

Skriptum

zu der Vorlesung

Grundzüge der Statistik

Teil B

von

Ralph Friedmann

Universität des Saarlandes WS 2003/2004

Inhaltsverzeichnis

1	Information aus Zufallsstichproben	1
2	Parameterpunktschätzung	5
3	Intervall-Schätzung	22
4	Hypothesentests auf der Grundlage von Stichprobeninformation	28
5	Chi-Quadrat-Tests auf Anpassung und Unabhängigkeit	55
6	Vergleich der Lage von Verteilungen mit unabhängigen Stichproben	66
7	Das klassische Regressionsmodell: Einführung	80
8	Einfache lineare Regression	88
9	Multiple lineare Regression	103
A	Einführung in die Matrizenrechnung	123

Kapitel 1

Information aus Zufallsstichproben

1.1 Grundgesamtheit und Zufallsstichprobe

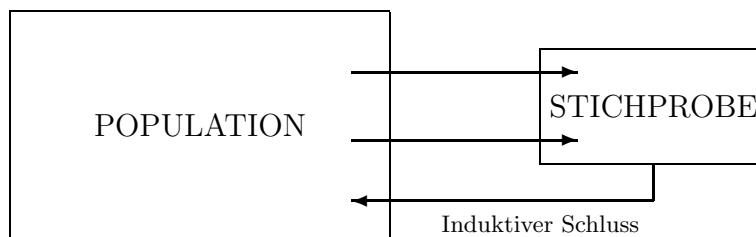


Abbildung 1.1: Induktiver Schluss

Bei einer klaren Abgrenzung einer Grundgesamtheit und eines statistischen Merkmals Y , das für jedes Element der Grundgesamtheit einen bestimmten Wert (reelle Zahl) annimmt, steht die Häufigkeitsverteilung der Merkmalswerte in der Grundgesamtheit (kurz: die Verteilung der Grundgesamtheit) prinzipiell fest. Es bedarf aber einer vollständigen Erhebung der Merkmalswerte aller Elemente der Grundgesamtheit, um diese Verteilung wirklich exakt kennen zu lernen. Eine vollständige Erhebung ist häufig, zumindest bei einem begrenzten Budget, nicht durchführbar. Stattdessen wird versucht, aus einer Stichprobe, die aus der Grundgesamtheit gezogen wird, auf die Verteilung der Grundgesamtheit zu schließen und dabei auch die Größe des Fehlers, mit dem zu rechnen ist, zu charakterisieren. Die Konstruktion und Analyse geeigneter Schlussverfahren ist der Gegenstand der **induktiven Statistik** (synonym: schließende Statistik, Inferenz-Statistik).

Im Unterschied zur mathematischen Deduktion - der Ableitung eines Satzes

mit logischen Schlüssen aus bestimmten Voraussetzungen - geht es bei der **induktiven Statistik** um einen Schluss von dem **Stichprobenergebnis** (also einem Teil der Grundgesamtheit) auf die **Verteilung der Grundgesamtheit** (also auf das Ganze). Oft ist man auch nicht an der gesamten Information über die Verteilung der Grundgesamtheit, sondern nur an einzelnen **Parametern** der Verteilung der Grundgesamtheit interessiert.

Die induktive Statistik hat nicht nur die Verteilung eines Merkmals Y in einer konkret vorhandenen, abgegrenzten Grundgesamtheit von Merkmalsträgern im Auge. Stattdessen kann auch die Annahme einer bestimmten, aber nicht (vollständig) bekannten, **stochastischen Gesetzmäßigkeit** den Ausgangspunkt einer Stichprobenuntersuchung bilden (Beispiele: Anzahl Y_S der bei einer Versicherung pro Woche gemeldeten Schadensfälle, Wartezeit Y_W zwischen dem Eingang von Aufträgen). Die unbekanntenen Verteilungen von Y_S und Y_W werden in solchen Fällen wieder als **Verteilung der Grundgesamtheit** betrachtet, für die aus einer Stichprobe Schlüsse gezogen werden. In den beiden genannten Beispielen liegt es nahe, für die Verteilung der Grundgesamtheit eine **parametrische Verteilungsannahme** zu treffen, etwa mit der Annahme, dass die Anzahl Y_S der gemeldeten Schadensfälle einer Poissonverteilung mit einem bestimmten (unbekannten) Verteilungsparameter λ_S folgt, die Wartezeit Y_W zwischen zwei Auftragseingängen einer Exponentialverteilung mit einem Parameter λ_W . In einer solchen Situation liegt der Stichprobenuntersuchung also die Annahme zugrunde, dass die Beobachtungswerte als Realisierungen einer Zufallsvariablen mit einer Verteilung aus einer bestimmten parametrischen Familie von Verteilungen (z.B. Normalverteilung, Exponentialverteilung, Poissonverteilung, Bernoulliverteilung) betrachtet werden können. Die Stichprobeninformation soll dann genutzt werden, um eine Aussage über die speziellen Parameterwerte der Verteilung zu machen.

Beispiele für Parameter der Grundgesamtheit

- **Einkommensstichprobe**

μ	Mittleres Einkommen i. d. Grundges.	}	endliche Grundgesamtheit
σ^2	Varianz d. Einkommens i. d. Grundges.		
- **Verpackungsmaschine: Normalverteiltes Füllgewicht von abgepackten Nudeln**
 (μ, σ^2) Parameter der Normalverteilung des Füllgewichts
- **Qualitätskontrolle: Gut/Schlecht-Prüfung, Bernoulli-Verteilung**
 p Ausschussanteil bei einem Produktionsprozess, Trefferwahrscheinlichkeit

Man beachte, dass eine **parametrische Verteilungsannahme** auch selbst wieder zum Gegenstand einer statistischen Untersuchung gemacht werden kann und dann bei entsprechendem Stichprobenergebnis gegebenenfalls verworfen werden kann.

1.2 Zufallsstichproben, Stichprobenraum und Stichprobenfunktionen

1.2.1 Verteilung der Stichprobenvariablen

Begrifflich ist zu unterscheiden zwischen der Verteilung der Grundgesamtheit (wir sagen: der Verteilung von Y) und der Verteilung der Zufallsvariablen X , die durch einen Stichprobenzug definiert wird. Wenn man z.B. bei der Auswahl eines Haushalts für eine Einkommensstichprobe dem Interviewer gestattet, den Haushalt aus seinem Freundeskreis auszuwählen, wird sich die Verteilung der Stichprobenvariablen X von der Verteilung von Y (also der Grundgesamtheit) unterscheiden. Wir betrachten aber nur solche Stichproben, bei denen die Verteilung der Stichprobenvariablen X mit der Verteilung von Y übereinstimmt. Solche Stichproben werden als **Zufallsstichproben** bezeichnet.

Bei einer endlichen Grundgesamtheit kann die reine Zufallsauswahl eines Elements in der Praxis unter Verwendung von Zufallszahlentabellen oder der Erzeugung von gleichverteilten (Pseudo-)Zufallszahlen mit einem Computer realisiert werden (z.B. wird jedem Element eine Zufallszahl zugeordnet und dann das Element mit der kleinsten Zufallszahl ausgewählt). Wenn man dagegen z.B. eine Stichprobe aus einem fortlaufenden Prozess entnimmt, ist die Übereinstimmung der Verteilung der Stichprobenvariablen X mit der Verteilung der Grundgesamtheit ein Teil der Annahme.

Eine Zufallsstichprobe vom Umfang n , bei der die n Stichprobenvariablen X_1, \dots, X_n unabhängig und jeweils identisch mit der Grundgesamtheit verteilt sind, heißt **einfache Zufallsstichprobe** vom Umfang n . Wir beschränken uns auf die Betrachtung von einfachen Zufallsstichproben.

1.2.2 Stichprobenraum und Stichprobenfunktionen

Die Realisierung der n Werte einer Stichprobe vom Umfang n , also (x_1, \dots, x_n) , wird als **Stichprobenergebnis** bezeichnet. Die Menge aller möglichen Stichprobenergebnisse heißt Stichprobenraum. Eine **Stichprobenfunktion** oder **Statistik** ist eine Funktion von Stichprobenvariablen X_1, \dots, X_n , deren Werte bzw. Realisierungen vollständig durch das Stichprobenergebnis (x_1, \dots, x_n)

numerisch bestimmt sind. So ist z.B. die Funktion

$$V = \frac{1}{n} \sum_i (X_i - \mu)^2$$

nur dann eine Stichprobenfunktion, wenn der Erwartungswert μ der Grundgesamtheit bekannt ist. Sonst betrachten wir diese Funktion nicht als eine Stichprobenfunktion, da sie vom unbekanntem Parameter μ abhängig ist und somit der realisierte Wert nicht aus dem Stichprobenergebnis berechnet werden kann. Dagegen ist

$$S^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$$

eine Stichprobenfunktion.

Man beachte, dass eine Stichprobenfunktion als Funktion von Zufallsvariablen auch selbst eine Zufallsvariable ist. Die Verteilung einer Stichprobenfunktion (und insbesondere deren Erwartungswert und Varianz) sind von Interesse und klar zu unterscheiden von der Verteilung der Grundgesamtheit (und deren Erwartungswert und Varianz).

Beispiele für Stichprobenfunktionen:

Stichprobenfunktion	Stichproben	Realisierung
$\bar{X} = \frac{1}{n} \sum_i X_i$	- Mittel	$x = \frac{1}{n} \sum_i x_i$
$S^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$	- Varianz	$s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$
$\hat{p} = \frac{1}{n} \sum_i X_i$	- Anteil	$\hat{p} = \frac{1}{n} \sum_i x_i$

Im letzten Fall wird eine Grundgesamtheit mit Bernoulli-Verteilung unterstellt, d.h. $P(Y = 1) = p$, $P(Y = 0) = 1 - p$, d.h. das Stichprobenmittel ist hier dasselbe wie der Stichprobenanteil der Treffer.

Stichprobenfunktionen werden vor allem verwendet zur

- Schätzung von Parametern der Grundgesamtheit („**Schätzfunktionen**“)
- Entscheidung über Ablehnung oder Annahme von Hypothesen über die Verteilung der Grundgesamtheit („**Teststatistiken**“)

Kapitel 2

Parameterpunktschätzung

In diesem Kapitel geht es um Stichprobenfunktionen, deren Realisierung bei einem gegebenen Stichprobenergebnis als Schätzwert für einen Verteilungsparameter der Grundgesamtheit verwendet werden sollen. Im ersten Abschnitt betrachten wir zwei verschiedene Konstruktionsprinzipien für Schätzfunktionen, die **Methode der Momente** und die **Maximum-Likelihood-Methode**. Anschließend definieren wir Eigenschaften von Schätzfunktionen (**Erwartungstreue, Effizienz und Konsistenz**), die wir als Maßstäbe zur Beurteilung von Schätzfunktionen verwenden können. Im dritten Abschnitt gehen wir dann noch einmal speziell auf die Verteilung von Schätzfunktionen für den Mittelwert μ , eines Anteilswerts bzw. einer Trefferwahrscheinlichkeit p und der Varianz σ^2 der Grundgesamtheit ein.

2.1 Konstruktion von Schätzfunktionen

Die beiden Konstruktionsprinzipien für Schätzfunktionen basieren auf einer Zusammenführung von theoretischer Information (einer Verteilungsannahme) und empirischer Information (Daten aus einer einfachen Zufallsstichprobe):

- (1) Die parametrische Verteilungsannahme legt fest, dass die Verteilung des untersuchten Merkmals Y in der Grundgesamtheit zu einer bestimmten parametrischen Verteilungsfamilie gehört. Die Angemessenheit des speziellen parametrischen Verteilungsmodells wird man in der Regel aus dem Sachzusammenhang begründen. Formal wird mit der parametrischen Verteilungsannahme eine Klasse von möglichen Verteilungsfunktionen $F_Y(\cdot; \theta_1, \dots, \theta_K)$ festgelegt; mit Θ bezeichnen wir diejenige Teilmenge des \mathbb{R}^K , die die zulässigen Parametervektoren $\theta = (\theta_1, \dots, \theta_K)$ enthält. Die Menge Θ wird als **Parameterraum** bezeichnet.

- (2) Die empirische Information wird durch eine einfache Zufallsstichprobe zu der Variablen Y vom Umfang n , (X_1, \dots, X_n) , mit dem Stichprobenergebnis (x_1, \dots, x_n) geliefert.

Bei gegebener Verteilungsannahme sollen aus der Stichprobe Schätzwerte für die unbekannt, wahren Parameter $\theta_1, \dots, \theta_K$ bestimmt werden. Mit den geschätzten Parametern $\hat{\theta}_1, \dots, \hat{\theta}_K$ ist dann die ganze Verteilung von Y geschätzt.

2.1.1 Methode der Momente

Man bezeichnet den Erwartungswert $E(Y^k)$ als das **k-te Moment** der Zufallsvariablen Y , wobei für diskrete bzw. stetige Zufallsvariablen Y gilt:

$$E(Y^k) = \sum_i p_i y_i^k \text{ bzw. } E(Y^k) = \int_{-\infty}^{\infty} y^k f_Y(y) dy.$$

Wir nehmen an, dass dieser Erwartungswert durch die Summe (für diskrete Zufallsvariable) bzw. durch das Integral (für stetige Zufallsvariable) endlich und eindeutig definiert ist. (Es gibt Verteilungen, für die das nicht gilt.) Wenn die Verteilung von Y zu einem parametrischen Verteilungstyp mit K Parametern $\theta_1, \dots, \theta_K$ gehört, dann sind die K Momente $E(Y^1), \dots, E(Y^K)$ abhängig von den Parametern $\theta_1, \dots, \theta_K$. Die Schätzung nach der Methode der Momente erfolgt so:

1. Die Parameter $\theta_1, \dots, \theta_K$ werden dargestellt als Funktionen der Momente $E(Y^1), \dots, E(Y^K)$,

$$\theta_k = g_k(E(Y^1), \dots, E(Y^K)), k = 1, \dots, K.$$

2. Die Schätzfunktionen werden gebildet, indem die theoretischen Momente $E(Y^k)$ durch die entsprechenden Stichprobenmomente $\overline{X^k} = \frac{1}{n} \sum_{i=1}^n X_i^k$ ersetzt werden,

$$\hat{\theta}_k = g_k(\overline{X^1}, \dots, \overline{X^K}), k = 1, \dots, K.$$

Beispiele

- (1) Y ist $B(1; p)$ -verteilt, $0 \leq p \leq 1$ (Dichotome Grundgesamtheit, Anteilsschätzung).

$$(1.1) \quad p = E(Y)$$

$$(1.2) \hat{p} = \bar{X} \text{ (Stichprobenanteil der Treffer)}$$

(2) Y ist poissonverteilt mit dem Parameter λ , $\lambda \in \mathbb{R}^+$.

$$(2.1) \lambda = E(Y)$$

$$(2.2) \hat{\lambda} = \bar{X}$$

(3) Y ist normalverteilt $N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ und $\sigma^2 \in \mathbb{R}^+$

$$(3.1) \begin{aligned} \mu &= E(Y) \\ \sigma^2 &= E(Y^2) - [E(Y)]^2 \end{aligned}$$

$$(3.2) \begin{aligned} \hat{\mu} &= \bar{X} \\ \hat{\sigma}^2 &= \bar{X}^2 - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i\right)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \end{aligned}$$

(4) Y ist stetig gleichverteilt über dem Intervall $[a, b]$, $a, b \in \mathbb{R}$, $a < b$

$$(4.1) \begin{aligned} E(Y) &= \frac{a+b}{2} \\ E(Y^2) - [E(Y)]^2 &= \text{Var}(Y) = \frac{(b-a)^2}{12} \end{aligned}$$

Die Auflösung dieser beiden Gleichungen nach a und b ergibt unter Berücksichtigung von $a < b$:

$$(4.2) \begin{aligned} a &= E(Y) - \sqrt{3(E(Y^2) - [E(Y)]^2)} \\ b &= E(Y) + \sqrt{3(E(Y^2) - [E(Y)]^2)} \\ \hat{a} &= \bar{X} - \sqrt{3(\bar{X}^2 - \bar{X}^2)} \\ \hat{b} &= \bar{X} + \sqrt{3(\bar{X}^2 - \bar{X}^2)} \end{aligned}$$

2.1.2 Maximum Likelihood Schätzung

Einführendes Beispiel: Anteilsschätzung

Betrachte eine dichotome Grundgesamtheit, für die der Anteilswert (bzw. die Trefferwahrscheinlichkeit) p geschätzt werden soll. Die Verteilung der Grundgesamtheit ist also die Verteilung der Indikatorvariablen Y mit $P_Y(Y = 1) = p$ und $P_Y(Y = 0) = 1 - p$, d.h. $Y \sim B(1; p)$. Der Parametervektor θ hat also nur eine Komponente, $\theta = p$, der Parameterraum Θ der zulässigen Werte für p ist das Intervall $[0; 1]$. (X_1, \dots, X_{10}) sei eine einfache Zufallsstichprobe zu $Y \sim B(1; p)$ und das Stichprobenergebnis sei $(1, 0, 0, 0, 1, 0, 0, 1, 0, 0)$. Die Wahrscheinlichkeit für das beobachtete Stichprobenergebnis ist von der unbekanntem Trefferwahrscheinlichkeit p abhängig. Wegen der Unabhängigkeit der Stichprobenvariablen X_1, \dots, X_n ist

$$\begin{aligned}
P\left((X_1, X_2, \dots, X_{10}) = (1, 0, 0, 0, 1, 0, 0, 1, 0, 0) \mid p\right) \\
= P_Y(X_1 = 1|p) \cdot P_Y(X_2 = 0|p) \cdot \dots \cdot P_Y(X_{10} = 0|p) = p^3(1-p)^7.
\end{aligned}$$

Es ist nicht möglich, aus der Stichprobe mit Sicherheit zu schließen, wie groß der wahre Anteil p in der Grundgesamtheit ist. Das beobachtete Stichprobenergebnis ist bei jedem Anteilswert p , $0 < p < 1$ möglich. Die **Maximum-Likelihood-Methode** (Abkürzung: ML-Methode, ML-Schätzung) wählt denjenigen Parameterwert als Schätzwert \hat{p} aus, bei dem die Wahrscheinlichkeit für das beobachtete Stichprobenergebnis am größten ist.

Die Funktion, die jedem Parameterwert p die Wahrscheinlichkeit **für das beobachtete Stichprobenergebnis** zuordnet, wird als **Likelihoodfunktion** $L(p)$ bezeichnet,

$$L : [0, 1] \rightarrow \mathbb{R}, \quad L(p) = p^3(1-p)^7. \quad (2.1)$$

Für den Maximum-Likelihood-Schätzwert \hat{p} muss also gelten:

$L(\hat{p}) = \max\{L(p); 0 \leq p \leq 1\}$. In den Randpunkten 0 und 1 ist $L(p)$ gleich Null und für $0 < p < 1$ ist $L(p) > 0$. Das Maximum liegt also nicht in einem Randpunkt. Die notwendige Bedingung für ein relatives Maximum in p , $0 < p < 1$ ist $L'(p) = 0$,

$$L'(p) = 3p^2(1-p)^7 - 7p^3(1-p)^6 = p^2(1-p)^6(3-10p) = 0, \quad (2.2)$$

woraus mit $0 < p < 1$ folgt $p = 0.3$. In $p = 0.3$ ist die zweite Ableitung der Likelihoodfunktion negativ, d.h. L nimmt in $\hat{p} = 0.3$ das Maximum an.

Da die Likelihoodfunktion L durch das Stichprobenergebnis (x_1, \dots, x_n) festgelegt wird (im betrachteten Beispiel dadurch, dass in $n = 10$ Stichprobenzügen genau 3 Treffer realisiert wurden), wird die Likelihoodfunktion auch ausführlicher notiert mit $L(p; x_1, \dots, x_n)$.

Der Maximum-Likelihood-Schätzwert \hat{p} kann oftmals rechnerisch einfacher bestimmt werden, indem man aus der Likelihoodfunktion die **Log-Likelihoodfunktion** $\ln L$ bildet und deren relatives Maximum bestimmt. Im Beispiel erhält man die Log-Likelihoodfunktion als

$$\ln L : [0, 1] \rightarrow \mathbb{R}, \quad \ln L(p) = \ln(p^3(1-p)^7) = 3 \ln(p) + 7 \ln(1-p). \quad (2.3)$$

Die logarithmische Transformation ist streng monoton, d.h. in dem Maximum-Likelihood-Schätzwert \hat{p} wird auch $\ln L(p)$ maximiert,

$$\ln L(\hat{p}) = \max_{0 < p < 1} \{\ln L(p)\}.$$

Die notwendige Bedingung für ein relatives Maximum ist

$$\frac{d\ln L}{dp} = \frac{3}{p} - \frac{7}{1-p} = 0, \quad (2.4)$$

woraus wieder der ML-Schätzwert $\hat{p} = 0.3$ folgt (die zweite Ableitung ist negativ in $p = 0.3$).

ML-Schätzung bei diskret verteilter Grundgesamtheit

Ausgehend von dem vorangegangenen Beispiel betrachten wir allgemeiner die Maximum-Likelihood-Schätzung in dem Fall einer diskreten Verteilung von Y , die von einem Parameter θ abhängig ist, wobei der Parameterraum Θ die Menge der zulässigen Parameterwerte enthält.

(1) **Verteilungsannahme:**

Y ist eine diskrete Zufallsvariable, deren Wahrscheinlichkeitsfunktion vom Parameterwert θ , $\theta \in \Theta$, abhängig ist,

$$P_Y(Y = a_i | \theta) = p_{i|\theta}, \text{ für die möglichen Werte } a_i, i = 1, 2, 3, \dots$$

(2) **Einfache Stichprobe:**

(X_1, \dots, X_n) ist eine einfache Stichprobe zu Y mit dem Stichprobenergebnis (x_1, \dots, x_n) .

(3) **Likelihoodfunktion:**

$$L(\theta) = L(\theta; x_1, \dots, x_n) = \prod_{i=1}^n P_Y(X_i = x_i | \theta)$$

(4) **Log-Likelihoodfunktion:**

$$\ln L(\theta) = \ln L(\theta; x_1, \dots, x_n) = \sum_{i=1}^n \ln(P_Y(X_i = x_i | \theta))$$

(5) **ML-Schätzwert $\hat{\theta}$:**

Der Maximum-Likelihood-Schätzwert $\hat{\theta}$ wird ermittelt als diejenige reelle Zahl $\hat{\theta} \in \Theta$, die die Likelihoodfunktion maximiert, d.h. für die gilt:

$$L(\hat{\theta}) \geq L(\theta) \text{ für alle } \theta \in \Theta.$$

In den meisten Fällen erhält man das Maximum als ein relatives Maximum, d.h. indem man die Likelihoodfunktion oder die Log-Likelihoodfunktion nach θ ableitet, die erste Ableitung gleich Null setzt und nach θ auflöst,

$$L'(\theta) = 0 \text{ bzw. } \frac{d\ln L}{d\theta} = 0,$$

und überprüft, dass die zweite Ableitung im relativen Extremum negativ ist.

Zwei weitere Beispiele zu diskret verteilter Grundgesamtheit

Beispiel: Poissonverteilte Grundgesamtheit

(1) **Verteilungsannahme:**

$$P_Y(Y = k|\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, 3, \dots$$

(2) **Einfache Stichprobe:**

(X_1, \dots, X_n) ist eine einfache Stichprobe zu Y mit dem Stichprobenergebnis (x_1, \dots, x_n) .

(3) **Likelihoodfunktion:**

$$L(\lambda) = \prod_{i=1}^n P_Y(X_i = x_i|\lambda) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$

(4) **Log-Likelihoodfunktion:**

$$\begin{aligned} \ln L(\lambda) &= \sum_{i=1}^n \ln(P_Y(X_i = x_i|\lambda)) = \sum_{i=1}^n (x_i \ln \lambda - \ln(x_i!) - \lambda) \\ &= \left(\sum_{i=1}^n x_i\right) \ln \lambda - \sum_{i=1}^n \ln(x_i!) - n\lambda \end{aligned}$$

(5) **ML-Schätzwert $\hat{\lambda}$:**

$$\frac{d\ln L}{d\lambda} = \frac{\sum_{i=1}^n x_i}{\lambda} - n = 0,$$

woraus sich nach λ auflöst ergibt:

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x},$$

Beispiel: Diskrete Gleichverteilung der natürlichen Zahlen von 1 bis M

(1) **Verteilungsannahme:**

Der unbekannte Parameter ist die natürliche Zahl M , also die größte Zahl, die die Zufallsvariable Y bei einer Realisierung annehmen kann. Der Parameterraum ist $\Theta = \mathbb{N}$.

$$P_Y(Y = k|M) = \frac{1}{M}, \quad k = 1, 2, 3, \dots, M$$

(2) **Einfache Stichprobe:**

(X_1, \dots, X_6) ist eine einfache Stichprobe vom Umfang $n = 6$ zu Y mit dem Stichprobenergebnis $(x_1, \dots, x_6) = (26, 3, 7, 18, 31, 14)$.

(3) **Likelihoodfunktion:**

$$L(M) = \prod_{i=1}^n P_Y(X_i = x_i|M) = \left(\frac{1}{M}\right)^n \text{ für jedes } M \geq 31.$$

Für jedes $M < 31$ ist $P_Y(X_5 = 31|M) = 0$, also auch $L(M) = 0$.

(4) **Log-Likelihoodfunktion:** Nicht erforderlich.

(5) **ML-Schätzwert \hat{M} :**

Die Likelihoodfunktion $L(M)$ (und ebenso die Log-Likelihoodfunktion) nimmt ihr Maximum in $M = 31$, dem Maximum der Stichprobenwerte, an. Somit ist der ML-Schätzwert $\hat{M} = \max_{i=1, \dots, n} \{x_i\} = 31$.

ML-Schätzung bei stetig verteilter Grundgesamtheit

Die Übertragung des ML-Schätzprinzips von der Anwendung bei diskreter Verteilung auf die Anwendung bei stetiger Verteilung geschieht praktisch einfach dadurch, dass anstelle der Wahrscheinlichkeitsfunktion die Dichtefunktion von Y verwendet wird.

Gegen die Übertragung des ML-Prinzips auf den Fall einer Zufallsvariablen Y , deren stetige Verteilung von einem Parameter θ abhängig ist, könnte man zunächst einwenden, dass jedes Stichprobenergebnis (x_1, \dots, x_n) die Wahrscheinlichkeit 0 besitzt, welchen Wert der Parameter θ auch annimmt. Betrachtet man aber anstelle der Punktwahrscheinlichkeit die Wahrscheinlichkeit

$$\begin{aligned} P\left(x_1 - \frac{\varepsilon}{2} \leq X_1 \leq x_1 + \frac{\varepsilon}{2}, \dots, x_n - \frac{\varepsilon}{2} \leq X_n \leq x_n + \frac{\varepsilon}{2} \mid \theta\right) \\ = \prod_{i=1}^n P\left(x_i - \frac{\varepsilon}{2} \leq X_i \leq x_i + \frac{\varepsilon}{2} \mid \theta\right) \end{aligned}$$

für beliebig kleines $\varepsilon > 0$, dann ist dieses Produkt von Intervallwahrscheinlichkeiten im allgemeinen nicht gleich Null und kann für kleines ε approximiert werden mit dem Produkt

$$\prod_{i=1}^n \varepsilon f_Y(x_i|\theta) = \varepsilon^n \prod_{i=1}^n f_Y(x_i|\theta).$$

Dies ist - in Abhängigkeit vom Parameter θ - die Wahrscheinlichkeit, dass ein Stichprobenergebnis in der Nähe des beobachteten Ergebnisses realisiert wird. Bei der Maximierung bezüglich θ spielt der Faktor ε^n keine Rolle und man braucht nur das Produkt der Dichtefunktionen zu maximieren.

Die ML-Schätzung wird also im Fall einer stetigen Verteilung der Grundgesamtheit folgendermaßen durchgeführt:

(1) **Verteilungsannahme:**

Y ist eine stetige Zufallsvariable, deren Dichtefunktion vom Parameterwert θ , $\theta \in \Theta$, abhängig ist,

$$f_Y(y|\theta), \text{ für } y \in \mathbb{R}$$

(2) **Einfache Stichprobe:**

(x_1, \dots, x_n) ist das Ergebnis einer einfachen Stichprobe zu Y .

(3) **Likelihoodfunktion:**

$$L(\theta) = L(\theta; x_1, \dots, x_n) = \prod_{i=1}^n f_Y(x_i|\theta)$$

(4) **Log-Likelihoodfunktion:**

$$\ln L(\theta) = \ln L(\theta; x_1, \dots, x_n) = \sum_{i=1}^n \ln(f_Y(x_i|\theta))$$

(5) **ML-Schätzwert $\hat{\theta}$:**

Der Maximum-Likelihood-Schätzwert $\hat{\theta}$ wird ermittelt als diejenige reelle Zahl $\hat{\theta} \in \Theta$, die die Likelihoodfunktion maximiert, d.h. für die gilt:

$$L(\hat{\theta}) \geq L(\theta) \text{ für alle } \theta \in \Theta.$$

In den meisten Fällen erhält man auch hier das Maximum als ein relatives Maximum, d.h. indem man die Likelihoodfunktion oder die Log-Likelihoodfunktion nach θ ableitet, die erste Ableitung gleich Null setzt und nach θ auflöst,

$$L'(\theta) = 0 \text{ bzw. } \frac{d \ln L}{d\theta} = 0,$$

und überprüft, dass die zweite Ableitung im relativen Extremum negativ ist.

Zwei Beispiele zu stetig verteilter Grundgesamtheit

Beispiel: Exponentialverteilte Grundgesamtheit

(1) **Verteilungsannahme:**

$$f_Y(y|\lambda) = \lambda e^{-\lambda y}, y \in \mathbb{R}^+ \text{ und } f_Y(y|\lambda) = 0 \text{ für } y < 0.$$

(2) **Einfache Stichprobe:**

(X_1, \dots, X_n) ist eine einfache Stichprobe zu Y mit dem Stichprobenergebnis (x_1, \dots, x_n) .

(3) **Likelihoodfunktion:**

$$L(\lambda) = \prod_{i=1}^n f_Y(x_i|\lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i}$$

(4) **Log-Likelihoodfunktion:**

$$\ln L(\lambda) = \sum_{i=1}^n \ln(f_Y(x_i|\lambda)) = n \ln \lambda - \lambda \sum_{i=1}^n x_i$$

(5) **ML-Schätzwert $\hat{\lambda}$:**

$$\frac{d \ln L}{d \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0,$$

woraus sich nach λ auflöst ergibt:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{x}},$$

Beispiel: Stetige Gleichverteilung über das Intervall $[0, \theta]$

(1) **Verteilungsannahme:**

Der unbekannte Parameter ist die reelle Zahl θ , der Parameterraum ist $\Theta = \mathbb{R}^+$. Die Dichtefunktion von Y ist

$$f_Y(y|\theta) = \begin{cases} \frac{1}{\theta} & \text{für } 0 \leq y \leq \theta, \\ 0 & \text{sonst} \end{cases}$$

(2) **Einfache Stichprobe:**

(X_1, \dots, X_n) ist eine einfache Stichprobe vom Umfang n zu Y mit dem Stichprobenergebnis (x_1, \dots, x_n) .

(3) **Likelihoodfunktion:**

$$L(\theta) = \prod_{i=1}^n f_Y(x_i|\theta) = \left(\frac{1}{\theta}\right)^n \text{ für jedes } \theta \geq \max\{x_i, i = 1, \dots, n\}.$$

Für jedes θ , das kleiner als der größte realisierte Stichprobenwert ist, ist die Dichtefunktion in diesem Wert gleich Null, also auch $L(\theta) = 0$.

(4) **Log-Likelihoodfunktion:** Nicht erforderlich.

(5) **ML-Schätzwert \hat{M} :**

Die Likelihoodfunktion $L(\theta)$ (und ebenso die Log-Likelihoodfunktion) nimmt ihr Maximum in dem Maximum der Stichprobenwerte an. Somit ist der ML-Schätzwert $\hat{\theta} = \max_{i=1, \dots, n}\{x_i\}$.

2.2 Eigenschaften von Schätzfunktionen

In den betrachteten Beispielen wurden für verschiedene Parameter Schätzfunktionen konstruiert, so z.B.

$\hat{\mu} = \bar{X}$ als Schätzfunktion für das Mittel μ der Grundgesamtheit.

$\hat{\sigma}^2 = \overline{X^2} - \bar{X}^2$ als Schätzfunktion für die Varianz σ^2 der Grundgesamtheit.

$\hat{\lambda} = \bar{X}$ mit einer einfachen Stichprobe aus einer poissonverteilten Grundgesamtheit als Schätzfunktion für den Parameter λ .

$\hat{\lambda} = 1/\bar{X}$ mit einer einfachen Stichprobe zu einer exponentialverteilten Grundgesamtheit als Schätzfunktion für den Parameter λ .

$\hat{a} = \bar{X} - \sqrt{3(\overline{X^2} - \bar{X}^2)}$ und $\hat{b} = \bar{X} + \sqrt{3(\overline{X^2} - \bar{X}^2)}$ als Schätzfunktionen für die Parameter a, b einer über dem Intervall $[a, b]$ stetig gleichverteilten Zufallsvariablen (konstruiert mit der Methode der Momente).

$\hat{\theta} = \max_{i=1, \dots, n}\{X_i\}$ als Schätzfunktion für den Parameter θ einer über dem Intervall $[0, \theta]$ stetig gleichverteilten Zufallsvariablen (konstruiert mit der Maximum-Likelihood-Methode).

Als Stichprobenfunktionen sind alle diese Schätzfunktionen Zufallsvariablen. Für ein beobachtetes Stichprobenergebnis (x_1, \dots, x_n) ergibt sich jeweils ein numerischer Schätzwert als Realisierung der Schätzfunktion. Im allgemeinen ist nicht mit einer Übereinstimmung zwischen dem wahren Parameterwert θ und dem realisierten Schätzwert $\hat{\theta}$ zu rechnen. Die Differenz zwischen

Schätzwert und wahren Parameterwert bezeichnen wir als **Stichprobenfehler**. Die Eigenschaften von Schätzfunktionen, die im Folgenden betrachtet werden, beziehen sich **nicht** auf den jeweiligen Stichprobenfehler, der bei einem gegebenen Stichprobenergebnis von der Schätzfunktion realisiert wird (wenn man den Stichprobenfehler bei einem speziellen Stichprobenergebnis bestimmen könnte, dann könnte man den Fehler direkt korrigieren). Sie beziehen sich auf die **Verteilung** des Stichprobenfehlers bzw. der Schätzfunktion, insbesondere auf Erwartungswert und Varianz der Schätzfunktion, $E(\hat{\theta})$ und $Var(\hat{\theta})$.

2.2.1 Erwartungstreue bzw. unverzerrte Schätzfunktionen

Eine Schätzfunktion $\hat{\theta} = g(X_1, \dots, X_n)$ für den Parameter θ heißt **erwartungstreu** (auch: **unverzerrt**), wenn $E(\hat{\theta}) = \theta$.

Die Differenz zwischen dem Erwartungswert $E(\hat{\theta})$ und dem wahren Parameter θ , $E(\hat{\theta}) - \theta$, heißt **Verzerrung** oder auch **Bias** der Schätzfunktion $\hat{\theta}$. Offensichtlich ist der Bias einer erwartungstreuen bzw. unverzerrten Schätzfunktion gleich Null.

Während der Stichprobenfehler $\hat{\theta} - \theta$ eine Zufallsvariable ist, die unterschiedliche Werte jeweils in Abhängigkeit vom Stichprobenergebnis realisiert, ist der Bias bzw. die Verzerrung gleich dem Erwartungswert des Stichprobenfehlers.

Beispiele

- (1) $\hat{\mu} = \bar{X}$ ist eine erwartungstreue Schätzfunktion für das Mittel μ der Grundgesamtheit, denn

$$E(\hat{\mu}) = E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu.$$

- (2) $\overline{X^2}$ ist eine erwartungstreue Schätzfunktion für das zweite Moment $E(Y^2)$ der Grundgesamtheit, denn

$$E(\overline{X^2}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i^2\right) = \frac{1}{n} \sum_{i=1}^n E(X_i^2) = E(Y^2).$$

- (3) Wenn zwischen einem Parameter θ und dem Erwartungswert $\mu = E(Y)$ der Grundgesamtheit eine lineare Beziehung besteht, dann ist die Schätzung von θ nach der Methode der Momente (also durch Ersetzung von μ durch \bar{X}) erwartungstreu.

- (4) Für eine dichotome Grundgesamtheit, $Y \sim B(1; p)$, gilt für den Anteilswert p die Beziehung $p = E(Y) = \mu$. Nach (3) ist also der Stichprobenanteil $\hat{p} = \bar{X}$ eine erwartungstreue Schätzfunktion für p .
- (5) In einer poissonverteilten Grundgesamtheit gilt mit (3) für den Parameter λ wegen $\lambda = E(Y) = \mu$, dass $\hat{\lambda} = \bar{X}$ erwartungstreu für λ ist.
- (6) Die Schätzfunktion für λ aus einer Stichprobe zu einer exponentialverteilten Grundgesamtheit, in der $E(Y) = 1/\lambda$, also $\lambda = 1/\mu$ gilt, ergibt sich nach der Methode der Momente als $\hat{\lambda} = 1/\bar{X}$. Sie ist nicht erwartungstreu.
- (7) Neben $\hat{\mu} = \bar{X}$ ist auch jedes gewogene arithmetische Mittel der Stichprobenwerte $\check{\mu} = \sum_{i=1}^n w_i X_i$ mit Gewichten w_i , die sich auf 1 summieren, $\sum_{i=1}^n w_i = 1$, eine erwartungstreue Schätzfunktion für das Mittel μ der Grundgesamtheit, denn

$$E(\check{\mu}) = E\left(\sum_{i=1}^n w_i X_i\right) = \sum_{i=1}^n w_i E(X_i) = \left(\sum_{i=1}^n w_i\right)\mu = \mu.$$

Da $\check{\mu}$ linear in den Stichprobenvariablen ist, wird jedes $\check{\mu}$ auch als eine **lineare erwartungstreue Schätzfunktion** für μ bezeichnet.

- (8) Die Schätzung der Varianz $\sigma^2 = E(Y^2) - (E(Y))^2$ der Grundgesamtheit mit der Schätzfunktion nach der Methode der Momente,

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

ist nicht erwartungstreu. Für den Erwartungswert $E(\hat{\sigma}^2)$ erhalten wir

$$\begin{aligned} E(\hat{\sigma}^2) &= E(\overline{X^2}) - E(\bar{X}^2) \\ &= E(Y^2) - E\left(\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n X_i X_j\right) \\ &= \sigma^2 + \mu^2 - \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n E(X_i X_j) \\ &= \sigma^2 + \mu^2 - \frac{1}{n^2} (n^2 \mu^2 + n \sigma^2) \\ &= \sigma^2 - \frac{1}{n} \sigma^2 \\ &= \frac{n-1}{n} \sigma^2 \end{aligned}$$

Bei der Umformung haben wir verwendet, dass für $i \neq j$ wegen der Unabhängigkeit der Stichprobenvariablen gilt $E(X_i X_j) = \mu^2$, während wir für $i = j, i = 1, \dots, n$ erhalten $E(X_i X_i) = E(Y^2) = \mu^2 + \sigma^2$.

- (9) Offensichtlich erhält man eine erwartungstreue Schätzfunktion für die Varianz σ^2 der Grundgesamtheit, indem man $\hat{\sigma}^2$ mit $\frac{n}{n-1}$ multipliziert. D.h. die Stichprobenvarianz S^2 ,

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

ist eine erwartungstreue Schätzfunktion für σ^2 .

2.2.2 Effizienz

Wenn es für einen Parameter θ unterschiedliche erwartungstreue Schätzfunktionen gibt, dann vergleichen wir die Qualität dieser Schätzfunktionen anhand ihrer Varianz. Die Varianz einer Schätzfunktion $\hat{\theta}$ ist zugleich die Varianz des Stichprobenfehlers $\hat{\theta} - \theta$. Die Wurzel der Fehlervarianz wird auch als **Standardfehler** der Schätzung bezeichnet. Der Standardfehler einer erwartungstreuen Schätzfunktion ist ein Maß für die Genauigkeit bzw. Unsicherheit der Schätzung. Wir ziehen diejenige Schätzfunktion vor, die einen kleineren Standardfehler, d.h. eine kleinere Varianz hat. Von zwei erwartungstreuen Schätzfunktionen $\hat{\theta}, \check{\theta}$ für θ bezeichnen wir $\hat{\theta}$ als **relativ effizient** (auch: effizienter, wirksamer) im Vergleich zu $\check{\theta}$, wenn $Var(\hat{\theta}) < Var(\check{\theta})$.

Beispiel

Nach Beispiel (7) aus dem vorigen Abschnitt ist neben $\hat{\mu} = \bar{X}$ auch jedes gewogene arithmetische Mittel der Stichprobenwerte $\check{\mu} = \sum_{i=1}^n w_i X_i$ mit Gewichten w_i , die sich auf 1 summieren, $\sum_{i=1}^n w_i = 1$, eine erwartungstreue Schätzfunktion für μ . Wir zeigen, dass $\hat{\mu} = \bar{X} = \sum_{i=1}^n \frac{1}{n} X_i$ in der Klasse der linearen erwartungstreuen Schätzfunktionen für μ effizient ist.

