The “exact” confidence limits for unknown probability in Bernoulli models

R.I. Andrushkiw
Department of Mathematical Sciences
and Center for Applied Mathematics and Statistics,
New Jersey Institute of Technology
Newark, NJ 07102

D.A. Klyushin, Yu.I. Petunin
Department of Cybernetics,
Kyiv National Taras Shevchenko University
Kyiv, Ukraine

M.Yu. Savkina
Institute of Mathematics,
National Academy of Sciences of Ukraine
Kyiv, Ukraine

CAMS Report 0405-25, Spring 2005
Center for Applied Mathematics and Statistics

THE "EXACT" CONFIDENCE LIMITS FOR UNKNOWN PROBABILITY IN BERNOULLI MODELS

R.I.Andrushkiw
Department of Mathematical Sciences and Center for Applied Mathematics and Statistics, New Jersey Institute of Technology, Newark NJ, USA
D.A.Klyushin, Yu.I.Petunin
Department of Cybernetics, Kyiv National Taras Shevchenko University, Kyiv,Ukraine
M.Yu. Savkina
Institute of Mathematics, National Academy of Sciences of Ukraine, Kyiv, Ukraine

Abstract. *The application of mathematical-statistical models in medical diagnostics often requires the construction of an "exact" confidence interval for the unknown probability p of success in Bernoulli models (so called binomial proportion, or proportion of population). This problem was considered in a number of papers (for example, see [1-5] and references cited there). The web-site BioMed Central gives more than 200 citations devoted to this theme. The purpose of our paper is to construct an "exact" confidence interval for unknown probability p of success in classical and generalized Bernoulli models.*

Keywords. Probability, exact confidence interval, Bernoulli models.

1. The Setting

Consider the following test of homogeneity for two populations. Let G_x and G_y be general populations with unknown continuous distribution functions $F_x(u)$ and $F_y(u)$, respectively. Let $x=(x_1,x_2,...,x_n)$ be a sample from G_x and $y=(y_1,y_2,...,y_m)$ be a sample from G_y. We want to test whether the unknown distribution functions $F_x(u)$ and $F_y(u)$ are the same (hypothesis H_0) or not (hypothesis H_1). If the hypothesis H_0 is true we have a homogeneous composite sample $x_1,x_2,...,x_n,y_1,...,y_m$, otherwise the composite sample is heterogeneous. For this purpose, introduce the variational series $x_{(1)}\le x_{(2)}\le ...\le x_{(n)}$, where $x_{(0)}=-\infty$ and $x_{(n+1)}=\infty$, and consider a random interval $I_{i,q}=\left(x_{(i)},x_{(i+q)}\right)$, with i and q fixed numbers, $0\le i\le n$, $1\le q\le n-i+1$. The scheme of trials is formulated in the following way: at the kth step ($k=1,2,...,m$) we test whether the sample value y_k belongs to the interval $I_{i,q}$ and obtain a set of events $A_k=\{y_k\in I_{i,q}\}, k=1,2,...,m$, where every event can occur with certain probability $p_k=P(A_k), k=1,...,m$. Let us introduce a random variable κ that is equal to the number of events A_k arising in m trials. If the hypothesis H_0 is true, then all probabilities p_k are the same and equal to

$$p_q=P(A_k|H_0)=\frac{q}{n+1}. \qquad (1)$$

This scheme is called the *generalized Bernoulli model*. In paper [6,7] the distribution of probabilities of random variable κ was determined:

$$P(\kappa=l|H_0)=\frac{C_{l+q-1}^{l}C_{m+n-l-q}^{m-l}}{C_{m+n}^{m}}, \qquad (2)$$

$$l=1,2,...,m$$

where n,m are sizes of samples x and y, respectively, q is a fixed number which is equal to the number of the order statistics in the interval $I_{i,q}$, C_r^s is the number of combinations of r elements taken s at a time.

