SCIENTIFIC PUBLICATIONS OF THE STATE UNIVERSITY OF NOVI PAZAR
SER. A: APPL. MATH. INFORM. AND MECH. vol. 12, 1 (2020), 37-46

Empirical Distribution Function as a Tool in Quality Control

V. Jevremović, A. Avdović

Abstract: In this paper, we are introducing two new methods for Quality Control. Those methods rely on using the Empirical distribution function of a given sample (or several samples). The first method we use for one sample, i.e. we can determine if each given sample is adequate. We use the second method to determine whether several samples are "in control" or not. For the sample to be "in control" state, the normal distribution is required.

Keywords: quality control, X-bar charts, empirical distribution function, Kolmogorovs distribution, control band

1 Introduction

Controlling quality is very important in many areas of application, and not only in controlling some production processes. We can follow traces of Quality control in everyday life all around us. There are appropriate methods for Quality control, even in teaching or when dealing with customer satisfaction. That is why after discovering Quality control many methods have been developed and improved. In this paper, we suggest the use of empirical distribution function (EDF) as a tool in Quality control. The properties of EDF are well known from the Glivenko-Kantelli theorem (central theorem of mathematical statistics) and Kolmogorovs theorem. Based on these properties, it is possible to use EDF as a tool in Quality control.

In the first part, the paper deals with some basic facts about Quality control. In the second part, we shall explain the main idea of the paper, i.e. how to use EDF as a tool in Quality control, and we shall illustrate the methods with numerical examples.

2 Quality control principles and methods

"Quality control is a procedure or set of procedures intended to ensure that a manufactured product or performed service adheres to a defined set of quality criteria or meets the re-

Manuscript received January 25, 2020. ; accepted April 11,2020.

Vesna Jevremović, Atif Avdović are with the State University of Novi Pazar, Department of Mathematical Sciences, Novi Pazar, Serbia

quirements of the client or customer" [5].

The word "Quality" is ancient, its origin dates back to the period BCE, but we can consider that the first exact approach to quality control, based on Probability and Statistics, was made by Walter Shewhart in 1924 during his work for Bell laboratories. He noticed that causes of variation in the quality of a product or process are of two kinds: common causes and assignable causes.

Common causes of variation are random ones that we cannot identify, and the differences in quality due to this kind of situation remain in acceptable limits. Assignable causes of variation are the ones that we can identify and eliminate.

2.1 Control charts

The control chart is a graph used in statistical process control that shows whether a sample of data falls within the normal range of variation. Each control chart consists of the central line (CL) and upper (UCL) and lower control limits (LCL). These limits separate common from assignable causes of variation. There is a difference in control charts for continuous random variables and the discrete ones. We shall deal with the continuous ones.

Control charts for variables are:

- X-bar charts which monitor the mean or average value of product characteristic,
- R-charts, which monitor the range of values of product characteristic in the sample,
- S-charts which monitor sample variance,
- CUSUM and EWMA charts in [2].

For our purpose, we give a short explanation for the X-bar charts. As stated previously, we use X-bar chart to check the average value of a characteristic. Since the normal distribution is assumed, we refer to the following theorems:

Theorem 2.1. *If for some characteristic X having a normal distribution $N(m,\sigma^2)$, we have a random sample of size n, $\vec{X} = (X_1, X_2, ..., X_n)$, then the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ has the normal distribution $N(m, \frac{\sigma^2}{n})$.*

Theorem 2.2. *If for some characteristic X having a normal distribution $N(m,\sigma^2)$, we have a random sample of size n, then the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is unbiased estimator for m, with minimal variance among all unbiased estimators for m.*

X-bar chart uses 6σ rule, which means that for normal distribution $N(m,\sigma^2)$ we have $P(X \in (m-3\sigma, m+3\sigma)) = 0.997$. This way the central line is m and upper and lower control lines are, respectively, $m-3\sigma$ and $m+3\sigma$. On the chart, we plot the means of samples in order of appearance. In figure 1, an example of a X-bar chart is given.

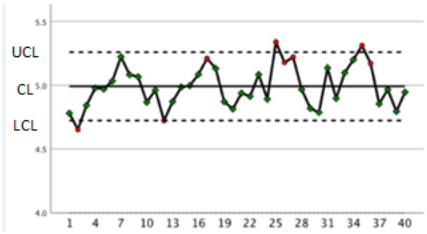

Figure 1: X-bar chart

Decision rule in Quality control when we use X-bar chart is that the process is in control if the sample means are all between UCL and LCL, and is out of control when at least one sample mean is greater than UCL, or smaller than LCL.

