*STATISTICS IN TRANSITION-new series, Spring 2013*
*Vol. 14, No. 1, pp. 129–138*

# A PROBABILISTIC SCHEME WITH UNIFORM CORRELATION STRUCTURE

**Raffaella Calabrese**[1]

## ABSTRACT

The probabilistic schemes with independence between the trials show different dispersion characteristics depending on the behaviour of the probabilities of the binary event in the trials. This work proposes a probabilistic scheme with uniform correlation structure that leads to different dispersion characteristics depending on the sign of the linear correlation. Finally, a hypothesis test is proposed to identify the type of the dispersion of the probabilistic scheme.

**Key words:** probabilistic scheme, uniform correlation, binary event.

## 1. Introduction

Binary events clustered into groups are analysed by the probabilistic schemes (Feller, 1968, p.146). Under the assumption of the independence between the trials, by changing the characteristics of the probabilities of the binary events the Bernoulli, Poisson, Lexis and Coolidge probabilistic schemes (Kendall, 1994, p.164) are defined. In this paper the above-mentioned schemes are analysed by highlighting how the different characteristic of the probabilities of the binary events lead to different dispersion properties. By removing the assumption of the independence of the trials, a probabilistic scheme with uniform correlation structure is proposed in this paper.

Analogously to the previous schemes, the dispersion of the proposed scheme can be normal, subnormal and supernormal, depending on whether the correlation is zero, negative or positive, respectively. Finally, a hypothesis test is proposed to verify the assumption of binomial dispersion.

The present paper is organized as follows. In the next section the probabilistic schemes with independence between the trials is analysed. In the following section a probabilistic scheme with unform correlation is proposed. Section

[1] Dynamic Labs Geary Institute University College Dublin.
E-mail: raffaella.calabrese@ucd.ie.

4 suggests a hypothesis test to identify the kind of dispersion of a probabilistic scheme. Finally, the last section contains some concluding remarks.

## 2. The probabilistic schemes with independence between the trials

Let us assume to be interested in attaining an event $A$ (success) in $k$ series of $n_j$ trials each with $j = 1, 2, \ldots, k$. For the subsequent results the assumption of independence between both the $k$ series and the $n_j$ trials of each series will be essential. Thus, let $A_{ji}$ be the Bernoulli random variable associated to the $i$-th trial of the $j$-th series, with $i = 1, 2, ..., n_j$ and $j = 1, 2, ..., k$

$$A_{ji} = \begin{cases} 1 & \text{the event } A \text{ occurs in the } i\text{-th trial of the } j\text{-th series} \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

having the following success and failure probabilities

$$P\{A_{ji} = 1\} = p_{ji} \qquad P\{A_{ji} = 0\} = 1 - p_{ji} = q_{ji}.$$

In addition, let us define the random variables $X_j = \sum_{i=1}^{n_j} A_{ji}$ which indicates the number of times the event $A$ occurs in the $n_j$ trials of the $j$-th series and $X = \sum_{j=1}^{k} \sum_{i=1}^{n_j} A_{ji}$ which represents the number of times the event $A$ occurs in the $n = \sum_{j=1}^{k} n_j$ trials. For the previous assumptions the $n$ indicator random variables $A_{ji}$ are thus mutually independent.

The relative frequency of the event $A$ in the $n_j$ trials of the $j$-th series can be represented through the random variable $\hat{p}_j = \frac{X_j}{n_j}$; while the relative frequency of the event $A$ on the total of the $n$ trials is $\hat{p} = \frac{X}{n} = \frac{1}{n} \sum_{j=1}^{k} \hat{p}_j n_j$; which coincides with the weighted arithmetic mean of the relative frequencies of the $k$ series with weights equal to $n_j$ .

