MODELING CATEGORICAL DATA IN FREQUENCY DOMAIN VIA MUTUAL INFORMATION APPROACH

Wale-Orojo O. A.[1], Soyinka A.T.[1], Apantaku F. S.[1], Atanda O. D.[2] and Akintunde A. A.[1]

[1] **Department of Statistics, Federal University of Agriculture Abeokuta, Ogun state Nigeria**
[2] **Department of Computer Science, Faculty of Natural and Applied Science, Nile university of Nigeria, Abuja**

Corresponding author: Oluwaseun Wale-Orojo, seunorojo@gmail.com

ABSTRACT: Phenomenon whose occurrences are described with count data in a multi-dimensional contingency table is common in medical and epidemiological studies. This study thus obtained the density function(s) of kth nonlinear vectors via mutual information approach and its measure of dependence from multivariate exponential power distribution which is a member of the multivariate elliptical contoured family. The obtained model which accommodates light and heavy-tailed member distributions depending on the shape parameter was used to establish results for kth nonlinear vectors from multivariate Laplace, normal, uniform, students' t, and Fisher distributions. The Multivariate dependency of live birth on maternal age as the number of pregnancies increases irrespective of maternal height advantage as claimed from the previous study was established from the theoretical model obtained.

KEYWORDS: Multivariate elliptical contoured distribution, exponential power distribution, mutual information, kth nonlinear density model, entropy, live birth, maternal height.

1. INTRODUCTION

Defining and obtaining a powerful measure of nonlinear statistical dependence among kth dependent vectors in frequency domain using mutual information approach is centered on obtaining a well-defined density function for the dependent vectors [5]. Owing that most real life problems have varying vectors (variables) whose individual outcomes has collective conditional joint effect on each other due to complicated interactions 'for instatnce molecular interactions in biological networks'; then a clear multivariate representation of vector interactions via a multivariate density function for kth dependent vectors is very necessary to define in clear terms total and conditional multivariate mutual information (MMI) dependence (Soyinka et al. (2017), [10], [6]). Though density functions for independent kth vectors are well developed in literatures for various continuous distributions; the need to develop a multivariate density function for kth dependent vectors is inevitable at this crucial point of our research development because it is frequently manifesting itself in cluster analysis, cryptography, data mining, networking and imaging and in various studies involving natural grouping of attributes in frequency domain [1][3]. Hence this study used the analogy between multivariate mutual information and symmetric covariance matrix to obtain an expression that can represent the squared radial function of the elliptical contoured distribution in terms of nonlinear dependent structural model (Soyinka et al. (2017)). The obtained multivariate nonlinear density model which is shape dependent was transformed into various member distributions of the exponential power distribution while creating a link to categorize the continuous function via degrees of freedom parameter. The measure of statistical dependence was then obtained as the entropy of the obtained transformed distributions. Though the interpretation of mutual information as a measure of dependence was once not straight forward [6], the use of well-defined distributional based multivariate mutual information density model has made easy the desired interpretation even for nonlinear dependence. So, in this study the obtained density model and its entropy are very useful in quantifying nonlinear dependence metric among qualitative and quantitative data set in frequency, discrete or continuous outcome from various distributions.

2. MATERIALS AND METHODS

2.1 Elliptical Density

Definition 1: Let $X \sim EC_k\left[\mu, \Sigma, \beta, h^{(k)}\right]$ be an elliptical contour (EC) random vector in a kth dimensional sets of real numbers $\Re^k$, with location vector $\mu \in \Re^k$, dispersion matrix $\Sigma \in \Re^{k\times k}$, shape parameter $\beta \in \Re$ and density generator h^k; then the probability density function (pdf) of the elliptical contour is

$$|\Sigma|^{-\frac{1}{2}} h^k \exp\left(-\frac{1}{\beta}\left[(X-\mu)^T \Sigma^{-1}(X-\mu)\right]^{\beta/2}\right);$$

$X \in \Re^k$. Here, the density generator h^k is a non-negative real valued function from $g(s) = \frac{\pi^{k/2} s^{\frac{k}{2}-1} h^k(s)}{\Gamma(k/2)}$, $s > 0$, which makes the function a valid probability density function Arellano-et-al (2012), Lindsey (1999). Note 's' is a squared radial function distributed as Laplace, normal and uniform density when $\beta = 1$, $\beta = 2$ and when β approaches ∞. However, note that for multivariate normally distributed ($\beta = 2$) vectors say X and Y with distinct categorical variates p and q respectively; the squared radial function's' is Chi–squared distributed with $p+q$ degree of freedom that is as $S_x \approx \chi_p^2$, $S_y \approx \chi_q^2$, $S_{xy} \approx \chi_{p+q}^2$ Arellano-et-al (2012), where S_{xy} is the joint distribution of random vectors X and Y.

