Phi-Koeffizient

Haben zwei Variablen jeweils nur zwei Ausprägungen, so handelt es sich um *dichotome* Variablen. Man unterscheidet *natürlich* dichotome und *künstlich* dichotome Variablen.

Von natürlicher Dichotomie sprechen wir dann, wenn eine Variable *nominalskaliert* ist und von Hause aus nur *zwei* Ausprägungen hat (Beispiel: Rauchen Sie - Ja, Nein; Geschlecht - männlich, weiblich).

Unter künstliche Dichotomie versteht man, dass eine Variable ursprünglich *intervallskaliert* und *normalverteilt* ist und erst im Nachhinein in zwei Klassen eingeteilt wurde (Beispiel: zwei Altersklassen).

Haben wir nun zwei dichotome Variablen, von denen *mindestens eine* natürlich dichotom ist, so lässt sich zwischen beiden ein Zusammenhang errechnen mit Hilfe des so genannten Phi-Koeffizienten.

Beispiel: Betrachten wir die beiden natürlich dichotomen Variablen Geschlecht (m, w) und Rauchen (ja, nein). Unsere Daten bestehen aus folgenden beiden Zahlenkolonnen:

Geschlecht	Rauchen Sie?
m	Ja
m	Ja
w	Nein
w	Ja
m	Nein

usw.

Zur Berechnung des Phi-Koeffizienten benötigen wir zunächst eine zweidimensionale Häufigkeitstabelle.

Denken Sie zurück an unsere ersten einfachen, eindimensionalen Häufigkeitstabellen.

Rauchen Sie?

	Wert	Häufigkeit	Prozent
Ja	1	14	46.7
Nein	2	16	53.3
	Gesamt	30	100.0

Eine zweidimensionale Häufigkeitstabelle ist nun nichts anderes als eine nochmalige Untergliederung der Frage "Rauchen Sie?" nach dem Geschlecht, also:

	m	w	Randsumme
JA	20(a)	10(b)	30
NEIN	30(c)	40(d)	70
Randsumme	50	50	100 (TOTAL)

Man nennt eine solche Anordnung von Häufigkeiten auch eine 2 mal 2 Felder-Tafel.

$$Phi = \frac{a \cdot d - b \cdot c}{\sqrt{(a+c) \cdot (b+d) \cdot (a+b) \cdot (c+d)}}$$

$$Phi = \frac{20 \cdot 40 - 10 \cdot 30}{\sqrt{(20+30) \cdot (10+40) \cdot (20+10) \cdot (30+40)}}$$

Man beachte: der Wurzelausdruck entspricht dem Produkt der Randsummen!!!

PHI = 0,22

Interpretation des Phi-Koeffizienten:

Um diesen Korrelationskoeffizienten sinnvoll interpretieren zu können, müssen wir uns in Erinnerung rufen, welchen Sinn die Berechnung einer Korrelation eigentlich hat. Wir wollen aufgrund der Werte der *einen* Variablen die Werte der *anderen* Variablen voraussagen.

Wie müsste nun unsere zweidimensionale Häufigkeitstabelle beschaffen sein, damit wir von der Eigenschaft, Raucher zu sein oder nicht, auf die Eigenschaft, männlich oder weiblich zu sein, schließen können?

Stellen Sie sich vor alle Raucher wären männlich ***und*** (!) alle Nichtraucher wären weiblich. In diesem Falle könnte man einfach aufgrund der Ausprägung der Variable "Rauchen Sie" auf das Geschlecht schließen. Haben wir die Frage "Rauchen Sie?" beantwortet, so wüssten wir bei einer perfekten Korrelation automatisch das Geschlecht. Nur in diesem Falle hätten wir eine perfekte Korrelation, also eine Korrelation = 1!

Alle Raucher sind männlich bedeutet: Es gibt keine rauchenden Frauen. Alle Nichtraucher sind weiblich bedeutet: es gibt keine nicht rauchenden Männer. Was dies für das Erraten des Geschlechts aufgrund der Variable Rauchen bedeutet, kann folgendes Beispiel illustrieren, wobei ich hoffe, dass Sie sich dabei amüsieren ;-)

Stellen Sie sich vor, wir würden über den ganzen Hörsaal eine Segelplane aufspannen und an jeder Stelle, an der sich ein Student bzw. eine Studentin befindet, wäre ein kleines Loch als Dunstabzug. Einer von uns würde nun oben auf dieser Segelplane herumspazieren (vorausgesetzt, die Plane ist genug stabil dafür).

Wären nun alle Raucher männlich und alle Nichtraucher weiblich, so wäre es ein leichtes, das Geschlecht des jeweiligen Studieren unter dem Dunstabzug zu erraten: Überall dort, wo ein Rauch aufsteigen würde, wäre ein männlicher Student und unter allen Löchern, aus denen kein Rauch aufstiege, befände sich eine Studentin!

Die Voraussage ist aber nur solange perfekt, als auch tatsächlich kein einziger männlicher Student Nichtraucher wäre *UND* keine einzige Studentin rauchen würde.