Sei $\check{\mu} = \sum_{i=1}^n w_i X_i$ erwartungstreue Schätzfunktion für μ und $\check{\mu} \neq \hat{\mu}$, also

$$w_i = \frac{1}{n} + \varepsilon_i \text{ mit } \sum_{i=1}^n \varepsilon_i = 0 \text{ und } \varepsilon_i \neq 0 \text{ für mindestens ein } i.$$

Dann gilt für die Varianz von $\check{\mu}$:

$$\begin{aligned} Var(\check{\mu}) &= Var(\sum_{i=1}^n w_i X_i) \\ &= \sum_{i=1}^n w_i^2 Var(X_i) \\ &= (\sum_{i=1}^n w_i^2) \sigma^2 \\ &= (\sum_{i=1}^n (\frac{1}{n} + \varepsilon_i)^2) \sigma^2 \\ &= (\sum_{i=1}^n (\frac{1}{n^2} + \frac{2}{n} \varepsilon_i + \varepsilon_i^2)) \sigma^2 \\ &= \frac{\sigma^2}{n} + 0 + (\sum_{i=1}^n \varepsilon_i^2) \sigma^2 \\ &= Var(\bar{X}) + (\sum_{i=1}^n \varepsilon_i^2) \sigma^2 > Var(\bar{X}) \end{aligned}$$

Das gewöhnliche arithmetische Mittel \bar{X} ist also eine bessere Schätzfunktion für den Mittelwert μ der Grundgesamtheit als andere Mittelwerte der Stichprobenvariablen, die nicht jeder Beobachtung das gleiche Gewicht geben, in dem Sinn, dass \bar{X} den kleinsten Standardfehler hat.

2.2.3 Konsistenz

Die Eigenschaft der **Konsistenz** einer Schätzfunktion zielt auf die Verteilung der Schätzfunktion wenn der Stichprobenumfang n immer größer wird. Da hier das Verhalten der Schätzfunktion in Abhängigkeit vom Stichprobenumfang n betrachtet wird, notieren wir eine Schätzfunktion mit $\hat{\theta}_n$. Die Konsistenz ist eine Eigenschaft, die auch von verzerrten Schätzfunktionen erfüllt werden kann. Wir bezeichnen eine Schätzfunktion $\hat{\theta}_n$ für θ als **konsistent im quadratischen Mittel**, wenn die beiden folgenden Bedingungen erfüllt werden:

$$(1) \lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta$$

$$(2) \lim_{n \rightarrow \infty} Var(\hat{\theta}_n) = 0$$

Die erste Bedingung ist von einer erwartungstreuen Schätzfunktion automatisch erfüllt. Wenn die Schätzfunktion nicht erwartungstreu ist, dann fordert die erste Bedingung, dass der Bias mit wachsendem n gegen Null konvergiert. Diese Bedingung ist zum Beispiel von der Schätzfunktion $\hat{\sigma}^2$, die oben unter Punkt (8) betrachtet wurde, erfüllt. Die zweite Bedingung fordert, dass auch der Standardfehler mit wachsendem Stichprobenumfang gegen Null geht.

2.3 Zur Verteilung der Schätzfunktionen \bar{X}, \hat{p} und S^2

2.3.1 Das Stichprobenmittel \bar{X}

(X_1, \dots, X_n) sei eine einfache Stichprobe zu Y mit $E(Y) = \mu$ und $Var(Y) = \sigma^2$. Unabhängig von einer speziellen Verteilungsannahme für Y gilt für das Stichprobenmittel $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$:

$$E(\bar{X}) = \mu, \tag{2.5}$$

sowie

$$\sigma_{\bar{X}}^2 = Var(\bar{X}) = \frac{\sigma^2}{n}, \quad \sigma_{\bar{X}} = \sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}. \tag{2.6}$$

Der **Standardfehler** $\sigma_{\bar{X}}$ des Stichprobenmittels ist also direkt proportional zur Standardabweichung σ des Merkmals in der Grundgesamtheit und umgekehrt proportional zum Stichprobenumfang n . Um beispielsweise den Standardfehler von \bar{X} um den Faktor $\frac{1}{10}$ zu reduzieren, muss der Stichprobenumfang von n auf $100n$ erhöht werden.

Mit der Tschebyscheffungleichung kann die Mindestwahrscheinlichkeit dafür angegeben werden, dass \bar{X} in ein vorgegebenes **Schwankungsintervall** fällt:

$$P(|\bar{X} - \mu| \leq k \frac{\sigma}{\sqrt{n}}) \geq 1 - \frac{1}{k^2}, \quad (2.7)$$

bzw.

$$P(|\bar{X} - \mu| \leq \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}. \quad (2.8)$$

Beispiel

Haushaltsstichprobe zur Schätzung des mittleren Einkommens. Die Stichprobengröße sei $n = 10000$, und die Standardabweichung des Einkommens in der Grundgesamtheit sei $\sigma = 1500$ [€], der Standardfehler von \bar{X} ist also $\sigma_{\bar{X}} = 15$ [€]. Dann wird (mit $k=2$) das Stichprobenmittel \bar{X} zumindest mit der Wahrscheinlichkeit von $1 - \frac{1}{k^2} = 0.75$ nicht mehr als 30 € vom mittleren Einkommen μ in der Grundgesamtheit abweichen:

$$P(|\bar{X} - \mu| \leq 30) \geq 0.75.$$

Wegen des großen Stichprobenumfangs kommt man aufgrund des Zentralen Grenzwertsatzes zu dem wesentlich schärferen Ergebnis, dass das Stichprobenmittel sogar mit Wahrscheinlichkeit von etwa 95 % in das **Schwankungsintervall** von $\mu - 30$ [€] bis $\mu + 30$ [€] fällt:

$$\bar{X} \overset{\bullet}{\sim} N(\mu, \sigma_{\bar{X}}^2), \text{ mit } \sigma_{\bar{X}} = \frac{1500}{\sqrt{n}} = 15,$$

und

$$P(|\bar{X} - \mu| \leq 30 = 2\sigma_{\bar{X}}) = 2\Phi(2) - 1 = 0.9544.$$

Für die Feststellung der Verteilung von \bar{X} unterscheiden wir allgemein, ob

- (1) die Grundgesamtheit normal verteilt ist,
- (2) die Grundgesamtheit nicht normal verteilt ist, aber der Stichprobenumfang groß ist ($n \geq 30$),

- (3) die Grundgesamtheit nicht normal verteilt ist und der Stichprobenumfang klein ist ($n < 30$).

Im Fall (1) ist \bar{X} **normal verteilt**:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right). \quad (2.9)$$

Im Fall (2) kann die Verteilung von \bar{X} mit der Normalverteilung **approximiert** werden; praktisch verfahren wir wie in Fall (1):

$$\bar{X} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right). \quad (2.10)$$

Im Fall (3) ist die Normalverteilung eine schlechte Approximation. In diesem Fall können Wahrscheinlichkeitsaussagen mit der Tschebyscheff-Ungleichung gemacht werden.

2.3.2 Der Stichprobenanteil \hat{p}

Die Verteilung einer dichotomen Grundgesamtheit, in der ein Anteil p von Elementen mit einer bestimmten Eigenschaft vorliegt bzw. mit einer Erfolgswahrscheinlichkeit p ein Treffer erzielt wird, repräsentieren wir durch die Verteilung der Indikatorvariablen Y , $Y \sim B(1; p)$. Der Stichprobenanteil \hat{p} von Treffern in einer einfachen Stichprobe (X_1, \dots, X_n) ist gegeben mit $\hat{p} = \bar{X}$. Die Anzahl der Treffer, $n\hat{p}$, ist binomialverteilt mit Erwartungswert np und Varianz $np(1-p)$. Für \hat{p} gilt somit:

$$E(\hat{p}) = p \quad (2.11)$$

und für die Varianz bzw. den Standardfehler von \hat{p} folgt

$$\begin{aligned} \text{Var}(\hat{p}) &= \sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}, \\ \sqrt{\text{Var}(\hat{p})} &= \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}. \end{aligned} \quad (2.12)$$

Wenn die Stichprobe ohne Zurücklegen gezogen wird, handelt es sich streng genommen nicht mehr um eine einfache Stichprobe und die Binomialverteilung für $n\hat{p}$ gilt nur approximativ. Das spielt dann eine Rolle, wenn der Umfang N der Grundgesamtheit nicht sehr groß ist im Vergleich zum Stichprobenumfang n . Als eine Regel kann man verwenden, dass bei $N < 20n$ der Standardfehler $\sigma_{\hat{p}}$ korrigiert (d.h. verkleinert) wird um den Faktor:

$$\sqrt{\frac{N-n}{N-1}}. \quad (2.13)$$

Bei hinreichend großem Stichprobenumfang n ,

$$np \geq 5 \text{ und } n(1-p) \geq 5, \quad (2.14)$$

kann die Verteilung von \hat{p} wegen des Zentralen Grenzwertsatzes durch die Normalverteilung approximiert werden

$$\hat{p} \overset{\bullet}{\sim} N(p, \sigma_{\hat{p}}^2). \quad (2.15)$$

Beispiel

Bei der Produktion von Rädern gelte ein Produkt als Ausschuss, wenn der Durchmesser außerhalb vorgegebener Grenzen liegt. Angenommen der normale Ausschussanteil beträgt $p = 5\%$.

Zur Qualitätskontrolle wird eine Stichprobe vom Umfang $n = 200$ geprüft. Wie groß ist bei normalem Produktionsverlauf die Wahrscheinlichkeit, dass mehr als 15 der 200 geprüften Räder schlecht sind (d.h. dass der Ausschussanteil \hat{p} größer als 7.5 % ist)?

Wir haben $np = 10$, $n(1-p) = 190$, also

$$\begin{aligned} \hat{p} &\overset{\bullet}{\sim} N\left(0.05, \frac{0.05 \cdot 0.95}{200}\right), \\ \sigma_{\hat{p}} &= \sqrt{\frac{0.05 \cdot 0.95}{200}} = 0.0154, \\ P\left(\hat{p} > \frac{15}{200}\right) &= 1 - P(\hat{p} \leq 0.075) = 1 - \Phi\left(\frac{0.075 - 0.05}{0.0154}\right) = 0.0526. \end{aligned}$$

Bei normaler Produktion tritt ein Ausschussanteil von 7.5% oder mehr nur mit einer Wahrscheinlichkeit von etwa 5.26% auf.

2.3.3 Die Stichprobenvarianz S^2

Eine **unverzerrte Schätzfunktion** für die Varianz σ^2 in der Grundgesamtheit ist (unabhängig von einer speziellen Verteilungsannahme) die **Stichprobenvarianz** :

$$S^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2. \quad (2.16)$$

Im Folgenden werden wir dann, wenn die Varianz σ^2 der Grundgesamtheit nicht bekannt ist, den Schätzwert s^2 verwenden, den die Stichprobenvarianz bei gegebenem Stichprobenergebnis liefert.

Kapitel 3

Intervall-Schätzung

Im vorangegangenen Kapitel wurden Schätzfunktionen $\hat{\theta} = g(X_1, \dots, X_n)$ für einen Parameter θ betrachtet, die zu jedem Stichprobenergebnis eine reelle Zahl $\hat{\theta} = g(x_1, \dots, x_n)$ als Schätzwert für θ liefern. Um die Unsicherheit der Schätzung anzuzeigen, sollte zu dem Schätzwert $\hat{\theta}$ auch eine Information über den **Standardfehler**, d.h. über die Standardabweichung der Schätzfunktion gegeben werden.

Wenn zum Beispiel für eine Grundgesamtheit mit dem Erwartungswert μ und der Varianz σ^2 aus einer einfachen Stichprobe der Erwartungswert μ mit $\hat{\mu} = \bar{X}$ geschätzt wird, kann man zusätzlich zum Schätzergebnis \bar{x} auch den Standardfehler $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ angeben. Wenn σ nicht bekannt ist, kann man stattdessen den geschätzten Standardfehler $s_{\bar{X}} = s/\sqrt{n}$ unter Verwendung der geschätzten Stichprobenstandardabweichung s angeben. Die Ergänzung des Punktschätzwerts durch den Standardfehler $\sigma_{\bar{X}}$ bzw. $s_{\bar{X}}$ gibt einen wichtigen Hinweis für die Beurteilung des Schätzergebnisses.

Etwas weitergehend betrachten wir im Folgenden die Möglichkeit, als Schätzergebnis für den Parameter θ ein Intervall $[\hat{\theta}_l; \hat{\theta}_u]$ anzugeben, wobei die Intervallgrenzen vom Stichprobenergebnis abhängen sollen, also

$$\hat{\theta}_l = g_l(x_1, \dots, x_n), \quad (3.1)$$

$$\hat{\theta}_u = g_u(x_1, \dots, x_n). \quad (3.2)$$

Als Funktionen der Stichprobenvariablen sind die Intervallgrenzen Zufallsvariablen. Wir bezeichnen ein solches Intervall $[\hat{\theta}_l; \hat{\theta}_u]$ als ein **Konfidenzintervall** zum **Konfidenzniveau** $1 - \alpha$ oder auch zur **Vertrauenswahrscheinlichkeit** $1 - \alpha$, wenn die zufälligen Intervallgrenzen den wahren Parameter

θ mit Wahrscheinlichkeit $1 - \alpha$ einschließen,

$$\begin{aligned} P(g_l(X_1, \dots, X_n) \leq \theta \leq g_u(X_1, \dots, X_n)) &= 1 - \alpha \\ &\text{bzw.} \\ P(\hat{\theta}_l \leq \theta \leq \hat{\theta}_u) &= 1 - \alpha. \end{aligned} \quad (3.3)$$

Bei der Konstruktion der Konfidenzintervalle werden wir außerdem dafür sorgen, dass bei jedem Stichprobenergebnis der Schätzwert für die Untergrenze des Intervalls nicht größer als der Schätzwert für die Obergrenze des Intervalls ist.

Das für ein gegebenes Stichprobenergebnis realisierte Intervall $[\hat{\theta}_l; \hat{\theta}_u]$ wird als **Schätzintervall** oder **Ergebnis der Intervallschätzung** bezeichnet. Da man den wahren Parameter θ typischerweise nicht kennt (sonst bräuchte man ihn nicht zu schätzen), weiß man im Einzelfall auch nicht, ob das Schätzintervall θ einschließt. Bekannt ist nur die Eigenschaft des Verfahrens, mit der festgelegten Vertrauenswahrscheinlichkeit $1 - \alpha$ (bzw. bei vielen wiederholten Stichprobenuntersuchungen: mit der relativen Häufigkeit $1 - \alpha$) Intervalle zu produzieren, die den wahren Parameter überdecken. Als Vertrauenswahrscheinlichkeiten $1 - \alpha$ sind 90 %, 95 %, und 99 % besonders gebräuchlich.

3.1 Konfidenzintervalle für μ

Die Grundgesamtheit sei verteilt mit Erwartungswert $E(Y) = \mu$ und Varianz $Var(Y) = \sigma^2$. Wir beschränken uns auf die beiden folgenden Fälle:

- (1) Große Stichprobe ($n \geq 30$)
- (2) Kleine Stichprobe ($n < 30$) aus normalverteilter Grundgesamtheit

Zu (1):

Für große Stichproben ist \bar{X} approximativ normalverteilt, also

$$P\left(\left|\frac{\bar{X} - \mu}{\sigma_{\bar{X}}}\right| \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha. \quad (3.4)$$

Daraus ergibt sich

$$P\left(|\bar{X} - \mu| \leq \sigma_{\bar{X}} z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha. \quad (3.5)$$

Wenn man die Ungleichung so umformt, dass \bar{X} eingeschlossen wird von einer festen Untergrenze und Obergrenze, dann erhält man das zentrale Schwankungsintervall für das Stichprobenmittel \bar{X} zur Wahrscheinlichkeit $1 - \alpha$. Dies

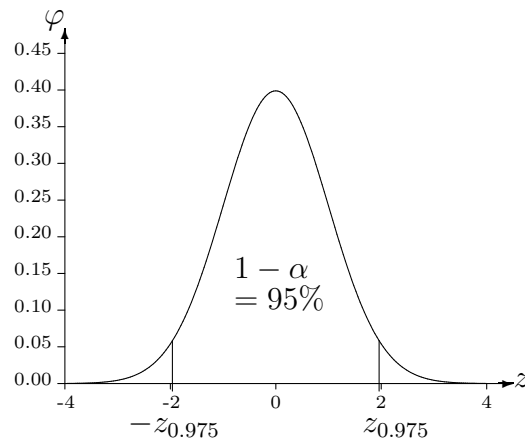


Abbildung 3.1: Standardnormalverteilung

ist zu unterscheiden vom Konfidenzintervall, das man erhält, wenn man die Ungleichung derart umformt, dass der feste (unbekannte) Parameter μ von zufälligen Intervallgrenzen eingeschlossen wird. Das Konfidenzintervall für μ zum Konfidenzniveau $1 - \alpha$ ist dann gegeben mit:

$$\left[\bar{X} - z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{X}}, \bar{X} + z_{1-\frac{\alpha}{2}} \sigma_{\bar{X}} \right]. \quad (3.6)$$

Für $(1 - \alpha)100$ % als 90 %, 95 %, 99 %:

Konfidenzniveau	Konfidenzintervall
90%	$[\bar{X} - 1.645\sigma_{\bar{X}}, \bar{X} + 1.645\sigma_{\bar{X}}]$
95%	$[\bar{X} - 1.960\sigma_{\bar{X}}, \bar{X} + 1.960\sigma_{\bar{X}}]$
99%	$[e\bar{X} - 2.576\sigma_{\bar{X}}, \bar{X} + 2.576\sigma_{\bar{X}}]$

Beachte, dass die Länge des Konfidenzintervalls wesentlich vom Standardfehler $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ und damit vom Stichprobenumfang n abhängt.

Wenn σ unbekannt ist, ersetzen wir bei der Berechnung des Konfidenzintervalls die unbekannt Standardabweichung σ durch s , die Realisierung der Stichprobenstandardabweichung $S = \sqrt{S^2}$,

$$S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2. \quad (3.7)$$

Im Hinblick auf den Ausgangspunkt der Herleitung des Konfidenzintervalls, die Formel (3.4), bedeutet die Verwendung von $S_{\bar{X}}$ anstelle von $\sigma_{\bar{X}}$, dass sich dadurch die Verteilung des standardisierten Stichprobenmittels

$$\frac{\bar{X} - \mu}{S_{\bar{X}}} \quad \text{versus} \quad \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \quad (3.8)$$

verändert. Bei einer großen Stichprobe können wir trotzdem als gute Approximation die Normalverteilung annehmen.

Zu (2): (Kleine Stichprobe ($n < 30$) aus **normalverteilter** Grundgesamtheit)

Bei kleinem Stichprobenumfang gewinnt die Unterscheidung, ob die Standardabweichung σ der Grundgesamtheit bekannt ist oder nicht, größere Bedeutung:

- Wenn σ bekannt ist (was selten der Fall ist), wird das Konfidenzintervall wie im Fall (1) bestimmt und das jeweilige Konfidenzniveau gilt sogar exakt.
- Wenn σ unbekannt ist und durch die Realisierung s der Stichprobenstandardabweichung S ersetzt wird, dann vernachlässigen wir bei kleiner Stichprobengröße den größeren Unterschied der Verteilungen der beiden standardisierten Stichprobenmittel in (3.8) nicht mehr. Während $\bar{X} - \mu$, standardisiert mit dem Standardfehler $\sigma_{\bar{X}}$, standardnormalverteilt ist, gilt bei Standardisierung mit dem geschätzten Standardfehler $S_{\bar{X}}$ die **t-Verteilung mit $n - 1$ Freiheitsgraden**,

$$\frac{\bar{X} - \mu}{S_{\bar{X}}} \sim t(n - 1). \quad (3.9)$$

Die t -Verteilung (und damit auch das Perzentil $t_{1-\frac{\alpha}{2};n-1}$) ist abhängig von der sogenannten Anzahl k der Freiheitsgrade, wobei hier $k = n - 1$ gilt. Mit wachsender Anzahl der Freiheitsgrade k nähert sich die t -Verteilung der Standardnormalverteilung. Allgemein ist die Dichte der t -Verteilung mit k Freiheitsgraden ebenso wie die der Standardnormalverteilung symmetrisch um Null. Die Streuung ist jedoch - durch die zusätzliche Schwankung des Nenners - größer als bei Standardnormalverteilung (vgl. Abbildung 3.2), so dass das Perzentil $t_{1-\frac{\alpha}{2};n-1}$ größer ist als $z_{1-\frac{\alpha}{2}}$. Für die gängigen Werte von α sind die Perzentile für die t -Verteilung mit k Freiheitsgraden tabelliert, insbesondere für kleines k , wo sie sich deutlich von denen der Normalverteilung unterscheiden. Damit erhält man aus

$$P\left(\left|\frac{\bar{X} - \mu}{S_{\bar{X}}}\right| \leq t_{1-\frac{\alpha}{2};n-1}\right) = 1 - \alpha \quad (3.10)$$

das Konfidenzintervall zum Konfidenzniveau $1 - \alpha$ als

$$[\bar{X} - t_{1-\frac{\alpha}{2};n-1}S_{\bar{X}}, \bar{X} + t_{1-\frac{\alpha}{2};n-1}S_{\bar{X}}], \quad (3.11)$$

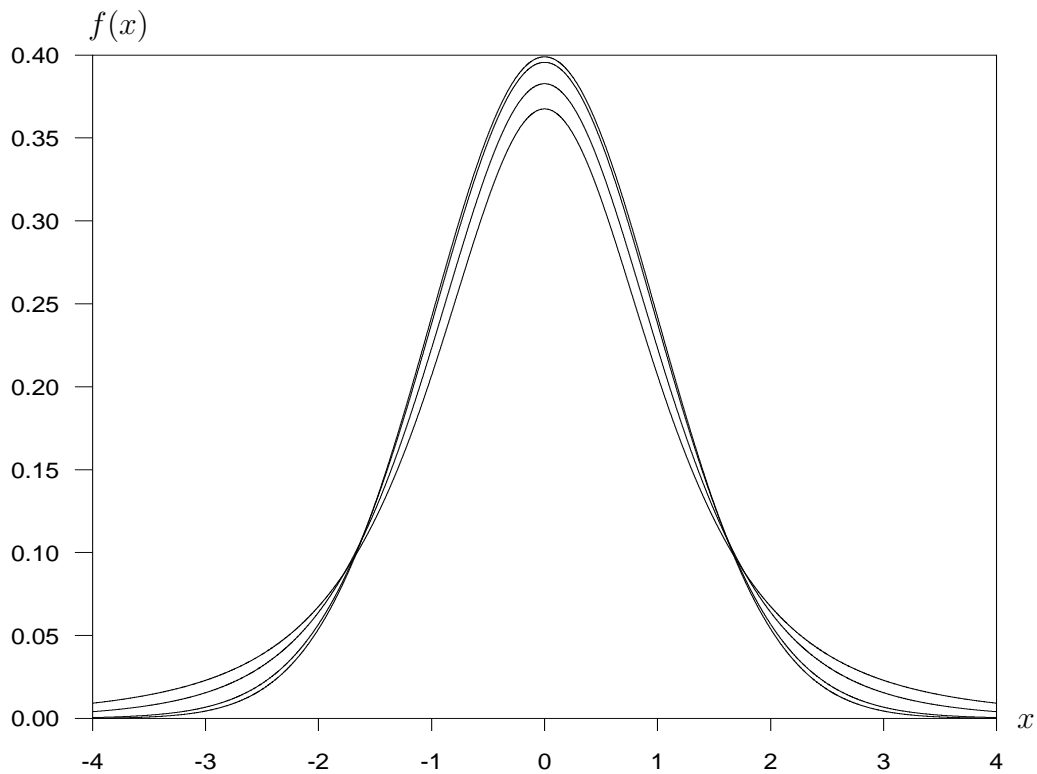


Abbildung 3.2: Dichtefunktionen von t -Verteilungen mit 3, 6 und 30 Freiheitsgraden im Vergleich zur Dichte der Standardnormalverteilung.

wobei $S_{\bar{X}} = \frac{S}{\sqrt{n}}$ und $t_{1-\frac{\alpha}{2}; n-1}$ das $(1 - \frac{\alpha}{2})$ -Perzentil der t -Verteilung mit $n - 1$ Freiheitsgraden ist.

Beispiel

Angenommen das Nettogewicht von 500 g -Müsli-Paketen ist normalverteilt.

Eine einfache Stichprobe ($n = 25$) ergibt: $\bar{X} = 495 [g]$, $s = 5 [g]$.

Mit $s_{\bar{X}} = \frac{s}{\sqrt{n}} = 1$ ist das 95 %-Schätzintervall unter Verwendung von $t_{0.975; 24} = 2.064$:

$$[492.94, 497.06].$$

3.2 Konfidenzintervalle für p (aus großen Stichproben)

Für große Stichproben ($n\hat{p} > 5$, $n(1 - \hat{p}) > 5$) ist der Stichprobenanteil \hat{p} approximativ normalverteilt $N(p, \frac{p(1-p)}{n})$.

Die Standardabweichung von \hat{p} wird geschätzt mit

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}. \quad (3.12)$$

$(1 - \alpha)$ -Konfidenzintervalle für p erhalten wir aus großen Stichproben näherungsweise unter Verwendung der Approximation der Verteilung von \hat{p} mit der Normalverteilung $N(p, s_{\hat{p}}^2)$ als:

$$[\hat{p} - z_{1-\frac{\alpha}{2}} s_{\hat{p}}, \hat{p} + z_{1-\frac{\alpha}{2}} s_{\hat{p}}]. \quad (3.13)$$

Kapitel 4

Hypothesentests auf der Grundlage von Stichprobeninformation

4.1 Einführung: Hypothese und Gegenhypothese, Entscheidung und zwei Fehlertypen

Zunächst einige Beispiele für Fragestellungen, die zur Formulierung eines Testproblems führen können:

- 1) Ist der Produktionsablauf störungsfrei?
- 2) Ist das von einer Großbäckerei angegebene Gewicht eines „500 g-Roggenbrots“ korrekt?
- 3) Hat eine bestimmte Maßnahme die Kundenzufriedenheit erhöht?
- 4) Sind die von einem PKW-Hersteller angegebenen Verbrauchswerte zutreffend?
- 5) Sind Aktienkursänderungen an zwei aufeinanderfolgenden Handelstagen unkorreliert?
- 6) Sind die mittleren Tagesrenditen an einer Börse für jeden Wochentag gleich groß?
- 7) Kann man für die Verteilung der Tagesrendite auf den DAX eine Normalverteilung unterstellen?

- 8) Hat die Ökosteuer den privaten Einsatz des PKW an Wochenenden verändert?
- 9) Tritt eine bestimmte Krankheit bei Beschäftigten einer Branche häufiger auf als in der Gesamtbevölkerung?

Zunächst ist die praktische Fragestellung umzusetzen in ein Entscheidungsproblem zwischen zwei Hypothesen über die Verteilung einer (ggf. mehrdimensionalen) Zufallsvariablen, zu der eine Stichprobe erhoben werden kann bzw. ein Stichprobenergebnis vorliegt. Wir setzen an dieser Stelle ein, d.h. wir gehen davon aus, dass als Hypothesen Aussagen über die Verteilung einer Zufallsvariablen formuliert sind.

4.1.1 Die Hypothesen

Wir unterscheiden zwei Hypothesen, die **Nullhypothese** (H_0) und die **Alternativhypothese** oder **Gegenhypothese** (H_1). Die beiden Hypothesen werden unterschiedlich gewichtet:

H_1 : In der Regel die Aussage, die man statistisch absichern möchte und für deren Akzeptanz man hohe Anforderungen an das Stichprobenergebnis stellt. Die Entscheidung für H_1 hat typischerweise erhebliche Konsequenzen, so dass man das Risiko einer Fehlentscheidung für H_1 kontrollieren will. Gegen H_0 und für H_1 entscheidet man sich nur, wenn man angesichts des Stichprobenbefunds ziemlich sicher ist, d.h. wenn ein Stichprobenergebnis wie das beobachtete Ergebnis bei Gültigkeit von H_0 sehr unwahrscheinlich ist, bei Gültigkeit von H_1 dagegen eher erklärlich ist.

H_0 : In der Regel die Aussage, der man von vornherein eine gewisse Glaubwürdigkeit zuspricht und die man dann als unwiderlegt betrachtet und beibehält, wenn das Stichprobenergebnis (auch) bei Gültigkeit von H_0 nicht überraschend ist.

Beispiel (Gewicht eines „500 g-Roggenbrots“)

Das Brotgewicht wird als normalverteilte Zufallsvariable Y mit dem Erwartungswert μ betrachtet. Eine Verbraucherorganisation will prüfen, ob die Herstellerangabe ($\mu = 500$ g) zutrifft oder das mittlere Gewicht kleiner als 500 g ist. Die Behauptung, dass die Herstellerangabe falsch ist, will man nur dann aufstellen, wenn man ziemlich sicher ist. Also:

$$H_0 : \mu \geq 500 \quad \text{gegen} \quad H_1 : \mu < 500. \quad (4.1)$$

Wenn umgekehrt die Großbäckerei aus Kostengründen wissen will, ob das mittlere Gewicht zu groß ist (um dann gegebenenfalls Einstellungen im Produktionsablauf zu verändern), wird sie das Testproblem anders formulieren, etwa:

$$H_0 : \mu \leq 500 \quad \text{gegen} \quad H_1 : \mu > 500. \quad (4.2)$$

Wenn die Hypothesen Teilmengen aus einer parametrischen Verteilungsfamilie spezifizieren, (wie im Beispiel: H_0 alle Normalverteilungen mit $\mu \leq 500$), spricht man von **parametrischen Tests**. Beispiele für nichtparametrische Tests lernen wir im fünften Kapitel kennenlernen. Zunächst betrachten wir parametrische Tests.

Die Nullhypothese und die Gegenhypothese zerlegen den Parameterraum Θ in zwei Teilmengen Θ_0 und Θ_1 . Dabei wird zugleich die Menge der in Betracht gezogenen Verteilungen für das Merkmal Y in zwei Teilmengen zerlegt. Eine Hypothese, die genau eine Verteilung festlegt, heißt **einfache Hypothese** (z.B. die folgende Nullhypothese bei einem Test auf den Mittelwert einer Normalverteilung mit bekannter Varianz: $H_0 : \mu = 500$, d.h. $\Theta_0 = \{500\}$). Eine Hypothese, die nicht einfach ist, heißt **zusammengesetzte Hypothese**.

Man unterscheidet zwischen **einseitigen** und **zweiseitigen** Tests. Wenn die Parametermenge, die man in der Gegenhypothese H_1 angibt, nur auf **einer Seite** der in H_0 angegebenen Parameter liegt, spricht man von einem **einseitigen Test**. Insbesondere liegt eine **linksseitiger Test** vor, wenn Θ_1 links von Θ_0 liegt (Beispiel: $H_0 : \mu \geq 500$, $H_1 : \mu < 500$) und ein **rechtsseitiger Test**, wenn Θ_1 rechts von Θ_0 liegt (Beispiel: $H_0 : \mu \leq 500$, $H_1 : \mu > 500$). Bei einem **zweiseitigen Test** liegt Θ_1 **zu beiden Seiten** von Θ_0 (Beispiel: $H_0 : \mu = 500$, $H_1 : \mu \neq 500$).

Die beiden durch die Nullhypothese und die Alternativhypothese gegebenen Teilmengen Θ_0 und Θ_1 sind in jedem Fall disjunkt, $\Theta_0 \cap \Theta_1 = \emptyset$. Besonders bei einseitigen Tests ist es aber auch möglich und sogar üblich, dass ein Teil des insgesamt zulässigen Parameterraums ausgeblendet wird und das Hypothesenpaar sich ganz auf den interessierenden Gegensatz konzentriert. So kann man die Hypothesen des linksseitigen Tests der Verbraucherorganisation auch formulieren mit

$$H_0 : \mu = 500 \quad \text{gegen} \quad H_1 : \mu < 500, \quad (4.3)$$

und die Hypothesen des rechtsseitigen Testproblems zur Kostenkontrolle mit

$$H_0 : \mu = 500 \quad \text{gegen} \quad H_1 : \mu > 500. \quad (4.4)$$

4.1.2 Prüfgröße, Entscheidungsregel und zwei mögliche Fehler

Die Entscheidung zwischen den Hypothesen H_0 und H_1 wird in Abhängigkeit vom Stichprobenergebnis getroffen. Im Brotgewicht-Beispiel wird das mittlere Gewicht \bar{X} aus einer einfachen Zufallsstichprobe zu Y eine Rolle spielen. Als **Prüfgröße** oder **Teststatistik** bevorzugt man aber solche Stichprobenfunktionen, die standardisiert sind, so dass man auf Verteilungstabellen zurückgreifen kann. Im Beispiel bietet sich als Teststatistik an:

$$Z = \frac{\bar{X} - 500}{\sigma_{\bar{X}}} = \sqrt{n} \frac{\bar{X} - 500}{\sigma}, \quad \text{wenn } \sigma \text{ bekannt ist} \quad (4.5)$$

und unter Verwendung der Stichprobenstandardabweichung $S = \sqrt{S^2}$ mit $S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$,

$$t = \frac{\bar{X} - 500}{S_{\bar{X}}} = \sqrt{n} \frac{\bar{X} - 500}{S}, \quad \text{wenn } \sigma \text{ unbekannt ist.} \quad (4.6)$$

Unter H_0 , bzw. wenn die Nullhypothese richtig ist mit $\mu = 500$, ist Z standardnormalverteilt bzw. t ist t -verteilt mit $n - 1$ Freiheitsgraden. Je nachdem welche Verteilung die Teststatistik hat, spricht man von einem **Gauß-Test** bei Verwendung der Teststatistik Z und von einem **t-Test** bei Verwendung der t -verteilten Teststatistik.

Vor Beobachtung des Stichprobenergebnisses wird eine Entscheidungsregel festgelegt, nach der dann in Abhängigkeit von der realisierten Prüfgröße entweder

- H_0 abgelehnt und für H_1 entschieden wird, oder
- H_0 nicht abgelehnt und somit beibehalten wird.

Bei dem Hypothesenpaar (4.1), das von der Verbraucherorganisation formuliert wurde,

$$H_0 : \mu \geq 500 \quad \text{gegen} \quad H_1 : \mu < 500$$

wird man H_0 ablehnen, wenn die Prüfgröße einen stark negativen Wert annimmt (d.h. das mittlere Brotgewicht deutlich unterhalb von 500 g liegt), im Fall der Kostenkontrolle (vgl. (4.2)),

$$H_0 : \mu \leq 500 \quad \text{gegen} \quad H_1 : \mu > 500$$

wird man H_0 ablehnen, wenn die Prüfgröße einen stark positiven Wert annimmt (d.h. das mittlere Brotgewicht deutlich oberhalb von 500 g liegt).

Da die Entscheidung auf dem Ergebnis der Zufallsstichprobe basiert, kann sie falsch sein. Je nachdem, welche der beiden Entscheidungen getroffen wurde, gibt es eine andere Fehlermöglichkeit: Den **Fehler erster Art** oder den **Fehler zweiter Art**.

		Tatsächliche Situation:	
		H_0 ist wahr	H_1 ist wahr
Entscheidung:	H_0 wird beibehalten	Korrekt	Fehler 2. Art
	H_0 wird abgelehnt	Fehler 1. Art	Korrekt

Tabelle 4.1: Zwei Fehlerarten

4.1.3 Signifikanz

Wir betrachten zunächst das obige **linksseitige Testproblem** der Verbraucherorganisation zu den Hypothesen

$$H_0 : \mu \geq 500 \quad \text{gegen} \quad H_1 : \mu < 500$$

und nehmen an, dass die Standardabweichung σ des Gewichts Y in dem Produktionsverfahren bekannt ist, d.h. wir verwenden als Teststatistik (4.5),

$$Z = \frac{\bar{X} - 500}{\sigma_{\bar{X}}} = \sqrt{n} \frac{\bar{X} - 500}{\sigma}.$$

Wenn man dann z.B. als Entscheidungsregel festlegt,

$$H_0 \text{ ablehnen, wenn } Z < -2,$$

d.h. H_0 wird abgelehnt, wenn das mittlere Gewicht mehr als zwei Standardfehler unterhalb der Gewichtsangabe von 500 g liegt, dann kann man die (maximale) Wahrscheinlichkeit bestimmen, einen Fehler 1. Art zu machen (also H_0 fälschlich abzulehnen). Diese Wahrscheinlichkeit wird mit α bezeichnet und heißt das **Signifikanzniveau** des Tests; α wird auch als **Irrtumswahrscheinlichkeit** bezeichnet.

Welchen Wert erhält man mit der festgelegten Entscheidungsregel für das Signifikanzniveau α ? Wenn H_0 gilt und genau $\mu = 500$ zutrifft, dann ist die Prüfgröße Z standardnormalverteilt und $P(Z \leq -2) = \Phi(-2) = 0.028$. Wenn sogar $\mu > 500$ gilt, dann wird die Wahrscheinlichkeitsverteilung von Z

noch weiter nach rechts verschoben und die Wahrscheinlichkeit $P(Z \leq -2)$ wird noch kleiner als 2.8 %. Bei der angegebenen Entscheidungsregel ist also die maximale Wahrscheinlichkeit für einen Fehler 1. Art und damit das **Signifikanzniveau** $\alpha = 2.8\%$.

Üblicherweise geht man umgekehrt vor und legt nicht zuerst die Entscheidungsregel fest, um dann das zugehörige Signifikanzniveau α auszurechnen, sondern legt zunächst das Signifikanzniveau α fest, um dann den **kritischen Wert** für die Entscheidungsregel zu bestimmen. Gebräuchlich sind vor allem die Werte 1%, 5% und 10% für das Signifikanzniveau α . In dem betrachteten Beispiel ergeben sich die entsprechenden kritischen Werte und Entscheidungsregeln als:

- $\alpha = 0.01$: H_0 ablehnen, wenn $Z < z_{0.01} = -2.236$
- $\alpha = 0.05$: H_0 ablehnen, wenn $Z < z_{0.05} = -1.645$
- $\alpha = 0.10$: H_0 ablehnen, wenn $Z < z_{0.10} = -1.282$

Zu dem jeweiligen Signifikanzniveau legt die entsprechende Entscheidungsregel einen sogenannten **Ablehnungsbereich** fest: Wenn die Prüfgröße in den Ablehnungsbereich fällt, dann ist H_0 abzulehnen.

Die Verbraucherorganisation wird also zunächst ein Signifikanzniveau festlegen (etwa: $\alpha = 0.05$). Wenn dann die Prüfgröße bei der erhobenen Stichprobe kleiner als -1.645 ist, wird sie behaupten, dass das Brotgewicht signifikant (deutlich) kleiner als 500 g ist. Es ist möglich, dass diese Behauptung falsch ist, aber das kommt bei dem angewandten Verfahren durchschnittlich nur einmal bei 20 Stichprobenerhebungen vor.

Die Wahl des Signifikanzniveaus (also der Wahrscheinlichkeit, H_0 irrtümlich zu verwerfen) sollte von den Kosten einer fälschlichen Ablehnung der Nullhypothese abhängen. Bei sehr hohen Kosten wird man α sehr klein wählen. Wie wir sehen werden, gibt es leider einen Trade-off zwischen den beiden Fehlerwahrscheinlichkeiten. Je kleiner man α wählt, um so eher läuft man Gefahr, H_0 auch dann beizubehalten, wenn H_0 falsch ist.

Zum Verständnis von statistischer Signifikanz ist es wichtig, eine „signifikante Abweichung“ von einer im Sachzusammenhang „bedeutenden Abweichung“ begrifflich abzugrenzen. Angenommen, das Brotgewicht streut mit $\sigma = 5$ g und das mittlere Brotgewicht ist tatsächlich $\mu = 499$ g. Diese Abweichung um 1 g wird man vielleicht nicht als bedeutend ansehen. Wenn aber der

Stichprobenumfang n nur groß genug gewählt wird (etwa: $n = 400$), und damit der Standardfehler $\sigma_{\bar{x}} = 5/\sqrt{n} = 0.25$ [g] sehr klein wird, dann ergibt sich bei einem beobachteten Stichprobenmittel $\bar{x} = 499$ die Prüfgröße $Z = -4$, d.h. die Unterschreitung der Gewichtsangabe (500 g) wird als statistisch hoch signifikant betrachtet.

Umgekehrt wird bei einem geringen Stichprobenumfang n (etwa: $n = 4$ und somit dem Standardfehler $\sigma_{\bar{x}} = 5/\sqrt{n} = 2.5$ [g]) die größere beobachtete Abweichung bei einem Stichprobenmittel von $\bar{x} = 496$ g nicht als statistisch signifikant zu dem Signifikanzniveau $\alpha = 0.05$ angesehen, denn die Prüfgröße Z nimmt dann den Wert $(496 - 500)/2.5 = -1.6$ an, fällt also nicht in den Ablehnungsbereich. Ob ein Stichprobenergebnis in signifikantem Gegensatz zur Nullhypothese steht, hängt nur davon ab, ob es in einen festgelegten „kritischen Bereich“ fällt, der bei Gültigkeit der Nullhypothese sehr unwahrscheinlich ist (nur mit Wahrscheinlichkeit α realisiert wird).

4.1.4 Die Gütefunktion und die Wahrscheinlichkeit für den Fehler 2. Art

Das Signifikanzniveau α - die maximale Wahrscheinlichkeit für den Fehler 1. Art - wird vorgegeben und damit wird der Ablehnungsbereich für die Prüfgröße festgelegt. Die Verteilung der Prüfgröße hängt von dem tatsächlichen Wert des unbekanntem Parameters (im Beispiel: μ) ab. Wenn der tatsächliche Wert mit dem in der Nullhypothese gegebenen (Rand-) Wert 500 g übereinstimmt, dann fällt die Prüfgröße nach Konstruktion mit Wahrscheinlichkeit α in den Ablehnungsbereich. Wenn dagegen μ kleiner als 500 g ist, verschiebt sich die Verteilung der Prüfgröße nach links und die Wahrscheinlichkeit, dass die Prüfgröße in den Ablehnungsbereich fällt, wird größer als α . Ideal wäre natürlich ein Test, der bei festgelegter Irrtumswahrscheinlichkeit α eine möglichst große Ablehnungswahrscheinlichkeit hat, wenn $\mu < 500$ g und somit H_0 falsch ist.

Wenn σ bekannt ist, kann man die Wahrscheinlichkeit, dass die Prüfgröße in den festgelegten Ablehnungsbereich fällt, für jeden hypothetischen Wert von μ ausrechnen. Die entsprechende Funktion $G(\mu)$ wird als **Gütefunktion** des Tests bezeichnet.

