5 Assoziationsmessung in Kontingenztafeln

5.1 Multivariate Merkmale

Gerade in der Soziologie ist die Analyse eindimensionaler Merkmale nur der allererste Schritt zur Beschreibung der Daten. Meist ist die Analyse von $Zusammenhängen$ zwischen Merkmalen von größerem Interesse. Beispiele für typische Fragestellungen:

- Beeinflusst das Geschlecht das Erwerbseinkommen?
- Gibt es einen Zusammenhang zwischen Schichtzugehörigkeit (als etwas veralteter, dennoch klassischer soziologischer Begriff) und Aggressionsneigung?
- Spielt die Stärke der Kirchenbindung eine Rolle bei der Parteienpräferenz?

Dazu werden an jeder Einheit $mehrere$ Merkmale erhoben und ihre Ausprägungen auch $gemeinsam$ analysiert (z.B. wird das Geschlecht der i-ten Person mit ihrem Einkommen in Beziehung gesetzt).

Für Merkmale X, Y, Z nennt man

- das Paar (X, Y) ein zweidimensionales (bivariates) Merkmal
- das Tripel (X, Y, Z) ein dreidimensionales (trivariates) Merkmal.

Allgemein spricht man von mehrdimensionalen Merkmalen.

$$\begin{aligned}(X, Y): \ \Omega &\longrightarrow (W_x \times W_y) \\ \omega &\longmapsto (X(\omega), Y(\omega))\end{aligned}$$

Statistische Zusammenhangsmaße messen die Stärke von Zusammenhängen.

Achtung: Statistische Zusammenhangsmaße...

- ...erlauben keine Aussagen über Kausalität!
- ...können nicht klären:
  - die Richtung des Zusammenhangs (was ist Ursache, was Wirkung?)
  - ob eine dritte, evtl. unbeobachtete Variable den Zusammenhang verursacht

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

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

Wir nehmen außerdem an, dass X und Y nur endlich viele (wenige), verschiedene Werte

$$a_1, \ldots, a_i, \ldots, a_k \quad \text{bzw.} \quad b_1, \ldots, b_j, \ldots b_m$$

annehmen können.

Anmerkung: In vielen Büchern (v.a. zur induktiven Statistik) wird statt $a_1, \ldots, a_k$ auch $x_1, \ldots, x_k$ und analog statt $b_1, \ldots, b_m$ auch $y_1, \ldots, y_m$ geschrieben. Bei uns aber sind die (x_i, y_i) Werte der Urliste, x_i also der Wert von X bei der i-ten Einheit.

Beispiel (fiktiv):

$$X \quad \text{Schichtzugehörigkeit} \begin{cases} 1, & \text{Unterschicht} \\ 2, & \text{Mittelschicht} \\ 3, & \text{Oberschicht} \end{cases}$$

$$Y \quad \text{latente Aggressivität} \begin{cases} 1, & \text{ja} \\ 2, & \text{nein} \end{cases}$$

Mögliche Urliste des zweidimensionalen Merkmals (X, Y):

$$(3,1),(2,2),(2,1),(3,1),(3,2),(3,1),(1,2),(1,1),(1,1),(1,2),(3,1),(3,1)$$

Einheit	X	Y
1	3	1
2	2	2
3	2	1
4	3	1
5	3	2
6	3	1
7	1	2
8	1	1
9	1	1
10	1	2
11	3	1
12	3	1

Achtung:

- Tupel sind – im Gegensatz zu Mengen – *geordnete* Anordnungen von Zahlen
- Die Tupel sind „gemeinsam indiziert“, d.h. die Werte in einem Tupel beziehen sich immer auf dieselbe Einheit i. Bei (x_i, y_i) sind x_i und y_i also die Ausprägungen derselben Einheit i (z.B. der Person i). Nur so können Zusammenhänge zwischen den Merkmalen sichtbar werden!

