

Computer-Aided Cytogenetic Method of Breast Cancer Diagnosis. Part II - Test Criteria

R.I.Andrushkiw

*Department of Mathematical Sciences and
Center for Applied Mathematics and Statistics,
New Jersey Institute of Technology, Newark, NJ,
USA*

Abstract. *In this part we describe the statistical test criteria which are used in Part I in the construction of computer-aided cytogenetic method of breast cancer diagnosis.*

Keywords: *breast cancer, fibroadenomatosis, buccal epithelium, discriminant analysis.*

1 The 3σ -rule

The empirical 3σ -rule, which is well known in mathematical statistics, states that for the overwhelming majority of commonly encountered random variables x the following inequality holds:

$$P(|x - m(x)| \geq 3\sigma(x)) \leq 0.05 \quad (1)$$

where $m(x)$ is the expectation and $\sigma(x)$ is the standard deviation of x . The value of the constant 0.05 is stipulated by the fact that in many applied sciences (for example, biology and medicine) the 5% significance level is the most widely used. The justification of the 3σ -rule was given in paper [1]. There also exist several different proofs of this empirical rule [2–4].

Theorem 1. *For all $k > 0$, the following inequality holds for an arbitrary random variable x having a unimodal distribution and finite variance $\sigma^2(x) > 0$*

$$P(|x - m(x)| \geq k\sigma(x)) \leq \frac{4}{9} \cdot \frac{1}{k^2}, \quad k \geq \sqrt{\frac{8}{3}} \quad (2)$$

2 The $3s$ -rule

In order to construct the confidence interval

D.A.Klyushin, K.N.Golubeva,
M.Pokoyovy, A.V.Romanov

*Kyiv National Taras Shevchenko University,
Kyiv, Ukraine*

containing the bulk of general population G with the help of Gauss-Vysochansky-Petunin inequality we must know the mathematical expectation $m(x)$ and variance $\sigma^2(x)$. Unfortunately, these characteristics usually are unknown. In this cases we can select a random sample x_1, x_2, \dots, x_n from the general population G and replace the unknown values $m(x)$ and $\sigma^2(x)$ by their estimations \bar{x} and s_n^2 respectively.

$$m(x) \approx \bar{x} = \frac{1}{n} \sum_{k=1}^n x_k,$$

$$\sigma^2(x) \approx s^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2.$$

These estimations have good properties. They are unbiased, i.e. their mathematical expectations coincide with the exact value of the estimated parameters $m(x)$ and $D(x)$:

$$m(\bar{x}) = m(x),$$

$$m(s^2(x)) = D(x).$$

In constructing the confidence interval J containing the bulk of the general population G on the basis of the sample x_1, x_2, \dots, x_n it is quite naturally to replace the mathematical expectation $m(x)$ and the variance $\sigma^2(x)$ by their estimations \bar{x} and s^2 respectively. So, we can formulate the so-called $3s$ -rule:

$$\hat{J} = (\bar{x} - 3s, \bar{x} + 3s),$$

where

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k, s^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2. \text{When}$$

n is large, this interval contains not less than 95% of the values from G . Now, let us consider the following question: under what n the 3s-rule holds. According to practical recommendations, the estimation \bar{x} almost coincides with $m(x)$ when $n \geq 30$, and $s^2(x) \approx D(x)$ when $n \geq 150$. But mathematical simulations show that the interval \hat{J} contains not less than 95% of the values from G when $n \geq 11$.

The 3s-rule is closely connected with the 3s₁-rule, which allows us to calculate a confidence interval for unknown mathematical expectation $m(x)$ on the basis of the sample x_1, x_2, \dots, x_n with significance level not exceeding 0.05. At first, consider the problem of the constructing of the confidence interval on the basis of 3σ-rule, in the case when the value of the random variable x and its variance $\sigma^2(x)$ are known. By virtue of the inequality (2) we have:

$$\begin{aligned} p(|x - m(x)| \leq 3\sigma(x)) &= \\ &= p(-3\sigma(x) \leq m(x) - x \leq 3\sigma(x)) = \\ &= p(x - 3\sigma(x) \leq m(x) \leq x + 3\sigma(x)) \geq 0.95 \end{aligned}$$

