# Fitting a bivariate normal distribution to a two-way contingency table using maximum likelihood estimation

Gretel Crafford*
University of Pretoria, Pretoria, South Africa - gretel.crafford@up.ac.za

Nico Crowther
University of Pretoria, Pretoria, South Africa - nico.crowther@up.ac.za

**Abstract**

The historic height data set (1885) of Sir Francis Galton which considered the relationship between the height of parents and their children is revisited. The data set consists of 928 cases and is categorised in a two-way contingency table. Although the data is only available in a grouped format, an underlying bivariate normal distribution is evident. Employing an iterative maximum likelihood (ML) procedure, the estimated bivariate cumulative relative frequencies are fitted to follow a cumulative bivariate normal distribution curve at the intersections of the upper class boundaries of the two-way contingency table. These estimates are referred to as the ML estimates under constraints of the bivariate normal probabilities. The five parameters of the bivariate normal distribution, namely the mean and variance of the two variables and the correlation coefficient can now be solved uniquely from these ML estimates under constraints. In the iterative procedure use is made of the fact that the marginal distributions are normally distributed and that the correlation coefficient can be uniquely expressed as a function of the total probability of the positive quadrant of the bivariate normal distribution. The conditions of the marginal and joint distributions are incorporated simultaneously in the ML estimation procedure. By obtaining the ML estimates of the underlying bivariate normal distribution the complete relationship between the two variables can be investigated. The slope of the estimated regression line is 0.63 which corresponds closely to that of Galton, who suggested a slope of two thirds.

**Keywords:** bivariate grouped data; maximum likelihood estimates under constraints.

## 1. Introduction

In the case where two grouped response variables are jointly normally distributed, it is often required to explain the relationship between these variables. Although the data is presented in a two-way contingency table and the data is now only available in a grouped format, it will be shown how to estimate the underlying bivariate normal distribution. The estimation of the bivariate normal distribution reveals the complete underlying continuous structure between the two variables.

## 2. Galton's height data

To illustrate how to fit a bivariate normal distribution to a two-way contingency table the height data of Sir Francis Galton is considered. Galton wanted to investigate the relationship between the height of the parent and the child. The original data is regrouped and presented in Table1.

**Table 1:** Sir Francis Galton's height data

| **PARENT** | **CHILD** $(y)$ | | | | | |
|---|---|---|---|---|---|---|
| $(x)$ | $(-\infty, 65]$ | $(65, 67]$ | $(67, 69]$ | $(69, 71]$ | $(71, \infty]$ | **Total** |
| $(-\infty, 66]$ | 32 | 27 | 26 | 15 | 3 | **103** |
| $(66, 68]$ | 33 | 70 | 97 | 74 | 15 | **289** |
| $(68, 69]$ | 19 | 41 | 65 | 69 | 25 | **219** |
| $(69, 70]$ | 17 | 21 | 47 | 58 | 40 | **183** |
| $(70, \infty]$ | 2 | 6 | 23 | 50 | 53 | **134** |
| **Total** | **103** | **165** | **258** | **266** | **136** | **928** |

## 3. Formulation

The vectors of upper class boundaries are

$$x = \begin{pmatrix} 66 \\ 68 \\ 69 \\ 70 \end{pmatrix} \quad \text{and} \quad y = \begin{pmatrix} 65 \\ 67 \\ 69 \\ 71 \end{pmatrix}.$$

It is assumed that the random vector of frequencies $\mathbf{f}$ has a multinomial distribution. The expected value and covariance matrix of the random vector of cumulative relative frequencies $\mathbf{p}$ is denoted by

$$\mathrm{E}(\mathbf{p}) = \boldsymbol{\pi} \quad \text{and} \quad \mathrm{Cov}(\mathbf{p}) = \mathbf{V}$$

## 4. Estimation

In order to fit a bivariate normal distribution the ML estimates of $\boldsymbol{\pi}$ will be fitted such that the cumulative relative frequencies $\pi_{ij}$ equal the bivariate normal probabilities $\phi_{ij} = P(X \leq x_i, Y \leq y_j)$ i.e.