Gemeinsame relative und absolute Häufigkeitsverteilung:

$$\begin{aligned} h_{ij} &= h(a_i, b_j),\ i = 1, \ldots, k,\ j = 1, \ldots, m, \\ &\quad \text{Anzahl von Beobachtungen mit } x = a_i \text{ und } y = b_j. \end{aligned}$$

$$\begin{aligned} f_{ij} &= h_{ij}/n = f(a_i, b_j),\ i = 1, \ldots, k,\ j = 1, \ldots, m, \\ &\quad \text{Anteil von Beobachtungen mit } x = a_i \text{ und } y = b_j. \end{aligned}$$

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

Kontingenztafel / Kontingenztabelle / Kreuztabelle:

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

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

Der Punkt steht für Summation über den entsprechenden Index. Es gilt also:

$$h_{i\bullet} = \sum_{j=1}^{m} h_{ij}, \quad \text{d.h. } h_{i\bullet} \text{ ist die } i\text{-te Zeilensumme}$$

$$h_{\bullet j} = \sum_{i=1}^{k} h_{ij}, \quad \text{d.h. } h_{\bullet j} \text{ ist die } j\text{-te Spaltensumme}$$

Man kann diese $Randhäufigkeiten$ (sie treten in der Kontingenztabelle am $Rand$ auf) zur getrennten Betrachtung von X und Y verwenden, denn es gilt

$$\begin{aligned} h_{i\bullet} &= h_{i1} + \ldots + h_{im} = h(a_i), \quad i = 1, \ldots, k, \\ h_{\bullet j} &= h_{1j} + \ldots + h_{kj} = h(b_j), \quad j = 1, \ldots, m. \end{aligned}$$

Somit ist $h_{i\bullet}$ die absolute Häufigkeit von a_i bei der Betrachtung des Merkmals X und $h_{\bullet j}$ die absolute Häufigkeit von b_j bei der Betrachtung des Merksmals Y.

Man nennt $\{h_{i\bullet}, i = 1, \ldots, k\}$ und $\{h_{\bullet j}, j = 1, \ldots, m\}$ die $Randverteilungen$ des Tupels (X, 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}	$f_{2\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
a_i	f_{i1}	$\cdots$	f_{ij}	$\cdots$	f_{im}	$f_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
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 den relativen Häufigkeiten $f_{ij} = \dfrac{h_{ij}}{n}$ und den *Randverteilungen*

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

und

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

Beispiel: Aggressivität und Schichtzugehörigkeit

X \ Y	1	2	

Beachte: Aus der gemeinsamen Verteilung kann man die Randverteilungen berechnen (aber nicht umgekehrt).

Beispiel: Parteipräferenz bei unterschiedlichem Einkommen (Befragung von 722 Personen, 3.5.-5.5.2004)

Einkommen * Parteipräferenz Kreuztabelle

Anzahl

		Parteipräferenz						
		SPD	CDU/CSU	Grüne	FDP	PDS	Sonstige	Gesamt
Einkommen	< 1500	55	88	27	15	15	10	210
	1500 < ... < 3000	119	151	28	21	17	14	350
	> 3000	26	89	26	15	0	6	162
Gesamt		200	328	81	51	32	30	722

Bemerkung:

Bei Vierfeldertafeln vereinfachen sich die Tabellen wesentlich, mit der Angabe der Häufigkeit in einer Zelle sind bei gegebenen Randhäufigkeiten auch die Häufigkeiten in den anderen Zellen bestimmt.

	b_1	b_2	
a_1	h_{11}	h_{12}	$h_{1\bullet}$
a_2	h_{21}	h_{22}	$h_{2\bullet}$
	$h_{\bullet 1}$	$h_{\bullet 2}$	n

z.B. gegeben h_{11}

$\Rightarrow h_{12} = h_{1\bullet} - h_{11}$ etc.

