Stochastik I

Robert Stelzer[1]

Institute of Financial Mathematics
Ulm University

Sommersemester 2018

[1] basierend auf einem Foliensatz von Markus Pauly

Kapitel 8:

Deskriptive Statistik - Bivariate Kennzahlen und Visualisierung

Ziel der deskriptiven Statistik

Ziel der deskriptiven Statistik ist es, erhobene Daten durch Kenngrößen, Tabellen und Grafiken zu ordnen und übersichtlich darzustellen. Dies ist insbesondere bei sehr großen Datensätzen sinnvoll und notwendig.

In der **Datenerhebung** werden die **Ausprägungen**/**Realisierungen** von verschiedenen **Merkmalen**/**Variablen** in einer **Grundgesamtheit** (**Population**) oder – falls diese zu groß ist – einem Teil der Grundgesamtheit (**Stichprobe**) erfasst.

Bemerkung: Die deskriptive Darstellung erfolgt dabei zunächst ohne genaue Annahmen an den datengenerierenden Prozess (DGP)!

Datenerfassung

Beispiel 8.1 (Erhebung von Studierendendaten): Am Anfang einer Vorlesung werden Alter, Semesteranzahl, Geschlecht, . . . von allen Studierenden erhoben, die im Hörsaal sind.

- **Grundgesamtheit:** Studierende in der Vorlesung
- **Merkmale:** Erhobene Variablen wie das Alter, Geschlecht, Semesteranzahl etc.
- **Ausprägung:** Werte eines Merkmals bei einem/r bestimmten Studierenden/r

Bemerkung: Dies sind nur Sprechweisen! Im Rahmen der W-Theorie kann man Merkmale auch als ZVe und Ausprägungen als Realisierungen auffassen.

Urliste

ID / lfd. Nr.	M_1 Alter	M_2 Semester	M_3 Geschlecht	M_4 ...
1	19	1	w	⋮
2	21	2	m	
3	20	1	m	
⋮	↑	↑	↑	
	A u s p r ä g u n g e n			

Merkmale: $M_1, M_2, M_3, \ldots$

Merkmalstypen

Es gibt verschiedene Typen von Merkmalen:

- **Qualitative Merkmale**: Es gibt weder eine natürliche Ordnung der Merkmalsausprägungen noch ist es möglich, Abstände zu messen.
  Man spricht von einem sog. **Nominalskalenniveau**
  Beispiele: Geschlecht, Familienstand, Steuerklasse, Blutgruppe.

- **Ordinale Merkmale:** Es gibt eine natürliche Ordnung, aber Abstände lassen sich nicht wirklich interpretieren.
  Man spricht von einem sog. **Ordinalskalenniveau**
  Beispiele: Noten, Güteklassen, militärischer Rang.

- **Quantitative Merkmale:** Es gibt eine natürliche Ordnung und Abstände lassen sich interpretieren.
  Man spricht von einem sog. **Kardinalskalenniveau**
  Beispiele: Alter, Einkommen, geometrische Größen wie Längen und Flächen, physikalische Größen wie Spannung und Stromstärke, Vektoren mit entsprechenden Einträgen.

Merkmalstypen nach erlaubten Operationen

Man teilt das Kardinalskalenniveau häufig noch auf in Intervall- und Verhältnisskalen und erhält dann 4 verschiedene Merkmalstypen, die sich u.a. im Hinblick auf ihre erlaubten mathematischen Operationen unterscheiden:

1. **Nominalskalenniveau:** Erlaubt sind $=, \neq$
2. **Ordinalskalenniveau:** Erlaubt sind $=, \neq, <, >$
3. **Intervallskalenniveau:** Ordnung auf einer Dimension möglich; die Abstände zwischen den Skalenpunkten sind gleich.
   Erlaubte mathematische Operationen: $=, \neq, <, >, +, -$

   Beispiele: Zeitskala (Datum), Temperaturskalen (Celsius, Fahrenheit), IQ-Werte
4. **Verhältnisskalenniveau:** Intervallskala mit einem festen, nicht willkürlichen Nullpunkt. Verhältnisse (halb oder doppelt so viel etc.) sind sinnvoll.
   Erlaubte mathematische Operationen: $=, \neq, <, >, +, -, *, \backslash$

   Beispiele: Reaktionszeit, Lebensalter (0–150 Jahre), Fläche, Volumen

Quantitative Merkmale

Bei quantitativen Merkmalen unterscheidet man zusätzlich noch zwischen

- **diskreten Merkmalen:** Besitzen eine abzählbare Anzahl von möglichen Ausprägungen. Beispiele: Anzahl Kinder, Haushaltsgröße, Anzahl Studenten in einer Vorlesung bzw. an einer Uni.

- **stetigen Merkmalen**: Können beliebig (überabzählbar) viele Ausprägungen annehmen. Beispiele: Körpergröße, Körpergewicht, Blutdruck.

Bemerkung: Auch wenn Merkmale wie Größe, Gewicht und Alter theoretisch beliebig genau gemessen werden können, so werden sie häufig nur mit einer bestimmten Genauigkeit gemessen (Größe: cm, Gewicht: kg, Alter: Jahre), zeigen in diesem Fall also nur eine bestimmte Anzahl an Ausprägungen. Man spricht hier deshalb auch manchmal von “pseudodiskreten” Merkmalen.

Exkurs: Datenerhebung

Datenerhebung erfolgt durch Beobachtung bzw. Messung von Realisierungen interessierender Merkmale bzw. Variablen bzw. Sachverhalten. Man unterscheidet zwischen

- einer *primärstatistischen Datenerhebung*: Die Daten werden für eine vorliegende Fragestellung neu erhoben
- einer *sekundärstatistischen Erhebung*: Die Daten sind bereits vorhanden und werden nur entnommen (z.B. aus Datenbanken)
- *Vollerhebung*: Messung des interessierenden Merkmals für jedes Element der Grundgesamtheit
- *Teilerhebung*: Messung des interessierenden Merkmals für eine Teilmenge (*Stichprobe*) der Grundgesamtheit.

Exkurs: Datenerhebung

Beispiel: Schätzung des durchschnittlichen Haushaltseinkommens einer 3-köpfigen Familie in Ulm

- Kann man auf Daten des Finanzamtes zurückgreifen $\Rightarrow$ sekundärstatistisch
- Führt man eine Telefonumfrage nach dem Einkommen unter ausgewählten Haushalten durch $\Rightarrow$ primärstatistisch

Teilerhebungen weren häufig verwendet, wenn Vollerhebungen nicht möglich oder mit einem zu hohen finanziellen bzw. zeitlichen Aufwand verbunden sind. Dies tritt beispielsweise bei *Qualitätskontrollen*, Prognosen von Wahlergebnissen durch Befragung von 5000 Wahlberechtigten aber auch in obigem Einkommensbeispiel auf. Dort wird man zudem auf das Problem *fehlender Daten* (keine Antwort etc.) treffen.

Exkurs: Datenerhebung

Hat man sich für eine primärstatistische Datenerhebung entschieden, so sollte im Vorfeld im Rahmen einer statistischen Versuchsplanung entschieden werden

- *wie* man *welche* Daten/Merkmale erhebt.
- Insbesondere sollten die interessierenden Größen (sog. *Zielvariablen*) und *Einflussfaktoren*, welche die Zielgrößen beeinflussen, festgelegt werden.
- Unabdingbar ist dabei eine möglichst genaue und eindeutig formulierte Fragestellung.

