Verteilungsfunktion F_2
$\vdots$	$\vdots$ $\vdots$
$X_1^{(k)}, \ldots, X_{n_k}^{(k)}$	unabhängig und identisch verteilt mit Verteilungsfunktion F_k

Hypothese
(a) $H_0 : F_1 = F_2 = \ldots = F_k$

Um die Hypothese zu testen, unterteilen wir zunächst die x-Achse in $m \geq 2$ disjunkte Intervalle

$$A_1 =]-\infty, z_1], A_2 =]z_1, z_2], \ldots, A_m =]z_{m-1}, \infty[$$

und bestimmen für jedes Intervall die Häufigkeiten

$$h_{ij} = \#\{X_l^{(i)} \,:\, X_l^{(i)} \in A_j\}\,, i = 1, \ldots, k,\, j = 1, \ldots, m$$

und bilde hierzu die Spaltensummen

$$h_{\cdot j} = h_{1j} + \ldots h_{kj}\,, j = 1, \ldots, m$$

Entscheidungsregel Betrachte als Testgröße

$$T(X_1^{(1)}, \ldots, X_{n_k}^{(k)}) = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(h_{ij} - \frac{n_i h_{\cdot j}}{n}\right)^2}{\frac{n_i h_{\cdot j}}{n}}$$

mit $n = n_1 + \ldots + n_k$.

Ablehnung, falls

$T > \chi^2_{(k-1)(m-1),1-\alpha}$.

Verbundene Stichproben

χ^2-Anpassungstest

Häufig ist man daran interessiert, ob die unbekannte Verteilung einer Grundgesamtheit gleich einer gegebenen hypothetischen Verteilung ist.
Dazu stellen wir uns vor, dass die Stichprobenvariablen $X_1, \ldots, X_n$ unabhängig und identisch verteilt sind mit einer Verteilungsfunktion F und wir stellen zu gegebener Verteilungsfunktion F_0 die

Hypothese $H_0 : F = F_0$

auf.

Im nächsten Schritt unterteilen wir die x-Achse in $k \geqslant 2$ disjunkte Intervalle

$$A_1 =]-\infty, z_1],\ A_2 =]z_1, z_2], \ldots, A_k =]z_{k-1}, \infty[$$

und bestimmen für jedes Intervall A_j

- die Anzahl h_j der in A_j liegenden Stichprobenwerte

$$h_j = \#\{x_i : x_i \in A_j\}$$

- die theoretische Wahrscheinlichkeit p_j, dass eine Stichprobenvariable X mit Verteilungsfunktion F_0 einen Wert in A_j annimmt

$$p_j = P(X \in A_j) = F_0(z_j) - F_0(z_{j-1})\,.$$

  Hierbei setzt man $F_0(z_0) = 0$ und $F_0(z_k) = 1$.

Hinweis: Ist der Wertebereich der Stichprobenvariablen endlich, etwa $\{a_1, \ldots a_k\}$, so kann man auf die Klassifizierung verzichten und h_j, p_j definieren durch

$$h_j = \#\{x_i : x_i = a_j\} \qquad p_j = P(X = a_j)\,.$$

Entscheidungsregel Betrachte die Testgröße

$$T(X_1, \ldots, X_n) = \sum_{j=1}^{k} \frac{(h_j - np_j)^2}{np_j} \quad \left(= \frac{1}{n} \sum_{j=1}^{n} \frac{h_j^2}{p_j} - n\right)\,.$$

Ablehnung, falls

$$T > \chi^2_{k-1,1-\alpha}\,.$$

Dieser Test hat dann das approximative Niveau α.

Hinweis Die Anzahl der Beobachtungen sollte mindestens so groß sein, dass $np_j \geq 5$ gilt für $j = 1, \ldots, k$.

