

Multivariate Statistik

2. Merkmale / Zufallsvariablen

- **2.1 Grundbegriffe**
 - Typen von Merkmalen bzw. Zufallsvariablen
 - Häufigkeits- bzw. Wahrscheinlichkeits**verteilung**
 - **Kumulierte** Häufigkeits- bzw. Wahrscheinlichkeitsverteilung
- 2.2 Kennzahlen diskreter Merkmale / Zufallsvariablen
 - Arithmetischer Mittelwert / Erwartungswert
 - Andere Mittelwerte: geometrischer / harmonischer Mittelwert
 - Median, Quantil, Modus, Varianz / Standardabweichung
- 2.3 **Stetige** Merkmale / Zufallsvariablen
 - Wahrscheinlichkeitsdichten / Dichtefunktion
 - Übertragung der diskreten Kennzahldefinitionen
- 2.4 Wichtige **Standardverteilungen**:
 - diskrete u. stetige Gleichverteilung
 - Binomialverteilung, Poissonverteilung
 - Exponentialverteilung, Normalverteilung
- **Multivariate Statistik**
 - **Korrelation zwischen Zufallsvariablen**
 - **lineare Regression**
 - **Rechengesetze für Erwartungswert und Varianz**

Zufallsexperimente mit mehreren Zufallsvariablen

Manche Zufallsexperimente bestimmen *mit einer einzigen Durchführung* des Experimentes mehrere Zufallsvariablen. Man spricht dann von **multivariater Statistik**. Oft interessiert man sich dann für einen möglichen Zusammenhang zwischen diesen Variablen.

Beispiel

10 zufällig herausgegriffene Passanten werden nach Körpergröße X und Gewicht Y befragt.

Zweidimensionale Stichproben

Betrachtet man n Datensätze mit je zwei Merkmalen, so spricht man von einer **zweidimensionalen Stichprobe** vom Umfang n . Man kann sie durch n **Wertepaare** $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ wiedergeben.

Beispiel

Zweidimensionale Stichprobe:

Von 50 Studierenden eines Semesters wird jeweils das Ergebnis der Probeklausur (x_i) und der Abschlussklausur (y_i) erfasst. ($i = 1..50$)

Keine zweidimensionale Stichprobe:

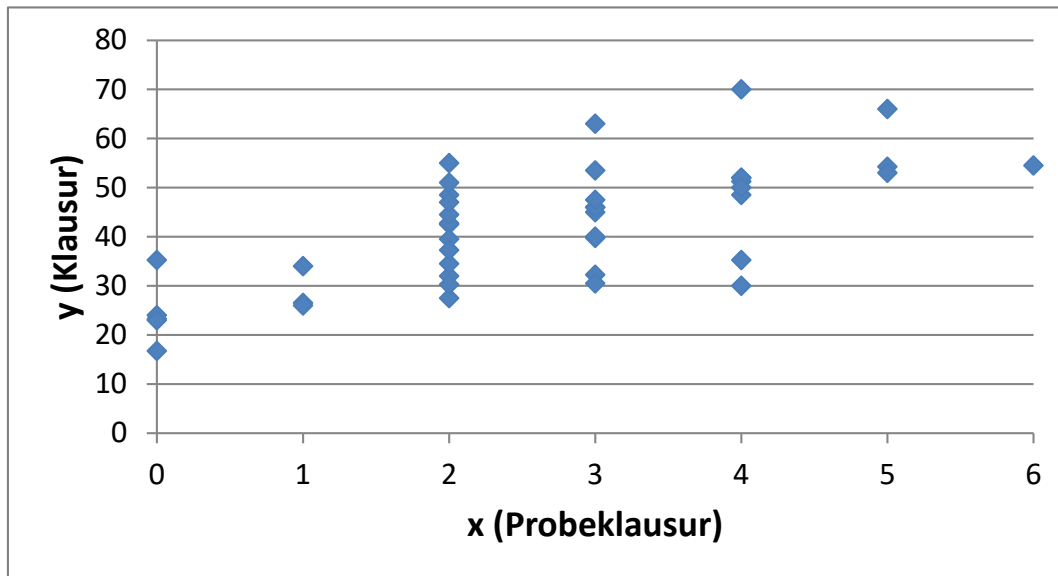
Die Probeklausur- und Endklausur-Ergebnisse von 50 Studierenden werden erfasst, aber die Zuordnung welche Ergebnisse jeweils zur selben Person gehören, wird nicht gespeichert.

Streudiagramme

Mit **Streudiagrammen** kann man Zusammenhänge zwischen **zwei intervall-skalierten** Merkmalen einer zweidimensionalen Stichprobe visualisieren.

Jede statistische Einheit wird als ein Punkt dargestellt. Die Werte der beiden Merkmale bestimmen dabei die x- und y-Koordinate des Punktes.

Streudiagramm



Tabelle

$x_1 \Rightarrow$ $x_2 \Rightarrow$	x (PK) [Punkte]	y (Klausur) [Punkte]	y_1 y_2
	5	66	
	4	50	
	0	24	
	4	30	
	1	34	
	3	45	
	2	47	
	3	40	
	2	51	
	3	46	
	0	23	
	

Unabhängigkeit von Zufallsvariablen

Für Zufalls-*Ereignisse* hatten wir „Unabhängigkeit“ bereits definiert. Nun also für Zufallsvariablen:

Definition 27.15

Zwei **Zufallsvariablen** X und Y des selben Zufallsexperimentes heißen voneinander *stochastisch unabhängig*, wenn Kenntnis des Wertes von X keine Information über Y liefert. (Die umgekehrte Richtung folgt dann automatisch).

Insbesondere sind dann alle *Ereignisse*, die man auf X definieren kann (z.B. „ $X = c$ “, „ $X < c$ “) von allen Ereignissen, die man auf Y definieren kann, stochastisch *unabhängig*.

Beispiel: Die Augenzahlen X_1 des ersten und X_2 des zweiten Wurfs eines Würfels sind unabhängig.

Bemerkung: Auf unabhängige Z-Variablen kann man also die Formeln für unabhängige Z-Ereignisse anwenden: Für **unabhängige** X und Y gilt also z.B:

$$P(X=x \cap Y=y) = P(X=x) \cdot P(Y=y)$$

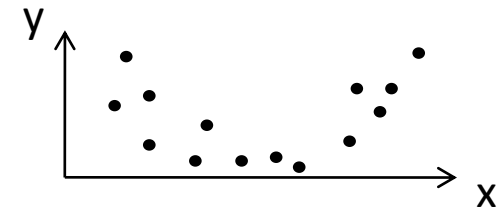
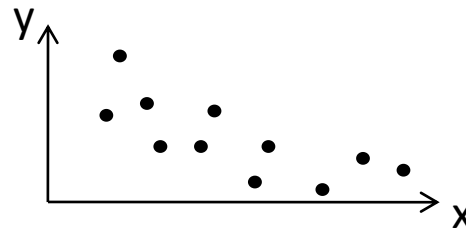
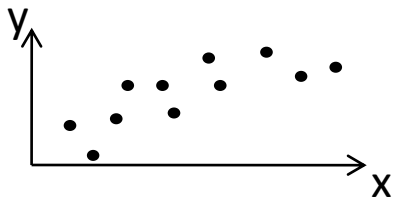
$$\text{und } P(X \leq x \cap Y \leq y) = P(X \leq x) \cdot P(Y \leq y)$$

Unabhängigkeit anschaulich

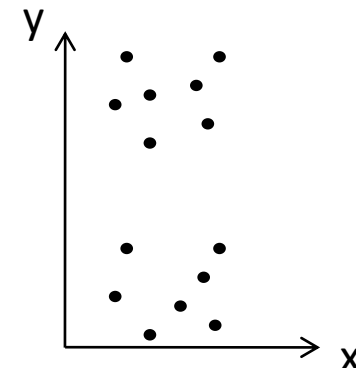
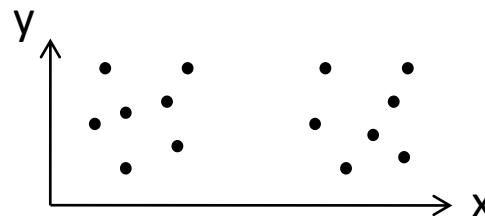
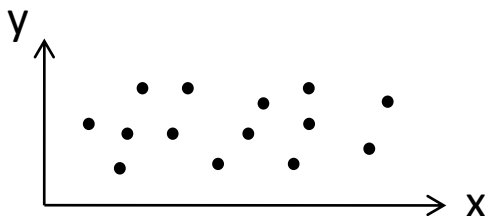
Zufallsvariable Y ist von Zufallsvariable X stochastisch unabhängig, wenn die Wahrscheinlichkeitsverteilung der Y -Werte, die in Verbindung mit einem bestimmten X -Wert vorkommen, für jeden X -Wert dieselbe ist.

Anders formuliert: Sind x und y unabhängig, so lässt der x -Wert keine Rückschlüsse auf den y -Wert zu.

Beispiele (X und Y abhängig)



Beispiele (X und Y unabhängig)



Lineare Regression

Definition

Bei der **Linearen Regression** sucht man eine Gerade $g(x) := k \cdot x + d$ mit der Eigenschaft, dass das mittlere **Fehlerquadrat** $MSE = \frac{1}{n} \sum_{i=1}^n (y_i - g(x_i))^2$ minimal wird.