Definition 2: Arellano et.al. (2012) noted that the elliptical mutual information index (total dependence) between X~$\left[\mu_x, \Sigma_{xx}, h^{(p)}\right]$ and Y~ $\left[\mu_y, \Sigma_{yy}, h^{(q)}\right]$is

$$I(X,Y) = E\left[log\left(h^{p+q} S_{xy}\right)\right] - E[log(h^p S_x)] - E\left[log\left(h^q S_y\right)\right] - \frac{1}{2} log(1 - \rho_{xy}^2) \quad (1)$$

Note that (1) has two components; the linear component $I_{Linear}(X,Y) = -\frac{1}{2}\log(1-\rho_{xy}^2)$ and the nonlinear component

$$I(X,Y)_{non-linear} = E\left[log(h^{p+q} S_{xy})\right] - E[log(h^p S_x)] - E\left[log(h^q S_y)\right].$$

The linear component is a measure of linear dependence/similarity/correlation co-efficient while the nonlinear component which ranges from $[0,\infty)$ accommodates the degrees of nonlinearity that defines real life dependence. Extending the result of Arenallo et al. (2012) on total dependence to kth dimensional random variables $X_1, X_2, ..., X_k$ from exponential power distribution we obtained a finite measure for multivariate mutual information total dependence as

$$I(X_1, X_2, \ldots, X_k) = \frac{1}{2} ln|\Sigma_{k\times k}| + \left[ln\, D + \frac{1}{\beta}\right] \quad (2)$$

where $D = \frac{\pi^{k/2}\Gamma\left(1+{}^{k}/_{\beta}\right)\beta^{{}^{k}/_{\beta}}}{\Gamma\left(1+{}^{k}/_{2}\right)}$, see Soyinka et al. (2017) for proof.

Hence in the multi-dimensional form, the nonlinear dependence $I(X_1, X_2, \ldots, X_k)_{nonlinear} = \left(ln\, D + {}^{1}/_{\beta}\right)$ is the absolute difference between total dependence and linear dependence.

$$I(X_1, X_2, \ldots, X_k)_{nonlinear} = I(X_1, X_2, \ldots, X_k) - \frac{1}{2} ln|\Sigma_{k\times k}| \quad (3)$$

Substituting (3) into the pdf in definition (1) owing that the multivariate mutual information and the covariance matrix structure has the relationship $\exp\left[I\left(X_1, X_2, .., X_k\right)\right] \propto \Sigma_{k\times k}$ [Soyinka et al. (2017), Reginald (2015); Adrenallo et al, (2012)]; then the pdf $f(X_1, X_2, ..., X_k)$ of kth nonlinear vectors from multivariate mutual information elliptical contour is

$$exp\left(-{}^{P}/_{2}\right) h^k \exp\left[-\frac{1}{\beta}\left([(X-\mu)\exp(-P)(X-\mu)]^{\beta/2}\right)\right];$$

where $P = I(X_1, X_2, ..., X_k) - \frac{1}{2}\ln\left|\Sigma_{k\times k}\right|$ (4)

where $h^k = D^{-1}$ is the normalizing constant.

Proposition 1: The measure of entropy $-E\left[\log f(X_1, X_2, ...X_k)\right]$ among kth nonlinear vectors from elliptical contoured distribution is

$$\frac{P}{2} - ln\, h^k + \frac{1}{\beta} exp\left(-\frac{P\beta}{2}\right) E[(X-\mu)^T (X-\mu)]^{\beta/2} \quad (5)$$

Proof: Take the logarithm of (4) and find it negative expectation. Note that the value of the function $E\left|(X-\mu)^T (X-\mu)\right|^{{}^{\beta}/_{2}}$ in (5) is dependent on the shape parameter $\beta \in \Re$ and it takes different distributions with varying β values. Next from (5) we establish the following results:

Corollary 1: The entropy of the kth nonlinear vectors from standardized Laplace distribution is

$$\frac{P}{2} - ln\left[\Gamma\left({}^{k}/_{2}\right)\Big/_{2\Gamma(k)}\right] + k\left[ln\sqrt{\pi} + exp\left(-{}^{P}/_{2}\right)\right] \quad (6)$$

Proof: Substitute $\beta = 1$ into (5) and simplify.