Dabei gehen Fachwissen zur Fragestellung und statistische Kenntnisse Hand in Hand! Wir erläutern das Vorgehen exemplarisch...

Exkurs: Datenerhebung

Beispiel 8.2 (Einkommen und Bildung):

- Frage von Interesse: Hat der Bildungsgrad einen Einfluss auf das Einkommen in Deutschland in der Altersklasse 31-40?
- Man sollte also in jedem Fall die Zielgröße “Einkommen” und den Einflussfaktor “Bildungsgrad” erheben
- Allerdings: Obige Größen werden u.U. auch von anderen Faktoren beeinflusst wie Branche, Berufserfahrung, Geschlecht, Körpergröße, soziale Herkunft etc.
- Die Identifizierung aller “relevanten” Merkmale, die erhoben werden sollten bedarf spezifischen Fachwissens.
- Wichtig: Wurden relevante Merkmale vergessen, so kann man zu falschen Schlüssen gelangen (Beispiel: Babys und Störche)
- Aus statistischer Sicht muss geklärt werden, wie viele Daten gesammelt werden müssen, um statistisch gesicherte Aussagen treffen zu können.
  Typisch dabei: Der benötigte Datenumfang ist um so größer, je mehr Faktoren erhoben werden und je mehr diese untereinander interagieren[2] (sich gegenseitig beeinflussen). Beispiel für Letzteres: Bildung und soziale Herkunft.

[2] Wesentliche Effekte können sich überlagern/ununterscheidbar sein bzw. sich abschwächen oder verstärken (Confounding, Antagonismen, Synergien).

Bivariate Daten

- In diesem Kapitel befassen wir uns zunächst mit bivariaten Daten der Form

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix},$$

  d.h. man hat bei n Beobachtungen jeweils (mind.) die zwei Merkmale X (1. Komponente) und Y (2. Komponente) beobachtet.[3]
- Die univariaten Kenngrößen lassen sich wie zuvor pro Merkmal berechnen
- Hier interessiert uns mehr die Darstellung der Abhängigkeit/Wechselwirkung zwischen den Merkmalen

[3] W-theoretisch liegen Realisierungen von n unabhängigen Kopien der ZVen $(X, Y)^\top$ vor

Bivariate Daten

- Für ein qualitatives und ein quantitatives Merkmal sind Kontingenztafeln ein einfaches Darstellungsmittel der absoluten Häufigkeiten:

Geschlecht X /Alter Y	18	19	20	...	36	Σ
weiblich	1	2	14	...	1	104
männlich	2	6	16	...	0	96
Σ	3	8	30	...	1	200

- Frage: Was ist hier qualitativ/quantitativ?
- Datenstruktur hier

$$\binom{w}{20}, \ldots, \binom{m}{27}.$$

- Solche Kontingenztabellen lassen sich allgemein für zwei Merkmale aufstellen, die höchstens diskret (endlich) ausgeprägt sind.

Allgemeine Kontingenztafeln

Definition 8.1 (Kontingenztafel):

- Merkmal X mit möglichen Ausprägungen $a_1, \ldots, a_K,\ K \in \mathbb{N}$
- Merkmal Y mit möglichen Ausprägungen $b_1, \ldots, b_J,\ J \in \mathbb{N}$
- Ist klar, was die Ausprägungen im Beispiel der letzten Folie sind?
- Bezeichne mit $n_{\ell j}$ die absolute Häufigkeit des Beobachtungspaars (a_ℓ, b_j), d.h. $n_{\ell j} = |\{1 \leq i \leq n : (x_i, y_i) = (a_\ell, b_j)\}|$.

Dann ist eine **Kontingenztafel** für die beiden Merkmale definiert als

X/Y	b_1	b_2	$\ldots$	b_J	Σ
a_1	n_{11}	n_{12}	$\ldots$	n_{1J}	$n_{1\bullet}$
a_2	n_{21}	n_{22}	$\ldots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
a_K	n_{K1}	n_{K2}	$\ldots$	n_{KJ}	$n_{K\bullet}$
Σ	$n_{\bullet 1}$	$n_{\bullet 2}$	$\ldots$	$n_{\bullet J}$	n

Die Werte $n_{\ell\bullet}$ bzw. $n_{\bullet j}$ geben dabei die ℓ-te Zeilen- bzw. j-te Spaltensumme an

Allgemeine Kontingenztafeln – Bemerkungen

- Ist im Spezialfall $J = K = 2$, so spricht man auch von einer **Vierfeldertafel**
- Die Merkmale dürfen auch beide qualitativ (Geschlecht und Raucherstatus) oder quantitativ sein.
- Erhält man die Ausprägungen der Merkmale durch **Kategorisierung** von stetigen Merkmalen, so müssen die Kategorien disjunkt sein. Dies lässt sich immer überprüfen, da die Gesamtsumme n ergeben muss.
- Das obige Problem erhält man aber auch schon bei qualitativen und ordinalen Merkmalen: So sind die Ausprägungen “hat mittlere Reife” und “hat einen Bachelor in Mathematik” des Merkmals Bildungsstand zunächst nicht disjunkt. Lösung: Man verwendet den “höchsten Bildungsstand/Abschluss”
- Kontingenztafeln beschreiben nur das Zusammenwirken von 2 Merkmalen! Sind weitere Merkmale relevant für die Ausprägungen, kann dies zu Fehlinterpretationen führen. Stichwort: **Simpson Paradoxon**!
- Anstelle der absoluten Häufigkeiten sind auch Kontingenztafeln mit relativen Häufigkeiten $r_{\ell j} = n_{\ell j}/n$ geläufig. Als Gesamtsumme erhält man hier unten rechts die $1 = 100\%$.
- Bei Vorliegen von 2 Merkmalen möchte man häufig das eine durch das andere erklären. In so einem Fall arbeitet man auch gerne mit sog. bedingten relativen Häufigkeiten und bedingten Kontingenztafeln. Wir erläutern dies zunächst an einem Beispiel...

Bedingte Kontingenztafeln

Beispiel 8.3 (Pestizide): Das US-amerikanische Landwirtschaftsministerium hat 2002 die Pestizidbelastung von Bio- und herkömmlichen Lebensmitteln untersucht. Dabei wurden stichprobenartig die folgenden absoluten Hfgkten ermittelt:[4]

	Pestizide		
Lebensmittel	enthalten	nicht enthalten	Gesamt
Bio	29	98	127
Nicht-Bio	19485	7086	26571
Gesamt	19514	7184	26698

Um dabei den Pestizidstatus (sog. **abhängige Variable**) durch den Lebensmitteltyp (sog. **erklärende Variable**) zu erklären, wurde folgende **bedingte Kontingenztafel** angegeben:

	Pestizide		
Lebensmittel	enthalten	nicht enthalten	Gesamt
Bio	0.23	0.77	1.00
Nicht-Bio	0.73	0.27	1.00

[4] Daten aus Food Additives and Contaminations 2002, 19:5, 427–446, vgl. auch Agresti und Franklin (2007, Abschnitt 3.1)

Bedingte Kontingenztafeln

Definition 8.2 (Bedingte Kontingenztafel):

- Merkmal X mit möglichen Ausprägungen $a_1, \ldots, a_K,\ K \in \mathbb{N}$
- Merkmal Y mit möglichen Ausprägungen $b_1, \ldots, b_J\ J \in \mathbb{N}$
- Bezeichne mit $n_{\ell j}$ die absolute Häufigkeit des Beobachtungspaars (a_ℓ, b_j), d.h. $n_{\ell j} = |\{1 \leq i \leq n : (x_i, y_i) = (a_\ell, b_j)\}|$.

