

Nachname:

Vorname:

Matrikelnummer:

--	--	--

### Formelsammlung zur Vorlesung

# Statistik I für Studierende der Soziologie, des Nebenfachs Statistik, der Medieninformatik und der Cultural Cognitive Linguistics

Prof. Dr. Thomas Augustin und Georg Schollmeyer  
(Brandt, Jansen, Plaß)

Wintersemester 2015/16

Es gelten die Hinweise zu den Klausurmodalitäten und erlaubten Hilfsmitteln auf der Veranstaltungshomepage! Insbesondere gilt:

Für die Verwendung in der Klausur im Wintersemester 2015/16 darf diese Version auf den bedruckten Seiten (und nur dort!) mit zusätzlichen handschriftlichen Kommentaren versehen werden, falls

1. die Formelsammlung in Originalgröße ausgedruckt wurde und
2. die Eintragungen in der eigenen Handschrift im Original vorgenommen wurden (keine Kopie!).

## Inhaltsverzeichnis

<b>1</b>	<b>Einführung und erste Grundbegriffe</b>	<b>1</b>
<b>2</b>	<b>Häufigkeitsverteilungen</b>	<b>2</b>
<b>3</b>	<b>Lage- und Streuungsmaße</b>	<b>4</b>
<b>4</b>	<b>Konzentrationsmessung</b>	<b>9</b>
<b>5</b>	<b>Assoziationsmessung in Kontingenztafeln</b>	<b>13</b>
<b>6</b>	<b>Korrelationsanalyse: Zusammenhangsanalyse stetiger Merkmale</b>	<b>20</b>

# 1 Einführung und erste Grundbegriffe

## 1.1 Vorbemerkungen zur Organisation, Bedeutung und Struktur der Veranstaltung

## 1.2 Was soll Statistik (nicht)?

## 1.3 Literatur

## 1.4 Grundbegriffe

**Notation** Merkmale werden typischerweise mit Großbuchstaben bezeichnet ( $X, Y, Z$ , etc.), Ausprägungen mit dem zugehörigen Kleinbuchstaben ( $x, y, z$ ). Der Wertebereich wird mit  $W_x, W_y, W_z$  bzw.  $W$  bezeichnet.

Formal ist jedes Merkmal eine Funktion.

$$\begin{aligned} X : \Omega &\rightarrow W \\ \omega &\mapsto X(\omega) \end{aligned}$$

### Merkmaltypen

- Stetige, quasi-stetige und diskrete Merkmale
- Skalenniveaus
- Qualitative und quantitative Merkmale

## 2 Häufigkeitsverteilungen

**Ausgangssituation** An  $n$  Einheiten  $\omega_1, \dots, \omega_n$  sei das Merkmal  $X$  beobachtet worden. Die *verschiedenen* potentiell möglichen Merkmalsausprägungen werden mit  $a_1, \dots, a_k$  bezeichnet.

### 2.1 Häufigkeiten

**Absolute Häufigkeiten der Merkmalsausprägungen** Für jedes  $a_j$ ,  $j = 1, \dots, k$ , bezeichnen  $h_j$  und  $h(a_j)$  die *absolute Häufigkeit* der Ausprägung  $a_j$ , d.h. die Anzahl der  $x_i$  aus  $x_1, \dots, x_n$  mit  $x_i = a_j$ .

Formal:

$$h_j := h(a_j) := |\{\omega \in \Omega \mid X(\omega) = a_j\}|.$$

Es gilt:

$$\sum_{j=1}^k h_j = n.$$

**Relative Häufigkeiten der Merkmalsausprägungen** Für jedes  $a_j$ ,  $j = 1, \dots, k$ , bezeichnen  $f_j$  und  $f(a_j)$  die *relative Häufigkeit* der Ausprägung  $a_j$ , also

$$f_j := f(a_j) := \frac{h_j}{n}.$$

$f_1, f_2, \dots, f_k$  nennt man die *relative Häufigkeitsverteilung*.

Es gilt:

$$\sum_{j=1}^k f_j = 1.$$

### Häufigkeitstabelle

$j$	$a_j$	$h_j$	$f_j$
1	$a_1$	$h_1$	$f_1$
2	$a_2$	$h_2$	$f_2$
3	$a_3$	$h_3$	$f_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$k$	$a_k$	$h_k$	$f_k$
$\Sigma$		$n$	1

## 2.2 Grafische Darstellung

## 2.3 Histogramm

## 2.4 Kumulierte Häufigkeiten und empirische Verteilungsfunktion

**Definition** Gegeben sei die Urliste  $x_1, \dots, x_n$  eines (mindestens) ordinalskalierten Merkmals mit der Häufigkeitsverteilung  $h_1, \dots, h_k$  bzw.  $f_1, \dots, f_k$ . Dann heißt

$$H(x) = \sum_{j:a_j \leq x} h(a_j) = \sum_{j:a_j \leq x} h_j$$

*absolute kumulierte Häufigkeitsverteilung* und

$$F(x) = \sum_{j:a_j \leq x} f(a_j) = \frac{1}{n} \sum_{j:a_j \leq x} h(a_j) = \frac{H(x)}{n}$$

*relative kumulierte Häufigkeitsverteilung* bzw. *empirische Verteilungsfunktion*.

### Gruppierte Daten

- $k$  Klassen  $[c_0, c_1), \dots, [c_{j-1}, c_j), \dots, [c_{k-1}, c_k]$ ,  $h_j$  Häufigkeit in  $j$ -ter Klasse,  $j = 1, \dots, k$
- Verwende bei einem  $x$  aus der Klasse  $[c_{j-1}, c_j)$  als Approximation für  $H(x)$  folgenden, aus der linearen Interpolation gewonnenen, Punkt:

$$H(x) \approx H(c_{j-1}) + \frac{h_j}{(c_j - c_{j-1})} \cdot (x - c_{j-1})$$

## 3 Lage- und Streuungsmaße

### 3.1 Arithmetisches Mittel und Varianz

**Definition (Arithmetisches Mittel)** Sei  $x_1, \dots, x_n$  die Urliste eines (mindestens) intervallskalierten Merkmals  $X$ . Dann heißt

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$$

das *arithmetische Mittel* der Beobachtungen  $x_1, \dots, x_n$ .

