awm

angewandte **w**irtschafts- und **m**edienpsychologie

Thema 05:
Multivariate Deskriptivstatistik

QM1, SoSe 22

HOCHSCHULE ANSBACH

Zusammenhang nominaler Variablen

Häufigkeitstabellen werden auch Kontingenztabellen genannt

Kontingenztabellen stellen Häufigkeiten der Kombinationen zweier (oder mehr) Variablen dar.

	No	Yes
Female	54	33
Male	97	60

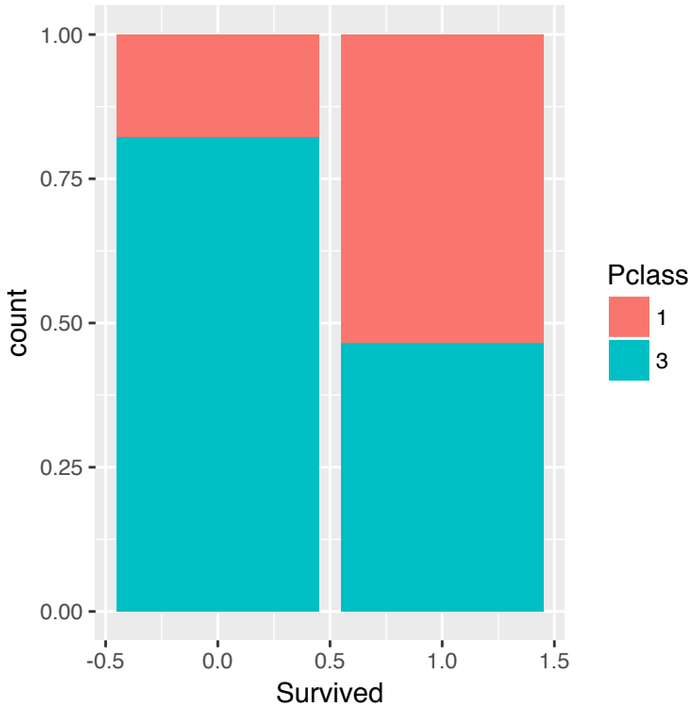
- Man teile Geschlecht in Männer vs. Frauen auf und Rauchstatus in Ja vs. Nein.
- Dann zähle man für jede (der 4) Gruppe aus, wie viele Fälle sich finden.
- Auf jeder der beiden Achsen (d.h. Zeilen und Spalten der Tabelle) wird eine der beiden Variablen dargestellt.
- Jede Zelle gibt die Häufigkeit einer Ausprägung eines Geschlechts und einer Ausprägung von Rauchstatus an.

	No	Yes
Female	35.76	35.48
Male	64.24	64.52
Total	100.00	100.00

- Man kann sich auch Prozentwerte oder Anteile (relative Häufkeiten) ausgeben lassen
- Dabei muss man sich festlegen, für welche der beiden Variable aufsummiert wird: für die Zeilenvariable oder die Spaltenvariable (oder beides).
- In diesem Beispiel (links) wird spaltenweise aufsummiert: Die Werte einer Spalte ergeben zusammen 100%.

Balkendiagramme und Zusammenhänge

- Hängt die Überlebensrate auf der Titanic vom Geschlecht ab? Anders gesagt: Sind Männer eher ertrunken als Frauen? Schauen wir uns ein Bild dazu an!

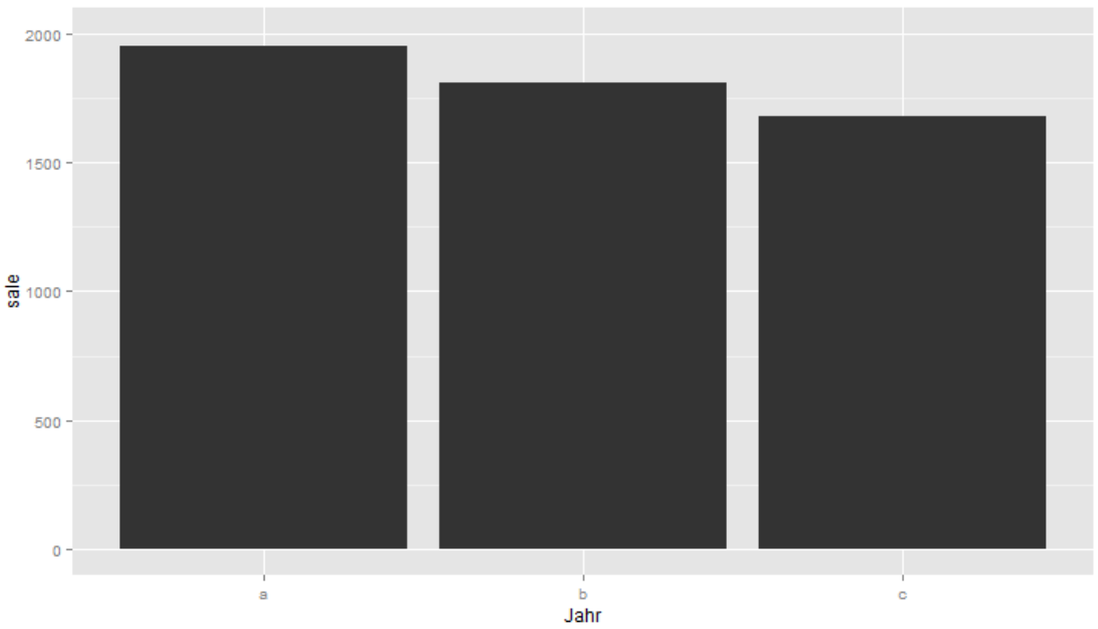


- Hm, bei den Überlebenden (Survived = 1) ist der Anteil der Passagiere der 1. Klasse recht hoch, etwa ein Drittel…
- Aber bei den Ertrunkenen (Survived = 0) ist der Anteil der 1. Klasse viel kleiner, gut 10%.
- Ah! Die Überlebensrate hängt von der Klasse ab; die beiden Variablen sind voneinander abhängig.
- Es gibt offenbar einen Zusammenhang zwischen Klasse und Überleben.

- Allgemein: Ist der Anteil von Ereignis "A" in allen Stufen der Variablen B *gleich*, so sind die beiden Variablen voneinander *unabhängig*. Ansonsten liegt ein *Zusammenhang* vor; die Variablen sind dann *abhängig* voneinander.
- Je größer die Unterschiede der Anteile, desto stärker der Zusammenhang.
- Dieses Maß (den Unterschied der Anteile bei dichotomen Variablen) bezeichnet man auch **Phi-Koeffizient** (φ, Φ).

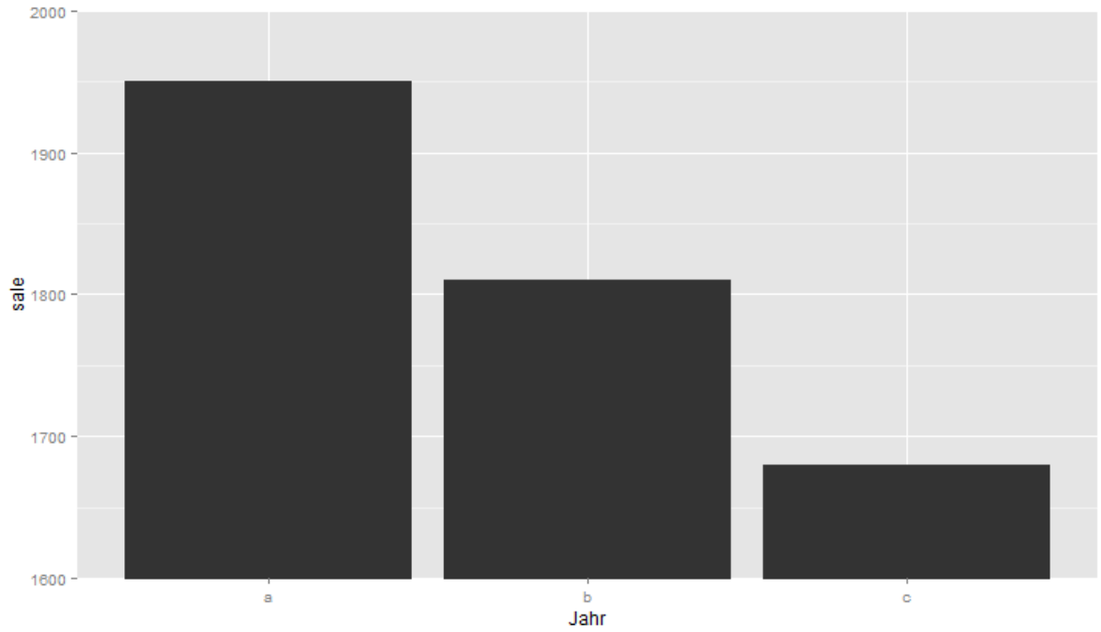
Wie man mit (Balken-)Diagrammen lügt

Meldung 1



Wir haben nur ein paar Dingens weniger verkauft. Halb so wild.

Meldung 2



Unsere Verkaufszahlen sind dramatisch eingebrochen!!!

Chancen-Verhältnis

Beispiel: Gibt es einen Zusammenhang von sozialem Status (Ihr Foto mit Maserati vs. Ihr Foto ohne Auto) und Dating-Erfolg auf einer Single-Online-Plattform?

