Tetrachoric Association Under The Assumption Of Normality

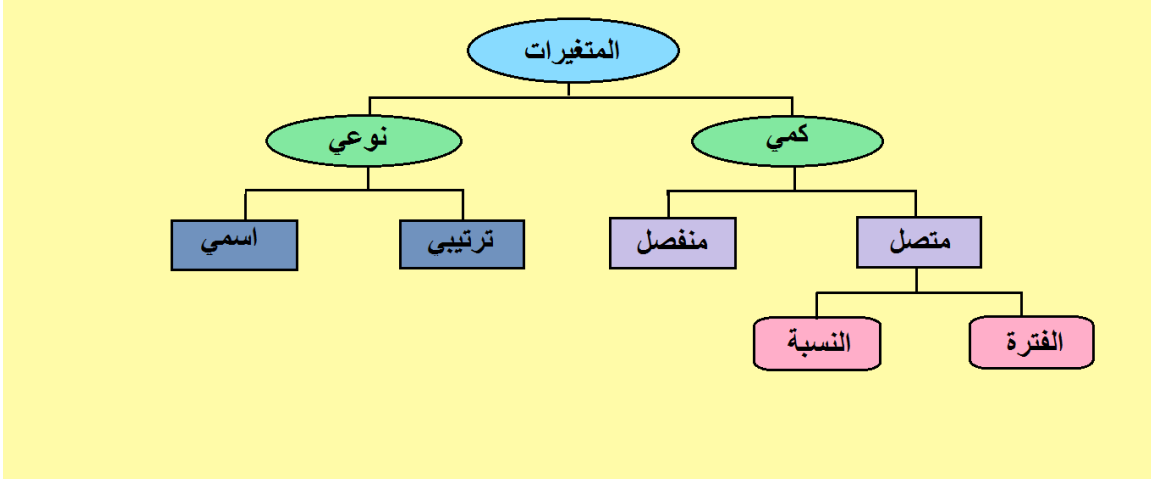
Dr. Ahmed S. El-Aloosy.
Baghdad College of Economic Sciences University

الخلاصة:

المتغيرات في الأحصاء مختلفة في الصفات فمنها ما هو كمي ومنها ماهو نوعي فالمتغيرات الكمية تخضع للقياس كالوزن و الطول .....الخ.و المتغيرات النوعية لا تخضع الى قياس بل تحتاج الى ترميز خلال اجراء التحليلات الأحصائية كالجنس و رتبة الولادة .....الخ. معامل الأرتباط بين المتغيرات كمي – كمي , كمي – نوعي , نوعي – نوعي , و كذ لك معامل الأرتباط في حالة المتغيرين إسمي – إسمي , إسمي – ترتيبي , ترتيبي – ترتيبي , حيث لا يوجد مؤشر واحد يمثل الارتباط بين المتغيرات المتعددة.

في هدا البحث استخدمنا الأرتباط الثلاثي (tetra choric) للجداول التوافقية (2X2)
بأفتراض أن التوزيعين الجانبين هما توزيع طبيعي معياري, و بأستخدام سلسلة هيرمت و خاصية
مهلر في التعامد استطعنا استنباط المؤشر أعلاه و وبتطبيق جداول بيرسون لأيجاد

$$r_t = \frac{h_n(x) \quad (x)}{n_i}$$



Introduction:

Numbers can't "talk" but they can tell as much as your human sources can. But just like with human sources, YOU HAVE TO ASK?

By the expression categorical data, we mean data which are presented in the form of frequencies falling into certain categories or classes. A categorized "variable" may simply be a convenient classification of a measurable variable into groups, in the manner already familiar to us. On the other hand it may not be expressible in terms of an underlying measurable variable at all. For example, we may classify by : (a) their height, (b) their color,(c) their favorite games. Here (a) is a categorization of measurable variable, but (b) and (c) are not. There is a further distinction between (b) and (c) , for hair color itself may be expressed on an order scale ,according to pigmentation from light to dark. This is not so for (c) .We refer to (b) as an ordered classification or categorization, and (c) as an unordered one.

