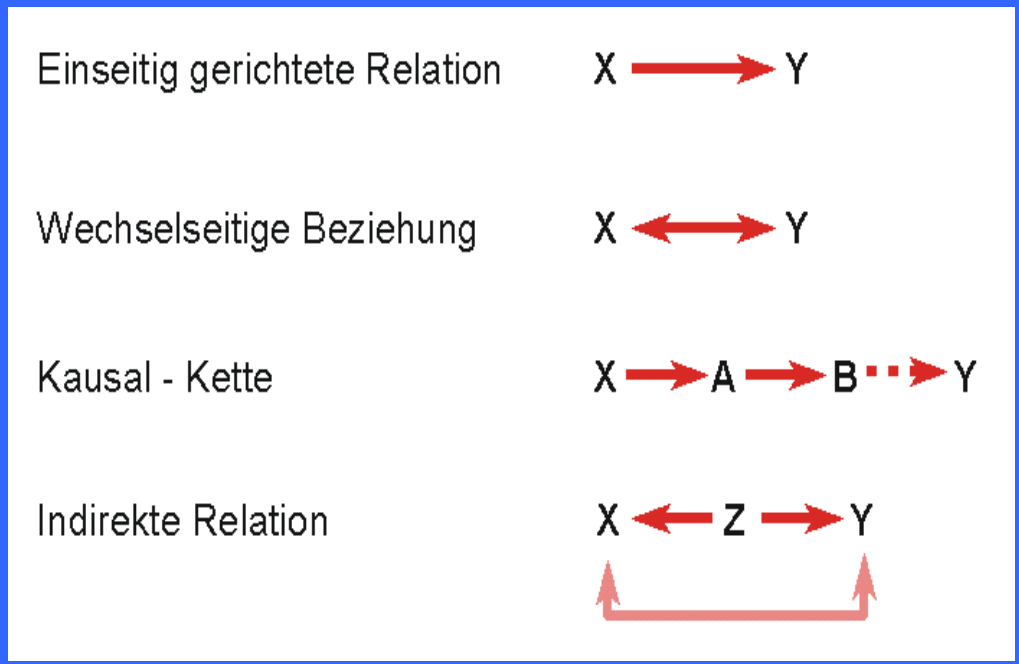
Statistik

- Grundlagen
- Charakterisierung von Verteilungen
- Einführung Wahrscheinlichkeitsrechnung
- Wahrscheinlichkeitsverteilungen
- Schätzen und Testen
- Korrelation
- Regression

Einführung

Die Analyse und modellhafte Beschreibung von Zusammenhängen zwischen zwei oder mehreren Variablen spielt in den Geowissenschaften eine sehr wichtige Rolle. Hat man einen statistisch signifikanten Zusammenhang zwischen zwei Zufallsvariablen X und Y gefunden, so ergibt sich zwangsläufig die Frage nach einer kausalen Beziehung zwischen beiden. Solche kausalen Beziehungen können ganz unterschiedlicher Art sein.



Kausalität

• Einseitig gerichtete Relation: Mit zunehmender Höhe über dem Meeresspiegel sinkt im allgemeinen die Lufttemperatur.

• Wechselseitige Beziehung (Rückkopplung): Auf vegetationsbedeckten Oberflächen wird die Bodenerosion vermindert, andererseits führt eine geringere Bodenerosion auch zu einer Verdichtung der Pflanzendecke.

• Kausal-Kette: Niederschlag führt zur Infiltration von Niederschlagswasser in den Boden, dies erhöht den Bodenwassergehalt, was wiederum zu einer verstärkten Perkolation in das Grundwasser und damit zur Grundwasserneubildung führt.

• Indirekte Relation, Scheinkorrelation: Die Höhenlage einer Messstation hat einen positiven Effekt auf die dort gemessene Niederschlagsmenge und zugleich einen negativen Effekt auf die Lufttemperatur; daraus resultiert eine indirekte (negative) Relation zwischen Niederschlag und Temperatur, die direkt (d.h. physikalisch) nicht vorhanden ist.

Grundbegriffe

Korrelationsanalyse: Sie dient nun dazu, Richtung und Stärke eines Zusammenhanges zwischen zwei oder mehr Variablen festzustellen. Die Korrelation betrachtet dabei im Gegensatz zur Regressionsanalyse beide Variablen gleichberechtigt. Man kann jedoch nicht eine kausale Beziehung beweisen oder im Sinne von Ursache-Wirkung erklären.

Korrelationskoeffizient: Die Stärke der linearen Assoziation der (kontinuierlichen) Variablen wird durch einen Korrelationskoeffizienten R festgestellt.

- Dazu werden beide Variablen X und Y, die eventuell in unterschiedlichen Maßeinheiten gemessen wurden, durch eine Z-Transformation standardisiert und vergleichbar gemacht.
- Der Wertebereich des Korrelationskoeffizienten liegt im Intervall [-1;1] mit:

  R = 0: es existiert kein linearer Zusammenhang.

  R = -1: es existiert ein perfekter negativer linearer Zusammenhang, je größer X wird, desto kleiner wird Y und umgekehrt.