Im Beispiel des **linksseitigen Tests** mit den Hypothesen

$$H_0 : \mu \geq 500 \quad \text{gegen} \quad H_1 : \mu < 500,$$

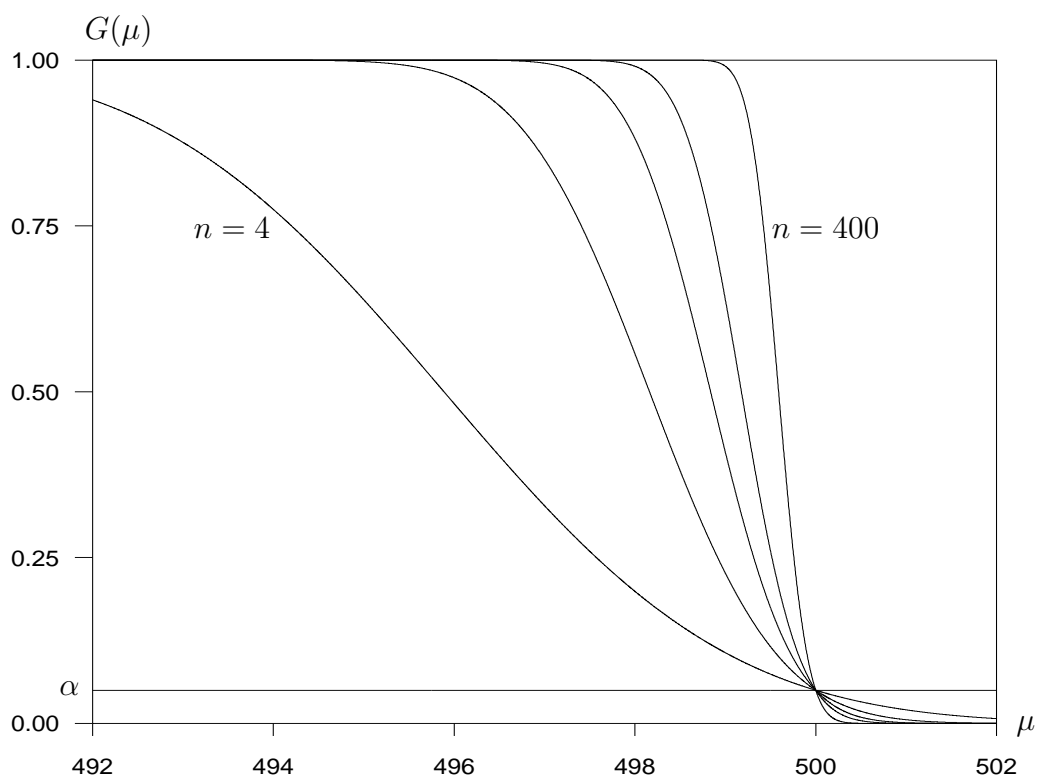


Abbildung 4.1: Gütefunktionen (Wahrscheinlichkeit der Ablehnung von H_0 als Funktion von μ) zum linksseitigen Test für $n = 4, 20, 50, 100$ und 400 .

ergibt sich

$$\begin{aligned}
 G(\mu) = P(Z < -1.645) &= P\left(\sqrt{n} \frac{\bar{X} - 500}{\sigma} < -1.645\right) \\
 &= P\left(\sqrt{n} \frac{\bar{X} - \mu + \mu - 500}{\sigma} < -1.645\right) \\
 &= P\left(\sqrt{n} \frac{\bar{X} - \mu}{\sigma} < \sqrt{n} \frac{500 - \mu}{\sigma} - 1.645\right) \\
 &= \Phi\left(-1.645 + \sqrt{n} \frac{500 - \mu}{\sigma}\right). \quad (4.7)
 \end{aligned}$$

Wie man an der Formel und in Abbildung 4.1 sieht, steigt für $\mu < 500$ die Wahrscheinlichkeit für die Ablehnung von H_0 mit zunehmendem Abstand zwischen μ und 500, und zwar um so steiler, je größer der Stichprobenumfang n ist.

Die Wahrscheinlichkeit für einen Fehler 2. Art, also nicht zu erkennen, dass $\mu < 500$ g und somit H_0 falsch ist, erhält man als die Gegenwahrscheinlichkeit

$1 - G(\mu)$. Man bezeichnet die **Wahrscheinlichkeit für den Fehler 2. Art** mit β , bzw. genauer $\beta(\mu)$, also

$$\beta(\mu) = 1 - G(\mu). \quad (4.8)$$

Im Folgenden werden wir die Durchführung des Gauß-Tests auf den Mittelwert einer Grundgesamtheit für den zweiseitigen Test, den linksseitigen und den rechtsseitigen Test darstellen. Danach kommen wir auch auf die Gütefunktion für diese Tests zurück.

4.2 Gauß-Test für den Mittelwert μ einer Grundgesamtheit

In diesem Abschnitt gehen wir davon aus, dass **zumindest** eine der beiden folgenden Voraussetzungen erfüllt ist:

Voraussetzung 1: Die Grundgesamtheit ist normalverteilt, $Y \sim N(\mu; \sigma^2)$, und σ ist bekannt.

Voraussetzung 2: Der Stichprobenumfang ist groß ($n \geq 30$).

Wenn nur die zweite Voraussetzung erfüllt ist, dann wird die Anwendung des Gauß-Tests mit dem zentralen Grenzwertsatz begründet und gilt nur approximativ.

Für das Stichprobenmittel \bar{X} aus einer einfachen Stichprobe zu Y gilt (unter Voraussetzung 1 exakt, sonst nur approximativ):

$$\bar{X} \sim N(\mu; \sigma_{\bar{X}}^2), \quad (4.9)$$

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}. \quad (4.10)$$

Das standardisierte Stichprobenmittel ist dann standardnormalverteilt,

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \sim N(0; 1). \quad (4.11)$$

Wenn Voraussetzung 2 erfüllt ist und σ^2 unbekannt ist, wird σ durch die Stichprobenstandardabweichung $S = \sqrt{S^2}$ (mit $S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$) ersetzt und man arbeitet mit der Approximation

$$Z = \frac{\bar{X} - \mu}{S_{\bar{X}}} \overset{\circ}{\sim} N(0; 1). \quad (4.12)$$

4.2.1 Zweiseitiger Test

(a) Hypothesen und Signifikanzniveau:

$$H_0 : \mu = \mu_0 \text{ gegen } H_1 : \mu \neq \mu_0 ; \text{ Signifikanzniveau: } \alpha. \quad (4.13)$$

(b) Teststatistik:

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \text{ oder } Z = \frac{\bar{X} - \mu_0}{S_{\bar{X}}} = \sqrt{n} \frac{\bar{X} - \mu_0}{S} \quad (4.14)$$

(c) Bei Gültigkeit von H_0 gilt:

$$Z \sim N(0; 1). \quad (4.15)$$

(d) Entscheidungsregel:

Bei dem festgelegten Signifikanzniveau α wird H_0 abgelehnt, wenn für den Stichprobenwert z der Teststatistik Z gilt

$$|z| > z_{1-\frac{\alpha}{2}}. \quad (4.16)$$

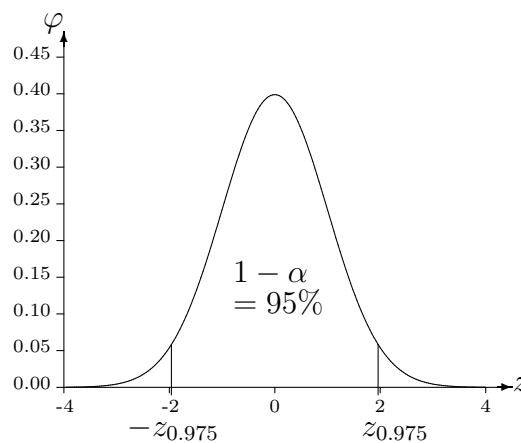


Abbildung 4.2: Zweiseitiger Gauß-Test

Der **Ablehnungsbereich** liegt auf zwei Seiten:

$$z > z_{1-\frac{\alpha}{2}} \text{ (rechte Seite),} \quad (4.17)$$

$$z < -z_{1-\frac{\alpha}{2}} \text{ (linke Seite).} \quad (4.18)$$

Die Werte $z_{1-\frac{\alpha}{2}}$ und $-z_{1-\frac{\alpha}{2}}$ heißen **kritische Werte**.

Beispiel: Qualitätskontrolle

Wenn sich die mittlere Länge von Stahlstiften in einer Stichprobe vom Umfang $n = 64$ signifikant ($\alpha = 0.05$) vom Sollwert $\mu = 10$ [cm] unterscheidet, dann soll die Produktion gestoppt werden, um die Einstellungen zu überprüfen. Angenommen die 64 Messwerte der Stichprobe ergeben $\bar{x} = 9.7$ [cm], $s = 4$ [cm].

Schritt 1: Hypothesen: $H_0 : \mu = 10$ gegen $H_1 : \mu \neq 10$. Signifikanzniveau: $\alpha = 0.05$.

Schritt 2: Die Teststatistik ist

$$Z = \frac{\bar{X} - 10}{S_{\bar{X}}} \quad \text{mit } S_{\bar{X}} = \frac{S}{\sqrt{n}}. \quad (4.19)$$

Bei Gültigkeit von H_0 ist die Teststatistik wegen $n = 64$ (große Stichprobe) approximativ standardnormalverteilt.

Schritt 3: Kritische Werte und Ablehnungsbereich: Zu $\alpha = 0.05$ ist $z_{1-\frac{\alpha}{2}} = 1.96$. Ablehnung von H_0 wenn $z > 1.96$ oder $z < -1.96$.

Schritt 4: Berechnung des realisierten Werts der Teststatistik:

$$z = \frac{9.7 - 10}{0.4/8} = \frac{-0.3}{0.05} = -6.1 \quad (4.20)$$

Schritt 5: Entscheidung: H_0 wird abgelehnt und die Produktion gestoppt, da

$$z = -6 < -1.96. \quad (4.21)$$

4.2.2 Rechtsseitiger Test

(a) Hypothesen und Signifikanzniveau:

$$H_0 : \mu \leq \mu_0 \quad \text{gegen} \quad H_1 : \mu > \mu_0 ; \quad \text{Signifikanzniveau: } \alpha \quad (4.22)$$

(b) Teststatistik:

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \quad \text{oder} \quad Z = \frac{\bar{X} - \mu_0}{S_{\bar{X}}} = \sqrt{n} \frac{\bar{X} - \mu_0}{S} \quad (4.23)$$

¹ Bei bekanntem σ würde man σ anstelle von s verwenden.

(c) Bei Gültigkeit von H_0 (speziell: $\mu = \mu_0$) gilt:

$$Z \sim N(0; 1). \quad (4.24)$$

(d) Entscheidungsregel:

Bei dem festgelegten Signifikanzniveau α ist der kritische Wert $z_{1-\alpha}$. H_0 wird abgelehnt, wenn für den Stichprobenwert z der Teststatistik Z gilt

$$z > z_{1-\alpha}. \quad (4.25)$$

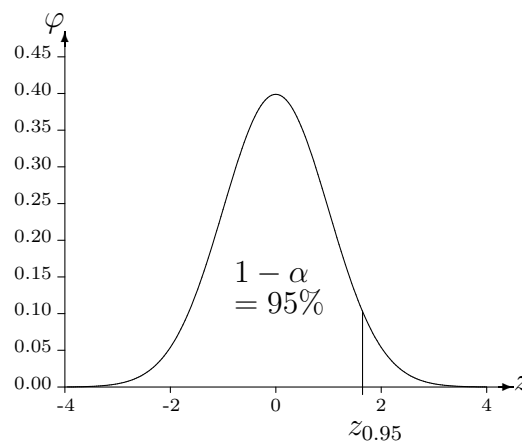


Abbildung 4.3: Rechtseitiger Test

Beispiel: Durchschnittliche Wohnfläche

Nach einer Mitteilung des Statistischen Bundesamtes betrug Anfang 1998 bei Mieterhaushalten die durchschnittliche Wohnfläche je Wohnung im früheren Bundesgebiet rund 71 qm. Mit einer Zufallsstichprobe von 400 Mietwohnungen im früheren Bundesgebiet soll bei einer Irrtumswahrscheinlichkeit von $\alpha = 5\%$ getestet werden, ob die mittlere Wohnfläche Anfang 2003 größer ist als im Jahr 1998. Die Zufallsstichprobe ergibt eine mittlere Wohnfläche von 73 qm und die Stichprobenstandardabweichung $s = 25$ qm, also den Standardfehler $s_{\bar{X}} = 25/20 = 1.25$ qm.

Schritt 1: Hypothesen: $H_0 : \mu \leq 71$ gegen $H_1 : \mu > 71$.
Signifikanzniveau: $\alpha = 0.05$.

Schritt 2: Die Teststatistik ist

$$Z = \frac{\bar{X} - 71}{S_{\bar{X}}} \quad \text{mit } S_{\bar{X}} = \frac{S}{\sqrt{n}}. \quad (4.26)$$

Bei Gültigkeit von H_0 (speziell: $\mu = 71$) ist die Teststatistik wegen $n = 400$ (große Stichprobe) approximativ standardnormalverteilt.

Schritt 3: Kritische Werte und Ablehnungsbereich: Zu $\alpha = 0.05$ ist $z_{1-\alpha} = 1.645$. Ablehnung von H_0 wenn $z > 1.645$.

Schritt 4: Berechnung des realisierten Werts der Teststatistik:

$$z = \frac{73 - 71}{1.25} = 1.6 \quad (4.27)$$

Schritt 5: Entscheidung: Die Teststatistik fällt mit $z = 1.6 < 1.645$ nicht in den Ablehnungsbereich, H_0 wird nicht abgelehnt. Die festgestellte Vergrößerung der mittleren Wohnfläche von 71 auf 73 qm ist nicht signifikant zum 5 %-Niveau.

4.2.3 Linksseitiger Test

(a) Hypothesen und Signifikanzniveau:

$$H_0 : \mu \geq \mu_0 \text{ gegen } H_1 : \mu < \mu_0 ; \text{ Signifikanzniveau: } \alpha \quad (4.28)$$

(b) Teststatistik:

$$Z = \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \sqrt{n} \frac{\bar{X} - \mu_0}{\sigma} \text{ oder } Z = \frac{\bar{X} - \mu_0}{S_{\bar{X}}} = \sqrt{n} \frac{\bar{X} - \mu_0}{S} \quad (4.29)$$

(c) Bei Gültigkeit von H_0 (speziell: $\mu = \mu_0$) gilt:

$$Z \sim N(0; 1). \quad (4.30)$$

(d) Entscheidungsregel:

Bei dem festgelegten Signifikanzniveau α ist der kritische Wert $z_\alpha = -z_{1-\alpha}$. H_0 wird abgelehnt, wenn für den Stichprobenwert z der Teststatistik Z gilt

$$z < -z_{1-\alpha}. \quad (4.31)$$

Beispiel: Verbraucherorganisation testet das mittlere Brotgewicht, vgl. Abschnitt 4.1.

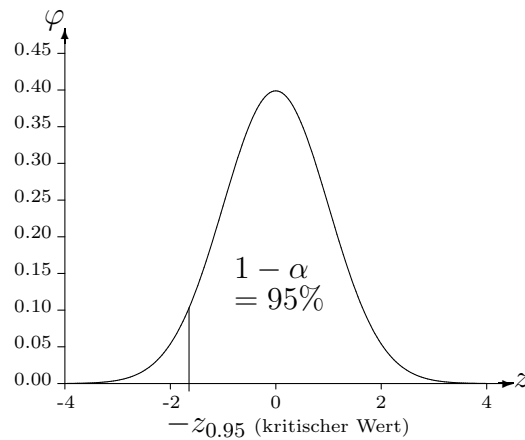


Abbildung 4.4: Linksseitiger Test

4.3 Testen unter Verwendung des p -Werts

In Statistik-Software-Paketen und in der Veröffentlichung von Testergebnissen wird oft ergänzend zum realisierten Wert der Teststatistik oder sogar anstelle dieses Werts der sogenannte p -Wert bzw. das **empirische Signifikanzniveau** angegeben. Darunter versteht man dasjenige Signifikanzniveau p , bei dem der realisierte Wert der Teststatistik gerade kritischer Wert wäre. Anders formuliert ist p gleich der Wahrscheinlichkeit, mit der - bei Gültigkeit der Nullhypothese - der beobachtete Wert der Teststatistik oder sogar noch unplausiblere Werte realisiert werden.

Ein sehr kleiner p -Wert (z.B. $p = 0.005$) bedeutet also, dass entweder H_0 gilt, aber ein dafür sehr untypisches (seltenes) Stichprobenergebnis realisiert wurde oder die Nullhypothese nicht richtig ist. Ein sehr kleiner p -Wert spricht insofern für die Ablehnung der Nullhypothese, genauer:

Äquivalent zur Durchführung des Tests zum Signifikanzniveau α ist die Entscheidungsregel:

$$H_0 \text{ ablehnen} \iff p < \alpha.$$

Bei einem Signifikanzniveau von $\alpha = 5\%$ und einem p -Wert von 0.005 wird also die Nullhypothese abgelehnt. Ein sehr kleiner p -Wert zeigt darüber hinaus an, dass man sich bei der Ablehnung sehr sicher ist (wesentlich sicherer, als das nominale Signifikanzniveau $\alpha = 5\%$ für die Ablehnung verlangt). Wenn dagegen der p -Wert größer ist (etwa: $p = 0.08$), ist festzustellen, dass bei einer zugelassenen Irrtumswahrscheinlichkeit von $\alpha = 5\%$ die Nullhypothese nicht verworfen werden kann, wohl aber, wenn man eine Irrtumswahrscheinlichkeit von $\alpha = 10\%$ in Kauf nimmt.

Beispiel: Reiselustige Senioren - Ausgaben für Pauschalreisen

Im Rahmen der in fünfjährigem Abstand durchgeführten **Einkommens- und Verbrauchsstichprobe** (EVS) werden Haushalte in Deutschland zu sämtlichen Einnahmen und Ausgaben befragt. Im Jahr 2003 ist es wieder soweit (Informationen dazu in der Rubrik „Haushaltsausstattung und -budget“ unter <http://www.destatis.de>). Hier beziehen wir uns auf Ergebnisse von 1998, die vom Statistisches Bundesamt am 15.03.2000 als Mitteilungen für die Presse veröffentlicht wurden und unter <http://www.destatis.de/presse/deutsch/pm2000/p0960024.htm> zu finden sind.

Unter der Position Freizeit, Unterhaltung und Kultur wurden auch Ausgaben für Pauschalreisen nachgewiesen. Darin finden sich alle pauschal gebuchten Reisen von der Tagesexkursion über längere Urlaubsreisen bis zur Safari. Selbst organisierte Reisen sind in dieser Position nicht enthalten. Wie das Statistische Bundesamt aufgrund von Ergebnissen der Einkommens- und Verbrauchsstichprobe mitteilt, geben Senioren einen überdurchschnittlichen Teil ihrer Verbrauchsausgaben für Pauschalreisen aus: Seniorenhaushalte in Deutschland verwendeten im ersten Halbjahr 1998 mehr als 4 % ihrer Ausgaben für den Privaten Verbrauch auf entsprechende Angebote der Reiseveranstalter. Haushalte von Pensionären wendeten für Pauschalreisen 1416 DM auf, Rentnerhaushalte 750 DM. Der Durchschnittswert der Ausgaben für Pauschalreisen aller Haushalte in Deutschland lag im ersten Halbjahr 1998 bei 720 DM und damit bei 3 % der Ausgaben des Privaten Verbrauchs. Zum Vergleich: Im gesamten Jahr 1993 buchten die Haushalte in Deutschland für 660 DM Pauschalreisen.

Nun zu unserem Thema, den p -Werten. Stellen Sie sich vor, dass man zwischen den im 5-jährigen Abstand durchgeführten großen Einkommens- und Verbrauchsstichproben Veränderungstendenzen statistisch abgesichert mit einer kleineren Stichprobe untersuchen möchte. Mit einer Stichprobe von 100 Haushalten soll geprüft werden, ob sich die durchschnittlichen Ausgaben für Pauschalreisen signifikant verändert haben ($H_0 : \mu = 720$ DM gegen $H_1 : \mu \neq 720$ DM, $\alpha = 0.10$). Zum Ergebnis des Tests wird angegeben: $\bar{x} = 740$ DM und ein p -Wert von $p = 0.20$. Interpretation: Eine Abweichung des Stichprobenmittels von $\mu_0 = 720$ DM um 20 DM (wie beobachtet) oder mehr tritt bei Gültigkeit von H_0 mit der Wahrscheinlichkeit von $p = 20\%$ auf. Die Veränderung ist also nicht signifikant zu dem vorgegebenen Niveau ($\alpha = 10\%$), H_0 wird beibehalten. (Spätestens nach Lektüre des restlichen Abschnitts sollten Sie in der Lage sein, aus den gegebenen Informationen die verwendete Standardabweichung der Pauschalreise-Ausgaben als 156 DM zu rekonstruieren!)

Im Folgenden betrachten wir, wie man zum realisierten Wert z der Teststatistik den entsprechenden p -Wert berechnet.

4.3.1 Zweiseitiger Test

Zu dem realisierten Wert z der Teststatistik, $z = (\bar{x} - \mu_0)/s_{\bar{X}}$, wird der p -Wert berechnet als $P(|Z| \geq |z|)$, also:

$$p = 2(1 - \Phi(|z|)). \quad (4.32)$$

Fall I: $|z| \leq z_{1-\frac{\alpha}{2}} \iff p \geq \alpha$ (keine Ablehnung von H_0)

Fall II: $|z| > z_{1-\frac{\alpha}{2}} \iff p < \alpha$ (Ablehnung von H_0)

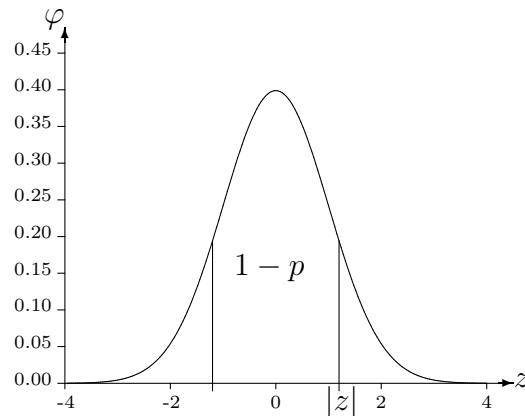


Abbildung 4.5: Fall I, Bsp.: $z = -1.2, \alpha = 0.10$

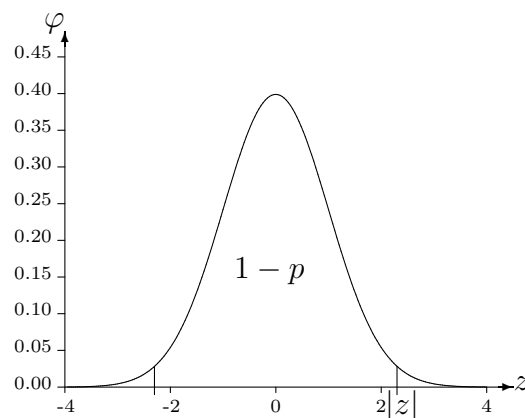


Abbildung 4.6: Fall II, Bsp.: $z = 2.3, \alpha = 0.10$

4.3.2 Rechtsseitiger Test

Beim rechtsseitigen Test ist der p -Wert gegeben durch die Wahrscheinlichkeit, dass die Teststatistik Z - wenn die Nullhypothese mit $\mu = \mu_0$ gilt - größer oder gleich dem realisierten Wert $z = (\bar{x} - \mu_0)/s_{\bar{X}}$ ist. Der p -Wert wird also berechnet als

$$p = 1 - \Phi(z). \quad (4.33)$$

Fall I: $z \leq z_{1-\alpha} \iff p \geq \alpha$ (keine Ablehnung von H_0)

Fall II: $z > z_{1-\alpha} \iff p < \alpha$ (Ablehnung von H_0)

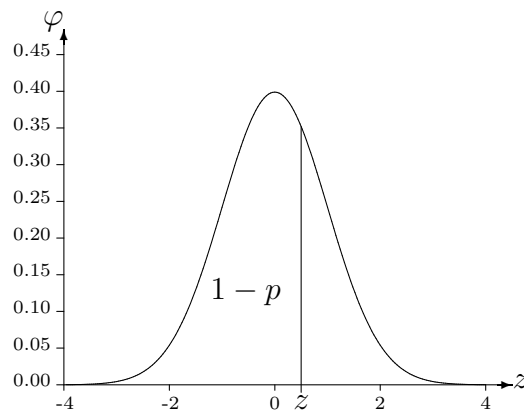


Abbildung 4.7: Fall I, Bsp.: $z = 0.5, \alpha = 0.10$

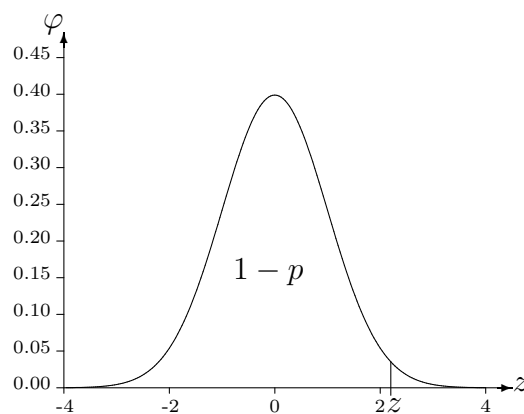


Abbildung 4.8: Fall II, Bsp.: $z = 2.15, \alpha = 0.10$

4.3.3 Linksseitiger Test

Ganz analog ist beim linksseitigen Test der p -Wert gegeben durch die Wahrscheinlichkeit, dass die Teststatistik Z - wenn die Nullhypothese mit $\mu = \mu_0$ gilt - kleiner oder gleich dem realisierten Wert $z = (\bar{x} - \mu_0)/s_{\bar{X}}$ ist. Der p -Wert wird also berechnet als

$$p = \Phi(z). \quad (4.34)$$

Fall I: $z \geq -z_{1-\alpha} \iff p \geq \alpha$ (keine Ablehnung von H_0)

Fall II: $z < -z_{1-\alpha} \iff p < \alpha$ (Ablehnung von H_0)

4.4 t -Test für den Mittelwert μ einer Grundgesamtheit

In diesem Abschnitt gehen wir von der folgenden Voraussetzung aus:

Voraussetzung: Die Grundgesamtheit ist normalverteilt, $Y \sim N(\mu; \sigma^2)$ und σ ist nicht bekannt.

Relevant ist der t -Test vor allem, wenn der Stichprobenumfang klein ist ($n < 30$). Auch bei großem Stichprobenumfang ist der t -Test adäquat, allerdings unterscheidet er sich dann nicht mehr nennenswert vom Gauß-Test. Unter der genannten Voraussetzung gilt für das Stichprobenmittel \bar{X} aus einer einfachen Stichprobe zu Y :

$$\bar{X} \sim N(\mu; \sigma_{\bar{X}}^2), \quad (4.35)$$

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}. \quad (4.36)$$

Das mit dem geschätzten Standardfehler $S_{\bar{X}} = S/\sqrt{n}$ standardisierte Stichprobenmittel, wobei $S = \sqrt{S^2}$ mit

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

ist dann t -verteilt mit $n - 1$ Freiheitsgraden,

$$\frac{\bar{X} - \mu}{S_{\bar{X}}} = \sqrt{n} \frac{\bar{X} - \mu}{S} \sim t(n-1). \quad (4.37)$$

Wenn $\mu = \mu_0$ gilt, ist somit die Teststatistik

$$t = \frac{\bar{X} - \mu_0}{S_{\bar{X}}} = \sqrt{n} \frac{\bar{X} - \mu_0}{S} \quad (4.38)$$

t -verteilt mit $n - 1$ Freiheitsgraden.

Die Durchführung des t -Tests unterscheidet sich von der Durchführung des Gauß-Tests nur in Schritt 2 und in Schritt 3:

Schritt 2: Als Verteilung der Teststatistik stellt man die t -Verteilung mit $n - 1$ Freiheitsgraden fest, wobei n der Stichprobenumfang ist.

Schritt 3: Die kritischen Werte werden der Tabelle der Perzentile der t -Verteilungen entnommen, und zwar

$$\begin{aligned} t_{1-\frac{\alpha}{2}; n-1} & \text{ für den zweiseitigen Test,} \\ t_{1-\alpha; n-1} & \text{ für den rechtsseitigen Test,} \\ -t_{1-\alpha; n-1} & \text{ für den linksseitigen Test.} \end{aligned}$$

4.5 Approximativer Gauß-Test für einen Anteilswert p in der Grundgesamtheit

In einer dichotomen Grundgesamtheit gilt für den Erwartungswert μ der Indikatorvariable Y mit $P(Y = 1) = p$ und $P(Y = 0) = 1 - p$ bekanntlich $\mu = p$. Tests auf den Anteilswert p sind also grundsätzlich zugleich Tests auf den Mittelwert. Wir betrachten wieder die drei Testvarianten:

Zweiseitiger Test: $H_0 : p = p_0$ gegen $H_1 : p \neq p_0$.

Rechtsseitiger Test: $H_0 : p \leq p_0$ gegen $H_1 : p > p_0$.

Linksseitiger Test: $H_0 : p \geq p_0$ gegen $H_1 : p < p_0$.

Da der Stichprobenanteil \hat{p} für große Stichproben - so groß, dass sowohl die erwartete Anzahl der Treffer (np_0) als auch die erwartete Anzahl der Fehlschläge ($n(1 - p_0)$) mindestens 5 ist - approximativ normalverteilt ist, kann approximativ ein Gauß-Test angewandt werden. Dabei wird für die Standardabweichung σ verwendet, dass mit dem hypothetischen Anteilswert p_0 für die Standardabweichung gelten muss $\sigma = \sqrt{p_0(1 - p_0)}$. Damit ergibt sich als Teststatistik:

$$Z = \sqrt{n} \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}}. \quad (4.39)$$

4.6 Verbundene Stichproben: Differenzentest

Während wir bisher Tests auf den Mittelwert **eines** Merkmals in der Grundgesamtheit betrachtet haben (bzw. zuletzt auf einen Anteilswert in einer dichotomen Grundgesamtheit), geht es in diesem Abschnitt um den Vergleich der Mittelwerte zweier Merkmale Y^A und Y^B . Dabei bedeutet der Begriff der **verbundenen Stichprobe** (englisch: **paired sample** oder **matched sample**), dass die beiden Merkmale nicht in zwei unabhängigen Stichproben sondern jeweils paarweise erhoben werden. Der Vergleich der Mittelwerte μ^A und μ^B erfolgt durch die Untersuchung der Differenz $\delta = \mu^A - \mu^B$ mit einer einfachen Stichprobe

$$X_1 = X_1^A - X_1^B, \dots, X_n = X_n^A - X_n^B$$

zur Zufallsvariablen $Y = Y^A - Y^B$. Durch die Differenzenbildung werden zufällige additive Effekte, die sich gleichermäßig auf beide Variablen X_i^A, X_i^B eines Paares (bzw. eines Merkmalsträgers) i auswirken, eliminiert. Beispiele für zwei verbundene Stichproben sind etwa:

- **Vergleich von Ernteerträgen bei verschiedenen Düngemitteln**
Eine verbundene Stichprobe erhält man, wenn man n Äcker jeweils zur Hälfte mit Düngemittel A und zur Hälfte mit Düngemittel B behandelt und die Erträge X_1^A, \dots, X_n^A und X_1^B, \dots, X_n^B erhebt. Effekte der unterschiedlichen Bodenqualität bei verschiedenen Äckern werden, soweit sie additiv den Ernteertrag eines Ackers i verschieben, durch die Differenzenbildung $X_i^A - X_i^B$ ausgeschaltet.
- **Untersuchung des Effekts einer Schulung von Bankangestellten im Schalterdienst**
Bei einer verbundenen Stichprobe würde man etwa die Anzahl der in einer Stunde bedienten Kunden bei n Angestellten **vor** der Schulung (X_i^A) und **nach** der Schulung ($X_i^B, i = 1, \dots, n$) erheben. Durch den Vergleich jeweils bei derselben Person soll der Effekt individueller Leistungsunterschiede reduziert werden. Wenn dagegen die Untersuchung der Kundenbedienung bei zwei unabhängig erhobenen Gruppen von n^A ungeschulten Angestellten und n^B geschulten Angestellten erfolgt, kann der Schulungseffekt durch unterschiedliche Leistungsfähigkeit der beiden Gruppen verwischt werden (was aber auch durch Vergrößerung des Stichprobenumfangs der beiden Zufallsstichproben immer stärker ausgeschaltet würde).
- **Vergleich der Länge des Zeigefingers und des Ringfingers bei n Personen**

Offensichtlich ist $\delta = \mu^A - \mu^B$ der Mittelwert der Differenzvariablen $Y = Y^A - Y^B$. Bei der Untersuchung der Unterschiedlichkeit der beiden Mittelwerte mit Hilfe der Differenzen betrachten wir wieder den zweiseitigen Test und einen einseitigen Test. Den einseitigen Test formulieren wir als rechtsseitigen Test, d.h. wir indizieren dasjenige Merkmal, dessen Mittelwert wir als den signifikant größeren nachweisen wollen (H_1) mit A :

Zweiseitiger Test:

$$H_0 : \mu^A = \mu^B \text{ bzw. } \delta = 0 \quad \text{gegen} \quad H_1 : \mu^A \neq \mu^B \text{ bzw. } \delta \neq 0.$$

Einseitiger Test:

$$H_0 : \mu^A \leq \mu^B \text{ bzw. } \delta \leq 0 \quad \text{gegen} \quad H_1 : \mu^A > \mu^B \text{ bzw. } \delta > 0.$$

Die Durchführung des Differenzentests unter Verwendung der einfachen Stichprobe der **Differenzen** X_1, \dots, X_n verläuft prinzipiell wie bei anderen einfachen Stichproben. Wenn die beiden Merkmale Y^A und Y^B normalverteilt sind (und damit auch deren Differenz Y), kann man die Standardabweichung der Differenzen wie üblich aus der Stichprobe schätzen und einen t -Test durchführen. Unter der Nullhypothese, genauer mit $\delta = 0$, ist die folgende Teststatistik t -verteilt mit $n - 1$ Freiheitsgraden:

$$t = \sqrt{n} \frac{\bar{X}}{S} \quad \text{mit} \quad S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \quad (4.40)$$

Bei großen Stichproben ($n \geq 30$) kann man - auch ohne Normalverteilungsannahme für die Differenzen - approximativ den Gauß-Test unter Verwendung der gleichen Teststatistik anwenden.

Beispiel: Vergleich der Montagezeit bei zwei Verfahren

Für den Zusammenbau eines Werkstücks wird eine neue Vorgehensweise in Betracht gezogen. Es werden $n = 7$ Arbeiter zufällig ausgewählt. Die von den einzelnen Arbeitern benötigte Arbeitszeit bei dem alten Verfahren wird erhoben. Nach einer Unterweisung in das neue Verfahren wird wieder die benötigte Arbeitszeit für den Zusammenbau erhoben. Die folgenden Tabelle gibt die gemessenen Zeiten (in Minuten) an:

Arbeiter	Altes Verfahren	Neues Verfahren	Differenz (x_i)
1	64	60	4
2	71	66	5
3	68	66	2
4	66	69	-3
5	73	63	10
6	62	57	5
7	70	62	8

Es wird angenommen, dass die benötigte Arbeitszeit bei beiden Verfahren normalverteilt ist, womit auch die Differenz normalverteilt ist.

Das neue Verfahren soll eingeführt werden, wenn es die benötigte Arbeitszeit signifikant verringert ($\alpha = 0.05$).

Schritt 1: $H_0 : \mu_{alt} = \mu_{neu}$ (bzw. $\delta = \mu_{alt} - \mu_{neu} = 0$)
gegen $H_1 : \mu_{alt} > \mu_{neu}$ (bzw. $\delta = \mu_{alt} - \mu_{neu} > 0$)
Signifikanzniveau: $\alpha = 0.05$.

Schritt 2: Die Teststatistik (4.40) ist hier

$$t = \sqrt{7} \frac{\bar{X}}{S} \quad \text{mit} \quad S = \sqrt{\frac{1}{6} \sum_{i=1}^7 (X_i - \bar{X})^2}$$

Bei Gültigkeit von H_0 ist die Teststatistik t -verteilt mit 6 Freiheitsgraden.

Schritt 3: Kritische Werte und Ablehnungsbereich: Zu $\alpha = 0.05$ ist $t(6)_{1-\alpha} = 1.943$. Ablehnung von H_0 (d.h. Einführung des neuen Verfahrens), wenn $t > 1.943$.

Schritt 4: Berechnung des realisierten Werts der Teststatistik:

$$s = \sqrt{\frac{7 \cdot \sum_{i=1}^7 x_i^2 - (\sum_{i=1}^7 x_i)^2}{7 \cdot 6}} = \sqrt{\frac{7 \cdot 243 - 31^2}{7 \cdot 6}} = 4.1975$$

$$t = \sqrt{7} \frac{\bar{x}}{s} = \sqrt{7} \frac{31/7}{4.1975} = 2.791$$

Schritt 5: Entscheidung: H_0 wird abgelehnt und das neue Verfahren eingeführt, da

$$t = 2.791 > 1.943.$$

4.7 Beziehung zwischen Konfidenzintervallen und Tests

Im 3. Kapitel wurde das Konzept des Konfidenzintervalls für einen Mittelwert μ (oder einen Anteilswert) zum Konfidenzniveau $1 - \alpha$ eingeführt. Wir haben zweiseitige, symmetrische Konfidenzintervalle um den Punktschätzer $\hat{\mu}$ betrachtet. Wie man leicht zeigen kann, besteht die folgende Beziehung zwischen einem solchen Konfidenzintervall zum Konfidenzniveau $1 - \alpha$ und dem zweiseitigen Test zum Signifikanzniveau α auf den Mittelwert (also mit den Hypothesen $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$):

- Das zu einem Stichprobenergebnis bestimmte Konfidenzintervall zum Konfidenzniveau $1 - \alpha$ stimmt überein mit der Menge aller derjenigen Parameterwerte μ_0 , für die die Nullhypothese des zweiseitigen Tests zum Signifikanzniveau α akzeptiert wird.

Beispiel: Monatliches Nettoeinkommen in Deutschland

Nach Angaben des Statistischen Bundesamts, die auf der letzten Einkommens- und Verbrauchsstichprobe (EVS) von 1998 beruhen, betrug das monatliche Nettoeinkommen pro Haushalt in Deutschland im ersten Halbjahr 1998 im Durchschnitt 5020 DM (Pressemitteilung vom 30. November 1999, vgl. <http://www.destatis.de/presse/deutsch/pm1999/p4100024.htm>). Zur Beurteilung der Lebensverhältnisse in einem bestimmten Wohngebiet wurde mit einer einfachen Stichprobe das durchschnittliche Einkommen pro Haushalt in diesem Wohngebiet untersucht. Die Stichprobe umfasste 36 Haushalte und das Stichprobenergebnis lieferte als mittleres Nettoeinkommen 6650 DM und die Standardabweichung $s = 2300$ DM in der Stichprobe. Damit ergibt sich zum Konfidenzniveau $1 - \alpha = 0.95$, unter Verwendung der Normalverteilung ($z_{0.975} = 1.96$) und des Standardfehlers $s_{\bar{x}} = s/6 = 383$ DM, als Ergebnis der Intervallschätzung das Intervall [5899 DM, 7401 DM].

Kann man aufgrund des vorliegenden Stichprobenergebnisses die Nullhypothese, dass es sich bei dem Wohngebiet um ein „repräsentatives Wohngebiet“ mit $\mu = 5020$ DM handelt, bei einer Irrtumswahrscheinlichkeit von $\alpha = 5\%$ verwerfen? (Ja, denn der Wert $\mu_0 = 5020$ liegt nicht in dem Konfidenzintervall.)

Analog zu dieser Interpretation des zweiseitigen Konfidenzintervalls als die Menge derjenigen Parameterwerte μ_0 , die beim zweiseitigen Test akzeptiert werden, können wir auch ein linksseitiges und ein rechtsseitiges Konfidenzintervall definieren. Das **linksseitige Konfidenzintervall** für μ zum Konfidenzniveau $1 - \alpha$ definieren wir als die Menge derjenigen Parameterwerte

μ_0 , für die die Nullhypothese des linksseitigen Tests

$$H_0 : \mu = \mu_0 \text{ gegen } H_1 : \mu < \mu_0 \quad (4.41)$$

zum Signifikanzniveau α beim Stichprobenmittel \bar{X} akzeptiert wird. Dies ergibt das Intervall $(-\infty; \bar{X} + \frac{\sigma}{\sqrt{n}}z_{1-\alpha}]$.

Beispiel: Für das mittlere Nettohaushaltseinkommens in dem untersuchten Wohngebiet erhält man aus den oben gegebenen Stichprobendaten als Ergebnis eines **linksseitigen Konfidenzintervalls** zum 95 %-Niveau, dass das mittlere Einkommen in diesem Wohngebiet bis zu 7280 DM betragen könnte (denn: $\bar{x} + (s/\sqrt{n})z_{1-\alpha} = 6650 + 383 \cdot 1.645 = 7280$).

Entsprechend ergibt sich das rechtsseitige Konfidenzintervall zum Konfidenzniveau $1 - \alpha$ als das Intervall $[\bar{X} - \frac{\sigma}{\sqrt{n}}z_{1-\alpha}; \infty)$.

Beispiel: Als Ergebnis eines **rechtsseitigen Konfidenzintervalls** zur Vertrauenswahrscheinlichkeit von 95 % erhält man aus den erhobenen Daten, dass das mittlere Einkommen in diesem Wohngebiet oberhalb von 6020 DM liegen sollte (denn: $\bar{x} - (s/\sqrt{n})z_{1-\alpha} = 6650 - 383 \cdot 1.645 = 6020$).

