Institut für Soziologie
Dipl. Soz. Maximilian Sonnauer

Methoden II

Zusammenhangsmaße für kategoriale und metrische Variablen

Organisatorisches

Prüfungsanmeldung Methoden II

- Die Anmeldung zur Prüfung läuft über die Tafelübung.
- Aus Einfachheitsgründen ist Maximilian Sonnauer als einziger Prüfer eingetragen.
- Unabhängig von Ihrem Übungstermin melden ALLE Teilnehmer*innen von Methoden II ihre Prüfung bei Maximilian Sonnauer an.

Programm

1. Wiederholung Chi² Unabhängigkeitstest

2. Zusammenhang bei kategorialen Variablen
   - Cramers V

3. Metrische Zusammenhänge grafisch darstellen
   - Streudiagramme
   - Nichtlineare Verfahren

4. Zusammenhangsmaße metrischer Variablen
   - Pearsons Korrelationskoeffizient r

Programm

1. Wiederholung Chi² Unabhängigkeitstest

2. Zusammenhang bei kategorialen Variablen
   - Cramers V

3. Metrische Zusammenhänge grafisch darstellen
   - Streudiagramme
   - Nichtlineare Verfahren

4. Zusammenhangsmaße metrischer Variablen
   - Pearsons Korrelationskoeffizient r

Wiederholung
χ^2-Unabhängigkeitstest

Welche Eigenschaften hat der χ^2-Unabhängigkeitstest?

- Anwendung bei kategorialen Variablen
- Vergleicht beobachtete Häufigkeiten mit erwarteten Häufigkeiten unter Unabhängigkeit

$$\chi^2 = \sum_k \sum_m \frac{(O_{km} - E_{km})^2}{E_{km}}$$

O_{km}: beobachtete Häufigkeiten
E_{km} : erwartete Häufigkeiten

Wiederholung
χ^2-Unabhängigkeitstest – fiktives Beispiel

Politische Ausrichtung		Geschlecht: weiblich	Geschlecht: männlich	
Rechts von der Mitte	$O_{1.}$	125	125	250
	$E_{1.}$	150	100	
Mitte	$O_{2.}$	250	200	450
	$E_{2.}$	270	180	
Links von der Mitte	$O_{3.}$	225	75	300
	$E_{3.}$	180	120	
		600	400	1000

$$\chi^2 = \frac{(125-150)\text{^}2}{150} + \frac{(125-100)\text{^}2}{100} + \frac{(250-270)\text{^}2}{270} + \frac{(200-180)\text{^}2}{180} + \frac{(225-180)\text{^}2}{180} + \frac{(75-120)\text{^}2}{120} = 42{,}25$$

Wiederholung
χ^2-Unabhängigkeitstest

Signifikanztest

Vergleich der berechneten Teststatistik (χ^2) mit dem kritischen Wert aus der Vergleichstabelle. Unabhängigkeit wird abgelehnt, wenn gilt:

$$\chi^2 = 42{,}25 > \chi^2_{1-\alpha}\ \underbrace{(df)}_{\substack{Freiheitsgrade \\ (Zeilen-1)(Spalten-1)}}$$

α entspricht dem festgelegten Signifikanzniveau (Fehler 1. Art). In der Tabelle muss für 1-α nachgeschlagen werden.

df \ P	0.005	0.01	0.025	0.05	0.1	0.9	0.95
1	0.00004	0.0002	0.001	0.004	0.016	2.706	3.842
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31

df = (3-1)(2-1) = 2

1-α = 0.95

Zusammenhangsmaß für nominalskalierte Variablen

Nachteile von χ^2

Keine Interpretation der *Stärke* eines Zusammenhangs, denn…

- die Höhe des Koeffizienten hängt von der Fallzahl ab.
- die Höhe des Koeffizienten hängt von der Anzahl der Zellen einer Tabelle ab, welche die Freiheitsgrade bestimmen (k – 1)*(m – 1).