$$\begin{pmatrix} \pi_{11} & \pi_{12} & \pi_{13} & \pi_{14} & \pi_{15} \\ \pi_{21} & \pi_{22} & \pi_{23} & \pi_{24} & \pi_{25} \\ \pi_{31} & \pi_{32} & \pi_{33} & \pi_{34} & \pi_{35} \\ \pi_{41} & \pi_{42} & \pi_{43} & \pi_{44} & \pi_{45} \\ \pi_{51} & \pi_{52} & \pi_{53} & \pi_{54} & \pi_{55} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} & \phi_{14} & \phi_{15} \\ \phi_{21} & \phi_{22} & \phi_{23} & \phi_{24} & \phi_{25} \\ \phi_{31} & \phi_{32} & \phi_{33} & \phi_{34} & \phi_{35} \\ \phi_{41} & \phi_{42} & \phi_{43} & \phi_{44} & \phi_{45} \\ \phi_{51} & \phi_{52} & \phi_{53} & \phi_{54} & \phi_{55} \end{pmatrix}$$

This will be done by making use of the maximum likelihood (ML) estimation procedure (Matthews and Crowther (1995)) given in the following theorem.

**Theorem 1 (ML estimation procedure)**
Consider a random vector of cumulative relative frequencies $\mathbf{p}$, belonging to the exponential family with

$$E(\mathbf{p}) = \boldsymbol{\pi} \quad \text{and} \quad \mathrm{Cov}(\mathbf{p}) = \mathbf{V}$$

The ML estimate of $\boldsymbol{\pi}$ under the constraints $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$ is obtained iteratively from

$$\hat{\boldsymbol{\pi}} = \mathbf{p} - (\mathbf{G}_\pi \mathbf{V})' (\mathbf{G}_p \mathbf{V} \mathbf{G}_\pi)^* \mathbf{g}(\mathbf{p}) \tag{1}$$

where $\mathbf{G}_\pi = \dfrac{\partial \mathbf{g}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}$ and $\mathbf{G}_p = \left.\dfrac{\partial \mathbf{g}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}\right|_{\pi=p}$ and $(\mathbf{G}_p \mathbf{V} \mathbf{G}_\pi)^*$ is a generalised inverse of $(\mathbf{G}_p \mathbf{V} \mathbf{G}_\pi)$.

□

The iterative procedure implies a double iteration over $\mathbf{p}$ and $\boldsymbol{\pi}$. The procedure starts with the observed or unrestricted ML estimate of $\boldsymbol{\pi}$, as the starting value for both $\mathbf{p}$ and $\boldsymbol{\pi}$. Convergence is first obtained over $\mathbf{p}$ using (1). The converged value of $\mathbf{p}$ is then used as the next value of $\boldsymbol{\pi}$, with convergence over $\mathbf{p}$ starting again at the observed $\mathbf{p}$. In this procedure $\mathbf{V}$ is recalculated for each new value of $\boldsymbol{\pi}$ in the iterative procedure. Convergence over $\boldsymbol{\pi}$ ultimately leads to $\widehat{\boldsymbol{\pi}}$, the ML estimate of $\boldsymbol{\pi}$ under constraints.

The vector of constraints, $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$, to fit the bivariate normal distribution will address simultaneuously the marginal and partial distributions of $x$ and $y$. The vector of constraints, $\mathbf{g}(\boldsymbol{\pi}) = \mathbf{0}$ is therefore implemented as

$$\mathbf{g}(\boldsymbol{\pi}) = \begin{pmatrix} \mathbf{g}_x(\pi) \\ \mathbf{g}_y(\pi) \\ \mathbf{g}_{xy}(\pi) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\pi}_x \\ \boldsymbol{\pi}_y \\ \boldsymbol{\pi}_{xy} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\Phi}_x \\ \boldsymbol{\Phi}_y \\ \boldsymbol{\Phi}_{xy} \end{pmatrix} \tag{2}$$

in the ML estimation procedure and will be explained according to the marginal and partial distributions of $x$ and $y$ in the next two sections.