**Alternative Berechnung basierend auf Häufigkeiten** Hat das Merkmal  $X$  die Ausprägungen  $a_1, \dots, a_k$  und die (relative) Häufigkeitsverteilung  $h_1, \dots, h_k$  bzw.  $f_1, \dots, f_k$ , so gilt:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^k a_j h_j = \sum_{j=1}^k a_j f_j.$$

**Definition (Varianz)** Sei  $x_1, \dots, x_n$  die Urliste eines intervallskalierten Merkmals  $X$ . Dann heißen

$$\tilde{s}_X^2 := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

die (*empirische*) *Varianz* oder *Stichprobenvarianz* und

$$\tilde{s}_X := \sqrt{\tilde{s}_X^2}$$

die *empirische Streuung*, *Stichprobenstreuung* oder *Standardabweichung von  $X$* .

**Alternative Berechnung basierend auf Häufigkeiten** Sind die Ausprägungen  $a_1, \dots, a_k$  mit (relativer) Häufigkeitsverteilung  $h_1, \dots, h_k$  bzw.  $f_1, \dots, f_k$  gegeben, so gilt

$$\tilde{s}_X^2 = \frac{1}{n} \sum_{j=1}^k h_j (a_j - \bar{x})^2 = \sum_{j=1}^k f_j (a_j - \bar{x})^2.$$

**Verschiebungssatz** Es gilt

$$\begin{aligned} \tilde{s}_X^2 &= \frac{1}{n} \sum_{i=1}^n x_i^2 - \left( \frac{1}{n} \sum_{i=1}^n x_i \right)^2 = \overline{x^2} - (\bar{x})^2, \\ &= \left( \frac{1}{n} \sum_{j=1}^k (a_j^2) \cdot h_j \right) - \left( \frac{1}{n} \sum_{j=1}^k a_j \cdot h_j \right)^2 \\ &= \sum_{j=1}^k (a_j^2) \cdot f_j - \left( \sum_{j=1}^k a_j \cdot f_j \right)^2 \end{aligned}$$

**Korrigierte empirische Varianz** Sei  $x_1, \dots, x_n$  die Urliste eines intervallskalierten Merkmals  $X$ . Dann heißt

$$s_X^2 := \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

die *korrigierte empirische Varianz* oder *korrigierte Stichprobenvarianz* von  $X$ .

**Satz (Arithmetisches Mittel und lineare Transformationen)** Gegeben sei die Urliste  $x_1, \dots, x_n$  eines (mindestens) intervallskalierten Merkmals  $X$ . Betrachtet wird das (linear transformierte) Merkmal  $Y = a \cdot X + b$  und die zugehörigen Ausprägungen  $y_1, \dots, y_n$ . Dann gilt für das arithmetische Mittel  $\bar{y}$  von  $Y$ :

$$\bar{y} = a \cdot \bar{x} + b.$$

**Satz (Varianz und lineare Transformationen)** Sei  $x_1, \dots, x_n$  die Urliste eines mindestens intervallskalierten Merkmals  $X$  mit  $\tilde{s}_X > 0$  und  $y_1, \dots, y_n$  die zugehörige Urliste des Merkmals  $Y = a \cdot X + b$ . Dann gilt

$$\tilde{s}_Y^2 = a^2 \cdot \tilde{s}_X^2$$

und

$$\tilde{s}_Y = |a| \cdot \tilde{s}_X.$$

**Definition (Arithmetisches Mittel bei gruppierten Daten)** Sei  $X$  ein intervallskaliertes Merkmal, das in gruppierter Form mit  $k$  Klassen  $[c_0, c_1), [c_1, c_2), \dots, [c_{k-1}, c_k]$  erhoben wurde. Mit  $h'_\ell$ ,  $\ell = 1, \dots, k$ , als absoluter Häufigkeit der  $\ell$ -ten Klasse,  $f'_\ell$  als zugehöriger relativer Häufigkeit und  $m_\ell := \frac{c_\ell + c_{\ell-1}}{2}$  als der jeweiligen Klassenmitte definiert man als *arithmetisches Mittel für gruppierte Daten*

$$\bar{x}_{\text{grupp}} := \frac{1}{n} \sum_{\ell=1}^k h'_\ell m_\ell = \sum_{\ell=1}^k f'_\ell m_\ell.$$

**Satz (Arithmetisches Mittel bei geschichteten Daten)** Zerfällt die Grundgesamtheit in  $z$  Schichten, so kann  $\bar{x}$  aus den Schichtmitteln  $\bar{x}^{(\ell)}$ ,  $\ell = 1, \dots, z$ , berechnet werden:

$$\bar{x} = \frac{1}{n} \sum_{\ell=1}^z n^{(\ell)} \bar{x}^{(\ell)}.$$

Dabei bezeichnet  $n^{(\ell)}$  die Anzahl der Elemente in der  $\ell$ -ten Schicht.

**Satz (Varianz bei geschichteten Daten) – Varianzzerlegung / Streuungszerlegung**

- Schicht  $1, \dots, \ell, \dots, z$
- Besetzungszahlen  $n^{(1)}, \dots, n^{(\ell)}, \dots, n^{(z)}$ ;  $\sum_{\ell=1}^z n^{(\ell)} = n$
- Mittelwerte  $\bar{x}^{(1)}, \dots, \bar{x}^{(\ell)}, \dots, \bar{x}^{(z)}$
- Varianzen  $\tilde{s}^{2(1)}, \dots, \tilde{s}^{2(\ell)}, \dots, \tilde{s}^{2(z)}$

Mit 
$$\tilde{s}_{\text{innerhalb}}^2 := \frac{1}{n} \sum_{\ell=1}^z n^{(\ell)} \tilde{s}^{2(\ell)}$$

sowie 
$$\tilde{s}_{\text{zwischen}}^2 := \frac{1}{n} \sum_{\ell=1}^z n^{(\ell)} (\bar{x}^{(\ell)} - \bar{x})^2$$

gilt 
$$\tilde{s}^2 = \tilde{s}_{\text{innerhalb}}^2 + \tilde{s}_{\text{zwischen}}^2.$$

**3.2 Median & Quantile**

**Definition (Median)** Gegeben sei die Urliste  $x_1, \dots, x_n$  eines (mindestens) ordinalskalierten Merkmals  $X$ . Jede Zahl  $x_{\text{med}}$  mit

$$\frac{|\{i | x_i \leq x_{\text{med}}\}|}{n} \geq 0.5 \quad \text{und} \quad \frac{|\{i | x_i \geq x_{\text{med}}\}|}{n} \geq 0.5$$

heißt *Median*.