Corollary 2: The entropy of the kth nonlinear vectors from standardized normal distribution is

$$\frac{P}{2} + \frac{k}{2}[ln\, 2\pi + exp(-P)] \quad (7)$$

Proof: Substitute $\beta = 2$ into (5) and simplify.

Corollary 3: The entropy of the kth nonlinear vectors from uniform distribution is

$$\frac{P}{2} + \frac{k}{2} ln\, \pi - ln\, \Gamma\left(1 + {}^{k}/_{2}\right) \quad (8)$$

Proof: Substitute a large value for β such that $\beta \to \infty$ in (5) and simplify.

2.1 *kth* Multivariate student's *t-distribution*

Proposition 2: Let $t \sim T_k[\mu, \Sigma, \beta]$ be a multivariate student's $t-distribution$ in a kth dimensional sets of real numbers $\Re^k$, with location vector $\mu \in \Re^k$, dispersion matrix $\Sigma \in \Re^{k\times k}$, shape parameter $\beta \in \Re$; then the pdf of kth nonlinear vectors from $t \sim T_k[\mu, \Sigma, \beta]$ is

$$f(t) = exp\left(-\frac{P}{2}\right)\frac{\Gamma\left(\frac{1}{\beta}+\frac{k}{2}\right)}{\left(\frac{2\pi}{\beta}\right)^{k/2}\Gamma\left(\frac{1}{\beta}\right)}\left(1+\frac{\beta}{2}(t^T t)\right)^{-\left(\frac{1}{\beta}+\frac{k}{2}\right)} \quad (9)$$

Proof: Obtain the probability density function of t for the relationship $t = Y\Big/\sqrt{\frac{\beta Z}{2}}$ supposing both $Z = {}^1\!/_\beta\,|(x-\mu)^T(exp[-P])(x-\mu)|^{\beta/2}$ and $Y \sim N_k(0, I_k)$ are independent.

Corollary 4: Similarly substituting $\frac{1}{r} = \frac{\beta}{2}$ in (9) we obtain the pdf of kth nonlinear vectors with r degrees of freedom from $t \sim T_k[\mu, \Sigma, \beta]$ as

$$f(t) = exp\left(-\frac{P}{2}\right)\left(\frac{\Gamma\left(\frac{r+k}{2}\right)}{(r\pi)^{k/2}\Gamma\left(\frac{r}{2}\right)}\right)\left(1+\frac{1}{r}\left(t^T t\right)\right)^{-\left(\frac{r+k}{2}\right)} \quad (10)$$

Proposition 3: The entropy of the kth nonlinear vectors from $t \sim T_k[\mu, \Sigma, \beta]$ is

$$\frac{P}{2} - log\left(\frac{\Gamma\left(\frac{r+k}{2}\right)}{\Gamma\left(\frac{r}{2}\right)}\right) + \frac{k}{2}ln(r\pi) + \left(\frac{r+k}{2}\right)\left[\psi\left(\frac{r+k}{2}\right) - \psi\left(\frac{r}{2}\right)\right] \quad (11)$$

Proof: Obtain $\left(-E\left[\log\left(f(t)\right)\right]\right)$ from (10), noting that $E\left(\ln\left[1+{}^{\left(t^T t\right)}\!/_r\right]\right)$ is a digamma function.

2.2 *kth* Multivariate *Fisher distribution*

Proposition 4: Let $X \sim EC_k[\mu_x, \Sigma_x, \beta_1, h^{(k)}]$ and $Y \sim EC_k[\mu_y, \Sigma_y, \beta_2, h^{(k)}]$ be two independent elliptical contoured (EC) random vector in a kth dimensional sets of real numbers $\Re^k$, with location vectors $\mu_x, \mu_y \in \Re^k$, dispersion matrices $\Sigma_x, \Sigma_y \in \Re^{k\times k}$, shape parameter $\beta_1, \beta_2 \in \Re$ and density generator h^{2k}; a random variable $q = \frac{\beta_1 y}{\beta_2 x}$ has the probability density function

