Kapitel 1

Schätzen und Testen

Gerhard Tutz, Skript LMU, 6. Mai 2014

1.1	Schätzung im Ein-Stichprobenfall	3
1.2	Einfache Tests im Einstichprobenfall	8
1.3	Schätzung im Mehr-Stichproben-Fall	9
1.4	Einfache Tests zum Erwartungswert im Zwei-Stichproben-Fall .	13

1.1 Schätzung im Ein-Stichprobenfall

Ein Zufallsvektor $\mathbf{x}$ mit Erwartungswert $\boldsymbol{\mu} = E(\mathbf{x})$ und Kovarianzmatrix $\boldsymbol{\Sigma}$ wird unabhängig und identisch verteilt wiederholt. Die resultierenden Vektoren $\mathbf{x}_1, \dots, \mathbf{x}_n$ lassen sich wieder in der Form der Datenmatrix anordnen

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \\ x_{n1} & \dots & x_{np} \end{pmatrix}.$$

Erwartungswert

Einen Schätzer für den Erwartungswert $\boldsymbol{\mu}$ erhält man durch das *arithmetische Mittel*.

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

Die Komponente $\bar{x}_j = \frac{1}{n} \sum_i x_{ij}$ entspricht dabei dem bekannten arithmetischen Mittel für das j te Merkmal.

Der Schätzer $\bar{\mathbf{x}}$ besitzt eine Art Schwerpunkteigenschaft. Es gilt

$$\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) = \sum_{i=1}^n \mathbf{x}_i - n\bar{\mathbf{x}} = \mathbf{0}.$$

Stellt man sich vor, daß die Vektoren $\mathbf{x}_i - \bar{\mathbf{x}}$ Kräfte darstellen, die auf $\bar{\mathbf{x}}$ wirken, dann ist das System im Gleichgewicht, da sich alle Kräfte aufheben (zu Null aufsummieren).

Eigenschaften des arithmetischen Mittels

Wie schon im univariaten Fall ist das arithmetische Mittel ein erwartungstreuer Schätzer. Für unabhängige und identisch verteilte $\mathbf{x}_1, \dots, \mathbf{x}_n$ mit $E(\mathbf{x}_i) = \boldsymbol{\mu}$ erhält man

$$E(\bar{\mathbf{x}}) = \boldsymbol{\mu}.$$

Das heißt, der Schätzer weist keine Verzerrung auf, er liegt „im Mittel“ richtig. Die Schätzgenauigkeit drückt sich in der Varianz des Schätzer aus, die durch

$$\text{cov}(\bar{\mathbf{x}}) = \frac{1}{n} \boldsymbol{\Sigma}$$

bestimmt ist. Für $n \rightarrow \infty$ verschwindet damit die Kovarianz, insbesondere konvergiert der Standardfehler für die j te Komponente $\text{var}(\bar{x}_j)$ (die Wurzel des j te Diagonalelementes von $\text{cov}(\bar{\mathbf{x}})$) gegen Null. Die Aussage ergibt sich unmittelbar aus

$$\text{cov}(\bar{\mathbf{x}}) = \text{cov}\left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{cov}(\mathbf{x}_i) = \frac{1}{n} \boldsymbol{\Sigma}.$$

Eine weitere Eigenschaft des Schätzers ist seine Variabilität im Vergleich zu anderen erwartungstreuen linearen Schätzern. Stellt $\hat{\boldsymbol{\mu}} = a_1 \mathbf{x}_1 + \dots + a_n \mathbf{x}_n$ einen linearen, erwartungstreuen ($E \hat{\boldsymbol{\mu}} = \boldsymbol{\mu}$) Schätzer dar, so gilt

$$E(\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|) \leq E(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|),$$

wobei $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ den quadratischen Abstand zwischen $\hat{\boldsymbol{\mu}}$ und $\boldsymbol{\mu}$ bezeichnet. In diesem Sinne ist $\bar{\mathbf{x}}$ der beste lineare, erwartungstreue Schätzer für den Erwartungswert.