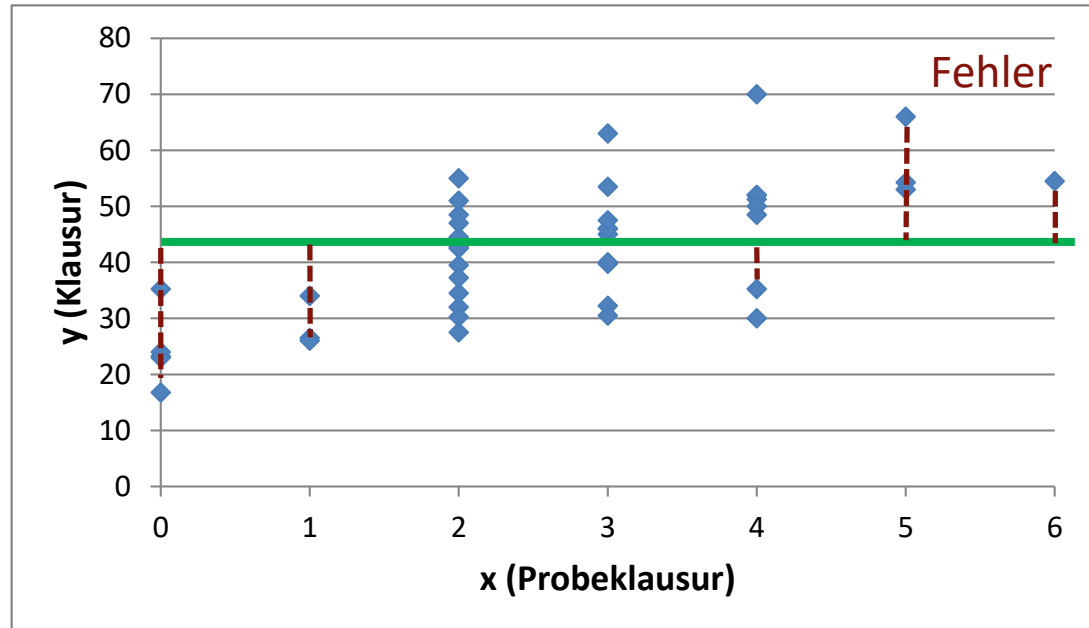
Bemerkung:

Wählt man die konstante Funktion

$$g(x) = \bar{y}$$

so ist **MSE** gleich der Varianz von y .

Es gibt aber meistens „bessere“ Geraden, für die **MSE** kleiner ausfällt.

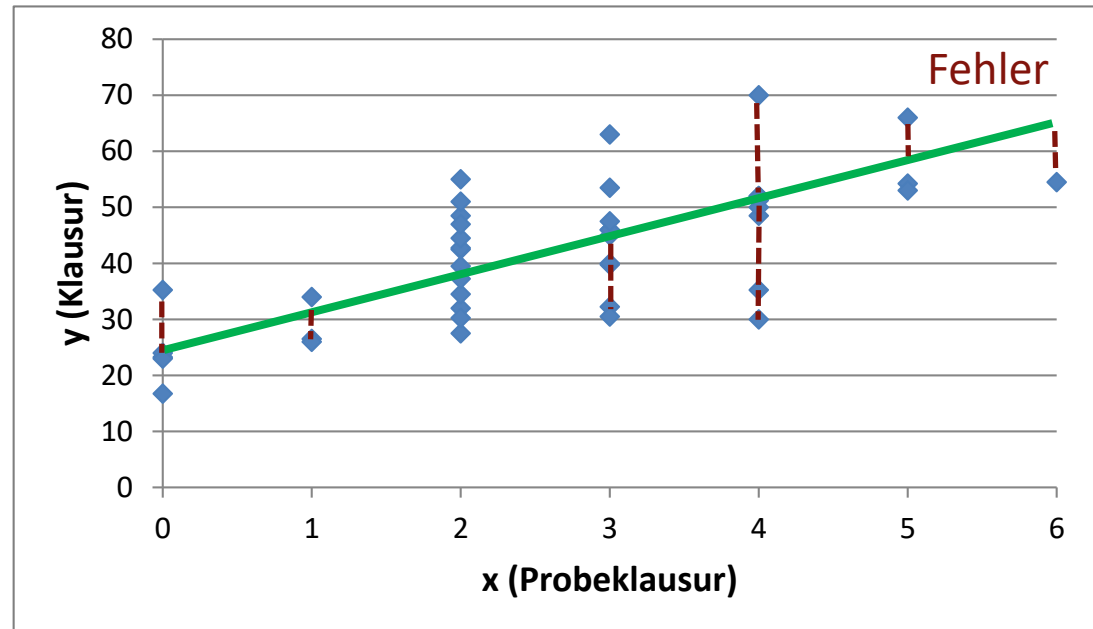


Lineare Regression

Definition

Bei der **Linearen Regression** sucht man eine Gerade $g(x) := k \cdot x + d$ mit der Eigenschaft, dass das mittlere **Fehlerquadrat** $MSE = \frac{1}{n} \sum_{i=1}^n (y_i - g(x_i))^2$ minimal wird.

Fällt MSE für die optimal gewählte Regressionsgerade z.B. um 60% geringer aus, als für die Gerade $g(x) := \bar{y}$, so sagt man, **60% der Varianz von y sei durch x linear erklärbar.**



Herleitung von Satz 25.21

Es werden k, d gesucht, so dass die Regressionsgerade $g(x) = k \cdot x + d$ die Summe der Fehlerquadrate minimiert:

$$\sum_{i=1}^n (y_i - g(x_i))^2 = \sum_{i=1}^n (y_i - (k \cdot x_i + d))^2 \stackrel{!}{=} \min$$

Partielle Ableitungen nach d bzw. k :

$$\begin{aligned} \frac{\partial}{\partial d} \sum_{i=1}^n (y_i - k \cdot x_i - d)^2 &= \sum_{i=1}^n 2 \cdot (y_i - k \cdot x_i - d) \cdot (-1) = -2 \cdot \left(\sum_{i=1}^n y_i - k \cdot \sum_{i=1}^n x_i - \sum_{i=1}^n d \right) = \\ &= -2 \cdot (n \cdot \bar{y} - k \cdot n \cdot \bar{x} - n \cdot d) = -2n \cdot (\bar{y} - (k \cdot \bar{x} + d)) \stackrel{!}{=} 0 \end{aligned}$$

Also muss gelten $d = \bar{y} - k\bar{x}$

$$\begin{aligned} \frac{\partial}{\partial k} \sum_{i=1}^n (y_i - k \cdot x_i - d)^2 &= \sum_{i=1}^n 2 \cdot (y_i - k \cdot x_i - d) \cdot (-x_i) = \\ &= 2 \cdot \left(-\sum_{i=1}^n x_i \cdot y_i + k \cdot \sum_{i=1}^n x_i^2 + d \sum_{i=1}^n x_i \right) \stackrel{!}{=} 0 \end{aligned}$$

Lässt sich mit einigem Rechnen umschreiben zu* $k = r_{x,y} \frac{s_y}{s_x}$

*Wobei $r_{x,y}$ die Korrelation zw. X und y ist (siehe nächste Folien)

Lineare Regression

Satz 25.21 (Regressionsgerade)

Die im Sinne des mittleren Fehlerquadrats MSE optimale Regressionsgerade $g(x) = kx + d$, ist gegeben durch

$$k = r_{x,y} \frac{s_y}{s_x} , \quad d = \bar{y} - k\bar{x} .$$

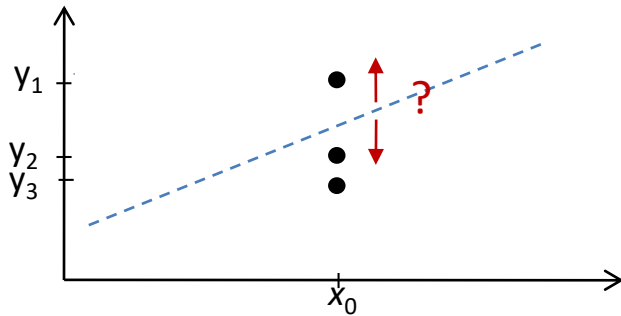
Dabei ist $r_{x,y}$ der empirische Korrelationskoeffizient (siehe nachfolgende Folien), \bar{x}, \bar{y} sind die arithmetischen Mittelwerte und s_x, s_y die Stichprobenstandardabweichungen der x – bzw. y –Werte.

Bemerkung:

1. Regression ist nicht symmetrisch in X und Y , d.h. Regression von X auf Y liefert eine andere Gerade als Regression von Y auf X .
2. Im Gegensatz zur **Interpolation** z.B. mit Splines wird **Regression** zum Schätzen des Y -Wertes aus dem X -Wert verwendet, wenn **keine perfekte Vorhersage möglich** ist.

Lineare Regression: Warum Fehler-*Quadrate*?

Angenommen es gibt zum selben Wert x_0 mehrere verschiedene y -Werte y_1, \dots, y_n in einer Stichprobe. Für welchen Wert $f(x_0)$ wird dann das mittlere Fehlerquadrat minimal, d.h. welcher Wert ist im Sinne der Summe der Fehlerquadrate der optimale Kompromiss zwischen y_1, \dots, y_n ?



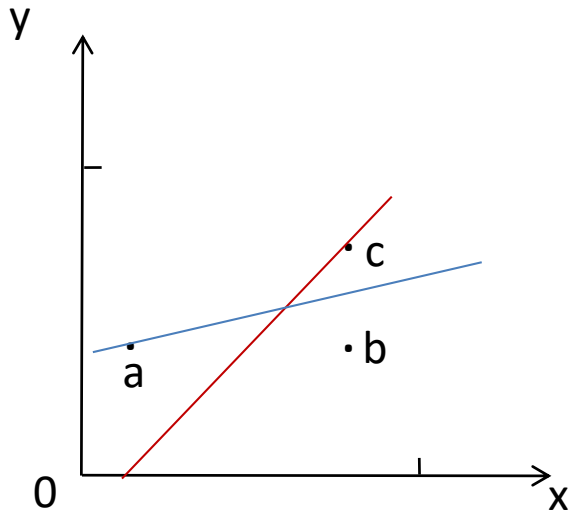
$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_0))^2 \stackrel{!}{=} \min$$

$$\frac{\partial}{\partial(f(x_0))} MSE \stackrel{!}{=} 0 \quad \Leftrightarrow \quad f(x_0) = \frac{1}{n} \sum_{i=1}^n y_i$$

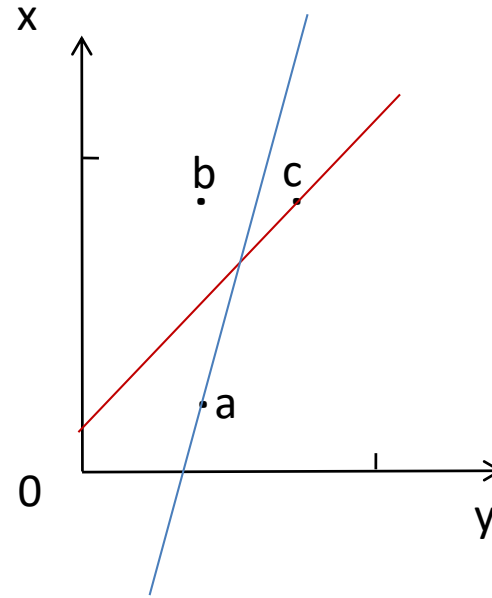
Das mittlere **Fehlerquadrat** wird also **minimal**, wenn der Wert der Regressionsfunktion gleich dem **Mittelwert** der y_1, \dots, y_n ist.

