Thema: χ2-basierte Zusammenhangsmaße, PRE-Maß Lambda, Kovarianz, Pearsons r

Formeln:

Erwartete Häufigkeiten $$f_{e(ij)} = \frac{Zeilensumme \times Spaltensumme}{n}$$

Residuum $$(f_{b(ij)} - f_{e(ij)})$$

Chi-Quadrat $$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{m} \frac{(f_{b(ij)} - f_{e(ij)})^2}{f_{e(ij)}}$$

Phi $$\phi = \sqrt{\frac{\chi^2}{n}}$$

Kontingenzkoeffizient C $$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad ; \quad C_{max} = \sqrt{\frac{R-1}{R}}$$

Cramér's V $$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{n * (R-1)}}$$

PRE-Maß λ $$\lambda = \frac{(Fehler_1 - Fehler_2)}{Fehler_1}$$

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

Pearsons r: $$r = \frac{\text{cov}(x,y)}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$\Rightarrow \quad s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \; ; \quad s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$$

Aufgaben:

1) χ^2-basierte Zusammenhangsmaße

Geben ist die Kreuztabelle der letzten Woche. Du hattest hierzu bereits einen Chi-Quadrat Wert von ($\chi^2 = 456{,}85$) errechnet, dann aber gemerkt, dass sich dein Ergebnis nur schlecht interpretieren lässt.

	Geschlecht		
Überleben d. Unglücks	Frau	Mann	Gesamt
Überlebt	$f_{o(ij)}$ = 344 $f_{e(ij)}$ = 151,83 Residuum = 192,17	$f_{o(ij)}$ = 367 $f_{e(ij)}$ = 559,17 Residuum = -192,17	711
Verstorben	$f_{o(ij)}$ = 126 $f_{e(ij)}$ = 318,17 Residuum = -192,17	$f_{o(ij)}$ = 1.364 $f_{e(ij)}$ = 1.171,83 Residuum =192,17	1.490
Gesamt	470	1.731	2.201

a) Mit welchen Zusammenhangsmaßen könntest du stattdessen arbeiten? Was wären Vor-und Nachteile der einzelnen Maßzahlen?

Da es sich um nominal skalierte Daten handelt können folgende Zusammenhangsmaße genutzt werden:

- Phi (ϕ): Kann nur für 2x2 Tabellen genutzt werden und besitzt einen Wertebereich von 0 bis 1
- Kontingenzkoeffizient C: variiert zwischen 0 und C_{max}
- Cramer's V: Ist am vielseitigsten einsetzbar. variiert zwischen 0 und 1.

b) Berechne alle sinnvoll nutzbaren Zusammenhangsmaße und interpretiere deine Ergebnisse!

2. $\phi = \sqrt{\frac{\chi^2}{n}}$

$= \sqrt{\frac{456{,}85}{2.201}}$

$\phi = \underline{0{,}46}$

3. $C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$

$= \sqrt{\frac{456{,}85}{456{,}85 + 2.201}}$

$C = \underline{0{,}41}$

$C_{max} = \sqrt{\frac{R-1}{R}}$

$= \sqrt{\frac{2-1}{2}}$

$C_{max} = \underline{0{,}71}$

4. $V = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{n*(R-1)}}$

$= \sqrt{\frac{456{,}85}{2.201*(2-1)}}$

$= \sqrt{\frac{456{,}85}{2.201}}$

$V = \underline{0{,}46}$

Hinweis: Bei 2x2 Tabellen sind Camer's V und Phi identisch.

Das χ^2-Maß berechnet über die gesamte Kreuztabelle hinweg, ob zwischen zwei Merkmalen ein Zusammenhang besteht ($\chi^2 > 0$) oder ob kein Zusammenhang vorliegt ($\chi^2 = 0$). Der berechnete χ^2 - Wert von 456,85 gibt an, dass ein Zusammenhang zwischen dem Überleben des Schiffsunglücks und dem Geschlecht der Passagiere besteht. Da das -Maß jedoch keine Aussage über der Stärke des Zusammenhangs treffen kann und zudem von der entsprechenden Fallzahl abhängig ist, ist es notwendig, weitere nominale Zusammenhangsmaße zu betrachten, die den entsprechenden -Wert normieren.

- Der entsprechende ϕ -Wert (0,46) gibt an, dass zwischen den beiden Merkmalen ein mittlerer Zusammenhang besteht (Wertebereich zwischen 0 und 1).

- Der Wert des Kontingenzkoeffizienten C (0,41) lässt auf einen eher starken Zusammenhang zwischen den beiden Variablen schließen, da die maximale Größe des Kontingenzkoeffizienten bei einer 2x2-Tabelle bei 0,71 liegt.
- Eine Betrachtung des Wertes von Cramér's V (0,46) lässt ebenfalls erkennen, dass ein starker Zusammenhang zwischen dem Überleben des Schiffsunglücks und dem Geschlecht der Passagiere vorliegt. In der Literatur herrschen jedoch unterschiedliche Auffassungen darüber, wann ein durch Cramér's V gemessener Zusammenhang als stark zu bezeichnen ist. Die unterschiedlichen Ergebnisse hinsichtlich der Stärke des Zusammenhangs zwischen den beiden Variablen sind demnach mit Vorsicht zu interpretieren.