Programm

1. Wiederholung Chi² Unabhängigkeitstest

2. Zusammenhang bei kategorialen Variablen
   - Cramers V

3. Metrische Zusammenhänge grafisch darstellen
   - Streudiagramme
   - Nichtlineare Verfahren

4. Zusammenhangsmaße metrischer Variablen
   - Pearsons Korrelationskoeffizient r

Cramers V

Cramers V

- Ausgangspunkt ist der Chi² Test
- Normierung des Wertebereichs zwischen 0 und 1.
- Aussagen über die *Stärke* eines Zusammenhangs möglich
- *Keine* Aussage über die Richtung des Zusammenhangs möglich (Vorzeichen ignorieren)

Zusammenhangsmaße für nominale Variablen: Cramers V

Formel:	$V = \sqrt{\frac{\chi^2}{N * \min(k-1, m-1)}}$	▪ k steht für „Zeile“ und m steht für „Spalte“ ▪ Man wählt entweder Anzahl der Spalten oder der Zeilen, je nachdem, was kleiner ist	
Eigenschaften:	Wertebereich von [0,1]		
	Interpretation	V = 0 →	kein Zusammenhang zwischen X und Y
		V = 1 →	perfekter Zusammenhang zwischen X und Y
Achtung:	Keine Interpretation des Vorzeichens möglich		
	Kann für k x m-Tabellen berechnet werden (wobei k=m sein darf)		

Cramers V
Stata-Syntax

Stata Syntax für Cramers V:

```
tabulate abhvar unabhvar, chi2 V
```

„chi2“ ruft den Chi-Quadrat Test auf [siehe letzte Woche]

Durch „V“ wird Cramers V angezeigt

Cramers V
Stata-Beispiel

Beispiel von letzter Woche:

H1: Amerikanische und ausländische Autos unterscheiden sich in ihrer Reparaturanfälligkeit.

Verwendete Variablen:
Dummy zu Herkunft: foreign

Anzahl der bisherigen Reparaturen: rep78

Cramers V
Stata-Beispiel

```
tab rep78 foreign, nofreq column
```

Repair Record 1978	Car type Domestic		Foreign	Total
1	4.17	>	0.00	2.90
2	16.67	>	0.00	11.59
3	56.25	>	14.29	43.48
4	18.75	<	42.86	26.09
5	4.17	<	42.86	15.94
Total	100.00		100.00	100.00

Unterdrückt die Anzeige der absoluten Häufigkeiten

Cramers V
Stata-Beispiel

```
tab rep78 foreign, nofreq column chi2 V
```

Repair Record 1978	Car type Domestic	Foreign	Total
1	4.17	0.00	2.90
2	16.67	0.00	11.59
3	56.25	14.29	43.48
4	18.75	42.86	26.09
5	4.17	42.86	15.94
Total	100.00	100.00	100.00

```
          Pearson chi2(4) =  27.2640   Pr = 0.000
               Cramér's V =   0.6286
```

Der Chi-Quadrat Wert ist mit einem p-Wert = 0.000 höchst signifikant.

Es muss davon ausgegangen werden, dass ein Unterschied besteht.

Cramers V ist mit 0.6286 mittel bis stark.

Programm

1. Wiederholung Chi² Unabhängigkeitstest

2. Zusammenhang bei kategorialen Variablen
   - Cramers V

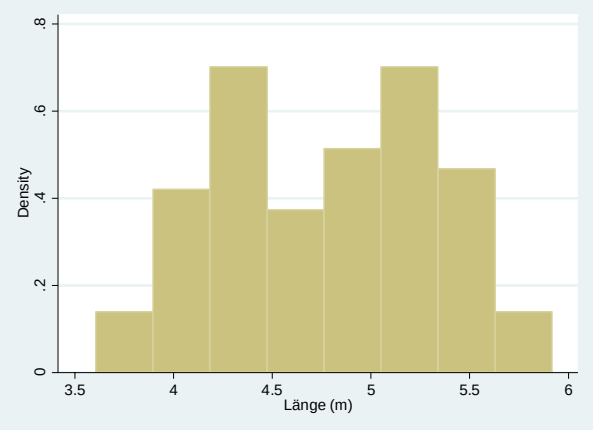
3. Metrische Zusammenhänge grafisch darstellen
   - Streudiagramme
   - Nichtlineare Verfahren