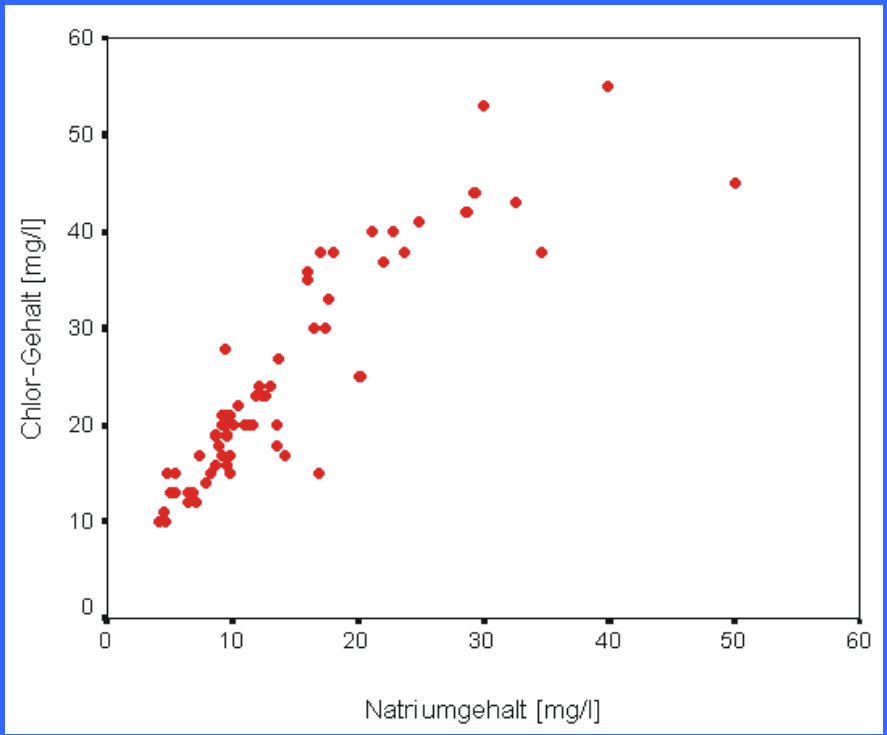
  R = 1: es existiert ein perfekter positiver linearer Zusammenhang, je größer X wird, desto größer wird auch Y und umgekehrt.

Grundbegriffe

Eigenschaften:

• Stehen beide Variablen allerdings in einem tatsächlich zu beobachtenden nicht-linearen Zusammenhang, so ist es trotzdem möglich, dass der Korrelationskoeffizient gleich 0 ist.

• Es ist ratsam, die Untersuchung der gemeinsamen Verteilung zweier stetiger Merkmale mit der Zeichnung einer Punktwolke (Scatterplot) zu beginnen, die Informationen auf einen Blick liefert.



Grundbegriffe

Eigenschaften:

• Bei der Korrelationsanalyse ist die Skalierung der beteiligten Zufallsvariablen (ZV) von wesentlicher Bedeutung (etwa beide nominalskaliert oder verschieden skalierte ZV, z.B. X ist metrisch, Y ist ordinal).

• Bei verschieden skalierten ZV kann man sich damit behelfen, die höherrangig skalierte ZV zunächst herab zu skalieren, um anschließend einen Korrelationskoeffizienten für die dann gleichartig skalierten ZV zu verwenden. Dadurch verliert man allerdings Informationen.

• Deshalb sind spezielle Korrelationskoeffizienten für verschieden skalierte ZV entwickelt worden.

Korrelationskoeffizienten

Produkt-Moment-Korrelationskoeffizient nach Pearson: Dieser Korrelationskoeffizient kann für normalverteilte Variablen verwendet werden, die mindestens intervallskaliert sind. Er berechnet sich zu:

