Mehrdimensionale Häufigkeitsverteilungen (1)

- Die Begriffe univariat und bivariat -

Von univariaten (eindimensionalen) statistischen Analysen spricht man, wenn pro Person nur ein Merkmal in tabellarischen oder grafischen Häufigkeitsverteilungen oder bei der Berechnung von Kennwerten (Mittelwert, Varianz etc.) Gegenstand der Betrachtung ist.

Bei bivariaten (zweidimensionalen) Analysen werden pro Person (Beobachtungs-einheit) zwei Merkmale betrachtet und deren gemeinsame Verteilung in Tabellen oder Grafiken beschrieben.

Die statistischen Kennwerte solcher bivariaten Verteilungen sind Maße des Zusammenhangs zwischen den beiden gemessenen Merkmalen.

Mehrdimensionale Häufigkeitsverteilungen (2)

- Beispiel einer bivariaten Häufigkeitsverteilung -

Häufigkeitsverteilungen der Merkmale ‚Schulnote Mathematik' und ‚erwartete Erfolgschance in der Statistikklausur' in einer Kreuztabelle:

			Mathematiknote			
Erfolgschance	1	2	3	4	5	*Gesamt*
sehr pessimistisch = 5				**1**		1
4		**3**	**4**		**1**	8
3	**1**	**5**	**10**	**7**		23
2	**6**	**14**	**7**	**2**	**1**	30
sehr optimistisch = 1	**4**	**6**	**1**		**1**	12
Gesamt	11	28	22	10	3	74

Mehrdimensionale Häufigkeitsverteilungen (3)

- Der Begriff des Zusammenhangs -

Aussagen über Zusammenhänge von zwei Merkmalen X und Y sind Aussagen, die sich entweder als „wenn-dann“- oder als „je-desto“-Aussagen formulieren lassen.

Beispiel:
Je besser die Abiturnote ist, desto weniger Semester wird jemand studieren.
(Merkmal X: Abiturnote; Merkmal Y: Studiendauer)

Möglichkeiten des linearen Zusammenhang zwischen zwei Merkmalen X und Y

- je größer X, desto größer Y (gleichläufig, positiv)
- je kleiner X desto kleiner Y (gleichläufig, positiv)
- je größer X desto kleiner Y (gegenläufig, negativ)
- je kleiner X desto größer Y (gegenläufig, negativ)
- kein Zusammenhang

Mehrdimensionale Häufigkeitsverteilungen (4)

- Exkurs: Datengrundlage der folgenden Analysen -

Stichprobe:	N=80 ErstsemesterInnen Psychologie
Merkmale:	Aussagen und Skalen aus einem Fragebogen zur Erfassung von Einstellungen gegenüber Statistik

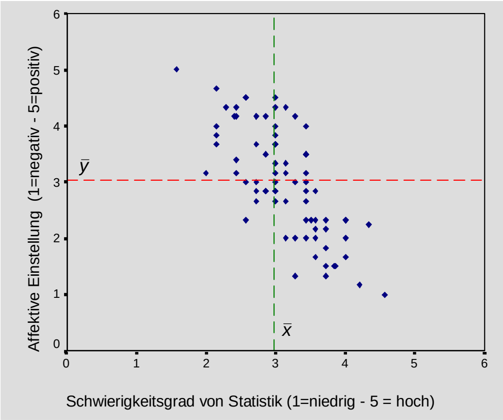
- Beispiel für die Definition einer Einstellungsskala
  Affektive Einstellung zu Statistik = aM der Antworten auf die Aussagen (Items)
  - Statistik find ich gut.
  - Ich fühle mich unwohl bei dem Gedanken, mich mit statistischen Fragestellungen befassen zu müssen.
  - Wenn ich an die Statistikklausur denke, wird mir schon ganz mulmig.
  - Die Statistik-Veranstaltung bedeutet Stress für mich.
  - Ich glaube, dass mir die Statistik-Veranstaltung Spass machen wird.
  - Schon der Gedanke an Statistik schreckt mich ab.
- Die Items sollten auf einer fünftstufigen Ratingskala (1-5) hinsichtlich des Grades der Zustimmung beantwortet werden.

Mehrdimensionale Häufigkeitsverteilungen (5)

- Das Streudiagramm (*scatterplot*) I -

Sind zwei Merkmale X und Y mindestens intervallskaliert, läßt sich die bivariate Verteilung als Punktewolke der Messwertpaare (x_i,y_i) aller Personen *i* in einem Streudiagramm darstellen.