4.8 Chi-Quadrat-Test für die Varianz

Neben dem Erwartungswert μ kann auch die Varianz σ^2 der Verteilung der Grundgesamtheit Gegenstand eines Tests sein. Zum Beispiel könnte in der Qualitätskontrolle interessieren, ob die Streuung einer Länge oder eines Gewichts signifikant größer als ein vorgegebener zulässiger Wert ist; man könnte untersuchen, ob sich die Varianz der täglichen Wechselkursänderungen zwischen Euro und USD mit der Ankündigung eines Wechsels im Amt des EZB-Präsidenten signifikant verändert. Für einen Test auf die Varianz wird der Vergleich der Stichprobenvarianz S^2 mit dem hypothetischen Wert σ_0^2 relevant sein. Als Verteilung der geeigneten Teststatistik ergibt sich die **Chi-Quadrat-Verteilung**, die wir zunächst kurz vorstellen:

Wenn die Zufallsvariablen Z_1, \dots, Z_m unabhängig und jeweils standardnormalverteilt sind, dann bezeichnet man die Verteilung der Summe der quadrierten Zufallsvariablen,

$$\chi^2(m) = Z_1^2 + \dots + Z_m^2 \quad (4.42)$$

als **Chi-Quadrat-Verteilung mit m Freiheitsgraden**. Offensichtlich kann $\chi^2(m)$ keine negativen Werte annehmen und hat den Erwartungswert m ; die

Varianz von $\chi^2(m)$ ist $2m$. Die Prozentpunkte der Chi-Quadrat-Verteilung, die wir als kritische Werte brauchen, z.B.

$$\chi^2(m)_{0.01}, \chi^2(m)_{0.05}, \chi^2(m)_{0.10}, \chi^2(m)_{0.90}, \chi^2(m)_{0.95}, \chi^2(m)_{0.99},$$

sind für verschiedene Freiheitsgrade m tabelliert.

Nach Definition der Chi-Quadrat-Verteilung ist für eine einfache Zufallsstichprobe X_1, \dots, X_n aus einer $N(\mu, \sigma^2)$ verteilten Grundgesamtheit die Summe

$$\frac{(X_1 - \mu)^2}{\sigma^2} + \dots + \frac{(X_n - \mu)^2}{\sigma^2} \quad (4.43)$$

χ^2 -verteilt mit n Freiheitsgraden. Wenn man anstelle von μ das Stichprobenmittel \bar{X} verwendet, kann man (nicht ganz einfach) zeigen, dass die entsprechende Summe χ^2 -verteilt mit $n - 1$ Freiheitsgraden ist. Das heißt,

$$\frac{(X_1 - \bar{X})^2}{\sigma^2} + \dots + \frac{(X_n - \bar{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \quad (4.44)$$

ist χ^2 -verteilt mit $n - 1$ Freiheitsgraden. Wenn man weiterhin anstelle von σ^2 einen hypothetischen Wert σ_0^2 verwendet, erhält man die im Folgenden verwendete Teststatistik, die unter H_0 , d.h. wenn $\sigma^2 = \sigma_0^2$ gilt, also χ^2 -verteilt mit $n - 1$ Freiheitsgraden ist.

Die Anwendung des folgenden χ^2 -Tests für die Varianz ist unter der Voraussetzung zulässig, dass die Grundgesamtheit, aus der eine einfache Stichprobe X_1, \dots, X_n vom Umfang n gezogen wird, (zumindest approximativ) normalverteilt ist. Außerdem gehen wir davon aus, dass μ nicht bekannt ist (sonst würde man natürlich μ verwenden und von n Freiheitsgraden der Teststatistik ausgehen). In der Übersicht des Testverfahrens betrachten wir nebeneinander den zweiseitigen und die beiden einseitigen Tests:

(a) Signifikanzniveau: α

$$\begin{array}{lll} \text{Zweiseitiger Test mit:} & H_0 : \sigma^2 = \sigma_0^2 \text{ gegen} & H_1 : \sigma^2 \neq \sigma_0^2 \\ \text{Rechtsseitiger Test mit:} & H_0 : \sigma^2 \leq \sigma_0^2 \text{ gegen} & H_1 : \sigma^2 > \sigma_0^2 \\ \text{Linksseitiger Test mit:} & H_0 : \sigma^2 \geq \sigma_0^2 \text{ gegen} & H_1 : \sigma^2 < \sigma_0^2. \end{array}$$

(b) Teststatistik:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{1}{\sigma_0^2} \left(\sum_{i=1}^n X_i^2 - \frac{1}{n} \left(\sum_{i=1}^n X_i \right)^2 \right)$$

- (c) Bei Gültigkeit von H_0 (speziell bei $\sigma^2 = \sigma_0^2$) gilt: $\chi^2 \sim \chi^2(n-1)$.
- (d) Entscheidungsregel:
 Bei dem festgelegten Signifikanzniveau α wird H_0 abgelehnt, wenn für den Stichprobenwert χ^2 der Teststatistik gilt:
- Zweiseitiger Test: $\chi^2 < \chi^2(n-1)_{\frac{\alpha}{2}}$ oder $\chi^2 > \chi^2(n-1)_{1-\frac{\alpha}{2}}$
- Rechtsseitiger Test: $\chi^2 > \chi^2(n-1)_{1-\alpha}$
- Linksseitiger Test: $\chi^2 < \chi^2(n-1)_{\alpha}$

Beispiel: Präzision einer neuen Abfüllmaschine

Für eine neue Abfüllmaschine wird geprüft, ob sie präziser als die alte Anlage arbeitet ($\alpha = 0.01$). Bei der alten Maschine beträgt die Standardabweichung des Füllgewichts um den eingestellten Wert 5 g. Bei fester Einstellung des Sollgewichts wird für eine Stichprobe von $n = 20$ jeweils das Füllgewicht gemessen (X_1, \dots, X_n). Es wird angenommen, dass das Füllgewicht normalverteilt ist. Aus dem Stichprobenergebnis wird als Stichprobenvarianz $s^2 = 20.14 \text{ g}^2$, also $s = 4.49 \text{ g}$, berechnet.

Schritt 1: Signifikanzniveau: $\alpha = 0.01$.

$$H_0 : \sigma^2 \geq 25.0 \quad \text{gegen} \quad H_1 : \sigma^2 < 25.0$$

Schritt 2: Die Teststatistik ist

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{19 \cdot S^2}{25}$$

Bei Gültigkeit von H_0 (speziell bei $\sigma^2 = 25$) ist die Teststatistik Chi-Quadrat-verteilt mit 19 Freiheitsgraden.

Schritt 3: Kritische Werte und Ablehnungsbereich:

Zu $\alpha = 0.01$ ist $\chi^2(19)_{\alpha} = 7.633$.

Ablehnung von H_0 (d.h. die neue Maschine arbeitet präziser) wenn $\chi^2 < 7.633$.

Schritt 4: Berechnung des realisierten Werts der Teststatistik:

$$\chi^2 = \frac{19s^2}{25} = 15.31$$

Vgl. dazu auch Abbildung 4.9.

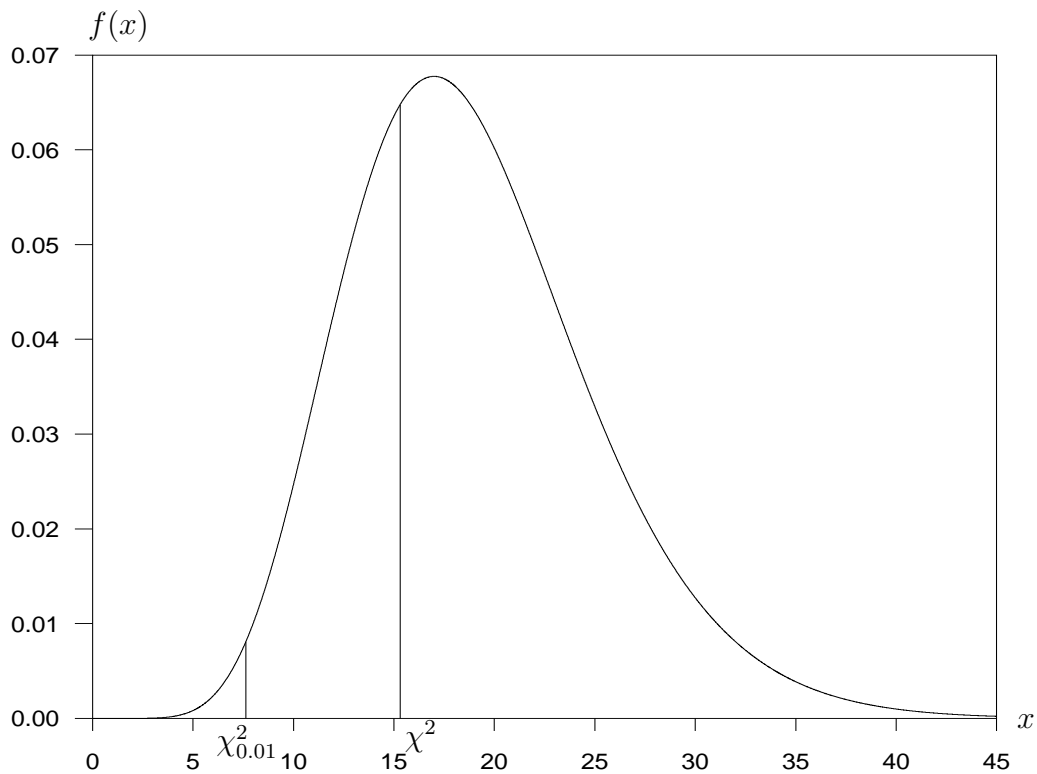


Abbildung 4.9: Dichte der χ^2 -Verteilung mit 19 Freiheitsgraden.

Schritt 5: Entscheidung: H_0 wird nicht abgelehnt, die Varianz des Füllgewichts ist bei der neuen Maschine nicht signifikant kleiner als bei der alten Anlage, der realisierte Wert der Teststatistik, 15.31, liegt **nicht** im Ablehnungsbereich (0; 7.633).

Unter Verwendung der Verteilungsfunktion der χ^2 -Verteilung mit 19 Freiheitsgraden erhält man für die Wahrscheinlichkeit

$$P(\chi^2 \leq 15.31 | H_0) = 29.7\%,$$

d.h. der p-Wert ist hier gleich 29.7 %, das Ergebnis ist fern von einer signifikanten Verbesserung der Genauigkeit der Verpackungsmaschine.

Kapitel 5

Chi-Quadrat-Tests auf Anpassung und Unabhängigkeit

5.1 Vergleich von beobachteten und erwarteten Häufigkeiten: Anpassungstests

Im vorigen Kapitel wurden Tests betrachtet, die eine Entscheidung zwischen alternativen Hypothesen über einen Verteilungsparameter vom beobachteten Stichprobenergebnis abhängig machen. Dabei wurde (insbesondere bei kleinen Stichproben) vorausgesetzt, dass die Verteilung von Y zu einer bestimmten Klasse von Verteilungen gehört (etwa dass Y normalverteilt ist). Im Verlauf dieses Kapitels werden wir eine Möglichkeit kennenlernen, eine solche Verteilungsannahme selbst zur Nullhypothese eines Tests zu machen, also z.B. in Abhängigkeit vom Stichprobenergebnis zu entscheiden, ob die Grundgesamtheit normalverteilt ist. Zunächst kommen wir zurück auf den Test auf einen Anteilswert p , den wir nicht nur als einen Parametertest sondern zugleich als einen Test auf eine hypothetische Verteilung verstehen können.

5.1.1 Anpassung an eine diskrete Verteilung mit k Ausprägungen

Im Fall des Tests auf einen Anteilswert p ist die Verteilungsannahme (dichotom verteilte Grundgesamtheit) unmittelbarer Ausdruck der dichotomen Natur des betrachteten Merkmals Y , das nur die beiden Ausprägungen 1 (Treffer) und 0 (Fehlschlag) hat. Der Test auf einen hypothetischen Anteilswert p_0 ist zugleich der Test auf die hypothetische diskrete Verteilung

$p_0, 1 - p_0$ des Merkmals Y . Für den Fall einer diskreten Verteilung mit nur zwei Ausprägungen haben wir also mit dem Test auf einen Anteilswert p zugleich einen Test, der die Anpassung der Stichprobenverteilung $\hat{p}, 1 - \hat{p}$ an eine hypothetische Verteilung $p_0, 1 - p_0$ untersucht. Wir erweitern die Fragestellung zunächst auf diskrete Verteilungen mit k Ausprägungen, wobei k größer als zwei sein kann, aber nicht sehr groß sein sollte.

Beispiele:

1. Einfache Zufallsstichprobe von Verkäufen von Zigaretten (k Marken): Sind die Marktanteile p_1, \dots, p_k unverändert gegenüber dem Vorjahr (p_1^0, \dots, p_k^0) ?
2. Einfache Zufallsstichprobe von 400 Bestellungen bei einem Versandhandel:
Ist der Eingang von Bestellungen über die fünf Arbeitstage (Montag bis Freitag) gleichverteilt, $(p_1^0, \dots, p_5^0) = (0.2, \dots, 0.2)$?
3. Einfache Zufallsstichprobe von Haushalten:
Hat sich die Einkommensverteilung auf k vorgegebene Einkommensklassen signifikant gegenüber der letzten Einkommens- und Verbrauchsstichprobe des Statistischen Bundesamts geändert?

Wir gehen also davon aus, dass das betrachtete Merkmal Y k **verschiedene Werte** annehmen kann. In den Beispielen:

1. k verschiedene Zigarettenmarken
2. $k = 5$ unterschiedliche Werkzeuge
3. k unterschiedliche Einkommensklassen

Nach einer vorgegebenen hypothetischen Verteilung mit den k Wahrscheinlichkeiten bzw. Anteilswerten $p_1^0, p_2^0, \dots, p_k^0$, die wir als Referenz betrachten, sind beim Stichprobenumfang n die **erwarteten Häufigkeiten** für die k Werte gegeben mit $np_1^0, np_2^0, \dots, np_k^0$. Diese **erwarteten Häufigkeiten** für die k Werte werden mit den **beobachteten Stichprobenhäufigkeiten** n_1, n_2, \dots, n_k verglichen. Als Maß für die Abweichung zwischen der theoretischen und der beobachteten Häufigkeitsverteilung wird die folgende Teststatistik verwendet:

$$\chi^2 := \sum_{i=1}^k \frac{(n_i - np_i^0)^2}{np_i^0}, \quad (5.1)$$

bzw. äquivalent dazu, das n -fache der mittleren quadrierten relativen Abweichung zwischen beobachteten und erwarteten Häufigkeiten

$$\chi^2 = n \sum_{i=1}^k p_i^0 \left(\frac{n_i - np_i^0}{np_i^0} \right)^2.$$

Beispiel: Verkaufszahlen für 6 Zigarettenmarken

Für eine Stichprobe von $n = 400$ verkauften Zigarettenpäckchen werden die nach den bekannten Marktanteilen im Vorjahr erwarteten Verkaufszahlen verglichen mit den in der Stichprobe realisierten Verkaufszahlen der 6 Marken.

Marke	1	2	3	4	5	6
Anteil p_i^0	0.40	0.30	0.10	0.10	0.05	0.05
np_i	160	120	40	40	20	20
n_i	140	130	45	36	23	26

Tabelle 5.1: Erwartete und beobachtete Häufigkeiten

$$\begin{aligned} \chi^2 = & \frac{(140 - 160)^2}{160} + \frac{(130 - 120)^2}{120} + \frac{(45 - 40)^2}{40} \\ & + \frac{(36 - 40)^2}{40} + \frac{(23 - 20)^2}{20} + \frac{(26 - 20)^2}{20} = 6.61 \end{aligned}$$

Ein „großer“ Wert von χ^2 liefert Evidenz gegen die Gültigkeit der spezifizierten Verteilung in der Grundgesamtheit. Um einen kritischen Wert (bei gegebenem Signifikanzniveau) angeben zu können, benötigen wir die Verteilung der Teststatistik bei Gültigkeit der spezifizierten Verteilung. Es kann gezeigt werden, dass die Verteilung der χ^2 -Teststatistik mit wachsendem Stichprobenumfang n gegen die

Chi-Quadrat Verteilung mit $k - 1$ Freiheitsgraden

konvergiert, wenn die Grundgesamtheit tatsächlich der hypothetischen Verteilung $p_1^0, p_2^0, \dots, p_k^0$ genügt. In dem Beispiel der Marktanteile der 6 Zigarettenmarken ist die Teststatistik also bei Gültigkeit der Nullhypothese approximativ $\chi^2(5)$ -verteilt. Damit die Approximation nicht zu schlecht ist, muss der Stichprobenumfang n groß genug sein, dass die erwartete Häufigkeit np_i^0 für jedes $i = 1, \dots, k$ mindestens 5 ist; vgl. unsere entsprechende Bedingung für die Anwendung des Gauß-Tests beim Test auf einen Anteilswert.

Für den speziellen Fall $k = 2$ können wir die Verteilungsaussage für die χ^2 -Teststatistik verifizieren. Mit den Bezeichnungen

$$\begin{aligned} p_1^0 &:= p_0 & \text{und} & & n_1 &= n\hat{p} \\ p_2^0 &:= 1 - p_0 & \text{und} & & n_2 &= n(1 - \hat{p}) \end{aligned}$$

folgt in diesem Fall leicht, dass die χ^2 -Teststatistik identisch ist mit dem **Quadrat** der approximativ **standardnormalverteilten** Teststatistik Z in (4.39), die beim Test auf den Anteilswert p_0 verwendet wurde, denn:

$$\begin{aligned} \chi^2 &= \frac{(n_1 - np_1^0)^2}{np_1^0} + \frac{(n_2 - np_2^0)^2}{np_2^0} \\ &= \frac{(n\hat{p} - np_0)^2}{np_0} + \frac{(n(1 - \hat{p}) - n(1 - p_0))^2}{n(1 - p_0)} \\ &= n \left(\frac{(\hat{p} - p_0)^2}{p_0} + \frac{(p_0 - \hat{p})^2}{1 - p_0} \right) \\ &= n \frac{(\hat{p} - p_0)^2}{p_0(1 - p_0)} \\ &= Z^2 \end{aligned}$$

Nach Definition (4.42) ist die Verteilung von $\chi^2 = Z^2$ die Chi-Quadrat-Verteilung mit 1 Freiheitsgrad.

Die Durchführung des Anpassungstests kann wieder in 5 Schritte gegliedert werden. Jeder Schritt wird jeweils konkretisiert für das oben erwähnte Beispiel 2, wobei angenommen wird, dass eine Stichprobe von $n = 400$ Bestellungen die folgende Verteilung der Eingänge auf die 5 Werkzeuge liefert

$$n_1 = 105, n_2 = 85, n_3 = 65, n_4 = 65, n_5 = 80.$$

Schritt 1: Festsetzung von α und Formulierung der Hypothesen:

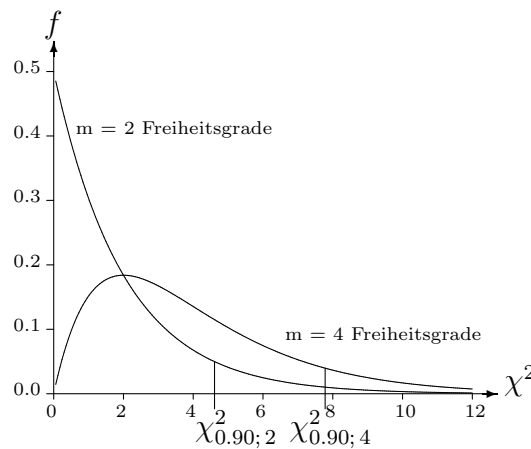
H_0 : Die Verteilung der diskreten Zufallsvariablen Y mit

k Ausprägungen y_1, \dots, y_k ist gegeben durch k

Wahrscheinlichkeiten p_1^0, \dots, p_k^0 mit $P(Y = y_i) = p_i^0, i = 1, \dots, k,$

H_1 : Y hat nicht die unter H_0 angegebene Verteilung.

(Beispiel: $\alpha = 0.10, H_0 : (p_1^0, \dots, p_5^0) = (0.2, \dots, 0.2)$, d.h. Gleichverteilung)

Abbildung 5.1: Dichtefunktion der χ^2 -Verteilung mit m Freiheitsgraden

Schritt 2: Prüfung des Stichprobenumfangs und Verteilung der Teststatistik:

Wenn

$$np_1^0 \geq 5, \dots, np_k^0 \geq 5, \quad (5.2)$$

dann ist die Teststatistik

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - np_i^0)^2}{np_i^0} \quad (5.3)$$

χ^2 -verteilt mit $k - 1$ Freiheitsgraden.

(Beispiel: $n = 400$ und $np_i^0 = 80 > 5$ für $i = 1, \dots, 5$, Teststatistik ist approximativ χ^2 -verteilt mit $k - 1 = 4$ Freiheitsgraden.)

Schritt 3: Bestimme den tabellierten kritischen Wert $\chi_{1-\alpha}^2$ zur χ^2 -Verteilung mit $k - 1$ Freiheitsgraden und den Ablehnungsbereich $(\chi_{1-\alpha}^2, \infty)$.

(Beispiel: Zu $\alpha = 0.10$ und $k - 1 = 4$ Freiheitsgraden ist der kritische Wert $\chi_{0.90;4}^2 = 7.779$, d.h. H_0 ablehnen, wenn $\chi^2 > 7.779$)

Schritt 4: Berechnung des Stichprobenwerts der Teststatistik:

(Beispiel: Mit der angegebenen Stichprobenverteilung der Bestellungen n_1, n_2, n_3, n_4, n_5 auf die 5 Werkzeuge und mit $np_i^0 = 80$ folgt: $\chi^2 = (25^2 + 5^2 + 15^2 + 15^2 + 0^2)/80 = 13.75$).

Schritt 5: Entscheidung:

H_0 wird abgelehnt, wenn $\chi^2 > \chi_{1-\alpha}^2$.

(Beispiel: H_0 wird abgelehnt, da $\chi^2 = 13.75 > 7.779$. Die Verteilungsfunktion liefert als p -Wert $p = P(\chi^2(4) > 13.75) = 0.00814$, d.h. die Nullhypothese der Gleichverteilung wird auch bei wesentlich geringerer Irrtumswahrscheinlichkeit als $\alpha = 0.10$ verworfen.)

Im folgenden verallgemeinern wir die Anwendung des χ^2 -Anpassungstests nacheinander in zwei Richtungen. Zum einen lassen wir zu, dass die hypothetische Verteilung nicht mehr vollständig spezifiziert ist unter H_0 , sondern noch von einem oder mehreren (r) unbekanntem Parametern abhängig ist. Wir entscheiden also anhand der Stichprobe über die Zugehörigkeit der Verteilung von Y zu einer gegebenen Familie von Verteilungen. Zum andern wenden wir den χ^2 -Anpassungstest nicht nur auf Verteilungen mit k Ausprägungen an, sondern auch auf Verteilungen mit unendlichem Wertebereich, insbesondere auch auf stetige Verteilungen. Dazu bilden wir Klassen und vergleichen die **erwarteten** Klassenhäufigkeiten mit den **beobachteten** Klassenhäufigkeiten.

5.1.2 Test auf Zugehörigkeit zu einer Familie von Verteilungen

Wenn die unter H_0 angegebene Verteilung noch von r unbekanntem Parametern - einem Parametervektor θ - abhängig ist, wird zunächst mit der vorliegenden Stichprobe eine Maximum-Likelihood-Schätzung der Parameter durchgeführt. Mit den Maximum-Likelihood-Schätzwerten wird aus der Familie von Verteilungen diejenige Verteilung ausgewählt, die (gemäß dem ML-Prinzip) am besten zu der Stichprobe passt. Der χ^2 -Test wird auf diese vollständig spezifizierte Verteilung angewandt. Durch die Schätzung der r Parameter aus den Stichprobendaten wird die Anzahl der Freiheitsgrade um r verringert. Die Teststatistik ist dann also nicht mehr mit $k - 1$ Freiheitsgraden verteilt, sondern χ^2 -verteilt mit $k - 1 - r$ Freiheitsgraden.

5.1.3 Anwendung auf Klassenhäufigkeiten

Von großer praktischer Bedeutung ist die Erweiterung des χ^2 -Anpassungstests auf diskrete Verteilungen mit unendlichem Wertebereich (z.B. Poissonverteilung) und auf stetige Verteilungen. Dazu wird der Wertebereich von Y in k disjunkte, aneinander angrenzende Intervalle unterteilt, $k \geq 2$.

Betrachten wir zunächst wieder den Fall, dass unter H_0 für Y eine Verteilung vollständig spezifiziert wird, z.B.

$$H_0 : Y \sim N(500; 10^2).$$

Damit können die k Intervallwahrscheinlichkeiten p_1^0, \dots, p_k^0 bei Gültigkeit der hypothetischen Verteilung berechnet werden. Für einen Stichprobenumfang von $n = 100$ nehmen wir eine Klassen-Unterteilung in 7 Intervalle an:

i	Intervall	p_i^0	np_i^0
1	$(-\infty, 486]$	0.0808	8.08
2	$(486, 493]$	0.1612	16.12
3	$(493, 498]$	0.1787	17.87
4	$(498, 502]$	0.1586	15.86
5	$(502, 507]$	0.1787	17.87
6	$(507, 514]$	0.1612	16.12
7	$(514, \infty)$	0.0808	8.08

Tabelle 5.2: Klassenwahrscheinlichkeiten und erwartete Klassenhäufigkeiten unter $H_0: Y \sim N(500; 10^2)$ mit $n = 100$.

Der χ^2 -Anpassungstest wird dann auf die beobachteten Klassenhäufigkeiten n_1, n_2, \dots, n_7 und die unter der hypothetischen Verteilung erwarteten Klassenhäufigkeiten np_1^0, \dots, np_7^0 angewandt. Die χ^2 -Teststatistik (5.1) ist approximativ χ^2 -verteilt mit $k - 1 = 6$ Freiheitsgraden.

Bei der Unterteilung in Intervalle sollte darauf geachtet werden, dass die Approximationsbedingung ($np_i^0 \geq 5$) erfüllt ist; gegebenenfalls können benachbarte Klassen zusammengelegt werden, um die Bedingung zu sichern. Offensichtlich ist für die Unterteilung in Intervalle ein beträchtlicher Spielraum gegeben; wenn die Approximationsbedingung erfüllt ist, liegt aber die Irrtumswahrscheinlichkeit für eine fälschliche Ablehnung der Verteilungshypothese unabhängig von der speziellen Klasseneinteilung beim gewählten Signifikanzniveau α . Bezüglich des Fehlers zweiter Art ist bei einer sehr groben Klasseneinteilung zu beachten, dass dadurch die implizierten Klassenwahrscheinlichkeiten nicht sehr spezifisch für die hypothetische Verteilung sind. Insofern ist die empirische Evidenz für die hypothetische Verteilung dann auch bei Annahme der Nullhypothese bei sehr grober Klasseneinteilung nicht sehr groß.

Wenn als Nullhypothese die Zugehörigkeit der Verteilung von Y zu einer parametrischen Familie von Verteilungen formuliert wird, z.B.

$$H_0 : Y \sim N(\mu, \sigma^2),$$

dann sind die Klassenwahrscheinlichkeiten abhängig von den unbekanntem Parametern, z.B. für die Klassen der Tabelle (5.2):

$$\begin{aligned}
P(Y \leq 486) &= \Phi\left(\frac{486-\mu}{\sigma}\right) &=: p_1(\mu, \sigma) \\
P(486 < Y \leq 493) &= \Phi\left(\frac{493-\mu}{\sigma}\right) - \Phi\left(\frac{486-\mu}{\sigma}\right) &=: p_2(\mu, \sigma) \\
P(493 < Y \leq 498) &= \Phi\left(\frac{498-\mu}{\sigma}\right) - \Phi\left(\frac{493-\mu}{\sigma}\right) &=: p_3(\mu, \sigma) \\
P(498 < Y \leq 502) &= \Phi\left(\frac{502-\mu}{\sigma}\right) - \Phi\left(\frac{498-\mu}{\sigma}\right) &=: p_4(\mu, \sigma) \\
P(502 < Y \leq 507) &= \Phi\left(\frac{507-\mu}{\sigma}\right) - \Phi\left(\frac{502-\mu}{\sigma}\right) &=: p_5(\mu, \sigma) \\
P(507 < Y \leq 514) &= \Phi\left(\frac{514-\mu}{\sigma}\right) - \Phi\left(\frac{507-\mu}{\sigma}\right) &=: p_6(\mu, \sigma) \\
P(514 < Y) &= 1 - \Phi\left(\frac{514-\mu}{\sigma}\right) &=: p_7(\mu, \sigma)
\end{aligned}$$

Die Maximierung der Likelihoodfunktion

$$L(\mu, \sigma; n_1, \dots, n_7) = [p_1(\mu, \sigma)]^{n_1} \cdot \dots \cdot [p_7(\mu, \sigma)]^{n_7} \quad (5.4)$$

liefert dann ML-Schätzer $\hat{\mu}, \hat{\sigma}$, mit denen man die Klassenwahrscheinlichkeiten und die erwarteten Klassenhäufigkeiten unter der Nullhypothese, dass die Grundgesamtheit normalverteilt ist, berechnen kann. Die resultierende χ^2 -Statistik ist dann nicht mehr approximativ $\chi^2(6)$ -verteilt, da $r = 2$ Parameter, μ und σ , aus der Stichprobe geschätzt wurden. Die Teststatistik ist approximativ χ^2 -verteilt mit $k - r - 1 = 7 - 2 - 1 = 4$ Freiheitsgraden.¹

5.2 Chi-Quadrat-Test auf Unabhängigkeit von zwei statistischen Variablen

Wir betrachten die gemeinsame Verteilung von zwei Zufallsvariablen X, Y mit k bzw. r verschiedenen Ausprägungen A_1, \dots, A_k bzw. B_1, \dots, B_r (gegebenenfalls auch k bzw. r verschiedenen Klassen). Auf der Grundlage einer zweidimensionalen einfachen Zufallsstichprobe $(X_1, Y_1), \dots, (X_n, Y_n)$ zu

¹Wenn man die ML-Schätzer nicht mit den klassierten Daten durch Maximierung von (5.4) bestimmt, sondern durch die üblichen ML-Schätzer aus den unklassierten Daten, im Fall der Normalverteilung also durch

$$\hat{\mu} = \bar{x} \quad \text{und} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

dann ist die resultierende Teststatistik (auch bei großem Stichprobenumfang n) nicht mehr $\chi^2(k - r - 1)$ -verteilt. Der kritische Wert $\chi_{1-\alpha}^2$ der Verteilung der Teststatistik liegt dann zwischen $\chi^2(k - 1 - r)_{1-\alpha}$ und $\chi^2(k - 1)_{1-\alpha}$, vgl. dazu Albrecht[1980], On the correct use of the Chi-Square Goodness-of-fit Test, Scand. Actuarial J. 149-160.

(X, Y) soll eine Entscheidung getroffen werden, ob X und Y in der Grundgesamtheit unabhängig verteilt sind. Der folgende Test, der auch als **Kontingenztest** bezeichnet wird, basiert darauf, dass man für alle Paare (i, j) vergleicht

- n_{ij} , die **beobachtete** gemeinsame Häufigkeit der Ausprägung (A_i, B_j) ,
- \tilde{n}_{ij} , die **bei Unabhängigkeit** (und den beobachteten Randhäufigkeiten) **erwartete** gemeinsame Häufigkeiten von (A_i, B_j) .

Als Vergleichsmaß wird wieder die χ^2 -Teststatistik verwendet.

Beispiel: Tägliche Fahrzeit (X) und Geschlecht (Y):

Angenommen eine einfache Zufallsstichprobe von $n = 2000$ Berufstätigen in Frankfurt/Main liefert die folgenden Daten einer Kontingenztabelle:

(X) Geschlecht	(Y) Fahrzeit (min)			Randhäufigkeit $n_{i\bullet}$
	$y < 30$	$30 \leq y < 60$	$y \geq 60$	
Männlich	524	455	221	1200
Weiblich	413	263	124	800
$n_{\bullet j}$	937	718	345	2000

Tabelle 5.3: Kontingenztabelle: **Beobachtete** Häufigkeiten

Unter Verwendung der relativen Randhäufigkeiten aus der Stichprobe als Schätzwerte für die Randwahrscheinlichkeiten erhalten wir die (geschätzten) **erwarteten** gemeinsamen Häufigkeiten bei Gültigkeit der Unabhängigkeit von X und Y mit

$$\tilde{n}_{ij} = n \cdot \frac{n_{i\bullet}}{n} \cdot \frac{n_{\bullet j}}{n} = \frac{n_{i\bullet} n_{\bullet j}}{n}. \quad (5.5)$$

In Tabelle 5.4 ist die Berechnung dieser erwarteten Häufigkeiten \tilde{n}_{ij} für das Beispiel angegeben. Die Kombination der beobachteten Häufigkeiten mit den erwarteten Häufigkeiten (in Klammern) wird mit Tabelle 5.5 dargestellt.

Die Teststatistik ist

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^r \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}}. \quad (5.6)$$

Die Anzahl der Freiheitsgrade für die χ^2 -Verteilung ist

$$(k - 1) \cdot (r - 1). \quad (5.7)$$

	$y < 30$	$30 \leq y < 60$	$y \geq 60$	$n_{i\bullet}$
Männl.	$\frac{937 \cdot 1200}{2000}$	$\frac{718 \cdot 1200}{2000}$	$\frac{345 \cdot 1200}{2000}$	1200
Weibl.	$\frac{937 \cdot 800}{2000}$	$\frac{718 \cdot 800}{2000}$	$\frac{345 \cdot 800}{2000}$	800
$n_{\bullet j}$	937	718	345	2000

Tabelle 5.4: Kontingenztabelle: **Erwartete** Häufigkeiten

	$y < 30$	$30 \leq y < 60$	$y \geq 60$	$n_{i\bullet}$
Männl.	524 (562.2)	455 (430.8)	221 (207.0)	1200
Weibl.	413 (374.8)	263 (287.2)	124 (138.0)	800
$n_{\bullet j}$	937	718	345	2000

Tabelle 5.5: Kontingenztabelle: Beobachtete und erwartete Häufigkeiten

In unserem Beispiel ist $(k - 1)(r - 1) = 2$ und

$$\chi^2 = \frac{(524 - 562.2)^2}{562.2} + \frac{(455 - 430.8)^2}{430.8} + \dots + \frac{(124 - 138)^2}{138} = 12.25.$$

Abschließend noch einmal die einzelnen Schritte der Durchführung des Kontingenztests:

Schritt 1: Festlegung von α und Formulierung der Hypothesen:

H_0 : X, Y sind unabhängig verteilt

(mit k bzw. r verschiedenen Ausprägungen)

H_1 : X, Y sind abhängig

Schritt 2: Überprüfe die Bedingung für eine „große Stichprobe“ für die χ^2 -Verteilung der Teststatistik χ^2 :

$$\tilde{n}_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n} \geq 5, \quad i = 1, \dots, k; \quad j = 1, \dots, r. \quad (5.8)$$

Schritt 3: Bestimme den kritischen Wert $\chi_{1-\alpha}^2$ für die χ^2 -Verteilung mit $(k-1) \cdot (r-1)$ Freiheitsgraden und den Ablehnungsbereich (H_0 ablehnen, wenn $\chi^2 > \chi_{1-\alpha}^2$)
(Im Beispiel: $\chi_{0.95;2}^2 = 5.99$)

Schritt 4: Berechne den Stichprobenwert der Teststatistik

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^r \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}}. \quad (5.9)$$

Schritt 5: Entscheidung (Beispiel: H_0 ablehnen, da $\chi^2 = 12.25 > 5.99 = \chi_{0.95;2}^2$. Fahrzeit und Geschlecht sind nicht unabhängig. Für den p -Wert erhält man $P(\chi^2(2) > 12.25) = 0.0029$.)

Kapitel 6

Vergleich der Lage von Verteilungen mit unabhängigen Stichproben

6.1 Zweistichproben-Tests

In diesem Abschnitt geht es um den Vergleich von zwei Populationen oder von zwei Verteilungen bezüglich des Erwartungswertes, bzw. des Medians (im dritten Abschnitt 6.1.3). Im Gegensatz zu dem Differenzentest bei verbundenen Stichproben gehen wir im Folgenden davon aus, dass zwei **unabhängige** einfache Stichproben

X_1, \dots, X_n zur Verteilung von X und

Y_1, \dots, Y_m zur Verteilung von Y

vorliegen.

Beispiel: Anfangsgehälter von männlichen und weiblichen Diplomkaufleuten

Um die mittleren Anfangsgehälter von Frauen und Männern nach Abschluss des BWL-Studiums zu vergleichen, wird eine einfache Stichprobe vom Umfang n aus der Grundgesamtheit der weiblichen Absolventen und unabhängig davon eine einfache Stichprobe vom Umfang m aus der Grundgesamtheit der männlichen Absolventen eines Abschlussjahrgangs gezogen und jeweils das erzielte Jahreseinkommen in der ersten Stelle nach dem Studienabschluss erhoben.

Wir betrachten das Testproblem für folgende Hypothesenpaare:

$$\begin{aligned} H_0 : \mu_X = \mu_Y & \quad \text{gegen} \quad H_1 : \mu_X \neq \mu_Y \quad \text{bzw.} \\ H_0 : \mu_X \leq \mu_Y & \quad \text{gegen} \quad H_1 : \mu_X > \mu_Y \quad \text{bzw.} \\ H_0 : \mu_X \geq \mu_Y & \quad \text{gegen} \quad H_1 : \mu_X < \mu_Y \end{aligned}$$

wobei μ_X und μ_Y die Erwartungswerte der beiden Verteilungen bezeichnen. Die Varianzen werden mit σ_X^2 und σ_Y^2 bezeichnet, die Stichprobenmittel und Stichprobenvarianzen entsprechend mit \bar{X}, \bar{Y} und S_X^2, S_Y^2 . Wegen der Unabhängigkeit der beiden Stichproben gilt dann für den Erwartungswert und die Varianz der Differenz der Stichprobenmittel:

$$E(\bar{X} - \bar{Y}) = \mu_X - \mu_Y \quad \text{und} \quad \text{Var}(\bar{X} - \bar{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m} \quad (6.1)$$

Wie beim Differenzentest ist die Differenz der Stichprobenmittel für den Vergleich von μ_X und μ_Y relevant; der Unterschied in der verwendeten Teststatistik ergibt sich aus der unterschiedlichen Varianz dieser Differenz bei unabhängigen Stichproben.

6.1.1 Zweistichproben-Gauß-Test

Der wichtigste Anwendungsfall des Zweistichproben-Gauß-Tests liegt vor, wenn

der Stichprobenumfang groß ist ($n \geq 30, m \geq 30$) und

die Varianzen σ_X^2 und σ_Y^2 **unbekannt** sind.

Dann gilt der Gauß-Test approximativ.

Zunächst betrachten wir vereinfachend die Teststatistik unter der stärkeren Voraussetzung, dass X und Y normalverteilt sind und die Varianzen σ_X^2, σ_Y^2 bekannt sind. Dann ist auch die Differenz der Stichprobenmittel normalverteilt und die Teststatistik

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \quad (6.2)$$

ist standardnormalverteilt, wenn H_0 mit $\mu_X = \mu_Y$ gilt. Diese Teststatistik ist natürlich nur dann anwendbar, wenn die Varianzen σ_X^2 und σ_Y^2 bekannt sind.

Wenn X und Y normalverteilt sind, die Varianzen der beiden Verteilungen aber **unbekannt** sind - was in der Regel der Fall sein wird - dann wird man an ihrer Stelle die Stichprobenvarianzen S_X^2 und S_Y^2 verwenden. Die resultierende Teststatistik ist approximativ t -verteilt, wobei die Anzahl der Freiheitsgrade von n, m und S_X^2, S_Y^2 abhängig ist (vgl. z.B. in Schlittgen). Wenn die beiden Stichprobenvarianzen S_X^2, S_Y^2 nicht zu stark voneinander abweichen (nicht mehr als im Verhältnis 1:3) dann genügen schon die Stichprobengrößen $n \geq 20, m \geq 20$, um den Gauß-Test approximativ anzuwenden. D.h. auch für die Teststatistik unter Verwendung der geschätzten Varianzen,

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}, \quad (6.3)$$

bestimmen wir den Ablehnungsbereich mit den Perzentilen $z_{1-\frac{\alpha}{2}}$ bzw. $z_{1-\alpha}$ der Standardnormalverteilung, wenn $n \geq 20, m \geq 20$.

Wenn X und Y nicht normalverteilt sind, entsteht nicht nur durch die Verwendung der geschätzten Varianzen, sondern auch durch die Anwendung des zentralen Grenzwertsatzes ein Approximationsfehler. Daher müssen in diesem Fall die Anforderungen an den Stichprobenumfang höher gesetzt werden ($n \geq 30, m \geq 30$), um die Verwendung der Standardnormalverteilung für die Bestimmung der kritischen Werte zu rechtfertigen.

Beispiel: Gehaltsniveau im Kredit- und Versicherungsgewerbe

Nach Angaben des Statistischen Bundesamtes (vgl. <http://www.destatis.de>) lag der mittlere Bruttomonatsverdienst von Angestellten im Kreditgewerbe im Jahr 2002 bei 3141 €, im Versicherungsgewerbe bei 3312 €. Angenommen diese Mittelwerte ergeben sich aus unabhängigen Stichproben vom Umfang $n = 500$ und $m = 400$, mit Stichprobenstandardabweichungen von 550 € und 500 €. Unterscheidet sich unter diesen Annahmen das Gehaltsniveau signifikant beim Signifikanzniveau $\alpha = 0.05$?

Wir bezeichnen die Bruttomonatsverdienste von Angestellten im Kreditgewerbe mit X und im Versicherungsgewerbe mit Y , also $\bar{x} = 3141, \bar{y} = 3312$ und $s_X = 550, s_Y = 500$.

Schritt 1: $H_0 : \mu_X = \mu_Y$ gegen $H_1 : \mu_X \neq \mu_Y$
Signifikanzniveau: $\alpha = 0.05$.

Schritt 2: Die Teststatistik (6.3) ist dann

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{500} + \frac{S_Y^2}{400}}}$$

Bei Gültigkeit von H_0 ist die Teststatistik approximativ standardnormalverteilt.

Schritt 3: Kritische Werte und Ablehnungsbereich:

Zu $\alpha = 0.05$ ist $z_{1-\frac{\alpha}{2}} = 1.96$. Ablehnung von H_0 , wenn $|z| > 1.96$.

