Using statistical techniques to validate the hypothesis of independence in control charts

Usando técnicas estatísticas para validar a hipótese da independência em gráficos de controle

Roberto Campos Leoni
Marcela Aparecida Guerreiro Machado
Fabrício Maciel Gomes
Herlandi de Souza Andrade
Nilo Antonio de Souza Sampaio

Abstract
The aim of this study was to identify and present the occurrence of autocorrelation in statistical data. To achieve this objective, a theoretical discussion was held on the subject and two exercises were also carried out using the R Statistical Software to show whether or not autocorrelation exists in statistical data. As a contribution of this research, a statistical technique using this software to show autocorrelation in data is pointed out.

Keywords: Statistical Data. Autocorrelation. R Software. Data Validation.

1 Doctor of Science, Academia Militar das Agulhas Negras, Rod. Pres. Dutra, km 30, Resende - RJ, CEP: 27534-970. E-mail: rcleoni@yahoo.com.br Orcid: https://orcid.org/0000-0001-6600-2963
2 Doctor of Engineering, Universidade Estadual Paulista “Julio de Mesquita Filho”, Dr. Ariberto Pereira da Cunha, 333, Portal das Colinas, Guaratinguetá. E-mail: marcela.freitas@unesp.br Orcid: https://orcid.org/0000-0002-0621-6932
3 Doctor of Mechanical Engineering, Escola de Engenharia de Lorena, Universidade de São Paulo (EEL - USP), Estrada Municipal do Campinho, No 100, Campinho Lorena - SP. E-mail: fmgomes@usp.br Orcid: https://orcid.org/0000-0001-7694-9835
4 Doctor of Aeronautical and Mechanical Engineering, Escola de Engenharia de Lorena, Universidade de São Paulo (EEL - USP), Estrada Municipal do Campinho, No 100, Campinho Lorena - SP. E-mail: herlandi@usp.br Orcid: https://orcid.org/0000-0003-3293-3991
5 Doctor of Mechanical Engineering, Universidade do Estado do Rio de Janeiro, Rod. Pres. Dutra, km 298, Polo Industrial, Resende - RJ, CEP: 27537-000. E-mail: nilo.samp@terra.com.br Orcid: https://orcid.org/0000-0002-6168-785X
Resumo
O objetivo deste estudo foi identificar e apresentar a ocorrência de autocorrelação em dados estatísticos. Para alcançar este objetivo, foi realizada uma discussão teórica sobre o assunto e dois exercícios também foram realizados usando o Software de Estatística de R para mostrar se existe ou não autocorrelação em dados estatísticos. Como uma contribuição desta pesquisa, uma técnica estatística usando este software para mostrar autocorrelação em dados é apontada.


Introduction

Almost a century ago, Shewhart (1926) proposed control charts. The basic assumption is that the measures of the quality characteristic being evaluated are independent and identically distributed. However, measures of the quality characteristic of neighboring items, according to the time at which they were produced, may show some degree of dependence between the observations. This dependence is called autocorrelation and creates challenges when using control charts.

Statistics is an essential tool and has been used in several applications and in all areas of human knowledge to improve processes, as decision making and to ensure that the environment is preserved (Cardoso, R.P; Sampaio, N.A.S; Reis, J.S.M; Silva, D.E.W; Barros, 2023; Cardoso et al., 2023; da Silva et al., 2021; Fonseca, D; Correa, M.P.O; Santos, R.R; Cardoso, R.P; Reis, J.S.M; Sampaio, 2023; Gomes et al., 2022; Mazza et al., 2023; Mendonça et al., 2023; Menezes et al., 2023; Reis et al., 2023; Reis, Espuny, Cardoso, Sampaio, de Barros, et al., 2022; Rezende et al., 2023; Rubert et al., 2023; Veloso et al., 2023; Yamada et al., 2023)

Many authors have addressed the problem of autocorrelation, especially in the manufacturing environment. This problem has become more serious in the last two decades with the advance of technology, as it is possible to measure the quality characteristics of practically every item produced. Pan and Jarret (2011) exemplify several of these processes in which autocorrelation is present. It is important to note that the autocorrelation discussed here is something inherent to the process and cannot be removed with human intervention.

