On the probability distribution of the classical Gras implication index between two binary random variables

Sur la distribution de probabilité de l'indice d'intensité d'implication classique de Gras entre deux variables aléatoires binnaires

PABLO GREGORI[1]
RAPHAËL COUTURIER[2]
RUBÉN PAZMIÑO-MAJÍ[3]

Abstract

In this contribution we study the behavior of the classical Gras implication index as a random variable, when applied to a couple of Bernoulli variables (X,Y), independent or not. We also show the effect of the conditional probability $p_{Y|X}$ on its probability distribution, and specially on its mean value and quartiles.

Keywords : *Binomial Model, Classical Implication Index, Conditional Probability, Multinomial Model.*

Résumé

Dans cette contribution nous étudions le comportement de l'indice d'implication classique de Gras comme une variable aléatoire quand celui-ci est associé à un couple de variables de Bernouilli (X,Y). Nous montrons également l'effet de la probabilité conditionnelle $p_{Y|X}$ sur sa distribution de probabilité, plus particulièrement sur sa moyenne et ses quartiles.

Mots-clés : *Modèle Binomiale, Implication Classique, Probabilité Conditionnelle, Modèle Multinomiale.*

Introduction

Statistical Implicative Analysis (SIA) provides the practitioner with several tools allowing to analyse and find patterns of the kind *association rules* and *cluster analysis* into samples of multivariate data. It was initiated in the PhD dissertation Gras (1979) and developed in Gras and Larher (1993) and Gras et al. (1996), in the context of Didactics of Mathematics. Since then, this topic has been growing in methodology, widening in scope from binary variables to categorical, discrete and continuous numerical, and even to fuzzy and vector-valued variables, and in applications areas,

[1] Institut de Matemàtiques de Castelló i Aplicacions (IMAC), Universitat Jaume I de Castellón, Castellón de la Plana, Spain, gregori@uji.es
[2] Université de Franche Comté, IUT Belfort-Montbéliard, Belfort, France, raphael.couturier@univ-fcomte.fr
[3] Facultad de Ciencias, Escuela Superior Politécnica de Chimborazo, Riobamba, Ecuador, ruben.pazmino@espoch.edu.ec

see : Gras et al. (1996), Gras et al. (2008), Orús et al. (2009), Régnier et al. (2012) and Gras et al. (2013). Most of these functionalities are implemented in CHIC in Couturier (2008).

We focus on one of the major topics of SIA: implication, and restrict ourselves to the analysis of the Gras implication (intensity) index in the classical version and under the binomial modelisation.

Two binary random variables *X* and *Y* can be seen as a couple of Bernoulli trials that can be independent or not. Both marginal success probabilities p_X and p_Y, plus the joint success probability p_{XY} are the three parameters required in order to specify completely the joint distribution of (X, Y).

The implication intensity $\varphi(X \to Y)$ can be seen as a random variable: for each realization of (X, Y) a value of $\varphi(X \to Y)$ is obtained, as a function of the number of successes observed in *X*, *Y*, and the number of counterexamples of the scrutinised rule $X \to Y$.

We have found in Barbu (2007), sect. 3, an interesting analysis of the distribution of the implication intensity (Poisson, classic and entropic) as well as the conditional probability for samples of several sizes. Therein, all possible values of $\varphi(X \to Y)$ were computed taking into account all the possible distributions of the individuals of the samples in the four cases $(X = 0, Y = 0)$, $(X = 0, Y = 1)$, $(X = 1, Y = 0)$ and $(X = 1, Y = 1)$. This procedure corresponds to the case of two independent Bernoulli trials, both with parameter 0.5. Then the distribution $\varphi(X \to Y)$ (classical, Poisson and entropic versions) has been numerical and exactly computed for 3 sample sizes (*n* = 100, 200, 2000). After that, goodness of fit tests to normal and lognormal distributions have been performed.