	Foto mit Maserati 🚗	***Foto ohne Auto 👱***	***Summe***
Likes 😊	*9*	*2*	***11***
Dislikes ☹️	*2*	*7*	***9***
Summe	***11***	***9***	***20***

Joachim will das Angenehme mit dem Nützlichen verbinden: Zu Forschungszwecken und um nicht mehr alleine zu sein, meldet er sich bei einer Single-Plattform an. Er erstellt 2 Accounts: Einmal *mit* einem Foto von sich vor einem Maserati (geliehen); einmal nur mit einem Foto von sich (*ohne* Auto). Dann schaut er, wie viele „Likes“ jedes seiner zwei Profile bekommt…

- Mit dem Maserati hat Joachim 9 „Likes“ und 2 „Dislikes“, also 9 zu 2 (9:2 oder 9/2).
- Das Foto ohne Maserati (Joachim pur) bekommt nur 2 Likes und 7 Dislikes, also 2 zu 7 (2:7 oder 2/7) — armer Joachim (an sich ein netter Kerl).
- *Mit* Maserati sind seine Chancen (Verhältnis der beiden Häufigkeiten, hier Like und Dislike) offenbar besser.
- Wie müsste das Verhältnis der Chancen sein, wenn die Frauen sich *nicht* vom Maserati beeinflussen lassen? — Genau gleich groß, also 1:1.
- Je *unterschiedlicher* die Chancen, desto *stärker* stehen die Variablen im Zusammenhang

Beispiele für dichotome Zusammenhänge nominaler Variablen

Vertrieb: Ein Verkaufsleiter lässt seine Verkäufer eine Monat im schicken Anzug verkaufen; einen zweiten Monat müssen sich alle Verkäufe im „Gammel-Look“ anziehen. Dann vergleicht er die Anzahl der Vertragsabschlüsse. Ob der Kleidungsstil mit dem Vertriebserfolg zusammen hängt?

Bei Verkauf mit Anzug im Vergleich zum Gammel-Look stehen die Chancen für Erfolg 12:9 (=1,33:1):

	Anzug	*Gammel-Look*	*Summe*
Kunde kauft	***3***	***2***	***5***
Kunde rennt weg	***27***	***24***	***51***
Summe	***30***	***26***	***56***

$$\frac{\frac{3}{27}}{\frac{2}{24}} = \frac{3 \cdot 24}{27 \cdot 2} = \frac{3 \cdot 4}{9 \cdot 1} = \frac{12}{9} = 12 : 9$$

Beispiele für dichotome Zusammenhänge nominaler Variablen

Personalentwicklung: Der Personalchef führt mit der Hälfte der Mitarbeiter im Bereich Fertigung eine aufwändige Schulung durch in der Hoffnung, die Produkte werden weniger Mängel behaftet. Dann prüft er, ob es einen Zusammenhang gibt zwischen der Schulung (geschult vs. nicht geschult) und der Produktqualität (ok vs. Mängel behaftet).

Mit Schulung sind die Chancen für gute Qualität (Produkt ist OK) 4 mal höher als ohne Schulung:

Schulung	*keine Schulung*	*Summe*
Produkt OK	***20***	***20***	***40***
Produkt Mängel behaftet	***10***	***40***	***50***
Summe	***30***	***60***	***90***

$$\frac{20}{10} : \frac{20}{40} = \frac{2}{1} : \frac{2}{4} = 2 : 0,5 = 4$$

Sind Raucher häufiger krank?

Gibt es einen Zusammenhang zwischen "Rauchen" und "Gesundheit"? Berechnen Sie das Odds Ratio von Gesundheit (zu Krankheit) von Nichtrauchern (im Verhältnis zu Rauchern!) in diesen fiktiven Daten.

	Raucher	krank	gesund
1	Raucher	24	1619
2	NRaucher	173	1321

Das Chancenverhältnis misst die Stärke eines Zusammenhangs

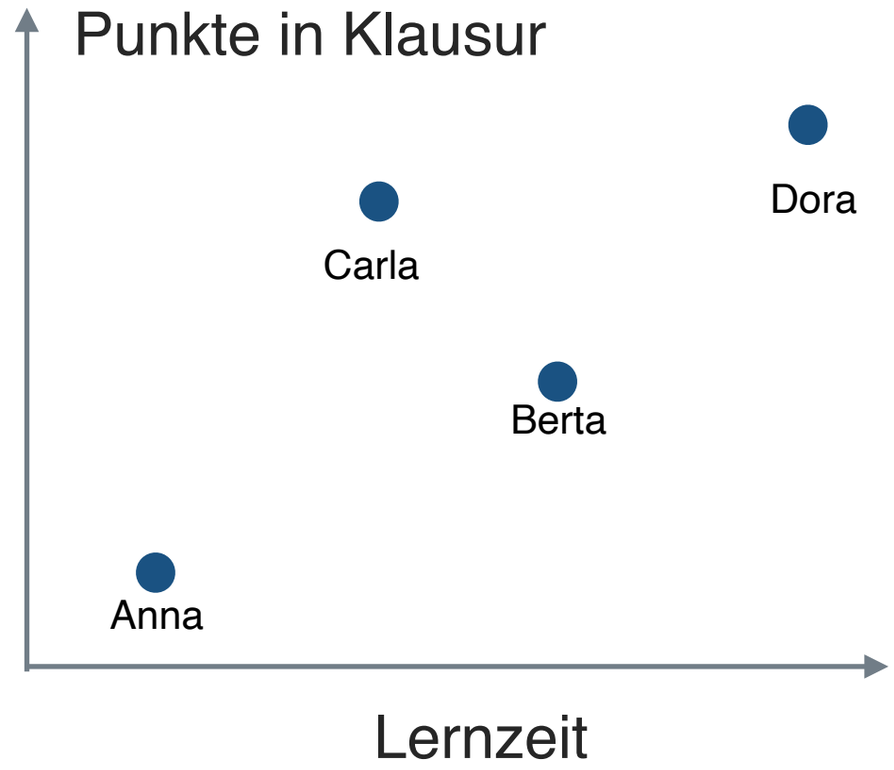
- Die Stärke des Zusammenhangs zweier dichotomer (zweiwertiger) nominaler Variablen lässt sich z.B. mit dem Chancenverhältnis bemessen. Eine gebräuchliche Bezeichnung lautet „Odds-Ratio“ (engl.: odds = Chance, Ratio = Verhältnis).
- Eine Chance ist definiert als das Verhältnis zweier absoluter Häufigkeiten (z.B. 9 zu 2) oder relativer Häufigkeiten (9/20 zu 2/20).
- Odds Ratio (OR) ist definiert als das Verhältnis zweier Chancen:
- Nimmt OR einen Wert von 1 an, so hängen die beiden Variablen nicht zusammen, sie sind unabhängig voneinander.
- Ist OR größer als 1, so ist der Zusammenhang positiv (gleichsinnig); ist OR kleiner als 1, ist der Zusammenhang negativ (gegensinnig).

$$OR = \frac{\frac{n_{11}}{n_{12}}}{\frac{n_{21}}{n_{22}}}$$

- Die untere Grenze von OR ist Null; die obere Grenze von OR ist Unendlich.
- Bei der Interpretation von OR muss die Anordnung der Kategorien beachtet werden, da der Wert von OR davon abhängt (Chancen Anzug vs. Gammel-Look ≠ Chancen Gammel-Look vs. Anzug). Die beiden Werte von OR sind jeweils der Kehrwert voneinander. Die Stärke des Zusammenhangs ist unabhängig davon gleich.
- OR ist unabhängig von der Stichprobengröße: Multipliziert man eine Zeile oder Spalte mit einem Faktor k, so ändert sich OR nicht. Der Grund ist, dass sich das Verhältnis nicht ändert, da sich der Faktor wieder aus dem Bruch herauskürzt.
- Eine Chance ist nicht dasselbe wie eine Wahrscheinlichkeit! Liegt die Chance eine Prüfung zu bestehen bei 30/3 (10:1), so beträgt die Wahrscheinlichkeit zu bestehen 30/(30+3) ≈ .91.
- Welches OR „hoch“ bzw. „stark“ ist, wird von vielen als subjektiv bzw. domänenabhängig betrachtet.

Zusammenhang metrischer Variablen

Gibt es einen Zusammenhang zw. Lernzeit und Klausurerfolg?