Unabhängige und abhängige Variable:

Hat man eine Vermutung über die Richtung einer potentiellen Wirkung, so bezeichnet man die Variablen entsprechend als *unabhängige* (wirkende, erklärende) und *abhängige* (bewirkte) Variable, z.B.:

möglicherweise:	Schicht	$\longrightarrow$	latente Aggresivität
eindeutig:	Geschlecht	$\longrightarrow$	Einkommen
allgemein:	unabhängige	$\longrightarrow$	abhängige Variable

In der Statistik ist es üblich, die unabhängige Variable mit X zu bezeichnen und die abhängige Variable mit Y, wie gewohnt ist dann Y eine Funktion von X.

Damit werden die Häufigkeitsverteilungen für feste Werte der unabhängigen Variablen in den Zeilen der Kontingenztabelle angegeben.

Vorsicht: In einigen Büchern wird entgegen dieser Konvention die unabhängige Variable in den Spalten und die abhängige in den Zeilen abgetragen.

5.2.2 Ökologischer Fehlschluss

Achtung: Hier wird der Begriff *ökologisch* im Sinne von *kollektiv* verwendet (Robinson, 1950).

Es gibt sehr viele gemeinsame Verteilungen, die zu denselben Randhäufigkeiten passen. Im Beispiel oben passt u.a.:

X \ Y	1	2	
1			4
2			2
3			6
	8	4	12

Man sieht also, wie wichtig es zur Feststellung potentieller Zusammenhänge ist, die *gemeinsame* Verteilung h_{ij} zu kennen, also tatsächlich die Paare (x_i, y_i) zu betrachten.

Der *unzulässige* Schluss

- von der Randverteilung auf Eigenschaften der gemeinsamen Verteilung,
- also von zwei univariaten Ergebnissen auf ein bivariates,
- von der Kollektiv- auf die Individualebene,

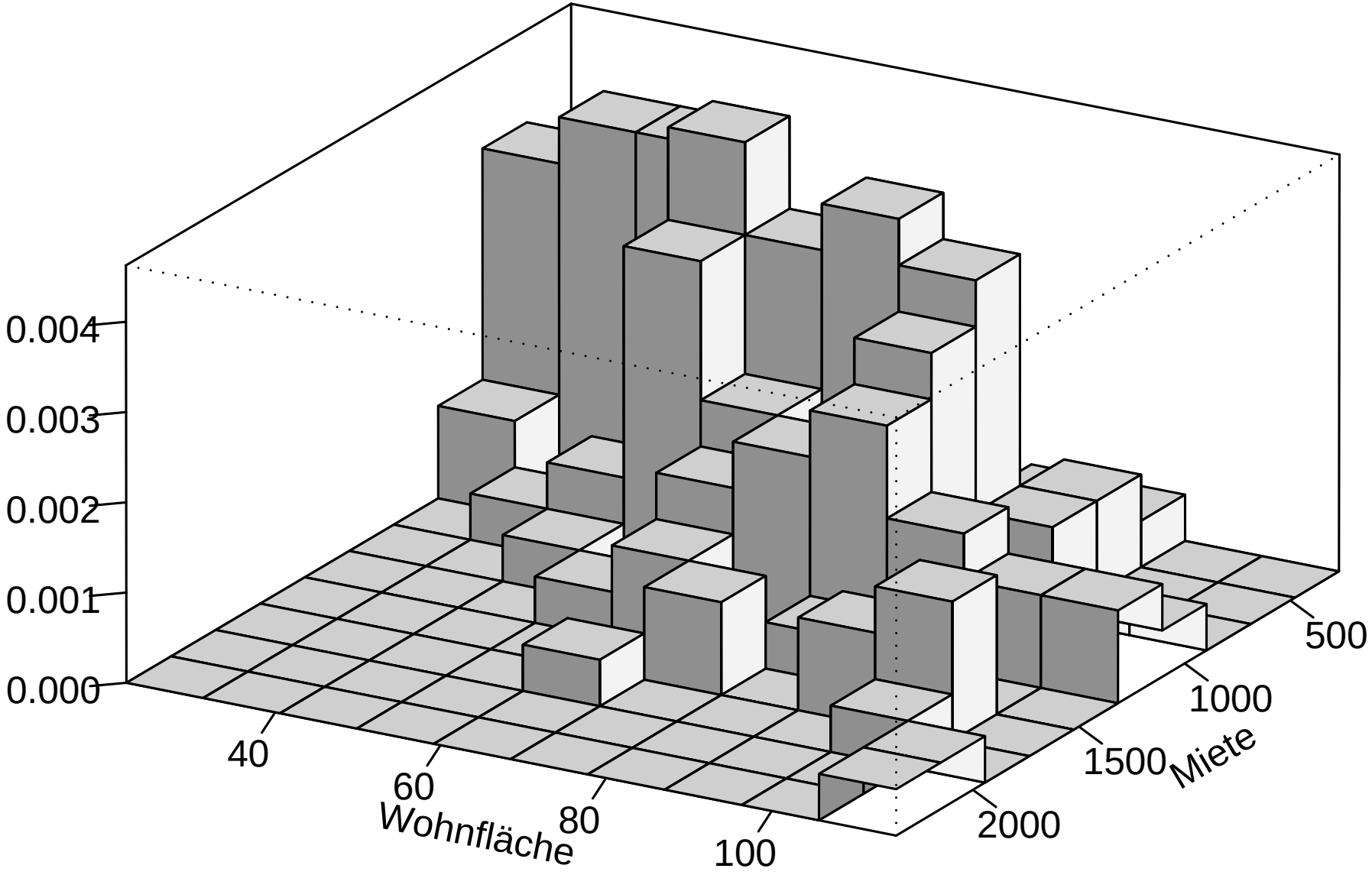
heißt *ökologischer Fehlschluss*.