There is a further point to be born in mind: on occasion, the two variables being investigated may simply be the same variable observed on two different occasions, e.g.,(before and after some events) or on the related samples e.g., father and son, husband and wife, etc..).We shall refer to such situation as one with identical categorization. Identical categorization may, of course, be of any of the types (a) ,(b)or (c).

Our interest in categorical data associated with two or more variables expressed in categorical form, there expression called a contingency table

Types of variables

(i) Qualitative, ex. Eye color, hair color etc.
(ii) Quantitative: ex. Weight, height etc.

Measures of Association:

In many research associations, the strength and nature of the dependence of variables is of central concern. No single measure adequately summarizes all possible types of association. Measures vary in their interpretation and in the way they define perfect and immediate association. These measures also differ in the way they are affected by various factors such as marginal, for example: many measures are "margins sensitive" in that they are influenced by the marginal distributions of rows and columns. Such measures reflect information about the margins along with information bout associations.

A particular measure may have a low value for a given table, not because the two variables are not related, but because they are not related in the way to which the measure is sensitive. No single measure is best for all situations, the type of data, the hypothesis of interest as well as the properties of various measures must be considered when selecting an index of association for a given table. It is not, however, reasonable to compute a large number of measures and then to report the most impressive as if it were the only one examined. We conclude the following:

1- Dependence of variables is of central concerned.
2- No single measure summarizes all possible types of associations.
3- Measures vary in their interpretation.

4- Measures are affected by various factors such as marginal that they are influenced by the marginal distributions of rows and columns.

Purpose: dependence of variables is of the central concern.

Properties:

- ❖ No single measure is best for all situations.
- ❖ Measures vary in their interpretations,
- ❖ Measures are affected by various factors.
- ❖ Measures are affected by type of variables such as nominal, ordinal, interval, and ratio.

Tetrachoric function:

The numerical evaluation of term like $h_n(x)\,\phi(x)$, may be carried out by:

- ❖ Directly by finding the polynomial $h_n(x)$ multiplied by $\phi(x)$ such as : $h_n(x)\,\phi(x)$.
- ❖ From certain tables of so-called Hermitian probability function with negative index. This give

$h_n(x) = (-D)^{n-1}\, e^{-1/2\, x2}$ to 10 decimal places from 0 to 7 .

or x = -7 (0.1) 6.5 .

British Association of Mathematical Table Vol.1, 1946.

- ❖ Similar function was tabulated by Karl Pearson in the form:

$$\tau_r(x) = \frac{h_{n-1}(x)\,\phi(x)}{\sqrt{n!}}$$

This was known as a tetrachoric function to estimate ρ from 2x2 table.

Estimation of Tetrachoric Correlation:

- ❖ Canonical forms Bivariate Normal Distribution; we want to look at the measure of association in 2x2 table where underline distribution is BSN (x, y, ρ), bivariate standard normal distribution:

$$f(x,y,\rho) = \frac{1}{2\pi\sigma x\,\sigma y\,(1-\rho)1/2} \exp\left[\frac{1}{2(1-\rho 2)}\left(\frac{(x-\mu)^2}{\sigma 2} - 2\rho\,\frac{(x-\mu)(y-\mu)}{\sigma\,\sigma} + \frac{(y-\mu)^2}{\sigma 2}\right)\right]$$

In order to do this we must look at series expansion for $\phi(x, y, \rho)$ in $\{f(x, y, \rho)$, where $\mu_x = \mu_y = 0$, $\sigma_x = \sigma_y = 1$ $\}$. And to facilitate this we have to look at Hermite Chebyshave Polynomial.