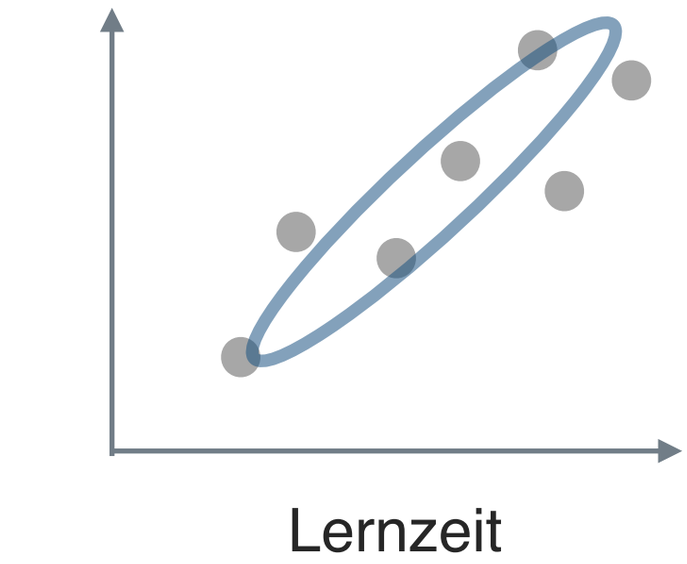
	Lernzeit	*Punkte in Klausur*
Anna	10	30
Berta	30	60
Carla	20	90
Dora	50	120

- Wir schauen uns das Diagramm an und sehen: Wer *viel gelernt* hat, hatte tendenziell eine *gute Note**. Wer *wenig* gelernt hatte, hatte eine *schlechte* Note (*wenig* Punkte).
- Es gibt also einen **Zusammenhang** zwischen *Lernzeit* und dem *Klausurerfolg*.
- Genauer gesagt: Wir haben uns eine Studentin angeschaut und geprüft, ob sich ihre beiden Werte „ähneln“.
- Hat die Studentin in *Lernzeit* einen *geringen* Wert, so erwarten wir auch einen *geringen* Wert in *Klausurerfolg*.
- Hat sie umgekehrt einen *hohen* Wert in Lernzeit erwarten wir einen *hohen* Wert im Klausurerfolg.
- Diese Prüfung machen wir für alle Studentinnen. Je ähnlicher die beiden Werte insgesamt, desto stärker ist der Zusammenhang.

*typisches Dozentenbeispiel….

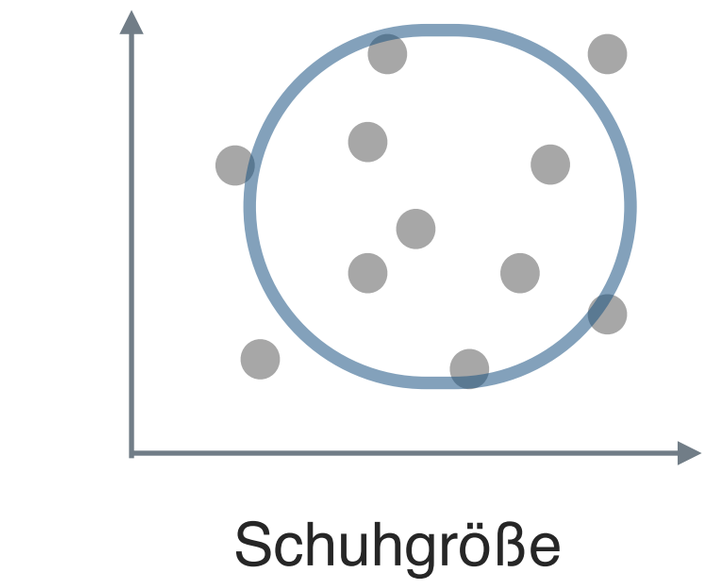
Zusammenhang: stark vs. schwach

Starker Zusammenhang



- Wenn Lernzeit mit Klausurerfolg stark zusammenhängt (stark korreliert ist), so geht *viel* Lernzeit systematisch (tendenziell) mit *viel* Klausurerfolg einher und *wenig* Lernzeit mit *wenig* Klausurerfolg.
- Im Diagramm bilden die Daten einer „*Zigarre*“ (schmale Ellipse)

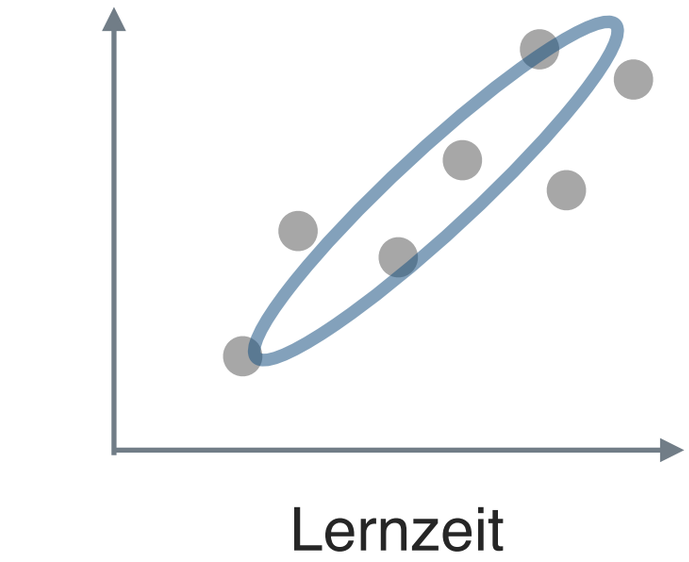
schwacher/kein Zusammenhang



- Wenn Schuhgröße wenig/nicht mit Gewissenhaftigkeit zusammenhängt (wenig/ nicht korreliert ist), so geht kleine Schuhgröße genauso mit wenig als auch mit viel Gewissenhaftigkeit einher. Für große Schuhgröße gilt das auch.
- Im Diagramm bilden die Daten eine „*Torte*“ (Kreis oder Quadrat)

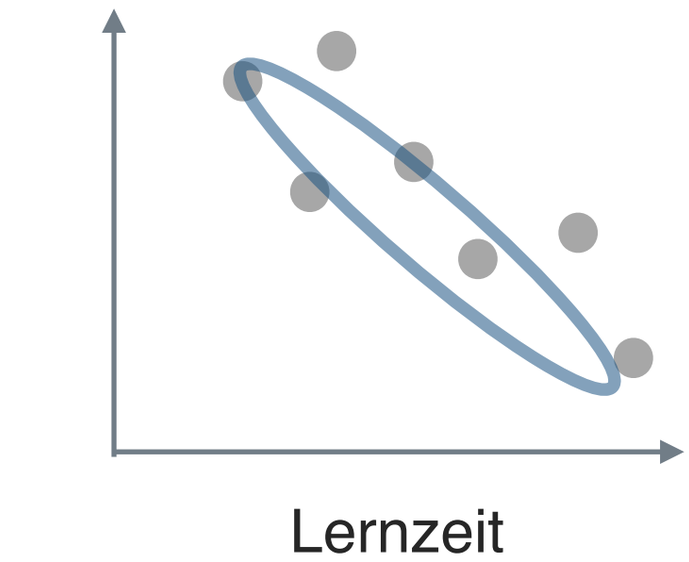
Zusammenhang: positiv vs. negativ

Positiver Zusammenhang



- *Hohe* Werte in Lernzeit sind mit *hohen* Werten im Klausurerfolg verbunden, bzw. geringer Werte mit geringen.
- *Gleichsinniger* Zusammenhang
- „Hoch-hoch und niedrig-niedrig“
- „Je mehr, desto mehr“
- Im Diagramm „steigt“ die Ellipse

Negativer Zusammenhang

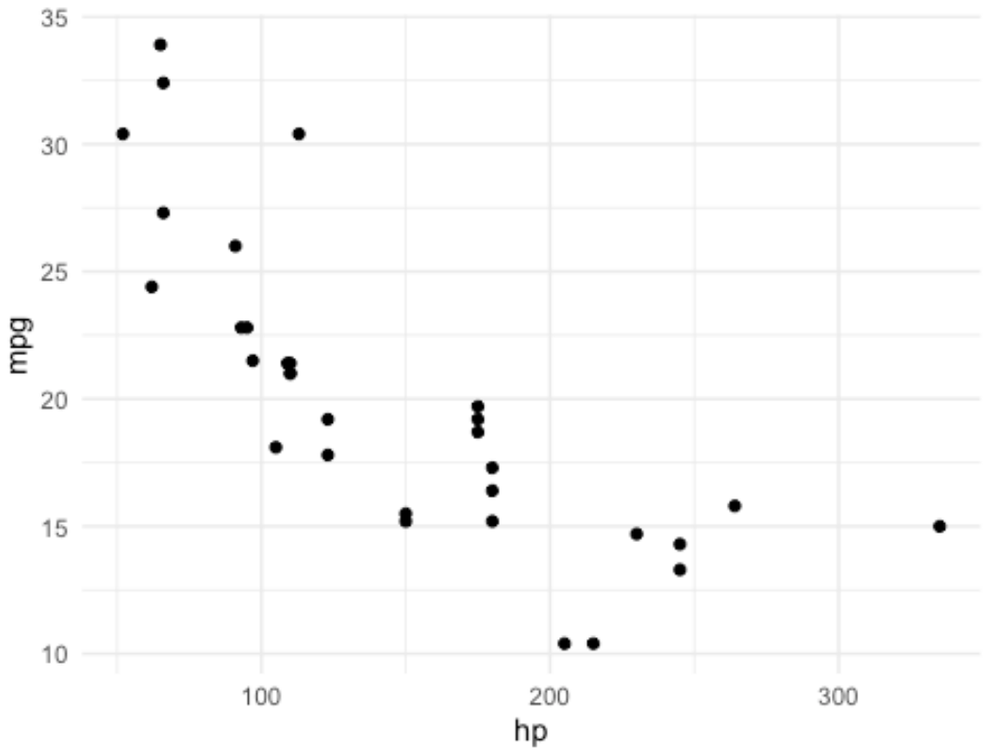


- Hohe Werte in Lernzeit sind mit *geringen* Werten in Freizeit verbunden, bzw. geringer Werte mit hohen.
- *Gegensinniger* Zusammenhang
- „Hoch-*niedrig* und niedrig-*hoch*“
- „Je mehr, desto *weniger*“
- Im Diagramm *sinkt* die Ellipse

