Zusammenhangsmaße II

Statistik I

Sommersemester 2009

$$\chi^2 = ?!?$$

Übung von Simone Reutzel

- *Heute* im HS1, altes ReWi-Haus

Zum Nachlesen

- Agresti/Finlay: Kapitel 8.5, 9.4
- Gehring/Weins: Kapitel 7.2-7.4
- Schumann: Kapitel 8.2

Was ist ein Zusammenhang?

- Zusammenhang = Verteilungen zweier Variablen nicht unabhängig voneinander
- Kategorien treten häufiger gemeinsam auf als bei rein zufälliger Verteilung zu erwarten
- Zusammenhangs*maße* quantifizieren Stärke und ggf. Richtung des Zusammenhangs
- Unterschiedliche Maße für unterschiedliche Skalenniveaus
- PRE-Maß λ betrachtet Verbesserung gegenüber naiver Vorhersage und ist asymmetrisch

Warum unterscheiden sich λ und V?

- Beide gehen von Kreuztabelle (bivariate Verteilung) aus
- V basiert auf χ^2 $\rightarrow$
  - Weichen interne/Randverteilungen voneinander ab $\rightarrow$
  - Alle Zellen werden berücksichtigt
  - Vergleichbar mit arithmetischem Mittel: gesamte Information wird genutzt
- λ betrachtet Modus der Randverteilung (abhängige Variable)
  - Vergleich mit Modus innerhalb von Subgruppen, definiert durch Ausprägungen unabhängige Variable (z. B. Ost/West)
  - Anschaulicher, aber weniger Information genutzt $\rightarrow$
- Wenn Modus in Randverteilung und in Gruppen identisch $\rightarrow$ $\lambda = 0$
- Höhe von λ hängt ab von Häufigkeit Modus vs. Häufigkeit aller anderen Kategorien zusammen innerhalb von Subgruppen

$\lambda = 0.57$ vs. $\lambda = 0$

	Kanzlerpräferenz		
Wahlabsicht	Merkel	Steinmeier	Σ
Union	335	15	350
SPD	25	320	345
Andere	84	102	186
Σ	444	437	881

$\lambda = 0.57$ vs. $\lambda = 0$

	Kanzlerpräferenz		
Wahlabsicht	Merkel	Steinmeier	Σ
Union	335	15	350
SPD	25	320	345
Andere	84	102	186
Σ	444	437	881

- Modus in den Subgruppen unterscheidet sich
- Innerhalb der Subgruppen Modus sehr häufig, alle anderen eher selten

$\lambda = 0.57$ vs. $\lambda = 0$

	Kanzlerpräferenz		
Wahlabsicht	Merkel	Steinmeier	Σ
Union	335	15	350
SPD	25	320	345
Andere	84	102	186
Σ	444	437	881

- Modus in den Subgruppen unterscheidet sich
- Innerhalb der Subgruppen Modus sehr häufig, alle anderen eher selten
- In der Randverteilung auch andere Antworten („Fehler“) häufig
- Große „Fehlerreduktion“, hohes λ

$\lambda = 0.57$ vs. $\lambda = 0$

	Kanzlerpräferenz		
Wahlabsicht	Merkel	Steinmeier	Σ
Union	335	15	350
SPD	25	320	345
Andere	84	102	186
Σ	444	437	881

- Modus in den Subgruppen unterscheidet sich
- Innerhalb der Subgruppen Modus sehr häufig, alle anderen eher selten
- In der Randverteilung auch andere Antworten („Fehler") häufig
- Große „Fehlerreduktion", hohes λ
- Einfluß der Kategorienbildung

$\lambda = 0.57$ vs. $\lambda = 0$

	Region		
Wahlabsicht	West	Ost	Σ
PDS	4	116	120
Andere	1572	606	2178
Σ	1576	772	2298

$\lambda = 0.57$ vs. $\lambda = 0$

	Region		
Wahlabsicht	West	Ost	Σ
PDS	4	116	120
Andere	1572	606	2178
Σ	1576	772	2298

