QUANTITATIVE METHODS IN ECONOMICS
Vol. XI, No. 1, 2010, pp.158 -164

STATISTICAL PROPERTIES OF A CONTROL DESIGN OF CONTROLS PROVIDED BY SUPREME CHAMBER OF CONTROL

Wojciech Zieliński
Katedra Ekonometrii i Statystyki SGGW
Nowoursynowska 159, PL-02-767 Warszawa
e-mail: wojtek.zielinski@statystyka.info

Abstract. In statistical quality control objects are alternatively rated. It is of interest to estimate a fraction of negatively rated objects. One of such applications is a quality control provided by Supreme Chamber of Control (NIK) to find out a percentage of abnormalities in the work among others of tax offices. Mathematical details of experimental designs for alternatively rated phenomena are given in Karliński (2003). There are given requirements for sample sizes, numbers of negative rates, accuracy of estimation and error risks. In the paper, some statistical properties of given plans are investigated.

Key words: statistical quality control, alternative rating, experimental design

One of the problem of the statistical quality control is the problem of the estimation of the fraction of defective products. Generally speaking, the products are alternatively rating and one is interested in estimation of a fraction of negatively rated objects. In this approach, the binomial statistical model is applied, i.e. if ξ is a random variable counting negative rated in a sample of size n, then ξ is binomially distributed

$$P_\theta\{\xi = x\} = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x = 0,1,\ldots,n,$$

where $\theta \in (0,1)$ is a probability of drawing a defective product. The aim of the statistical quality control is to estimate θ.

In many norms and books devoted to different applications there are given exacts designs of experiments, i.e. requirements for ample sizes, number of negative rates in the sample, accuracy of estimation and error risks. One of such applications are quality controls provided by Supreme Chamber of Control, the goal of which is finding abnormalities in tax offices. Karliński (2003) gives mathematical details of such controls. There are given methods of providing experiments and rules of statistical inference. In what follows, statistical properties of given experimental designs are investigated.

The aim of a control is the interval estimation of a percentage of the defective objects. NIK guidelines are based on approximate solutions. There is given a method of calculating a minimal sample size:

$$n = \frac{N u_\alpha^2 \theta(1-\theta)}{\varepsilon^2 (N-1) + u_\alpha^2 \theta(1-\theta)},$$

where N is the size of the population, u_α is the critical value of the standard normal distribution for $1-\alpha$ confidence level, ε is the given accuracy of estimation and θ is the real (assumed in control) theoretical percentage of defective objects. Let the number of observed defective objects in a sample be m .Then

$$e = u_\alpha \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n} \frac{N-n}{N-1}}$$

and an interval is obtained

$$\left(\hat{\theta} - e, \hat{\theta} + e\right).$$

The interval is considered as a confidence interval for the fraction of failures.

Let us investigate the statistical properties of the above method. In what follows it is assumed that the population is infinite. Under the assumption the given formulae take on the form

$$n = \frac{u_\alpha^2 \theta(1-\theta)}{\varepsilon^2} \quad \text{and} \quad e = u_\alpha \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}.$$

The formulae are derived from asymptotic approximations of e Binomial distribution. Central Limit Theorem states that for large n we have

$$P_\theta\{\xi \le k\} \approx \Phi\left(\frac{\xi - n\theta}{\sqrt{n\theta(1-\theta)}}\right),$$

where $\Phi(\cdot)$ denote the cumulative distribution function of the standard normal distribution $N(0,1)$. Hence, it assumed that $(\xi - n\theta)/\sqrt{n\hat{\theta}(1-\hat{\theta})}$ is asymptotically normal $N(0,1)$, it means that

$$P_\theta\left\{\frac{|\hat{\theta}-\theta|}{\sqrt{\hat{\theta}(1-\hat{\theta})}}\sqrt{n}\le u_\alpha\right\}\approx 1-\alpha$$

From the above we obtain given earlier interval.

Taking $1-\alpha=0.95$, $\varepsilon=0.05$ and $\theta=0.05$ we obtain

$$u_{0.05}=1.96, n=73, e=0.0499967.$$

The number of failures in the sample is a random variable ξ distributed as Binomial $B(73,0.05)$. Numerical results which demonstrate statistical properties of the above confidence are given In the Table 1.