In this contribution we write down the formula of the probability function of $\varphi(X \to Y)$ and its expectation for general sample size and marginal and joint success probabilities, as well as R scripts for its practical computation. We also give hints on how to simulate a bivariate binary random variable with known marginal success probabilities and a given value (in the long run) implication index. We consider that this is an important result since it allows the consideration of $\varphi(X \to Y)$ as a populational statistic, not only a mere sample statistic, which is, as far as we know, a new conception up to our knowledge.

Notation and definitions

The joint distribution of a binary random variable (X,Y) is completely determined by the joint probability table (completed with the marginal probabilities) shown in Table 1. For instance, it is enough either to fix only three joint probabilities or the two marginal probabilities plus one joint probability. Another possibility is giving the two marginal probabilities p_X and p_Y plus the conditional probability $p_{Y|X} := \frac{p_{XY}}{p_X}$, which is related to the degree of association between *X* and *Y*. An individual for which *X* = 1 and *Y* = 0 is considered to be a counterexample to the rule $X \rightarrow Y$, since it holds the hypothesis of the rule, but not the thesis.

Table 1. Notation for the joint probability table for the bivariate binary random variable (X,Y).

		Y		
		0	1	Margin *X*
X	0	$p_{\overline{X}\,\overline{Y}}$	$p_{\overline{X}Y}$	$p_{\overline{X}}$
	1	$p_{X\overline{Y}}$	p_{XY}	p_X
	Margin Y	$p_{\overline{Y}}$	p_Y	1

If we consider the process of sampling from (X,Y) with size *n*, we can consider the random frequency table given in Table 2. The symbol $N_{X\overline{Y}}$ denotes the random number of counterexamples to the rule, found in the generic sample of size *n*.

Table 2. Notation for the random joint frequency table for generic samples of size n of the bivariate binary random variable (X,Y).

		Y		
		0	1	Margin X
X	0	$N_{\overline{X}\,\overline{Y}}$	$N_{\overline{X}Y}$	$N_{\overline{X}}$
	1	$N_{X\overline{Y}}$	N_{XY}	N_X
	Margin Y	$N_{\overline{Y}}$	N_Y	*n*

One particular realization of the random joint frequency table is denoted as shown in Table 3. Hence, the symbol $n_{X\bar{Y}}$ denotes the observed number of counterexamples to the rule, found in the particular sample of size *n*.

Table 3. Notation for a particular realization of the frequency table of a sample of size n of the bivariate binary random variable (X,Y).

		Y		
		0	1	Margin X
X	0	$n_{\bar{X}\bar{Y}}$	$n_{\bar{X}Y}$	$n_{\bar{X}}$
	1	$n_{X\bar{Y}}$	n_{XY}	n_X
	Margin Y	$n_{\bar{Y}}$	n_Y	n

The classical Gras implication index $\varphi(X \to Y)$ is a sample statistic, which aims at measuring the interest of the rule $X \to Y$: it accounts for how surprisingly small is the observed number of counterexamples $n_{X\bar{Y}}$ found in the particular sample given by the frequency table shown in Table 3, when the statistical independence is taken for granted. In particular we consider the classical implication index for binary variables, which can be defined as

$$\varphi(X \to Y) := P(N_{X\bar{Y}} > n_{X\bar{Y}}).$$

Distribution of the classical Gras implication index

Our modelisation of the random binary variables, with a fixed sample size *n*, leads us to use the binomial model for $N_{X\bar{Y}}$, i.e. $N_{X\bar{Y}} \sim Bin(n, \frac{n_X n_{\bar{Y}}}{n^2})$.

Therefore, the four random variables in the random joint frequency table shown in Table 2 form a random vector ($N_{\bar{X}\bar{Y}}, N_{\bar{X}Y}, N_{X\bar{Y}}, N_{XY}$) which follows the multinomial distribution of parameters (n, $p_{\bar{X}\bar{Y}}, p_{\bar{X}Y}, p_{X\bar{Y}}, p_{XY}$). Consequently, $\varphi(X \to Y)$ varies at every sample (of size *n*) from (X,Y), thus it is a random variable.