- Modus in Randverteilung und beiden Subgruppen identisch

$\lambda = 0.57$ vs. $\lambda = 0$

	Region		
Wahlabsicht	West	Ost	Σ
PDS	4	116	120
Andere	1572	606	2178
Σ	1576	772	2298

- Modus in Randverteilung und beiden Subgruppen identisch
- Keine Fehlerreduktion, $\lambda = 0$

$\lambda = 0.57$ vs. $\lambda = 0$

	Region		
Wahlabsicht	West	Ost	Σ
PDS	4	116	120
Andere	1572	606	2178
Σ	1576	772	2298

- Modus in Randverteilung und beiden Subgruppen identisch
- Keine Fehlerreduktion, $\lambda = 0$
- Aufteilung „Andere“ $\rightarrow \lambda > 0$ möglich

$\lambda = 0.57$ vs. $\lambda = 0$

	Region		
Wahlabsicht	West	Ost	Σ
PDS	4	116	120
Andere	1572	606	2178
Σ	1576	772	2298

- Modus in Randverteilung und beiden Subgruppen identisch
- Keine Fehlerreduktion, $\lambda = 0$
- Aufteilung „Andere“ $\rightarrow \lambda > 0$ möglich
- Aber nur wenn anderer Modus innerhalb von Subgruppe (einfache Mehrheit für PDS)

Warum sind Maße für ordinale Daten wichtig?

- Viele politikwissenschaftlich interessante Variablen ordinal
  - Umfragedaten: „stimme voll zu", „stimme zu", „lehne ab" „lehne voll ab"
  - Freedom house ranking: „free", „partly free", „not free"
  - Politische Parteien: „links", „Mitte", „rechts"
  - Effektivität internationaler Umweltschutzregime: „none", „some" ...
- Berechnung nominaler Maße (V, λ) möglich
- Aber: Information über Ordnung der Kategorien wird ignoriert

Was ist ein ordinaler Zusammenhang?

- Zwei ordinale Variablen $\longrightarrow$ Richtung
  - Mehr x, mehr y; weniger x, weniger y $\longrightarrow$ positiver Zusammenhang
  - Mehr x, weniger y; weniger x, mehr y $\longrightarrow$ negativer Zusammenhang

Was ist ein ordinaler Zusammenhang?

- Zwei ordinale Variablen $\longrightarrow$ Richtung
  - Mehr x, mehr y; weniger x, weniger y $\longrightarrow$ positiver Zusammenhang
  - Mehr x, weniger y; weniger x, mehr y $\longrightarrow$ negativer Zusammenhang
- Wie mißt man das?

Was ist ein ordinaler Zusammenhang?

- Zwei ordinale Variablen $\longrightarrow$ Richtung
  - Mehr x, mehr y; weniger x, weniger y $\longrightarrow$ positiver Zusammenhang
  - Mehr x, weniger y; weniger x, mehr y $\longrightarrow$ negativer Zusammenhang
- Wie mißt man das?
- Vergleich von Paaren von Beobachtungen

Was ist ein ordinaler Zusammenhang?

- Zwei ordinale Variablen $\longrightarrow$ Richtung
  - Mehr x, mehr y; weniger x, weniger y $\longrightarrow$ positiver Zusammenhang
  - Mehr x, weniger y; weniger x, mehr y $\longrightarrow$ negativer Zusammenhang
- Wie mißt man das?
- Vergleich von Paaren von Beobachtungen

Land A
hohes Bildungsniveau
hohes Wirtschaftswachstum

Land B
niedriges Bildungsniveau
niedriges Wirtschaftswachstum

+

Zusammenhangsmaß γ

- Systematisiert Paarvergleich
- Symmetrisch
- Kann ebenfalls als PRE-Maß interpretiert werden
- Vorzeichen

Wie wird γ berechnet?

- Paare von Objekten bilden

Wie wird γ berechnet?

