3. Vorlesung Statistik I
Deskriptive Statistik II
Korrelation

Lineare Zusammenhänge zwischen Variablen

Vorbemerkung

- Bislang haben wir Maßzahlen kennengelernt, die der Beschreibung **einer** Variable dienen:
  - Lagemaße: Mittelwert, Median, Modus und Quantile.
  - Streuungsmaße: Varianz, Standardabweichung und Interquartilsabstand.

- Jetzt werden wir uns mit Maßzahlen beschäftigen, die den (linearen) **Zusammenhang zweier Variablen** beschreiben.
- Wir beschränken uns hierbei auf **metrische Variablen**.

Zusammenhang zweier Variablen I

- Was meinen wir damit, wenn wir sagen, dass zwei Variablen „zusammenhängen“?
- Variablen können auf verschiedene Arten miteinander zusammenhängen. In den meisten Fällen meinen wir damit aber, dass zwischen den beiden Variablen eine Beziehung der Form **„je ..., desto …“** besteht.

- Beispiele:
  - **Je** mehr eine Student*in lernt, **desto** mehr Punkte erreicht sie in der Klausur.

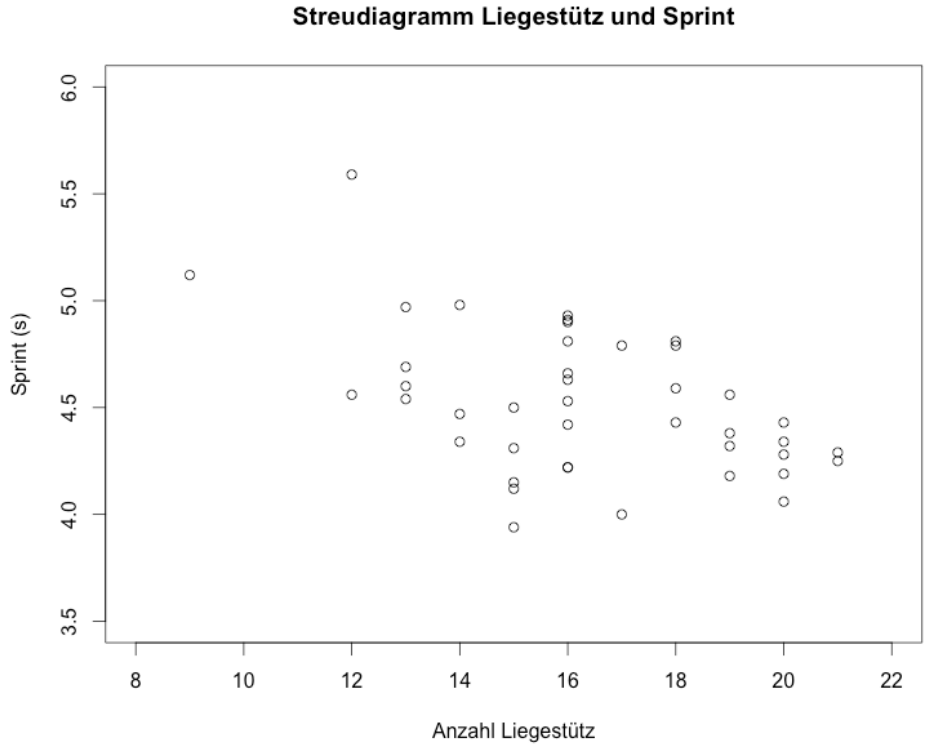
    bzw.:
  - **Je** weniger eine Student*in lernt, **desto** weniger Punkte erreicht sie in der Klausur.
  - **Je** mehr Liegestützen eine Schüler*in schafft, **desto** geringer ist ihre Sprintzeit.

    bzw.:
  - **Je** weniger Liegestützen eine Schüler*in schafft, **desto** höher ist ihre Sprintzeit.

- Hinweis: Ein häufiger Spezialfall von Zusammenhängen dieser Art sind **lineare Zusammenhänge**. Die in dieser Vorlesung behandelten Zusammenhangsmaße dienen streng genommen nur der Beschreibung von linearen Zusammenhängen.

Zusammenhang zweier Variablen II

- Die graphische Darstellung des Zusammenhangs zweier Variablen geschieht meist mithilfe eines **Streudiagramms**.
- Ein Streudiagramm stellt die Wertepaare (x_i, y_i) für $i = 1, \dots, n$ dar, wobei x, y zwei Variablen sind.
- Im Folgenden wird das Streudiagramm der Variablen „Anzahl Liegestütz“ und „Sprint“ dargestellt. Jedem Schüler kann ein Wertepaar (x_i, y_i) zugeordnet werden:



Zusammenhang zweier Variablen III

Wichtige Aspekte:

- **Richtung** des Zusammenhangs.
- **Stärke** des Zusammenhangs.
- **Einheitsunabhängigkeit** des Zusammenhangs.

Wichtige Aspekte

- Wichtige Aspekte:
  - **Richtung des Zusammenhangs.**
  - Stärke des Zusammenhangs.
  - Einheitsunabhängigkeit des Zusammenhangs.

Richtung des Zusammenhangs

- Wir können **gleichgerichtete** und **entgegengerichtete** Zusammenhänge unterscheiden.

- gleichgerichtete Zusammenhänge:
  - **Je höher** der Wert auf der einen Variable, **desto höher** der Wert auf der anderen Variable.

    bzw.:
  - **Je niedriger** der Wert auf der einen Variable, **desto niedriger** der Wert auf der anderen Variable.

- entgegengerichtete Zusammenhänge:
  - **Je höher** der Wert auf der einen Variable, **desto niedriger** der Wert auf der anderen Variable.

    bzw.:
  - **Je niedriger** der Wert auf der einen Variable, **desto höher** der Wert auf der anderen Variable.

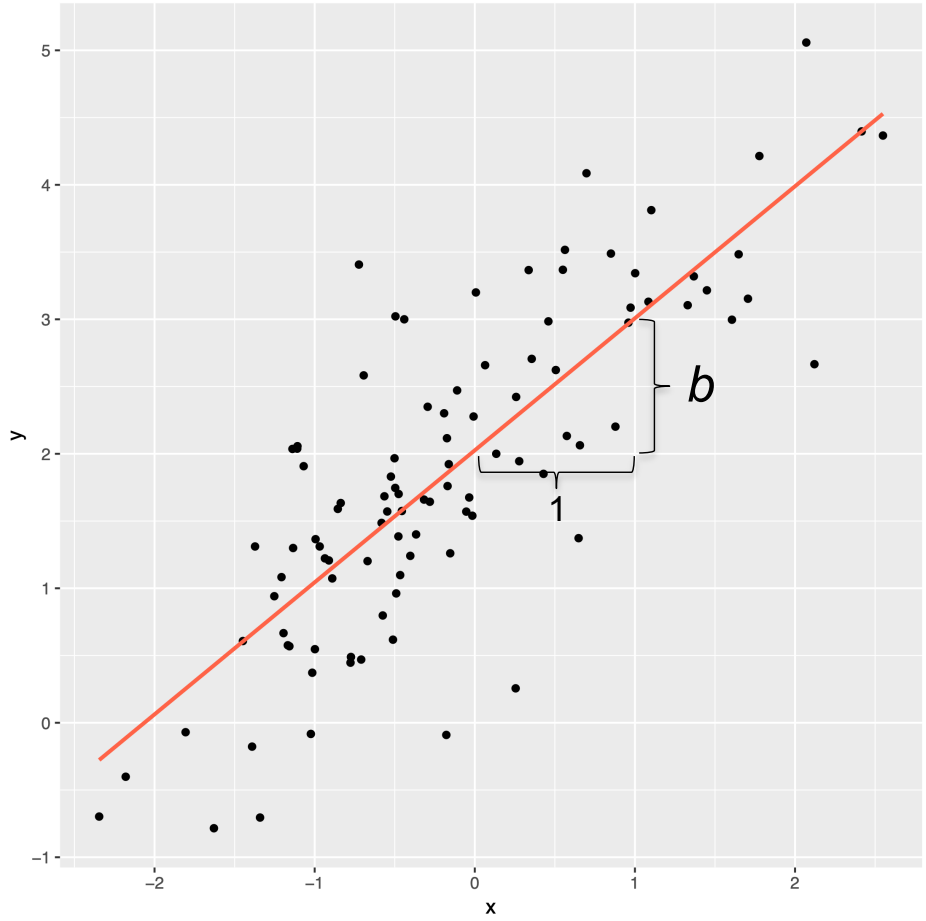
- Natürlich kann es auch sein, dass zwischen zwei Variablen überhaupt kein Zusammenhang besteht.

Wichtige Aspekte

- Wichtige Aspekte:
  - Richtung des Zusammenhangs.
  - **Stärke des Zusammenhangs.**
  - Einheitsunabhängigkeit des Zusammenhangs.

