Vorlesung: Statistik I für Studierende der Statistik, Mathematik & Informatik

Dozent: Fabian Scheipl
Material: H. Küchenhoff
LMU München

Multivariate Statistik

Einführung

In den meisten Anwendungen und in den Beispielen werden an jeder Einheit *gleichzeitig* mehrere Merkmale $X, Y, Z, \ldots$ erhoben:

$\Rightarrow$ **mehrdimensionale** oder **multivariate** Daten

- Grundgesamtheit
- mit Einheiten $i = 1, \ldots, n$
- Werte (x_i, y_i, z_i) der **Merkmale** (X, Y, Z)

Daten

- Daten $(x_1, y_1, z_1), \ldots, (x_i, y_i, z_i), \ldots, (x_n, y_n, z_n)$
  (Im weiteren: Meistens nur 2 Merkmale X, Y)
- Fragestellungen:
  - $X \leftrightarrow Y$: (Wie) hängen X und Y zusammen?
    Assoziation, **Korrelation**
  - $X \rightarrow Y$: (Wie) beeinflusst X das (Ziel-)Merkmal Y?
    Regression

Diskrete und gruppierte Merkmale

- Darstellung, Präsentation von (zwei) diskreten Merkmalen X und Y mit den Ausprägungen

$$\begin{aligned} a_1, \ldots, a_k \quad & \text{für } X \\ b_1, \ldots, b_m \quad & \text{für } Y \end{aligned}$$

- Skalenniveau von X, Y beliebig; X, Y können auch gruppierte metrische Merkmale sein.
- Benutzt wird nur das Nominalskalenniveau der Merkmale.

Kontingenztabellen

Sonntagsfrage: **“Welche Partei würden Sie wählen, wenn am nächsten Sonntag Bundestagswahlen wären?”**

Üblicherweise in Prozent (%) angegeben.
Für einen gegebenen Befragungszeitraum ergab sich folgende Tabelle:

	CDU/CSU	SPD	FDP	Grüne	Rest	$\sum$
Männer	33	35	4	6	22	100
Frauen	40	29	6	10	15	100
gesamt	37	32	5	8	18	100

Aus den ersten beiden Zeilen ergibt sich, dass die Parteipräferenzen für Männer und Frauen unterschiedlich sind.

Sonntagsfrage

In der angegebenen Tabelle sind die ursprünglichen Daten bereits in Prozenten für die geschlechtsspezifischen Populationen angegeben. Die Rückrechnung auf auf 435 Männer und 496 Frauen ergibt:

	CDU/CSU	SPD	FDP	Grüne	Rest	$\sum$
Männer	144	153	17	26	95	435
Frauen	200	145	30	50	71	496
$\sum$	344	298	47	76	166	931

Beispiel: Arbeitslosigkeit

Zwei Merkmale:

- X Ausbildungsniveau mit den Kategorien
  - "keine Ausbildung",
  - "Lehre",
  - "fachspezifische Ausbildung"
  - "Hochschulabschluß"
- Y Dauer der Arbeitslosigkeit mit den Kategorien
  - "Kurzzeitarbeitslosigkeit" ($\leq$ 6 Monate),
  - "mittelfristige Arbeitslosigkeit" (7–12 Monate),
  - "Langzeitarbeitslosigkeit" ($\geq$ 12 Monate)

Arbeitslosigkeit

	Kurzzeit-arbeitslosigkeit	mittelfristige Arbeitslosigkeit	Langzeit-arbeitslosigkeit	$\sum$
k.A.	86	19	18	123
Lehre	170	43	20	233
Fachspez.	40	11	5	56
Hochschule	28	4	3	35
$\sum$	324	77	46	447

Ausbildungsspezifische Dauer der Arbeitslosigkeit für männliche Deutsche

Allgemeine Darstellung

Kontingenztafel der absoluten Häufigkeiten:

Eine $(k \times m)$-**Kontingenztafel der absoluten Häufigkeiten** besitzt die Form

	b_1	$\ldots$	b_m	
a_1	h_{11}	$\ldots$	h_{1m}	$h_{1\cdot}$
a_2	h_{21}	$\ldots$	h_{2m}	$h_{2\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
a_k	h_{k1}	$\ldots$	h_{km}	$h_{k\cdot}$
	$h_{\cdot 1}$	$\ldots$	$h_{\cdot m}$	n

Notation

$h_{ij} = h(a_i, b_j)$ die absolute Häufigkeit der Kombination (a_i, b_j),

$h_{1\cdot}, \ldots, h_{k\cdot}$ die Randhäufigkeiten von X
$h_{\cdot 1}, \ldots, h_{\cdot m}$ die Randhäufigkeiten von Y

mit $h_{i\cdot} = \sum_{j=1}^{m} h_{ij}$, $h_{\cdot j} = \sum_{i=1}^{k} h_{ij}$.

Die Kontingenztabelle gibt die gemeinsame Verteilung der Merkmale X und Y in absoluten Häufigkeiten wieder.

Kontingenztafel der relativen Häufigkeiten

Die $(k \times m)$-**Kontingenztafel der relativen Häufigkeiten** hat die Form

	b_1	$\dots$	b_m	
a_1	f_{11}	$\dots$	f_{1m}	$f_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
a_k	f_{k1}	$\dots$	f_{km}	$f_{k\cdot}$
	$f_{\cdot 1}$	$\dots$	$f_{\cdot m}$	1

Notation

$f_{ij} = h_{ij}/n$ die relative Häufigkeit der Kombination (a_i, b_j),

$f_{i\cdot} = \sum_{j=1}^{m} f_{ij} = h_{i\cdot}/n, \quad i = 1, \ldots, k$, die relativen Randhäufigkeiten zu X,

$f_{\cdot j} = \sum_{i=1}^{k} f_{ij} = h_{\cdot j}/n, \quad j = 1, \ldots, m$, die relativen Randhäufigkeiten zu Y.

Die Kontingenztabelle gibt die gemeinsame Verteilung von X und Y wieder.

Bedingte Häufigkeiten

Zusammenhang zwischen X und Y aus **gemeinsamen** Häufigkeiten h_{ij} bzw. f_{ij} schwer ersichtlich.

Deshalb: Blick auf **bedingte** Häufigkeiten
$\Rightarrow$ Verteilung des einen Merkmals für festgehaltenen Wert des zweiten Merkmals

Bedingte Häufigkeiten: Beispiel

Sonntagsfrage nach Geschlecht:

	CDU/CSU	SPD	FDP	Grüne	Rest	$\sum$
Männer	33	35	4	6	22	100
Frauen	40	29	6	10	15	100

Prozentzahlen für Parteipräferenz in den Schichten (Subpopulationen) "weibliche Wähler", "männliche Wähler"
= **bedingte relative Häufigkeiten für Parteipräferenzen gegeben das Geschlecht**

Bedingte relative Häufigkeitsverteilung

Die **bedingte Häufigkeitsverteilung von Y unter der Bedingung** $X = a_i$, kurz $Y|X = a_i$, ist bestimmt durch

$$f_Y(b_1|a_i) = \frac{h_{i1}}{h_{i\cdot}}, \ldots, f_Y(b_m|a_i) = \frac{h_{im}}{h_{i\cdot}}.$$

Die **bedingte Häufigkeitsverteilung von X unter der Bedingung** $Y = b_j$, kurz $X|Y = b_j$, ist bestimmt durch

$$f_X(a_1|b_j) = \frac{h_{1j}}{h_{\cdot j}}, \ldots, f_X(a_k|b_j) = \frac{h_{kj}}{h_{\cdot j}}.$$

Bemerkung

Wegen

$$\frac{h_{i1}}{h_{i\cdot}} = \frac{h_{i1}/n}{h_{i\cdot}/n} = \frac{f_{i1}}{f_{i\cdot}}$$

gilt auch

$$f_Y(b_1|a_i) = \tfrac{f_{i1}}{f_{i\cdot}}, \ldots, f_Y(b_m|a_i) = \tfrac{f_{im}}{f_{i\cdot}}$$

$$f_X(a_1|b_j) = \tfrac{f_{1j}}{f_{\cdot j}}, \ldots, f_X(a_k|b_j) = \tfrac{f_{kj}}{f_{\cdot j}}.$$

Merksatz:
Bedingte Häufigkeitsverteilungen werden durch Division der h_{ij} bzw. f_{ij} durch die entsprechende Zeilen- bzw. Spaltensumme gebildet.

Beispiel: Sonntagsfrage

- Zeile $X = a_1 =$ Männer

  Bedingte Häufigkeiten für Männer ($X = a_1$):

  1.Zeile / Randhäufigkeit für Männer

  $\frac{h(a_1,b_1)}{h(a_1)} = f(b_1|a_1), \ldots, \frac{h(a_j,b_j)}{h(a_1)} = f(b_j|a_1), \ldots$

  $\frac{144}{435} \approx 33\%, \quad \frac{153}{435} \approx 35\%$ usw.
- Zeile $X = a_2 =$ Frauen analog, z.B. $\frac{200}{496} \approx 40\%$ usw.

Bedingte und gemeinsame Häufigkeiten

Man kann auch umgekehrt aus bedingten Häufigkeiten und Randhäufigkeiten die gemeinsamen Häufigkeiten ausrechnen. Bei der Sonntagsfrage ist die Tabelle der bedingten Häufigkeiten gegeben und dazu die Randhäufigkeiten

$$h(a_1) = 435 \text{ Männer}, \quad h(a_2) = 496 \text{ Frauen}; \quad n = 931.$$

$$\begin{aligned} & h(a_1) \cdot f(b_1|a_1) = h(a_1, b_1) \\ \Rightarrow \quad & \\ & 435 \cdot 33\% \approx 144 \quad \text{usw.} \end{aligned}$$

So wurde die Tabelle der gemeinsamen Häufigkeiten h_{ij} rekonstruiert.

Beispiel: Arbeitslosigkeit

$f(b_j|a_i), \quad X = a_i, \;\; i = 1, \ldots, 4$ Ausbildungsniveau

z.B. $\frac{86}{123} = 0.699$, $\frac{19}{123} = 0.154, \ldots$

$\frac{170}{233} = 0.730, \ldots$

usw.

Für festgehaltenes Ausbildungsniveau ($X = a_i$) erhält man die relative Verteilung über die Dauer der Arbeitslosigkeit durch die folgende Tabelle.

Bedingte Verteilung

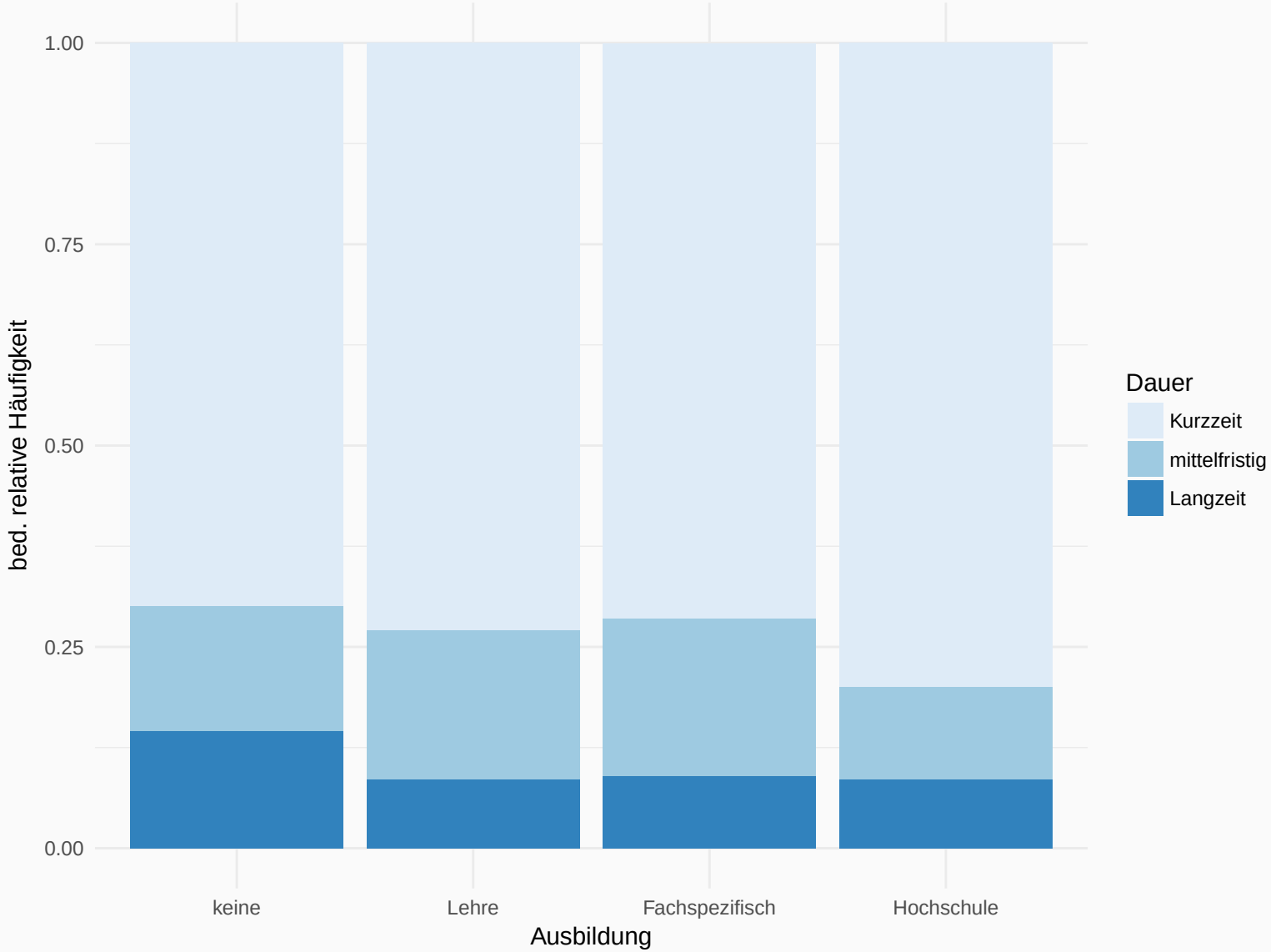
	Kurzzeit-arbeitslosigkeit	mittelfristige Arbeitslosigkeit	Langzeit-arbeitslosigkeit	
Keine Ausb.	0.699	0.154	0.147	1
Lehre	0.730	0.184	0.086	1
Fachspez. Aus.	0.714	0.197	0.089	1
Hochschula.	0.800	0.114	0.086	1

- Bedingen auf das Ausbildungsniveau:
  $\Rightarrow$ Verteilung der Dauer der Arbeitslosigkeit für die Subpopulationen "Keine Ausbildung", "Lehre", usw.
- Verteilungen lassen sich nun miteinander vergleichen

$\Rightarrow$ Relative Häufigkeit für Kurzzeitarbeitslosigkeit ist in der Subpopulation "Hochschulabschluß" mit 0.8 am größten.

Darstellung der bedingten Verteilung

Zum Beispiel Stapeldiagramme:



Zusammenhangsanalyse in Kontingenztabellen

Bisher: Tabellarische / graphische Präsentation
Jetzt: Maßzahlen für Stärke des Zusammenhangs zwischen X und Y.

Chancen und relative Chancen:

- Zunächst 2×2 - Kontingenztafel

		Y		
		1	2	
X	1	h_{11}	h_{12}	$h_{1\cdot}$
	2	h_{21}	h_{22}	$h_{2\cdot}$
		$h_{\cdot 1}$	$h_{\cdot 2}$	n

Chancen (Odds)

- Wir betrachten die Merkmale X und Y zunächst asymmetrisch: Die Ausprägungen von X definieren (hier 2) Subpopulationen, Y ist das interessierende binäre Merkmal in diesen Subpopulationen
- Unter einer **Chance** (*odds*) versteht man nun das **Verhältnis zwischen dem Auftreten von** $Y = 1$ **und** $Y = 2$.
- **bedingte Chance**: Verhältnis zwischen dem Auftreten von $Y = 1$ und $Y = 2$ **in einer Subpopulation** $X = a_i$**.**

Chancenverhältnis (Odds Ratio)

- Die (empirische) **bedingte Chance** für festes $X = a_i$ ist bestimmt durch

$$\gamma(1,2|X = a_i) = \frac{h_{i1}}{h_{i2}}.$$

- Ein sehr einfaches Zusammenhangsmaß stellen die empirischen **relativen Chancen** (*Odds Ratio*) dar, die gegeben sind durch

$$\gamma(1,2|X = 1, X = 2) = \frac{\gamma(1,2|X = 1)}{\gamma(1,2|X = 2)} = \frac{h_{11}/h_{12}}{h_{21}/h_{22}} = \frac{h_{11}h_{22}}{h_{21}h_{12}},$$

d.h. $\gamma(1,2|X = 1, X = 2)$ ist das Verhältnis zwischen der Chance für $Y = 1$ gegen $Y = 2$ in der 1. Population ($X = 1$, 1. Zeile) zu der entsprechenden Chance in der 2. Population ($X = 2$, 2. Zeile).

Beispiel: Dauer der Arbeitslosigkeit

Beschränkt man sich jeweils nur auf zwei Kategorien der Merkmale Ausbildungsniveau und Dauer der Arbeitslosigkeit, erhält man beispielsweise die Tabelle

	Kurzzeit-arbeitslosigkeit	Mittel- und langfristige Arbeitslosigkeit
Fachspezifische Ausbildung	40	16
Hochschulabschluß	28	7

Daraus ergibt sich für Personen mit fachspezifischer Ausbildung die "Chance", kurzzeitig arbeitslos zu sein im Verhältnis dazu mittel– oder längerfristig arbeitslos zu sein, durch

$$\gamma(1,2|\text{fachspezifisch}) = \frac{40}{16} = 2.5.$$

Beispiel: Dauer der Arbeitslosigkeit

Für Arbeitslose mit Hochschulabschluß erhält man

$$\gamma(1,2|\text{Hochschulabschluß}) = \frac{28}{7} = 4.$$

Für fachspezifische Ausbildung stehen die "Chancen" somit $5:2$, für Arbeitslose mit Hochschulabschluß $4:1$.

Man erhält für fachspezifische Ausbildung und Hochschulabschluß die relativen Chancen (Odds Ratio)

$$\gamma(1,2|\text{fachsp. Ausbildung, Hochschule}) = \frac{2.5}{4} = 0.625 = \frac{40 \cdot 7}{16 \cdot 28}$$

Interpretation "Odds Ratio"

- Wegen der spezifischen Form $\gamma(1,2|X=1,X=2) = (h_{11}h_{22})/(h_{21}h_{12})$ werden die relativen Chancen auch als **Kreuzproduktverhältnis** bezeichnet. Es gilt

  $\gamma = 1$ Chancen in beiden Subpopulationen gleich

  $\gamma > 1$ Chance in Subpopulation $X = 1$ größer als in Subpopulation $X = 2$

  $\gamma < 1$ Chance in Subpopulation $X = 1$ niedriger als in Subpopulation $X = 2$.