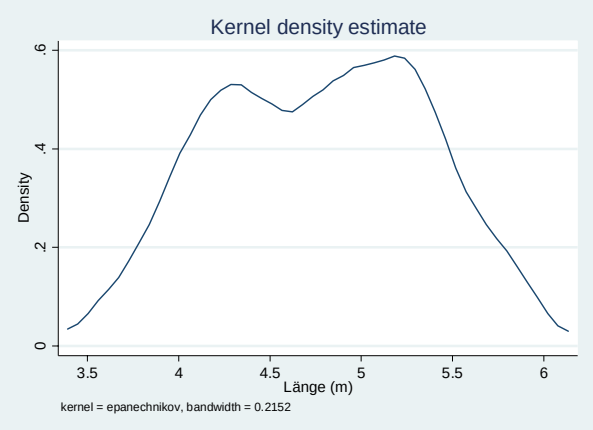
4. Zusammenhangsmaße metrischer Variablen
   - Pearsons Korrelationskoeffizient r

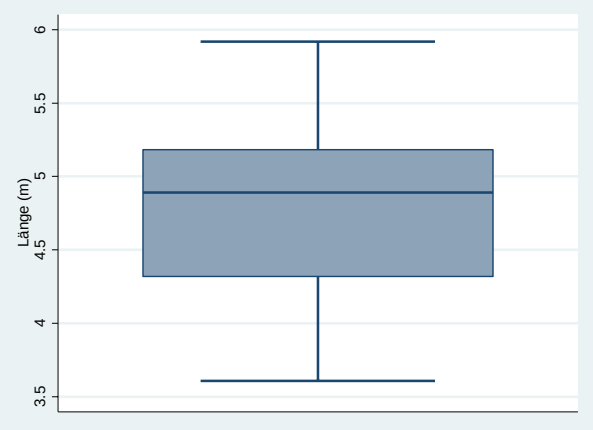
Metrische Zusammenhänge grafisch darstellen

- Bisher wurden *univariate* grafische Verfahren betrachtet

Beispiel: Variable Länge







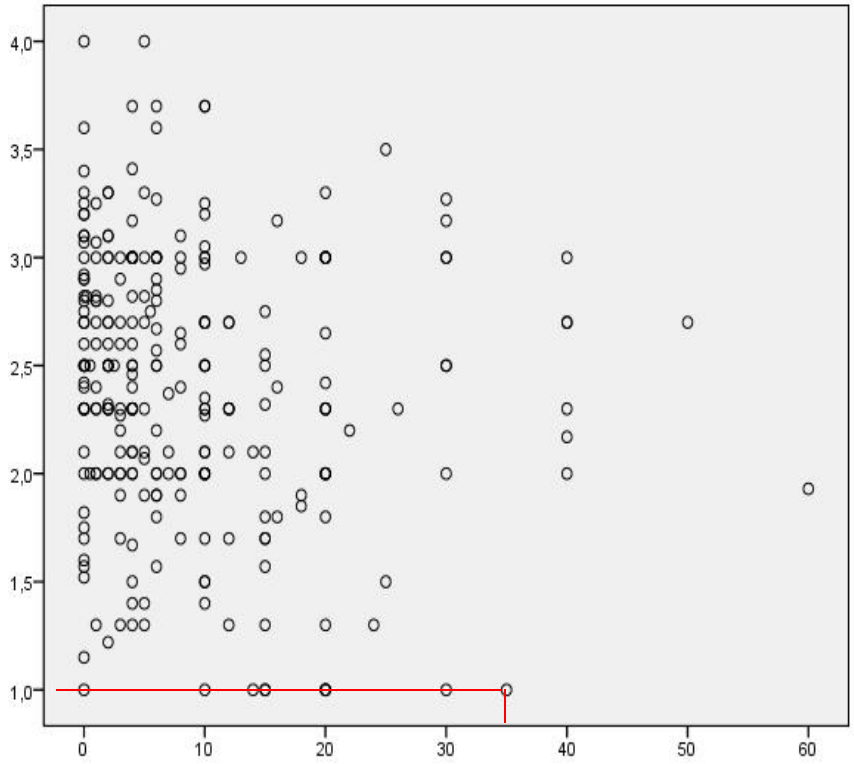
- Diese Grafen zeigen immer nur Verteilung *einer* Variable an

Metrische Zusammenhänge grafisch darstellen

- Eine Kernkompetenz empirischer Sozialforscher*innen ist die Fähigkeit relevante Ergebnisse grafisch darzustellen.

