Statistik – Bivariate Deskriptivstatistik

Zusammenhangsmaße

Nach der Untersuchung vorliegender Daten im Umfeld der unvariaten Deskriptivstatistik ist es üblich, dass man auch den Zusammenhang von Variablen untersucht.

Wie schon im univariaten Fall kommen auch hier Kennwerte, Grafiken und Tabellen zum Einsatz.

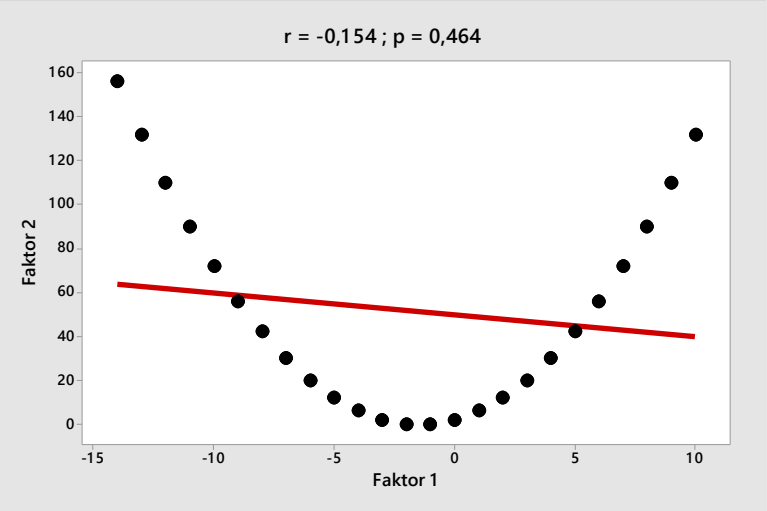
Frage: Gibt es erkennbare Zusammenhänge zwischen zwei (bivariat) oder mehr Variablen (multivariat)?

Korrelationskoeffizient r nach Bravais-Pearson

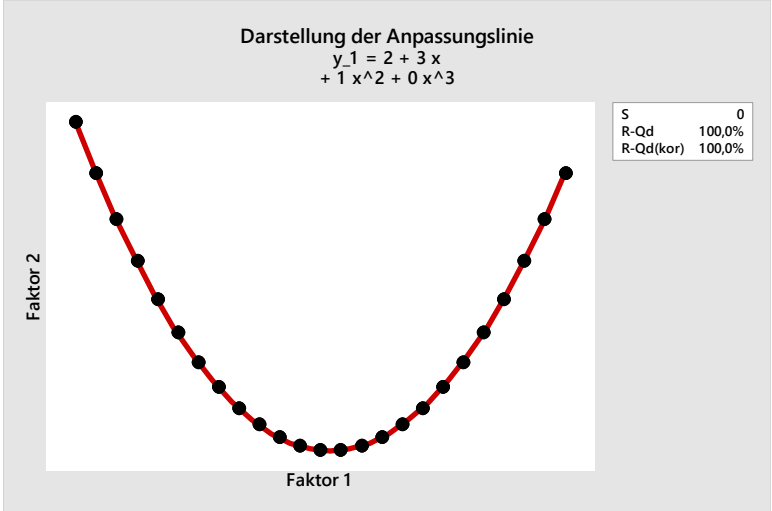
- Kommt zum Einsatz bei metrisch skalierten Merkmalen
- Richtung und Stärke des funktionalen Zusammenhangs werden identifiziert
- Ergebnisse können nicht zur Vorhersage weiterer Werte genutzt werden

Korrelationskoeffizient r nach Bravais-Pearson

- Ausschließlich zur Untersuchung von linearen Zusammenhängen zwischen zwei Variablen geeignet
- Nicht-lineare Zusammenhänge werden nicht erkannt (stattdessen Regressionsanalyse)