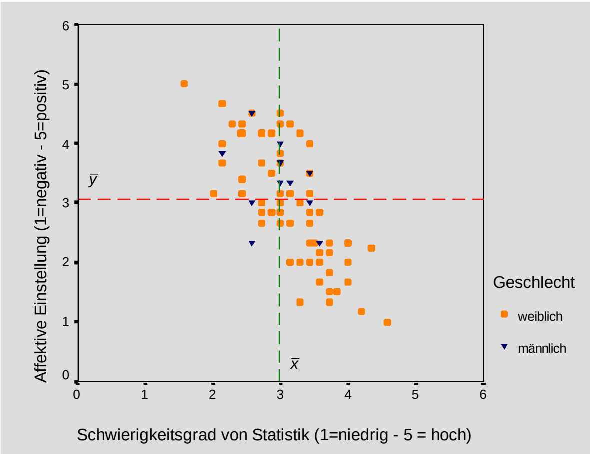
Beispiel:
Streudiagramm der Merkmale Schwierigkeitsgrad (X) und affektive Einstellung zu Statistik (Y); N=79



Mehrdimensionale Häufigkeitsverteilungen (6)

- Das Streudiagramm (*scatterplot*) II -

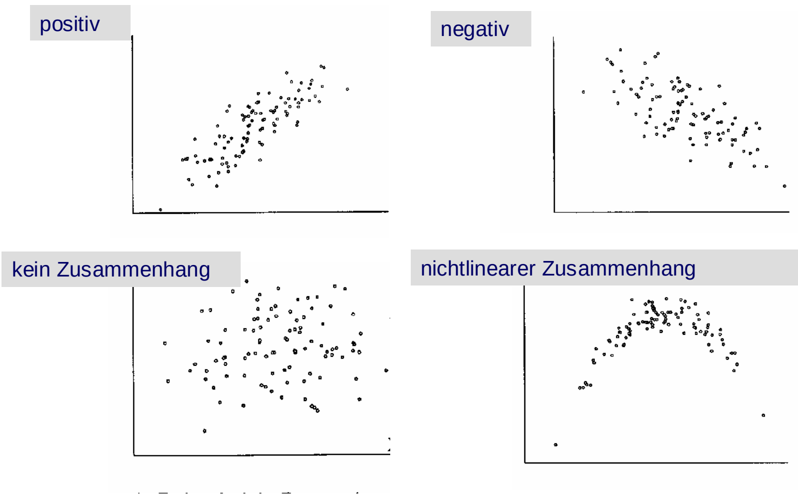
Darstellung des Zusammenhangs der Merkmale Schwierigkeitsgrad (X) und affektive Einstellung zu Statistik (Y) aufgeschlüsselt nach Geschlecht; N=79



Mehrdimensionale Häufigkeitsverteilungen (7)

- Das Streudiagramm (*scatterplot*) III -

Streudiagramme für den Zusammenhang zweier Merkmale X und Y



Mehrdimensionale Häufigkeitsverteilungen (8)

- Zusammenhangsmaße -

In Abhängigkeit des Skalennveaus der erhobenen Daten lassen sich Zusammenhänge mit unterschiedlichen Maßzahlen darstellen.
Die wichtigsten Zusammenhangsmaße sind für

- ⇨ intervallskalierte Daten
  Kovarianz und Korrelation
- ⇨ ordinalskalierte Daten
  Rangkorrelationskoeffizient
- ⇨ nominalskalierte Daten
  Phi-Koeffizient

Die Kovarianz (1)

- Definition -

Die Kovarianz ist definiert als das durchschnittliche Produkt der Abweichungen von korrespondierenden Messwerten x_i und y_i von ihrem jeweiligen Mittelwert $\bar{x}$ bzw.$\bar{y}$.
X und *Y* sollten zur sinnvollen Interpretation der Kovarianz mindestens intervallskalierte Variablen sein.