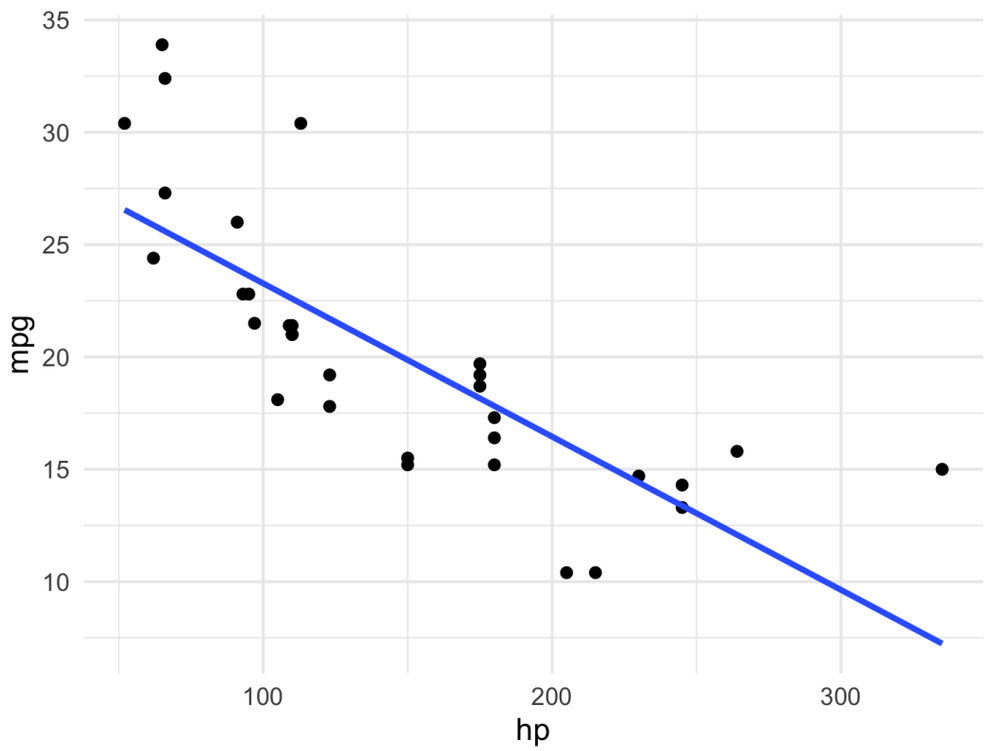
Streudiagramm zeigt Zusammenhänge

- Ein *Streudiagramm* zeigt den *Zusammenhang zweier metrischer Variablen* an; z.B. zwischen PS-Zahl (hp) und Spritverbrauch (mpg)
- Eine *Regressionsgerade* zeigt den *linearen* Trend des Zusammenhangs von X und Y

Streudiagramm



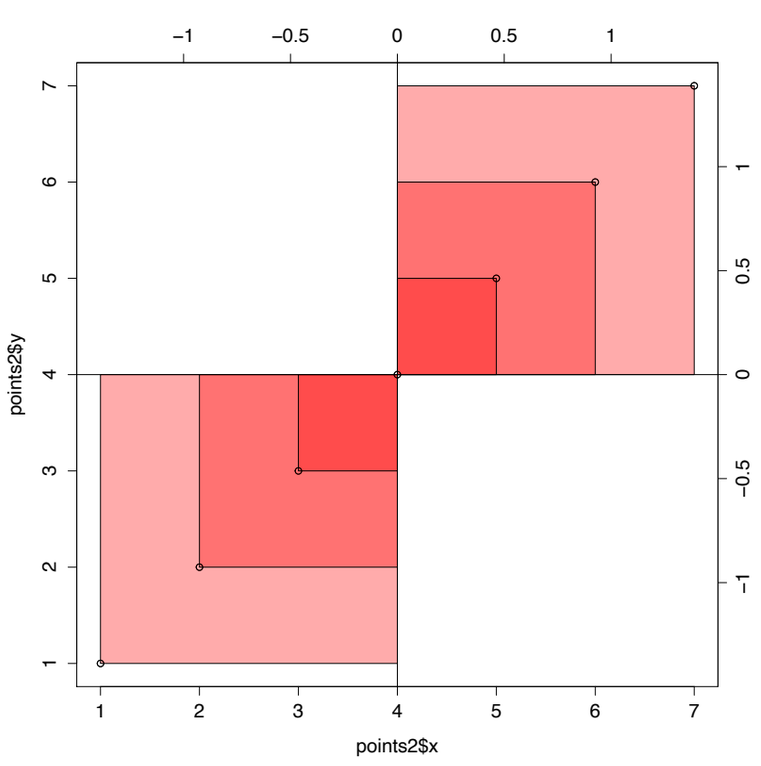
Mit Trendgerade



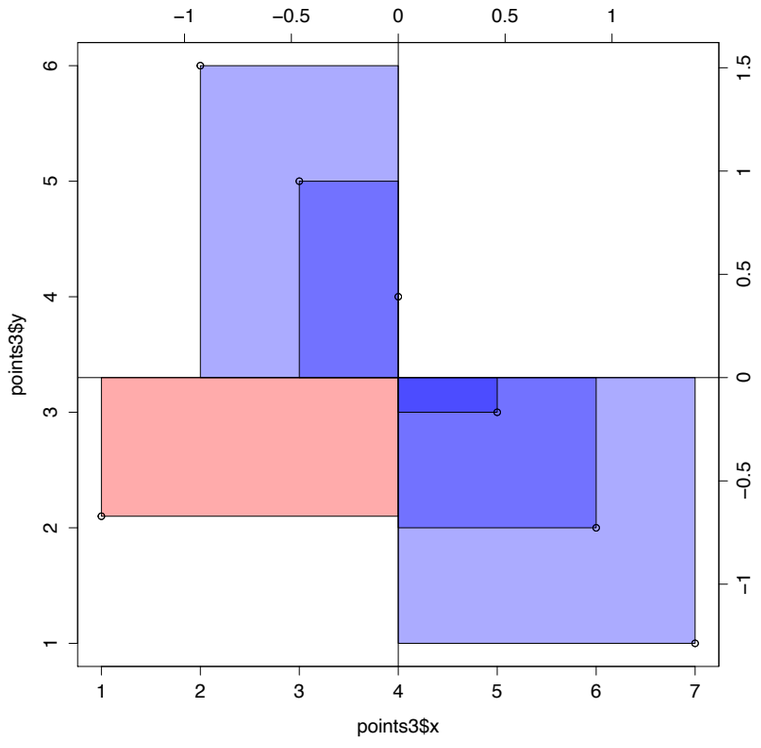
Abweichungsrechtecke

Mittelwerts-Rechtecke als Maß für den Zusammenhang

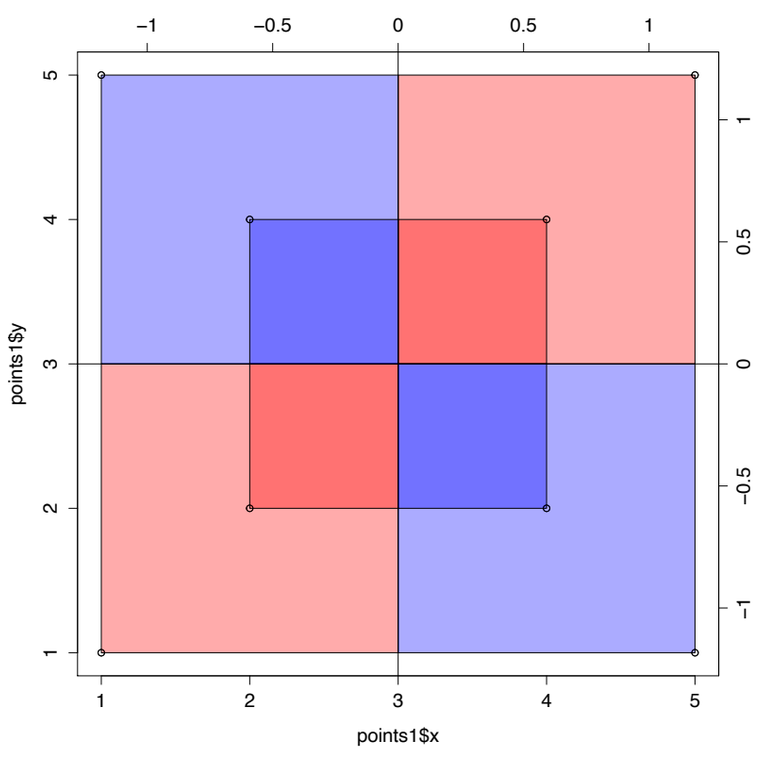
… sind ein Weg, der Stärke des (linearen) Zusammenhangs zweier metrischer Variablen eine Zahl zuzuweisen. Dabei wird für jeden Punkt das Rechteck gebildet, welches der jeweilige Punkt mit der $\bar{X}$-Mittellinie und der $\bar{Y}$-Mittellinie aufspannt.