Korrelation



Regression

Korrelationskoeffizient r nach Bravais-Pearson

Berechnung des Korrelationskoeffizienten

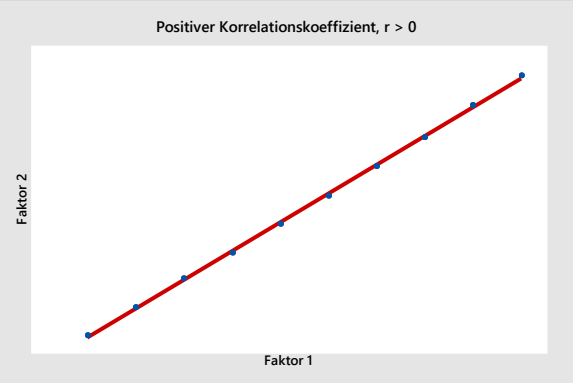
$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} * \sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

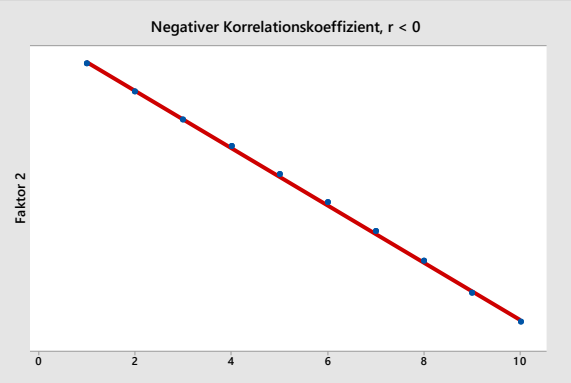
(x_i, y_i) : i-te Ausprägung eines metrisch skalierten Merkmals
n : Anzahl der Wertepaare

Der Korrelationskoeffizient r nach Bravais-Pearson liegt im Bereich $-1 \leq r_{XY} \leq +1$

Korrelationskoeffizient r nach Bravais-Pearson

Deutung: Das Vorzeichen



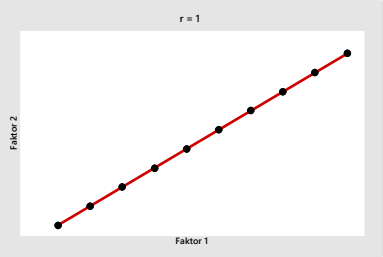


- $r > 0$
- Beide Faktoren entwickeln sich gleichsinnig, wächst Faktor 1, so wächst auch Faktor 2
- Beispiel: Größe und Schuhgröße

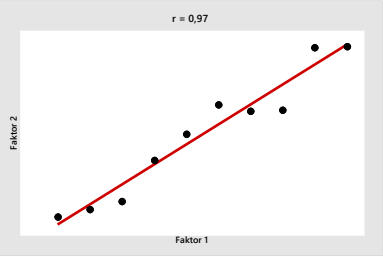
- $r < 0$
- Die Faktoren entwickeln sich gegensinnig, wächst Faktor 1, so sinkt Faktor 2 und umgekehrt
- Beispiel: Temperatur und Anzahl Skiurlauber

Korrelationskoeffizient r nach Bravais-Pearson

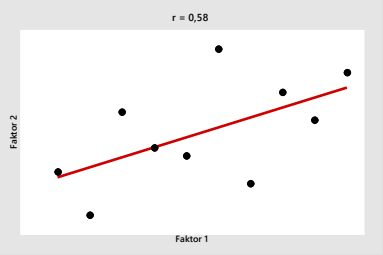
Deutung: Der Betrag



$$0,7 < |r| \leq 1,0$$ **Klare Korrelation**



$$0,3 \leq |r| \leq 0,7$$ **Unklare Korrelation**



$$0,0 \leq |r| < 0,3$$ **Keine Korrelation**

Korrelationskoeffizient r nach Bravais-Pearson

Anmerkungen (gilt auch für andere Korrelationskoeffizienten)