Stärke des Zusammenhangs I

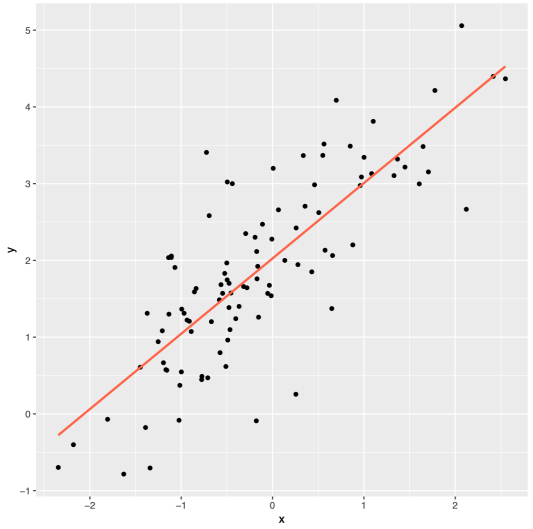
- Durch die Punktewolke jedes Streudiagramms lässt sich eine Gerade mit Steigung *b* ziehen, die die „Ausrichtung“ der Punktewolke bestmöglich beschreibt:

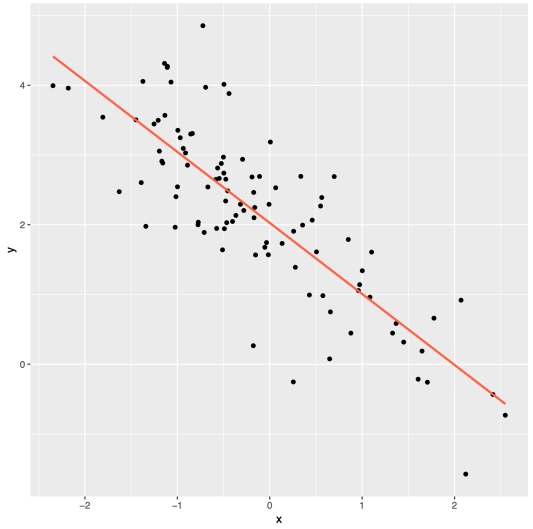


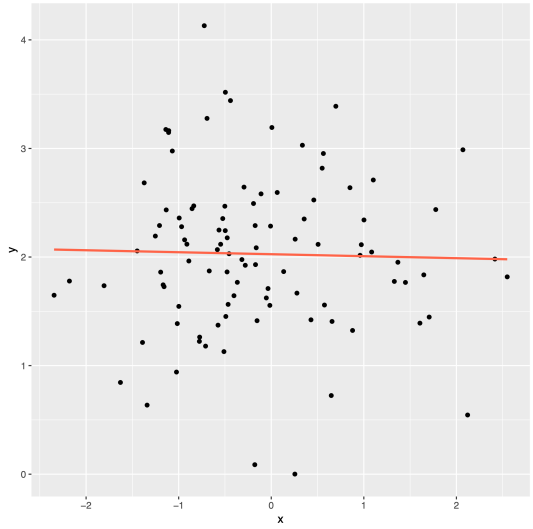
Stärke des Zusammenhangs II

- Bemerkung I: Was „bestmöglich“ genau bedeutet und wie man die Steigung der Gerade berechnen kann, werden wir erst in Statistik II besprechen.

- Bemerkung II: Bezug zur Richtung des Zusammenhangs:
  - Gleichgerichteter Zusammenhang → steigende Gerade / positive Steigung
  - Entgegengerichteter Zusammenhang → fallende Gerade / negative Steigung
  - Kein Zusammenhang → flache Gerade / Steigung nahe Null

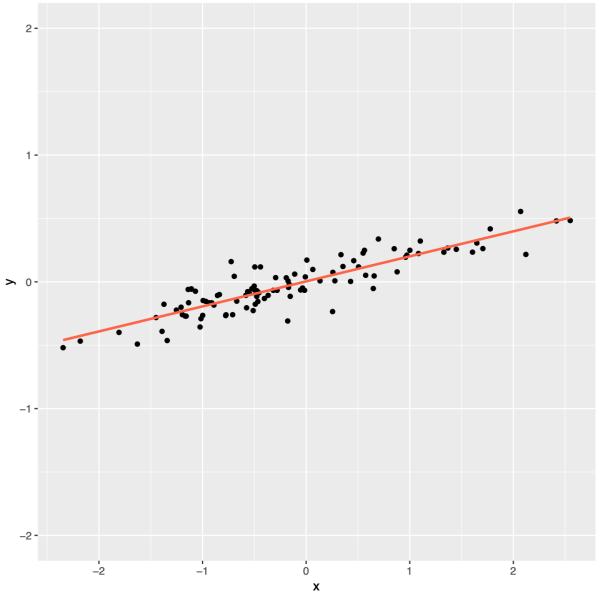


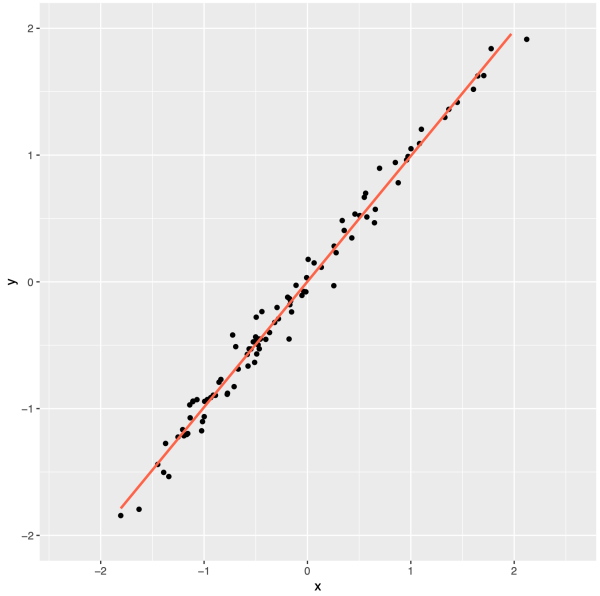


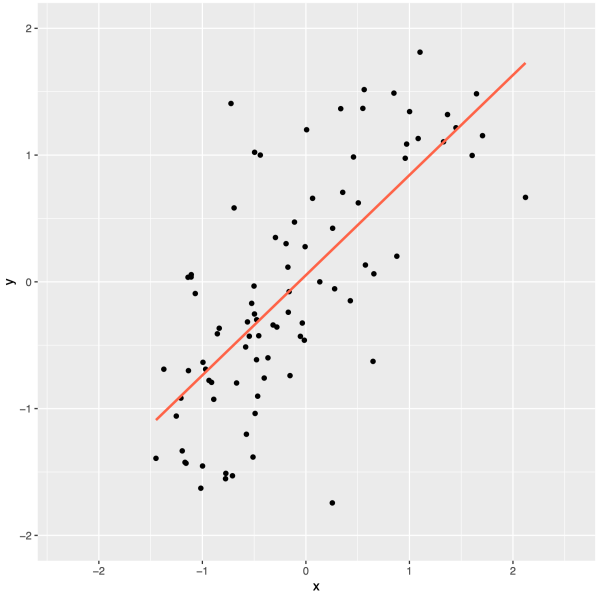


Stärke des Zusammenhangs III

Hinweis: Die Beispiele gehen davon aus, dass in jeder Grafik die gleichen Variablen abgebildet sind (z.B. Zusammenhang zwischen Motivation und Lerndauer in 3 Klassen)







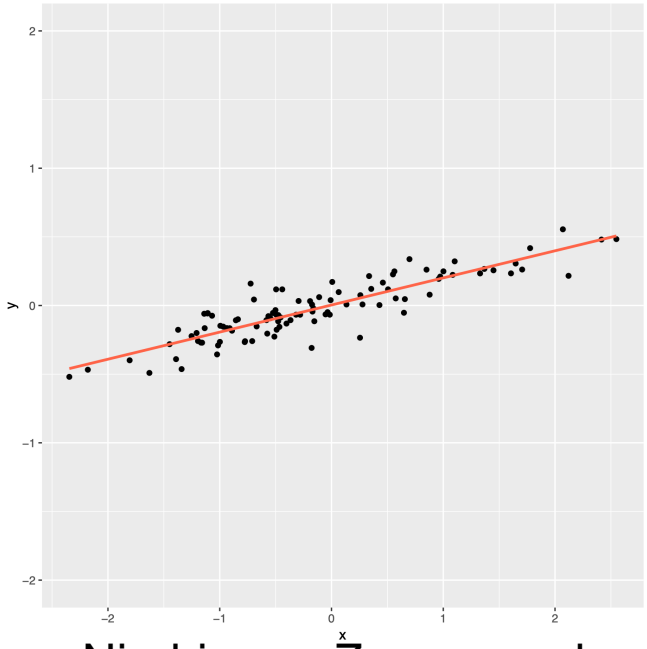
- Welches Streudiagramm zeigt den stärksten Zusammenhang?
- Welches den niedrigsten?
- Warum?

Stärke des Zusammenhangs IV

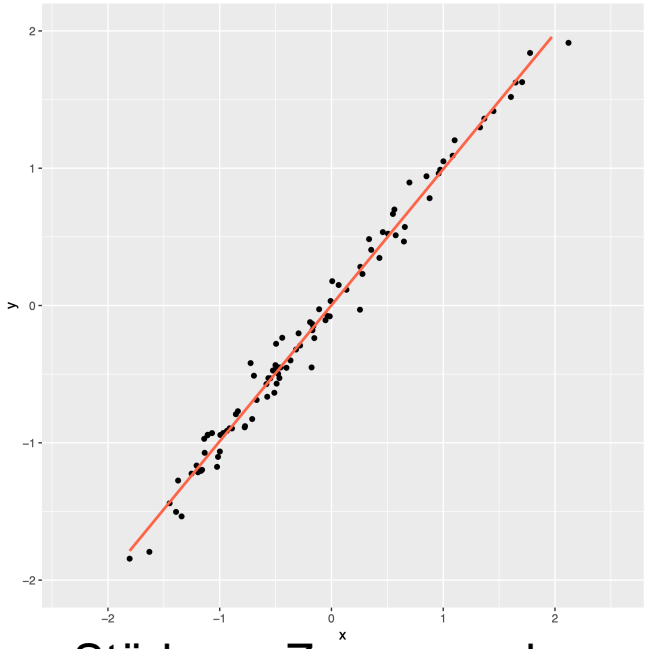
- Zwei Aspekte der Stärke des Zusammenhangs:
  - **Steigung** der durch das Streudiagramm gezogenen Geraden.
  - **Streuung** der Messwerte um diese Gerade.

Stärke des Zusammenhangs V

- Bei ähnlicher Streuung der Messwerte um die Gerade würden wir sagen, dass ein **umso stärkerer Zusammenhang** vorliegt, **je höher die Steigung** der Gerade im Betrag ist (gleiche Messeinheiten bei der Erfassung der Variablen vorausgesetzt).