- Die relativen Chancen geben somit an, um welchen Faktor sich die Chancen in den beiden Subpopulationen unterscheiden

Symmetrie

- Für die Kontingenztafel

$$\boxed{\begin{array}{cc} h_{11} & h_{12} \\ h_{21} & h_{22} \end{array}}$$

ist das *Kreuzproduktverhältnis* (*relative Chance* oder *Odds Ratio*) bestimmt durch

$$\gamma = \frac{h_{11}/h_{12}}{h_{21}/h_{22}} = \frac{h_{11}h_{22}}{h_{21}h_{12}}.$$

- Die asymmetrische Betrachtung der Merkmale X und Y wird aufgehoben

Fall - Kontroll - Studien

Beispiel: Morbus Alzheimer und Genetik

	ApoE3	ApoE4	Summe
Kontrolle	2258	803	3061
Fall	593	620	1213
	2851	1423	4274

$$OR = \frac{593/620}{2258/803} = 0.34$$

$\Rightarrow$ Chance für ApoE3 bei Fällen um den Faktor 3 niedriger als bei Kontrollen

$\overset{?}{\Rightarrow}$ ApoE4 Risiko-Faktor für Morbus Alzheimer

Fall - Kontroll - Studien

Zentrale Argumentation:

Odds Ratio ist symmetrisches Maß

d.h. Chancenverhältnis für Auftreten von ApoE4 bei Kontrolle zu Auftreten von ApoE4 bei Fällen

Person ist krank bei ApoE3
zu
Person ist krank bei ApoE4

$\Rightarrow$ Interpretation als **Risikofaktor** zulässig

Verallgemeinerung auf $k \times m$ Kontingenztafeln, Anmerkungen

- Verallgemeinerung des Verfahres auf mehr als zwei Ausprägungen mindestens eines Merkmals: Man beschränkt sich auf jeweils zwei Zeilen $X = a_i$ und $X = a_j$ und zwei Spalten $Y = b_r$ und $Y = b_s$ und die zugehörigen vier Zellen einer $(k \times m)$-Kontingenztafel.
- Verwendung einer Referenzkategorie
- Statt Odds Ratio wird oft der logarithmierte Odds Ratio verwendet

Anwendung: Apolipoprotein E und Morbus Alzheimer

Etablierter Zusammenhang zwischen Apolipoprotein Eε4 und Morbus Alzheimer

Daten aus Metaanalyse

ApoE genotype	$\varepsilon2\varepsilon2$	$\varepsilon2\varepsilon3$	$\varepsilon2\varepsilon4$	$\varepsilon3\varepsilon3$	$\varepsilon3\varepsilon4$	$\varepsilon4\varepsilon4$
Clinical controls	27	425	81	2258	803	71
Clinical Alzheimer	7	74	41	593	620	207
PM controls	3	75	18	358	120	8
PM Alzheimer	1	20	17	249	373	97

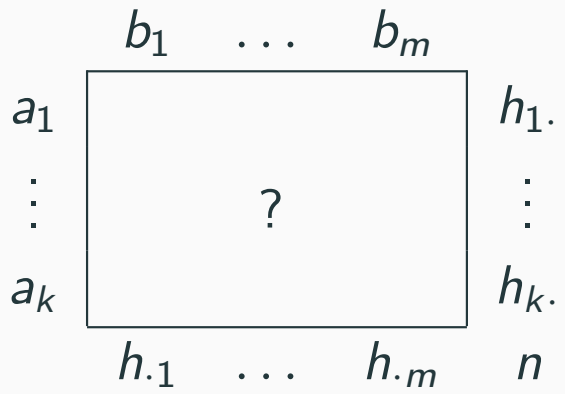
Anwendung: Apolipoprotein E und Morbus Alzheimer

OR im Vergleich zu $\varepsilon 3\varepsilon 3$ (Referenz)

ApoE genotype	$\varepsilon 2\varepsilon 2$	$\varepsilon 2\varepsilon 3$	$\varepsilon 2\varepsilon 4$	$\varepsilon 3\varepsilon 3$	$\varepsilon 3\varepsilon 4$	$\varepsilon 4\varepsilon 4$
OR (klinisch)	1	0.7	2.94	1	2.94	11.1
OR (post mortem)	0.5	0.4	1.4	1	4.5	17.4

Kontingenz- und χ^2-Koeffizient

Ausgangspunkt: Wie sollten gemeinsame Häufigkeiten $\tilde{h}_{ij}$ bzw. $\tilde{f}_{ij}$ verteilt sein, damit - bei vorgegebenen Randverteilungen - die Merkmale X und Y als "empirisch unabhängig" angesehen werden können?



Empirische Unabhängigkeit

Idee: X und Y "empirisch unabhängig"
$\Leftrightarrow$ Bedingte relative Häufigkeiten

$$f_Y(b_1|a_i), \ldots, f_Y(b_m|a_i), \quad i = 1, \ldots, k$$

sind in jeder Schicht $X = a_i$ identisch, d.h. unbeeinflusst von a_i.

Formal:

$$\begin{aligned}
f_Y(b_1|a_1) &= f(b_1), \ldots, f_Y(b_m|a_1) = f_Y(b_m) \\
f_Y(b_1|a_2) &= f(b_1), \ldots, f_Y(b_m|a_2) = f_Y(b_m) \\
\vdots &= \vdots \\
f_Y(b_1|a_k) &= f(b_1), \ldots, f_Y(b_m|a_k) = f_Y(b_m)
\end{aligned}$$

Bsp: Empirische Unabhängigkeit

	b_1	b_2	b_3	
a_1	10	20	30	60
a_2	20	40	60	120
	30	60	90	180

$$
\begin{aligned}
f_Y(b_1|a_1) &= f_Y(b_1|a_2) = f_Y(b_1) = \frac{1}{6} \\
f_Y(b_2|a_1) &= f_Y(b_2|a_2) = f_Y(b_2) = \frac{1}{3} \\
f_Y(b_3|a_1) &= f_Y(b_3|a_2) = f_Y(b_3) = \frac{1}{2}
\end{aligned}
$$

Bemerkung: Lokale Odds Ratios sind alle 1

Bsp: Empirische Unabhängigkeit

Wie sehen also die **unter empirischer Unabhängigkeit erwarteten** (absoluten und relativen) Häufigkeiten $\tilde{h}_{ij}$ und $\tilde{f}_{ij}$ aus?

$$
\begin{aligned}
& f_Y(b_1|a_i) = f(b_1), \ldots, f_Y(b_m|a_i) = f_Y(b_m), \ i = 1, \ldots, k \\
\Leftrightarrow & \frac{\tilde{h}_{ij}}{h_{i\cdot}} = \frac{h_{\cdot j}}{n} \\
\Leftrightarrow & \tilde{h}_{ij} = \frac{h_{i\cdot} h_{\cdot j}}{n} \\
\Leftrightarrow & \tilde{f}_{ij} = f_{i\cdot} f_{\cdot j}
\end{aligned}
$$

Unabhängigkeitstabelle

Idee:

Vergleiche für jede Zelle (i, j) $\tilde{h}_{ij}$ mit tatsächlich beobachteten h_{ij}

$\Rightarrow \chi^2$-**Koeffizient** ist bestimmt durch

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

Eigenschaften des χ^2-Koeffizienten

- $\chi^2 \in [0, \infty)$
- $\chi^2 = 0 \Leftrightarrow X$ und Y **empirisch unabhängig**
- χ^2 groß $\Leftrightarrow$ starker Zusammenhang
- χ^2 klein $\Leftrightarrow$ schwacher Zusammenhang
- **Nachteil:** χ^2 hängt vom Stichprobenumfang n und von der Dimension der Tafel ab, Wert schwierig interpretierbar

Kontingenzkoeffizient und korrigierter Kontingenzkoeffizient

Weitere Normierung $\Rightarrow$ **Kontingenzkoeffizient**

Der Kontingenzkoeffizient ist bestimmt durch

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

und besitzt den Wertebereich $K \in \left[0, \sqrt{\frac{M-1}{M}}\right]$, wobei $M = \min\{k, m\}$.

Der **korrigierte Kontingenzkoeffizient** ergibt sich durch

$$K^* = K / \sqrt{\frac{M-1}{M}}$$

mit dem Wertebereich $K^* \in [0, 1]$.

Eigenschaften des Kontingenzkoeffizienten

- Es wird nur die *Stärke* des Zusammenhangs gemessen, nicht die Richtung wie beim Odds Ratio.
- Vorsicht ist geboten bei einem Vergleich von Kontingenztafeln gleicher Zellenzahl mit stark unterschiedlichen Stichprobenumfängen, da χ^2 mit wachsendem Stichprobenumfang wächst, beispielsweise führt eine Verzehnfachung von h_{ij} und $\tilde{h}_{ij}$ zu zehnfachem χ^2.
- Sämtliche Maße benutzen nur das Nominalskalenniveau von X und Y.

Beispiel: Sonntagsfrage

Für die Kontingenztafel aus Geschlecht und Parteipräferenz für das Beispiel der Sonntagsfrage erhält man die in der folgenden Tabelle wiedergegebenen zu erwartenden Häufigkeiten $\tilde{h}_{ij}$.

	CDU/CSU	SPD	FDP	Grüne	Rest	
Männer	160.7 (144)	139.2 (153)	22.0 (17)	35.5 (26)	77.6 (95)	435
Frauen	183.3 (200)	158.8 (145)	25.0 (30)	40.5 (50)	88.4 (71)	496
	344	298	47	76	166	

Zu erwartende Häufigkeiten $\tilde{h}_{ij}$ und tatsächliche Häufigkeiten h_{ij} (in Klammern)

Interpretation: Wiesen Geschlecht und Parteipräferenz keinen Zusammenhang auf, wären 160.73 die CDU/CSU präferierende Männer zu erwarten gewesen, tatsächlich wurden aber nur 144 beobachtet.

Interpretation:

Insgesamt hier

- $\chi^2 = 20.1$
- $K = 0.15$
- $K^* = 0.21$

$\Rightarrow$ mittelstarke Abhängigkeit zwischen Geschlecht & Parteienpräferenz

Spezialfall: (2×2)-Tafel

Für den Spezialfall einer (2×2)-Tafel

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	

erhält man χ^2 aus

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}.$$

Beispiel: Arbeitslosigkeit

Aus der Kontingenztafel

	Mittelfristige Arbeitslosigkeit	Langfristige Arbeitslosigkeit	
Keine Ausbildung	19	18	37
Lehre	43	20	63
	62	38	100

erhält man also unmittelbar

$$\chi^2 = \frac{100(19 \cdot 20 - 18 \cdot 43)^2}{37 \cdot 63 \cdot 62 \cdot 38} = 2.8$$

und $K = 0.17,\ K^* = 0.23.$

Mehrdimensionale Kontingenztabellen

Beispiel: Überleben beim Titanic-Untergang

- Mehrere diskrete Merkmale: Geschlecht, Klasse, Kind/ Erwachsene, Überleben (Ja/Nein)
- Darstellung durch geeignete bedingte und marginale Verteilungen
- Berechnung von Odds-Ratio zweier Merkmale bedingt auf ein drittes Merkmal
- Graphische Darstellung durch Mosaik-Plot

Beispiel: Überlebende bei Titanic

```
str(Titanic)
```

```
##  table [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118 154 ...
##  - attr(*, "dimnames")=List of 4
##   ..$ Class   : chr [1:4] "1st" "2nd" "3rd" "Crew"
##   ..$ Sex     : chr [1:2] "Male" "Female"
##   ..$ Age     : chr [1:2] "Child" "Adult"
##   ..$ Survived: chr [1:2] "No" "Yes"
```

```
apply(Titanic, MAR = c(4, 1), FUN = sum)
```

```
##          Class
## Survived 1st 2nd 3rd Crew
##      No  122 167 528  673
##      Yes 203 118 178  212
```

Beispiel: Überlebende bei Titanic

Bedingte Verteilung Survived|Class:

```
apply(Titanic, FUN = sum, MAR = 1)
```

```
##  1st  2nd  3rd Crew
##  325  285  706  885
```

```
apply(Titanic, FUN = sum, c(1, 4))/apply(Titanic, FUN = sum, 1)
```

```
##       Survived
## Class    No  Yes
##   1st  0.38 0.62
##   2nd  0.59 0.41
##   3rd  0.75 0.25
##   Crew 0.76 0.24
```

Beispiel: Überlebende bei Titanic

```
apply(Titanic, FUN = sum, c(2, 4))
```

```
##          Survived
## Sex         No Yes
##   Male    1364 367
##   Female   126 344
```

Bedingte Verteilung:

```
apply(Titanic, FUN = sum, c(2, 4))/apply(Titanic, FUN = sum, c(2))
```

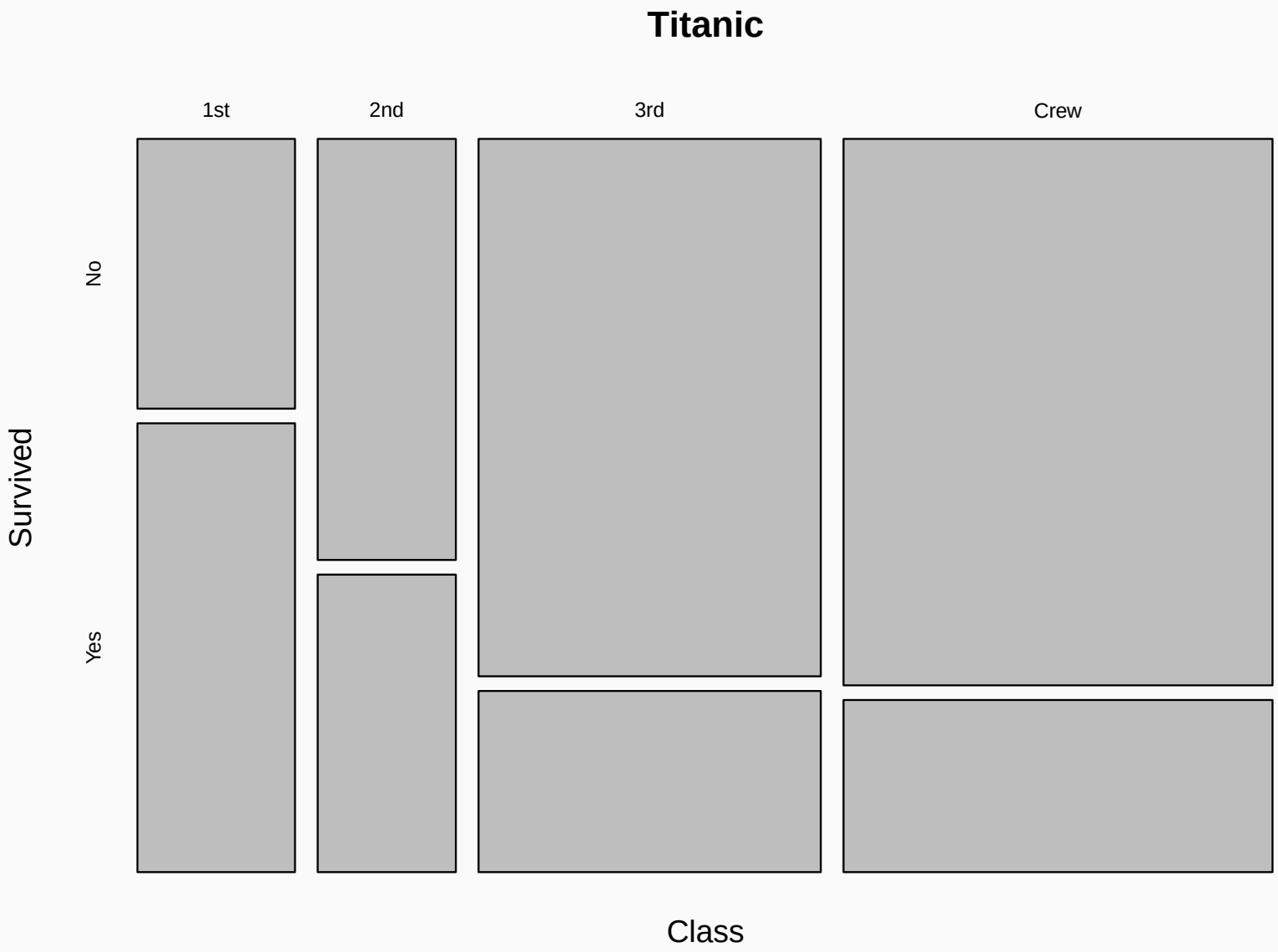
```
##          Survived
## Sex         No  Yes
##   Male    0.79 0.21
##   Female 0.27 0.73
```

Überlebens-Chance Männer: $\frac{367}{1364} \approx 0.27 \approx 1:3$
Überlebens-Chance Frauen: $\frac{344}{126} \approx 2.7 \approx 3:1$
Chancenverhältnis: 10