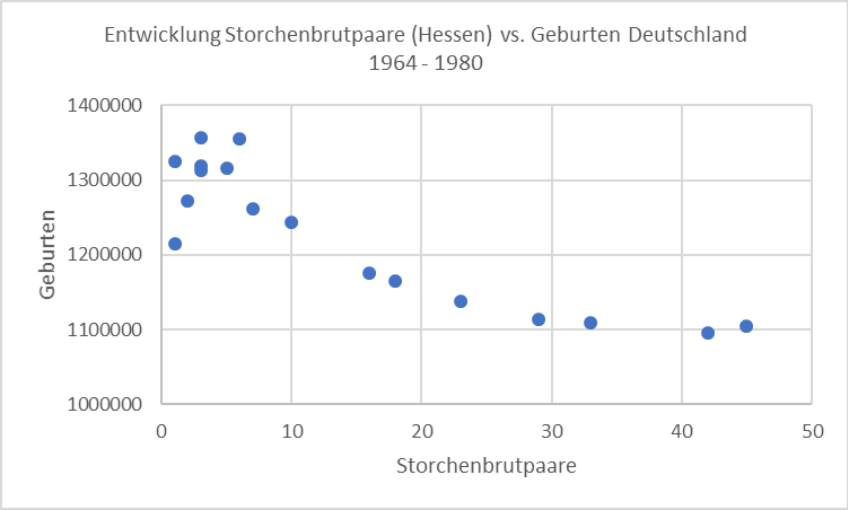
- Der Koeffizient gibt die reine Datenlage wieder
- Eine hohe Korrelation bedeutet nicht zwingend, dass es einen funktionalen Zusammenhang gibt (Nonsens-Korrelation: Nicolas Cage-Filme und Ertrinken im Pool)
- Der Korrelationskoeffizient macht keine Aussage, was hier Ursache und Folge ist (Temperatur und Skitouristen)
- Manchmal korrelieren zwei Merkmale, weil beide von einer dritten Größe abhängen, die nicht erkannt wird (Lurking Variables, Störfaktoren)

Korrelationskoeffizient r nach Bravais-Pearson

Beispiel: Rückgang der Storchenpopulation führt zu sinkender Geburtenrate

	Mensch	Storch
Mensch	1.0000000	-0.8828401
Storch	-0.8828401	1.0000000

Starke negative Korrelation!!!



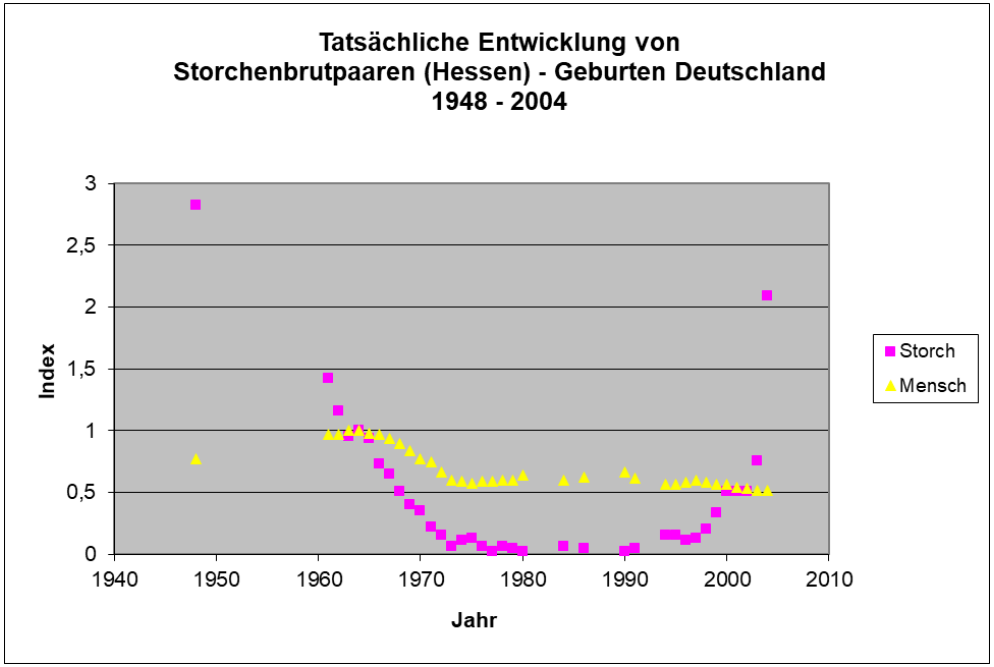
Korrelationskoeffizient r nach Bravais-Pearson

Beispiel: Rückgang der Storchenpopulation führt zu sinkender Geburtenrate

Die Realität sieht natürlich anders aus, wenn man größere Zeiträume betrachtet!