Pioneering works such as Bagshaw and Johnson (1975) were already concerned with evaluating the effect of autocorrelation on control charts. Various monitoring strategies have
been proposed over the last few decades, with the AR (1) and ARMA(1,1) models being the most widely used to describe the behavior of data from an autocorrelated process. Once autocorrelation is detected, two strategies can be employed to monitor a process: specific graphs are used to monitor the process (Apley and Tsung, 2002; Kalagonda and Kulkarni, 2004) or removing the effect of autocorrelation using the residuals of a model fitted to the data set (Mason and Young, 2002).

Although control charts have developed and become one of the most widely used monitoring techniques, their incorrect application is still no exception. A survey conducted by Alwan (1995) with a sample of 235 applications involving control charts, collected from sources from which one would expect a considerable degree of sophistication in mastering the technique and using this tool, showed that in 86% of the cases there was some kind of violation of the fundamental hypotheses and, more often than not, a failure to recognize serial dependence in the process.

The starting point for dealing with the problem of autocorrelation is to detect it. The independence assumption is rarely assessed in practice and the presence of autocorrelation can generate false alarms when a characteristic is monitored with control charts (Mason et al., 1996; Mason and Young, 2002). In this context, this article presents a broad review of the topic of autocorrelation and introduces statistical techniques for evaluating the independence assumption using the R software (R DEVELOPMENT CORE TEAM, 2014). The R software is free and well known in the academic community.

This paper is structured as follows: in section 2 a broad review of control charts applied to autocorrelated processes is presented. In section 4 some techniques for detecting the presence of autocorrelation are described and illustrated through examples using the R software. Section 5 presents some final considerations.

**Theoretical Referential**

The assumption of independence of observations when violated causes misleading results in the form of too many false alarms. Montgomery (2004) describes that basically all processes are governed by inertial elements and when the interval between taking samples is small in relation to these forces, the observations show correlation over time. An analytical demonstration of autocorrelation can be represented by a simple system of a tank with volume V and inlet and outlet material flows.
Using statistical techniques to validate the hypothesis of independence in control charts

Figure 1: Tank with volume $v$ and inlet and outlet material flows
Source: Adapted from Montgomery (2004), 2023.

Figure 1 illustrates a tank in which $w_t$ and $x_t$ represent the concentration of a certain material in the inlet and outlet streams at an instant $t$. Assuming homogeneity within the tank, the relationship between the inlet and outlet concentrations is given by:

$$x_t = w_t - T \frac{dx}{dt}$$  \hspace{1cm} (1)

in which:

$$T = \frac{\text{Tank volume}}{\text{Ratio between incoming and outgoing material flow}}$$

when the flow has a step-like variation $w_0$ at time $t=0$, the concentration at the outlet can be modeled by:

$$x_t = w_0 \left(1 - e^{-\frac{t}{T}}\right)$$  \hspace{1cm} (2)

as we observe in practice at equally spaced time intervals, the concentration at the outlet is represented by:

$$x_t = x_{t-1} + \left(w_t - x_{t-1}\right) \left(1 - e^{-\frac{\Delta t}{T}}\right)$$  \hspace{1cm} (3)
the concentration of the outflow depends on the concentration of the inflow and the sampling interval. Assuming that the observations are uncorrelated, the autocorrelation in will be:

\[ \rho = \frac{\Delta t}{T} \]  

If , the observations on the output concentration will always be correlated with each other (see figure 2).

\[ \Delta t/T \]

**Figure 2: Correlation between successive values of \( \Delta t \).
Source: Authors, 2023**

Another pioneering analytical model can be found in Yule (1927) who used the dynamic movement of a pendulum as inspiration for the formulation of an autoregressive model that explains time dependence in a time series. If a pendulum with mass \( m \) under the influence of gravity is in equilibrium and is suddenly struck by a single impulse force, it begins to execute a harmonic motion and the frequency of this motion depends on the length of the pendulum, the mass, the impulse and friction force and the viscosity of the medium in which the pendulum is located. The forces affecting a pendulum in motion are shown in Figure 3.

**Figure 3. A simple pendulum in motion.
Source: Authors, 2023.**
Figure 4 illustrates the harmonic motion of a pendulum $x(t)$ in equilibrium position at the instant of time when $t=0$.

![Figure 4: Harmonic motion of a pendulum as a function of time.](image)

This movement is approximately described by a second-order linear differential equation with constant coefficients.

$$\frac{d^2 Z}{dt^2} + \gamma \frac{dZ}{dt} + kZ = A\tau(t)$$  \hspace{1cm} (5)

Where

is an impulse function that forces the pendulum out of equilibrium and $A$ is the size of the impact that sets the pendulum in motion.