- „Gut gestaltete Graphiken sind nämlich mitunter die einfachste und zugleich wirkungsvollste Möglichkeit zur Analyse und Kommunikation statistischer Information.“ (Bauer 2010)

Streudiagramme



- Wertepaare der zwei Variablen werden als Punktwolke (Scatterplot) im Koordinaten-system abgebildet.
- Mögliches Muster der Punkte kann Aufschluss über Beziehung zwischen den Variablen geben
- Achtung: Grafische Verfahren sollten immer mit statistischen Tests kombiniert werden

Streudiagramm
Stata-Syntax

Stata Syntax für Streudiagramme:

```
graph twoway scatter abhvar unabhvar
```

zweidimensionale Grafik

scatterplot

Oder (Kurzform)

```
scatter abhvar unabhvar
```

Streudiagramme

Beispiel mit dem Auto-Datensatz:

H2: Je länger ein Auto ist, desto größer ist der Wendekreis.

Variablen:

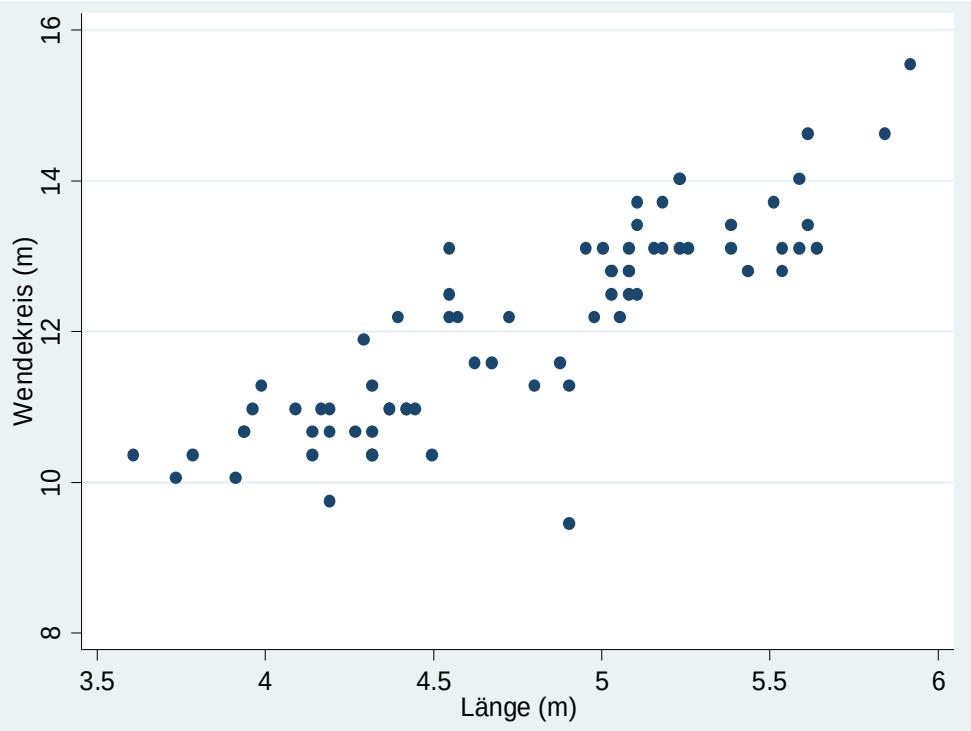
UV: Länge der Autos in Metern (laenge, bereits erstellt)

AV: Wendekreis in Metern (wend)

```
gen wend = turn*0.3048
lab var wend"Wendekreis (m)"
```

Streudiagramme

```
graph twoway scatter wend laenge
```



- Tendenziell linearer Zusammenhang erkennbar
- Längere Autos haben einen größeren Wendekreis