Niedrigerer Zusammenhang

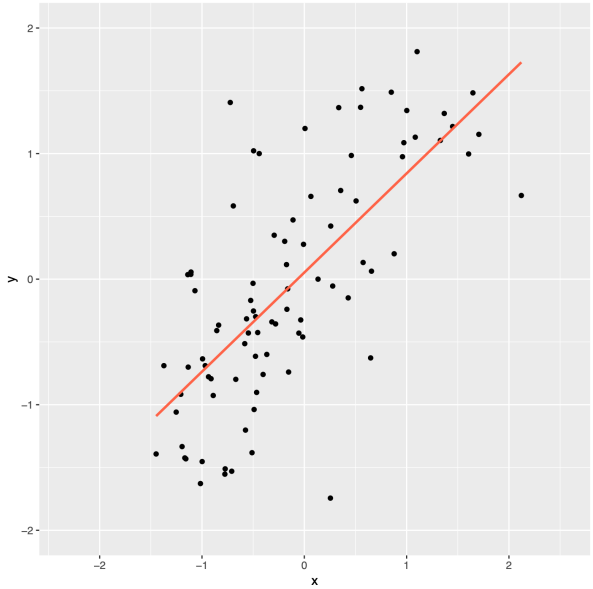


Stärkerer Zusammenhang

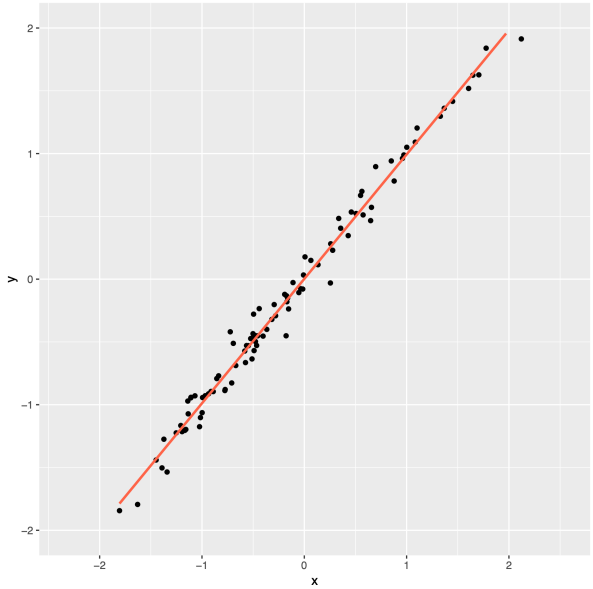
- Begründung: Eine Veränderung in der Variable x um 1 geht in diesem Fall mit einer vergleichsweise größeren Veränderung in y einher.
- Interessanter Extremfall: Steigung gleich Null → kein Zusammenhang

Stärke des Zusammenhangs VI

- Bei ähnlicher Steigung der Gerade würden wir sagen, dass ein **umso stärkerer Zusammenhang** vorliegt, **je geringer die Streuung** der Messwerte um die Gerade ist (gleiche Messeinheiten bei der Erfassung der Variablen vorausgesetzt).



Niedrigerer Zusammenhang



Stärkerer Zusammenhang

- Begründung: Bessere Vorhersage der Variablenwerte einzelner Merkmalsträger auf einer Variable bei Kenntnis ihrer Ausprägung auf der anderen Variable möglich.
- Interessanter Extremfall: Alle Punkte liegen auf der Gerade -> Perfekte Vorhersage für jede Person möglich.

Wichtige Aspekte

- Wichtige Aspekte:
  - Richtung des Zusammenhangs.
  - Stärke des Zusammenhangs.
  - **Einheitsunabhängigkeit des Zusammenhangs.**

Unabhängigkeit von der Einheit

- Wenn wir vom Zusammenhang zweier Variablen sprechen, gehen wir sinnvollerweise davon aus, dass die Richtung und die Stärke dieses Zusammenhangs nicht von der Einheit der Messinstrumente abhängt, mit der die beiden Variablen erfasst werden.

- Beispielsweise würden wir sagen, dass der Zusammenhang zwischen Körpergröße und Gewicht gleich bleibt, wenn wir die beiden Variablen in Zentimeter und Milligramm statt in Meter und Kilogramm erfassen.

Anforderungen an Maßzahlen

- Aus den bisherigen Überlegungen ergibt sich:

- Sinnvolle Maßzahlen, die den (linearen) Zusammenhang zwischen zwei Variablen beschreiben, sollten die folgenden Anforderungen erfüllen:

  - Sie sollten die **Richtung** des Zusammenhangs abbilden.

  - Sie sollten die **Stärke** des Zusammenhangs abbilden.

  - Sie sollten **unabhängig von der Einheit** der Variablen sein.

Zwischengliederung

- Bislang:
  - Lineare Zusammenhänge zwischen Variablen

- Jetzt:
  - Kovarianz

Kovarianz

Motivation

- Wie können wir die **Richtung** eines Zusammenhangs (gleichgerichtet vs. entgegengerichtet) durch das **Vorzeichen** einer Maßzahl (+ vs. -) ausdrücken?

- Zunächst betrachten wir das Wertepaar (x_i, y_i) einer einzelnen Merkmalsträger*in.

- Wir suchen eine Größe, die **positiv** ist, falls die beiden Messwerte x_i und y_i in die **gleiche Richtung** von ihrem jeweiligen Mittelwert abweichen und **negativ**, falls sie in die **entgegengesetzte Richtung** von ihrem jeweiligen Mittelwert abweichen.

- Mögliche Lösung:

$$L_i = (x_i - \bar{x})\,(y_i - \bar{y})$$

  - Falls x_i und y_i beide jeweils größer oder beide jeweils kleiner als ihre Mittelwerte sind, ist $L_i > 0$.
  - Falls x_i kleiner und y_i größer als der jeweilige Mittelwert ist oder x_i größer und y_i kleiner als ihre jeweiligen Mittelwerte sind, ist $L_i < 0$.

Kovarianz I

- Wie können wir aus den Größen L_i der einzelnen Merkmalsträger*innen eine Maßzahl für die gesamte Messwertreihe konstruieren?
- Idee: Bildung des Mittelwerts der L_i:

$$\frac{1}{n}\sum_{i=1}^{n} L_i$$

- Diese Maßzahl wird (empirische) **Kovarianz** genannt:

$$cov_{emp}(x,y) = \frac{1}{n}\sum_{i=1}^{n} L_i = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})$$

Kovarianz II

- Bemerkung I: Die Kovarianz ist **symmetrisch**, d.h. die Kovarianz von x und y entspricht der Kovarianz zwischen y und x:

$$cov_{emp}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = cov_{emp}(y,x)$$

- Bemerkung II: Die **Kovarianz einer Variable mit sich selbst** ist gleich der **Varianz der Variable**:

$$cov_{emp}(x,x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s_{emp}^2$$

- Bemerkung III: Der **Wertebereich** der Kovarianz ist **unbeschränkt**. Sie kann Werte zwischen $-\infty$ und $+\infty$ annehmen.

Zur Erinnerung

- Sinnvolle Maßzahlen, die den (linearen) Zusammenhang zwischen zwei Variablen beschreiben, sollten die folgenden Anforderungen erfüllen:
  - Sie sollten die **Richtung** des Zusammenhangs abbilden.
  - Sie sollten die **Stärke** des Zusammenhangs abbilden.
  - Sie sollten **unabhängig von der Einheit** der Variablen sein.

Kovarianz – Richtung des Zusammenhangs

$$cov_{emp}(x,y) = \frac{1}{n}\sum_{i=1}^{n} L_i = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})$$

- $cov_{emp}(x,y) > 0$: Die **positiven** L_i **überwiegen** die negativen L_i in der Summe $\sum_{i=1}^{n} L_i$. Es überwiegen **gleichgerichtete** Abweichungen vom Mittelwert.
- $cov_{emp}(x,y) < 0$: Die **negativen** L_i **überwiegen** die positiven L_i in der Summe $\sum_{i=1}^{n} L_i$. Es überwiegen **entgegengerichtete** Abweichungen vom Mittelwert.
- $cov_{emp}(x,y) = 0$: **Positive und negative** L_i **heben sich** in der Summe $\sum_{i=1}^{n} L_i$ **auf**. Es liegt **kein Zusammenhang** vor.
- Die Kovarianz drückt somit die Richtung des Zusammenhangs durch ihr Vorzeichen aus.

Kovarianz – Stärke des Zusammenhangs I

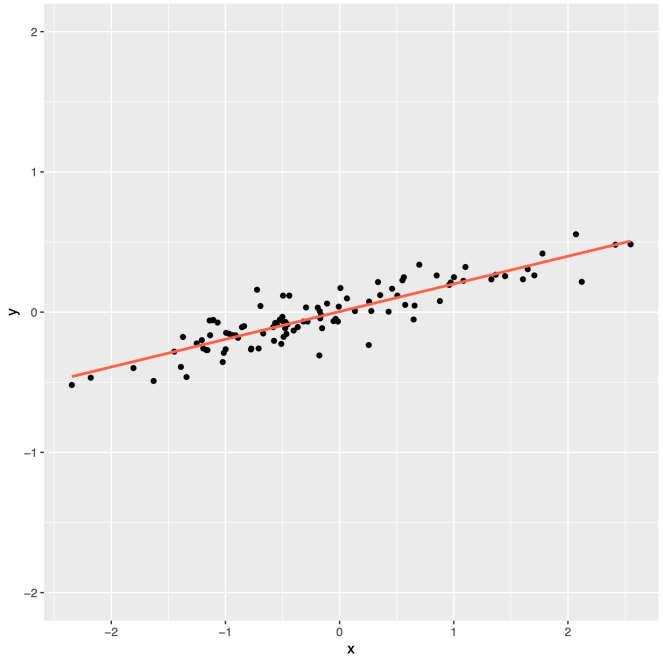
- Man kann mathematisch zeigen, dass folgender Zusammenhang zwischen der Steigung *b* der durch das Streudiagramm gezogenen Geraden und der Kovarianz besteht:

$$cov_{emp}(x, y) = b \cdot s^2_{x\,emp}$$

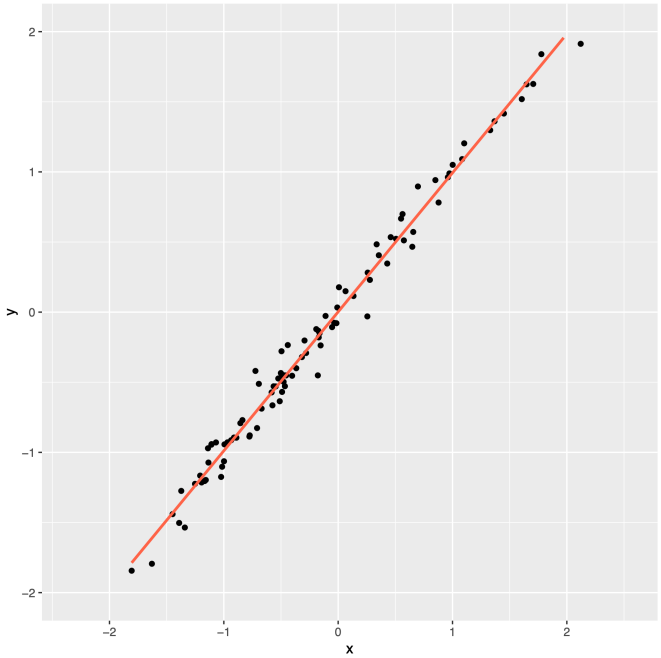
- An dieser Gleichung können zwei wichtige Eigenschaften der Kovarianz abgelesen werden:
  - Je höher die Steigung im Betrag, desto höher die Kovarianz im Betrag.
  - Bei höherer Streuung der Variablen um die Gerade bleibt die Kovarianz jedoch unverändert (bei gleicher Varianz der x-Variable).

- Das heißt:
  - Die Kovarianz kann nur einen Aspekt der Stärke des Zusammenhangs abbilden: Die Steigung der Gerade (und auch dies nur bei gleicher Einheit der Variablen).
  - Der andere Aspekt – die Streuung um die Gerade – wird durch die Kovarianz nicht berücksichtigt.
  - Vorsicht: Obwohl $cov_{emp}(x, y) = cov_{emp}(y, x)$ **muss nicht** gelten $b_{x \to y} = b_{y \to x}$.

Kovarianz – Stärke des Zusammenhangs II

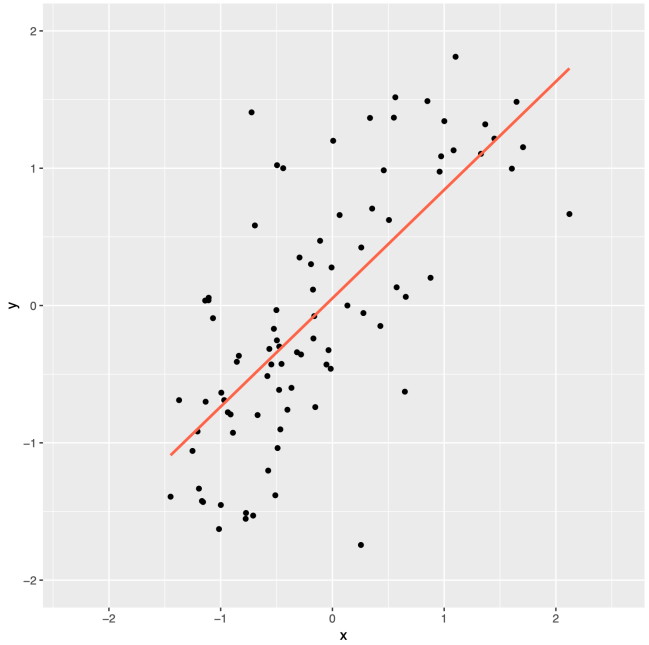


$cov_{emp}(x, y) = 0.20$

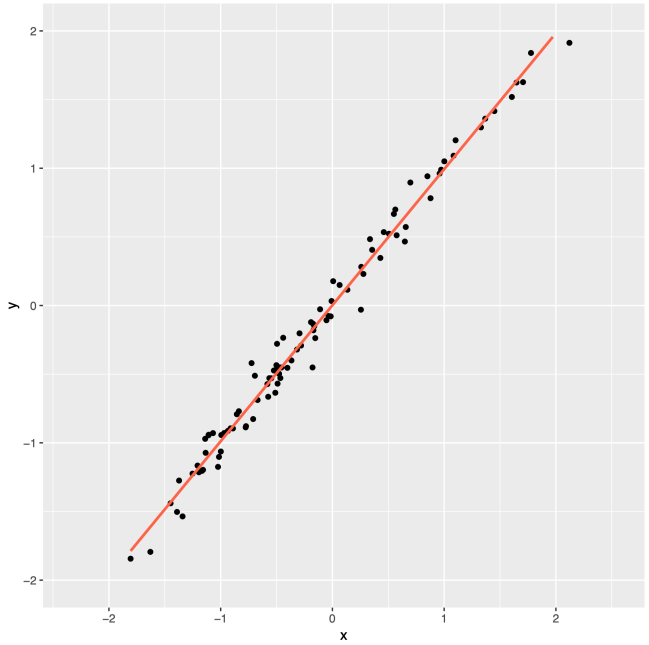


$cov_{emp}(x, y) = 1.01$

Kovarianz – Stärke des Zusammenhangs III



$cov_{emp}(x, y) = 1.00$



$cov_{emp}(x, y) = 1.01$

Kovarianz – Einheitsunabhängigkeit

- Die Höhe der Kovarianz hängt von der Einheit der Variablen ab.

- Ist beispielsweise die Kovarianz der Variablen Größe und Gewicht in Meter und Kilogramm gleich 10, dann ist die Kovarianz derselben Variablen nach Umrechnung in Zentimeter und Gramm gleich 1.000.000.

Fazit Kovarianz

- Sinnvolle Maßzahlen, die den (linearen) Zusammenhang zwischen zwei Variablen beschreiben, sollten die folgenden Anforderungen erfüllen:

  - Sie sollten die **Richtung** des Zusammenhangs abbilden.
    - ➢ Ist durch die Kovarianz erfüllt.

  - Sie sollten die **Stärke** des Zusammenhangs abbilden.
    - ➢ Ist durch die Kovarianz nicht (bzw. nur teilweise) erfüllt.

  - Sie sollten **unabhängig von der Einheit** der Variablen sein.
    - ➢ Ist durch die Kovarianz nicht erfüllt.

Zwischengliederung

- Bislang:
  - Lineare Zusammenhänge zwischen Variablen
  - Kovarianz

- Jetzt:
  - Standardisierung von Variablen
  - Korrelation

Standardisierung von Variablen

Standardisierung der Variablen

- Um das Problem der Einheitsabhängigkeit der Kovarianz zu lösen, werden wir die beiden Variablen **standardisieren**.

- Es wird sich zeigen, dass dadurch auch das Problem der fehlenden Abbildbarkeit der Stärke des Zusammenhangs gelöst wird.

- Standardisierung bedeutet, dass die einzelnen Messwerte derart transformiert werden, dass die resultierenden transformierten Messwerte einen vorgegebenen Mittelwert und eine vorgegebene Varianz aufweisen.

- Es gibt zahlreiche Möglichkeiten, Variablen zu standardisieren. Wir wählen die sogenannte **z-Standardisierung.**

z-Standardisierung - Definition

- Die Transformation der z-Standardisierung ist für jeden Messwert x_i wie folgt definiert:

$$z_i = \frac{x_i - \bar{x}}{s_{emp\,x}}$$

- Wir ziehen also von jedem Messwert x_i den Mittelwert $\bar{x}$ ab, teilen diese Differenz durch die Standardabweichung $s_{emp\,x}$ und erhalten die transformierten Messwerte z_i.
- Dies führt dazu, dass der Mittelwert der z-standardisierten Messwerte stets 0 und die Standardabweichung stets 1 ist:

$$\bar{z} = 0$$
$$s_{emp\,z} = 1$$

- Wichtig: Durch die z-Standardisierung ändert sich die Richtung der Abweichungen der einzelnen Messwerte vom jeweiligen Mittelwert nicht.
- Interpretation: Ein z-Wert gibt damit an, wie viele Standardabweichungen ein Wert vom Mittelwert abweicht.

z-Standardisierung - Beispiel

- Wir haben folgende Messwerte vorliegen:

$$x_1 = 1 \quad x_2 = 3 \quad x_3 = 2 \quad x_4 = 2$$

- Als Mittelwert dieser Messwerte ergibt sich $\bar{x} = 2$ und als Standardabweichung ergibt sich $s_{emp} \approx 0.71$

- Damit können wir alle Messwerte in z-standardisierte Werte umrechnen:

x_i	***Berechnung z_i***	***Interpretation***
$x_1 = 1$	$z_1 = \frac{x_1 - \bar{x}}{s_{emp}} = \frac{1-2}{0.71} \approx -1.41$	Der Wert x_1 weicht um -1.41 Standardabweichungen vom Mittelwert $\bar{x}$ ab.
$x_2 = 3$	$z_2 = \frac{x_2 - \bar{x}}{s_{emp}} = \frac{3-2}{0.71} \approx 1.41$	Der Wert x_2 weicht um 1.41 Standardabweichungen vom Mittelwert $\bar{x}$ ab.
$x_3 = 2$	$z_3 = \frac{x_3 - \bar{x}}{s_{emp}} = \frac{2-2}{0.71} = 0$	Der Wert x_3 weicht um 0 Standardabweichungen vom Mittelwert $\bar{x}$ ab.
$x_4 = 2$	$z_4 = \frac{x_4 - \bar{x}}{s_{emp}} = \frac{2-2}{0.71} = 0$	Der Wert x_4 weicht um 0 Standardabweichungen vom Mittelwert $\bar{x}$ ab.