The use of X-bar control chart is a repeated testing procedure, at each step, we test the hypothesis $H_0(m = m_0)$, where m_0 means the target value for the mean of characteristic we control. Thats why a false alarm can appear. This happens when the process is in control, but the given sample has the mean out of control limits. This is the reason why we choose small probabilities for the false alarm to appear. The most often 0.05, or 0.01, or, to be in accordance with the 6σ rule, we set the probability for a false alarm to be 0.003. We denote the probability of this mistake with α, and we also know it by the name *probability of the first type error*, and *threshold of significance*.

There is another possible mistake, false positive - which occurs when the process is out of control, but the sample mean is between control limits.

Walter Shewharts ideas about control charts have been in use since 1924, many of them improved by the time, but X-bar charts are still in use. X-bar chart is not complicated for use, even without a computer. The assumptions about the distribution and variance of the monitored characteristic have to be complete. We can check it using appropriate statistical methods/tests. The construction of X-bar charts in real situations is explained in [2], as well as in [3].

3 Empirical Distribution Function in Quality Control

3.1 Basic Facts about EDF

For some characteristic X lets have a random sample of size n i.e. n-dimensional random variable $\vec{X} = (X_1, , X_n)$. Empirical distribution function is given with

$$F_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x); x \in \mathbb{R},$$

where I represents the indicator variable. Thats why $F_n^*(x)$ is the rate of sample elements less than or equal to x, and $nF_n^*(x)$ has the binomial distribution $B(n, F(x))$, where $F(x)$ is the theoretical distribution function for the characteristic X. Since $E(F_n^*(x)) = F(x)$ we have that EDF is unbiased estimator for $F(x)$. EDF is step function, with values in $[0,1]$, having discontinuities in points different from the ones in sample. Sufficient statistic for EDF is the set of order statistics $X_{(1)}, , X_{(n)}$, thats why EDF gives us maximal number of information from a sample.

No such obvious properties are given in two very important following theorems [1]:

Theorem 3.1 ((Glivenko-Kantelli theorem)**.** *The EDF converges uniformly to the distribution function F(x) with probability 1, i.e.,*

$$P(\sup_{x\in\mathbb{R}} |F_n^*(x) - F(x)| \to 0; n \to +\infty) = 1.$$

This theorem allows us to use the EDF of a sample as a good approximation of $F(x)$. This way, we can conclude that EDF is an estimator for $F(x)$. There is more. EDF obtained from a sample converges to the distribution function of the corresponding characteristics. Also, the distribution of $\sup_{x\in\mathbb{R}} |F_n^*(x) - F(x)|$ is independent of the distribution $F(x)$. Kolmogorov stated this in his theorem.

Theorem 3.2 (Kolmogorov's theorem)**.** *If the distribution function for a characteristic X is continuous, and if*

$$D_n = \sup_{x\in\mathbb{R}} |F_n^*(x) - F(x)|.$$

Then:

$$P\left\{\sqrt{n}D_n \leq t\right\} \to \sum_{i=-\infty}^{+\infty} (-1)^k e^{-2k^2t^2}; \qquad n \to +\infty.$$

This distribution, known as Kolmogorovs distribution, is independent of $F(x)$. The sample size is said to represent the degrees of freedom for the statistic D_n, which has quantiles given in tables, from which we reproduce one part that we shall use in our examples.

n	α=0.01	α=0.05	α=0.1	α=0.15	α=0.2
1	0.995	0.975	0.95	0.925	0.9
2	0.929	0.842	0.776	0.726	0.684
3	0.828	0.708	0.642	0.597	0.565
4	0.733	0.624	0.564	0.525	0.494
5	0.669	0.565	0.51	0.474	0.446
6	0.618	0.521	0.47	0.436	0.41
7	0.577	0.486	0.438	0.405	0.381
8	0.543	0.457	0.411	0.381	0.358
9	0.514	0.432	0.388	0.36	0.339
10	0.49	0.41	0.368	0.342	0.322
12	0.45	0.375	0.338	0.313	0.295
20	0.356	0.294	0.264	0.246	0.231
50	0.23	0.19	0.17	0.16	0.15
Over 50	$\frac{1.63}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.07}{\sqrt{n}}$

Table 1: Quantiles for Kolmogorov's distribution

For preset α, we obtain a band in which EDF must be, for us to accept the hypothesis $H_0(F(x) = F_0(x))$ with the threshold of significance α.

$$F(x) - d_{n;\alpha} \leq F_n^*(x) \leq F(x) - d_{n;\alpha}. \tag{1}$$

Since the Kolmogorovs test is very powerful and gives good results even for small sample size, we shall use (1) as a criterion in Quality control.