$$R = \frac{S_{xy}}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

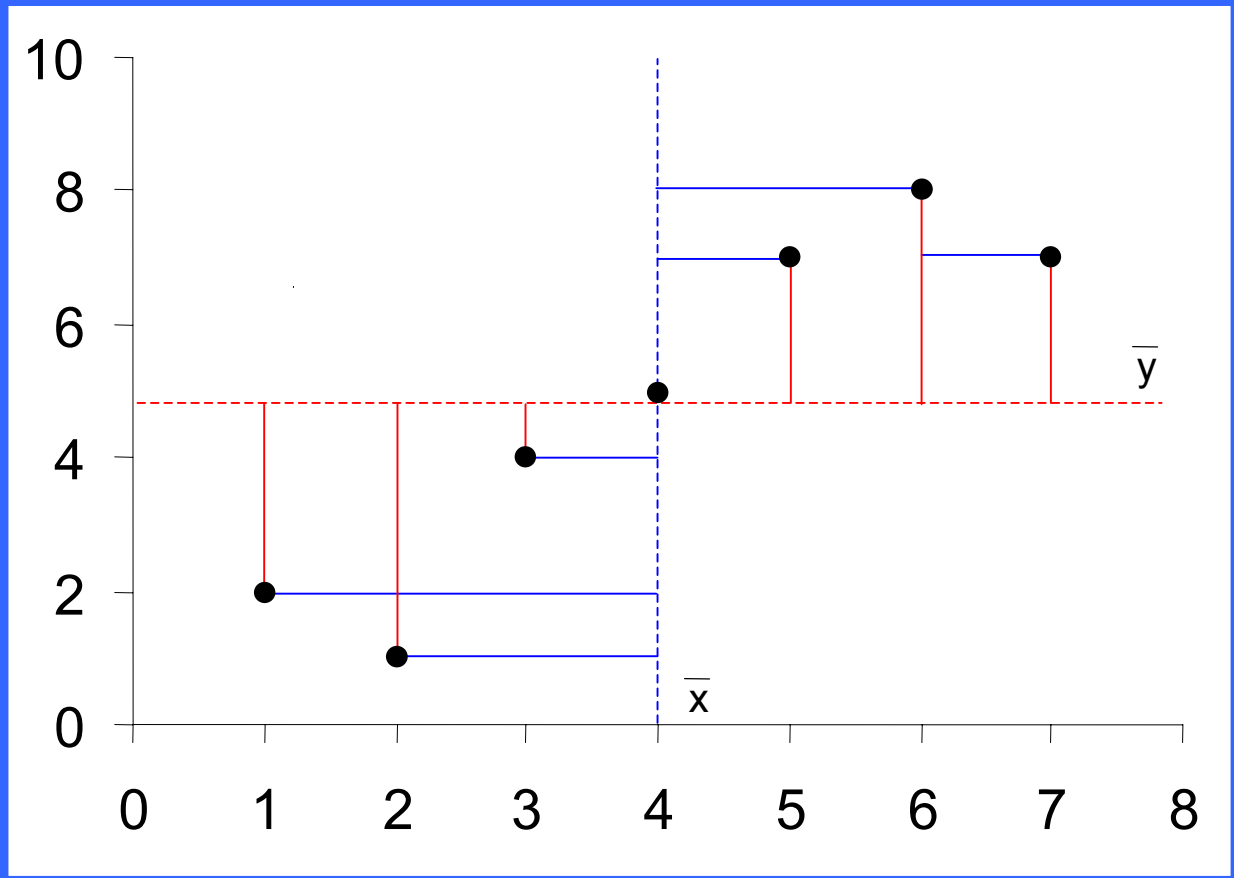
mit x_i, y_i Messwerte, $\bar{x}$, $\bar{y}$ arithmetische Mittel, S_{xy} Kovarianz der Variablen X und Y.

Eigenschaften:

- Die Kovarianz kann als Maß für den Zusammenhang zwischen zwei Variablen benutzt werden. Da sie allerdings von der Größenordnung der Variablen X und Y abhängt, wird sie noch normiert.
- Der Korrelationskoeffizient unterscheidet nicht zwischen abhängiger und unabhängiger Variable.
- Der Korrelationskoeffizient ändert sich nicht, wenn jeweils alle Wert der Variablen X oder Y linear transformiert werden.