Kommen zwei Eigenschaften (verschiedene Merkmale) häufig vor, heißt dies nicht notwendig, dass sie gemeinsam häufig vorkommen.

5.2.3 Grafische Darstellung der gemeinsamen Verteilung

Verschiedene Darstellungsarten, z.B.

- als 3D-Säulendiagramm der gemeinsamen Häufigkeiten h_{ij}
- als „normale“ Säulendiagramme nach einer Variable aufgespalten, d.h. für jeden Wert a_i von X werden jeweils die Häufigkeiten h_{ij} bzw. f_{ij} aufgetragen.


0.004
0.003
0.002
0.001
0.000
40
60
80
100
Wohnfläche
500
1000
1500
2000
Miete

5.2.4 Bedingte Häufigkeitsverteilungen

Beispiel:

Habilitationen nach Geschlecht und Fach

Grundgesamtheit: alle Habilitationen 1993

Geschlecht: X
Fächergruppe: Y

		Sprachw. Kulturw.	Rechtsw. Wirts., Soz.	Naturw.	Kunst	Medizin	
		1	2	3	4	5	
weiblich	1	51	20	30	4	44	149
männlich	2	216	92	316	10	433	1067
		267	112	346	14	477	1216

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

$$f_Y(b_1|a_i) := \frac{h_{i1}}{h_{i\bullet}} = \frac{h(a_i, b_1)}{h(a_i)}, \quad \ldots, \quad f_Y(b_m|a_i) := \frac{h_{im}}{h_{i\bullet}} = \frac{h(a_i, b_m)}{h(a_i)}$$

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

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

$$f_X(a_1|b_j) := \frac{h_{1j}}{h_{\bullet j}} = \frac{h(a_1, b_j)}{h(b_j)}, \quad \ldots, \quad f_X(a_k|b_j) := \frac{h_{kj}}{h_{\bullet j}} = \frac{h(a_k, b_j)}{h(b_j)}$$

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

Im Beispiel:

$$f_X(\text{Frau} \mid \text{Habil. in Kunst}) =$$

$$f_X(\text{Frau} \mid \text{Habil. in Naturw.}) =$$

Zu unterscheiden von

$$f_{13} = \frac{h_{13}}{n} = \frac{30}{1216} \approx$$

bzw.

$$f_Y(\text{Habil. in Kunst}|\text{Frau}) =$$

Die Verwechslung von gemeinsamer und bedingter Verteilung bzw. verschiedener bedingter Verteilungen ist eine häufige Fehlerquelle.