Korrelation

Korrelation I

- Die (Pearson-) **Korrelation** der Variablen x und y kann nun als **Kovarianz der jeweils z-standardisierten Variablen** definiert werden:

$$r_{xy} = cov_{emp}(z_x, z_y) = \frac{1}{n}\sum_{i=1}^{n}(z_{x_i} - \bar{z}_x)(z_{y_i} - \bar{z}_y) = \frac{1}{n}\sum_{i=1}^{n} z_{x_i} \cdot z_{y_i}$$

- Einsetzen der Definition der z-Standardisierung ergibt die alternative Formel:

$$r_{xy} = \frac{1}{n}\sum_{i=1}^{n} z_{x_i} \cdot z_{y_i} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_{emp\,x}}\right)\left(\frac{y_i - \bar{y}}{s_{emp\,y}}\right) = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_{emp\,x} \cdot s_{emp\,y}}$$

Korrelation II

- Bemerkung I: Die Korrelation ist **symmetrisch**:

$$r_{xy} = \frac{1}{n}\sum_{i=1}^{n} z_{x_i} \cdot z_{y_i} = \frac{1}{n}\sum_{i=1}^{n} z_{y_i} \cdot z_{x_i} = r_{yx}$$

- Bemerkung II: Die Korrelation entspricht der Steigung b_z der Gerade durch das Streudiagramm der z-standardisierten Variablen:

$$r_{xy} = b_z$$

- Bemerkung III: Der **Wertebereich** der Korrelation ist **beschränkt**. Die Korrelation kann nur Werte zwischen -1 und 1 annehmen (-1 und 1 eingeschlossen).

Zur Erinnerung

- Sinnvolle Maßzahlen, die den (linearen) Zusammenhang zwischen zwei Variablen beschreiben, sollten die folgenden Anforderungen erfüllen:
  - Sie sollten die **Richtung** des Zusammenhangs abbilden.
  - Sie sollten die **Stärke** des Zusammenhangs abbilden.
  - Sie sollten **unabhängig von der Einheit** der Variablen sein.

Korrelation – Richtung des Zusammenhangs

- Aus der alternativen Formel und der Definition der Kovarianz ergibt sich der folgende Zusammenhang zwischen Korrelation und Kovarianz:

$$r_{xy} = \frac{cov_{emp}(x, y)}{s_{emp\,x} \cdot s_{emp\,y}}$$

- Da $s_{emp\,x}$ und $s_{emp\,y}$ stets positiv sind, hat die Korrelation das gleiche Vorzeichen wie die Kovarianz.

- Wie wir gesehen haben, kann am Vorzeichen der Kovarianz die Richtung des Zusammenhangs abgelesen werden. Diese Eigenschaft überträgt sich somit auf die Korrelation:

  $r_{xy} > 0$: Es überwiegen **gleichgerichtete** Abweichungen vom Mittelwert.

  $r_{xy} < 0$: Es überwiegen **entgegengerichtete** Abweichungen vom Mittelwert.

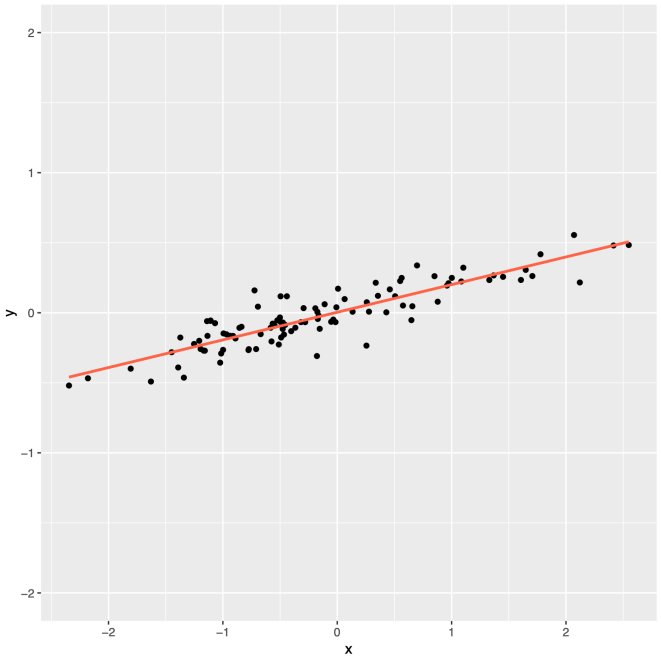
  $r_{xy} = 0$: Es liegt **kein Zusammenhang** vor.

- Die Korrelation drückt somit die Richtung des Zusammenhangs durch ihr Vorzeichen aus.

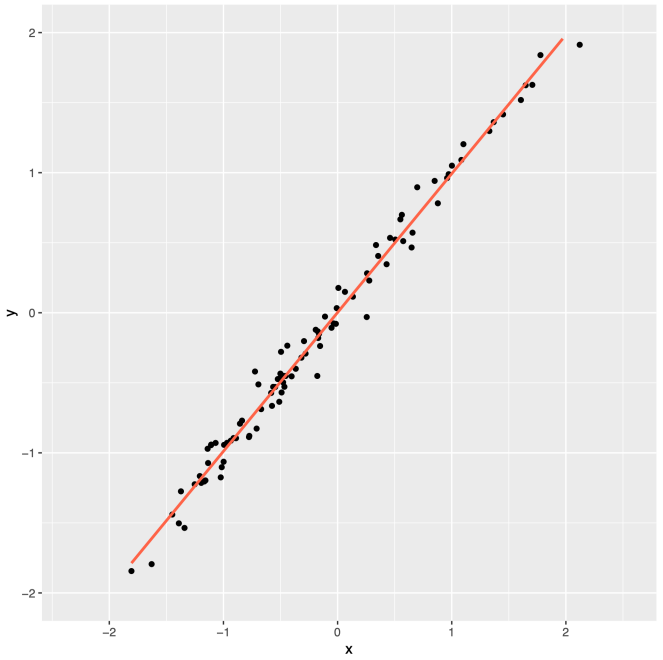
Korrelation – Stärke des Zusammenhangs I

- Man kann mathematisch zeigen, dass die Korrelation vom **Verhältnis** der **Steigung** der Gerade zur **Streuung** der Messwerte um die Gerade im Streudiagramm abhängt.
- Je größer dieses Verhältnis, desto höher fällt die Korrelation im Betrag aus.
- Das heißt, sowohl die Steigung der Gerade als auch die Streuung der Messwerte um die Gerade wird durch die Korrelation berücksichtigt.
- Extremfall: Liegen alle Punkte auf der Gerade, nimmt die Korrelation einen Wert von 1 bei positiver und einen Wert von -1 bei negativer Steigung an.
- Die Korrelation drückt somit die Stärke des Zusammenhangs durch ihre Höhe im Betrag aus: Je näher die Korrelation an 1 oder -1 liegt, desto stärker ist der Zusammenhang der beiden Variablen.

Korrelation – Stärke des Zusammenhangs II

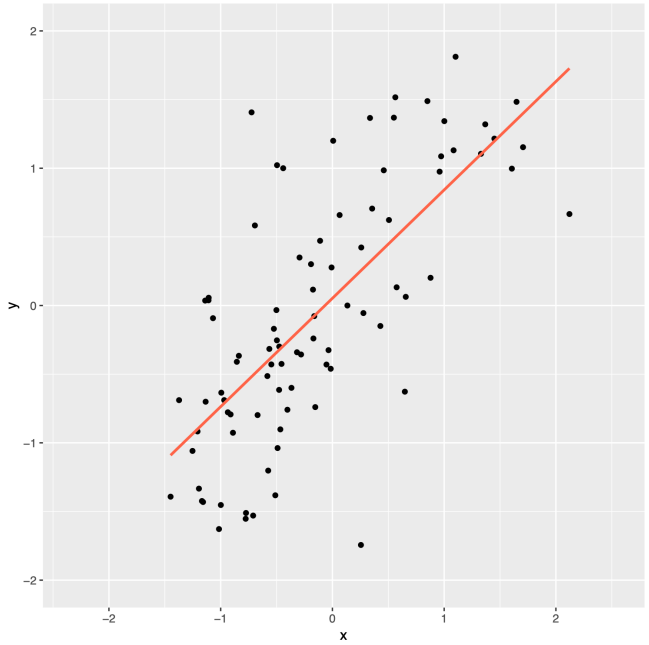


$r_{xy} = 0.89$

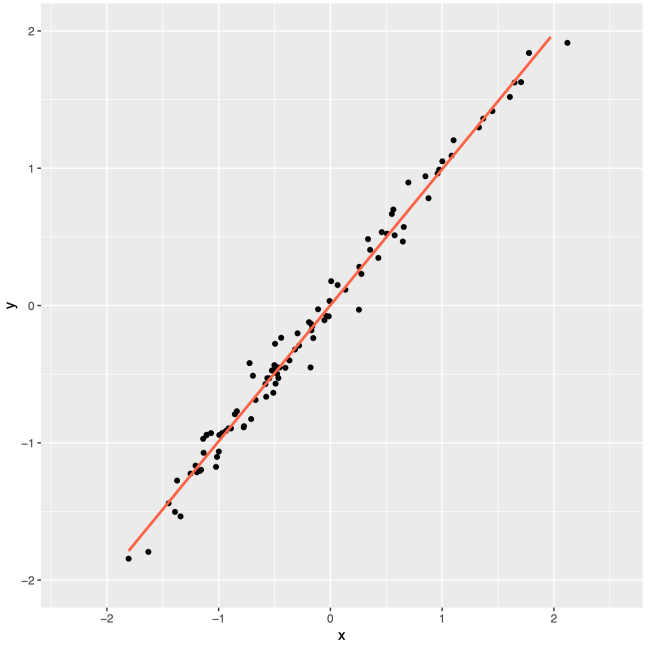


$r_{xy} = 0.99$

Korrelation – Stärke des Zusammenhangs III



$r_{xy} = 0.81$



$r_{xy} = 0.99$

Korrelation – Einheitsunabhängigkeit

- Die Höhe der Korrelation hängt nicht von der Einheit der Variablen ab.