- ❖ Hermite Chebyshave Polynomial (H.C.P) properties: we define the set of Hermite Chebyshave Polynomial

$$h_n(x)\,\phi(x) = (-1)^n \frac{d^n \phi(x)}{d^n x} \quad n = 0,1,2,\ldots\ldots..$$

Where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-1/2\, X\, 2}$

Example: (i) $n = 0$ $h_0(x) = 1$

(ii) $n = 1$ $h_1(x)\,\phi(x) = (-1)^1(-x)\dfrac{e^{-1/2\,x^2}}{\sqrt{2\pi}} = x\,\phi(x)$

then $h_1(x) = x$,

(iii) $n = 2$ $h_2(x) = x^2 - 1$

Hence in general term of $h_n(x)$ of H.C.P is

Therefore $h_0(x) = 1$ $h_1(x) = x$

$h_2(x) = x^2 - 1$ $h_3(x) = x^3 - 3x$

$h_4(x) = x^4 - 6x^2 + 3$ $h_5(x) = x^5 - 0x^3 + 15x$

$h_6(x) = x^6 - 15x^4 + 45x^3 - 15$

$$h_n(x) = x^n - \frac{n(n-1)}{2\,.\,1!}x^{n-1} + \frac{n(n-1)(n-2)(n-3)}{22\;\;2!}x^{n-2} - \ldots$$

$$\ldots\ldots..+(-1)^n \frac{n(n-1)(n-2)\ldots\ldots(n-2r+1)}{2r\quad r!}x^{n+r} + \ldots\ldots\ldots..$$

Properties of H.C.P. we have:

theorem (1) $h_{n+1}(x) = h_n(x) - n\,h_{n-1}(x)$

theorem (2) $h'_n(x) = n\,h_{n-1}(x)$

theorem (3) $\{h_n(x)\}$ orthogonal to $\phi(x) = \dfrac{1}{\sqrt{2\pi}}\,e^{-1/2\,x^2}$

Meaning of orthagonality:

$$I = \int_{-\infty}^{\infty} h_r(x)\,h_s(x)\,\phi(x)\,dx = \delta_{r\,s} = \begin{cases} 0 & \text{if } r =/= s \\ r_! & \text{if } r = s \end{cases}$$

(Proof is available).

Normalized or orthonormal H.C.P 's

(1) $$H_n(x) = \frac{h_n(x)}{\sqrt{n!}}$$

(2) $$\int_{-\infty}^{\infty} H_r(x)\,H_s(x)\,\phi(x) = \delta_{r\,s} = \begin{cases} 0 & r =/= s \\ 1 & r = s \end{cases}$$

Now ; If r =s

$$r! = \int_{-\infty}^{\infty} h_r(x)\, h_s(x)\, \phi(x)\, dx$$

- Mehler Identity:

Mehler (1866) showed that the bivariate standard normal BSN(0 , 0 , 1 , 1 , ρ) :

$$\Phi(x, y, \rho) = \frac{1}{2\pi (1-\rho^2)^{1/2}} \exp[-1/2(1-\rho^2)\{x^2 - 2\rho xy + y^2\}]$$

Could be expressed in the following canonical form involving H.C.B. such as:

$$\Phi(x, y, \rho) = \phi(x)\,\phi(y)\left(\sum \rho^n H_n(x)\, H_n(y)\right)$$

$$= \phi(x)\phi(y)\left(1 + \sum \rho^n H_n(x)\, H_n(y)\right)$$

$$= \phi(x)\,\phi(y)\left(\sum \rho^n h_n(x)\, h_n(y)\right)$$

The proof of Mehler Identity is based on the fact that there is a one to one correspondence between probability distribution and the characteristic function.
To show that Mehler Identity can be used to calculate product moment about the origin of $\phi(x,y,\rho)$ by showing that $E(x^2 y^2) = 1 + 2\rho^2$
We have:

$$E(x^2 y^2) = \int\int x^2 y^2\, \phi(x, y, \rho)\, dx\, dy$$

$$= \int\int (h_2(x)+1)(h_2(y)+1)\left(1+\sum \rho^n h_n(x)\, h_n(y)\right) \phi(x)\phi(y)\, dxdy$$

$$= \int\int \phi(x)\,\phi(y)dxdy + \int\int h_2(x)\, h_2(y)\, \phi(x)\,\phi(y)\, dxdy +$$

