Robust Statistical Methods for Applications in Quality Control

María Belén Allasia*
Universidad Nacional de Rosario, Rosario, Argentina mallasia@fcecon.unr.edu.ar

Fernanda Mendez
Universidad Nacional de Rosario, Rosario, Argentina nandixx@hotmail.com

Marta Quaglino
Universidad Nacional de Rosario, Rosario, Argentina mquaglino@fcecon.unr.edu.ar

Marta Ruggieri
Universidad Nacional de Rosario, Rosario, Argentina mruggieri@express.com.ar

Liliana Severino
Universidad Nacional de Rosario, Rosario, Argentina lilianaseve@gmail.com

Abstract

In this article it is presented some approaches to the theory of robust estimation, particularly useful in quality control area, field in which they are potentially useful. The idea of robustness is associated with insensitivity to small deviations from the assumptions, guaranteeing the same efficiency than conventional methods if they are satisfied. In this paper it is described and compared the position classical estimator: sample mean, and robust estimators: Median, Trimmed Mean, Huber M-estimator, Bisquare M-estimator. The disadvantages entailed by the classical approach are shown when the optimal conditions are not given, demonstrating the advantages of robust estimators in a practical quality control application dataset from a metallurgical company of Gran Rosario, through the calculation of different estimators. Moreover, in the processes that meet the assumptions required for a classical statistical analysis, it is proved that the use of robust estimators have the same status as the classics. In addition, in both subprocesses studied, the confidence intervals obtained with the classical estimator are wider. So, if a quality control study is performed a posteriori considering these limits, it would be much more liberal in terms of precision of the chosen method, implying a risk of missing some observations which might suggest that the process is no longer under control. In these situations, it was proved that robust estimators provide a more appropriate notion of habitual behavior of data. It is expected that future productivity observations will be evaluated in control charts where the limits will be determined based on robust estimators. The importance of these proposed methods is the great potential of application they have, because, although there are substantial methodological contributions, their use in the exercise of statistical analysis is not yet widespread.

Keywords: Robust Methods, Statistical Inference, Statistical Quality Control.

1. Introduction

All statistical methods are based in part on observations and, explicitly or implicitly, on a number of assumptions about the underlying situation. Generally, these assumptions point to the formalization of what the statistician knows about data analysis or modeling problem he faces and at the same time, aims to make manageable the resulting model both from the theoretical and computational point of view. However, it is known that the resulting formal models are simplifications of reality and its validity is, at best, approximate.

Even in the simplest cases, there are assumptions about randomness, independence and the distribution of the observations. For example, it is common to assume that the data the data are normally distributed.

The assumption of normality has been the framework for classical statistical methods, which are based on the premise that "the assumption of normality is exactly fulfilled". The main reason that such distribution is assumed is that for many real situations gives an approximate representation. At the

same time, it is quite convenient because it allows to derivate explicit formulas for optimal statistical methods: maximum likelihood, likelihood ratio test, sampling distribution of estimators t

In practice, it often happens that the behavior of the distribution data set has an "approximately" normal distribution. The main discrepancy may be caused by a small proportion of observations that deviate from data concentration. These atypical observations are called outliers and may appear due to different reasons, such as: errors in measurement instruments, variations in the condition under which the data were gathered, errors in data transmission or transcription procedures.

Standard procedures do not always provide a suitable tool since they are optimal only when the assumptions are exactly met and even a small deviation in the data distribution may distort the conclusions reached.

In most practical applications, the underlying distribution of the observations can only be determined "approximately". One way to determine approximate distributions is to consider contaminated environments according to distribution function:

$$\mathcal{F}_{\theta\varepsilon} = \{F \in \mathcal{F} / F = (1-\varepsilon)F_\theta + \varepsilon G,\ \ G \in \mathcal{G}\ \ \varepsilon \in [0,1]\} \qquad (1)$$

where $\mathcal{G}$

The notion of robustness is associated with "insensitivity to small deviations from the assumptions", ensuring the same efficiency of conventional methods when the assumptions are accomplished.

Robust statistical methods have a long history, at least from the end of the nineteenth century. The most important progress in this area occurred in the 1960s and early 1970s, when this approach has begun to have an impact beyond the domain of the robustness specialists. Since then, it seems to be a growing general awareness about the dangers caused by the outliers and the unreliability of the assumptions of classical statistical models. However, they remain underused and unknown, even by most of applied statisticians, data analysts and scientists which could may profit from them.

This work aims towards progress in the study of robust methods, potentially useful in multiple scenarios for continuous improvement. Within its objectives has a deepen study of properties of Statistical Methods for On-Line and Off-Line Quality Control, in order to analyze its performance in unconventional situations such as non-normality data or the presence of atypical observations, frequent situations on the industrial applications area.