	Mensch	Storch
Mensch	1.000000	0.411727
Storch	0.411727	1.000000

Schwache positive Korrelation



Korrelationskoeffizient r nach Bravais-Pearson

Übung Korrelationskoeffizient nach Bravais-Pearson

Befragter	Größe X [m]	Gewicht Y [kg]
1	1,87	72
2	1,70	60
3	1,80	73
4	1,84	74
5	1,78	72
6	1,80	70
7	1,72	62
8	1,76	70
9	1,86	80
10	1,77	67

Ihnen liegen Daten von 10 Personen vor. Bestimmen Sie den Korrelationskoeffizienten nach Bravais-Pearson.

Deuten Sie die Ergebnisse.

Korrelationskoeffizient r nach Bravais-Pearson

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1,87	72			
2	1,70	60			
3	1,80	73			
4	1,84	74			
5	1,78	72			
6	1,80	70			
7	1,72	62			
8	1,76	70			
9	1,86	80			
10	1,77	67			
Summe					

Korrelationskoeffizient r nach Bravais-Pearson

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1,87	72	3,4969	5184	134,64
2	1,70	60	2,8900	3600	102,00
3	1,80	73	3,2400	5329	131,40
4	1,84	74	3,3856	5476	136,16
5	1,78	72	3,1684	5184	128,16
6	1,80	70	3,2400	4900	126,00
7	1,72	62	2,9584	3844	106,64
8	1,76	70	3,0976	4900	123,20
9	1,86	80	3,4596	6400	148,80
10	1,77	67	3,1329	4489	118,59
Summe	17,9	700	32,0694	49306	1255,59

Korrelationskoeffizient r nach Bravais-Pearson

Mittelwerte: $\overline{x} = 1,79$

$\overline{y} = 70,0$

$$s_{XY} = 1255,59 - 10 * 1,79 * 70,0 = 2,5900$$

$$s_X = \sqrt{32,0694 - 10 * 1,79^2} = 0,168522995$$

$$s_Y = \sqrt{49306 - 10 * 70^2} = 17,492855685$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{2,59}{0,168522995 * 17,49285568} = 0,8785771$$

Es liegt eine starke positive Korrelation vor, d.h. steigt die Größe, steigt auch das Gewicht.

Korrelationskoeffizient r nach Bravais-Pearson

Ergebnisse aus R

	Gewicht.Y..kg.	Größe.X..m.
Gewicht.Y..kg.	1.0000000	**0.8785771**
Größe.X..m.	**0.8785771**	1.0000000

Korrelationskoeffizient nach Spearman

- Korrelationskoeffizient für mindestens ordinal skalierte Daten
- Ordinal skalierte Merkmalsausprägungen haben in der Regel eine eindeutige Rangfolge
- Einführung einer Rangfolge für die jeweiligen Variable (klein nach groß)
- Für gleiche Merkmalsausprägungen bildet man den Rang als arithmetisches Mittel
- Bildung von Rangdifferenzen R_i

Korrelationskoeffizient nach Spearman

$$r_s = \frac{\sum_{i=1}^{n}(R(x_i) - \overline{R(x)})(R(y_i) - \overline{R(y)})}{\sqrt{\sum_{i=1}^{n}\left(R(x_i) - \overline{R(x)}\right)^2 * \sum_{i=1}^{n}\left(R(y_i) - \overline{R(y)}\right)^2}}$$

bzw. $r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$

mit $d_i = R(x_i) - R(y_i)$, falls keine Bindungen auftreten (mehrere Merkmalsausprägungen sind gleich, wir wählen das arithmetische Mittel der betroffenen Ränge)

Es gilt: $-1 \leq r_s \leq +1$

Korrelationskoeffizient nach Spearman

Deutung

- $r > 0$ Positive Korrelation (große x-Werte bedeuten große y-Werte)
- $r \approx 0$ Keine Korrelation
- $r < 0$ Negative Korrelation (große x-Werte bedeuten kleine y-Werte

Korrelationskoeffizient nach Spearman

Übung Korrelationskoeffizient nach Spearman

Befragter	Größe X [m]	Gewicht Y [kg]
1	1,87	72
2	1,70	60
3	1,80	73
4	1,84	74
5	1,78	72
6	1,80	70
7	1,72	62
8	1,76	70
9	1,86	80
10	1,77	67