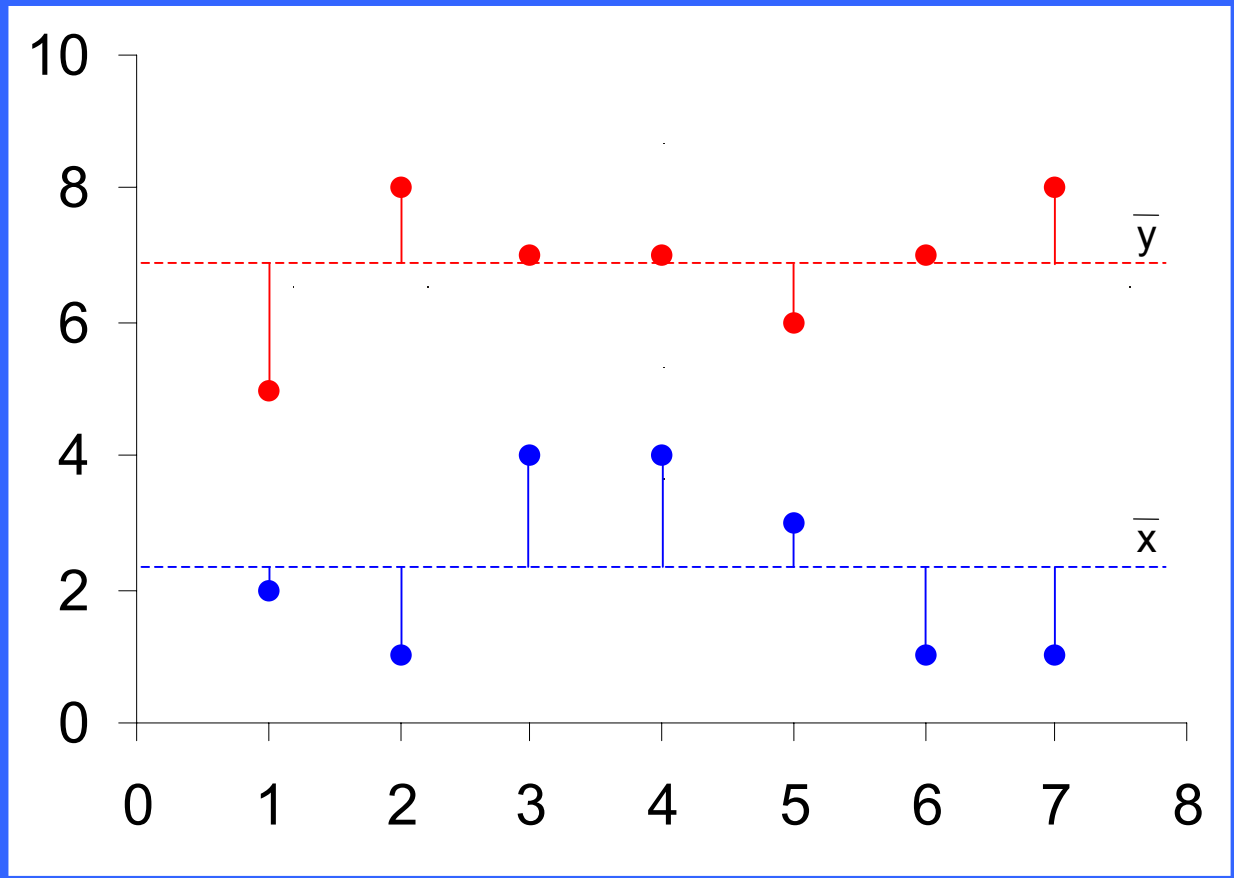
Korrelationskoeffizienten

Beispiel: Zwei Merkmale (etwa Länge, Breite) eines Proxies.



Korrelationskoeffizienten

Beispiel: Ein Merkmal (etwa Länge) zweier Proxies.



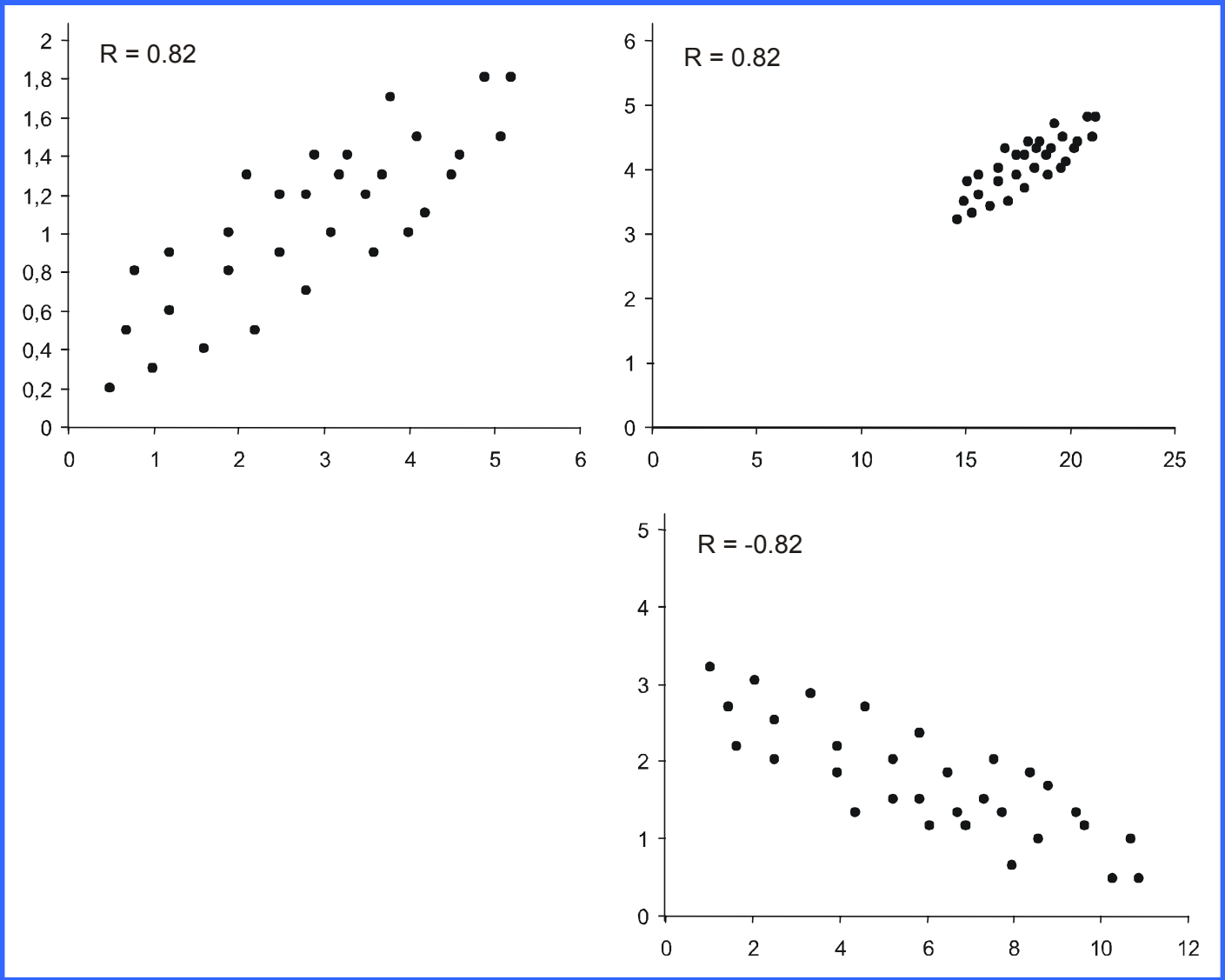
Korrelationskoeffizienten

Bezeichnung des Zusammenhangs:

Bezeichnung	Korrelationskoeffizient R
sehr stark	$0.87 \leq \lvert R \rvert \leq 0.99$
stark	$0.71 \leq \lvert R \rvert \leq 0.86$
mittel	$0.50 \leq \lvert R \rvert \leq 0.70$
schwach	$\lvert R \rvert < 0.50$

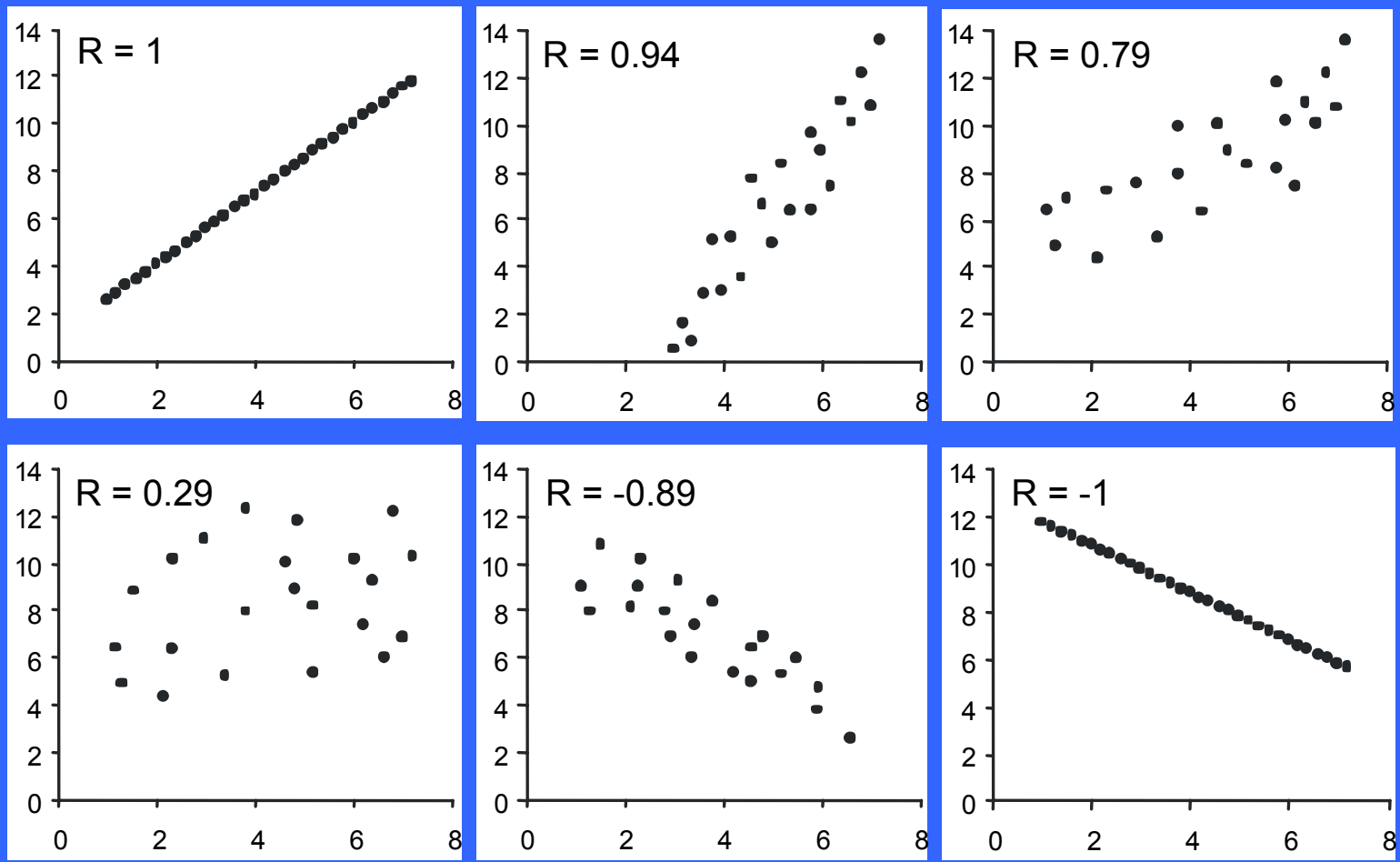
Korrelationskoeffizienten

Beispiel: Zusammenhang zwischen Länge und Breite.



