TEKA Kom. Mot. Energ. Roln., 2005, 5, 205–211

ESTIMATION OF A SMALL FRACTION UNDER NORMALITY

Joanna Tarasińska

Department of Applied Mathematics, Agricultural University of Lublin

Summary. We are interested in the fraction *p* of units for which a certain normally distributed characteristic X exceeds a permissible value L. When *p* and the sample size *n* are small, the fraction in the sample can not be used as the estimator of *p*. The aim of the paper is to encourage the practitioners-non statisticians to use in such a situation different estimators than simple „fraction in the sample".

Key words: normal distribution, estimator of a fraction, robustness

INTRODUCTION

In many situations we have a random variable X which is normally distributed ($X \sim N(\mu,\sigma^2)$) and we are interested in an estimation of the fraction of units for which the event $\{X > L\}$ happens. L can be, for example, the maximal permissible value of X and in such a case we want to estimate the fraction of defective units. It is a problem of an estimation of the probability $p = \Pr(X > L)$. Having the random sample $X_1 X_2,\ldots X_n$ we can estimate *p* just by the fraction of defective units in the sample, it means $\tilde{p} = \frac{k}{n}$, where *k* is the number of X_i being greater than *L*.

Such an estimator ignores the fact of normality of *X*. Additionally, it needs large sample size when *p* is small. Let us consider for example $p \approx 0.05$ and $n = 10$. $\tilde{p}$ in such a case is absolutely useless. It is known that there exist better estimators.

Considering $p = \Pr(X > L) = \Phi\left(\frac{\mu - L}{\sigma}\right)$ we have for example the maximum likelihood estimator [Patel and Read 1996]:

$$\hat{p} = \Phi\left(\frac{\overline{X} - L}{S}\right), \tag{1}$$

where $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$, $S^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$. $\Phi(\cdot)$ is the cumulative distribution function read from normal tables.

There also exists the "best" unbiased estimator of p which has the smallest variance in the class of unbiased estimators. It can be calculated [Lieberman and Resnikoff 1995, Patel and Read 1996] by the formula

$$\hat{\hat{p}}=\begin{cases}0 & \text{if } a<0\\ I_a\left(\frac{n}{2}-1,\frac{n}{2}-1\right) & \text{if } 0\le a\le 1\\ 1 & \text{if } a>1\end{cases}, \tag{2}$$

where $a=0.5\left[1+\frac{\sqrt{n}(\overline{X}-L)}{(n-1)S^*}\right]$, $S^{*2}=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$,

$I_a(p,q)=B^{-1}(p,q)\int_0^a t^{p-1}(1-t)^{q-1}dt$ is the incomplete beta function ratio and $B(p,q)$ is the complete beta function $B(p,q)=\int_0^1 t^{p-1}(1-t)^{q-1}dt$.

So, contrary to $\hat{p}$, $\hat{\hat{p}}$ demands rather troublesome calculations.

It is easy to find a formula for (1) and (2) in the situation when $p=\Pr(X<L)$. In such a case we have $p=\Phi\left(\frac{L-\mu}{\sigma}\right)$, $\hat{p}=\Phi\left(\frac{L-\overline{X}}{S}\right)$, $\hat{\hat{p}}$ is the same as in (2) with $a=0.5\left[1+\frac{\sqrt{n}(L-\overline{X})}{(n-1)S^*}\right]$.

Example (theoretical one, the idea taken from Bowker and Lieberman 1959, p.57:

The clearance between the external shaft diameter and the internal bearing diameter can be assumed to be normally distributed. The minimum permissible clearance is 0.005 inches.

For a random sample of 5 pairs of shaft and mating bearing we get the following measurements of clearance (in inches): 0.0080, 0.0079, 0.0140, 0.0081, 0.0094.

We have

$\overline{X}=0{,}00948$, $S\approx 0{,}002325$, $S^*\approx 0{,}002599$, $a=0{,}01828$ so $\hat{p}=0{,}027$ and $\hat{\hat{p}}=0{,}004$.