Nichtlineare Verfahren

- Auf der Basis von Streudiagrammen kann ein erster Eindruck über Zusammenhänge gewonnen werden
- Nichtlineare (oder nonparametrische) Verfahren unterteilen Punktewolke in einzelne Abschnitte auf der X-Achse
- In den jeweiligen Abschnitten werden lokale Werte errechnet
  - Gibt Einblick in Verlauf eines Zusammenhangs, ohne weitere Überlegungen hinsichtlich Verteilung etc. zu benötigen

Median Bands

- Unterteilung der Daten in *k* Abschnitte
- Für jeden Abschnitt wird jeweils ein lokaler Median gebildet

Stata Syntax:

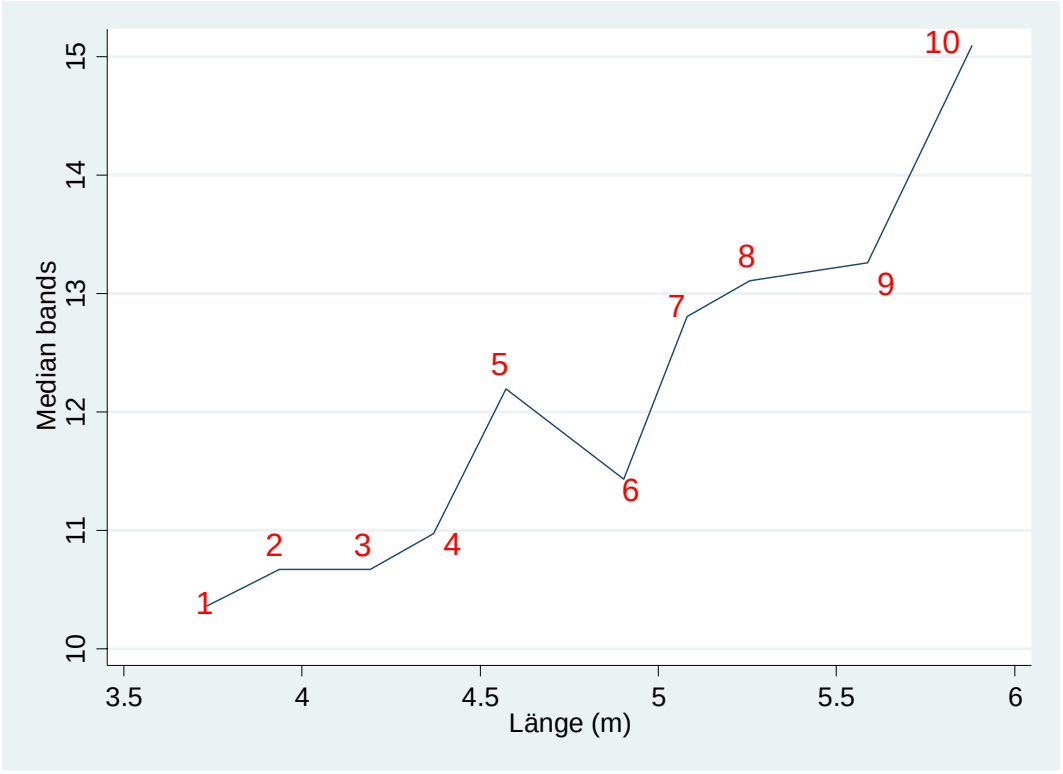
```
graph twoway mband AV UV , bands(k)
```

Anzahl der Abschnitte

Nichtlineare Verfahren

Zurück zum Beispiel

```
graph twoway mband wend laenge, bands(10)
```



- Linear-positiver Zusammenhang scheint sich zu bestätigen
- In Abschnitt 6 jedoch „Delle“ nach unten