An approach to the analysis of the distribution of $\varphi(X \to Y)$ under uniform random sampling of individuals falling at the four cases of the joint frequency table (i.e., $p_{\bar{X}\bar{Y}} = p_{\bar{X}Y} = p_{X\bar{Y}} = p_{XY} = 0.25$) has been performed in Barbu (2007). In that work several hypothesis tests on the goodness of fit to Gaussian and logGaussian distribution were conducted on the different implication indices (classical, Poisson and entropic).

Here we derive the general formula of the probability function of $\varphi(X \to Y)$ and of its expectation, and give R scripts for fast numerical computation.

On the one hand, and conditioned to a sample of size *n*, the probability function for the vector of absolute frequencies is:

$$P(N_{\bar{X}\bar{Y}} = n_{\bar{X}\bar{Y}}, N_{\bar{X}Y} = n_{\bar{X}Y}, N_{X\bar{Y}} = n_{X\bar{Y}}, N_{XY} = n_{XY})$$

$$= \frac{n!}{n_{\bar{X}\bar{Y}}!\,n_{\bar{X}Y}!\,n_{X\bar{Y}}!\,n_{XY}!}\, p_{\bar{X}\bar{Y}}{}^{n_{\bar{X}\bar{Y}}}\, p_{\bar{X}Y}{}^{n_{\bar{X}Y}}\, p_{X\bar{Y}}{}^{n_{X\bar{Y}}}\, p_{XY}{}^{n_{XY}} \qquad (1)$$

The value of $\varphi(X \to Y)$ conditioned to ($N_{\bar{X}\bar{Y}} = n_{\bar{X}\bar{Y}},\ N_{\bar{X}Y} = n_{\bar{X}Y},\ N_{X\bar{Y}} = n_{X\bar{Y}},\ N_{XY} = n_{XY}$) is:

$$\varphi(X \to Y) = 1 - F(n, \frac{n_X n_{\bar{Y}}}{n^2})(n_{X\bar{Y}})$$

where $F_{(n,p)}$ represents the cumulative distribution function of the binomial model with parameters $n \in N$ and $p \in (0,1)$. Hence, the probability function for the random variable $\varphi(X \to Y)$ can be written as:

$$f(\varphi_0) := P(\varphi(X \to Y) = \varphi_0)$$

$$= \sum_{n_{\bar{X}\bar{Y}} + n_{\bar{X}Y} + n_{X\bar{Y}} + n_{XY} = n} \frac{n!}{n_{\bar{X}\bar{Y}}!\,n_{\bar{X}Y}!\,n_{X\bar{Y}}!\,n_{XY}!}\, p_{\bar{X}\bar{Y}}{}^{n_{\bar{X}\bar{Y}}}\, p_{\bar{X}Y}{}^{n_{\bar{X}Y}}\, p_{X\bar{Y}}{}^{n_{X\bar{Y}}}\, p_{XY}{}^{n_{XY}} \qquad (2)$$

where the summation corresponds to every vector ($n_{\bar{X}\bar{Y}}, n_{\bar{X}Y}, n_{X\bar{Y}}, n_{XY}$) of nonnegative integers such that $n_{\bar{X}\bar{Y}} + n_{\bar{X}Y} + n_{X\bar{Y}} + n_{XY} = n$ and such that $\varphi_0 = 1 - F(n, \frac{n_X n_{\bar{Y}}}{n^2})(n_{X\bar{Y}})$.

For the computations, all partitions of *n* into four integers (where one or several of them might be null) are listed. For each partition, the value of φ_0 and its probability are calculated. Finally, probabilities are aggregated for repeated values of φ_0.

For the expectation of $\varphi(X \to Y)$, one can use the expression of $\varphi(X \to Y)$ conditioned to values ($n_{\bar{X}\bar{Y}}, n_{\bar{X}Y}, n_{X\bar{Y}}, n_{XY}$), and the definition to get:

$$E(\varphi(X \to Y))$$

$$= \sum_{n_{\bar{X}\bar{Y}} + n_{\bar{X}Y} + n_{X\bar{Y}} + n_{XY} = n} \left(1 - F(n, \frac{n_X n_{\bar{Y}}}{n^2})(n_{X\bar{Y}})\right) \frac{n!}{n_{\bar{X}\bar{Y}}!\,n_{\bar{X}Y}!\,n_{X\bar{Y}}!\,n_{XY}!}\, p_{\bar{X}\bar{Y}}{}^{n_{\bar{X}\bar{Y}}}\, p_{\bar{X}Y}{}^{n_{\bar{X}Y}}\, p_{X\bar{Y}}{}^{n_{X\bar{Y}}}\, p_{XY}{}^{n_{XY}}$$