Das asymptotische Verhalten von $\bar{\mathbf{x}}$ ergibt sich aus dem *multivariaten zentralen Grenzwertsatz*. Dieser besagt, daß für identisch verteilte unabhängige Zufallsvariablen $\mathbf{x}_1, \dots, \mathbf{x}_n$ mit $E(\mathbf{x}_i) = \boldsymbol{\mu}$, $\text{cov}(\mathbf{x}_i) = \boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}$ positiv definit, ohne weitere Annahmen an die Verteilungsform asymptotisch immer eine Normalverteilung resultiert, genauer

$$\sqrt{n}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

d.h. $\sqrt{n}(\bar{\mathbf{x}} - \boldsymbol{\mu})$ konvergiert nach Verteilung gegen eine zentrierte Normalverteilung. Das rechtfertigt für großes n die Aussage

$$\bar{\mathbf{x}} \text{ approximativ } N\left(\boldsymbol{\mu}, \frac{1}{n} \boldsymbol{\Sigma}\right)\text{-verteilt.}$$

Kovarianz

Die Schätzung für die Kovarianzmatrix läßt sich ebenso aus den bekannten Schätzern für die Kovarianz gewinnen. Der Momentenschätzer für die Kovarianz des r ten und s ten Merkmals ist bestimmt durch

$$\hat{\sigma}_{rs} = \frac{1}{n} \sum_{i=1}^n (x_{ir} - \bar{x}_r)(x_{is} - \bar{x}_s).$$

Daraus ergibt sich die Schätzmatrix $\hat{\Sigma} = (\hat{\sigma}_{rs})$. In Matrixschreibweise besitzt die Matrix $\hat{\Sigma}$ die einfache Form

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T,$$

wobei der Vektor $(\mathbf{x}_i - \bar{\mathbf{x}})$ auf der rechten Seite einmal als Spaltenvektor und einmal als Zeilenvektor erscheint, so daß das Produkt eine Matrix bildet. Der Schätzer $\hat{\sigma}_{rs}$ für die Kovarianz $\text{cov}(x_r, x_s)$ ist verzerrt. Eine Korrektur ist einfach zu erreichen, indem die Normierung $1/n$ modifiziert wird zu $1/(n-1)$. Man erhält die *empirische Kovarianzmatrix*

$$S = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \frac{n}{n-1} \hat{\Sigma}$$

Eigenschaften der empirischen Varianz

Für identisch verteilte unabhängige $\mathbf{x}_1, \dots, \mathbf{x}_n$ mit $\text{cov}(\mathbf{x}_i)$ erhält man mit S einen erwartungstreuen (unverzerrten) Schätzer,

$$E(S) = \Sigma,$$

somit ist S erwartungstreu und $\hat{\Sigma} = \frac{n-1}{n} S$ unterschätzt tendenziell die Kovarianzen und Varianzen der Komponenten.

Wenn die Vektoren normalverteilt sind, $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \Sigma)$, gilt zusätzlich, daß die Erwartungswertschätzung unabhängig von der Kovarianzschätzung ist,

$\bar{\mathbf{x}}$ und S sind unabhängig.

Für normalverteilte Merkmale läßt sich auch der Verteilungstyp bestimmen. Da der Schätzer selbst eine Matrix darstellt, ergibt sich eine Matrixverteilung, genauer die Wishart-Verteilung. Diese ist weitgehend auf den Fall der empirischen Kovarianzmatrix zugenschnitten. Für die Definition der *Wishart-Verteilung* legt man unabhängige p -dimensionale Zufallsvektoren $\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_m$, $\tilde{\mathbf{x}}_i \sim N(\boldsymbol{\theta}, \Sigma)$, zugrunde. Dann ist die Matrix

$$M = \sum_{i=1}^m \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T$$

Wishart-verteilt mit den Parametern Σ und m , kurz $M \sim W_p(\Sigma, m)$.