Man kann den Funktionswert $f(x_0)$ einer Regressionsgeraden deshalb als **prognostizierten Mittelwert** der an der Stelle x_0 zu erwartenden y -Werte interpretieren.

Unsymmetrie der Regression



Blau: Regression von x nach y .
 y als Funktion von x
Vorhersage von y aus x



Rot: Regression von y nach x .
 x als Funktion von y
Vorhersage von x aus y

Es macht einen Unterschied, ob man eine Regression von x auf y berechnet oder eine von y auf x . Man erhält unterschiedliche Geraden!

Stichproben-Kovarianz

Sei $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ eine zweidimensionale Stichprobe.

Definition

Die (unkorrigierte) **Stichproben-Kovarianz** oder **empirische Kovarianz** zwischen zwei quantitativen Merkmalen X und Y ist definiert als:

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

Diese Formel gilt für gleichberechtigte Datensätze einer Stichprobe. Geht man stattdessen von einer **Häufigkeitsverteilung** aus, so sind die Summanden analog zur Varianzberechnung mit ihrer **relativen Häufigkeit** zu **gewichten**.

Kovarianz zwischen Zufallsvariablen

Definition

Die **Kovarianz** zwischen zwei Zufallsvariablen X und Y desselben Zufallsexperimentes ist definiert als:

$$\text{Cov}(X, Y) = \sigma_{X, Y} \quad := \quad \sum_{x_i} \sum_{y_j} (x_i - E(X)) \cdot (y_j - E(Y)) \cdot P(X = x_i \cap Y = y_j)$$

Dabei laufen die Summen über alle Realisationen x_i der Zufallsvariablen X , in Kombination mit alle Realisationen y_j der Zufallsvariablen Y .

„ $\text{Cov}(X, Y)$ “ und „ $\sigma_{X, Y}$ “ sind gleichbedeutende Schreibweisen.

Korrelation

Definition

Die **empirische Korrelation** zweier Merkmale X und Y ist definiert als

$$r_{X,Y} := \frac{s_{X,Y}}{s_X \cdot s_Y}$$

(auch: *empirischer Korrelationskoeffizient* oder **Stichprobenkorrelation**)

Die **Korrelation** zwischen zwei quantitativen Zufallsvariablen X und Y desselben Zufallsexperimentes ist definiert als:

$$\rho_{X,Y} := \frac{\sigma_{X,Y}}{\sigma_X \cdot \sigma_Y}$$

wobei σ_X bzw. σ_Y die Standardabweichung von X bzw. Y ist.

Die Korrelation berechnet sich also durch eine bestimmte Art von **Normierung** aus der Kovarianz.

Beispiel: Berechnung der Stichproben-Korrelation aus einem Datensatz

Nebenstehende Tabelle zeigt für eine Stichprobe von 4 Studierenden jeweils die Punktezahl in der Probeklausur (x) und Abschlussklausur (y).

Wie berechnet man die Korrelation zwischen x und y?

x (Probekl.)	y (Klausur)
5	66
4	50
0	24
4	30

1. Mittelwerte von x und y berechnen :

$$\bar{x} = \frac{5+4+0+4}{4} = 3.25 ; \quad \bar{y} = \frac{66+50+24+30}{4} \approx 42.5$$

2. Kovarianz berechnen :

$$s_{x,y} = \frac{1}{4} \cdot [(5-3.25) \cdot (66-42.5) + (4-3.25) \cdot (50-42.5) + (0-3.25) \cdot (24-42.5) + (4-3.25) \cdot (30-42.5)] \approx 24.4$$

3. Empirische Standardabweichungen berechnen :

$$s_x = \sqrt{\frac{1}{4} \cdot [(5-3.25)^2 + (4-3.25)^2 + (0-3.25)^2 + (4-3.25)^2]} \approx 1.92$$

$$s_y = \sqrt{\frac{1}{4} \cdot [(66-42.5)^2 + (50-42.5)^2 + (24-42.5)^2 + (30-42.5)^2]} \approx 16.6$$

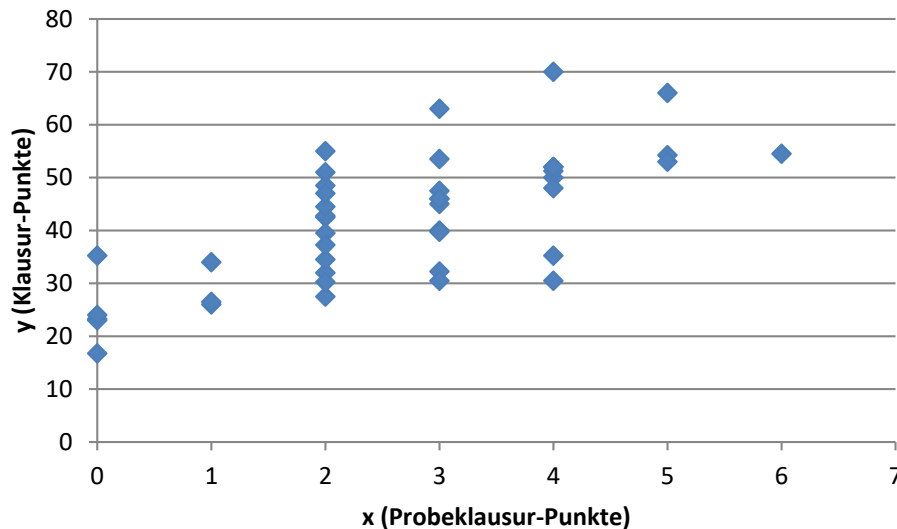
4. Korrelation berechnen :

$$r_{x,y} = \frac{s_{x,y}}{s_x \cdot s_y} \approx \frac{24.4}{1.92 \cdot 16.6} \approx 0.76$$

Beispiel (Fortsetzung)

Beispiel

Für den vollständigen Datensatz erhält man mit Computerhilfe:



Empirische Korrelation:

$$r_{x,y} \approx 0.70$$

Mittelwerte: $\bar{x} = 2.6$; $\bar{y} \approx 41.9$

Empirische Standardabweichungen: $s_x \approx 1.48$; $s_y \approx 12.31$

Empirische Kovarianz: $s_{x,y} \approx 12.8$

Skalierungsunabhängigkeit der Korrelation

Beispiel:

Gleicher Sachverhalt:

PK in *Punkten*

x (PK) [Punkte]	y (Klausur) [Punkte]
5	66
4	50
0	24
4	30
1	34
3	45
2	47
3	40
2	51
3	46
0	23

Kovarianz: **13.5**
Std-Abw: 1.56 12.3
Korrelation: **0.86**

PK in *Anteil*

(von max. 6 Punkten)

x (PK) [Anteil]	y (Klausur) [Punkte]
0.833	66
0.667	50
0	24
0.667	30
0.167	34
0.5	45
0.333	47
0.5	40
0.333	51
0.5	46
0	23

2.25
0.260 12.3
0.86

Die Kovarianz hängt von der Skalierung der Merkmale ab, die Korrelation nicht!

Kennzahlen für den Zusammenhang zweier Variablen

Zusammenhang zwischen der empirischen Korrelation zweier Merkmale X und Y und der Regressionsgeraden von X auf Y :

1. Die **Steigung** der (im Sinne des MSE) optimalen Regressionsgerade von X auf Y , die das mittlere Fehlerquadrat MSE minimiert, ist $r_{x,y} \cdot \frac{s_y}{s_x}$
2. **Definition:** Die **durch X erklärte Varianz von Y** : $g_{XY} := s_y^2 - MSE$
(Um wieviel geringer streuen die Y -Werte um die Regressionsgerade als um ihren Mittelwert, vgl. Folie 2-167)
3. **Definition: Bestimmtheitsmaß b :**

$$b_{XY} := \frac{\text{Durch } X \text{ erklärte Varianz von } Y}{s_y^2} = 1 - \frac{MSE}{s_y^2}$$

4. **Satz:** Man kann zeigen, dass $b_{XY} = r_{XY}^2$.
Das Quadrat der Korrelation besagt also, welcher Anteil der Varianz von Y durch X erklärbar ist, d.h. um welchen Faktor die Schwankungen der Y -Werte um die optimale Regressiongerade geringer ist, also um ihren Mittelwert.

Eigenschaften der Korrelation

Satz

Die Stichproben-Korrelation bzw. die Korrelation von Zufallsvariablen liegt immer im Intervall $[-1; 1]$ und ...

- ist **1** genau dann wenn alle Datenpunkte ***exakt auf einer Geraden mit positiver Steigung liegen.***
- ist **-1** genau dann wenn alle Datenpunkte ***exakt auf einer Geraden mit negativer Steigung liegen.***

Definition

Man sagt, Zufallsvariable X und Y sind **korreliert**, wenn $\rho_{x,y} \neq 0$

Satz

Wenn X und Y **stochastisch unabhängig** sind, dann sind sie **unkorreliert**.

(Die Umkehrung gilt nicht!)

