Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Shoukri et al., J Biomet Biostat 2014, 5:1
DOI: 10.4172/2155-6180.1000186

Research Article Open Access

Testing of Gender Differences on Sib-Sib Correlations for Binary Traits: Likelihood Based Inference with Application to Arterial Blood Pressures Data

Mohamed Shoukri[1,3]*, K Collison[2,3] and F Al-Mohanna[1,2,3]

[1]National Biotechnology Center, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia
[2]Department of Cell Biology, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia
[3]College of Medicine Al-Faisal University, Riyadh, Saudi Arabia

Abstract

Estimation of measures of familial aggregation is considered the first step in establishing whether a specified disease has a genetic component. Population based family study designs areusually used to estimate correlations among siblings. When the trait of interest is quantitative (e.g. blood pressure, body mass index, blood glucose level) testing the effect of gender differences on sib-sib correlations is achieved using the likelihood method of estimation under the assumption of multivariate normality. When the trait of interest is measured on the binary scale testing the equality of a brother-brother and sister-sister correlation is more complex. In this paper we develop likelihood-based inference procedures for this purpose which may beapplied to nuclear family data.

Keywords: Family studies; Multivariate beta-binomial distribution; Likelihood estimation; Score test

Introduction

Population based family studies have been used in genetic epidemiology to assess the association of environmental risk factors with disease and to quantify the aggregation of cases within families. These types of studies integrate statistical methods and classical epidemiology to analyze the correlations among family members who share the same genetic and environmental background. There are several advantages in using family study designs. The use of extended pedigrees or even nuclear families enhances the statistical power for gene discovery. Clinical characteristics common to family members may also be used to increase information by defining subgroups of families for analysis such as in the investigation of familial aggregation of components of metabolic syndrome [1]. Another important feature of family studies in contrast to studies of unrelated individuals is the issue of internal control. The analysis of traits of interest for family members should account for both genetic background factors and environmental exposures. A common example is the case of monozygotic twins where maximum genetic control is achieved. It is well-known that nuclear family members tend to have relatively similar environmental conditions, diet, and perhaps levels of physical activity. The familial aggregation of much chronic and infectious disease is also well documented. For example, results from recent studies have shown that pathogens causing Th1 diseases are passed from parents to child. Information accessed on December 23-2012 from (http://bacteriality.com/2008/07/31/hpv/) reveals that some of the chronic bacterial species that cause inflammatory illness can remain alive in breast milk and thus be passed from mother to child through breast feeding. Growing evidence suggests that the Th1 pathogens, rather than genetic mutation, are the driving force behind this familial aggregation. Although their role is unclear, researchers have also found a relationship between bacterial infection and cancer [2]. This chain of reasoning provides a possible explanation for the aggregation of cancer in families. A study conducted in 2010 using the PET scanner to examine the prevalence of plaque in brains (which is the hallmark ofAlzheimer's disease) found that a child's level of plaque is consistent with the corresponding levels of their fathers and in particular of their mothers, even years before the child's diagnosis [3].

In many population-based family studies, interest is focused in detecting gender differences in the risk of developing a chronic disease. For example, a recent study [4] aiming at examining sex-specific associations between cardiovascular risk factors and type 2 diabetes mellitus showed that there are gender-related dissimilarities that are apparently involved in disease development. Another study conducted on a sample of families from South Australia [5] found that men and women face different challenges in the management of diabetes and its associated complications.

One of the major limitations of the above studies is that the comparisons between males and females were based on parallel group designs, and consequently suffer from the lack of control over possible confounding. Another limitation is that the lack of a reference population makes the problem of statistical inference (estimation and hypothesis testing) less meaningful. A further methodological challenge that faces researchers is that estimates of trait correlations, specifically the intra-class correlations for males and females, are themselves correlated. Although studies on comparing sib-sib correlations have been of frequent interest [6,7] comparisons among these correlations have been usually made descriptively. When traits are measured on the continuous scale, Donner et al. [8] developed several procedures for comparing the sib-sib correlations among males and females, including likelihood-based tests, while assuming that the underlying mechanism generating the data is multivariate normal. When the trait of interest is measured the binary scale, efficient methods for comparing sib-

***Corresponding author:** Mohamed Shoukri, National Biotechnology Center-King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia, Tel: +966-509491454; E-mail: shoukri@kfshrc.edu.sa

Received January 16, 2014; **Accepted** February 22, 2014; **Published** February 28, 2014

Citation: Shoukri M, Collison K, Al-Mohanna F (2014) Testing of Gender Differences on Sib-Sib Correlations for Binary Traits: Likelihood Based Inference with Application to Arterial Blood Pressures Data. J Biomet Biostat 5: 186. doi:10.4172/2155-6180.1000186

sib correlations characterizing males and female have not yet been developed.

This paper has a threefold objective. First, we develop a multivariate probability distribution for the vector of binary observations based on a random sample of independent sib-ships. The vector will be split into two sub-clusters, separating female responses from male responses. Second we construct the likelihood function of the sample as based on the joint distribution of the created sub-clusters. This allows us to develop score and Wald chi-square-tests of significance that compare the levels of similarity among males and females from the same family. Finally, we illustrate our procedures using published arterial blood pressures data.