Schritt 4: Berechnung des realisierten Werts der Teststatistik:

$$z = \frac{3141 - 3312}{\sqrt{\frac{550^2}{500} + \frac{500^2}{400}}} = -4.88$$

Schritt 5: Entscheidung: H_0 wird abgelehnt (das Gehaltsniveau im Kreditgewerbe und im Versicherungsgewerbe unterscheidet sich signifikant), da $|z| = 4.88 > 1.96$.

6.1.2 Zweistichproben-t-Test

Der folgende Zweistichproben-t-Test ist angebracht, wenn die folgenden Voraussetzungen angenommen werden können:

X und Y sind normalverteilt und

die (unbekannten) Varianzen σ_X^2 und σ_Y^2 sind gleich.

Mit der Annahme der Normalverteilung und gleicher Varianz von X und Y geht es also um die Frage, ob X und Y die gleiche Verteilung haben (unter $H_0 : \mu_X = \mu_Y$), oder ob sich die Verteilung von X und Y (nur) durch eine Lageverschiebung unterscheiden. Die Annahme, dass die Varianz in beiden Populationen gleich ist, führt zur Schätzung dieser Varianz σ^2 unter Verwendung der beiden unabhängigen Stichproben mit dem gepoolten Schätzer

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2}{n + m - 2} = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n + m - 2}. \quad (6.4)$$

Wenn beide Stichproben gleich groß sind ($n = m$), vereinfacht sich dieser Ausdruck zu

$$S^2 = \frac{S_X^2 + S_Y^2}{2}.$$

Allgemein ist der gepoolte Schätzer ein gewogenes Mittel aus den beiden getrennten Schätzfunktionen S_X^2 und S_Y^2 . Als Teststatistik erhalten wir unter Verwendung des gepoolten Varianzschätzers S^2 an Stelle von S_X^2 und S_Y^2 in (6.3) die Teststatistik

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S^2}{n} + \frac{S^2}{m}}} = \sqrt{\frac{nm}{n+m}} \cdot \frac{\bar{X} - \bar{Y}}{S}. \quad (6.5)$$

Wenn die Nullhypothese mit $\mu_X = \mu_Y$ gilt, dann ist diese Prüfgröße t -verteilt mit $n + m - 2$ Freiheitsgraden. Als kritische Werte sind also die Quantile der $t(n + m - 2)$ -Verteilung zu verwenden, bzw. für großes $n + m$ auch näherungsweise die der $N(0; 1)$ -Verteilung.

Beispiel: Wirksamkeit einer Werbeaktion

Neben einer bundesweiten Radio/TV-Werbekampagne für ein Shampoo wird regional begrenzt eine zusätzliche Werbeaktion durchgeführt. Deren Wirksamkeit soll durch den Vergleich der prozentualen Absatzänderung in einer Woche gegenüber dem durchschnittlichen wöchentlichen Absatz im Vormonat untersucht werden. Es sei

X die prozentuale Absatzänderung in Supermärkten, die nicht von der zusätzlichen Werbeaktion berührt werden,

Y die prozentuale Absatzänderung unter Einfluss der Sonderaktion.

Es wird angenommen, dass X und Y als normalverteilte Zufallsvariablen mit einheitlicher Standardabweichung σ betrachtet werden können. Die Absatzänderungen werden in einer Stichprobe von 10 Supermärkten ohne Sonderaktion und in einer Stichprobe von 5 Supermärkten mit Sonderaktion erhoben. Die beiden unabhängigen Stichproben ergeben

$$\begin{aligned}\bar{x} &= 6.5 & s_X^2 &= 20.25 & (n &= 10) \\ \bar{y} &= 8.0 & s_Y^2 &= 23.04 & (m &= 5)\end{aligned}$$

Zum Signifikanzniveau $\alpha = 0.05$ soll untersucht werden, ob die Sonderaktion zu einer signifikanten zusätzlichen Absatzerhöhung (bzw. zu einer Verschiebung der Verteilung von Y nach rechts) geführt hat.

Schritt 1: $H_0 : \mu_X = \mu_Y$ gegen $H_1 : \mu_X < \mu_Y$

Signifikanzniveau: $\alpha = 0.05$.

Schritt 2: Die Teststatistik ist

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S^2}{10} + \frac{S^2}{5}}} = \frac{\bar{X} - \bar{Y}}{\sqrt{0.3S^2}}$$

mit

$$S^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} = \frac{9S_X^2 + 4S_Y^2}{13}.$$

Bei Gültigkeit von H_0 ist die Teststatistik t -verteilt mit $n + m - 2 = 13$ Freiheitsgraden.

Schritt 3: Kritische Werte und Ablehnungsbereich:

Es wird ein linksseitiger Test durchgeführt. Zu $\alpha = 0.05$ ist $t_{1-\alpha;13} = 1.771$; Ablehnung von H_0 , wenn $t < -1.771$.

Schritt 4: Berechnung des realisierten Werts der Teststatistik:

$$s^2 = \frac{9s_X^2 + 4s_Y^2}{13} = \frac{9 \cdot 20.25 + 4 \cdot 23.04}{13} = 21.1085.$$

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{0.3s^2}} = \frac{6.5 - 8.0}{\sqrt{0.3 \cdot 21.1085}} = -0.596$$

Schritt 5: Entscheidung: H_0 wird nicht abgelehnt (die Absatzerhöhung durch die Sonderaktion ist nicht signifikant), da $t = -0.596 > -1.771$.

6.1.3 Rangsummentest von Wilcoxon

Wenn die Annahme der Normalverteilung für X und Y nicht gerechtfertigt ist und der Stichprobenumfang klein ist, sollte weder der Zweistichproben-Gauß-Test noch der Zweistichproben-t-Test angewandt werden. Für diesen Fall betrachten wir den verteilungsfreien **Rangsummentest von Wilcoxon**. Dabei geht es um die Entscheidung, ob X und Y die gleiche Verteilung besitzen oder sich (nur) bezüglich der Lage unterscheiden. Den Lageunterschied kann man durch unterschiedliche Mediane erfassen, die wir im folgenden mit $\tilde{\mu}_X$ bzw. $\tilde{\mu}_Y$ bezeichnen, so dass wir unter der Voraussetzung des gleichen Verteilungstyps testen

$$H_0 : \tilde{\mu}_X = \tilde{\mu}_Y \quad \text{gegen} \quad H_1 : \tilde{\mu}_X \neq \tilde{\mu}_Y \quad \text{bzw.}$$

$$H_0 : \tilde{\mu}_X = \tilde{\mu}_Y \quad \text{gegen} \quad H_1 : \tilde{\mu}_X < \tilde{\mu}_Y \quad \text{bzw.}$$

$$H_0 : \tilde{\mu}_X = \tilde{\mu}_Y \quad \text{gegen} \quad H_1 : \tilde{\mu}_X > \tilde{\mu}_Y.$$

Die Konstruktion des Tests beruht vor allem auf kombinatorischen Überlegungen. Zwei unabhängige Stichproben vom Umfang n zu X und vom Umfang m zu Y können unter der Nullhypothese als eine einfache Stichprobe vom Umfang $n+m$ aus **einer** Population betrachtet werden. Alle $n+m$ Stichprobenwerte werden aufsteigend der Größe nach geordnet. Damit erhält jedes der Ergebnisse x_1, \dots, x_n einen Rangwert und die **Summe dieser Rangwerte** wird mit R_X bezeichnet. Entsprechend ist R_Y die **Rangwertsumme** der m Realisationen von Y in der geordneten Gesamtstichprobe. Da jeder Rangwert $1, \dots, n+m$ einmal angenommen wird, gilt insgesamt

$$R_X + R_Y = \sum_{i=1}^{n+m} i = \frac{(n+m)(n+m+1)}{2}. \quad (6.6)$$

Unter der Nullhypothese ist der Erwartungswert des Rangwerts für jede einzelne Stichprobenvariable gleich dem mittleren Rangwert

$$\frac{R_X + R_Y}{n + m} = \frac{n + m + 1}{2},$$

womit sich für die Rangwertsumme R_X der Erwartungswert

$$E(R_X) = \frac{n(n + m + 1)}{2} \quad (6.7)$$

ergibt. Wenn die Verteilung von X nach rechts verschoben ist ($\tilde{\mu}_X > \tilde{\mu}_Y$), wird man mit einem höheren Wert von R_X rechnen müssen, bei einer Verschiebung nach links werden die Stichprobenvariablen zu X eher die unteren Rangplätze besetzen und man wird mit einem kleineren Wert von R_X rechnen.

Für die Varianz von R_X kann man zeigen, dass unter der Nullhypothese gilt

$$\text{Var}(R_X) = \frac{n \cdot m \cdot (n + m + 1)}{12}. \quad (6.8)$$

Bei **großem Stichprobenumfang** kann die Verteilung der Rangwertsumme R_X durch eine Normalverteilung approximiert werden, so dass man approximativ einen Gauß-Test für die standardisierte Rangsumme als Teststatistik Z anwenden kann,

$$Z = \frac{R_X - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}}. \quad (6.9)$$

Für den Fall einer **kleinen Stichprobe** betrachten wir die Lösung des Testproblems mit Hilfe der Kombinatorik im Folgenden anhand eines Beispiels.

Beispiel: Anfangsgehälter von männlichen und weiblichen Diplomkaufleuten

Für eine einfache Stichprobe von 3 Diplom-Kauffrauen ($n = 3$, Anfangsgehalt: X) und 7 Diplomkaufmännern ($m = 7$, Anfangsgehalt: Y) wird jeweils das Anfangsgehalt erhoben. Dann gibt es

$$\binom{n + m}{n} = \binom{10}{3} = 120$$

unterschiedliche Möglichkeiten für die Belegung von 3 Rangplätzen durch die $n = 3$ Frauen in der (ansteigend) geordneten Liste der Anfangsgehälter. Wenn die Nullhypothese gilt (d.h. die Verteilung der Anfangsgehälter ist für

Frauen und Männer gleich), dann ist jede dieser 120 möglichen Folgen gleich wahrscheinlich (mit Wahrscheinlichkeit $\frac{1}{120}$). Wir wollen die Nullhypothese testen gegen die Hypothese, dass die Verteilung der Anfangsgehälter für Frauen nach links verschoben ist, d.h. dass $H_1 : \tilde{\mu}_X < \tilde{\mu}_Y$ gilt.

Angenommen die Stichprobe liefert die folgende Einordnung der Frauen und Männer: $(x\ y\ x\ y\ x\ y\ y\ y\ y)$, d.h. die X -Rangplätze 1,3 und 5 und die Rangwertsumme $R_X = 1 + 3 + 5 = 9$. Wir bestimmen den p -Wert für dieses Stichprobenergebnis als die Wahrscheinlichkeit

$$p = P(R_X \leq 9 | H_0). \quad (6.10)$$

Wenn die berechnete Wahrscheinlichkeit p kleiner als 10% ist, dann verwerfen wir die Nullhypothese und betrachten den Gehaltsabschlag für Frauen als signifikant zum Niveau $\alpha = 10\%$. Um die Wahrscheinlichkeit p zu berechnen, bilden wir die Liste aller möglichen X -Ränge mit niedriger Rangwertsumme R_X , beginnend mit dem Extremfall, dass die drei niedrigsten Gehälter an Frauen gezahlt werden:

X – Ränge	$R_X = r$	$P(R_X = r)$
(1,2,3)	6	$\frac{1}{120}$
(1,2,4)	7	$\frac{1}{120}$
(1,3,4),(1,2,5)	8	$\frac{2}{120}$
(1,3,5),(1,2,6),(2,3,4)	9	$\frac{3}{120}$
(1,2,7),(1,3,6),(1,4,5),(2,3,5)	10	$\frac{4}{120}$
...

Tabelle 6.1: Niedrige X -Rangwertsummen und Wahrscheinlichkeiten

Für den p -Wert zum Stichprobenergebnis $R_X = 9$ folgt damit:

$$p = P(R_X \leq 9 | H_0) = \frac{7}{120} = 0.0583$$

Zum vorgegebenem Signifikanzniveau $\alpha = 10\%$ wird die Nullhypothese („das Gehalt für Frauen ist im Mittel nicht niedriger“) also abgelehnt, der Gehaltsabschlag ist bei dieser Irrtumswahrscheinlichkeit signifikant. Zum Signifikanzniveau von 5% kann die Nullhypothese nicht verworfen werden.

Die kleine Stichprobe liefert keine ausreichende empirische Evidenz, um die Nullhypothese mit dieser geringen Irrtumswahrscheinlichkeit zu verwerfen.

Obwohl der Stichprobenumfang in dem betrachteten Beispiel klein ist, wollen wir zum Vergleich auch die Approximation mit dem Gauß-Test betrachten. Für $n = 3, m = 7$ erhalten wir mit (6.7) und (6.8)

$$E(R_X) = \frac{n(n+m+1)}{2} = 16.5, \quad \text{Var}(R_X) = \frac{nm(n+m+1)}{12} = 19.25.$$

Mit dem angenommenen Stichprobenergebnis, bei dem die 3 Frauen die Rangplätze 1,3 und 5 in der (ansteigend geordneten) Gehaltsliste belegt hatten, ist die realisierte Rangwertsumme $R_X = 9$ und wir erhalten für die Realisierung der Teststatistik (6.9):

$$z = \frac{9 - 16.5}{\sqrt{19.25}} = -1.71$$

Für den linksseitigen Test gegen $H_1 : \tilde{\mu}_X < \tilde{\mu}_Y$ ist der kritische Wert zum Signifikanzniveau $\alpha = 0.05$ gegeben mit $-z_{1-\alpha} = -1.645$. Unter Verwendung des approximativen Gauß-Tests ist die Nullhypothese also zum 5%-Niveau abzulehnen, da $-1.71 < -1.645$.

Der p -Wert zum realisierten Wert $z = -1.71$ der Teststatistik ergibt sich beim linksseitigen Gauß-Test mit $P(Z \leq -1.71) = \Phi(-1.71) = 0.0446$. Der Vergleich mit dem exakten p -Wert, den wir oben als 0.0583 berechnet haben, zeigt, dass die Approximation selbst bei der recht kleinen Stichprobe einen guten Anhaltspunkt gibt.

6.2 Einfache Varianzanalyse

Im ersten Abschnitt dieses Kapitels haben wir zwei Variablen X und Y (bzw. auch **ein Merkmal** in **zwei** Populationen) mit zwei unabhängigen Stichproben untersucht. In den betrachteten Beispielen orientierte sich die Differenzierung in zwei Verteilungen jeweils an zwei Ausprägungen einer qualitativen Variablen (die auch als „Stufen“ eines „Faktors“ bezeichnet werden):

- die Differenzierung der Verteilung der Anfangsgehälter nach dem Faktor „Geschlecht“ der Diplom-Kaufleute mit den Stufen „männlich“ und „weiblich“,
- die Differenzierung der Verteilung der Gehälter nach dem Faktor „Branche“ mit den Stufen „Kreditgewerbe“ und „Versicherungsgewerbe“,

- die Differenzierung der Verteilung der Absatzveränderungen nach dem Faktor „Werbeeinsatz“ mit den Stufen „keine Sonderaktion“ und „Sonderaktion“.

Im folgenden verallgemeinern wir diese Fragestellung auf den Effekt eines Faktors mit k Stufen auf die Lage einer Verteilung und differenzieren eine Gesamtpopulation entsprechend in k Variablen bzw. Teilpopulationen mit k unabhängigen Stichproben. Dies führt zum Konzept der einfachen (auch: ein-faktoriellen) Varianzanalyse, das wir zunächst ausführlich anhand eines Beispiels betrachten.

6.2.1 Einführendes Beispiel zur Varianzanalyse

In der Außendienststelle einer Versicherung werden die Kunden von drei ($k = 3$) Mitarbeitern bedient. Mit unabhängigen Stichproben vom Umfang n_1, n_2 bzw. n_3 werden Bedienungszeiten X_1, X_2, X_3 an den entsprechenden drei Servicepunkten erhoben:

$$x_{11}, x_{12}, \dots, x_{1n_1} \quad (\text{Mittel: } \bar{x}_1)$$

$$x_{21}, x_{22}, \dots, x_{2n_2} \quad (\text{Mittel: } \bar{x}_2)$$

$$x_{31}, x_{32}, \dots, x_{3n_3} \quad (\text{Mittel: } \bar{x}_3)$$

Wir differenzieren also die Gesamtpopulation der Bedienungszeiten nach dem Faktor „Mitarbeiter“, der drei Stufen annimmt. Unter der Annahme, daß die Bedienungszeiten X_1, X_2, X_3 an den drei Servicepunkten (zumindest approximativ) **normalverteilt** sind mit der **gleichen Varianz**,

$$X_1 \sim N(\mu_1; \sigma^2), \quad X_2 \sim N(\mu_2; \sigma^2), \quad X_3 \sim N(\mu_3; \sigma^2),$$

soll getestet werden, ob die Erwartungswerte gleich sind und somit die Gesamtstichprobe von $n = n_1 + n_2 + n_3$ Bedienungszeiten als eine einfache Stichprobe aus **einer** normalverteilten Population betrachtet werden kann, oder ob zumindest zwei Erwartungswerte unterschiedlich sind, d.h.:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad \text{gegen}$$

$$H_1 : \text{mind. zwei Erwartungswerte sind unterschiedlich}$$

Das Testverfahren kann in folgender Weise als **Analyse der Varianz der Gesamtstichprobe** verstanden werden:

Die Summe der quadrierten Abweichungen um das **Gesamtstichprobenmittel** \bar{x} (SS , Sum of Squares) kann stets zerlegt werden in die **Streuung**

zwischen den Gruppen (SB , Squares Between) und die Streuung innerhalb der Gruppen (SW , Squares Within):

$$SS = \sum_{i=1}^{n_1} (x_{1i} - \bar{x})^2 + \sum_{i=1}^{n_2} (x_{2i} - \bar{x})^2 + \sum_{i=1}^{n_3} (x_{3i} - \bar{x})^2 = SB + SW$$

mit

$$SB = \sum_{j=1}^3 n_j (\bar{x}_j - \bar{x})^2$$

$$SW = \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2 + \sum_{i=1}^{n_3} (x_{3i} - \bar{x}_3)^2,$$

denn z.B. für die erste Gruppe der Bedienungszeiten x_{1i} gilt:

$$\begin{aligned} \sum_{i=1}^{n_1} (x_{1i} - \bar{x})^2 &= \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1 + \bar{x}_1 - \bar{x})^2 \\ &= \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 + n_1 (\bar{x}_1 - \bar{x})^2 \\ &\quad + 2(\bar{x}_1 - \bar{x}) \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1), \end{aligned}$$

wobei der letzte Term gleich Null ist; entsprechend für die zweite und dritte Gruppe.

Wenn die Streuung zwischen den Gruppen (SB) groß ist gegenüber der Streuung innerhalb der Gruppen (SW), spricht dies gegen die Nullhypothese eines einheitlichen Erwartungswerts. Es spricht dafür, daß der Faktor „Mitarbeiter“ einen Effekt auf die Bedienungszeit hat. Um einen kritischen Wert für die Relation zwischen SB und SW zu erhalten, braucht man die Verteilung einer geeigneten Teststatistik bei Gültigkeit der Nullhypothese (= kein Effekt des Faktors „Mitarbeiter“ auf die Bedienungszeit). Man kann zeigen, daß bei Gültigkeit der Nullhypothese für die durch die einheitliche Varianz σ^2 dividierten Stichprobenfunktionen SB und SW gilt:

$\frac{SB}{\sigma^2}$ ist χ^2 -verteilt mit $k - 1 = 2$ Freiheitsgraden,

$\frac{SW}{\sigma^2}$ ist χ^2 -verteilt mit $n - k = n - 3$ Freiheitsgraden.

Weiterhin sind (was keineswegs unmittelbar offensichtlich ist) SB und SW unabhängig verteilt. Als Teststatistik wird der Quotient dieser beiden unabhängig χ^2 -verteilten Zufallsvariablen, jeweils geteilt durch die Anzahl der Freiheitsgrade, verwendet, wobei sich die unbekannte Varianz σ^2 herauskürzt:

$$F = \frac{SB/(k-1)}{SW/(n-k)} = \frac{SB}{SW} \cdot \frac{(n-k)}{(k-1)}. \quad (6.11)$$

Die Verteilung dieser Teststatistik ist die **F-Verteilung mit $k - 1$ und $n - k$ Freiheitsgraden**, $F(k - 1; n - k)$. Prozentpunkte der F-Verteilung sind tabelliert. Wenn die Nullhypothese zum Signifikanzniveau α getestet wird, wird H_0 abgelehnt, wenn der Stichprobenwert der Test-Statistik F größer als das $(1 - \alpha)$ -Quantil der $F(k-1;n-k)$ -Verteilung ist,

$$F > F_{k-1;n-k;1-\alpha}.$$

Die Ablehnung von H_0 bedeutet, daß ein signifikanter Effekt des betrachteten Faktors auf die Lage der Verteilung vorliegt, so daß die Erwartungswerte nicht alle gleich sind.

Die Größen, die bei der Durchführung des Tests zu berechnen sind, werden oft in der folgenden Tabelle zur Varianzanalyse (Tabelle 15.1) zusammengestellt (die auch als ANOVA-Tabelle bezeichnet wird, nach: ANalysis Of VAriance).

Wenn man die einfache Varianzanalyse auf den Fall von $k = 2$ Variablen

Streuungsursache	Freiheitsgrade	Quadratsumme	Mittleres Quadrat
Faktor	$k - 1$	SB	$\frac{SB}{k - 1}$
Zufallsfehler	$n - k$	SW	$\frac{SW}{n - k}$
Summe	$n - 1$	SS	

Tabelle 6.2: ANOVA-Tabelle

bzw. Populationen unter Verwendung von zwei unabhängigen Stichproben anwendet, dann ist der Test äquivalent zur Anwendung des zweiseitigen Zweistichproben-t-Tests.

Die Methode der einfachen (auch: ein-faktoriellen) Varianzanalyse kann verallgemeinert werden, indem der Einfluß mehrerer qualitativer „Faktoren“ auf die Lage einer Verteilung untersucht wird. In dem Beispiel könnte man in Betracht ziehen, daß neben dem Mitarbeiter etwa die Versicherungsart einen Effekt auf die mittlere Bedienungszeit hat. In einer etwas komplexeren Modellbildung kann dabei auch eine mögliche Interaktion verschiedener Faktoren getestet werden, etwa ob für die mittlere Bedienungszeit das Zusammenspiel von speziellem Mitarbeiter und bestimmter Versicherungsart relevant ist.

Wir fassen im folgenden Abschnitt das Testverfahren der einfachen Varianzanalyse noch einmal zusammen.

6.2.2 Übersicht zur einfachen Varianzanalyse

Voraussetzung:

- (1) Die Variablen X_1, \dots, X_k sind normalverteilt mit gleicher Varianz σ^2 ,

$$X_j \sim N(\mu_j; \sigma^2), \quad j = 1, \dots, k,$$

- (2) Zu den k Variablen werden k unabhängige einfache Stichproben gezogen, jeweils mit den folgenden Stichprobenvariablen:

$$X_{11}, X_{12}, \dots, X_{1n_1}$$

$$X_{21}, X_{22}, \dots, X_{2n_2}$$

...

$$X_{k1}, X_{k2}, \dots, X_{kn_k}$$

Der Gesamtstichprobenumfang ist $n = n_1 + n_2 + \dots + n_k$.

Durchführung der Varianzanalyse:

Schritt 1: Signifikanzniveau α festlegen; Hypothesen formulieren:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k \quad \text{gegen}$$

$$H_1 : \text{mind. zwei Erwartungswerte sind unterschiedlich}$$

Schritt 2: Die Teststatistik ist

$$F = \frac{SB/(k-1)}{SW/(n-k)} = \frac{SB}{SW} \cdot \frac{(n-k)}{(k-1)}$$

mit

$$SB = \sum_{j=1}^k n_j (\bar{X}_j - \bar{X})^2$$

$$SW = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ji} - \bar{X}_j)^2.$$

Bei Gültigkeit von H_0 ist die Teststatistik F -verteilt mit $k-1$ und $n-k$ Freiheitsgraden.

Schritt 3: Kritische Werte und Ablehnungsbereich:

Zum festgelegten Signifikanzniveau α ist der kritische Wert das $1 - \alpha$ -Quantil der F -Verteilung mit $k-1$ und $n-k$ Freiheitsgraden,

$$F_{k-1; n-k; 1-\alpha}.$$

Ablehnung von H_0 , wenn $F > F_{k-1; n-k; 1-\alpha}$.

Schritt 4: Berechnung des realisierten Werts der Teststatistik:

Die Berechnung der realisierten Werte von SB und SW , wird in der Regel vereinfacht mit

$$SB = \sum_{j=1}^k n_j \bar{X}_j^2 - n \bar{X}^2$$
$$SW = \sum_{j=1}^k \sum_{i=1}^{n_j} X_{ji}^2 - \sum_{j=1}^k n_j \bar{X}_j^2.$$

Einsetzen der berechneten Werte von SB und SW in $F = \frac{SB}{SW} \cdot \frac{(n-k)}{(k-1)}$ liefert die Stichprobenrealisierung der Teststatistik.

Schritt 5: Entscheidung:

H_0 wird abgelehnt (d.h. die Mittelwerte unterscheiden sich signifikant), wenn

$$F > F_{k-1; n-k; 1-\alpha}.$$

Während man die Varianzanalyse interpretieren kann als die Untersuchung, ob die verschiedenen Ausprägungen einer **qualitativen** Variablen (die k Stufen des betrachteten Faktors) die Lage der Verteilung einer anderen Variablen beeinflussen, werden wir in den nächsten Kapiteln mit der **Regressionsanalyse** dieser Frage nicht nur für qualitative Variablen, sondern für beliebige erklärende Variablen nachgehen. Dabei wird es nicht nur darum gehen, **ob** eine oder mehrere erklärende Variablen eine andere beeinflussen (z.B. das Einkommen und Alter die monatlichen Ausgaben für Kosmetika), sondern auch darum, diesen Einfluß zu quantifizieren.

Kapitel 7

Das klassische Regressionsmodell: Einführung

7.1 Problemstellung

Ziel der folgenden Abschnitte ist es, einfache Instrumente bereitzustellen, mit denen regelmäßige Beziehungen zwischen ökonomischen Variablen anhand von wirtschaftsstatistischen Daten untersucht werden können. Genauer beschränken wir uns auf Fragestellungen der folgenden Struktur:

1. Eine bestimmte ökonomische Variable wird als abhängige, zu erklärende Variable betrachtet, z.B.
 - die Bruttoinvestitionen in einem Wirtschaftszweig,
 - der gesamtwirtschaftliche private Konsum,
 - die Mietausgaben privater Haushalte,
 - die Nachfrage nach Kaffee,
 - der Wechselkurs US-Dollar/DM.

Für die betrachtete abhängige Variable, die mit y bezeichnet wird, sollen Lageverschiebungen bzw. unterschiedliche Werte, die diese Variable zu verschiedenen Zeitpunkten (oder in verschiedenen Perioden) bzw. für verschiedene Merkmalsträger annimmt, darauf zurückgeführt werden, daß wesentliche Einflußfaktoren x_1, x_2, \dots sich verändert haben. Die Variation der abhängigen Variablen y soll zerlegt werden in eine systematische Veränderung, die durch eine Funktion der Einflußgrößen x_1, x_2, \dots beschrieben wird, und in zufällige Schwankungen. Dieser Untersuchungsansatz wird als **Regression** von y auf die erklärenden Variablen x_1, x_2, \dots bezeichnet.

Bei der theoretischen Untersuchung der Abhängigkeit einer Variablen y von einer bestimmten Einflußgröße x wird der interessierende Zusammenhang oft durch die „ceteris paribus-Klausel“ isoliert, d.h. durch die Annahme, daß die übrigen Einflußgrößen und Rahmenbedingungen unverändert bleiben. In einem Gedankenexperiment kann eine solche Betrachtungsweise - zumindest soweit die übrigen Einflußgrößen nicht zwangsläufig durch eine Änderung von x tangiert werden - sehr hilfreich sein. Bei der Analyse von empirischen Daten y_1, \dots, y_T für die abhängige Variable y muß dagegen berücksichtigt werden, daß die beobachteten unterschiedlichen Werte der abhängigen Variablen bei unterschiedlichen Konstellationen zahlreicher Einflußgrößen realisiert wurden und nicht etwa nur die isolierten Veränderungen einer speziellen interessierenden Einflußgröße zugrundeliegen. Die vielfältigen Einflußfaktoren, auf deren Veränderungen die Variation der abhängigen Variablen beruht, werden in unterschiedlicher Weise im Modell berücksichtigt. Zum einen wird - gestützt auf wirtschaftstheoretische Ergebnisse oder Hypothesen - festgelegt, welches die wesentlichen Einflußvariablen x_1, \dots, x_K für die Variation der abhängigen Variablen sind. Mit dieser Festlegung derjenigen Variablen, die in den Erklärungsansatz $f(x_1, \dots, x_K)$ eingehen, wird auf der anderen Seite die Hypothese verknüpft, daß die übrigen Einflußgrößen insgesamt keinen systematischen Beitrag zur Erklärung der Veränderung der y -Werte liefern. Formal wird der Gesamteinfluß, den die „übrigen Einflußgrößen“ auf die Variation von y haben, mit einer additiven Störgröße u erfaßt, so daß für jeden Beobachtungsfall t , $t = 1, \dots, T$, gilt

$$y_t = f(x_{t1}, x_{t2}, \dots, x_{tK}) + u_t$$

Die Störgröße u_t wird als Zufallsvariable mit $E(u_t) = 0$, $Var(u_t) = \sigma^2$ und $Cov(u_t, u_s) = 0$ für $t \neq s$ angenommen. Damit werden auch die beobachteten y -Werte als Realisationen von Zufallsvariablen aufgefaßt¹. Bei gegebenen Werten x_{t1}, \dots, x_{tK} der erklärenden Variablen, bzw. bei nichtstochastischen erklärenden Variablen, stimmt der Erklärungsansatz $f(x_{t1}, \dots, x_{tK})$ mit dem Erwartungswert von y_t überein. Die Variablen y_t haben also für jedes t die gleiche Varianz, während der Erwartungswert von y_t als Funktion der erklärenden Variablen unterschiedlich ist.

Die Auswahl einer Einflußgröße als eine wesentliche erklärende Variable ist natürlich nur dann sinnvoll, wenn diese Einflußgröße im

¹Im Bezug auf u_t , y_t übernehmen wir bei der Behandlung des Regressionsmodells die in der ökonomischen Literatur verbreitete vereinfachende aber ungenaue Schreibweise, die nicht zwischen Zufallsvariablen und deren Realisationen unterscheidet.

Untersuchungszusammenhang unterschiedliche Werte annimmt, also tatsächlich eine Variable und nicht ein konstanter Parameter ist. Betrachten wir z.B. die Untersuchung der Nachfrage privater Haushalte nach Kaffee, dann werden grundsätzlich das Haushaltseinkommen und der Kaffeepreis die Nachfrage beeinflussen. Bei einer Querschnittsuntersuchung über verschiedene Haushalte, bei der das Preisniveau für alle Beobachtungsfälle gleich ist, wird dann das Einkommen wesentliche erklärende Variable für die unterschiedliche Nachfrage sein, während der Einfluß des Preises sich in der Gestalt der Nachfragefunktion bzw. in deren Parametern niederschlagen wird.

2. Im Bezug auf die erklärenden Variablen x_1, \dots, x_K wird angenommen, daß diese Variablen nicht ihrerseits wieder von der Zufallsvariablen y oder von der Störgröße u beeinflußt werden. Dieses Postulat der Modell-**Exogenität** der erklärenden Variablen bedeutet eine gewichtige Einschränkung. Vor allem bei makroökonomischen Zusammenhängen - z.B. einer makroökonomischen Konsumfunktion, die den gesamtwirtschaftlichen Konsum als Funktion des verfügbaren Einkommens erklärt - wird man eher von **gegenseitigen** Abhängigkeiten ausgehen müssen. Dementsprechend sind auch Methoden entwickelt worden, um interdependente Modellstrukturen zu untersuchen. Im folgenden werden wir jedoch solche interdependenten Modellstrukturen nicht betrachten und an der vereinfachenden Annahme exogener erklärender Variablen festhalten. Es bedeutet dann keine wesentliche Einschränkung mehr, wenn die Daten für die erklärenden Variablen als gegebene, deterministische Größen betrachtet werden.
3. Das im folgenden betrachtete multiple Regressionsmodell ist ein **lineares** Modell, d.h. für den Erklärungsansatz $f(x_1, \dots, x_K)$ wird eine lineare Funktion unterstellt, so daß

$$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_K x_{tK} + u_t$$

gilt.

Mit dieser Festlegung des Funktionstyps hängt die Gestalt des Erklärungsansatzes nur noch von den unbekanntem Koeffizienten β_i , $i = 1, \dots, K$ ab. Für die Parameter $\beta_1, \beta_2, \dots, \beta_K$ und σ^2 werden anhand der Daten für die erklärenden Variablen und für die abhängige Variable Schätzwerte bestimmt. Die Festlegung auf den linearen Funktionstyp stellt wiederum eine erhebliche Einschränkung dar. Allerdings ist zu beachten, daß nichtlineare Beziehungen manchmal

durch Transformationen (z.B. log-Transformation einer Cobb-Douglas-Produktionsfunktion) oder durch Approximationen linearisiert werden können.

Im übrigen bleiben die zu besprechenden Methoden anwendbar, so lange das Modell linear in den Parametern β_1, \dots, β_K bleibt, also z.B. in dem Modell

$$y_t = \beta_1 z_{t1} + \beta_2 z_{t2} + \dots + \beta_K z_{tK} + u_t$$

wobei $z_{tk} = \log(x_{tk})$ (oder eine andere nichtlineare Transformation der ursprünglichen erklärenden Variablen x_{t1}, \dots, x_{tK}). In einem solchen Modell wird sich natürlich die Interpretation der Parameter β_1, \dots, β_K als Kennziffern für den Einfluß der erklärenden Variablen x_1, \dots, x_K ändern.

Schließlich sei noch darauf hingewiesen, daß auch der Einfluß qualitativer Variablen mit dem linearen Ansatz modelliert werden kann (und damit auch die Fragestellung der Varianzanalyse im Rahmen des linearen Regressionsmodells verfolgt werden kann). Mit Hilfe einer 0,1-Variablen D_t (meist als Dummy-Variable bezeichnet), die den Wert 1 annimmt, wenn eine bestimmte Bedingung erfüllt ist und den Wert 0 sonst, wird der Einfluß des Vorliegens dieser Bedingung auf das Niveau von y mit dem Koeffizienten von D_t erfaßt. (Bsp.: Dummy-Variable für zwei alternative politische Konstellationen; Saison-Dummies für unterschiedliche Quartale bei einer Zeitreihenuntersuchung mit Quartalsdaten.) Häufig stellen solche Dummy-Variablen aber auch nur einen Notbehelf dar, wenn die Variation der abhängigen Variablen mit den eigentlichen erklärenden Variablen nur unzureichend erklärt werden kann.

Im folgenden werden im einzelnen die Annahmen des linearen Regressionsmodells formuliert und Folgerungen aus diesen Annahmen für die Eigenschaften von Schätzfunktionen, für Tests von Parameterrestriktionen und für Prognoseschätzungen gezogen. Jede Anwendung dieses Modellrahmens und der Schlußfolgerungen auf einen konkret spezifizierten Erklärungsansatz für eine ökonomische Variable y und auf die entsprechenden Daten basiert immer auf der Arbeitshypothese, daß die Modellannahmen auch in dem Anwendungsfall gültig sind. Insbesondere steht vor der Parameterschätzung die Arbeitshypothese, daß die spezifizierte Gesetzmäßigkeit mit bestimmten Parametern (den unbekanntem „wahren“ Parametern) vorliegt und es nur darum geht, diese wahren Parameter aus den Daten, die dieser Gesetzmäßigkeit unterworfen sind, möglichst gut zu rekonstruieren, Entscheidungen über die Gültigkeit

von Hypothesen über die wahren Parameter zu treffen usw. Anhand des empirischen Befunds ist es aber auch möglich, die Arbeitshypothese selbst Tests zu unterwerfen und gegebenenfalls zu verwerfen. Dann stellt sich die Frage nach einer neuen Spezifikation (z.B. anderen erklärenden Variablen) oder dem Verzicht auf bestimmte nicht haltbare Modellannahmen.

7.2 Schreibweise und Modellannahmen

Eine lineare Beziehung zwischen einer abhängigen Variablen y und $K - 1$ erklärenden Variablen x_2, x_3, \dots, x_K wird für T Beobachtungsfälle (Zeitpunkte oder Merkmalsträger) notiert mit

$$y_t = \beta_1 + \beta_2 x_{t2} + \dots + \beta_K x_{tK} + u_t, \quad t = 1, \dots, T$$

bzw. mit der Scheinvariablen $x_{t1} = 1$ für alle t :

$$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_K x_{tK} + u_t, \quad t = 1, \dots, T$$

Diese Gleichung heißt auch **Regressionsgleichung**, y heißt **Regressand** und x_2, \dots, x_K werden als **Regressoren** bezeichnet. Der Parameter β_1 ist das Absolutglied und β_2, \dots, β_K sind die Steigungskoeffizienten; insgesamt wird der Vektor $\beta = (\beta_1, \dots, \beta_K)'$ als **Vektor der Regressionskoeffizienten** bezeichnet. Die Störgröße u_t und die abhängige Variable y_t sind Zufallsvariablen. Beobachtet werden die Realisationen von y_t und die Daten x_{t2}, \dots, x_{tK} für $t = 1, \dots, T$. Die Regressionskoeffizienten sind feste, unbekannte Parameter, so daß die unbeobachtbaren Störgrößen u_t auch nicht aus den Daten $y_t, x_{t2}, \dots, x_{tK}$ berechnet werden können.

Als Sonderfall ergibt sich mit $K = 2$ die „**einfache Regressionsgleichung**“

$$y_t = \beta_1 + \beta_2 x_{t2} + u_t ;$$

bei $K > 2$ wird von **multipler Regression** gesprochen. Für alle T Beobachtungsfälle kann die Regressionsgleichung in Matrixschreibweise geschrieben werden als

$$y = X\beta + u$$

mit

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \dots & \dots & \dots & \dots \\ x_{T1} & x_{T2} & \cdots & x_{TK} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix}$$

Die $(T \times K)$ -Matrix X , die als erste Spalte die Scheinvariable enthält, heißt **Regressormatrix**; vor allem, wenn man die erklärenden Variablen in einem geplanten Experiment festsetzen kann, um möglichst viel über deren Wirkung auf y herauszufinden, wird X auch als **Designmatrix** bezeichnet. (Bei Anwendungen auf ökonomische Variablen besteht im allgemeinen keine Möglichkeit zu experimentellen Untersuchungen.)

Modellannahmen

(A1) X ist eine exogen gegebene, nichtstochastische $(T \times K)$ -Matrix mit vollem Spaltenrang, d.h. $r(X) = K$.

(A2) Die Momentenmatrix der Regressoren konvergiert mit wachsendem T gegen eine **endliche, nichtsinguläre** Matrix Q :

$$\frac{1}{T} X'X = \begin{bmatrix} 1 & \frac{1}{T} \sum x_{t2} & \cdots & \frac{1}{T} \sum x_{tK} \\ \frac{1}{T} \sum x_{t2} & \frac{1}{T} \sum x_{t2}^2 & \cdots & \frac{1}{T} \sum x_{t2}x_{tK} \\ \vdots & \vdots & & \vdots \\ \frac{1}{T} \sum x_{tK} & \frac{1}{T} \sum x_{tK}x_{t2} & \cdots & \frac{1}{T} \sum x_{tK}^2 \end{bmatrix} \longrightarrow Q$$

(A3) Die Varianzen von y_1, \dots, y_T sind alle gleich und für $t \neq s$ sind y_t und y_s unkorreliert, bzw. für die Störgrößen gilt:

$$\begin{aligned} E(u_t) &= 0 \\ \text{Var}(u_t) &= \sigma^2 \quad (\text{konstant für alle } t) \\ \text{Cov}(u_t, u_s) &= 0 \quad \text{für alle } t \neq s. \end{aligned}$$

(A4) Die Störgrößen sind unabhängig identisch normalverteilt, $u_t \sim N(0, \sigma^2)$ (Verstärkung von (A3)).

Erläuterungen und Folgerungen:

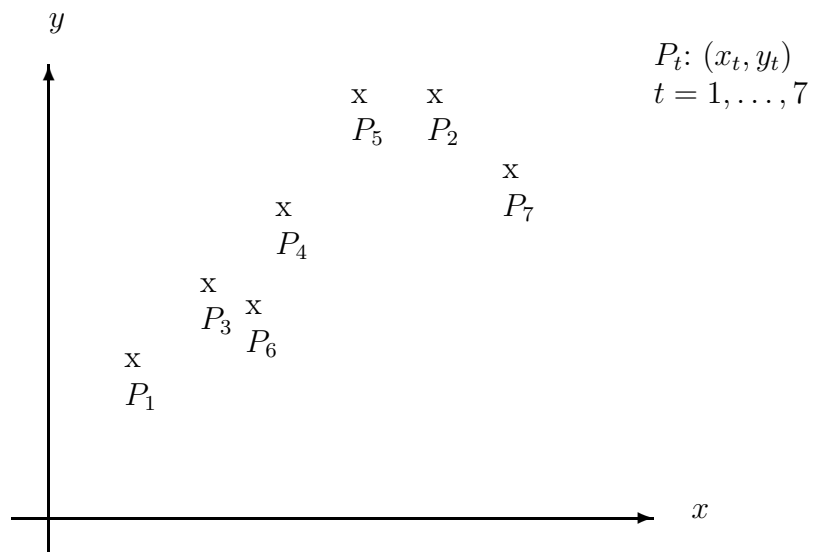
Zu (A1): Ausgeschlossen werden mit der Exogenitätsannahme Abhängigkeiten der Regressoren von y ; insbesondere werden verzögerte Werte y_{t-1} als erklärende Variablen für y_t ausgeschlossen.