Ihnen liegen Daten von 10 Personen vor. Bestimmen Sie den Korrelationskoeffizienten nach Spearman

Deuten Sie die Ergebnisse

(Spearman funktioniert auch bei metrischen Skalen)

Korrelationskoeffizient nach Spearman

i	x_i	Rang x_i	y_i	Rang y_i	$R(x_i) - \overline{R(x)}$	$R(y_i) - \overline{R(y)}$
1	1,87		72			
2	1,70		60			
3	1,80		73			
4	1,84		74			
5	1,78		72			
6	1,80		70			
7	1,72		62			
8	1,76		70			
9	1,86		80			
10	1,77		67			
	$\overline{R(x)}$		$\overline{R(y)}$			

Korrelationskoeffizient nach Spearman

i	x_i	Rang x_i	y_i	Rang y_i	$R(x_i) - \overline{R(x)}$	$R(y_i) - \overline{R(y)}$
1	1,87	10	72	6,5	4,5	1
2	1,70	1	60	1	-4,5	-4,5
3	1,80	6,5	73	8	1	2,5
4	1,84	8	74	9	2,5	3,5
5	1,78	5	72	6,5	-0,5	1
6	1,80	6,5	70	4,5	1	-1
7	1,72	2	62	2	-3,5	-3,5
8	1,76	3	70	4,5	-2,5	-1
9	1,86	9	80	10	3,5	4,5
10	1,77	4	67	3	-1,5	-2,5
	$\overline{R(x)}$	5,5	$\overline{R(y)}$	5,5		

Korrelationskoeffizient nach Spearman

$$r_s = \frac{68,75}{\sqrt{82 * 81,5}} = 0,8409825$$

Es liegt eine starke positive Korrelation vor, d.h. steigt die Größe, steigt auch das Gewicht.

Ergebnisse aus R

	Gewicht.Y..kg.	Größe.X..m.
Gewicht.Y..kg.	1.0000000	**0.8409825**
Größe.X..m.	**0.8409825**	1.0000000

Kreuztabellen

- Andere Namen: Kontingenztafeln oder –tabellen
- Besonders geeignet für qualitative oder kategoriale Variablen
- Darstellung von absoluten oder relativen Häufigkeiten der Kombination bestimmter Merkmalsausprägung
- Verknüpfung von Merkmalen („und“)
- Zusätzlich können Randhäufigkeiten gebildet werden

Kreuztabellen

Übung Kreuztabellen

Ihnen liegen aus verschiedenen Bundesländern Daten zur Religionszugehörigkeit vor:

- NRW: RK 42%; P 28%; M 8%; Sonstige 22%
- HH: RK10%; P 30%; M 8%; Sonstige 52%
- BY: RK 55%; P 21%; M 4%; Sonstige 20%

Erstellen Sie eine Kreuztabelle.

Kreuztabellen

Übung Kreuztabellen

Bundesland	K	P	M	Sonstige	Σ
NRW	42	28	8	22	100
HH	10	30	8	52	100
BY	55	21	4	20	100
Σ	107	79	20	94	

Die Auswertung erfolgt grafisch wie rechnerisch zu einem späteren Zeitpunkt.