Models

Suppose that we have a random sample of k sib-ships, where each sib-ship constitutes a cluster. Let $y_i=(y_{i1},y_{i2},\ldots,y_{ib_i},x_{i1}, x_{i2},\ldots x_{is_i})^T$ denote the vector of observations from the i^{th} cluster, where b_i=number of brothers in the i^{th} family, s_i=number of sisters in the i^{th} family, $n_i=b_i+s_i$=sibship size of the i^{th} family, $b=\sum_{i=1}^{k} b_i$= number of brothers in the sample of *k* families, $s=\sum_{i=1}^{k} s_i$=number of sisters in the sample of *k* families, and N=b+s= number of siblings in the *k* families. It is clear then that each cluster (sib-ship) is naturally divided into two sub-clusters, one cluster represents brothers and the other sub-cluster represents sisters.

Let $y_{ij}=1(0)$ denote the presence (absence) of a trait observed on the j^{th} brother from the i^{th} family (j=1,2,…b_i;i=1,2,…k). Similarly, let $x_{ij}=1(0)$ denote the presence (absence) of this trait as observed on the j^{th} sister in the ith family (j=1,2,…s_i;i=1,2,…k). Let $\lambda_{ib}=p(y_{ij}=1\,|\,\lambda_{ib})$ denote the probability that a randomly selected brother from the i^{th} family is classified as having the trait of interest, and let $1-\lambda_{ib}=p(y_{ij}=0\,|\,\lambda_{ib})$. Moreover let $P(x_{ij}=1\,|\,\lambda_{is})=\lambda_{is}$, and $P(x_{ij}=0\,|\,\lambda_{is})=1-\lambda_{is}$. We initially assume that the distribution of the brothers' scores is conditionally independent of the distribution of the sisters' scores. To introduce the correlation among brothers within the ith family we shall assume that λ_{ib} is an element of a random sample obtained from a beta distribution with parameters (α_b,β_b) so that $\mu_b=E(\lambda_{ib})=\frac{\alpha_b}{\alpha_b+\beta_b}$, $Var(\lambda_{ib})=\frac{\alpha_b\beta_b}{(\alpha_b+\beta_b)^2(1+\alpha_b+\beta_b)}=\rho_b\mu_b(1-\mu_b)$,

Where $\rho_b=(1+\alpha_b+\beta_b)^{-1}$.

Similarly

$$\mu_s=E(\lambda_{is})=\frac{\alpha_s}{\alpha_s+\beta_s}\quad Var(\lambda_{is})=\frac{\alpha_s}{(\alpha_s+\beta_s)^2(1+\alpha_s+\beta_s)}=\rho_s\mu_s(1-\mu_s)\quad ,$$

where $\rho_s=(1+\alpha_s+\beta_s)^{-1}$.

We can show that the population common intraclass correlation among brothers in the samesub-cluster is:

$$\text{Corr}\,(y_{ij}, y_{ij}')=\rho_b,$$

and the common intraclass correlation among sisters in the other sub-cluster is:

$$\text{Corr}\,(x_{ij}, x_{ij}')=\rho_s$$

$i\neq j'=a,2,\ldots b_i$, and $m\neq l=1,2,\ldots s_i$ for all i=1,2,…k.

We further define the interclass correlation among brothers and sisters as:

Corr $(y_{ij}, x_{il})=\rho_{12}$ i=1,2,…k, j=1,2,…b_i and l = 1,2,…s_i.

Note that, because of the exchangeability condition, the unconditional distribution of $y_{ib}=\sum_{j=1}^{b_i} y_{ij}$ is that of a beta-binomial distribution with:

$$E(y_{ib})=b_i\mu_b \tag{1}$$

$$\sigma_{b_i}^2=Var(y_{ib})=b_i\mu_b(1-\mu_b)[1+(b_i-1)\rho_b] \tag{2}$$

Similarly, the unconditional distribution of $x_{is}=\sum_{j=1}^{s_i} x_{ij}$ will be that of a beta-binomial distribution with

$$E(x_{is})=s_i\mu_s \tag{3}$$

$$\sigma_{si}^2=Var(x_{is})=s_i\mu_s(1-\mu_s)[1+(s_i-1)\rho_s] \tag{4}$$

Details may be found in references [9-11].

The beta-binomial probability distributions of y_{ib} and x_{is} are given respectively as:

$$p(y_i)=\binom{b_i}{y_i}\frac{\prod_{j=o}^{y_i^*}(\mu_b+j\theta_b)\prod_{j=o}^{b_i^*-y_i}(1-\mu_b+j\theta_b)}{\prod_{j=0}^{b_i^*}(1+j\theta_b)}\quad y_i=0,1,2,\ldots b_i \tag{5}$$

$$p(x_i)=\binom{s_i}{x_i}\frac{\prod_{j=o}^{x_i^*}(\mu_s+j\theta_s)\prod_{j=o}^{s_i^*-y_i}(1-\mu_s+j\theta_s)}{\prod_{j=0}^{s_i^*}(1+j\theta_s)}\quad x_i=0,1,2,\ldots s_i \tag{6}$$

($\alpha^*=a-1$), $\vartheta_b=\rho_b/(1-\rho_b)$, with a similar transformation for $\vartheta_s=\rho_s/(1-\rho_s)$.