3.2 EDF for Each Sample in Quality Control

Testing procedure using X-bar charts in Quality control presumes that the distribution of monitored characteristic is normal (Gaussian) distribution $N(m, \sigma^2)$. X-bar chart tests the hypothesis $H_0(m = m_0)$. Using EDF, we shall include this test in the Quality control procedure, and do even more - we shall check if the distribution of monitored characteristic is normal.

Based on (1) we obtain a *band*:

$$\max\{0, F(x) - d_{n;\alpha}\} \leq F_n^*(x) \leq \min\{1, F(x) - d_{n;\alpha}\} \tag{2}$$

where, using mentioned presumption, $F(x)$ is cumulative distribution function (CDF) for normal distribution $N(m, \sigma^2)$. Values $d_{n;\alpha}$ can be found in table 1. We represent this situation in figure 2.

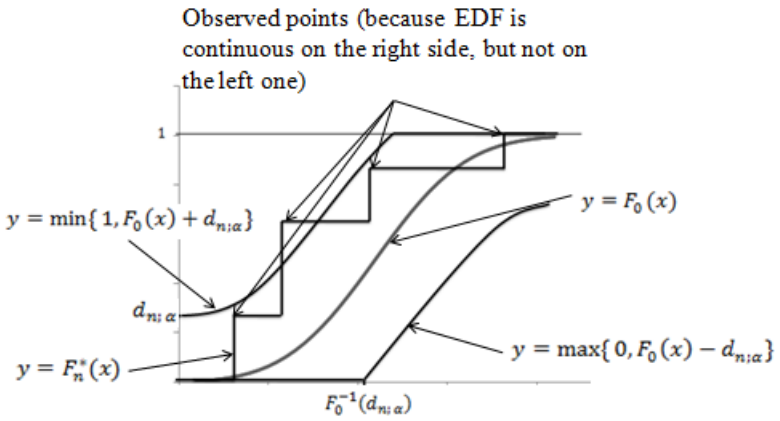

Figure 2: EDF for one sample

If we only use what we have mentioned so far, we can check the normality of monitored characteristics by checking the validity of (2) for each element. Next, we need a rule or criteria based on which we apply the procedure. We suggest one possible way by computing:

$$T = 1 - \frac{1}{n}\sum_{i=1}^{n} I(\max\{0, F(x_i) - d_{n;\alpha}\} \leq F_n^*(x_i) \leq \min\{1, F(x_i) - d_{n;\alpha}\}). \tag{3}$$

If $T < \alpha$ we accept H_0, i.e. X has a normal distribution $N(m, \sigma^2)$. Otherwise, we reject H_0. We may consider the hypothesis $H_0(m = m_0)$ tested as well. Notice that $T \cdot 100\%$ represents the percent of data lying outside the band (2).

Example 3.1: To control the quantity of protein in milk, we take four 100 gram packages of milk for 12 times from the production line. Measurements have yielded the following results:

Sample serial no.	Amount of protein (grams)
1	3.04, 3.12, 3.12, 3.22
2	3.09, 3.13, 3.21, 3.18
3	3.10, 3.18, 3.21, 3.18
4	3.04, 3.11, 3.17, 3.06
5	3.13, 3.12, 3.11, 3.07
6	3.15, 3.05, 3.14, 3.18
7	3.11, 3.21, 3.22, 3.13
8	3.06, 3.07, 3.17, 3.22
9	3.05, 3.19, 3.18, 3.20
10	3.08, 3.20, 3.21, 3.09
11	3.05, 3.14, 3.22, 3.08
12	3.19, 3.18, 3.21, 3.06

Table 2: Amount of proteins in 100g milk packaging

According to the given data, verify whether the conditions of the production process are satisfactory, considering average protein intake, 3.2g, with an aloud standard deviation of 0.06g.