The purpose of our paper is to construct the "exact" confidence interval $I(\kappa)$ containing the probability (2) on the basis of the value of the random variable κ. The word "exact" means that the significance level of this confidence interval does not exceed a given number β (as a rule, $\beta=0.05$). With the help of this interval it is possible to propose the following test of hypotheses H_0 and H_1:

1) On the basis of sample x construct the variational series and take a random interval $I_{i,q}$ where *i* and *q* fixed numbers.
2) Calculate the statistics κ which is equal to the number of the elements of the sample y which fall into the interval $I_{i,q}$.
3) On the basis of the statistics κ construct the confidence interval $I(\kappa)$ with the significance level 2β.
4) If the interval $I(\kappa)$ does not cover the probability p_q, the hypothesis H_0 is rejected, otherwise this hypothesis is not rejected.

2. Construction of "Exact" Confidence Interval

Let us construct the interval $I(\kappa)$. Let x be an arbitrary integer discrete random variable with the distribution $p_n(k) = p(x = k)$, $k = 0,1,...,n$. This random variable generate the function $\varphi_x(k)$ defined on the set $M_x = \{0,1,...,n\}$ by the formula $\varphi_x(k) = p_k(x)$. Consider an arbitrary segment $I = \{k : k_0 \le k \le k_1, 0 \le k_0, k_1 \le n\} \subset M$. We call the segment I a domain of monotonicity of function $\varphi_x(k)$, if the condition $k \ge i$, $(k, i \in I)$ implies that $\varphi(k) \ge \varphi(i)$. The integer random variable x is called unimodal, if its range M_x can be represented as a union of one or two domains of monotonicity of the function $\varphi_x(k)$. For example, a random variable with binomial distribution and a random variable with distribution (2) are unimodal.

Remark 1. The concept of unimodal integer discrete random variable when $n > 2$ differs from the concept of unimodal random variable proposed by A.Khinchin. Indeed, by Khinchin, a distribution function $F_x(u)$ of the unimodal random variable x is convex on the ray $(-\infty, x_0)$, where x_0 is a mode of x and is concave on the ray (x_0, ∞). Therefore, $F_x(u)$ can have no more than one break point, but a discrete random variable has a staircase function, hence if $n > 2$ the number of break points is more than two.

In addition to the integer discrete random variable x with the distribution $p_n(k) = p(x = k)$, $k = 0,1,...n$, let us consider a continuous random variable y with the following density of probabilities

$$f(u) = \begin{cases} 0, \text{ if } u \le 1, \\ p_n(k), \text{ if } k \le u \le k+1, k = 0,1,...,n, \\ 0, \text{ if } u \ge n+1. \end{cases}$$

We shall call this function *inducing continuous random variable*. The random variable y induces the random variable x with the help of function $x = \lfloor y \rfloor = Int(y)$ mapping every value $y \in R^1$ to its integer part $\lfloor y \rfloor = Int(y)$. Equality in the above formula has the following implication. Denote by E_x a random trial producing a random variable x with given function of probability $p_n(k)$ and denote by E_z an independent random trial producing the random variable z uniformly distributed on $[0,1]$. In a compound random trial $E_y = (E_x, E_z)$ the random variable $y = x + z$ has the density function $f(u)$, and the random variable $Int(y)$ takes the value k, if and only if $x = k$ as a result of the trial E_z. Therefore, we can consider that the compound random trial E_y produces the random values x and y, such that not only the distributions of x and $Int(y)$ are the same, but also the values x and $Int(y)$ are the same.