The above set-up assumes that the three sibling correlations ρ_b, ρ_s, and ρ_{bs} are constant among families in the parent population. With this assumption our main interest is in developing a likelihood-based approach for testing several hypotheses that are inherently related. We first construct a bivariate distribution based on the marginal distributions given in (5) and (6) which includes all the parameters of interest. However a different approach is needed to construct the bivariate distribution of the sibling scores characterized by the interclass correlation. This approach, developed by Sarmanov [12] and Lancaster [13], is known as Positive Dependence byExpansion (PDE). Danaher [14] proposed a simplified and flexible form of the distribution basedon Lancaster's representation.

We are interested in testing the following hypotheses:

1- $H_0: \rho_{12}=0$

2- $H_0: \rho_b=\rho_s$

As a first step, we follow Lancaster [13] in constructing a bivariate distribution by joining the marginal distributions given in (5) and (6). The resulting representation is given by:

$$p(x_i,y_i)=p(x_i)\cdot p(y_i)\left[1+\rho_{12}\left(\frac{y_i-b_i\mu_b}{\sigma_{b_i}}\right)\left(\frac{x_i-s_i\mu_s}{\sigma_{s_i}}\right)\right] \tag{7}$$

$y_i=0,1,2,\ldots b_i$, and $x_i=0,1,2,\ldots s_i$. Direct computations show that $\text{Corr}(x_i,y_i)=\rho_{12}$. The sum of theright hand-side of equation (7) over all the possible values of (x_i,y_i) is one. Therefore, the equationrepresents a proper bivariate probability distribution with parameters vector $\varphi=(\mu_b,\mu_s,\rho_b,\rho_s,\rho_{12})'$.

Methods

Our inferences on the parameters of interest are based on the

likelihood principle. To test the hypothesis $H_0:\rho_{12}=0$ against the alternative $H_1:\rho_{12}>0$, we assume that a random sample of k sib-ships is available. The log- likelihood function of the sample is given by:

$$l=\sum_{i=1}^{k}\log p(x_i,y_i)$$
$$=\sum_{i=1}^{k}\{\log p(x_i)+\log p(y_i)+\log[1+\rho_{12}H(x_i,y_i)]\},$$

Where, $H(x_i,y_i)=\left(\frac{y_i-b_i\mu_b}{\sigma_{b_i}}\right)\left(\frac{x_i-s_i\mu_s}{\sigma_{s_i}}\right)$. The score function $u=\frac{\partial l}{\partial \rho 12}|\,\rho 12=0$ is given by:

$$u=\sum_{i=1}^{k}(y_i-b_i\mu_b)(x_i-s_i\mu_s)(\sigma_{bi}\sigma_{si})^{-1}=\sum_{i=1}^{k}u_i$$

By the central limit theorem and for fixed b_i and s_i, the distribution of each component u_i tends uniformly to the standard normal distribution under H_0 as $k\to\infty$. Moreover, we can show that under H_0 that the statistic $s^2=u^2/k$ will be asymptotically distributed as chi-square with one degree of freedom [15]. This statistic is the locally most powerful one-sided test of $H_0:\rho_{12}=0$ against $H_1:\rho_{12}>0$. Full details of the proof can be found [15,16]. Moran [17] showed that if the remaining parameters are replaced by any consistent estimators under the null hypothesis the asymptotic properties of this test statistic will be preserved. Such consistent estimators can be either the maximum likelihood (MLE) or the moment estimators.

In Table 1 we provide estimates of the sample sizes (number of sib-sips) needed to detect thedeparture from $H_0:\rho_{12}=0$ in the direction of a two sided alternative under several scenarios. We limited our computations to the balanced case with equal response rates. It can be seen from Table 1 that when we have a small departure from the null hypothesis, a large sample is needed, regardless of the sib-ship sizes. Moreover, when the response rates are far from their boundary values (0, 1), a substantially smaller number of sib-ships are needed. Tables 2 and 3 present the empirical powers, when the design is balanced (number of brothers equals number of sisters within the same family), for $\mu_b=\mu_s=0.1$ and $\mu_b=\mu_s=0.5$, respectively. It is clear that the power increases with the increase in the number of sib-ships, and is unaffected by sib-ship sizes. Again, a noticeable increase in the power is achieved

				$\mu_b=\mu_s=.1$				$\mu_b=\mu_s=.5$		
ρ_b	ρ_s	ρ_{12}	b=s	2	5	10	b=s	2	5	10
.2	.2	.1		715	697	693		612	612	612
.2	.5	.1		715	613	613		612	612	612
.5	.5	.5		44	44	44		22	22	22
.8	.8	.5		45	41	41		22	22	22
.8	.8	.8		21	21	21		7	7	7

Table 1: Sample size requirements for Type I error rate 5% and power 80%. To test the hypothesis $H_0:\rho_{12}=0$.