Wäre im vorliegenden Schätzproblem aus den Beobachtungen $\mathbf{x}_1, \dots, \mathbf{x}_n$, $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \Sigma)$ der Erwartungswert bekannt, würde man Σ schätzen durch Einzentrieren der $\mathbf{x}_i$ mit

$$\hat{\Sigma}_\mu = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T.$$

Entsprechend der Definition ist $n\hat{\Sigma}_\mu$ Wishart-verteilt mit $n\hat{\Sigma}_\mu \sim W_p(\Sigma, n)$. In der empirischen Kovarianzmatrix S ist jedoch $\boldsymbol{\mu}$ durch den Schätzer $\bar{\mathbf{x}}$ ersetzt. Der Effekt ist, daß nicht n unabhängige Vektoren die Schätzung bestimmen, sondern nur $n - 1$, man erhält

$$(n - 1)S = n\hat{\Sigma} = \sum_{i=1}^n n(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \sim W_p(\Sigma, n - 1).$$

Für den univariaten Spezialfall $p = 1$ erhält man $(n - 1)\hat{\sigma}^2 \sim W_1(\sigma^2, n - 1)$ und die Wishart-Verteilung $W_1(\sigma^2, n - 1)$ ist identisch zu $\sigma^2\chi^2(n - 1)$, d.h. der χ^2 -Verteilung mit $n - 1$ Freiheitsgraden, multipliziert mit σ^2 . Für weitere Eigenschaften der Wishart-Verteilung siehe Appendix.

Grundlegende Schätzer

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$$

$$S = \frac{1}{n - 1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

$$\hat{\Sigma} = \frac{n - 1}{n} S$$

Eigenschaften

$$E(\bar{\mathbf{x}}) = \boldsymbol{\mu}, \quad \text{var}(\bar{\mathbf{x}}) = \frac{1}{n} \Sigma,$$

$$E(S) = \Sigma$$

Wenn $\mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma)$:

$$\bar{\mathbf{x}} \sim N(\boldsymbol{\mu}, \Sigma/n)$$

$$(n - 1)S = n\hat{\Sigma} \sim W_p(\Sigma, n - 1)$$

Schätzung von Kovarianzen von Gruppen von Kovariablen

Betrachtet man zwei Merkmale $\mathbf{x}$ und $\mathbf{y}$ beliebiger Dimension mit den Stichprobenvariablen $(\mathbf{y}_i, \mathbf{x}_i)$, $i = 1, \dots, n$ erhält man die üblichen erwartungstreuen Schätzer für die Kovarianzen zwischen Komponenten aus $\mathbf{y}$ und aus $\mathbf{x}$ in Matrixform durch

$$S_{\mathbf{y}\mathbf{x}} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{x}_i - \bar{\mathbf{x}})^T,$$

wobei $\bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i$, $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$.

Matrixdarstellungen

Das arithmetische Mittel läßt sich mit dem „Einser“-Vektor $\mathbf{I}^T = (1, \dots, 1)$ auch mit Hilfe der Beobachtungsmatrix $\mathbf{X}$ darstellen durch

$$\bar{\mathbf{x}} = \mathbf{X}^T \mathbf{I}.$$

Eine alternative Darstellung der Kovarianzmatrix ergibt sich aus der zentrierten Datenmatrix $\mathbf{X}_c$, in der von den Beobachtungen jeweils der Mittelwert abgezogen ist,

$$\mathbf{X}_c = \begin{pmatrix} \mathbf{x}_1^T - \bar{\mathbf{x}}^T \\ \vdots \\ \mathbf{x}_n^T - \bar{\mathbf{x}}^T \end{pmatrix}.$$

Die Beobachtungen der j -ten Variable, die der j -ten Spalte entsprechen summieren sich zu Null auf. Einfache Matrixrechnung ergibt, daß die empirische Kovarianzmatrix von der Form

$$S = \frac{1}{n-1} \mathbf{X}_c^T \mathbf{X}_c.$$

ist.

Die Datenmatrix für die zentrierten Daten (mit den Zeilen $\mathbf{x}_i - \bar{\mathbf{x}}$) ergibt sich damit durch

$$\mathbf{X}_c = \mathbf{X} - \frac{1}{n} \mathbf{I} \mathbf{I}^T \mathbf{X} = (\mathbf{I} - \frac{1}{n} \mathbf{I} \mathbf{I}^T) \mathbf{X} = \mathbf{C} \mathbf{X},$$

wobei $\mathbf{C} = \mathbf{I} - \frac{1}{n} \mathbf{I} \mathbf{I}^T$ die Zentrierungsmatrix bezeichnet. $\mathbf{C}$ ist symmetrisch und idempotent ($\mathbf{C}^2 = \mathbf{C}$).