perfekter *positiver* Zusammenhang



starker *negativer* Zusammenhang

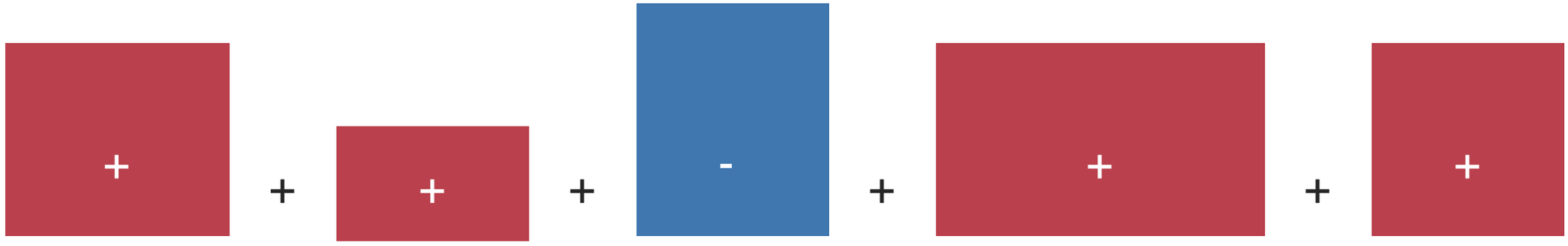


kein Zusammenhang

- In Summe viel rot, *wenig* blau: starker positiver Zusammenhang
- In Summe viel blau, wenig rot: starker negativer Zusammenhang
- In Summe etwa gleich viel rot wie blau: kein/ wenig Zusammenhang

Die Summe der Fläche ist ein Maß für den Zusammenhang

Beispiel:

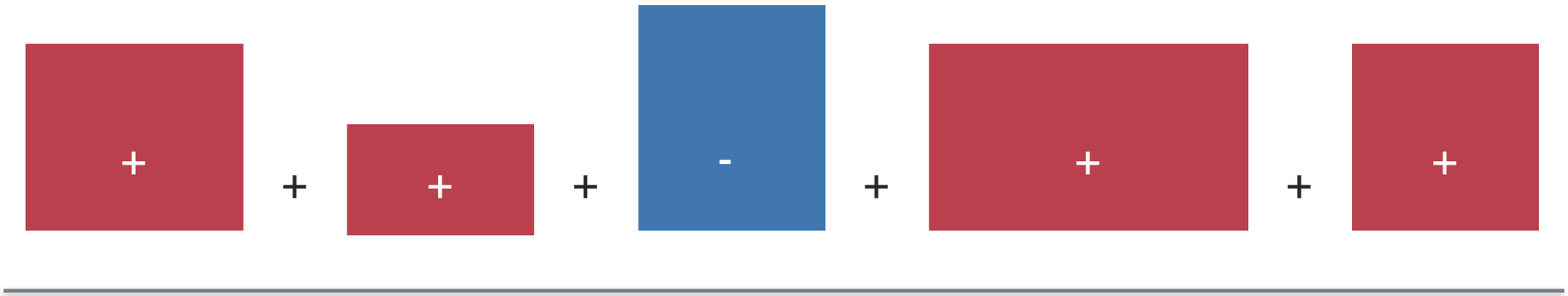


=

Die Summe der 5 Abweichungs-Rechtecke entspricht dem unteren Quadrat*

*Augenmaß ;)

Teilen durch n liefert das durchschnittliche Rechteck

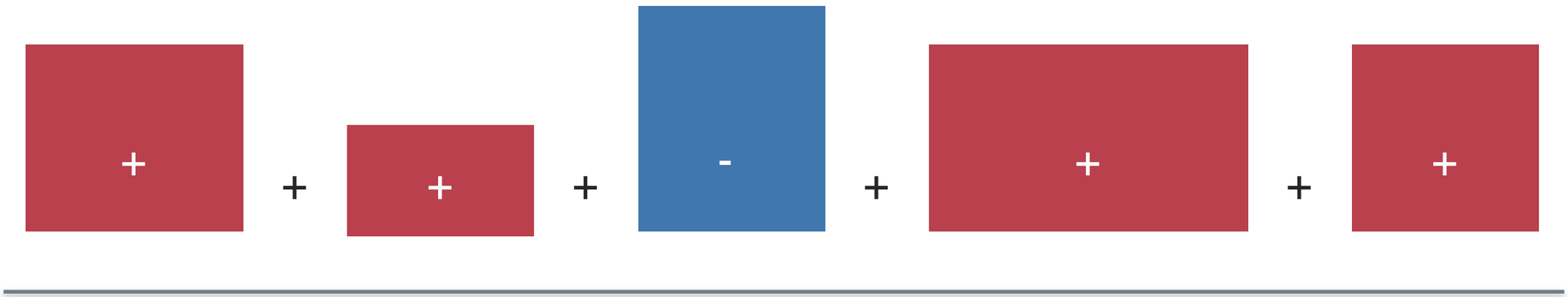


Der Durchschnitt der 5 Abweichungs-Rechtecke entspricht dem unteren Rechteck*

- Dieses „**durchschnittliche Rechteck**“ bezeichnet man als **Kovarianz**
- Der Durchschnitt hat im Gegensatz zur Summe den Vorteil, dass er **unabhängig** ist von der **Anzahl** der Werte; somit ist die Kovarianz ein Maß, das *unabhängig* ist von der Anzahl der Elemente

*Augenmaß ;)

Voilà: Die Kovarianz



5

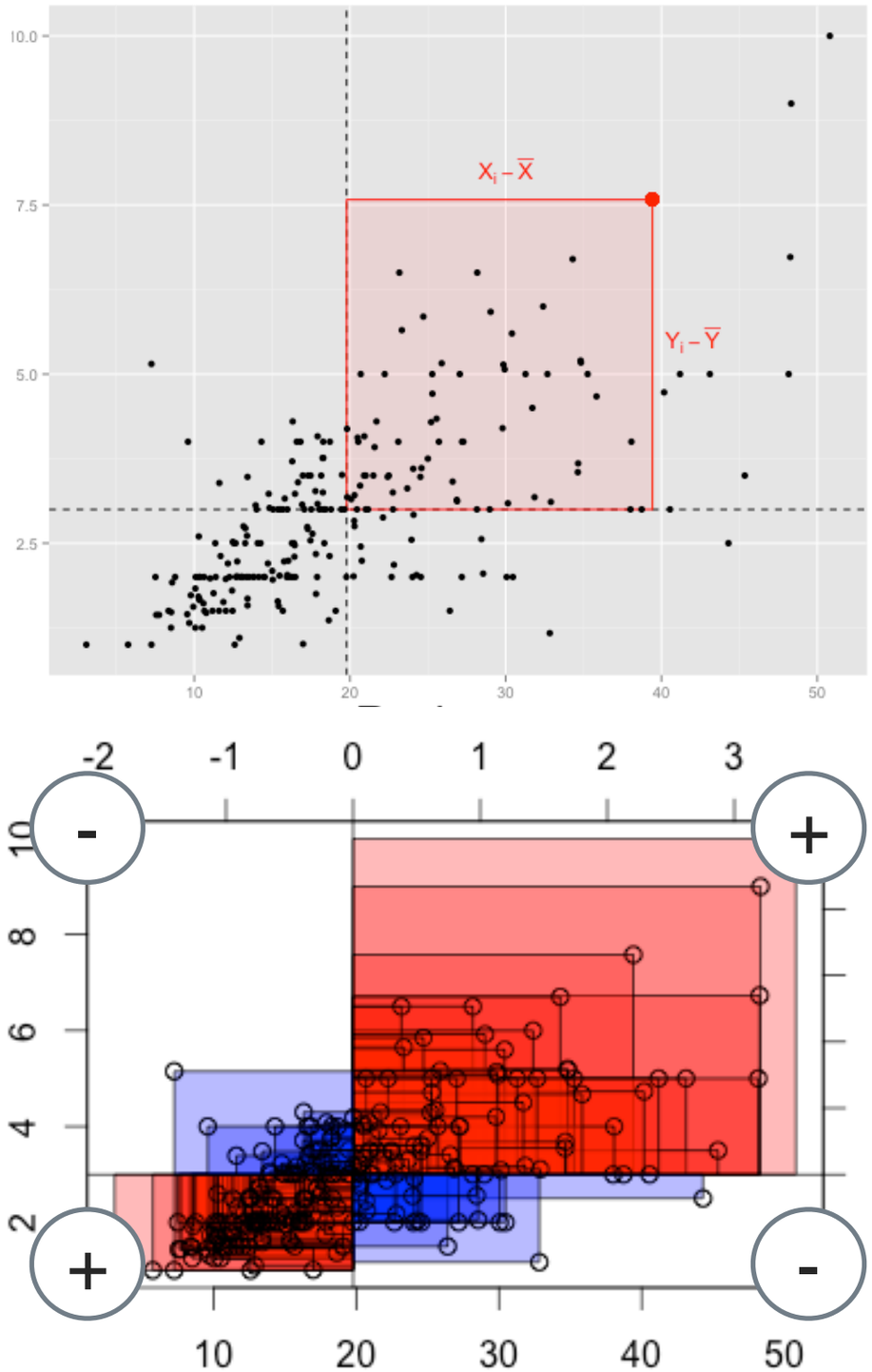
=

Der Durchschnitt der 5 Abweichungs-Rechtecke entspricht dem unteren Rechteck

- Dieses „**durchschnittliche Rechteck**“ bezeichnet man als **Kovarianz**
- Der Durchschnitt hat im Gegensatz zur Summe den Vorteil, dass er **unabhängig** ist von der **Anzahl** der Werte; somit ist die Kovarianz ein Maß, das *unabhängig* ist von der Anzahl der Elemente

Die Kovarianz als Zusammenhangsmaß

Die Kovarianz (Kov) bezeichnet die Fläche des *durchschnittlichen Rechtecks*, d.h. das durchschnittliche Produkt der Abweichungen von $\bar{X}$ und $\bar{Y}$.