- Paare von Objekten bilden
  - *Konkordantes* Paar $A \leftrightarrows B$: B hat mehr von x (z. B. Bildung) und mehr von y (z. B. politisches Interesse) als A

Wie wird γ berechnet?

- Paare von Objekten bilden
  - *Konkordantes* Paar $A \leftrightarrows B$: B hat mehr von x (z. B. Bildung) und mehr von y (z. B. politisches Interesse) als A
  - *Diskonkordantes* Paar $A \leftrightarrows B$: B hat mehr von x (z. B. Bildung) als A, aber *weniger* von y (z. B. politisches Interesse) als A

Wie wird γ berechnet?

- Paare von Objekten bilden
  - *Konkordantes* Paar $A \leftrightarrows B$: B hat mehr von x (z. B. Bildung) und mehr von y (z. B. politisches Interesse) als A
  - *Diskonkordantes* Paar $A \leftrightarrows B$: B hat mehr von x (z. B. Bildung) als A, aber *weniger* von y (z. B. politisches Interesse) als A
- γ: Verhältnis konkordante – diskonkordante Paare

Wie wird γ berechnet?

- Paare von Objekten bilden
  - *Konkordantes* Paar $A \leftrightarrows B$: B hat mehr von x (z. B. Bildung) und mehr von y (z. B. politisches Interesse) als A
  - *Diskonkordantes* Paar $A \leftrightarrows B$: B hat mehr von x (z. B. Bildung) als A, aber *weniger* von y (z. B. politisches Interesse) als A
- γ: Verhältnis konkordante – diskonkordante Paare
  - Konkordante Paare überwiegen: positiver Zusammenhang
  - Diskonkordante Paare überwiegen: negativer Zusammenhang

Wie wird γ berechnet?

- Paare von Objekten bilden
  - *Konkordantes* Paar $A \leftrightarrows B$: B hat mehr von x (z. B. Bildung) und mehr von y (z. B. politisches Interesse) als A
  - *Diskonkordantes* Paar $A \leftrightarrows B$: B hat mehr von x (z. B. Bildung) als A, aber *weniger* von y (z. B. politisches Interesse) als A
- γ: Verhältnis konkordante – diskonkordante Paare
  - Konkordante Paare überwiegen: positiver Zusammenhang
  - Diskonkordante Paare überwiegen: negativer Zusammenhang
- Paare mit identischen Werten für eine oder beide Variablen: „ties“

Wie wird γ berechnet?

- Paare von Objekten bilden
  - *Konkordantes* Paar $A \leftrightarrows B$: B hat mehr von x (z. B. Bildung) und mehr von y (z. B. politisches Interesse) als A
  - *Diskonkordantes* Paar $A \leftrightarrows B$: B hat mehr von x (z. B. Bildung) als A, aber *weniger* von y (z. B. politisches Interesse) als A
- γ: Verhältnis konkordante – diskonkordante Paare
  - Konkordante Paare überwiegen: positiver Zusammenhang
  - Diskonkordante Paare überwiegen: negativer Zusammenhang
- Paare mit identischen Werten für eine oder beide Variablen: „ties“
- Werden bei γ ignoriert

Wie berechnet man die Anzahl der Paare?

- Kreuztabelle aufstellen
- Anordnung der Kategorien: niedrigste oben/links, höchste unten/rechts
  - Bildung konkordanter Paare möglich mit allen Objekten rechts und unterhalb der eigenen Zelle
  - Bildung diskonkordanter Paare möglich mit Objekten in den Zellen links und unterhalb der eigenen Zelle
- Zahl der möglichen Paare für zwei Zellen: Produkt der Häufigkeiten beider Zellen
- Z. B. 10 Personen in der ersten Zelle, 5 in der zweiten Zelle, 10×5 verschiedene Paare möglich
- Auf diese Weise Summe aller konkordanten und dann aller diskonkordanten Paare bilden

Beispiel γ: Bildung und Nationalstolz

- Was erwarten Sie?