The variables defined in this way show therefore the following expectations and variances:

$$\mathbb{E}(\hat{p}_j) = \frac{1}{n_j} \sum_{i=1}^{n_j} p_{ji} \tag{2.2}$$

$$\mathbb{V}(\hat{p}_j) = \frac{1}{n_j^2} \sum_{i=1}^{n_j} p_{ji}(1 - p_{ji}) \tag{2.3}$$

$$\mathbb{E}(\hat{p}) = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} p_{ji} \tag{2.4}$$

$$\mathbb{V}(\hat{p}) = \frac{1}{n^2} \sum_{j=1}^{k} \sum_{i=1}^{n_j} p_{ji}(1 - p_{ji}) \tag{2.5}$$

Thus, probabilistic schemes with independence between both the trials and the series require carrying out $k$ series of $n_j$ trials each.

These schemes can be classified according to the conditions under which these trials are performed, which influence the probability of success $p_{ji}$.

### 2.1. The Bernoulli probabilistic scheme

In the Bernoulli probabilistic scheme the assumption is made that the probability of success is constant from trial to trial and from series to series

$$p_{ji} = p \qquad \text{with} \quad i = 1, 2, ..., n_j \quad \textit{and} \quad j = 1, 2, ..., k.$$

Under such conditions the indicator random variables $A_{ji}$ are independent and identically distributed with common parameter $p$.

The expectation and the variance of the relative frequency $\hat{p}$, for the calculation of which it is advisable to determine the mathematical expectation and the variance of $\hat{p}_j$, in a Bernoulli scheme are

(2.6)

$$\mathbb{E}(\hat{p}_j) = p \qquad \mathbb{V}(\hat{p}_j) = \frac{p(1-p)}{n_j}$$

(2.7)

$$\mathbb{E}(\hat{p}) = p \qquad \mathbb{V}(\hat{p}) = \frac{p(1-p)}{n^2} \sum_{j=1}^{k} n_j = \frac{pq}{n}$$

To analyse the dispersion of the Bernoulli scheme, the following quantity is computed

$$\mathbb{E}\left[\sum_{j=1}^{k} (\hat{p}_j - p)^2 n_j\right] = kpq. \tag{2.8}$$

The Bernoulli scheme is defined as *normal dispersion* scheme (Feller, 1968, p.146).

In this probabilistic scheme the relative frequency $\hat{p}_j$ of the $j$-th series can be approximated with a normal having mean and variance given by the (2.6). This means that the random quantity

$$\sum_{j=1}^{k} \frac{(\hat{p}_j - p)^2}{pq} n_j, \tag{2.9}$$

approximates, as $n_j$ diverges, to a chi-square with $k$ degree of freedom. Similar considerations applied to the relative frequency $\hat{p}$, having expectancy and

variance given by the equations (2.7), enable one to state that the following random variable

$$\frac{(\widehat{p}-p)^2}{pq}n \tag{2.10}$$

can be approximated, as *n* diverges, to a chi-square with one degree of freedom. In the random quantities defined by the expressions (2.9) and (2.10) the probability of success *p*, whose value is usually unknown, is included. For this reason, it is advisable to modify the above said random quantities so that they become functions of known parameters.

The following relation is deduced from the decomposition of the deviance

$$\sum_{j=1}^{k}(\widehat{p}_j-\widehat{\overline{p}})^2 n_j = \sum_{j=1}^{k}(\widehat{p}_j-p)^2 n_j - n(\widehat{\overline{p}}-p)^2.$$

Dividing both members of the previous equation by the factor *pq* we obtain

$$\sum_{j=1}^{k}\frac{(\widehat{p}_j-\widehat{\overline{p}})^2}{pq}n_j = \sum_{j=1}^{k}\frac{(\widehat{p}_j-p)^2}{pq}n_j - \frac{(\widehat{\overline{p}}-p)^2}{pq}n.$$

Because of the associative property of the random variable chi-square, we can deduce that the following expression

$$\sum_{j=1}^{k}\frac{(\widehat{p}_j-\widehat{\overline{p}})^2}{pq}n_j \tag{2.11}$$

can be approximated, as the number of occurrences *nj* diverges, to a chi-square with ($k$ − 1) degrees of freedom. From the convergence in probability of the relative frequency $\hat{p}$ to the unknown parameter *p* and by applying Slutsky's theorem (Cramer, 1996, pp. 254-255) we observe that the random quantity

$$\sum_{j=1}^{k}\frac{(\widehat{p}_j-\widehat{\overline{p}})^2}{pq}n_j\frac{pq}{\widehat{\overline{p}}\widehat{\overline{q}}} = \sum_{j=1}^{k}\frac{(\widehat{p}_j-\widehat{\overline{p}})^2}{\widehat{\overline{p}}\widehat{\overline{q}}}n_j, \tag{2.12}$$

converges in distribution, as the number of occurrences $n_j$ diverges, to a chi-square random variable with ($k$ − 1) degrees of freedom.