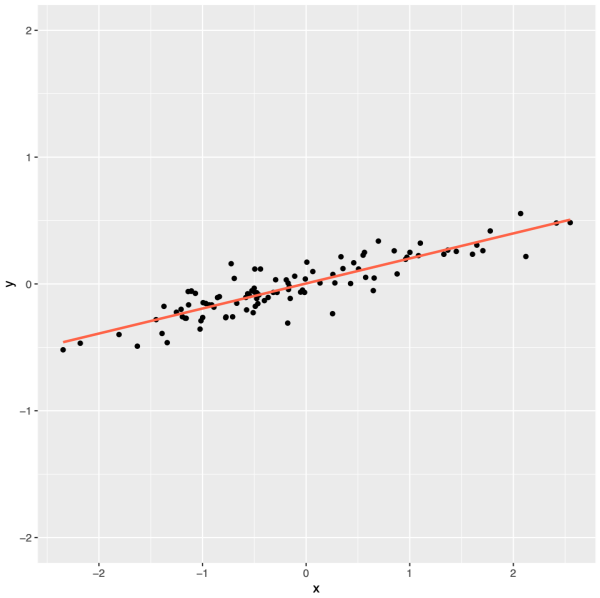
- Ist beispielsweise die Korrelation der Variablen Größe und Gewicht in Meter und Kilogramm gleich 0.9, dann ist die Korrelation derselben Variablen umgerechnet in Zentimeter und Gramm immer noch gleich 0.9.

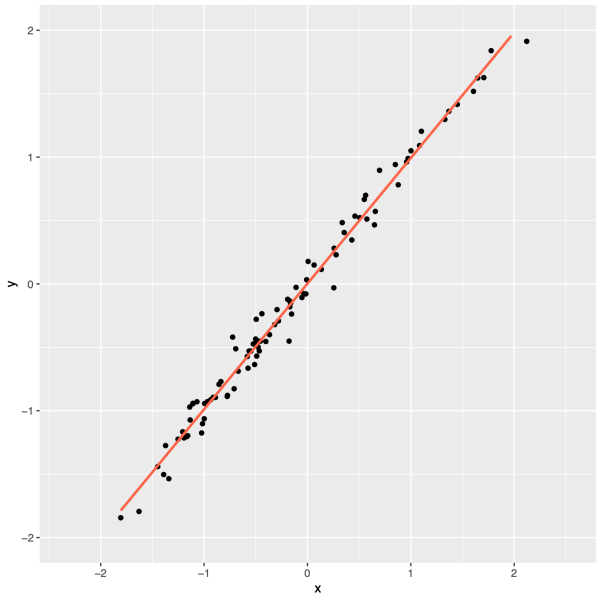
Fazit Korrelation

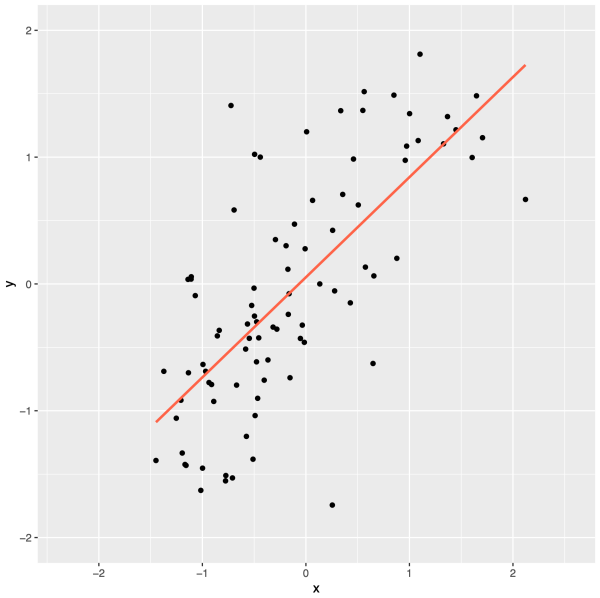
- Sinnvolle Maßzahlen, die den (linearen) Zusammenhang zwischen zwei Variablen beschreiben, sollten die folgenden Anforderungen erfüllen:

  - Sie sollten die **Richtung** des Zusammenhangs abbilden.
    - ➢ Ist durch die Korrelation erfüllt.
  - Sie sollten die **Stärke** des Zusammenhangs abbilden.
    - ➢ Ist durch die Korrelation erfüllt.
  - Sie sollten **unabhängig von der Einheit** der Variablen sein.
    - ➢ Ist durch die Korrelation erfüllt.

Nochmal zur Erinnerung







- Welches Streudiagramm zeigt den höchsten Zusammenhang?
- Korrelationen:
  - links: $r_{xy} = 0.89$
  - mittig: $r_{xy} = 0.99$
  - rechts: $r_{xy} = 0.81$

Weitere Korrelationsmaße (Auswahl)

- Neben der hier besprochenen Pearson-Korrelation, die den Zusammenhang zweier metrischer Variablen beschreibt, gibt es zahlreiche weitere Maßzahlen, die den Zusammenhang zwischen nicht-metrischen Variablen beschreiben, z.B:
  - **Phi-Koeffizient**: Zwei nominale Variablen
  - **Spearman-Rangkorrelation**: Zwei ordinale Variablen
  - **Punktbiseriale Korrelation**: Eine nominale und eine metrische Variable

- Da die Verwendung dieser Korrelationsmaße im Vergleich zur Pearson-Korrelation eher selten ist, verzichten wir hier auf ihre formale Darstellung und verweisen auf entsprechende Fachliteratur.

Zwischengliederung

- Bislang:
  - Lineare Zusammenhänge zwischen Variablen
  - Kovarianz
  - Standardisierung von Variablen
  - Korrelation

- Jetzt:
  - Häufige Fehler bei der Interpretation der Korrelation

Häufige Fehler bei der Interpretation der Korrelation

Häufige Fehler I

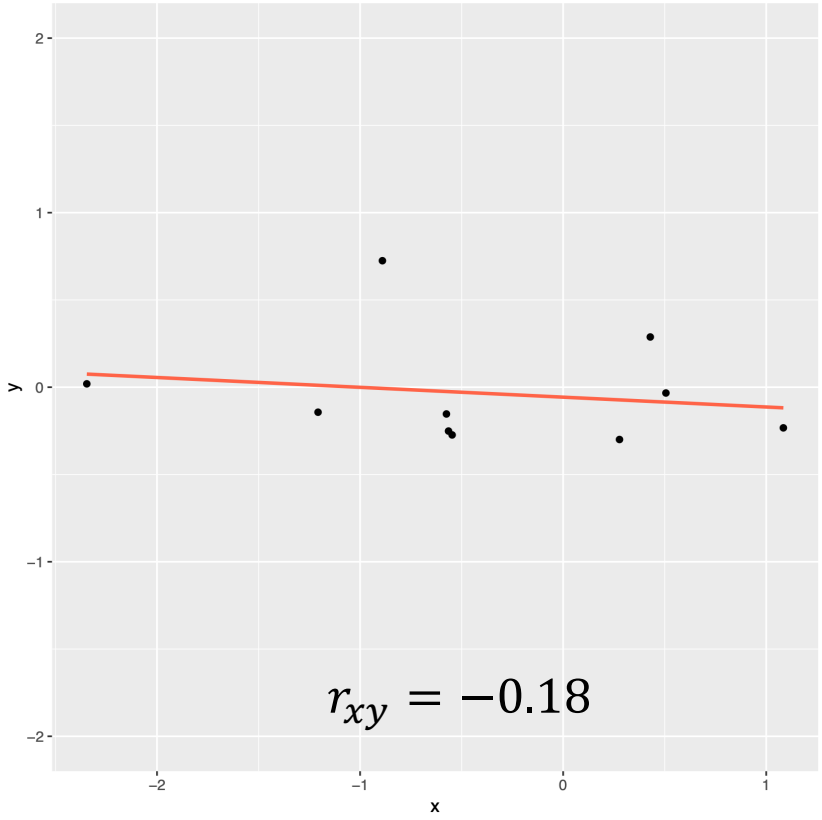
- Die Ausreißersensitivität der Korrelation wird nicht beachtet.
- Die Korrelation wird für die Beschreibung eines nonlinearen Zusammenhangs verwendet.
- Aus der Korrelation werden Aussagen über einzelne Personen abgeleitet.
- Aus der Korrelation werden kausale Aussagen abgeleitet.

Häufige Fehler II

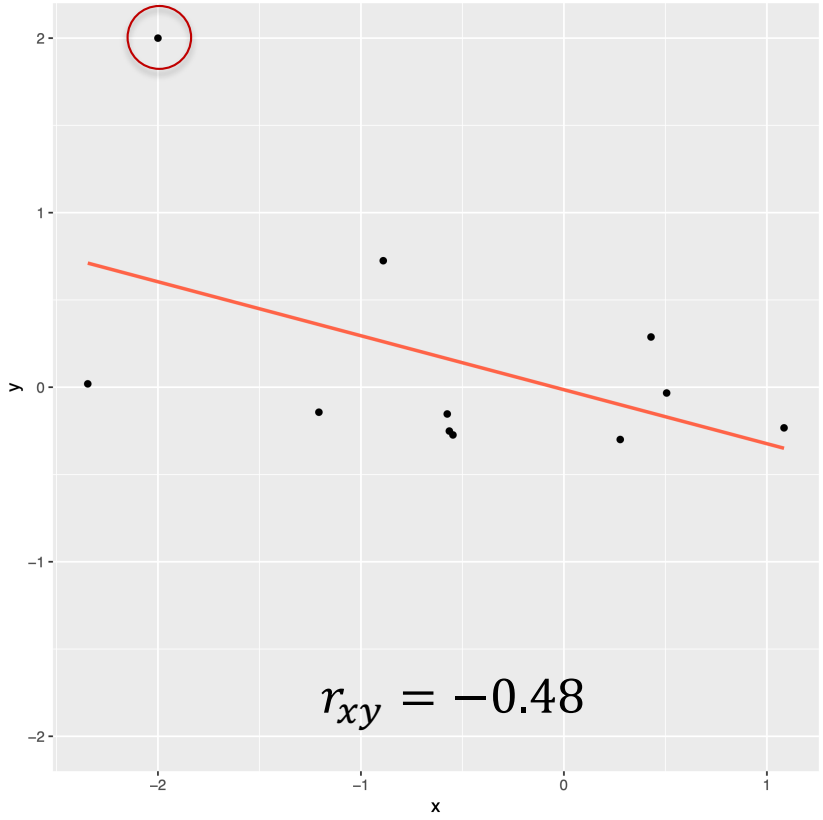
- **Die Ausreißersensitivität der Korrelation wird nicht beachtet.**
- Die Korrelation wird für die Beschreibung eines nonlinearen Zusammenhangs verwendet.
- Aus der Korrelation werden Aussagen über einzelne Personen abgeleitet.
- Aus der Korrelation werden kausale Aussagen abgeleitet.

