Vorlesung: Statistik I

Prof. Dr. Simone Abendschön

11. Vorlesung am 18.01.24: Abschluss bivariate Zusammenhänge, Spearman‘s Rho und Eta-Koeffizient

Plan für heute

- **Gemeinsame Planung Sitzung 8.2.**

Abschluss bivariate Analyseverfahren

- **Zusammenhangsmaß zwischen ordinalskalierten Merkmalen: Spearman‘s Rho**
- **Eta-Koeffizient als Beispiel für ein sog. „PRE-Maß“**

Wie soll die die letzte Sitzung im Semester inhaltlich gestaltet werden?

Folgende Möglichkeiten:

- Normal weiter im Stoff, Wiederholung Statistik I im SoSe einplanen
- Wiederholung durch Vorbereitung Ihrer erstellter „Klausurfragen"
- Wiederholung wichtigste Inhalte durch mich, evtl. Quizformat

Abstimmung über BigBlueButton am Ende der heutigen Sitzung

Bivariate Zusammenhänge…

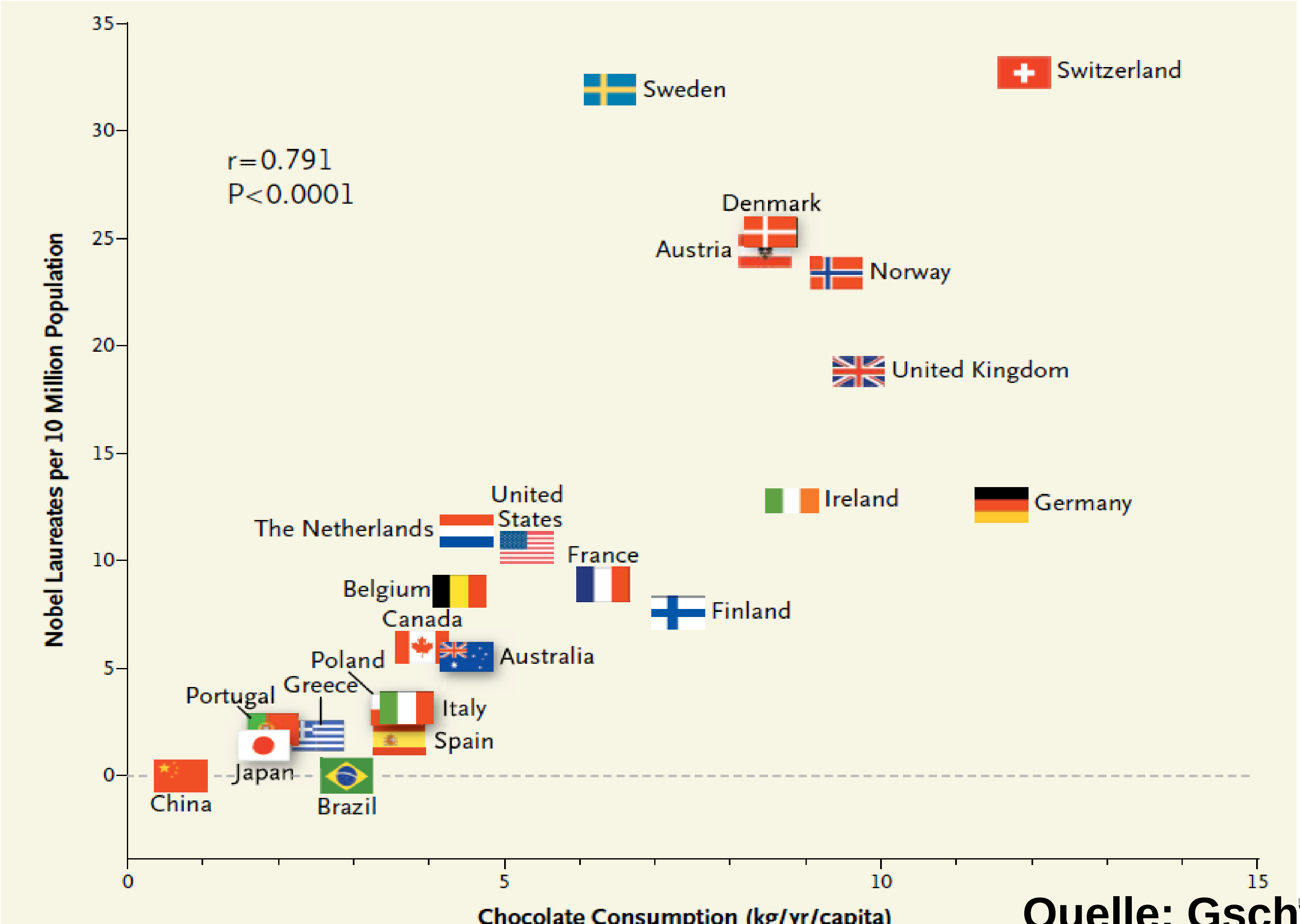
…berechnen

Messniveau	nominal	ordinal	metrisch
nominal	**Chi-Quadrat, Cramer‘s V** Lambda C	**Cramer‘s V** Lambda C	**Eta-Koeffizient** Mittelwertvergleich (t-test)
ordinal	**Cramer‘s V** Lambda C	**Spearman‘s Rho;** (Kendalls Tau B gamma	**Spearman‘s Rho;** (Kendalls Tau B) gamma
metrisch	Eta-Koeffizient Mittelwertvergleich (t-test)	**Spearman‘s Rho;** (Kendalls Tau B) gamma	**Pearson‘s r**

Wichtig

- **Korrelationsberechnung braucht inhaltliche Theorie bzw. Plausibilität!**
- **Korrelation ≠ Kausalität!**

Theorie? Plausibilität?



Quelle: Gschwend 2013

Plan für heute

- **Gemeinsame Planung Sitzung 8.2.**

Abschluss bivariate Analyseverfahren

- **Zusammenhangsmaß zwischen ordinalskalierten Merkmalen: Spearman‘s Rho**
- **Eta-Koeffizient als Beispiel für ein sog. „PRE-Maß“**

Spearman‘s Rho

- Auch für zwei **ordinalskalierte Merkmale** lassen sich (positive und negative) Zusammenhänge berechnen

- Es empfehlen sich zunächst auch hier **Kreuztabellen** zur Anschauung der kombinierten Häufigkeiten

- **Spearman‘s Rho als sog. „Rangkorrelationskoeffizient“:** Merkmalsausprägungen zweier Merkmale werden in jeweils geordnete Rangfolgen gebracht und in Beziehung gesetzt