Konvention: Bei Vermutung über Richtung des Zusammenhangs betrachtet man vorwiegend die bedingte Verteilung der abhängigen Variablen gegeben die festen Werte der unabhängigen Variable. In diese Richtung geht ja auch die „Prognose“! Man kennt den Wert der unabhängigen Variablen und will Aussagen über die abhängige machen.

Beispiel: Gesucht: Bedingte Verteilung der Fächergruppen gegeben das Geschlecht, d.h. $f_Y(b_j|a_i)$ für verschiedene i.

		Sprachw. Kulturw.	Rechtsw. Wirts., Soz.	Naturw.	Kunst	Medizin	
		1	2	3	4	5	
weiblich	1	51	20	30	4	44	149
männlich	2	216	92	316	10	433	1067
		267	112	346	14	477	1216

$Y : b_j$ / $X : a_i$		Sprachw. Kulturw.	Rechtsw. Wirts., Soz.	Naturw.	Kunst	Medizin
		1	2	3	4	5
weiblich	1					
männlich	2					

$$f_Y(\text{Rechtsw.}|\text{weiblich}) =$$
$$f_Y(\text{Kunst}|\text{männlich}) =$$

Gesucht: Bedingte Verteilung des Geschlechts gegeben die Fachgruppe,
d.h. $f_X(a_i|b_j)$ für verschiedene j.

b_j / a_i		Sprachw. Kulturw. 1	Rechtsw. Wirts., Soz. 2	Naturw. 3	Kunst 4	Medizin 5
weiblich	1					
männlich	2					

$$f_X(\text{weiblich}|\text{Rechtsw.}) =$$

$$f_X(\text{männlich}|\text{Kunst}) =$$

Nochmals zur Interpretation:

1. $f_X(\text{weiblich}|\text{Medizin}) =$
2. $f_Y(\text{Medizin}|\text{weiblich}) =$
3. $f_{15} = f(\text{Medizin und weiblich}) =$

Es liegt jeweils eine andere Grundgesamtheit zu Grunde:

Bedingte Verteilungen werden „automatisch“ durch relative Häufigkeiten ausgedrückt. Für die Berechnung gilt

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

und analog

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

Beispiel: Parteipräferenzen

Einkommen * Parteipräferenz Kreuztabelle

% von Einkommen

		Parteipräferenz						
		SPD	CDU/CSU	Grüne	FDP	PDS	Sonstige	Gesamt
Einkommen	< 1500	26.2%	41.9%	12.9%	7.1%	7.1%	4.8%	100.0%
	1500 < ... < 3000	34.0%	43.1%	8.0%	6.0%	4.9%	4.0%	100.0%
	> 3000	16.0%	54.9%	16.0%	9.3%		3.7%	100.0%
Gesamt		27.7%	45.4%	11.2%	7.1%	4.4%	4.2%	100.0%

Einkommen * Parteipräferenz Kreuztabelle

% von Parteipräferenz

		Parteipräferenz						
		SPD	CDU/CSU	Grüne	FDP	PDS	Sonstige	Gesamt
Einkommen	< 1500	27.5%	26.8%	33.3%	29.4%	46.9%	33.3%	29.1%
	1500 < ... < 3000	59.5%	46.0%	34.6%	41.2%	53.1%	46.7%	48.5%
	> 3000	13.0%	27.1%	32.1%	29.4%		20.0%	22.4%
Gesamt		100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Beispiel: A. Quatember (Institut für angewandte Statistik, Linz):
Unsinn in den Medien - Vom allzu sorglosen Umgang mit Daten (I):

Man nehme kritisch Stellung zu dem folgenden Zeitungsausschnitt!

