



UNIVERSITAT  
POLITÈCNICA  
DE VALÈNCIA



UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Estadística e Investigación Operativa  
Aplicadas y Calidad

Gráficos de control no paramétricos para la correlación  
basados en el estadístico de Kendall: diseño e  
implementación

Trabajo Fin de Máster

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

AUTOR/A: Ibañez Sirvent, Neus

Tutor/a: Giner Bosch, Vicent

Cotutor/a externo: CASTAGLIOLA, PHILIPPE

CURSO ACADÉMICO: 2021/2022

## Resumen

Este trabajo tiene como objetivos el diseño, implementación y análisis de gráficos de control para la correlación entre dos variables continuas, no necesariamente normales, basados en la monitorización del estadístico ‘tau’ de Kendall. La naturaleza de este estadístico, que transforma las observaciones en rangos, permite catalogar este gráfico como no paramétrico o independiente de las distribuciones subyacentes.

Se realiza el diseño de un gráfico de control tipo Shewhart, basado en el coeficiente de correlación de Kendall, a partir del estudio de su distribución en el muestreo, y modelando la relación entre las dos variables continuas implicadas mediante cópulas.

La implementación del gráfico diseñado se lleva a cabo en lenguaje R. Se diseñan y ejecutan experiencias computacionales con el fin de medir la aptitud de la técnica desarrollada para detectar diferentes grados de asociación entre las dos variables de interés.

Por último, se ilustra la utilidad de este gráfico mediante un ejemplo numérico.

Mediante el análisis y estudio de las experiencias numéricas obtenidas se deduce la influencia del tamaño muestral, las cópulas empleadas y el coeficiente de correlación a detectar en la medida de desempeño del gráfico de control. Además de la influencia de estos factores, también se observa una interacción entre la cópula utilizada y el coeficiente de correlación a detectar, y así como entre el tamaño muestral y el coeficiente de correlación a detectar.

**Palabras clave :** gráficos de control; métodos no paramétricos; correlación de Kendall; estadísticos de rangos; simulación Montecarlo; cópulas

## Resum

Aquest treball té com objectius el diseny, implementació i anàlisi de gràfics de control per a la correlació entre dues variables contínues, no necessàriament normals, basats en la monitorització del estadístic ‘tau’ de Kendall. La naturalesa d’aquest estadístic, que transforma les observacions en rangs, permet catalogar aquest gràfic com no paramètric o independent de les distribucions subyacents.

Es realitza el diseny d’un gràfic de control de tipus Shewhart, basat en el coeficient de correlació de Kendall, a partir de l’estudi de la seua distribució en el mostreig, i modelant la relació entre les dues variables contínues implicades mitjançant còpules.

La implementació del gràfic disenyat es duu a terme amb el llenguatge R. Es dissenya i executen experiències computacionals amb la finalitat de mesurar la capacitat de la tècnica desenvolupada per a detectar diferents graus d’associació entre les dues variables d’interès.

Per últim, s’il·lustra la utilitat d’aquest gràfic mitjançant un exemple numèric.

Mitjançant l’anàlisi i l’estudi de les experiències numèriques obtingudes es dedueix la influència del tamany mostral, les còpules emprades i el coeficient de correlació a detectar en la mesura d’acompliment del gràfic de control. A més de la influència d’aquests factors, també s’observa una interacció entre la còpula utilitzada i el coeficient de correlació a detectar, així com entre la mida mostral i el coeficient de correlació a detectar.

**Paraules clau:** gràfics de control; mètodes no paramètrics; correlació de Kendall; estadístics de rangs; simulació Montecarlo; còpules

## Abstract

The objectives of this work are the design, implementation and analysis of control charts for the correlation between two continuous variables, not necessarily normal, based on the monitoring of Kendall's 'tau' statistic. The nature of this statistic, which transforms observations into ranges, allows us to classify this chart as nonparametric or independent of the underlying distributions.

The design of a Shewhart-type control chart based on Kendall's correlation coefficient is based on the study of its sampling distribution and modeling the relationship between the two continuous variables involved by means of copulas.

The implementation of the designed chart is carried out in R language. Computational experiments are designed and executed in order to measure the aptitude of the developed technique to detect different degrees of association between the two variables of interest.

Finally, the usefulness of this chart is illustrated by means of a numerical example.

Through the analysis and study of the numerical experiences obtained, the influence of the sample size, the copulas used and the correlation coefficient to be detected on the performance measurement of the control chart is deduced. In addition to the influence of these factors, an interaction is also observed between the copula used and the correlation coefficient to be detected, as well as between the sample size and the correlation coefficient to be detected.

**Keywords :** control charts; nonparametric methods; Kendall's correlation; rank statistics; Monte Carlo simulation; copulas;

# Índice general

<b>1. Introducción</b>	<b>6</b>
1.1. Motivación y alcance . . . . .	6
1.2. Objetivos . . . . .	7
1.3. Estructura del trabajo . . . . .	7
<b>2. Marco teórico</b>	<b>8</b>
2.1. Gráficos de Control . . . . .	8
2.1.1. Diseño gráfico de control . . . . .	8
2.1.2. Medidas de desempeño de los gráficos de control . . . . .	10
2.2. Gráficos de Control no paramétricos . . . . .	11
2.3. Coeficiente de correlación de Pearson . . . . .	11
2.4. Coeficiente de correlación de Kendall . . . . .	12
2.4.1. Coeficiente poblacional . . . . .	12
2.4.2. Coeficiente muestral . . . . .	13
2.4.3. Ventajas del coeficiente correlación de Kendall frente al coeficiente de Pearson . . . . .	16
2.5. Cópulas . . . . .	17
2.5.1. Cálculo de $\tau$ mediante cópulas . . . . .	18
2.6. Simulación Montecarlo . . . . .	20
<b>3. Desarrollo de un gráfico de control no paramétrico para la correlación basado en el estadístico de Kendall</b>	<b>21</b>
3.1. Diseño del gráfico. Obtención de los límites de control . . . . .	21
3.2. Obtención medidas de desempeño del gráfico . . . . .	22
3.2.1. Proceso bajo control . . . . .	22
3.2.2. Proceso fuera de control . . . . .	22
3.3. Diseño de experimentos . . . . .	24
<b>4. Experiencias computacionales</b>	<b>25</b>
4.1. Implementación . . . . .	25
4.1.1. Software y hardware empleado . . . . .	25
4.1.2. Implementación en R . . . . .	25
4.2. Análisis a realizar . . . . .	30
4.3. Resultados y discusión . . . . .	31
4.3.1. Resultados de la experiencia numérica . . . . .	31
4.3.2. Análisis de los resultados obtenidos . . . . .	35
<b>5. Ejemplo práctico</b>	<b>47</b>
5.1. Proceso bajo control . . . . .	47

5.2. Proceso fuera de control . . . . .	49
5.2.1. Coeficiente de correlación a detectar positivo . . . . .	49
5.2.2. Coeficiente de correlación a detectar negativo . . . . .	50
<b>6. Conclusiones y trabajo futuro</b>	<b>51</b>
6.1. Trabajo futuro . . . . .	52
<b>A. Anexos</b>	<b>55</b>
A.1. Código R . . . . .	55
A.2. Archivo Excel con los resultados numéricos . . . . .	55
A.3. Código Mathematica . . . . .	55

# 1. Introducción

## 1.1. Motivación y alcance

El análisis de correlación es una de las metodologías estadísticas descriptivas más utilizadas en la mayoría de las investigaciones aplicadas, cuyo objetivo consiste en estimar el nivel de asociación entre las variables objeto de estudio; es decir, pretende estimar el grado de dependencia entre ellas. Es común analizar si las variables están correlacionadas y, en caso de estarlo, se procede a generar el modelo de regresión con el objetivo de estudiar su relación.

En este contexto, los coeficientes más utilizados para determinar la correlación entre variables son: el de Pearson, Spearman y Kendall. Este trabajo se centra en estudiar la posible correlación entre variables no necesariamente normales. Por ello, el coeficiente de correlación objeto de estudio será el de Kendall, que, junto con el de Spearman, es más adecuado que el de Pearson ante la posible falta de normalidad. En concreto, el coeficiente de Kendall se utiliza para estudiar la relación entre variables cualitativas de tipo ordinal, o bien variables continuas cuyos valores observados son reemplazados por valores discretos ordinales.