**5. Marginal distributions of** $x$ and $y$

Since the cumulative relative frequencies of $x$ has to follow a cumulative normal distribution curve at the standardised upper class boundaries $\mathbf{z}_x$ it follows that

$$\begin{array}{ccc} \boldsymbol{\pi}_x & = & \boldsymbol{\Phi}_x \\ \begin{pmatrix} \pi_{15} \\ \pi_{25} \\ \pi_{35} \\ \pi_{45} \end{pmatrix} & = & \begin{pmatrix} \phi_{15} \\ \phi_{25} \\ \phi_{35} \\ \phi_{45} \end{pmatrix} \end{array}$$

and therefore

$$\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi}_x) = \mathbf{z}_x = \mathbf{X}\boldsymbol{\alpha}_x$$

where $\mathbf{X} = \begin{pmatrix} \mathbf{x} & \mathbf{-1} \end{pmatrix}$ and $\boldsymbol{\alpha}_x = \begin{pmatrix} \frac{1}{\sigma_x} \\ \frac{\mu_x}{\sigma_x} \end{pmatrix}$ may be considered as the vector of natural parameters. Since

$$\boldsymbol{\alpha}_x = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi}_x)$$

it follows that the vector of standardised upper class boundaries of $\mathbf{x}$ namely

$$\mathbf{z}_x = \mathbf{P}_X\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi}_x), \quad \mathbf{P}_X = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$$

is the projection of $\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi}_x)$ on the vector space generated by the columns of $\mathbf{X}$.
Similarly, since the cumulative relative frequencies of $y$ has to follow a cumulative normal distribution curve at the standardised upper class boundaries $\mathbf{z}_y$ it follows that

$$\begin{array}{ccc} \boldsymbol{\pi}_y & = & \boldsymbol{\Phi}_y \\ \begin{pmatrix} \pi_{51} & \pi_{52} & \pi_{53} & \pi_{54} \end{pmatrix}' & = & \begin{pmatrix} \phi_{51} & \phi_{52} & \phi_{53} & \phi_{54} \end{pmatrix}' \end{array}$$

and therfore

$$\mathbf{z}_y = \mathbf{P}_Y\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi}_y), \quad \mathbf{P}_Y = \mathbf{Y}\left(\mathbf{Y}'\mathbf{Y}\right)^{-1}\mathbf{Y}'$$

is the projection of $\boldsymbol{\Phi}^{-1}(\boldsymbol{\pi}_y)$ on the vector space generated by the columns of $\mathbf{Y} = \begin{pmatrix} \mathbf{y} & \mathbf{-1} \end{pmatrix}$.

## 6. Joint distribution of $x$ and $y$

The domain of the standard normal distribution

$$f(z_x, z_y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[z_x^2 - 2\rho z_x z_y + z_y^2\right]\right\}$$

is divided in 4 so-called quadrants by the lines $z_x = 0$ and $z_y = 0$. The corresponding volumes are indicated in Table 2.

**Table 2:** The four quadrants of the bivariate normal distribution

| Quadrant | Region | Volume |
|---|---|---|
| $Q_1$ | $-\infty < z_x < 0 \quad -\infty < z_y < 0$ | $\text{VOL1} = \iint_{Q_1} f(z_x, z_y)\, dz_x dz_y$ |
| $Q_2$ | $-\infty < z_x < 0 \quad 0 \le z_y < \infty$ | $\text{VOL2} = \iint_{Q_2} f(z_x, z_y)\, dz_x dz_y$ |
| $Q_3$ | $0 \le z_x < \infty \quad -\infty < z_y < 0$ | $\text{VOL3} = \iint_{Q_3} f(z_x, z_y)\, dz_x dz_y$ |
| $Q_4$ | $0 \le z_x < \infty \quad 0 \le z_y < \infty$ | $\text{VOL4} = \iint_{Q_4} f(z_x, z_y)\, dz_x dz_y$ |