$$\int\int h_2(x)\,\phi(x)\phi(y)dxdy + \int\int h_2(y)\phi(y)\,\phi(x)dxdy +$$

$$\int\int \left(h_2(x)\, h_2(y) + h_2(x) + h_2(y) + 1\right)\sum \frac{\rho^n}{n!} h_n(x)\, h_n(y)\, \phi(x)\phi(y)$$

$$= 1 + 2\rho^2$$

Similarly

$E(x^3 y^5) = E(h_3(x) + 3h_1(x))(h_5(y)+10h_3(y)+15h_1(y)) = 45\rho + 60\rho^3$

And

$$E(x^6 y^6) = E(h_6(x)+15(h_4(x)+6h_2(x)+3) - 45(h_2(x)+1)+15) * (h6(x)+15(h4(x)+6h2(x)+3) - 45(h2(x)+1)+15)$$

$$= 225+\rho^6/6!+\rho^4/4!+\rho^2/2!$$

From these results we are able to construct a table for the product moments

For $E(x^i y^j) = \int\int x^i y^j \phi(x,y,\rho)\, dxdy$ as the following:

$E(x^i y^j)$

i, j	1	2	3	4	5
1	ρ	0	3ρ	0	15ρ
2	0	$1 + 2\rho^2$	0	$3+12\rho^2$	0
3	3ρ	0	$9\rho+6\rho^3$	0	$45\rho+60\rho^3$
4	0	$3+12\rho^2$	0	$9+72\rho^2+24\rho^4$	0
5	15ρ	0	$45\rho+60\rho^3$	0	$225+600\rho^3+120\rho^5$

- ❖ Tetrachoric estimation of 2X2 contingency table frequency:

1. We know that bivariate normal implies the marginal are also normal.
2. Fit normal to marginal.
3. Values (h, k) are not necessarily the centers for marginal they are point of dichotomy.

We have $\Phi(h) = p(x \le h) = \int \phi(x)dx = \dfrac{n_{1.}}{n..}$

$\Phi(k) = p(y \le k) = \int \phi(y)\, dy = \dfrac{n_{.1}}{n..}$

n_{11}	n_{12}	$n_{1.}$
n_{21}	n_{22}	$n_{2.}$
$n_{.1}$	$n_{.2}$	$n..$

Pearson 1900 derived the tetrachoric estimator of ρ by equating the relative frequency in any cell to its expected value i.e. consider n_{22} say:

$$\frac{n_{22}}{n..} = E\left(\frac{n_{22}}{n..}\right) = \frac{1}{n..} E(n_{22}) = \frac{1}{n..} n..\, p_{22} = p_{22}$$

or $p_{22} = \int\int \phi(x,y,\rho)\, dx\, dy$

$$= \int\int \phi(x)\,\phi(y)\,(1+ \Sigma\, \rho^i\, h_i(x)\, h_i(y))\, dx\, dy$$

$$= \int \phi(y)\, dy \int \phi(x)dx + \Sigma \left(\frac{\rho^i}{i!} \left(\int h_i(y)\,\phi(y)\, dy\right) \left(\int h_i(x)\phi(x)dx\right)\right)$$

$$= (1 - \Phi(k))(1- \Phi(h)) + \Sigma\, (\rho^i$$

$$\left(\int h_i(y)\phi(y)dy \right) \left(\int h_i(x)\phi(x)dx \right)$$

$$= \left(1 - \frac{n_{.1}}{n..}\right)\left(1 - \frac{n_{1.}}{n..}\right) + \Sigma_{i!} \frac{\rho^i}{i!} \left(-h_{i-1}(y)\phi(y) \right) \left(-h_{i-1}(x)\phi(x) \right)$$

$$= \left(1 - \frac{n_{.1}}{n..}\right)\left(1 - \frac{n_{1.}}{n..}\right) + \Sigma_{i!} \frac{\rho^i}{i!} h_{i-1}(h)\, h_{i-1}(k)\phi(h)\phi(k)$$

$$\frac{n_{22}}{n..} - \frac{n_{2.}\, n_{.2}}{n..^2} = \phi(h)\phi(k) \Sigma_{i!} \frac{\rho^i}{i!} h_{i-1}(h) h_{i-1}(k)$$

$$\text{or} \quad w(\rho) = \frac{n_{11}\, n_{22} - n_{12}\, n_{21}}{n..^2} = \phi(h)\, \phi(k) \Sigma_{i!} \frac{\rho^i}{i!} h_{i-1}(h)\, h_{i-1}(k)$$

