UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Estadística e Investigación Operativa Aplicadas y Calidad

Diseño de gráficos de control no paramétricos para el coeficiente de correlación de Spearman

Trabajo Fin de Máster

Máster Universitario en Ingeniería de Análisis de Datos, Mejora de Procesos y Toma de Decisiones

AUTOR/A: Palomar Yarritu, Iñigo

Tutor/a: Giner Bosch, Vicent

Cotutor/a externo: CASTAGLIOLA, PHILIPPE

CURSO ACADÉMICO: 2021/2022

Resumen

En este trabajo, se aborda el diseño e implementación de gráficos de control para monitorizar el grado de asociación entre dos variables continuas, no necesariamente normales, a través del coeficiente de correlación de Spearman, r_s. Debido a la naturaleza de este estadístico, basado en la transformación de las observaciones en rangos, estos gráficos pueden considerarse no paramétricos, en el sentido de que la distribución en el muestreo de r_s no depende de la distribución de la variable bidimensional objeto de interés bajo la hipótesis de independencia, situación que será considerada como el estado bajo control.

En primer lugar, se estudia el marco teórico correspondiente, haciendo hincapié en las propiedades de la distribución en el muestreo de r_s. A partir de este contexto, se diseña un esquema de monitorización tipo Shewhart basado en dicho estadístico.

A continuación, se implementa este esquema en lenguaje R y se diseñan y ejecutan experiencias computacionales, basadas en simulación Montecarlo, orientadas a evaluar el comportamiento del gráfico bajo diferentes condiciones iniciales y, en concreto, su potencia para detectar diferentes niveles de interdependencia entre las dos variables objeto de interés.

Por último, se ejemplifica la aplicación del gráfico diseñado mediante un caso práctico elaborado ad-hoc.

Las experiencias numéricas llevadas a cabo revelan la importancia del tamaño muestral para la detección de correlación. Sin embargo, no se han obtenido evidencias estadísticas que nos indiquen que las distribuciones de X e Y, las desviaciones típicas y las cópulas empleadas influyan en la capacidad de detección de correlación del gráfico de control.

Palabras clave gráficos de control; métodos no paramétricos; correlación de Spearman; estadísticos de rangos; simulación Montecarlo; cópulas.

Abstract

In this paper, we address the design and implementation of control charts to monitor the degree of association between two continuous variables, not necessarily normally distributed, through Spearman's correlation coefficient, r_s. Due to the nature of this statistic, which is based on the transformation of observations into ranges, these charts can be considered nonparametric, in the sense that the sample distribution of r_s does not depend on the distribution of the two-dimensional variable of interest under the hypothesis of independence, a situation that will be considered as the state under control.

First, the corresponding theoretical framework is studied, with emphasis on the properties of the distribution in r_s sampling. From this context, a Shewhart-type monitoring scheme based on this statistic is designed.

Then, this scheme is implemented in R language and computational experiments, based on Montecarlo simulation, are designed and executed to evaluate the behavior of the chart under different initial conditions and, in particular, its power to detect different levels of interdependence between the two variables of interest.

Finally, the application of the designed chart is exemplified by means of an ad-hoc case study.

The numerical experiments carried out reveal the importance of the sample size for the detection of correlation. However, no statistical evidence has been obtained to indicate that the distributions of X and Y, the standard deviations and the copulas used influence the correlation detection capacity of the control chart.

Keywords control charts; nonparametric methods; Spearman's correlation; rank statistics; Montecarlo simulation; copulas.

Índice general

Capítulo 1

Introducción

1.1. Antecedentes y motivación

Los gráficos de control son herramientas muy potentes y versátiles para monitorizar y mejorar la calidad de los procesos de una empresa, proporcionando una determinante ventaja competitiva. En ocasiones, y en el contexto de una realidad necesariamente multivariante, resulta de interés valorar el grado de asociación o correlación entre las diferentes variables que definen la calidad de un producto (conocidas como *características de calidad*). El deseo de las empresas por controlar la relación entre las características de calidad de sus productos, se ve, a menudo, obstaculizado por la ausencia de normalidad o la falta de certeza sobre la distribución de cada una de las características.

En paralelo a lo que sucede en el campo de la inferencia estadística, una alternativa válida ante la falta de conocimiento sobre la distribución de las características de calidad objeto de interés es el uso de gráficos de control no paramétricos. Este campo de estudio ha despertado el interés de investigadores, que han proporcionado estudios, tales como los trabajos de Langenberg y Iglewicz (1986), Alloway y Raghavachari (1991) [1] y Yourstone y Zimmer (1992) [2], por nombrar solo algunos.

En un contexto no paramétrico, herramientas conocidas, tales como el coeficiente de correlación de Pearson, no resultan adecuadas, ya que su distribución en el muestreo, incluso en el supuesto de independencia, depende de la distribución de las variables implicadas. Además, la correlación de Pearson se trata de un estadístico cuyo valor, en general, puede verse bastante afectado por la presencia de valores anómalos o valores

extremos, propios de distribuciones asimétricas. En la literatura existen otros estadísticos, basados en rangos, orientados a cuantificar el grado de relación o asociación entre dos variables aleatorias, sin necesidad de suponer normalidad. Uno de ellos es el *coeficiente de correlación de Spearman*. Hasta donde conocemos, no existe ningún trabajo publicado en el que se aborde el diseño de un gráfico de control para la correlación basado en este estadístico.

Con el objeto de proporcionar una metodología no paramétrica eficaz capaz de realizar la monitorización de la correlación entre dos variables aleatorias mediante el índice de correlación de Spearman y el cálculo de su desempeño, se motiva este trabajo fin de máster.

1.2. Objetivos

Este trabajo pretende alcanzar los siguientes objetivos:

(i) Afianzar la base teórica necesaria para la construcción de un gráfico de control no paramétrico para la monitorización del índice de correlación de Spearman entre dos variables aleatorias.

(ii) Proporcionar una metodología eficaz para calcular el límite de control de gráfico y su desempeño tanto bajo control como fuera de control.

(iii) Estudiar el efecto de los factores implicados en la construcción del gráfico de control.

(iv) Ejemplificar la metodología desarrollada mediante la construcción de un gráfico de control y el cálculo del desempeño.

1.3. Estructura del trabajo

Sin tener en cuenta la introducción, el trabajo consta de cinco capítulos más.

En el segundo capítulo, se introducen los conceptos matemáticos necesarios para la comprensión de la propuesta de trabajo realizada.

En el tercer capítulo, se presenta y se desarrolla la propuesta realizada en el trabajo: un gráfico de control para detectar la correlación entre pares de variables continuas. En concreto, se detalla el procedimiento para el cálculo del límite de control del gráfico y se describen las herramientas proporcionadas para el cálculo del desempeño del gráfico de control en función de diferentes condiciones.

En el cuarto capítulo, se muestran los resultados de aplicar la metodología propuesta para diferentes condiciones iniciales fijadas por el proceso en sí mismo o por el usuario final, y se evalúa el efecto de los diferentes factores involucrados.

En el quinto capítulo, se construye un gráfico de control como ejemplo numérico donde se aplica la metodología propuesta para calcular el límite de control y evaluar el desempeño del mismo.

En el sexto y último capítulo, se muestran las conclusiones que pueden extraerse del trabajo realizado y las posibles futuras líneas de investigación.

Capítulo 2

Marco teórico

Antes de comenzar con el desarrollo de nuestra propuesta, es necesario introducir los fundamentos teóricos necesarios. Entre otros conceptos, en este capítulo se define de forma adecuada el coeficiente de correlación de Spearman y se presentan conceptos básicos del control estadístico de la calidad y, en particular, los gráficos de control y el cálculo del desempeño de los mismos.

2.1. Covarianza y correlación

Cada vez que se analicen al menos dos variables de forma simultánea referidas a una misma población, será de gran interés conocer si los valores de una variable cambian de manera consistente a los valores de la otra variable, o por el contrario, si no hay ninguna relación que los asocie. Este fenómeno se conoce como la covarianza entre dos variables.

Definición 2.1.1 (Covarianza)**.** Definimos la *covarianza* de las variables X e Y como:

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

Donde E es el operador que define la esperanza matemática y μ_X y μ_Y son la media de las variables X e Y, respectivamente. Como caso particular, la covarianza de una variable consigo misma es la varianza de dicha variable.

Demostración.

$$Cov(X, X) = E[(X - \mu_X)(X - \mu_X)] = Var(X).$$

□

El problema de utilizar la covarianza para medir la relación entre dos variables aleatorias es que, habitualmente, las variables están expresadas en distintas unidades de medida. Por ello, la covarianza de las variables podría no ser interpretable, entre otros inconvenientes.. Para resolver este problema, se introduce el coeficiente de correlación de Pearson, que no es más que la estandarización de la covarianza.[3]

Definición 2.1.2 (Coeficiente de correlación poblacional de Pearson)**.** Definimos el *coeficiente de correlación poblacional de Pearson* de las variables X e Y como:

$$\rho_{XY} = \rho_{YX} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}. \tag{2.1}$$

Donde σ_X y σ_Y son las desviaciones típicas de X e Y, respectivamente. Como observamos en la Ecuación 2.1, el coeficiente de correlación de Pearson es adimensional, lo que facilita la interpretación.

Definición 2.1.3 (Coeficiente de correlación muestral de Pearson)**.** Dados n pares de datos $\{(x_i, y_i)\}_{i=1}^n$ definimos el *coeficiente de correlación muestral de Pearson* de las variables X e Y como:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.. \tag{2.2}$$

Dado que el coeficiente de correlación de Pearson no es más que una estandarización de la covarianza, este varía en el intervalo $[-1, 1]$, donde el signo del mismo define el sentido de la relación entre las variables.

- si $|\rho| = 1 \Rightarrow$ Correlación perfecta entre las variables, es decir, dependencia total entre las variables.
- si $0 < |\rho| < 1 \Rightarrow$ Existe una correlación entre las variables.
- si $\rho = 0 \Rightarrow$ No existe una correlación lineal entre las variables. Esto no implica que no estén relacionadas de ninguna forma, esta relación podría ser no lineal.

Ahora bien, es importante destacar que no es recomendable utilizar el coeficiente de correlación de Pearson para calcular la relación entre dos variables cualesquiera. Esto se da, puesto que, la distribución en el muestreo de r depende de la distribución de las variables X e Y, incluso bajo la hipótesis de que sean independientes, por lo que, su valor se ve afectado por la presencia de valores anómalos, lo cual lo hace poco apropiado para evaluar la correlación en el caso de variables no normales.

En algunas situaciones y contextos, las variables presentes en los problemas de la vida cotidiana no siguen una distribución de probabilidad reconocible, y mucho menos, normal. Por ello, para solucionar este problema y poder medir la relación entre variables sin asumir su distribución normal, existen estadísticos como, la *tau de Kendall*, la *Gamma de Goodman y Kruskal* o la que estudiaremos con mayor detalle la *rho de Spearman* [4].

2.2. Coeficiente de correlación de Spearman

El coeficiente de correlación de Spearman es un coeficiente no paramétrico alternativo al coeficiente de correlación de Pearson cuando las variables cuyo grado de asociación se desea estudiar no cumplen la hipótesis de normalidad o, sencillamente, no se tiene certeza sobre ello.

Para el cálculo del coeficiente de correlación de Spearman, r_S, en primer lugar, debemos ordenar los n valores de la variable X de menor a mayor. De esta manera, denotaremos como R_i al rango del valor X_i, $\forall i \in \{1, \ldots, n\}$. De manera análoga, denotamos S_i al rango del valor Y_i, $\forall i \in \{1, \ldots, n\}$. Por esta razón, se dice que el coeficiente de correlación de Spearman es un estadístico basado en rangos.

Una vez obtenidos los rangos de las variables a analizar, R_i y S_i, obtenemos el coeficiente de correlación a través de la fórmula del coeficiente de correlación de Pearson reemplazando X_i por R_i y Y_i por S_i.

En el caso del coeficiente de correlación de Spearman, no existe un equivalente poblacional. Solo está definido a nivel muestral.

Definición 2.2.1. Dados n pares de datos $\{(x_i, y_i)\}_{i=1}^{n}$ de una población continua biva-

riante, con una función de distribución conjunta F_{XY} y unas distribuciones marginales F_X y F_Y. Denotamos como r_S al coeficiente de correlación por rangos de Spearman que se define de la siguiente manera [5]:

$$r_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}. \tag{2.3}$$

Donde n es el tamaño de las variables X e Y y $D_i = S_i - R_i \ \forall i \in \{1, \ldots, n\}$.