Dann heißt $r_{b_j|a_\ell} = n_{\ell j}/n_{\ell\bullet}$ **bedingte relative Häufigkeit** von $Y = b_j$ gegeben $X = a_\ell$ und eine **bedingte Kontingenztafel** beider Merkmale ist definiert als

X/Y	b_1	b_2	$\ldots$	b_J	Σ
a_1	$n_{11}/n_{1\bullet}$	$n_{12}/n_{1\bullet}$	$\ldots$	$n_{1J}/n_{1\bullet}$	1.00
a_2	$n_{21}/n_{2\bullet}$	$n_{22}/n_{2\bullet}$	$\ldots$	$n_{2J}/n_{2\bullet}$	1.00
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
a_K	$n_{K1}/n_{K\bullet}$	$n_{K2}/n_{K\bullet}$	$\ldots$	$n_{KJ}/n_{K\bullet}$	1.00

Bedingte Kontingenztafeln und Assoziation

- Bedingte Kontingenztafeln werden zur Assoziations- bzw. Zusammenhangsanalyse (auch Abhängigkeitsanalyse) der erklärenden und abhängigen Variable verwendet.
- Im obigen Beispiel würde auf den ersten Blick einen starken Zusammenhang zwischen beiden Merkmalen vermuten.
- Um dies statistisch genauer zu beschreiben betrachten wir das Folgende

Definition 8.3 (χ^2-Assoziationsmaß): In der Situation von Definition 4.1 sei $N_{\ell j} = \frac{n_{\ell\bullet} n_{\bullet j}}{n}$. Dann definiert man das sog. **χ^2-Assoziationsmaß** zwischen X und Y als

$$\chi^2 = \sum_{\ell=1}^{K} \sum_{j=1}^{J} \frac{(n_{\ell j} - N_{\ell j})^2}{N_{\ell j}}.$$

Man kann zeigen: Sind X und Y (als ZVe) unabhängig, so konvergiert χ^2 für $n \to \infty$ in Verteilung gegen eine χ^2_ν-Verteilung mit $\nu = (J-1)(K-1)$ Freiheitsgraden. Dies führt im Rahmen der Testtheorie auf den sog. χ^2-Test und die Sprechweise: Ist $\chi^2 > \chi^2_{\nu;0.95}$, so sagt man, dass X **und** Y (zum Niveau 5% signifikant) **assoziiert** bzw. abhängig sind. Dabei bezeichnet $\chi^2_{\nu;0.95}$ das 95%-Quantil der χ^2_ν-Verteilung.

Bedingte Kontingenztafeln und Assoziation

Beispiel 8.3 (Fortsetzung):
Wie bereits nach Blick auf die bedingte Kontingenztafel vermutet

	Pestizide		
Lebensmittel	enthalten	nicht enthalten	Gesamt
Bio	0.23	0.77	1.00
Nicht-Bio	0.73	0.27	1.00

sind Lebensmittel und Pestizide (signifikant) miteinander assoziiert bzw. abhängige Merkmale.

- Genauer: Man erhält hier $\chi^2 = 163.87 > 3.84 = \chi^2_{1;0.95}$.

Ein quantitatives Merkmal

- Kontingenztafeln werden typischerweise verwendet, um zwei Merkmale mit qualitativen oder ordinalen Ausprägungen gegenüberzustellen.
- Ist jedoch, wie im anfänglichen Beispiel zur Erhebung von Alter und Geschlecht von Studierenden ein Merkmal quantitativ, so eignen sich Boxplots mitunter besser zur Darstellung.
- Die Ausprägungen des nominalen Merkmals (hier Geschlecht) werden hier zur Gruppenbildung verwendet.

Zwei quantitative Merkmale – Scatterplots

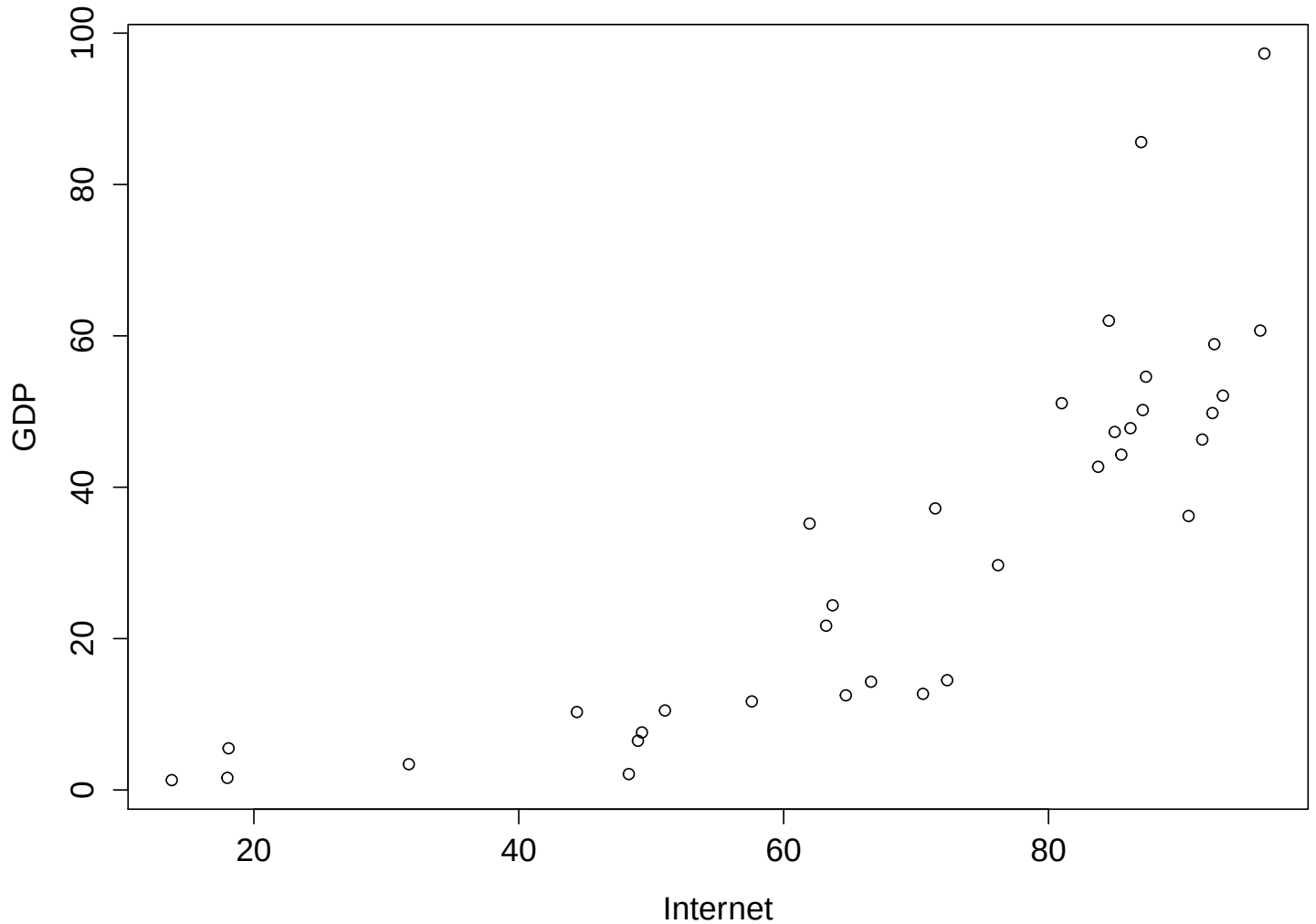
- Sind die beiden Merkmale X und Y quantitativ, so eignen sich **Scatterplots** (Streudiagramme) zur graphischen Darstellung.
- Dabei trägt man gegebene Daten