Die empirische Kovarianzmatrix läßt sich dann in Matrixschreibweise darstellen durch

$$\begin{aligned} S &= \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \frac{1}{n-1} \mathbf{X}_c^T \mathbf{X}_c \\ &= \frac{1}{n-1} \mathbf{X}^T \mathbf{C}^T \mathbf{C} \mathbf{X} = \frac{1}{n-1} \mathbf{X}^T \mathbf{C} \mathbf{X}. \end{aligned}$$

Für die empirische Kovarianzmatrix für Gruppen von Variablen in Matrixform benutzt man die Matrizen

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_n^T \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}.$$

Man erhält dann aus den zentrierten Matrizen $\mathbf{Y}_c, \mathbf{X}_c$ die Darstellung

$$S_{yx} = \frac{1}{n-1} \mathbf{Y}_c^T \mathbf{X}_c.$$

1.2 Einfache Tests im Einstichprobenfall

Ausgangspunkt sind unabhängige Beobachtungen $\mathbf{x}_1, \dots, \mathbf{x}_n$, die aus einer Normalverteilung gezogen sind, $\mathbf{x}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Test auf Erwartungswert bei bekannter Varianz

Untersucht wird, ob der tatsächliche (und unbekante) Erwartungswert $\boldsymbol{\mu}$ einem vorgegebenen Wert $\boldsymbol{\mu}_0$ entspricht. Das zu untersuchende Hypothesenpaar aus Nullhypothese und Alternativhypothese ist gegeben durch

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0,$$

wobei vorerst $\boldsymbol{\Sigma}$ als bekannt vorausgesetzt wird.

Eine Teststatistik ist gegeben durch

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0), \quad (1.1)$$

die bei Gültigkeit von H_0 χ^2 -verteilt ist mit p Freiheitsgraden, $T^2 \sim \chi^2(p)$. Entsprechend erhält man einen α -Signifikanztest durch

$$H_0 \text{ ablehnen, falls } T^2 > \chi_{1-\alpha}^2(p),$$

wobei $\chi_{1-\alpha}^2(p)$ das $(1-\alpha)$ -Quantil der χ^2 -Verteilung mit p Freiheitsgraden bezeichnet.

Man beachte, daß das Testproblem der multivariaten Struktur entsprechend immer „zweiseitig“ angelegt ist. Während man in eindimensionalen Testproblemen auch einseitige Testprobleme (Abweichung nach unten bzw. oben) betrachtet, ist hier die Alternative immer $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$. T^2 läßt sich als (quadrierte) Verallgemeinerung des eindimensionalen Gauß-Tests verstehen (siehe die folgenden Bemerkungen zur Teststatistik).

Zur Teststatistik

Für eindimensionales Merkmal $\mathbf{x} = x$ kennt man den Gauß-Test

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\text{cov}(\bar{x})}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}},$$

der unter H_0 standardnormalverteilt ist. t ist die Standardisierung von $\bar{x}$, da unter H_0 $\tilde{x} \sim N(\mu_0, \sigma^2/n)$ gilt.

Da im multivariaten Fall $\bar{\mathbf{x}} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}/n)$ läßt sich $\bar{\mathbf{x}}$ in Matrixform normieren durch $n^{1/2} \boldsymbol{\Sigma}^{-1/2}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$, wobei $\boldsymbol{\Sigma}^{-1/2}$ eine Wurzel von $\boldsymbol{\Sigma}$ bezeichnet (siehe ???).

Wenn H_0 wahr ist, gilt $n^{1/2} \boldsymbol{\Sigma}^{-1/2}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Eine eindimensionale Teststatistik erhält man indem, man statt des Vektors $n^{1/2} \boldsymbol{\Sigma}^{-1/2}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$ dessen quadrierte Länge betrachtet. Man erhält

$$T^2 = \| n^{1/2} \boldsymbol{\Sigma}^{-1/2}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \|^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}^{-1/2}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0).$$

Als α -Konfidenzintervall für $\boldsymbol{\mu}$ benutzt man

$$\{\boldsymbol{\mu} : n(\boldsymbol{\mu} - \bar{\mathbf{x}})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{x}}) \leq \chi_{1-\alpha}^2(p)\},$$

das einen Ellipsoiden mit Mittelpunkt $\bar{\mathbf{x}}$ darstellt. Das Konfidenzintervall enthält somit alle $\boldsymbol{\mu}$, so daß die Mahalanobisdistanz zwischen $\boldsymbol{\mu}$ und $\bar{\mathbf{x}}$ (gebildet mit $(\boldsymbol{\Sigma}/n)^{-1}$) kleiner als $\chi_{1-\alpha}^2(p)$ ist.