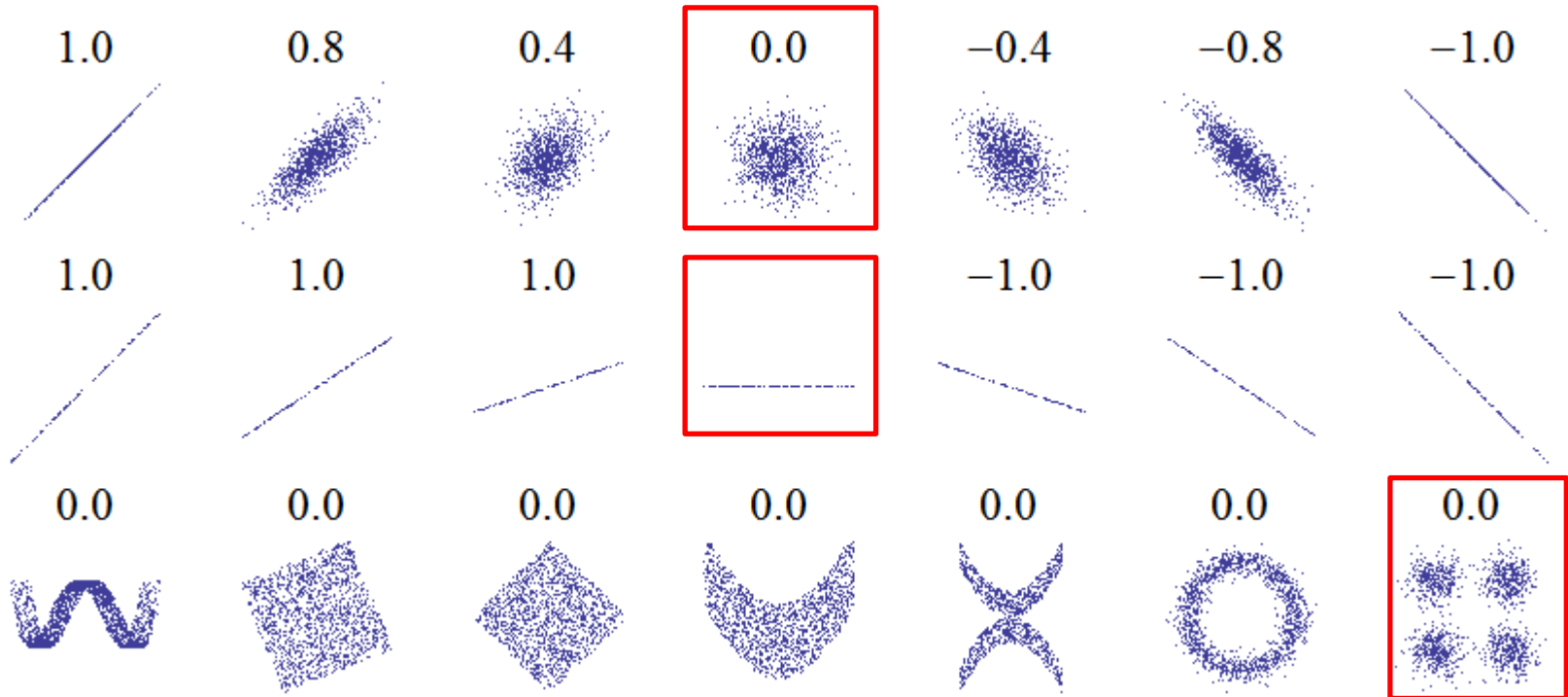
Anschaulich:

Die Korrelation misst, welcher Anteil der Varianz beider Variablen durch einen **linearen** Zusammenhang zur jeweils anderen Variablen erklärbar ist.

(vgl. 2-167 und 2-177)

Korrelation (Beispiele)

Streudiagramme und die zugehörige empirische Korrelation:



Da die Korrelation skalierungsinvariant ist, kommt es auf die Beschriftung der Achsen nicht an.

Nur in den rot eingerahmten Fällen sind die Variablen **unabhängig**.

Zusammenfassung wichtiger Formeln

Konkreter Datensatz

arithmetischer Mittelwert

$$\bar{x} = \frac{1}{n} \cdot (x_1 + \dots + x_n)$$

Stichproben-Varianz

$$s^2 = \frac{1}{n} \cdot \sum_{x_k} (x_k - \bar{x})^2$$

Stichproben-Covarianz

$$s_{x,y} = \frac{1}{n} \cdot \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})$$

Stichproben-Korrelation

$$r_{X,Y} = \frac{s_{X,Y}}{s_X \cdot s_Y}$$

x_k durchläuft die
 n Datensätze der Stichprobe

Zufallsvariable

Erwartungswert

$$\mu_X = E(X) = \sum_{x_i} x_i \cdot P(X = x_i)$$

Varianz

$$\sigma^2(X) = \sum_{x_i} (x_i - \mu)^2 \cdot P(X = x_i)$$

Covarianz

$$\begin{aligned} \sigma_{X,Y} &= Cov(X,Y) = \\ &= \sum_{x_i} \sum_{y_j} (x_i - \mu_X) \cdot (y_j - \mu_Y) \cdot P(X = x_i \cap Y = y_j) \end{aligned}$$

Korrelation

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X) \cdot \sigma(Y)}$$

x_i durchläuft die **möglichen Werte von X**

Beispiel: Berechnung der Kovarianz aus einer Wahrscheinlichkeitsverteilung

Die gemeinsame Wahrscheinlichkeitsverteilung der Zufallsvariablen X und Y sei durch die folgende Tabelle gegeben. Berechnen Sie die Kovarianz von X und Y .

Wahrscheinlichkeit	<i>X=150</i>	<i>X=-1000</i>	
<i>Y=130</i>	88%	7%	95,0%
<i>Y=-2000</i>	2%	3%	5,0%
	90,0%	10,0%	

Lösung:

Schritt 1:

$$E(X) = 150 \cdot 0.9 + (-1000) \cdot 0.1 = 35$$

$$E(Y) = 130 \cdot 0.95 + (-2000) \cdot 0.05 = 23.5$$

Schritt 2:

$$\begin{aligned} \text{Cov}(X, Y) &= (150 - 35) \cdot (130 - 23.5) \cdot 0.88 + (-1000 - 35) \cdot (130 - 23.5) \cdot 0.07 + \\ &\quad (150 - 35) \cdot (-2000 - 23.5) \cdot 0.02 + (-1000 - 35) \cdot (-2000 - 23.5) \cdot 0.03 \approx \\ &\approx 61238 \end{aligned}$$

Regression: Beispiel

Beispielaufgabe (vgl. Beispiel zur Korrelation)

Nebenstehende Tabelle zeigt für eine Stichprobe von 4 Studierenden jeweils die Punktezahl in der Probeklausur (x) und Abschlussklausur (y).

Bestimmen Sie die Regressionsgerade von x nach y und zeichnen Sie sie in ein Streudiagramm ein.

x (Probekl.)	y (Klausur)
5	66
4	50
0	24
4	30

1. Bereits zuvor berechnet (Folie 2-170):

$$\bar{x} = 3.25 ; \quad \bar{y} \approx 42.5$$

$$s_x \approx 1.92 ; \quad s_y \approx 16.6$$

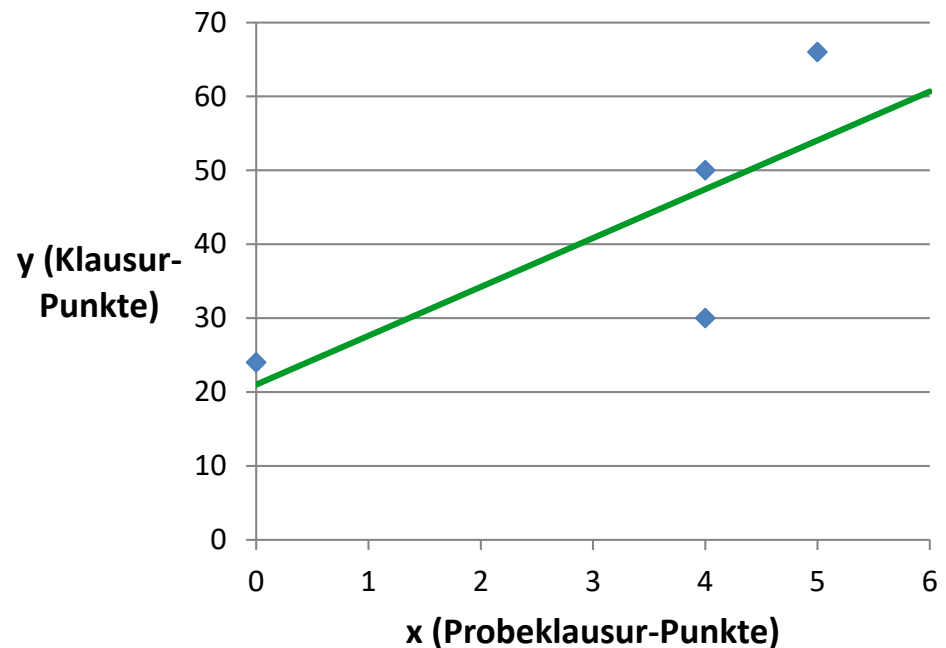
$$r_{x,y} \approx 0.76$$

$$2. \quad k = r_{x,y} \cdot \frac{s_y}{s_x} \approx 6.61 \quad (\text{nach Satz 25.21})$$

$$d = \bar{y} - k \cdot \bar{x} = 42.5 - 6.61 \cdot 3.25 \approx 21.0$$

3. Regressionsgerade:

$$g(x) = 6.61 \cdot x + 21.0$$



Regression (Beispiel, Fortsetzung)

Realistisches Beispiel

Mit einem größeren Datensatz erhält man:

$$\bar{x} = 2.6 \quad ; \quad \bar{y} \approx 41.9$$

$$s_x \approx 1.48 \quad ; \quad s_y \approx 12.31$$

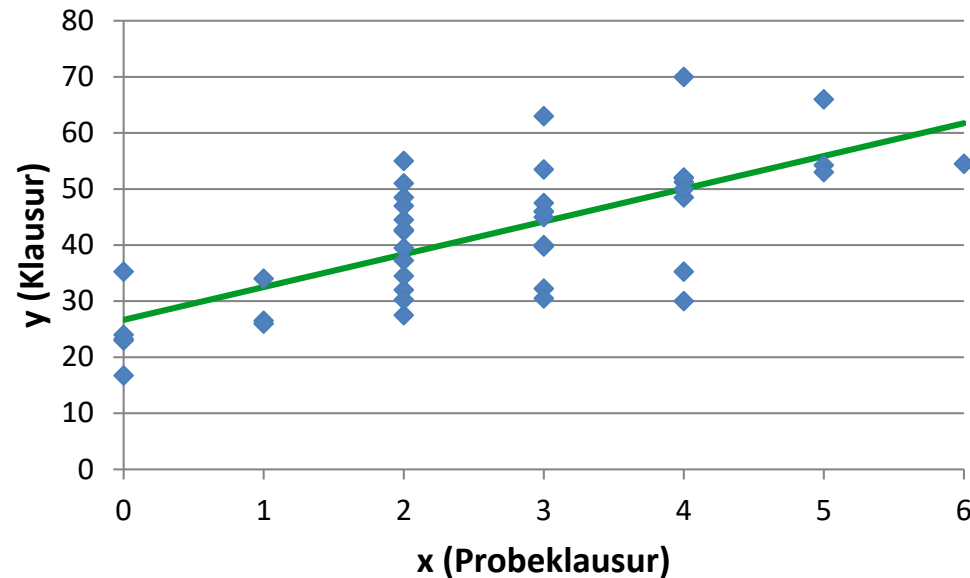
$$s_{x,y} \approx 12.8 \quad ; \quad r_{x,y} \approx 0.70$$