Let x be a random variable with binomial distribution, and y be a continuous random variable inducing x. In a Bernoulli model the mathematical expectation $m(y)$ and variance $\sigma^2(y)$ are as follows:

$$m(y) = \int_0^{n+1} uf(u)\,du = \sum_{k=0}^{n} \int_k^{k+1} uf(u)\,du =$$

$$= \sum_{k=0}^{n} p_n(k) \int_k^{k+1} u\,du = \frac{1}{2} \sum_{k=0}^{n} (2k+1)\, p_n(k) =$$

$$= \sum_{k=0}^{n} k p_n(k) + \frac{1}{2} \sum_{k=0}^{n} p_n(k) = np + \frac{1}{2};$$

$$m\left(y^2\right)=\sum_{k=0}^{n}p_k\int_{k}^{k+1}u^2du=$$

$$=\frac{1}{3}\sum_{k=0}^{n}p_k\left(3k^2+3k+1\right)$$

$$=\sum_{k=0}^{n}k^2p_k+\sum_{k=0}^{n}kp_k+\frac{1}{3}\sum_{k=0}^{n}p_k=$$

$$=m\left(x^2\right)+m\left(x\right)+\frac{1}{3}=$$

$$=\left(\sigma^2\left(x\right)+\left(m\left(x\right)\right)^2\right)+np+\frac{1}{3}=$$

$$=npq+\left(np\right)^2+np+\frac{1}{3}.$$

$$\sigma^2\left(y\right)=npq+\frac{1}{12},$$

$$\sigma\left(y\right)=\sqrt{npq+\frac{1}{12}}\quad.$$

In the generalized Bernoulli model we have

$$m\left(x\right)=mp_q,$$

$$\sigma^2\left(x\right)=\frac{\left(m+n+1\right)m}{n+2}p_q\left(1-p_q\right)$$

Therefore, in the generalized Bernoulli model

$$m\left(y\right)=mp_q+\frac{1}{2},$$

$$\sigma^2\left(y\right)=\frac{\left(m+n+1\right)m}{n+2}p_q\left(1-p_q\right)+\frac{1}{12}.$$

Consider an arbitrary fixed confidence interval $\left(a,b\right)$ containing the bulk of G_y with significance level α. Since $Int\left(y\right)$ is a non-decreasing function, it follows that the random event $A_y=\left\{y\in\left[a,b\right]\right\}=\left\{a\le y\le b\right\}$ implies the random event $A_x=\left\{x=Int\left(y\right)\in\left[Int\left(a\right),Int\left(b\right)\right]\right\}$. Therefore, the significance level of the closed confidence interval $\left[Int\left(a\right),Int\left(b\right)\right]$ for the bulk of G_x does not exceed α.

Moreover,

$$\left[Int\left(a\right),Int\left(b\right)\right]\subset\left[a-1,b\right],$$

and hence

$$p\left(x\in\left[Int\left(a\right),Int\left(b\right)\right]\right)\le p\left(x\in\left[a-1,b\right]\right).$$

Therefore, the significance level of the confidence interval $\left[a-1,b\right]$ for the bulk of G_x also does not exceed α:

$$p\left\{x\notin\left[a-1,b\right]\right\}\le\alpha.$$

It is easy to see that the integer discrete random variable x is unimodal if and only if inducing continuous random variable y is unimodal in the sense of Khinchin. For such random variables y the Gauss-Vysochanskij-Petunin inequality holds [8]

$$p\left(\left|y-m\left(y\right)\right|\ge\lambda\sigma\left(y\right)\right)\le\frac{4}{9}\frac{1}{\lambda^2},$$

where $\lambda>\sqrt{\frac{8}{3}}$.

Therefore, the significance level of the confidence interval

$$\left[m\left(y\right)-\lambda\sigma\left(y\right),m\left(y\right)+\lambda\sigma\left(y\right)\right]$$

covering the bulk of G_y does not exceed $\alpha=\frac{4}{9}\frac{1}{\lambda^2}$.

In particular, when $\lambda=3$ we have $\alpha=\frac{4}{81}<0.05$.