Spearman‘s Rho

Vorgehen:

1) Merkmalsausprägungen zweier Merkmale werden in jeweils geordnete Rangfolgen gebracht

2) Anschließend werden die jeweiligen Rangpositionen der gepaarten Merkmalsausprägungen in Beziehung gesetzt

3) Rho mit Formel berechnen

$$\rho = 1 - \frac{6 * \sum_{i=1}^{n} d_i^2}{n * (n^2 - 1)}$$

$$\text{wobei } d_i = R(x_i) - R(y_i) \text{ (Differenz der Ränge)}$$

Kann Werte zwischen -1 und +1 annehmen, wobei 0 keinen Zusammenhang bedeutet

Spearman‘s Rho

- **Positiver Koeffizient** weist auf einen gleichgerichteten bzw. positiven Zusammenhang hin
- **Negativer Koeffizient** auf einen negativen Zusammenhang

Beispiel: Schulbildung und politisches Interesse (beide ordinal skaliert)

➢ **Interpretation positiver Wert:** Mit steigendem Bildungsabschluss lässt sich auch ein höheres politisches Interesse beobachten

➢ **Interpretation negativer Wert**: Mit steigendem Bildungsabschluss lässt sich ein niedrigeres politisches Interesse beobachten

Spearman‘s Rho: Beispiel

- **Zusammenhang zwischen sozialer Schicht und Gesundheitszustand** (vgl. auch Lehrbrief S. 60f.)

- **Hypothese**: je höher die soziale Schicht desto besser auch der Gesundheitszustand

ID	Soziale Schicht	Gesundheitszustand
1	Mittelschicht (3)	Schlecht (2)
2	Arbeiterschicht (2)	Sehr schlecht (1)
3	Oberschicht (5)	Sehr gut (5)
4	Obere Mittelschicht (4)	Gut (4)
5	Unterschicht (1)	Zufriedenstellend (3)

Quelle: Eigene Darstellung

Spearman‘s Rho: Beispiel

- Schritt 1: Überführung der Werte auf den Merkmalen in Rangpositionen - Arbeitstabelle

ID	Soziale Schicht (x)	Gesundheitszustand (y)	$d_i = R(x_i) - R(y_i)$
1	3	2	1
2	2	1	1
3	5	5	0
4	4	4	0
5	1	3	-2

Quelle: Eigene Darstellung

„Rangplätze“

Spearman‘s Rho: Beispiel

- Überführung der Werte auf den Merkmalen in Rangpositionen – Arbeitstabelle

ID	Soziale Schicht (x)	Gesundheitszustand (y)	$d_i = R(x_i) - R(y_i)$
1	3	2	1
2	2	1	1
3	5	5	0
4	4	4	0
5	1	3	-2