Teorema 2.2.1. *El coeficiente de correlación de Spearman no es más que el coeficiente de correlación de* Pearson *reemplazando X_i e Y_i por R_i y S_i respectivamente [6].*

Demostración. Para demostrar la formula del coeficiente de correlación de Spearman, primero debemos recordar ciertas propiedades de la varianza y la covarianza para una muestra.

- $Cov(X,Y) = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$.
- $Var(X) = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$.

De esta manera, podemos reescribir la Ecuación 2.1 del coeficiente poblacional como la Ecuación 2.2 del coeficiente muestral:

$$\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Ahora, sustituimos X_i e Y_i por R_i y S_i respectivamente, obteniendo la siguiente expresión:

$$\frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2}}. \tag{2.4}$$

Donde, por una parte:

$$\bar{R} = \bar{S} = \sum_{i=1}^{n} \frac{R_i}{n} = \sum_{i=1}^{n} \frac{S_i}{n} = \frac{1 + 2 + \cdots + n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2},$$

y, por otra parte:

$$\sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} S_i^2 = \frac{(2n+1)(n+1)n}{6}.$$

Vamos ahora a desarrollar las expresiones por partes para simplificar la comprensión de la demostración.

$$\sum_{i=1}^{n}(R_i - \bar{R})^2 = \sum_{i=1}^{n} R_i^2 - n\bar{R}^2 = \frac{(2n+1)(n+1)n}{6} - n\left(\frac{n+1}{2}\right)^2 =$$

$$\frac{(2n+1)(n+1)n}{6} - \frac{n(n+1)^2}{4} = \frac{2(2n+1)(n+1)n - 3n(n+1)^2}{12} =$$

$$\frac{4n^3 + 6n^2 + 2n - (3n^3 + 6n^2 + 3n)}{12} = \frac{n^3 - n}{12}.$$

Análogamente,

$$\sum_{i=1}^{n}(S_i - \bar{S})^2 = \frac{n^3 - n}{12}.$$

Finalmente, vamos a obtener la expresión $\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})$ en función de D_i, donde $D_i = S_i - R_i$.

$$D_i = S_i - R_i \Rightarrow \sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n}(S_i - R_i)^2 = \sum_{i=1}^{n}[(S_i - \bar{S}) - (R_i - \bar{R})]^2 =$$

$$\sum_{i=1}^{n}(S_i - \bar{S})^2 + \sum_{i=1}^{n}(R_i - \bar{R})^2 - 2\sum_{i=1}^{n}(S_i - \bar{S})(R_i - \bar{R}) \Rightarrow$$

$$\sum_{i=1}^{n}(S_i - \bar{S})(R_i - \bar{R}) = \frac{\sum_{i=1}^{n}(S_i - \bar{S})^2 + \sum_{i=1}^{n}(R_i - \bar{R})^2 - \sum_{i=1}^{n} D_i^2}{2} =$$

$$\frac{2\frac{n^3 - n}{12} - \sum_{i=1}^{n} D_i^2}{2} = \frac{n^3 - n}{12} - \sum_{i=1}^{n} \frac{D_i^2}{2}.$$

De esta manera, reescribimos la Ecuación 2.4 como:

$$\frac{\sum_{i=1}^{n}(R_i-\bar{R})(S_i-\bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i-\bar{R})^2}\sqrt{\sum_{i=1}^{n}(S_i-\bar{S})^2}} = \frac{\frac{n^3-n}{12}-\sum_{i=1}^{n}\frac{D_i^2}{2}}{\sqrt{\frac{n^3-n}{12}}\sqrt{\frac{n^3-n}{12}}} = 1-\frac{6\sum_{i=1}^{n}D_i^2}{n(n^2-1)}.$$

□

Por ello, al igual que el coeficiente de correlación de Pearson, el coeficiente de correlación de Spearman oscila en el intervalo $[-1,1]$, donde valores mas lejanos al 0 implican mayor grado de asociación.

A continuación, veamos el cálculo del coeficiente de correlación de Spearman mediante un ejemplo hipotético para comprender mejor los conceptos [7].

Ejemplo 2.2.1. Se recogen en la Tabla 2.1 las calificaciones de 10 alumnos en dos materias distintas de la Politècnica de València. Se desea conocer si existe algún tipo de correlación entre las calificaciones obtenidas.

Alumno	Calificación en la materia X (X_i)	Calificación en la materia Y (Y_i)
1	0,1	0,7
2	1,2	1,7
3	1	0,5
4	2,2	2,1
5	0,5	0
6	8	7,2
7	8,3	8,7
8	9	9,2
9	9,5	10
10	8,7	8,9

Cuadro 2.1: Calificaciones en las materias X e Y de 10 alumnos de la UPV.

Lo primero que vamos a hacer es observar si los datos siguen una distribución normal. Para ello, vamos a observar los histogramas, realizaremos un papel probabilístico normal sobre los datos y utilizaremos el test de normalidad de Shapiro-Wilk.

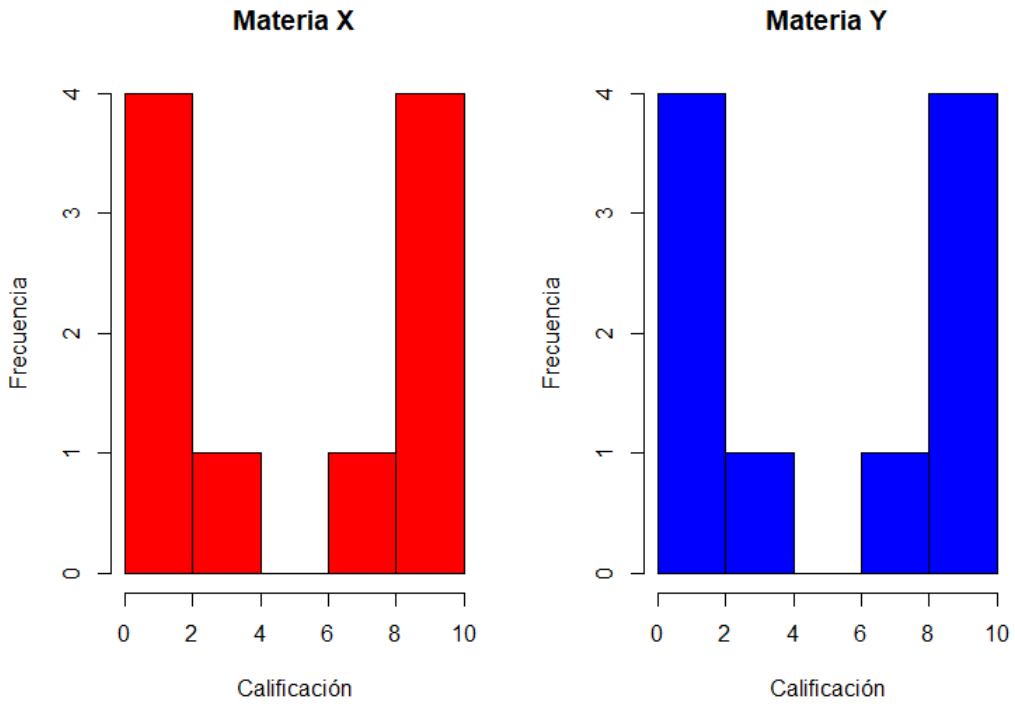


Figura 2.1: Histogramas de las calificaciones de las materias X e Y.

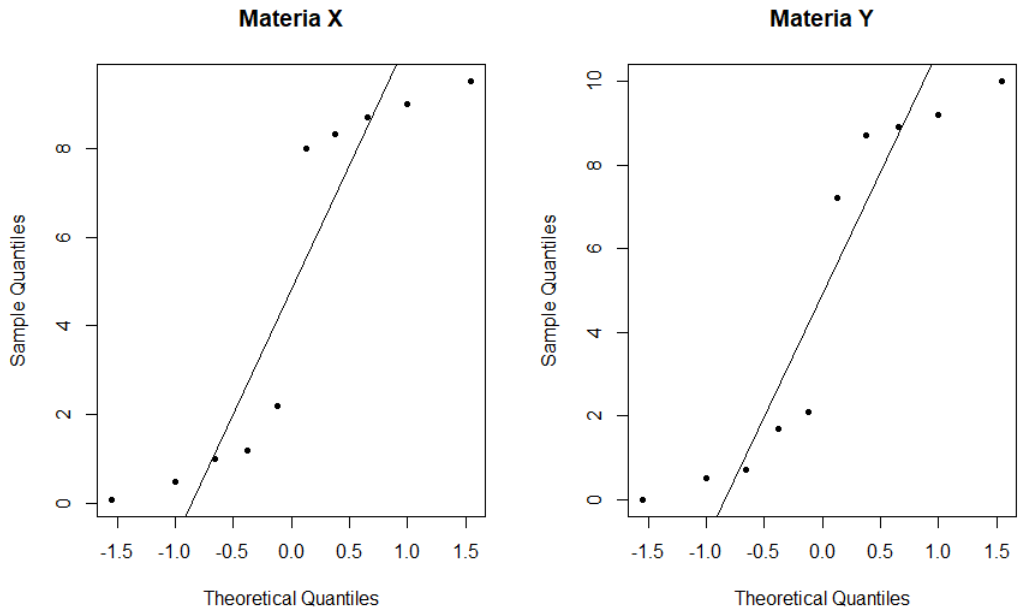


Figura 2.2: Papel probabilístico normal de las calificaciones de las materias X e Y.

Analizando tanto los histogramas como los papeles probabilísticos normales pode-

mos llegar a la conclusión de que las calificaciones de los alumnos en las materias X e Y no siguen una distribución normal. Para asegurarnos, hemos realizado el test de normalidad de *Shapiro-Wilk*. Para comprobar la normalidad de los datos, el test utiliza el siguiente contraste de hipótesis.

$$\begin{cases} H_0: \text{ la distribución es normal} \\ H_1: \text{ la distribución no es normal} \end{cases}$$

Donde obtenemos unos P-valores de 0,01013 y 0,02444 para las calificaciones de las materias X e Y respectivamente. Puesto que estos P-valores son menores a 0,05, podemos rechazar con un 95 % de acierto la hipótesis nula asegurándonos que los datos no siguen una distribución normal.

Dado que no se cumple la normalidad de los datos y no conocemos su distribución, no sería correcto utilizar la correlación de Pearson para determinar la relación entre las calificaciones de las dos materias. Para ello, utilizaremos la correlación de Spearman.

En primer lugar completaremos una tabla con los parámetros necesarios para el cálculo del coeficiente de correlación. Para ello, tal y como ya hemos explicado, necesitaremos obtener los valores de R_i, S_i y D_i. Después, sustituyendo los valores obtenidos en la Ecuación 2.3, obtendremos el coeficiente de correlación de Spearman.

Finalmente, graficaremos las calificaciones de los 10 alumnos en las dos materias para contrastar los resultados obtenidos.

Alumno	X_i	Y_i	R_i	S_i	D_i	D_i^2
1	0,1	0,7	1	3	2	4
2	1,2	1,7	4	4	0	0
3	1	0,5	3	2	-1	1
4	2,2	2,1	5	5	0	0
5	0,5	0	2	1	-1	1
6	8	7,2	6	6	0	0
7	8,3	8,7	7	7	0	0
8	9	9,2	9	9	0	0
9	9,5	10	10	10	0	0
10	8,7	8,9	8	8	0	0

Cuadro 2.2: Método de Spearman para el cálculo de la correlación.

Entonces, utilizando la Ecuación 2.3 tenemos que,

$$r_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)} = 1 - \frac{6(4+1+1)}{10(10^2-1)} = 1 - \frac{36}{990} = 1 - 0,03636 = \mathbf{0,96364}.$$

Como hemos obtenido un coeficiente de correlación de Spearman de **0,96364**, un valor muy cercano al 1, aseguramos que hay una alta correlación positiva entre las calificaciones obtenidas por los 10 alumnos de la UPV en la materia X y en la materia Y. Podemos contrastar la correlación obtenida observando la Figura 2.3.