**Definition (Quantile)** Gegeben sei die Urliste  $x_1, \dots, x_n$  eines (mindestens) ordinalskalierten Merkmals  $X$  und eine Zahl  $0 < \alpha < 1$ . Jede Zahl  $x_\alpha$  mit

$$\frac{|\{i | x_i \leq x_\alpha\}|}{n} \geq \alpha \quad \text{und} \quad \frac{|\{i | x_i \geq x_\alpha\}|}{n} \geq 1 - \alpha$$

heißt  $\alpha \cdot 100\%$ -Quantil.

**Spezielle Quantile**

- Median:  $x_{0.5} = x_{\text{med}}$ .
- Quartile:  $x_{0.25}, x_{0.75}$ .
- Dezile:  $x_{0.1}, x_{0.2}, \dots, x_{0.8}, x_{0.9}$ .

**Alternative Definition** des Medians über die *geordnete* Urliste

$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ :

$$x_{\text{med}} := \begin{cases} \frac{1}{2} \left( x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} \right) & \text{für } n \text{ gerade} \\ x_{(\frac{n+1}{2})} & \text{für } n \text{ ungerade} \end{cases}$$

**Satz (Verhalten unter Transformation)** Sei  $x_1, x_2, \dots, x_n$  die Urliste eines (mindestens) ordinalskalierten Merkmals  $X$  und  $g$  eine monotone Funktion.

- i) Ist  $x_{\text{med}}$  ein Median von  $X$ , so gilt mit  $y_1 = g(x_1), \dots, y_n = g(x_n)$  als Urliste des Merkmals  $Y = g(X)$ :

$$y_{\text{med}} = g(x_{\text{med}})$$

ist ein Median von  $Y$ .

- ii) Fordert man zusätzlich, dass  $g(\cdot)$  monoton steigend ist, so gilt die entsprechende Aussage für beliebige Quantile.

Bei gruppierten Daten gilt für alle  $\alpha \in (0, 1)$  und alle  $\alpha$ -Quantile  $x_\alpha$ : Die Gruppe, in der  $x_\alpha$  liegt, ist ein  $\alpha$ -Quantil für das gruppierte Merkmal  $X_{\text{grupp}}$ .

### 3.3 Modus

**Definition** Sei  $x_1, \dots, x_n$  die Urliste eines nominalskalierten Merkmals mit den Ausprägungen  $a_1, \dots, a_k$  und der Häufigkeitsverteilung  $h_1, \dots, h_k$ , so heißt  $a_{j^*}$  *Modus*  $x_{\text{mod}}$  genau dann, wenn  $h_{j^*} \geq h_j$ , für alle  $j = 1, \dots, k$ .

### 3.4 Ein kurzer Vergleich der Lagemaße und einige Bemerkungen

### 3.5 Geometrisches und harmonisches Mittel

**Definition (Geometrisches Mittel)** Sei  $\Omega = \{0, \dots, n\}$  eine Menge von Zeitpunkten und  $b_0, b_1, \dots, b_n$  mit  $b_i := B(i)$  die Urliste eines Merkmals  $B$ .

Für  $i = 1, \dots, n$  heißt

$$x_i = \frac{b_i}{b_{i-1}}$$

der  $i$ -te *Wachstumsfaktor* und

$$r_i = \frac{b_i - b_{i-1}}{b_{i-1}} = x_i - 1$$

die  $i$ -te *Wachstumsrate*.

Dann bezeichnet man

$$\bar{x}_{\text{geom}} := \left( \prod_{i=1}^n x_i \right)^{\frac{1}{n}} = (x_1 \cdot x_2 \cdot \dots \cdot x_n)^{\frac{1}{n}}$$

als das *geometrische Mittel der Wachstumsfaktoren*  $x_1, \dots, x_n$ .

Es gilt

$$b_n = b_0 \cdot (\bar{x}_{\text{geom}})^n.$$

**Definition (Harmonisches Mittel)** Sei  $x_1, \dots, x_n$  mit  $x_i \neq 0$  für alle  $i$  die Urliste eines verhältnisskalierten Merkmals  $X$ . Dann heißt

$$\bar{x}_{\text{har}} := \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}}$$

das *harmonische* Mittel der  $x_1, \dots, x_n$ .

### 3.6 Weitere Streuungsmaße

**Variationskoeffizient** Ist  $\bar{x} > 0$ , so heißt die Größe

$$v_X := \frac{\tilde{s}_X}{\bar{x}}$$

*Variationskoeffizient* des Merkmals  $X$ .

**Inter-Quartils-Abstand** Sind  $x_{0.25}$  und  $x_{0.75}$  das obere und das untere Quartil eines Merkmals, so heißt

$$d_{QX} := x_{0.75} - x_{0.25}$$

der *Interquartilsabstand*.

**Median-Absolute-Deviation** Der Median der Werte  $|x_i - x_{\text{med}}|$ ,  $i = 1, \dots, n$ , heißt Median-Absolute-Deviation von  $X$  ( $MAD_X$ ).

**Spannweite** Die Größe

$$R_X := x_{(n)} - x_{(1)}$$

heißt *Spannweite* von  $X$ .

### 3.7 Boxplot

## 4 Konzentrationsmessung

### Durchgängige Annahmen in diesem Kapitel

- $X$  sei ein *verhältnisskaliertes* Merkmal (mit Urliste  $x_1, \dots, x_n$ )
- $x_i \geq 0$ , für alle  $i = 1, \dots, n$ , und  $\sum_{i=1}^n x_i > 0$  (d.h. mindestens ein Wert ist von Null verschieden)
- Betrachtet werden die der Größe nach geordneten Daten:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

### 4.1 Relative Konzentrationsmessung

#### 4.1.1 Lorenzkurve

**Definition** Sei

$$u_j := \frac{j}{n}$$

und

$$v_j := \frac{\sum_{i=1}^j x_{(i)}}{\sum_{i=1}^n x_{(i)}} = \frac{\sum_{i=1}^j x_{(i)}}{\sum_{i=1}^n x_{(i)}}$$

dann heißt die stückweise lineare Kurve durch die Punkte  $(0, 0)$ ,  $(u_1, v_1)$ ,  $(u_2, v_2)$ ,  $\dots$ ,  $(u_n, v_n)$  =  $(1, 1)$  *Lorenzkurve*.