Differential equations are used to describe the dynamic behavior of processes as a function of time. However, time series data is usually sampled (observed) in discrete units of time, for example, every hour or every minute. Yule (1927) showed that the differential equation described in (5) can be rewritten as:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + e_t$$  \hspace{1cm} (6)

Therefore, if observed in discrete time, the oscillatory behavior of a pendulum can be described by equation (6) which is known as a second order autoregressive time series model and the value at time instant $t$ depends on the observations and, the autoregressive parameters and a random error.

In many continuous and discrete processes, autocorrelation is present and reduces the performance of the control chart, as they are based on the assumption that the data from the
observations of the quality characteristic are independent. On a production line, observations of the quality characteristic close together in time tend to be more similar than those further apart. Models from the ARIMA family are used to describe these processes. This section presents the most commonly used models and a broad review of the work associated with each model.

The first-order autoregressive model - AR(1) is represented by:

$$X_{ij} - \mu_0 = \phi(X_{i,j-1} - \mu_0) + \epsilon_{i,j} = \{1, 2, ..., n\} \text{ e, } i = 1, 2, ...$$

(7)

Where

the t-th observation of the i-th sample is autocorrelated with the (j-1)-th observation, through the autocorrelation coefficient $\phi$. Table 1 shows articles dealing with control charts for autocorrelated processes based on model (7).

The second-order autoregressive model - AR(2) is represented by:

$$X_{ij} - \mu_0 = \phi_1(X_{i,j-1} - \mu_0) + \phi_2(X_{i,j-2} - \mu_0) + \epsilon_{i,j} \text{ para } j = \{1, 2, ..., n\} \text{ e, } i = 1, 2, ...$$

(8)

The j-th observation of the i-th sample ($X_{(i,j)}$) is autocorrelated with the (j-1)-th ($X_{(i,j-1)}$) and (j-2)-th ($X_{(i,j-2)}$) observations, through the autocorrelation coefficients $\phi_1$ and $\phi_2$ respectively. Table 2 shows papers dealing with control charts for autocorrelated processes based on model (8).

The first-order autoregressive moving average model - ARMA(1,1), which is equivalent to a first-order autoregressive model AR(1) with additional random error, is represented by:

$$X_{ij} - \mu_0 = \phi(X_{i,j-1} - \mu_0) + \theta \epsilon_{i,j-1} + \epsilon_{i,j} \text{ para } j = \{1, 2, ..., n\} \text{ e, } i = 1, 2, ...$$

(9)

The i-th observation has a linear correlation with the (j-1)-th observation ($X_{(i,j-1)}$) and the (j-1)-th random error ($\epsilon_{(i,j-1)}$), where $\phi$ is the autocorrelation coefficient between observations and $\theta$ is the oscillation coefficient of the sample mean related to the additional autocorrelated error. Studies related to model (9) are presented in Table 3
Using statistical techniques to validate the hypothesis of independence in control charts

Reference | Research Objective
---|---
Bagshaw and Johnson (1975) | Demonstrate the effect of autocorrelation on CUSUM graphs.
Montgomery and Mastrangelo (1991) | Research the strategy of modifying the opening of the control limits of the $\overline{X}$ graph to improve performance.
Alwan and Radson (1992) | Study the impact of autocorrelation on the $\alpha$ error (Type I) of the $\overline{X}$ graph.
Runger and Willemain (1995) | Demonstrate that the $X$ and CUSUM graphs perform better when compared to methods based on time series from the ARIMA family.
Costa and Claro (2008) | Study the performance of the $\overline{X}$ graph with double sampling.
Costa and Castagliiola (2011) | Apply a systematic sampling strategy to the $\overline{X}$ graph to reduce the effect of autocorrelation.
Lwin (2011) | Study the effect of parameter estimation on EWMA graphs.
Costa and Machado (2011) | Propose the Markov chain approach to study the performance of $\overline{X}$ graphs with variable parameters and double sampling.
Linet et al. (2011) | Propose a model for real-time monitoring in order to recognize unnatural patterns in the presence of autocorrelation.

Table 1: Articles related to the first-order autoregressive model - AR(1)
Source: Authors, 2023

Reference | Research Objective
---|---
Vasilopoulos and Stamboulis (1978) | Apply the strategy of increasing the openness of the control limits of the $\overline{X}$ graph to improve statistical performance.
English et al. (2000) | Evaluate the performance of the $\overline{X}$ and EWMA graphs.