Nichtlineare Verfahren

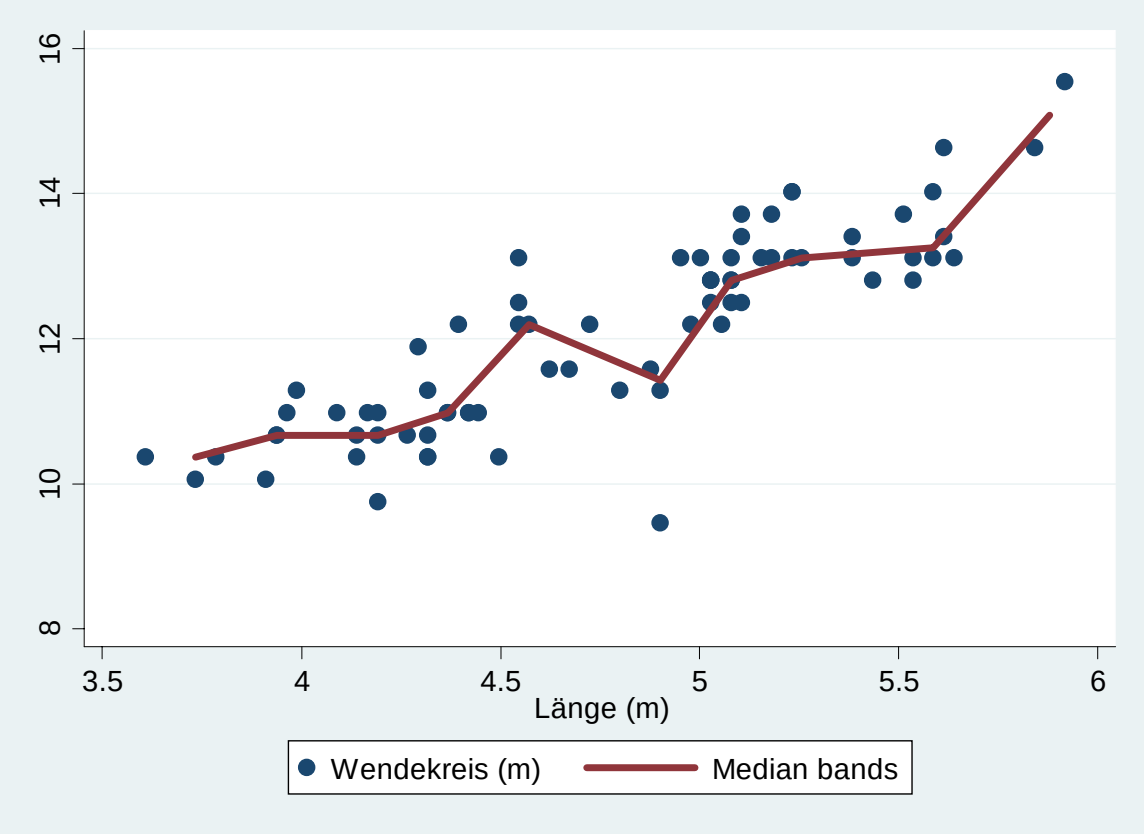
- In Stata können mehrere twoway-Grafiken miteinander kombiniert werden

- Zur Kombination verwendet man

  || (also zweimal den Operator für „oder“)

Nichtlineare Verfahren

```
graph twoway scatter wend laenge || mband wend laenge ,bands(10)
```



Programm

1. Wiederholung Chi² Unabhängigkeitstest

2. Zusammenhang bei kategorialen Variablen
   - Cramers V

3. Metrische Zusammenhänge grafisch darstellen
   - Streudiagramme
   - Nichtlineare Verfahren

4. Zusammenhangsmaße metrischer Variablen
   - Pearsons Korrelationskoeffizient r

Zusammenhangsmaße: Kovarianz und Korrelationskoeffizient

Fragestellung:

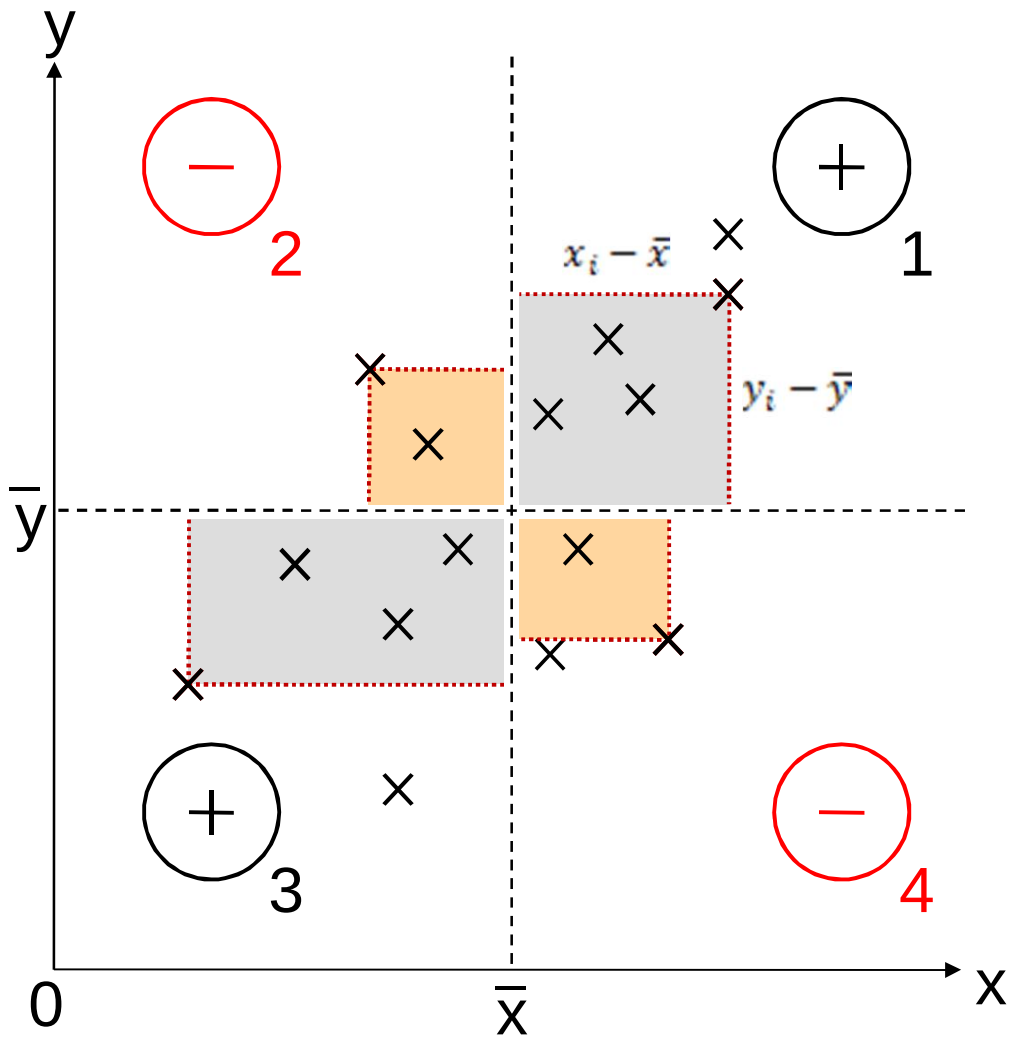
Besteht ein linearer Zusammenhang zwischen zwei metrischen Variablen?

Wie stark ist dieser Zusammenhang?

Anforderungen von Kovarianz und Korrelationskoeffizient nach Pearson:

- Beide Variablen müssen metrisch skaliert (Intervall-, Verhältnis- oder Absolutskala) sein
- Ein linearer Zusammenhang musst unterstellt werden können

Zusammenhangsmaße für metrische Variablen: Kovarianz



Berechnung der Kovarianz:

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

$(x_i - \bar{x}) * (y_i - \bar{y})$

- ist *positiv* im 1. und 3. Quadranten
- ist *negativ* im 2. und 4. Quadranten

Zusammenhangsmaße für metrische Variablen: Korrelation

Formel:

$$r = \frac{cov(x,y)}{\sqrt{var(y)} \cdot \sqrt{var(x)}}$$

Eigenschaften:

Wertebereich von [-1,+1]

Interpretation

r = -1	Perfekter, negativ-linearer Zusammenhang
↑	*Je größer X, desto kleiner Y*
r = 0	Kein linearer Zusammenhang.
↓	*Je größer X, desto größer Y*
r = 1	Perfekter, positiv-linearer Zusammenhang.