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$$

  in ein kartesisches Koordinatensystem ein und erhält eine **Punktewolke**.
- Wir betrachten beispielhaft die Darstellung der Merkmale *BIP* und *Internetreichweite* (pro 100 Personen) von 35 verschiedenen Nationen[5]

[5] Daten aus dem Jahre 2014, entnommen von databank.worldbank.org

Scatterplot: BIP ~ Internet



Spezielle Scatterplots: QQ-Plots

- Scatterplots können auch verwendet werden, um graphisch zu analysieren, ob
  (a) die Daten eines univariaten Merkmals zu einer vorgegebenen Verteilung gehören oder
  (b) ob zwei Gruppen ähnliche Verteilungen besitzen.
- Im zweiten Fall geht man zu den geordneten Statistiken der ersten Gruppe $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ und zweiten Gruppe $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ über und erzeugt ein Streudiagramm von

$$\begin{pmatrix} x_{(1)} \\ y_{(1)} \end{pmatrix}, \ldots, \begin{pmatrix} x_{(n)} \\ y_{(n)} \end{pmatrix}.$$

- Liegen diese nahe an der Winkelhalbierenden, so kann man (sehr wahrscheinlich) von ähnlichen Verteilungen ausgehen.

QQ-Plot zur Verteilungsanalyse

- Im ersten Fall a) stellt man sich also die Frage, ob ein vorliegender univariater Datensatz $x_1, \ldots, x_n$ eine gegebene Verteilungsfunktion F besitzt.
- Dies ist insbesondere von Interesse, wenn man statistische Verfahren verwendet, die auf einer bestimmten Verteilungsannahme (wie bspw. der Normalverteilung) des DGP beruhen.
- Stammen die Daten wirklich von F, so sollten die sortierten Beobachtungen $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ sich verhalten wie[6] $(F^{-1}(\frac{1}{n+1}), F^{-1}(\frac{2}{n+1}), \ldots, F^{-1}(\frac{n}{n+1}))$, wobei F^{-1} die sog. inverse Verteilungsfunktion bezeichnet. Man erstellt deshalb einen Scatterplot von

$$\begin{pmatrix} F^{-1}(\frac{1}{n+1}) \\ x_{(1)} \end{pmatrix}, \ldots, \begin{pmatrix} F^{-1}(\frac{n}{n+1}) \\ x_{(n)} \end{pmatrix} \in \mathbb{R}^2.$$

[6] Mathematische Begründung?

QQ-Plot zur Verteilungsanalyse und Transformationen

Da man häufig nur wissen möchte, ob die Daten zu einer Verteilungsfamilie (wie den Normalverteilungen) gehören, führt man QQ-Plots häufig auch mit transformierten Daten durch. Am geläufigsten sind QQ-Plots mit zentrierten oder standardisierten Daten.

Definition 8.4 (Standardisierte Beobachtungen): Für einen gegebenen Datensatz $x_1, \ldots, x_n$ eines quantitativen Merkmals mit arithmetischem Mittel (AM) $\overline{x}_n$ und empirischer Streuung (ES) $s_x > 0$ heißen

(i) $(x_i - \overline{x}_n), 1 \leq i \leq n$, die zentrierten Beobachtungen und

(ii) $\frac{(x_i - \overline{x}_n)}{s_x}, 1 \leq i \leq n$, die standardisierten Beobachtungen.

BEM 1. Die zentrierten Daten haben ein AM von 0 und die standardisierten Daten zusätzlich eine ES von 1.

QQ-Plot zur Verteilungsanalyse und Transformationen

BEM 2. QQ-Plots mit standardisierten Beobachtungen werden häufig verwendet, um mit der standardisierten[7] Verteilung zu vergleichen.
3. Der Wert

$$z_{x_i} = \frac{(x_i - \overline{x}_n)}{s_x}$$

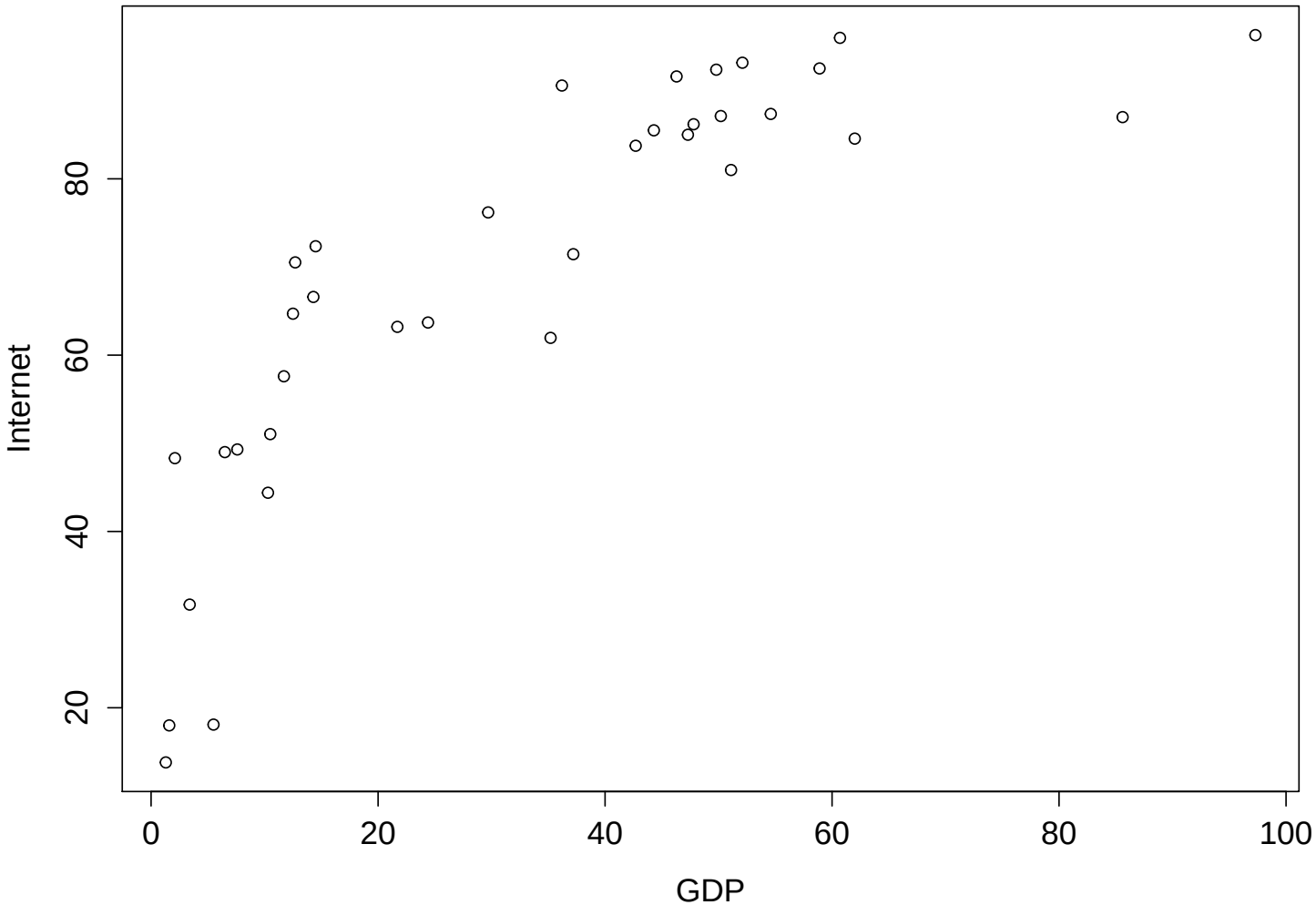
heißt auch *z*-score von x_i.