2. Material and Methods

2.1. Position Estimators

2.1.1. Sample Mean

In the position parametric model

$$x_i = \mu + u_i \quad (i = 1, \dots, n) \qquad (2)$$

$$x_1, x_2, \dots, x_n \overset{iid}{\sim} F \ \ con\ F \in P_\mu = \{F_\mu / F_\mu(x) = F_0(x-\mu)\} \qquad (3)$$

it is assumed that F_0 is the distribution function of a Normal distribution: $N(0, \sigma^2)$ -with known σ^2-, and if it is used the classical method of maximum likelihood

$$\hat{\mu}_n = \arg\max_\mu L(x_1, x_2, \dots, x_n; \mu) = \arg\max_\mu \prod_{i=1}^{n} f_\mu(x_i) \qquad (4)$$

it is obtained as estimator $\hat{\mu}_n = \frac{1}{n}\sum x_i = \bar{x}_n$, the sample mean.

However, in many practical applications it can be ensured that measurement errors are approximately normally distributed, at most. Therefore, imports the behavior of the $\bar{x}_n$ estimator under this situation. Considering a contaminated environment as defined in (1), whereas the observations come from a normal distribution with probability $(1-\varepsilon)$, and an unknown mechanism (G) with probability ε:

$$F = (1-\varepsilon)F_\mu + \varepsilon G \quad (5)$$

where $F_\mu = N(\mu, \sigma^2)$ and G can be any distribution. For example, if G is another normal with more variance or a different average, F is called a mixture of normal.

So, if $E_G(x) = E_{F_\mu}(x) = \mu$, then $Var_F(\bar{x}_n) = \frac{Var_F(x)}{n} = \frac{(1-\varepsilon)\sigma^2 + \varepsilon Var_G(x)}{n}$. This reflects the extreme sensitivity of $\bar{x}$ to a contamination of size ε, which may largely increase its variance ($Var_G(x)$ could be extremely high or even infinite).

2.1.2. Sample Median

Considering the ordered sample observations, $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, the sample median $(\tilde{x})$ is given by:

$$\tilde{x}_n = \begin{cases} x_{(k)} & if\ n\ is\ an\ odd\ number \\ \dfrac{x_{(k)} + x_{(k+1)}}{2} & if\ n\ is\ an\ even\ number \end{cases} \quad (6)$$

where $k = \left[\frac{n+1}{2}\right]$.

Conceptually, the median is the value of the variable which leaves 50% of the observations beneath itself, ie $P(x \leq \tilde{x}) = F(\tilde{x}) = 0.50$

2.1.3. Trimmed Mean

It is a position estimator that removes a proportion of the minor and major sample observations. Being $\alpha \in \left[0; \frac{1}{2}\right)$ $m = [(n-1)\alpha]$ α is defined as

$$\bar{x}_\alpha = \frac{1}{n-2m} \sum_{i=m+1}^{n-m} x_{(i)} \quad (7)$$

ie, the first and last m discard $\alpha = 0$ and $\alpha \to 0.5$ ely.

2.1.4. M-Estimators

Considering the position model (2) and assuming that F_μ x_i $f_\mu = F'_\mu$

$$L(x_1, x_2, \ldots, x_n; \mu) = \prod_{i=1}^{n} f_\mu(x_i) = \prod_{i=1}^{n} f_0(x_i - \mu) \quad (8)$$

where f_0 u_i

The maximum likelihood estimator (MLE) of μ (8):

$$\hat{\mu}_{MV} = \hat{\mu}(x_1, x_2, \ldots, x_n) = \arg\max_\mu L(x_1, x_2, \ldots, x_n; \mu) \quad (9)$$

If f_0 and considering:

$$\rho = -\log f_0(u) + \log f_0(0) \quad (10)$$

solving equation (9) is equivalent to:

$$\hat{\mu}_n = \arg\min_\mu \sum_{i=1}^{n} \rho(x_i - \mu) \quad (11)$$

[1]Note: $[x]$ is the integer part function, meaning the major integer value less or equal than x.