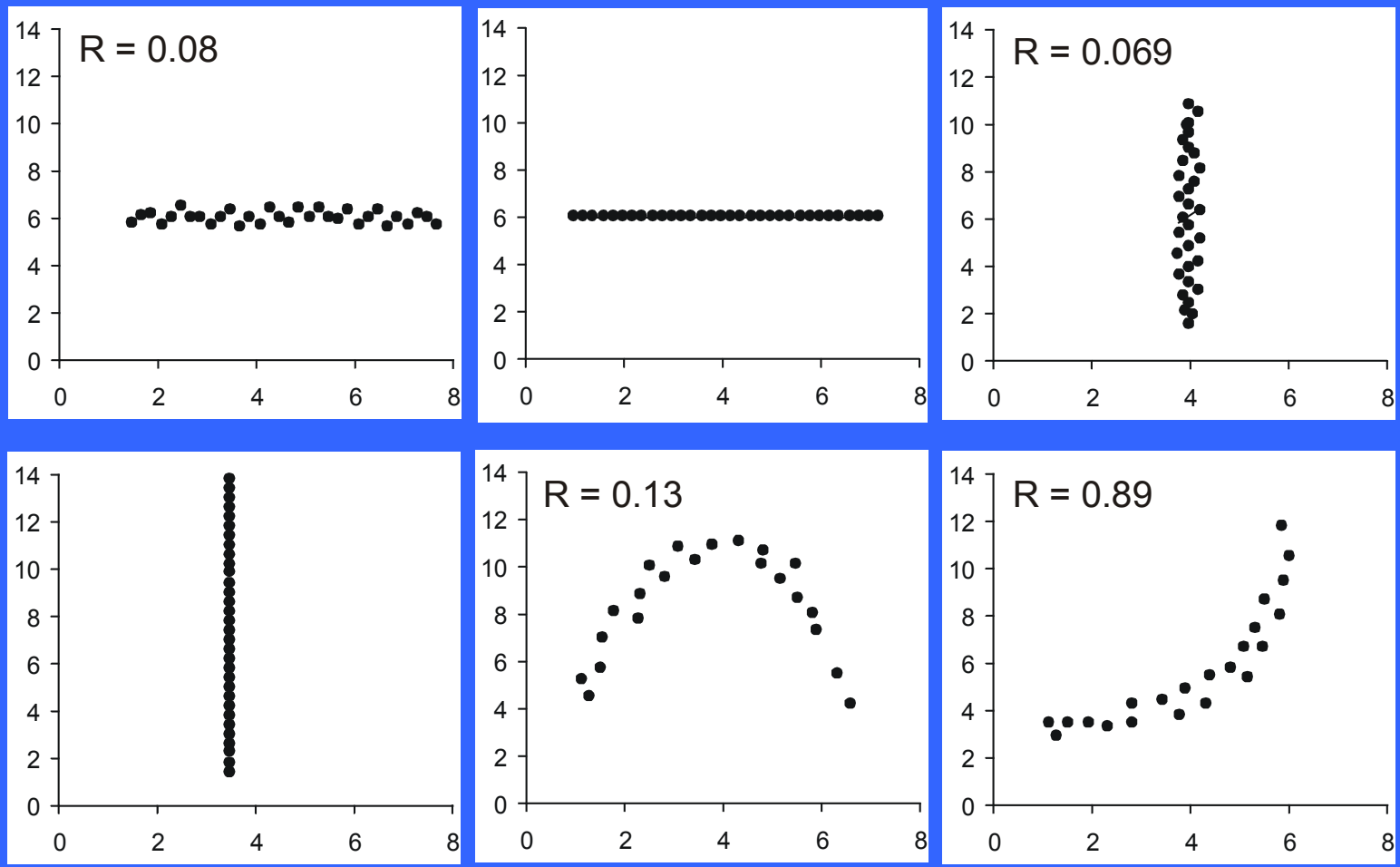
Korrelationskoeffizienten

Beispiel: Zusammenhang zwischen Länge und Breite.



Korrelationskoeffizienten

Beispiel: Zusammenhang zwischen Länge und Breite.



Korrelationskoeffizienten

Bestimmtheitsmaß: Das Quadrat R^2 des Korrelationskoeffizienten heißt Bestimmtheitsmaß B. Es lässt sich interpretieren als der prozentuale Anteil der Streuung der einen Variable, die durch die andere Variable erklärt werden kann (und umgekehrt). Es ist also das Verhältnis von erklärter Varianz zur Gesamtvarianz.

• Ist etwa $R = 0.7$, dann ist $B = R^2 = 0.49$, d.h. 49 % der Streuung der Y-Werte werden durch die Streuung der X-Werte in einem linearen Zusammenhang erklärt.