Zusammenhangsmaße für metrische Variablen: Korrelation

Vorteile

- Symmetrisches Maß, d.h. abhängige und unabhängig Variablen können vertauscht werden.
- Das Korrelationsmaß informiert darüber, wie nah die Datenpunkte um eine lineare Gerade streuen, welche an die Beobachtungen angepasst ist.
- Die Richtung des Zusammenhangs lässt sich interpretieren.

Nachteile

- Das Korrelationsmaß ist ausreißeranfällig, d.h. sehr große Werte haben ein starken Einfluss auf die Maßzahl
- Das Maß informiert nicht darüber, wie stark bei Veränderung der einen Variable die Veränderung der anderen Variable ist (gibt keinen Aufschluss über Kausalität).

Zusammenhangsmaß Korrelation
Stata-Syntax

Stata Syntax für Korrelationen:

correlate *abhvar unabvar*

Zusammenhangsmaß Korrelation
Stata-Beispiel

Zurück zum Beispiel:

```
correlate wend laenge
```

	wend	laenge
wend	1.0000	
laenge	0.8643	1.0000

- Es gibt eine stark positive Korrelation zwischen der Länge eines Autos und dem Wendekreis

Aber: Ist dieser Zusammenhang auch überzufällig (signifikant)?

Zusammenhangsmaß Korrelation
Stata-Beispiel

Für detailliertere Korrelationstabellen verwenden wir den Befehle pairwise-correlation:

pwcorr *abhvar unabvar , sig*

pwcorr *wend laenge , sig*

	wend	laenge
wend	1.0000	
laenge	0.8643	1.0000
	0.0000	

„sig“ zeigt das Signifikanzniveau jeder Variable an

p < 0,0000:

→ Zusammenhang ist höchst signifikant

→ Längere Autos haben einen signifikant größeren Wendekreis

Übersicht der Zusammenhangsmaße

	Griechischer Buchstabe	Skalenniveau	Tabellen größe	Wertebe-reich	Sym-metrisch	PRE-Maß	Stärke/Rich-tung
Chi²	X^2	nominal					-
Phi	Φ	nominal	2*2 k*m	[-1, 1] [0, 1]	Ja	Nein	Stärke
Cramers V	V	nominal	k*m	[0, 1]	Ja	Nein	Stärke
Lambda	λ	nominal	k*m	[0, 1]	Nein	Ja	Stärke
Gamma	γ	ordinal	k*m	[-1, 1]	Ja	Ja	Stärke Richtung
Kendall's Tau b	τ_b	ordinal	k*m	[-1, 1]	Ja	Nein	Stärke Richtung
Kendall's Tau c	τ_c	ordinal	k*m	[-1, 1]	Ja	Nein	Stärke Richtung
eta und eta²	η η^2	UV: nominal/ordinal AV: metrisch		[0, 1]	Nein	eta²: ja eta: nein	Stärke Richtung
Korrelation	r	metrisch		r: [-1;1]	Ja	r: nein	Stärke Richtung

Befehlsübersicht

Cramers V:	**tab**ulate AV UV , chi2 **V**
Scatterplot:	graph twoway **scatter** AV UV if X==
Median Bands:	graph twoway **mband** AV UV, bands(k)
komb. Grafen:	graph twoway plot1 || plot2
Korrelation:	**corr**elate AV UV
Korr. detailiert:	**pwcorr** AV UV , sig

Übungsaufgaben

Sie vermuten, dass der Zusammenhang zwischen der Länge von Autos und dem Wendekreis sich je nach der Herkunft der Autos unterscheidet.

Überprüfen Sie diese Vermutung.

1) Erstellen Sie vier Variablen um folgende Unterscheidung treffen zu können:

	Herkunft	
	Domestic	Foreign
Wendekreis	*wend_dom*	*wend_for*
Länge	*laenge_dom*	*laenge_for*

2) Unterziehen Sie beide Herkunftsgruppen einer grafischen Analyse

3) Stützen Sie die grafische Analyse mittels einer geeigneten statistischen Maßzahl

4) Interpretieren Sie die Ergebnisse inhaltlich