Mosaik-Plot

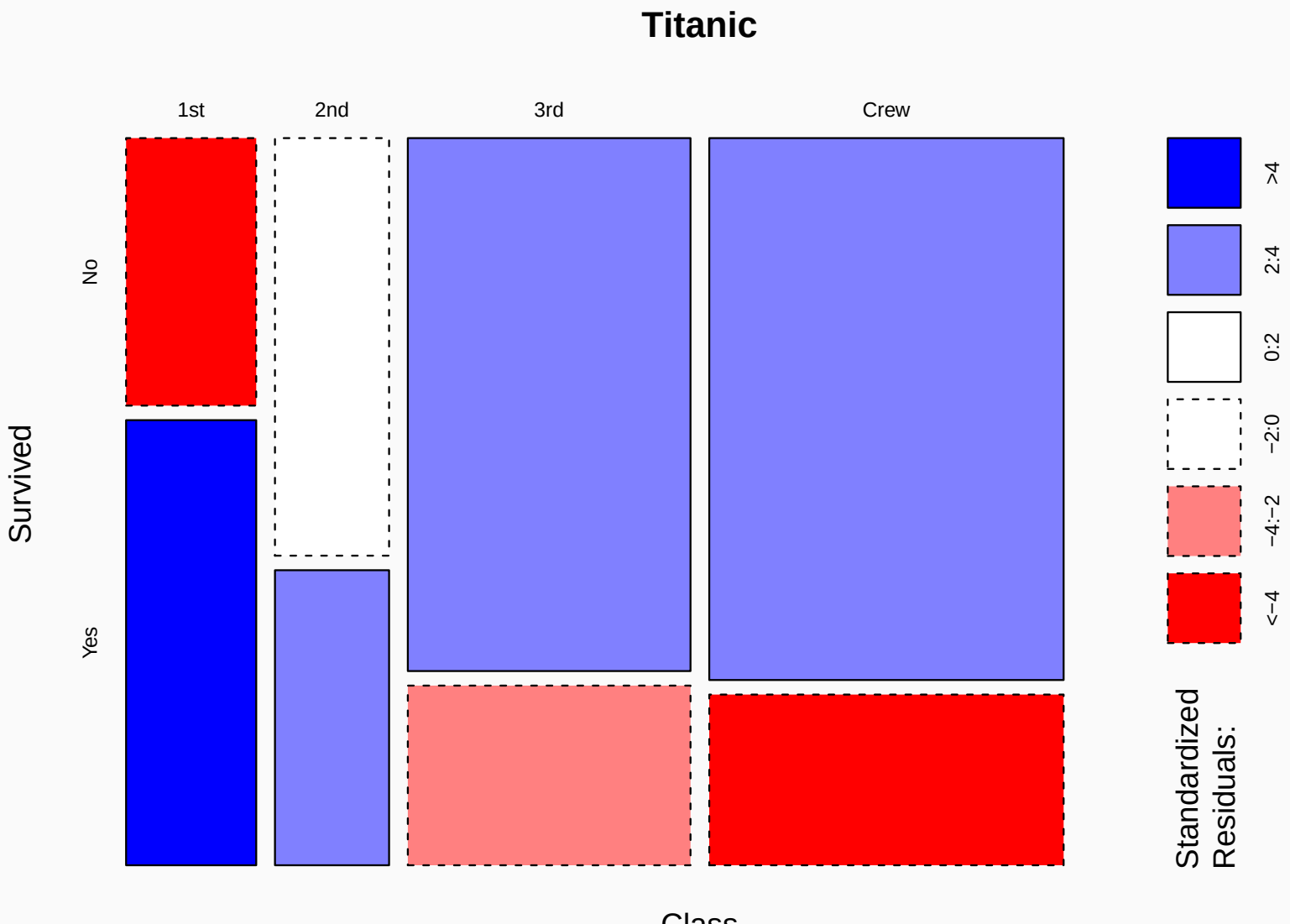
- Flächentreue Darstellung von gemeinsamen Häufigkeiten
- Aufteilung schrittweise
- Zuerst Einflussgröße, zum Schluss nach Zielgröße aufteilen
- Gut geeignet für mehrkategoriale ordinale Daten
- Auch für höhere Dimensionen geeignet

Beispiel: Überlebende bei Titanic

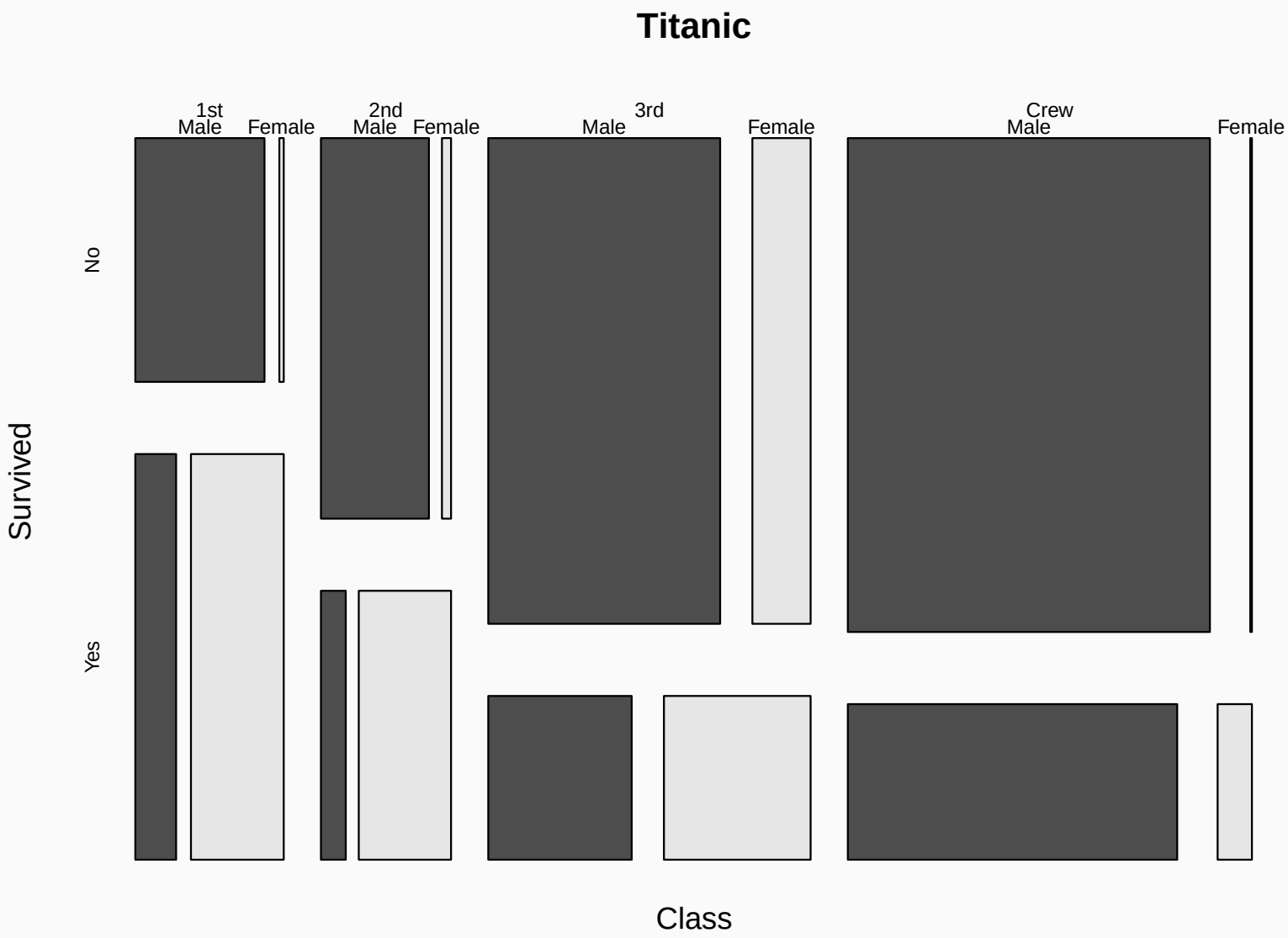


Beispiel: Überlebende bei Titanic

Zellen eingefärbt nach $\frac{(h_{ij} - \tilde{h}_{ij})}{\sqrt{\tilde{h}_{ij}}}$ (*standardized pearson residuals*):



Beispiel: Überlebende bei Titanic



Maße zur Übereinstimmung

Problemstellung:

- Stimmen zwei oder mehrere Beobachter in ihrer Einschätzung überein? (auch: *(inter)rater agreement*)
- Beispiel: Zwei Professoren beurteilen die Referate oder Seminararbeiten von Studenten. Stimmen Sie in ihrer Bewertung (durch Noten) überein?
- Beispiel: Stimmen zwei oder mehrere Ärzte in ihrer Diagnose überein?
- "Wahre" Diagnose oder Bewertung (*gold standard*) unbekannt

Medizinisches Beispiel

Eine Klinik macht computertomografische Aufnahmen (CT–Bilder) oder Röntgenaufnahmen von sogenannten Kalkschultern, also Schultern, die Kalkablagerungen aufweisen.
Mehrere Ärzte sollen anhand von 56 Patienten beurteilen, um welchen Typ es sich bei diesen Ablagerungen handelt (sog. Gärtner–Skala, ordinal):

- **Typ 1**: Aufbauphase (kann Monate oder Jahre dauern).
  Der Patient hat chronische Beschwerden, Schmerztherapie und Krankengymnastik bringen keine Linderung, eine Operation muss erwogen werden.
- **Typ 2**: Beginnende Auflösungsphase.
- **Typ 3**: Auflösung des Kalkdepots (kann Wochen oder Monate dauern).
  Behandlung meist konservativ durch Krankengymnastik und Schmerztherapie.

Medizinisches Beispiel (Fortsetzung)

- Die Beobachtungen lassen sich für zwei Ärzte in einer quadratischen Kontingenztafel veranschaulichen
- Die folgenden Kontingenztafeln geben die Ergebnisse für jeweils zwei verschiedene Ärzte bei den gleichen 56 Patienten wieder

Medizinisches Beispiel (Fortsetzung)

Tabelle 1: 3×3–Tafel für die Einschätzung von Arzt A und Arzt B

	Arzt A			
Arzt B	1	2	3	$\sum$
1	8	20	3	31
2	2	15	1	18
3	0	5	2	7
$\sum$	10	40	6	56

Medizinisches Beispiel (Fortsetzung)

Tabelle 2: 3×3–Tafel für die Einschätzung von Arzt C und Arzt D

	Arzt C			
Arzt D	1	2	3	$\sum$
1	27	6	3	36
2	2	6	3	11
3	0	4	5	9
$\sum$	29	16	11	56

Medizinisches Beispiel (Fortsetzung)

- Vollständige Übereinstimmung in der Einschätzung des Patienten liegt vor, wenn beide Ärzte den gleichen Typ (1, 2 oder 3) zuordnen
- *Bemerkung:* Gleiche Einstufung bedeutet nicht unbedingt, dass diese auch *richtig* im Sinne einer validen Einstufung ist. Beide können sich irren!

Übliche Maßzahlen?

- Üblich bei Kontingenztafeln: χ^2, Kontingenzkoeffizient, etc.
- Man geht von vornherein davon aus, dass ein Zusammenhang bezüglich der Bewertung der Beobachter besteht
- Beobachter führen Bewertung *unabhängig voneinander durch*, d.h. kein Beobachter kennt die Bewertung des oder der anderen Beobachter, die Beurteilungen hängen aber zusammen, da sie jeweils am gleichen Subjekt (hier: Patient) durchgeführt werden

Medizinisches Beispiel (Fortsetzung)

- Für die Tabelle der Ärzte A und B erhält man $(8 + 15 + 2) = 25$ vollständige Übereinstimmungen
- Für die Tabelle der Ärzte C und D erhält man $(27 + 6 + 5) = 38$ vollständige Übereinstimmungen
- Kann man daraus sofort schließen, dass Übereinstimmung von C und D größer ist als von A und B?
  $\rightarrow$ Im Prinzip ja, allerdings müssen wir beachten, dass ein **gewisser Teil der Übereinstimmung zufällig** sein kann

Maßzahl: Kappa–Koeffizient

- Kappa–Koeffizient nach Cohen (1960) dient zur Messung der Übereinstimmung in quadratischen $I \times I$–Kontingenztafeln
- Verwendet lediglich die Beobachtungen, bei denen eine vollständige Übereinstimmung vorliegt, also die Hauptdiagonale der Kontingenztafel

Allgemeine Darstellung

Tabelle 3: Schema einer $I \times I$-Kontingenztafel. Die fettgedruckten Häufigkeiten liegen auf der Diagonalen und werden zur Berechnung von Kappa verwendet.

		Beobachter 1					
		1		i		I	$\sum$
	1	$\mathbf{n_{11}}$	$\cdots$	n_{1i}	$\cdots$	n_{1I}	n_{1+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
Beobachter 2	i	n_{i1}	$\cdots$	$\mathbf{n_{ii}}$	$\cdots$	n_{iI}	n_{i+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	I	n_{I1}	$\cdots$	n_{Ii}	$\cdots$	$\mathbf{n_{II}}$	n_{I+}
	$\sum$	n_{+1}	$\cdots$	n_{+i}	$\cdots$	n_{+I}	n

Allgemeine Darstellung

- Kappa–Koeffizient berücksichtigt darüber hinaus die *zufällige Übereinstimmung* die man auch bekäme wenn die Einschätzungen der Beobachter keinerlei Zusammenhang aufwiesen.
- Wir berechnen daher zwei Größen:
  - Relativer Anteil der Übereinstimmung beider Beobachter:
    $f_o = \sum_{i=1}^{I} f_{ii} = \sum_{i=1}^{I} \frac{n_{ii}}{n} = \frac{\sum_{i=1}^{I} n_{ii}}{n}$
  - Erwartete zufällige Übereinstimmung, wenn kein Zusammenhang bestünde: äquivalent zur Bestimmung der sogenannten erwarteten relativen Häufigkeiten unter Unabhängigkeit bei χ^2:
    $f_e = \sum_{i=1}^{I} f_{i+} f_{+i} = \sum_{i=1}^{I} \frac{n_{i+}}{n} \frac{n_{+i}}{n} = \sum_{i=1}^{I} \frac{n_{i+} n_{+i}}{n^2} = \frac{\sum_{i=1}^{I} n_{i+} n_{+i}}{n^2}$

Berechnung von Kappa (Fortsetzung)

Der Kappa–Koeffizient ist definiert durch

$$\kappa = \frac{f_o - f_e}{1 - f_e} \ , \tag{1.1}$$

mit $f_o = \sum_{i=1}^{I} f_{ii}$ und $f_e = \sum_{i=1}^{I} f_{i+} f_{+i}$

Interpretation von Kappa

- Der Zähler ist die Differenz aus der beobachteten Übereinstimmung und der unter Zufälligkeit zu erwartenden Übereinstimmung. Dies ist damit ein Maß für die über die Zufälligkeit hinausgehende Übereinstimmung der Beobachter (*chance-corrected agreement*)
- Die Eins im Nenner stellt die maximal mögliche relative Häufigkeit für Übereinstimmung dar, nämlich wenn alle Beobachtungen auf der Diagonalen der Kontingenztafel liegen und sämtliche Nebendiagonalen nur Nullen enthalten
- Der Kappa–Koeffizient ist damit ebenfalls Eins, wenn alle Beobachtungen auf der Diagonalen der Kontingenztafel liegen und sämtliche Nebendiagonalen nur Nullen enthalten. Er kann auch negativ werden, wenn zum Beispiel keine Übereinstimmung da ist (im Extremfall: Nullen auf der Diagonalen)
- Der Kappa–Koeffizient ist Null, wenn für alle $i = 1, \ldots, I$ gilt: $f_{ii} = f_{i+} f_{+i}$, das heißt, wenn exakte Unabhängigkeit in der beobachteten Tafel vorliegt

Medizinisches Beispiel (Fortsetzung): Arzt A und B

$$
\begin{aligned}
f_o &= \frac{8+15+2}{56} = \frac{25}{56} \\
\text{und } f_{11} &= \frac{31 \cdot 10}{56^2} \\
f_{22} &= \frac{18 \cdot 40}{56^2} \\
f_{33} &= \frac{7 \cdot 6}{56^2} \\
f_e &= \frac{31 \cdot 10 + 18 \cdot 40 + 7 \cdot 6}{56^2} \\
&= \frac{1072}{56^2} \\
\text{Also } \kappa &= \frac{\frac{25}{56} - \frac{1072}{56^2}}{1 - \frac{1072}{56^2}} = 0.16
\end{aligned}
$$

Medizinisches Beispiel (Ergebnis)

- Entsprechend: Arzt D und C: $\kappa = 0.45$
- Die Einschätzung von A und B ist „schwach übereinstimmend“
- Die Einschätzung von C und D ist „mäßig übereinstimmend“

Veranschaulichung

Zwei mögliche Ergebnisse für die Einschätzung zweier Beobachter für 20 Objekte bezüglich eines dichotomen Merkmals:

Tabelle 4: Problematik von Kappa

	1	0
1	10	1
0	0	9

	1	0
1	18	1
0	0	1

In beiden Fällen erhält man eine Übereinstimmung in 19 von 20 Objekten, oder einen Wert von $f_o = 0.95$. Allerdings ist Kappa für die linke Tafel 0.9 und für die rechte Tafel nur 0.64

Erklärung

- Die Randverteilungen der beiden Beobachter unterscheiden sich dabei jeweils nur gering, d.h. sie scheinen gut kalibriert zu sein (daran liegt es also nicht)
- Offenbar ist die Prävalenz (d.h. der Grundanteil in der untersuchten Population/Stichprobe) in der rechten Tafel für die Ausprägung “1” wesentlicher größer ist als für die Ausprägung “0”. Damit ist aber auch die zufällige Übereinstimmung wahrscheinlicher! Genau dieser Effekt wird bei Kappa berücksichtigt und korrigiert
- Befürworter von Kappa sehen diesen Effekt als wünschenswert an

Kritik an Kappa

Zwei Aspekte werden vermischt:

- Die Beobachter können einen Bias aufweisen, das heisst, die Nicht–Übereinstimmung beruht darauf, dass zum Beispiel Lehrer 1 generell bessere Noten vergibt als Lehrer 2. Man sagt dann auch, dass die Beobachter nicht *kalibriert* sind
- Die Beobachter schätzen die Subjekte verschieden ein. Die Nicht–Übereinstimmung beruht darauf, dass Beobachter 1 zum Beispiel Subjekt 1 höher einstuft als Subjekt 2, Beobachter 2 dagegen Subjekt 2 höher als Subjekt 1.
  Beispiel: Lehrer 1 gibt Schüler 1 eine bessere Note als Schüler 2, Lehrer 2 dagegen gibt Schüler 2 eine bessere Note als Schüler 1.

Aspekt 2 ist der eigentlich uns interessierende Aspekt, während Aspekt 1 (Bias) nach Möglichkeit durch Kalibrierung vermieden werden sollte

Veranschaulichung der Kritik an Kappa

Tabelle 5: Problematik von Kappa

9	3	5	7
5	3	1	7

In der linken Tafel ergibt sich ein Kappa von 0.13, in der rechten Tafel ein Kappa von 0.26, obwohl wieder in beiden Fällen 12 von 20 Objekten ($f_o = 0.6$) übereinstimmend eingestuft wurden

Veranschaulichung der Kritik an Kappa

Erklärung:

- Die Randverteilungen in der rechten Tafel divergieren wesentlich stärker als in der linken Tafel (d.h. hier kann ein Kalibrierungsproblem vorliegen)
- Rechte Tafel:
  - Beobachter 1: $((5+7)/20, (1+7)/20) = (0.6, 0.4)$
  - Beobachter 2: $((5+1)/20, (7+7)/20) = (0.3, 0.7)$
- Linke Tafel:
  - Beobachter 1: $(12/20, 8/20) = (0.6, 0.4)$
  - Beobachter 2: $(14/20, 6/20) = (0.7, 0.3)$

Fazit: Beobachter müssen unbedingt kalibriert werden!