Ausreißersensitivität der Korrelation I

- Ausreißer können die Korrelation (vor allem bei einer geringen Anzahl an Merkmalsträger*innen) stark verzerren:



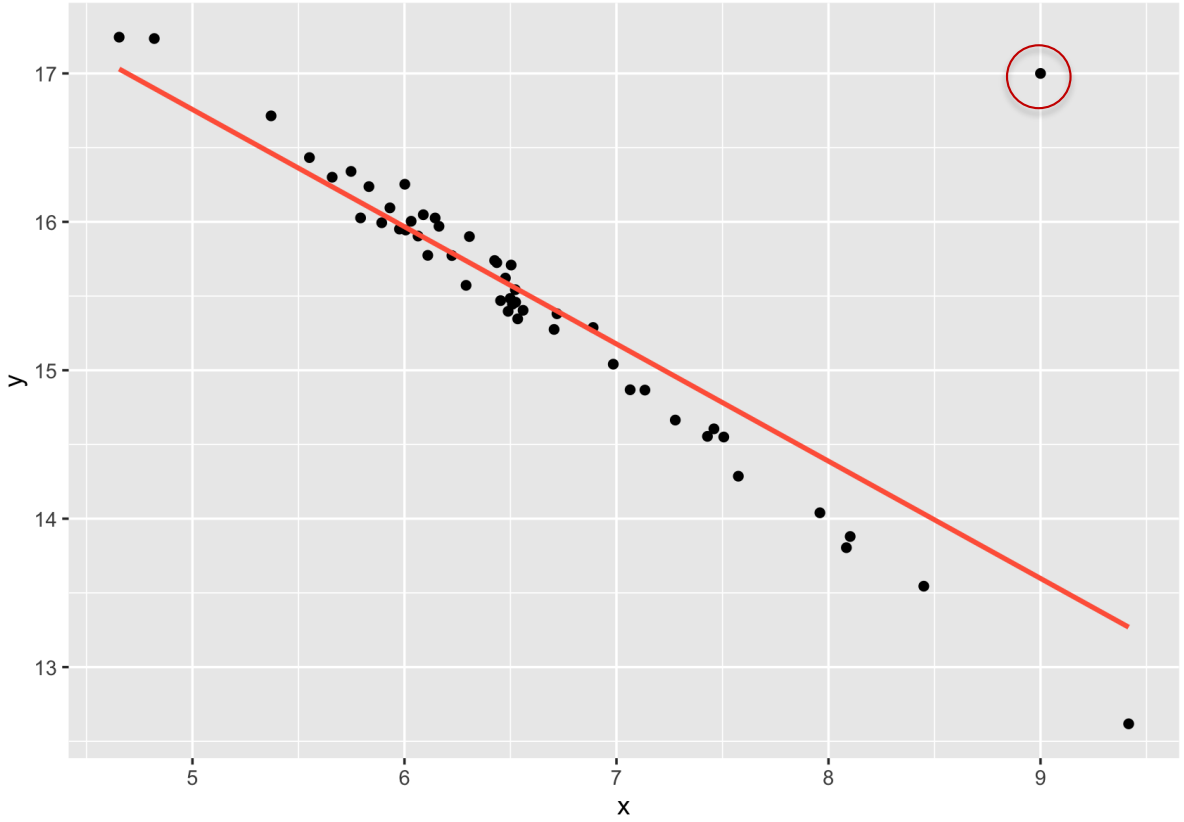
ohne Ausreißer



mit Ausreißer

Ausreißersensitivität der Korrelation II

- Wichtig: Ausreißer sollten nur aus den Daten entfernt werden, falls sich herausstellt, dass sie durch einen Dateneingabefehler entstanden sind (z.B. in Excel vertippt).
- Häufig ist es interessant, sich die Ausreißer genauer anzuschauen:



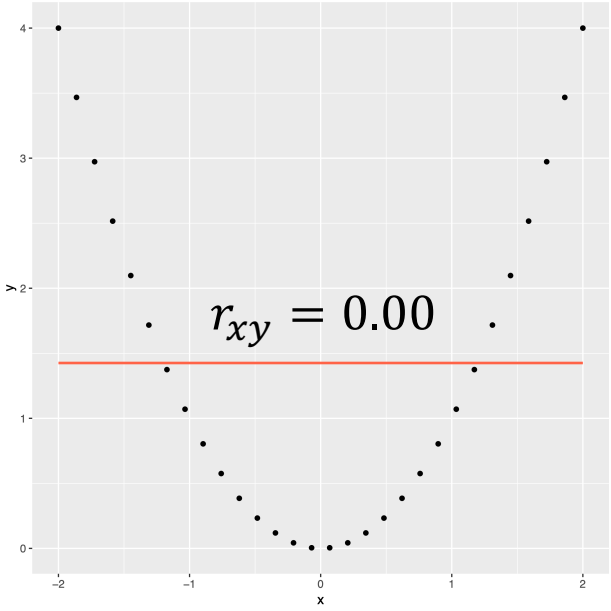
- Bevor man eine Korrelation inhaltlich interpretiert, sollte man sich also wenn möglich immer erst das Streudiagramm anschauen, um mögliche Ausreißer zu erkennen.

Häufige Fehler III

- Die Ausreißersensitivität der Korrelation wird nicht beachtet.
- **Die Korrelation wird für die Beschreibung eines nonlinearen Zusammenhangs verwendet.**
- Aus der Korrelation werden Aussagen über einzelne Personen abgeleitet.
- Aus der Korrelation werden kausale Aussagen abgeleitet.

Nonlineare Zusammenhänge

- Die Korrelation ist eine Maßzahl für den **linearen** Zusammenhang zweier Variablen (also für eine spezielle „je …, desto …“ - Beziehung).
- Nicht alle Arten von Zusammenhängen haben diese Form. Es können auch **nonlineare** Zusammenhänge zwischen Variablen bestehen. Häufig ist dies leicht im Streudiagramm zu sehen.
- Liegt zwischen zwei Variablen ein nonlinearer Zusammenhang vor, ist die Korrelation nicht die geeignete Maßzahl. Beispielsweise ist die Korrelation bei einem perfekten quadratischen Zusammenhang gleich 0:



- Bevor man eine Korrelation inhaltlich interpretiert, sollte man sich also wenn möglich immer erst das Streudiagramm anschauen, um zu erkennen, falls ein nonlinearer Zusammenhang vorliegt.

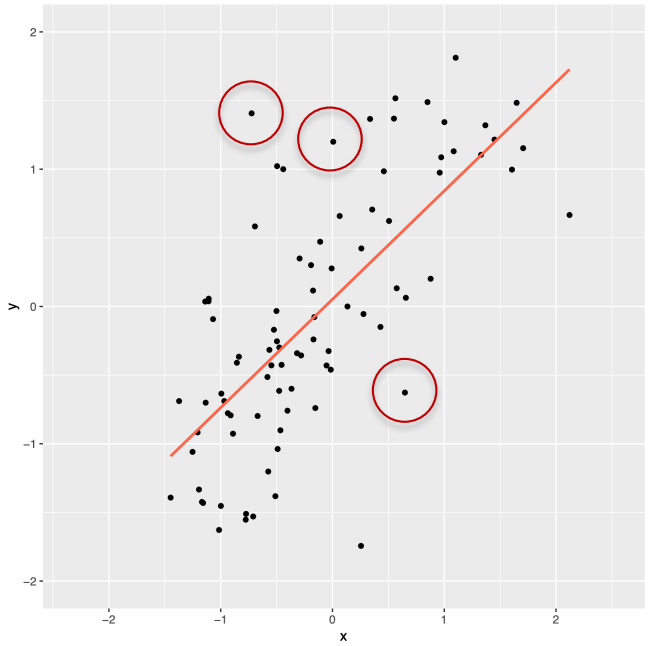
https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Häufige Fehler IV

- Die Ausreißersensitivität der Korrelation wird nicht beachtet.
- Die Korrelation wird für die Beschreibung eines nonlinearen Zusammenhangs verwendet.
- **Aus der Korrelation werden Aussagen über einzelne Personen abgeleitet.**
- Aus der Korrelation werden kausale Aussagen abgeleitet.

Aussagen über einzelne Personen

- Die Korrelation ist ein Maß, dass sich auf eine **Reihe von Messwerten** bezieht.
- Sie gibt hierbei lediglich die **durchschnittlichen** gleich- bzw. entgegengerichteten Abweichungen der Messwerte vom Mittelwert an.
- Es können daher auf der Basis der Korrelation **keine direkten Aussagen über die Abweichungen einzelner Personen** getroffen werden:

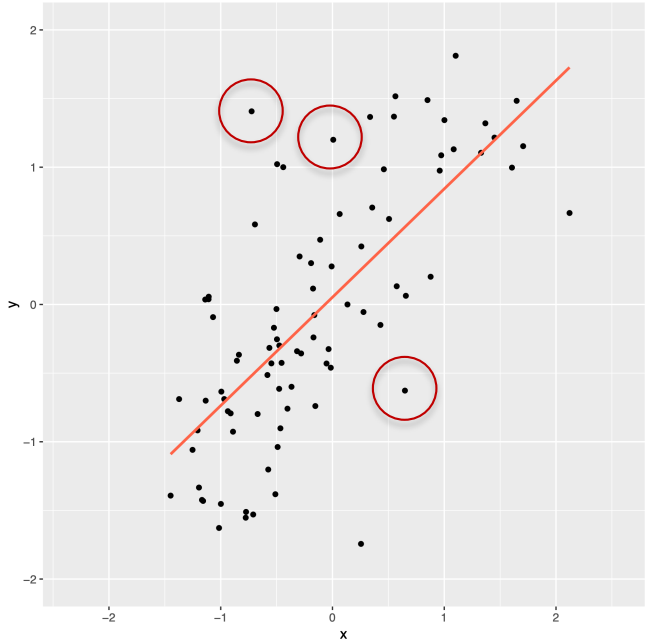


- Unrealistische Ausnahme: Alle Punkte liegen auf der Geraden.