Beispiel γ: Bildung und Nationalstolz

- Was erwarten Sie?

	niedrig	mittel	hoch
gar nicht stolz	115	128	181
nicht sehr stolz	267	267	209
ziemlich stolz	731	463	260
sehr stolz	416	138	58

- Wo sind die konkordanten Paare?

Beispiel γ: Bildung und Nationalstolz

- Was erwarten Sie?

	niedrig	mittel	hoch
gar nicht stolz	115	128	181
nicht sehr stolz	267	267	209
ziemlich stolz	731	463	260
sehr stolz	416	138	58

- Wo sind die konkordanten Paare?
- Z. B. hier . . .

Beispiel γ: Bildung und Nationalstolz

- Was erwarten Sie?

	niedrig	mittel	hoch
gar nicht stolz	115	128	181
nicht sehr stolz	267	267	209
ziemlich stolz	731	463	260
sehr stolz	416	138	58

- Wo sind die konkordanten Paare?
- Z. B. hier ... oder hier ...

Beispiel γ: Bildung und Nationalstolz

- Was erwarten Sie?

	niedrig	mittel	hoch
gar nicht stolz	115	128	181
nicht sehr stolz	267	267	209
ziemlich stolz	731	463	260
sehr stolz	416	138	58

- Wo sind die konkordanten Paare?
- Z. B. hier ... oder hier ... oder hier

Beispiel γ: Bildung und Nationalstolz

- Was erwarten Sie?

	niedrig	mittel	hoch
gar nicht stolz	115	128	181
nicht sehr stolz	267	267	209
ziemlich stolz	731	463	260
sehr stolz	416	138	58

- Wo sind die konkordanten Paare?
- Z. B. hier ... oder hier ... oder hier
- Diskonkordante Paare ...

Beispiel γ: Bildung und Nationalstolz

- Was erwarten Sie?

	niedrig	mittel	hoch
gar nicht stolz	115	128	181
nicht sehr stolz	267	267	209
ziemlich stolz	731	463	260
sehr stolz	416	138	58

- Wo sind die konkordanten Paare?
- Z. B. hier . . . oder hier . . . oder hier
- Diskonkordante Paare . . . oder hier

Beispiel γ: Bildung und Nationalstolz

- Was erwarten Sie?

	niedrig	mittel	hoch
gar nicht stolz	115	128	181
nicht sehr stolz	267	267	209
ziemlich stolz	731	463	260
sehr stolz	416	138	58

- Konkordant:

Beispiel γ: Bildung und Nationalstolz

- Was erwarten Sie?

	niedrig	mittel	hoch
gar nicht stolz	115	128	181
nicht sehr stolz	267	267	209
ziemlich stolz	731	463	260
sehr stolz	416	138	58

- Konkordant:
  - $115 \times (267+209+463+260+138+58) = 160\,425$
  - $128 \times (209+260+58) = 67\,456$
  - $267 \times (463+260+138+58) = 245\,373$
  - $267 \times (260+58) = 84\,906$
  - $731 \times (138+58) = 143\,276$
  - $463 \times 58 = 26\,854$

Beispiel γ: Bildung und Nationalstolz

- Was erwarten Sie?

	niedrig	mittel	hoch
gar nicht stolz	115	128	181
nicht sehr stolz	267	267	209
ziemlich stolz	731	463	260
sehr stolz	416	138	58

- Konkordant: 728 290
  - $115 \times (267{+}209{+}463{+}260{+}138{+}58) = 160\,425$
  - $128 \times (209{+}260{+}58) = 67\,456$
  - $267 \times (463{+}260{+}138{+}58) = 245\,373$
  - $267 \times (260{+}58) = 84\,906$
  - $731 \times (138{+}58) = 143\,276$
  - $463 \times 58 = 26\,854$

Beispiel γ: Bildung und Nationalstolz

- Was erwarten Sie?

	niedrig	mittel	hoch
gar nicht stolz	115	128	181
nicht sehr stolz	267	267	209
ziemlich stolz	731	463	260
sehr stolz	416	138	58