Además de, por supuesto, en la ingeniería y en la industria, uno de los campos en los que más se suele utilizar este tipo de coeficiente de correlación es en el de la psiquiatría. Un ejemplo se puede apreciar en el artículo [1], en el cual se estudia el efecto del medio a los suicidios que ocurren en España mediante el coeficiente de correlación de Kendall. La aplicación de este coeficiente de correlación es muy amplia, siendo utilizado de manera práctica en cualquier campo de las ciencias donde se realicen mediciones y se quiera observar las variaciones cruzadas entre individuos. En el ámbito educativo, es destacable la aplicación del coeficiente de Kendall en el trabajo [2].

Como se acaba de destacar, monitorizar la correlación entre las diferentes variables implicadas en un proceso industrial puede ser de interés para asegurar la calidad del mismo. Por ello, resulta natural plantearse el uso de gráficos de control para realizar un seguimiento a lo largo del tiempo de dicha correlación.

Hasta la fecha, no hay publicados trabajos en los que se aborde la monitorización de la correlación entre dos características de calidad continuas, no necesariamente normales, mediante el coeficiente de correlación de Kendall.

Tal y como se ha mencionado previamente, el coeficiente de Kendall se utiliza para estudiar la relación entre variables cualitativas de tipo ordinal, o bien variables continuas cuyos valores observados son reemplazados por valores discretos ordinales. En el caso de calcular el coeficiente de Kendall para variables que ya son ordinales, se puede dar el caso de empates,

es decir, el caso  $(x_i - x_j)(y_i - y_j) = 0$ . Sin embargo, cuando se calcula a partir de variables  $X$  e  $Y$  continuas, el caso de empates es imposible teóricamente. Este estudio se centra en el caso de variables  $X$  e  $Y$  continuas, y, por tanto, no se va a tener en cuenta el caso de empates.

## 1.2. Objetivos

El objetivo principal de este trabajo consiste en diseñar e implementar un gráfico de control para el coeficiente de correlación de Kendall de dos variables continuas no necesariamente normales, así como evaluar su eficiencia.

Este objetivo principal se puede descomponer en los siguientes objetivos específicos:

- Diseño e implementación del gráfico.
- Validación y evaluación del gráfico a partir de experiencias numéricas.

## 1.3. Estructura del trabajo

El resto del trabajo se estructura como sigue. En el **capítulo 2**, se detallan los conceptos teóricos y notaciones básicas que aparecerán a lo largo del trabajo. Además, se presentan y justifican las herramientas de las que se hará uso en el desarrollo de esta propuesta (gráficos de control no paramétricos y coeficiente de correlación de Kendall).

Posteriormente, en el **capítulo 3** se presenta la metodología desarrollada para monitorizar el coeficiente de correlación de Kendall y se detallan los pasos a seguir para diseñar e implementar este gráfico de control.

En el **capítulo 4** se muestra la implementación en R del procedimiento expuesto en el capítulo anterior, así como los resultados obtenidos al aplicar el gráfico propuesto a una batería de ejemplos, junto con un análisis de estos resultados.

A continuación, el **capítulo 5** muestra un ejemplo práctico en el que se hace uso de la metodología propuesta.

Finalmente, en el **capítulo 6** se exponen las principales conclusiones que se obtienen a lo largo de este trabajo.

## 2. Marco teórico

En este capítulo se presentan los conceptos y notaciones básicas que se utilizan en este trabajo. Todos los conceptos expuestos y desarrollados en este capítulo se han extraído, principalmente, de [3], [4], [5] y [6].

### 2.1. Gráficos de Control

Tal y como se vio en la asignatura de Control Estadístico de la Calidad del presente máster, y como describe Montgomery (1997) [3], el principal objetivo de los gráficos de control es comprobar si el proceso está o no bajo control, es decir, ver si un proceso trabaja bajo la presencia de fuentes de variabilidad asignables o especiales.

#### 2.1.1. Diseño gráfico de control

Un gráfico de control típico presenta los siguientes elementos:

- Una línea central (CL), la cual representa el valor objetivo (o deseado) o el valor medio (o, en algunos casos, mediano) de la característica de calidad del proceso en caso de ausencia de fuentes de variabilidad asignables o especiales.
- Los límites de control superior e inferior (UCL y LCL), que representan el rango de valores donde se espera que se mueva, *casi siempre*, el valor del estadístico objeto de monitorización cuando el proceso está bajo control.

Un ejemplo de este tipo de gráficos se puede observar en la figura 2.1.

Si los puntos se localizan dentro de los límites de control, se supone que el proceso está bajo control y no es necesaria ninguna acción. Por otro lado, en caso de que un punto se localice fuera de estos límites de control, entonces significa que el proceso está fuera de control debido a ciertas causas asignables (o bien, se trata de una falsa alarma).

Por tanto, un gráfico de control puede ser visto como la aplicación repetida de un test de hipótesis a lo largo del tiempo y dibujado en un gráfico. Un punto que se localiza dentro de los límites de control es equivalente a aceptar que el proceso está bajo control, por tanto, se acepta la hipótesis nula. Y si se sale de estos límites de control, entonces se rechaza la hipótesis nula.

$$\begin{cases} H_0 : \text{Proceso bajo control} \\ H_1 : \text{Proceso fuera de control} \end{cases}$$

Se tienen los dos errores de los test de hipótesis:

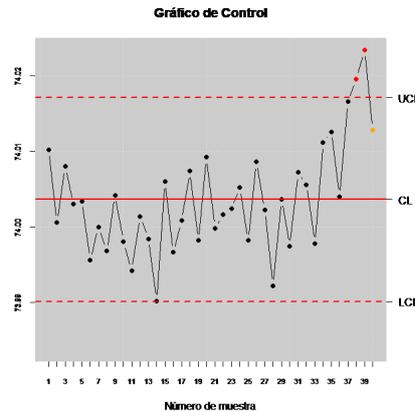


Figura 2.1.: Gráfico de Control.

- Error Tipo I:  $H_0$  es cierta, pero la rechazamos.
- Error Tipo II:  $H_0$  es falsa, pero la aceptamos.

Y, por tanto, se tiene  $\alpha = P(\text{Error Tipo I})$  y  $\beta = P(\text{Error Tipo II})$ .

La elección de los límites de control (LCL y UCL) es una de las decisiones críticas que deben tomarse al diseñar un gráfico de control, dado que si se toman los límites de control muy alejados de la línea central, se disminuye el riesgo de un error de tipo I, (en otras palabras, el riesgo de que un punto caiga más allá de los límites de control) indicando que un punto se encuentra fuera de control cuando no existe una causa asignable, mientras que, por otra parte, aumentará el riesgo de un error de tipo II, es decir, el riesgo de que un punto caiga entre los límites de control cuando el proceso está realmente fuera de control. Si se desplazan los límites de control más cerca de la línea central, se obtiene el efecto contrario: el riesgo de error de tipo I aumenta, mientras que el riesgo de error de tipo II disminuye.

Por consiguiente, se tiene que los límites de control superior e inferior que delimitan la región de aceptación de la hipótesis nula vienen determinados por la probabilidad de error de tipo I,  $\alpha$ . Por consiguiente, la probabilidad de caer fuera de estos límites es  $\alpha$  y los límites UCL y LCL satisfacen:

$$P(LCL \leq x \leq UCL) = 1 - \alpha$$

Equivalentemente, puede decirse que  $\alpha$  es la probabilidad de que un punto se salga de los límites cuando el proceso está controlado.

Algunas veces se hace uso de dos gráficos de control que se definen como unilaterales, es decir, con solo un límite de control (el superior o el inferior), en lugar de los dos límites en el mismo gráfico. Esta suele ser la elección lógica cuando la distribución en el muestreo del estadístico objeto de interés no es simétrica (por ejemplo, es el caso de la desviación típica muestral,  $s_{n-1}$ ) o, simplemente, solo es de interés detectar incrementos (o decrementos). De hecho, esto sucederá con la monitorización del coeficiente de correlación de Kendall, como se verá más adelante.

## 2. Marco teórico

### 2.1.2. Medidas de desempeño de los gráficos de control

En esta subsección se presentan las principales formas de medir la eficiencia, rendimiento o desempeño de un gráfico de control, en cuanto a su aptitud o potencia para detectar en poco tiempo situaciones de falta de control.