To check the quality of milk, we can use X-bar chart. Since we assume normal distribution to apply X-bar chart, we need to know if samples of size four (Amount of protein (grams) column), are the samples of characteristic $X : N(3.2, 0.06^2)$. We shall check this assumption using EDF, and, this way, we eliminate the necessity of using X-bar charts.

Let's see what the quality of milk is for the first sample:

We obtain EDF:

$$F_4^* = \begin{cases} 0; & x < 3.04 \\ 0.25; & 3.04 \leq x < 3.12 \\ 0.75; & 3.12 \leq x < 3.22 \\ 1; & x \geq 3.22 \end{cases} .$$

We use (2) and we get:

$$\max\{0, F(x) - d_{4;0.05}\} \leq F_4^*(x) \leq \min\{1, F(x) - d_{4;0.05}\}$$

as a condition that needs to be satisfied for sufficient number of first sample's elements. By looking up in Table 1 we get $d_{4;0.05} = 0.642$. F is the distribution function of $N(3.2, 0.06^2)$.

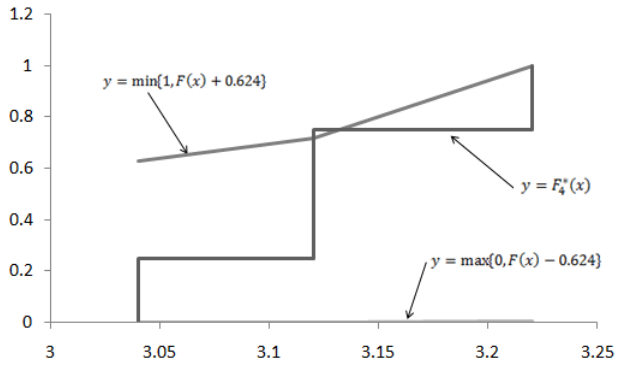

Figure 3: Quality control of the first sample

If we compute T in (3) for the first day we get $T = 0.25$ (i.e., 25 percent of elements does not fulfill the condition (2)). Since $T = 0.25 > 0.05 = \alpha$, we conclude that first sample does not have normal distribution $N(3.2, 0.06^2)$.
According to this sample, the quality of milk does not meet the prescribed standard.

If we complete the examination, doing the same procedure, we get that the samples for the first day, the fourth day, and the fifth day do not meet the prescribed standard for the average amount of proteins in milk packaging within the limits of allowed standard deviation.

Therefore we suggest revising the production process (especially for the mentioned days).

3.3 EDF in Quality Control using Means of all Samples

Using EDF in Quality control can also be done if we use means of samples taken in different periods. We shall now illustrate how to perform it. If k samples $x_{i1}, x_{i2}, ..., x_{in}; i = 1, ..., k$ are given, we compute the mean for each of them: $\bar{x}_{n1}, \bar{x}_{n2}, ..., \bar{x}_{nk}$ and take them as one sample of size k and we determine the EDF $F_k^*(x)$ for it. Then for each element of that sample we check:

$$\max\{0, \bar{F}(\bar{x}_{ni}) - d_{k;\alpha}\} \leq F_k^*(\bar{x}_{ni}) \leq \min\{1, \bar{F}(\bar{x}_{ni}) - d_{k;\alpha}\}; i = 1, ..., k \tag{4}$$

where $\bar{F}(x)$ is CDF for normal $N(m, \frac{\sigma^2}{n})$ distribution. Hence, based on Theorem 2.1, each sample's distribution is normal distribution $N(m, \sigma^2)$.

In other words, we can construct a control chart based on EDF and Kolmogorov's test statistic's properties, for which graphical representation is:

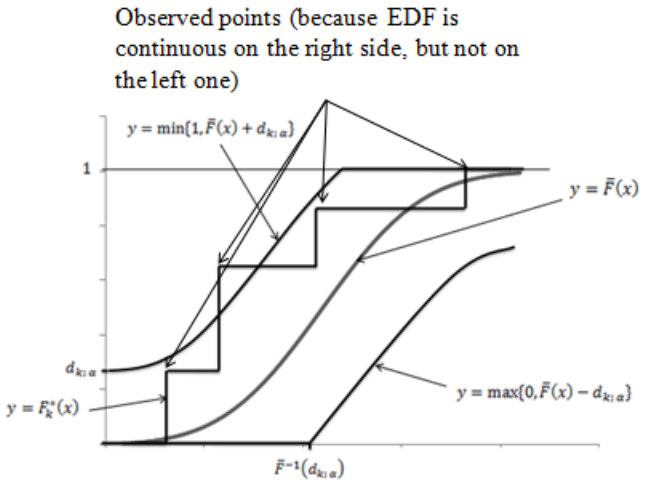

Figure 4: EDF in Quality control using means of all samples

UCL is graph of function $y = \max\{0, \bar{F}(\bar{x}) - d_{k;\alpha}\}$ and LCL is graph of function $y = \min\{1, \bar{F}(\bar{x}) - d_{k;\alpha}\}$.