Several authors have compared $\hat{p}$ and $\hat{\hat{p}}$ [Zacks and Eden 1966, Brown and Rutemiller 1973, Gertsbakh and Winterbottom 1991] taking into consideration their MSE (mean squared error) and bias of $\hat{p}$. It turns out for example that, for $p\approx 0.05$, $\hat{p}$ is nearly unbiased.

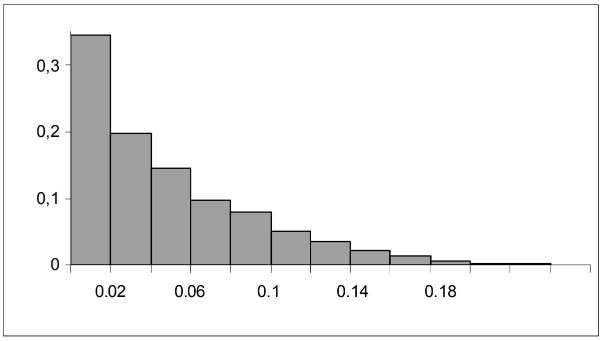

Fig. 1. The histogram for $\hat{p}$, $n = 10$, $p = 0.05$

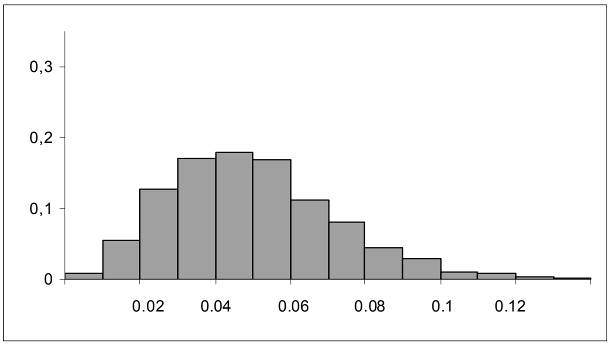

Fig. 3. The histogram for $\hat{p}$, $n = 50$, $p = 0.05$

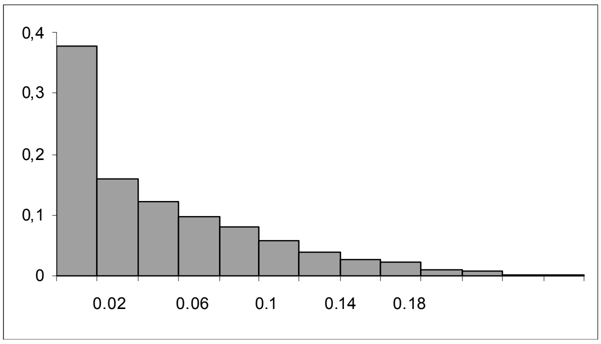

Fig. 2. The histogram for $\hat{\hat{p}}$, $n = 10$, $p = 0.05$

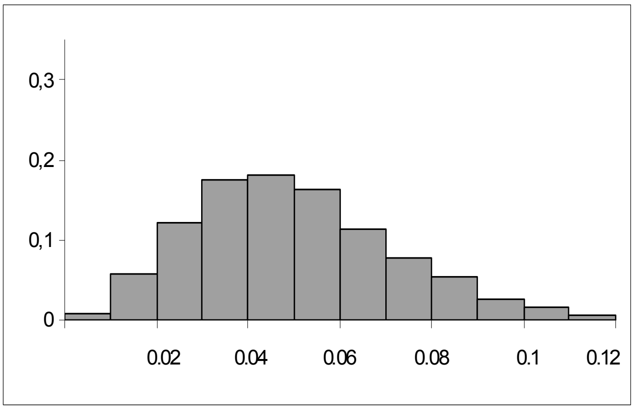

Fig. 4. The histogram for $\hat{\hat{p}}$, $n = 50$, $p = 0.05$

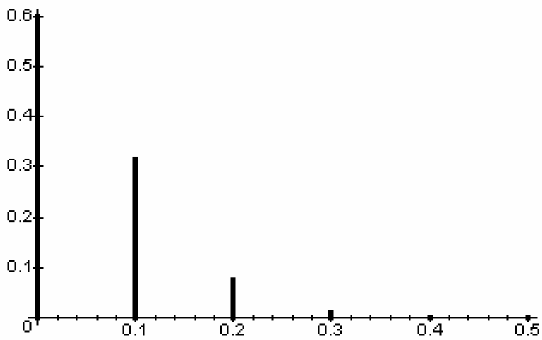

Fig. 5. The distribution of $\widetilde{p}$, n=10, p=0.05

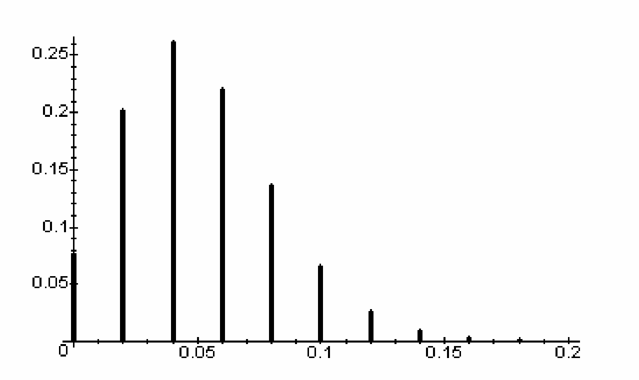

Fig. 6. The distribution of $\widetilde{p}$, $n = 50$, $p = 0.05$

Of course, MSE does not say everything about the distribution. To check whether the distributions of $\hat{p}$ and $\hat{\hat{p}}$ differ much or not, some simulations were done.

For $n = 10$ and 50, $p = 0.05$ five thousands random samples from standard normal distribution were generated and $\hat{p}$ and $\hat{\hat{p}}$ were computed (with $L = \Phi^{-1}(1-p)$). Their histograms are presented in Figures 1,2,3 and 4. They can be compared with the distribution of $\tilde{p}$ given in the Figures 5 and 6. Of course $\Pr\left(\tilde{p} = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k}$.