$$f(q) = exp\left(-\frac{1}{2}[P_1 - P_2]\right)\frac{\Gamma\left(k\left[\frac{1}{\beta_1}+\frac{1}{\beta_2}\right]\right)}{\Gamma\left(\frac{k}{\beta_1}\right)\Gamma\left(\frac{k}{\beta_2}\right)}\frac{\left(\frac{\beta_2}{\beta_1}\right)^{\frac{1}{\beta_1}}q^{\frac{k}{\beta_1}-1}}{\left(1+\frac{\beta_2}{\beta_1}q\right)^{k\left[\frac{1}{\beta_1}+\frac{1}{\beta_2}\right]}} \quad (13)$$

Proof: Obtain the probability density function for the relationship $q = \frac{\beta_1 X}{\beta_2 Y}$ supposing $X \sim EC_k[\mu_x, \Sigma_x, \beta_1, h^{(k)}]$ and $Y \sim EC_k[\mu_y, \Sigma_y, \beta_2, h^{(k)}]$ are independent owing that $\left|\Sigma_{k\times k}\right|_{x+y}^{-1/2} = \left|\Sigma_{k\times k}\right|_x^{-1/2} + \left|\Sigma_{k\times k}\right|_y^{-1/2}$ for $I[(X_1, X_2, \ldots X_k)|(Y_1, Y_2, \ldots Y_k)]$.

Corollary 4: Substituting $\frac{1}{\beta} = \frac{r}{2}$ into (13) where r_1 and r_2 are degrees of freedom and β_1 and β_2 are shape parameters for random variables X and Y respectively then we obtain

$$f(q) = exp\left(-\frac{1}{2}[P_1 - P_2]\right)\frac{\Gamma\left(\frac{k}{2}[r_1+r_2]\right)\left(\frac{r_1}{r_2}\right)^{\frac{kr_1}{2}}}{\Gamma\left(\frac{kr_1}{2}\right)\Gamma\left(\frac{kr_2}{2}\right)}\frac{q^{\frac{kr_1}{2}-1}}{\left(1+\frac{r_1}{r_2}q\right)^{\frac{k}{2}[r_1+r_2]}} \quad (14)$$

Proposition 5: The entropy of the kth nonlinear vectors from Fisher distribution is

$$\frac{1}{2}(P_1 - P_2) - ln\left[\frac{\Gamma\left(\frac{k}{2}[r_1+r_2]\right)\left(\frac{r_1}{r_2}\right)^{\frac{kr_1}{2}}}{\Gamma\left(\frac{kr_1}{2}\right)\Gamma\left(\frac{kr_2}{2}\right)}\right] + \left(\frac{kr_2}{2}+1\right)\left[\psi\left(\frac{k}{2}\langle r_1+r_2\rangle\right) - \psi\left(\frac{kr_1}{2}\right)\right] \quad (15)$$

Proof: Obtain $\left(-E\left[\log\left(f(q)\right)\right]\right)$ from (14), owing that q is beta distributed.

3. RESULTS

Restricting our application to kth nonlinear multivariate student's *t-distribution* vectors via mutual information dependence, the tabular representation of live/still birth with maternal age as the number of pregnancy increases despite variation in maternal height was examined among the

population of pregnant women that attended the facility of General Hospital Ijaye Abeokuta, Ogun state, Nigeria for antenatal as in tables 1 and 2.

Table 1: Distribution of Live and still birth in a cross tabulation of maternal age and height

	Live Birth (Y)			Still Birth (Y)		
Age (X)	Maternal Height			Maternal Height		
	Low	Average	High	Low	Average	High
18	0	3	9	-	-	-
22	7	0	61	1	0	3
27	12	4	111	3	0	11
32	8	1	83	2	2	13
37	4	2	58	0	1	8
42	2	0	10	0	0	3
	$H(X) = 1.543062,$ $H(Y) = 0.4183516,$ $H(X,Y) = 1.9380436,$ $I(X,Y) = 0.02337,$ $\lvert\Sigma\rvert = 2.5$			$H(X) = 1.430414301,$ $H(Y) = 0.610261,$ $H(X,Y) = 1.956137,$ $I(X,Y) = 0.0845387,$ $\lvert\Sigma\rvert = 3.5$		