In the case of the classical Bernoulli model put

$$a=m\left(y\right)-\lambda\sigma\left(y\right)=np+\frac{1}{2}-\lambda\sqrt{npq+\frac{1}{12}},$$

$$b=m\left(y\right)+\lambda\sigma\left(y\right)=np+\frac{1}{2}+\lambda\sqrt{npq+\frac{1}{12}}.$$

On the basis of the previous reasoning we have that the confidence interval $\left[a-1,b\right]$ covers the bulk of the random variable x with binomial distribution, i.e. the confidence interval

$$I=\left[np-\frac{1}{2}-\lambda\sqrt{npq+\frac{1}{12}},np-\frac{1}{2}+\lambda\sqrt{npq+\frac{1}{12}}\right]$$

has the significance level, which does not exceed $\alpha=\frac{4}{9}\frac{1}{\lambda^2}$ when $\lambda>\sqrt{\frac{8}{3}}$.

The random event $\left\{x\in I\right\}$ can be rewritten in the following form:

$$\left|x-np\right|\le\frac{1}{2}+\lambda\sqrt{npq+\frac{1}{12}}.$$

Thus, in the Bernoulli model

$$p\left(\left|h-p\right|\le\frac{1}{2n}+\frac{\lambda}{n}\sqrt{npq+\frac{1}{12}}\right)\ge1-\alpha.$$

To construct the confidence interval for the unknown probability p on the basis of the proportion h in the Bernoulli model consisting of n

trials consider two functions depending on $p \in [0,1]$:

$$\varphi(p) = |h - p|$$

and

$$\psi(p) = \frac{1}{2n} + \frac{\lambda}{n}\sqrt{np(1-p) + \frac{1}{12}}.$$

Let

$$\tilde{\psi}(p) = \sqrt{np(1-p) + \frac{1}{12}}, \quad p \in R^1.$$

In easy to see, that the graph of the function $\tilde{\psi}(p)$, $p \in R^1$ is the upper half of the ellipse E passing through the points

$$A = \left(\frac{1}{2n}\left(n + \sqrt{\frac{n}{3} + n^2}\right), 0\right),$$

$$B = \left(\frac{1}{2}, \sqrt{\frac{1}{12n} + \frac{1}{4}}\right),$$

$$C = \left(\frac{1}{2n}\left(n - \sqrt{\frac{n}{3} + n^2}\right), 0\right),$$

$$D = \left(\frac{1}{2}, -\sqrt{\frac{1}{12n} + \frac{1}{4}}\right)$$

with the center at the point $\left(\frac{1}{2}, 0\right)$. The graph of $\psi(p)$ is constructed on the basis of restriction of the graph of $\tilde{\psi}(p)$ to the segment $[0,1]$ by stretching or compressing its graph by a factor $\frac{\lambda}{n}$ and shifting by $\frac{1}{2n}$.

Therefore, the graph of the function $\psi(p)$ which does not depend on h is an arc of an ellipse Γ passing through the points $(0, \psi(0))$, $\left(\frac{1}{2}, \psi\left(\frac{1}{2}\right)\right)$, $(1, \psi(1))$, such that the function $\psi(p)$ achieves its minimum at the point $p = \frac{1}{2}$ and is symmetrical with respect to that point.

The lower confidence limit p_1 is a root of the quadratic equation

$$\left(1 + \frac{\lambda^2}{n}\right)p^2 - \left(\frac{\lambda^2}{n} - \frac{1}{n} + 2h\right)p + h^2 - \frac{h}{n} + \frac{1}{4n^2}\left(1 - \frac{\lambda^2}{3}\right) = 0. \quad (3)$$

If $h > \psi(0) = \frac{1}{2n} + \frac{\lambda}{n\sqrt{12}}$, then the lower confidence limit p_1 is the least root of (3). If $h \le \psi(0)$, then $p_1 = 0$.

Similarly, the upper confidence limit p_2 is a root of the equation

$$\left(1 + \frac{\lambda^2}{n}\right)p^2 - \left(\frac{\lambda^2}{n} + \frac{1}{n} + 2h\right)p + h^2 + \frac{h}{n} + \frac{1}{4n^2}\left(1 - \frac{\lambda^2}{3}\right) = 0. \quad (4)$$

If $1 - h > \psi(1)$, then the upper confidence limit p_2 is the largest root of (4). If $1 - h \le \psi(1)$, then $p_2 = 1$.