Of course it can be seen from Fig. 5 that $\tilde{p}$ is completely useless in the case of small sample size.

Table 1 contains the MSE and bias of $\hat{p}$ calculated from simulations. The MSE for $\hat{p}$ was calculated by the formula $MSE = \frac{1}{5000}\sum_{i=1}^{5000}(\hat{p}_i - 0.05)^2$, bias by the formula $\frac{1}{5000}\sum_{i=1}^{5000}\hat{p}_i - 0.05$. The MSE for $\hat{\hat{p}}$ is equal to the variance of $\hat{\hat{p}}_i$ because $\hat{\hat{p}}$ is unbiased. From Table 1 it can be seen that $\hat{p}$ is superior to $\hat{\hat{p}}$ when MSE is the criterion.

Table 1. The MSE's and bias of $\hat{p}$

	$\hat{p}$		$\hat{\hat{p}}$
	MSE	bias	MSE
n = 10	0.002100	-0.016	0.002662
n = 50	0.000491	0	0.000495

ROBUSTNESS OF $\hat{p}$ AND $\hat{\hat{p}}$ TO DEVIATIONS FROM NORMALITY

Both estimates $\hat{p}$ and $\hat{\hat{p}}$ can be used when X is normally distributed. But what happens if not? Let us assume $X \sim \mu + \sigma \cdot t_3$, where t_3 is Student's t distribution with three degrees of freedom. In such a case the variance of X is three times larger than under normality. Of course now $\hat{\hat{p}}$ is not the best unbiased estimator and $\hat{p}$ is not the maximum likelihood one.

What are their properties? How much worse are they? To answer these questions 5000 samples of size $n = 10$ and $n = 50$ were generated in the case $p = 0.05$.

The Figures 7 and 8 present the histograms of $\hat{p}$ and $\hat{\hat{p}}$.