Table 2: Effect of maternal age and rth number of pregnancy on the likelihood of live birth

Maternal Age	Number of Previous Pregnancy / Degree of freedom						
	$r = 1$	$r = 2$	$r = 5$	$r = 7$	$r = 10$	$r = 15$	$r = 20$
18	0.101	0.1285	0.1502	0.15365	0.1554	0.1558	0.1556
22	0.5463	0.34454	0.0703	0.0279	0.0086	0.0018	0.0005
27	0.6316	0.3211	0.0357	0.0097	0.0018	0.0002	2.51×10^{-5}
32	0.6010	0.3329	0.0472	0.015	0.0034	0.0005	8.9×10^{-5}
37	0.5268	0.3458	0.0789	0.0335	0.0113	0.0027	9×10^{-4}
42	0.1135	0.14	0.1536	0.1529	0.1503	0.1463	0.1435

Discussion of Numerical Results

The mutual information kth nonlinear multivariate student's *t-distribution* random variable used in this study shows various patterns in live birth with maternal age; the most common among the patterns was however a gradual decrease in the likelihood of live birth as maternal age increases. This decreasing trend was later found to show some distortion when the maternal age is beyond 38 years. As a result of this distortion, the study further considers the likelihood of live birth among women between the age of 38 – 45 years only, viz a viz the number of previous pregnancies they have heard before. This however reveals that after the fifth previous pregnancy, the likelihood of live birth begins to drop in agreement to the initial decreasing trend. Various graphs were used to further explain the workability of the theoretical model to the application. See appendix.

CONCLUSION

This study established the methodology to obtain a dependent density model among multi-dimensional nonlinear vectors whose interaction are arranged in a contingency set up; such that each of the vectors is represented in at most ordinal scale and the outcome of interest is in frequency. The shape parameter of the desired density model was proven to be proportional to the degree of freedom in a chi-square distribution and thus was required to link the methodology of obtaining the desired density model for kth nonlinear vectors with frequency outcome to that of existing measured outcome. Variables (vectors) with frequency outcome are common in medical and epidemiological studies where prevalence and incidence of an event over a period of time is usually the interest of the researcher in a case-control or cohort study. Thus, in addition to obtaining a measure of nonlinear dependence among multi-dimensional vectors, to reveal the level of researcher's interest as a point estimate, this study also established the graphical relationship between variables in categorical set up to reveal true association. The measure of nonlinear dependence established in this study is more sensitive to nonlinearity among variables unlike the correlation in the past that is highly insensitive to nonlinear variables.

Recommendation

This study recommends that every woman should not go beyond five pregnancies. Likewise, the study noted that the best age to start child bearing is between 22 and 38 years.

REFERENCES

[1] **Faes L. and Nollo G.** (2011). Multivariate frequency domain analysis of causal interactions in physiological time series. INTECH Open Access Publisher, 2011.

[2] **Han T. S.** (1980). Multiple Mutual Information and Multiple Interactions in Frequency Data Information and Control, volume 46(1); page 26-45.

[3] **Javier, W. R. & Gupta, A. K.** (2009). Mutual information for certain multivariate distributions. *Far East J. Theor. Stat.* 29, 39–51.

[4] **Lindsey J. K.** (1999). Multivariate Elliptically Contoured Distributions for Repeated Measurements. Biometrics 55, 1277-1280.

[5] **Malladi R., Johnson D. H., Kalamangalam G., Tandon N., and Aazhang B.** "Measuring cross-frequency coupling using mutual information and its application to epilepsy," in Cosyne Abstracts, Salt Lake City, USA, 2017.

[6] **Reginald Smith** (2015). A mutual information approach to calculating non linearity. The ISI journal for rapid dissemination of statistical research. ArXiv: 1512.0075Ovl.