Hence, it follows that the interval $J = (x - 3\sigma(x), x + 3\sigma(x))$ is a random confidence interval for unknown mathematical expectation $m(x)$ with significance level 0.05 (by virtue of 3σ-rule). In prevalent number of cases we can put $x = \bar{x}$, so that

$$\begin{aligned} m(\bar{x}) &= m\left(\frac{1}{n} \sum_{k=1}^n x_k\right) = \frac{1}{n} m\left(\sum_{k=1}^n x_k\right) = \\ &= \frac{1}{n} \sum_{k=1}^n m(x_k) = m(x), \\ \sigma^2(\bar{x}) &= D(\bar{x}) = D\left(\frac{1}{n} \sum_{k=1}^n x_k\right) = \\ &= \frac{1}{n^2} \sum_{k=1}^n D(x_k) = \frac{\sigma^2}{n}. \end{aligned}$$

Therefore, the significance level of the confidence interval $\left(\bar{x} - 3\frac{\sigma}{\sqrt{n}}, \bar{x} + 3\frac{\sigma}{\sqrt{n}}\right)$ does not exceed 0.05, i.e.

$$p\left(m(x) \in \left(\bar{x} - 3\frac{\sigma(x)}{\sqrt{n}}, \bar{x} + 3\frac{\sigma(x)}{\sqrt{n}}\right)\right) \geq 0.95$$

It is easy to see that the following estimation of the variance of the sample mean is unbiased, and has the same properties as the estimation $s^2(x)$:

$$s_1^2(\bar{x}) = \frac{1}{n} s^2(x) = \frac{1}{n(n-1)} \sum_{k=1}^n (x_k - \bar{x})^2.$$

Replacing $\sigma^2(\bar{x})$ by its estimation $s_1^2(\bar{x})$, we obtain the 3s₁-rule that states that the confidence interval

$$J_1 = \left(\bar{x} - \frac{3s(x)}{\sqrt{n}}, \bar{x} + \frac{3s(x)}{\sqrt{n}}\right)$$

contains unknown mathematical expectation $m(x)$ with the probability not exceeding 0.95, when n is large.

Since the estimation $s^2(x)$ has practically the same value as $\sigma^2(x)$ if $n \geq 150$, we can assume that the estimation $s_1^2(\bar{x})$ coincides with the variance $\sigma^2(\bar{x})$ and that the 3s₁-rule holds when $n \geq 150$. Nevertheless, this rule may be applied even for $n \geq 11$.

In mathematical statistics samples are classified by their size: 1) small samples, when $n \leq 30$; 2) middle samples, when $30 < n < 150$, and 3) large samples, when $n \geq 150$. To summarize, we can state that the 3s and 3s₁-rules hold for middle and large samples, and even for small samples, if their size exceeds $n = 11$.

3 Confidence intervals and order statistics

Suppose G is some general population with unknown distribution function $F(u)$, x_1, x_2, \dots, x_n is a sample obtained from G as the result of a simple random sampling, and x

is an element from G which does not depend on the sample x_1, x_2, \dots, x_n . Let

$$x_{(1)} \leq \dots \leq x_{(i)} \leq \dots \leq x_{(j)} \leq \dots \leq x_{(n)}$$

be a variational series of the sample x_1, x_2, \dots, x_n , and let $x_{(i)}$ be the i th order statistics. The basic aim of this section is the construction of the most accurate confidence interval (a, b) , $a < b$, containing the bulk of general population G , where $a(x_1, x_2, \dots, x_n)$ and $b(x_1, x_2, \dots, x_n)$ are two arbitrary Borel-measured functions of the sample values x_1, x_2, \dots, x_n .