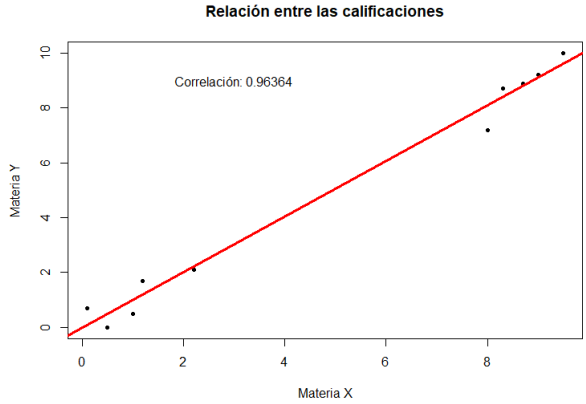


Figura 2.3: Gráfico de dispersión de las calificaciones de las materias X e Y.

En este ejemplo hemos supuesto que no hay valores coincidentes en ninguna de las variables a analizar, este tipo de problemas se denominan "sin empates". Si existiesen valores coincidentes, se pondría el promedio de los rangos que hubiesen sido asignados si no hubiese coincidencias, por ejemplo:

X_i	R_i
0,1	1,5
1,2	3,5
1,2	3,5
2,2	5
0,1	1,5
8	6,5
8	6,5
9	9,5
9	9,5
8,7	8

Cuadro 2.3: Método de Spearman para el cálculo de la correlación con empates.

En este documento, abordaremos únicamente problemas sin empates, puesto que el espacio en el que trabajaremos será el espacio real ($\mathbb{R}$) donde, como sabemos, no se pueden dar los empates.

2.3. Control estadístico de la calidad

2.3.1. Breve historia sobre la calidad

Antes de la revolución industrial, la producción de bienes era elaborada por artesanos especializados, quienes trabajaban toda su vida para perfeccionar estos escasos pero perfectos productos a los que, orgullosos, ponían sus nombres. Sin embargo, el aumento exponencial de la demanda, la producción en línea y la aparición de nuevos sistemas de producción cambiaron completamente el concepto de la producción [8].

Desde ese momento, el artesano dejaba de perfeccionar los productos pasando a ser un trabajador de fábrica más. El objetivo dejaba de ser fabricar productos perfectos, sino fabricar en masa, disminuyendo enormemente la calidad de estos, ya que, se descuidaban las características que satisfacían a los consumidores.

No fue hasta mediados de la década de 1920 que *Walter Shewhart* [9], investigador de *Bell Laboratories*, hizo un descubrimiento en la mejora de la producción: observó que aunque la variación en la fabricación de los productos era inevitable, esta variación podía vigilarse y controlarse utilizando la estadística. Para ello, Shewhart desarrollo el *gráfico de control*; una gráfica con la que se podía determinar cuándo la variación de alguno de los procesos de producción excedía los limites aceptables [10].

Esta mejora que estudió Shewhart establece los estándares de calidad de forma cuantitativa y objetiva que todas las empresas centraban sus esfuerzos en cumplir.

A raíz de estos trabajos, *W. Edwards Deming*, alumno de Shewhart en la década de 1950, comenzó a estudiar la calidad desde un punto de vista más subjetivo, afirmando que: “La calidad puede ser medida sólo en términos de los consumidores” (1986). Deming consideró que había que definir la calidad como “uniformidad alrededor del objetivo”, de forma que la variación del objetivo supondría una desconfianza en los

clientes y por ello un incremento del coste. De esta forma, la calidad dejaba de ser obtener productos que se ajustaran a los límites establecidos, sino la mejora continua de estos para así reducir las variaciones alrededor del objetivo.

Deming: *"Un grado previsible de uniformidad y confiabilidad a bajo costo y adecuado para el mercado."*

Esta conclusión se reitera en la afirmación de Montgomery (2012) [11]: "La calidad es inversamente proporcional a la variabilidad". Por ello, una reducción en la variabilidad en el proceso de producción se comenzó a considerar una mejora de la calidad. Así, comenzaron a desarrollarse nuevas técnicas de control estadístico de la calidad, entre las que se encuentras los gráficos de control, que analizaremos con detalle en este trabajo.

2.3.2. Planificación, control y mejora de la calidad

Las empresas, con el fin de asegurar que los productos que ofertan cumplen las características de calidad requeridas, llevan a cabo un conjunto de procesos que se conocen como *Ingeniería de la calidad*. Estos procesos pueden estructurarse en tres etapas diferentes: *Planificación*, *Control* y *Mejora* de la calidad [11].

- **Planificación de la calidad.** En esta etapa, la empresa determina las características óptimas de sus productos, de manera que cumpla con los requerimientos establecidos por los consumidores.
- **Control estadístico de la calidad.** Forma el conjunto de procesos necesarios para asegurar el cumplimiento de las características establecidas en la etapa de la *planificación de la calidad*.
- **Mejora de la calidad.** Esta etapa se constituye por procesos cuyo propósito consiste en reducir la variabilidad de las características de los productos en su creación.

En este trabajo nos centraremos en la etapa del *Control estadístico de la calidad*. En el *Control estadístico de la calidad* diferenciamos tres procesos que nos ayudan a monitorizar las características del producto: la *medición* de los parámetros que deseamos controlar, la *comparación* de los valores obtenidos con la norma y la *corrección* en caso de que sea necesario.

2.3.3. Control estadístico de la calidad

El *Control estadístico de la calidad* es la parte del control de la calidad donde se utilizan técnicas estadísticas para medir y mejorar la calidad de los procesos. Pueden distinguirse diversos tipos de control estadístico de la calidad, tales como:

- Control de procesos
- Inspección por muestreo

El control de procesos, también conocido como control de la fabricación, tiene como objetivo hacer un seguimiento del funcionamiento de los sistemas para que se aseguren las condiciones de los productos establecidas.

La inspección por muestreo o control de recepción se realiza sobre partidas o lotes de unidades recibidas, ya sean materias primas, semielaboradas o acabadas, con el propósito de inspeccionar si se verifican las especificaciones establecidas (Navarrete, 1998) [12].

En el control estadístico de la calidad se utilizan las siguientes herramientas estadísticas:

- Estadística descriptiva.
- Variables estadísticas.
- Distribuciones.
- Test de hipótesis.
- Análisis de la varianza.
- Estudio de la desviación típica muestral en poblaciones normales
- Estudio de la distribución del rango muestral dentro del contexto de los estadísticos ordenados.

Posteriormente, se introducen técnicas de *control estadístico de procesos* para estabilizar el proceso y reducir su variabilidad.

2.3.4. Control estadístico de procesos

Para que un producto cumpla con los requerimientos establecidos deberá fabricarse mediante un proceso que sea estable o bajo control, es decir, que el proceso tenga poca variabilidad.

El control estadístico de procesos, *SPC* por sus siglas en inglés, se define como la aplicación de técnicas estadísticas con el fin de estabilizar y mejorar la capacidad del proceso mediante la reducción de la variabilidad.

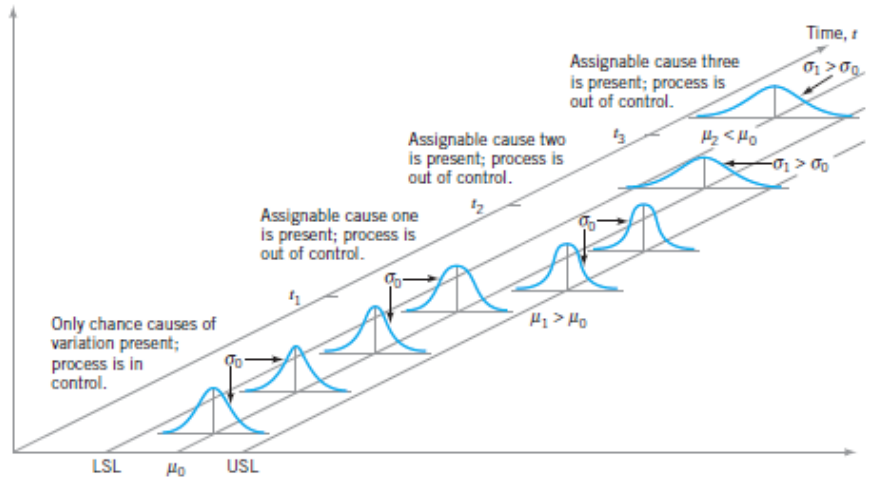


Figura 2.4: Ejemplo de procesos bajo control y fuera de control. (Montgomery, 2012).

El *SPC* es uno de los desarrollos tecnológicos más importantes del siglo 20, es fácil de utilizar, tiene un significante impacto y puede aplicarse a cualquier tipo de proceso (Montgomery, 2012) [11]. Según Montgomery (2012) [11], las siete herramientas principales que se utilizan para realizar el *SPC*, conocidas como **las siete magníficas** son:

- **El histograma o el diagrama de tallo y hoja.** Representan gráficamente la variabilidad existente entre los datos.
- **Hoja de control.** O también llamadas hojas de recogida de datos son formas estructuradas que recopilan información. Las hojas de control son diseñadas en base a las características de los datos que se necesitan para controlar uno o varios procesos.

- **Diagrama de pareto.** Permite clasificar gráficamente la información de mayor a menor relevancia, con el objetivo de reconocer los problemas más importantes presentes en el proceso.
- **Diagrama de causa-efecto.** O diagrama de *Ishikawa* permite identificar los factores potenciales que contribuyen a un problema, para después realizar acciones. correctoras.
- **Diagrama de concentración de defectos.** Es una representación visual, normalmente un diagrama, que muestra todos los defectos del proceso que se analiza.
- **Diagrama de dispersión.** Es la representación gráfica de dos variables para un conjunto de datos. Es decir, analizamos la relación entre dos variables, observando el grado de correlación que presentan.
- **Gráfico de control.** Es un diagrama que monitoriza un estadístico muestral, ubicado en una serie cronológica.

De los siete instrumentos descritos para realizar el *SPC*, profundizaremos en los *gráficos de control*, que pueden ser considerados como la herramienta estrella del SPC.

2.4. Gráficos de Control

Los gráficos de control introducidos por Shewhart (1931) [9] consisten en la monitorización de un estadístico muestral a lo largo del tiempo, realizándose mediciones del estadístico y observando si estas mediciones se encuentran dentro de un intervalo de valores cuyos límites son llamados límites de control. Estos límites se calculan como resultado de observar y caracterizar el comportamiento del estadístico objeto de monitorización cuando se asume que el proceso asociado está bajo control (es decir, es estable). Vemos a continuación un ejemplo de un gráfico de control por Montgomery (2012) [11].

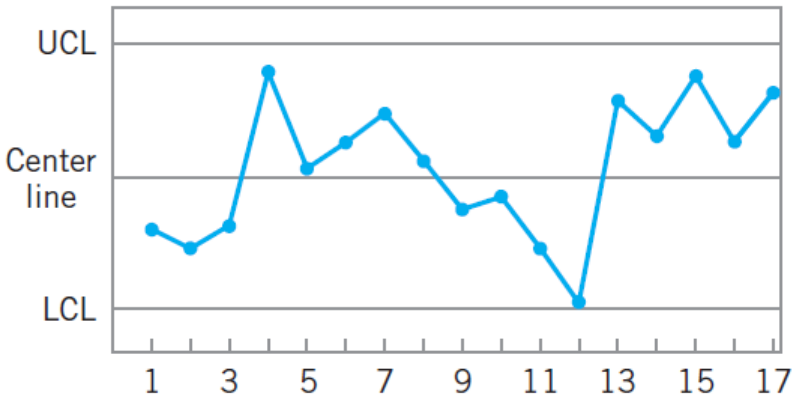


Figura 2.5: Ejemplo gráfico de control. (Montgomery, 2012).

En la Figura 2.5, observamos la monitorización de un estadístico muestral. El valor esperado del estadístico viene representado por la linea central del gráfico (CL por sus siglas en ingles). Además, se observan los límites de control superior e inferior (UCL y LCL, respectivamente).

Para realizar la monitorización, se calcula el estadístico a partir de una muestra aleatoria de n mediciones y se representa el valor mediante un punto en el gráfico. Cuando uno de los puntos sale de los límites de control, se dice que hay una señal o alarma de falta de control.