Korrelationskoeffizient R	**Bestimmtheitsmaß $B = R^2$**
$R = 0.87$ $R = 0.71$ $R = 0.50$	$B = 0.75$ $B = 0.50$ $B = 0.25$

Korrelationskoeffizienten

Eigenschaften:

- Ausreißer können die Korrelation und das Bestimmtheitsmaß stark beeinträchtigen.
- Zusammengesetzte homogene Stichproben jeweils ohne nennenswerte Korrelation können eine Korrelation vortäuschen.
- Viele Variablen die in den Geowissenschaften untersucht werden, sind nicht voneinander unabhängig, daher ergeben sich oftmals höhere Korrelationskoeffizienten, etwa die Abnahme der Temperatur mit steigender Höhe. Daher besteht ein formaler Zusammenhang.
- Die Aussagekraft der Korrelationskoeffizienten hängt auch von der Anzahl der betrachteten Werte ab. Bei einer großem Anzahl von Werten ist die Aussagekraft des Korrelationskoeffizienten auch größer.
- Bei einem Signifikanztest wird ein anhand der Stichprobe berechneter Korrelationskoeffizienten mit einem theoretischen Wert verglichen und entschieden ob sich diese signifikant voneinander unterscheiden oder nicht.

Korrelationskoeffizienten

Test des Produkt-Moment-Korrelationskoeffizient auf Signifikanz: wurde anhand der GG keine Korrelation zwischen den normalverteilten Variablen X und Y festgestellt, die STP zeigt allerdings eine messbare Korrelation R, so kann mit einer Irrtumswahrscheinlichkeit α geprüft werden, ob R signifikant von 0 verschieden ist. Entsprechend werden die Hypothesen formuliert:

H_0: Es besteht kein Unterschied zwischen R und 0.

H_A: Es besteht ein signifikanter Unterschied zwischen R und 0.

• Mit einem statistischen Testverfahren passend zur Fragestellung wird die Prüfgröße bestimmt. Zudem erhält man aus der dazugehörenden Tabelle den Schwellenwert für den Annahmebereich und den Ablehnungsbereich. Der Tabellenwert hängt dabei von der Irrtumswahrscheinlichkeit α und den Freiheitsgraden n-2 ab.

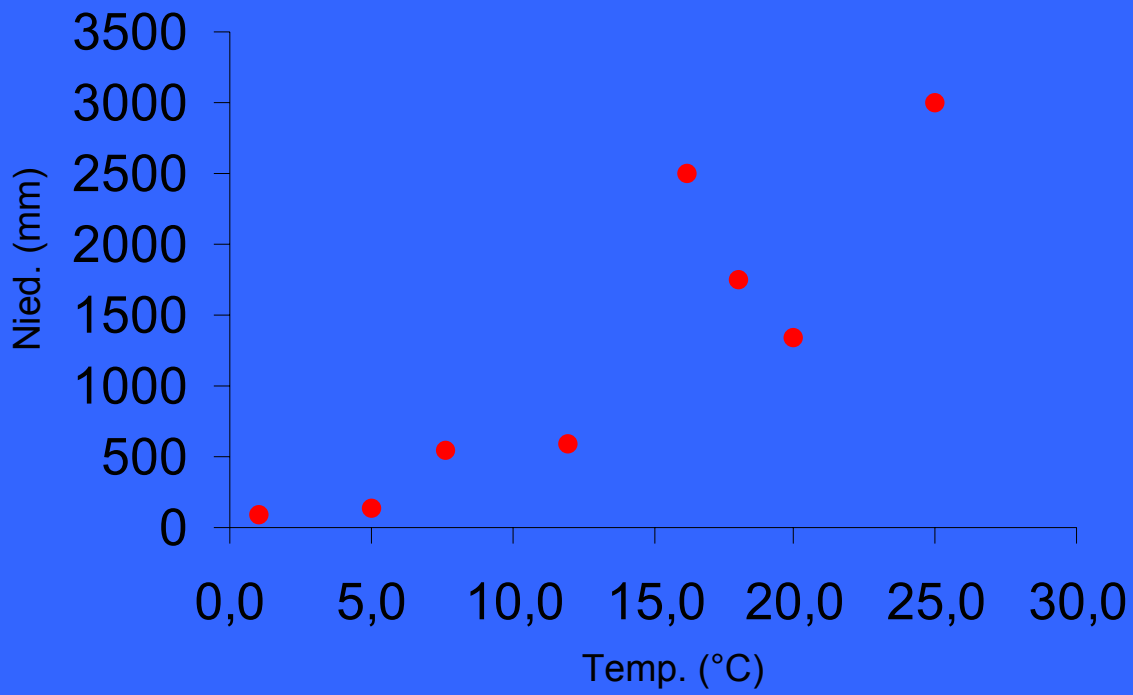
Korrelationskoeffizienten

Beispiel: Zusammenhang zwischen Temperatur und Niederschlag in einem globalen meridionalen Schnitt. Der Korrelationskoeffizient beträgt hierbei R = 0.88 und das Bestimmtheitsmaß R^2 = 0.77.

Station	Temp. (°C)	Nied. (mm)
1	1,0	100
2	5,0	140
3	7,6	550
4	12,0	580
5	20,0	1330
6	18,0	1740
7	16,2	2500
8	25,0	3000

Korrelationskoeffizienten

Beispiel: Zusammenhang zwischen Temperatur und Niederschlag in einem globalen meridionalen Schnitt.