(3)

where the summation corresponds to every vector $(n_{\bar{X}\bar{Y}}, n_{\bar{X}Y}, n_{X\bar{Y}}, n_{XY})$ of nonnegative integers such that $n_{\bar{X}\bar{Y}} + n_{\bar{X}Y} + n_{X\bar{Y}} + n_{XY} = n$.

Let us note that these formulae can be written in terms of the size *n* and any three independent parameters of the joint probability table of the bivariate binary random variable (X,Y). For instance, the triplet (p_X , p_Y , p_{XY})or the triplet (p_X , p_Y , $p_{Y|X}$). If the marginal probabilities p_X and p_Y and the sample size *n* are to be fixed, then one can see the effect of the parameter $p_{Y|X}$ (confidence of the rule $X \to Y$) on the complete distribution of $\varphi(X \to Y)$ and on its expected value $E(\varphi(X \to Y))$, giving a wider picture than the one pointed out in the foundational papers of SIA (where a graph shows how both the implication index and the conditional probability decrease as the number of counterexamples increase). We shall show these graphical relations in the next section, as well as provide the R scripts in the appendix.

Graphics and conclusions

The first fact that we have found is the atypical behavior of the distribution of $\varphi(X \to Y)$. It is obviously a discrete variable with values in [0,1], with a sample space with size increasing with *n*, and its distribution behaves in an unusual way: it is more or less bell shaped and it is not piecewise monotone, in the sense that there is one mode (maximum probability) and the values which are the farthest from the mode have lower probability. The plot is reminiscent of some kind of chaotic behavior or fractal structure (more easily seen for large values of *n* such as 100). We think it is caused by the partitions of the integer *n*, which show a recurrence structure, and if each of the four numbers in the partition has an effect of different order of magnitude on the probabilities, it can explain the result, as can be seen on any plot of Figure 1.

A second interesting point is a more thorough description of the relation between conditional probability and implication intensity, that can be seen only through particular values of the parameters. Until now, only simulated samples did this job: for each sample, the values were computed and plotted one against the other. Now we can plot the true distribution of $\varphi(X \to Y)$ as a function of the true conditional probability $p_{Y|X}$.

We show those features in Figure 1: the distribution of $\varphi(X \to Y)$ for fixed marginal probabilities $p_X = 0.5$ and $p_Y = 0.5$, and we move the dependence parameter $p_{Y|X}$ and the sample size *n*, in order to show the effects of each of these values.

Barbu (2007) has performed some goodness-of-fit test to normality, with some positive results. Here we can see that it must be done with caution, since even for large *n*, the shape of the probability function is highly peaked.