### 2.2. The Poisson probabilistic scheme

In 1830 Poisson formalized the scheme of repeated trials in conditions of independence with probabilities of success $p_{ji}$ varying from trial to trial within the same series. The probabilistic scheme called after this author considers constant from series to series both the partial means

$$\overline{p}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}p_{ji} = \overline{p},$$

with $j = 1, 2, \ldots, k$, and the variances

$$\mathbb{V}_j(p_{ji}) = \frac{1}{n_j}\sum_{i=1}^{n_j}(p_{ji} - \overline{p}_j)^2 = \sigma_j^2(p) = \sigma^2(p)$$

between the probabilities of the trials of each series[1], with $j = 1, 2, \ldots, k$.

By considering the deviation $\lambda_{ji}$ between the probability of success of the $i$-th trial of the $j$-th series and the overall mean

$$\lambda_{ji} = p_{ji} - \overline{p} \qquad i = 1, 2, ..., n_j \quad \text{and} \quad j = 1, 2, ..., k \tag{2.13}$$

we obtain

$$\begin{aligned}\mathbb{V}(X_j) &= \sum_{i=1}^{n_j} p_{ji} - \sum_{i=1}^{n_j} p_{ji}^2 \\ &= n_j\overline{p} - n_j\overline{p}^2 - n_j\sigma^2(p) - 2\overline{p}\sum_{i=1}^{n_j}\lambda_{ji} \\ &= n_j\overline{p}(1-\overline{p}) - n_j\sigma^2(p).\end{aligned}$$

Like in the case of the Bernoulli scheme, to analyse the dispersion we calculate the expectation of the weighted sum of the deviations squared between the relative frequencies $\hat{p}_j$ and the overall average probability $\overline{p}$, with weight equal to the number of occurrences $n_j$

$$\mathbb{E}\left[\sum_{j=1}^{k}(\hat{p}_j - \overline{p})^2 n_j\right] = \sum_{j=1}^{k}\frac{\mathbb{V}(X_j)}{n_j} = k\overline{p}(1-\overline{p}) - k\sigma^2(p). \tag{2.14}$$

Comparing this result with the outcome obtained (2.8) in the Bernoulli scheme with constant success probability equal to $\overline{p}$, we understand why the Poisson scheme is defined as *subnormal dispersion scheme* (Kendall, 1996, p.166). The dispersion of the Poisson scheme, therefore, depends on the variability $\sigma^2(p)$ among the probabilities of a series, in particular, the higher it is, the lower will the expectation of the 'deviation' of the relative frequencies $\hat{p}_j$ be.

### 2.3. The Lexis probabilistic scheme

In 1876 Lexis proposed the following probabilistic scheme that was named after him, in which the probabilities of success $p_{ji}$ stay constant within the same series $p_{ji} = p_j$, with $i = 1, 2, ..., n_j$ and $j = 1, 2, \ldots, k$, but vary from series to series[2].

[1] The features of Poisson's probabilistic scheme coincide with those of a stratified sampling scheme in which a single sample unit is extracted from each population layer (Cochran, 1953, chapter 5).

[2] A. Lexis probabilistic scheme represents a particular sampling scheme in two stages (Cochran, 1953, p. 274).

Thus, let

$$\overline{p} = \sum_{j=1}^{k} p_j \frac{n_j}{n}$$

be the average probability of success, obtained as the arithmetic mean of the probabilities of success of the individual series with weights equal to the number of occurrences $n_j$ , and let

$$\sigma^2(p_j) = \sum_{j=1}^{k} (p_j - \overline{p})^2 \frac{n_j}{n}$$

be the variance among the probabilities of the different series.