If we truncate w (ρ); we have a polynomial equation in ρ, the solution is the characteristic estimate r_t of ρ that is :

$$w(\rho) = \Sigma_{i!} \frac{\rho^i}{i!} h_{i-1}(h)\, h_{i-1}(k)\phi(h)\, \phi(k)$$

❖ Maximum likelihood 0f ρ 1n 2X2 table under BSND:

We want to maximize L where:

$$L = \Pi\Pi\, p_{ij}^{n_{ij}}$$

to do this we need estimate of p_{ij} , we have already shown that

$$p_{22} = \int\int \phi(x,y,\rho)\, dxdy$$

$$= \frac{n_{2.}\, n_{.2}}{n..^2} + w(\rho) \qquad \text{or} \quad Pij = \frac{n_{ij}\, n_{ij}}{n..} w(\rho) \quad \text{for all i,j .}$$

now consider $\log L \;\alpha\; \Sigma\Sigma\, n_{ij} \log p_{ij}$

Therefore $\frac{d \log L}{d\rho} = \Sigma\Sigma \frac{n_{ij}}{p_{ij}} \frac{d \log p_{ij}}{d\rho}$

put $\frac{d \log L}{d\rho} = 0$ to get :

$$\frac{n_{11}}{p_{11}}\frac{dp_{11}}{d\rho} + \frac{n_{21}}{p_{12}}\frac{dp_{12}}{d\rho} + \frac{n_{21}}{p_{21}}\frac{dp_{21}}{d\rho} + \frac{n_{22}}{p_{22}}\frac{dp_{22}}{d\rho} = 0$$

To solve for ρ we need to find expression for the p_{ij} and $\frac{dp_{ij}}{d\rho}$;

Recall that $p_{ij} = \frac{n_{i.}\, n_{.j}}{n_{..}^{2}} +\text{-} w(\rho)$ and $p'_{ij} = \frac{dp_{ij}}{d\rho} = +\text{-}\, w'(\rho)$

For all i=1, 2 and j = 1, 2

We have then: $\left(\frac{n_{11}}{P_{11}} - \frac{n_{12}}{p_{12}} - \frac{n_{21}}{p_{21}} + \frac{n_{22}}{p_{22}}\right) w'(\rho) = 0$

Since w' (ρ) = 0 then: $\left(\frac{n_{11}}{P_{11}} - \frac{n_{12}}{p_{12}} - \frac{n_{21}}{p_{21}} + \frac{n_{22}}{p_{22}}\right) = 0$

If we replace p_{ij} in this equation and simplifying we get:

$\frac{ni.\, n.j}{n^2} + w(\rho) = \frac{n_{ij}}{n} + u(\rho)$, To obtain the following equation:

$$n^3\, u^3(\rho) + 2n(n_{11}\, n_{22} - n_{12}\, n_{21})\, u^2(\rho) - (n_{11}\, n_{22}(n_{12} + n_{21}) + n_{12}\, n_{21}\,(n_{11} + n_{22}))u(\rho) = 0$$

The maximum likelihood equation admits three values for ρ that is ρ_1 , ρ_2 , and ρ_3 Where ρ_1 , is obtained from u(ρ) = 0 , while ρ_2 and ρ_3 are obtained from the second part of the solution. Thus there three values for the MLE. of ρ , we can show that ρ_2 and ρ_3 are terminal maximal, then ρ_1 can be proved easily that it is the MLE for ρ .