Test auf Erwartungswert bei unbekannter Varianz

Ist die Kovarianz unbekannt, ist $\boldsymbol{\Sigma}$ durch eine Schätzung zu ersetzen. Man betrachtet

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0).^1$$

Man erhält dann unter H_0 eine F -verteilte Größe für die Transformation

$$\frac{n-p}{(n-1)p} T^2 \sim F(p, n-p).$$

H_0 wird abgelehnt, wenn die Größe das Quantil der F -Verteilung (mit Freiheitsgraden p und $n-p$) $F_{1-\alpha}(p, n-p)$ überschreitet. Das zugehörige Konfidenzintervall ist bestimmt durch

$$\{\boldsymbol{\mu} : \frac{(n-p)n}{(n-1)p} (\boldsymbol{\mu} - \bar{\mathbf{x}})^T \mathbf{S}^{-1}(\boldsymbol{\mu} - \bar{\mathbf{x}}) \leq F_{1-\alpha}(p, n-p)\}$$

und stellt wiederum einen Ellipsoid mit Mittelpunkt $\bar{\mathbf{x}}$ dar.

1.3 Schätzung im Mehr-Stichproben-Fall

Zufallsvektoren in partitionierten Grundgesamtheiten

Wird ein Merkmalsvektor in verschiedenen Grundgesamtheiten (Klassen, Gruppen, Bedingungen) untersucht, ist die Verteilung i.a. in diesen Grundgesamtheiten unterschiedlich. Bezeichnet $\mathbf{x}_r$ den Merkmalsvektor in der r ten Grundgesamtheit, dann ergeben sich mit

$$\text{der Dichte } f(\mathbf{x} | r), \boldsymbol{\mu}_r = \mathbb{E}(\mathbf{x}_r) \text{ und } \boldsymbol{\Sigma}_r = \text{cov}(\mathbf{x}_r)$$

potentiell unterschiedliche Dichten, Erwartungswerte und Kovarianzmatrizen.

¹hier konnte ich was nicht lesen: (Tunb) ???

Besitzt die r te Grundgesamtheit die a-priori-Wahrscheinlichkeit $p(r)$ ($\sum_{r=1}^g p(r) = 1$) so erhält man für die Ziehung aus der Grundgesamtheit (Zufallsvektor $\mathbf{x}$) eine *Mischverteilung* mit der Dichte

$$f(\mathbf{x}) = \sum_{r=1}^g p(r) f(\mathbf{x} | r)$$

und dem Erwartungswert

$$\boldsymbol{\mu} = E(\mathbf{x}) = \sum_{r=1}^g p(r) \boldsymbol{\mu}_r.$$

Für die Kovarianz $\text{cov}(\mathbf{x})$ erhält man die wichtige *Streuungszerlegung*

$$\text{cov}(\mathbf{x}) = \boldsymbol{\Sigma}_w + \boldsymbol{\Sigma}_b, \quad (1.2)$$

wobei

$$\boldsymbol{\Sigma}_w = \sum_{r=1}^g p(r) \boldsymbol{\Sigma}_r$$

als gewichtete Summe der Varianzen in den Grundgesamtheiten die Kovarianz innerhalb der Gruppen/Klassen bezeichnet (w für within/innerhalb) und

$$\boldsymbol{\Sigma}_b = \sum_{r=1}^g p(r) (\boldsymbol{\mu}_r - \boldsymbol{\mu})(\boldsymbol{\mu}_r - \boldsymbol{\mu})^T$$

die Kovarianz zwischen den Gruppen/Klassen (b für between) bezeichnet.

Die Kovarianz zerfällt somit in die Variabilität innerhalb der Gruppen/Klassen und zwischen diesen. Die Kovarianz innerhalb erfasst, wie stark das Merkmal innerhalb der Gruppen um den jeweiligen Erwartungswert $\boldsymbol{\mu}_r$ schwankt, während die Kovarianz zwischen den Gruppen die Variabilität der Klassenmittelwerte $\boldsymbol{\mu}_r$ um das Gesamtmittel $\boldsymbol{\mu}$ wiedergibt.