Erweiterungen von Kappa

- Ein weiterer Nachteil von Kappa ist, dass nur die Diagonale berücksichtigt wird
- Wenn die Bewertungsskala ordinal ist und sehr viele verschiedene Merkmalsausprägungen besitzt, so liegen aufeinanderfolgende Ausprägungen oft nicht so weit auseinander
- Wenn zwei Beobachter sich nur gering in der Bewertung unterscheiden, so sollte eine Maßzahl dies auch berücksichtigen können
- Eine solche Maßzahl ist das *gewichtete Kappa*

Gewichtetes Kappa

- Das gewichtete Kappa wurde von Cohen (1968) vorgeschlagen (**Cohen's Kappa**)
- Formal gehen alle Zellen der Kontingenztafel in die Berechnung ein
- Die Zellen auf der Hauptdiagonalen erhalten das höchste Gewicht (in der Regel Gewicht Eins), während die anderen Zellen ein geringeres Gewicht erhalten
- Idee: Zellen deren Einträge schlechte Übereinstimmung repräsentieren werden niedriger gewichtet.

Definition: Gewichtetes Kappa

Das gewichtete Kappa ist definiert als

$$\kappa_w = \frac{f_o^* - f_e^*}{1 - f_e^*}, \tag{1.2}$$

mit

$$\begin{aligned} f_o^* &= \sum_{i=1}^{I} \sum_{j=1}^{I} w_{ij} f_{ij} \\ f_e^* &= \sum_{i=1}^{I} \sum_{j=1}^{I} w_{ij} f_{i\cdot} f_{\cdot j} \end{aligned}$$

Dabei wird f_o^* wie beim ungewichteten Kappa als relativer Anteil der Übereinstimmung beider Beobachter aufgefasst, während f_e^* die zufällige Übereinstimmung darstellt, wenn kein Zusammenhang bestehen würde

Wahl der Gewichte w_{ij}

Zwei populäre Vorschläge sind

$$w_{ij} = 1 - \frac{(i-j)^2}{(I-1)^2} \tag{1.3}$$

und

$$w_{ij}^* = 1 - \frac{|i-j|}{I-1} \ . \tag{1.4}$$

Gewichte w_{ij} und w_{ij}^* im 3×3-Fall

Tabelle 6: Gewichte w_{ij} einer 3×3–Tafel

1.0	0.75	0.0
0.75	1.0	0.75
0.0	0.75	1.0

Tabelle 7: Gewichte w_{ij}^* einer 3×3–Tafel

1.0	0.5	0.0
0.5	1.0	0.5
0.0	0.5	1.0

Die Zellen der größten Nichtübereinstimmung (Zelle $(1, 3)$ und $(3, 1)$ bei einer 3×3–Tafel) werden in beiden Fällen mit 0 gewichtet, die Zellen auf der Diagonalen mit Gewicht 1

Gewichte w_{ij} und w_{ij}^* im 4×4-Fall

Tabelle 8: Gewichte w_{ij} einer 4×4–Tafel

1.0	0.89	0.56	0.0
0.89	1.0	0.89	0.56
0.56	0.89	1.0	0.89
0.0	0.56	0.89	1.0

Tabelle 9: Gewichte w_{ij}^* einer 4×4–Tafel

1.0	0.67	0.33	0.0
0.67	1.0	0.67	0.33
0.33	0.67	1.0	0.67
0.0	0.33	0.67	1.0

Medizinisches Beispiel (Fortsetzung): Arzt A und B

Verwendung der Gewichte gemäß Formel (1.3):

$$
\begin{aligned}
nf_o &= 8 \cdot 1.0 + 20 \cdot 0.75 + 3 \cdot 0.0 \\
&\quad + 2 \cdot 0.75 + 15 \cdot 1.0 + 1 \cdot 0.75 \\
&\quad + 0 \cdot 0.0 + 5 \cdot 0.75 + 2 \cdot 1.0 \\
\text{und damit } f_o &= \frac{46}{56} = 0.8214, \\
\text{sowie } n^2 f_e &= 31 \cdot 10 \cdot 1.0 + 31 \cdot 40 \cdot 0.75 + 31 \cdot 6 \cdot 0.0 \\
&\quad + 18 \cdot 10 \cdot 0.75 + 18 \cdot 40 \cdot 1.0 + 18 \cdot 6 \cdot 0.75 \\
&\quad + 7 \cdot 10 \cdot 0.0 + 7 \cdot 40 \cdot 0.75 + 7 \cdot 6 \cdot 1.0 \\
\text{also } f_e &= \frac{2428}{56^2} = 0.7742 \\
\text{und } \kappa_w &= \frac{0.8214 - 0.7742}{1 - 0.7742} = 0.21
\end{aligned}
$$

Medizinisches Beispiel (Fortsetzung): Arzt A und B

Verwendung der Gewichte gemäß Formel (1.4):

$$
\begin{aligned}
nf_o &= 8 \cdot 1.0 + 20 \cdot 0.5 + 3 \cdot 0.0 \\
&\quad + 2 \cdot 0.5 + 15 \cdot 1.0 + 1 \cdot 0.5 \\
&\quad + 0 \cdot 0.0 + 5 \cdot 0.5 + 2 \cdot 1.0 \\
\text{und damit } f_o &= \frac{39}{56} = 0.6964, \\
\text{sowie } n^2 f_e &= 31 \cdot 10 \cdot 1.0 + 31 \cdot 40 \cdot 0.5 + 31 \cdot 6 \cdot 0.0 \\
&\quad + 18 \cdot 10 \cdot 0.5 + 18 \cdot 40 \cdot 1.0 + 18 \cdot 6 \cdot 0.5 \\
&\quad + 7 \cdot 10 \cdot 0.0 + 7 \cdot 40 \cdot 0.5 + 7 \cdot 6 \cdot 1.0
\end{aligned}
$$

Medizinisches Beispiel (Fortsetzung), Arzt A und B

Also

$$f_e = \frac{1976}{56^2} = 0.6301$$

Damit erhalten wir

$$\kappa_{w^*} = \frac{0.6964 - 0.6301}{1 - 0.6301} = 0.18$$

Für dieses Beispiel erhalten wir also

$$\kappa = 0.16 < \kappa_{w^*} = 0.18 < \kappa_w = 0.21$$

Medizinisches Beispiel (Fortsetzung), Arzt C und D

Wir erhalten unter Verwendung der Gewichte aus (1.3)

$$\kappa_w = 0.601$$

und unter Verwendung der Gewichte aus (1.4)

$$\kappa_{w^*} = 0.525$$

Auch hier erhalten wir

$$\kappa = 0.445 < \kappa_{w^*} = 0.525 < \kappa_w = 0.601$$

Cohens Kappa:

- Nützliches Maß zur Übereinstimmung bei binären und metrischen Merkmalen
- Im ordinalen Fall sollte gewichtetes Kappa verwendet werden
- Zusätzlich sollte Kalibrierung geprüft werden

Zusammenhänge zwischen metrischen Merkmalen

Darstellung des Zusammenhangs: Korrelation und Regression

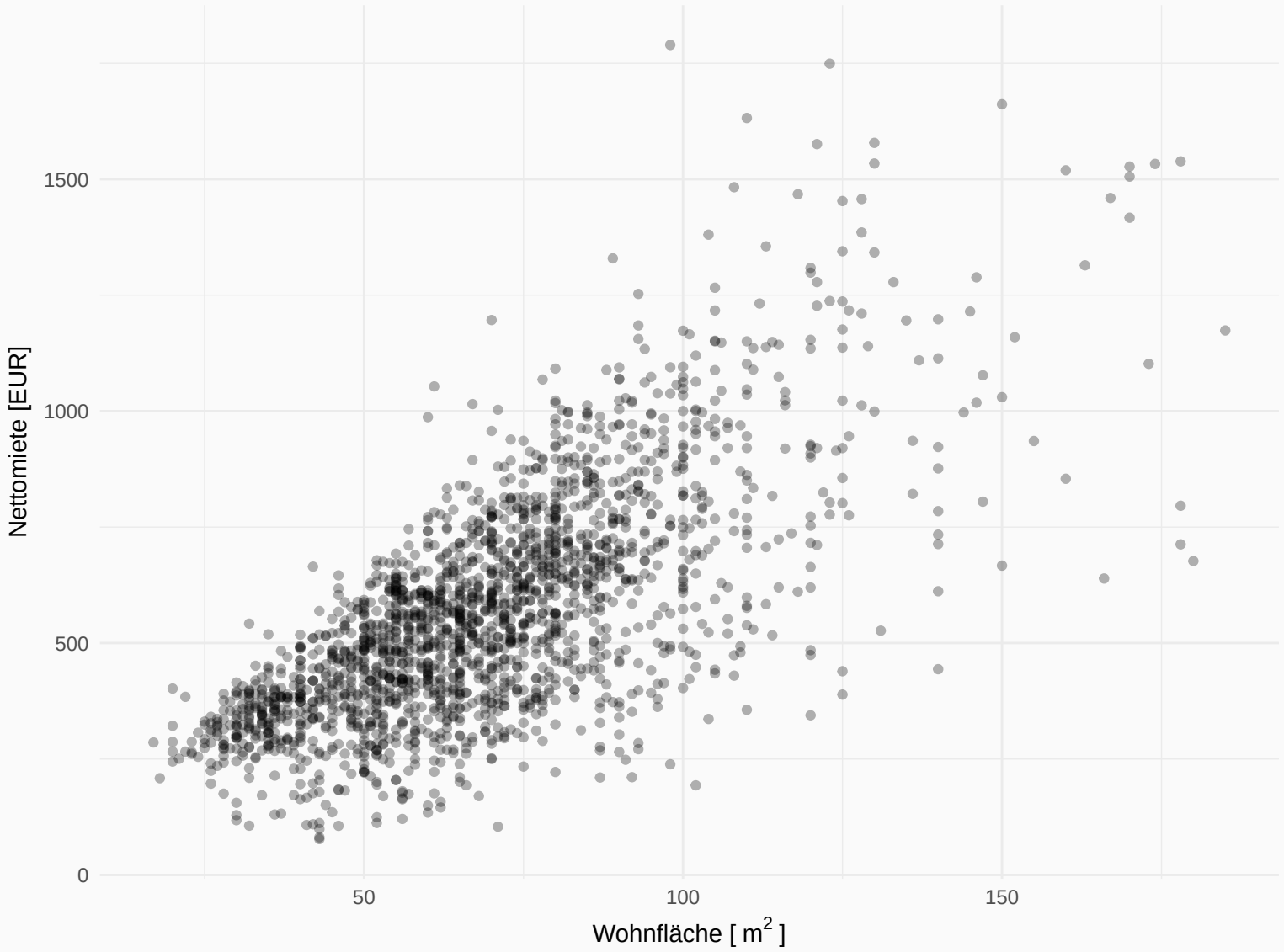
Daten liegen zu zwei metrischen Merkmalen vor: \ Datenpaare (x_i, y_i), $i = 1, \ldots, n$

Beispiel:
X: Wohnfläche $[m^2]$
Y: Nettomiete [€]

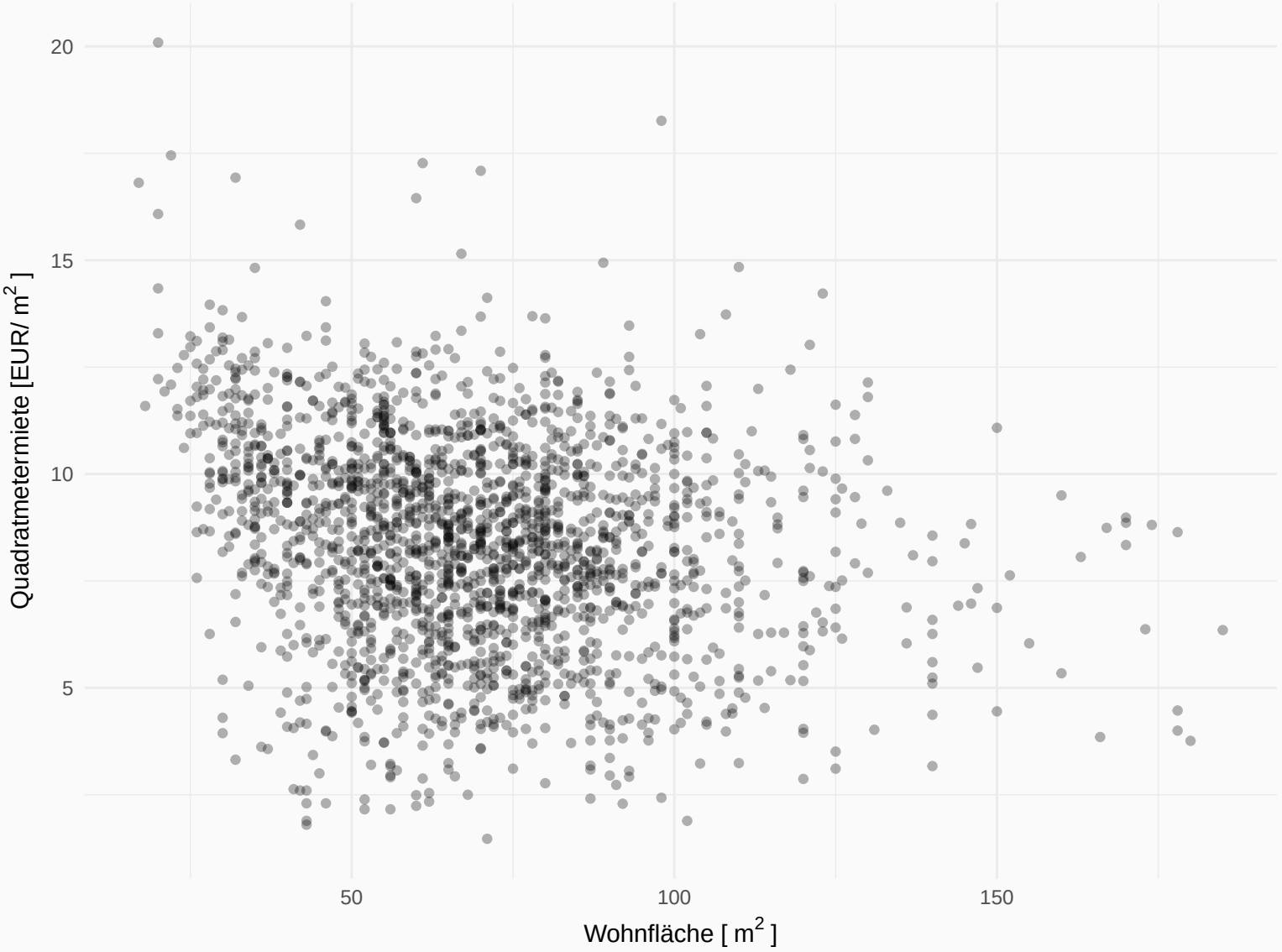
Frage:
Gibt es einen Zusammenhang zwischen diesen Merkmalen?
Wie lässt sich dieser Zusammenhang beschreiben?

Einfachste graphische Darstellung: Streudiagramm.
Die Datenpaare entsprechen Punkten in der Ebene ("Punktwolke")

Beispiel 1: Streudiagramm



Beispiel 2

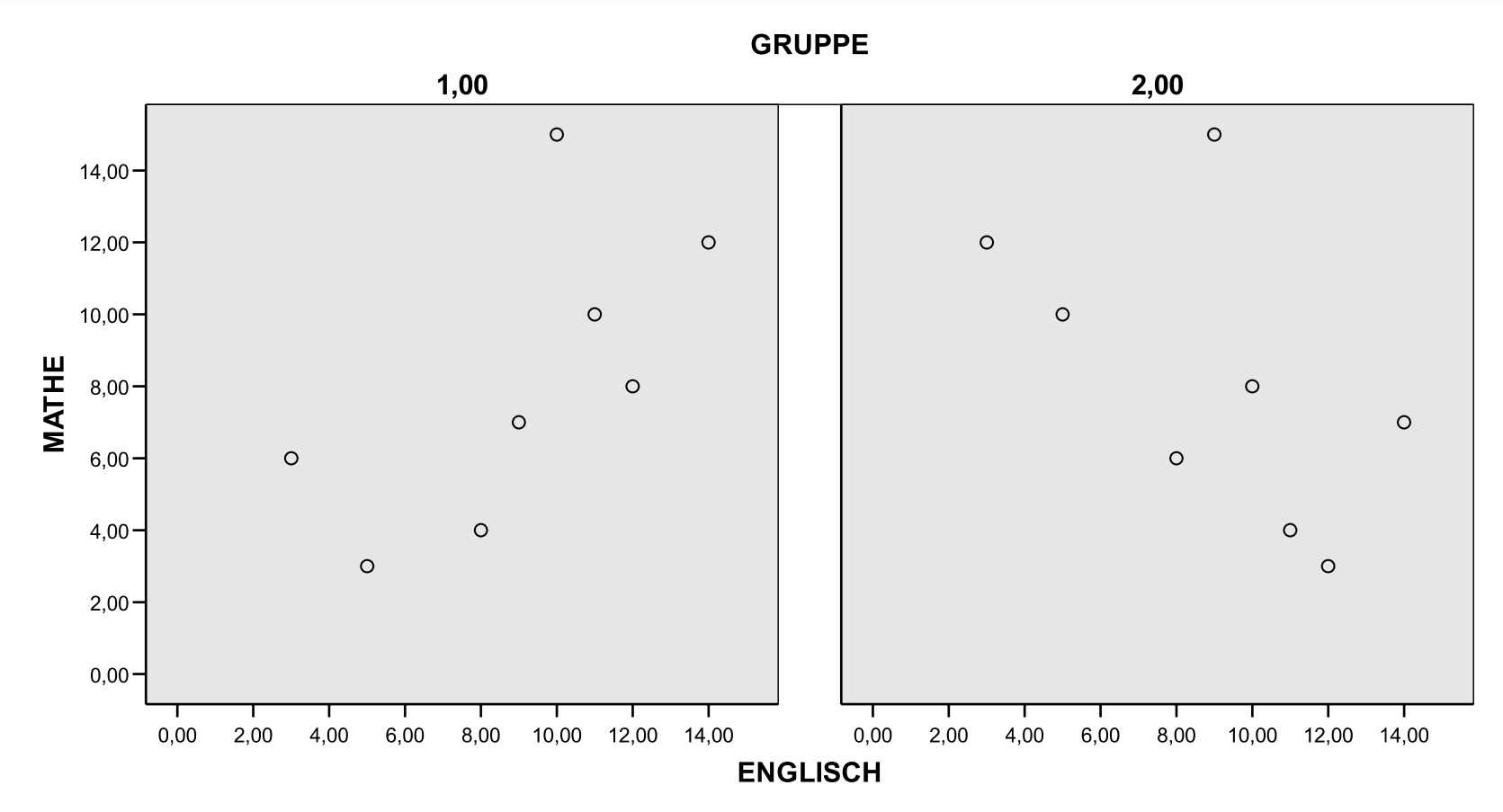


Beispiel 3

Punkte in Englisch und Mathematik

	Gruppe 1		Gruppe 2	
Schüler	Englisch	Mathe	Englisch	Mathe
1	14	12	10	8
2	9	7	8	6
3	5	3	3	12
4	3	6	5	10
5	11	10	14	7
6	8	4	9	15
7	10	15	11	4
8	12	8	12	3
Mittelwert	9.0	8.1	9.0	8.1
Standardabweichung	3.6	4.1	3.6	4.1

Beispiel 3 (Streudiagramm SPSS)



Kovarianz

Maß für den Zusammenhang der beiden Merkmale:

Daten: $(x_i, y_i), \quad i = 1, \ldots, n$

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Beachte:

- Summand i positiv, falls x_i und y_i relativ zum Mittelwert das gleiche Vorzeichen haben.
- Für S_{xx} ergibt sich die Varianz von X: $S_{xx} = S_x^2$
- Die Kovarianz hängt sowohl von der Streuung als auch von dem Zusammenhang der beiden Merkmale ab.