Let us introduce the notions of reliability of an arbitrary confidence interval $J = (a, b)$ containing the bulk of the general populations. Let $a(u_1, u_2, \dots, u_n)$ and $b(u_1, u_2, \dots, u_n)$ be two arbitrary (Borel) functions satisfying for every $u \in R^1$ the following inequality:

$$a(u_1, u_2, \dots, u_n) \leq b(u_1, u_2, \dots, u_n).$$

Using these functions and sample x_1, x_2, \dots, x_n we can construct a random confidence interval $J = (a(u_1, u_2, \dots, u_n), b(u_1, u_2, \dots, u_n))$ for the bulk of the general population G . Suppose, that the random variables $a(u_1, u_2, \dots, u_n)$ and $b(u_1, u_2, \dots, u_n)$ have the mathematical expectations $m(a)$ and $m(b)$, respectively. We shall call the reliability $\alpha(a, b)$ of the confidence interval J its significance level:

$$\alpha(a, b) = p(x \in (a, b)),$$

Theorem 2. *If G is a general population with continuous distribution $F(u)$, then the reliability level of the confidence interval $(x_{(i)}, x_{(j)})$ is equal to $\frac{j-i}{n+1}$.*

4 Ellipsoid of minimal volume enclosing the set $M = \{X_k\}_{k=1, \dots, N} \subset R^n$.

Consider the following algorithm for constructing an ellipsoid of minimal volume

enclosing the set of point $M = \{X_k\}_{k=1, \dots, N} \subset R^n$

Let us describe the algorithm in the case of R^2 . At the first stage of the algorithm we select the pair of the points X_i and X_j with maximal distance between them

$$\rho(X_i, X_j) = \text{diam}\{X_k\}_{k=1, \dots, N}.$$

Then the points X_i and X_j are connected by the segment $a = [X_i, X_j]$ and the coordinate system is rotated so that the abscissa becomes parallel to the segment a . Then we construct the minimal rectangle P containing the set M with sides which are parallel to coordinate axes of the new coordinate system. At the next stage we compress the plane along the abscissa so that the rectangle P transforms to the square K , and construct a circle C of minimal radius ρ centered at the point U , which corresponds to the intersection of diagonals of the square K containing all points of the set M :

$$\rho = \max_{k=1, \dots, N} \rho(U, X_k)$$

At the last stage we perform an inverse transformation: expansion of the plane transforming the square K into the rectangle P and the circle C into the ellipse E containing the set M . This ellipse is considered as an approximation of the ellipse having minimal area.

The construction of the ellipsoid having minimal volume containing the set M in R^3 is performed in the following way. As in the case of R^2 , we first select the pair of points X_i, X_j with maximal distance (the ends of the diameter of the set M). Let $a = [X_i, X_j]$ be the line segment joining the points X_i, X_j and pass through the ends of the segment a two planes, β and γ , which are perpendicular to the segment a . Consider the orthogonal projection of the set M on the plane β and denote this set by M_β . Then with the help of the method described above we construct the minimal rectangle P_β on the plane β , containing the set M_β whose side is parallel to the segment a . The rectangle P_β and the

segment a determine the parallelepiped $P = P_\beta \times a$ containing the set M . Then we compress the space in the direction which is parallel to the segment a so that the parallelepiped P transforms to the cube K . At the next stage we construct the ball C of minimal radius centered at the point U , which corresponds to the intersection of the diagonals of the cube K , containing the transformed compressed set M . At the final stage we transform the cube K into a parallelepiped P , using the inverse transformation (extension) of the space, and obtain from the ball C an ellipsoid E which approximates the ellipsoid of minimal volume.

For higher dimensions the construction of the confidence ellipses is analogous.

Now, let us show that the confidence level of such ellipsoids is equal to $\frac{n}{n+1}$. Indeed, if

the centers of these ellipsoids are fixed, then the random variables $\rho(O, X_i)$ are independent and identically distributed. On the basis of results obtained in Section 3, the probability of falling out of the values $\rho(O, X_i)$ from the maximal order statistics is equal to $\frac{1}{n+1}$. Hence, the confidence level of

this ellipsoid is $\frac{n}{n+1}$.

5 References

- [1] D.F. Vysochanskij and Yu.I. Petunin, "Justification of the 3-sigma rule for unimodal distribution", *Theor. Probability and Mathem. Statistics*, Vol. 21, pp. 25-36, 1980.
- [2] F. Pukelsheim, "The three sigma rule", *Amer. Statist.*, Vol. 48, pp. 88-91, 1994
- [3] T. Sellke, "Generalized Gauss-Chebyshev inequalities for unimodal random variables", *Metrika*, Vol. 43, pp. 107-121, 1996
- [4] S. Dharmadhikari and K. Joag-dev Unimodality, *Convexity, and Applications*, Academic Press, New York, 1988.