Wie müsste nun eine derartige zweidimensionale Häufigkeitstabelle beschaffen sein, damit sie diese Voraussetzung überhaupt erfüllen kann?

Versuchen wir zunächst, alle Raucher in dem Feld "männlich" (keine Raucher im Feld weiblich!) und alle Nichtraucher in dem Feld "weiblich" (keine Nichtraucher im Feld männlich) unterzubringen. (Denken Sie daran, dass in unserer gesamten Untersuchung 30 Raucher und 70 Nichtraucher vorkommen):

	m	w	Randsumme
JA	30(a)	0(b)	30
NEIN	0(c)	70(d)	70
Randsumme	30(50)	70(50)	100(TOTAL)

Diese Tabelle ist aber falsch. In Klammern stehen die richtigen Angaben, wie sie sich aufgrund unserer empirischen Untersuchung ergeben haben (nämlich eine Verteilung beim Geschlecht von 50:50). Richtig dagegen ist, dass wir einen Anteil von 30 Rauchern und 70 Nichtrauchern haben(macht in der Gesamtsumme 100 befragte Personen),

aufgrund den Häufigkeiten in den vier Feldern müssten wir jedoch auch einen gleichen Anteil von Frauen und Männern haben, um einen perfekten Zusammenhang von der Variable „Rauchen“ und dem „Geschlecht“ überhaupt erzielen zu können!

Wir bräuchten also auch 30 Frauen und 70 Männer, um überhaupt zu einer perfekten Korrelation zu kommen!

Allgemein gilt: Damit eine perfekte Korrelation überhaupt möglich ist, muss das Verhältnis der Ausprägungen der Variablen x gleich sein dem Verhältnis der Ausprägungen der Variable y (Beispiel: 30: 70 beim Geschlecht und ex aequo bei der Frage "Rauchen Sie?")

Nun sind aber die Randsummen bereits durch die Stichprobenerhebung festgelegt. Wir haben eben in unserem Kollektiv von 100 Personen zufällig 30 Raucher und 70 Nichtraucher erwischt, die sich auf 50 Frauen und 50 Männer verteilen!

Welches maximale Phi können wir daher aufgrund der vorgegebenen Randsummen bekommen?

Bringen wir wieder alle Raucher unter der Rubrik "männlich" unter. Sind alle Raucher männlichen Geschlechts, so gibt es 0 Personen weiblichen Geschlechts, die rauchen.

Umgekehrt müssten alle Nichtraucher weiblichen Geschlechts sein. Es dürften also keine nicht rauchenden Männer geben. Das geht sich aber aufgrund der Randsummen nicht aus. Wir können aufgrund der vorgegebenen Randsummen nur danach trachten, möglichst wenige Frauen bei den Rauchern bzw. umgekehrt möglichst viele Frauen bei den Nichtrauchern unterzubringen, also:

	m	w	Randsumme
JA	30(a)	0(b)	30
NEIN	20(c)	50(d)	70
Randsumme	50	50	100(TOTAL)

Dies ist - aufgrund der vorgegebenen Randsummen - der bestmögliche Zusammenhang, den wir bei einer Verteilung von 30 : 70 bei den Rauchern bzw. von 50 : 50 beim Geschlecht bekommen können. Wir wissen in dem konkreten Fall zwar, daß jemand, der Raucher ist, auch männlichen Geschlechts sein muss (es gibt in dieser Verteilung keine rauchenden Frauen),

alle Raucher sind zwar Männer, aber das umgekehrte ist nicht der Fall: Wissen wir von jemandem, dass er Nichtraucher ist, so ist eine exakte Zuteilung zu einem bestimmten Geschlecht nicht möglich. Auf unser obiges Beispiel mit der Segelplane bezogen bedeutet dies nun, dass wir aus einer solchen Tabelle (die wir ja nur konstruiert haben, um den maximal möglichen Zusammenhang zu erzielen) folgendes herauslesen können: Steigt irgendwo ein Rauch auf, so wissen wir mit Sicherheit, dass es ein Mann sein wird (da alle unsere 30 Raucher männlichen Geschlechts sind). Wo aber kein Rauch aufsteigt, können wir nur einen Trend ablesen: dass es sich nämlich eher um einen Frau als einen Mann handeln wird (die 70 Nichtraucher sind mehrheitlich weiblich, nämlich insgesamt 50).

Der langen Rede kurzer Sinn: Wir können im Falle der vorliegenden Randsummen nur einen *Trend* zwischen der Frage "Rauchen Sie?" und dem Geschlecht feststellen. Eine exakte Voraussage ist dahingegen schon aufgrund der empirisch vorgegebenen Randverteilung gar nicht möglich.

Kodieren wir die Ausprägungen der beiden natürlich dichotomen Variablen mit "0" und "1", so entspricht der Phi-Koeffizient rechnerisch dem Produkt-Moment-Koeffizienten. Phi wird manchmal daher auch als Vier-Felder-Produkt-Moment-Korrelation bezeichnet.