We can interpret the results the same way (or the similar way) we do at X-bar charts, or some other chart type ([3] page 88).

Example 3.2: The goal is to perform quality control of data given in example 3.1 using EDF in Quality control using means of all samples. We observe 12 samples of size 4, given in Table 2.

We obtain a sample of 12 means: 3.125, 3.1525, 3.1675, 3.095, 3.1075, 3.13, 3.1675, 3.13, 3.155, 3.145, 3.1225, 3.16 and we compute EDF for this sample:

$$F_{12}^{*}(x)=\begin{cases}0; & x<3.095\\ 0.08; & 3.095\leq x<3.1075\\ 0.17; & 3.1075\leq x<3.1225\\ \ldots & \\ 1; & x\geq 3.1675\end{cases}.$$

We use (4), and we get that

$$\max\{0,\bar{F}(\bar{x}_{ni})-d_{12;0.05}\}\leq F_{k}^{*}(\bar{x}_{ni})\leq\min\{1,\bar{F}(\bar{x}_{ni})-d_{12;0.05}\};i=1,\ldots,12$$

must be valid for each $\bar{x}_{ni}$ for quality standards to be fulfilled. By looking up in the Table 1 we get $d_{12;0.05}=0.375$. $\bar{F}$ is the cumulative distribution function of normal $N(3.2,\frac{0.06^2}{4})$ distribution. By checking our condition, for each mean from the sample, we construct our control chart:

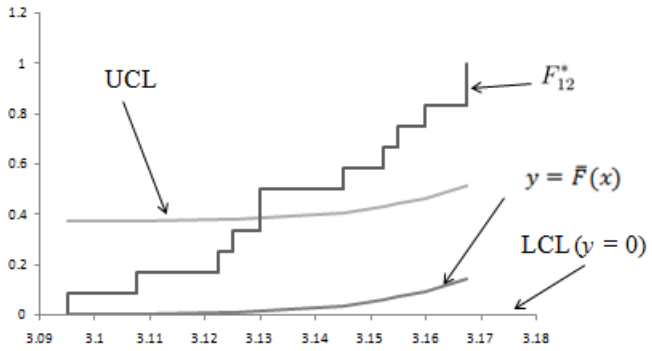

Figure 5: Quality control using means of all samples

We conclude that condition is not satisfied for eight means in the sample (this procedure is stricter than the one that we use to control quality for each sample), i.e. since the values of EDF for mentioned means are above the UCL, the quantity of protein in milk is not large enough to meet the standard. The monitored production process is not under control.

4 Concluding remarks

In this paper, we have introduced EDF as a tool in Quality control, and we have shown through examples how to use EDF in Quality control both for analyzing the production for a given period, and for successive samples during the production. The procedure is different from the one given in [6].

Further research in this area will deal with some other graphical methods in descriptive statistics that can be useful in Quality control. Also, there is a possibility to research the use of EDF in Quality control through simulation studies to estimate Average run length (expected number of samples before the first alarm), under the condition that process is in control state.

A more detailed list of references is available in [2].

References

[1] V. JEVREMOVIĆ, *Verovatnoa i statistika*, Matematiki fakultet Univerziteta u Beogradu, Beograd, 2014. (in Serbian)

[2] M. MINIĆ, *Kontrola kvaliteta - kontrolne karte i njihove osobine*, In: Vesna Jevremovi, Jovan Malii, *Izabrana poglavlja statistike - zbornik radova*, Departman za matematike nauke, Dravni Univerzitet u Novom Pazaru, 2018. (in Serbian)

[3] J. S. OAKLAND, *Statistical Process Control*, Butterworth - Heinemann, Oxford, 2003.

[4] http://whatis.techtarget.com/definition/quality-control-QUALITY CONTROL

[5] S. BAKIR, *A non-parametric Shewhart-type quality control chart for monitoring broad changes in a process distribution*, Int. journal of quality statistics and reliability, 2012, DOI: 10.1155/2012/147520