Table 1.

m	$\left(\hat{\theta}-e,\hat{\theta}+e\right)$		$P_{0.05}\{\xi=m\}$
0	(-0.049997,0.049997)	0	0.0236
1	(-0.036298,0.063695)	1	0.0909
2	(-0.022599,0.077394)	1	0.1722
3	(-0.008901,0.091093)	1	0.2144
4	(0.004798,0.104791)	1	0.1975
5	(0.018496,0.118490)	1	0.1435
6	(0.032195,0.132188)	1	0.0856
7	(0.045894,0.145887)	1	0.0431
8	(0.059592,0.159586)	0	0.0187
9	(0.073291,0.173284)	0	0.0071
10	(0.086990,0.186983)	0	0.0024
⋮	⋮	⋮	⋮

In the column before last one denotes that the obtained interval covers the estimated value 0.05. Note that, for small values of m the left end of the interval is a negative number. Moreover, multiplying the last columns the real confidence level is obtained: 0.9471. This is smaller value than nominal 0.95. In consequence, by the application of the above method more correct populations are considered as wrong ones. Note that the expected length is 0.1 and equals given accuracy ε of estimation.

Clopper and Pearson (1934) give the confidence interval for θ, based on the exact distribution of ξ. Because

$$P_\theta\{\xi \le x\} = \beta(n-x, x+1; 1-\theta) \quad \text{and} \quad P_\theta\{\xi \ge x\} = \beta(x, n-x+1; \theta),$$

where $\beta(a,b;\cdot)$ denotes the CDF of Beta distribution with parameters (a,b), hence the confidence interval has the form $\left(\theta_L(x), \theta_U(x)\right)$, where

$$\theta_L(x) = \beta^{-1}\left(x, n-x+1; \frac{\alpha}{2}\right), \qquad \theta_U(x) = \beta^{-1}\left(x+1, n-x; 1-\frac{\alpha}{2}\right).$$

For $x = 0$ we take $\theta_L(0) = 0$, and for $x = n$ is taken $\theta_U(n) = 1$.

Here $\beta^{-1}(a,b;\cdot)$ denotes the quantile of the Beta distribution with parameters (a,b).

In the Table 2 there are given analogous results as in the Table 1, but for Clopper—Pearson confidence interval.

Table 2.

m	(left, right)		$P_{0.05}\{\xi = m\}$
0	(0.0000,0.0493)	0	0.0236
1	(0.0003,0.0740)	1	0.0909
2	(0.0033,0.0955)	1	0.1722
3	(0.0086,0.1154)	1	0.2144
4	(0.0151,0.1344)	1	0.1975
5	(0.0226,0.1526)	1	0.1435
6	(0.0308,0.1704)	1	0.0856
7	(0.0394,0.1876)	1	0.0431
8	(0.0485,0.2046)	1	0.0187
9	(0.0580,0.2212)	0	0.0071
10	(0.0677,0.2375)	0	0.0024
⋮	⋮	⋮	⋮

True confidence level of the classical confidence interval equals 0.9659, so it is higher than nominal 0.95. Unfortunately expected length of the confidence interval is 0.1119 and is bigger than postulated precision $2\varepsilon = 0.1$. To gain the accuracy, the sample size should be enlarged. It is easy to calculate that minimal sample size is $n = 90$. The results of calculations for that sample size are given in the Table 3.

Table 3.

m	(left, right)		$P_{0.05}\{\xi = m\}$
0	(0.0000,0.0402)	0	0.0099
1	(0.0003,0.0604)	1	0.0468
2	(0.0027,0.0780)	1	0.1097
3	(0.0069,0.0943)	1	0.1694
4	(0.0122,0.1099)	1	0.1939
5	(0.0183,0.1249)	1	0.1755
6	(0.0249,0.1395)	1	0.1309
7	(0.0318,0.1537)	1	0.0827
8	(0.0392,0.1677)	1	0.0451
9	(0.0468,0.1814)	1	0.0216
10	(0.0546,0.1949)	0	0.0092
11	(0.0626,0.2082)	0	0.0035
⋮	⋮	⋮	⋮

The true confidence level equals 0.9756, and its expected length is 0.0998. It is seen that expected length is smaller than required. For the sample of the size $n = 89$ expected length of the confidence interval equals 0.1004, which is a little bigger that assumed. To obtain precision exactly 2ε, the randomization should be applied in the following way

$$n = \begin{cases} 89, \text{ with probability } 0.3461 \\ 90, \text{ with probability } 0.6539 \end{cases}$$

Expected length equals now

$$0.0998 \cdot 0.3461 + 0.1004 \cdot 0.6539 = 0.1.$$

Of course, drawing sample size should be done before realization of the proper experiment. Any random number generator may be applied, for example the one in Excel.