Bravais-Pearson-Korrelationskoeffizient

Der Bravais-Pearson-Korrelationskoeffizient r_{xy} ergibt sich aus den Daten $(x_i, y_i), i = 1, \ldots, n$ durch

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2)}} = \frac{S_{xy}}{S_x S_y}$$

Wertebereich: $-1 \leq r_{xy} \leq 1$

$r_{xy} > 0$	positive Korrelation, gleichsinniger linearer Zusammenhang, Tendenz: Werte (x_i, y_i) um eine Gerade positiver Steigung liegend
$r_{xy} < 0$	negative Korrelation, gleichsinniger linearer Zusammenhang, Tendenz: Werte (x_i, y_i) um eine Gerade negativer Steigung liegend
$r_{xy} = 0$	keine Korrelation, unkorreliert, kein linearer Zusammenhang

Punkte in Englisch und Mathematik

Gruppe 1:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{9.57}{3.641} = 0.65$$

Gruppe 2:

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{-8.29}{3.6 \cdot 4.1} = -0.56$$

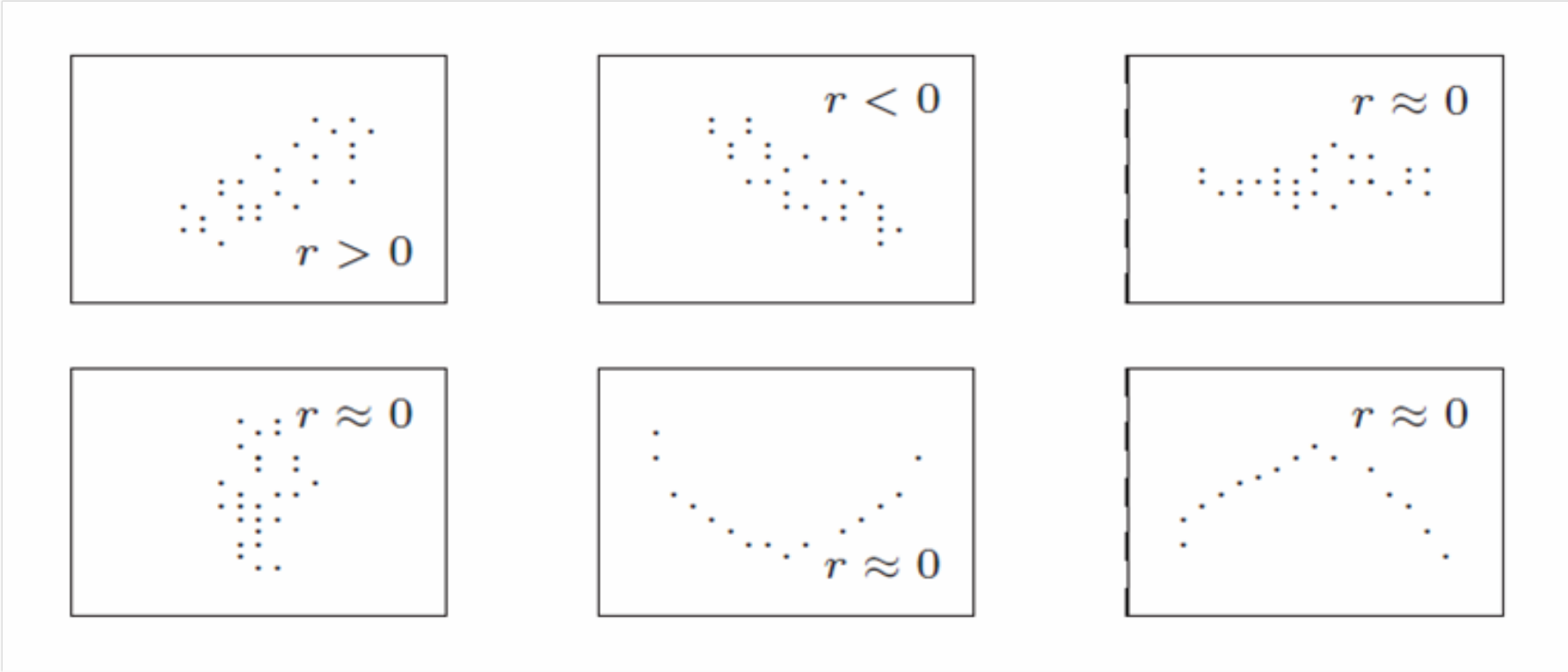
Gruppe 1: positiver linearer Zusammenhang
Gruppe 2: negativer linearer Zusammenhang

Eigenschaften des Korrelationskoeffizienten

- Maß für die Stärke des **linearen** Zusammenhangs
- Ändert sich nicht bei linearen Transformationen
- Symmetrisch (Korrelation zwischen x und y = Korrelation zwischen y und x)
- Positive Korrelation bedeutet: Je größer x, desto größer im Durchschnitt y
- Korrelation = +1 oder -1, falls die Punkte genau auf einer Geraden liegen
- Korrelation = 0 bedeutet keinen linearen Zusammenhang (aber nicht Unabhängigkeit)
- Korrelation (und Kovarianz) empfindlich gegenüber Ausreißern

Eigenschaften von r_{xy}

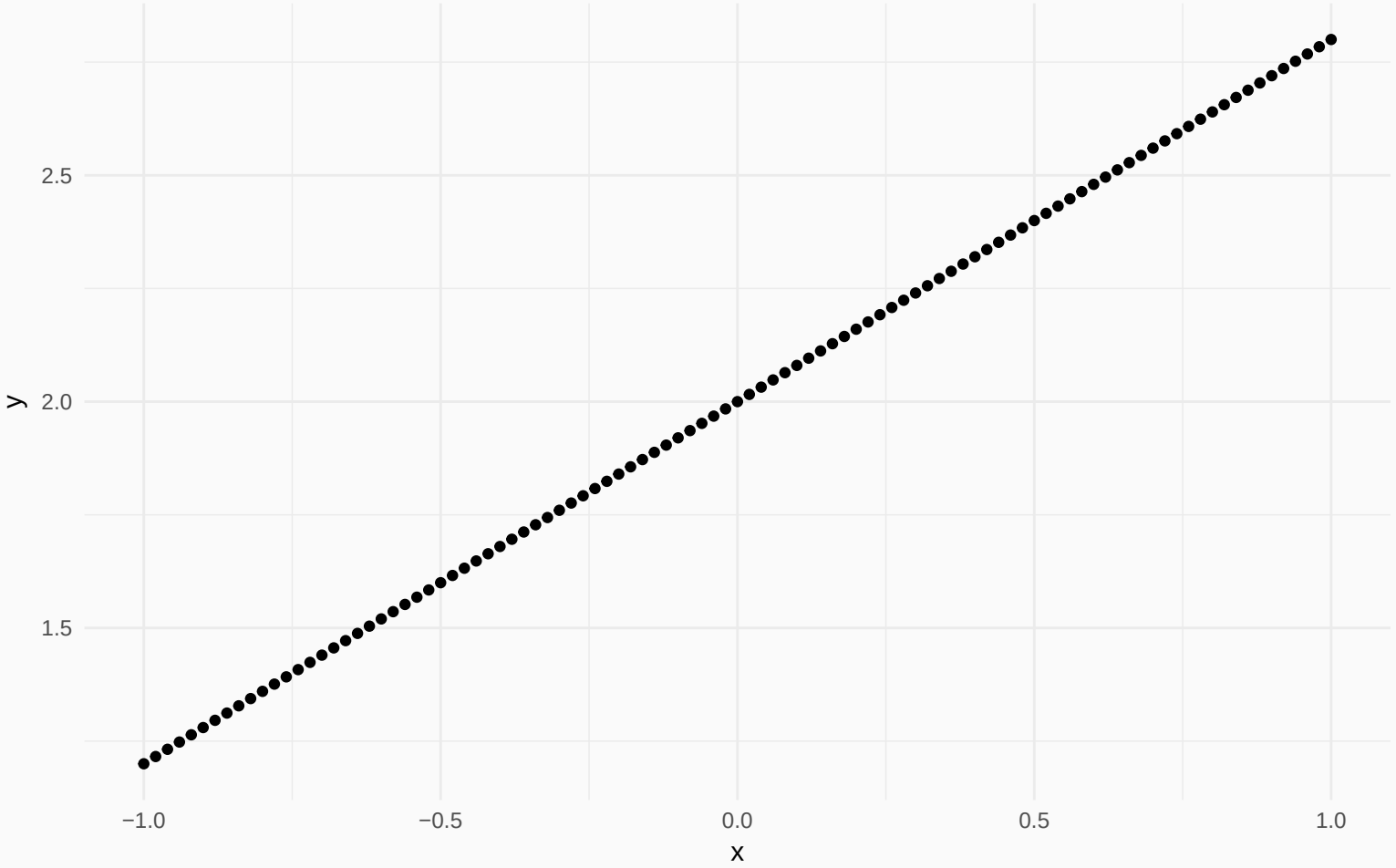
Misst nur Stärke des **linearen** Zusammenhangs:



Punktkonfigurationen und Korrelationskoeffizienten (qualitativ)

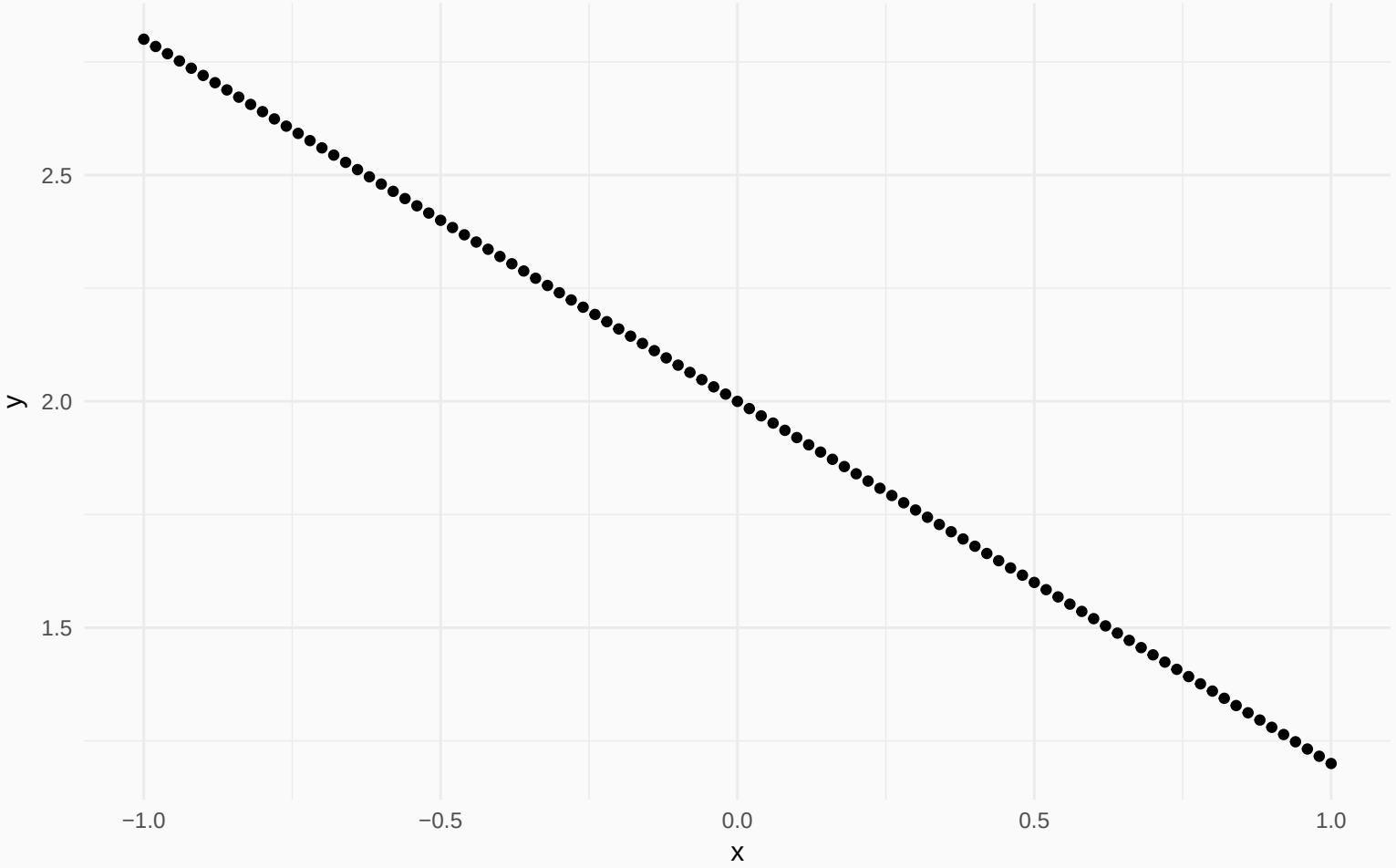
Beispiele: exakte & verrauschte Zusammenhänge

Beispiel 1: Lineare (unverrauschte) Funktion, $y = 0.8x + 2.0$, 101 equidistante Stützstellen im Intervall [-1,1], $r_{xy} = ?$



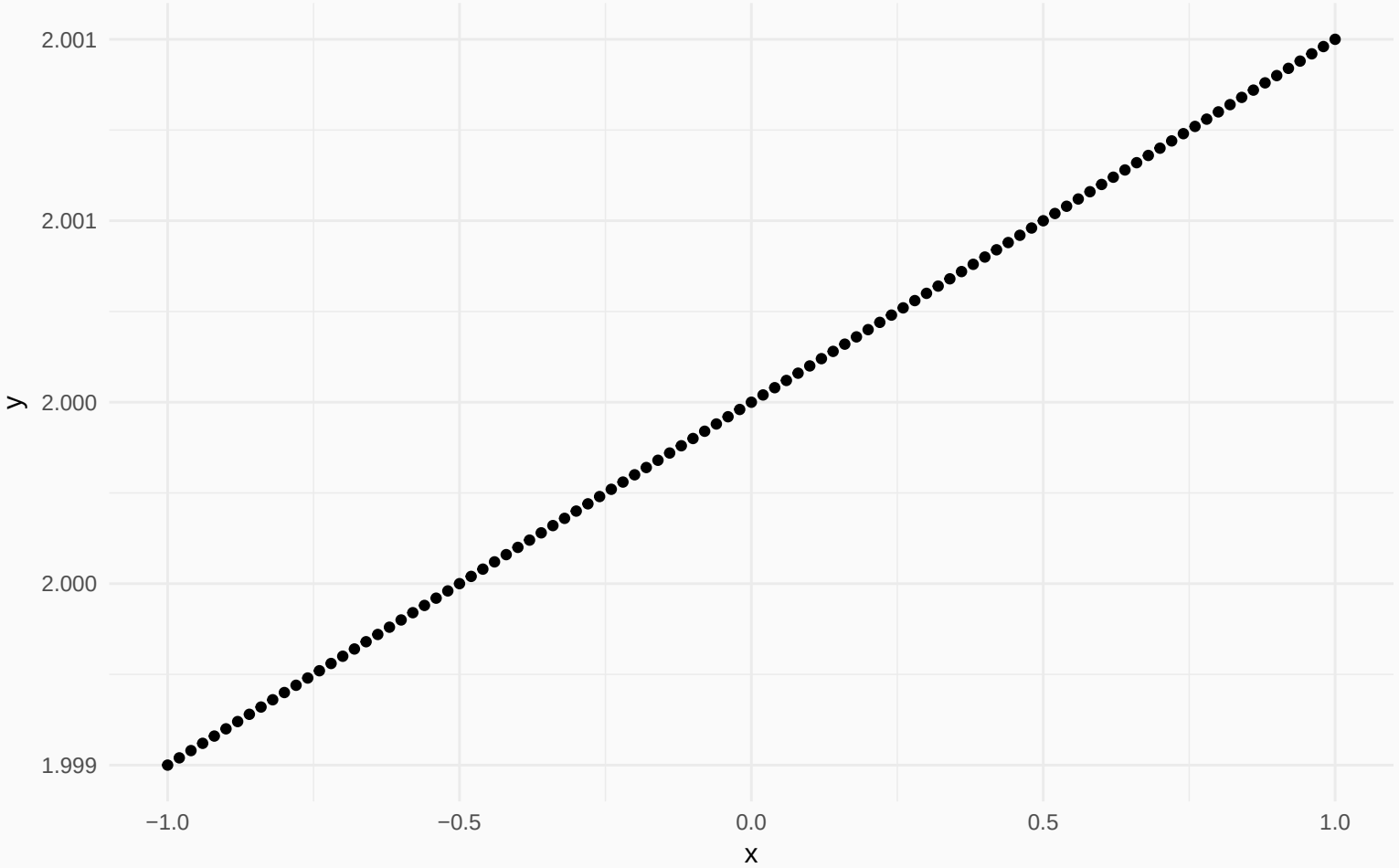
Beispiele: exakte & verrauschte Zusammenhänge

Beispiel 2: Lineare (unverrauschte) Funktion, $y = -0.8x + 2.0$, 101 equidistante Stützstellen im Intervall [-1,1], $r_{xy} = ?$