Quelle: Eigene Darstellung

$$\rho = 1 - \frac{6 * \sum_{i=1}^{n} d_i^2}{n * (n^2 - 1)} = \frac{6 * (1^2 + 1^2 + 0^2 + 0^2 + -2^2)}{5 * (5^2 - 1)} = 1 - \frac{6 * 6}{120} = 1 - 0{,}3 = 0{,}7$$

Spearman‘s Rho: Interpretation

Rangkorrelationskoeffizient ($\|r_{SP}\|$)	Interpretation
$\leq 0{,}05$	kein Zusammenhang
$> 0{,}05$ bis $\leq 0{,}20$	schwacher Zusammenhang
$> 0{,}20$ bis $\leq 0{,}50$	mittelstarker Zusammenhang
$> 0{,}50$ bis $\leq 0{,}70$	starker Zusammenhang
$> 0{,}70$	sehr starker Zusammenhang

Quelle: Eigene Darstellung

Spearman‘s Rho: Hinweis

- **Bei größeren Stichproben**: mindestens zwei oder mehr Fälle weisen die gleiche Merkmalsausprägung auf → Rangpositionen mit sogenannten Bindungen (engl. ties)
- **Pragmatisches Vorgehen**: Den verbundenen Beobachtungseinheiten wird das arithmetische Mittel derjenigen Rangplätze zugewiesen, die ohne Bindungen vergeben worden wären
- **Achtung: Übersteigt die Anzahl der gebundenen Rangplätze den Anteil von 20% aller Ränge, muss Spearman's Rho mit einer komplexeren Formel berechnet werden (siehe Lehrbrief; Statistikprogramm nutzt automatisch die korrekte Formel)**

Spearman's Rho: Hinweis

ID	Soziale Schicht	Gesundheitszustand
1	Mittelschicht (3)	Schlecht (2)
2	Arbeiterschicht (2)	Sehr schlecht (1)
3	Oberschicht (5)	Sehr gut → 4,5 für beide
4	Obere Mittelschicht (4)	
5	Unterschicht (1)	Zufriedenstellend (3)

Quelle: Eigene Darstellung

Mittelwert aus 4. und 5. Rangplatz

Spearman‘s Rho: Übung

Bildungsabschluss und Einkommens-gruppe:

Bitte berechnen Sie auf Basis der Urliste Spearman‘s Rho

ID	Bildung	Einkommen
1	Hauptschule (1)	Mittel (2)
2	Realschule (2)	Niedrig (1)
3	Abitur (3)	Hoch (3)
4	Realschule (2)	Hoch (3)
5	Studium (4)	Sehr hoch (4)

Plan für heute

- Gemeinsame Planung Sitzung 8.2.

Abschluss bivariate Analyseverfahren

- Zusammenhangsmaß zwischen ordinalskalierten Merkmalen: Spearman‘s Rho
- **Eta-Koeffizient (auch Eta-Quadrat) als Beispiel für ein sog. „PRE-Maß“**

Was sind „PRE-Maße“?

- PRE: „**Proportional Reduction of Error**“
- Verschiedene PRE-Maße in der Statistik, folgen der gleichen **Ausgangsfrage** und **Logik**
- **Ausgangsfrage**: Wie gut können die Werte einer abhängigen Variable durch die Werte einer unabhängigen Variable vorhergesagt bzw. erklärt werden?
- Setzen eine **„gerichtete“ Hypothese** voraus, d.h. wir formulieren einen Zusammenhang zwischen aV und uV
- Weisen Werte zwischen 0 und 1 auf (je höher desto besser erklärt die uV)

Bestimmung PRE-Maße

Schrittweises Vorgehen:

1) Wie lautet die Prognose des Wertes der abhängigen Variable **ohne** Kenntnis der unabhängigen Variablen? (→ Vorhersagefehler E1)
2) Prognose des Wertes der abhängigen Variable **mit** Kenntnis der Verteilung der unabhängigen Variable → (Vorhersagefehler E2)
3) Ermittlung des PRE-Maßes und Aussage, ob die Vorhersage durch die unabhängige Variable verbessert wurde. PRE = (E1 - E2) / E1

Je nach PRE-Maß werden die Fehler dabei unterschiedlich berechnet

PRE-Maß $\eta 2$ (Eta-Koeffizient oder Eta-Quadrat)

Wie gut lässt sich die Varianz eines abhängigen metrischen Merkmals durch ein unabhängiges Merkmal erklären?