- Die Größe der *Kovarianz* ist *abhängig* von der *Skalierung* der Variablen; transformiert man z.B. eine Variable [EURO] in [Cent], so wird die Kovarianz 100fach vergrößert
- Die Kovarianz hat *keinen* minimalen und auch keinen maximalen Wert
- Da sie „*ungedeckelt*“ ist, ist oft schwer (oder gar nicht) zu sagen, ob die Kovarianz groß ist
- Eine *große positive* (negative) Kovarianz zeigt einen *starken positiven* (negativen) Zusammenhang an
- Eine kleine positive (negative) Kovarianz zeigt einen schwachen positiven (negativen) Zusammenhang an

Die Kovarianz als durchschnittliches Rechteck

Die **Kovarianz** ist ein **Maß** für den **linearen Zusammenhang** zweier Variablen. Sie berechnet sich wie folgt:

- **Für jeden Punkt** im Streudiagramm berechnet man die **Abstände** zu den **Mittelwerten** der Variable **X** bzw. **Y**.
- Für **jeden Punkt** kann man die **Fläche** eines **Rechteck** berechnen als Produkt der Abweichung des X-Wertes bzw. Y-Wertes vom Mittelwert von X bzw. von Y.
- Dabei kann ein Punkt in einen der **vier Quadranten** fallen, die um die Mittelwert herum liegen (s. Diagramm vorherige Seite):
  - Für Punkte in den Quadranten I und III ergeben sich positive Vorzeichen für das „Rechteck“ (rote Rechtecke);
  - für die Quadranten II und IV ergeben sich negative Vorzeichen (blaue Rechtecke).
- **Summiert** man nun die **Flächen** der **Rechtecke aller Punkte** auf auf, ergibt sich dann ein *betragsmäßig* hoher Wert, wenn die Rechtecke *vor allem* in den Quadranten I und III liegen (positiv korreliert sind) *oder auch* wenn die Rechtecke vor allen in den Quadranten II und IV liegen (negativ korreliert sind).
- Sind die Rechtecke über alle vier Quadranten etwa gleichmäßig verteilt, ergibt sich eine kleine Summe (nahe) Null (oder sogar exakt Null).
- Teilt man diese Summe der Rechtecke durch die Anzahl der Rechtecke, so erhält man das „**durchschnittliche Rechteck**“; diesen Wert bezeichnet man als Kovarianz.

Korrelationskoeffizient r: „Mittleres z-Rechteck“

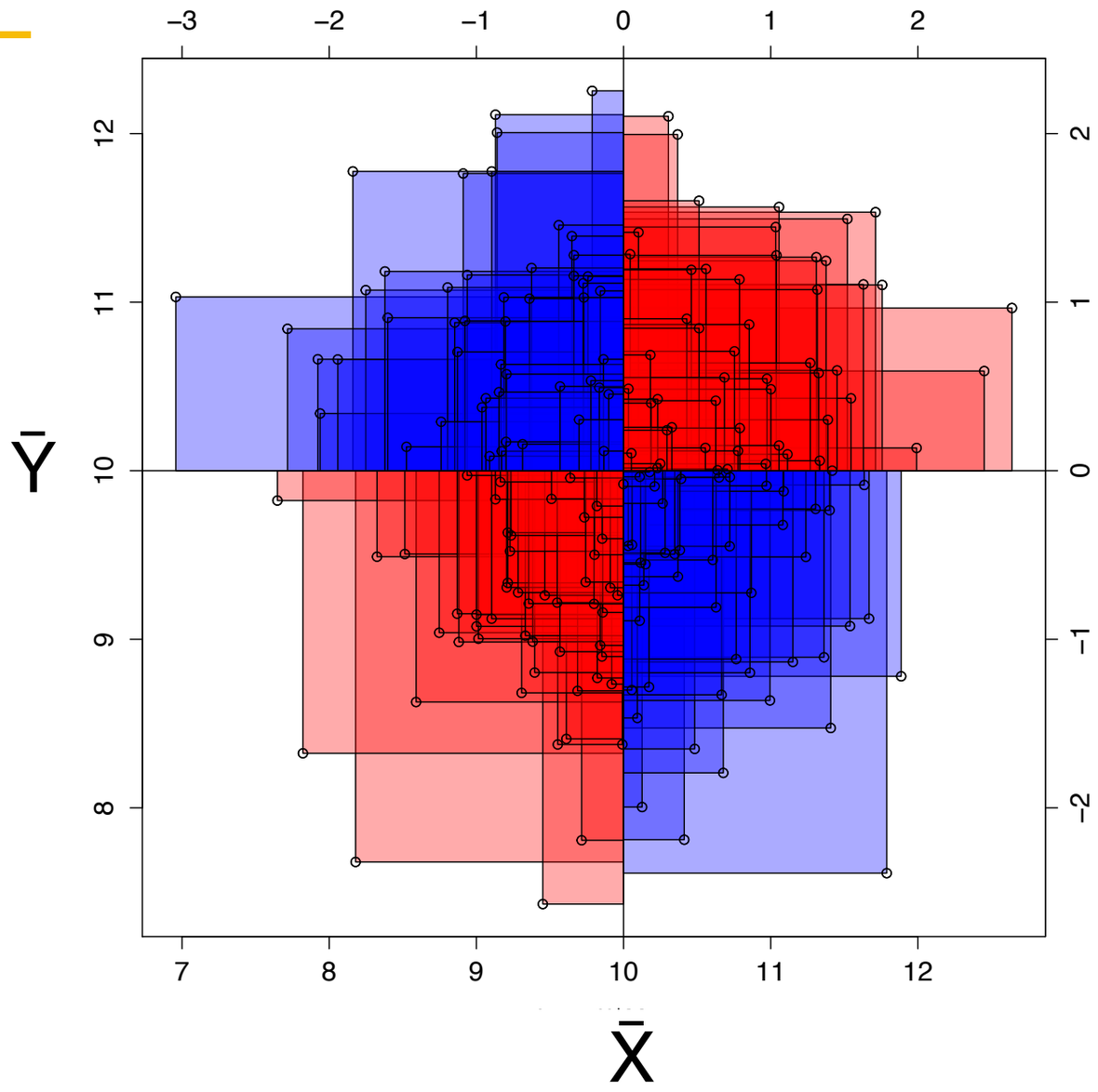
- Der **Korrelationskoeffizient** *r* nach K. Pearson löst das Problem, dass die Kovarianz schwer interpretierbar ist; der **Wertebereich** reicht von **-1** (perfekte negative lineare Korrelation) bis **+1** (perfekte positive lineare Korrelation); eine Korrelation von r = 0 bedeutet kein linearer Zusammenhang.
- Bei der Korrelation werden **beide Variablen z-transformiert**, dadurch wird die Stärke (und Richtung) des Zusammenhangs unabhängig von der Skalierung (Varianz) der Variablen.
- Entsprechend berechnet sich die **Korrelation** als **„durchschnittliches Rechteck“ der z-transformierten Variablen.**

$$r_{XY} = \frac{1}{n} \sum z_x z_y = \varnothing(z_x \cdot z_y)$$

- Der Korrelationskoeffizient *r* ist (anders als die Kovarianz!) gegenüber Maßstabsunterschieden in den untersuchten Merkmalen unempfindlich.
- Nach einer gängigen groben **Faustregel** von J. Cohen gilt **r ≈ ±0.1** als „**schwach**“, **r ≈ ±0.3** als „**mittel**“ und ab **r ≈ ±0.5** als „**stark**“.

Vertiefung

Bei r = 0 gleichen sich die Rechtecke aus



200 Punkte mit r ≈ 0

- Bei r ≈ 0 ist die Fläche der Abweichungs-Rechtecke, wenn man sie pro Quadrant aufsummiert, etwa gleich groß
- Addiert man die Abweichungs-Rechtecke (unter Beachtung der Vorzeichen: rot = positiv; blau = negativ), so beträgt die Summe in etwa (oder genau) Null
- Damit ist die Korrelation (und die Kovarianz) etwa bzw. genau Null:

$$\sum(\Delta X \cdot \Delta Y) = 0$$

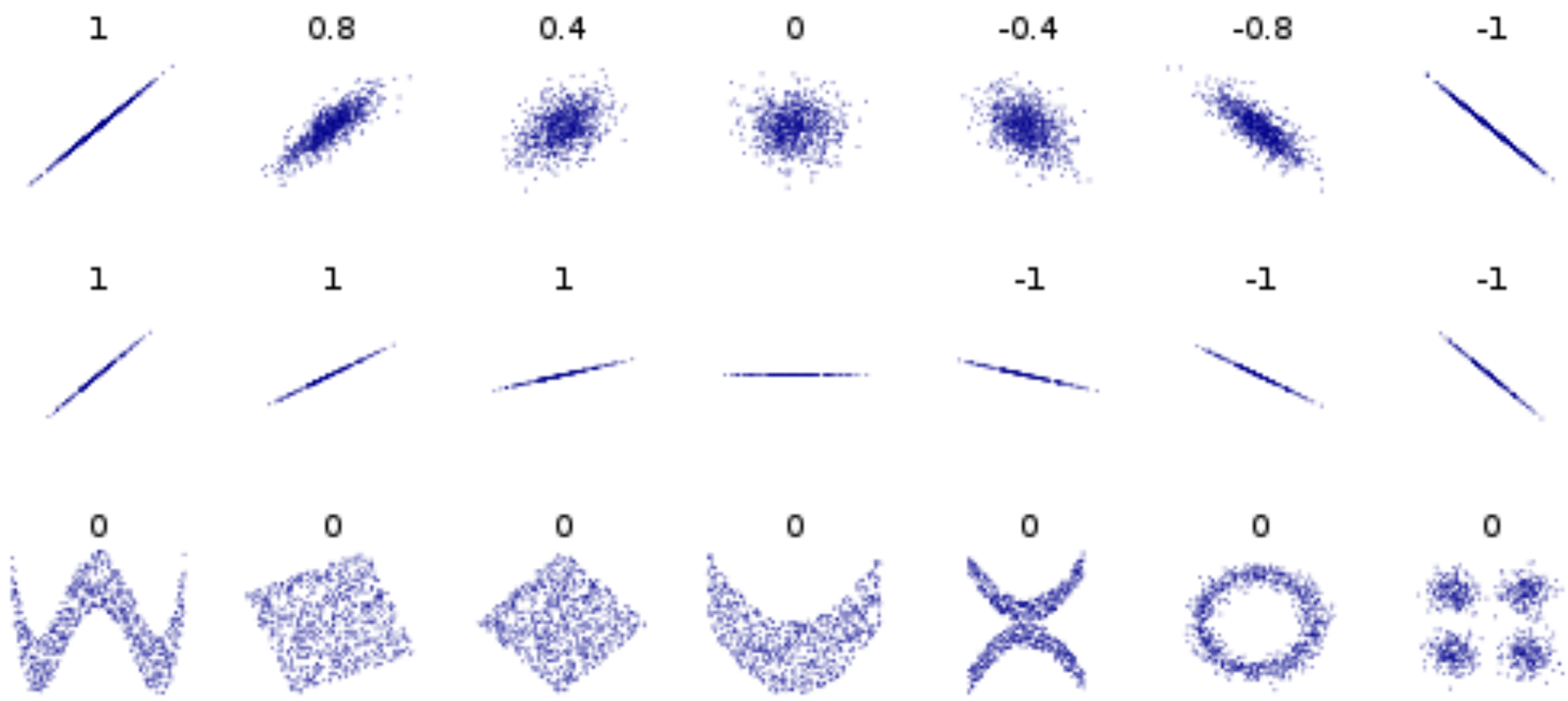
$$\Leftrightarrow \varnothing(\Delta X \cdot \Delta Y) = 0$$

$$\Leftrightarrow Kov = 0$$

$$\Leftrightarrow r = 0$$

Zusammenhänge können in Stärke und Richtung variieren

Beispiele für Korrelationskoeffizienten verschiedener Stärke und Richtung:



Faustregel:

- **Starke** Korrelation: Punktewolke ähnelt einer „**Zigarre**“ (schmale Ellipse)
- **Schwache** Korrelation: Punktewolke ähnelt einer „**Torte**“ (Kreis oder breite Ellipse)

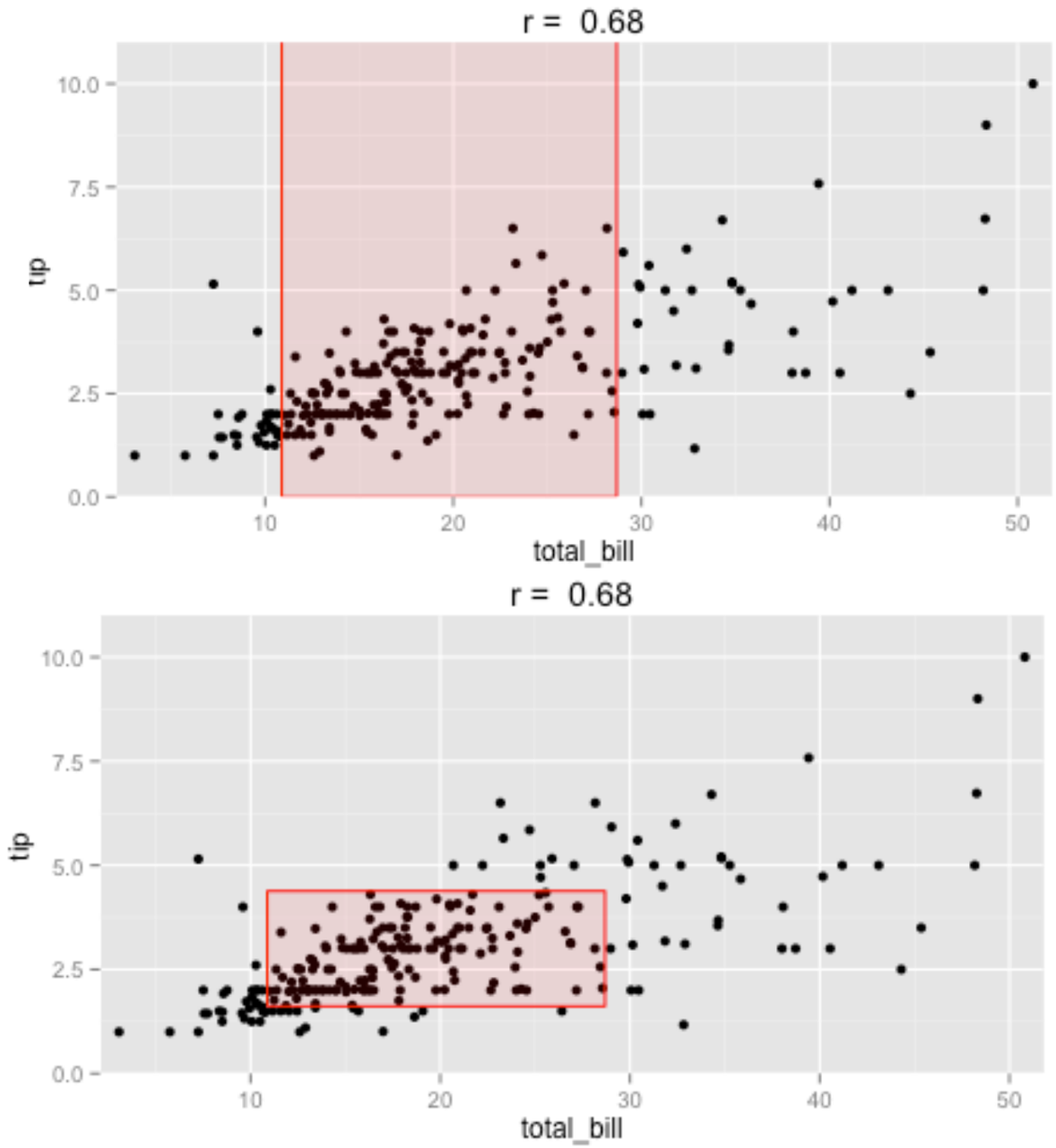
Merke: Die Korrelation misst *nur* die Stärke des **linearen** Zusammenhangs

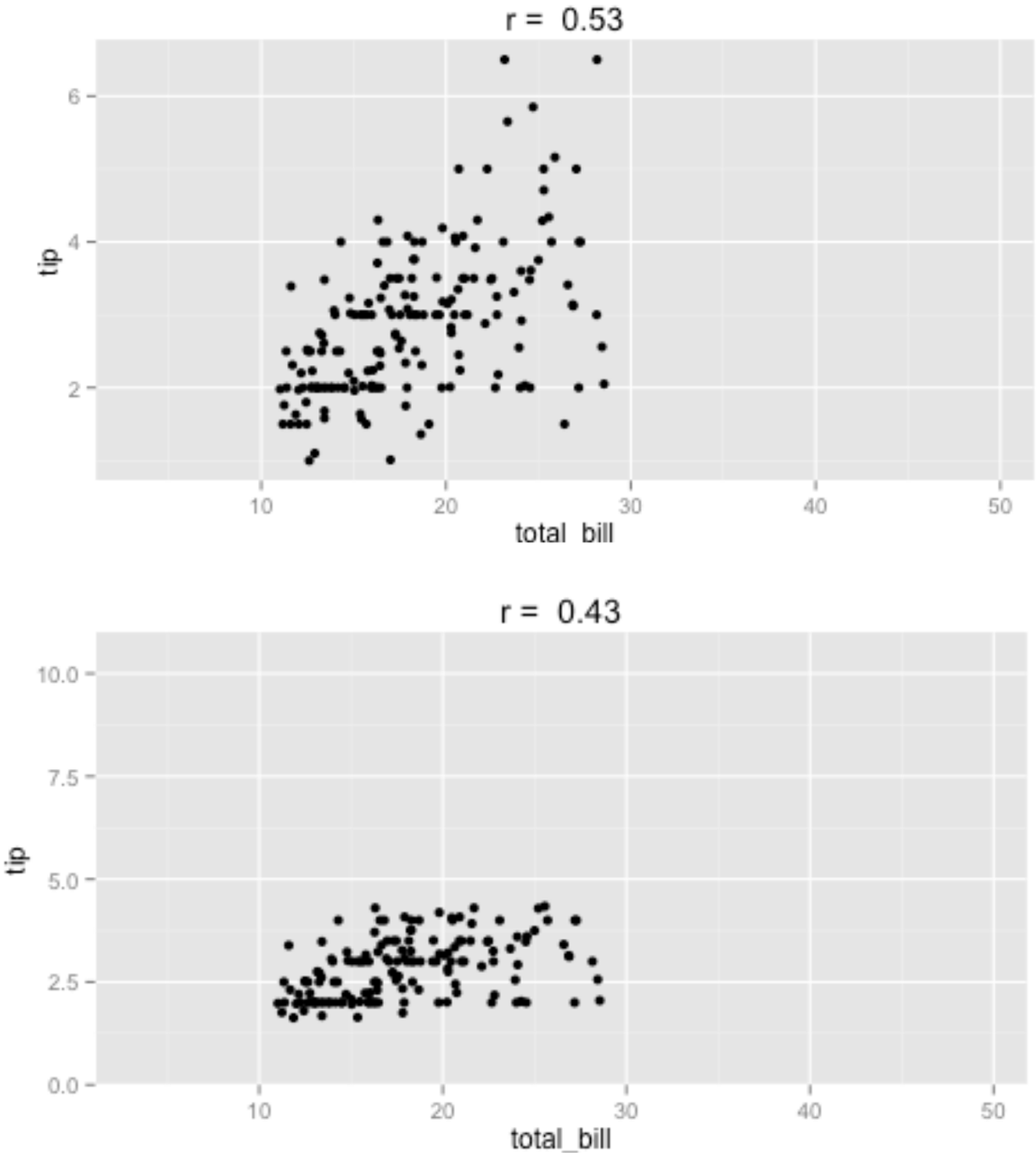
Visualisierung: http://rpsychologist.com/d3/correlation/