- Diskonkordant:

Beispiel γ: Bildung und Nationalstolz

- Was erwarten Sie?

	niedrig	mittel	hoch
gar nicht stolz	115	128	181
nicht sehr stolz	267	267	209
ziemlich stolz	731	463	260
sehr stolz	416	138	58

- Diskonkordant:
  - $128 \times (267 + 731 + 416) = 180\,992$
  - $181 \times (267 + 267 + 731 + 463 + 416 + 138) = 413\,042$
  - $267 \times (731 + 416) = 306\,249$
  - $209 \times (731 + 463 + 416 + 138) = 365\,332$
  - $463 \times 416 = 192\,608$
  - $260 \times (416 + 138) = 144\,040$

Beispiel γ: Bildung und Nationalstolz

- Was erwarten Sie?

	niedrig	mittel	hoch
gar nicht stolz	115	128	181
nicht sehr stolz	267	267	209
ziemlich stolz	731	463	260
sehr stolz	416	138	58

- Diskonkordant: 1 602 263
  - $128 \times (267 + 731 + 416) = 180\,992$
  - $181 \times (267 + 267 + 731 + 463 + 416 + 138) = 413\,042$
  - $267 \times (731 + 416) = 306\,249$
  - $209 \times (731 + 463 + 416 + 138) = 365\,332$
  - $463 \times 416 = 192\,608$
  - $260 \times (416 + 138) = 144\,040$

Beispiel γ: Bildung und Nationalstolz

- Was erwarten Sie?

	niedrig	mittel	hoch
gar nicht stolz	115	128	181
nicht sehr stolz	267	267	209
ziemlich stolz	731	463	260
sehr stolz	416	138	58

$$\gamma = \frac{N_c - N_D}{N_C + N_D} = \frac{728290 - 1602263}{728290 + 1602263} = -0.375$$

Interpretation von γ

- $N_c = N_D \rightarrow \gamma = 0$

Interpretation von γ

- $N_c = N_D \rightarrow \gamma = 0$
- γ meist relativ hoch
- $\frac{75-25}{75+25} = 0.5$

Interpretation von γ

- $N_c = N_D \rightarrow \gamma = 0$
- γ meist relativ hoch
- $\frac{75-25}{75+25} = 0.5$
- Ties werden komplett ignoriert
- Potentiell viel verlorene Information
  - Irreführende Werte möglich
  - γ reagiert u. U. nicht auf Veränderungen in Tabelle

Was sind ties?

	niedrig	mittel	hoch
gar nicht stolz	115	128	181
nicht sehr stolz	267	267	209
ziemlich stolz	731	463	260
sehr stolz	416	138	58

- Paare mit identischem Wert für Spaltenvariable (x) – „tied in x“, T_x
- Paare mit identischem Wert für Zeilenvariable (y) – „tied in y“, T_y
- Paare mit identischen Werten für beide Variablen – relativ uninteressant
- Verschiedene alternative Koeffizienten berücksichtigen ties

τ_b

$$\tau_b = \frac{N_C - N_D}{\sqrt{(N_C + N_D + T_x) \times (N_C + N_D + T_Y)}}$$

τ_b

$$\tau_b = \frac{N_C - N_D}{\sqrt{(N_C + N_D + T_x) \times (N_C + N_D + T_Y)}}$$

- Identisch mit γ wenn keine ties

τ_b

$$\tau_b = \frac{N_C - N_D}{\sqrt{(N_C + N_D + T_x) \times (N_C + N_D + T_Y)}}$$

- Identisch mit γ wenn keine ties
- Ansonsten kleiner als γ weil Nenner größer

τ_b

$$\tau_b = \frac{N_C - N_D}{\sqrt{(N_C + N_D + T_x) \times (N_C + N_D + T_Y)}}$$

- Identisch mit γ wenn keine ties
- Ansonsten kleiner als γ weil Nenner größer
- Somers' D – asymmetrische Variante von τ_b