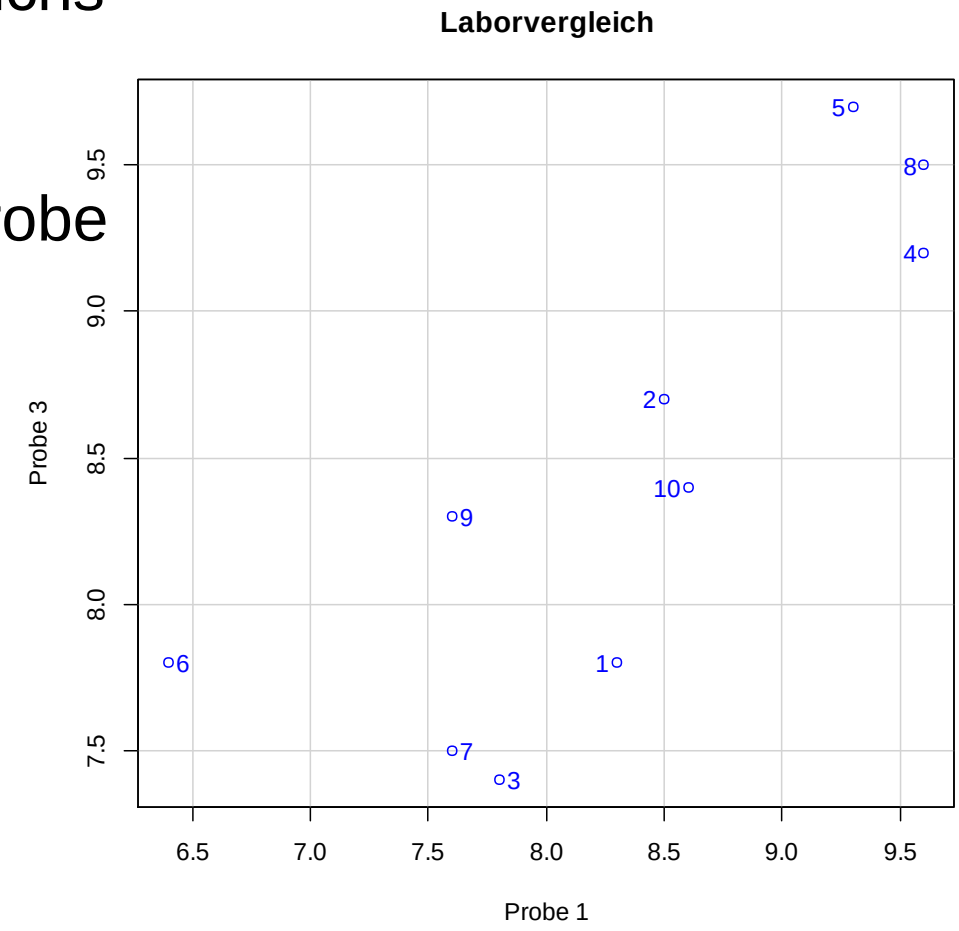
Streudiagramme

- Gemeinsame Verteilung der Werte von zwei oder drei Variablen
- Die zusammengehörenden Werte der Variablen werden gegeneinander aufgetragen
- Die Lage einzelner Wertekombinationen und deren Häufung bzw. Fehlen lässt mögliche Zusammenhänge erkennen
- Ein funktionaler Zusammenhang kann nicht abgelesen werden, es ist nicht einmal erkennbar, ob der Zusammenhang tatsächlich besteht