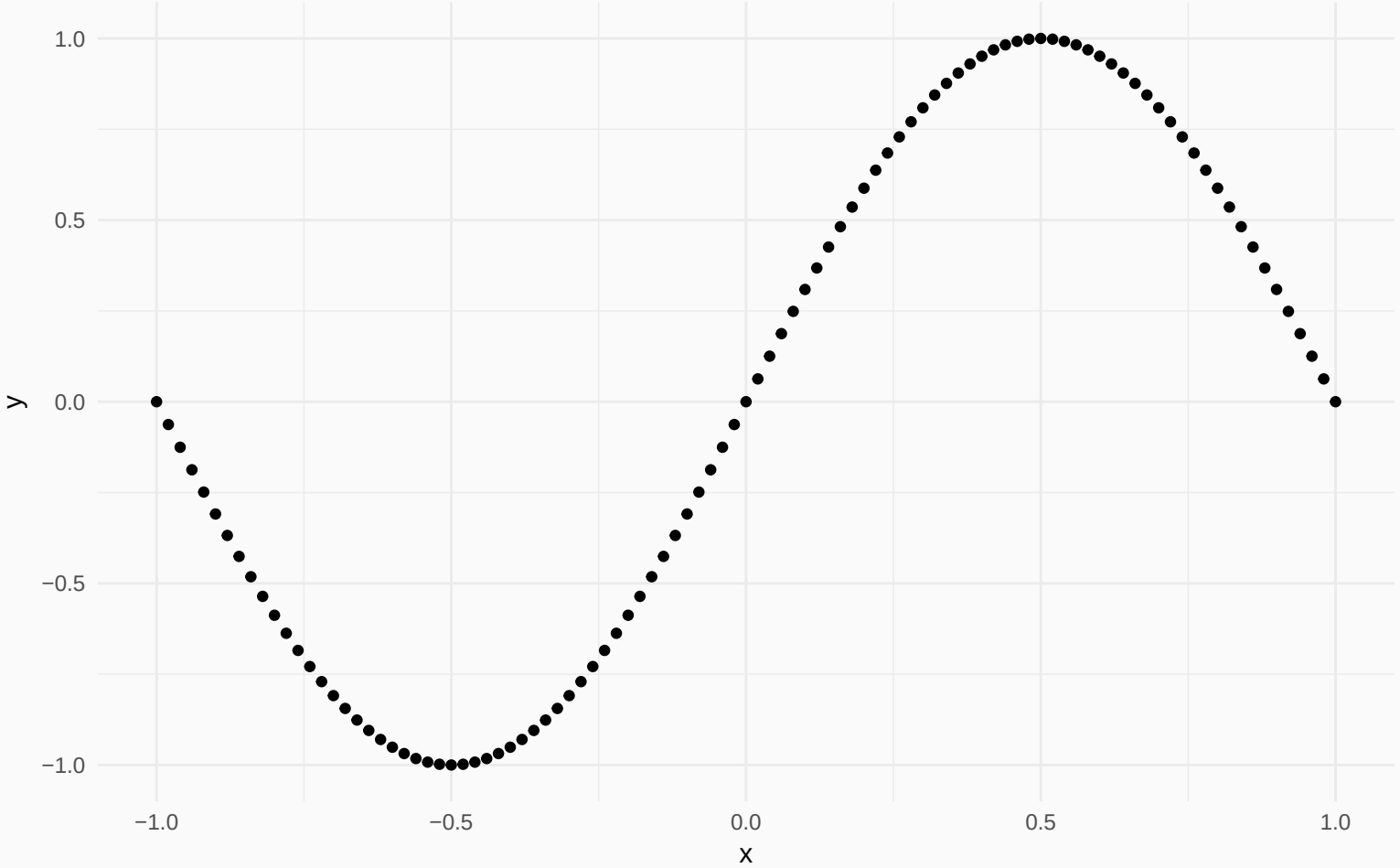
Beispiele: exakte & verrauschte Zusammenhänge

Beispiel 3: Lineare (unverrauschte) Funktion, $y = 0.001x + 2.0$, 101 equidistante Stützstellen im Intervall [-1,1], $r_{xy} = ?$



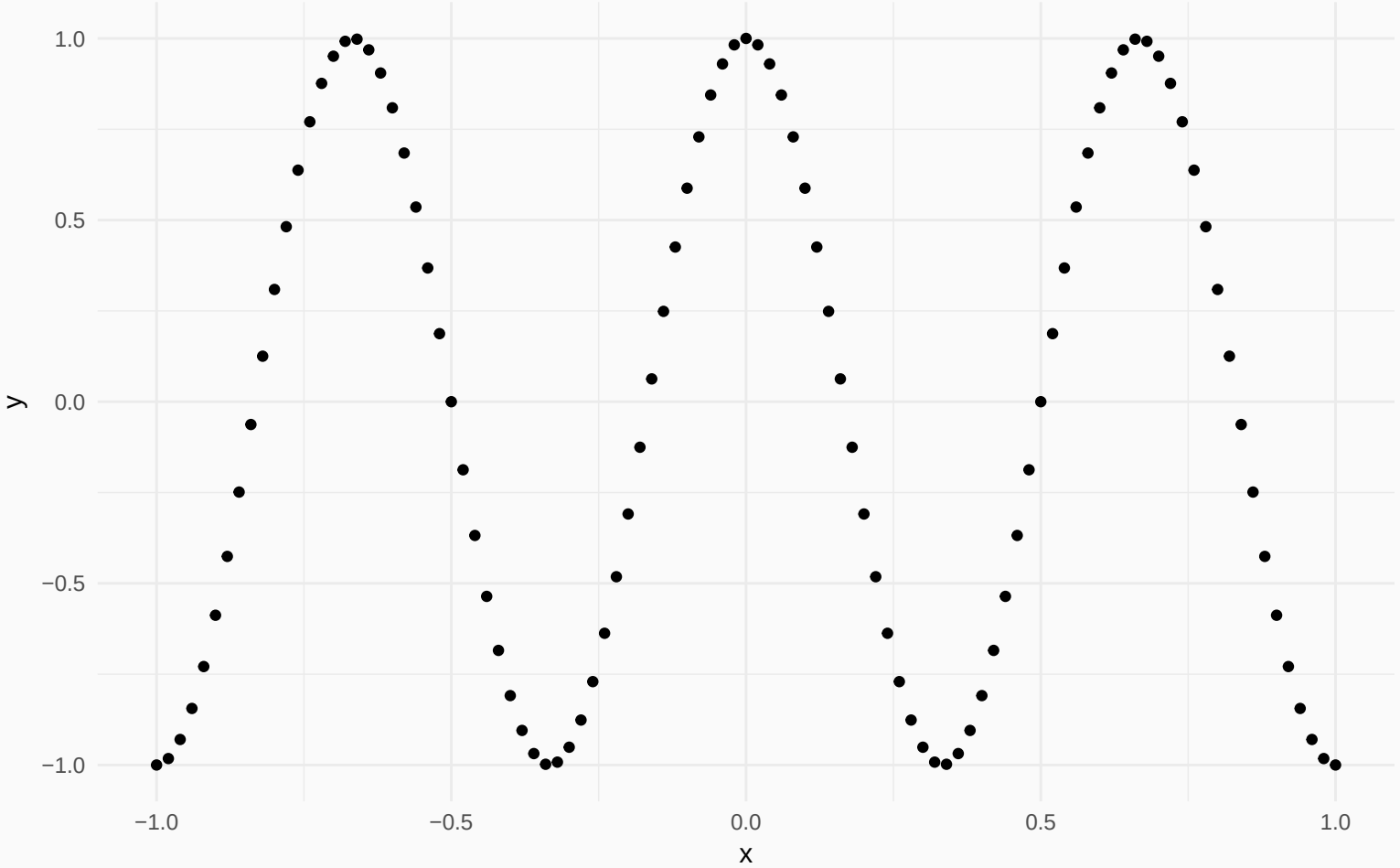
Beispiele: exakte & verrauschte Zusammenhänge

Beispiel 4: Periodische (unverrauschte) Funktion, $y = \sin(\pi x)$, 101 equidistante Stützstellen im Intervall $[-1, 1]$, $r_{xy} = ?$



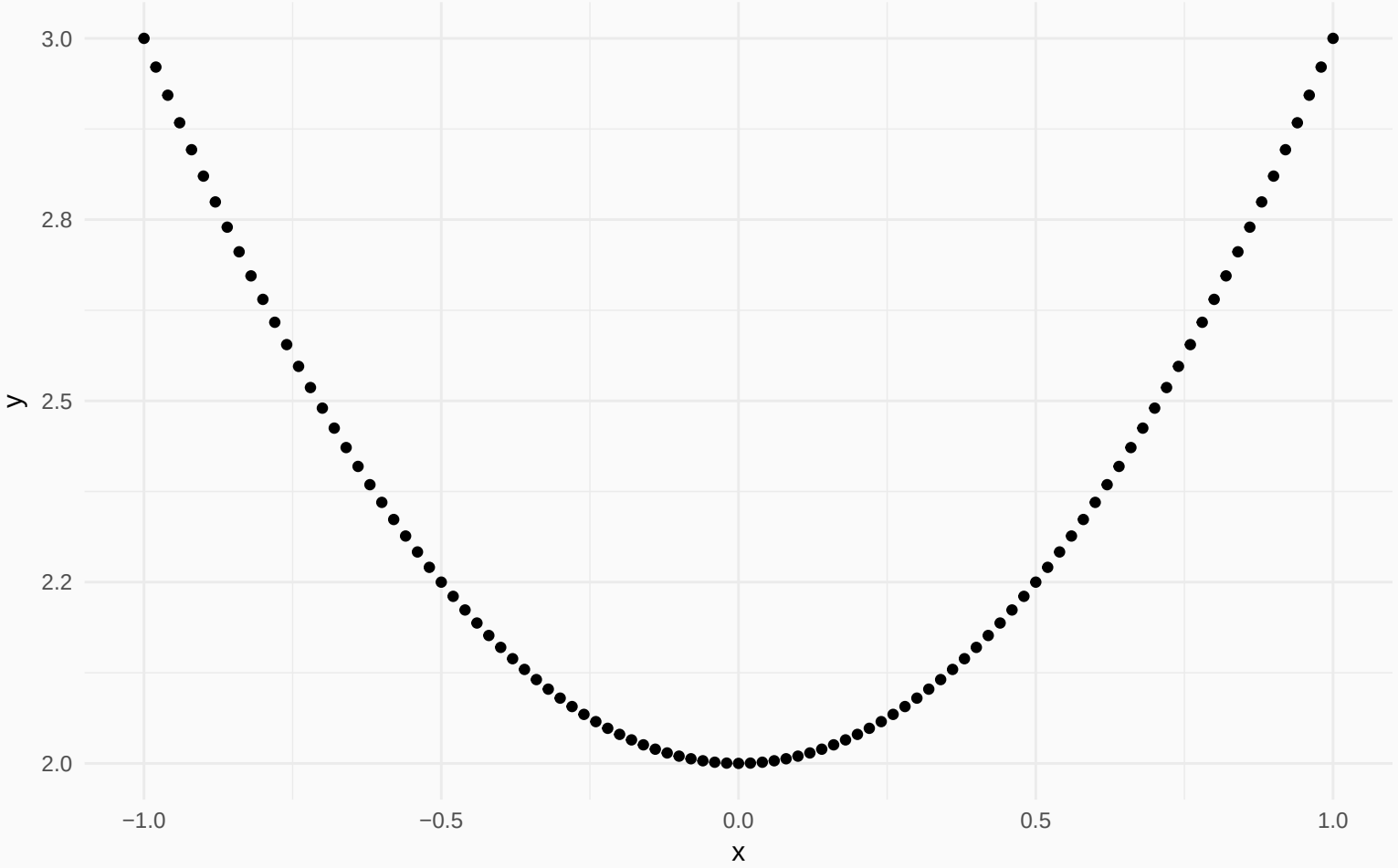
Beispiele: exakte & verrauschte Zusammenhänge

Beispiel 5: Periodische (unverrauschte) Funktion, $y = \cos(3\pi x)$, 101 equidistante Stützstellen im Intervall $[-1, 1]$, $r_{xy} = ?$



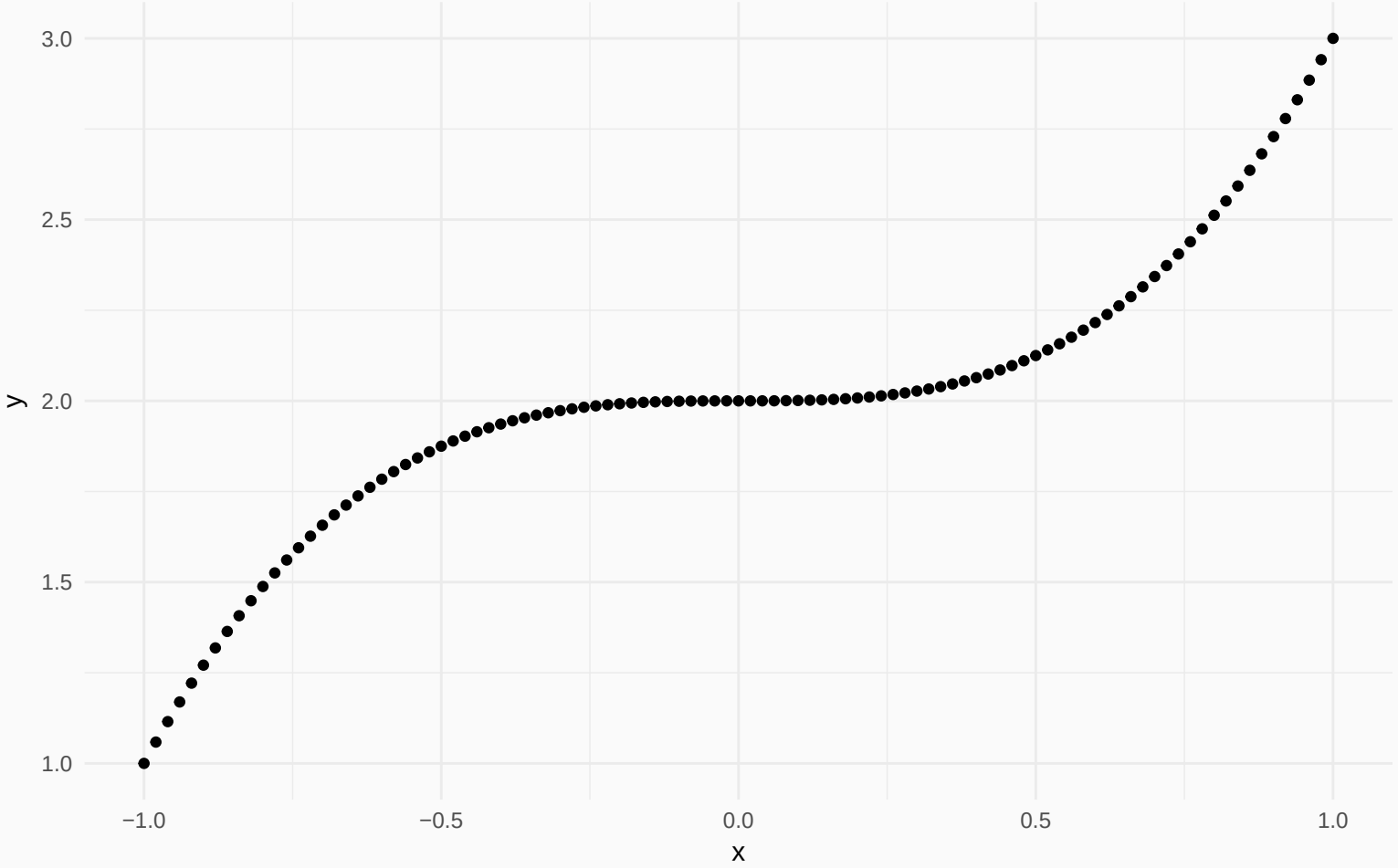
Beispiele: exakte & verrauschte Zusammenhänge

Beispiel 6: Quadratische (unverrauschte) Funktion, $y = x^2 + 2.0$, 101 equidistante Stützstellen im Intervall $[-1, 1]$, $r_{xy} = ?$



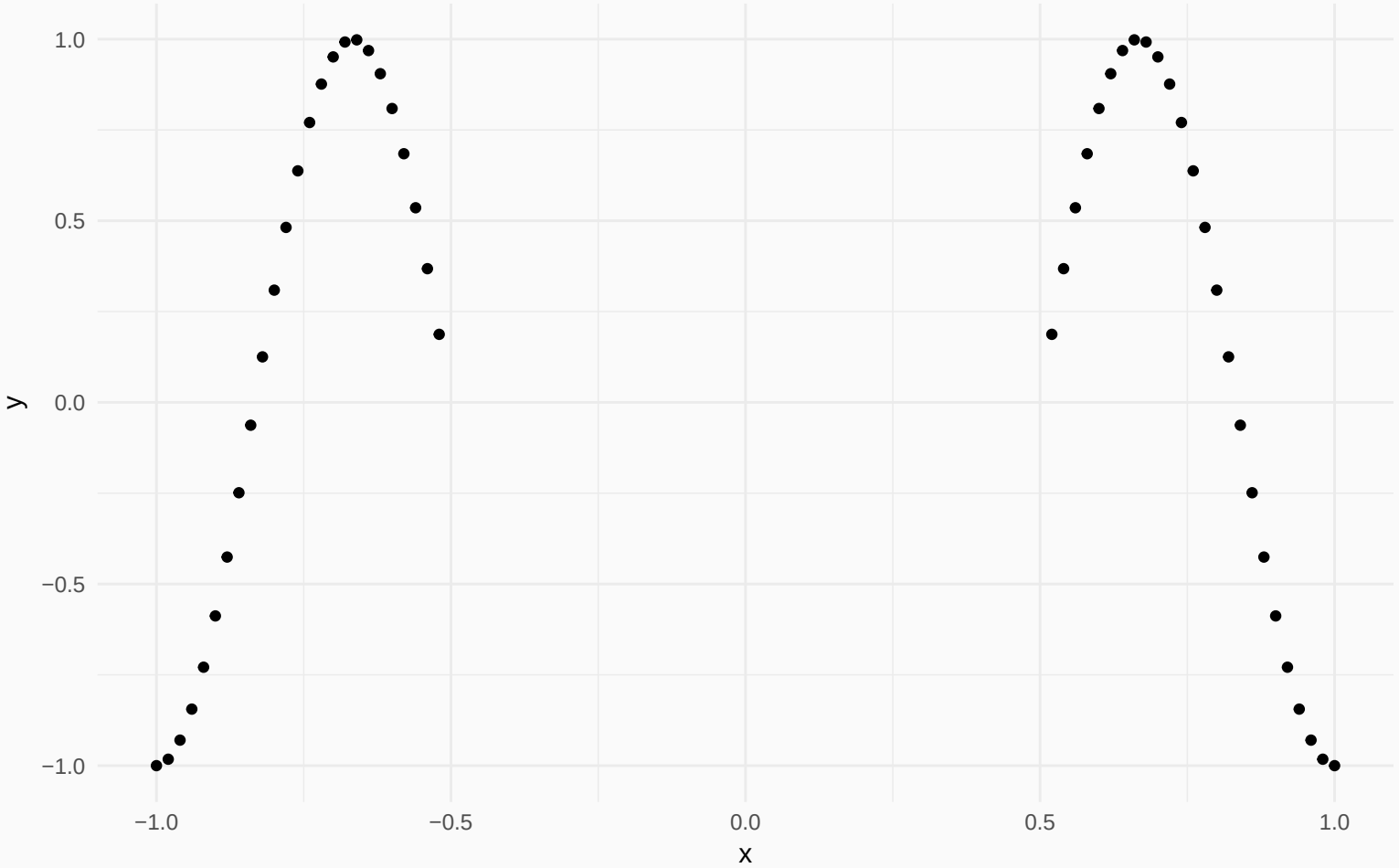
Beispiele: exakte & verrauschte Zusammenhänge