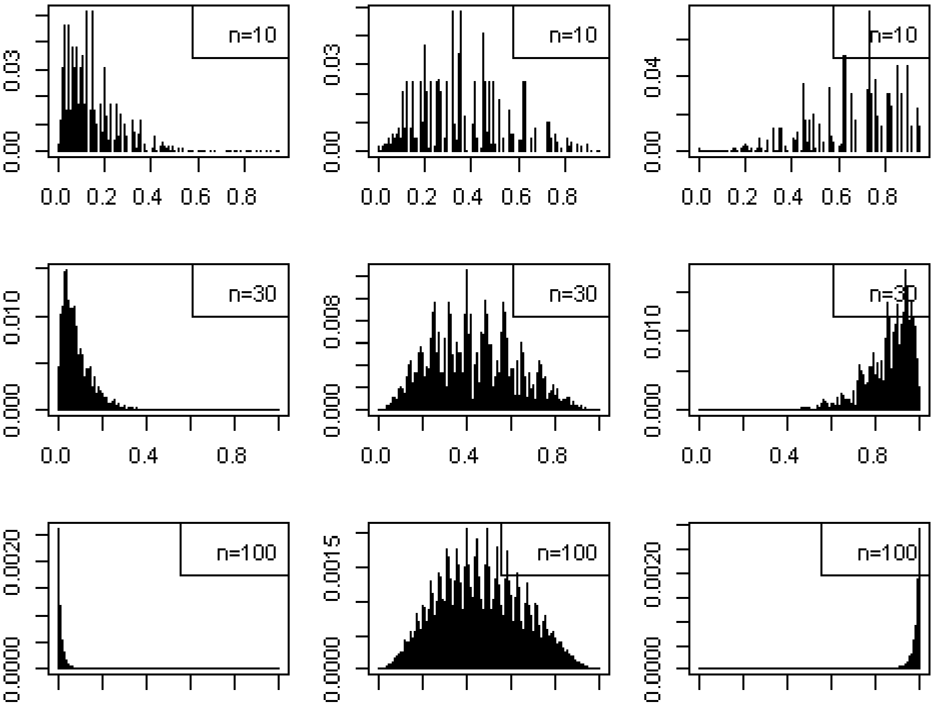

Figure 1. Distribution of $\varphi(X \to Y)$ for p_X =0.5, p_Y =0.5. From left to right, $p_{Y|X}$ =0.25, 0.5, 0.75. (i.e. We move from low to high confidence). From up to down n = 10, 30, 100, we move from smaller to larger samples. Thus we show the effect of these two parameters on the distribution of $\varphi(X \to Y)$.

This is a surprising feature of $\varphi(X \to Y)$, since the study of topics like confidence intervals is based on the fact that distributions are bell shaped or have monotone density functions, and in those cases the sets of most likely values are really intervals. In this case the set of most probable values cannot be included in intervals, and if one wants to give estimations of the true value of the mean, i.e. $E(\varphi(X \to Y))$, it is not easy to consider more than point estimations.

Another useful plot is to simplify the previous plot (of the complete distribution of $\varphi(X \to Y)$) and show the effect of the conditional probability on the expected values and quartiles of the implication index. We show in Figure 2 some examples for different parameter values.

Focusing on the expectation, Equation 3, if p_X, p_Y and the sample size n are to be fixed, $E(\varphi(X \to Y))$ remains as a function of p_{XY}. If one is able to solve, at least numerically, the equation $\varphi_0 = E(\varphi(X \to Y))$, with unknown p_{XY}, then the simulation problem of producing samples of couples of Bernoulli variables with fixed marginal parameters and fixed size, and with a prescribed classical Gras implication index in the long run, is solved. In this case, one can see the implication index as another parameter of the distribution, together with the joint probabilities.