El gráfico de control se diseña de manera que, cuando el proceso permanece bajo control, (es decir, que la característica de calidad a estudiar sigue una distribución constante en el tiempo o que la variabilidad del proceso solo se ve afectada por causas aleatorias), la probabilidad de obtener una señal de falta de control sea muy pequeña. A esta probabilidad se le conoce como *tasa de falsa alarma*. Además, cuando el proceso está fuera de control, es decir, la variabilidad del proceso se ve afectada por causas asignables o no aleatorias, la probabilidad de obtener una señal de falta de control debe ser alta.

En resumen, podríamos decir que un gráfico de control actúa al igual que un contraste de hipótesis que se ejecuta cada vez que se estudia una nueva muestra.

2.4.1. Desempeño de un gráfico de control

Para evaluar la eficiencia o la potencia de un gráfico de control estándar o de tipo Shewhart, podemos utilizar la *longitud promedio de racha* o *average run length* (ARL) del gráfico de control. En esencia, el ARL es el número promedio de puntos que se deben dibujar para obtener un punto que indique una señal de falta de control (Montgomery, 2012) [11].

Si las observaciones del proceso no están correlacionadas, entonces, para cualquier gráfico de control de tipo Shewhart, se puede calcular el ARL de la siguiente manera:

$$ARL = \frac{1}{p}.$$

Donde p es la probabilidad de que cualquier punto salga de los límites de control. En particular, si el proceso está bajo control, p es la probabilidad de falsa alarma, que se denota como α. Por lo que, en este caso, tenemos que,

$$ARL_0 = \frac{1}{\alpha}.$$

En cambio, cuando el proceso esta fuera de control, la probabilidad de obtener un punto fuera de los límites de control se mide mediante la potencia estadística del proceso, es decir, la probabilidad de rechazar la hipótesis nula correctamente. En este caso, tenemos que,

$$ARL_1 = \frac{1}{Potencia} = \frac{1}{1-\beta}.$$

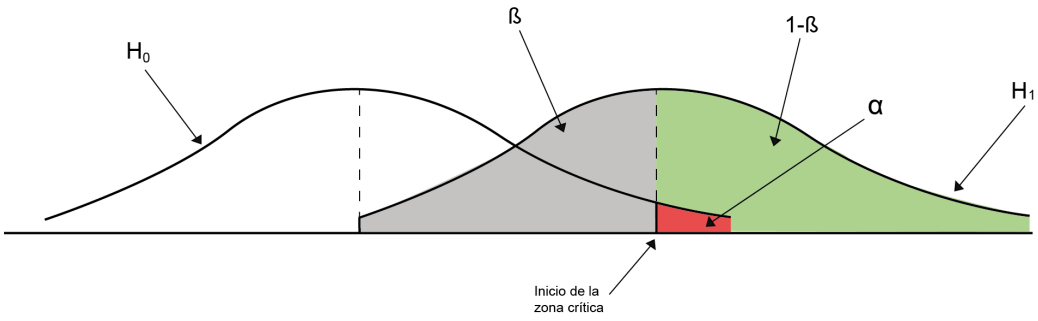


Figura 2.6: Gráfico de probabilidad de error de I y II especie. Elaboración propia.

Capítulo 3

Diseño de un gráfico de control para monitorizar la correlación de Spearman

Una vez presentado los antecedentes teóricos del trabajo, vamos a desarrollar la metodología llevada a cabo para diseñar un gráfico de control que permita detectar la correlación entre dos variables continuas a través de el coeficiente de correlación de Spearman. Asimismo, se mostrará cómo evaluar la eficiencia, potencia o desempeño de este gráfico.

Las tareas que se han realizado para completar el trabajo pueden agruparse en tres etapas diferentes:

- **Etapa 1:** Calcular el límite de control superior, UCL, para distintos tamaños muestrales y diferentes probabilidades de error de primera especie (α).

- **Etapa 2:** Calcular el desempeño del gráfico de control (ARL_0 y ARL_1) teniendo en cuenta los siguientes parámetros de entrada: n : tamaño muestral, α : probabilidad de error de primera especie, r_S : valor esperado de la correlación de Spearman muestral, F_X y F_Y : distribuciones de las variables X e Y, respectivamente, Cópula: cópula utilizada para modelar el comportamiento conjunto de X e Y y finalmente, $NMAX$: cantidad de puntos fuera de control necesarios para la estimación del ARL_1 mediante simulación Montecarlo.

- **Etapa 3:** Estudiar el efecto de cada uno de los factores para el cálculo del ARL_1 realizando un ANOVA sobre la tabla de efectos cruzados del proceso.

Cabe aclarar que, en paralelismo con los contrastes de hipótesis no paramétricos existentes en la literatura basados en el estadístico de Spearman (Hollander, 2014 [5] y Seskin, 2011 [13]), el gráfico que diseñaremos parte de la premisa que la situación que se considera bajo control es la de *independencia* de las variables X e Y objeto de interés, y que, por tanto, una señal de falta de control se asociará con la existencia de correlación entre ambas.

3.1. Etapa 1: Cálculo del UCL

Por simplicidad, en este trabajo únicamente vamos a monitorizar correlaciones no negativas, es decir, valores comprendidos en el intervalo [0,1], por esta razón, el gráfico de control que vamos a diseñar será un gráfico de control "unilateral", es decir, tan sólo necesitaremos el UCL para monitorizar el proceso.

Para calcular el UCL del gráfico de control, tenemos que tener en cuenta la definición del mismo,

$$\alpha = P(U_t > UCL).$$

Donde U_t es la distribución teórica del coeficiente de correlación de Spearman para un determinado tamaño muestral. Puesto que esta distribución no es trivial, vamos a hacer uso de las funciones "pspearman" del paquete estadístico "pspearman" desarrollado por VanDeWiel (2001) [14] y "pSpearman" del paquete estadístico "SuppDists" utilizado por Hollander (2014) [5]. Ambas funciones han sido implementadas en el software estadístico *R* y son capaces de dar el valor exacto de la CDF y la inversa de la CDF de r_s bajo el supuesto de independencia estadística entre X e Y, y por eso es el instrumento adecuado para calcular UCL.

De esta manera, estas funciones nos proporcionan los valores exactos del UCL para un valor de α y un tamaño muestral $n < 100$ deseado. Para profundizar más sobre la distribución en el muestreo del coeficiente de correlación de Spearman, pueden consultarse los trabajos de (Hollander, 2014 [5] y Seskin, 2011 [13]).

3.2. Etapa 2: Cálculo del desempeño del gráfico de control

En esta sección se busca proporcionar las herramientas necesarias para el cálculo del ARL_0 y ARL_1 del gráfico de control. En primer lugar, calcular el desempeño del gráfico de control cuando el proceso está bajo control (se supone la independencia de las variables) resulta ser trivial, puesto que,

$$ARL_0 = \frac{1}{\alpha}.$$

En cambio, cuando el proceso está fuera de control (existencia de correlación), el cálculo del desempeño no es trivial puesto que no se dispone de la función de distribución necesaria para el cálculo de la potencia estadística (Hollander, 2014) [5]. Por ello, el cálculo del desempeño del gráfico en este caso se realizará mediante simulación Montecarlo.

Para poder realizar la simulación, debemos introducir algunos conceptos matemáticos.

3.2.1. Cópulas

En estadística, una cópula es, en esencia, una distribución multivariante (bivariante, en nuestro caso), con la característica de que las distribuciones marginales siguen una distribución uniforme entre 0 y 1. Combinada con un conjunto de variables aleatorias $(X_1, ..., X_p)$ referidas a una misma población, una cópula $(U_1, ..., U_p) \sim C$ permite describir el comportamiento multivariante de dichas variables. La cópula se *conecta* a las variables a partir de la distribución de cada una de ellas. En concreto, cada valor posible x_i de X_i está conectado con un valor $u_i \in [0, 1]$ correspondiente a la componente i-ésima de la cópula C de la siguiente forma: $x_i = F_{X_i}^{-1}(u_i)$, siendo F_{X_i} la función de distribución acumulada de X_i (o sea, su distribución marginal). De esta forma, la cópula termina siendo una estructura que *encapsula* toda la información acerca del comportamiento conjunto de las diferentes variables consideradas.

En este trabajo, utilizaremos una familia de cópulas llamadas cópulas *Arquimedianas*. Esta familia de cópulas tiene la característica de poder modelar la dependencia o

correlación entre una dimensión alta de variables aleatorias utilizando un único parámetro θ. En particular, utilizaremos las siguientes tres cópulas Arquimedianas.

Cópula de Frank

La cópula de *Frank* [15] se define como,

$$C(u_1, u_2|\theta) = -\frac{1}{\theta} ln\left(1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1}\right).$$

Donde $\theta \in \mathbb{R}\backslash\{0\}$. Para la cópula de Frank, se puede probar que el coeficiente de correlación de Kendall, τ, esta relacionado con el parámetro de dependencia θ mediante la siguiente ecuación:

$$\tau = 1 + \frac{4(D_1(\theta) - 1)}{\theta}. \tag{3.1}$$

Donde $D_1(\theta)$ se corresponde a la *función de Debye* de primera clase,

$$D_1(\theta) = \frac{1}{\theta}\int_0^\theta \frac{t}{e^t - 1} dt.$$

Ahora, nuestro objetivo es calcular el ARL del gráfico de control cuando se supone correlación de Spearman que valga, en promedio, r_S. Sabemos que hay una relación directa entre el valor del parámetro θ que determina el grado de relación entre las dos componentes de la cópula y el coeficiente de correlación de Kendall, determinada por la Ecuación 3.1. Además, existe una relación aproximada, asintótica, entre los valores muestrales de la τ de Kendall y la r_S de Spearman, que se aproxima a $\frac{\tau}{r_S} \approx 0{,}67$ (Sheskin, 2011) [13]. Por ello, asumir un valor para r_S se puede traducir, de forma aproximada, en un valor para θ.

Una vez obtenido el parámetro de dependencia θ, en función de la correlación esperada, r_S, podemos generar los valores de las variables aleatorias (u_1, u_2) de la siguiente manera [16]:

(i) Generar dos valores aleatorios independientes (u_1, v_2) de una distribución uniforme [0,1].

(ii) Establecer $u_2 = -\frac{1}{\theta} ln\left(1 + \frac{v_2(1 - e^{-\theta})}{v_2(e^{-\theta u_1} - 1) - e^{-\theta u_1}}\right)$.

Cópula de Clayton

La cópula de *Clayton* [17] se define como,

$$C(u_1, u_2|\theta) = max(0, u_1^{-\theta} + u_2^{-\theta} - 1)^{-\frac{1}{\theta}}.$$

Donde $\theta \in [-1, \infty)\backslash\{0\}$. Para la cópula de Clayton, se puede probar que el coeficiente de correlación de Kendall, τ, esta relacionado con el parámetro de dependencia θ mediante la siguiente ecuación:

$$\tau = \frac{\theta}{\theta + 2} \leftrightarrow \theta = \frac{2\tau}{1 - \tau}.$$

Una vez obtenido el parámetro de dependencia θ, en función de la correlación esperada, r_S, podemos generar los valores de las variables aleatorias (u_1, u_2) de la siguiente manera:

(i) Generar dos valores aleatorios independientes (u_1, v_2) de una distribución uniforme [0,1].

(ii) Establecer $u_2 = \left(u_1^{-\theta}(v_2^{\frac{-\theta}{1+\theta}} - 1) + 1\right)^{\frac{-1}{\theta}}$.

Cópula de Gumbel

La cópula de *Gumbel* [18] se define como,

$$C(u_1, u_2|\theta) = exp\left(-\left((-ln(u_1))^\theta + (-ln(u_2))^\theta\right)^{\frac{1}{\theta}}\right).$$

Donde $\theta \in [1, \infty)$. Para la cópula de Gumbel, se puede probar que el coeficiente de correlación de Kendall, τ, esta relacionado con el parámetro de dependencia θ mediante la siguiente ecuación:

$$\tau = 1 - \frac{1}{\theta} \leftrightarrow \theta = \frac{1}{1 - \tau}.$$

Una vez obtenido el parámetro de dependencia θ, en función de la correlación esperada, r_S, podemos generar los valores de las variables aleatorias (u_1, u_2) de la siguiente manera:

(i) Generar dos valores aleatorios independientes (v_1, v_2) de una distribución uniforme [0,1].