Beispiel 7: Kubische (unverrauschte) Funktion, $y = x^3 + 2.0$, 101 equidistante Stützstellen im Intervall $[-1, 1]$, $r_{xy} = ?$



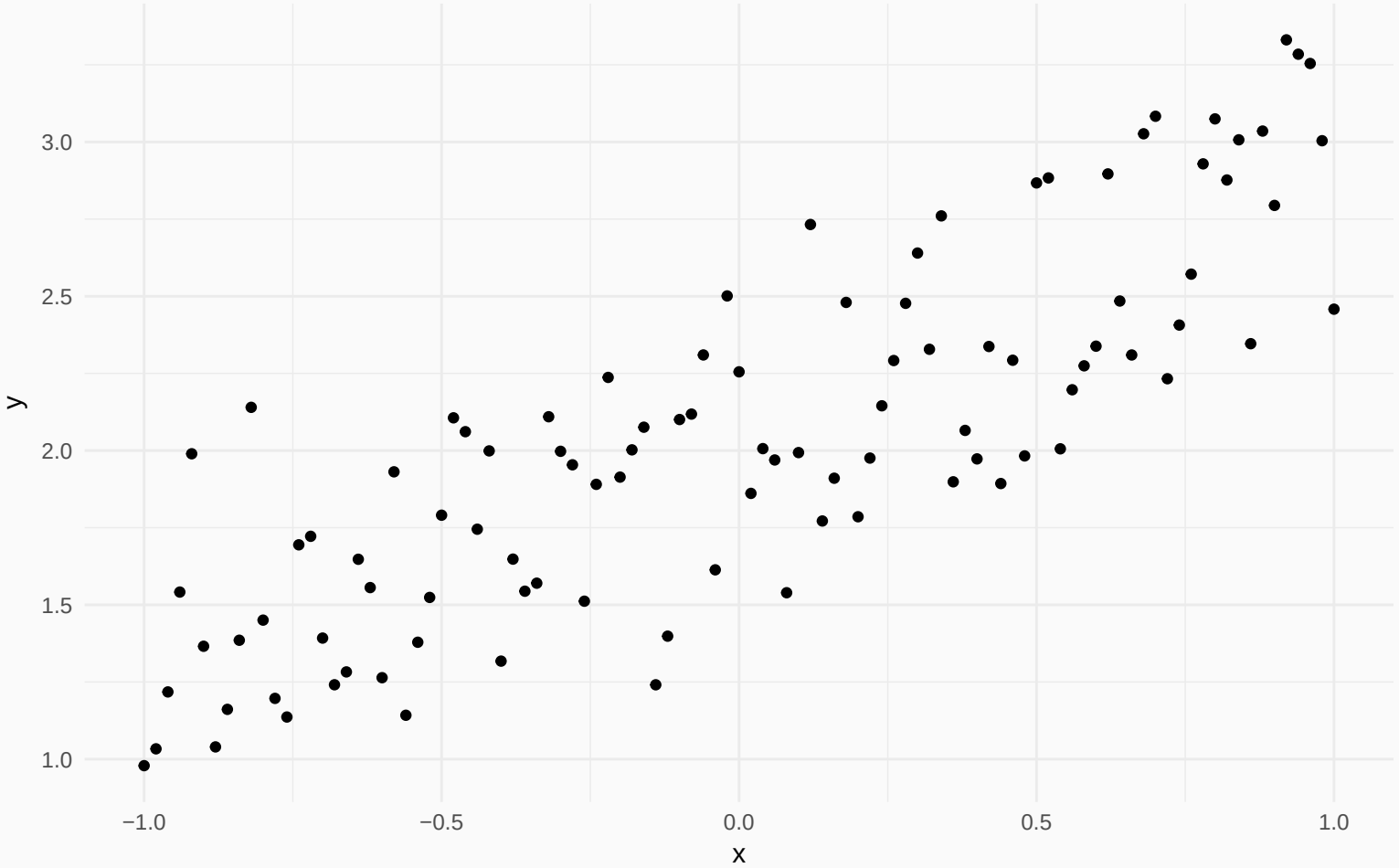
Beispiele: exakte & verrauschte Zusammenhänge

Beispiel 8: Abschnittweise definierte (unverrauschte) Funktion $y = \sin(2\pi x)$, 50 und 51 equidistante Stützstellen in den Intervallen $[-1, -0.5]$ und $[0.5, 1]$, $r_{xy} = ?$



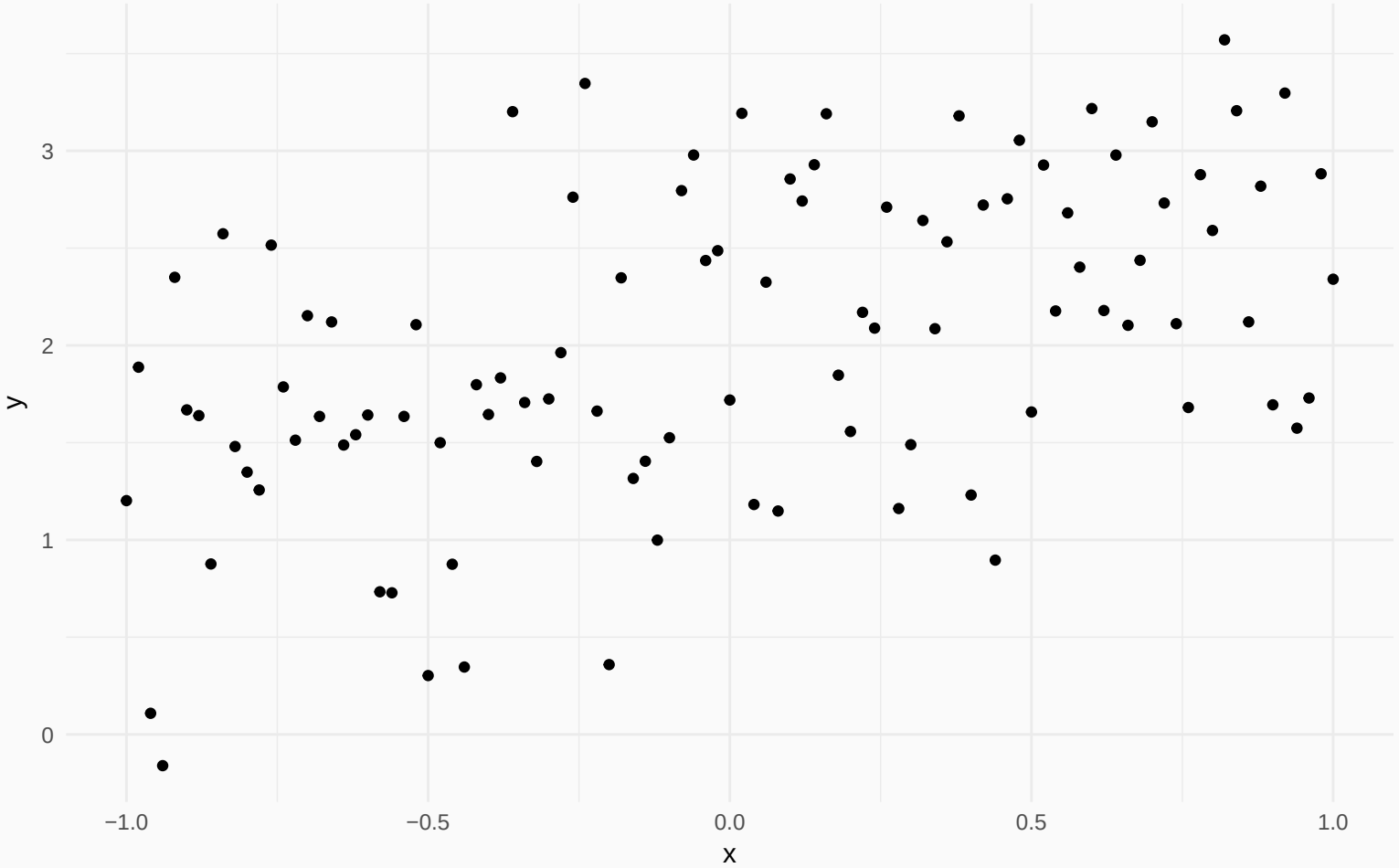
Beispiele: exakte & verrauschte Zusammenhänge

Beispiel 9: Lineare, schwach verrauschte Funktion,
$y = 0.8x + 2.0 + N(\mu = 0, \sigma^2 = 0.1)$, 101 equidistante Stützstellen im Intervall [-1,1], $r_{xy} = ?$



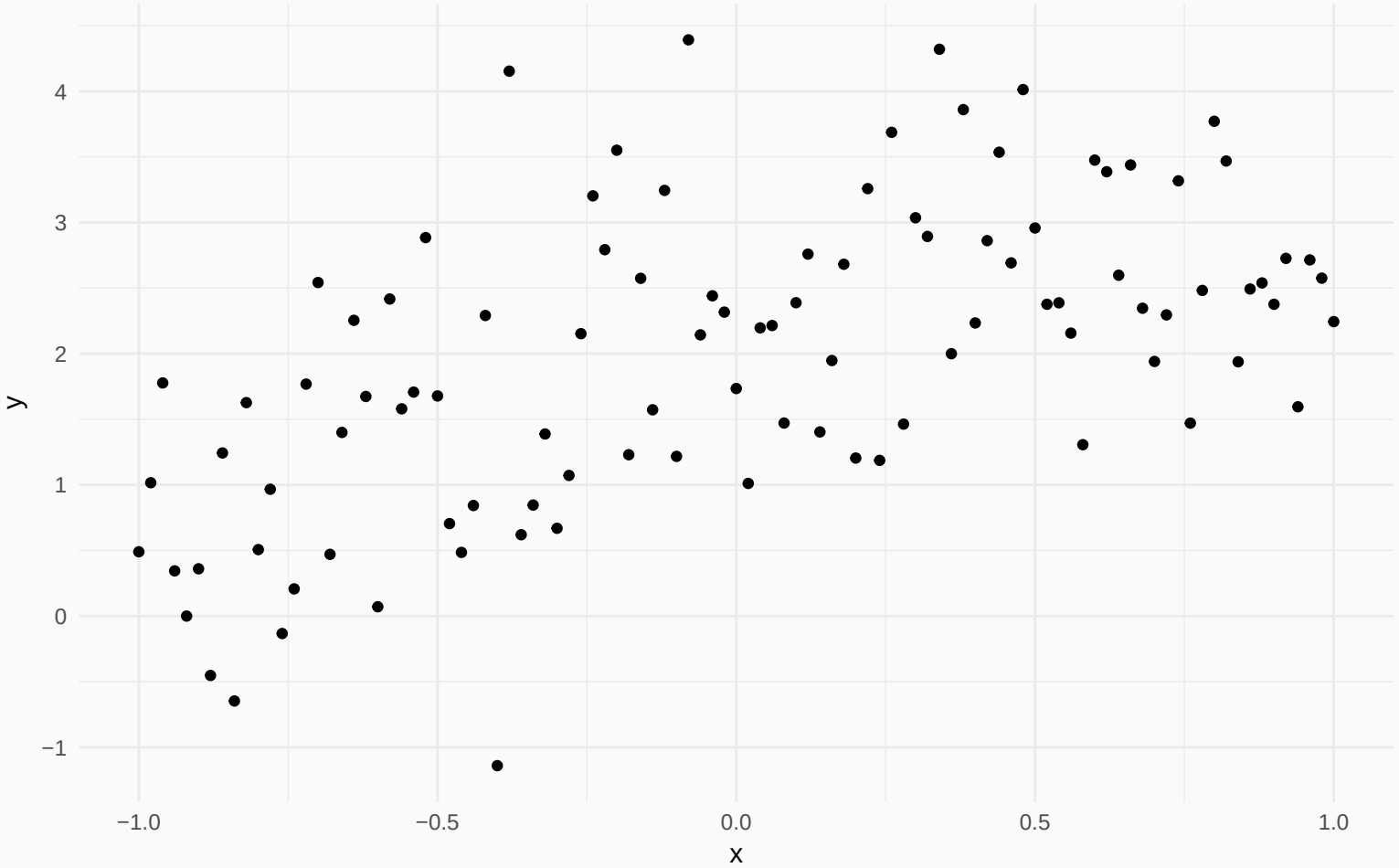
Beispiele: exakte & verrauschte Zusammenhänge

Beispiel 10: Lineare, stärker verrauschte Funktion,
$y = 0.8x + 2.0 + N(\mu = 0, \sigma^2 = 0.1)$, 101 equidistante Stützstellen im Intervall [-1,1], $r_{xy} =$?



Beispiele: exakte & verrauschte Zusammenhänge

Beispiel 10: Lineare, stark verrauschte Funktion, $y = 0.8x + 2.0 + N(\mu = 0, \sigma^2 = 1)$, 101 equidistante Stützstellen im Intervall [-1,1], $r_{xy} = ?$



Lineare Transformationen

- Bei exakten lineare Zusammenhängen gilt:

$$r_{xy} = +1 \text{ bzw. } -1 \quad \Leftrightarrow \quad Y = aX + b \text{ mit } b > 0 \text{ bzw. } b < 0$$

- Lineare Transformationen
  $\tilde{X} = a_X X + b_X, \tilde{Y} = a_Y Y + b_Y, a_X, a_Y \neq 0$:
  r_{xy} Korrelationskoeffizient zwischen X und Y
  $\tilde{r}_{xy}$ Korrelationskoeffizient zwischen $\tilde{X}$ und $\tilde{Y}$

$$\Rightarrow \begin{array}{lll} \tilde{r}_{xy} = r_{xy} & \Leftrightarrow & a_X, a_Y > 0 \text{ oder } a_X, a_Y < 0 \\ \tilde{r}_{xy} = -r_{xy} & \Leftrightarrow & a_X > 0, a_Y < 0 \text{ oder } a_X < 0, a_Y > 0. \end{array}$$

Vektor-Darstellung

Definiere die zentrierten Datenvektoren

$$\boldsymbol{x}_c = (x_1 - \bar{x}, \ldots, x_i - \bar{x}, \ldots, x_n - \bar{x})'$$
$$\boldsymbol{y}_c = (y_1 - \bar{y}, \ldots, y_i - \bar{y}, \ldots, y_n - \bar{y})'$$

$\Rightarrow r_{xy} = \frac{\boldsymbol{x}_c' \boldsymbol{y}_c}{||\boldsymbol{x}_c|| \, ||\boldsymbol{y}_c||}$, mit $||\,.\,||$ euklidische Norm.

Aus der Cauchy-Schwarz-Ungleichung folgt

$$|\boldsymbol{x}_c' \boldsymbol{y}_c| \leq ||\boldsymbol{x}_c|| \, ||\boldsymbol{y}_c||,$$

d.h. $-1 \leq r \leq +1$.

Spearman-Korrelationskoeffizient, Rang-Korrelationskoeffizient

X, Y (mindestens) ordinal

Idee: Gehe von Werten $x_i, \quad i = 1, \ldots, n$ und $y_i, \quad i = 1, \ldots, n$ über zu ihren Rängen.

$$x_{(1)} \leq \ldots x_{(i)} \ldots \leq x_{(n)}$$

$$rg(x_{(i)}) = i,$$

analog für $y_{(1)}, \ldots, y_{(n)}$.

Beispiel

x_i	2.3	7.1	1.0	2.1
$rg(x_i)$	3	4	1	2

bei Bindungen (ties):

x_i	2.3	7.1	1.0	2.1	2.3
$rg(x_i)$	3.5	5	1	2	3.5

$\Rightarrow$ Durchschnittsrang $\frac{3+4}{2} = 3.5$ vergeben.

Vorgehen

- Urliste der Größe nach durchsortieren
- $\Rightarrow$ Ranglisten $rg(x_i), rg(y_i), i = 1, \ldots, n$ vergeben (bei ties: Durchschnittsränge)

Idee: Berechne den **Korrelationskoeffizienten nach Bravais-Pearson für die Ränge** statt für die Werte der Urliste.