**Berechnung über die Häufigkeiten** Sind die relativen/absoluten Häufigkeiten  $f_1, \dots, f_k$  bzw.  $h_1, \dots, h_k$  der der Größe nach geordneten Merkmalsausprägungen  $a_1 < a_2 < \dots < a_k$  gegeben, so gilt für  $j = 1, \dots, k$

$$u_j = \sum_{l=1}^j \frac{h_l}{n} = \sum_{l=1}^j f_l = F(a_j)$$

und

$$v_j = \frac{\sum_{l=1}^j h_l \cdot a_l}{\sum_{l=1}^k h_l \cdot a_l} = \frac{\sum_{l=1}^j f_l \cdot a_l}{\sum_{l=1}^k f_l \cdot a_l}.$$

**Berechnung bei klassierten Daten** Bei klassierten Daten mit den Klassen  $[c_0, c_1), [c_1, c_2), \dots, [c_{k-1}, c_k]$  und Klassenmitten  $m_l = \frac{c_{l-1} + c_l}{2}$  (mit  $l = 1, \dots, k$ ) verwendet man als Approximation

$$v_j = \frac{\sum_{l=1}^j h_l \cdot m_l}{\sum_{l=1}^k h_l \cdot m_l} = \frac{\sum_{l=1}^j f_l m_l}{\sum_{l=1}^k f_l m_l}.$$

#### 4.1.2 Gini-Koeffizient

**Definition** Gegeben sei die geordnete Urliste  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  eines verhältnisskalierten Merkmals  $X$ . Dann heißt

$$G := \frac{2 \cdot \sum_{i=1}^n i \cdot x_{(i)}}{n \sum_{i=1}^n x_i} - \frac{n+1}{n}$$

*Gini-Koeffizient* und

$$G^{norm} := \frac{n}{n-1} \cdot G$$

*normierter Gini-Koeffizient* (Lorenz-Münzner-Koeffizient).

**Bemerkung** Betrachtet man die geordneten Ausprägungen  $a_1 < a_2 < \dots < a_k$  mit den Häufigkeiten  $h_1, h_2, \dots, h_k$ , so gilt

$$G = \frac{\sum_{l=1}^k (u_{l-1} + u_l) f_l \cdot a_l}{\sum_{l=1}^k f_l \cdot a_l} - 1 = \frac{\sum_{l=1}^k (u_{l-1} + u_l) h_l \cdot a_l}{\sum_{l=1}^k h_l \cdot a_l} - 1 = 1 - \sum_{l=1}^k f_l (v_{l-1} + v_l)$$

mit

$$u_j = \frac{1}{n} \sum_{l=1}^j h_l \quad \text{und} \quad u_0 := 0.$$

#### 4.1.3 Quantilsbezogene relative Konzentrationsmessung

Sei  $0 =: \alpha_0 < \alpha_1 < \dots < \alpha_l < \dots < \alpha_{q-1} < 1 =: \alpha_q$  eine Einteilung der Abszisse und  $z_l^*$  derjenige Merkmalsanteil, der auf die  $l$ -te Quantilsgruppe entfällt. Dann ergibt sich die Kurve durch die Punkte  $(u_l^*, v_l^*)$  mit

$$u_l^* = \alpha_l \quad \text{und} \quad v_l^* = \sum_{r \leq l} z_r^*$$

### Berechnung des Gini-Koeffizienten

Unter der Annahme, dass in der jeweiligen Quantilsgruppe alle Einkommen gleich sind, so hat man Häufigkeitsdaten mit den Ausprägungen  $a_1, a_2, \dots, a_k$  vorliegen, d.h.  $a_l$  ist der Wert in der  $l$ -ten Quantilsgruppe und man erhält

$$\begin{aligned}
 G^* &= \frac{\sum_{l=1}^k (u_{l-1}^* + u_l^*) f_l^* \cdot a_l}{\sum_{l=1}^k f_l^* \cdot a_l} - 1 \\
 &= \sum_{l=1}^k (u_{l-1}^* + u_l^*) \cdot \frac{f_l^* \cdot a_l}{\sum_{l=1}^k f_l^* \cdot a_l} - 1 \\
 &= \left( \sum_{l=1}^k (u_{l-1}^* + u_l^*) \cdot z_l^* \right) - 1 \\
 &= 1 - \sum_{l=1}^q f_l^* (v_{l-1}^* + v_l^*)
 \end{aligned}$$

mit

$$f_l^* := \alpha_l - \alpha_{l-1}, \quad l = 1, \dots, q$$

#### 4.1.4 Einige weitere quantilsbasierte Maße

##### Robin-Hood-Index

- Äquidistante Einteilung der Abszisse
- Wie viel müsste den Reichen weggenommen werden, um zu einer Konzentration von 0 zu kommen?
- Ermittle für jede Quantilsgruppe mit einem Anteil von höchstens  $\alpha = \frac{1}{q}$  den Abstand ihres Anteils zu  $\alpha$ !
- Aufaddieren der positiven Abstände liefert den *Robin-Hood-Index*.

##### Quantilverhältnisse

Bilde das Verhältnis von  $(1 - \alpha)$ - und  $\alpha$ -Quantil, zum Beispiel:

$$\frac{x_{0.9}}{x_{0.1}} \quad \text{Dezilverhältnis (falls } x_{0.1} > 0 \text{)}.$$

## 4.2 Absolute Konzentrationsmessung

**Definition (Konzentrationsrate)** Sei  $0 \leq x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  die geordnete Urliste eines verhältnisskalierten Merkmals mit  $\sum_{i=1}^n x_i > 0$ . Mit

$$p_{(i)} := \frac{x_{(i)}}{\sum_{j=1}^n x_j}$$

heißt

$$CR_g := \sum_{i=n-g+1}^n p_{(i)}$$

*Konzentrationsrate* (vom Grade  $g$ ).

**Definition (Herfindahl-Index)** Sei  $0 \leq x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  die geordnete Urliste eines verhältnisskalierten Merkmals mit  $\sum_{i=1}^n x_i > 0$ . Mit

$$p_{(i)} := \frac{x_{(i)}}{\sum_{j=1}^n x_j}$$

heißt

$$H := \sum_{i=1}^n p_{(i)}^2 = \sum_{i=1}^n p_i^2$$

*Herfindahl-Index*.

**Bemerkung** Die Größe  $1 - H$  wird auch als *Rae-Index* bezeichnet.  $\frac{1}{H}$  heißt *Zahl der effektiven Parteien (Marktteilnehmer)*.