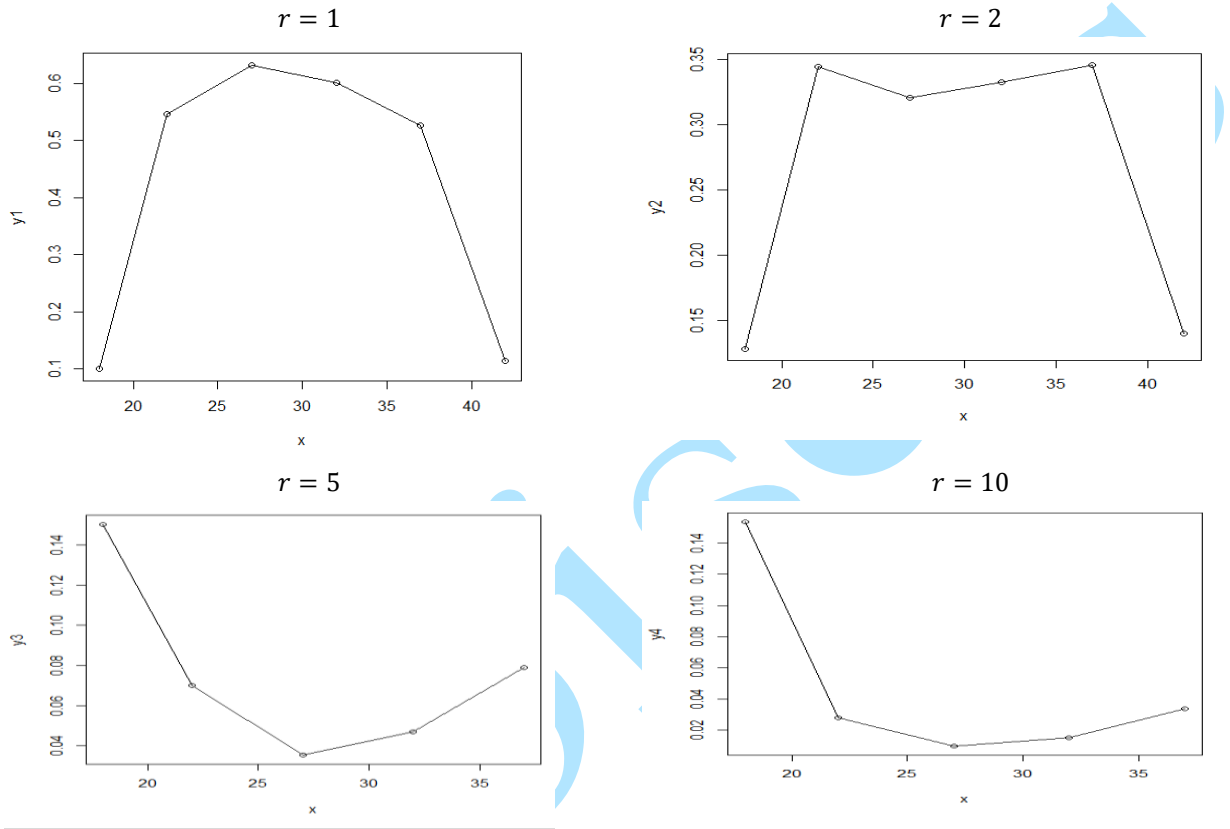
[7] **Reinaldo B. Arellano-valley, Javier E. Contreras-reyes and Marc G. Genton** (2012). Shanon Entropy and mutual information for multivariate and skew-elliptical distribution. Scandinavian Journal of statistics. Theory and applications. Wiley Publishing limited.

[8] **Shanon C. E.** (2001). A mathematical theory of communication. Bell sys. Tech.J., 27(3):379-423.

[9] **Sunil Srinivasa** (2003). Review on Multivariate Mutual Information.
www.nd.edu>tutorial>sunil.

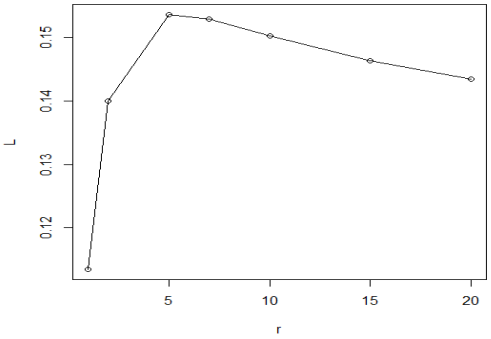
[10] **ThoHoan Pham, TuBao Ho, QuynhDiep Nguyen, Dang Hung Tran and Van Hoang Nguyen** (2012). Multivariate Mutual Information measures for Discovering Biological Networks.

Appendix



$y1, y2, y3, y4$: Likelihood; x : Maternal age; r : Number of previous pregnancy

Likelihood of life birth beyond 38 years in pregnant women (L)



Number of previous Pregnancy (r)