χ^2-Test auf Unabhängigkeit (Kontingenztest)

Ausgangspunkt des Tests auf Unabhängigkeit ist die Frage, ob zwei Merkmale X und Y in einer gegebenen Grundgesamtheit voneinander unabhängig sind oder nicht. Es ist also ein statistischer Test zu konstruieren, der aufgrund einer zweidimensionalen Stichprobe

$$(x_1, y_1), \ldots, (x_n, y_n)$$

entscheidet, ob die folgende

Hypothese H_0: X und Y sind unabhängig

angenommen werden kann oder nicht.
Wie beim χ^2-Anpassungstest unterteilen wir die x-Achse in $k \geq 2$ disjunkte Intervalle

$$A_1 =]-\infty, z_1],\ A_2 =]z_1, z_2], \ldots, A_k =]z_{k-1}, \infty[$$

und die y-Achse in $l \geq 2$ disjunkte Intervalle

$$B_1 =]-\infty, \tilde{z}_1],\ B_2 =]\tilde{z}_1, \tilde{z}_2], \ldots, B_l =]\tilde{z}_{l-1}, \infty[.$$

Hierzu stellen wir dann die zughörige Kontingenztabelle mit Randhäufigkeiten auf

	y				
x	B_1	B_2	$\ldots$	B_l	
A_1	h_{11}	h_{12}	$\ldots$	h_{1l}	$h_{1\cdot}$
A_2	h_{21}	h_{22}	$\ldots$	h_{2l}	$h_{2\cdot}$
$\vdots$	$\vdots$			$\vdots$	$\vdots$
A_k	h_{k1}	h_{k2}	$\cdots$	h_{kl}	$h_{k\cdot}$
	$h_{\cdot 1}$	$h_{\cdot 2}$	$\ldots$	$h_{\cdot l}$	

und bilden die Größe

$$\tilde{h}_{ij} := \frac{h_{i\cdot} h_{\cdot j}}{n};\ 1 \leq i \leq k,\ 1 \leq j \leq l\,.$$

Begründung Unter der Hypothese sind die Merkmale X und Y unabhängig und damit

$$P(X \in A_i, Y \in B_j) = P(X \in A_i)\, P(Y \in B_j)\,. \tag{3.5}$$

Bei großer Stichprobenlänge n sollte zudem die relative Häufigkeit in der Nähe der theoretischen Wahrscheinlichkeit liegen, also

$$\frac{h_{ij}}{n} \sim P(X \in A_i, Y \in B_j)\,,$$

$$\frac{h_{i\cdot}}{n} \sim P(X \in A_i) \quad \text{und} \quad \frac{h_{\cdot j}}{n} \sim P(Y \in B_j)\,.$$

Ingesamt sollte also gelten

$$\frac{h_{ij}}{n} \sim P(X \in A_i, Y \in A_j) = P(X \in A_i)P(Y \in A_j) \sim \frac{h_{i\cdot}}{n} \cdot \frac{h_{\cdot j}}{n}\,,$$

also $h_{ij} \sim \tilde{h}_{ij}$.

Folglich führen wir nun einen χ^2-Anpassungstest gegen die Produktverteilung $\tilde{h}_{ij}/n$ durch und bilden dementsprechend die Testgröße

$$T(X_1, \ldots, X_n) = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(h_{ij} - \tilde{h}_{ij}\right)^2}{\tilde{h}_{ij}} = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{h_{ij}^2}{\tilde{h}_{ij}} - n\,.$$

Entscheidungsregel

Ablehnung, falls $T > \chi^2_{(k-1)(l-1),1-\alpha}$.

Bemerkung (zur Anzahl der Freiheitsgrade)

Da pro Zeile (bzw. Spalte) eine der Häufigkeiten h_{ij} von den übrigen $l-1$ (bzw $k-1$) Häufigkeiten über die entsprechende Randhäufigkeit $h_{i\cdot}$ (bzw. $h_{\cdot j}$) abhängt, ergibt sich als Anzahl der Freiheitsgrade in der Testgröße

$$kl - l - k + 1 = (k-1)(l-1)\,.$$

Literatur

[1] G. Bamberg, F. Baur, M. Krapp, Statistik, 13. Auflage, R. Oldenbourg Verlag, 2007.

[2] L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz, Statistik, 6. Auflage, Springer Verlag, 2007.

Weitere Literatur

[3] J. Bleymüller, G. Gehlert, H. Gülicher, Statistik für Wirtschaftswissenschaftler, 14. Auflage, Verlag Vahlen, 2004.

[4] L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz, A. Caputo, S. Lang, Arbeitsbuch Statistik, 4. Auflage, Springer Verlag, 2004.

[5] J. Schira, Statistische Methoden der VWL und BWL: Theorie und Praxis, 2. Auflage, Pearson Studium, 2005.