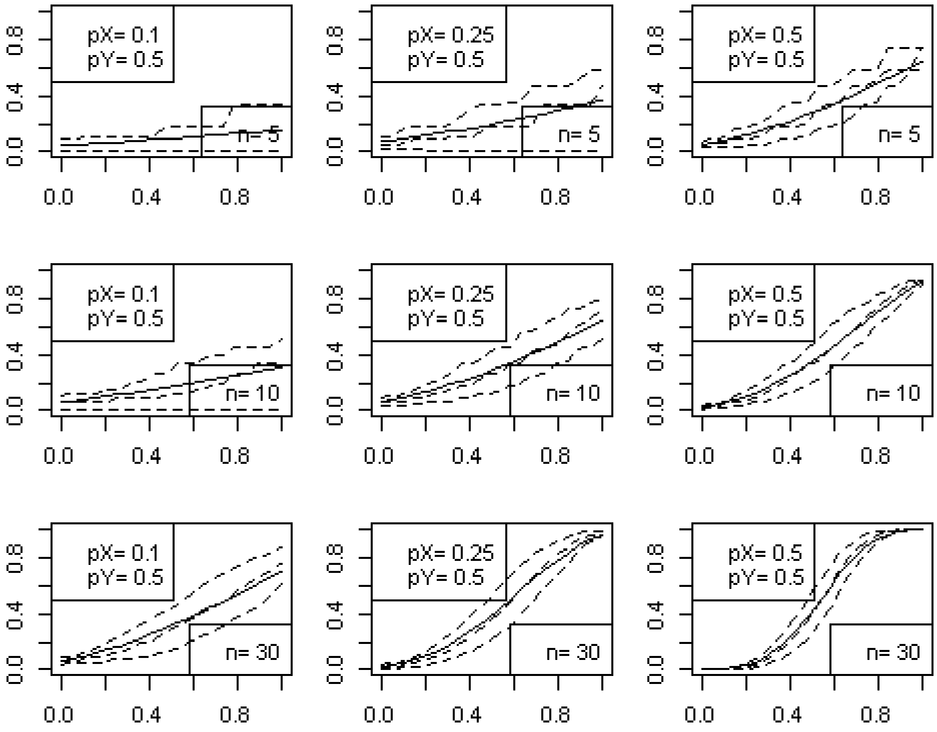

Figure 2. Effect of the conditional probability $p_{Y|X}$ on the mean value (continuous line) and quartiles (dashed lines) of $\varphi(X \to Y)$ for fixed p_Y =0.5 and varying p_X from left to right ($p_X = 0.1, 0.25, 0.5$), and the sample size from up to down ($n = 5,10,30$).

The problem of conducting hypothesis tests on the true value of the classical Gras implication index between two binary variables remains a difficult one, because of the behavior of the distribution. We shall devote our future efforts in this direction.

Références

BARBU, E. (2007), Hiérarchie cohésitive (ou implicative). Technical report, Ecole Polytechnique de l'Université de Nantes, mémoire de DEA.

COUTURIER, R. (2008), Statistical implicative analysis. In CHIC: Cohesive Hierarchical Implicative Classification, volume 127 of Studies in Computational Intelligence, pages 41-52. Springer.

GRAS, R. (1979), Contribution a l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques. PhD thesis, Thèse d'Etat, Université de Rennes 1.

GRAS, R., S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn, and A. Totohasina (1996), *L'implication Statistique*, La Pensée Sauvage.

GRAS, R., and A. Larher (1993), L'implication statistique, une nouvelle méthode d'analyse de données. *Mathématiques, Informatique et Sciences Humaines*, **120**, 5-31.

GRAS, R., J.-C. Régnier, C. Marinica, and F. Guillet (2013), *L'analyse statistique implicative Méthode exploratoire et confirmatoire à la recherche de causalités*, Cépaduès Editions.

GRAS, R., E. Suzuki, F. Guillet, and F. Spagnolo (2008), *Statistical Implicative Analysis, Theory and Applications*, volume 127 of Studies in Computational Intelligence. Springer.

ORÚS, P., L. Zamora, and P. Gregori (2009), *Teoria y Aplicaciones del Análisis Estadístico Implicativo.*, Departamento de Matemáticas de la Universitat Jaume I de Castellón.

REGNIER, J.-C., M. Bailleul, and R. Gras (2012), L'analyse statistique implicative: de l'exploratoire au confirmatoire. Université de Caen.

Appendix

Here is the R code. Long lines have been cut for the printing.