En este estudio se va a trabajar con un gráfico de control clásico o tipo Shewhart, lo cual presupone que cada muestra es independiente de las anteriores y presenta límites de control y tamaño muestral constantes a lo largo del tiempo. Bajo este supuesto, el cálculo de una de las principales medidas de desempeño, como es la longitud media de racha (o *Average Run Length*, ARL) es bastante sencillo.

El ARL representa la cantidad promedio de puntos o muestras que hay que graficar hasta detectar un punto fuera de los límites de control. Asumiendo que el proceso se encuentra bajo control estadístico, la fórmula es:

$$ARL_0 = \frac{1}{\alpha} = \frac{1}{\alpha_{sup} + \alpha_{inf}} \quad (2.1)$$

Siendo

- $\alpha_{sup}$ : La probabilidad de que un punto caiga fuera del límite de control superior (UCL) cuando el proceso está bajo control.
- $\alpha_{inf}$ : La probabilidad de que un punto caiga fuera del límite de control inferior (LCL) cuando el proceso está bajo control.

Este cálculo del  $ARL_0$ , se basa en que el punto que cae fuera de los límites es considerado como una falsa alarma.

El cálculo de  $ARL_0$  se basa en una distribución binomial, dado que  $\alpha_{inf}$  y  $\alpha_{sup}$  se pueden calcular mediante la utilización de una probabilidad binomial acumulada, tal y como se expone en las ecuaciones (2.2) y (2.3).

$$\alpha_{sup} = P(x > n \cdot UCL|p) = 1 - \sum_{x=0}^{nUCL} \binom{n}{x} p^x (1-p)^{n-x} \quad (2.2)$$

$$\alpha_{inf} = P(x > n \cdot LCL|p) = \sum_{x=0}^{nLCL} \binom{n}{x} p^x (1-p)^{n-x} \quad (2.3)$$

Siendo  $p$  la porción de elementos defectuosos que produce el proceso.

Siguiendo el mismo razonamiento, en el caso de que el proceso esté fuera de control, el ARL se puede calcular como:

$$ARL_1 = \frac{1}{1 - \beta} \quad (2.4)$$

Siendo  $\beta$  el riesgo de que un punto caiga dentro de los límites de control cuando el proceso está realmente fuera de control.

Este parámetro se utiliza para estudiar la efectividad o eficiencia del gráfico de control. Lo deseado sería un  $ARL_0$  lo más elevado posible para tener pocas falsas alarmas y un  $ARL_1$  lo más pequeño posible para poder detectar más rápido cuando el proceso está fuera de control.

Al igual que sucede en otras herramientas de inferencia estadística, una forma natural de elevar la potencia de un gráfico de control (o, equivalentemente, disminuir su  $ARL_1$ ) sin empeorar su  $ARL_0$  es incrementar el tamaño muestral. Sin embargo, esto no siempre está al alcance del usuario final, dado que, habitualmente, el tamaño muestral a utilizar está fijado de antemano.

## 2.2. Gráficos de Control no paramétricos

En la sección anterior, se ha visto qué es un gráfico de control y cómo, para su realización, se debe estudiar la distribución probabilística del proceso. Muchos de los gráficos más conocidos asumen la normalidad de la variable aleatoria continua subyacente. En la práctica, sin embargo, no siempre es posible verificar la hipótesis de normalidad, o directamente esta no se cumple. Si se diseña un gráfico asumiendo que las distribuciones son normales cuando no lo son, entonces el comportamiento del gráfico puede no ser el esperado (mayor  $ARL_1, \dots$ ).

En [7] se habla de las pruebas no paramétricas como pruebas libres de distribución o pruebas de suposición. En concreto, una prueba estadística o un gráfico de control puede considerarse como no paramétrico si, cuando el proceso se encuentra bajo control (equivalentemente, cuando se cumple  $H_0$ ), la distribución en el muestreo del estadístico objeto de monitorización no depende de la distribución de las variables aleatorias subyacentes [6]. En este tipo de pruebas, no es necesario el supuesto de que las distribuciones con las que se trabaja sean normales.

Según [7] se toma como regla general que las pruebas estadísticas inferenciales que evalúan los datos categóricos/nominales y datos ordinales/de rango se clasifican como pruebas no paramétricas, en el sentido indicado previamente, mientras que las pruebas que evalúan datos de intervalo o proporción se clasifican como pruebas paramétricas. Dado que en este trabajo se van a evaluar datos de rango mediante el estadístico de Kendall, los gráficos de control que construiremos pueden considerarse no paramétricos.

De hecho, es fácil demostrar (tal y como se verá más adelante) que la distribución en el muestreo de la ‘tau’ de Kendall no depende de la distribución de las variables aleatorias implicadas,  $X$  e  $Y$ , bajo la hipótesis de independencia [8], lo cual justifica el enfoque de este trabajo.

## 2.3. Coeficiente de correlación de Pearson

En este trabajo se va a tratar con el coeficiente de correlación de Kendall, aunque es necesario tener claro previamente qué es un coeficiente de correlación. Habitualmente, cuando se habla de coeficiente de correlación se está hablando del coeficiente de correlación de Pearson, este coeficiente se suele representar por la letra  $\rho$  y es una medida de dependencia entre dos variables aleatorias cuantitativas.[9]

Este coeficiente se define como:

$$\rho_{ZW} = \frac{Cov(Z,W)}{\sigma_Z \sigma_W}$$

donde  $\sigma_Z$  y  $\sigma_W$  son las desviaciones típicas de  $Z$  y  $W$  (siendo  $Z$  y  $W$  variables aleatorias cuantitativas), y  $Cov(Z, W)$  es la covarianza de estas dos variables, la cual también es una

## 2. Marco teórico

medición de la dependencia de estas variables, aunque no es fiable dado que su valor depende de la escala de medición de cada variable. En cambio, el coeficiente de correlación de Pearson es una especie de estandarización de la covarianza.

Este coeficiente de correlación satisface la desigualdad  $-1 \leq \rho \leq 1$ . Se considera que dos variables son dependientes (linealmente) cuando  $|\rho| \approx 1$ . De hecho, el signo de este coeficiente de correlación es igual al signo de la covarianza. Por lo tanto, tenemos estos casos:

- Si  $\rho_{ZW} > 0$  esto indica que  $W$  aumenta a medida que  $Z$  aumenta. En el caso particular de  $\rho_{ZW} = 1$  esto implica una correlación perfecta, con todos los puntos cayendo en una recta con pendiente positiva.
- Si  $\rho_{ZW} < 0$  implica una disminución en  $W$  cuando  $Z$  aumenta. En el caso particular de  $\rho_{ZW} = -1$  también implica una correlación perfecta pero todos los puntos cayendo en una recta con pendiente negativa.
- Si  $\rho_{ZW} = 0$  implica que no hay correlación lineal entre estas dos variables.

Además, tenemos que si  $Z$  y  $W$  son variables aleatorias independientes, entonces  $\rho_{ZW} = 0$ . Por tanto, tenemos que las variables aleatorias independientes deben de ser no correlacionadas; sin embargo, el recíproco no es cierto.

## 2.4. Coeficiente de correlación de Kendall

### 2.4.1. Coeficiente poblacional

La  $\tau$  de Kendall mide el grado de asociación entre dos variables aleatorias continuas, que hemos denominado  $X$  e  $Y$ , con respecto al orden relativo de todos los posibles pares de elementos. A continuación, se explica el procedimiento mediante el que obtenemos este parámetro.

Una característica notable del coeficiente de Kendall es que reporta valores más bajos con respecto a los coeficientes de Spearman y Pearson en aquellas situaciones donde se analizan las asociaciones lineales con la misma intensidad (sin la presencia de valores atípicos). Este resultado no se puede traducir como que el coeficiente de Kendall es menos preciso que los otros dos, sino que tau se determina a partir de valores ranqueados, e inclusive reporta la misma tendencia en datos con distribuciones monótonas no lineales, en comparación con Spearman.

Se tienen  $n$  observaciones bivariantes  $(X_1, Y_1), \dots, (X_n, Y_n)$ , donde el conjunto de pares  $(X, Y)$  es un conjunto cualquiera de variables continuas.