  Interpretationshilfe für Cramér's V:
  0,0 bis 0,1 – kein Zusammenhang
  0,1 bis 0,2 – schwacher Zusammenhang
  0,2 bis 0,3 – mittlerer Zusammenhang
  ab 0,3 – starker Zusammenhang

2) PRE-Maß Lambda (λ)

a) Lege für die obige Tabelle fest, welche die unabhängige und welche die abhängige Variable ist!

UV: Geschlecht (Spalte)
AV: Überleben des Schiffsunglücks (Zeile)

b) Berechne Lambda!

Ausgangspunkt:
Wie gut können die Werte einer abhängigen Variablen durch die Werte einer unabhängigen Variable vorhergesagt werden? Hinweis: **λ** ist ein PRE-Maß für nominale Variablen.

Schritt 1: Prognose des Wertes der abhängigen Variable ohne Kenntnis der unabhängigen Variable.

Die beste Prognose für ein nominales Merkmal ist der Modus, da bei der Prognose des Modus die wenigsten Fehler gemacht werden. Der Modus hinsichtlich des Überlebens des Schiffsunglücks liegt bei der Kategorie „verstorben". Ohne Kenntnis des Geschlechts der Passagiere ist die bestmögliche Vorhersage des Überlebens des Schiffsunglücks somit die Kategorie „verstorben". Prognostizieren wir, dass die Passagiere der Titanic das Schiffsunglück nicht überlebt haben, liegen wir in 1.490 Fällen richtig, denn so viele Personen sind bei dem Schiffsunglück tatsächlich verstorben. In 711 Fällen, das sind die Personen, die das Schiffsunglück überlebt haben, irren wir uns. Die Summe der Fehler bei der abhängigen Variable ohne Berücksichtigung einer unabhängigen Variable wird Fehler1 genannt. Dieser beträgt in unserem Fall 711.

Schritt 2: Prognose des Wertes der abhängigen Variable mit Kenntnis der unabhängigen Variable.

Zur Prognose des Überlebens des Schiffsunglücks wird das Geschlecht der Passagiere herangezogen. Für jede Ausprägung der unabhängigen Variable wird der Wert der abhängigen Variable nun getrennt prognostiziert. Für die 470 Passagiere, die weiblich sind, prognostizieren wir ein Überleben des Schiffsunglücks, weil die Kategorie „überlebt" am häufigsten genannt wird (Modalkategorie). In 344 Fällen liegen wir mit dieser Prognose richtig, in 126 Fällen irren wir uns. Das sind diejenigen Passagiere, die trotz ihres weiblichen Geschlechts, das Schiffsunglück nicht überlebt haben.

Unsere Prognose für die 1.731 Passagiere, die männlich sind, lautet dagegen, dass diese das Schiffsunglück nicht überlebt haben, da die Kategorie „verstorben" am häufigsten genannt wird (Modalkategorie). Hier liegen wir bei 1.364 Fällen richtig, dagegen irren wir uns bei 367 Personen.

Das sind diejenigen Passagiere, die trotz ihres männlichen Geschlechts, das Schiffsunglück überlebt haben. Die Summe der Fehler, die wir trotz Berücksichtigung der unabhängigen Variable begehen, nennen wir Fehler2. Dieser beträgt in unserem Fall 493 (126 + 367).

Schritt 3: Ermittlung des PRE-Maßes **λ**

$$\lambda = \frac{(Fehler1 - Fehler2)}{Fehler\ 1} = \frac{(711-493)}{711} = 0{,}31$$

Interpretation:
Inwieweit wurde die Vorhersage durch Einbezug der unabhängigen Variable verbessert? Die Fehler bei der Prognose des Überlebens des Schiffunglücks werden durch die Kenntnis des Geschlechts der Passagiere um 31% verringert.

3) Kovarianz

Du hast gerade in der Mensa mit einer Freundin darüber diskutiert, dass ein Kommilitone nur sehr selten zu Seminaren erscheint und trotzdem immer gute Noten bekommt. Ihr beschließt mit Hilfe einer Erhebung unter 5 KommilitonInnen herauszufinden, ob die Häufigkeit der Anwesenheit in einem Seminar mit der Seminarnote zusammenhängt. Eure Befragung bringt euch folgende Ergebnisse: Steffen: 4 mal da, 5 Punkte; Mathias: 9 mal da, 11 Punkte; Aggi: 2mal da, 9 Punkte; Maria: 12 mal da, 14 Punkte; Tanja: 9 mal da 15 Punkte.