				$\mu_b=\mu_s=.1$										
				k=25			k=50				k=100			
ρ_b	ρ_s	ρ_{12}	b=s	2	5	10	b=s	2	5	10	b=s	2	5	10
0	0	0		.051	.051	.051		.051	.051	.051		.051	.051	.051
.2	.2	0		.051	.051	.051		.051	.051	.051		.051	.051	.051
.5	.5	.2		.336	.334	.334		.440	.440	.440		.594	.595	.595
.8	.8	.5		.661	.661	.660		.820	.820	.820		.950	.950	.950
.8	.8	.8		.830	.830	.830		.950	.950	.950		.990	.990	.990

Table 2: Power Calculations for testing $H_0:\rho_{12}=0$.

				$\mu_b=\mu_s=.5$										
				k=25			k=50				k=100			
ρ_b	ρ_s	ρ_{12}	b=s	2	5	10	b=s	2	5	10	b=s	2	5	10
0	0	0		.051	.051	.051		.051	.051	.051		.051	.051	.051
.2	.2	0		.051	.051	.051		.051	.051	.051		.051	.051	.051
.5	.5	.2		.336	.334	.334		.489	.489	.489		.643	.643	.643
.8	.8	.5		.661	.661	.660		.820	.820	.820		.950	.950	.950
.8	.8	.8		.839	.839	.839		.985	.985	.985		.999	.999	.999

Table 3: Power Calculations for testing $H_0:\rho_{12}=0$.

ρ_b	ρ_s	ρ_{12}	b	s	k=25	k=50	k=100
.5	.5	.2	2	4	.257	.409	.643
.5	.5	.2	10	5	.258	.411	.645
.8	.8	.5	10	2	.841	.986	.999
.8	.8	.5	6	3	.840	.985	.999
.2	.2	.2	2	4	.260	.409	.650

Table 4: Power calculations for testing $H_0:\rho_{12}=0$, $\mu_b=\mu_s=.5$, Under unbalanced design.

when the response rates are far from their boundary values. In Table 4 we show that the effect of unbalanced design (number of brothers does not equal number of sisters within the same family) on the power is nottangible.

Maximum Likelihood Estimation

The MLE's of the model parameters are obtained by simultaneously solving the likelihood equations:

$$\frac{\partial l}{\partial \mu_b}=0,\frac{\partial l}{\partial \mu_s}=0,\frac{\partial l}{\partial \theta_b}=0,\frac{\partial l}{\partial \theta_s}=0,\frac{\partial l}{\partial \rho_{12}}=0$$

We obtain the variance-covariance matrix Σ of the MLE's by inverting the matrix of the negative of the second partial derivatives of ℓ with respect to the five parameters. The two matrices are given as:

$$I=\begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22}\end{bmatrix},\text{ and }\sum=I^{-1}=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}.$$

Here, Σ is the variance-covariance matrix of $(\hat{\mu}_b,\hat{\mu}_s,\hat{\theta}_b,\hat{\theta}_s,\hat{\rho}_{12})$. To find the elements of Σ we use the method of matrix partitioning [18].

The matrix I_{11} is a 2×2 and symmetric, $I_{12}=I'_{21}$ is a 2×3 matrix and I22 is a 3×3 diagonal matrix. The elements of the covariance matrix are given in closed form as:

$$\sum_{11}=(I_{11}-I_{12}I_{22}^{-1}\ I_{22})^{-1}$$
$$\sum_{22}=I_{22}^{-1}+I_{22}^{-1}I_{21}\ \Sigma_{11}I_{12}I_{22}^{-1}$$
$$\sum_{12}=(I_{22}^{-1}\ I_{12}\Sigma_{22})$$
$$\sum_{21}=(I_{22}^{-1}\ I_{21}\Sigma_{22})$$

In the Appendix we provide expressions for the elements of I and Σ.

Hypothesis testing

In this section we develop an approach for testing the effect of gender differences on sib-sib correlations. If the trait of interest is normally distributed, the sib means and the sib-sib correlations are orthogonal to each other, implying that the expected value of the second partial derivatives of the likelihood function with respect to the mean

and correlation is zero [19]. This orthogonality also implies that the maximum likelihood estimators of these parameters are asymptotically independent. Therefore, as in Donner et al. [8] we can test the equality of ρ_b and ρ_s independent of the values of μ_b and μ_s. For the bivariate beta-binomial distribution, the orthogonality condition is not satisfied and therefore we propose an omnibus test in the form:

$$H_o : \mu_b = \mu_s \cap \rho_b = \rho_s \quad (8)$$

This hypothesis takes into account the correlations among all the estimated parameters. Let $\psi = (\mu_b, \mu_s, \rho_b, \rho_s)'$ and consider the affine transformation.