## 5 Assoziationsmessung in Kontingenztafeln

### 5.1 Multivariate Merkmale

### 5.2 Kontingenztafeln und bedingte Verteilungen

#### 5.2.1 Gemeinsame Verteilung, Randverteilung, Kontingenztafel

Betrachtet wird ein zweidimensionales Merkmal  $(X, Y)$  bestehend aus den diskreten Merkmalen  $X$  und  $Y$  und die zugehörige Urliste  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . Wir wollen ferner annehmen, dass  $X$  und  $Y$  nur endlich viele („wenige“) verschiedene Werte  $a_1, \dots, a_i, \dots, a_k$  bzw.  $b_1, \dots, b_j, \dots, b_m$  annehmen können.

#### Gemeinsame relative und absolute Häufigkeitsverteilung

$$h_{ij} = h(a_i, b_j), \quad i = 1, \dots, k, \quad j = 1, \dots, m$$

Anzahl von Beobachtungen mit  $x = a_i$  und  $y = b_j$ .

$$f_{ij} = h_{ij}/n = f(a_i, b_j), \quad i = 1, \dots, k, \quad j = 1, \dots, m$$

Anteil von Beobachtungen mit  $x = a_i$  und  $y = b_j$ .

Man nennt  $(h_{ij})$  und  $(f_{ij})$ ,  $i = 1, \dots, k, j = 1, \dots, m$ , die *gemeinsame Verteilung* von  $(X, Y)$  in absoluten bzw. relativen Häufigkeiten.

#### Kontingenztafel / Kontingenztafel / Kreuztafel

Darstellung der Häufigkeiten in Form einer  $(k \times m)$ -dimensionalen Häufigkeitstabelle

- Kontingenztafel der absoluten Häufigkeitsverteilung:

	$b_1$	$\dots$	$b_j$	$\dots$	$b_m$	
$a_1$	$h_{11}$	$\dots$	$h_{1j}$	$\dots$	$h_{1m}$	$h_{1\bullet}$
$a_2$	$h_{21}$	$\dots$	$h_{2j}$	$\dots$	$h_{2m}$	$h_{2\bullet}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
$a_i$	$h_{i1}$	$\dots$	$h_{ij}$	$\dots$	$h_{im}$	$h_{i\bullet}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
$a_k$	$h_{k1}$	$\dots$	$h_{kj}$	$\dots$	$h_{km}$	$h_{k\bullet}$
	$h_{\bullet 1}$	$\dots$	$h_{\bullet j}$	$\dots$	$h_{\bullet m}$	$n$

mit den *Randverteilungen*

$$h_{i\bullet} = h_{i1} + \dots + h_{im} = h(a_i), \quad i = 1, \dots, k, \quad \text{für } X$$

und

$$h_{\bullet j} = h_{1j} + \dots + h_{kj} = h(b_j), \quad j = 1, \dots, m, \quad \text{für } Y.$$

- Kontingenztafel der relativen Häufigkeitsverteilung:

	$b_1$	$\cdots$	$b_j$	$\cdots$	$b_m$	
$a_1$	$f_{11}$	$\cdots$	$f_{1j}$	$\cdots$	$f_{1m}$	$f_{1\bullet}$
$a_2$	$f_{21}$	$\cdots$	$f_{2j}$	$\cdots$	$f_{2m}$	$f_{2\bullet}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
$a_i$	$f_{i1}$	$\cdots$	$f_{ij}$	$\cdots$	$f_{im}$	$f_{i\bullet}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
$a_k$	$f_{k1}$	$\cdots$	$f_{kj}$	$\cdots$	$f_{km}$	$f_{k\bullet}$
	$f_{\bullet 1}$	$\cdots$	$f_{\bullet j}$	$\cdots$	$f_{\bullet m}$	1

mit der relativen Häufigkeiten  $f_{ij} = \frac{h_{ij}}{n}$  und den *Randverteilungen*

$$f_{i\bullet} = \frac{h_{i\bullet}}{n} = f_{i1} + \dots + f_{im} = f(a_i), \quad i = 1, \dots, k, \quad \text{für } X$$

und

$$f_{\bullet j} = \frac{h_{\bullet j}}{n} = f_{1j} + \dots + f_{kj} = f(b_j), \quad j = 1, \dots, m, \quad \text{für } Y.$$

### 5.2.2 Ökologischer Fehlschluss

### 5.2.3 Graphische Darstellung der gemeinsamen Verteilung

### 5.2.4 Bedingte Häufigkeitsverteilungen

**Definition** Seien  $h_{i\bullet} > 0$  und  $h_{\bullet j} > 0$  für alle  $i, j$ . Für jedes  $i = 1, \dots, k$  heißt

$$f_{Y|X}(b_1|a_i) := \frac{h_{i1}}{h_{i\bullet}} = \frac{h(a_i, b_1)}{h(a_i)}, \quad \dots, \quad f_{Y|X}(b_m|a_i) := \frac{h_{im}}{h_{i\bullet}} = \frac{h(a_i, b_m)}{h(a_i)}$$

*bedingte (relative) Häufigkeitsverteilung* von  $Y$  unter der *Bedingung*  $X = a_i$ .

Analog heißt für jedes  $j = 1, \dots, m$

$$f_{X|Y}(a_1|b_j) := \frac{h_{1j}}{h_{\bullet j}} = \frac{h(a_1, b_j)}{h(b_j)}, \quad \dots, \quad f_{X|Y}(a_k|b_j) := \frac{h_{kj}}{h_{\bullet j}} = \frac{h(a_k, b_j)}{h(b_j)}$$

*bedingte (relative) Häufigkeitsverteilung* von  $X$  unter der *Bedingung*  $Y = b_j$ .

**Bemerkung** Bedingte Verteilungen werden immer als relative Häufigkeiten ausgedrückt. Für die Berechnung gilt

$$f_{X|Y}(a_i|b_j) = \frac{h_{ij}}{h_{\bullet j}} = \frac{\frac{h_{ij}}{n}}{\frac{h_{\bullet j}}{n}} = \frac{f_{ij}}{f_{\bullet j}}$$

und analog

$$f_{Y|X}(b_j|a_i) = \frac{h_{ij}}{h_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}}.$$

### 5.3 (Empirische) Unabhängigkeit und $\chi^2$

#### 5.3.1 (Empirische) Unabhängigkeit

**Definition** Die beiden Komponenten  $X$  und  $Y$  eines bivariaten Merkmals  $(X, Y)$  heißen voneinander (*empirisch*) *unabhängig*, falls für alle  $i = 1, \dots, k$  und  $j = 1, \dots, m$

$$f_{Y|X}(b_j|a_i) = f_{\bullet j} = f(b_j) \quad (1)$$

und

$$f_{X|Y}(a_i|b_j) = f_{i\bullet} = f(a_i) \quad (2)$$

gilt.