Table 2: Articles related to the second-order autoregressive model - AR(2).
Source: Authors, 2023

Reference | Research Objective
---|---
Bagshaw and Johnson (1975) | Evaluate the effect of autocorrelation on CUSUM graphs.
Harris and Ross (1991) | Evaluate the effect of autocorrelation on the CUSUM and EWMA graphs.
Wardell et al. (1994) | Study the properties of the special-cause control chart proposed by Alwan and Roberts (1988).
Apley and Tsung (2002) | Present the guidelines for using the autoregressive $T^2$ control chart.
Apley and Lee (2003) | Employ the strategy of increasing the openness of the control limits of the EWMA chart to improve statistical performance.
Costa and Claro (2008) | Study the performance of the $X$ graph with double sampling.
Pacella and Semeraro (2007) | Use a recurrent neural network to detect the process of mean change in autocorrelated processes.
Soleimani et al. (2009) | To study the performance of $T^2$ and EWMA graphs when the effect of autocorrelation is ignored.
Mertens et al. (2009) | Apply the CUSUM graph to monitor autocorrelated processes.

Table 3: Articles related to the first-order autoregressive moving average model - ARMA(1,1).
Source: Authors, 2023

The hypothesis that the mean oscillates was introduced by Reynolds Jr. et al. (1996) who broke with the paradigm that the process mean is a variable that can only take on two values: a target value and a value outside the target resulting from the occurrence of a special cause. Since then, the monitoring of processes with oscillating averages has been studied by various researchers, see Table 4. .
Using statistical techniques to validate the hypothesis of independence in control charts

Reference
Reynolds Jr. et al. (1996) Present the properties of the $X^{-}$ chart with variable sample size and interval.
VanBrackle Jr. and Reynolds (1997) Present how high levels of autocorrelation and mean oscillation affect the performance of EWMA and CUSUM charts.
Lu e Reynolds Jr. (1999a, 1999b) Propose the combined use of the residual Shewhart chart with EWMA for the original observations.
Jiang et al. (2000) Proposing a new chart called the Autoregressive Moving Average (ARMA Chart), Jiang and Tsui (2001) later investigated the strategy of increasing the openness of the control limits of the ARMA chart to improve statistical performance.
Lu and Reynolds (2001) Show that in the presence of autocorrelation, the CUSUM charts offer the same performance as the EWMA chart.
Zou et al. (2008) Propose a method called Variable Sampling Rate at Fixed Times (VSRFT).
Lin et al. (2012) Carry out economic planning on the ARMA graph.
Franco et al. (2012) Carry out economic planning for the graph $\bar{X}$.

Table 4. Articles related to the model in which the process mean oscillates
Source: Authors, 2023.

Research Method

This paper can be classified as an applied research, as it aims to provide improvements in the current literature, with normative empirical objectives, aiming at the development of policies and strategies to improve the current condition (Bertrand & Fransoo, 2002; Reis, Espuny, Cardoso, Sampaio, Barros, et al., 2022; Sampaio et al., 2010). The problem approach is quantitative, as is the modeling and simulation research method. The research steps were carried out following the sequence shown in Fig.1.

- **Step 1**: The Experimental data was selected from the book of (B et al., 2020). This choice was based on the fact that this book brings interesting data to be treated with the respective statistical analyses.
- **Step 2**: A Discussion was held on Statistical Techniques that have Autocorrelation
- **Step 3**: Two Examples were solved using R Software to show whether or not there is autocorrelation between the data.
- **Step 4**: The interpretation of the results was performed.
• Step 5: Scientific writing of the Results and Discussion were performed
• Step 6: The conclusions presented at the end of this paper are drawn from the results obtained in the previous steps.

![Diagram showing the steps of the search method](Source: Authors (2023)).

Results and Discussions

4.1 Identification of Autocorrelation:

Autocorrelation is identified using statistical techniques that assess the independence between observations of the quality characteristic you want to monitor. In this section we present the Ljung-Box statistic (Ljung and Box, 1978) and the autocorrelation graph introduced by Box and Jenkins (1976). The R software was used to exemplify the techniques described here.