[7] mit Erwartungswert 0 und Varianz 1; typisch $N(0, 1)$

Scatterplots

- Intuitiv kann man an einem Scatterplot eine Assoziation zwischen zwei Merkmalen ablesen: Verteilt sich die Punktewolke gleichmäßig über das Quadrat, so ist eine Assoziation eher unwahrscheinlich; andernfalls könnten Zusammenhänge bestehen.
- Um dies genauer zu untersuchen betrachten wir nochmals den Scatterplot der Merkmale BIP (X) und Internetreichweite pro 100 Personen (Y) von 35 verschiedenen Nationen.
- Ziel wäre es nun eine Assoziation zwischen den beiden quantitativen Merkmalen zu beschreiben. Da wir dabei das BIP als erklärende Variable für die Zielgröße Internet auffassen wollen, spiegeln wir den Scatterplot noch einmal, d.h. wir vertauschen die Koordinatenachsen.

Scatterplot: BIP ~ Internet



- Die zu Beginn des Kapitels vorgestellte Assoziationsanalyse mittels des χ^2-Assoziationsmaßes erfordert eine Kategorisierung der Daten, wobei Genauigkeit und detaillierte Information verloren gehen.
- Aus diesem Grund führen wir im Folgenden den **Pearson-Korrelationskoeffizienten** als Assoziationsmaß für zwei quantitative Merkmale ein.
- Zur Motivation betrachten wir zunächst die z-scores der jeweiligen Merkmalsausprägungen

$$z_{x_i} = \frac{(x_i - \overline{x}_n)}{s_x} \quad \text{und} \quad z_{y_i} = \frac{(y_i - \overline{y}_n)}{s_y}, \; 1 \leq i \leq n.$$

- Sind die beiden Komponenten der Beobachtungen $\binom{x_i}{y_i}$ jeweils unabhängig (d.h. die Merkmale X und Y nicht assoziiert), so würde man erwarten, dass das arithmetische Mittel aller z-score Produkte $z_{x_i} z_{y_i}$ ungefähr 0 ergibt. Die letzte Überlegung führt auf...

Empirische Korrelation

Definition 8.5 (Empirische Korrelation): Der **empirische (Pearson) Korrelationskoeffizient** von n Beobachtungen

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$$

des bivariaten Merkmals (X, Y) ist gegeben durch

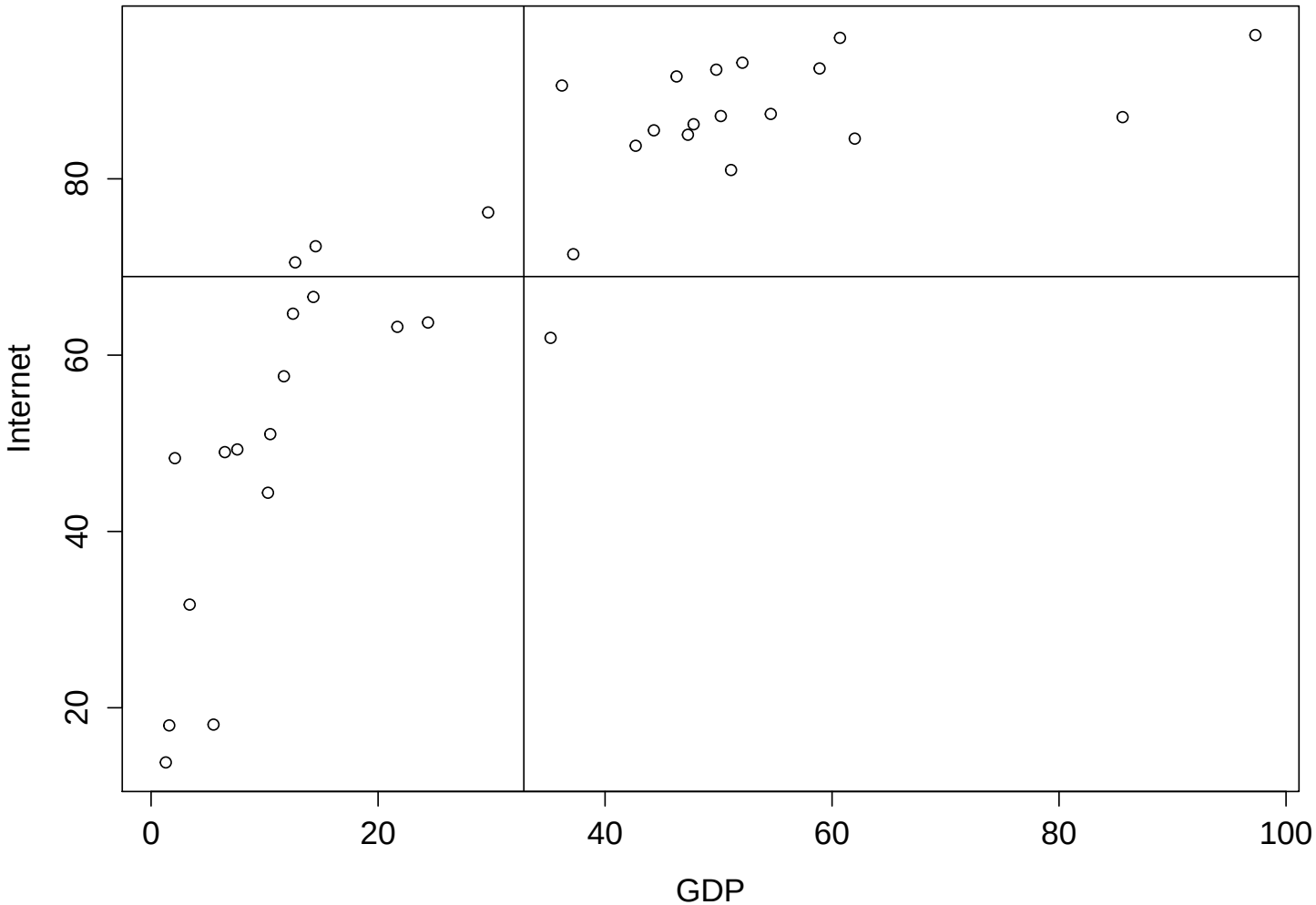
$$r_{xy} := \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} z_{y_i} = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{s_x s_y} =: \frac{s_{xy}}{s_x s_y}.$$

Der Zähler s_{xy} heißt dabei **empirische Kovarianz.**

BEM: Für den Wert r_{xy} ist es irrelevant, welches Merkmal die erklärende und welches die abhängige Variable darstellt.