Daraus

$$k = r_{x,y} \cdot \frac{s_y}{s_x} \approx 0.70 \cdot \frac{12.31}{1.48} \approx 5.84$$

$$d = \bar{y} - k \cdot \bar{x} \approx 41.9 - 5.84 \cdot 2.6 \approx 26.7$$

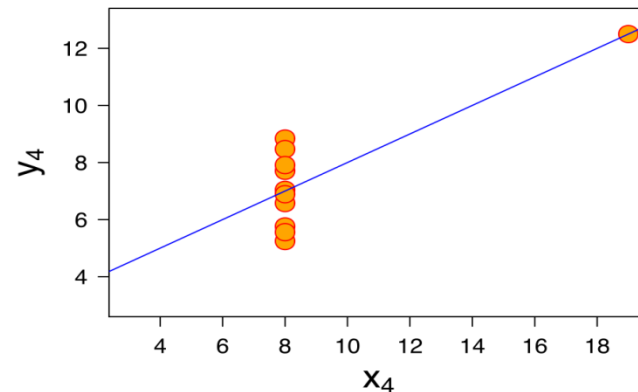
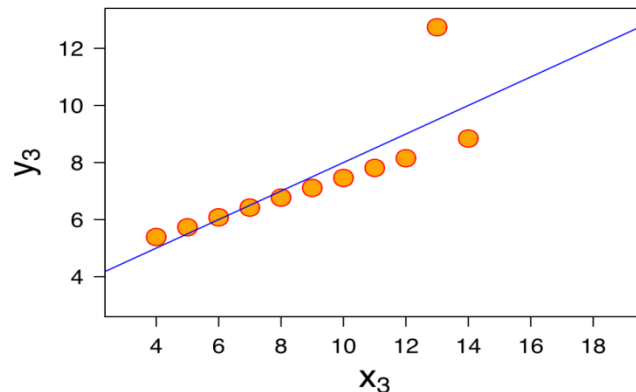
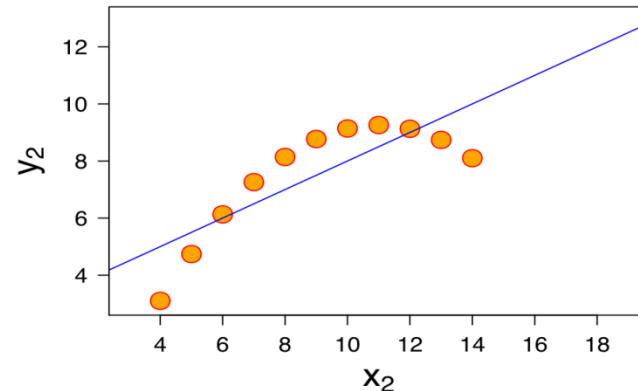
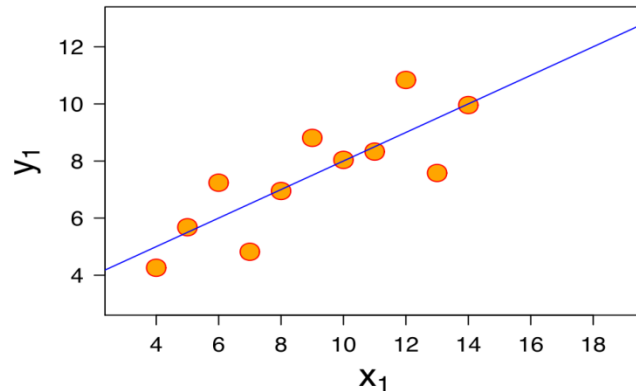
Also Regressionsgerade **$g(x) = 5.84 \cdot x + 26.7$**



Wieviel Abschlusspunkte wird man also bei jemandem mit 1 PK-Punkt prognostizieren?

→ $g(1) \approx \underline{32.5}$. Diese Prognose basiert darauf, wie viele Abschlussklausur-Punkte andere Studierenden mit 1 PK-Punkt erzielt haben, aber auch Studierende mit 0 oder 2 oder 3 PK-Punkten haben die Prognose beeinflusst.

Grenzen der linearen Regression (Beispiele)



Mittelwerte, Standardabweichungen, Kovarianz und die Regressionsgerade sind in allen vier Fällen identisch

Lineare Regression mit mehreren Eingangsvariablen

Regression bei mehreren Eingangsvariablen lässt sich in Vektorschreibweise darstellen:

Sei \vec{x}_i für den i -ten Datensatz der **Vektor der Eingangsvariablen** und \vec{k} der **Steigungs-Vektor** der Regressionsgeraden.

Lineare Regression mit mehreren Eingangsvariablen

Man sucht eine Funktion $g(\vec{x}) := \vec{x} \cdot \vec{k} + d$ (genauer: man sucht \vec{k}, d) mit der Eigenschaft,

dass das mittlere Fehlerquadrat $MSE = \frac{1}{n} \sum_{i=1}^n (y_i - g(\vec{x}_i))^2$ minimal wird.

Dabei ist: $\vec{x} \cdot \vec{k}$ das Skalarprodukt der Vektoren \vec{x} und \vec{k} ,
und \vec{x}_i der Vektor der Eingangsvariablen im i -ten Datensatz.

Formaler Trick, um die Darstellung weiter zu vereinfachen:

Ergänzt man den Eingangsvektor um eine Dimension, die bei allen Datensätzen den Wert 1 hat, vereinfacht sich die Geradengleichung zu $g(\vec{x}) := \vec{x} \cdot \vec{k}$.

Lineare Regression mit mehreren Eingangsvariablen

Sei x_{ij} der Wert der j -ten Eingangsvariablen ($j = 1..m$) im i -ten Datensatz ($i = 1..n$), und y_i der Wert der Zielvariablen im i -ten Datensatz.

In Matrixschreibweise können die Daten dann so dargestellt werden:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix} \quad \vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Satz: Der optimale Parametervektor \vec{w} ,

für den die Regressionsgerade $g(\vec{x}) := \vec{k} \cdot \vec{x}$

das mittlere Fehlerquadrat auf dem durch X und \vec{y} gegebenen Datensatz,

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - g(\vec{x}_i))^2 = \frac{1}{n} |\vec{y} - X \cdot \vec{k}|$$

minimiert, lässt sich berechnen als $\vec{w} = (X^t \cdot X)^{-1} \cdot X^t \cdot \vec{y}$

Bemerkung: Falls alle Eingangsvariablen auf Erwartungswert 0 normiert sind, ist $X^t \cdot X$ die **Kovarianzmatrix** der Eingangsvariablen.

Kennzahlen für den Zusammenhang mehrerer Variablen

Man bezeichnet oft die Korrelation zwischen den Variablen x_i und x_j mit $r_{i,j}$.

Die $m \times m$ **Korrelationsmatrix** stellt die Korrelationen jeder Variable mit jeder anderen Variable dar, wobei der Wert in Zeile i und Spalte j die Korrelation zwischen den Merkmalen x_i und x_j darstellt:

$$R := \begin{pmatrix} 1 & -0.5 & -0.6 \\ -0.5 & 1 & 0.9 \\ -0.6 & 0.9 & 1 \end{pmatrix}$$

Analog dazu gibt es auch eine **Kovarianzmatrix**.

Die Kennzahlen der Regression mit nur einer Eingangsvariablen lassen sich auf den Fall mehrerer Eingangsvariablen analog verallgemeinern:

Kennzahlen für den Zusammenhang mehrerer Variablen

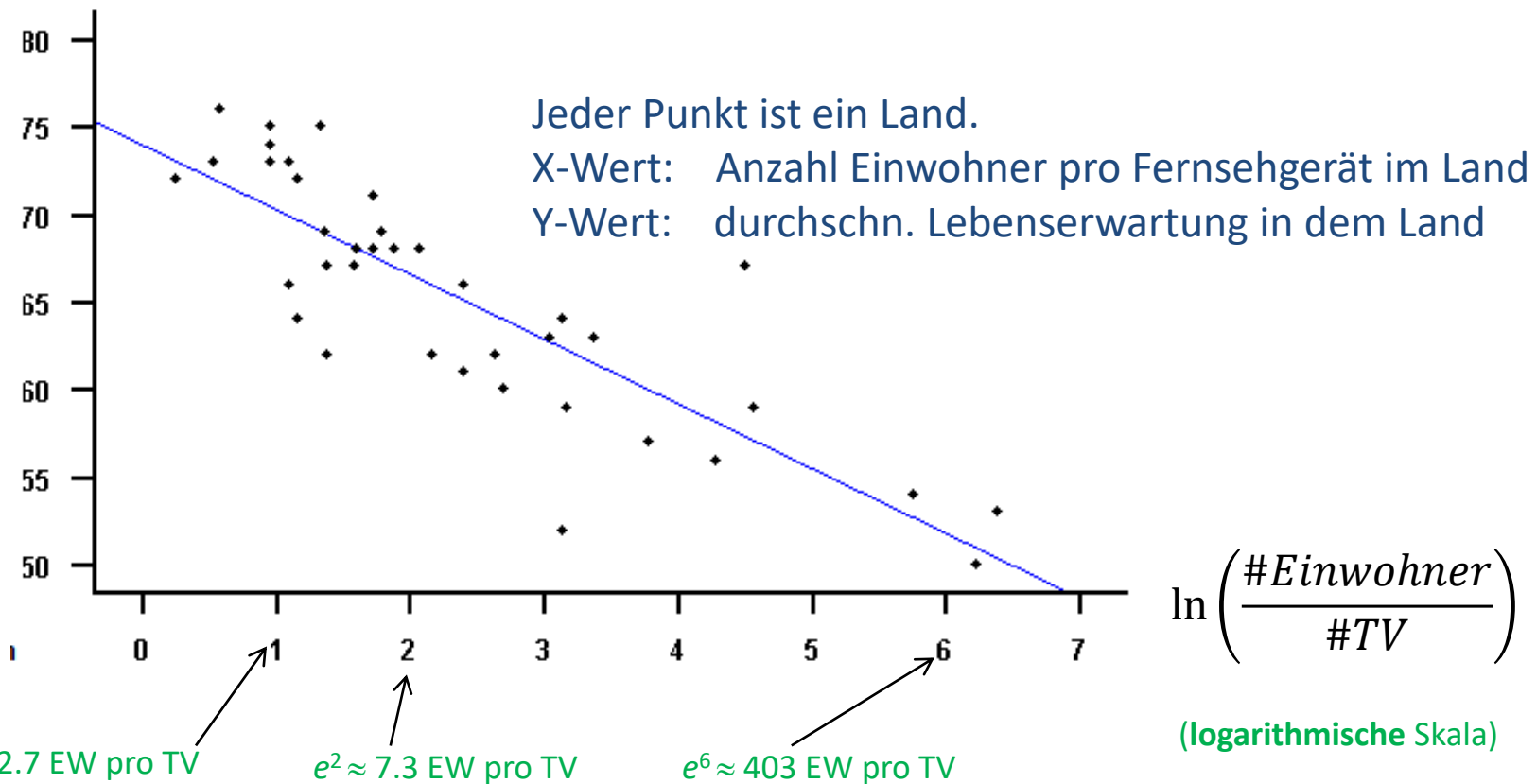
Kennzahlen zum Beschreiben des Zusammenhangs zwischen einem Vektor aus mehreren Eingangsvariablen \vec{x} und einer Ausgangsvariablen y in einem Datensatz :

1. **Vektor \vec{k} der Steigungen** der Regressionsgeraden von den Eingangs- auf die Ausgangsvariable (R: `lm(y~x) ;`)
2. **Mittleres Fehlerquadrat** der Regressionsgeraden : *MSE*
(R: `reg<-lm(y~x) ; mean(summary(reg)$residuals^2)`)
3. **Korrelationsvektor** zwischen der Ausgangs- und den einzelnen Eingangsvariablen: $\vec{r_{xy}}$
(R: `cor(x, y)`)
Wie stark korrelieren die einzelnen Eingangsvariablen mit der Zielvariablen?
4. **Bestimmtheitsmaß b** :

$$b_{xy} := \frac{\text{Durch } x \text{ erklärte Varianz von } y}{\text{Varianz}(y)} = 1 - \frac{MSE}{\text{Varianz}(y)}$$

Scheinkorrelationen (1)

Lebenserwartg. m.



→ Korreliert, es besteht aber kein kausaler Zusammenhang

Scheinkorrelationen (2)

Satz

Aus einer Korrelation zweier Merkmale x und y folgt nicht, dass ein direkter kausaler Zusammenhang bestehen muss.

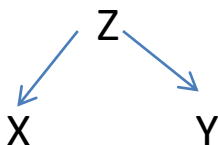
Beispiel:

Im Beispiel von der vorigen Folie sind X (Fernseher) und Y (Lebenserwartung) **korreliert**.

Kausale Fragestellung:

Ändert sich die Lebenserwartung, wenn wir einem Land mehr Fernsehgeräte schenken?

Mögliches kausales Modell

- (1) : $X \longrightarrow Y$ (Fernsehen ist gesund)
- (2) : $X \longleftarrow Y$ (alte Leute kaufen Fernseher)
- (3) :  (Wohlstand beeinflusst beides)

Prognose

- Lebenserwartung steigt
- Lebenserwartung bleibt gleich
- Lebenserwartung bleibt gleich

Alle drei Modelle erklären die beobachtete Korrelation. **Welches kausale Modell zutrifft kann ohne Vorwissen nicht aus den Daten beurteilt werden, die durch rein passive Beobachtung gewonnenen wurden.**

Zusammenfassung Regression

- Die Stichprobenkorrelation zweier Merkmale beschreibt, wie nahe die Punkte des Streudiagramms an einer optimal hindurchgelegten Geraden liegen – wobei der Abstand im Vergleich zur Standardabweichung der beiden Merkmale bewertet wird.
 - Korrelation +1 bzw. -1: Die Punkte liegen perfekt auf einer steigenden bzw. fallenden Geraden.
 - Korrelation 0: Die Steigung der optimalen Geraden ist 0 –lineare Regression bringt also nichts.
- Die Regressionsgerade beschreibt (so gut das mit einer Geraden möglich ist), wie sich der Mittelwert des einen Merkmals abhängig vom anderen verändert.
- Statistisch unabhängig \Rightarrow unkorreliert
Die Umkehrung gilt NICHT! (da evtl. nichtlineare Zusammenhänge bestehen)
- Aus einer Korrelation folgt NICHT, dass ein direkter kausaler Zusammenhang bestehen muss.

Rechengesetze für Erwartungswert und Varianz

Beispiel:

In einer Beratungsfirma bestehen die Projektteams im Durchschnitt aus 5 Mitarbeitern und die durchschnittliche Projektdauer beträgt 70 Arbeitstage.

Was kann man daraus über die im Schnitt pro Projekt benötigte Anzahl an Personentagen schließen?

Rechengesetze für Erwartungswert und Varianz

Satz: 27.28 Für Zufallsvariablen X, Y des selben Zufallsexperimentes, und Konstante a, b (die also nicht vom Ausgang des Experimentes abhängen) gilt:

$$(E1) \quad E(aX + b) = a \cdot E(X) + b$$

$$(V1) \quad V(aX + b) = a^2 \cdot V(X)$$

$$(E2) \quad E(X + Y) = E(X) + E(Y)$$

$$(V2) \quad V(X + Y) = V(X) + V(Y) + 2Cov(X, Y)$$

$$(E3) \quad E(X \cdot Y) = E(X) \cdot E(Y) + Cov(X, Y)$$

$$(C1) \quad Cov(aX, Y) = Cov(X, aY) = a \cdot Cov(X, Y)$$

$$(C2) \quad Cov(X, X) = V(X)$$

Falls X von Y unabhängig ist, vereinfachen sich die Formeln, da dann $Cov(X, Y) = 0$.

E1, E2, V1 und V2 gelten analog auch für mehr als zwei Zufallsvariablen X_1, \dots, X_n , insbesondere:

$$(E4) \quad E(a_1X_1 + a_2X_2 + \dots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \dots + a_nE(X_n)$$

Falls die X_1, \dots, X_n paarweise voneinander stochastisch **unabhängig** sind gilt zusätzlich:

$$(V3) \quad V(a_1X_1 + a_2X_2 + \dots + a_nX_n) = a_1^2V(X_1) + a_2^2V(X_2) + \dots + a_n^2V(X_n)$$

Vorsicht: Für die Standardabweichung gelten diese Rechenregeln nur indirekt über die zugehörige Varianz!

Herleitung der Rechenregeln für den Erwartungswert

Beweis der Formel $E(aX + b) = a E(X) + b$:

Seien x_1, \dots, x_n die Realisationen von X , dann gilt:

$$\begin{aligned} E(aX + b) &= \sum_i (a x_i + b) P(X = x_i) = a \sum_i x_i P(X = x_i) + b \underbrace{\sum_i P(X = x_i)}_{=1} = \\ &= a E(X) + b \end{aligned}$$

Beweis der Formel $E(X + Y) = E(X) + E(Y)$:

Seien x_1, \dots, x_n die Realisationen von X , und y_1, \dots, y_m die Realisationen von Y . Im Folgenden verwenden wir die Abkürzungen $p_{i,j} := P(X = x_i \cap Y = y_j)$:

$$\begin{aligned} E(X + Y) &= \sum_i \sum_j (x_i + y_j) p_{i,j} = \sum_i \sum_j x_i p_{i,j} + \sum_i \sum_j y_j p_{i,j} = \\ &= \sum_i x_i \sum_j p_{i,j} + \sum_j y_j \sum_i p_{i,j} = \\ &= \sum_i x_i \cdot P(X = x_i) + \sum_j y_j \cdot P(Y = y_j) = E(X) + E(Y) \end{aligned}$$

Herleitung der Rechenregeln für den Erwartungswert

Satz 27.28 Seien X und Y zwei **unabhängige** Zufallsvariablen. Dann ist der Erwartungswert des Produkts gleich dem Produkt der Erwartungswerte:

$$E(X \cdot Y) = E(X) \cdot E(Y).$$

Warum? Wir betrachten wieder nur den diskreten Fall. Die Werte von X bzw. Y seien x_i bzw. y_j mit zugehörigen Wahrscheinlichkeiten p_i bzw. q_j . Wegen der Unabhängigkeit von X und Y tritt der Wert $x_i y_j$ mit der Wahrscheinlichkeit $P(X = x_i \text{ und } Y = y_j) = P(X = x_i) \cdot P(Y = y_j) = p_i q_j$ auf. Daher ist

$$E(X \cdot Y) = \sum_i \sum_j (x_i y_j) (p_i q_j) = \sum_i \sum_j (x_i p_i) (y_j q_j) = \left(\sum_i x_i p_i \right) \left(\sum_j y_j q_j \right) = E(X) E(Y).$$

Herleitung der Rechenregel für Varianzen

Satz 27.34 Sei X eine Zufallsvariable und a, b beliebige reelle Zahlen. Dann ist die Varianz der Zufallsvariablen $Y = aX + b$

$$\text{Var}(Y) = a^2 \text{Var}(X).$$

Für die Standardabweichung folgt: $\sigma_Y = |a| \sigma_X$.

Die Verschiebung um b kümmert die Varianz also nicht (im Gegensatz zum Erwartungswert, für den ja $E(aX + b) = a E(X) + b$ gilt). Insbesondere ist also $\text{Var}(X + b) = \text{Var}(X)$ und $\text{Var}(aX) = a^2 \text{Var}(X)$.