Se dice que dos parejas  $(x_i, y_i)$  y  $(x_j, y_j)$  son **concordantes** si  $(x_i - x_j)(y_i - y_j) > 0$  y que son **discordantes** cuando  $(x_i - x_j)(y_i - y_j) < 0$ .

En este trabajo no se tiene en cuenta el caso de empates,  $((x_i - x_j)(y_i - y_j) = 0)$ , dado que vamos a suponer que  $X$  e  $Y$  son variables continuas.

Teniendo estos conceptos en cuenta, la  $\tau$  de Kendall es la diferencia entre la probabilidad de ser concordante y la probabilidad de ser discordante. Equivalentemente:

$$\tau = P[(X - X')(Y - Y') \geq 0] - P[(X - X')(Y - Y') \leq 0]$$

donde  $(X', Y')$  y  $(X, Y)$  son independientes.

### 2.4.2. Coeficiente muestral

Para definir la  $\tau$  muestral de Kendall ( $\hat{\tau}$ ) se deben definir previamente unos parámetros necesarios:

- Considerando una muestra de tamaño  $n$  y sabiendo que si hay un conjunto de  $n$  elementos distintos, el número de todos los posibles subconjuntos de dicho conjunto que tengan  $m$  elementos, sin importar el orden de dichos  $m$  elementos, viene dado por la siguiente ecuación:

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}$$

Dado que en este caso se pretende agrupar en parejas las  $n$  observaciones bivariantes, se tienen:

$$\binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{n(n-1)}{2} \text{ parejas}$$

- Ahora para las  $\frac{n(n-1)}{2}$  parejas se calcula un estadístico que proporciona un 1 si los pares son concordantes y un  $-1$  si son discordantes. Teniendo en cuenta que  $1 \leq i \leq j \leq n$  el estadístico es:

$$Q((X_i, Y_i), (X_j, Y_j)) = \begin{cases} 1 & \text{si } (X_i, Y_i)(X_j, Y_j) > 0 \\ -1 & \text{si } (X_i, Y_i)(X_j, Y_j) < 0 \end{cases}$$

- A partir del estadístico anterior se calcula el estadístico K de Kendall:

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^n Q((X_i, Y_i), (X_j, Y_j)) \quad (2.5)$$

Por tanto, teniendo en cuenta estos dos parámetros, la  $\tau$  muestral de Kendall, que se define mediante  $\bar{K}$ , se calcula siguiendo la ecuación (2.6).

$$\hat{\tau} = \frac{K}{\frac{n(n-1)}{2}} \quad (2.6)$$

De esta manera, se aprecia que  $\hat{\tau}$  es una forma de estandarizar el parámetro K. Además, cabe destacar que este estadístico muestral presenta una distribución discreta, es decir, solo es distinto de cero una cantidad finita de puntos.

Se puede notar que la fórmula para calcular  $\hat{\tau}$  que se muestra en la ecuación (2.6) es semejante a la media de los estadísticos  $Q((X_i, Y_i), (X_j, Y_j))$  de todos los pares de la muestra.

## 2. Marco teórico

Equivalentemente, tenemos que la  $\tau$  de Kendall muestral viene determinada por

$$\hat{\tau} = \frac{c - d}{c + d} = \frac{c - d}{n}$$

donde  $c$  es la cantidad de parejas concordantes y  $d$  es la cantidad de parejas discordantes de un conjunto de  $n$  observaciones del conjunto  $(X, Y)$ .

**Ejemplo 1** *Se tiene una muestra de 10 países, de los cuales se tiene información contenida en dos variables, el índice de desigualdad de género y el índice de desarrollo humano(HDI), ambos medidos en el año 2019, obtenidos de “Human Development Reports“. Por tanto, en este ejemplo, se va a medir la correlación mediante el coeficiente de correlación de Kendall muestral  $\hat{\tau}$  de estos dos índices a partir de los datos de la siguiente tabla.*

País	Gender Inequality Index	Human Development Index
Spain	0,070	0,904
Haiti	0,636	0,510
China	0,168	0,761
United Kingdom	0,118	0,932
Saudi Arabia	0,250	0,854
Cuba	0,304	0,783
Mozambique	0,523	0,456
Afghanistan	0,655	0,511
Greece	0,116	0,888
Poland	0,115	0,880

Se denotará por el parámetro  $Y$  a la variable índice de desarrollo humano(HDI) y por  $X$  a la variable índice de desigualdad de género.

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^n Q((X_i, Y_i), (X_j, Y_j)) = -32$$

$$\hat{\tau} = \frac{K}{\frac{n(n-1)}{2}} = -0,7111111$$

En consecuencia, dado que el valor obtenido de  $\hat{\tau}$  asociado a esta muestra es superior a 0,5 podemos intuir que estos dos índices no son independientes, por tanto, existe correlación entre ellos. Además, al ser  $\hat{\tau} < 0$  podemos intuir que estos dos índices están correlacionados inversamente, lo cual tiene sentido debido a la definición de los mismos.

La utilidad de la  $\tau$  va a ser para poder ver si las muestras  $X$  e  $Y$  están correlacionadas o son independientes. Veamos según  $\tau$  la posible relación entre estas dos muestras:

- Si  $\tau > 0$  entonces se tiene que  $X$  e  $Y$  están correlacionadas positivamente.
- Si  $\tau < 0$  entonces  $X$  e  $Y$  están correlacionadas negativamente.
- Si  $X$  e  $Y$  son independientes, entonces  $\tau = 0$  , teniendo en cuenta que el recíproco no siempre es cierto .

**Distribución de  $\hat{\tau}$** 

Como se ha mencionado previamente,  $\hat{\tau}$  es una estandarización del parámetro  $K$ , tal como se puede apreciar en la ecuación (2.6). Teniendo esto en cuenta, siguiendo los pasos indicados en [8], se va a estudiar la distribución de  $K$  bajo la hipótesis nula  $H_0$ . Sea  $R_i$  el rango  $X_i$  en la clasificación conjunta de  $X_1, \dots, X_n$  y que  $S_i$  es el rango de  $Y_i$  en la clasificación conjunta de  $Y_1, \dots, Y_n$ . Es evidente que para calcular  $K$  (2.5) basta con conocer las  $R$  y las  $S$ . Se utiliza este hecho para ilustrar cómo se puede obtener la distribución de  $K$  bajo el supuesto de  $H_0$ . Sin pérdida de generalidad, se toma  $R_1 = 1, \dots, R_n = n$ . Entonces, bajo la hipótesis nula, se tienen todas las configuraciones posibles de  $n!$   $(S_1, S_2, \dots, S_n)$  y son igualmente probables, lo que implica que cada una tiene una probabilidad  $(\frac{1}{n!})$ .

Para ilustrar las afirmaciones anteriores se considera el caso para  $n = 4$ . En el Cuadro 2.1 se muestran las  $4! = 24$  configuraciones posibles  $(S_1, S_2, S_3, S_4)$ , los valores asociados de  $K$ , y las probabilidades nulas correspondientes.

$(R_1, R_2, R_3, R_4)$	$(S_1, S_2, S_3, S_4)$	Null Probability	K
(1, 2, 3, 4)	(1, 2, 3, 4)	$\frac{1}{24}$	6
(1, 2, 3, 4)	(1, 2, 4, 3)	$\frac{1}{24}$	4
(1, 2, 3, 4)	(1, 3, 2, 4)	$\frac{1}{24}$	4
(1, 2, 3, 4)	(1, 3, 4, 2)	$\frac{1}{24}$	2
(1, 2, 3, 4)	(1, 4, 2, 3)	$\frac{1}{24}$	2
(1, 2, 3, 4)	(1, 4, 3, 2)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(2, 1, 3, 4)	$\frac{1}{24}$	4
(1, 2, 3, 4)	(2, 1, 4, 3)	$\frac{1}{24}$	2
(1, 2, 3, 4)	(2, 3, 1, 4)	$\frac{1}{24}$	2
(1, 2, 3, 4)	(2, 3, 4, 1)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(2, 4, 1, 3)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(2, 4, 3, 1)	$\frac{1}{24}$	-2
(1, 2, 3, 4)	(3, 1, 2, 4)	$\frac{1}{24}$	2
(1, 2, 3, 4)	(3, 1, 4, 2)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(3, 2, 1, 4)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(3, 2, 4, 1)	$\frac{1}{24}$	-2
(1, 2, 3, 4)	(3, 4, 1, 2)	$\frac{1}{24}$	-2
(1, 2, 3, 4)	(3, 4, 2, 1)	$\frac{1}{24}$	-4
(1, 2, 3, 4)	(4, 1, 2, 3)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(4, 1, 3, 2)	$\frac{1}{24}$	-2
(1, 2, 3, 4)	(4, 2, 1, 3)	$\frac{1}{24}$	-2
(1, 2, 3, 4)	(4, 2, 3, 1)	$\frac{1}{24}$	-4
(1, 2, 3, 4)	(4, 3, 1, 2)	$\frac{1}{24}$	-4
(1, 2, 3, 4)	(4, 3, 2, 1)	$\frac{1}{24}$	-6

Cuadro 2.1.: Configuraciones posibles  $(S_1, S_2, S_3, S_4)$ .