Formal lässt sich der Zufallsvektor $\mathbf{x}$ als Abbildung $\mathbf{x} : \Omega \rightarrow \mathbb{R}$ verstehen. Betrachtet man eine zweite diskrete Zufallsvariable $Y : \Omega \rightarrow \{1, \dots, g\}$, die die Klassen repräsentiert, dann läßt sich Ω als partitioniert in $\Omega_1, \dots, \Omega_g$ darstellen, d.h. $\Omega = \Omega_1 \cup \dots \cup \Omega_g$ mit $\Omega_r = \{\omega : Y(\omega) = r\}$. Die Zufallsvariable $\mathbf{x}_r$ entspricht dann der Einschränkung von $\mathbf{x}$ auf Ω bzw. der Zufallsvariable $\mathbf{x}$ unter der Bedingung $Y = r$, kurz $\mathbf{x} | Y = r$. Die Streuungszerlegung entspricht dann der üblichen Zerlegung

$$\text{cov}(\mathbf{x}) = E_Y(\text{cov } E(\mathbf{x}|Y)) + \text{cov}_Y(E(\mathbf{x}|Y)).$$

Ziehung aus unterschiedlichen Grundgesamtheiten

Betrachtet werden Beobachtungen aus g Grundgesamtheiten (Klassen, Gruppen). Die Beobachtungen aus der r ten Grundgesamtheit $\mathbf{x}_{r1}, \dots, \mathbf{x}_{rn_r}$ seien wieder unabhängig und identisch verteilt, d.h.

$\mathbf{x}_{r1}, \dots, \mathbf{x}_{rn_r}$ besitzen Erwartungswert $\boldsymbol{\mu}_r$ und Kovarianz $\boldsymbol{\Sigma}_r$.

Darüberhinaus nimmt man an, daß die Beobachtungen nicht nur innerhalb der Klassen, sondern auch über diese hinweg unabhängig sind. Die Beobachtungen zur r ten Klassen lassen sich wieder zu einer Beobachtungsmatrix $\mathbf{X}_r$ zusammenfassen, die Zusammenstellung aller $\mathbf{X}_1, \dots, \mathbf{X}_g$ ergibt dann die gesamte Beobachtungsmatrix $\mathbf{X}$:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_g \end{pmatrix}, \quad \mathbf{X}_r = \begin{pmatrix} \mathbf{x}_{r1}^T \\ \vdots \\ \mathbf{x}_{rn_r}^T \end{pmatrix}.$$

Als natürliche Schätzung für Erwartungswert und Varianz in der r ten Klasse ergeben sich die folgenden Schätzer

Arithmetisches Mittel und empirische Kovarianz in der i ten
Kovarianz r ten Grundgesamtheit

$$\bar{\mathbf{x}}_r = \sum_{i=1}^{n_r} \mathbf{x}_{ri},$$

$$\mathbf{S}_r = \frac{1}{n_r - 1} \sum_{i=1}^{n_r} (\mathbf{x}_{ri} - \bar{\mathbf{x}}_r)(\mathbf{x}_{ri} - \bar{\mathbf{x}}_r)^T, \quad \hat{\boldsymbol{\Sigma}}_r = \frac{n_r - 1}{n_r} \mathbf{S}_r.$$

Als Gesamtmittel ergibt sich mit $n = n_1 + \dots + n_g$

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{r=1}^g \sum_{i=1}^{n_r} \mathbf{x}_{ri} = \sum_{r=1}^g \frac{n_r}{n} \bar{\mathbf{x}}_r.$$

Empirische Streuungszerlegung und gepoolte Varianz

Der Streuungszerlegung in der Grundgesamtheit (1.2) entspricht die *empirische Streuungszerlegung*

$$\sum_{r=1}^g \sum_{i=1}^{n_r} (\mathbf{x}_{ri} - \bar{\mathbf{x}})(\mathbf{x}_{ri} - \bar{\mathbf{x}})^T = \sum_{r=1}^g \sum_{i=1}^{n_r} (\mathbf{x}_{ri} - \bar{\mathbf{x}}_r)(\mathbf{x}_{ri} - \bar{\mathbf{x}}_r)^T + \sum_{r=1}^g n_r (\bar{\mathbf{x}}_r - \bar{\mathbf{x}})(\bar{\mathbf{x}}_r - \bar{\mathbf{x}})^T,$$