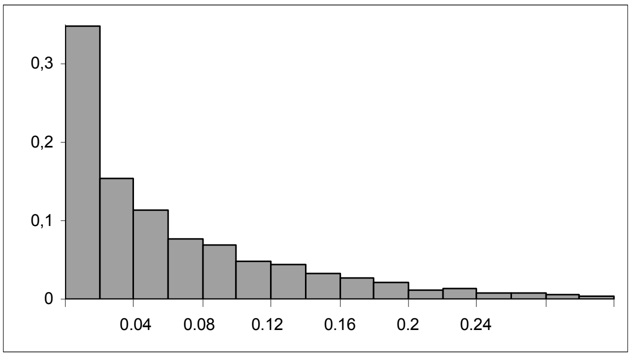

Fig. 7. The histogram for $\hat{p}$, $n = 10$, $p = 0.05$, without normality

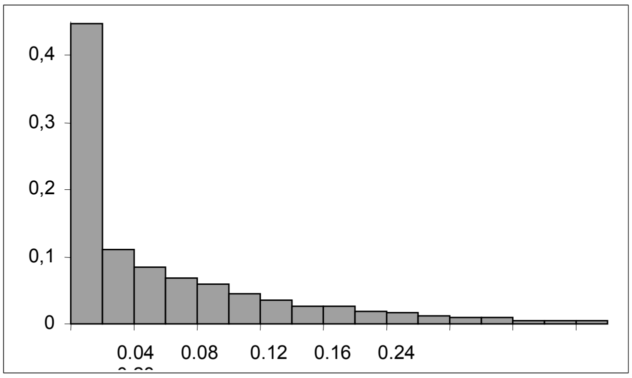

Fig. 8. The histogram for $\hat{\hat{p}}$, $n = 10$, $p = 0.05$, without normality

Table 2. The MSE's and biases without normality

	$\hat{p}$		$\hat{\hat{p}}$	
	MSE	bias	MSE	bias
n = 10	0.006337	+0.0169	0.01960	+0.0129
n = 50	0.003155	+0.0229	0.01776	+0.0224

So, $\hat{p}$ has got less mean square error and can be considered as better than $\hat{\hat{p}}$ when the probability which is to be estimated is near 0.05.

LARGE SAMPLE SIZE

When sample size n is large enough, the estimate $\tilde{p}$ can be used. Let us compare it with $\hat{p}$. Let us assume we are interested in he probability of attaining the relative error not greater than a certain acceptable value ε. That is let us compare the probabilities

$\Pr\left(\left|\frac{\hat{p}-p}{p}\right| \le \varepsilon\right)$ and $\Pr\left(\left|\frac{\tilde{p}-p}{p}\right| \le \varepsilon\right)$. Table 3 gives the results for n = 200, p = 0.05 and ε = 0.1, 0.2, 0.3.

$\Pr\left(\left|\frac{\hat{p}-p}{p}\right| \le \varepsilon\right)$ is calculated under assumption of normality using normal approximation to non-central t distribution ([5]). $\Pr\left(\left|\frac{\tilde{p}-p}{p}\right| \le \varepsilon\right)$ does not depend on the distribution of X and is calculated using binomial probability.

Table 3. Comparison of $\hat{p}$ and $\tilde{p}$, $n = 200, p = 0.05$

ε	0.1	0.2	0.3
$\Pr\left(\left\|\frac{\hat{p}-p}{p}\right\| \le \varepsilon\right)$	0.34	0.63	0.82
$\Pr\left(\left\|\frac{\tilde{p}-p}{p}\right\| \le \varepsilon\right)$	0.37	0.58	0.75

So, when sample size is large enough to use $\tilde{p}$ just **this** estimator should be preferable as it is as good as $\hat{p}$ under normality and, additionally, it is completely independent upon the distribution of X.

REFERENCES

Bowker A.H., Lieberman G.J., 1959: Engineering Statistics. Prentice-Hall Inc.

Brown G.G., Rutemiller H.C., 1973: The efficiencies of maximum likelihood and minimum variance unbiased estimators of fraction defective in the normal case. Technometrics,15, 849-855.

Gertsbakh I., Winterbottom A., 1991: Point and interval estimation of normal tail probabilities. Communications in Statistics-A20 (4), 1497-1514.

Lieberman G.J., Resnikoff G.J., 1955: Sampling plans for inspection by variables. Journal of the American Statistical Association 50, 457-516.

Patel J.K., Read C.B. ,1996: Handbook of the normal distribution. 1996, Marcel Dekker inc., New York.

Zacks S., Eden M., 1966: The efficiences In small samples of the maximum likelihood and best unbiased estimators of reliability functions. Journal of the American Statistical Association, 61, 1033-1051.