Die Rangbedingung $r(X) = K$ bedeutet, daß die $(T \times 1)$ -Spaltenvektoren x_1, x_2, \dots, x_K linear unabhängig sind. Notwendige Bedingung dafür ist, daß $T \geq K$ gilt. Mit $r(X) = K$ folgt auch $r(X'X) = K$, d.h. die $(K \times K)$ -Matrix $X'X$ ist nichtsingulär.

- Zu (A2): Diese Annahme spielt eine Rolle bei der Untersuchung von Eigenschaften von Schätzfunktionen, wenn der Stichprobenumfang T groß ist. Sie schließt z.B. aus, daß eine Variable langfristig (für $T \rightarrow \infty$) einem linearen Trend folgt. Während die Rangbedingung in (A1) direkt für die Datenmatrix X überprüft werden kann, bleibt (A2) eine eher theoretische Annahme.
- Zu (A3): Mit der Annahme (A3) wird der unsystematische, nicht prognostizierbare Charakter der Störgröße erfaßt. Mit $E(u_t) = 0$ ist der Erwartungswert der Störgrößen (nicht aber der y_t) konstant. Die konstante Varianz σ^2 ist neben dem Vektor β ein weiterer unbekannter Parameter des Modells.
- Zu (A4): Die Normalverteilungsannahme für die Störgröße wird bei der Maximum-Likelihood-Schätzung verwendet. Auch die Verteilung von Schätzfunktionen und Teststatistiken basiert bei kleinen Stichproben auf der Normalverteilungsannahme.

7.3 Die Methode der Kleinsten Quadrate

Im folgenden gehen wir davon aus, daß ein lineares Regressionsmodell mit (A1), (A2), (A3) die stochastische Gesetzmäßigkeit darstellt, nach der zu gegebener $(T \times K)$ -Regressormatrix X Beobachtungswerte y_1, \dots, y_T der abhängigen Variablen realisiert worden sind.



Wenn ein einfaches Regressionsmodell ($K = 2$) vorliegt,

$$y_t = \beta_1 + \beta_2 x_t + u_t \quad t = 1, \dots, T,$$

dann ist also damit zu rechnen, daß die mit den Daten gegebenen Punkte $(x_1, y_1), \dots, (x_T, y_T)$ um die unbekannte Regressionsgerade $E(y) = \beta_1 + \beta_2 x$ streuen.

Aufgrund zweier Informationsbestandteile, nämlich

1. der Daten,
2. der Arbeitshypothese, daß ein lineares Modell zugrundeliegt,

soll die unbekannte Regressionsgerade (d.h. β_1 und β_2) geschätzt werden. Da nach 2. die Beobachtungspunkte um die unbekannte wahre Regressionsgerade streuen, werden wir umgekehrt aus den Datenpunkten diejenige Gerade als Regressionsgerade schätzen, von der die Datenpunkte einen minimalen Abstand haben, genauer, für die

$$\sum_{t=1}^T (y_t - (\hat{\beta}_1 + \hat{\beta}_2 x_t))^2$$

minimal ist. ($\hat{\beta}_1$ bzw. $\hat{\beta}_2$ sind hierbei Absolutglied bzw. Steigung der geschätzten Geraden.)

Wenn ein Regressionsmodell mit $K = 3$ angenommen wird,

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + u_t \quad t = 1, \dots, T,$$

ist ganz analog damit zu rechnen, daß die mit den Daten gegebenen Punkte (x_{t2}, x_{t3}, y_t) , $t = 1, \dots, T$, um die (unbekannte) Regressionsebene $E(y) = \beta_1 + \beta_2 x_2 + \beta_3 x_3$ streuen. Diese Regressionsebene bzw. die Parameter $\beta_1, \beta_2, \beta_3$ werden dann durch $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ so geschätzt, daß

$$\sum_{t=1}^T (y_t - (\hat{\beta}_1 + \hat{\beta}_2 x_{t2} + \hat{\beta}_3 x_{t3}))^2$$

minimal ist.

Dieses Anpassungsprinzip der **Methode der Kleinsten Quadrate, KQ-Methode** (auch: **Least Squares, LS**) wenden wir im nächsten Abschnitt zur Bestimmung einer Schätzfunktion $\hat{\beta}(y_1, \dots, y_T)$ des Parametervektors β im einfachen Regressionsmodell an.

Kapitel 8

Einfache lineare Regression

8.1 Daten und Modell

Für T Zeitpunkte (oder: T Querschnittseinheiten) liegen für gegebene Werte der unabhängigen Variablen x die Beobachtungswerte der Realisierungen der abhängigen Variablen y vor:

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_T, y_T). \quad (8.1)$$

Die Datenpaare können als Punkte in einem **Streuungsdiagramm** in der $x - y$ -Ebene dargestellt werden.

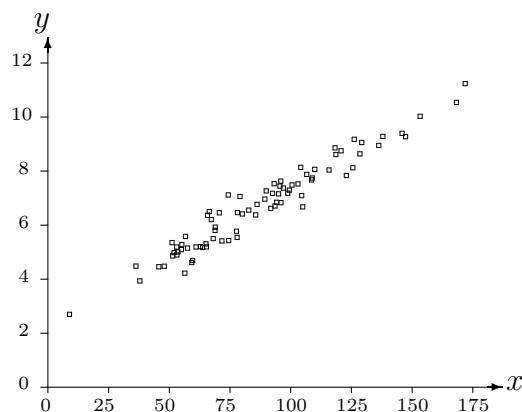


Abbildung 8.1: Streuungsdiagramm

Im einfachen linearen Regressionsmodell werden die Werte y_1, \dots, y_T als Realisierungen von T paarweise unkorrelierten Zufallsvariablen mit den Erwartungswerten $\beta_1 + \beta_2 x_t$, $t = 1, \dots, T$ und gleicher Varianz σ^2 betrachtet. Es wird also angenommen, dass die (unbeobachteten) Erwartungswerte auf einer

Geraden liegen:

$$y_t = \beta_1 + \beta_2 x_t + u_t, \quad t = 1, \dots, T, \quad (8.2)$$

wobei β_1 als **Absolutglied**, **Achsenabschnitt** oder **Konstante** bezeichnet wird und β_2 als **Geradensteigung**.

Unter Verwendung der Matrixschreibweise können die T Beobachtungsgleichungen mit einer Gleichung zusammengefasst werden zu:

$$y = X\beta + u$$

mit der Regressormatrix

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_T \end{bmatrix}$$

Wenn wir die Annahmen (A1) bis (A3) speziell für das einfache Regressionsmodell wiederholen, erhalten wir:

(A1) Die gegebene (deterministische) Regressormatrix hat den Rang $K = 2$. In Anbetracht der ersten Spalte, die aus lauter Einsen besteht, ist diese Rangbedingung gleichbedeutend damit, dass von den Werten x_1, \dots, x_T mindestens zwei verschieden voneinander sind.

(A2) Die sogenannte Momentenmatrix ergibt sich hier als

$$\frac{1}{T} X'X = \begin{bmatrix} 1 & \frac{1}{T} \sum x_t \\ \frac{1}{T} \sum x_t & \frac{1}{T} \sum x_t^2 \end{bmatrix}.$$

Es wird angenommen, dass diese Matrix mit wachsendem Stichprobenumfang T gegen eine endliche, nichtsinguläre Matrix konvergiert.

(A3) (unverändert)

$$E(u_t) = 0$$

$$Var(u_t) = \sigma^2$$

$$Cov(u_t, u_s) = 0 \quad \text{für alle} \quad t \neq s \quad (\text{keine Autokorrelation}).$$

8.2 Kleinst-Quadrate-Schätzung der Regressionsgeraden

8.2.1 Bestimmung der KQ-Schätzfunktion

Durch jede Festlegung von numerischen Werten β_1^0 (Absolutglied) und β_2^0 (Steigung) für die unbekannt Parameter β_1, β_2 wird eine Gerade festgelegt. Zu jedem Datenpunkt (x_t, y_t) kann dann die (auch als **Residuum** bezeichnete) Abweichung e_t^0 zwischen dem beobachteten Wert y_t und dem zu x_t gehörigen Wert $y_t^0 = \beta_1^0 + \beta_2^0 x_t$ auf der Geraden bestimmt werden:

$$e_t^0 = y_t - (\beta_1^0 + \beta_2^0 x_t) \quad , \quad t = 1, \dots, T.$$

Wenn sämtliche T Residuen zu einem Vektor zusammengefasst werden, erhält man diesen $(T \times 1)$ -Vektor von Residuen auch in Matrix-Notation als

$$e^0 = Y - X\beta^0$$

Die Schätzwerte $\hat{\beta}_1, \hat{\beta}_2$ für die unbekannt wahren Parameter β_1, β_2 heißen **KQ-Schätzwerte** (Schätzwerte nach der Methode der Kleinsten Quadrate, LS-Schätzwerte), wenn für die mit $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)'$ berechneten Residuen

$$e_t = y_t - (\hat{\beta}_1 + \hat{\beta}_2 x_t) \quad , \quad t = 1, \dots, T.$$

gilt:

$$\sum_{t=1}^T e_t^2 = \min_{\beta^0} \sum_{t=1}^T (e_t^0)^2$$

Satz: Wenn (A1) gilt, dann sind die KQ-Schätzwerte $\hat{\beta}_1, \hat{\beta}_2$ eindeutig bestimmt und es ist:

$$\hat{\beta}_2 = \frac{\sum_{t=1}^T (x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^T (x_t - \bar{x})^2} \quad \text{und} \quad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}. \quad (8.3)$$

Für die Berechnung des Schätzwerts für die Steigung, $\hat{\beta}_2$, kann man äquivalent auch die folgende Formel verwenden:

$$\hat{\beta}_2 = \frac{\sum_{t=1}^T x_t y_t - T \bar{x} \bar{y}}{\sum_{t=1}^T x_t^2 - T \bar{x}^2} = \frac{T \sum_{t=1}^T x_t y_t - \left(\sum_{t=1}^T x_t\right) \left(\sum_{t=1}^T y_t\right)}{T \sum_{t=1}^T x_t^2 - \left(\sum_{t=1}^T x_t\right)^2}.$$

Unter Verwendung der Matrizeschreibweise kann der KQ-Vektor $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)'$ auch kurz geschrieben werden als

$$\hat{\beta} = (X'X)^{-1}X'y \quad (8.4)$$

Bevor der Satz bewiesen wird, illustrieren wir die Anwendung an einem Beispiel.

Beispiel: Ausgaben in Abhängigkeit vom Haushaltseinkommen

Für eine Zufallsstichprobe von $T = 7$ Haushalten liegen die monatlichen Ausgaben für Nahrungs- und Genussmittel (y , in 100 €) und das jeweilige Haushaltseinkommen (x , in 100 €) vor. Die Tabelle enthält diese Daten zusammen mit einigen Zwischenergebnissen, aus denen dann sofort die beiden Schätzwerte $\hat{\beta}_2$ und $\hat{\beta}_1$ berechnet werden.

y_t	9	15	7	11	5	8	9	$\sum y_t = 64$	$\sum y_t^2 = 646$
x_t	35	49	21	39	15	28	25	$\sum x_t = 212$	$\sum x_t^2 = 7222$
$x_t y_t$	315	735	147	429	75	224	225	$\sum x_t y_t = 2150$	

$$\hat{\beta}_2 = \frac{T \sum_{t=1}^T x_t y_t - \left(\sum_{t=1}^T x_t\right) \left(\sum_{t=1}^T y_t\right)}{T \sum_{t=1}^T x_t^2 - \left(\sum_{t=1}^T x_t\right)^2} = \frac{15050 - 212 \cdot 64}{50554 - (212)^2} = \frac{1482}{5610} = 0.2642$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} = 1.1422$$

Die geschätzte Regressionsgerade ist also gegeben durch:

$$\hat{y} = 1.1422 + 0.2642 x.$$

Als Residuen e_1, \dots, e_7 erhält man mit $e_t = y_t - \hat{y}_t$ (gerundet):

$$e_t : \quad -1.389 \quad 0.912 \quad 0.310 \quad -0.446 \quad -0.105 \quad -0.540 \quad 1.253.$$

Beweis des Satzes: Notwendige Bedingungen für ein relatives Minimum der Summe der quadrierten Residuen

$$\sum_{t=1}^T (e_t^0)^2 = \sum_{t=1}^T (y_t - \beta_1^0 - \beta_2^0 x_t)^2 =: g(\beta_1^0, \beta_2^0)$$

erhält man, indem die partiellen Ableitungen von g nach β_1^0 und β_2^0 gleich Null gesetzt werden. Die beiden partiellen Ableitungen sind:

$$\frac{\partial g}{\partial \beta_1^0} = -2 \sum_{t=1}^T (y_t - \beta_1^0 - \beta_2^0 x_t)$$

$$\frac{\partial g}{\partial \beta_2^0} = -2 \sum_{t=1}^T (y_t - \beta_1^0 - \beta_2^0 x_t) x_t$$

Als notwendige Bedingungen für die KQ-Schätzwerte $\hat{\beta}_1, \hat{\beta}_2$ erhält man somit die beiden folgenden linearen Gleichungen in $\hat{\beta}_1, \hat{\beta}_2$, die auch als **System der Normalgleichungen** für die KQ-Schätzwerte bezeichnet werden:

$$\sum_{t=1}^T (\hat{\beta}_1 + \hat{\beta}_2 x_t) = \sum_{t=1}^T y_t \quad (8.5)$$

$$\sum_{t=1}^T (\hat{\beta}_1 x_t + \hat{\beta}_2 x_t^2) = \sum_{t=1}^T x_t y_t \quad (8.6)$$

bzw.

$$T \hat{\beta}_1 + \sum x_t \hat{\beta}_2 = \sum y_t$$

$$\sum x_t \hat{\beta}_1 + \sum x_t^2 \hat{\beta}_2 = \sum x_t y_t.$$

Die Lösung der Normalgleichungen ist dann gegeben durch:

$$\hat{\beta}_1 = \bar{y} - \bar{x} \hat{\beta}_2$$

$$\hat{\beta}_2 = \frac{\sum_{t=1}^T x_t y_t - T \bar{x} \bar{y}}{\sum_{t=1}^T x_t^2 - T \bar{x}^2}$$

Man beachte, dass $\hat{\beta}_2$ und damit auch $\hat{\beta}_1$ nicht eindeutig definiert sind, wenn sämtliche Werte x_t übereinstimmen; dann sind Zähler und Nenner von $\hat{\beta}_2$ gleich Null. Dieser Fall wird durch die Annahme (A1) ausgeschlossen.

Die alternativen Schreibweisen für $\hat{\beta}_2$ ergeben sich durch einfache Umformungen. Die Darstellung der KQ-Schätzwerte in matrizieller Schreibweise kann man unter Verwendung der weiter oben berechneten Matrix

$$\frac{1}{T} X'X = \begin{bmatrix} 1 & \frac{1}{T} \sum x_t \\ \frac{1}{T} \sum x_t & \frac{1}{T} \sum x_t^2 \end{bmatrix}$$

auch leicht bestätigen. Das **System der Normalgleichungen** für die KQ-Schätzwerte ergibt sich damit in Matrixschreibweise als:

$$X'X\hat{\beta} = X'y.$$

Aus dieser Matrixgleichung in $\hat{\beta}$ erhält man durch Multiplikation mit $(X'X)^{-1}$

$$\hat{\beta} = (X'X)^{-1}X'y$$

die eindeutig bestimmte Lösung der Normalgleichungen. Wegen (A1) ist der Rang von X gleich 2 und damit ist auch der Rang der (2×2) -Matrix $X'X$ gleich 2, d.h. die Inverse von $X'X$ ist definiert.

Hinreichende Bedingung für ein Minimum ist dann, dass die Hesse-Matrix der zweiten Ableitungen positiv definit ist. Man erhält als Hesse-Matrix die Matrix $2X'X$ und bestätigt leicht, dass diese Matrix positiv definit ist. Damit ist bewiesen, dass mit $\beta^0 = \hat{\beta}$ die Summe der quadrierten Residuen tatsächlich minimiert wird.

8.2.2 Streuungszerlegung und Bestimmtheitsmaß R^2

Die beiden Normalgleichungen

$$\begin{aligned} \sum_{t=1}^T (y_t - (\hat{\beta}_1 + \hat{\beta}_2 x_t)) &= 0 \\ \sum_{t=1}^T (y_t - (\hat{\beta}_1 + \hat{\beta}_2 x_t)) x_t &= 0 \end{aligned}$$

besagen, dass für die KQ-Residuen e_t gilt:

$$\sum e_t = 0 \quad (\Rightarrow \bar{y} = \bar{\hat{y}}) \quad (8.7)$$

$$\sum e_t x_t = 0 \quad (8.8)$$

Daraus folgt mit

$$\sum \hat{y}_t e_t = \sum (\hat{\beta}_1 + \hat{\beta}_2 x_t) e_t = \hat{\beta}_1 \sum e_t + \hat{\beta}_2 \sum x_t e_t$$

auch unmittelbar

$$\sum \hat{y}_t e_t = 0 \quad (8.9)$$

Weiterhin ergibt sich damit, ähnlich wie im Fall der Varianzanalyse, eine Zerlegung der gesamten Streuung der y_t -Werte um \bar{y} in die durch die Variable x erklärte Streuung und in die zufallsbedingte, unerklärte Streuung der KQ-Residuen e_t um deren Mittelwert 0:¹

$$\sum (y_t - \bar{y})^2 = \sum (\hat{y}_t - \bar{y})^2 + \sum e_t^2. \quad (8.10)$$

oder äquivalent

$$\frac{1}{T} \sum (y_t - \bar{y})^2 = \underbrace{\frac{1}{T} \sum (\hat{y}_t - \bar{y})^2}_{\text{erklärte Varianz}} + \underbrace{\frac{1}{T} \sum e_t^2}_{\text{unerklärte Varianz}} \quad (8.11)$$

Die Gültigkeit dieser Zerlegung folgt unter Beachtung von $\bar{y} = \bar{\hat{y}}$ und

$$y_t - \bar{y} = y_t - \hat{y}_t + \hat{y}_t - \bar{y} = e_t + \hat{y}_t - \bar{y}$$

mit:

$$\begin{aligned} \sum (y_t - \bar{y})^2 &= \sum (e_t + \hat{y}_t - \bar{y})^2 \\ &= \sum e_t^2 + \sum (\hat{y}_t - \bar{y})^2 + 2 \sum (\hat{y}_t - \bar{y})e_t. \end{aligned}$$

Dabei ist der letzte, gemischte Term wegen (8.9) und (8.7) gleich Null.

Aufbauend auf der Varianzzerlegung in erklärte und unerklärte Varianz definiert man das **multiple Bestimmtheitsmaß** R^2 als den Anteil der (durch x) erklärten Varianz an der gesamten Varianz der Beobachtungswerte y_1, y_2, \dots, y_T :

$$R^2 := \frac{\sum (\hat{y}_t - \bar{y})^2}{\sum (y_t - \bar{y})^2} = 1 - \frac{\sum (e_t^2)}{\sum (y_t - \bar{y})^2}, \quad 0 \leq R^2 \leq 1. \quad (8.12)$$

Im einfachen linearen Regressionsmodell ist R^2 numerisch identisch mit dem quadrierten Korrelationskoeffizienten r_{XY}^2 , das heißt

$$R^2 = r_{XY}^2, \quad (8.13)$$

¹Man beachte, dass es bei der in diesem Abschnitt vorgenommenen Betrachtung der Varianzzerlegung von y nicht um die bedingte Varianz σ^2 der Zufallsvariablen y_t bei gegebenem x_t geht, sondern um die Varianz der zu den unterschiedlichen x_t -Werten realisierten Beobachtungswerte y_1, y_2, \dots, y_T der abhängigen Variablen.

denn

$$\begin{aligned}
 r_{XY}^2 &= \frac{[\sum(x_t - \bar{x})(y_t - \bar{y})]^2}{\sum(x_t - \bar{x})^2 \cdot \sum(y_t - \bar{y})^2} \\
 &= \hat{\beta}_2 \cdot \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sum(y_t - \bar{y})^2} \\
 &= \frac{\sum(\hat{\beta}_2 x_t - \hat{\beta}_2 \bar{x})(y_t - \bar{y})}{\sum(y_t - \bar{y})^2} \\
 &= \frac{\sum(\hat{y}_t - \bar{\hat{y}})(y_t - \bar{y})}{\sum(y_t - \bar{y})^2} \\
 &= \frac{\sum(\hat{y}_t - \bar{\hat{y}})(e_t + \hat{y}_t - \bar{\hat{y}})}{\sum(y_t - \bar{y})^2} \\
 &= \frac{\sum(\hat{y}_t - \bar{\hat{y}})^2}{\sum(y_t - \bar{y})^2} \\
 &= R^2.
 \end{aligned}$$

Mit dieser Beziehung ist auch nachträglich gezeigt, dass der Korrelationskoeffizient zwischen -1 und $+1$ liegt. Außerdem kann die Größenordnung des Korrelationskoeffizienten durch diese Beziehung besser interpretiert werden. Beispielsweise impliziert ein Korrelationskoeffizient von 0.5 , dass die dem $x - y$ -Streudiagramm angepasste Regressionsgerade nur 25% der Streuung der y_t -Werte erklärt, während 75% der Streuung auf die Schwankungen um die Regressionsgerade entfällt. In dem oben betrachteten Beispiel (Ausgaben in Abhängigkeit vom Haushaltseinkommen) berechnet man für das multiple Bestimmtheitsmaß den Wert $R^2 = 0.92$, d.h. 92% der Streuung der Ausgaben wird durch das unterschiedliche Einkommen erklärt.

8.3 Die Wahrscheinlichkeitsverteilung der Schätzfunktionen $\hat{\beta}_1$ und $\hat{\beta}_2$

Im vorigen Abschnitt wurde die Kleinst-Quadrate-Schätzung im wesentlichen deskriptiv im Hinblick auf die Anpassung einer Geraden in einem Streudiagramm behandelt; es wurde nur die Annahme (A1), nicht aber die Verteilungsannahmen (A3) bzw. (A4) verwendet. Im folgenden ändert sich die Perspektive insofern, als wir den Stichprobencharakter der beobachteten y_t -Werte betonen und die Formeln zu KQ-Schätzung von β_1 und β_2 als Stichprobenfunktionen, d.h. als Funktionen von Zufallsvariablen y_1, y_2, \dots, y_T betrachten. Die berechneten KQ-Schätzwerte $\hat{\beta}_1$ und $\hat{\beta}_2$ sind damit als Realisierungen von Zufallsvariablen anzusehen, die bei unterschiedlichen Stich-

probenergebnissen auch unterschiedliche Werte annehmen. Jetzt steht nicht mehr die Anpassungsgüte im Vordergrund, sondern die Beurteilung der Frage, in welcher Beziehung die Verteilung von $\hat{\beta}_1$ und $\hat{\beta}_2$ zu den unbekanntem wahren Parametern β_1 und β_2 steht, insbesondere die Fragen:

- Sind $\hat{\beta}_1$ und $\hat{\beta}_2$ erwartungstreu für β_1 und β_2 ?
- Wie groß ist die Varianz von $\hat{\beta}_1$ und $\hat{\beta}_2$?
- Wie können Parametertests auf hypothetische Parameterwerte für β_1 und β_2 durchgeführt werden?

8.3.1 Erwartungstreue von $\hat{\beta}_1$ und $\hat{\beta}_2$

Die Schätzfunktionen

$$\hat{\beta}_2 = \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{\sum(x_t - \bar{x})^2}, \quad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

sind lineare Funktionen in den Zufallsvariablen y_1, y_2, \dots, y_T . Um die Erwartungswerte der beiden Schätzfunktionen zu bestimmen, betrachten wir die Erwartungswerte von y_t und \bar{y} . Aufgrund der Modellannahme (A3) gilt bei gegebenem Wert für x_t

$$\begin{aligned} E(y_t) &= \beta_1 + \beta_2 x_t, \\ E(\bar{y}) &= \frac{1}{T} \sum E(y_t) = \frac{1}{T} \sum (\beta_1 + \beta_2 x_t) = \beta_1 + \beta_2 \bar{x} \end{aligned}$$

also

$$E(y_t - \bar{y}) = \beta_2(x_t - \bar{x}).$$

Daraus folgt durch Einsetzen in die Schätzfunktionen:

$$E(\hat{\beta}_2) = \beta_2, \quad \text{und} \quad E(\hat{\beta}_1) = \beta_1.$$

Somit sind $\hat{\beta}_1$ und $\hat{\beta}_2$ **erwartungstreu Schätzer** für β_1 und β_2 .

8.3.2 Die Varianz von $\hat{\beta}_1$ und $\hat{\beta}_2$

Um die Varianz des Schätzers des Steigungskoeffizienten, $Var(\hat{\beta}_2)$, zu bestimmen, stellen wir zunächst fest, dass wir die Schätzfunktion $\hat{\beta}_2$ äquivalent auch als

$$\hat{\beta}_2 = \frac{\sum(x_t - \bar{x})y_t}{\sum(x_t - \bar{x})^2} \tag{8.14}$$

schreiben können (denn $\sum(x_t - \bar{x})\bar{y} = 0$). Damit erhält man, da nach (A3) die Zufallsvariablen y_1, y_2, \dots, y_T paarweise unkorreliert sind und jeweils die gleiche Varianz σ^2 haben:

$$\text{Var}(\hat{\beta}_2) = \frac{\sum(x_t - \bar{x})^2 \sigma^2}{(\sum(x_t - \bar{x})^2)^2} = \frac{\sigma^2}{\sum(x_t - \bar{x})^2}. \quad (8.15)$$

Wie man sieht, wird die Varianz von $\hat{\beta}_2$ mit wachsender Varianz σ^2 (d.h. wachsender Streuung der Werte um die wahre Regressionsgerade) größer und mit wachsender Summe $\sum(x_t - \bar{x})^2$ kleiner. Wegen

$$\sum(x_t - \bar{x})^2 = T \cdot \frac{\sum(x_t - \bar{x})^2}{T}$$

nimmt die Varianz von $\hat{\beta}_2$ also mit wachsendem Stichprobenumfang T oder wachsender Streuung der unabhängigen Variablen x_t ab.

Für die Varianz des Schätzers des Absolutglieds, $\text{Var}(\hat{\beta}_1)$, ergibt sich etwas aufwändiger (wenn man auf die Matrixschreibweise verzichtet):

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2 \sum x_t^2 / T}{\sum(x_t - \bar{x})^2}. \quad (8.16)$$

Wenn die Annahme (A4) (Normalverteilung der Störgrößen bzw. der y_t) gilt oder wenn der Stichprobenumfang T groß ist, sind $\hat{\beta}_1$ und $\hat{\beta}_2$ normalverteilt, jeweils um den wahren Parameter β_1 bzw. β_2 mit der Varianz (8.16) bzw. (8.15).

Nach dem Satz von **Gauß-Markov**, den wir hier nicht beweisen, sind die KQ-Schätzfunktionen effizient in der Klasse aller erwartungstreuen Schätzfunktionen, die linear in den Stichprobenvariablen y_1, y_2, \dots, y_T sind, d.h. es gibt keine anderen Schätzfunktionen in dieser Klasse, die eine kleinere Varianz haben.

Die gegebenen Varianzen hängen noch von dem unbekanntem Parameter σ^2 ab. Wie gezeigt werden kann, ist die folgende Schätzfunktion eine **erwartungstreue** Schätzfunktion für σ^2 :

$$\widehat{\text{Var}}(u_t) = \hat{\sigma}^2 = \frac{1}{T-2} \sum_{t=1}^T e_t^2. \quad (8.17)$$

Rechnerisch bequemer (da man nicht die einzelnen Residuen ausrechnen muss) ist die folgende äquivalente Formel:

$$\hat{\sigma}^2 = \frac{T \sum y_t^2 - (\sum y_t)^2 - \hat{\beta}_2 (T \sum x_t y_t - \sum x_t \sum y_t)}{T(T-2)}. \quad (8.18)$$

Damit ergeben sich die geschätzten Varianzen der beiden Schätzfunktionen $\hat{\beta}_1$ und $\hat{\beta}_2$ als

$$\hat{\sigma}_{\hat{\beta}_2}^2 = \frac{\hat{\sigma}^2}{\sum(x_t - \bar{x})^2} \quad \text{und} \quad \hat{\sigma}_{\hat{\beta}_1}^2 = \frac{\hat{\sigma}^2 \sum x_t^2 / T}{\sum(x_t - \bar{x})^2}. \quad (8.19)$$

Beispiel: Ausgaben in Abhängigkeit vom Haushaltseinkommen

Mit den gegebenen Daten erhält man für die Schätzung von σ^2 :

$$\begin{aligned} \hat{\sigma}^2 &= \frac{T \sum y_t^2 - (\sum y_t)^2 - \hat{\beta}_2(T \sum x_t y_t - \sum x_t \sum y_t)}{T(T-2)} \\ &= \frac{7 \cdot 646 - 64^2 - 0.2642 \cdot (7 \cdot 2150 - 212 \cdot 64)}{7 \cdot 5} = 0.9844 \\ \hat{\sigma} &= 0.992 \end{aligned}$$

Damit berechnen wir die geschätzten Standardabweichungen der Koeffizientenschätzer mit:

$$\begin{aligned} \hat{\sigma}_{\hat{\beta}_2} &= \sqrt{\widehat{Var}(\hat{\beta}_2)} = \frac{\hat{\sigma}}{\sqrt{\sum x_t^2 - T\bar{x}^2}} = \frac{0.992}{\sqrt{801.4286}} = 0.0350 \\ \hat{\sigma}_{\hat{\beta}_1} &= \sqrt{\widehat{Var}(\hat{\beta}_1)} = \frac{\hat{\sigma} \sqrt{\sum x_t^2 / T}}{\sqrt{\sum x_t^2 - T\bar{x}^2}} = \hat{\sigma}_{\hat{\beta}_2} \sqrt{\sum x_t^2 / T} = 1.126 \end{aligned}$$

Zusammenfassend erhalten wir also für die beiden Koeffizientenschätzer und deren (geschätzte) Standardabweichung:

$$\begin{aligned} \hat{\beta}_1 &= 1.1422, \quad \hat{\sigma}_{\hat{\beta}_1} = 1.126 \\ \hat{\beta}_2 &= 0.2642, \quad \hat{\sigma}_{\hat{\beta}_2} = 0.035. \end{aligned}$$

Abschließend geben wir in matrizieller Notation die beiden Varianzen und die Kovarianz der beiden Schätzfunktionen zusammengefasst in der Varianz-Kovarianz-Matrix von $\hat{\beta}$ an:

$$\begin{aligned} V(\hat{\beta}) &:= \begin{bmatrix} Var(\hat{\beta}_1) & Cov(\hat{\beta}_1, \hat{\beta}_2) \\ Cov(\hat{\beta}_2, \hat{\beta}_1) & Var(\hat{\beta}_2) \end{bmatrix} = \sigma^2 (X'X)^{-1} \\ &= \sigma^2 \begin{bmatrix} T & \sum x_t \\ \sum x_t & \sum x_t^2 \end{bmatrix}^{-1} \\ &= \frac{\sigma^2}{T \sum x_t^2 - (\sum x_t)^2} \begin{bmatrix} \sum x_t^2 & -\sum x_t \\ -\sum x_t & T \end{bmatrix}. \end{aligned}$$

Wenn der unbekannte Parameter σ^2 durch die erwartungstreue Schätzung der Varianz $\hat{\sigma}^2$ ersetzt wird, erhält man die geschätzte Varianz-Kovarianz-Matrix von $\hat{\beta}$:

$$\begin{aligned}\hat{V}(\hat{\beta}) &= \begin{bmatrix} \widehat{Var}(\hat{\beta}_1) & \widehat{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \widehat{Cov}(\hat{\beta}_2, \hat{\beta}_1) & \widehat{Var}(\hat{\beta}_2) \end{bmatrix} = \hat{\sigma}^2 (X'X)^{-1} \\ &= \frac{\hat{\sigma}^2}{T \sum x_t^2 - (\sum x_t)^2} \begin{bmatrix} \sum x_t^2 & -\sum x_t \\ -\sum x_t & T \end{bmatrix}.\end{aligned}$$

Mit den Daten des obigen Beispiels berechnet man als geschätzte Varianz-Kovarianz-Matrix

$$\hat{V}(\hat{\beta}) = \begin{bmatrix} \widehat{Var}(\hat{\beta}_1) & \widehat{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \widehat{Cov}(\hat{\beta}_2, \hat{\beta}_1) & \widehat{Var}(\hat{\beta}_2) \end{bmatrix} = \begin{bmatrix} 1.27 & -0.0372 \\ -0.0372 & 0.00123 \end{bmatrix}.$$

8.3.3 Konfidenzintervalle und Tests für die Parameter β_1 und β_2

Konfidenzintervalle und Test basieren auf der t -Verteilung mit $T - 2$ Freiheitsgraden der mit den geschätzten Standardabweichungen standardisierten Schätzfunktionen

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \quad \text{und} \quad \frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}_{\hat{\beta}_2}}.$$

Das $(1 - \alpha)$ -**Konfidenzintervall** für β_2 ist gegeben durch

$$\left[\hat{\beta}_2 - t_{1-\frac{\alpha}{2}; T-2} \cdot \hat{\sigma}_{\hat{\beta}_2}, \hat{\beta}_2 + t_{1-\frac{\alpha}{2}; T-2} \cdot \hat{\sigma}_{\hat{\beta}_2} \right], \quad (8.20)$$

entsprechend für das Absolutglied β_1 .

Beispiel: Ausgaben in Abhängigkeit vom Haushaltseinkommen

Konfidenzintervall für die Steigung ($1 - \alpha = 0.95$):

$$\begin{aligned}\left[\hat{\beta}_2 \pm t_{0.975; T-2} \cdot \hat{\sigma}_{\hat{\beta}_2} \right] &= [0.2642 \pm 2.571 \cdot 0.035] = [0.2642 \pm 0.0900] \\ &= [0.17, 0.35].\end{aligned}$$

Konfidenzintervall für das Absolutglied:

$$\begin{aligned}\left[\hat{\beta}_1 \pm t_{0.975; T-2} \cdot \hat{\sigma}_{\hat{\beta}_1} \right] &= [1.1422 \pm 2.571 \cdot 1.126] = [1.1422 \pm 2.8949] \\ &= [-1.75, 4.04].\end{aligned}$$

Entsprechend werden **Tests** auf hypothetische Parameterwerte als t -Tests durchgeführt. Wir betrachten einen solchen Test für den Regressionskoeffizienten β_2 , der den Einfluss der erklärenden Variablen x misst. Wir testen also

$$H_0 : \beta_2 = \beta_2^0 \text{ gegen } H_1 : \beta_2 \neq \beta_2^0.$$

Als Teststatistik verwenden wir

$$t = \frac{\hat{\beta}_2 - \beta_2^0}{\hat{\sigma}_{\hat{\beta}_2}}. \quad (8.21)$$

Wenn H_0 gilt, dann ist die Teststatistik t -verteilt mit $T - 2$ Freiheitsgraden. H_0 wird also abgelehnt, wenn

$$|t| > t_{1-\frac{\alpha}{2}; T-2}. \quad (8.22)$$

Um zu überprüfen, ob die erklärende Variable einen signifikanten Beitrag zur Erklärung der Variation der abhängigen Variablen leistet, wird

$$H_0 : \beta_2 = 0 \text{ gegen } H_1 : \beta_2 \neq 0$$

getestet. Der realisierte Wert der Teststatistik

$$t = \frac{\hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_2}}$$

wird oft (z.B. in Statistik-Programm-Paketen) als t -Wert des Koeffizienten bezeichnet. Offensichtlich gibt der t -Wert an, welches Vielfache der (geschätzten) Standardabweichung der Schätzwert beträgt.

Beispiel: Ausgaben in Abhängigkeit vom Haushaltseinkommen

Die t -Werte für β_1 und β_2 sind:

$$\begin{aligned} \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} &= \frac{1.1422}{1.126} = 1.01, \\ \frac{\hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_2}} &= \frac{0.2642}{0.0350} = 7.55. \end{aligned}$$

Wie man leicht feststellt, ist zum Signifikanzniveau $\alpha = 0.05$ das Absolutglied nicht signifikant von Null verschieden, während der Koeffizient β_2 , der den Einfluss des Einkommens auf die Ausgaben misst, hoch signifikant ist.

8.3.4 Bedingte Prognose für y bei gegebenem Wert von x

Unter der Annahme der Gültigkeit des linearen Modells für die Abhängigkeit der Variablen y von der erklärenden Variablen x können wir auf der Grundlage der geschätzten Parameter Prognosen bzw. Prognose-Intervalle für die Realisierung von y_0 bei gegebenem Wert $x = x_0$ konstruieren. Nach dem linearen Modell gilt für y_0 :

$$y_0 = \beta_1 + \beta_2 x_0 + u_0 \quad \text{mit } E(u_0) = 0, \text{Var}(u_0) = \sigma^2. \quad (8.23)$$

Unter Verwendung der erwartungstreuen Schätzfunktionen $\hat{\beta}_1, \hat{\beta}_2$ verwenden wir als **Punktprognose**

$$\hat{y}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_0, \quad (8.24)$$

also den zu x_0 gehörigen Wert von y auf der geschätzten Regressionsgeraden. Der Prognosefehler, den wir mit v_0 bezeichnen, wird zum einen durch die Verwendung der Schätzwerte (anstelle der unbekanntenen wahren Parameter) verursacht, zum anderen durch die nicht prognostizierbare zufällige Abweichung u_0 von der wahren Regressionsgeraden:

$$\begin{aligned} v_0 = y_0 - \hat{y}_0 &= \beta_1 + \beta_2 x_0 + u_0 - \hat{\beta}_1 - \hat{\beta}_2 x_0 \\ &= u_0 - ((\hat{\beta}_1 - \beta_1) + (\hat{\beta}_2 - \beta_2)x_0). \end{aligned} \quad (8.25)$$

Für den Erwartungswert des Prognosefehlers ergibt sich

$$E(v_0) = E(u_0) - (E(\hat{\beta}_1 - \beta_1) + (E(\hat{\beta}_2 - \beta_2))x_0) = 0. \quad (8.26)$$

Die beiden Fehlerkomponenten (Störgröße und Schätzfehler) sind unkorreliert, da die Stichprobenvariablen y_1, \dots, y_T , die in die Schätzfunktionen eingehen, unkorreliert mit u_0 sind. Die Varianz des Prognosefehlers ergibt sich somit als die Summe der Varianzen der beiden Fehlerkomponenten:²

$$\begin{aligned} \text{Var}(u_0) &= \sigma^2 \\ \text{Var}((\hat{\beta}_1 - \beta_1) + (\hat{\beta}_2 - \beta_2)x_0) &= \sigma^2 \left(\frac{1}{T} + \frac{(x_0 - \bar{x})^2}{\sum (x_t - \bar{x})^2} \right), \end{aligned}$$

also

$$\text{Var}(v_0) = \sigma_0^2 = \sigma^2 \left(1 + \frac{1}{T} + \frac{(x_0 - \bar{x})^2}{\sum (x_t - \bar{x})^2} \right). \quad (8.27)$$

²In die Berechnung der Varianz des Schätzfehlers gehen nicht nur die beiden Varianzen, sondern auch die Kovarianz von $\hat{\beta}_1$ und $\hat{\beta}_2$ ein.

Wie man sieht, wird die Varianz des Prognosefehlers um so größer, je weiter x_0 von \bar{x} entfernt ist.

Bei Gültigkeit der Normalverteilungsannahme (A4) ist auch der Prognosefehler v_0 normalverteilt, also $v_0 \sim N(0, \sigma_0^2)$. Um ein Prognose-Intervall um die Punktprognose zu konstruieren, muss die unbekannte Varianz σ^2 durch deren erwartungstreue Schätzfunktion ersetzt werden, womit sich die geschätzte Varianz des Prognosefehlers v_0 ergibt:

$$\widehat{Var}(v_0) = \hat{\sigma}_0^2 = \hat{\sigma}^2 \left(1 + \frac{1}{T} + \frac{(x_0 - \bar{x})^2}{\sum(x_t - \bar{x})^2} \right). \quad (8.28)$$

Der durch die geschätzte Standardabweichung $\hat{\sigma}_0 = \sqrt{\hat{\sigma}_0^2}$ dividierte Prognosefehler ist t -verteilt mit $T-2$ Freiheitsgraden,

$$\frac{v_0}{\hat{\sigma}_0} = \frac{y_0 - \hat{y}_0}{\hat{\sigma}_0} \sim t(T-2). \quad (8.29)$$

Damit ergibt sich als Prognose-Intervall zum Niveau $1-\alpha$ (d.h. das Prognose-Intervall wird den realisierten Wert von y_0 mit Wahrscheinlichkeit $1-\alpha$ enthalten) als

$$\left[\hat{y}_0 - t_{1-\frac{\alpha}{2}; T-2} \cdot \hat{\sigma}_0, \hat{y}_0 + t_{1-\frac{\alpha}{2}; T-2} \cdot \hat{\sigma}_0 \right], \quad (8.30)$$

wobei

$$\hat{y}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_0.$$

Beispiel: Ausgaben in Abhängigkeit vom Haushaltseinkommen

Es wird ein Prognose-Intervall für die Ausgaben für Nahrungs- und Genussmittel eines Haushalts mit dem Einkommen $x_0 = 5000$ [€] zum 95 %-Niveau bestimmt. Die Punktprognose für die Ausgaben y_0 beträgt:

$$\hat{y}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_0 = 1.1422 + 0.2642 \cdot 50 = 14.3522 [100 \text{ €}] = 1435.22 [\text{€}].$$

Die geschätzte Varianz des Prognosefehlers berechnet man als:

$$\begin{aligned} \widehat{Var}(v_0) = \hat{\sigma}_0^2 &= \hat{\sigma}^2 \left(1 + \frac{1}{T} + \frac{(x_0 - \bar{x})^2}{\sum(x_t - \bar{x})^2} \right) \\ &= 0.9844 \left(1 + \frac{1}{7} + \frac{(50 - \frac{212}{7})^2}{801.4286} \right) = 1.628 \end{aligned}$$

Mit $t_{0.975; 5} = 2.571$ ist $t_{0.975; 5} \cdot \hat{\sigma}_0 = 3.2802$, so dass sich als 95%-Prognose-Intervall ergibt:

$$\begin{aligned} &\left[\hat{y}_0 - t_{1-\frac{\alpha}{2}; T-2} \cdot \hat{\sigma}_0, \hat{y}_0 + t_{1-\frac{\alpha}{2}; T-2} \cdot \hat{\sigma}_0 \right] \\ &= [11.07, 17.63] [100 \text{ €}] \quad \text{bzw.} \quad [1107, 1763] [\text{€}]. \end{aligned}$$

Kapitel 9

Multiple lineare Regression

Das multiple lineare Regressionsmodell

$$y_t = \beta_1 + \beta_2 x_{t2} + \dots + \beta_K x_{tK} + u_t$$

mit den Annahmen (A1) bis (A4) ist im 7. Kapitel bereits vorgestellt worden. Im Unterschied zum einfachen linearen Regressionsmodell, in dem mit $K = 2$ neben der Scheinvariablen (Absolutglied) nur eine echte erklärende Variable zugelassen wurde, gehen wir im folgenden allgemeiner von $K \geq 2$ aus und betrachten insbesondere mit $K > 2$ die Abhängigkeit des Erwartungswerts $E(y)$ von mehreren unabhängigen Variablen x_2, \dots, x_K .¹ Die zweckmäßige Notation des multiplen Regressionsmodells für alle T Beobachtungen erfolgt in matrizieller Schreibweise als

$$y = X\beta + u$$

unter Verwendung von

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1K} \\ 1 & x_{22} & \cdots & x_{2K} \\ \vdots & & & \vdots \\ 1 & x_{T2} & \cdots & x_{TK} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}.$$

¹Die multiple lineare Regression einer abhängigen Variablen y auf die Regressoren x_{t2}, \dots, x_{tK} ist zu unterscheiden von $K - 1$ einfachen Regressionen der abhängigen Variablen y auf jeweils einen dieser Regressoren.