https://en.wikipedia.org/wiki/Correlation_and_dependence

Einschränkung des Ranges verringert i.d.R. die Korrelation

- Verringert man in einer oder beiden Variablen den Wertebereich (und damit die Streuung), so verringert sich zumeist auch die Höhe der Kovarianz und damit auch der Korrelation.



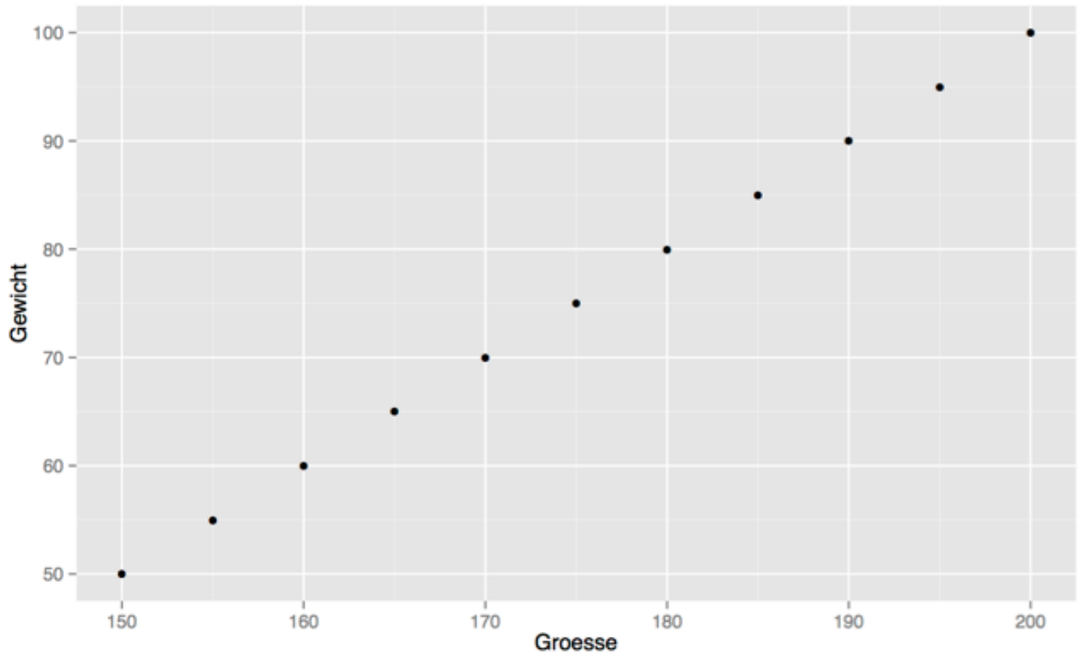


Der rot markierte Bereich wird herausgeschnitten

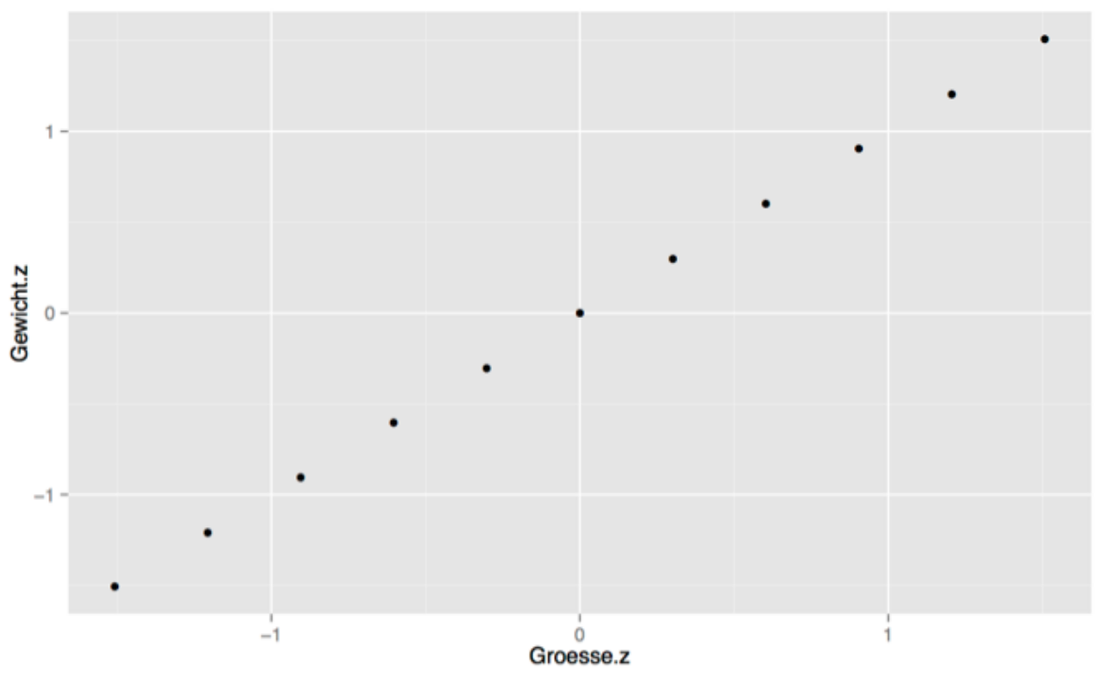
Die Korrelation dieser Teil-Daten ist geringer

Was ist eine perfekte Korrelation?

- Liegen *alle Messwertpaare* (Punkte) auf einer *Geraden* (Steigung ≠ 0), so ist die Korrelation *perfekt* (r = +1 bzw. r = -1); die Gerade muss dabei *nicht* durch den Ursprung gehen und kann eine beliebige Steigung (≠0) aufweisen. Jedem x_i wird dabei genau ein y_i zugeordnet und umgekehrt.
- Dies *gilt auch für z-transformierte Werte*; z-transformierte Werte liegen auf einer Geraden, die durch den Ursprung geht und eine Steigung von 1 hat, wenn die Korrelation perfekt ist.
- Ist die Varianz bei einer oder beiden Variablen Null, so ist die Korrelation nicht definiert (da r = Kov(XY)/(sd(x)*sd(y))); manchmal wird aber dann auch r=0 angegeben.
- Eine perfekte Korrelation bedeutet *nicht*, dass die $x_i = y_i$.



perfekte Korrelation von Größe und Gewicht



dieselben Daten, z-standardisiert

Achtung – Lieblingsfehler

Wenn zwei Variablen korrelieren, heißt das nicht (unbedingt), dass es einen ursächlichen (kausalen) Zusammenhang gib!

- Die Anzahl der Störche pro Landkreis korreliert mit der Anzahl der Babies in diesem Landkreis.
- Trotzdem: Wer Kinder hat, kann aus eigener Erfahrung bestätigen, dass es Störche *nicht* die Ursache von Babies sind…
- Also: Auch wenn „Babies und Störche“ korrelieren, heißt das nicht (unbedingt), dass es es einen kausalen Zusammenhang gibt! Kann sein, muss aber nicht…
- Die Korrelation ist also eine notwendige, aber keine ausreichende Bedingung für einen Kausalzusammenhang.

Spielen Sie das Correlation Game!

Nette Sammlung weiterer Scheinkorrelationen: http://scheinkorrelation.jimdo.com

Guter TED-Vortrag zum Thema: https://www.youtube.com/watch?v=8B271L3NtAw

Abschluss

Hinweise

- Dieses Dokument steht unter der Lizenz CC-BY 3.0.
- Autor: Sebastian Sauer
- Für externe Links kann keine Haftung übernommen werden.
- Dieses Dokument entstand mit reichlicher Unterstützung vieler Kolleginnen und Kollegen aus der FOM. Vielen Dank!
- Dieses Dokument baut in Teilen auf auf dem Skript zu quantitative Methoden des ifes-Instituts der FOM-Hochschule.