Definition: Spearmans Korrelationskoeffizient

Der *Korrelationskoeffizient nach Spearman* ist definiert durch

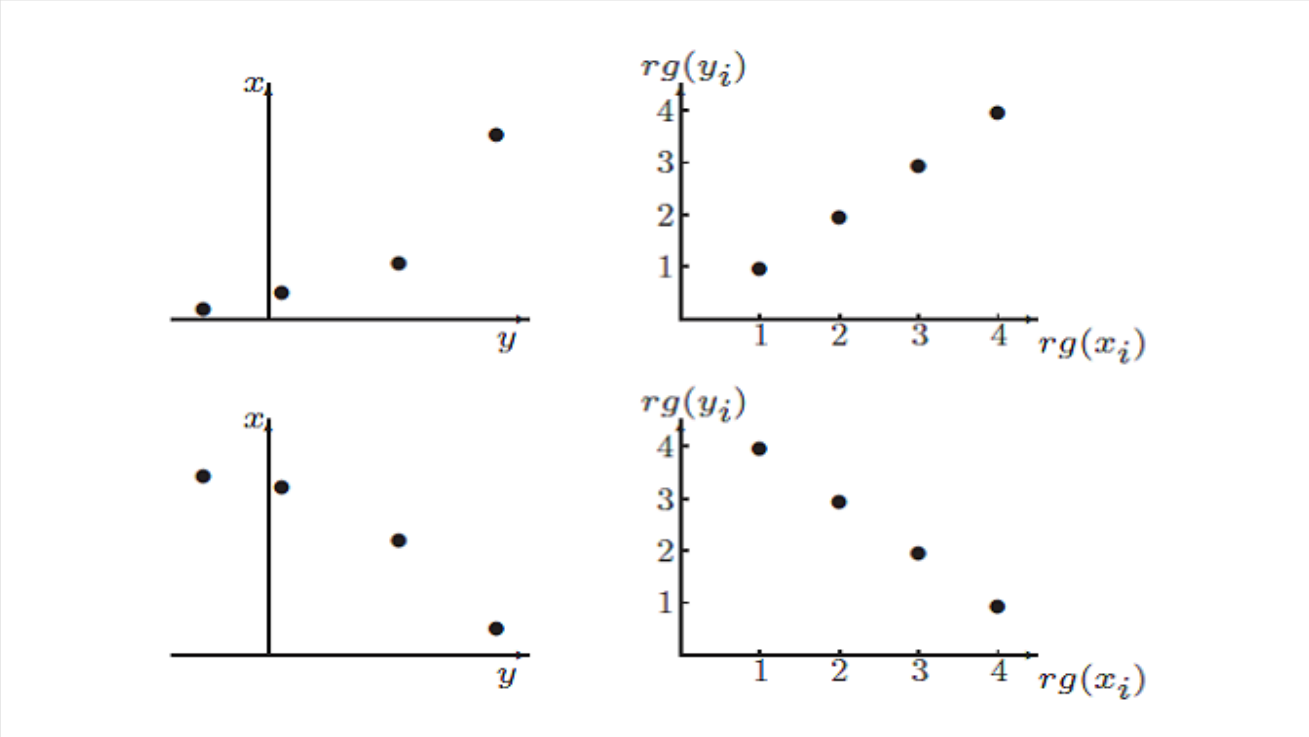
$$r_{SP} = \frac{\sum (rg(x_i) - \bar{rg}_X)(rg(y_i) - \bar{rg}_Y)}{\sqrt{\sum (rg(x_i) - \bar{rg}_X)^2 \sum (rg(y_i) - \bar{rg}_Y)^2}}.$$

Wertebereich: $-1 \leq r_{SP} \leq 1$

Interpretation

- $r_{SP} > 0$ gleichsinniger monotoner Zusammenhang,
  Tendenz: x groß $\Leftrightarrow$ y groß, x klein $\Leftrightarrow$ y klein
- $r_{SP} < 0$ gegensinniger monotoner Zusammenhang,
  Tendenz: x groß $\Leftrightarrow$ y klein, x klein $\Leftrightarrow$ y groß
- $r_{SP} \approx 0$ kein monotoner Zusammenhang

Extremfälle



$r_{SP} = 1$ (oben) und $r_{SP} = -1$ (unten)

Spearmans Korrelationskoeffizient misst Stärke **monotoner** (auch nichtlineare) Zusammenhänge!

Bemerkungen:

- Rechentechnische Vereinfachungen:

$$\bar{rg}_X = \frac{1}{n}\sum_{i=1}^{n} rg(x_i) = \frac{1}{n}\sum_{i=1}^{n} i = (n+1)/2,$$
$$\bar{rg}_Y = \frac{1}{n}\sum_{i=1}^{n} rg(y_i) = \frac{1}{n}\sum_{i=1}^{n} i = (n+1)/2.$$

  Rechentechnisch günstige Version von r_{SP}:

Daten: $(x_i, y_i),\ i = 1, \ldots, n, \quad x_i \neq x_j,\ y_i \neq y_j$ für alle i, j

Rangdifferenzen: $d_i = rg(x_i) - rg(y_i)$

$$r_{SP} = 1 - \frac{6\sum d_i^2}{(n^2-1)n}$$

Voraussetzung: keine Bindungen

Monotone Transformationen

$\tilde{X} = g(X)$ g streng monoton, $\tilde{Y} = h(Y)$ h streng monoton

$\Rightarrow r_{SP}(\tilde{X}, \tilde{Y}) = r_{SP}(X, Y)$,

- wenn g *und* h monoton wachsend
- bzw.~g *und* h monoton fallend sind.

$r_{SP}(\tilde{X}, \tilde{Y}) = -r_{SP}(X, Y)$

- wenn g monoton wachsend *und* h monoton fallend,
- bzw. g monoton fallend *und* h monoton wachsend sind.

Kendall's Tau

Betrachte Paare von Beobachtungen (x_i, y_i) und (x_j, y_j)

Ein Paar heißt:

konkordant, falls $x_i < x_j$ und $y_i < y_j$
oder $x_i > x_j$ und $y_i > y_j$

diskordant, falls $x_i < x_j$ und $y_i > y_j$
oder $x_i > x_j$ und $y_i < y_j$

N_C: Anzahl der konkordanten Paare
N_D: Anzahl der diskordanten Paare

$$\tau_a = \frac{N_C - N_D}{n(n-1)/2}$$ Kendall's Tau

Andere Varianten

- Goodman & Kruskal γ-Koeffizient

$$\gamma = \frac{N_C - N_D}{N_C + N_D}$$

- Somers D wird typischerweise verwendet wenn Y binär ist
  T_x: Anzahl der Paare mit ungleichem y und gleichem x
  ("Ties" = Bindungen)

$$D_{xy} := \frac{N_C - N_D}{N_C + N_D + T_x} = \frac{N_C - N_D}{\text{Anzahl Paare mit ungleichem y}}$$

Kendall's τ, Spearman's r_{sp}}

Beispiel:

					τ	r_{sp}
rg X	1	2	3	4	0.33	0.6
rg Y	2	1	4	3		
rg X	1	2	3	4	0.33	0.4
rg Y	1	3	4	2		

r_{sp} bestraft Abweichung stärker als τ

Unterschiede Kendall's τ, Spearman's r_{sp}

- r_{sp} verwendet Abstände auf der Rang-Skala
- τ orientiert sich an Paarvergleichen
- τ hat theoretische Entsprechung
- τ in der Regel kleiner als r_{sp}

Dichotome und stetige Merkmale: Punktbiseriale Korrelation

Korrelations-Koeffizient zwischen dichotomen (binären) und metrischem Merkmal

$X \in \{0, 1\}$; $\quad Y$ metrisch

$$r_{XY} = \frac{\bar{Y}_1 - \bar{Y}_0}{\widetilde{S}_Y} \cdot \sqrt{\frac{n_0 n_1}{N^2}}$$

$\bar{Y}_0$ Mittelwert bei $X = 0$,
$\bar{Y}_1$ Mittelwert bei $X = 1$

Entspricht normiertem Abstand der Gruppenmittelwerte.

Dichotome und stetige Merkmale

- Beispiel Kredit Scoring: Die Kreditwürdigkeit wird mit einem Scorewert gemessen ("Schufa"-Score)
  Dieser Scorewert X soll auf seine Prognosegüte geprüft werden.
  Variable Y=1 bedeutet Eintrag nach 1.5 Jahren (Default); Y=0 kein Eintrag
- Beispiel: Blutserum Konzentration und stress-induzierte Herzinfarkte
  X: Marker für Herzinfarkt
  Y: Infarkt während der WM (Gruppenphase)

ROC-Kurve

Jetzt Y dichotome Zielgröße und X metrische Einflussgröße

$Y = 1 \longrightarrow$ Ausfall (krank)
$Y = 0 \longrightarrow$ kein Ausfall (gesund)

Prognose $\hat{y}_i$ auf Basis von diagnostischem Score x und Schwellenwert c:

$$\hat{y}_i = 1 \Leftrightarrow x_i \geq c$$

Sensitivität und Spezifität

Anteil korrekt positiver Prognosen für positive = Sensitivität

$$f(\hat{Y} = 1|Y = 1) \quad = \quad f(x \geq c|Y = 1) = S_1(c)$$

Anteil korrekt negativer Prognosen für negative = Spezifität

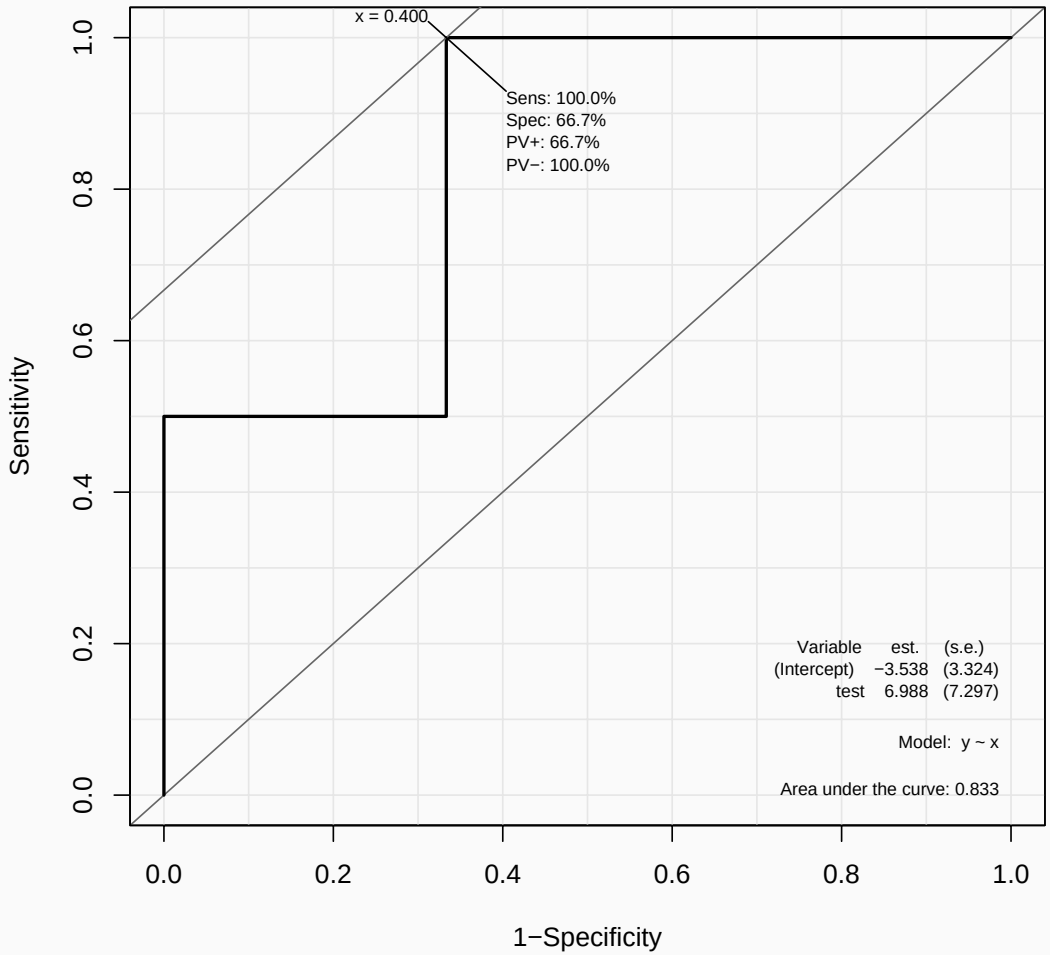
$$f(\hat{Y} = 0|Y = 0) = 1 - f(x \geq c|Y = 0) = 1 - S_0(c)$$

Falsch Positiv = 1- Spezifität

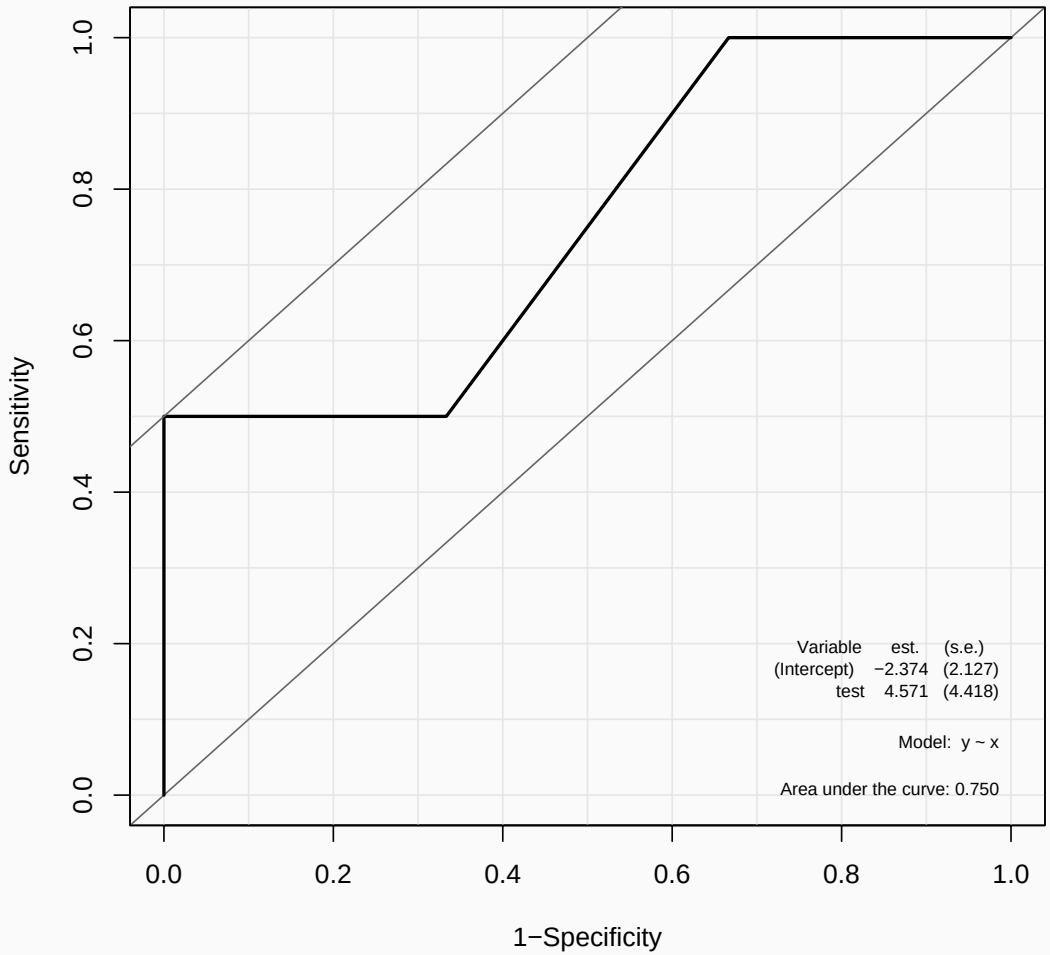
$$f(\hat{Y} = 1|Y = 0) = f(x \geq c|Y = 0) = S_0(c)$$

Die ROC-Kurve besteht aus den Punkten $(S_0(c), S_1(c))$

Beispiel für ROC-Kurve



Beispiel für ROC-Kurve mit Bindung



Maß zur Bewertung der Kurve: AUC

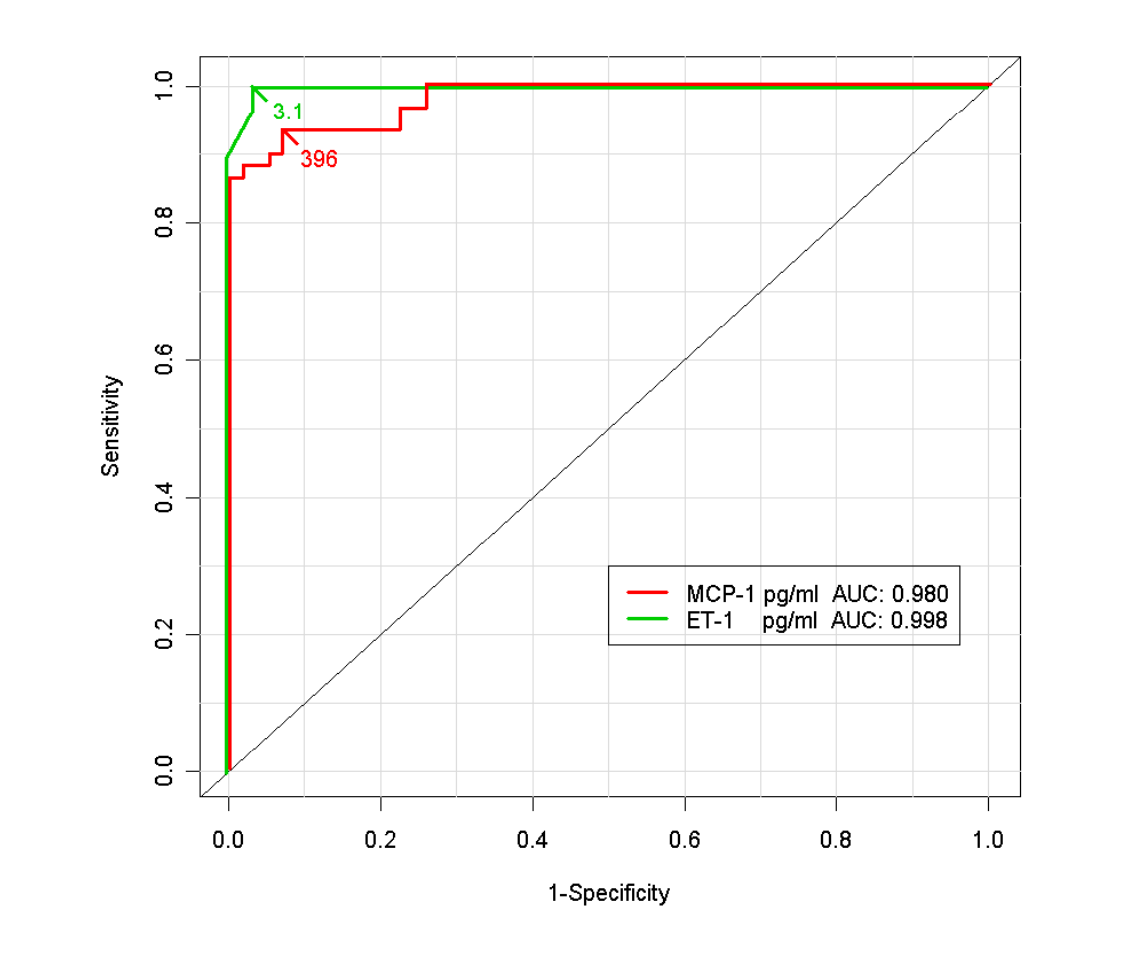
$$AUC = \int_{t=0}^{1} ROC(t)dt \tag{1.5}$$

Dies stellt die Fläche unter der Kurve dar.

Es gilt: $AUC = \frac{N_C + 0.5 N_E}{N}$

Dabei bezeichnet N_C die Anzahl der konkordanten Paare, N_E die Anzahl der identischen Paare, und N die Anzahl der Paare mit unterschiedlichem Y.

Beispiel: Stress induzierter Herzinfarkt



Korrelationsmatrix

Bei mehr als zwei Merkmalen werden die Korrelationen häufig in Form einer Matrix dargestellt.

Auf der Hauptdiagonalen stehen 1er.

Die Matrix ist symmetrisch.

$$\begin{pmatrix} 1 & r_{xy} & r_{xz} \\ r_{xy} & 1 & r_{yz} \\ r_{xz} & r_{yz} & 1 \end{pmatrix}$$