Remark 2. Note, that $p_1 \le h \le p_2$, so that the proportion of successes always lies in the confidence interval $[p_1, p_2]$.

For the generalized Bernoulli model a similar reasoning gives the following quadratic equation for the lower confidence limit:

$$\left(1 + \frac{(m+n+1)\lambda^2}{(n+2)m}\right)p^2 + \left(\frac{1}{m} - \frac{(m+n+1)\lambda^2}{(n+2)m} - 2h\right)p + h^2 - \frac{h}{m} + \frac{1}{4m^2}\left(1 - \frac{\lambda^2}{3}\right) = 0 \quad (5)$$

If $h > \frac{1}{2m} + \frac{\lambda}{m\sqrt{12}} = \gamma$, then the lower confidence limit p_1 for the generalized Bernoulli model is the least root of (5). If $h \le \gamma$, then $p_1 = 0$.

Similarly, the upper confidence limit p_2 for the generalized Bernoulli model is the root of the quadratic equation

$$\left(1+\frac{(m+n+1)\lambda^2}{(n+2)m}\right)p^2 -$$
$$-\left(\frac{1}{m}+\frac{(m+n+1)\lambda^2}{(n+2)m}+2h\right)p+ \quad (6)$$
$$+h^2+\frac{h}{m}+\frac{1}{4m^2}\left(1-\frac{\lambda^2}{3}\right)=0$$

If $1-h>\gamma$, then the upper confidence limit p_2 is the largest root of (6). If $1-h\le\gamma$, then $p_2=1$. By virtue of the previous results the significance level of the confidence interval does not exceed $\frac{4}{9}\frac{1}{\lambda^2}$ (in particular, 0.05 for $\lambda=3$).

3. References

[1] Petunin Yu. I., Klyushin D.A., Andrushkiw R.I., Ganina K.P., Boroday N.V., Computer-Aided Differential Diagnosis of Breast Cancer and Fibroadenomatosis based on Malignancy Associated Changes in Buccal Epithelium, Automedica 2001; 19(3-4): 135-164.

[2] Brown L.D., Cai T.T., DasGupta A. Interval Estimation for a Binomial Proportion. Stat. Sci. 2001; 2(16): 101-133.

[3] Petunin Yu. I., Klyushin D.A., Andrushkiw R.I., Ganina K.P., Boroday N.V. Analysis of Malignancy-Associated DNA Changes in the Nuclei of Buccal Epithelium in the Pathology of the Thyroid and Mammary Glands, Annals of the New York Academy of Sciences 2002; 980: 1-12.

[4] [4] Yoo S., David H., Revisiting Clopper-Pirson. Technical Report 2002-05. Department of Statistics and Statistical Laboratory. Iowa University; 2002.

[5] Andrushkiw R.I., Klyushin D.A., Petunin Yu. I., Lysyuk V., Boroday N.V., Diagnosis of Breast Cancer by the Modified Nearest Neighbor Recognition Method. In: F. Valafar, editor. Proceedings of the International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences; 2002 Jun 24- 27; Las Vegas, Nevada, USA; p. 176-189.

[6] Matveychuk S.A., Petunin Yu.I. Generalized Bernoulli schemes in variance statistics. Part I. Ukr. Mat. Journal 1991; 43(4): 518-528.

[7] Matveychuk S.A., Petunin Yu.I. Generalized Bernoulli schemes in variance statistics. Part II. Ukr. Mat. Journal 1991; 43(6): 779-785.

[8] Vysochanskij D.F., Petunin Yu.I. Justification of the 3σ rule for unimodal distribution. Theor. Probab. and Math. Stat. 1989; 21: 25-36.