Somers' D

$$D_{yx} = \frac{N_C - N_D}{N_C + N_D + T_Y} \quad \text{mit } y \text{ als abhängiger Variable}$$

- Identisch mit γ wenn keine ties
- Ansonsten kleiner als γ weil Nenner größer
- Somers' D – asymmetrische Variante von τ_b

Somers' D

$$D_{yx} = \frac{N_C - N_D}{N_C + N_D + T_Y} \quad \text{mit } y \text{ als abhängiger Variable}$$

- Identisch mit γ wenn keine ties
- Ansonsten kleiner als γ weil Nenner größer
- Somers' D – asymmetrische Variante von τ_b
- Zahlreiche andere Maße

Warum nominale/intervallskalierte Variable?

- Viele Variablen (näherungsweise) intervallskaliert
- Z. B. Einkommen, Testwerte, LRS ...
- Vergleich über Gruppen (Geschlecht, Länder etc.)
- Zusammenhangsmaß: η bzw. η^2
- PRE-Maß/asymmetrisch
- Kann Gruppenzugehörigkeit Vorhersage der abhängigen Variablen verbessern?

Wie wird η berechnet?

- Beste Prognose für intervallskalierte Variable?

Wie wird η berechnet?

- Beste Prognose für intervallskalierte Variable?
- Arithmetisches Mittel
  - Summe einfache Abweichungen $= 0$
  - Summe quadrierte Abweichungen minimal

Wie wird η berechnet?

- Beste Prognose für intervallskalierte Variable?
- Arithmetisches Mittel
  - Summe einfache Abweichungen $= 0$
  - Summe quadrierte Abweichungen minimal
- Maß für Vorhersagefehler: SAQ

Wie wird η berechnet?

- Beste Prognose für intervallskalierte Variable?
- Arithmetisches Mittel
  - Summe einfache Abweichungen $= 0$
  - Summe quadrierte Abweichungen minimal
- Maß für Vorhersagefehler: SAQ
- Reduktion SAQ bei Kenntnis der Gruppenzugehörigkeit?
  - Vorhersage globaler Mittelwert
  - Vs. Vorhersage Gruppenmittelwerte

Wie wird η berechnet?

- SAQ_{gesamt}: Summe der quadrierte Abweichungen vom *gemeinsamen* Durchschnitt
- $SAQ_{Kategorien}$: Summe der Summen der quadrierten Abweichungen vom jeweiligen Gruppendurchschnitt

η^2

$$\eta^2 = \frac{SAQ_{gesamt} - SAQ_{Kategorien}}{SAQ_{gesamt}}$$

Ein Beispiel: FN 2004 (drei Regionen)

Region	x	$(x - \bar{x}_{kat})$	$(x - \bar{x}_{kat})^2$	$(x - \bar{x}_{ges})$	$(x - \bar{x}_{ges})^2$
Basse-Normandie	12.6	−1.9	3.4	−2.9	8.3
$\bar{x}_{B-N} = 14.5$	14.1	−0.4	0.1	−1.4	2.0
	16.7	2.2	5.0	1.2	1.5
Σ		0.0	8.6		
Limousin	8.3	−0.9	0.8	−7.2	52.0
$\bar{x}_L = 9.2$	9.2	0.0	0.0	−6.3	40.3
	10.1	0.9	0.8	−5.4	29.6
Σ		0.0	1.6		
Picardie	24.1	1.2	1.5	8.6	73.5
$\bar{x}_P = 22.9$	24.9	2.1	4.2	9.4	88.6
	19.6	−3.3	10.8	4.1	16.6
Σ		0.0	16.5		
$\Sigma\Sigma$			26.7	0.0	312.4

$$\bar{x}_{ges} = 15.5$$

Ein Beispiel: FN 2004 (drei Regionen)