In order to compare the dispersion of the Bernoulli scheme with constant probability of success equal to $\bar{p}$ with the one of the Lexis scheme, the following deviation is considered. We obtain

$$\mathbb{E}\left[\sum_{j=1}^{k} (\hat{p}_j - \overline{p})^2 n_j\right] = k\overline{p}(1-\overline{p}) + (1-2\overline{p}) \sum_{j=1}^{k} \lambda_j + \sum_{j=1}^{k} (p_j - \overline{p})^2 (n_j - 1).$$

Defining $\widetilde{p} = \frac{1}{k} \sum_{j=1}^{k} p_j$ as the simple (not weighted) mean of the probabilities of success of the various series, we can rewrite the previous equation

$$\mathbb{E}\left[\sum_{j=1}^{k} \left(\frac{X_j}{n_j} - \overline{p}\right)^2 n_j\right] = k\overline{p}(1-\overline{p}) + k(1-2\overline{p})(\widetilde{p} - \overline{p}) + \sum_{j=1}^{k} (p_j - \overline{p})^2 (n_j - 1). \tag{2.15}$$

As the number of occurrences $n_j$ diverges, the last summation tends to infinity. Now, it is possible to compare the result obtained with that (2.8) obtained previously in the Bernoulli scheme with constant probability of success equal to $\bar{p}$. Therefore, the Lexis scheme shows a *supernormal* dispersion (Kendall, 1996, p. 166).

### 2.4. The Coolidge probabilistic scheme

Finally, let us consider the probabilistic scheme proposed by Coolidge in 1921, which represents a generalization of the schemes of repeated trials examined before, since the probabilities of success $p_{ji}$ are free to vary both from trial to trial and from series to series.

To determine the properties of the random variable $X$ associated to the Coolidge scheme we associate to each series the random variable $X_j$ of the Poisson probabilistic scheme and then go ahead with mixing the $k$ variables

determined with weights $n_j$. This method enables to use some of the results obtained previously.

Following the same method used for the Poisson probabilistic scheme, the deviation (2.13) is used to obtain the following result

$$\mathbb{V}(X_j) = n_j\overline{p}_j - n_j\overline{p}^2 - \sum_{i=1}^{n_j}(p_{ji} - \overline{p}_j)^2 - n_j(\overline{p}_j - \overline{p})^2 - 2\overline{p}n_j(\overline{p}_j - \overline{p}).$$

At this point we calculate the expectation of the mixture of the $k$ random variables $\hat{p}_j$ with weights $n_j$ by obtaining

$$\mathbb{E}\left[\sum_{j=1}^{k}(\hat{p}_j - \overline{p})^2 n_j\right] =$$
$$= \sum_{j=1}^{k}\overline{p}_j - k\overline{p}^2 - \sum_{j=1}^{k}\sigma_j^2(p) + \sum_{j=1}^{k}(\overline{p}_j - \overline{p})^2(n_j - 1) - 2\overline{p}\sum_{j=1}^{k}(\overline{p}_j - \overline{p}). \tag{2.16}$$

Depending on various assumptions of different probabilistic schemes, the previous expression includes as special cases the results obtained for the Bernoulli, Poisson and Lexis schemes.

To be able to make some considerations on the above result we should consider a Coolidge scheme made up of $k$ series, all having a constant number of occurrences equal to $m$; under such circumstances the quantity (2.16) becomes

$$\mathbb{E}\left[\sum_{j=1}^{k}\left(\frac{X_j}{m} - \overline{p}\right)^2 m\right] = k\overline{p}(1-\overline{p}) + (m-1)\sum_{j=1}^{k}(\overline{p}_j - \overline{p})^2 - \sum_{j=1}^{k}\sigma_j^2(p).$$

In the Coolidge scheme, the last two addenda of the previous equation are both not equal to zero, but since the two summations $\sum_{j=1}^{k}(\overline{p}_j - \overline{p})^2$ and $\sum_{j=1}^{k}\sigma_j^2(p)$ have the same magnitude, as $m$ diverges, the positive component prevails on the negative one, thus obtaining a supernormal dispersion scheme, also in the case in which the assumptions of the Lexis probabilistic scheme are not met. Therefore, in order for a phenomenon with subnormal dispersion to manifest itself, both the assumptions of the Poisson scheme need to be met, i.e. the probabilities must vary within the same series, while the average probabilities $\bar{p}_j$ and the variances $\sigma_j^2(p)$ have to remain constant from series to series. To find supernormal dispersion instead, it is not necessary that the probabilities of success remain constant from trial to trial in each series, as long as they vary from series to series.