In both cases, i.e. for sample size 89 and 90, the real confidence level is greater than nominal one. This inconvenience may be suppressed by second randomization. This randomization is applied in the construction of the confidence interval. If the number of observed failures is m two random numbers u_1, u_2 from uniform distribution $U(0,1)$ are drawn and ends of the confidence interval are obtained by solving two equations:

$$\textit{left}\quad u_1\beta(m-1,n-m+2;\theta)+(1-u_1)\beta(m,n-m+1;\theta)=0.025,\quad (L)$$
$$\textit{right}\quad u_2\beta(m+1,n-m;\theta)+(1-u_2)\beta(m+2,n-m-1;\theta)=0.975.\quad (R)$$

In that way, obtained confidence interval has the confidence level exactly 0.95.
In practice, the above method may be realized as follows. In the spreadsheet Excel three numbers are drawn according to the uniform $U(0,1)$ distribution:

$$u_0=0.3820, u_1=0.1006, u_2=0.5964.$$

Number u_0 is used to drawing a sample size: because u_0 is greater than 0.3461, hence the sample size is $n=90$.

In the experiment $m=10$ defective objects were observed. The two following equations are solved (Addin Solver in Excelu may be used):

$$\textit{left}\quad 0.1006\beta(9,82;\theta)+0.8994\beta(10,81;\theta)=0.025,\quad (L)$$
$$\textit{right}\quad 0.5964\beta(11,80;\theta)+0.4036\beta(12,79;\theta)=0.975.\quad (R)$$

The interval is obtained

$$(0.0535, 0.2012).$$

Drawn numbers u_0, u_1 and u_2 should be added to the report.
As it was mentioned, all calculations may be done in Excel. There are following useful functions.
BETADISTRIBUTION(x;alfa;beta;A;B): where alfa and beta are the parameters of the distribution. The function gives a values of CDF at point x. Numbers A and B defines a support of the distribution: default values are 0 and 1.

BETAINV(probability;alpha;beta;A;B): where alfa and beta are parameters of the distribution. The function gives the probability quantile of the Beta distribution. Numbers A and B defines a support of the distribution: default values are 0 and 1.

The confidence interval in the binomial model may be calculated in the following way.

	A	B
1	100	Sample size
2	10	Number of succeses
3	0.95	Confidence level
4	=IF(A2=0;0;BETAINV((1-A3)/2;A2;A1-A2+1))	Left end
5	=IF(A2=A1;1;BETAINV((1+A3)/2;A2+1;A1-A2))	Wright end

In cells A4 and A5 the values of the left and the right end of the interval are obtained.

The worksheet calculating a randomized confidence interval is a little bit complicated.

	A	B
1	100	Sample size
2	10	Number of succeses
3	0.95	Confidence level
4	0.1006	u1
5	0.5964	u2
6	0.1	Left end
7	=A4* BETAINV (A6;A2-1;A1-A2+2) +(1-A4)* BETAINV (A6;A2;A1-A2+1)-(1-A3)/2	equation (L)
8	0.1	Right end
9	=A5* BETAINV (A8;A2+1;A1-A2) +(1-A5)* BETAINV (A8;A2+2;A1-A2-1)-(1+A3)/2	equation (R)

To obtain the randomized confidence interval the Addin Solver should be used twice. Firstly, the goal is cell A7 by changing A6, secondly the goal is the cell A9 by changing A8. In cells A6 and A8 ends of randomized confidence interval are calculated. Numbers in A4 and A5 are obtained from a random number generator (Addis DataAnalysis in Excel).

More information on the confidence level may be fund in Zieliński (2009), and on randomized confidence intervals in Bartoszewicz (1996).

REFERENCES

Bartoszewicz, J. 1996: Wykłady ze statystyki matematycznej, wyd. II, PWN Warszawa.

Clopper C. J., Pearson E. S. 1934: The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial, Biometrika 26, 404-413.

Karliński W. 2003: http://nik.gov.pl/docs/kp/kontrolapanstwowa-200305.pdf

Zieliński R. 2009: Przedział ufności dla frakcji, Matematyka Stosowana 10(51), 51-67.