Fall	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
Steffen	4	5	-3,2	-5,8	18,56	10,24	33,64
Mathias	9	11	1,8	0,2	0,36	3,24	0,04
Aggi	2	9	-5,2	-1,8	9,36	27,04	3,24
Maria	12	14	4,8	3,2	15,36	23,04	10,24
Tanja	9	15	1,8	4,2	7,56	3,24	17,64
	36	54			51,2	66,8	64,8

a) Überprüfe mit Hilfe der Kovarianz, ob ein Zusammenhang zwischen den beiden Variablen besteht!
n = 5
$\bar{x}$ = (4+9+2+12+9)/5 = 36/5 = 7,2
$\bar{y}$ = (5+11+9+14+15)/5 = 54/5 = 10,8
Kovarianz:

$$\text{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n} = \frac{51{,}2}{5} = 10{,}24$$

Die Kovarianz ist ungleich 0, somit korrelieren hohe Werte der X-Variablen mit hohen Werten der Y-Variablen (und niedrige mit niedrigen Werten). Es besteht also ein Zusammenhang zwischen der Häufigkeit der Seminarbesuche und der Seminarnote.

b) Zeichne einen Scatterplot!
Der Scatterplot lässt einen starken positiven Zusammenhang erahnen. Es ist zu erkennen, dass hohe Werte der X-Variablen mit hohen Werten der Y-Variablen einhergehen.

c) Berechne und interpretiere Pearson's r!

Berechnung: Pearson's r:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} = \frac{66,8}{5} = 13,36 \; ; \; s_y^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n} = \frac{64,8}{5} = 12,96$$

$$r = \frac{\mathrm{cov}(x,y)}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{10,24}{\sqrt{13,36*12,96}} = 0,78$$

Interpretation: Es liegt ein starker Zusammenhang zwischen der Häufigkeit der Seminarbesuche und der Seminarnote vor.

4) χ^2-basierte Zusammenhangsmaße

Gegeben ist die Kreuztabelle aus Aufgabe 1. Die Zahl der Fälle sowie die Angaben der absoluten Häufigkeiten wurden im Vergleich zur Aufgabe 1 jedoch verdoppelt.

	Geschlecht		
Überleben d. Unglücks	Frau	Mann	Gesamt
Überlebt	$f_{o(ij)}$ = 688 $f_{e(ij)}$ = 303,65 Residuum = 384,35	$f_{o(ij)}$ = 734 $f_{e(ij)}$ = 1.118,35 Residuum = -384,35	1.422
Verstorben	$f_{o(ij)}$ = 252 $f_{e(ij)}$ = 636,35 Residuum = -384,35	$f_{o(ij)}$ = 2.728 $f_{e(ij)}$ = 2.343,65 Residuum = 384,35	2.980
Gesamt	940	3.462	4.402

a) Berechnet die folgenden nominalen Zusammenhangsmaße:

1. (Chi-Quadrat)
2. (Cramér's V)

1. $$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{m} \frac{(f\,b_{(ij)} - f\,e_{(ij)})^2}{f\,e_{(ij)}}$$

$$= \frac{(384{,}35)^2}{303{,}65} + \frac{(-384{,}35)^2}{1.118{,}35} + \frac{(-384{,}35)^2}{636{,}35} + \frac{(384{,}35)^2}{2.343{,}65}$$

$$\chi^2 = \underline{913{,}77}$$

2. $$V = \sqrt{\frac{\chi^2}{\chi^2_{\max}}} = \sqrt{\frac{\chi^2}{n*(R-1)}}$$

$$= \sqrt{\frac{913{,}77}{4.402*(2-1)}}$$

$$= \sqrt{\frac{913{,}77}{4.402}}$$

$$V = \underline{0{,}46}$$

b) Was fällt Euch bei einem Vergleich mit den entsprechenden Werten aus Aufgabe 1 auf?
Bei einem Vergleich der beiden χ^2-Werte fällt auf, dass sich mit Verdopplung der absoluten Häufigkeiten auch der entsprechende χ^2- Wert verdoppelt hat. Dies liegt daran, dass χ^2 u.a. abhängig ist von den absoluten Häufigkeiten in den Zellen. Somit führt eine Verdopplung der absoluten Häufigkeiten schließlich zu einer Verdopplung von χ^2. Die prozentuale Verteilung in den Zellen bleibt hingegen unverändert. Bei einem Vergleich der beiden Werte von Cramér's V lässt sich erkennen, dass eine Verdopplung der absoluten Häufigkeiten nicht zu einer Verdopplung von Cramér's V geführt hat. Dies liegt daran, dass Cramér's V den χ^2-Wert normiert, indem dieser durch den maximal erreichbaren χ^2-Wert dividiert wird. In einer 2x2-Tabelle kann χ^2 somit maximal so groß sein wie die Zahl der Beobachtungen n.