- **Abhängige Variable** muss mind. (pseudo-) Intervallskalenniveau aufweisen
- **Unabhängige Variable** kann beliebiges (aber ggf. gruppiertes) anderes Niveau aufweisen

PRE-Maß η2 (Eta-Quadrat)

- Berechnung basiert auf **Quadratsummen**
- Quadratsumme stellt die **Summe der quadrierten Abweichungen vom arithmetischen Mittelwert** eines Merkmals dar
- **Verschiedene Quadratsummen werden unterschieden:**
  - **„Quadratsumme Gesamt“:** Entspricht der Summe aller quadrierten Abweichungen vom Mittelwert der (abhängigen) Variable **(Vorhersagefehler E1)**.
  - **„Quadratsumme innerhalb“**: Wird die unabhängige Gruppenvariable berücksichtigt, wird innerhalb der Gruppen berechnet, wie stark die Gruppenmittelwerte vom Mittelwert abweichen **(Vorhersagefehler E2)**.
  - **„Quadratsumme zwischen“:** „Quadratsumme Gesamt“- „Quadratsumme innerhalb“

PRE-Maß $\eta 2$ (Eta-Quadrat)

PRE = (E1 - E2) / E1

$$\eta 2 = \frac{\text{Quadratsumme Gesamt} - \text{Quadratsumme innerhalb}}{\text{Quadratsumme Gesamt}} = \frac{\text{Quadratsumme Zwischen}}{\text{Quadratsumme Gesamt}}$$

Beispiel PRE-Maß $\eta 2$ (Eta-Quadrat)

Politisches Wissen von Schüler*innen mit und ohne Lernbeeinträchtigung

ID	uV: Lernbeeinträchtigung	Politisches Wissen (y_i)
1	1 (Nein)	15
2	2 (Ja)	10
3	2 (Ja)	2
4	2 (Ja)	9
5	1 (Nein)	7
6	1 (Nein)	16
7	1 (Nein)	14
8	2 (Ja)	6
9	2 (Ja)	4
10	1 (Nein)	12

Quelle: Eigene Darstellung

Arithmetisches Mittel: 9,5

Beispiel PRE-Maß $\eta 2$ (Eta-Quadrat)

Politisches Wissen von Schüler*innen mit und ohne Lernbeeinträchtigung

ID	uV: Lernbeeinträchtigung	Politisches Wissen (y_i)
1	1 (Nein)	15
2	2 (Ja)	10
3	2 (Ja)	2
4	2 (Ja)	9
5	1 (Nein)	7
6	1 (Nein)	16
7	1 (Nein)	14
8	2 (Ja)	6
9	2 (Ja)	4
10	1 (Nein)	12

Quelle: Eigene Darstellung

Arithmetisches Mittel: 9,5

PRE-Maß η2 (Eta-Quadrat)

Schritt 1

- Arithmetisches Mittel als bester Vorhersagewert, wenn keine unabhängige Variable berücksichtigt wird (Quadratsumme gesamt)

Schritt 2

- Berücksichtigung uV (Quadratsumme innerhalb der Gruppen - Gruppenmittelwerte berücksichtigen, quadrierte Abweichungen berechnen und addieren)

Schritt 3

- Ermittlung Eta-Koeffizient mit Formel

PRE-Maß $\eta 2$ (Eta-Quadrat) - Arbeitstabelle

ID	Lern-beeinträchtigung	Politisches Wissen (y_i)	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	1 (Nein)	15	5,5	30,25
2	2 (Ja)	10	0,5	0,25
3	2 (Ja)	2	–7,5	56,25
4	2 (Ja)	9	–0,5	0,25
5	1 (Nein)	7	–2,5	6,25
6	1 (Nein)	16	6,5	42,25
7	1 (Nein)	14	4,5	20,25
8	2 (Ja)	6	–3,5	12,25
9	2 (Ja)	4	–5,5	30,25
10	1 (Nein)	12	2,5	6,25
		$\sum = 95$		$\sum = 204,5$
		$\bar{y} = 9,5$		

Quelle: Eigene Darstellung

E1: Quadratsumme Gesamt

PRE-Maß $\eta 2$ (Eta-Quadrat)

ID	Lern-beeinträchtigung	Politisches Wissen (y_i)	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	1 (Nein)	15	2,2	4,84
5	1 (Nein)	7	−5,8	33,64
6	1 (Nein)	16	3,2	10,24
7	1 (Nein)	14	1,2	1,44
10	1 (Nein)	12	−0,8	0,64
		$\sum = 64$		$\sum = 50,8$
		$\bar{y} = 12,8$		