9.1 Kleinst-Quadrate Schätzung der Parameter $\beta_1, \beta_2, \dots, \beta_K$

9.1.1 Anpassung der Regressionsebene

Die Kleinst-Quadrate Schätzwerte $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_K$ sind definiert als diejenigen Parameterwerte, die die Summe der quadrierten Residuen minimieren:

$$\min_{\beta^0} \sum_{t=1}^T \left(y_t - (\beta_1^0 + \beta_2^0 x_{t2} + \dots + \beta_K^0 x_{tK}) \right)^2.$$

Das Minimierungsproblem wird gelöst durch den Vektor

$$\hat{\beta} = (X'X)^{-1} X'y \quad (9.1)$$

wobei wie bisher X' die Transponierte zur Matrix X bezeichnet.

$\hat{\beta}$ ist eine lineare Funktion der Zufallsvariablen y_1, \dots, y_T , und für die Realisierungen y_1, \dots, y_T aus der Stichprobe ist es ein Vektor mit K numerischen Werten, d.h. den Parameterschätzwerten $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_K$. Es gibt zahlreiche statistische und ökonometrische Softwarepakete, mit denen die Kleinst-Quadrate-Schätzwerte berechnet werden können.

Geometrisch wird bei $K = 3$ den Punkten (x_{t2}, x_{t3}, y_t) eine Regressionsebene so gut wie möglich angepasst; auch bei $K > 3$ spricht man von der angepassten Regressionsebene. Für die Punkte $(x_{t2}, \dots, x_{tK}, \hat{y}_t)$ auf der Regressionsebene sind die angepassten Werte \hat{y}_t gegeben als

$$\hat{y}_t = \hat{\beta}_1 + \hat{\beta}_2 x_{t2} + \dots + \hat{\beta}_K x_{tK}$$

und die Kleinst-Quadrate-Residuen sind gegeben durch

$$e_t = y_t - \hat{y}_t, \quad t = 1, \dots, T.$$

Wie in der einfachen Regression gilt

$$\sum_t e_t = 0, \quad (\text{und somit: } \sum y_t = \sum \hat{y}_t \quad \text{und} \quad \bar{y} = \bar{\hat{y}}) \quad (9.2)$$

$$\sum_t e_t x_{ti} = 0, \quad i = 2, \dots, K. \quad (9.3)$$

Damit gilt auch wieder die Streuungszersetzung

$$\sum (y_t - \bar{y})^2 = \sum (\hat{y}_t - \bar{\hat{y}})^2 + \sum e_t^2. \quad (9.4)$$

oder äquivalent

$$\frac{1}{T} \sum (y_t - \bar{y})^2 = \underbrace{\frac{1}{T} \sum (\hat{y}_t - \bar{\hat{y}})^2}_{\text{erklärte Varianz}} + \underbrace{\frac{1}{T} \sum e_t^2}_{\text{unerklärte Varianz}} \quad (9.5)$$

9.1.2 Bestimmtheitsmaß und korrigiertes Bestimmtheitsmaß

Das **multiple Bestimmtheitsmaß**, bezeichnet mit R^2 , wird wie im einfachen Regressionsmodell definiert als der Anteil der erklärten Varianz an der gesamten Varianz der Beobachtungswerte y_1, \dots, y_T :

$$R^2 = \frac{\sum(\hat{y}_t - \bar{y})^2}{\sum(y_t - \bar{y})^2}, \quad 0 \leq R^2 \leq 1. \quad (9.6)$$

Äquivalent kann das multiple Bestimmtheitsmaß R^2 wegen (9.4) auch geschrieben werden als

$$R^2 = 1 - \frac{\sum e_t^2}{\sum(y_t - \bar{y})^2}. \quad (9.7)$$

Beim Vergleich des multiplen Bestimmtheitsmaßes R_I^2 , das man aus den Daten zu einem Modell (I) berechnet, mit dem multiplen Bestimmtheitsmaß R_{II}^2 , das man bei Berücksichtigung einer zusätzlichen erklärenden Variablen erhält, also

$$R_I^2 \text{ zu } y_t = \beta_1 + \beta_2 x_{t2} + \dots + \beta_K x_{tK} + u_t,$$

$$R_{II}^2 \text{ zu } y_t = \beta_1 + \beta_2 x_{t2} + \dots + \beta_K x_{tK} + \beta_{K+1} x_{t, K+1} + u_t,$$

ist der Wert R_{II}^2 für das zweite, erweiterte Modell stets mindestens so groß wie R_I^2 (ganz unabhängig davon, ob die zusätzliche Variable vom Sachzusammenhang her sinnvoll ist). Denn da mit den Parameterschätzwerten nach dem Prinzip der Kleinsten Quadrate die Summe der quadrierten Residuen minimiert wird, gilt

$$\sum e_{t,II}^2 \leq \sum \left(y_t - (\hat{\beta}_1 + \hat{\beta}_2 x_{t2} + \dots + \hat{\beta}_K x_{tK} + 0 \cdot x_{t, K+1}) \right)^2 = \sum e_{t,I}^2,$$

wobei $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_K$ die KQ-Schätzwerte im Modell (I) bezeichnen soll. Das multiple Bestimmtheitsmaß R^2 eignet sich daher nicht als ein Kriterium, um eine Entscheidung über die Einbeziehung der zusätzlichen erklärenden Variablen in das Modell zu treffen.² Das im folgenden vorgestellte **korrigierte Bestimmtheitsmaß** \overline{R}^2 enthält dagegen durch die Berücksichtigung der Anzahl der Parameter eine „Bestrafung“ für die Einbeziehung einer

²Grundsätzlich entspricht es unserer Vorgehensweise am ehesten, eine derartige Entscheidung mit Hilfe eines Test der Alternativen $H_0 : \beta_{K+1} = 0$ gegen $H_1 : \beta_{K+1} \neq 0$ zu treffen.

zusätzlichen Variablen, so dass es grundsätzlich möglich ist, dass das korrigierte Bestimmtheitsmaß des kurzen Modells (I) größer als das des erweiterten Modells (II) ist.

Das (durch Berücksichtigung der Freiheitsgrade) **korrigierte Bestimmtheitsmaß** $\overline{R^2}$ ist definiert als

$$\overline{R^2} = 1 - \frac{\frac{1}{T-K} \sum e_t^2}{\frac{1}{T-1} \sum (y_t - \bar{y})^2}, \quad (9.8)$$

oder äquivalent:

$$\overline{R^2} = 1 - (1 - R^2) \frac{T-1}{T-K}. \quad (9.9)$$

Das korrigierte Bestimmtheitsmaß $\overline{R^2}$ ist offensichtlich stets kleiner oder gleich eins; es kann auch (bei sehr schlechter Anpassung, genauer bei $R^2 < \frac{K-1}{T-1}$) negative Werte annehmen. Die auf der Streuungszerlegung basierende Interpretation als Anteil der erklärten Streuung an der gesamten Streuung gilt nicht mehr für das korrigierte Bestimmtheitsmaß.

9.1.3 Der Erklärungsbeitrag der einzelnen Variablen

In der deskriptiven Auswertung der multiplen linearen Regression stellt sich neben der Erfassung des Erklärungsbeitrags der unabhängigen Variablen insgesamt (mit dem multiplen Bestimmtheitsmaß R^2) auch die Frage nach dem relativen Beitrag der einzelnen Variablen. Die absolute Größe der einzelnen Schätzwerte $\hat{\beta}_2, \dots, \hat{\beta}_K$ ist dafür schon wegen deren Abhängigkeit von dem gewählten Maßstab für die einzelnen Variablen nicht geeignet. Wenn man z.B. die Maßeinheit für die abhängige Variable y von [1 €] in [1000 €] verändert, werden alle Beobachtungswerte y_1, \dots, y_T mit 0.001 multipliziert und damit auch die resultierenden Schätzwerte $\hat{\beta}_2, \dots, \hat{\beta}_K$. Wenn man für eine einzelne erklärende Variable x_k den Maßstab verändert, z.B. wieder von der Einheit [1 €] zur Einheit [1000 €] übergeht und damit die Beobachtungswerte x_{1k}, \dots, x_{Tk} mit 0.001 multipliziert, dann vergrößert diese Transformation den Schätzwert $\hat{\beta}_k$ um den Faktor 1000 (so dass das Produkt $\hat{\beta}_k x_{tk}$ unverändert bleibt).

Um bei einem Vergleich der Größenordnung der einzelnen Schätzwerte $\hat{\beta}_2, \dots, \hat{\beta}_K$ die Maßstabsabhängigkeit auszuschalten, werden gelegentlich auch die folgenden sogenannten **BETA-Koeffizienten** angegeben (z.B. in SPSS):

$$\text{BETA}_k = \hat{\beta}_k \cdot \frac{s_{x_k}}{s_y} = \hat{\beta}_k \cdot \frac{\sqrt{\sum (x_{tk} - \bar{x}_k)^2}}{\sqrt{\sum (y_t - \bar{y})^2}}, \quad k = 2, \dots, K. \quad (9.10)$$

Diese BETA-Koeffizienten ergeben sich als Kleinst-Quadrate-Schätzwerte, wenn man alle Variablen y, x_2, \dots, x_K standardisiert, also den jeweiligen Mittelwert der T Beobachtungen subtrahiert und durch die jeweilige Standardabweichung dividiert (das Absolutglied verschwindet mit der Standardisierung aller Variablen).

In dem Spezialfall, dass die erklärenden Variablen x_2, \dots, x_K paarweise unkorreliert sind (wovon in der Regel **nicht** ausgegangen werden kann), ergibt sich die folgende Zerlegung des Bestimmtheitsmaßes R^2 , die in diesem besonderen Fall eine weitergehende Interpretation der BETA-Koeffizienten als Maß für den Erklärungsbeitrag der einzelnen Variablen zulässt:

$$R^2 = \text{BETA}_2^2 + \text{BETA}_3^2 + \dots + \text{BETA}_K^2. \quad (9.11)$$

Diese Zerlegung von R^2 ergibt sich, **wenn die Regressoren paarweise unkorreliert sind**, folgendermaßen:

$$\begin{aligned} \sum (y_t - \bar{y})^2 &= \sum (\hat{y}_t - \bar{y})^2 + \sum e_t^2 \\ &= \sum \left(\hat{\beta}_2(x_{t2} - \bar{x}_2) + \dots + \hat{\beta}_K(x_{tK} - \bar{x}_K) \right)^2 + \sum e_t^2 \\ &= \hat{\beta}_2^2 \sum (x_{t2} - \bar{x}_2)^2 + \dots + \hat{\beta}_K^2 \sum (x_{tK} - \bar{x}_K)^2 + \sum e_t^2, \end{aligned}$$

wegen der angenommenen Unkorreliertheit der Regressoren, also

$$s_y^2 = \hat{\beta}_2^2 s_{x_2}^2 + \dots + \hat{\beta}_K^2 s_{x_K}^2 + s_e^2,$$

und somit

$$\begin{aligned} 1 &= \left(\frac{\hat{\beta}_2 s_{x_2}}{s_y} \right)^2 + \dots + \left(\frac{\hat{\beta}_K s_{x_K}}{s_y} \right)^2 + \frac{s_e^2}{s_y^2} \\ &= \text{BETA}_2^2 + \dots + \text{BETA}_K^2 + \frac{s_e^2}{s_y^2}, \end{aligned}$$

woraus unmittelbar (9.11) folgt. Neben dem Punkt, dass diese Zerlegung nur gilt, wenn die unabhängigen Variablen unkorreliert sind, ist der rein deskriptive Charakter dieser Analyse mit Hilfe der BETA-Koeffizienten hervorzuheben. Aus der Größe eines BETA-Werts kann nicht geschlossen werden, ob die entsprechende erklärende Variable im statistischen Sinn einen signifikanten Beitrag zur Erklärung der Variation von y leistet. Um derartige Aussagen zu treffen, muss man von der deskriptiven Analyse zur Verteilung von $\hat{\beta}$ unter den stochastischen Annahmen des linearen Regressionsmodells übergehen.

9.2 Die Wahrscheinlichkeitsverteilung von $\hat{\beta}$

Unser Interesse richtet sich jetzt wieder auf die Schätzfunktion $\hat{\beta}$ als ein Vektor von Zufallsvariablen, bei dem für $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_K$ in Abhängigkeit von den realisierten Stichprobenwerten y_1, \dots, y_T bestimmte Schätzwerte realisiert werden.

Zunächst halten wir fest, dass die Kleinst-Quadrate-Schätzfunktionen der Parameter β_i , $i = 1, \dots, K$ **erwartungstreu** sind:³

$$E(\hat{\beta}_i) = \beta_i, \quad i = 1, \dots, K.$$

Weiterhin ist $\hat{\beta}$ effizient, d.h. die Varianzen von $\hat{\beta}_i$, $i = 1, \dots, K$, oder auch von linearen Funktionen der $\hat{\beta}_i$, $i = 1, \dots, K$, sind so klein wie möglich in dem Sinn, dass keine andere unverzerrte Schätzfunktion, die eine lineare Funktion der Variablen y_1, \dots, y_T ist, eine kleinere Varianz besitzt; dies ist die Aussage des Satzes von **Gauß-Markov**.

Die **Varianzen** $Var(\hat{\beta}_i)$, $i = 1, \dots, K$, sind gegeben durch die entsprechenden K Elemente der Hauptdiagonalen der Matrix

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1}, \quad (9.12)$$

abgekürzt geschrieben auch mit $Var(\hat{\beta}_i) = \sigma^2(X'X)^{ii}$.

Das (i, j) -Element dieser Matrix gibt die Kovarianz von $\hat{\beta}_i$ und $\hat{\beta}_j$ an, abgekürzt mit $Cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2(X'X)^{ij}$.

³Der Beweis der folgenden Ergebnisse für $E(\hat{\beta}_i)$ und $Var(\hat{\beta}_i)$ basiert vor allem auf zwei Regeln für den Erwartungswertvektor $E(z)$ und die Varianz-Kovarianz-Matrix $V(z)$ einer Lineartransformation $z = Ay + b$ eines Zufallsvektors y . Es gilt (mit deterministischer Transformationsmatrix A und deterministischem Vektor b):

$$\begin{aligned} E(Ay + b) &= A E(y) + b \\ V(Ay + b) &= A V(y) A' \end{aligned}$$

Für die Lineartransformation $\hat{\beta} = (X'X)^{-1}X'y$ des Vektors y mit $E(y) = X\beta$ und $V(y) = \sigma^2I_T$ folgt dann mit $A = (X'X)^{-1}X'$, $A' = X(X'X)^{-1}$ und $b = 0$:

$$\begin{aligned} E(\hat{\beta}) &= E((X'X)^{-1}X'y) \\ &= (X'X)^{-1}X'E(y) = (X'X)^{-1}X'X\beta \\ &= \beta \\ V(\hat{\beta}) &= V((X'X)^{-1}X'y) \\ &= (X'X)^{-1}X'V(y)X(X'X)^{-1} = \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

Eine erwartungstreue Schätzfunktion für den unbekanntem Parameter σ^2 ist $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{1}{T - K} \sum e_t^2. \quad (9.13)$$

Wenn man σ^2 durch die Schätzfunktion $\hat{\sigma}^2$ ersetzt, erhält man die erwartungstreue geschätzten Varianzen und Kovarianzen als Elemente der Matrix

$$\hat{V}(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}. \quad (9.14)$$

Statistik-Programme geben häufig zu den Parameterschätzwerten die jeweilige geschätzte Standardabweichung und den sogenannten t -Wert des Koeffizienten an:

$$\hat{\beta}_i, \hat{\sigma}_{\hat{\beta}_i}, \frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}} \text{ (t-Wert)}, \quad (9.15)$$

wobei $\hat{\sigma}_{\hat{\beta}_i}$ für die geschätzte Standardabweichung von $\hat{\beta}_i$, also für $\sqrt{\widehat{Var}(\hat{\beta}_i)}$ steht.

9.3 Konfidenzintervalle und Tests für einzelne Koeffizienten und Restriktionen

Wir betrachten im folgenden Konfidenzintervalle und t -Tests für einen beliebigen einzelnen Koeffizienten β_i . Die Verfahren und Ergebnisse können aber leicht übertragen werden auf eine einzelne lineare Funktion von mehreren Koeffizienten, also z.B. auf den Test zu den Hypothesen

$$H_0 : \beta_2 - \beta_3 = 0 \quad (\text{d.h. Gleichheit der beiden Koeffizienten})$$

$$H_1 : \beta_2 - \beta_3 \neq 0.$$

oder zu

$$H_0 : \beta_2 + \beta_3 + \beta_4 = 1$$

$$H_1 : \beta_2 + \beta_3 + \beta_4 \neq 1.$$

Für solche Anwendungen benötigt man aber nicht nur die Varianzen der einzelnen Koeffizientenschätzer, sondern auch die Kovarianzen, um die Varianz der Linearkombination der Schätzer zu bestimmen; die oben angegebene Varianz-Kovarianz-Matrix enthält diese Information. Man hat also in den beiden Beispielen zu berechnen:

$$\widehat{Var}(\hat{\beta}_2 - \hat{\beta}_3) = \widehat{Var}(\hat{\beta}_2) + \widehat{Var}(\hat{\beta}_3) - 2\widehat{Cov}(\hat{\beta}_2, \hat{\beta}_3).$$

bzw.

$$\widehat{Var}\left(\sum_{i=2}^4 \hat{\beta}_i\right) = \sum_{i=2}^4 \widehat{Var}(\hat{\beta}_i) + 2\left(\widehat{Cov}(\hat{\beta}_2, \hat{\beta}_3) + \widehat{Cov}(\hat{\beta}_2, \hat{\beta}_4) + \widehat{Cov}(\hat{\beta}_3, \hat{\beta}_4)\right).$$

Der Einfachheit halber beschränken wir uns in der folgenden Darstellung auf einzelne Koeffizienten. Der grundlegende Ausgangspunkt für die Konfidenzintervalle und gleichermaßen für die Tests ist die **Normalverteilung** der $\hat{\beta}_i$ mit Erwartungswert β_i und dem i -ten Hauptdiagonalelement der Matrix $\sigma^2(X'X)^{-1}$ als Varianz. Die Normalverteilung der $\hat{\beta}_i$ resultiert aus der Normalverteilungsannahme (A4) für die abhängige Variable y_t , kann aber bei großem Stichprobenumfang T als Approximation auch ohne diese Annahme verwendet werden. Durch die Verwendung der Schätzfunktion $\hat{\sigma}^2$ anstelle der unbekanntenen Varianz σ^2 ergibt sich als Verteilung von

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\widehat{Var}(\hat{\beta}_i)}} = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}}$$

die t -Verteilung mit $T - K$ Freiheitsgraden.

9.3.1 Konfidenzintervalle für die Koeffizienten

Das $(1 - \alpha)100\%$ -Konfidenzintervall für β_i ist gegeben durch:

$$[\hat{\beta}_i - t_{1-\frac{\alpha}{2}; T-K} \cdot \hat{\sigma}_{\hat{\beta}_i}, \hat{\beta}_i + t_{1-\frac{\alpha}{2}; T-K} \cdot \hat{\sigma}_{\hat{\beta}_i}].$$

9.3.2 Parametertests

Ein wichtiger Sonderfall ist die Entscheidung, ob die Variable x_i einen signifikanten Einfluss auf die Veränderung der abhängigen Variablen y hat, also – zu gegebenem Signifikanzniveau α – der Test zu:

$$H_0 : \beta_i = 0 \quad \text{gegen} \quad H_1 : \beta_i \neq 0$$

Die Teststatistik ist hier

$$t = \frac{\hat{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}}$$

Unter H_0 ist die Teststatistik t -verteilt mit $T - K$ Freiheitsgraden, d.h. der kritische Wert ist

$$t_{1-\frac{\alpha}{2}; T-K}$$

und H_0 wird abgelehnt (d.h. ein signifikanter Einfluss der Variablen x_i bejaht), wenn

$$|t| > t_{1-\frac{\alpha}{2}; T-K}.$$

Vielfach geben Statistik-Programmpakete zu einer Regression neben den Schätzwerten $\hat{\beta}_i$, deren Standardabweichungen und den t -Werten zum Test auf $\beta_i = 0$ auch den p -Wert zu diesem Test an (auch unter der Bezeichnung: Signifikanzniveau, empirisches Signifikanzniveau, oder marginales Signifikanzniveau).

Ganz entsprechend kann allgemeiner auch der t -Test auf einen beliebigen hypothetischen Wert β_i^0 , also zu

$$H_0 : \beta_i = \beta_i^0 \text{ gegen } H_1 : \beta_i \neq \beta_i^0 \quad (9.16)$$

mit der Teststatistik

$$t = \frac{\hat{\beta}_i - \beta_i^0}{\hat{\sigma}_{\hat{\beta}_i}} \quad (9.17)$$

durchgeführt werden.

Während hier der Ablehnungsbereich unverändert mit

$$\left(-\infty, -t_{1-\frac{\alpha}{2}; T-K}\right) \cup \left(t_{1-\frac{\alpha}{2}; T-K}, \infty\right) \quad (9.18)$$

gegeben ist, ändert sich der Ablehnungsbereich für einen einseitigen Test. Bei einem rechtsseitigen Test, also zu

$$H_0 : \beta_i \leq \beta_i^0 \text{ gegen } H_1 : \beta_i > \beta_i^0 \quad (9.19)$$

ist der Ablehnungsbereich

$$\left(t_{1-\alpha; T-K}, \infty\right), \quad (9.20)$$

bei einem linksseitigen Test, also zu

$$H_0 : \beta_i \geq \beta_i^0 \text{ gegen } H_1 : \beta_i < \beta_i^0 \quad (9.21)$$

ist der Ablehnungsbereich

$$\left(-\infty, -t_{1-\alpha; T-K}\right). \quad (9.22)$$

9.3.3 Beispiel

Abhängige Variable y : KFZ-Versicherungsprämie pro Monat
 Unabhängige Variablen x_2 : Fahrpraxis (in Jahren)
 x_3 : Anzahl der verschuldeten Verkehrsunfälle
 während der letzten drei Jahre

Es wird angenommen, dass ein multiples lineares Regressionsmodell mit (A1) und (A4) gilt. Eine Stichprobe vom Umfang 12 ergibt:

$$\hat{y}_t = 55.138 - 1.3736 x_{t2} + 8.053 x_{t3}$$

$$(\pm 7.309) \quad (\pm 0.4885) \quad (\pm 1.307)$$

Die Standardabweichungen sind in Klammern angegeben.
 Die t -Werte sind:

$$t_1 = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{55.138}{7.309} = 7.54,$$

$$t_2 = \frac{\hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_2}} = -\frac{1.3736}{0.4885} = -2.81,$$

$$t_3 = \frac{\hat{\beta}_3}{\hat{\sigma}_{\hat{\beta}_3}} = \frac{8.053}{1.307} = 6.16.$$

Für $\alpha = 0.05$ ist $t_{1-\frac{\alpha}{2}, T-K} = t_{0.975, 9} = 2.262$, womit alle Koeffizienten signifikant von Null verschieden sind.

Die Frage, ob der Preis eines zusätzlichen Unfalls größer als 6 Geldeinheiten für die monatliche Versicherungsprämie ist (d.h. $\beta_3 > 6$), wird mit dem rechtsseitigen Test (9.19), (9.20) entschieden, also:

$$H_0 : \beta_3 \leq 6 \text{ gegen } H_1 : \beta_3 > 6, \alpha = 0.05$$

$$t = \frac{\hat{\beta}_3 - 6}{\hat{\sigma}_{\hat{\beta}_3}} = \frac{8.053 - 6}{1.307} = 1.57$$

$$\text{Kritischer Wert: } t_{0.95; 9} = 1.833$$

H_0 wird nicht abgelehnt, $\hat{\beta}_3$ ist nicht signifikant größer als 6.

9.4 F -Test auf Signifikanz des Erklärungsansatzes

Ähnlich wie bei der Varianzanalyse kann mit einem F -Test entschieden werden, ob die erklärenden Variablen x_2, \dots, x_K mit dem linearen Ansatz insgesamt einen signifikanten Beitrag zur Erklärung der Variation der y -Werte

liefern. Das Prüfmaß für diesen Test vergleicht die erklärte Streuung auf der Regressionsebene mit der unerklärten, zufälligen Streuung um die Regressionsebene. Große Werte von F führen zur Ablehnung der Nullhypothese, dass die Variablen keinen Einfluss hätten.

Formal wird im Modell

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_K x_K + u$$

mit den Annahmen (A1) und (A4) ein Test zu den Hypothesen

$$\begin{aligned} H_0 : \beta_i = 0, i = 2, \dots, K \\ \text{gegen} \\ H_1 : \beta_i \neq 0 \text{ für mindestens ein } i, i \in \{2, \dots, K\} \end{aligned} \quad (9.23)$$

zum Signifikanzniveau α durchgeführt.

Die Test-Statistik ist

$$\begin{aligned} F &= \frac{\sum(\hat{y}_t - \bar{y})^2}{\sum e_t^2} \cdot \frac{T - K}{K - 1} \\ &= \frac{R^2}{1 - R^2} \cdot \frac{T - K}{K - 1}. \end{aligned} \quad (9.24)$$

Wenn H_0 gültig ist, dann folgt diese Teststatistik F der F -Verteilung mit $k_1 = K - 1$ und $k_2 = T - K$ Freiheitsgraden.

Der kritische Wert ist:

$$F_{1-\alpha; K-1, T-K}.$$

H_0 ist abzulehnen, wenn

$$F > F_{1-\alpha; K-1, T-K}. \quad (9.25)$$

In Statistik-Programmpaketen wird zur linearen Regression meist auch der realisierte Wert der Teststatistik F zu diesem Test auf Signifikanz des gesamten Erklärungsansatzes angegeben; in der Regel zusammen mit dem p -Wert, d.h. mit der Wahrscheinlichkeit, dass eine F -verteilte Zufallsvariable mit $k_1 = K - 1$ und $k_2 = T - K$ Freiheitsgraden einen Wert annimmt, der größer als der realisierte Wert der Teststatistik ist. Häufig (bei einem brauchbaren Modell) ist dieser p -Wert nahe bei Null. Die Nullhypothese, dass $\beta_2 = \dots = \beta_K = 0$ und somit der Erklärungsansatz untauglich ist, kann abgelehnt werden, wenn der p -Wert kleiner als das vorgegebene Signifikanzniveau α ist.

9.5 Entscheidung zwischen zwei linearen Modellen

Während der Test im vorigen Abschnitt verstanden werden kann als eine Entscheidung zwischen dem Modell

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_K x_K + u$$

und dem reduzierten Modell

$$y = \beta_1 + u$$

betrachten wir jetzt etwas allgemeiner die Entscheidung zwischen den Modellen

$$H_0 : y = \beta_1 + \beta_2 x_2 + \dots + \beta_{K_0} x_{K_0} + u$$

$$H_1 : y = \beta_1 + \beta_2 x_2 + \dots + \beta_{K_0} x_{K_0} + \dots + \beta_K x_K + u$$

Anders formuliert betrachten wir auf der Grundlage des allgemeineren Modells (H_1) die Frage, ob die $m := K - K_0$ Variablen x_{K_0+1}, \dots, x_K insgesamt einen signifikanten Beitrag zur Erklärung der Veränderungen der abhängigen Variablen y leisten.

Wir formulieren als Hypothesenpaar

$$H_0 : \beta_i = 0, \quad i = K_0 + 1, \dots, K$$

gegen

$$H_1 : \beta_i \neq 0, \quad \text{für mindestens ein } i \in \{K_0 + 1, \dots, K\}$$

(9.26)

Die Teststatistik vergleicht $\sum e_{t,0}^2$ mit $\sum e_t^2$ mit

$$F = \frac{(\sum e_{t,0}^2 - \sum e_t^2) / m}{\sum e_t^2 / (T - K)} \quad (9.27)$$

bzw.

$$F = \frac{\sum e_{t,0}^2 - \sum e_t^2}{\sum e_t^2} \cdot \frac{T - K}{m} \quad \stackrel{H_0}{\sim} F_{m, T-K}, \quad (9.28)$$

mit $m := K - K_0$; $e_{t,0}$ bezeichnet die Kleinst-Quadrate-Residuen unter H_0 und e_t die Kleinst-Quadrate-Residuen im unrestringierten Modell.

Mit diesem Prüfmaß wird die Anpassung unter der Restriktion (gemessen mit $\sum e_{t,0}^2$) verglichen mit der Anpassung, die bei Einbeziehung der zusätzlichen Variablen erreicht wird (gemessen mit $\sum e_t^2$). Durch die Einbeziehung der

zusätzlichen Variablen kann die Anpassung natürlich nicht verschlechtert werden, d.h. es ist stets $\sum e_t^2 \leq \sum e_{t,0}^2$. Wenn die Nullhypothese gilt, wird aber die Summe der quadrierten Residuen durch die zusätzlichen Variablen nicht wesentlich kleiner als im H_0 -Modell sein, d.h. das Prüfmaß F wird dann mit großer Wahrscheinlichkeit einen kleinen Wert annehmen.⁴

Unter der Nullhypothese ist das Prüfmaß F -verteilt mit $k_1 = m$ und $k_2 = T - K$ Freiheitsgraden.⁵ Zum Signifikanzniveau α ist der kritische Wert

$$F_{1-\alpha; m, T-K}.$$

Die Nullhypothese (d.h. der Ausschluss der m Variablen x_{K_0+1}, \dots, x_K) wird also verworfen, wenn

$$F > F_{1-\alpha; m, T-K}. \quad (9.29)$$

Als Sonderfall dieses Tests erhält man den zuvor betrachteten F -Test auf Signifikanz des gesamten Erklärungsansatzes.

9.6 Vergleich der Modellauswahl mit F -Test und mit korrigiertem Bestimmtheitsmaß

Wir betrachten wie im letzten Abschnitt

$$\begin{aligned} H_0 : y &= \beta_1 + \beta_2 x_2 + \dots + \beta_{K_0} x_{K_0} + u \\ H_1 : y &= \beta_1 + \beta_2 x_2 + \dots + \beta_{K_0} x_{K_0} + \dots + \beta_K x_K + u. \end{aligned}$$

Auswahl mit dem F -Test:

Wähle Modell (H_0), d.h. schließe die Variablen x_{K_0+1}, \dots, x_K aus, wenn die Nullhypothese

$$H_0 : \beta_{K_0+1} = \dots = \beta_K = 0,$$

nicht abgelehnt wird, d.h. wenn

$$F \leq F_{1-\alpha; m, T-K}.$$

⁴Mit der Streuungszersetzung in beiden Modellen (H_0) und (H_1) sieht man leicht, dass die Differenz $\sum e_{t,0}^2 - \sum e_t^2$ auch geschrieben werden kann als $\sum (\hat{y}_t - \bar{y})^2 - \sum (\hat{y}_{t,0} - \bar{y})^2$, d.h. als der zusätzliche Beitrag zur erklärten Streuung, den die m Variablen x_{K_0+1}, \dots, x_K leisten.

⁵Man kann nämlich zeigen, dass das Prüfmaß unter der Nullhypothese der Quotient aus zwei unabhängig χ^2 -verteilten Zufallsvariablen ist, die jeweils durch die Anzahl der Freiheitsgrade m bzw. $T - K$ dividiert werden.

Auswahl mit dem $\overline{R^2}$ -Kriterium:

Wähle Modell (H_0), wenn das korrigierte Bestimmtheitsmaß für dieses Modell ($\overline{R_0^2}$) größer ist als für das vollständige Modell ($\overline{R^2}$).

Diese Bedingung ist genau dann erfüllt, wenn das Prüfmaß F kleiner als Eins ist, denn

$$\begin{aligned} \overline{R_0^2} > \overline{R^2} &\Leftrightarrow \frac{\sum e_{t,0}^2}{T - K_0} < \frac{\sum e_t^2}{T - K} \\ &\Leftrightarrow \frac{\sum e_{t,0}^2}{\sum e_t^2} < 1 + \frac{K - K_0}{T - K} \\ &\Leftrightarrow \frac{\sum e_{t,0}^2 - \sum e_t^2}{\sum e_t^2} < \frac{m}{T - K} \\ &\Leftrightarrow F < 1 \end{aligned}$$

Da für gebräuchliches Signifikanzniveau α der kritische Wert $F_{1-\alpha; m, T-K}$ wesentlich größer als 1 ist, wird mit dem $\overline{R^2}$ -Kriterium das lange Modell (H_1) wesentlich stärker favorisiert als beim F -Test. Beim F -Test wird also mehr empirische Evidenz für die Relevanz der zusätzlichen Variablen verlangt, um für die Einbeziehung dieser Variablen zu entscheiden.

9.7 Verallgemeinerung: F-Test auf m lineare Restriktionen

In den beiden vorangegangenen Abschnitten haben wir m Nullrestriktionen

$$\beta_{K_0+1} = 0, \dots, \beta_K = 0$$

betrachtet. Der F -Test ist auch zur Entscheidung über die Gültigkeit anderer linearer Restriktionen in einem linearen Regressionsmodell

$$y = \beta_1 + \beta_2 x_2 + \dots + \beta_K x_K + u$$

anwendbar⁶, wie z.B. auf die beiden Restriktionen:

$$\begin{aligned} \beta_1 &= \beta_2 \text{ bzw. } \beta_1 - \beta_2 = 0 \text{ und} \\ \beta_3 &+ 2\beta_4 = 1. \end{aligned}$$

⁶Durch eine neue Parametrisierung kann man jede lineare Restriktion auch als eine Nullrestriktion für einen Parameter schreiben.

Für m unabhängige lineare Restriktionen, die geschrieben werden können als

$$\begin{aligned} c_{11}\beta_1 + c_{12}\beta_2 + \dots + c_{1K}\beta_K &= d_1 \\ c_{21}\beta_1 + c_{22}\beta_2 + \dots + c_{2K}\beta_K &= d_2 \\ &\vdots \\ c_{m1}\beta_1 + c_{m2}\beta_2 + \dots + c_{mK}\beta_K &= d_m \end{aligned}$$

wird der F -Test zur Nullhypothese, dass diese Restriktionen gelten, gegen die Hypothese, dass mindestens einer dieser Restriktionen nicht gilt, in der gleichen Weise wie oben durchgeführt mit der Teststatistik:

$$F = \frac{\sum e_{t,0}^2 - \sum e_t^2}{\sum e_t^2} \cdot \frac{T-K}{m}, \quad (9.30)$$

wobei $e_{t,0}$ die Residuen bei einer Kleinst-Quadrate-Schätzung unter Berücksichtigung der m Restriktionen bezeichnet, während e_t wie üblich die Residuen der Kleinst-Quadrate-Schätzung (ohne Auferlegung von Restriktionen) bezeichnet. Wenn H_0 gilt, dann folgt F der F -Verteilung mit $k_1 = m$ und $k_2 = T - K$ Freiheitsgraden. H_0 wird also abgelehnt, wenn $F > F_{1-\alpha; m, T-K}$.

9.8 Einfache Varianzanalyse und F -Test im multiplen linearen Regressionsmodell

Die Fragestellung der einfachen Varianzanalyse kann auch im Rahmen des multiplen Regressionsmodells formuliert werden. Wenn der Effekt eines Faktors mit K Stufen auf die Lage der Verteilung einer Variablen y untersucht werden soll, betrachten wir das folgende Regressionsmodell:

$$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_K x_{tK} + u_t$$

mit den **Dummyvariablen**

$$x_{t1} = \begin{cases} 1, & \text{wenn Faktorstufe 1 vorliegt bei Beobachtung } t \\ 0, & \text{sonst} \end{cases}$$

$$x_{t2} = \begin{cases} 1, & \text{wenn Faktorstufe 2 vorliegt bei Beobachtung } t \\ 0, & \text{sonst} \end{cases}$$

usw. bis

$$x_{tK} = \begin{cases} 1, & \text{wenn Faktorstufe } K \text{ vorliegt bei Beobachtung } t \\ 0, & \text{sonst} \end{cases}$$

Mit der üblichen Annahme (A4) für die Störgrößen u_t , d.h. die Störgrößen sind unabhängig identisch $N(0; \sigma^2)$ -verteilt, impliziert dieses Modell, dass die abhängige Variable in dem Fall, dass der Faktor die Stufe k annimmt, normalverteilt ist mit Erwartungswert β_k und Varianz σ^2 . Insgesamt liegt zu den verschiedenen Faktorstufen eine Stichprobe vom Umfang T von unabhängigen Beobachtungen vor. Als Schätzwerte für die Parameter $\beta_k, k = 1, \dots, K$ nach der Methode der Kleinsten Quadrate erhält man leicht mit der Formel

$$\hat{\beta} = (X'X)^{-1}X'y$$

die jeweiligen arithmetischen Mittel der y -Werte zur Faktorstufe k , was wir mit

$$\hat{\beta}_k = \bar{y}_{(k)} ; \quad k = 1, \dots, K \quad (9.31)$$

notieren.

Die untersuchte Nullhypothese in der einfachen Varianzanalyse ist die Hypothese, dass der Erwartungswert von y unabhängig von der jeweiligen Faktorstufe stets gleich ist, also

$$H_0 : \beta_i = \beta_1, i = 2, \dots, K \quad \text{gegen} \quad H_1 : \beta_i \neq \beta_1 \text{ für mind. ein } i.$$

Im linearen Regressionsmodell testen wir diese $m = K - 1$ linearen Restriktionen wie im letzten Abschnitt besprochen mit einem F -Test als Entscheidung zwischen den Modellen

$$\begin{aligned} H_0 : y_t &= \beta_1(x_{t1} + \dots + x_{tK}) + u_t = \beta_1 + u_t \\ H_1 : y_t &= \beta_1 x_{t1} + \dots + \beta_K x_{tK} + u_t \end{aligned}$$

Unter der Nullhypothese ist der Kleinst-Quadrate-Schätzwert für β_1 gleich dem arithmetischen Mittel \bar{y} über alle T Beobachtungen. Damit ergibt sich für die Teststatistik

$$F = \frac{\sum e_{t,0}^2 - \sum e_t^2}{\sum e_t^2} \cdot \frac{T - K}{K - 1},$$

im einzelnen:

$$\begin{aligned} \sum e_t^2 &= \sum (y_t - \hat{y}_t)^2 = SW, \\ \sum e_{t,0}^2 &= \sum (y_t - \bar{y})^2 = SW + SB \end{aligned}$$

und somit

$$\begin{aligned} F &= \frac{\sum e_{t,0}^2 - \sum e_t^2}{\sum e_t^2} \cdot \frac{T - K}{K - 1} \\ &= \frac{SB}{SW} \cdot \frac{T - K}{K - 1}. \end{aligned}$$

Die einfache Varianzanalyse kann also auch als ein Spezialfall im Rahmen des multiplen linearen Regressionsmodells mit geeignet definierten Dummyvariablen behandelt werden.

9.9 Eine Bemerkung über Multikollinearität und die Varianzen von $\hat{\beta}_1, \dots, \hat{\beta}_K$

Im multiplen Regressionsmodell

$$y_t = \beta_1 + \beta_2 x_{t2} + \dots + \beta_K x_{tK} + u_t$$

mit den Annahmen (A1) und (A4) wird die Präzision der Parameterschätzung durch hohe Korrelation zwischen den unabhängigen Variablen beeinträchtigt. Bei Anwendungen des linearen Regressionsmodells auf wirtschaftsstatistische, nicht-experimentelle Daten ist man häufig damit konfrontiert, dass zwischen den unabhängigen Variablen starke korrelative Beziehungen bestehen; man spricht dann von **Multikollinearität**. Solange keine exakte lineare Abhängigkeit zwischen den Regressoren x_1, x_2, \dots, x_K besteht, bedeutet Multikollinearität keine Verletzung der Annahmen des Regressionsmodells. (Den Fall exakter linearer Abhängigkeit, der die Rangbedingung für die Regressormatrix X verletzt, wollen wir hier nicht betrachten.)