(ii) Establecer $w\left(1-\frac{ln(w)}{\theta}\right)=v_2$ y resolverlo de manera numérica para $0<w<1$.

(iii) Establecer $u_1=exp[v_1^{\frac{1}{\theta}}ln(w)]$ y $u_2=exp[(1-v_1)^{\frac{1}{\theta}}ln(w)]$.

3.2.2. Distribuciones de X e Y

Una vez que hemos obtenido los valores de (u_1,u_2) utilizando las cópulas definidas, podemos generar los valores correspondientes de X e Y utilizando la inversa de sus respectivas distribuciones. Es decir,

$$x_1=F_X^{-1}(u_1) \quad \text{y} \quad y_1=F_Y^{-1}(u_2).$$

Donde F_X y F_Y son las funciones de distribución de X e Y, respectivamente. En particular, en este trabajo, analizaremos las funciones de distribución $N(a_0,b_0)$ y $Weibull(a_0,b_0)$, donde $\mu_0=10$ y $\sigma_0\in\{1,2,5\}$.

Con el objetivo de obtener la media y la desviación típica deseada para la distribución de Weibull, se han calculado los parámetros necesarios mediante la búsqueda de raíces de las siguientes ecuaciones [19].

$$\frac{\Gamma\left(\frac{2}{a_0}+1\right)}{\Gamma^2\left(\frac{1}{a_0}+1\right)}=\left(\frac{\sigma_0}{\mu_0}\right)^2+1,$$

y

$$b_0=\frac{\mu_0}{\Gamma\left(\frac{1}{a_0}+1\right)}.$$

Distribución	a_0	b_0	μ_0	σ_0
Normal	10	1	10	1
	10	2	10	2
	10	5	10	5
Weibull	12,1534	10,4304	10	1
	5,7974	10,7998	10	2
	2,1013	11,2906	10	5

Cuadro 3.1: Parámetros de las distribuciones Normal y Weibull para las diferentes medias y desviaciones típicas.

3.2.3. Experimentación

Con el objetivo de obtener una estimación robusta del índice de desempeño cuando el proceso está fuera de control, ARL_1, se procede a realizar una simulación del proceso. Como ya se ha mencionado en la Sección 3.2, la simulación se realiza puesto que no se puede modelizar analíticamente la distribución en el muestreo del estadístico r_S cuando no se da la independencia entre las dos variables objeto de estudio.

Para ello, se construye una muestra de X e Y de tamaño n utilizando el procedimiento descrito anteriormente. Después, se calcula el índice de correlación de Spearman de las dos variables aleatorias, y se determina si es mayor o menor al límite de control.

Este proceso se repite tantas veces como se desee, contabilizando la cantidad de pruebas realizadas y la cantidad de prueba que obtienen un índice de correlación fuera del límite de control. Mediante pruebas computacionales previas se ha determinado que un valor de 30.000 para $NMAX$ es adecuado para obtener estimaciones del ARL_1 suficientemente precisas, convergentes al segundo decimal.

3.3. Etapa 3: Diseño de experimentos

Con el objetivo de evaluar la influencia de diferentes parámetros en el comportamiento de nuestro gráfico, se ha realizado un diseño de experimentos (Ridge y Kudenko, 2010) [20]. Para ello, se realiza un análisis de la varianza (ANOVA) utilizando el software *Statgraphics*.

En este análisis, únicamente vamos a tener en cuenta los efectos simples de los factores, es decir, no vamos a introducir ninguna interacción en el modelo debido a su difícil interpretación. Más concretamente, los factores a analizar son los siguientes:

- **Tamaño muestral n:** 2 factores, 10 y 30.
- **Correlación esperada r_S:** 5 factores, 0.1, 0.3, 0.5, 0.7 y 0.9.
- **Distribuciones de X e Y:** 3 factores, Normal y Normal (N,N), Normal y Weibull (N,W) y Weibull y Weibull (W,W).

- **Desviaciones típicas de X e Y:** 9 factores, (1,1), (1,2), (1,5), (2,1), (2,2), (2,5), (5,1), (5,2), (5,5).
- **Tipo de cópula:** 3 factores, Frank, Clayton y Gumbel.

Como es habitual en diseño de experimentos, para cada uno de los factores que se sospecha que pueden tener influencia en la aptitud del gráfico para detectar la presencia de correlación, se han seleccionado niveles representativos y suficientemente diferentes entre sí.

Destacamos que el objetivo de este diseño de experimentos es explicar mejor qué está pasando con los datos obtenidos, es decir, poder describir mejor los resultados y no tanto inferir sobre ellos. Los resultados de este ANOVA se muestran y analizan el Capítulo 4.

Capítulo 4

Experiencia computacional

En este capítulo, se detallan los resultados obtenidos al aplicar la metodología descrita para evaluar el desempeño de un gráfico de control para la monitorización de la correlación de Spearman de dos variables aleatorias. Se destaca que para la totalidad de la experimentación, se ha utilizado una probabilidad de error de primera especie de $\alpha = 0{,}0027$, por ser un valor habitual o de referencia en gráficos de control (en el contexto de la monitorización de la media de una variable aleatoria normal univariante).

En el Apéndice A, se detallan los algoritmos diseñados para realizar la experimentación del trabajo. En particular, estos algoritmos han sido diseñados en el software estadístico *R*.

4.1. Resultados de la Etapa 1

El objetivo de la primera etapa es proporcionar el límite de control superior para los diferentes tamaños de muestra y probabilidades de error de primera especie.

Al ejecutar el algoritmo A.1, se obtienen de manera genérica los límites de control superiores para cualquier tamaño muestral y probabilidad de primera especie. Además, se ha calculado el ARL_0 para cada caso.

α	ARL_0	$n = 10$	$n = 15$	$n = 20$	$n = 25$	$n = 30$
0.0005	2000	0.89696	0.77678	0.69248	0.61923	0.57107
0.001	1000	0.87272	0.74821	0.66090	0.5892	0.54215
0.0025	400	0.82424	0.69821	0.61127	0.54461	0.49988
0.0027	370.370	0.82424	0.69464	0.60676	0.54076	0.49588
0.005	200	0.78787	0.65178	0.56917	0.50615	0.46384
0.01	100	0.73939	0.60178	0.52105	0.46346	0.42335
0.025	40	0.64242	0.51964	0.44586	0.39769	0.36195
0.05	20	0.55757	0.44464	0.3796	0.33769	0.30678
0.1	10	0.44848	0.35178	0.29849	0.26615	0.2413
0.15	6.66667	0.36363	0.2875	0.24285	0.21692	0.19644
0.2	5	0.30303	0.23392	0.19774	0.17692	0.15995

Cuadro 4.1: Limite de control superior (UCL) y ARL_0 en función del tamaño muestral y α.

Tal y como se observa en la Tabla 4.1, al aumentar el tamaño muestral, se disminuye el límite de control para una misma probabilidad de error de primera especie. Esto se da, puesto que, al aumentar el tamaño muestral, el cálculo de la correlación será mucho más preciso. De la misma manera, al aumentar la probabilidad de error de primera especie, se disminuye enormemente el UCL y, así mismo, el ARL_0.

4.2. Resultados de la Etapa 2

Una vez calculado el desempeño cuando el gráfico está bajo control, el siguiente paso es calcular el desempeño cuando el gráfico está fuera de control, es decir, cuando existe una correlación entre las variables.

En esta sección se presentan los resultados obtenidos en el cálculo del ARL_1 en función de los factores detallados en la Sección 3.3. Al ejecutar el Algoritmo A.2. En primer lugar, se realiza el cálculo del parámetro de dependencia θ de las diferentes cópulas en función de la correlación r_S utilizando el procedimiento descrito en la Sección 3.2.1.

r_S	τ	θ Frank	θ Clayton	θ Gumbel
0.1	0.067	0.60520	0.14362	1.07181
0.3	0.201	1.87083	0.50312	1.25156
0.5	0.335	3.32576	1.00751	1.50375
0.7	0.469	5.19950	1.76647	1.88323
0.9	0.603	8.01028	3.03778	2.51889

Cuadro 4.2: Valores de θ de las diferentes cópulas en función de la correlación r_S.

Una vez calculados los valores del parámetro de dependencia θ para las diferentes cópulas y correlaciones, vamos a visualizar gráficamente la función de densidad de cada una de las cópulas tomando unas distribuciones marginales $N(10,5)$ y $W(2{,}1013, 11{,}2906)$ con una correlación esperada de $r_S = 0{,}5$ utilizando el software *Mathematica*.

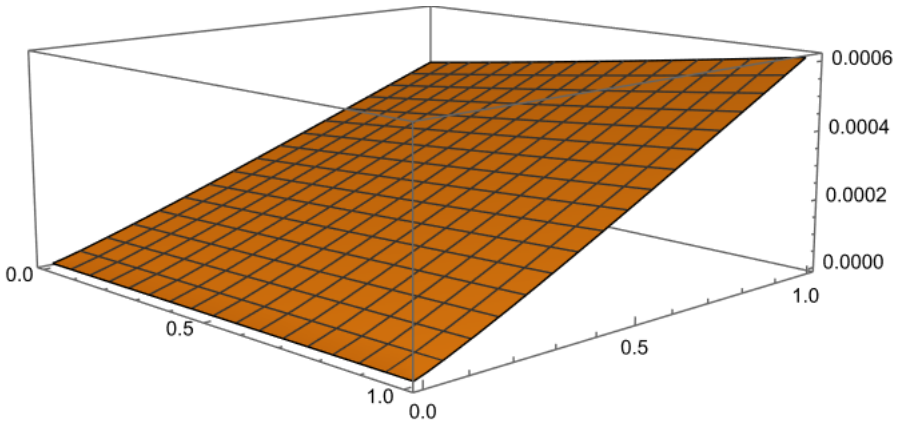


Figura 4.1: CDF utilizando la cópula de Frank.

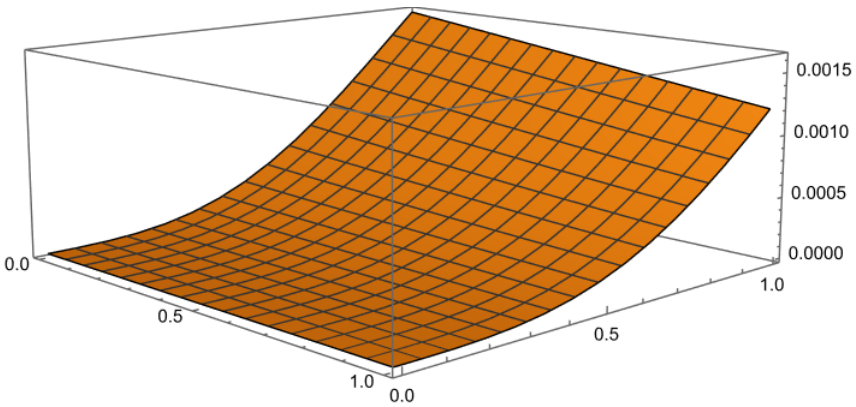


Figura 4.2: CDF utilizando la cópula de Clayton.

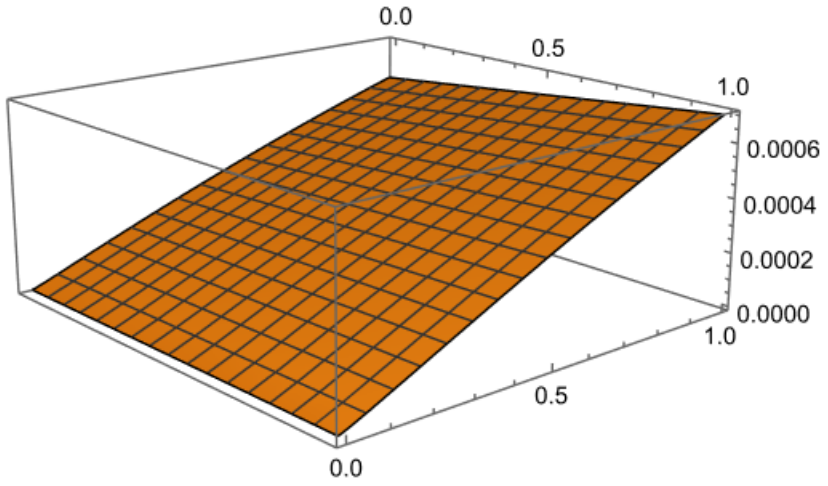


Figura 4.3: CDF utilizando la cópula de Gumbel.