Region	x	$(x - \bar{x}_{kat})$	$(x - \bar{x}_{kat})^2$	$(x - \bar{x}_{ges})$	$(x - \bar{x}_{ges})^2$
Basse-Normandie	12.6	−1.9	3.4	−2.9	8.3
$\bar{x}_{B-N} = 14.5$	14.1	−0.4	0.1	−1.4	2.0
	16.7	2.2	5.0	1.2	1.5
Σ		0.0	8.6		
Limousin	8.3	−0.9	0.8	−7.2	52.0
$\bar{x}_{L} = 9.2$	9.2	0.0	0.0	−6.3	40.3
	10.1	0.9	0.8	−5.4	29.6
Σ		0.0	1.6		
Picardie	24.1	1.2	1.5	8.6	73.5
$\bar{x}_{P} = 22.9$	24.9	2.1	4.2	9.4	88.6
	19.6	−3.3	10.8	4.1	16.6
Σ		0.0	16.5		
$\Sigma\Sigma$			26.7	0.0	312.4

$\bar{x}_{ges} = 15.5$

$\eta^2 = \frac{312.4 - 26.7}{312.4} = 0.91$

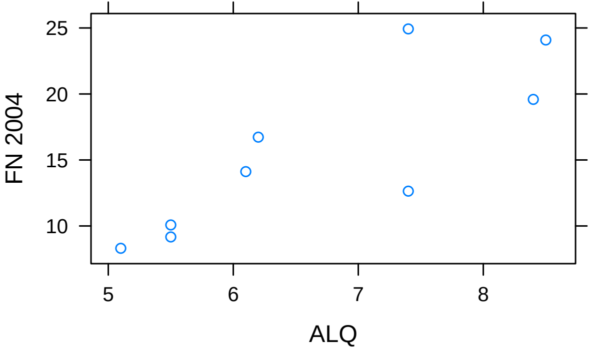
Warum zwei intervallskalierte Variablen?

- *Wenn* Intervallskalierung plausible Annahme, eine der häufigsten Konstellationen
- LRS $\leftrightarrows$ Konservatismus-Skala
- Proportionalität des Wahlsystems $\leftrightarrows$ (effektive) Zahl der Parteien
- (Relative) Größe des tertiären Sektors $\leftrightarrows$ Anteil der Postmaterialisten
- Arbeitslosenquote $\leftrightarrows$ Stimmenanteil Extreme Rechte

Was ist die Logik von Pearson's r (Korrelationskoeffizient)?

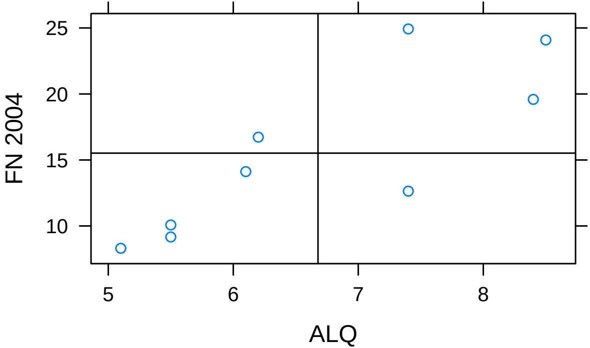
- Beste unkonditionale Vorhersage: arithmetisches Mittel; SAQ = Vorhersagefehler
- *Zwei* intervallskalierte Variablen $\rightarrow$ zwei Mittelwerte, zwei SAQ
- Visualisierung durch Streudiagramm
  - Positiver Zusammenhang: überdurchschnittliche Werte von x, überdurchschnittliche Werte von y und umgekehrt
  - Negativer Zusammenhang: überdurchschnittliche Werte von x, *unter*durchschnittliche Werte von y und umgekehrt
- Starke Zusammenhänge mit bloßem Auge erkennbar
- r: quantifiziert Stärke, symmetrisch, PRE-Maß