H_0:Aψ=0 versus H_1:Aψ=δ>0

The matrix A has 2 rows and 4 columns and is specified as:

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

To test the stated hypothesis, an omnibus test statistic is constructed, using the asymptotic distributional properties of the MLE of ψ. From Serfling [20], the MLE $\hat{\psi}$ has the asymptotic distribution:

$$\sqrt{k}(\hat{\psi} - \psi) \vec{d} N(O, V) \ .$$

where V is the variance-covariance matrix of $\hat{\psi}$ and is obtained by deleting the 5th row and the 5th column of Σ. Letting H=Aψ, the question reduces to testing H_0: H=δ>0.

$\hat{H} = A\hat{\psi}$ is therefore distributed as $\hat{H} \vec{d} N(H, AVA^T)$.

Hence, from Graybill [21] the quadratic form:

$$Q(\in) = K\hat{H}^T (AV\,A^T)^{-1} \hat{H} \quad (9)$$

is asymptotically non-central chi-square with 2 degrees of freedom and non-centrality parameter

$\epsilon = KH^T(AV\,AT)^{-1}\,H$

Moreover, ϵ=0 if and only if H_0 is true. Hence referring Q(0) to the table of chi-square distribution with 2 degrees of freedom, H_0: H=0 is rejected if Q(0) exceeds the tabulated value of a chi-square with 2 degrees of freedom at the chosen level of significance.

Example: Mial and Oldham's blood pressure data

The data used for illustration here are obtained from a survey that aimed at assessing the levels of similarity in systolic and diastolic blood pressure among family members living within 25 miles of Rhonda Fach Valley in South Wales and published by Miall and Oldham [22] previously analyzed [23,24]. Observations were made on parents and their offspring, with each observation consisting of systolic and diastolic blood pressures measured to the nearest 5 mmHg. However among 250 sampled families, only 204 contained information on brothers and sisters. Furthermore, because of the impossibly low systolic blood pressure (15 mmHg) for one daughter, another family was omitted leaving 203 families for the analysis. Since these data weregiven on a continuous scale, we dichotomized the observations such that for an individual whose systolic/diastolic blood levels above 130/80, the assigned binary score was 1, otherwise, set as 0. The results of the data analysis are summarized in Tables 5 and 6. Table 5 shows the maximum likelihood of the model parameters, together with their standard errors. Table 6 displays the variances and covariances among the estimated parameters using the expression in the Appendix.

The null hypothesis H_0: ρ_{12}=0 tested against the one-sided alternative

Parameter	Estimate ± SE
μ_b	.294 ± 0.028
ρ_b	.200 ± 0.077
μ_s	.217 ± 0.026
ρ_s	.274 ± 0.071
ρ_{12}	.195 ± .095

Table 5: Estimates of the model of parameters ± standard error.

	$\hat{\mu}_b$	$\hat{\mu}_s$	$\hat{\rho}_b$	$\hat{\rho}_s$	$\hat{\rho}_{12}$
$\hat{\mu}_b$	$.8\times10^{-3}$	$-.15\times10^{-3}$	$.1\times10^{-3}$	0	$.19\times10^{-3}$
$\hat{\mu}_s$		$.7\times10^{-3}$	0	$.1\times10^{-3}$	$.13\times10^{-3}$
$\hat{\rho}_b$			$.6\times10^{-2}$	$.2\times10^{-3}$	$.83\times10^{-3}$
$\hat{\rho}_s$				$.5\times10^{-2}$	$.88\times10^{-3}$
$\hat{\rho}_{12}$					$.90\times10^{-2}$

Table 6: Variance-Covariance matrix of the estimates.

Parameter	Systolic	Diastolic
ρ_b	0.146 ± 0.073	0.163 ± .073
ρ_s	0.32 ± 0.069	0.248 ± .070
ρ_{12}	0.178 ± 0.054	0.215± .052

Table 7: Estimates of Correlation parameters for original data.

H_1: ρ_{12}>0 is rejected as s^2=4.48 (p-value=.034). The Wald one degree of freedom chi-square test *w*=(0.195/0.095)2=4.21, leading to the same conclusion as the score test. To test the null hypothesis (8) we use the statistic givenin (9). Direct computation shows that Q(0)=3.246, (p=0.197). Therefore, we conclude that there is no sufficient evidence to support the claim of gender differences in the distribution of hypertension basedon this data set. An equally important hypothesis to be tested is whether gender has influence on the dependence structure. That is we need to test whether within gender correlations are the same as across gender correlation. This hypothesis can be easily formulated as:

H_0:ρ_b=ρ_s=ρ_{12}. We may then formulate the simple contrast $\hat{T} = \hat{\rho}_b + \hat{\rho}_s - 2\hat{\rho}_{12}$. Under the null hypothesis, T is asymptotically unbiased, that is E(T)=0, and $\text{var}(\hat{T}) = \text{var}(\hat{\rho}_b) + \text{var}(\hat{\rho}_s) + 2\,\text{cov}(\hat{\rho}_b, \hat{\rho}_s) + 4\,\text{var}(\hat{\rho}_{12}) - 4\,\text{cov}(\hat{\rho}_b, \hat{\rho}_{12}) - 4\,\text{cov}(\hat{\rho}_s, \hat{\rho}_{12})$.