Streudiagramme

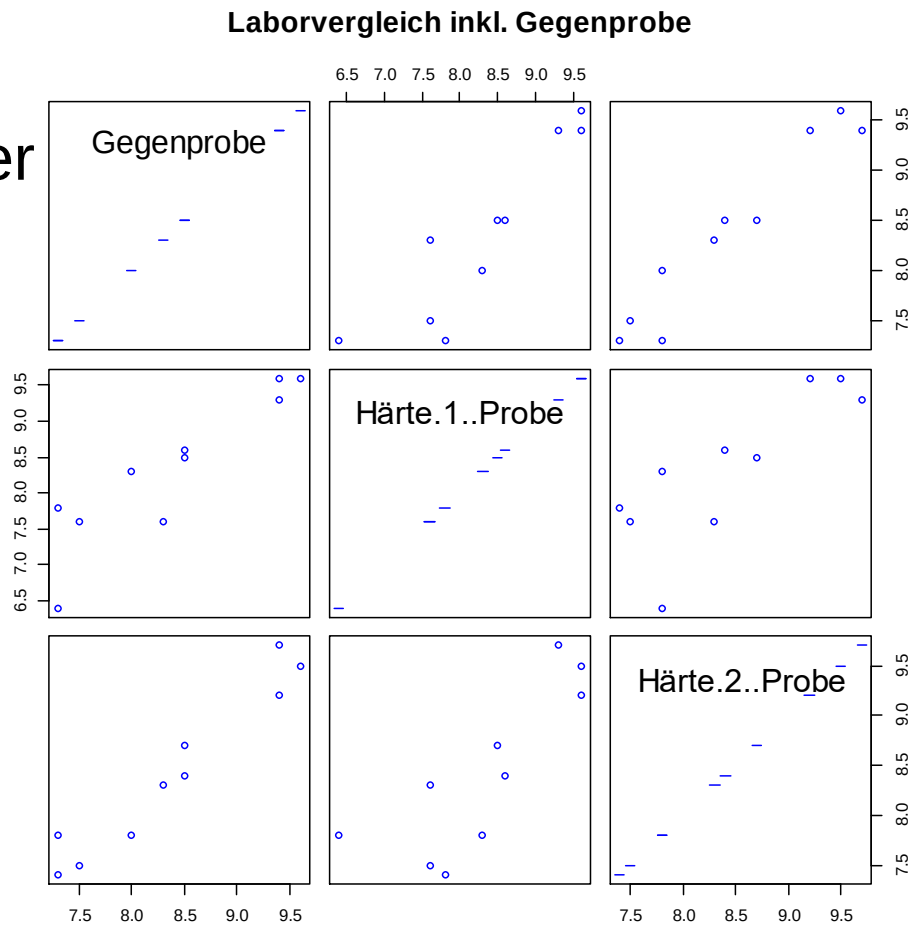
Beispiel eines Laborvergleichs

Erzielen die Labore 1-10
in der erste und zweiten Probe
ähnliche Ergebnisse?



Streudiagramme

- Liegt eine höhere Anzahl an Variablen vor, kann der Vergleich auch mittels einer Streudiagramm-Matrix durchgeführt werden



Gruppierte Balkendiagramme

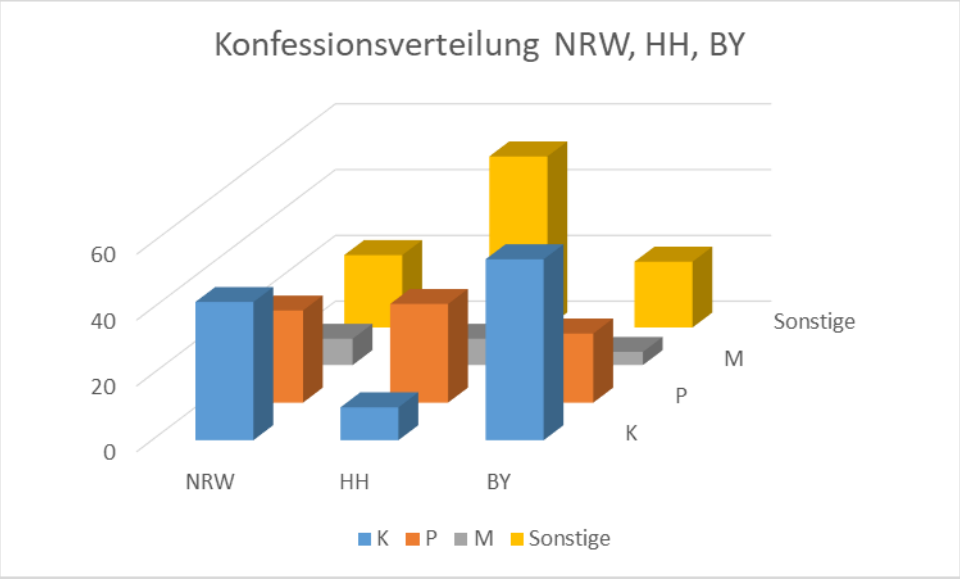
- Grafische Gegenüberstellung von qualitativen Merkmalen (nominal oder ordinal)
- Gegenstück zur Kreuztabelle
- Balkendiagramme sind besonders für diskrete Merkmale geeignet
- Stetige Merkmale können ggf. in Klassen eingeteilt werden
- Verschiedene Darstellungsformen sind möglich (3D, gestapelte Balken)

Gruppierte Balkendiagramme

Im Diagramm wird offensichtlich, bei welchen Kombinationen der betrachteten Merkmale Spitzen oder Senken zu erwarten sind

Wie zu erwarten ist der Anteil der Katholiken in Bayern sehr hoch

Überraschend ist der Hohe Anteil *Sonstige* in Hamburg (Sonstige enthält auch Konfessionslose)



Gruppierte Balkendiagramme

Verschiedene Gruppierungsmöglichkeiten