ALQ und FN 2004 in 9 Départements



$\bar{x} = 6.7; \bar{y} = 15.5$

ALQ und FN 2004 in 9 Départements

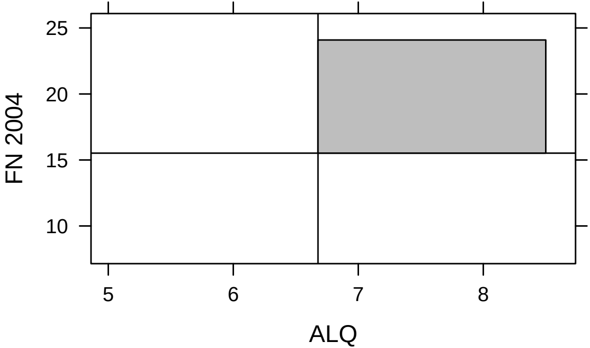


$\bar{x} = 6.7; \bar{y} = 15.5$

Wie läßt sich die *gemeinsame* Streuung zweier Variablen erfassen?

- Ein Fall ist nicht mit dem Mittelwert identisch
- Für eine Variable: quadrierte Abweichung vom Mittelwert $\bar{x}$
- Für zwei Variablen: Produkt der Abweichungen von *beiden* Mittelwerten $\bar{x}, \bar{y}$ – „Abweichungsprodukt“
  - Positives Abweichungsprodukt: bezüglich beider Variablen über- oder unterdurchschnittlich
  - Negatives Abweichungsprodukt: überdurchschnittlich bei einer, unterdurchschnittlich bei anderer Variablen
  - Vgl. mit γ

Abweichungsprodukt für einen Fall



$$AP = (x_i - \bar{x}) \times (y_i - \bar{y}) = (8.5 - 6.7) \times (24.1 - 15.5) = 15.6$$

Abweichungsprodukte für alle Fälle: Summe Abweichungsprodukte

No.	x_i	$x_i - \bar{x}$	y_i	$y_i - \bar{y}$	$(x_i - \bar{x}) \times (y_i - \bar{y})$
2	8.5	1.8	24.1	8.6	15.6
14	7.4	0.7	12.6	−2.9	−2.1
19	5.1	−1.6	8.3	−7.2	11.4
23	5.5	−1.2	9.2	−6.3	7.5
50	6.1	−0.6	14.1	−1.4	0.8
60	7.4	0.7	24.9	9.4	6.8
61	6.2	−0.5	16.7	1.2	−0.6
80	8.4	1.7	19.6	4.1	7.0
87	5.5	−1.2	10.1	−5.4	6.4
		0.0		0.0	52.8

Kovarianz und Korrelation

- Kovarianz $= \frac{SAP}{n} = \frac{52.8}{9} = 5.9 = \text{cov}(x, y)$
  - Wertebereich $-\infty; +\infty$
  - Vgl. χ^2
  - Abhängig von Stärke und Skalierung
- Kovarianz durch Produkt beider Standardabweichungen teilen $\rightarrow$ Korrelationskoeffizient r

Korrelationskoeffizient

$$r = \frac{\text{cov}(x, y)}{s_x \times s_y} = \frac{\frac{SAP}{n}}{\sqrt{\frac{SAQ_x}{n}} \times \sqrt{\frac{SAQ_y}{n}}} = \frac{SAP}{\sqrt{SAQ_x \times SAQ_y}}$$

Kovarianz und Korrelation II

Korrelationskoeffizient

$$\frac{SAP}{\sqrt{SAQ_x \times SAQ_y}} = \frac{52.8}{\sqrt{13.1 \times 312.2}} = \frac{52.8}{64.1} = 0.82$$

- Bringt Kovarianz auf standardisierten Wertebereich
- Wertebereich [-1;+1]
- 0 = kein Zusammenhang
- r^2 symmetrisches PRE-Maß
- Wie stark reduziert Kenntnis von x Vorhersagefehler von y (SAQ) und umgekehrt?
- $\eta^2 \leftrightarrows r^2$, unterschiedliche Skalenniveaus/Berechnungsvorschriften

Zusammenfassung

- Zusammenhang – gemeinsame Verteilung zweier Variablen
- Vielzahl von Zusammenhangsmaßen
- Skalenniveau beachten