If the exact distribution is known, the maximum likelihood estimator can be used, which is "optimal" in the sense that it has minimal asymptotic variance. Generally F_0 is approximately known, so F_0 (1964) defined the M-estimators for the position model as in (11) , where the ρ function f_0 properties (see Maronna, 2006)

Considering that $\rho' = \psi$, $S(\mu) = \sum_{i=1}^{n} \rho(x_i - \mu)$ can be found through their critical points, so $\hat{\mu}_n$

$$\sum_{i=1}^{n} \psi(x_i - \hat{\mu}) = 0 \qquad (12)$$

Furthermore, a position M-estimator can be viewed as a weighted average. In most cases of interest $\psi(0) = 0$ $\psi'(0)$, ψ neal at the origin. Then, the weight function is:

$$W(x) = \begin{cases} \dfrac{\psi(x)}{x} & si\ x \neq 0 \\ \psi'(0) & si\ x = 0 \end{cases} \qquad (13)$$

Equation (13) can be written as:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \qquad w_i = W(x_i - \hat{\mu}) \qquad (14)$$

As in general $W(x)$ is a non-increasing $|x|$ receive smaller weights. Notice that, although (14) appears to be an explicit expression for $\hat{\mu}$, the weights on the right side of the equation also depend $\hat{\mu}$

3. Results

A dataset of quality application from a metallurgical company of Gran Rosario has been analyzed.

To illustrate the use of the studied position estimators, two subprocesses data corresponding to saw cutting and head wrought were considered. There were 78 complete observations in terms of number of parts and work time, properly recorded in the studied period.

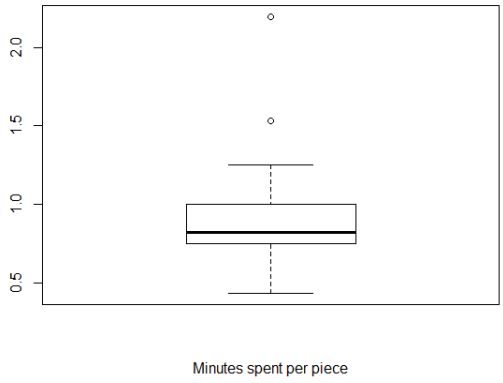

Figure 3. Number of minutes spent per piece on the saw cutting subprocess.

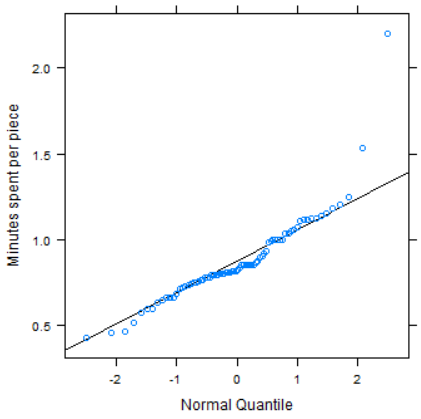

Figure 4. Normal Probability Plot Minutes spent per piece on the saw cutting subprocess

As it can be seen, this dataset does not meet the necessary assumptions for a classical statistical analysis. The graphs above show that the distribution of time spent on saw cutting subprocess presents right side asymmetry, so it could not be assumed that it comes from a normal distribution. In the modified box-plot (Figure 1) it is revealed that two values are potential outliers.

The following table shows the calculated position estimates and their 95% confidence intervals:

Table 1. Position estimators for minutes spent per piece on the saw cutting subprocess

	$\hat{\mu}$	$\hat{v}$	$\sqrt{\hat{v}(\hat{\mu})}$	$IC_{90\%}(\mu)$	
Mean	0.8775	0.0600	0,0277	0.8318	0.9230
Median	0.8205	∞	∞	−∞	∞
Trimmed Mean	0.8612	0.0654	0,0290	0.8136	0.9089
Huber M-estimator	0.8567	0.0275	0,0188	0.8258	0.8876
Bisquare M-estimator	0.8494	0.0216	0,0166	0.8220	0.8767

Regarding the point estimate of the position parameter, the sensitivity of classical estimator in the presence of positive extreme values is revealed. The sample mean gives a value of 0.8775 minutes per part, which is higher than the values obtained for the remaining estimators (even higher than the upper limit of the 95% confidence interval of Bisquare M-estimator). Moreover, the magnitude of the confidence intervals of M-estimators is considerably less, demonstrating that they are more accurate estimators in this scenario of industrial production.

By studying the minutes per piece used for the head wrought per side with the customer logo subprocess revealed again that the observations might not fit the classical assumptions. They vary in 1.2619 minutes per piece range, while central 50% data is concentrated in a 0.09 minutes per piece range only. Figure 3 shows the existence of atypical observations in both right and left sides of the distribution. It particularly evidences that there is an outlier at the right side of the distribution with a much higher magnitude than the others, so it is predicted that the classical estimator will be affected by this observation, giving a higher value than the remaining estimators. Given these characteristics of the variable and considering the Normal Probability Plot (Figure 4), it is concluded that it does not meet the normality assumption required for a classical analysis.