Observando los gráficos de las funciones de densidad, podemos apreciar como aunque las distribuciones marginales pueden ser siempre la misma Normal y la misma Weibull, la cópula que usemos influye mucho en la forma de la distribución. El código utilizado para obtener las gráficas de las funciones de densidad se encuentra en la Sección A.4 del Apéndice A

Ahora, vamos a realizar la simulación para el cálculo del ARL_1. En la Tabla 4.3 se muestra el ARL_1 y la probabilidad de error de segunda especie, β, obtenido para 10 combinaciones de factores diferentes. La tabla completa con la totalidad de los resultados obtenidos se encuentra en la Sección B.1 del Apéndice B.

r_S	Dist.X	Dist.Y	Cópula	ARL_1(n=10)	β (n=10)	ARL_1(n=30)	β (n=30)
0,1	N(10,1)	N(10,1)	Frank	211,554	0,99527	86,08533	0,98838
0,1	N(10,2)	W(5.7974,10.7998)	Clayton	199,03933	0,99497	81,85133	0,98778
0,3	N(10,1)	N(10,1)	Frank	57,244	0,98253	8,75666	0,88580
0,3	W(12.1534,10.4304)	W(12.1534,10.4304)	Gumbel	53,28533	0,98123	8,70133	0,88507
0,5	N(10,5)	W(12.1534,10.4304)	Clayton	16,372	0,93892	2,212	0,54792
0,5	N(10,1)	N(10,2)	Frank	17,99533	0,94443	2,142	0,53314
0,7	N(10,1)	N(10,1)	Frank	6,412	0,84404	1,124	0,11032
0,7	N(10,1)	W(12.1534,10.4304)	Gumbel	6,09933	0,83604	1,17533	0,14917
0,9	N(10,1)	N(10,1)	Frank	2,47066	0,59525	1,00066	0,00066
0,9	W(2.1013,11.2906)	W(2.1013,11.2906)	Clayton	2,53667	0,60578	1,00733	0,00727

Cuadro 4.3: Muestra de los resultados del ARL_1 para las diferentes combinaciones de factores.

A simple vista, analizando la Tabla 4.3, no es posible determinar el efecto que produce cada factor en el cálculo del ARL_1. Con el fin de determinar el efecto de cada factor, se ha realizado un análisis ANOVA de los resultados obtenidos.

4.3. Resultados de la Etapa 3

Los resultados obtenidos por el análisis ANOVA son de valor descriptivo, más que inferencial, con el objetivo principal de describir los resultados que obtenemos de las experiencias numéricas del cálculo del ARL_1.

En primer lugar, vamos a analizar la significación estadística de cada factor.

Analysis of Variance for ARL1 - Type III Sums of Squares

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
MAIN EFFECTS					
A:n	238372,	1	238372,	425,87	0,0000
B:r_s	1,86098E6	4	465246,	831,20	0,0000
C:Distrib	4,97187	2	2,48594	0,00	0,9956
D:Desv.Tip	16,7755	8	2,09694	0,00	1,0000
E:Copula	331,214	2	165,607	0,30	0,7440
RESIDUAL	342553,	612	559,727		
TOTAL (CORRECTED)	2,44226E6	629			

Figura 4.4: Análisis ANOVA sobre el ARL_1.

Antes de analizar los efectos de los factores debemos determinar si el modelo ANOVA es válido. Para ello, deben cumplirse los supuestos de independencia, normalidad y homocedasticidad de los residuos.

4.3.1. Validación del ANOVA

Independencia

Para analizar la independencia de los residuos, vamos a realizar un gráfico de dispersión de los mismos.

Observando el Figura 4.5, está claro que los residuos no son independientes y siguen alguna especie de patrón. Por ello, no se cumple es supuesto de independencia.

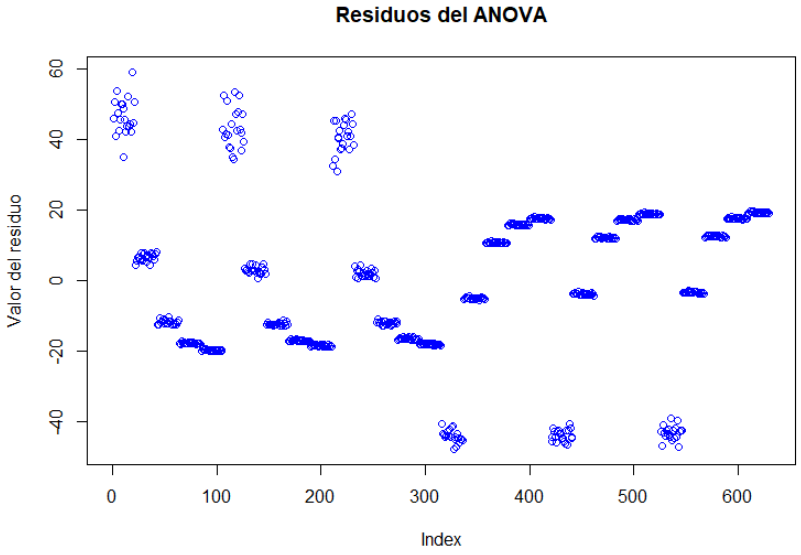


Figura 4.5: Gráfico de dispersión de los residuos.

Normalidad

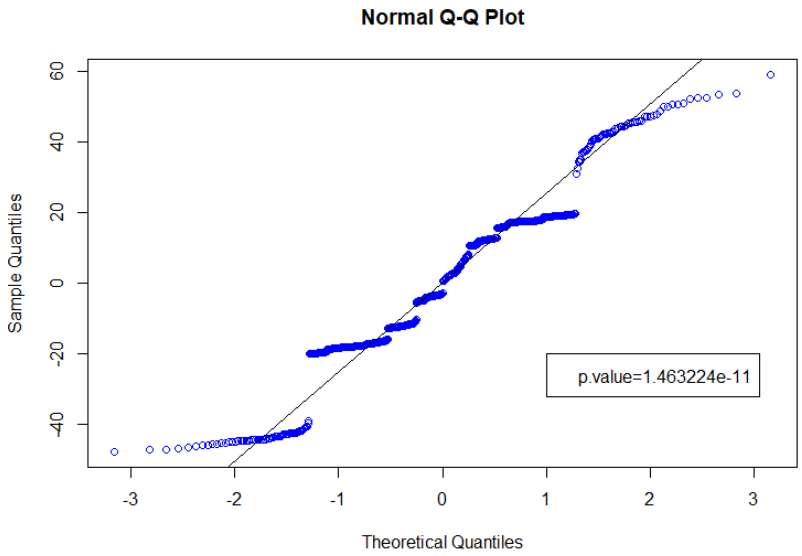


Figura 4.6: Papel probabilístico normal de los residuos.

Analizando tanto el papel probabilístico normal como el P-Valor obtenido en el test de Shapiro Wilk, podemos asegurar que los residuos no siguen una distribución normal. De nuevo, no se cumple el supuesto de normalidad.

Homocedasticidad

Analicemos ahora si la varianza de los residuos es constante en los diferentes niveles de los factores. Para ello, vamos a realizar el test de Bartlett para cada factor. Donde se toma como hipótesis nula del test la igualdad de las varianzas.

	n	r_S	Distribuciones	Desv.Típ	Cópula
P-Valor	0.9159	0	0.9875	0.9999	0.7246

Cuadro 4.4: P-Valor del test de Bartlett para los diferentes factores

Observando los P-Valores de la Tabla 4.4, observamos que la varianza de los residuos no es constante en función de la correlación esperada, r_S. En particular, vamos a observar un gráfico de cajas y bigotes de los residuos en función de la correlación esperada.

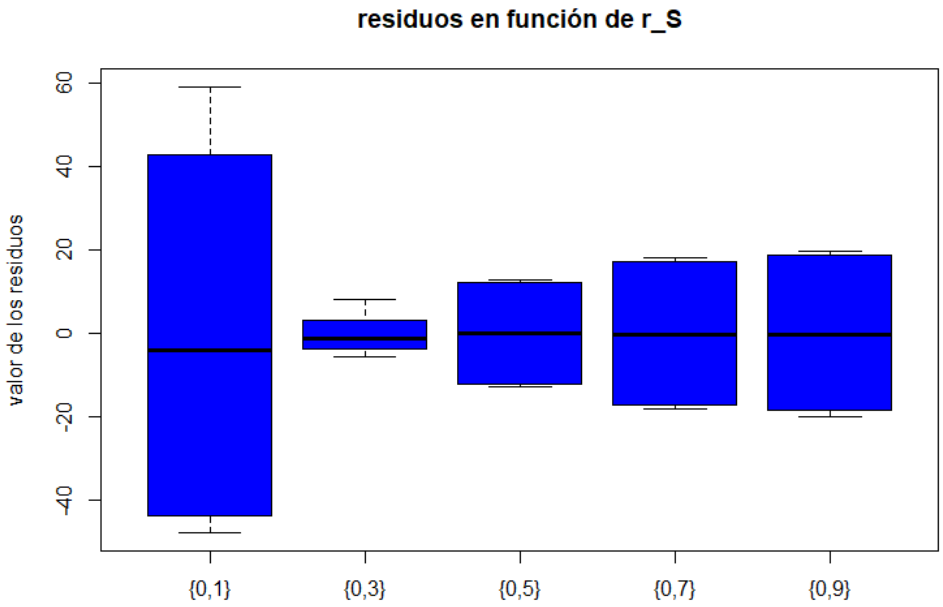


Figura 4.7: Varianza de los residuos en función de la correlación esperada.

La varianza de los residuos cuando hay una correlación de 0.1, es mucho mayor a la varianza del resto de niveles del factor. Por ello, tampoco se cumple el supuesto de homocedasticidad. Se puede observar la varianza de los residuos en función del resto de factores en la Sección B.2 del Apéndice B.

Dado que no se cumplen las hipótesis del ANOVA, no podemos tomar como válido el modelo. Por ello, vamos a realizar el análisis no paramétrico de Kruskall-Wallis.

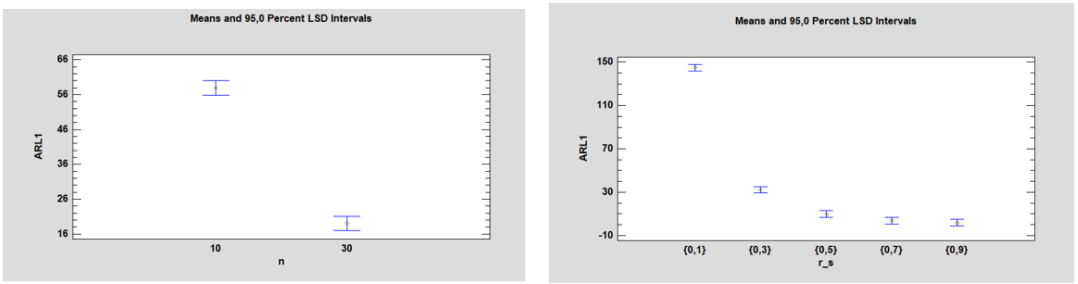
	n	r_S	Distribuciones	Desv.Típ	Cópula
P-Valor	0	0	0.9993	1	0.9608

Cuadro 4.5: P-Valor del test de Kruskall-Wallis para los diferentes factores.

Destacamos que los resultados obtenidos por el análisis no paramétrico de Kruskall-Wallis son similares a los resultados obtenidos por el modelo ANOVA. En ambos análisis, el tamaño muestral, n, y la correlación esperada, r_S, resultan ser altamente significativas. En cambio, las distribuciones, las desviaciones típicas y las cópulas no resultan ser estadísticamente significativas.

4.3.2. Interpretación de los resultados

Con el fin de analizar con mayor detalle el efecto de cada factor, volvemos al ANOVA paramétrico con carácter descriptivo más que inferencial, realizando los intervalos LSD de Fisher al 95 %.



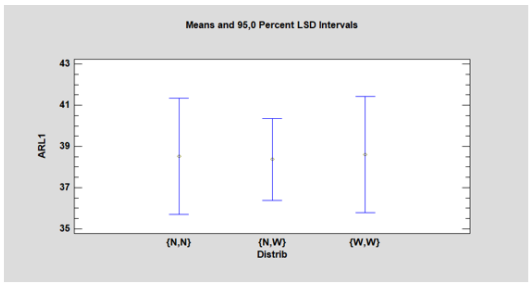
(a) Intervalos LSD para tamaño muestral. (b) Intervalos LSD para correlación.