#### Satz

- Es genügt, entweder (1) oder (2) zu überprüfen: Mit einer der beiden Beziehungen gilt auch die andere.
- $X$  und  $Y$  sind genau dann empirisch unabhängig, wenn für alle  $i = 1, \dots, k$  und alle  $j = 1, \dots, m$  gilt:

$$f_{ij} = f_{i\bullet} \cdot f_{\bullet j}. \quad (3)$$

- Gleichung (3) ist äquivalent zu

$$h_{ij} = \frac{h_{i\bullet} \cdot h_{\bullet j}}{n}.$$

#### 5.3.2 $\chi^2$ -Abstand, $\chi^2$ -Koeffizient

**Definition** Mit

$$\tilde{h}_{ij} := \frac{h_{i\bullet} \cdot h_{\bullet j}}{n}.$$

wird definiert:

$$\chi^2 := \sum_{i=1}^k \sum_{j=1}^m \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}.$$

#### Alternative Berechnung von $\chi^2$ in Vierfeldertafeln

$$\chi^2 = \frac{n \cdot (h_{11}h_{22} - h_{12}h_{21})^2}{h_{1\bullet}h_{2\bullet}h_{\bullet 1}h_{\bullet 2}} \quad (4)$$

### 5.3.3 $\chi^2$ -basierte Maßzahlen

#### Kontingenzkoeffizient nach Pearson

$$K := \sqrt{\frac{\chi^2}{n + \chi^2}}$$

#### Korrigierter Kontingenzkoeffizient

$$K^* := \frac{K}{K_{\max}}$$

mit

$$K_{\max} := \sqrt{\frac{\min\{k, m\} - 1}{\min\{k, m\}}}.$$

#### Kontingenzkoeffizient nach Cramér (Cramér's V)

$$\begin{aligned} V &= \sqrt{\frac{\chi^2}{n \cdot (\min\{k, m\} - 1)}} \\ &= \sqrt{\frac{\chi^2}{\text{maximaler Wert}}} \end{aligned}$$

**Phi-Koeffizient  $\Phi$**  (entspricht dem Cramér's V bei der Vierfeldertafel ( $k = m = 2$ ))

$$V = \sqrt{\frac{\chi^2}{n \cdot (\min\{k, m\} - 1)}} = \sqrt{\frac{\chi^2}{n}}.$$

mit (4) ergibt sich also

$$\Phi = \left| \frac{h_{11}h_{22} - h_{12}h_{21}}{\sqrt{h_{1\bullet}h_{2\bullet}h_{\bullet 1}h_{\bullet 2}}} \right|.$$

Lässt man die Betragsstriche weg, so erhält man den *signierten Phi-Koeffizienten* oder *Punkt-Korrelationskoeffizienten*

$$\Phi_s = \frac{h_{11}h_{22} - h_{12}h_{21}}{\sqrt{h_{1\bullet}h_{2\bullet}h_{\bullet 1}h_{\bullet 2}}},$$

der häufig ebenfalls als *Phi-Koeffizient* bezeichnet wird.

### 5.4 Weitere Methoden für Vierfeldertafeln

**Risiko** Aus der medizinischen Statistik kommend wird die bedingte relative Häufigkeit  $f(b_j|a_i)$  oft auch als *Risiko* für  $b_j$  unter Bedingung  $a_i$  bezeichnet:

$$R(b_j|a_i) := f_{Y|X}(b_j|a_i) = \frac{h_{ij}}{h_{i\bullet}} \quad i, j = 1, 2.$$

**Relatives Risiko** Für eine Vierfelder-Tafel heißt

$$RR(b_1) := \frac{f_{Y|X}(b_1|a_1)}{f_{Y|X}(b_1|a_2)} = \frac{h_{11}/h_{1\bullet}}{h_{21}/h_{2\bullet}}$$

relatives Risiko für  $b_1$ .

**Prozentsatzdifferenz** Die Größe

$$d\%(b_j) := (f_{Y|X}(b_j|a_1) - f_{Y|X}(b_j|a_2)) \cdot 100, \quad j = 1, 2$$

heißt *Prozentsatzdifferenz* für  $b_j$ .

**Odds** Die Größe

$$O(b_1|a_i) := \frac{R(b_1|a_i)}{1 - R(b_1|a_i)} \quad i = 1, 2$$

heißt *Odds* oder *Chance* von  $b_1$  unter der Bedingung  $a_i$ .

**Odds Ratio (Kreuzproduktverhältnis)** Es gilt:

$$OR(b_1) := \frac{O(b_1|a_1)}{O(b_1|a_2)} = \frac{h_{11} \cdot h_{22}}{h_{12} \cdot h_{21}}$$

**Yules Q** Die Größe

$$Q := \frac{h_{11} \cdot h_{22} - h_{12} \cdot h_{21}}{h_{11} \cdot h_{22} + h_{12} \cdot h_{21}}$$

heißt Yules  $Q$ .

### 5.5 PRE-Maße (Prädiktionsmaße)

**Definition (Prädiktionsmaß)** PRE = Proportional Reduction in Error

$$PRE = \frac{E_1 - E_2}{E_1} = 1 - \frac{E_2}{E_1}$$

wobei

$E_1$  : Vorhersagefehler bei Modell 1

$E_2$  : Vorhersagefehler bei Modell 2

#### Guttmans Lambda

$$\lambda_Y = \frac{\left( \sum_{i=1}^k \max_j(h_{ij}) \right) - \max_j(h_{\bullet j})}{n - \max_j(h_{\bullet j})}$$

$$\lambda_X = \frac{\left( \sum_{j=1}^m \max_i(h_{ij}) \right) - \max_i(h_{i\bullet})}{n - \max_i(h_{i\bullet})}$$

$$\lambda = \frac{\sum_{i=1}^k \max_j(h_{ij}) + \sum_{j=1}^m \max_i(h_{ij}) - \max_j(h_{\bullet j}) - \max_i(h_{i\bullet})}{2n - \max_j(h_{\bullet j}) - \max_i(h_{i\bullet})}$$