Die Kovarianz ist

- positiv bei positivem linearen Zusammenhang
- (ungefähr) Null, bei keinem linearen Zusammenhang
- negativ bei gegenläufigem linearem Zusammenhang.

$$s_{xy} = \operatorname{cov}(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})\cdot(y_i - \bar{y})}{N}$$

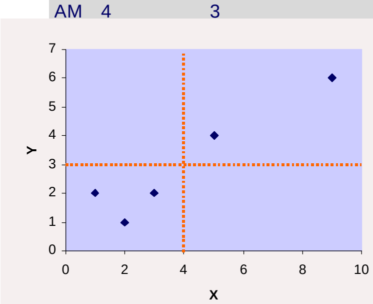
$$= \frac{\sum_{i=1}^{N} x_i \cdot y_i}{N} - \bar{x}\cdot\bar{y}$$

Die Kovarianz (2)

- Beispiel I -

1. (Hohe) *positive* Kovarianz

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})\cdot(y-\bar{y})$
	2	1	-2	-2	4
	1	2	-3	-1	3
	9	6	5	3	15
	5	4	1	1	1
	3	2	-1	-1	1
Σ	20	15			24
AM	4	3			



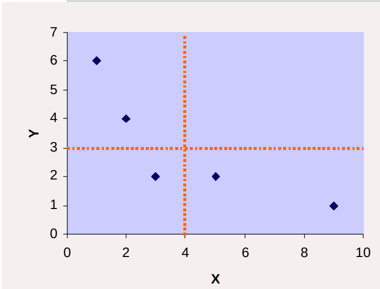
$$s_{xy} = \operatorname{cov}(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})\cdot(y_i - \bar{y})}{N} = 4.8$$

Die Kovarianz (3)

- Beispiel II -

2. (Hohe) *negative* Kovarianz

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})\cdot(y-\bar{y})$
	2	4	-2	1	-2
	1	6	-3	3	-9
	9	1	5	-2	-10
	5	2	1	-1	-1
	3	2	-1	-1	1
Σ	20	15			-21
AM	4	3			



$$s_{xy} = \text{cov}(x,y) = \frac{\sum_{i=1}^{N}(x_i-\bar{x})\cdot(y_i-\bar{y})}{N} = -4.2$$

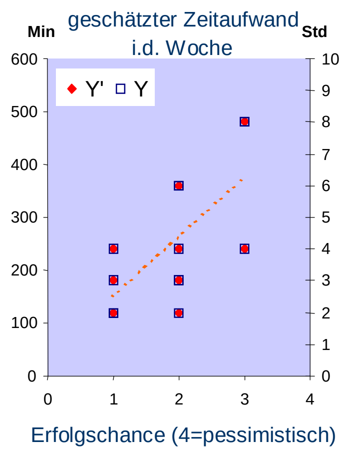
Die Kovarianz (4)

- Die Maßstabsabhängigkeit der Kovarianz I -

Die Kovarianz bleibt bei Lineartransformationen nicht unverändert. Verändert man z.B. die Maßeinheit eines Merkmals, so verändert sich die Höhe der Kovarianz.

Beispiel: N=13 männliche Erstsemester

Subjektive Erfolgschance für die Statistikklausur X	geschätzter Zeitaufwand für Statistik (Std/Woche) Y	geschätzte Zeitaufwand Statistik (min/W Y'
3	8	480
2	4	240
3	4	240
1	2	120
2	6	360
3	8	480
2	3	180
1	4	240
2	2	120
1	2	120
1	3	180
2	3	180
2	6	360



Die Kovarianz (5)

- Die Maßstabsabhängigkeit der Kovarianz II -

Berechnung der Kovarianzen von X (Erfolgschance) mit Y (geschätzter Zeitaufwand in Std./Woche) bzw. Y' (geschätzter Zeitaufwand in min/Woche):

$$\mathrm{cov}(X,Y) = 1.103$$
$$\mathrm{cov}(X,Y') = 66.154 = 60 \cdot \mathrm{cov}(X,Y)$$

Allgemein gilt für das Verhalten der Kovarianz bei linearer Transformation der Merkmale X bzw. Y in X' bzw. Y' mit $X' = a_1 + b_1 \cdot X$ bzw. $Y' = a_2 + b_2 \cdot Y$ $\Rightarrow$

$$\mathrm{cov}(X',Y') = b_1 \cdot b_2 \cdot \mathrm{cov}(X,Y)$$

Die Maßstabsabhängigkeit macht die Kovarianz zu einem uneindeutigen und daher wenig geeigneten Zusammenhangsmaß.

Sie hat ihre Bedeutung in der Statistik vor allem für die Vorhersage von Werten, die anhand einer Regressionsgleichung geschätzt wurden bzw. zur Schätzung der Parameter in Pfadmodellen (hierzu später).