Aussagen über einzelne Personen II

- Um zu betonen, dass keine direkten Aussagen über die Abweichungen einzelner Personen getroffen werden können, könnten wir bei der Beschreibung einer Korrelation sprachlich präziser sein, zum Beispiel:

  „Je mehr eine Student*in lernt, desto mehr Punkte erreicht sie **im Mittel** in der Klausur.“



Häufige Fehler V

- Die Ausreißersensitivität der Korrelation wird nicht beachtet.
- Die Korrelation wird für die Beschreibung eines nonlinearen Zusammenhangs verwendet.
- Aus der Korrelation werden Aussagen über einzelne Personen abgeleitet.
- **Aus der Korrelation werden kausale Aussagen abgeleitet.**

Korrelation und Kausalität I

- **Auf Basis einer Korrelation kann (ohne starke zusätzliche Annahmen) keine Aussage über einen Ursache-Wirkung-Zusammenhang getroffen werden.**
- Es existieren stets mehrere Erklärungen für den beobachteten Zusammenhang, die von unterschiedlichen kausalen Zusammenhängen ausgehen.
- Ausnahme: Experimenteller Versuchsaufbau oder explizite kausale Annahmen.
- Diese Themen sind Gegenstand der „kausalen Inferenz“, mit der wir uns in Statistik 2 genauer beschäftigen werden.

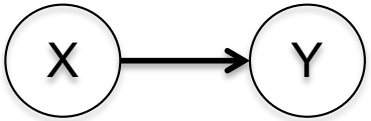
Korrelation und Kausalität II

- Weil “je ..., desto …“ immer ein bisschen nach Kausalität klingt, könnten wir bei der Beschreibung einer Korrelation sprachlich präziser sein, zum Beispiel:

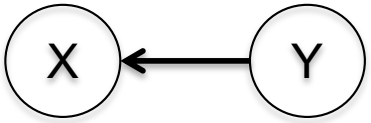
  „Für Studierende die mehr gelernt haben, **beobachten wir** im Mittel mehr Punkte in der Klausur **(und umgekehrt)**.“

Korrelation und Kausalität III

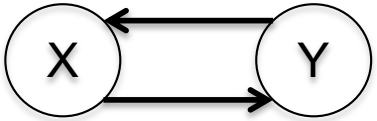
- Im Folgenden werden beispielhaft fünf mögliche kausale Erklärungen für eine beobachtete Korrelation zwischen den Variablen x und y dargestellt:



x beeinflusst y.



y beeinflusst x.

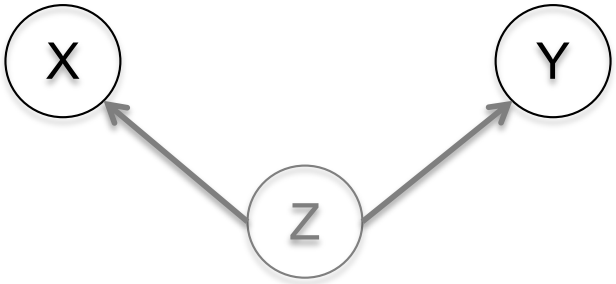


x und y beeinflussen sich (über die Zeit hinweg) wechselseitig.

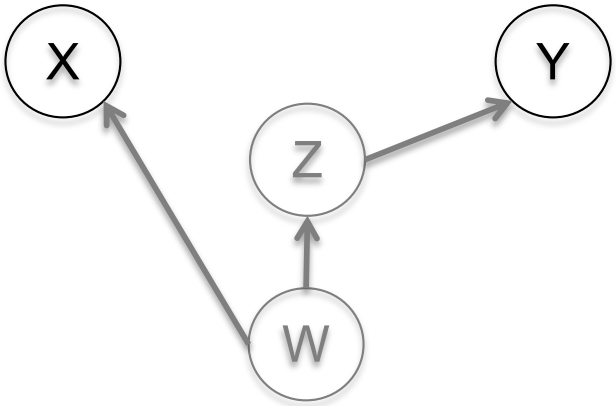
z.B.

x = depressive Stimmung
y = Schlafmangel

Korrelation und Kausalität IV



x und y werden durch eine Drittvariable z beeinflusst.



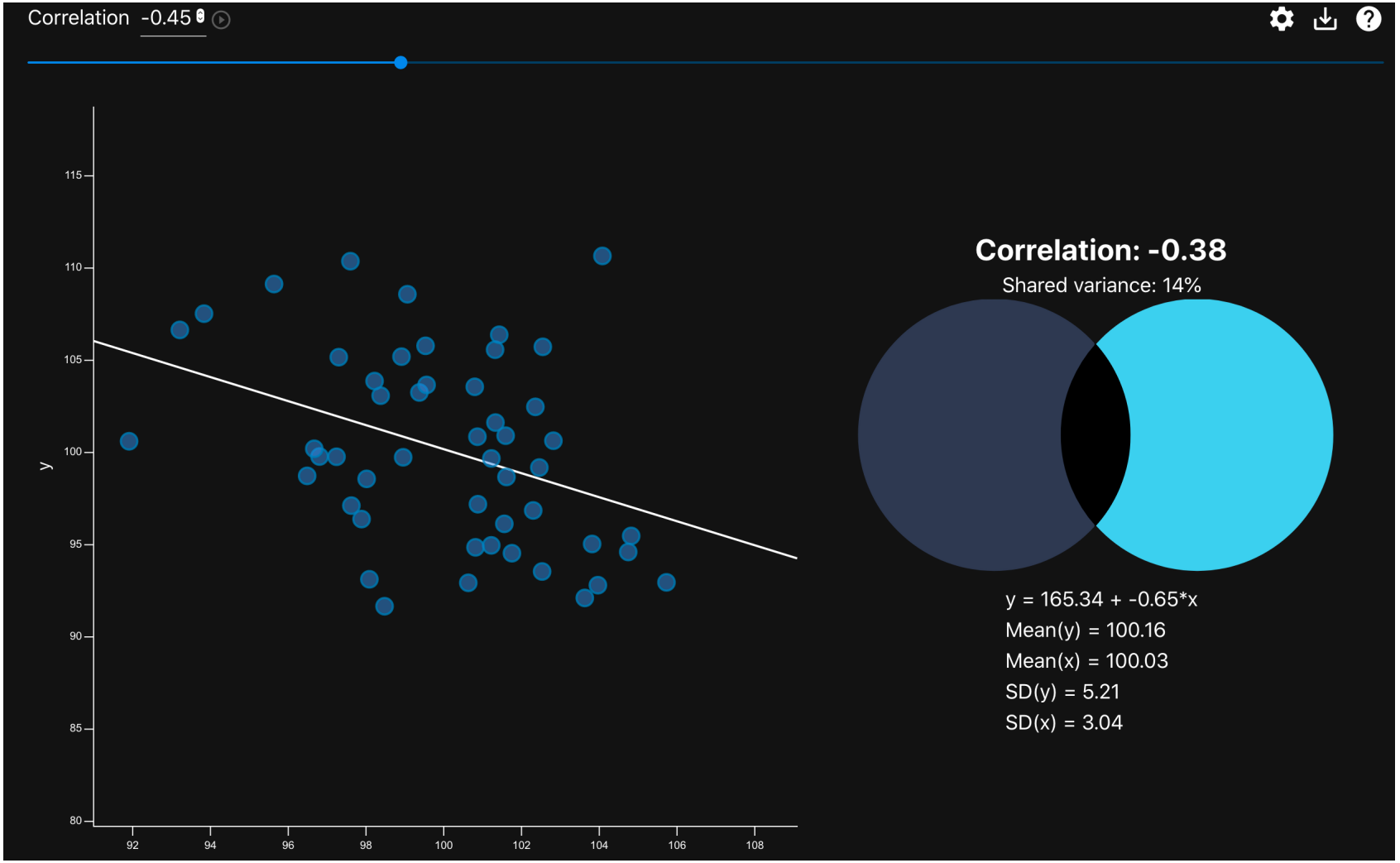
Eine vierte Variable w beeinflusst y über z indirekt und x direkt.

z.B.

x = depressive Stimmung
y = Schlafmangel
z = neurobiologische Veränderung
w = Stress

Ein Gefühl für Korrelationen bekommen

https://rpsychologist.com/correlation/



https://www.guessthecorrelation.com/

Zusammenfassung

- Lineare Zusammenhänge zwischen metrischen Variablen lassen sich mit der **Kovarianz** oder der **Korrelation** beschreiben.
- Kovarianz und Korrelation sind bei **gleichgerichteten** Zusammenhängen positiv, bei **gegengerichteten** Zusammenhängen negativ.
- Die **Stärke** eines Zusammenhangs bezieht sich sowohl auf die **Steigung** einer gedachten Geraden durch die Punktewolke, als auch auf die durchschnittliche **Abweichung** der Punkte von dieser Geraden.
- Die Kovarianz ist dabei jedoch **einheitsabhängig** und kann die Stärke des Zusammenhangs **nur unvollständig** wiederspiegeln.
- Die Korrelation liegt im Wertebereich von -1 bis +1 und gibt sowohl die **Richtung** als auch die **Stärke** des Zusammenhangs **einheitsunabhängig** an.
- Mit der z-Standardisierung können wir Variablen in **einheitsunabhängige** z-Werte transformieren. Z-Werte geben dann an, wie viele Standardabweichungen über (positiv) oder unter (negativ) dem Mittelwert ein einzelner Variablenwert liegt.
- Auch wenn Kausalität zu einer Korrelation führen kann, **bedeutet eine Korrelation umgekehrt nicht automatisch, dass eine direkte Kausalität vorliegt**.