Denn: $\text{Var}(aX + b) = E((aX + b - a\mu - b)^2) = E((aX - a\mu)^2) = E(a^2(X - \mu)^2) = a^2 E((X - \mu)^2) = a^2 \text{Var}(X)$, wobei wir die Linearität des Erwartungswerts aus Satz 27.23 verwendet haben.

Beispiel: Rechenregeln für den Erwartungswert

Beispiel 27.29 Erwartungswert des Produktes von Zufallsvariablen

Betrachten Sie den Wurf zweier Würfel und sei $X_1 = \text{Augenzahl des ersten Würfels}$ bzw. $X_2 = \text{Augenzahl des zweiten Würfels}$. Die zugehörigen Erwartungswerte sind (siehe Beispiel 27.24) $E(X_1) = E(X_2) = 3.5$. Berechnen Sie:

- a) $E(X_1 \cdot X_2)$ b) $E(X_1 \cdot X_1)$

Lösung

- a) Natürlich kann man die komplette Wahrscheinlichkeitsverteilung von $Y = X_1 \cdot X_2$ bestimmen und den Erwartungswert dann gemäß seiner Definition berechnen. Da X_1 von X_2 stochastisch unabhängig ist, also $\text{Cov}(X_1, X_2) = 0$ geht es nach (E3) auch einfacher:

$$E(X_1 \cdot X_2) = E(X_1) \cdot E(X_2) = 3.5 \cdot 3.5 = 12.25$$

- b) X_1 ist von sich selbst natürlich nicht stochastisch unabhängig, insofern hilft (E3) diesmal nicht weiter. Zum Berechnen nach der Definition sind alle 6 möglichen Fälle zu berücksichtigen:

$$E(X_1 \cdot X_1) = \frac{1^2}{6} + \frac{2^2}{6} + \cdots + \frac{6^2}{6} = 15 \frac{1}{6}$$

Beispiele

Beispiel:

- a) Ein Würfel wird geworfen. Berechnen Sie den Erwartungswert der Augenzahl.
- b) Berechnen Sie den Erwartungswert der Augensumme von zwei Würfeln mit Hilfe der Rechengesetze und des Ergebnisses aus (a).
- c) Bei einem Spiel zahlen Sie 13 Cent Einsatz. Sie würfeln mit 2 Würfeln und bekommen $2 \cdot (\text{Augensumme} - 1)$ Cent ausbezahlt. Berechnen Sie den Erwartungswert Ihres Gewinns.
- d) Bei einem Spiel zahlen Sie 13 Cent Einsatz. Sie würfeln mit 2 Würfeln und bekommen $(\text{Augensumme}^2 - 49)$ Cent ausbezahlt. Berechnen Sie den Erwartungswert Ihres Gewinns.

Lösung:

a) Sei Z_1 die Augenzahl eines Würfels: $E(Z_1) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \dots + \frac{1}{6} \cdot 6 = \underline{\underline{3.5}}$

b) Sei Z_1 die Augenzahl des ersten, Z_2 die des zweiten Würfels und Z die Augensumme:

$$E(Z) = E(Z_1 + Z_2) = E(Z_1) + E(Z_2) = 3.5 + 3.5 = \underline{\underline{7}}$$

c) Sei G der Gewinn.

Ansatz mit Rechengesetzen:

$$E(G) = E(2 \cdot Z - 2 - 13) = 2 \cdot E(Z) - 15 = 2 \cdot 7 - 15 = \underline{\underline{-1 \text{ Cent}}}$$

d) Man kann nicht einfach den EW in die Formel einsetzen, deshalb:

Entweder über alle 36 Fälle gehen: (Ergebnis aus Aufgabe 27.3 benutzen):

$$E(G) = \frac{1}{36} \cdot (2^2 - 49 - 13) + \frac{2}{36} \cdot (3^2 - 49 - 13) + \frac{3}{36} \cdot (4^2 - 49 - 13) + \dots + \frac{6}{36} \cdot (7^2 - 49 - 13) + \dots + \frac{1}{36} \cdot (12^2 - 49 - 13) = -\underline{\underline{7.2 \text{ Cent}}}$$

$$\text{Oder: } E(G) = E(Z^2 - 49 - 13)$$

$$= (\text{nach E1}) \quad E(Z^2) - 49 - 13$$

$$= (\text{nach E3}) \quad (E(Z) \cdot E(Z) + \text{Cov}(Z, Z)) - 49 - 13 = 7 \cdot 7 + V(Z) - 49 - 13 =$$

$$= (\text{nach c2, } V(Z)=5.83, \text{ s. Folie 127}) \quad 49 + 5.83 - 49 - 13 = -\underline{\underline{7.2 \text{ Cent}}}$$

Beispiel (Rechengesetze)

Was bedeuten Ausdrücke wie $E(a \cdot X + b)$?

Alles, was zu jedem Ergebnis des Zufallsexperiments eine Zahl liefert, ist eine Zufallsvariable. Daher kann man Erwartungswert und Varianz ausrechnen.

Beispiel:

Bei einem Spiel zahlen Sie 13 Cent Einsatz. Sie würfeln mit 2 Würfeln und bekommen $(2 \cdot \text{Augensumme} - 2)$ Cent ausbezahlt. Berechnen Sie Erwartungswert und Varianz Ihres Gewinns.

Lösungsweg 1 (ohne Verwendung der Rechenregeln)

Der Gewinn $G := 2 \cdot Z - 2 - 13 = 2 \cdot Z - 15$ bei jedem Spiel ist genauso eine Zufallsvariable wie die Augensumme Z . Die zugehörige W-Verteilung ist:

x_i	2	3	4	5	6	7	8	9	10	11	12	
$p_i := P(Z=x_i)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	$\sum p_i = 1$
$g_i := 2 \cdot x_i - 2 - 13$	-11	-9	-7	-5	-3	-1	1	3	5	7	9	

Erwartungswert und Varianz von G werden wie gewohnt berechnet:

$\mu := E(G) =$	$-11 \cdot 1/36 +$	$(-9) \cdot 2/36 +$	$(-7) \cdot 3/36$	$+$	\dots	\dots	\dots	\dots	$+$	$7 \cdot 2/36 +$	$9 \cdot 1/36$	<u>= -1</u>
$(g_i - \mu)^2 =$	$(-11 - (-1))^2$	$(-9 - (-1))^2$	$(-7 - (-1))^2$			\dots				$(7 - (-1))^2$	$(9 - (-1))^2$	
	100	64	36	16	4	0	4	16	36	64	100	
V(G) =	$100 \cdot 1/36 +$	$64 \cdot 2/36 +$	$36 \cdot 3/36 +$	$+$	\dots	\dots	\dots	\dots	$+$	$64 \cdot 2/36 +$	$100 \cdot 1/36$	<u>= 23.3</u>

Lösungsweg 2 (mit Rechenregeln)

In jedem Spiel lassen sich die folgenden Zufallsvariablen beobachten:

$Z :=$ Augensumme, $G :=$ Gewinn, $Z_i :=$ Augenzahl des i -ten Würfels

Ansatz:

$$E(G) = E(2 \cdot Z - 2 - 13) \stackrel{E1}{=} 2 \cdot E(Z) - 15 \quad (*)$$

$$E(Z) = E(Z_1 + Z_2) \stackrel{E2}{=} E(Z_1) + E(Z_2) \quad (**)$$

$$E(Z_1) = E(Z_2) = 1 \cdot 1/6 + 2 \cdot 1/6 + \dots + 6 \cdot 1/6 = 3.5 \quad (\text{EW eines Würfels})$$

$$\text{Also} \quad E(G) \stackrel{*}{=} 2 \cdot E(Z) - 15 \stackrel{**}{=} 2 \cdot (3.5 + 3.5) - 15 = \underline{\underline{-1}}$$

Ein Spiel kostet Sie also im Mittel 1 Cent Verlust.

Analog für die Varianz:

$$V(Z_1) = V(Z_2) = (1-3.5)^2 \cdot 1/6 + (2-3.5)^2 \cdot 1/6 + \dots + (6-3.5)^2 \cdot 1/6 \approx 2.917 \quad (\text{Var eines Würfels})$$

$V2$, da X_1 und X_2 unabhängig

$$V(Z) = V(Z_1 + Z_2) \Rightarrow V(Z_1) + V(Z_2) + 2 \cdot 0 \approx 2.917 + 2.917 = 5.83$$

$$V(G) = V(2 \cdot Z - 2 - 13) \stackrel{V1}{=} 2^2 \cdot V(Z) \approx \underline{\underline{23.3}}$$

$V1$

Dieser Ansatz lässt sich problemlos auch bei 5 Würfeln anwenden. Ohne Rechenregeln wären 5 Würfeln dagegen extrem aufwendig!

Beispiel (wie vorher, aber 5 Würfel):

Bei einem Spiel zahlen Sie 40 Cent Einsatz. Sie würfeln mit 5 Würfeln und bekommen ($3 \cdot \text{Augensumme} - 15$) Cent ausbezahlt.

Berechnen Sie den Erwartungswert Ihres Gewinns.

$X :=$ Augensumme, $X_i :=$ Augenzahl beim i -ten Würfel

$G :=$ Gewinn

Gesucht: $E(G) = E(3 \cdot X - 15 - 40) \stackrel{E1}{=} 3 \cdot (E(X)) - 55$

Es gilt $X = X_1 + X_2 + \dots + X_5$, also auch

$E(X) = E(X_1 + X_2 + \dots + X_5) \stackrel{E2 \text{ mehrfach bzw. } E4}{=} E(X_1) + E(X_2) + \dots + E(X_5) = 3.5 \cdot 5 = 17.5$

Also $E(G) = 3 \cdot E(X) - 55 = 3 \cdot 17.5 - 55 = -2.5$

Ein Spiel kostet Sie also im Mittel 2.5 Cent.