Therefore $w(\rho_1) = \frac{n_{11}\, n_{22} - n_{12}\, n_{21}}{n^2}$ and $\rho_1 = \tau(x)$.

Example(1) :

Find the tetrachoric estimator of ρ in the following table of classification of milk (gallon's) and age of cows:

Age years				
Milk Gallon's		3-5	> 5	Total
	8-18	1407	1078	2485
	> 18	881	1546	2427
	Total	2288	2624	4912

Pearson's tables for statistician and Biometrician, Vol.2 ,(1931) are tabulated where n_{22} corresponds to the smallest frequency. If n_{22} is not in the desired position change rows and columns, odd numbers if change effects of the sign of $\tau_{r(x)}$.

We have

Age years				
Milk Gallon's		> 5	3-5	Total
	8-18	1078	1407	2485
	>18	1546	881	2427
	Total	2624	2288	4912

$$\frac{n_{22}}{n} = \frac{881}{4912} = 0.1794$$

$$\phi(k) = \frac{n_{.1}}{n} = \frac{2624}{4912} = 0.542 \longrightarrow \text{hence} \quad k = 0.11$$

$$\phi(h) = \frac{n_{1.}}{n} = \frac{2485}{4912} = 0.5059 \longrightarrow \text{hence} \quad h = 0.015$$

The tables are entered with triple $\left(\frac{n_{22}}{n}, h, k\right) = (0.1794, 0.11, 0.015)$

Therefore $\rho t = 0.32$. which correspond to $\varphi = 0.201$ according to X^2

Example (2) :

The following table summarizes hypothetical ratings by two raters on presence (+) or absence (-) of schizophrenia.

	Rater 2		
Rater 1	+	- _	total
+	40	10	50
-	20	30	50
total	60	40	100

for these data ,the tetrachoric correlation $\rho t = 0.6072$,which is much larger than pearson correlation of $r = 0.4082$ calculated for same data.

References:

Alan AG., Categorical Data Analysis: Applied Statistics. Wiley, 1990

Bishop YMM, Fienberg SE, Holland PW. Discrete multivariate analysis: Theory and practice. Cambridge, Massachusetts: MIT Press, 1975

Brown MB. Algorithm AS 116 : The tetra choric correlation and its standard error. Applied Statistics, 1977, 26 , 343-351.

Drasgow F. Polychoric and polyserial correlations. In Kotz L, Johnson NL (Eds.), Encyclopedia of statistical sciences. Vol. 7 (pp. 69-74). New York: Wiley, 1988

Fedoryuk, M.V. (2001), "Hermite function" in Hazewinkel, Michiel, Encyclopedia of Mathematics, Springer, ISBN 978-1-55608-010-4

Harris B. Tetrachoric correlation coefficient. In Kotz L, Johnson NL (Eds.), Encyclopedia of statistical sciences. Vol. 9 (pp. 223-225). New York: Wiley, 1988

Hutchinson TP . Kappa muddles together two sources of disagreement: Tetrachoric correlation is better. Research in Nursing and Health. 1993, 16, 313-315

Hutchinson TP . Assessing the health of plants: Simulation helps us understand observer disagreements. Environ metrics, 2000, 11, 305-314.

Joreskog KG, Sorbom, D. PRELIS User's Manual, Version 2. Chicago: Scientific Software, Inc., 1996.

Koornwinder, Tom H.; Wong, Roderick S. C.; Koekoe, Roelof; Swarttouw, Rene' F .(2010), 'Orthogonal Polynomials", in Olver, Frank W . J.;Lozier, Daniel M.; Biosvert, Roland F .; Clark, Charles W., NIST Handbook of Mathematical Functions, Cambridge University Press, ISBN 978-0521192255, MR 2723248

Whittaker, E. T.; Watson, G. N. (1962), 4th, ed., A Course of Modern Analysis, London: Cambridge University Press

Temme, Nico, Special Functions: An Introduction to the Classical Functions of Mathematical Physics, Wiley, New York, 1996