Therefore, asymptotically $G = \hat{T}^2 / \text{var}(\hat{T})$ has a chi-square distribution with one-degree of freedom. From the data, G=0.185, and a p-value=0.911. Therefore, based on this data we conclude that the correlations within gender are the same across genders.

Remarks

Originally the Miall and Oldham's data are measured on the continuous scale. Under the assumption of multivariate normality Mian and Shoukri [24] used the MLE to produce the following estimates for the within gender and across gender correlations. We summarize the results in Table 7. For testing H_0:ρ_b=ρ_s, a one degree of freedom chi-square test statistic is 61, with p-value<0.00001. Similar to the above approach, H_0:ρ_{12}=0 is rejected (p-value<0.00001). Similarly H_0:ρ_b=ρ_s=ρ_{12} has a chi-square value=26.48, with p-value<0.00001.

It is clear that the dichotomization resulted in a reduction in

the efficiency of the maximum likelihood estimates of the sibling correlations, which would result in a substantial loss of power of detecting departure from the null hypotheses of interest. This issue has been a subject of discussion by many authors [25,26]. Loss of power and sensitivity to the choice of the cut-off point are the price to pay due to discretization. However, selecting a cut-off point is not a matter of concern to statisticians but is based on clinical expertise. For example, components of what is known as metabolic syndrome (obesity, triglyceride, high density lipoprotein, blood pressures level, and blood glucose levels) are all measured on the continuous scale. However, communicating the clinical diagnoses of the components of the syndrome are based on the cutoff points recommended by the WHO, or the International Diabetes Federation. In this paper weused the WHO definition of hypertension 130/80 when we dichotomized the blood pressures data.

Discussion

Estimation of measures of family resemblance is considered the first step prior to investigating whether the variation in the distribution of the trait of interest is may be attributed to genetic factors. Similarly, detection of gender differences may be important to identify sex-linked traits. Establishing a statistical significance may provide the quantitative bases to study the distribution of the traits at the molecular level. The major contribution of this paper is the application of likelihood methods to a constructed bivariate beta-binomial distribution. This has allowed us to establish a score test for the goodness of fit of the model. A second finding is that testing for gender differences in the sib-sib correlations can be established in a relatively simple way, e.g. without computing the more complicated likelihood ratio test. Our limited scale computations showed that we need to sample a large number of families to retain reasonable power for the test statistic. Moreover, we showed that the power of the test of significance for the interclass correlation is quite insensitive to variations in the sub-cluster sizes. The implication is that as long as we have a sufficient number of families in the sample, the actual sib-ship sizes become less important. The model (7) is quite flexible. For example it easily allows for inclusion of covariates measured at the sub-cluster level. This can be done by employing a suitable transformation on the response probabilities similar to the case of a non-linear mixed model for binary responses. One important assumption of the present model is that it assumes that the correlation parameters are constant in the sampled population. This assumption may not be tenable in cases where some siblings are reared together and some reared apart. It should also be noted that there is a large number of statistical models used to fit clustered binary data, to name but a few, the Generalized Estimating Equations (GEE) which is a semi parametric approach, and the General Linear Mixed Models (GLIMMIX). These models are geared towards estimation of the regression coefficients, treating the correlation structure as nuisance. The application of the GEE can be problematic for the analysis of family data. In fact Crowder [27] demonstrated that the parameters involved in working correlation matrix are subject to "uncertainty of definition which can lead to a breakdown of the asymptotic properties of the estimators". On the other hand, the GLIMMIX does not readily produce estimates of correlations at each level of hierarchy. More seriously, there are some concerns regarding the approximation of the variance covariance matrix of the estimated parameters. In genetic epidemiology, clustering of traits is usually measured by a set of familial correlations, and such correlations become the population target parameters of interest. The model developed in this paper is constructed to address these issues. Finally, it should be noted however, that further research is needed to investigate the asymptotic properties of the test statistics that we developed when the sub-cluster sizes are much larger than the number of clusters, a problem of common occurrence in community-based studies.

Acknowledgement

The authors are thankful to the constructive comments made by two anonymous reviewers.