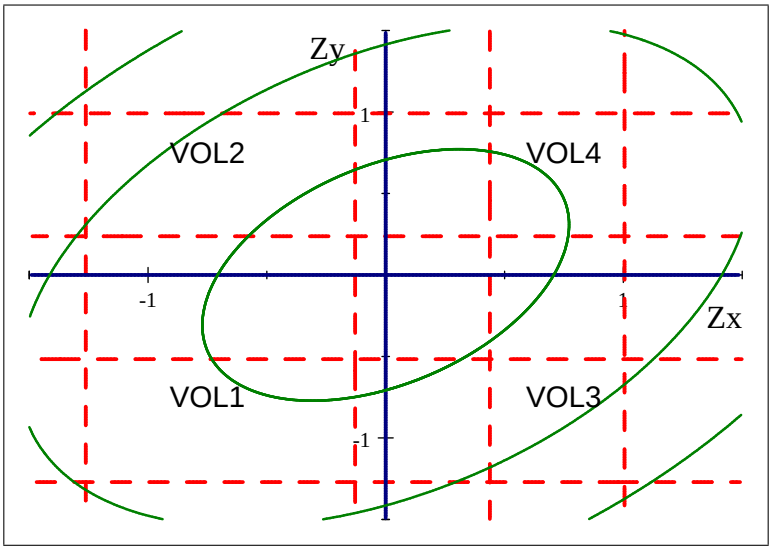
From *Sheppard's theorem on median dichotomy (1898)* (Kendall and Stuart (1958)) it follows that the probability or the total volume of the positive quadrant $Q_4$ may be expressed in terms of the correlation coefficient

$$\frac{\arcsin\rho}{2\pi} = \text{VOL}\,4 - \frac{1}{4}$$

From the symmetry of the bivariate normal distribution it follows that

$$\rho = \sin\left(\frac{\pi}{2}\left[(\text{VOL}\,1 + \text{VOL}\,4) - (\text{VOL}\,2 + \text{VOL}\,3)\right]\right) \tag{3}$$

**Figure 1:** The correlation coefficient can be expressed in terms of the positive quadrant




It now follows that

$$\boldsymbol{\pi}_{xy} = \boldsymbol{\Phi}_{xy}$$

$$\text{vec}\begin{pmatrix} \pi_{11} & \pi_{12} & \pi_{13} & \pi_{14} \\ \pi_{21} & \pi_{22} & \pi_{23} & \pi_{24} \\ \pi_{31} & \pi_{32} & \pi_{33} & \pi_{34} \\ \pi_{41} & \pi_{42} & \pi_{43} & \pi_{44} \end{pmatrix} = \text{vec}\begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} & \phi_{14} \\ \phi_{21} & \phi_{22} & \phi_{23} & \phi_{24} \\ \phi_{31} & \phi_{32} & \phi_{33} & \phi_{34} \\ \phi_{41} & \phi_{42} & \phi_{43} & \phi_{44} \end{pmatrix}$$

where $\phi_{ij}$, for $i,\ j = 1,2,3,4$ are the standardised bivariate normal probabilities with correlation coefficient $\rho$ determining the shape of the bivariate normal distribution. The value of the correlation coefficient $\rho$ may be uniquely estimated from the relative frequencies by repeating

$$\widehat{\rho} = \sin\left(\frac{\pi}{2}\left[\left(\widehat{\text{VOL1}} + \widehat{\text{VOL4}}\right) - \left(\widehat{\text{VOL2}} + \widehat{\text{VOL3}}\right)\right]\right)$$

until convergence, starting from, say $\widehat{\rho} = 0$.