#### Goodmans und Kruskals Tau

$$\tau_Y = \frac{\sum_{j=1}^m \sum_{i=1}^k \frac{f_{ij}^2}{f_{i\bullet}} - \sum_{j=1}^m f_{\bullet j}^2}{1 - \sum_{j=1}^m f_{\bullet j}^2}$$

$$\tau_X = \frac{\sum_{i=1}^k \sum_{j=1}^m \frac{f_{ij}^2}{f_{\bullet j}} - \sum_{i=1}^k f_{i\bullet}^2}{1 - \sum_{i=1}^k f_{i\bullet}^2}$$

$$\tau = \frac{\sum_{j=1}^m \sum_{i=1}^k \frac{f_{ij}^2}{f_{i\bullet}} + \sum_{i=1}^k \sum_{j=1}^m \frac{f_{ij}^2}{f_{\bullet j}} - \sum_{j=1}^m f_{\bullet j}^2 - \sum_{i=1}^k f_{i\bullet}^2}{2 - \sum_{j=1}^m f_{\bullet j}^2 - \sum_{i=1}^k f_{i\bullet}^2}$$

## 5.6 Zusammenhangsanalyse bivariater ordinaler Merkmale

**Definition (Konkordante Paare)** Gegeben sei die Urliste eines bivariaten Merkmals  $(X, Y)$ , wobei  $X$  und  $Y$  jeweils ordinales Skalenniveau besitzen. Ein Paar  $(i, j), i \neq j$ , von Einheiten mit den Ausprägungen  $(x_i, y_i)$  und  $(x_j, y_j)$  heißt

a) *konkordant* (gleichläufig), falls entweder

$$(x_i > x_j \text{ und } y_i > y_j) \quad \text{oder} \quad (x_i < x_j \text{ und } y_i < y_j)$$

gilt.

b) *diskordant* (gegenläufig), falls entweder

$$(x_i > x_j \text{ und } y_i < y_j) \quad \text{oder} \quad (x_i < x_j \text{ und } y_i > y_j)$$

gilt.

c) *ausschließlich in  $X$  gebunden*, falls

$$(x_i = x_j \text{ und } y_i \neq y_j)$$

d) *ausschließlich in  $Y$  gebunden*, falls

$$(x_i \neq x_j \text{ und } y_i = y_j)$$

e) *in  $X$  und  $Y$  gebunden*, falls

$$(x_i = x_j \text{ und } y_i = y_j)$$

Ferner bezeichne

- $C$  die Anzahl der konkordanten Paare,
- $D$  die Anzahl der diskordanten Paare,
- $T_X$  die Anzahl der Paare mit Bindungen ausschließlich in  $X$ ,
- $T_Y$  die Anzahl der Paare mit Bindungen ausschließlich in  $Y$ ,
- $T_{XY}$  die Anzahl der Paare mit Bindungen in  $X$  und  $Y$ .

**Definition ( $\tau_a$ ,  $\tau_b$  und  $\gamma$  für ordinale Daten)** Die Zusammenhangsmaße für ordinale Daten heißen

$$\tau_a := \frac{C - D}{\frac{n(n-1)}{2}}$$

Kendalls *Tau a*,

$$\tau_b := \frac{C - D}{\sqrt{(C + D + T_X) \cdot (C + D + T_Y)}}$$

Kendalls *Tau b* und

$$\gamma := \frac{C - D}{C + D}$$

Goodmans und Kruskals *Gamma*.

## 5.7 Drittvariablenkontrolle

## 6 Korrelationsanalyse: Zusammenhangsanalyse stetiger Merkmale

### 6.1 Korrelationsanalyse

#### 6.1.1 Streudiagramm, Kovarianz- und Korrelationskoeffizienten

#### 6.1.2 Streudiagramme (Scatterplots)

#### 6.1.3 Kovarianz und Korrelation

**Definition** Gegeben sei ein bivariates Merkmal  $(X, Y)$  mit metrisch skalierten Variablen  $X$  und  $Y$  mit  $\tilde{s}_X^2 > 0$  und  $\tilde{s}_Y^2 > 0$ . Dann heißen

$$\text{Cov}(X, Y) := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

(empirische) Kovarianz von  $X$  und  $Y$ ,

$$\varrho(X, Y) := \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

(empirischer) Korrelationskoeffizient nach Bravais und Pearson von  $X$  und  $Y$ , und

$$R_{XY}^2 := (\varrho(X, Y))^2$$

Bestimmtheitsmaß von  $X$  und  $Y$ .

#### Verschiebungssatz

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}$$

und damit

$$\varrho(X, Y) = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sqrt{\sum_{i=1}^n x_i^2 - n \bar{x}^2} \cdot \sqrt{\sum_{i=1}^n y_i^2 - n \bar{y}^2}}.$$

**Transformation**  $\varrho(X, Y)$  und  $R_{XY}^2$  sind invariant gegenüber streng monoton steigenden linearen Transformationen. Genauer gilt mit  $\tilde{X} := a \cdot X + b$  und  $\tilde{Y} := c \cdot Y + d$

$$\varrho(\tilde{X}, \tilde{Y}) = \varrho(X, Y) \quad \text{falls } a \cdot c > 0$$

und

$$\varrho(\tilde{X}, \tilde{Y}) = -\varrho(X, Y) \quad \text{falls } a \cdot c < 0.$$

#### 6.1.4 Weitere Korrelationskoeffizienten

**Punkt-Korrelationskoeffizienten** Anwendung des Korrelationskoeffizienten nach Bravais-Pearson auf dichotome nominale Merkmale

- dichotome nominale Merkmale kodiert mit 0 und 1
- Der Punkt-Korrelationskoeffizienten ist identisch zu  $\Phi$  aus Kapitel 5.3

**Punkt-biseriale Korrelation** Anwendung des Korrelationskoeffizienten nach Bravais-Pearson auf eine dichotome und eine metrischen Variablen.

#### Rangkorrelationskoeffizient nach Spearman

- Wir betrachten ein bivariates Merkmal  $(X, Y)$ , wobei  $X$  und  $Y$  nur ordinalskaliert sind, aber viele unterschiedlichen Ausprägungen besitzen.
- Man rechnet statt mit Beobachtungen  $(x_i, y_i)_{i=1, \dots, n}$  mit Rängen  $(\text{rg}(x_i), \text{rg}(y_i))_{i=1, \dots, n}$ . Dabei ist

$$\text{rg}(x_i) = j : \iff x_i = x_{(j)},$$

- Liegen keine Bindungen vor, rechnet man direkt mit den Rängen.
- Liegen Bindungen vor, so nimmt man den Durchschnittswert der in Frage kommenden Ränge.