Im folgenden wird der Effekt von Multikollinearität auf die Varianzen der Schätzfunktionen $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_K$ quantifiziert. Um Multikollinearität zu messen, betrachten wir für jede echte unabhängige Variable x_i das Bestimmtheitsmaß R_i^2 der (rein deskriptiv interpretierten) Regression von x_i auf die übrigen $K - 1$ unabhängigen Variablen $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_K$. Für jede unabhängige Variable x_i ($i = 2, \dots, K$) ergibt diese Regression eine geschätzte Regressionsgleichung

$$\hat{x}_{ti} = \hat{\alpha}_1^i x_{t1} + \dots + \hat{\alpha}_{i-1}^i x_{t,i-1} + \hat{\alpha}_{i+1}^i x_{t,i+1} + \dots + \hat{\alpha}_K^i x_{tK}. \quad (9.32)$$

Uns interessieren im Folgenden nicht die $K - 1$ Kleinst-Quadrate-Schätzwerte $\hat{\alpha}_1^i, \dots, \hat{\alpha}_{i-1}^i, \hat{\alpha}_{i+1}^i, \dots, \hat{\alpha}_K^i$, sondern nur das jeweilige Bestimmtheitsmaß für die Regressionsgleichung, das wir mit R_i^2 bezeichnen. Wenn R_i^2 nahe Null

ist, dann liegt keine nennenswerte Korrelation zwischen x_i und den anderen unabhängigen Variablen vor. Wenn R_i^2 dagegen nahe bei Eins liegt, besteht eine hohe "Kollinearität" zwischen x_i und den anderen unabhängigen Variablen. Eine Konsequenz daraus ist, dass der Erklärungsbeitrag $\beta_i x_{ti}$ der Variablen x_i im Modell für y nahezu vollständig auch von den anderen Variablen übernommen werden könnte. Wie die folgende Darstellung der Varianz von $\hat{\beta}_i$ (die hier nicht bewiesen wird) zeigt, führt dementsprechend auch ein hohes Bestimmtheitsmaß R_i^2 zu einer großen Varianz von $\hat{\beta}_i$. Es gilt:

$$\text{Var}(\hat{\beta}_i) = \frac{\sigma^2}{T} \frac{1}{s_{x_i}^2} \frac{1}{1 - R_i^2} \quad (9.33)$$

wobei $s_{x_i}^2 = \frac{1}{T} \sum (x_{it} - \bar{x}_i)^2$.

Diese Darstellung der Varianz von $\hat{\beta}_i$ zeigt leicht den Einfluss

- der Varianz σ^2 der Störgröße
- der Anzahl T der Beobachtungen
- der Streuung $s_{x_i}^2$ der erklärenden Variablen x_i
- der mit R_i^2 gemessenen Kollinearität zwischen x_i und den anderen erklärenden Variablen.

9.10 Prognosen

Neben der Untersuchung von Beziehungen zwischen ökonomischen Variablen dient die Regressionsanalyse auch der Prognoseschätzung für die abhängige Variable. Wir setzen voraus, dass (A1) und (A4) für das lineare Regressionsmodell

$$y_t = \beta_1 + \beta_2 x_{t2} + \dots + \beta_K x_{tK} + u_t$$

für $t = 1, \dots, T$ und für den Prognosezeitpunkt $t = T+1$ gelten. Insbesondere gehen wir also von der Annahme aus, die erklärenden Variablen seien auch für den Prognosezeitpunkt gegeben. Aus diesem Grund wird die Prognose für y_{T+1} auch als (durch die gegebenen erklärenden Variablen) **bedingte Prognose** bezeichnet.

Wir untersuchen als naheliegende Prognosefunktion für

$$\begin{aligned} y_{T+1} &= \beta_1 + \beta_2 x_{T+1,2} + \dots + \beta_K x_{T+1,K} + u_{T+1} \\ &= x'_{(T+1)} \beta + u_{T+1} \end{aligned}$$

die Zufallsvariable

$$\hat{y}_{T+1} = x'_{(T+1)} \hat{\beta} \quad . \quad (9.34)$$

Als Prognosefehler v definieren wir die Differenz

$$v = \hat{y}_{T+1} - y_{T+1} = x'_{(T+1)}(\hat{\beta} - \beta) - u_{T+1} \quad . \quad (9.35)$$

Der Prognosefehler ist also die Differenz von zwei Zufallsvariablen.

Die wichtigsten Eigenschaften der Prognosefunktion $\hat{y}_{T+1} = x'_{(T+1)} \hat{\beta}$ sind:

- \hat{y}_{T+1} ist eine in $y = (y_1, \dots, y_T)'$ **lineare** Prognosefunktion, denn $\hat{y}_{T+1} = x'_{(T+1)}(X'X)^{-1}X'y$.
- \hat{y}_{T+1} ist eine **unverzerrte** Prognosefunktion für y_{T+1} , d.h. der Erwartungswert des Prognosefehlers ist Null,

$$\begin{aligned} E(v) &= E(x'_{(T+1)}(\hat{\beta} - \beta) - u_{T+1}) \\ &= x'_{(T+1)}(E(\hat{\beta}) - \beta) - E(u_{T+1}) \\ &= 0 \quad . \end{aligned}$$

- Die Varianz des Prognosefehlers $Var(v)$ ist, da $\hat{\beta}$ als Funktion von u_1, \dots, u_T stochastisch unabhängig von u_{T+1} ist und es sich bei β um einen deterministischen Vektor handelt, gegeben durch

$$\begin{aligned} Var(v) &= Var(x'_{(T+1)}(\hat{\beta} - \beta) - u_{T+1}) \\ &= x'_{(T+1)}V(\hat{\beta})x_{(T+1)} + Var(u_{T+1}) \\ &= \sigma^2(x'_{(T+1)}(X'X)^{-1}x_{(T+1)} + 1) \quad . \quad (9.36) \end{aligned}$$

Wenn das Regressionsmodell ein Absolutglied enthält, dann kann man $Var(v)$ auch äquivalent umformen in

$$Var(v) = \sigma^2(x^*_{(T+1)})'(X^*X^*)^{-1}x^*_{(T+1)} + \frac{1}{T} + 1$$

wobei mit * jeweils bezeichnet wird, dass die Variablen um den Mittelwert zentriert werden, also

$$x^*_{(T+1)}' := \left[x_{T+1,2} - \bar{x}_2 \quad \dots \quad x_{T+1,K} - \bar{x}_K \right]$$

und

$$X^* := \begin{bmatrix} x_{12} - \bar{x}_2 & \dots & x_{1K} - \bar{x}_K \\ \vdots & & \vdots \\ x_{T2} - \bar{x}_2 & \dots & x_{TK} - \bar{x}_K \end{bmatrix} .$$

Diese Darstellung liefert für den Fall $K = 2$ das gleiche Ergebnis zur Varianz des Prognosefehlers wie im Kapitel zur einfachen Regression.

Im wesentlichen auf der Grundlage des Gauß-Markov-Theorems kann gezeigt werden, dass \hat{y}_{T+1} die beste lineare unverzerrte Prognosefunktion für y_{T+1} ist, d.h. dass der Prognosefehler v die kleinste Varianz im Vergleich mit der Prognosefehlervarianz jeder anderen linearen unverzerrten Prognosefunktion hat.

Abschließend bestimmen wir unter Verwendung der Normalverteilungsannahme (A4) ein Prognose-Intervall für y_{T+1} .

Der Prognosefehler v ist eine lineare Funktion der normalverteilten Zufallsvariablen y_1, \dots, y_T, y_{T+1} , und somit ist v normalverteilt, also

$$\hat{y}_{T+1} - y_{T+1} \sim N(0, \sigma_v^2)$$

mit der Bezeichnung $\sigma_v^2 := \text{Var}(v)$. Folglich gilt:

$$\frac{(\hat{y}_{T+1} - y_{T+1})}{\sigma_v} \sim N(0, 1).$$

Die Ersetzung von σ_v durch $\hat{\sigma}_v$, wobei

$$\hat{\sigma}_v^2 = \hat{\sigma}^2(x'_{(T+1)}(X'X)^{-1}x_{(T+1)} + 1), \quad (9.37)$$

führt zu der t -verteilten Zufallsvariablen:

$$\frac{(\hat{y}_{T+1} - y_{T+1})}{\hat{\sigma}_v} \sim t_{(T-K)} .$$

Damit ist zu vorgegebener Wahrscheinlichkeit $1 - \alpha$:

$$P\left(\left|\frac{\hat{y}_{T+1} - y_{T+1}}{\hat{\sigma}_v}\right| \leq t_{T-K, 1-\frac{\alpha}{2}}\right) = 1 - \alpha .$$

Als Prognose-Intervall zum Niveau $1 - \alpha$ erhält man daraus

$$[\hat{y}_{T+1} - \hat{\sigma}_v \cdot t_{T-K, 1-\frac{\alpha}{2}}, \hat{y}_{T+1} + \hat{\sigma}_v \cdot t_{T-K, 1-\frac{\alpha}{2}}] .$$

Ein derart konstruiertes Prognose-Intervall überdeckt mit Wahrscheinlichkeit $1 - \alpha$ den Wert y_{T+1} der abhängigen Variablen.

Anhang A

Einführung in die Matrizenrechnung

A.1 Grundlegende Definitionen und einige einfache Operationen

A.1.1 Matrix, Vektor, Skalar

Eine **Matrix** ist ein rechteckiges Feld von reellen Zahlen, die in Zeilen und Spalten angeordnet sind. Eine Matrix mit m Zeilen und n Spalten hat die **Dimension** $(m \times n)$.

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \quad \text{auch kurz } A = [a_{ij}]$$

a_{ij} : Element in Zeile i , Spalte j

Falls erforderlich, wird in der Bezeichnung der Matrix auch die Dimension angegeben, für eine $(m \times n)$ -Matrix: $A_{m \times n}$. Zwei Matrizen A, B sind gleich ($A = B$), wenn sie die gleiche Dimension haben und elementweise gleich sind. $(m \times 1)$ -Matrizen bzw. die einzelnen Spalten von $A_{m \times n}$ heißen **(Spalten-) Vektoren**, $(1 \times n)$ - Matrizen bzw. die einzelnen Zeilen von $A_{m \times n}$ heißen **(Zeilen) - Vektoren**. (1×1) - Matrizen, also einzelne reelle Zahlen, heißen auch **Skalare**.

Eine Matrix $A_{m \times n}$ heißt **quadratisch**, wenn $m = n$ gilt; dann wird n auch als **Ordnung der Matrix** bezeichnet. Die Elemente $a_{11}, a_{22}, \dots, a_{nn}$ von $A_{n \times n}$ heißen **Diagonalelemente** (bzw. Diagonale). Eine quadratische $(n \times n)$ -Matrix A heißt

- **Diagonalmatrix**, wenn $a_{ij} = 0$ für alle $i \neq j$;

$$\text{Schreibweise: } A = \text{diag}(a_1, \dots, a_n) = \begin{bmatrix} a_1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_n \end{bmatrix}$$

- **Einheitsmatrix**, wenn $A = \text{diag}(1, 1, \dots, 1) =: I_n$
- **(untere) Dreiecksmatrix**, wenn $a_{ij} = 0$ für alle $i < j$
- **(obere) Dreiecksmatrix**, wenn $a_{ij} = 0$ für alle $i > j$
- **symmetrisch**, wenn $a_{ij} = a_{ji}$ für alle i, j .

Für eine quadratische Matrix A heißt die skalare Funktion

$$\text{tr}(A) = \sum_i a_{ii}$$

die **Spur (Trace) von A**.

A.1.2 Transponieren einer Matrix

Zur $(m \times n)$ -Matrix $A = [a_{ij}]$ heißt die $(n \times m)$ -Matrix $A' = [a'_{ij}]$ mit $a'_{ij} = a_{ji}$ die **Transponierte von A**.

Beispiel:

$$A = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}, \quad A' = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

Es gilt: $(A')' = A$.

Für quadratische Matrizen gilt:

- $\text{tr}(A) = \text{tr}(A')$
- $A = A'$ genau dann, wenn A symmetrisch ist.

A.1.3 Teilmatrizen und Partitionierung einer Matrix

Durch waagrechte und senkrechte Trennungslinien kann eine Matrix in verschiedene Teilmatrizen **partitioniert** werden, etwa

$$A_{m \times n} = \left[\begin{array}{cc} A_{11} & A_{12} \\ m_1 \times n_1 & m_1 \times n_2 \\ A_{21} & A_{22} \\ m_2 \times n_1 & m_2 \times n_2 \end{array} \right], \quad \begin{array}{l} m = m_1 + m_2 \\ n = n_1 + n_2 \end{array}$$

$$\text{z.B. } A = \left[\begin{array}{c|ccc} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ \hline 9 & 0 & 1 & 2 \end{array} \right] = \left[\begin{array}{cc} A_{11} & A_{12} \\ A_{21} & A_{22} \end{array} \right]$$

$$\text{mit } A_{11} = \begin{bmatrix} 1 \\ 5 \end{bmatrix}, \quad A_{12} = \begin{bmatrix} 2 & 3 & 4 \\ 6 & 7 & 8 \end{bmatrix} \\ A_{21} = \begin{bmatrix} 9 \end{bmatrix}, \quad A_{22} = \begin{bmatrix} 0 & 1 & 2 \end{bmatrix}$$

Transponieren einer partitionierten Matrix

$$\left[\begin{array}{ccc} A & B & C \\ D & E & F \end{array} \right]' = \left[\begin{array}{cc} A' & D' \\ B' & E' \\ C' & F' \end{array} \right]$$

A.1.4 Addition von Matrizen gleicher Dimension

Wenn A und B gleiche Dimensionen haben, dann ist die Addition von A und B definiert durch:

$$A + B = [a_{ij}] + [b_{ij}] := [a_{ij} + b_{ij}]$$

Es gilt:

$$\begin{aligned} A + B &= B + A \\ (A + B) + C &= A + (B + C) \\ (A + B)' &= A' + B' \end{aligned}$$

und – falls A und B quadratisch sind –

$$\operatorname{tr}(A + B) = \operatorname{tr}(A) + \operatorname{tr}(B).$$

Für die **Nullmatrix** $\mathbf{0}$, d.h. die Matrix, deren Elemente alle gleich Null sind, gilt

$$A + \mathbf{0} = A$$

A.1.5 Skalare Multiplikation

Wenn α ein Skalar ist und A eine Matrix, dann ist die **skalare Multiplikation** einer Matrix definiert durch:

$$\alpha A = \alpha[a_{ij}] := [\alpha a_{ij}]$$

Es gelten die Regeln:

$$\begin{aligned}\alpha A &= A\alpha \\ (\alpha A)' &= \alpha A' \\ (\alpha + \beta)A &= \alpha A + \beta A \\ (\alpha\beta)A &= \alpha(\beta A) = \beta(\alpha A)\end{aligned}$$

Insbesondere gilt mit $\alpha = -1$: $(-1)A = [-a_{ij}]$ Damit wird die **Subtraktion** für gleich dimensionierte Matrizen definiert durch:

$$A - B = A + (-1)B = [a_{ij} - b_{ij}]$$

A.2 Matrizenmultiplikation

A.2.1 Definition der Multiplikation

Für eine $(m \times n)$ -Matrix A und eine $(r \times s)$ -Matrix B mit $n = r$ ist das **Produkt** AB definiert als die $(m \times s)$ -Matrix C mit

$$c_{ij} = \sum_{k=1}^n a_{ik}b_{kj}; \quad i = 1, \dots, m; \quad j = 1, \dots, s$$

Für $n \neq r$ ist das Produkt AB nicht definiert.

Beispiele:

(i)

$$\text{Preisvektor } p = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix}, \text{ Mengenvektor } q = \begin{bmatrix} q_1 \\ \vdots \\ q_n \end{bmatrix}$$

$$p'q = (p_1 \cdots p_n) \begin{bmatrix} q_1 \\ \vdots \\ q_n \end{bmatrix} = \sum_{i=1}^n p_i q_i$$

(ii)

$$\text{Preismatrix } P = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & & \vdots \\ p_{T1} & \cdots & p_{Tn} \end{bmatrix}, \text{ Mengenvektor } q = \begin{bmatrix} q_1 \\ \vdots \\ q_n \end{bmatrix}$$

$$Pq = \begin{bmatrix} \sum p_{1i} q_i \\ \vdots \\ \sum p_{Ti} q_i \end{bmatrix}.$$

(iii) Lineares Gleichungssystem:

$$\begin{array}{ccccccc} a_{11}x_1 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ \vdots & & & & \vdots & & \vdots \\ a_{m1}x_1 & + & \cdots & + & a_{mn}x_n & = & b_m \end{array}$$

In Matrixschreibweise:

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

(iv)

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 0 \\ 1 & 2 & 0 \\ 4 & 5 & 1 \end{bmatrix} = \begin{bmatrix} 8 & 8 & 1 \\ 10 & 7 & 1 \end{bmatrix}$$

A.2.2 Multiplikation mit speziellen Matrizen

Sei A eine $(m \times n)$ -Matrix, $A = [a_{ij}]$.

(i) Multiplikation mit einer Nullmatrix.

$$0_{r \times m} A_{m \times n} = 0_{r \times n}$$

$$A_{m \times n} 0_{n \times s} = 0_{m \times s}$$

Aber z.B. auch:

$$\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} 2 & -4 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad (\text{also } AB = 0 \\ \not\Rightarrow A = 0 \vee B = 0)$$

(ii) Multiplikation mit der (passenden) Einheitsmatrix:

$$IA = A$$

$$AI = A$$

Aber z.B. auch:

$$\begin{bmatrix} 2 & 2 \\ 3 & 7 \end{bmatrix} \begin{bmatrix} 2 & -4 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & -4 \\ -1 & 2 \end{bmatrix} \quad (\text{also } BA = A \not\Rightarrow B = I)$$

(iii) Multiplikation mit einer Diagonalmatrix:

von links: $\text{diag}(d_1, \dots, d_m) A = [d_i a_{ij}]$, (Zeile i mit d_i multipliziert)

von rechts: $A \text{diag}(d_1, \dots, d_n) = [a_{ij} d_j]$, (Spalte j mit d_j multipliziert)

A.2.3 Regeln für die Matrizenmultiplikation

- (i) Assoziativgesetz: $(AB)C$ ist genau dann definiert, wenn auch $A(BC)$ definiert ist und dann gilt:

$$(AB)C = A(BC)$$

- (ii) Distributivgesetz:

$$\begin{aligned}(A + B)C &= AC + BC \\ C(A + B) &= CA + CB,\end{aligned}$$

wieder vorausgesetzt, dass die Ausdrücke erklärt sind.

- (iii) Im allgemeinen gilt nicht die Kommutativität
1. Fall: AB ist definiert und BA ist nicht definiert, wenn $A_{m \times n}$, $B_{n \times s}$ und $s \neq m$ gilt.
 2. Fall: AB ist definiert und BA ist definiert, aber unterschiedlicher Dimension, wenn $A_{m \times n}$, $B_{n \times m}$ und $n \neq m$ gilt.
Dann ist AB eine $(m \times m)$ -Matrix und BA eine $(n \times n)$ -Matrix. Insbesondere ist für eine $(m \times n)$ -Matrix A das Produkt AA' eine $(m \times m)$ -Matrix, $A'A$ jedoch eine $(n \times n)$ -Matrix.

Für $(n \times 1)$ -Vektoren, a, b ($a' = (a_1, \dots, a_n)$, $b' = (b_1, \dots, b_n)$) gilt:

$$a'b = \sum a_i b_i \quad \text{und} \quad ba' = \begin{bmatrix} b_1 a_1 & \dots & b_1 a_n \\ \vdots & & \vdots \\ b_n a_1 & \dots & b_n a_n \end{bmatrix}$$

3. Fall: AB ist definiert und BA ist definiert und beide Produkte sind von gleicher Dimension, wenn A, B, AB, BA quadratische $(n \times n)$ -Matrizen sind. Dann ist im allgemeinen AB verschieden von BA .

Beispiel:

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 5 & 3 \\ 3 & 7 \end{bmatrix} = \begin{bmatrix} 10 & 6 \\ 3 & 7 \end{bmatrix}$$

und

$$\begin{bmatrix} 5 & 3 \\ 3 & 7 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 10 & 3 \\ 6 & 7 \end{bmatrix}$$

Es gibt auch Beispiele, wo $AB = BA$ gilt, z.B. für zwei Diagonalmatrizen A, B und allgemeiner für zwei symmetrische Matrizen A, B , wenn deren Produkt AB auch symmetrisch ist (vgl. 4.2.4).

- (iv) Matrizenmultiplikation und skalare Multiplikation:
Für einen Skalar α und Matrizen A, B gilt

$$\alpha(AB) = (\alpha A)B = A(\alpha B).$$

Diese Regel gilt natürlich auch, wenn der Skalar als Matrizenprodukt $b'c$ entsteht, also z.B. für vier $(n \times 1)$ -Vektoren a, b, c, d :

$$(ab')(cd') = a(b'c)d' = (b'c)ad'.$$

A.2.4 Transponieren eines Produktes

Wenn AB erklärt ist, gilt

$$(AB)' = B'A',$$

denn für das i, j -Element berechnet man auf beiden Seiten

$$\sum_k a_{jk}b_{ki} = \sum_k b_{ki}a_{jk}$$

Das Ergebnis kann unmittelbar verallgemeinert werden für mehr als zwei Faktoren, etwa

$$(ABC)' = (A(BC))' = (BC)'A' = C'B'A'$$

A.2.5 Spur eines Produkts

Wenn AB erklärt und eine quadratische Matrix ist, also $A_{m \times n}, B_{n \times m}$, dann ist $tr(AB)$ und $tr(BA)$ definiert und

$$tr(AB) = \sum_{i=1}^m \sum_{j=1}^n a_{ij}b_{ji} = \sum_{j=1}^n \sum_{i=1}^m b_{ji}a_{ij} = tr(BA)$$

Daraus folgt unmittelbar: $tr(ABC) = tr(BCA) = tr(CAB)$, wenn $tr(ABC)$ erklärt ist. Auch bei mehr als drei Faktoren bleibt die Spur des Produkts bei zyklischer Vertauschung der Faktoren unverändert.

Für zwei $(m \times n)$ -Matrizen A, B gilt

$$tr(A'B) = \sum_{i=1}^m \sum_{j=1}^n a_{ij}b_{ij},$$

d.h. $tr(A'B)$ ergibt sich durch Summation der Produkte $a_{ij}b_{ij}$ über alle Felder (i, j) . Insbesondere

$$tr(A'A) = \sum_i \sum_j a_{ij}^2$$

A.2.6 Multiplikation partitionierter Matrizen

Wenn AB erklärt ist und A, B so partitioniert werden, dass die Zusammenfassung von Spalten von A mit der Zusammenfassung von Zeilen von B korrespondiert, gilt die Produktdefinition auch, wenn sie auf die Teilmatrizen anstelle der Elemente angewandt wird. **Beispiel:**

$$[A_1 : A_2] \cdot \begin{bmatrix} B_1 \\ \dots \\ B_2 \end{bmatrix} = A_1 B_1 + A_2 B_2 \quad ,$$

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}$$

A.3 Multiplikation spezieller Matrizen

A.3.1 Summation und Zentrierung durch Matrizenmultiplikation

Sei ι der $(n \times 1)$ -Vektor, dessen Elemente alle gleich 1 sind, und x ein beliebiger $(n \times 1)$ -Vektor. Dann gilt:

$$\iota'x = (1, 1, \dots, 1) \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix} = \sum x_i,$$

und

$$\iota'x/n = \sum x_i/n = \bar{x}$$

Der Vektor ι heißt **Einsvektor** oder **Summationsvektor der Ordnung n** (und wird, falls erforderlich, mit Angabe der Ordnung als ι_n bezeichnet).

Weiterhin heißt die $(n \times n)$ -Matrix $J = \iota \iota'$, deren Elemente alle gleich 1 sind, **Einsmatrix** oder **Summationsmatrix**. Es gilt dann:

$$\iota'x = Jx = \begin{bmatrix} \sum x_i \\ \vdots \\ \sum x_i \end{bmatrix}, \quad \frac{1}{n}Jx = \begin{bmatrix} \bar{x} \\ \vdots \\ \bar{x} \end{bmatrix} \quad \text{und} \quad \left(I - \frac{1}{n}J\right)x = \begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix} =: x^*$$

Die Multiplikation mit der Matrix $(I - \frac{1}{n}J)$ transformiert x also in den zentrierten Vektor x^* , dessen Element x_i^* die Abweichung von x_i vom Mittelwert \bar{x} angibt. Dementsprechend wird die Matrix

$$M = I - \frac{1}{n}J$$

auch als **Zentrierungsmatrix** bezeichnet. Die Summe der x_i^* ist gleich Null, matriziell:

$$\iota'x^* = \iota'(I - \frac{1}{n}J)x = (\iota' - \frac{\iota'\iota'}{n})x = (\iota' - \iota')x = 0,$$

denn $\iota'\iota = n$.

Für jeden Vektor x muß wegen $\overline{x^*} = 0$ gelten, dass die Zentrierung von x^* , also Mx^* , wieder x^* liefert, d.h.

$$MMx = Mx.$$

Da diese Gleichung für jeden Vektor x , insbesondere also auch für die Spalten der Einheitsmatrix I gilt, folgt $MM = M$. Dies rechnet man auch direkt nach, denn wegen

$$\frac{\iota'\iota'}{n^2} = \frac{\iota n \iota'}{n^2} = \frac{\iota'}{n}$$

gilt

$$MM = (I - \frac{\iota'}{n})(I - \frac{\iota'}{n}) = I - \frac{\iota'}{n} - \frac{\iota'}{n} + \frac{\iota'}{n} = I - \frac{\iota'}{n} = M.$$

A.3.2 Matrizenpotenzen und idempotente Matrizen

Für eine Matrix A ist das Produkt $AA = A^2$ genau dann definiert, wenn A quadratisch ist. Dann ist A^k für jede natürliche Zahl k definiert; für $k = 0$ wird $A^0 := I$ definiert. Wie das Beispiel der Matrix M im vorigen Abschnitt zeigt, gibt es außer der Nullmatrix und der Einheitsmatrix weitere quadratische Matrizen mit der Eigenschaft, dass

$$A^2 = AA = A$$

und somit $A^k = A$ für alle natürlichen Zahlen k . Eine quadratische Matrix A mit dieser Eigenschaft $AA = A$ heißt **idempotent**. Wenn A idempotent ist, dann ist auch die Matrix $I - A$ idempotent, denn dann gilt

$$(I - A)(I - A) = I - A - A + A = I - A.$$

A.3.3 Quadratische Formen

Für einen $(n \times 1)$ -Vektor x heißt eine gewichtete Summe über alle Produkte $x_i x_j$, $i, j = 1, \dots, n$, eine **quadratische Form in x** :

$$\sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j.$$

Beispiel für $n = 2$:

$$\begin{aligned} a_{11}x_1^2 + a_{12}x_1x_2 + a_{21}x_2x_1 + a_{22}x_2^2 &= 5x_1^2 + 2x_1x_2 + 4x_2x_1 + x_2^2 \\ &= 5x_1^2 + 3x_1x_2 + 3x_2x_1 + x_2^2. \end{aligned}$$

Wie man an dem Beispiel erkennt, kann man wegen $x_i x_j = x_j x_i$ jede quadratische Form auch mit Gewichten a_{ij} darstellen, für die gilt:

$$a_{ij} = a_{ji}, \quad \text{d.h.} \quad A = [a_{ij}] \quad \text{ist symmetrisch.}$$

Im folgenden wird die Koeffizientenmatrix A von vornherein als **symmetrisch** angenommen. Unter Verwendung der symmetrischen $(n \times n)$ -Matrix $A = [a_{ij}]$ kann man die quadratische Form auch als Matrizenprodukt schreiben:

$$x'Ax = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j.$$

Für $A = I$ ergibt sich insbesondere:

$$x'Ix = x'x = \sum_{i=1}^n x_i^2.$$

Man bezeichnet $\sqrt{x'x}$ als **Norm des Vektors x** . Offensichtlich gilt, dass $x'x$ und somit die Norm von x nur dann gleich Null ist, wenn x der Nullvektor ist.

Für eine symmetrische **idempotente** Matrix A gilt

$$x'Ax = x'A'Ax = (Ax)'Ax,$$

z.B. mit der in A.3.1. eingeführten Zentrierungsmatrix M

$$x'Mx = (Mx)'Mx = x^*x^* = \sum_i (x_i - \bar{x})^2 \quad .$$

Wenn x in allen Elementen konstant ist ($x = \alpha \iota$) – und nicht nur wenn x der Nullvektor ist – gilt also $x'Mx = 0$

A.3.4 Positiv definite Matrizen

Eine quadratische Form $x'Ax$ heißt **positiv definite quadratische Form**, wenn

$$x'Ax > 0 \text{ für alle } x \neq 0.$$

Dann heißt auch die symmetrische Matrix A **positiv definit**.

Beispiele:

(i) : $A = I$ ist positiv definit;

(ii) : $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ ist positiv definit, da

$$x'Ax = 2x_1^2 + 2x_1x_2 + 2x_2^2 = x_1^2 + x_2^2 + (x_1 + x_2)^2 > 0 \text{ für alle } x \neq 0;$$

(iii): M ist nicht positiv definit, da $\iota'M\iota = 0$.

Eine quadratische Form $x'Ax$ heißt **positiv semidefinite quadratische Form**, wenn

$$x'Ax \geq 0 \text{ für alle } x$$

$$\text{und } x'Ax = 0 \text{ für ein } x \neq 0.$$

Dann heißt auch die symmetrische Matrix A **positiv semidefinit**.

Beispiel: M ist positiv semidefinit.

Wenn A positiv definit oder positiv semidefinit ist, heißt es auch zusammenfassend, dass A **nichtnegativ definit** ist.

Jede idempotente symmetrische Matrix A ist nichtnegativ definit, denn mit $y := Ax$ ist

$$x'Ax = x'AAx = x'A'Ax = (Ax)'Ax = y'y = \sum y_i^2 \geq 0.$$

A.4 Die Determinante einer quadratischen Matrix

Bei der Bestimmung der Inversen zu einer quadratischen Matrix A spielt die Determinante von A , kurz bezeichnet mit $\det(A)$ oder $|A|$, eine wichtige Rolle.

Die Determinante von A ist eine reelle Zahl (ein Skalar), die als Funktion aller Matrixelemente berechnet wird. Wir definieren zunächst für $n = 1, 2, 3$:

$$\begin{aligned}
 A = a_{11} & \longrightarrow \det(A) = a_{11} \\
 A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} & \longrightarrow \det(A) = a_{11}a_{22} - a_{12}a_{21} \\
 A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} & \longrightarrow \det(A) = a_{11} \det(A_{11}) - a_{12} \det(A_{12}) \\
 & \qquad \qquad \qquad + a_{13} \det(A_{13})
 \end{aligned}$$

Dabei ist A_{ij} definiert als die quadratische Teilmatrix von A , die entsteht, wenn aus A die Zeile i und Spalte j gestrichen werden, also

$$\det(A_{11}) = a_{22}a_{33} - a_{32}a_{23}$$

$$\det(A_{12}) = a_{21}a_{33} - a_{31}a_{23}$$

$$\det(A_{13}) = a_{21}a_{32} - a_{31}a_{22}$$

Allgemein heißt $\det(A_{ij})$, wobei A_{ij} die Teilmatrix von A ist, die entsteht, wenn Zeile i und Spalte j entfernt werden, der **Minor zu a_{ij}** . Als **Kofaktor c_{ij}** zu a_{ij} bezeichnet man dann

$$c_{ij} = (-1)^{i+j} \det(A_{ij})$$

Die Definition der Determinante einer (3×3) -Matrix unter Rückgriff auf die Kofaktoren, also auf Determinanten von (2×2) -Matrizen wird übertragen zur rekursiven Definition der **Determinante einer $(n \times n)$ -Matrix A** :

$$\det(A) = \sum_{j=1}^n a_{1j}c_{1j} = \sum_{j=1}^n (-1)^{1+j} a_{1j} \det(A_{1j}).$$

Diese Definition wird auch als „**Entwicklung der Determinante nach der ersten Zeile**“ bezeichnet. Allgemein ergibt sich der gleiche Wert, wenn nach einer beliebigen Zeile oder Spalte entwickelt wird, d.h. es gilt auch

$$\begin{aligned}
 \det(A) &= \sum_{j=1}^n a_{ij}c_{ij} \quad (\text{für beliebige Zeile } i) \\
 &= \sum_{i=1}^n a_{ij}c_{ij} \quad (\text{für beliebige Spalte } j)
 \end{aligned}$$

Wichtige Eigenschaften für $(n \times n)$ -Matrizen A, B :

$$\det(A) = \det(A')$$

Wenn B aus A entsteht durch Vertauschen von zwei Zeilen (bzw. Spalten), gilt:

$$\det(B) = -\det(A) \quad .$$

Wenn in A zwei Zeilen (bzw. Spalten) gleich oder proportional sind, dann gilt:

$$\det(A) = 0 \quad .$$

Wenn die Matrix A mit einem Skalar δ multipliziert wird, dann gilt:

$$\det(\delta A) = \delta^n \det(A).$$

Multiplikationssatz für Determinanten:

$$\det(AB) = \det(A)\det(B).$$

Wenn A eine Dreiecksmatrix ist, dann gilt:

$$\det(A) = a_{11} \cdot a_{22} \cdot \dots \cdot a_{nn}.$$

Wenn die Matrix eine Block-Dreiecksmatrix ist, dann gilt:

$$\det \begin{bmatrix} A & C \\ 0 & B \end{bmatrix} = \det \begin{bmatrix} A & 0 \\ D & B \end{bmatrix} = \det(A)\det(B).$$

A.5 Die Inverse einer quadratischen Matrix

So wie zu einer reellen Zahl $a \neq 0$ eindeutig die reelle Zahl $a^{-1} = 1/a$ existiert, so dass $aa^{-1} = 1$, soll zu einer quadratischen Matrix A eine Matrix $B = A^{-1}$ bestimmt werden, so dass $AB = BA = I$. Notwendige Bedingung für die Existenz einer solchen **inversen Matrix** ist nicht nur, dass $A \neq 0$ gilt, sondern dass auch $\det(A) \neq 0$ ist. Dies folgt sofort aus dem Multiplikationssatz für Determinanten; denn für $AB = I$ ist

$$\det(A)\det(B) = \det(AB) = \det(I) = 1,$$

also muß $\det(A) \neq 0$ gelten. Diese Bedingung ist auch hinreichend. Für eine $(n \times n)$ -Matrix A mit $\det(A) \neq 0$ ist die Inverse A^{-1} gegeben mit

$$A^{-1} = (\det(A))^{-1}C'$$

wobei $C = [c_{ij}]$ die Matrix der Kofaktoren ist. Die Inverse ist eindeutig bestimmt, und es gilt:

$$AA^{-1} = A^{-1}A = I.$$

Wenn für eine quadratische Matrix A die Determinante gleich Null ist und die Inverse von A somit nicht existiert, heißt A **singulär**; wenn A^{-1} existiert, heißt A **nichtsingulär**. Da die Matrizenmultiplikation im allgemeinen nicht kommutativ ist, ist es wichtig, die Umkehrung der Reihenfolge der Faktoren bei der Invertierung eines Produkts zu beachten. Für nichtsinguläre Matrizen A, B gilt:

$$\begin{aligned} (A')^{-1} &= (A^{-1})' && \text{(somit ist } A^{-1} \text{ symmetrisch,} \\ &&& \text{wenn } A \text{ symmetrisch ist)} \\ (\alpha A)^{-1} &= (1/\alpha)A^{-1} \\ (AB)^{-1} &= B^{-1}A^{-1} && \text{(denn } ABB^{-1}A^{-1} = I) \end{aligned}$$

Für eine nichtsinguläre partitionierte Matrix A

$$A = \begin{bmatrix} E & F \\ G & H \end{bmatrix}$$

mit quadratischen nichtsingulären Teilmatrizen E und H kann die entsprechend partitionierte Inverse A^{-1} berechnet werden als

$$A^{-1} = \begin{bmatrix} E^{-1}(I + FD^{-1}GE^{-1}) & -E^{-1}FD^{-1} \\ -D^{-1}GE^{-1} & D^{-1} \end{bmatrix}$$

mit $D = H - GE^{-1}F$.

Beispiele:

(i)

$$\begin{aligned} A &= \begin{bmatrix} 2 & 3 \\ 1 & 5 \end{bmatrix} ; \quad \det(A) = 7 ; \quad C = \begin{bmatrix} 5 & -1 \\ -3 & 2 \end{bmatrix} \\ A^{-1} &= \begin{bmatrix} 5/7 & -3/7 \\ -1/7 & 2/7 \end{bmatrix} \end{aligned}$$

(ii)

$$A = \begin{bmatrix} 2 & 3 & 1 \\ 1 & 0 & 2 \\ -1 & 5 & 1 \end{bmatrix} ; \quad \det A = -24$$

$$C = \begin{bmatrix} \begin{vmatrix} 0 & 2 \\ 5 & 1 \end{vmatrix} & - \begin{vmatrix} 1 & 2 \\ -1 & 1 \end{vmatrix} & \begin{vmatrix} 1 & 0 \\ -1 & 5 \end{vmatrix} \\ - \begin{vmatrix} 3 & 1 \\ 5 & 1 \end{vmatrix} & \begin{vmatrix} 2 & 1 \\ -1 & 1 \end{vmatrix} & - \begin{vmatrix} 2 & 3 \\ -1 & 5 \end{vmatrix} \\ \begin{vmatrix} 3 & 1 \\ 0 & 2 \end{vmatrix} & - \begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} & \begin{vmatrix} 2 & 3 \\ 1 & 0 \end{vmatrix} \end{bmatrix} = \begin{bmatrix} -10 & -3 & 5 \\ 2 & 3 & -13 \\ 6 & -3 & -3 \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} 10/24 & 2/24 & -6/24 \\ 3/24 & -3/24 & 3/24 \\ -5/24 & 13/24 & 3/24 \end{bmatrix}$$

A.6 Lineare Unabhängigkeit und Rang einer Matrix

A.6.1 Lineare Unabhängigkeit von Vektoren

Es seien x_1, \dots, x_k Vektoren der gleichen Ordnung ($n \times 1$) und $\alpha_1, \dots, \alpha_k$ reelle Zahlen. Dann heißt der Vektor y ,

$$y = \alpha_1 x_1 + \dots + \alpha_k x_k$$

eine **Linearkombination der Vektoren** x_1, \dots, x_n .

Eine Menge von k ($n \times 1$)-Vektoren x_1, \dots, x_k heißt **linear unabhängig**,

wenn keiner der Vektoren als Linearkombination der restlichen Vektoren darzustellen ist. Äquivalent dazu ist das Kriterium :

$$\sum_{i=1}^k \alpha_i x_i = 0 \quad \implies \quad \alpha_i = 0 \quad \text{für alle } i = 1, \dots, k$$

Offensichtlich sind z.B. die n Spaltenvektoren e_1, \dots, e_n der $(n \times n)$ -Einheitsmatrix linear unabhängig. Wenn die Menge dieser n Vektoren um irgendeinen weiteren $(n \times 1)$ -Vektor $a = (a_1, a_2, \dots, a_n)'$ ergänzt wird, sind die entsprechenden $n + 1$ Vektoren nicht mehr linear unabhängig, denn es ist

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = a_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + a_2 \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \dots + a_n \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} = \sum_{i=1}^n a_i e_i$$

Allgemein ist eine Menge von k Vektoren, deren Anzahl k größer als die Dimension n ist, stets **linear abhängig**, d.h. nicht linear unabhängig. Jede Menge von n linear unabhängigen $(n \times 1)$ -Vektoren, also z.B. die Vektoren e_1, \dots, e_n , heißt **Basis** für den Raum der $(n \times 1)$ -Vektoren. Wenn b_1, \dots, b_n eine Basis im Raum der $(n \times 1)$ Vektoren bilden, dann läßt sich jeder $(n \times 1)$ -Vektor in eindeutiger Weise als Linearkombination der Basisvektoren darstellen. Die Konzepte gelten entsprechend für $(1 \times n)$ -Zeilenvektoren.

A.6.2 Rang einer Matrix

Für eine $(m \times n)$ -Matrix A heißt die Maximalzahl linear unabhängiger Spaltenvektoren der **Spaltenrang von A** , die Maximalzahl linear unabhängiger Zeilenvektoren heißt der **Zeilenrang von A** . Es kann gezeigt werden, dass der Zeilenrang und der Spaltenrang einer Matrix gleich sind; diese Zahl wird als der **Rang der Matrix A** definiert und mit $r(A)$ bezeichnet. Für den Rang gelten die folgenden Aussagen:

- (i) $r(A) = r(A')$
- (ii) $r(A) \leq \min\{m, n\}$
- (iii) $r(0) = 0$
- (iv) $r(I_n) = n$
- (v) $r(AB) \leq \min\{r(A); r(B)\}$

- (vi) $r(A'A) = r(A)$ Für quadratische $(n \times n)$ -Matrizen gilt:
- (vii) $r(A) < n \Leftrightarrow \det(A) = 0$
- (viii) Für idempotente Matrizen kann der Rang besonders einfach bestimmt werden. Es gilt für idempotente Matrizen: $r(A) = \text{tr}(A)$

Beispiel:

Sei X eine beliebige $(T \times K)$ -Matrix mit T Beobachtungen für K Variablen mit $r(X) = K$. Damit ist die Matrix $M = I_T - X(X'X)^{-1}X'$ idempotent (nachprüfen: $M^2 = M$) und es gilt:

$$\begin{aligned} r(M) &= \text{tr}(M) = \text{tr}(I_T - X(X'X)^{-1}X') = \text{tr}(I_T) - \text{tr}(X(X'X)^{-1}X') \\ &= T - \text{tr}(X'X(X'X)^{-1}) = T - \text{tr}(I_K) \\ &= T - K \end{aligned}$$

A.7 Orthogonale Vektoren und Matrizen

Zwei Vektoren a und b , $a \neq b$, heißen **orthogonal**, wenn $a'b = 0$.

Eine $(n \times n)$ -Matrix P nennen wir orthogonal, wenn je zwei verschiedene Spalten zueinander orthogonal sind und jede Spalte die Norm 1 hat, d.h. wenn $P'P = I$.

Beispiel:

$$P = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} \end{bmatrix} \quad \text{ist orthogonal.}$$

Weiterführende Literatur zur Matrizenrechnung:

Graybill, F.A.: Matrices with Applications in Statistics, 2nd edition, Belmont, 1983

Searle, S.S.: Matrix Algebra useful for Statistics, New York, 1982