Así, por ejemplo, la probabilidad de que  $K = 2$  bajo la hipótesis nula  $H_0$  es  $\frac{5}{24}$ , dado que cuando se produce cualquiera de los cinco resultados  $(S_1, S_2, S_3, S_4) \in \{(1, 3, 4, 2), (1, 4, 2, 3),$

## 2. Marco teórico

$(2, 1, 4, 3), (2, 3, 1, 4), (3, 1, 2, 4)$  y cada uno de estos resultados tiene probabilidad nula  $\frac{1}{24}$ . Simplificando, obtenemos la distribución de  $K$  bajo la hipótesis nula.

Posible valor de $K$	Probabilidad bajo $H_0$
-6	$\frac{1}{24}$
-4	$\frac{3}{24}$
-2	$\frac{5}{24}$
0	$\frac{6}{24}$
2	$\frac{5}{24}$
4	$\frac{3}{24}$
6	$\frac{1}{24}$

Cuadro 2.2.: Distribución de  $K$  bajo  $H_0$ .

Mediante el Cuadro 2.2, se puede obtener, por ejemplo, la probabilidad de que  $K$  sea mayor o igual que 2 de la siguiente manera:

$$P_0(K \geq 2) = P_0(K = 2) + P_0(K = 4) + P_0(K = 6) = \frac{5}{24} + \frac{3}{24} + \frac{1}{24} = \frac{3}{8}$$

Nótese que se ha obtenido la distribución nula de  $K$  sin especificar la distribución de las poblaciones independientes subyacentes  $X$  e  $Y$ . Por ello, los procedimientos de prueba basados en  $K$  se denominan procedimientos libres de distribución. Por consiguiente, la distribución de muestreo de  $\tau$  es independiente de las distribuciones de  $X$  e  $Y$ .

Tal como se ha podido ver, la obtención de la distribución de  $K$  bajo control presentará una elevada complejidad cuando se den tamaños de muestra elevados, debido al alto número de combinaciones a realizar, por lo que se suele recurrir a paquetes de programación que ya tienen la función de distribución implementada, como por ejemplo el paquete "SuppDists" de RStudio [10], en el cual esta implementada esta función de distribución.

### 2.4.3. Ventajas del coeficiente correlación de Kendall frente al coeficiente de Pearson

El coeficiente de correlación de Pearson se basa en los siguientes supuestos:

1. La muestra de  $n$  sujetos para la que se calcula el valor  $\rho$  se selecciona aleatoriamente de la población que representa. Además, cuanto mayor sea  $n$ , más preciso será el coeficiente de correlación. Por consiguiente, cuando se trabaje con muestras pequeñas, será recomendable utilizar otro coeficiente de correlación de tipo no paramétrico (como el coeficiente de Correlación de Kendall).
2. El valor de este coeficiente de correlación es independiente de las unidades usadas para medir las variables.
3. No es un estadístico robusto, dado que es sensible a los valores atípicos de las observaciones. Cuando hay valores extremos en alguna de las variables, el valor del coeficiente de correlación puede verse alterado de forma importante.

4. El nivel de medición en el que se basa cada una de las variables es un dato de intervalo o tasa.
5. Se presupone normalidad en la distribución bivalente de las dos variables en análisis. En el caso de que esta condición no se cumpla, sería recomendable la utilización de otro coeficiente de correlación (el Coeficiente de Correlación de Kendall, por ejemplo).
6. El coeficiente de correlación mide solo la relación lineal, pero puede que las dos variables a estudiar tengan una relación diferente. Por ejemplo, al representarlas gráficamente se puede encontrar que tienen una relación curvilínea, y no en lineal. En este caso, el coeficiente de correlación de Pearson, que mide la relación lineal, daría un valor pequeño, a pesar de que en realidad las dos variables tienen asociación entre sí (solo que esta es de tipo no lineal).

Teniendo en cuenta que para utilizar el coeficiente de correlación de Pearson se deben satisfacer estos supuestos, mientras que el coeficiente de correlación de Kendall se ve poco afectado ante la presencia de un número pequeño de valores atípicos, no requiere del cumplimiento del supuesto de normalidad y, como ya se ha mencionado, se puede utilizar para cualquier tamaño de muestra. Además, el coeficiente de correlación de Kendall no solamente captura la relación lineal.

Por consiguiente, en el caso de no satisfacer alguno de estos supuestos, sería recomendable la utilización del coeficiente de correlación de Kendall.

## 2.5. Cópulas

En este trabajo se modelizará el comportamiento conjunto de dos o más variables continuas no independientes entre sí mediante cópulas (véase la sección 4.2).

Una cópula bidimensional es una función  $C : I^2 \rightarrow I = [0, 1]$  la cual satisface:

1.  $C(t, 0) = C(0, t) = 0$
2.  $C(t, 1) = C(1, t) = t \forall t \in I$
3.  $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0 \forall u_1, v_1, u_2, v_2 \in I$  tal que  $u_1 \leq u_2$  i  $v_1 \leq v_2$

Es decir, una cópula es la restricción a  $I^2$  de una distribución bivalente continua, la frontera de la cual es uniforme en  $I$ .

En particular, una cópula arquimediana es aquella que se puede expresar de la siguiente manera [5]:

$$C_\phi(x, y) = \phi^{-1}\{\phi(x) + \phi(y)\}, \quad 0 < x, y < 1\}$$

Para alguna función convexa decreciente  $\phi(t)$  tal que  $\phi(1) = 0$ . Además, por consenso se tiene que  $\phi^{-1}(t) = 0$  siempre que  $t \geq \phi(0)$ . Estas condiciones son necesarias y suficientes para que  $C_\phi(x, y)$  sea una función de distribución.

## 2. Marco teórico

**Proposición 1** Sean  $X$  e  $Y$  variables aleatorias uniformes cuya función de dependencia  $C(x,y)$  es de la forma  $\phi^{-1}\{\phi(x) + \phi(y)\}$  para alguna función convexa decreciente  $\phi$  definida en el intervalo  $(0, 1]$  satisfaciendo  $\phi(1) = 0$ . Sea  $U = \frac{\phi(X)}{\phi(X)+\phi(Y)}$ ,  $V = C(X, Y)$  y  $\lambda(v) = \frac{\phi(v)}{\phi'(v)}$  con  $0 < v \leq 1$ . Entonces:

1.  $U$  presenta una distribución uniforme en el intervalo  $(0, 1)$ .
2.  $V$  es distribuido mediante la siguiente función  $K(v) = v - \lambda(v)$  en el intervalo  $(0, 1)$ .
3.  $U$  y  $V$  son variables aleatorias independientes.

La existencia de una función  $\phi$  que satisface las tres propiedades anteriores implica que la siguiente igualdad se satisface  $C(x, y) = \phi^{-1}\{\phi(x) + \phi(y)\}$  en todo su dominio.

Las cópulas van a ser de utilidad para poder calcular  $\tau$  cuando se encuentre fuera de control. En la Tabla 2.3 se pueden observar distintos ejemplos de cópulas Arquimedianas bivariantes.