Since, in empirical terms, the average probabilities $\bar{p}_j$ and the variances $\sigma_j^2(p)$ are seldom constant from series to series, it is obvious why a minor number of phenomena displays hypo-binomial dispersion, a property of the Poisson scheme, if compared to those with hyperbinomial dispersion, which mostly follow the Coolidge probabilistic scheme and only to a small extent the Lexis scheme.

## 3. A probabilistic scheme with uniform correlation between the trials

Into a probabilistic scheme, in which the goal is always that of obtaining an event $A$ (success) in $k$ series of $n_j$ trials each with $j = 1, 2, \ldots, k$, we introduce at this point the assumption of dependence between the $n_j$ trials of each series, maintaining the assumption of independence between the $k$ series, though.

Since the following analysis focuses on the relationships of dependence between the variables, we assume to simplify matters that the probability of success $p$ is constant from trial to trial and from series to series. Let us consider the case in which the (linear) dependence between each pair of random variables $A_{ji}$ and $A_{jl}$, with $i \neq l$, of the $j$-th series, manifests itself in a uniform way

$$r(A_{ji}, A_{jl}) = \rho \quad i \neq l; \quad i, l = 1, 2, \ldots, n_j \quad \text{and} \quad j = 1, 2, \ldots, k,$$

by obtaining

$$Cov(A_{jl}, A_{ji}) = \rho(1 - p)p.$$

From the assumption of independence between the series we deduce that

$$r(A_{ji}, A_{sl}) = 0 \quad j \neq s; \quad j, s = 1, 2, \ldots, k \text{ and } i = 1, 2, \ldots, n_j ; \; l = 1, 2, \ldots, n_s.$$

As for the case of independence between the trials, the following variance is computed

$$V(X_j) = n_j p(1 - p) + n_j(n_j - 1)\,\rho p(1 - p) \qquad j = 1, 2, \ldots, k.$$

It follows that

$$\mathbb{E}\left[\sum_{j=1}^{k} (\hat{p}_j - p)^2 n_j\right] = kpq + \rho pq(n - k). \tag{3.1}$$

Comparing this result to those determined previously in the various probabilistic schemes with independence between trials, we obtain the following relation between the linear correlation coefficient $\rho$ and the dispersion of the probabilistic scheme considered:

- if $\rho > 0$ the dispersion is supernormal, same behaviour as for the Lexis scheme;
- if $\rho = 0$ the dispersion is normal, same behaviour as for the Bernoulli scheme;
- if $\rho < 0$ the dispersion is subnormal, same behaviour as for the Poisson scheme.

An estimator of the linear correlation coefficient $\rho$ is

$$\widehat{\rho} = \frac{\sum_{j=1}^{k}\left[(\hat{p}_j - \widehat{p})^2 \frac{n_j}{n}\right] - \frac{k\widehat{p}\widehat{q}}{n}}{\widehat{p}\widehat{q}\left(1 - \frac{k}{n}\right)}.$$

If the number of trials $n$ is very high in comparison to $k$, we can approximate the previous equation as follows

$$\widehat{\rho} \simeq \frac{\sum_{j=1}^{k}\left[\left(\hat{p}_j - \widehat{p}\right)^2 \frac{n_j}{n}\right]}{\widehat{pq}}.$$

It should be noted that the numerator of this proportion represents the variability of the relative frequencies $\hat{p}_j$, whereas the denominator consists of the variability of the indicator random variable $A_{ji}$ in the Bernoulli probabilistic scheme with constant probability of success equal to $p$.