References

1. Park HS, Park JY, Cho SI (2006) Familial aggregation of the metabolic syndrome in Korean families with adolescents. Atherosclerosis 186: 215-221.
2. Mager DL (2006) Bacteria and cancer: cause, coincidence or cure? A review. J Transl Med 4: 14.
3. Mosconi L, Rinne JO, Tsui WH, Berti V, Li Y, et al. (2010) Increased fibrillar amyloid-{beta} burden in normal individuals with a family history of late-onset Alzheimer's. Proc Natl Acad Sci U S A 107: 5949-5954.
4. Meisinger C, Thorand B, Schneider A, Stieber J, Döring A, et al. (2002) Sex differences in risk factors for incident type 2 diabetes mellitus: the MONICA Augsburg cohort study. Arch Intern Med 162: 82-89.
5. Grant J, Hicks N, Taylor A, Chittleborough C, Phillips P (2009) Gender specific epidemiology of diabetes: a representative cross-sectional study. International Journal for Equity in Health 8: 10.1186/1475-9276-8-6.
6. Roberts DF, Billewicz WZ, McGregor IA (1978) Heritability of stature in a West African population. Ann Hum Genet 42: 15-24.
7. Martorell R, Yarbrough C, Himes JH, Klein RE (1978) Sibling similarities in number of ossification centers of the hand and wrist in a malnourished population. Hum Biol 50: 73-81.
8. Donner A, Koval JJ, Bull S (1984) Testing the effect of sex differences on sib-sib correlations. Biometrics 40: 349-356.
9. Cox DR, Snell E (1989) Analysis of Binary Data. Chapman & Hall, London, UK.
10. Prentice RL (1988) Correlated binary regression with covariates specific to each binary observation. Biometrics 44: 1033-1048.
11. Crowder M (1978) Beta- binomial ANOVA for proportions. Applied Statistics 27: 34-37.
12. Sarmanov OV (1966) Generalized normal correlation and two-dimensional Frechet classes. Doklady (Soviet Mathematics) 168: 596-599.
13. Lancaster HO (1969) The chi-squared distribution. Wiley, New York.
14. Danaher PJ (1991) A canonical expansion model for multivariate media exposure distributions: generalization of the duplication of viewing law. Journal of Marketing Research 28: 361-367.
15. Koziol J (1979) A smooth test for bivariate independence. Sankhya: The Indian Journal of Statistics, Volume 41: 260-269.
16. Neyman J (1959) Optimal asymptotic tests of composite hypotheses. In: Grenander V (ed) Probability and Statistics: The Harold Cramer Volume 213-234, Wiley: New York.
17. Moran (1970) PAP "On Asymptotically Optimal Tests of Composite Statistical Hypotheses". Biometrika 57: 47-55.
18. Graybill FA (1983) Matrices with applications in statistics. (2nd edn), Duxbury Classic Series.
19. Cox DR, Reid N (1987) Parameter orthogonality and approximate conditional a. inference (with discussion). J Roy Statist Soc Ser B 49: 1-39.
20. Serfling RJ (1983) Approximation Theory of Mathematical Statistics. Wiley, New York.
21. Graybill F (1976) Theory and Application of the Linear Models. Duxbury, North Scituate, Massachusetts.
22. Miall WE, Oldham PD (1955) A study of arterial blood pressure and its inheritance in a sample of the general population. ClinSci (Lond) 14: 459-488.
23. Shoukri MM, ElDali A, Donner A (2012) Measures of family resemblance for

binary traits: likelihood based inference. Int J Biostat 8: 20.

24. Mian IU, Shoukri MM (1997) Statistical analysis of intraclass correlations from multiple samples with applications to arterial blood pressure data. Stat Med 16: 1497-1514.

25. Royston P, Altman DG, Sauerbrei W (2006) Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 25: 127-141.

26. http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/CatContinuous

27. Crowder M (1995) On the use of a working correlation matrix in using generalized linear models for repeated measures. Biometrika 82: 407-410.

APPENDIX: Elements of the variance-covariance matrix of the MLE 's of the parameters of the bivariate-beta binomial distribution.

The following are the negative of the second partial derivatives of the log-likelihood function with respect to the five parameters:

$$i_{\mu_b^2} = \sum_{i=1}^{k}\left[\sum_{j=0}^{y_i^*}(\mu_b + j\theta_b)^{-2} + \sum_{j=0}^{b_i^*-y_i}(1-\mu_b + j\theta_b)^{-2}\right]$$

$$+\sum_{i=1}^{k} H^2(y_i, x_i) b_i^2 w_i^2\,(x_i - s_i\mu_s)^2$$

where $H_i^{-1}(\rho_{12}) = 1 + \rho_{12}\left(\frac{y_i - b_i\mu_b}{\sigma_{b_i}}\right)\left(\frac{x_i - s_i\mu_s}{\sigma_{s_i}}\right)$

$$i_{\mu_{bs}} = -\sum_{i=1}^{k} H_i^2(\rho_{12}) b_i s_i \left(\sigma_{b_i}\sigma_{s_i}\right)^{-1}\rho_{12}$$

$$i_{\mu_b\theta_b} = \sum_{i=1}^{k}\left[\sum_{j=0}^{y_i^*}(\mu_b + j\theta_b)^{-2} + \sum_{j=0}^{b_i^*-y_i} j(1-\mu_b + j\theta_b)^{-2}\right]$$

$$i_{\mu_b\rho_{12}} = \sum_{i=1}^{k}\left[H_i^2(\rho_{12}) b_i (x_i - s_i\mu_s)\left(\sigma_{b_i}\sigma_{s_i}\right)^{-1}\right]$$