Cópula	$\phi$	range of $\theta$	$-\lambda$	$\tau$
Clayton	$\frac{t^{-\theta}-1}{\theta}$	$(0, \infty)$	$\frac{t(1-t^\theta)}{\theta}$	$\frac{\theta}{\theta+2}$
Frank	$\log\left(\frac{1-\exp(-\theta)}{1-\exp(-\theta t)}\right)$	$(-\infty, \infty)$	$\frac{1-\exp(-\theta t)}{\theta \exp(-\theta t)} \log\left(\frac{1-\exp(-\theta)}{1-\exp(-\theta t)}\right)$	$1 + \frac{4(D_1(\theta)-1)}{\theta}$
Gumbel	$(-\log(t))^{\theta+1}$	$[0, \infty)$	$-t \cdot \frac{\log(t)}{\theta+1}$	$\frac{\theta}{\theta+1}$
Bivariate Pareto	$t^{-1/\theta} - 1$	$(0, \infty)$	$\theta(t^{-1+\frac{1}{\alpha}} - t)$	$\frac{1}{2\theta+1}$

Cuadro 2.3.: Principales cópulas arquimedianas bivariantes y su relación con el coeficiente  $\tau$ .

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

### 2.5.1. Cálculo de $\tau$ mediante cópulas

Muchas medidas de dependencia son inalterables bajo transformaciones estrictamente crecientes de las variables aleatorias. Y como se tiene que la cópula  $C$  de un par de variables aleatorias  $X$  e  $Y$  también es invariante bajo transformaciones estrictamente crecientes de  $X$  e  $Y$ , varias medidas de dependencia invariantes en escala se pueden expresar a partir de las cópulas. Una de estas medidas es el coeficiente de correlación de Kendall denominado  $\tau$ , que se ha introducido en la sección 2.4.1, se puede expresar mediante cópulas de la forma que se muestra en la ecuación (2.7), siendo  $C$  una cópula Arquimediana generada por un cierta  $\phi$ .

$$\tau = 4 \cdot \int_0^1 \int_0^1 C(u, v)C(u, v) dudv - 1 = 4 \cdot E(C(U, V)) - 1 \quad (2.7)$$

Genest y Mackay (1986) demostraron una forma de calcular la Ecuación 2.7, mediante la función  $\phi$  a partir de la cual se genera la cópula. Esta igualdad viene presentada en la ecuación (2.8).

$$\tau = 4 \cdot E(C(U, V)) - 1 = 4 \cdot \int_0^1 \frac{\phi(t)}{\phi'(t)} dt + 1 \quad (2.8)$$

**Ejemplo 2** (Cópula de Clayton)

La cópula de Clayton viene determinada por la función  $\phi(t) = \frac{t^{-\theta}-1}{\theta}$  y se tiene  $\phi'(t) = \frac{1}{\theta} \cdot (-\theta) \cdot t^{-\theta-1} = -t^{-\theta-1}$  y sustituyendo en la ecuación (2.8) se obtiene el resultado dado en la ecuación (2.9).

$$\tau = 4 \cdot \int_0^1 \frac{\phi(t)}{\phi'(t)} dt = 4 \cdot \int_0^1 \frac{(t^{-\theta} - 1)/\theta}{-t^{-\theta-1}} dt = 4 \cdot \left( \frac{-2}{\theta + 2} \right) + 1 = \frac{\theta}{\theta + 2} \quad (2.9)$$

Y de forma análoga, se calcula para el resto de cópulas Arquimedianas obteniendo las ecuaciones de  $\tau$  expuestas en la Tabla 2.3.

El parámetro  $\theta$  que aparece en las cópulas es una medida de relación entre las variables  $U$  y  $V$ , este parámetro también va a ser de utilidad para generar variables aleatorias tal como se va a poder ver más adelante. Entonces, a partir de las fórmulas que se muestran en la última columna de la Tabla 2.3 para  $\tau$  se procede a calcular el parámetro  $\theta$ :

- **Cópula de Frank.** En el caso de la cópula de Frank se tiene que la relación entre el parámetro  $\theta$  y  $\tau$  viene determinada por la ecuación (2.10).

$$\tau = 1 + \frac{4(D_1(\theta) - 1)}{\theta} \quad \text{tal que} \quad D_1(\theta) = \frac{1}{\theta} \int_0^\theta \frac{t}{e^t - 1} dt \quad (2.10)$$

Por tanto, para poder calcular  $\theta$  se van a realizar mediante la utilización de métodos numéricos.

- **Cópula de Clayton.** En el caso de la cópula de Clayton se tiene que la relación entre los parámetros  $\theta$  y  $\tau$  viene determinada por la ecuación (2.11).

$$\tau = \frac{\theta}{\theta + 2} \implies \theta = \frac{2\tau}{1 - \tau} \quad (2.11)$$

- **Cópula de Gumbel.** En el caso de esta cópula se tiene que la relación entre los parámetros  $\theta$  y  $\tau$  viene expresada mediante la ecuación (2.12).

$$\tau = \frac{\theta}{\theta + 1} \implies \theta = \frac{1}{1 - \tau} \quad (2.12)$$

Ahora, a partir de estos valores de  $\theta$  que se han obtenido, se van a generar valores aleatorios de cada una de las cópulas arquimedianas con las que estamos trabajando.

- **Cópula de Frank**

1. El primer paso consistirá en generar dos variables aleatorias uniformes e independientes  $(u_1, v_2)$ .
2. Ahora, a partir de las variables aleatorias generadas, se calcula  $u_2$  mediante la ecuación (2.13).

$$u_2 = -\frac{1}{\theta} \left( 1 + \frac{v_2(1 - e^{-\theta})}{v_2(e^{-\theta u_1} - 1) - e^{\theta u_1}} \right) \quad (2.13)$$

## 2. Marco teórico

### ▪ Cópula de Clayton

1. El primer paso va a consistir en generar dos variables aleatorias uniformes e independientes  $(u_1, v_2)$ .
2. Ahora a partir de las variables aleatorias generadas se calcula

$$u_2 = [u_1^{-\theta}(v_2^{-\theta/(1-\theta)} - 1) + 1]^{-1/\theta} \quad (2.14)$$

### ▪ Cópula de Gumbel

1. El primer paso consiste en generar dos variables aleatorias uniformes e independientes  $(v_1, v_2)$ .
2.  $K_c(w) = w(1 - \frac{\ln(w)}{\theta}) = v_2$  y teniendo en cuenta que  $0 < w < 1$ , obtenemos un valor de  $w$  mediante el método de la sección Áurea ("Golden Section Search"). Ahora, a partir de estas variables, se obtiene el par  $(u_1, u_2)$  de la siguiente manera:

$$u_1 = \exp(v_1^{1/\theta} \ln(w)) \quad (2.15)$$

$$u_2 = \exp((1 - v_1)^{1/\theta} \ln(w)) \quad (2.16)$$

## 2.6. Simulación Montecarlo

El método de simulación de Montecarlo (MC) está basado en la generación de números aleatorios. El procedimiento es simular, mediante modelos matemáticos, el comportamiento aleatorio de sistemas reales no dinámicos.

La clave de la simulación MC consiste en crear un modelo matemático del sistema, proceso o actividad que se quiere analizar, identificando aquellas variables (inputs del modelo) cuyo comportamiento aleatorio determina el comportamiento global del sistema. Una vez identificados dichos inputs o variables aleatorias, se lleva a cabo un experimento consistente en generar muestras aleatorias (valores concretos) para dichos inputs y, a continuación, analizar el comportamiento del sistema ante los valores generados. Tras repetir  $n$  veces este experimento, se obtendrán  $n$  observaciones sobre el comportamiento del sistema, lo cual será de utilidad para entender el funcionamiento del mismo.

### 3. Desarrollo de un gráfico de control no paramétrico para la correlación basado en el estadístico de Kendall

#### 3.1. Diseño del gráfico. Obtención de los límites de control

Ahora se unirá lo visto en los capítulos anteriores, dado que lo que se pretende es realizar un gráfico de control suponiendo que la situación bajo control es que  $X$  e  $Y$  sean independientes y fuera de control que sean dependientes. Por tanto, el test de hipótesis asociado al gráfico que se diseñará es:

$$\begin{cases} H_0 : X \text{ e } Y \text{ independientes} \\ H_1 : X \text{ e } Y \text{ dependientes} \end{cases}$$

Dado que  $\tau$  de Kendall es una medida de dependencia no paramétrica, es este el parámetro a graficar y, por tanto, el test de hipótesis será el siguiente:

$$\begin{cases} H_0 : \tau = 0 \\ H_1 : \tau \neq 0 \end{cases}$$

es equivalente al test anterior dado que si  $X$  e  $Y$  son independientes  $\implies \tau = 0$  y, por tanto, se tiene que si  $\tau \neq 0 \implies X$  e  $Y$  son dependientes. No tenemos que si  $\tau = 0$  entonces  $X$  e  $Y$  son independientes, pero dado que la hipótesis alternativa es equivalente en ambos tests, se puede asumir que los resultados van a ser los mismos.