kurz

$$\mathbf{T} = \mathbf{W} + \mathbf{B}. \quad (1.3)$$

Dabei bezeichnet

$$\mathbf{T} = \sum_{r=1}^g \sum_{i=1}^{n_r} (\mathbf{x}_{ri} - \bar{\mathbf{x}})(\mathbf{x}_{ri} - \bar{\mathbf{x}})^T$$

die *totale Streuungsmatrix* (häufig auch totale SSP-Matrix für sum of squares and products), die die Streuung aller Messungen um das globale Mittel wiedergibt,

$$\mathbf{W} = \sum_{r=1}^g \sum_{i=1}^{n_r} (\mathbf{x}_{ri} - \bar{\mathbf{x}}_r)(\mathbf{x}_{ri} - \bar{\mathbf{x}}_r)^T = \sum_{r=1}^g (n_r - 1) \mathbf{S}_r = \sum_{r=1}^g n_r \hat{\boldsymbol{\Sigma}}_r$$

die *Streuung innerhalb der Gruppen* (within SSP-Matrix) und

$$\mathbf{B} = \sum_{r=1}^g n_r (\bar{\mathbf{x}}_r - \bar{\mathbf{x}})(\bar{\mathbf{x}}_r - \bar{\mathbf{x}})^T$$

die *Streuung zwischen den Gruppen* (between SSP-Matrix), d.h. die Variation der Gruppenmittelwerten $\bar{\mathbf{x}}_r$ um $\bar{\mathbf{x}}$.

Man beachte, daß die Zerlegung (1.3) Abweichungsquadratsummen wiedergeben, ohne daß durch die Anzahl der Terme geteilt wird. Teilt man durch n , so ergibt sich die Zerlegung $\mathbf{T}/n = \mathbf{W}/n + \mathbf{B}/n$ und mit $\hat{\boldsymbol{\Sigma}} = \mathbf{T}/n$ die zur Zerlegung der Grundgesamtheiten analoge Form

$$\hat{\boldsymbol{\Sigma}} = \sum_{r=1}^g \frac{n_r}{n} \hat{\boldsymbol{\Sigma}}_r + \sum_{r=1}^g \frac{n_r}{n} (\bar{\mathbf{x}}_r - \bar{\mathbf{x}})(\bar{\mathbf{x}}_r - \bar{\mathbf{x}})^T.$$

Allerdings wurden dabei die verzerrten Momentenschätzer $\hat{\boldsymbol{\Sigma}}_r$ für $\boldsymbol{\Sigma}_r$ verwendet.

Eine gelegentlich getroffene Annahme ist die, daß die Kovarianzen in jeder Subpopulation identisch ist, d.h.

$$\boldsymbol{\Sigma}_1 = \dots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}. \quad (1.4)$$

$\boldsymbol{\Sigma}$ ist dann identisch mit der Kovarianz innerhalb der Klassen $\boldsymbol{\Sigma}_W = \boldsymbol{\Sigma}$. Eine Schätzung für $\boldsymbol{\Sigma}$ erhält man durch die gepoolte empirische Kovarianz

$$\mathbf{S} = \frac{1}{n-g} \sum_{r=1}^g (n_r - 1) \mathbf{S}_r = \frac{1}{n-g} \mathbf{W}.$$

$\mathbf{S}$ ist unter der Annahme (1.4) ein erwartungstreuer Schätzer für $\boldsymbol{\Sigma}$, der für normalverteilte Merkmale einer Wishart-Verteilung folgt, $(n-g)\mathbf{S} \sim W_p(\boldsymbol{\Sigma}, n-g)$.

Darüberhinaus sind dann $\mathbf{W}$ und $\mathbf{B}$ aus der Streuungszersetzung unabhängig. Gilt für normalverteilte Merkmale $\mu_1 = \dots = \mu_g$, dann ist auch $\mathbf{B}$ Wishart-verteilt $\mathbf{B} \sim W_p(\boldsymbol{\Sigma}, n-g)$.