Quelle: Kronen-Zeitung, 15.07.2000

Beispiel: A. Quatember:
Unsinn in den Medien - Vom allzu sorglosen Umgang mit Daten (II):

Wiens Schüler fallen öfter durch

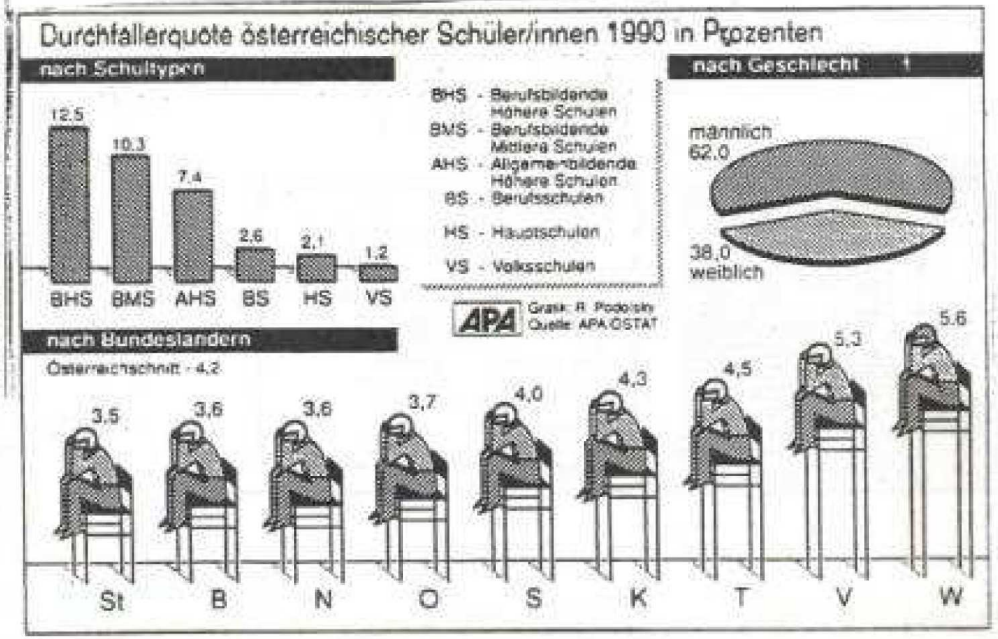
DER STANDARD: 8.5.1992

Mädchen bleiben viel seltener sitzen, sagt das Statistische Zentralamt

Wien – In Wien und Vorarlberg fallen um ein Drittel mehr Schüler durch, als in der Steiermark, Niederösterreich oder im Burgenland. Laut jüngster Erhebung des Österreichischen Statistischen Zentralamtes (ÖSTAT) liegen die „Durchfallerquoten" dieser beiden Länder klar über jenen anderer Bundesländer.

Die Ursachen ortet das ÖSTAT in der unterschiedlichen Leistungsbeurteilung in den Ländern. Im Bundesdurchschnitt dürfen jährlich 4,2 Prozent der Schüler nicht „aufsteigen".

Die Mädchen – oft zahlenmäßig überlegen – stellen nur etwas mehr als ein Drittel der „Sitzenbleiber". Bei den Burschen dagegen erreichen 62 Prozent ihr Klassenziel nicht. Die meisten „Durchfaller" gibt es an Berufsbildenden Höheren Schulen: Nicht einmal jeder Zehnte schafft den Aufstieg. In Höheren Technischen Lehranstalten bleiben mehr als 15 Prozent „sitzen". Die Allgemeinbildenden Höheren Schulen verzeichnen bundesweit 7,4 Prozent „Sitzenbleiber". (APA)



Quelle: Der Standard, 08.05.1992