• Autocorrelation function and the Ljung-Box statistic:

The autocorrelation graph, also known as a correlogram, was presented by Box and Jenkins (1976). It is based on statistics:

\[ r_h = \frac{C_h}{C_0} - 1 \leq r_h \leq 1 \]  

(10)
Using statistical techniques to validate the hypothesis of independence in control charts

where:

$$C_h = \frac{\sum_{t=1}^{n-h}(x_t-\bar{x})(x_{t+h}-\bar{x})}{n}$$  \quad (11)

And

$$C_0 = \frac{\sum_{t=1}^{n}(x_t-\bar{x})^2}{n}$$  \quad (12)

Where

\( n \) is the total amount of data and \( h \) is the interval between observations (lag).

The results of (10) are plotted on a graph (correlogram) and control limits (CL) are used to assess whether there is evidence of autocorrection. The limits are:

$$LC = \pm \frac{Z_{1-\alpha/2}}{\sqrt{n}}$$  \quad (13)

With

\( Z_{(1-\alpha/2)} \) = standard normal score; \( \alpha \) = interval significance level.

The Ljung-Box test evaluates the hypothesis of independence of a time series. It is based on statistics:

$$Q^* = n(n+2)\sum_{k=1}^{H}(n-h)^{-1}r_h^2 \quad k=1,2,\ldots,K$$  \quad (14)

Where

\( Q^* \sim \chi^2_1 \); \( n \) is the number of observations available in the series; \( r_h \) is the \( h \)-th autocorrelation coefficient and \( H \) is the lag between observations. Large values for \( Q^* \) indicate autocorrelation in the set of observations. For more details see Ljung and Box (1978).

**Example 1:** Consider the data from an accelerated tire test. The loss of resistance by abrasion in gram/hour was evaluated and the results obtained were: 372; 206; 175; 154; 136;
112; 55; 45; 221; 166; 164; 113; 82; 32; 228; 196; 128; 97; 64; 249; 219; 186; 155; 114; 341; 340; 283; 267; 215; 148.

The sequence of commands to be executed in R to assess whether the data is autocorrelated:

```
Loss=c(372,206,175,154,136,112,55,45,221,166,164,113,82,32,228,196,128,97,64,249,219,186,155,114,341,340,283,267,215,148)
acf(perda,ci=0.95)  #ci= confidence level
Box.test(loss, lag = 1, type = c("Ljung-Box"), fitdf = 0)
```

The results are illustrated in Figure 5. Both methods presented in Figure 5 indicate, at a 5% significance level, that there is evidence of autocorrelation.

**Example 2:** The temperatures of a chemical bath were measured at regular intervals of 3 minutes, the results of which were: 250.9; 252.2; 251.8; 252.8; 254.6; 256.7; 256.9; 255.6; 254.7; 255.2; 254.9; 256.6; 254.9; 256.3; 258.6; 257.4; 258.7; 259.2; 259.2; 259.8; 259.0; 258.6; 258.0; 258.1; 256.2; 257.6; 256.3; 256.0; 255.9; 254.4; 254.5; 254.1; 254.0; 254.9; 255.2; 255.9; 255.6; 256.4; 256.4; 256.5.

Sequence of commands to be executed in R:
temperatura=c(250.9,252.2,251.8,256.7,256.9,255.6,254.7,254.9,256.6,254.9,256.3,258.6,256.9,254.4,254.1,254.9,255.2,255.9,255.6,256.4,256.4,256.5)

acf(temperatura,ci=0.95) #ci=nível de confiança
Box.test(temperatura, lag = 1, type = c("Ljung-Box"), fitdf = 0)

The results are illustrated in Figure 6. In this example there is strong evidence of autocorrelation in the temperature data of a chemical bath. The Ljung-Box statistic has a p-value very close to zero. The correlogram also provides evidence of autocorrelation.

![Figure 6 - R software output screen - Chemical bath temperature data](source)

Source: Authors, 2023.

**Conclusion**

This article has presented some statistical techniques for assessing the presence of autocorrelation in univariate processes and a broad review of control charts in the presence of autocorrelation. The autocorrelation observed in industrial processes, i.e. that inherent in the process, is typically due to the presence of inertial elements that limit the variability between observations close together on the time scale. Whatever technique is used to monitor a characteristic that is autocorrelated, checking whether a process is autocorrelated is fundamental to the correct use of a control chart.
References


MONTGOMERY, D. C. *Introduction to Statistical Quality Control*, 5


Using statistical techniques to validate the hypothesis of independence in control charts


Using statistical techniques to validate the hypothesis of independence in control charts


Submetido em: 10.11.2023
Aceito em: 12.12.2023

Revista Gestão e Secretariado (GeSec), São Paulo, SP, v. 14, n. 12, 2023, p. 21321-21339.