Vorsicht: Für Varianz und Standardabweichung sind die Regeln nicht so einfach:

$$V(G) = V(3 \cdot X - 15 - 40) = 3^2 \cdot V(X_i)$$

Beispiel (Erwartungswert, Standardabweichung, Kovarianz)

Sie sind zwei Geschäfte eingegangen, die beide ein gewisses Risiko beinhalten. Im Normalfall machen Sie einen moderaten Gewinn, mit einer geringen Wahrscheinlichkeit aber einen hohen Verlust:

Verlustwahrsch.		Gewinn im Normalfall [GE]	Gewinn im Verlustfall [GE]
A	10%	150	-1000
B	10%	150	-1000

Sie interessieren sich für den Gesamtgewinn G aus beiden Geschäften und gehen zunächst davon aus, dass der Verlauf beider Geschäfte voneinander stochastisch **unabhängig** ist:

a) Erstellen Sie eine Tabelle der Wahrscheinlichkeitsfunktion von G .

b) Berechnen Sie $E(G)$ und $\sigma(G)$ aus der in (a) erstellten Tabelle.

c) Seien G_A und G_B der Gewinn (bzw. Verlust) aus Geschäft A bzw. B.

Berechnen Sie $E(G_A)$, $E(G_B)$, $\sigma(G_A)$, $\sigma(G_B)$, $\text{Cov}(G_A, G_B)$ und daraus mit Hilfe der Rechengesetze $E(G)$ und $\sigma(G)$.

Abweichend von a-c nehmen Sie nun keine Unabhängigkeit mehr an, sondern dass die gemeinsame Wahrscheinlichkeitsverteilung zu beiden Geschäfte wie folgt aussieht:

d) Wiederholen Sie (c) unter diesen veränderten Annahmen.

e) Wiederholen Sie a-b unter diesen veränderten Annahmen.

Beispiel (Lösung)

a) Für alle möglichen Werte g_i von G die Wahrscheinlichkeit ermitteln:

Wahrscheinlichkeit	$G_A=150$	$G_A=-1000$	
$G_B=150$	$0.9 * 0.9 = \underline{81\%}$	$0.1 * 0.9 = \underline{9\%}$	90,0%
$G_B=-1000$	$0.9 * 0.1 = \underline{9\%}$	$0,1 * 0,1 = \underline{1\%}$	10,0%
	90,0%	10,0%	

Für die Wahrscheinlichkeitsverteilung von G müssen nur noch die zugehörigen Werte von G ausgerechnet werden:

da unabhängig

g		$P(G=g)$	
300	$=150+150$	81,0%	$= 0.9 * 0.9$
-850	$= - 1000 + 150$	18,0%	$= 0.1 * 0.9$
	$= 150 - 1000$		$= 0.9 * 0.1$
-2000	$= - 1000 - 1000$	1,0%	$= 0.1 * 0.1$

b)

$$E(G) := \sum_i g_i \cdot P(G = g_i) = 300 * 0.81 - 850 * 0.18 - 2000 * 0.01 = 70 [\text{GE}]$$

$$V(G) := \sum_i (g_i - E(G))^2 \cdot P(G = g_i) = (300 - 70)^2 * 0.81 + (-850 - 70)^2 * 0.18 + (-2000 - 70)^2 * 0.01 \approx 238050 [\text{GE}^2]$$

$$\sigma(G) = \sqrt{V(G)} \approx \sqrt{238050} \approx 488 [\text{GE}]$$

Beispiel (Lösung)

	Ausfallwahrsch	Gewinn im Normalfall [GE]	Gewinn bei Ausfall [GE]
A	10%	150	-1000
B	10%	150	-1000

Wahrscheinlichkeit	$G_A = 150$	$G_A = -1000$	
$G_B = 150$	$0.9 * 0.9 = 81\%$	$0.1 * 0.9 = 9\%$	90,0%
$G_B = -1000$	$0.1 * 0.9 = 9\%$	$0.1 * 0.1 = 1\%$	10,0%
	90,0%	10,0%	

$$c) \quad E(G_A) = E(G_B) = 150 \cdot 0.9 + (-1000) \cdot 0.1 = 35$$

$$V(G_A) = V(G_B) = (150 - 35)^2 \cdot 0.9 + (-1000 - 35)^2 \cdot 0.1 = 119\,025 \text{ [GE}^2\text{]}$$

$$\sigma(G_A) = \sigma(G_B) = 345 \text{ [GE]}$$

$$\begin{aligned} \text{Cov}(G_A, G_B) &= (150 - 35) \cdot (150 - 35) \cdot 0.81 + (-1000 - 35) \cdot (150 - 35) \cdot 0.09 + \\ &\quad (150 - 35) \cdot (-1000 - 35) \cdot 0.09 + (-1000 - 35) \cdot (-1000 - 35) \cdot 0.01 = 0 \end{aligned}$$

Bemerkung: Die Berechnung von $\text{Cov}(G_A, G_B)$ hätte man sich sparen können, denn **unabhängige Zufallsvariablen haben immer Kovarianz Null**

Mit den Rechengesetzen:

$$E(G) = E(G_A + G_B) = E(G_A) + E(G_B) = 35 + 35 = 58.5 \text{ [GE]}$$

$$\begin{aligned} V(G) &= V(G_A + G_B) = V(G_A) + V(G_B) + 2\text{Cov}(G_A, G_B) = 119\,025 + 119\,025 + 2 \cdot 0 = \\ &= 238\,050 \text{ [GE}^2\text{]} \end{aligned}$$

$$\sigma(G) = \sqrt{V(G)} \approx 588 \text{ [GE]}$$

Beispiel (Forts.)

- d) Das Eintreten der Verluste bei Geschäft A und B werde nun **nicht mehr als unabhängig** angenommen, sondern die gemeinsame Wahrscheinlichkeitsverteilung sei durch nebenstehende Tabelle angegeben.

Wahrscheinlichkeit	$G_A=150$	$G_A=-1000$	
$G_B=150$	90%	0%	90,0%
$G_B=-1000$	0%	10%	10,0%
	90,0%	10,0%	

Damit ergibt sich nebenstehende (geänderte) Wahrscheinlichkeitsverteilung:

g		$P(G=g)$
300	$=150+150$	90,0%
-850	$= -1000 + 150$ $= 150 - 1000$	0,0%
-2000	$= -1000 - 1000$	10,0%

Trotzdem unverändert gegenüber (c):

$$P(\text{Verlust mit A}) = P(\text{Verlust mit B}) = 10\%$$

$$E(G_A) = E(G_B) = 35 [\text{GE}] \quad V(G_A) = V(G_B) = 119\,025 [\text{GE}^2] \quad \sigma(G_A) = \sigma(G_B) = 345 [\text{GE}]$$

$$\begin{aligned} \text{Neu ist: } \text{Cov}(G_A, G_B) &= (150 - 35) \cdot (150 - 35) \cdot 0.90 + (150 - 35) \cdot (-1000 - 35) \cdot 0 + \\ &\quad (-1000 - 35) \cdot (150 - 35) \cdot 0 + (-1000 - 35) \cdot (-1000 - 35) \cdot 0.10 \approx 119\,025 [\text{GE}^2] \end{aligned}$$

Mit den Rechengesetzen:

$$E(G) = E(G_A + G_B) = E(G_A) + E(G_B) = 70 [\text{GE}] \quad (\text{E2 gilt auch für abhängige Variablen})$$

$$\begin{aligned} V(G) &= V(G_A + G_B) = V(G_A) + V(G_B) + 2 \text{Cov}(G_A, G_B) = 119\,025 + 119\,025 + 2 \cdot 119\,025 = \\ &= 476\,100 [\text{GE}^2] \end{aligned}$$

$$\sigma(G) = \sqrt{V(G)} \approx \sqrt{476\,100} \approx 690 [\text{GE}]$$

Beispiel (Forts.)

- e) Das Eintreten der Verluste bei Geschäft A und B wird nun **nicht mehr als unabhängig** angenommen, sondern die gemeinsame Wahrscheinlichkeitsverteilung sei durch nebenstehende Tabelle angegeben.

Wahrscheinlichkeit	$G_A=150$	$G_A=-1000$
$G_B=150$	90%	0%
$G_B=-1000$	0%	10%

Damit ergibt sich nebenstehende (geänderte) Wahrscheinlichkeitsverteilung:

g		$P(G=g)$
300	$=150+150$	90,0%
-850	$= -1000 + 150$ $= 150 - 1000$	0,0%
-2000	$= -1000 - 1000$	10,0%

Lösung: Analog zu (a), nur werden die Wahrscheinlichkeiten der möglichen Realisationen von G nicht berechnet, sondern direkt aus der Tabelle abgelesen:

$$E(G) := \sum_i g_i \cdot P(G = g_i) = 300 \cdot 0.9 - 850 \cdot 0 - 2000 \cdot 0.10 = 70 [\text{GE}]$$

Beobachtung: Gleicher Erwartungswert wie bei (b) wg. Rechengesetz E2

Berechnung der Standardabweichung analog zu (b):

$$V(G) := (300 - 70)^2 \cdot 0.90 + (-2000 - 70)^2 \cdot 0.10 \approx 476\,100 [\text{GE}^2]$$

$$\sigma(G) = \sqrt{V(G)} \approx \sqrt{476\,100} \approx 690 [\text{GE}]$$

Zusammenfassung des Beispiels

In Teilaufgabe d/e war die gemeinsame Wahrscheinlichkeitsverteilung gegenüber a,b,c so verändert, dass häufiger in *beiden* Geschäften Verluste oder in *beiden* Gewinne anfielen, als im Fall von Unabhängigkeit. Die Verlustwahrscheinlichkeit in jedem Geschäft für sich alleine war unverändert.

--> Kovarianz (und damit Korrelation) ist positiv.

Die Summe /der Gesamtgewinn

- nimmt häufiger Extremwerte an
- hat höhere Varianz / Standardabweichung
- hat unveränderten Erwartungswert

als bei Unabhängigkeit.

Was Sie gelernt haben sollten

- Kovarianzen und Korrelationen sowohl zu Datensätzen einer Stichprobe als auch aus Wahrscheinlichkeits- bzw. Häufigkeitsverteilungen berechnen.
- Regressionsgeraden bestimmen und interpretieren.
- Den Unterschied zwischen Korrelation und kausalem Zusammenhang kennen.
- Erwartungswert von Zufallsvariablen, die über Formeln definiert sind
 - mit den Rechenregeln aus den Erwartungswerten der zugrundeliegenden Zufallsvariablen berechnen.
 - direkt aus der Wahrscheinlichkeitsverteilung berechnen.