Figura 4.8: Intervalos LSD al 95 %.

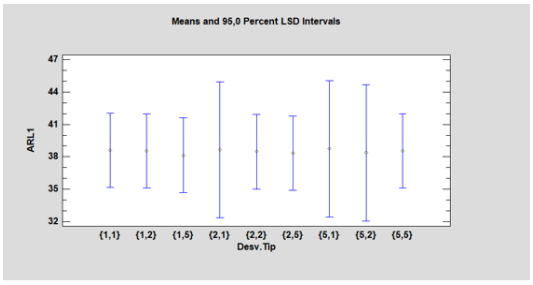
Se aprecia claramente cómo el aumento del tamaño muestral reduce significativamente el ARL_1 del gráfico de control, por lo que, cuanto mayor sea el tamaño muestral, mayor capacidad de detección de correlación tendrá el gráfico de control.

Otro resultado esperable es el comportamiento del ARL_1 en función de la correlación. Cuanto mayor es la correlación a detectar, menor el ARL_1. Sin embargo, un resultado destacable es el comportamiento de la disminución del ARL_1 en función del aumento de la correlación. Vemos como disminuye enormemente al pasar de una correlación de 0.1 a 0.3, en cambio, de ahí en adelante, el ARL_1 parece no disminuir. Por

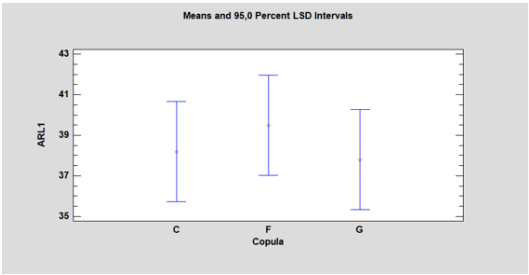
ello, el ARL_1 parece seguir un comportamiento con descenso exponencial.



(a) Intervalos LSD para las distribuciones.



(b) Intervalos LSD para las desviaciones típicas.



(c) Intervalos LSD para las cópulas.

Figura 4.9: Intervalos LSD al 95 %.

Respecto a los factores que no resultan ser estadísticamente significativos, observando la Figura 4.9, vemos que tanto la media como la desviación típica del ARL_1 resultan ser prácticamente idénticas independientemente del tipo de distribución que sigan X e Y y las desviaciones típicas que presenten.

Respecto a las cópulas utilizadas, podemos determinar que el uso de la cópula de Frank proporciona unos valores de ARL_1 ligeramente superiores, pero, de nuevo, no resulta ser un cambio estadísticamente significativo.

Los resultados numéricos de las medias y las desviaciones típicas de cada nivel de los factores podemos observarlo con mayor detalle en la Tabla B.5 del Apéndice B.

Con esto termina el capítulo 4, donde se han proporcionado tablas numéricas para los valores del límite de control en función de la probabilidad de error de primera

especie y del tamaño muestral y para los valores del ARL_1 es función de la correlación, distribuciones de X e Y y cópulas descritas. Además, se ha logrado comprender el efecto de cada factor en el cálculo del ARL_1.

Capítulo 5

Ejemplo numérico

En este capítulo vamos a utilizar los resultados obtenidos en el Capítulo 4 para construir un ejemplo numérico final y así mostrar de forma sencilla el comportamiento del gráfico de control no paramétrico [21].

Para este ejemplo, vamos a fijar los siguientes parámetros:

- Probabilidad de error de primera especie de 0.05 ($\alpha = 0{,}05$).
- Tamaño muestral de 25 ($n = 25$).
- Distribución de X Normal con media 5 y desviación típica 3 ($X \sim N(5, 3)$).
- Distribución de Y Weibull con media 10 y desviación típica 5 ($Y \sim Weibull(2{,}1013, 11{,}2906)$).
- Cópula de Clayton.

5.1. Proceso bajo control

En primer lugar, vamos a ver el comportamiento del gráfico de control cuando el proceso está bajo control, es decir, no hay correlación esperada ($r_S = 0$).

En primer lugar, observando la Tabla 4.1, determinamos que debemos establecer el límite de control superior en 0.33769 ($UCL = 0{,}33769$). Una vez establecido el límite de control superior, se han monitorizado 1000 ensayos o pruebas independientes.

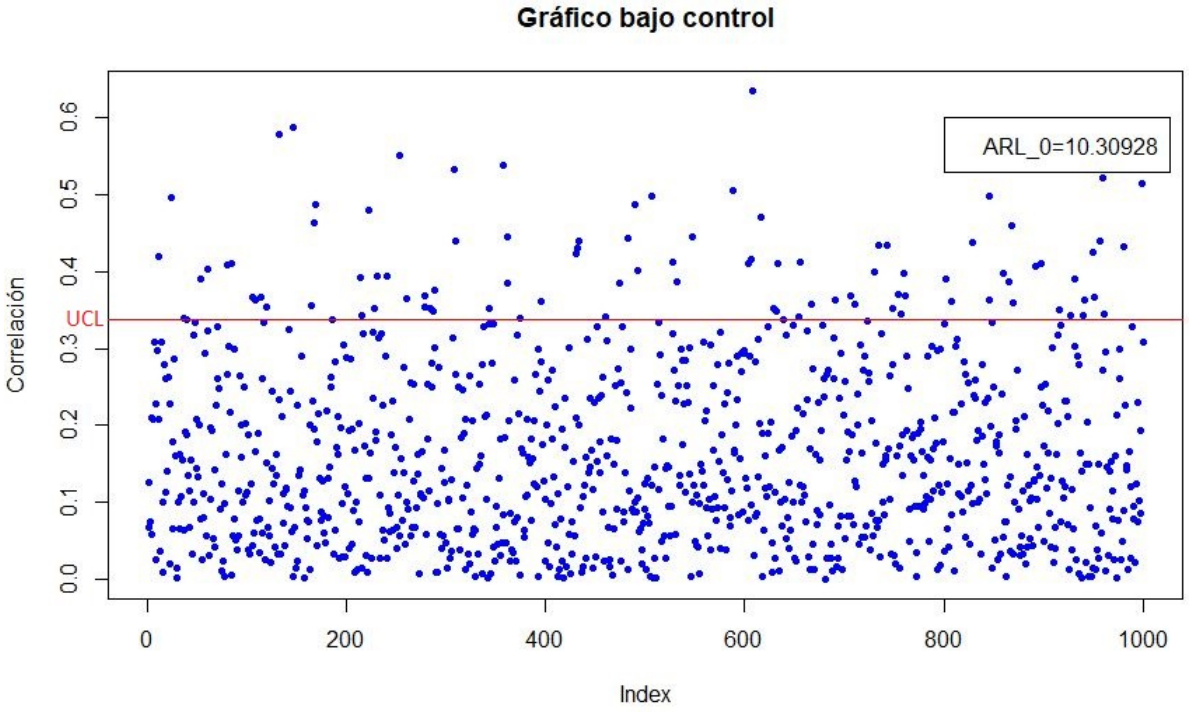


Figura 5.1: Gráfico bajo control ($r_S = 0$).

Cada punto en el gráfico muestra una medición de correlación de Spearman para una muestra de 25 valores de X e Y. Puesto que, el proceso está bajo control, se observa que casi la mayoría de las pruebas realizadas se comprenden en los límites establecidos. Además, se espera un ARL_0 teórico de $1/\alpha = 20$. Sin embargo, obtenemos un ARL_0 observado de 10,3.

5.2. Proceso fuera de control

Veamos ahora el comportamiento del gráfico cuando existe una correlación entre las variables. En concreto, vamos a suponer que hay una correlación de $r_S = 0{,}7$. Como hemos dicho, utilizaremos la cópula de Clayton para obtener los valores de X e Y.

De nuevo, se han monitorizado 1000 ensayos o pruebas independientes y establecido un límite de control superior de 0.33769.

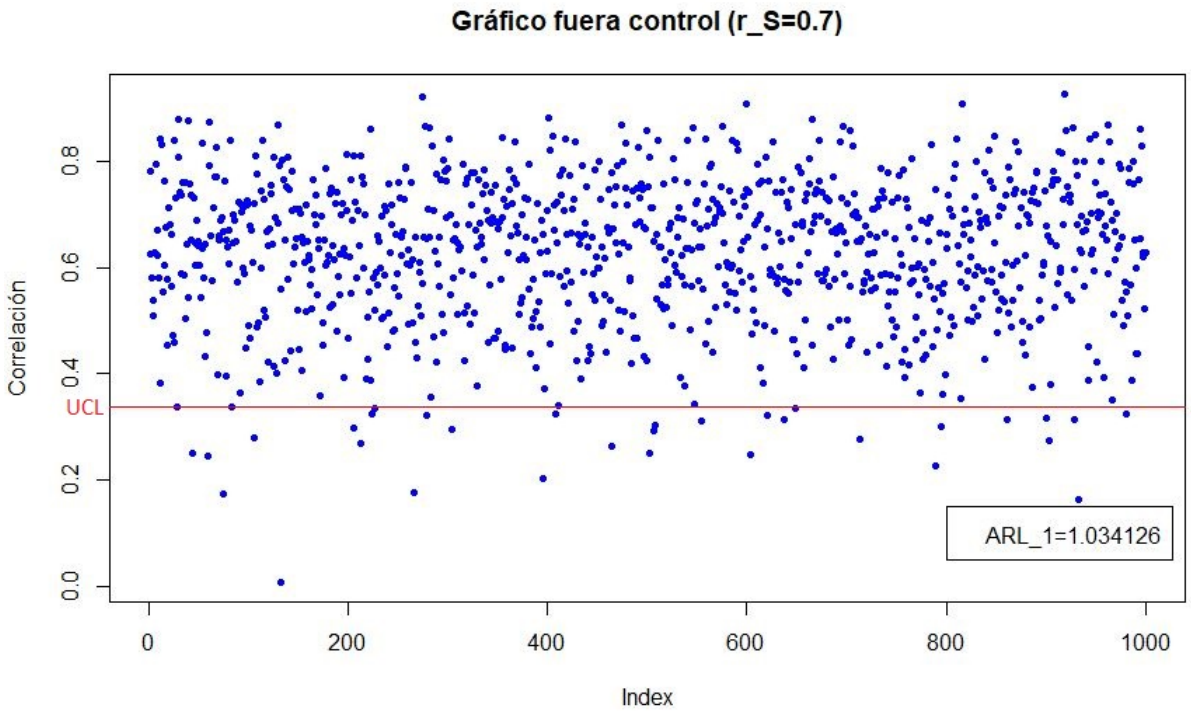


Figura 5.2: Gráfico fuera de control ($r_S = 0{,}7$).

En este caso, el gráfico de control logra detectar a la perfección la existencia de correlación entre las variables. El gráfico con tan solo 1 prueba detecta la falta de control ($ARL_1 = 1{,}03$).

Capítulo 6

Conclusiones y trabajo futuro

En este capítulo presentaremos las conclusiones obtenidas en la realización de este estudio y en qué ámbitos se podría seguir investigando.

6.1. Conclusiones

En cuanto al presente trabajo, se han cumplido los objetivos propuestos. Se ha expuesto de manera sucinta la teoría de interés para el trabajo, así como la relación entre el coeficiente de correlación de Spearman y Pearson, el control estadístico de la calidad y, en particular, los gráficos de control y la forma de medir su desempeño.

Se ha propuesto una metodología eficaz para calcular el límite de control superior de un gráfico de control para la detección de correlación entre dos variables en función del tamaño muestral y la probabilidad de error de primera especie. Además, se ha logrado un procedimiento mediante simulación para obtener el desempeño del gráfico de control tanto bajo control como fuera de control (existencia de correlación) haciendo uso de la teoría de las cópulas.

Esta metodología se ha llevado a cabo para diferentes tamaños muestrales, correlaciones esperadas, distribuciones de X e Y, desviaciones típicas de X e Y y tipos de cópulas resumido en la Sección 3.3. Respecto a los resultados numéricos, se ha proporcionado una tabla numérica con los valores del límite de control superior y el desempeño del gráfico de control tanto bajo control como fuera de control para la totalidad de combinaciones posibles entre los factores con el fin de facilitar una herramienta informativa.