En el apartado 2.1.1, se introduce un gráfico de control clásico o tipo Shewhart y se indican los elementos que constituyen este tipo de gráfico: los límites de control (superior y/o inferior) y la línea central, la cual representa el valor deseado o esperado para el estadístico graficado en el estado bajo control.

En este trabajo, se considerará que la situación bajo control es la *independencia* de las variables objeto de estudio ( $X$  e  $Y$ ), es decir,  $\tau = 0$ .

El cálculo de los límites de control se realiza de la manera explicada en las ecuaciones (3.1) y (3.2), siendo  $\alpha$  el nivel de significación y  $F$  la función de distribución del estadístico  $\hat{\tau}$  bajo la hipótesis nula ( $\tau = 0$ ).

$$F(LCL) = P(\hat{\tau} \leq LCL \mid \tau = 0) = \alpha \implies LCL = F^{-1}(\alpha) \quad (3.1)$$

$$F(UCL) = P(\hat{\tau} \leq UCL \mid \tau = 0) = 1 - \alpha \implies UCL = F^{-1}(1 - \alpha) \quad (3.2)$$

### 3. Desarrollo de un gráfico de control no paramétrico para la correlación basado en el estadístico de Kendall

Tal y como puede deducirse de las ecuaciones 3.1 y 3.2, se está asumiendo el diseño de dos gráficos unilaterales. En concreto, se utilizará el gráfico con UCL cuando la correlación  $\tau$  a detectar sea mayor que cero; y el gráfico con el LCL cuando la situación fuera de control que se desee detectar sea  $\tau < 0$ .

Ya se conocen los elementos necesarios y la forma del gráfico de control, a continuación se va a proceder a obtener las medidas de desempeño de este gráfico.

## 3.2. Obtención medidas de desempeño del gráfico

El objetivo de este apartado es proporcionar los métodos con los que obtener las medidas de desempeño del gráfico, tanto para cuando el proceso se encuentra bajo control como para el caso en el que el proceso se encuentra fuera de control.

### 3.2.1. Proceso bajo control

El cálculo del desempeño del gráfico de control bajo el supuesto de que las variables objeto de estudio son independientes (en otras palabras, cuando el proceso se encuentra bajo control) se realiza, a partir de un valor asumido para la tasa de falsa alarma  $\alpha$ , mediante la fórmula

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

que ya había sido expuesta en la ecuación 2.1.

### 3.2.2. Proceso fuera de control

Cuando el proceso se encuentra fuera de control (las variables objeto de estudio están correlacionadas), el cálculo del  $ARL_1$  no se puede calcular con facilidad mediante la fórmula introducida en la ecuación 2.4. Esto es debido a que, tal y como ya se explicó en el apartado 2.4.2, no se dispone de la función de distribución del estadístico  $\hat{\tau}$  y esta función de distribución es fundamental para el cálculo de la potencia estadística. En consecuencia, el cálculo de esta medida de desempeño del gráfico ( $ARL_1$ ) se obtiene mediante simulación Montecarlo, tal y como se explica a continuación.

En primer lugar, la situación de falta de control (es decir, la correlación, asociación o dependencia entre  $X$  e  $Y$ , representada por un valor  $\tau \neq 0$ ) se va a modelizar mediante un modelo de cópulas, tal como se ha presentado en la sección 2.5. En dicha sección, también se explica y ejemplifica el cálculo de  $\tau$  a partir de las cópulas.

A partir de un valor dado de  $\tau \neq 0$  a detectar, y asumiendo una cópula concreta, se obtiene el valor del parámetro  $\theta$  (que representa, dentro de la cópula, el grado de asociación entre sus componentes) mediante las ecuaciones (2.11), (2.10) o (2.12), dependiendo de que cópula se utilice. Este parámetro  $\theta$  se calcula con el objetivo de generar valores (pseudo)aleatorios de la cópula  $(u_1, u_2) \in [0, 1] \times [0, 1]$  a partir de los algoritmos expuestos en las ecuaciones (2.14), (2.13) o (2.15) y (2.16), dependiendo de con qué cópula se esté utilizando en el estudio.

### 3.2. Obtención medidas de desempeño del gráfico

Estos valores (pseudo)aleatorios de la cópula  $(u_1, u_2)$  se van a utilizar para generar valores aleatorios de las variables  $X$  e  $Y$ , fijando previamente las distribuciones de cada una de estas dos variables aleatorias. Estos valores se van a obtener de la forma determinada en la ecuación (3.3) siendo  $F_X$  y  $F_Y$  las funciones de distribución de  $X$  e  $Y$  respectivamente.

$$(x_1, y_1) = (F_X^{-1}(u_1), F_Y^{-1}(u_2)) \quad (3.3)$$

De este modo se calcula una muestra con  $n$  valores aleatorios de  $(X, Y)$  y se procede a calcular el  $\hat{\tau}$  de esta muestra haciendo uso de la fórmula (2.6).

Una vez se haya obtenido la  $\tau$  de Kendall muestral,  $\hat{\tau}$ , se comprueba si se encuentran dentro de los límites de control (en otras palabras, si  $\hat{\tau}$  es mayor o menor que los límites de control) expresados en las ecuaciones (3.1) y (3.2), dependiendo si el valor de  $\tau$  a detectar presenta un valor negativo o positivo, respectivamente (recuérdese que el planteamiento propuesto es diseñar un gráfico unilateral, es decir, con solo un límite de control, el superior o el inferior).

- **Test de hipótesis unilateral derecha.**

$$\begin{cases} H_0 : \tau = 0 \\ H_1 : \tau > 0 \end{cases}$$

En este caso, se rechaza  $H_0$  cuando  $\hat{\tau} \geq UCL$ .

- **Test de hipótesis unilateral izquierda.**

$$\begin{cases} H_0 : \tau = 0 \\ H_1 : \tau < 0 \end{cases}$$

En este caso, se rechaza  $H_0$  cuando  $\hat{\tau} \leq LCL$ .

Como ya se ha mencionado previamente, un gráfico de control es la representación gráfica de un test de hipótesis. Por lo tanto, en este trabajo se estudiarán dos gráficos de control con un límite de control cada uno, el uso de cada uno de estos gráficos de control se decidirá al observar el signo de la  $\tau$  a detectar.

Finalmente, se procede a calcular el promedio de puntos que salen fuera de los límites de control, es decir, el  $ARL_1$  para cada uno de los casos. El cálculo de esta medida de desempeño,  $ARL_1$ , se va a realizar mediante simulación, para ello se van a generar valores aleatorios para  $\tau$  muestral tal y como se ha explicado al inicio de esta sección. Tras repetir  $N$  veces este experimento se calcula el  $ARL_1$  de la siguiente manera:

$$ARL_1 = \frac{N}{\text{puntos fuera control}}$$

Para este trabajo se ha elegido un  $N = 100000$ , dado que con este valor se obtienen estimaciones suficientemente precisas del  $ARL_1$ . De esta manera, asumiendo valores para  $\alpha$ ,  $n$ ,  $\tau \neq 0$  y la cópula, se puede obtener mediante simulación una estimación del rendimiento del gráfico bajo esas condiciones.

### 3.3. Diseño de experimentos

En las dos secciones anteriores se ha desarrollado la metodología para diseñar (sección 3.1) y evaluar el desempeño (sección 3.2) del gráfico propuesto en este trabajo. Además, en la sección actual, se propone una forma de analizar y evaluar la influencia de los diferentes parámetros ( $\alpha$ ,  $n$ ,  $\tau \neq 0$  y la cópula) sobre el rendimiento del gráfico medido a través de su  $ARL_1$ . En concreto, se realizará un diseño de experimentos, tomando varios valores representativos para cada uno de los parámetros o factores.

Sobre los resultados de dicho experimento, se realizará un análisis descriptivo y un ANOVA para ver qué parámetros son los más influyentes sobre la velocidad de detectar los puntos que salen fuera de los límites de control; en otras palabras, los más influyentes sobre el  $ARL_1$ . Esto se muestra en el capítulo 4.