#### Definition (Rangkorrelationskoeffizient nach Spearman)

$$\varrho_S(X, Y) := \frac{\sum_{i=1}^n \text{rg}(x_i) \cdot \text{rg}(y_i) - n \left( \frac{n+1}{2} \right)^2}{\sqrt{\sum_{i=1}^n (\text{rg}(x_i))^2 - n \left( \frac{n+1}{2} \right)^2} \sqrt{\sum_{i=1}^n (\text{rg}(y_i))^2 - n \left( \frac{n+1}{2} \right)^2}}$$

heißt (empirischer) *Rangkorrelationskoeffizient nach Spearman*.

Liegen keine Bindungen vor, so gilt

$$\varrho_S(X, Y) = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n(n^2 - 1)},$$

wobei  $d_i := \text{rg}(x_i) - \text{rg}(y_i)$ .

## 6.2 Regressionsanalyse I: Die lineare Einfachregression

**Definition** Gegeben seien zwei metrische Merkmale  $X$  und  $Y$  und das Modell der linearen Einfachregression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n.$$

Dann bestimme man  $\hat{\beta}_0$  und  $\hat{\beta}_1$  so, dass mit

$$\begin{aligned} \hat{\varepsilon}_i &:= y_i - \hat{y}_i \\ &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \end{aligned}$$

das Kleinste-Quadrate-Kriterium

$$\sum_{i=1}^n \hat{\varepsilon}_i^2$$

minimal wird. Die optimalen Werte  $\hat{\beta}_0$  und  $\hat{\beta}_1$  heißen *KQ-Schätzungen*,  $\hat{\varepsilon}_i$  bezeichnet das  $i$ -te (geschätzte) *Residuum*.

**Satz** Für die KQ-Schätzer gilt:

$$\text{a) } \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{Cov}(X, Y)}{\tilde{s}_X^2} = \varrho(X, Y) \frac{\tilde{s}_Y}{\tilde{s}_X}$$

$$\text{b) } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x},$$

$$\text{c) } \sum_{i=1}^n \hat{\varepsilon}_i = 0.$$

### 6.2.1 Modellanpassung: Bestimmtheitsmaß und Residualplots

**Streuungszerlegung**

$$SQT = SQR + SQE$$

mit

- $SQT := \sum_{i=1}^n (y_i - \bar{y})^2$   
(Gesamtstreuung / Gesamtvariation der  $y_i$ : „sum of squares total“)
- $SQR := \sum_{i=1}^n (\hat{y}_i - y_i)^2$   
(Residualstreuung / Residualvariation: „sum of squared residuals“).
- $SQE := SQT - SQR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$   
(durch das Regressionsmodell erklärte Streuung: „sum of squares explained“)

**Bestimmtheitsmaß**

$$\frac{SQT - SQR}{SQT} = \frac{SQE}{SQT}.$$

Es gilt:

$$\frac{SQE}{SQT} = R_{XY}^2.$$

**6.3 Multiple lineare Regression****Modellgleichung**

$$y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i.$$

Dabei bezeichnet  $x_{i1}$  den für die  $i$ -te Beobachtung beobachteten Wert der Variablen  $X_1$ ,  $x_{i2}$  den Wert der Variablen  $X_2$ , usw.

**KQ-Prinzip** Bestimme  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$  so, dass mit

$$\hat{\varepsilon}_i = y_i - \hat{y}_i := y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_p x_{pi})$$

der Ausdruck

$$\sum_{i=1}^n \hat{\varepsilon}_i^2$$

minimal wird.

**Bestimmtheitsmaß**

$$R^2 = \frac{SQE}{SQT}$$

**Korrigiertes Bestimmtheitsmaß**

$$\tilde{R}^2 := 1 - \frac{n-1}{n-p-1}(1 - R^2)$$

**6.3.1 Schema eines Computer-Outputs einer multiplen Regression**

	Estimate	Std. Dev.	t	Sig.
(Intercept)	$\hat{\beta}_0$	$\hat{\sigma}_0$	$T_0$	p-Wert
$X_1$	$\hat{\beta}_1$	$\hat{\sigma}_1$	$T_1$	"
$X_2$	$\hat{\beta}_2$	$\hat{\sigma}_2$	$T_2$	"
$\vdots$	$\vdots$	$\vdots$	$\vdots$	"
$X_p$	$\hat{\beta}_p$	$\hat{\sigma}_p$	$T_p$	"

## 6.4 Nominale Einflussgrößen in Regressionsmodellen, Varianzanalyse

**Dichotome Kovariable** Dichotome Variablen können, sofern sie mit 0 und 1 (wichtig!) kodiert sind, ebenfalls als Einflussgrößen zugelassen werden.

**Dummykodierung** Mache aus einer kategorialen Variablen mit  $k$  Ausprägungen  $(k - 1)$  Variablen mit den Ausprägungen 0 und 1. Diese  $k - 1$  *Dummyvariablen* dürfen dann in der Regression verwendet werden.

**Interaktionseffekte** Wechselwirkung zwischen Kovariablen lassen sich durch den Einbezug des Produkts als zusätzliche Kovariable modellieren

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \cdot x_{2i} + \varepsilon_i$$

**Varianzanalyse** Ist ein nominales Merkmal  $X$  mit insgesamt  $k$  verschiedenen Ausprägungen die einzige unabhängige Variable, so führt die Regressionsanalyse mit den entsprechenden  $k - 1$  Dummyvariablen auf die sogenannte (einfaktorielle) *Varianzanalyse*: Das zugehörige Bestimmtheitsmaß wird üblicherweise mit  $\eta^2$  bezeichnet:

$$\eta^2 = \frac{SQE}{SQT} = \frac{\sum_{j=1}^k n_j (\bar{y}_j - \bar{y})^2}{\sum_{j=1}^k \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2}.$$

$\eta^2$  und  $\eta = \sqrt{\eta^2}$  werden auch als Maße für den Zusammenhang zwischen einer metrischen Variable und einer nominalen Variable verwendet.

## 6.5 Korrelation und „Kausalität“