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 permissable 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 permissable 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 $\hat{p} = \frac{k}{n}$, 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 $n = 10$. $\hat{p}$ 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]:

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

(1)
where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, $S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$. $\Phi()$ 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

$$\hat{p} = \left\{ \begin{array}{ll} 0 & \text{if } a < 0 \\ I_a \left( \frac{n}{2} - 1, \frac{n}{2} - 1 \right) & \text{if } 0 \leq a \leq 1 \\ 1 & \text{if } a > 1 \end{array} \right., \quad (2)$$

where $a = 0.5 \left[ 1 + \frac{\sqrt{n} (\bar{X} - L)}{(n-1)S^*} \right]$, $S^* = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$, $I_a(p,q) = B^{-1}(p,q) \int_{0}^{a} t^{p-1}(1-t)^{q-1} dt$ 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 $\hat{p}$, $\hat{p}$ 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 $\hat{p} = \Phi \left( \frac{L - \mu}{\sigma} \right)$, $\hat{p} = \Phi \left( \frac{L - \bar{X}}{S} \right)$, $\hat{p}$ is the same as in (2) with

$$a = 0.5 \left[ 1 + \frac{\sqrt{n} (\bar{X} - L)}{(n-1)S^*} \right].$$

**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 $\bar{X} = 0.00948$, $S \approx 0.002325$, $S^* \approx 0.002599$, $a = 0.01828$ so $\hat{p} = 0.027$ and $\hat{p} = 0.004$.

Several authors have compared $\hat{p}$ and $\hat{p}$ [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. 2. The histogram for $\hat{p}$, $n = 10$, $p = 0.05$

Fig. 3. The histogram for $\hat{p}$, $n = 50$, $p = 0.05$
Fig. 4. The histogram for $\hat{p}, n = 50, p = 0.05$

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

Fig. 6. The distribution of $\hat{p}, n = 50, p = 0.05$
ESTIMATION OF A SMALL FRACTION UNDER NORMALITY

Of course, MSE does not say everything about the distribution. To check whether the distributions of \( \hat{p} \) and \( \tilde{p} \) 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 \( \tilde{p} \) 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 \( \text{MSE} = \frac{1}{5000} \sum_{i=1}^{5000} (\hat{p}_i - 0.05)^2 \), bias by the formula \( \frac{1}{5000} \sum_{i=1}^{5000} \hat{p}_i - 0.05 \). The MSE for \( \hat{p} \) is equal to the variance of \( \hat{p}_i \) because \( \hat{p} \) is unbiased. From Table 1 it can be seen that \( \hat{p} \) is superior to \( \tilde{p} \) when MSE is the criterion.

<table>
<thead>
<tr>
<th>( n )</th>
<th>( \hat{p} ) MSE</th>
<th>( \hat{p} ) bias</th>
<th>( \tilde{p} ) MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.002100</td>
<td>-0.016</td>
<td>0.002662</td>
</tr>
<tr>
<td>50</td>
<td>0.000491</td>
<td>0</td>
<td>0.000495</td>
</tr>
</tbody>
</table>

ROBUSTNESS OF \( \hat{p} \) AND \( \tilde{p} \) TO DEVIATIONS FROM NORMALITY

Both estimates \( \hat{p} \) and \( \tilde{p} \) 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 \( \tilde{p} \) 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 \( \tilde{p} \).
So, \( \hat{p} \) has got less mean square error and can be considered as better than \( \tilde{p} \) 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 the probability of attaining the relative error not greater than a certain acceptable value \( \varepsilon \). That is let us compare the probabilities of attaining the relative error not greater than \( \varepsilon \) with the sampling distribution of \( \hat{p} \) and \( \tilde{p} \).
\[ \Pr \left( \left\| \hat{p} - p \right\| \leq \varepsilon \right) \quad \text{and} \quad \Pr \left( \left\| \tilde{p} - p \right\| \leq \varepsilon \right) \]. Table 3 gives the results for \( n = 200, p = 0.05 \) and \( \varepsilon = 0.1, 0.2, 0.3 \).

\[ \Pr \left( \left\| \hat{p} - p \right\| \leq \varepsilon \right) \] is calculated under assumption of normality using normal approximation to non-central \( t \) distribution ([15]). \( \Pr \left( \left\| \tilde{p} - p \right\| \leq \varepsilon \right) \) does not depend on the distribution of \( X \) and is calculated using binomial probability.

<table>
<thead>
<tr>
<th>( \varepsilon )</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Pr \left( \left| \hat{p} - p \right| \leq \varepsilon \right) )</td>
<td>0.34</td>
<td>0.63</td>
<td>0.82</td>
</tr>
<tr>
<td>( \Pr \left( \left| \tilde{p} - p \right| \leq \varepsilon \right) )</td>
<td>0.37</td>
<td>0.58</td>
<td>0.75</td>
</tr>
</tbody>
</table>

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