Korrelationskoeffizienten

Rang-Korrelationskoeffizient nach Spearman: Der Rangkorrelationskoeffizient ρ_S (Rho-S) nach Spearman ist eine statische Maßzahl für den Zusammenhang zwischen zwei ordinal-skalierten ZV und ihrer monotonen Abhängigkeit. Er wird für Daten verwendet, die z.B. aus Bewertungen oder Befragungen hervorgegangen und nicht quantifizierbar sind, aber nach Rängen geordnet werden können. Dabei wird eine Auswertung der Rangzahlen der STP-Paare $(x_i;y_i)$ einer verbundenen STP vorgenommen, die paarweise miteinander verglichen werden.

Die ursprünglich gemessenen Daten beider Merkmale werden durch ihre Rangzahl ersetzt, die man dadurch erhält, dass die Daten beider Merkmale getrennt nach der Größe geordnet werden. Der kleinste Wert erhält die Rangzahl 1, der nächste die 2, der größte die Rangzahl n (n Stichprobenumfang).

Sind die Merkmalsausprägungen mehrerer Objekte gleich, dann wird aus den zugehörenden Rangplätzen das arithmetische Mittel gebildet und dieser Mittelwert allen Objekten mit dieser Ausprägung zugewiesen. Der Koeffizient ρ_S berechnet sich zu:

$$\rho_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n-1)}$$

mit $d_i = x_i - y_i$ und x_i, y_i Rangplätze, n Umfang der STP

Korrelationskoeffizienten

Eigenschaften:

• Geringe Rangdifferenzen weisen auf einen gleichsinnigen (positiven) Zusammenhang hin: je größer X, desto größer auch Y und umgekehrt.

• Starke Rangdifferenzen weisen auf einen ungleichsinnigen (negativen) Zusammenhang hin: je größer X, desto kleiner auch Y und umgekehrt.

• Wie der Produkt-Moment-Korrelationskoeffizient nimmt auch der Rangkorrelationskoeffizient nur Werte zwischen -1 und 1 an. Der Grenzfall 1 liegt vor, wenn nach aufsteigender Anordnung der Messreihe A automatisch auch die zugehörigen Werte der Messreihe B aufsteigend geordnet sind. Der Grenzfall -1 liegt vor, wenn sie absteigend geordnet sind.

• In der Praxis wird der Rangkorrelationskoeffizient häufig auch dann angewendet, wenn:

eine ordinal- und eine metrisch-skaliere ZV vorliegt: Dann muss die metrische ZV zunächst auf Ordinal-Niveau herunterskaliert werden.

beide ZV metrisch-skaliert sind: Dann müssen beide ZV zunächst auf Ordinal-Niveau herunterskaliert werden. Man verliert dabei zwar an Informationsgehalt, vermeidet aber mögliche Probleme bzgl. der Bi-Normalverteilung, die beim Korrelationskoeffizienten R nach Pearson vorausgesetzt werden muss.

Korrelationskoeffizienten

Beispiel: Baumkataster.

Anhand eines Baumkatasters soll untersucht werden, ob ein statistisch signifikanter Zusammenhang (α = 0.05) zwischen dem Baumalter und dem Schädigungsgrad der Bäume besteht. Der Schädigungsgrad ist ordinal-skaliert; das (metrische, jedoch klassifizierte) Baumalter wird hier nur als Ordinal-Information benutzt.

Korrelationskoeffizienten

Beispiel: Salzkonzentration.

Es soll der Zusammenhang zwischen Abfluss und Salzkonzentration in einem Vorfluter anhand einer Probenahme bestimmt werden.

Datum	Abfluß Q (l/s) X	gel. Salze K (mg/l) Y	Rangplatz Q	Rangplatz K	Rangdifferenz d	d^2
03.06.95	0.4	34.2	01	01	0	0
07.10.95	2.2	40.4	02	04	- 2	4
11.03.95	2.3	38.5	03	02	+1	1
04.11.95	3.9	39.4	04	03	+1	1
16.12.95	4.3	44.8	05	06	- 1	1
18.11.95	4.4	47.0	06	08	- 2	4
25.02.95	5.1	42.6	07	05	+2	4
28.01.95	8.0	45.6	08	07	+1	1
14.01.95	8.1	52.8	09	11	- 2	4
26.08.95	8.3	48.7	10	09	+1	1
30.12.95	9.7	50.5	11	10	+1	1
SUMME						22

$\rho_S = -0.2$

Korrelationskoeffizienten

Test des Rang-Korrelationskoeffizient auf Signifikanz: Der statistische Test erfolgt üblicherweise als Signifikanztest gegen 0 (jedoch gibt es auch Tests für andere Sollwerte). Man prüft also, ob in der GG ein von 0 statistisch signifikant abweichender Zusammenhang besteht:

H_0: Es besteht kein Unterschied zwischen ρ_S und 0.

H_A: Es besteht ein signifikanter Unterschied zwischen ρ_S und 0.

• Der Signifikanztest erfolgt mittels der t-Verteilung nach dem üblichen Entscheidungsprinzip.