Insgesamt ergibt sich für normalverteilte Merkmale mit $\Sigma_1 = \dots = \Sigma_g$ und $\mu_1 = \dots = \mu_g$ die Zerlegung in unabhängige Komponenten.

$$\begin{array}{rcl} \mathbf{I} & = & \mathbf{B} \quad + \quad \mathbf{W} \\ W_p(\Sigma, n-1) & & W_p(\Sigma, n-g) \quad W_p(\Sigma, g-1) \\ p=1: \sigma^2 \chi^2(n-1) & & \sigma^2 \chi^2(n-g) \quad \sigma^2 \chi^2(g-1) \end{array}$$

1.4 Einfache Tests zum Erwartungswert im Zwei-Stichproben-Fall

Vergleich von Erwartungswerten bei unabhängigen Stichproben

Unter der Voraussetzung unabhängiger Beobachtungen

$$\begin{array}{ll} \mathbf{x}_{11}, \dots, \mathbf{x}_{1n_1} & \text{aus } N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \\ \mathbf{x}_{21}, \dots, \mathbf{x}_{2n_2} & \text{aus } N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}), \end{array}$$

untersucht man das Testproblem

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \quad (1.5)$$

Eine Teststatistik ist gegeben durch

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2),$$

wobei $\mathbf{S}$ die gepoolte Kovarianzmatrix darstellt. Unter H_0 ergibt sich eine F -verteilte Größe für die Transformation

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2 \sim F(p, n_1 + n_2 - p - 1).$$

H_0 wird abgelehnt, wenn die Größe das entsprechende $(1-\alpha)$ -Quantil der F -Verteilung, $F_{1-\alpha}(p, n_1 + n_2 - p - 1)$ überschreitet.

Vergleich von Erwartungswerten bei verbundenen Stichproben

Im Gegensatz zum Fall unabhängiger Stichproben werden nun die Paare von Beobachtungsvektoren (x_{1i}, x_{2i}) an einer statistischen Einheit erhoben. Beispiele dafür sind

- physiologische Messgrößen an Individuen vor (x_{1i}) und nach (x_{2i}) einer Behandlung,
- ökonomische Kenngrößen einer Firma in zwei aufeinanderfolgenden Jahren,
- Meßwerte, die unter zwei experimentellen Bedingungen an demselben Individuum erhoben werden.

Man will nun testen, ob die Erwartungswerte der beiden Messungen identisch sind oder nicht. Mit $\mu_1 = E(x_i)$, $\mu_2 = E(x_{2i})$ erhält man das Hypothesenpaar

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \quad (1.6)$$

Vorausgesetzt werden unabhängige Wiederholungen (x_{1i}, x_{2i}) , $i = 1, \dots, n$ die einer multivariaten Normalverteilung folgen. Während im Fall unabhängiger Beobachtungen die Unabhängigkeit von x_{1i} und x_{2i} (d.h. innerhalb der Paare) gefordert wird, besitzt die Kovarianz $\text{cov}(x_{1i}, x_{2i})$ im allgemeinen keine Diagonalstruktur.

Das Testproblem läßt sich einfach auf ein Ein-Stichprobenproblem zurückführen, indem man als Beobachtungen die Differenzen $\mathbf{d}_i = \mathbf{x}_{1i} - \mathbf{x}_{2i}$ betrachtet. Diese sind wiederum normalverteilt mit $N(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_d)$, wobei $\boldsymbol{\Sigma}_d$ die meist unbekannte Kovarianz von $\mathbf{x}_{1i} - \mathbf{x}_{2i}$ bezeichnet. Für das zu (1.6) äquivalente Hypothesenpaar

$$H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = 0 \quad H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq 0$$

läßt sich unmittelbar die Teststatistik T^2 aus (1.1) verwenden, wobei die Beobachtungen durch $\mathbf{d}_1, \dots, \mathbf{d}_n$ gegeben sind. Man erhält unter H_0

$$\frac{(n-p)n}{(n-1)p} \bar{\mathbf{d}}^T \mathbf{S}_d \bar{\mathbf{d}} \sim F(p, n-p),$$

wobei $\bar{\mathbf{d}} = \sum_i \mathbf{d}_i/n$, $\mathbf{S}_d = \sum_i (\mathbf{d}_i - \bar{\mathbf{d}})(\mathbf{d}_i - \bar{\mathbf{d}})^T/(n-1)$.