- Man rechnet leicht nach, dass im Internetbeispiel gilt: $r_{xy} = 0.826$.
- Wie hat man diesen Wert zu interpretieren?
- Um dies heuristisch zu beantworten schauen wir uns nochmals genauer an, wie dieser Wert entsteht.
- Dazu zeichnen wir in den Scatterplot zusätzliche zwei Linien, die die AMe der beiden Merkmale widerspiegeln
- Hier: $\overline{x} = 32.85$ und $\overline{y} = 68.91$.
- Dies teilt den Scatterplot in vier Quadranten.

Scatterplot: BIP ~ Internet



- Die Beobachtungen im zweiten und dritten Quadranten (oben rechts und unten links) führen zu positiven Werten von $z_{x_i} z_{y_i}$ und tragen somit zu einer **positiven (empirischen) Korrelation** der beiden Merkmale bei.
- Die Beobachtungen in den beiden anderen Quadranten führen zu negativen Werten und tragen somit zu einer **negativen (empirischen) Korrelation** der beiden Merkmale bei.
- Der Wert der empirischen Korrelation setzt sich nun durch die Verteilung der beiden Fälle gewichtet mit deren Entfernung zum Punkt

$$\begin{pmatrix} \overline{x} \\ \overline{y} \end{pmatrix}$$

beider AM zusammen.
- Da im Beispiel der erste Fall überwiegt (31:4) und die Punkte im zweiten Fall auch näher an $\left(\frac{\overline{x}}{\overline{y}}\right)$ liegen, erhalten wir hier eine (stark) positive Korrelation der beiden Merkmale.

Bemerkung 8.1 (Eigenschaften und Interpretation): (i) Für den empirischen Korrelationskoeffizienten r_{xy} zweier Merkmale kann man zeigen:

- $r_{xy} \in [-1, 1]$
- $r_{xy} = 1$
  $\Leftrightarrow y_i = a + bx_i$ für alle $i = 1, \ldots, n$ und geeignete $a \in \mathbb{R}, b > 0$.
- $r_{xy} = -1$
  $\Leftrightarrow y_i = a + bx_i$ für alle $i = 1, \ldots, n$ und geeignete $a \in \mathbb{R}, b < 0$.

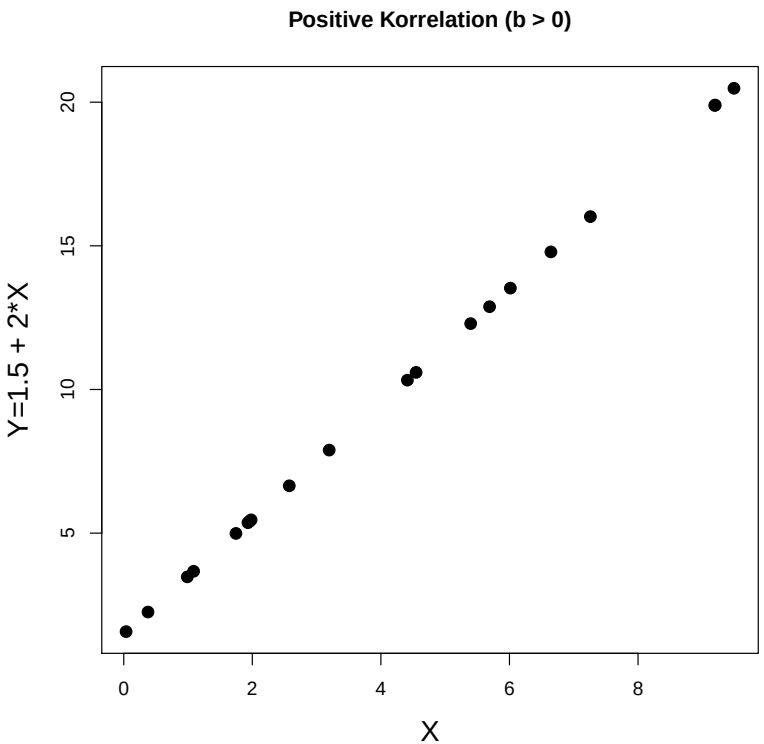
Durch r_{xy} wird also die Stärke des linearen Zusammenhangs der beiden Merkmale gemessen!
(ii) Die Merkmale X und Y heißen

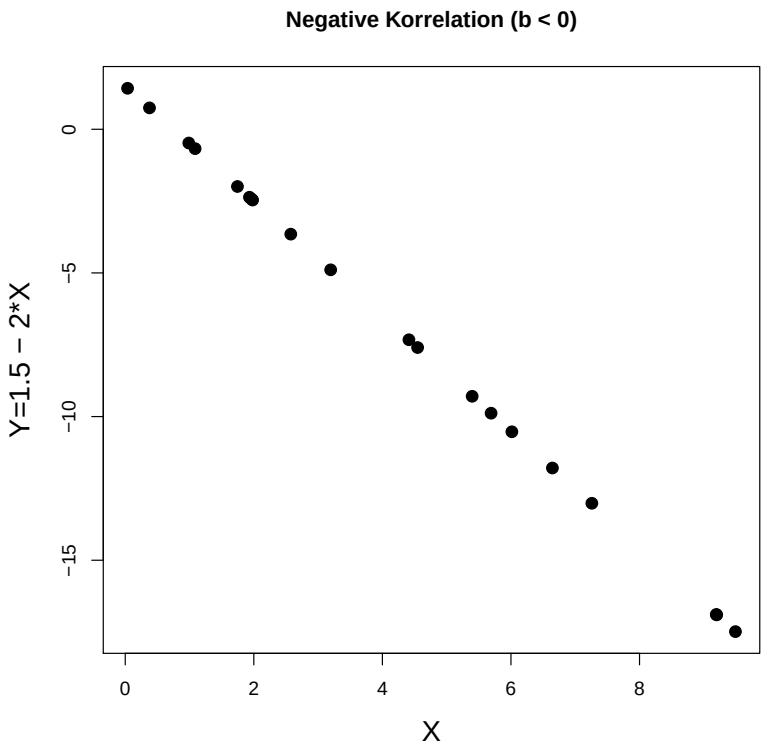
- **(empirisch) unkorreliert**, falls $r_{xy} = 0$
- schwach positiv [negativ] korreliert, falls $r_{xy} \in (0, 0.5)$ [(-0.5,0)]
- stark positiv [negativ] korreliert, falls $r_{xy} \in (0.8, 1)$ [(-1,-0.8)]. ₃

Bemerkung 4.1 (Eigenschaften und Interpretation): (iii) Beliebte Interpretation bei stark positiver [negativer] Korrelation:
Steigende Werte von X gehen mit steigenden [fallenden] Werten von Y einher.
(iv) Die empirische Kovarianz s_{xy} im Zähler von r_{xy} ist auch eine Kenngröße für den linearen Zusammenhang zweier Merkmale. Aufgrund der Normiertheit wird aber in der Regel r_{xy} verwendet.
(v) In Textbüchern ist auch die abkürzende Notation $\widehat{Cor}(X, Y)$ für r_{xy} geläufig.