## 7. Results

Employing the iterative ML estimation procedure, the ML estimate under constraints, $\hat{\boldsymbol{\pi}}$ is obtained and can be given in matrix form as

$$\widehat{\boldsymbol{\Pi}} = \begin{pmatrix} 0.027496 & 0.0588046 & 0.0852946 & 0.097142 & 0.100295 \\ 0.073349 & 0.1920246 & 0.3278041 & 0.4115897 & 0.444574 \\ 0.08886 & 0.2495666 & 0.4551269 & 0.5992973 & 0.666717 \\ 0.096148 & 0.2817205 & 0.5376930 & 0.7351208 & 0.841613 \\ 0.09912 & 0.2981769 & 0.5898979 & 0.8374953 & 1 \end{pmatrix}$$

The expected frequencies with an underlying bivariate normal distribution is now

$$\mathbf{M} = \begin{pmatrix} 25.52 & 29.05 & 24.58 & 10.99 & 2.93 \\ 42.55 & 81.08 & 101.42 & 66.76 & 27.68 \\ 14.39 & 39.00 & 64.76 & 56.04 & 31.96 \\ 6.76 & 23.08 & 46.78 & 49.42 & 36.26 \\ 2.76 & 12.51 & 33.17 & 46.56 & 51.98 \end{pmatrix}$$

and can be compared with the observed frequencies in Table 1. The ML estimates of the first four parameters of the bivariate normal distribution namely $\mu_x$, $\sigma_x$, $\mu_y$, and $\sigma_y$ can be found in Table 3.

**Table 3:** The ML estimates under constraints of the marginal distributions

| $\mathbf{x}$ | $\hat{\boldsymbol{\pi}}_x$ | $\boldsymbol{\Phi}^{-1}(\hat{\boldsymbol{\pi}}_x)$ | $\hat{\boldsymbol{z}}_x = \boldsymbol{P}_X\boldsymbol{\Phi}^{-1}(\hat{\boldsymbol{\pi}}_x)$ | $\begin{pmatrix}\hat{\mu}_x \\ \hat{\sigma}_x\end{pmatrix}$ |
|---|---|---|---|---|
| $\begin{pmatrix}66\\68\\69\\70\end{pmatrix}$ | $\begin{pmatrix}0.100295\\0.444574\\0.666717\\0.841613\end{pmatrix}$ | $\begin{pmatrix}-1.27987\\-0.13938\\0.430865\\1.001111\end{pmatrix}$ | $\begin{pmatrix}-1.27987\\-0.13938\\0.430865\\1.001111\end{pmatrix}$ | $\begin{pmatrix}68.244\\1.754\end{pmatrix}$ |
| | | | | |
| $\mathbf{y}$ | $\hat{\boldsymbol{\pi}}_y$ | $\boldsymbol{\Phi}^{-1}(\hat{\boldsymbol{\pi}}_y)$ | $\hat{\mathbf{z}}_y = \mathbf{P}_Y\boldsymbol{\Phi}^{-1}(\hat{\boldsymbol{\pi}}_y)$ | $\begin{pmatrix}\hat{\mu}_y \\ \hat{\sigma}_y\end{pmatrix}$ |
| $\begin{pmatrix}65\\67\\69\\70\end{pmatrix}$ | $\begin{pmatrix}0.099\,12\\0.298\,18\\0.589\,90\\0.837\,50\end{pmatrix}$ | $\begin{pmatrix}-1.28659\\-0.52965\\0.227282\\0.984216\end{pmatrix}$ | $\begin{pmatrix}-1.28659\\-0.52965\\0.227282\\0.984216\end{pmatrix}$ | $\begin{pmatrix}68.399\\2.642\end{pmatrix}$ |

From Table 3 it is clear that the ML estimates under constraints $\hat{\boldsymbol{\pi}}_x$ follow a standardised normal distribution curve at $\mathbf{x}$ and therefore $\boldsymbol{\Phi}^{-1}(\hat{\boldsymbol{\pi}}_x)$ is already in the vector space generated by the columns of $\mathbf{X}$. The same conclusion can be made with regard to $\hat{\boldsymbol{\pi}}_y$.