Se ha estudiado el comportamiento del ARL_1, observado que los factores que muestran una mayor influencia en el cálculo son el tamaño muestral y la correlación esperada, r_S, disminuyendo considerablemente con el aumento de estos factores. Para el resto de factores no se han mostrado diferencias significativas en la media del ARL_1.

Finalmente, a partir de la teoría y metodología desarrollada, se ha construido un gráfico de control no paramétrico donde se ha calculado el límite de control superior y el desempeño del gráfico de control tanto bajo control ($r_S = 0$) como fuera de control ($r_S = 0{,}7$), resultando ser muy eficaz a la hora de detectar la correlación existente.

6.2. Trabajo futuro

A partir de la investigación realizada en este trabajo, se podría ampliar en las siguientes direcciones:

- La primera sería completar el estudio tomando correlaciones negativas, puesto que, como las distribuciones marginales que estamos manejando no tienen por qué ser simétricas, en este momento no sabemos exactamente qué habría pasado en situaciones fuera de control en las que la θ de la cópula (o, equivalentemente, la r_S) fuesen negativas.

- La segunda sería ampliar las opciones del estudio explorando diferentes distribuciones de X e Y con diferentes medias y desviaciones típicas y desarrollar nuevas cópulas para determinar si alguna plantea mejoras en el cálculo del desempeño del gráfico.

- La última sería investigar en la construcción de gráficos de control con memoria, así como EWMA o CUSUM [22]. El beneficio esperado sería la obtención de gráficos más eficientes, es decir, con ARL_1 más pequeño que el gráfico de Shewhart que hemos desarrollado.

Apéndice A

Algoritmos

A.1. Límite de control

El algoritmo utilizado para calcular los límites de control genéricos empleando el software R es el siguiente: `https://n9.cl/offat`

A.2. Experimentación

El algoritmo utilizado para realizar la experimentación empleando el software R es el siguiente: `https://n9.cl/wb6en`

A.3. Código completo

El algoritmo utilizado para realizar el trabajo completo en el software R es el siguiente: `https://n9.cl/8klob`

A.4. Gráficas CDF cópulas

El código utilizado para realizar los gráficos de las funciones de densidad para las diferentes cópulas en el software *Mathematica* es el siguiente: `https://n9.cl/7bo5t`

Apéndice B

Tablas y gráficos

B.1. Tabla completa de ARL_1

La tabla completa con los valores obtenidos del ARL_1 para cada combinación de los factores es la siguiente: `https://n9.cl/wz4ck`

B.2. Gráficos de las varianzas

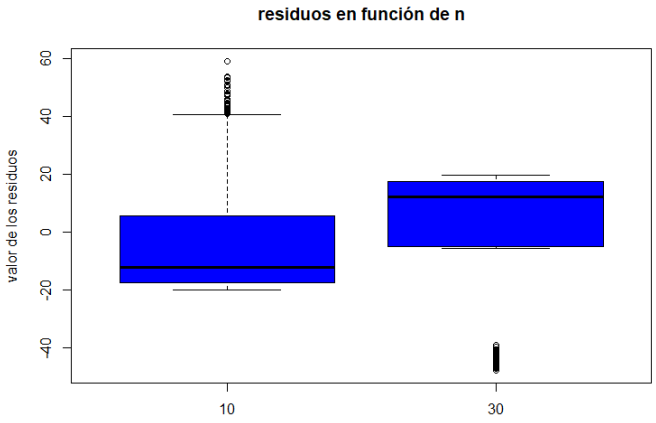


Figura B.1: Varianza de los residuos en función del tamaño muestral.

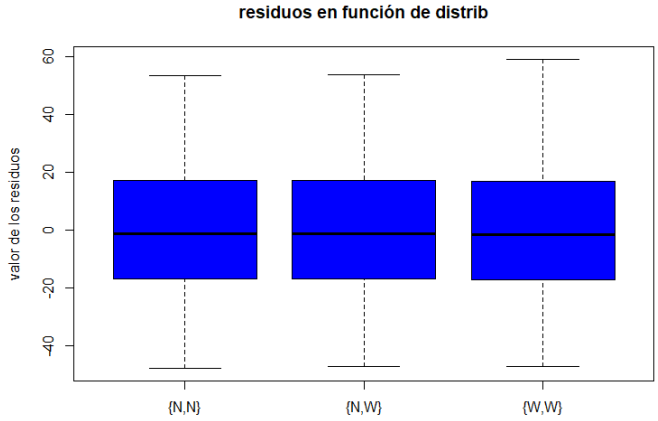


Figura B.2: Varianza de los residuos en función de la distribución de X e Y.

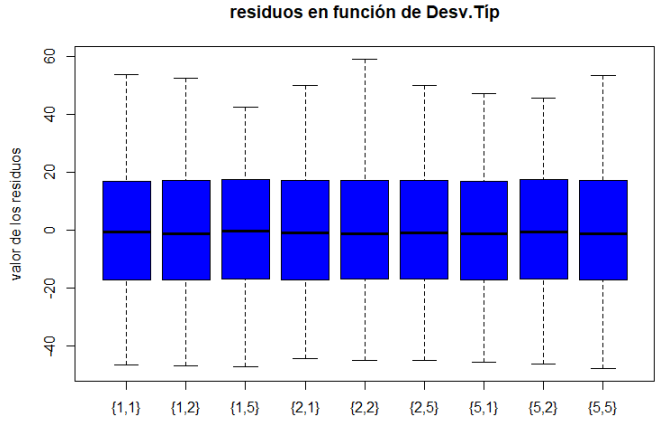


Figura B.3: Varianza de los residuos en función de la desviación típica.

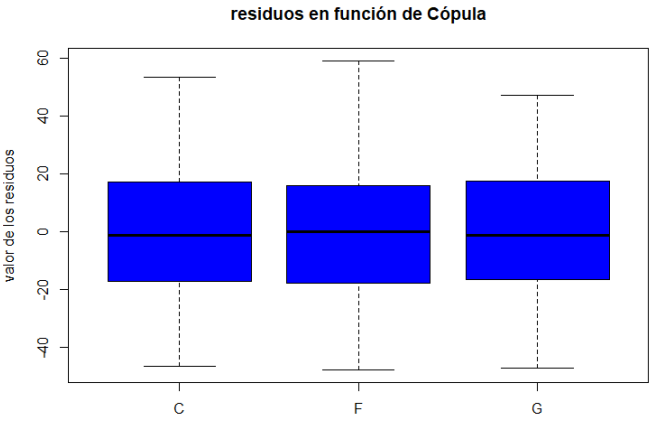


Figura B.4: Varianza de los residuos en función de la cópula.

B.3. Tabla de medias

Table of Least Squares Means for ARL1 with 95,0% Confidence Intervals

			Stnd.	Lower	Upper
Level	*Count*	*Mean*	*Error*	*Limit*	*Limit*
GRAND MEAN	630	38,4925			
n					
10	315	57,9441	1,50682	54,9908	60,8975
30	315	19,0408	1,50682	16,0874	21,9941
r_s					
{0,1}	126	145,019	2,22168	140,664	149,373
{0,3}	126	32,1808	2,22168	27,8264	36,5353
{0,5}	126	9,75513	2,22168	5,40071	14,1096
{0,7}	126	3,6817	2,22168	-0,672719	8,03613
{0,9}	126	1,82596	2,22168	-2,52846	6,18039
Distrib					
{N,N}	180	38,5089	2,0362	34,518	42,4997
{N,W}	270	38,3676	1,43981	35,5456	41,1896
{W,W}	180	38,6009	2,0362	34,61	42,5918
Desv.Tip					
{1,1}	90	38,5949	2,49383	33,707	43,4827
{1,2}	90	38,5481	2,49383	33,6603	43,436
{1,5}	90	38,1476	2,49383	33,2598	43,0354
{2,1}	30	38,6683	4,55309	29,7444	47,5922
{2,2}	90	38,4741	2,49383	33,5863	43,362
{2,5}	90	38,3403	2,49383	33,4525	43,2282
{5,1}	30	38,7375	4,55309	29,8135	47,6614
{5,2}	30	38,3644	4,55309	29,4405	47,2883
{5,5}	90	38,5568	2,49383	33,669	43,4446
Copula					
C	210	38,1886	1,77734	34,7051	41,6721
F	210	39,4925	1,77734	36,009	42,9761
G	210	37,7962	1,77734	34,3127	41,2798

Figura B.5: Tabla de medias del ARL_1 de cada factor.

Bibliografía

[1] James A. Alloway Jr. and M. Raghavachari. Control chart based on the hodges-lehmann estimator. *Journal of Quality Technology*, 23(4):336–347, 1991.

[2] Steven Yourstone and William Zimmer. Non-normality and the design of control charts for averages*. *Decision Sciences*, 23:1099 – 1113, 06 1992.

[3] Hernández-Lalinde J., C. F. Espinoza, J. E. Rodríguez, R. J. G. Chacón, S. C. A. Toloza, T. M. K. Arenas, S. S. M. Carrillo, and P. V. J. Bermúdez. Sobre el uso adecuado del coeficiente de correlación de Pearson. *Archivos Venezolanos de Farmacología y Terapéutica*, 37(5):586–601, 2018.

[4] Helena Chmura Kraemer. The non-null distribution of the spearman rank correlation coefficient. *Journal of the American Statistical Association*, 69(345):114–117, 1974.

[5] Eric Chicken. Myles Hollander, Douglas A.Wolfe. *Nonparametric Statistical Methods*, volume 1999. Wiley, 2014.

[6] Luis Restrepo and Julián Gónzales. From Pearson to Spearman. *Revista Colombiana de Ciencias Pecuarias*, 20(2):183–192, 2007.

[7] A Badii, O P Guillen, Serrato Lugo, and J J Aguilar Garnica. Correlación No-Paramétrica y su Aplicación en la Investigaciones Científica Non-Parametric Correlation and Its Application in Scientific Research. *International Journal of Good Conscience Agosto*, 9(2):31–40, 2014.

[8] Alberto Rodríguez and Jesús Rodríguez. Control estadístico de la calidad de un servicio mediante Gráficas X y R. *Politica y Cultura*, pages 151–169, 2009.

[9] W.A. Shewhart. *Economic Control of Quality of Manufactured Product*. PhD thesis, 1931.

[10] Elodia Vives Besalduch. CONTROL DE CALIDAD: control estadístico de procesos. Technical report.

[11] Douglas C. Montgomery. *Introduction to Statistical Quality Control*, volume 7. Wiley, 2012.

[12] E. Navarrete. *Control estadístico de la calidad control estadístico de procesos*. Ed. Adhara, 1998.

[13] David J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. 2011.

[14] M. A. Van De Wiel and A. Di Bucchianico. Fast computation of the exact null distribution of Spearman's ρ and Page's L statistic for samples with and without ties. *Journal of Statistical Planning and Inference*, 92(1-2):133–145, 2001.

[15] M.J. Frank. On the simultaneous associativity of f(x, y) and x + y - f(x, y). *Aequationes mathematicae*, 19:194–226, 1979.

[16] Phillipe Castagliola. Procedure to Generate Uniform Random Variates from Each Copula. *Options*, 4:1–3, 2006.

[17] D. G. CLAYTON. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 04 1978.

[18] E.J. GUMBEL. Les valeurs extrêmes des distributions statistiques. 2:115–158, 1935.

[19] Dorra Rahali, Philippe Castagliola, Hassen Taleb, and Michael Boon Chong Khoo. Evaluation of Shewhart time-between-events-and-amplitude control charts for correlated data. *Quality and Reliability Engineering International*, 37(1):219–241, 2021.

[20] Mike Preuss Thomas Bartz-Beielstein, Marco Chiarandini, Luis Paquete. *Experimental methods for the analysis of Optimization Algorithms*, volume 1999. Springer, 2010.

[21] S. Chakraborti, P. Van Der Laan, and S. T. Bakir. Nonparametric control charts: An overview and some results. *Journal of Quality Technology*, 33(3):304–315, 2001.

[22] Gemai Chen, Smiley W. Cheng, and Hansheng Xie. Monitoring process mean and variability with one EWMA chart. *Journal of Quality Technology*, 33(2):223–233, 2001.