## 4. A hypothesis test for the dispersion of a probabilistic scheme

To find out whether a test meets the assumptions of the Bernoulli scheme we propose to consider the following ratio

$$\frac{\sum_{j=1}^{k}\left[\left(\hat{p}_j - \widehat{p}\right)^2 \frac{n_j}{n}\right]}{(k\widehat{pq})n^{-1}} = \frac{\sum_{j=1}^{k}\left[\left(\hat{p}_j - \widehat{p}\right)^2 n_j\right]}{k\widehat{pq}}. \tag{4.1}$$

As already previously mentioned, the numerator of the ratio (4.1) represents the variability of the relative frequencies $\hat{p}_j$, and, as we can see from the result (2.8), in Bernoulli's probabilistic scheme the expectation of the numerator and the denominator coincide. This means that if the ratio (4.1) is close to unity, then the test taken into consideration meets the assumptions of the Bernoulli probabilistic scheme[1]. Since from the equation (3.1) the expectation of the numerator of the ratio (4.1) appears to be smaller than the denominator, we deduce that if this proportion is sizeably smaller than one, we should, instead, be inclined to use a probabilistic scheme with subnormal dispersion, that is the Poisson scheme, or a scheme with uniform negative correlation between the indicator random variables of each series. If, finally, this proportion is sizeably bigger than one, for the relations (2.15) and (3.1), a scheme with supernormal dispersion is preferred, namely the Lexis scheme or a scheme with uniform positive correlation between the trials of each series. It has to be pointed out that in the latter case, in which the value of the Lexis divergence quotient obtained is considerably higher than one, we might also take into consideration the Coolidge scheme, since it approximately displays supernormal dispersion as the number of tests of each series diverges.

Defining a significance level equal to $\alpha$, we observe that if the value obtained by the test statistics

$$\sum_{j=1}^{k} \frac{\left(\frac{X_j}{n_j} - \widehat{p}\right)^2}{\widehat{pq}} n_j$$

[1] It has to be pointed out that in this case we might also consider a probabilistic scheme with dependence and uncorrelation ($\rho = 0$) between the trials, which means that between the indicator random variables of each series there is a tie of dependence of the non-linear kind, but given the rarity of the case we prefer to disregard this possibility.

is included within the values assumed by the quantiles of the $\alpha/2$ and $(1 - \alpha/2)$ orders of a random variable chi-square with $(k - 1)$ degrees of freedom, then we accept the null hypothesis that the test considered fits the Bernoulli scheme. If instead the value assumed by the test statistics is higher than the quantile of the $(1 - \alpha/2)$ order of a random variable chi-square with $(k - 1)$ degrees of freedom, then we accept the alternative hypothesis and choose either a Lexis scheme or a positive (uniform) correlation scheme between the trials of each series. In the latter case, in which the value assumed by the aforementioned test statistics is lower than the quantile of the $\alpha/2$ order of a random variable chi-square with $(k - 1)$ degrees of freedom, we always accept the alternative hypothesis which, however, consists in the Poisson scheme or in a scheme with negative (uniform) correlation between the trials of each series.

## 5. Conclusion remarks

The probabilistic schemes (Bernoulli, Poisson, Lexis and Coolidge) with independence between the trials show different dispersion properties. By introducing a uniform correlation structure between the trials, a new probabilistic scheme is proposed. By changing the type of correlation, the suggested scheme shows the same dispersion characteristics of the probabilistic schemes analysed in the literature. To identify the type of the dispersion of the probabilistic scheme, a hypothesis test is proposed.

## REFERENCES

COCHRAN, W. G., (1953). Sampling techniques, Wiley.

CRAMER, H., (1996). Mathematical Methods of Statistics. Princeton University Press, Princeton.

FELLER, W., (1968). An introduction to Probability Theory and Its Applications. Vol. I John Wiley & Sons, New York.

JOHNSON, N. L., KEMP A. W. and KOTZ S., (2005). Univariate Discrete Distributions. Wiley, New York.

JOHNSON, N. L., KEMP A. W. and KOTZ S., (1969). Discrete Distributions. Houghton Mifflin, Boston.

KENDALL, S., (1994). The Advanced Theory of Statistics. Vol. I. Hafner Publishing Company, New York.