Listing 1. Code for computing the density of $\varphi(X \to Y)$.

```
phi0 = function(x) {
  # computes the value of phi0 for a particular sample
  # with x[1] in nXnY, x[2] in nXY, x[3] in XnY and x[4] in XY
  return(1-pbinom(q=x[3], size=sum(x),
     prob=((x[3]+x[4])*(x[1]+x[3]))/((sum(x))^2)))
}
rvgrasphi = function(pX=0.5, pY=0.5, pXY=NULL, pYgivenX=NULL, n=10) {
  # density (actually probability function) and expectation for the
  # Gras implication index of two Bernoulli variables
  # X and Y of parameters pX and pY and joint success
  # probability pXY (or conditional probability pYgivenX).
  # It returns a list of two components:
  # $f = the values of phi and their probability
  # $E = the expected value
  require(partitions) # needs the package
  if( is.null(pXY) ) {
   pXY = pX * pYgivenX
  } else {
   pYgivenX = pXY/pX
  }
  pnXnY = 1 - pX - pY + pXY
  pnXY = pY - pXY
  pXnY = pX - pXY
  # PROBABILITY FUNCTION FORMULA f(x) := Pr(Phi=x)
  # f(x) = sum_[nn : phi(nn)=x] prob(NN=nn)
  # where nn are all the possible 4 joint absolute frequencies
  # Compute all phi.nn, and sum probabilities of
  # repeated values
  nn = compositions(n,4)
  # computation of prob(NN=nn)
  pr.nn = apply(X=nn, MAR=2, FUN='dmultinom', size=n,
                  prob=c(pnXnY, pnXY, pXnY, pXY))
  # computation of phi(nn)
  phi.nn = apply(X=nn, MAR=2, FUN='phi0')
  phi.values = sort(phi.nn)[c((1:(length(phi.nn)-1))
             [as.logical(sign(diff(sort(phi.nn))))], length(phi.nn))]
  phi.prob = diff(c(0,cumsum(pr.nn[order(phi.nn)])[c((1:(length(phi.nn)-1))
               [as.logical(sign(diff(sort(phi.nn))))], length(phi.nn))]))
  Ephi = sum(phi.nn * pr.nn)
  result = list( f=data.frame(phi=phi.values, fphi=phi.prob), E=Ephi )
  return( result )
}
```

Listing 2. Code for computing the conditional probability on the mean and quartiles.

```
pX1=0.5; pY1=0.5
n1=c(10,30,100)
par(mar=c(3, 2, 1, 1) + 0.1 )
layout(matrix(1:9, 3, 3))
pX = c(0.1, 0.25, 0.5)
pY = 0.5
n = c(5, 10, 30)
for( i in 1:3) {
  # range of P(Y|X)
  pXYmin = max(c((pX[i] + pY)-1, 0))
  pXYmax = min(c(pX[i],pY))
  pYgivenXmin = pXYmin/pX[i]
  pYgivenXmax = pXYmax/pX[i]
  p = seq(fr=pYgivenXmin, to=pYgivenXmax, len=20)
  for( j in 1:3 ) {
    # sample size
    Ephi = numeric(0)
    f25 = numeric(0)
    f50 = numeric(0)
    f75 = numeric(0)
    for( pYgivenX in p ) {
      pXY = pX[i] * pYgivenX
      pnXnY = 1 - (pX[i] + pY) + pXY
      pnXY = pY - pXY
      pXnY = pX[i] - pXY
      nn = compositions(n[j],4)
      thismultinom = function(x) {
        return( dmultinom(x=x, size=n[j], prob=c(pnXnY, pnXY, pXnY, pXY)) )
      }
      # prob(NN=nn)
      pr.nn = apply(X=nn, MAR=2, FUN='thismultinom')
      thisphi = function(x) {
        return(1-pbinom(q=x[3],size=n[j], prob=((x[3]+x[4])*(x[1]+x[3]))/((n[j])^2)))
      }
      # phi(nn)
      phi.nn = apply(X=nn, MAR=2, FUN='thisphi')
      Ephi = c(Ephi, sum(phi.nn * pr.nn))
      f25 = c(f25, sort(phi.nn)[
                which( cumsum(pr.nn[order(phi.nn)]) > 0.25 )[1] ])
      f50 = c(f50, sort(phi.nn)[ which( cumsum(pr.nn[order(phi.nn)]) > 0.50 )[1] ])
      f75 = c(f75, sort(phi.nn)[ which( cumsum(pr.nn[order(phi.nn)]) > 0.75 )[1] ])
    }
    plot(x=p, y=Ephi, type='l', ylim=c(0,1), xlab='P(Y|X)', ylab='Phi')
    points(x=p, y=f25, type='l', lty=2)
    points(x=p, y=f50, type='l', lty=2)
    points(x=p, y=f75, type='l', lty=2)
    legend(x='bottomright', legend=paste('n=',n[j], collapse=''))
    legend(x='topleft', legend=c(paste('pX=', pX[i], collapse=''),
              paste('pY=',pY, collapse='')))
  }
}
```