Korrelation von 1 und -1

Realisierungen von $Y = a + bX$

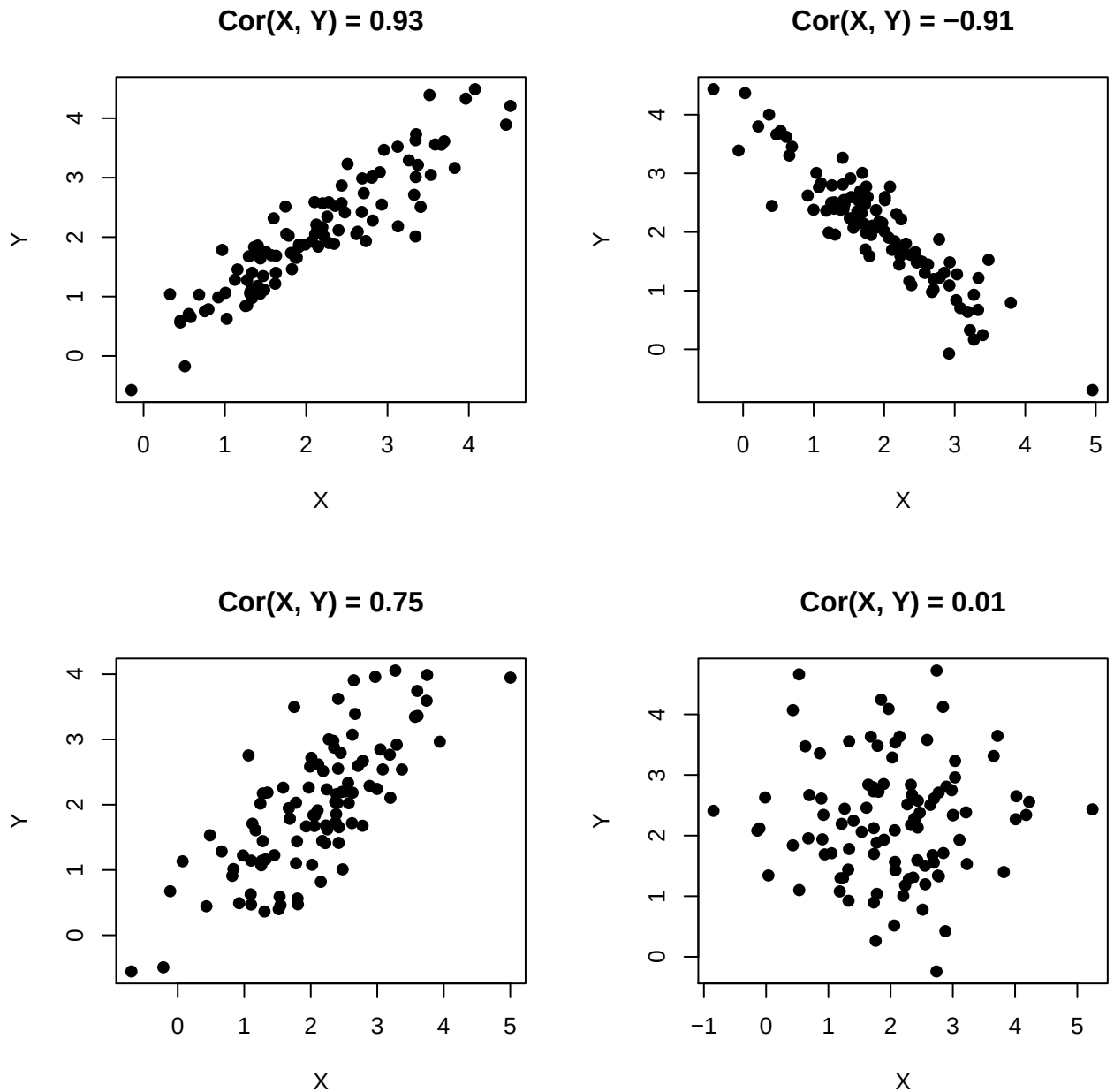




$r_{xy} = 1$

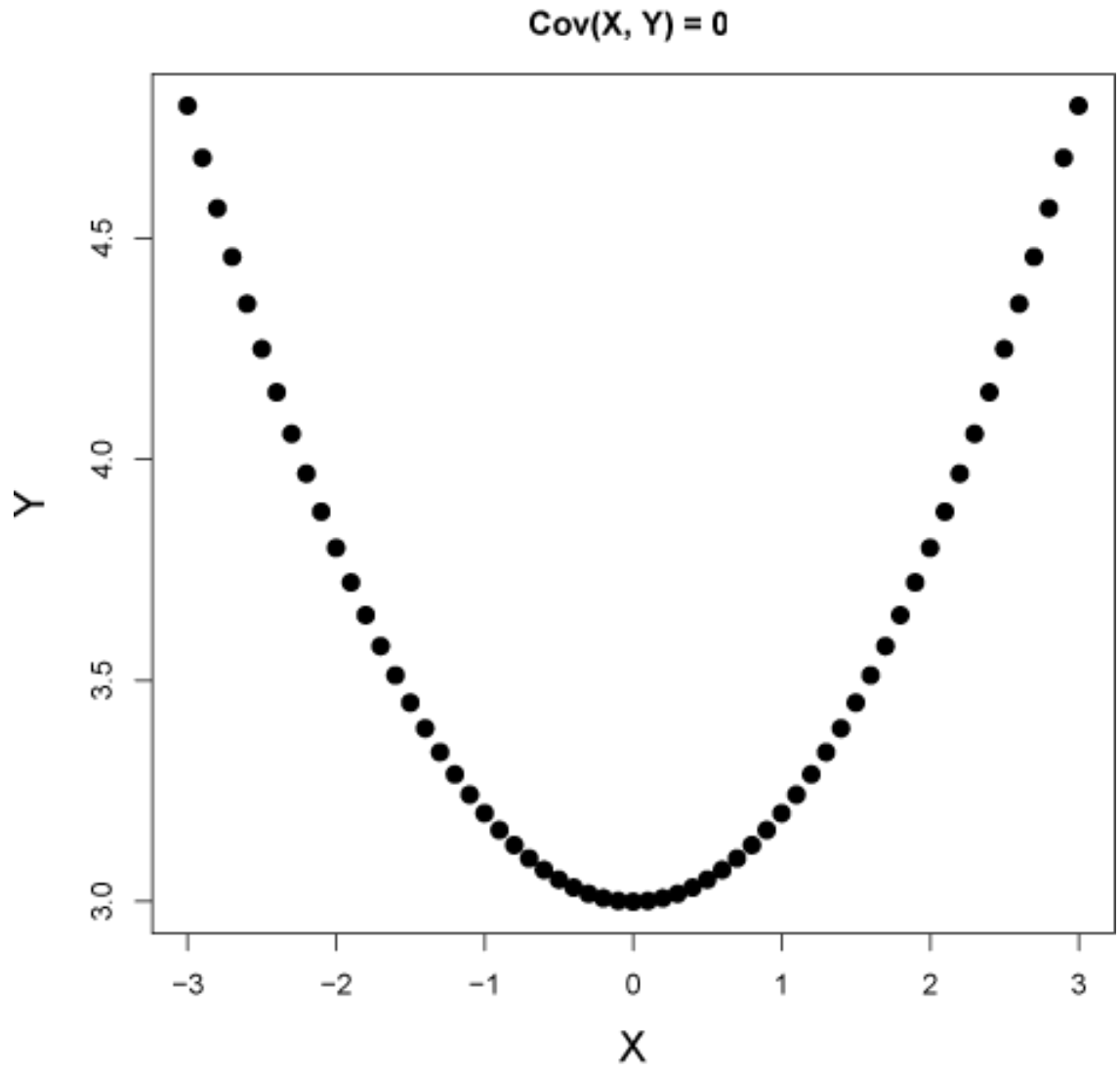
$r_{xy} = -1$

Korrelationen



Korrelationen

Merke nochmals: Die empirische Korrelation (und auch die empirische Kovarianz) misst nur die **lineare** Abhängigkeit zweier Merkmale!