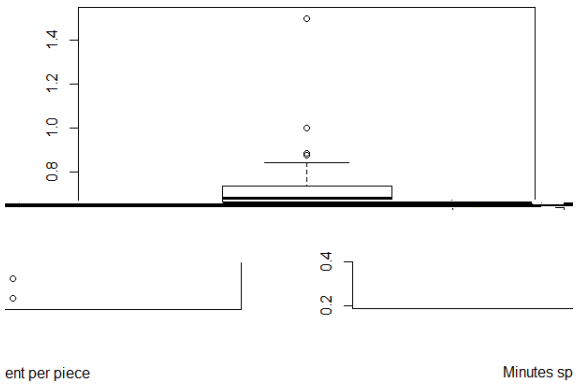

Figure 5. Number of minutes spent per piece on the head wrought per side with the customer logo subprocess

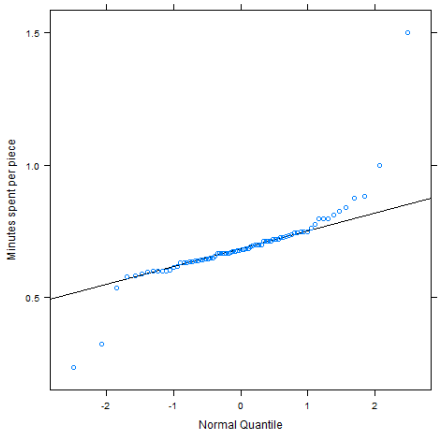

Figure 6. Normal Probability Plot - Minutes spent per piece on the head wrought per side with the customer logo subprocess

By calculating the position estimates and comparing them, the same observations made for the previously studied subprocess are obtained. The sample mean estimator is affected by the outliers magnitude and its confidence interval is wider than the remaining confidence intervals estimations. Therefore, the sample mean is a less accurate estimate.

The above observations lead to think that it would be inappropriate to conduct a statistical control process based on the sample mean as a position estimator, since it is not adequate to describe the central position of the data in the studied scenarios.

Table 2. Position estimators for minutes spent per piece on the head wrought per side with the customer logo subprocess

	$\hat{\mu}$	$\hat{v}$	$\sqrt{\hat{v}(\hat{\mu})}$	$IC_{90\%}(\mu)$	
Mean	0,6931	0,0186	0,0154	0,6677	0,7185
Median	0,6796	∞	∞	$-\infty$	∞
Trimmed Mean	0,6855	0,0226	0,0170	0,6575	0,7136
Huber M-estimator	0,6844	0,0044	0,0075	0,6721	0,6967
Bisquare M-estimator	0,6848	0,0033	0,0065	0,6741	0,6955

4. Conclusions

It has been show, by studying two subprocesses of a particular piece production from a metallurgical company of Gran Rosario, that the results for variables that record the manufacture time spent per piece does not often exhibit behaviors that can suit the normality assumption. This is significantly reflected when position estimates are calculated to evaluate the accuracy of the production process. As can be noted, in the first subprocess studied, the sample mean estimator is approximately equal to the upper alert limit of the Bisquare M-estimator 95% confidence interval. In addition, in both subprocesses the confidence intervals obtained to the sample mean are wider. So, if a quality control study is performed a posteriori considering these limits, it would be much more liberal in terms of precision of the chosen method, implying a risk of missing some observations which might suggest that the process is no longer under control.

In these situations, robust estimators are recommended because they provide a more appropriate notion of habitual behavior of data. It is expected that future productivity observations will be evaluated in control charts where both alert and action limits will be determined according to the Huber or Bisquare M-estimators normal distribution quantiles.

5. References

Hampel, F. (1968). Contributions to the theory of robust estimation. PhD. Thesys, University of California, Berkeley.

Hampel, F. (1974). The Influence Curve and Its role in robust estimation. The Annals of Statistics, 69, 383-393.

Hampel, FR, Ronchetti, EM, Rousseeuw, PJ, & Stahel, WA (1986) Robust Statistics. The Approach Based on Influence Functions New York. John Wiley & Sons.

Huber, P. (1964). Robust estimation of a location parameter. The Annals of Mathematical Statistics, 35, 73-101.

Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the Fifth Berkeley Symposium on Probability and Statistics Mathematics. 1, pp. 221-233. University of California Press.

Huber, P., & Ronchetti, E. (2009). Robust Statistics (Second ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Maronna, R., Martin, R., & Yohai, V. (2006) Robust Statistics. Theory and Methods Chichester, West Sussex, England. John Wiley and Sons, Ltd.

Montgomery, D. (2005). Introduction to Statistical Quality Control (Fifth ed.). John Wiley & Sons, Inc.

Bartes Prat, A., Tort Llabrés Martorell, X., Grima Cintas, P., & Pozueta Fernández, L. 2000). Statistical methods. Control and quality improvement Mexico. Alfaomega Group Editor.

Tuckey, J. (1970). Exploratory Data Analysis. Mimeographed Preliminary Edition.

Tukey, J. (1962). The future of data analysis. The Annals of Mathematical Statistics, 33, 1-67.