The estimated 4 volumes are given in Table 4 and it is clear that the property of symmetry is satisfied.

**Table 4:** The ML estimates of the 4 volumes

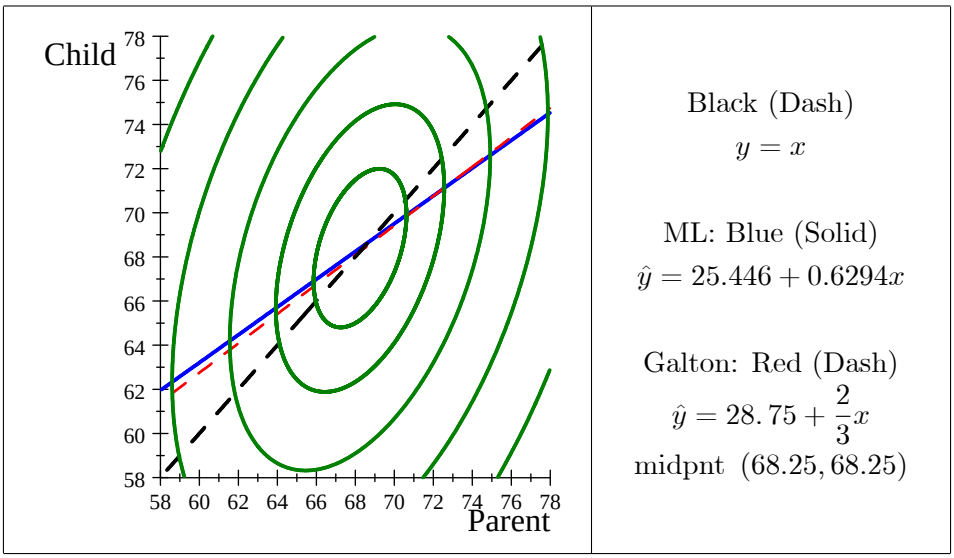
| **Quadrant** | $\widehat{VOL}$ |
|---|---|
| $Q_1 : z_x < 0, z_y < 0$ | $\widehat{VOL1} = 0.318588$ |
| $Q_2 : z_x < 0, z_y > 0$ | $\widehat{VOL2} = 0.181412$ |
| $Q_3 : z_x > 0, z_y < 0$ | $\widehat{VOL3} = 0.181412$ |
| $Q_4 : z_x > 0, z_y > 0$ | $\widehat{VOL4} = 0.318588$ |

The estimated correlation coefficient is

$$\hat{\rho} = sin\left(\frac{\pi}{2}\left[\left(\widehat{VOL1} + \widehat{VOL4}\right) - \left(\widehat{VOL2} + \widehat{VOL3}\right)\right]\right) = 0.4177351$$

yielding the regression equation $\hat{y}_{y|x} = 25.446 + 0.629\,41x$. The results obtained by making use of the ML estimation technique compares very well with that of Francis Galton. See Figure 2.

**Figure 2:** Regression equations of Galton and the ML estimation procedure




## 8. Conclusions

By employing the iterative ML estimation procedure it is possible to fit a bivariate normal distribution to data summarised in a two-way contingency table with an underlying bivariate normal distribution. This will enable us to explain the relationship between the two variables effectively.

## References


Galton F. (1886). Regression towards mediocrity in hereditary stature. Journal of the Anrhropological Institute, 15 pp.246-63.

Kendall, M.G., & Stuart, A. (1958) The advanced theory of statistics. Griffin.

Matthews G. B., & Crowther, N.A.S. (1995). A maximum likelihood estimation procedure when modelling in terms of constraints. South African Statistical Journal, 31, pp.161-184.

Crowther N.A.S., Crafford G., & Matthews G.B. (2006). Fitting distributions to grouped data using a maximum likelihood approach. South African Statistical Journal, 40 (1), pp.99-122.

Crafford G., & Crowther N.A.S. (2009). Linear models for grouped data. South African Statistical Journal, 43 (2), pp.151-176.