Quelle: Eigene Darstellung

Quadratsumme Gruppe „keine Lernbeeinträchtigung“

PRE-Maß $\eta 2$ (Eta-Quadrat)

ID	Lern-beeinträchtigung	Politisches Wissen (y_i)	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
2	2 (Ja)	10	3,8	14,44
3	2 (Ja)	2	–4,2	17,64
4	2 (Ja)	9	2,8	7,84
8	2 (Ja)	6	–0,2	0,04
9	2 (Ja)	4	–2,2	4,84
		$\sum = 31$		$\sum = 44{,}8$
		$\bar{y} = 6{,}2$		

Quelle: Eigene Darstellung

Quadratsumme Gruppe innerhalb „mit Lernbeeinträchtigung“

Schritt 2 Zusammenfassung

Quadratsumme innerhalb (E2): 50,8+44,8= 95,6

PRE-Maß $\eta 2$ (Eta-Quadrat)

$$\eta 2 = \frac{\text{Quadratsumme Gesamt} - \text{Quadratsumme innerhalb}}{\text{Quadratsumme Gesamt}} = \frac{\text{Quadratsumme Zwischen}}{\text{Quadratsumme Gesamt}}$$

E2: 44,8+50,8=95,6

$$\eta 2 = \frac{E1 - E2}{E1} = \frac{204{,}5 - (50{,}8 + 44{,}8)}{204{,}5} = \frac{204{,}5 - 95{,}6}{204{,}5} = \frac{108{,}9}{204{,}5} = 0{,}53$$

Schätzfehler kann um 53% vermindert werden, wenn wir die uV (in unserem Beispiel: Lernbeeinträchtigung) berücksichtigen

PRE-Maß $\eta 2$ (Eta-Quadrat)

- Kann Werte zwischen 0 und 1 annehmen
- Lässt sich prozentual oder als Effektstärke interpretieren

Tabelle 41: Interpretation von Eta-Quadrat

Eta-Quadrat	Interpretation
< 0,01	kein Effekt
0,01 bis < 0,06	kleiner Effekt
0,06 bis < 0,14	mittlerer Effekt
≥ 0,14	großer Effekt

Quelle: Eigene Darstellung

Übung: PRE-Maß $\eta 2$ (Eta-Quadrat)

ID	Taschengeld pro Monat in Euro	Gruppe
1	3	1
2	9	2
3	10	1
4	4	2
5	2	1
6	4	1
7	3	2
8	6	1
9	5	2
10	6	2

Sie wollen untersuchen, ob ältere Grundschulkinder mehr Taschengeld bekommen als jüngere.

Sie haben dazu bei 10 Kindern die in der Tabelle abgetragenen Summen ermittelt. Zur Prüfung der Hypothese haben Sie zwei Kinder-Gruppen gebildet:

Gruppe 1: Klasse 1 und 2 vs.
Gruppe 2: Klasse 3 und 4 –

Berechnen Sie Eta-Quadrat und interpretieren Sie das Ergebnis!

Zusammenfassung bivariate Analysen

Was können bivariate Analysen leisten?

- Beschreibung der **Zusammenhänge** zwischen 2 Variablen
- **Kreuztabellen** und die damit verbundenen Maße (Chi-Quadrat, Cramers V) können Auskunft über Zusammenhänge zwischen nominalen Variablen liefern
- **Korrelationskoeffizienten** (Spearman/Pearson) beschreiben die Stärke „gleichsinniger“ oder gegenläufiger Zusammenhänge zweier Variablen
- **PRE-Maße** (z.B. Eta^2) geben an, inwieweit eine unabhängige Variable durch Einbezug einer weiteren Variable „besser“ erklärt werden kann

Was können bivariate Analysen nicht leisten?

- Überprüfen, inwieweit eine abhängige Variable noch von weiteren Variablen abhängig ist (nächstes Semester)
- Mehrere unabhängige Variablen in ihrem gemeinsamen Einfluss auf eine abhängige Variable untersuchen (nächstes Semester)
- Kausalitätsrichtung bestimmen

Wie soll die die letzte Sitzung im Semester inhaltlich gestaltet werden?

Folgende Möglichkeiten:

- Normal weiter im Stoff, Wiederholung Statistik I im SoSe einplanen
- Wiederholung durch Vorbereitung Ihrer erstellter „Klausurfragen“
- Wiederholung wichtigste Inhalte durch mich, evtl. Quizformat

Abstimmung über BigBlueButton jetzt gleich