Korrelationen

- Bei den zuletzt simulierten Daten liegt offenbar ein quadratischer Zusammenhang zwischen den Merkmalen vor
- Solche oder andere kompliziertere Zusammenhänge (z.B. über andere Kurven definierte) können i.a. nicht vom empirischen Korrelationskoeffizienten erkannt werden
- **Merkregel:** Wenn man r_{xy} berechnet, sollte man sich immer auch den zugehörigen Scatterplot der Daten anschauen!

Ränge und Korrelation von ordinalen Merkmalen

Möchte man die Korrelation bzw. Assoziation von ordinal skalierten Merkmalen untersuchen, so verwendet man häufig eine sog. **Rangtransformation** der Daten:
Definition 8.6 (Ränge): (a) Sei $x_1, \ldots, x_n$ die *paarweise verschiedenen* Ausprägungen eines Merkmals X. Der **Rang** der i-ten Beobachtung ist dann definiert als

$$Rg(x_i) = \#\{1 \leq j \leq n : x_j \leq x_i\}$$

und gibt die Position von x_i in der geordneten Statistik $x_{(1)} < \cdots < x_{(n)}$ wieder.
(b) Tritt der Wert x_i in der Urliste mehrfach (hier k-fach) auf, d.h. es gilt für ein $r > 0$: $x_{(r-1)} < x_i = x_{(r)} = x_{(r+1)} \cdots = x_{(r+k-1)} < x_{(r+k)}$, so teilt man die Ränge dieser Beobachtungen fair auf und definiert

$$Rg(x_i) = \frac{1}{k} \sum_{j=0}^{k-1} (r + j) = r + \frac{k-1}{2}$$

als den (Mittel)Rang von x_i.
(c) Die Abbildung $(x_1, \ldots, x_n) \mapsto (Rg(x_1), \ldots, Rg(x_n))$ heißt **Rangtransformation**.

Ränge: Beispiele

Bestimmen Sie die Ränge der folgenden Urlisten:

- $(x_1, x_2, x_3, x_4, x_5) = (7, 10, 3, 5, 1)$ sowie $(y_1, y_2, y_3, y_4, y_5) = (2, 3, 4, 1, 2)$
- Lösungen:
- $(Rg(x_1), Rg(x_2), Rg(x_3), Rg(x_4), Rg(x_5)) = (4, 5, 2, 3, 1)$ sowie
- $(Rg(y_1), Rg(y_2), Rg(y_3), Rg(y_4), Rg(y_5)) = (2.5, 4, 5, 1, 2.5)$
- Im zweiten Beispiel mit Bindungen (d.h. gleichen Werten; hier die 2) teilt man sich also die erzielten Punkte.
- **BEM:** Für bivariate Merkmale berechnet man die Rangtransformation komponentenweise, d.h. hier

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_5 \\ y_5 \end{pmatrix} \mapsto \begin{pmatrix} Rg(x_1) \\ Rg(y_1) \end{pmatrix}, \ldots, \begin{pmatrix} Rg(x_5) \\ Rg(y_5) \end{pmatrix} = \begin{pmatrix} 4 \\ 2.5 \end{pmatrix}, \ldots, \begin{pmatrix} 1 \\ 2.5 \end{pmatrix}.$$

Korrelation von ordinalen Merkmalen

Definition 8.7 (Spearman-Rang-Korrelationskoeffizient): Für n gegebene Beobachtungen

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$$

des bivariaten Merkmals (X, Y) mit komponentenweiser Rangtransformation

$$\begin{aligned} Rg(x) &= (Rg(x_1), \ldots, Rg(x_n)) \\ Rg(y) &= (Rg(y_1), \ldots, Rg(y_n)) \end{aligned}$$

ist der **Spearman-Rang-Korrelationskoeffizient** $\rho_{x,y}$ (auch **Spearman's-ρ**) der bivariaten Werte von (X, Y) gegeben durch den empirischen (Pearson) Korrelationskoeffizienten der rangtransformierten Werte, d.h.

$$\rho_{xy} := r_{Rg(x),Rg(y)}.$$

Bemerkung 8.2 (Eigenschaften und Interpretation): (i) Der Einsatz von Rängen ist beliebt, da diese invariant unter monotonen Transformationen der Daten sind und robust ggü Ausreißern sind.
(ii) Für Spearman's ρ_{xy} von zwei Merkmalen (X, Y) kann man zeigen:

- $\rho_{xy} \in [-1, 1]$
- $\rho_{xy} = 1$
  $\Leftrightarrow Rg(y_i) = Rg(x_i)$ für alle $i = 1, \ldots, n$.
- $\rho_{xy} = -1$
  $\Leftrightarrow Rg(y_i) = n + 1 - Rg(x_i)$.

Allgemeiner als r_{xy} ist ρ_{xy} ein Maß für eine mögliche Monotoniebeziehung zwischen den beiden Merkmalen (Genauer: es wird nicht nur der lineare; sondern ein monoton funktionaler Zusammenhang gemessen).

Bemerkung 8.2 (Eigenschaften und Interpretation): (iii) analog zur Interpretation von r_{xy} spricht man von schwach (stark) positiv (negativ) rangkorrelierten Daten bzw. einem schwach (stark) positiven (negativen) monotonen Zusammenhang.

(iv) Eine weitere verwandte Größe zur Messung der Assoziation von ordinalen Merkmalen ist der Kendall-Rang-Korrelationskoeffizient (**Kendall's tau**).

(v) Das obige Vorgehen mittels Rangtransformation aus einem statistischen Verfahren für quantitative Merkmale eines für ordinale zu gewinnen funktioniert nicht immer. Hier gibt es z.B. im Rahmen von sog. 2-faktoriellen Varianzanalysen Gegenbeispiele.

Bemerkungen zu Assoziationsmaße

- Wir haben in diesem Kapitel 3 Assoziationsmaße zur Untersuchung zweier Merkmale kennengelernt:
  - Das χ^2-Assoziationsmaß, welches insbesondere auch für qualitativ ausgeprägte Merkmale geeignet ist.
  - Der Spearman-(Rang-)Korrelationskoeffizient ρ_{xy}, der sich für mind. ordinal skalierte Merkmale eignet.
  - Der empirische Korrelationskoeffizient r_{xy}, welcher nur für quantitativ ausgeprägte Merkmale geeignet ist.
- Alle drei Größen geben Auskunft darüber, ob es einen (irgendwie gearteten[8]) Zusammenhang zwischen zwei Merkmalen gibt. **BEM:** Hierbei werden die Merkmale nicht nach abhängiger und erklärender Variable unterschieden!
- U.a. auch deshalb ist der häufig verwendete Schluss von Assoziation auf Kausalität i.a. falsch:

Eine Assoziation (bzw. Korrelation) impliziert keine Kausalität!

[8] z.B. linearen bei r_{xy}