$$i_{\rho_{12}^2} = \sum_{i=1}^{k}\left[H_i(\rho_{12})\left(\frac{y_i - b_i\mu_b}{\sigma_{b_i}}\right)\left(\frac{x_i - s_i\mu_s}{\sigma_s}\right)\right]^2$$

$$i_{\theta_b^2} = \sum_{i=1}^{k}\left[\sum_{j=0}^{y_i^*} j^2\,(\mu_b + j\theta_b)^{-2} + \sum_{j=0}^{b_i^*-y_i} j^2\,(1-\mu_b + j\theta_b)^{-2} - \sum_{j=0}^{b_i^*} j^2\,(1 + j\theta_b)^{-2}\right]$$

$$\frac{\partial^2 l}{\partial\mu_s\partial\theta_b} = \frac{\partial^2 l}{\partial\mu_b\partial\theta_s} = \frac{\partial^2 l}{\partial\theta_b\partial\theta_s} = \frac{\partial^2 l}{\partial\theta_b\partial\rho_{12}} = \frac{\partial^2 l}{\partial\theta_s\partial\rho_{12}} = 0$$

The other second partial derivations are obtained by symmetry, on replacing (b, y) with (s, x)

Elements of the information matrix are:

$$I = \begin{bmatrix} i_{\mu_b^2} & i_{\mu_{bs}} & i_{\mu_b\theta_b} & 0 & i_{\mu_b\rho_{12}} \\ & i_{\mu_s^2} & 0 & i_{\mu_s\theta_s} & i_{\mu_s\rho_{12}} \\ & & i_{\theta_b^2} & 0 & 0 \\ & & & i_{\theta_s^2} & 0 \\ & & & 0 & i_{\rho_{12}^2} \end{bmatrix}$$

When the matrix is partitioned:

$$I_{11} = \begin{bmatrix} i_{\mu_b^2} & i_{\mu_{bs}} \\ i_{\mu_{sb}} & i_{\mu_s^2} \end{bmatrix}$$

$$I_{12} = \begin{bmatrix} i_{\mu_b\theta_b} & 0 & i_{\mu_b\rho_{12}} \\ 0 & i_{\mu_s\theta_s} & i_{\mu_s\rho_{12}} \end{bmatrix} = I_{21}^{'}$$

$$I_{22} = \begin{bmatrix} i_{\theta_b^2} & 0 & 0 \\ 0 & i_{\theta_s^2} & 0 \\ 0 & 0 & i_{\rho_{12}^2} \end{bmatrix}$$

The variance covariance matrix is given in a partitioned form.

$$I_{11}^{-1} = \Delta^{-1} \begin{bmatrix} i_{\mu_s^2} & -i_{\mu_{bs}} \\ -i_{\mu_{sb}} & i_{\mu_b^2} \end{bmatrix}$$

Where $\Delta = i_{\mu_b^2} i_{\mu_s^2} - \left(i_{\mu_{sb}}\right)^2$.

$$I_{22}^{-1} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

Where

$$\lambda_1^{-1} = i_{\theta_b^2}, \quad \lambda_2^{-1} = i_{\theta_s^2}, \quad \lambda_3^{-1} = i_{\rho_{12}^2}$$

$$\Sigma_{11} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

Where

$$a_{11} = \Delta_{11}^{-1} \left(i_{\mu_s^2} - \lambda_2 i_{\mu_s\theta_s} - \lambda_3 i_{\mu_s\rho_{12}}^2\right)$$

$$a_{12} = a_{21} = \Delta_{11}^{-1} \left(\lambda_3 i_{u_b\rho_{12}} i_{\mu_s}\rho_{12} - i_{\mu_{bs}}\right)$$

$$a_{22} = \Delta_{11}^{-1} \left(i_{\mu_b^2} - \lambda_1 i_{\mu_b\theta_b}^2 - \lambda_3 i_{\mu_b\rho_{12}}^2\right)$$

and, $\Delta_{11} = \left(i_{\mu_s^2} - \lambda_2 i_{\mu_s\theta_s}^2 - \lambda_3 i_{\mu_s\rho_{12}}^2\right)\left(i_{\mu_b^2} - \lambda_1 i_{\mu_b\theta_b}^2 - \lambda_3 i_{\mu_b\rho_{12}}^2\right) - \left(\lambda_3 i_{\mu_b\rho_{12}} i_{\mu_s\rho_{12}} - i_{\mu_{bs}}\right)^2$

Note that, since $\rho_b = \frac{\theta_b}{1-\theta_b}$, the variance of $\hat{\rho}$ is obtained by the delta method as

$$\text{var}(\hat{\rho}_b) = \left(1 - \hat{\theta}_b\right)^{-2} \text{var}\left(\hat{\theta}_b\right)$$

where$\hat{\theta}$ is the MLE of θ, and var$\left(\hat{\theta}\right)$ is the asymptotic variance of $\hat{\theta}$ obtained from the inverted observation matrix.