## 4. Experiencias computacionales

En el capítulo 3, se ha mostrado cómo construir un gráfico de control para el coeficiente de correlación de Kendall, y cómo evaluar su eficiencia.

En el presente capítulo, se dan detalles de cómo implementar la propuesta del capítulo anterior. Se ha elegido para ello el lenguaje R. También se muestran los resultados de aplicar este gráfico sobre una batería de ejemplos representativa, con el fin de evaluar la eficiencia del gráfico para detectar diferentes situaciones de falta de control bajo diversas condiciones iniciales.

### 4.1. Implementación

En esta sección se detalla la implementación informática del gráfico propuesto en el capítulo 3. En concreto, se muestra cómo calcular el límite de control inferior o superior (recordemos que el planteamiento es usar un gráfico unilateral), y cómo determinar el ARL, todo ello para diversos valores de los parámetros tipo de cópula, tamaño muestral, valor de  $\tau$  a detectar, distribuciones de  $X$  e  $Y$  y desviaciones típicas de  $X$  e  $Y$ .

#### 4.1.1. Software y hardware empleado

Para la realización de este trabajo, se ha utilizado el lenguaje de programación R versión 4.1.3, en su entorno de desarrollo integrado (IDE) de software libre RStudio versión 2022.02.3+492. Se ejecutaron en un sistema operativo Windows 10 Home versión 19043.1766 con procesador de 2.6GHz Intel Core i7 y memoria RAM de 8GB.

#### 4.1.2. Implementación en R

El fichero de R con la implementación completa de todo el código que se va a mostrar a continuación se encuentra disponible en el anexo A.1.

Los límites de control con los que se va a llevar a cabo el control estadístico del coeficiente de correlación  $\tau$ , que han sido presentados en las ecuaciones (3.2) y (3.1), se van a calcular mediante la función `qKendall` del paquete `SuppDists` [10] como ya se mencionó en la sección 2.4.2, esto se puede observar implementado en el Código 4.1.

```
LCL<-qKendall(p=alpha,N=n,lower.tail = TRUE)
```

#### 4. Experiencias computacionales

```
UCL<-qKendall(p=1-alpha,N=n,lower.tail = TRUE)
```

Código 4.1: Cálculo del límite de control inferior o superior para un gráfico unilateral tipo Shewhart para el coeficiente de correlación de Kendall con nivel de significación  $\alpha$  y tamaño muestral  $n$ .

Como se mencionó en la sección 2.5, existe una relación entre  $\tau$  y las cópulas, obteniendo de esta manera una relación entre  $\tau$  y el parámetro  $\theta$ . En concreto, las cópulas que se van a utilizar en estas experiencias numéricas son las de Clayton, Frank y Gumbel. Las fórmulas que permiten calcular  $\theta$  a partir de la  $\tau$  a detectar, vienen determinadas por las ecuaciones (2.10), (2.11) y (2.12). Estas tres ecuaciones han sido implementadas mediante el Código 4.2.

$\tau$ .

```
D1<-function(theta){
  integ<-integrate(t,lower=0,upper=theta)
  return(iffelse(theta == 0, 0,(1/theta)*integ$value))}
tau.Frank<-function(theta,tau){
  (1+4*(D1(theta)-1)/theta-tau)
}
theta.valor<-function(tau,C){
  if(C=="Frank"){
    theta=uniroot(function(x) tau.Frank(x,tau),c(-40,40)
  )$root }
  else{
    if (C=="Clayton"){
      theta=(2*tau)/(1-tau)}
    else{
      if (C=="Gumbel"){
        theta=1/(1-tau)}
      else { cat('La cópula introducida:', C, ' no es correcta,
  las opciones son Frank/Clayton/Gumbel')}
    }
  }
  return(theta)
}
```

Código 4.2: Cálculo del parámetro  $\theta$  correspondiente a las cópulas de Frank

A partir de estos valores de  $\theta$  que se han obtenido mediante el código 4.2, se van a generar valores aleatorios de cada una de las cópulas arquimedianas con las que se esté trabajando.

Para generar valores aleatorios mediante la cópula de Frank, se utilizará el algoritmo explicado en la ecuación (2.13) de la sección 2.5. Este algoritmo es implementado en R mediante el Código 4.3.

```
Frank<-function(theta,n){
  u2=0
```

```

u1<-runif(n,min=0,max=1)
v2<-runif(n,min=0,max=1)
for (i in 1:n){
  u2.2<-(-1/theta)*log(1+(v2[i]*(1-exp(-theta)))/(v2[i]*(exp(-theta*u1[i])-1)-exp(-theta*u1[i]))))
  u2<-append(u2,u2.2)}
u2<-u2[-1]
u<-cbind(u1,u2)
return(u)}

```

Código 4.3: Generación de valores aleatorios de la cópula de Frank.

Para generar valores aleatorios mediante la cópula de Clayton, se utilizará el algoritmo explicado en la ecuación (2.14) de la sección 2.5. Este algoritmo es implementado en R mediante el Código 4.4.

```

Clayton<-function(theta,n){
  u2=0
  u1<-runif(n,min=0,max=1)
  v2<-runif(n,min=0,max=1)
  for (i in range(1:n)){
    u2.2<-((u1[i]^(-theta))*(v2[i]^(-theta/(1+theta))-1)+1)^(-1/theta))
    u2<-append(u2,u2.2)}
  u<-cbind(u1,u2[-1])
  return(u)}

```

Código 4.4: Generación de valores aleatorios de la cópula de Clayton.

Para generar valores aleatorios mediante la cópula de Clayton, se utilizará el algoritmo explicado en las ecuaciones (2.15) y (2.16) de la sección 2.5. Este algoritmo es implementado en R mediante el Código 4.5.

```

Gumbel<-function(theta,n){
  v1<-runif(n,min=0,max=1)
  v2<-runif(n,min=0,max=1)
  u1=0
  u2=0
  for (i in 1:n){
    w<-golden.section.search(function(w){w*(1-log(w)/theta)-v2[i]},0,1)
    u1.1<-exp(v1[i]^(1/theta)*log(w))
    u2.1<-exp((1-v1[i])^(1/theta)*log(w))
    u1<-append(u1,u1.1)
    u2<-append(u2,u2.1)}
  u<-cbind(u1[-1],u2[-1])
  return(u)
}

```

#### 4. Experiencias computacionales

```
}
```

Código 4.5: Generación de valores aleatorios de la cópula de Gumbel.

A continuación, se detalla cómo generar en R valores aleatorios de las variables X e Y a partir de las distribuciones de estas, de la cópula y del valor de  $\theta$ . La función mediante la cual se van a generar estos valores es la que se presenta en el Código 4.6.

```
Generador.aleatorio<-function(xdist,ydist,C,theta,y1,y2,x1,x2,n){  
  # y1,y2,x1,x2 son los parámetros necesarios para las ←  
  distribuciones de y y de x respectivamente.  
  u<-Rand.val.copula(theta,C,n)  
  if (xdist=="Normal" & ydist=="Normal") {  
    x=0  
    y=0  
    for (i in 1:n){  
      x.1<-qnorm(u[i,1],mean=x1,sd=x2)  
      y.1<-qnorm(u[i,2],mean=y1,sd=y2)  
      x<-append(x,x.1)  
      y<-append(y,y.1)}  
  } else if (xdist=="Weibull" & ydist=="Normal") {  
    x=0  
    y=0  
    for (i in 1:n){  
      x.1<-qweibull(u[i,1],shape=x1,scale=x2)  
      y.1<-qnorm(u[i,2],mean=y1,sd=y2)  
      x<-append(x,x.1)  
      y<-append(y,y.1)}  
  } else if (xdist=="Weibull" & ydist=="Weibull") {  
    x=0  
    y=0  
    for (i in 1:n){  
      x.1<-qweibull(u[i,1],shape=x1,scale=x2)  
      y.1<-qweibull(u[i,2],shape=y1,scale=y2)  
      x<-append(x,x.1)  
      y<-append(y,y.1)}  
  } else if (xdist=="Normal" & ydist=="Weibull") {  
    x=0  
    y=0  
    for (i in 1:n){
```

```

x.1<-qnorm(u[i,1],mean=x1,sd=x2)
y.1<-qweibull(u[i,2],shape=y1,scale=y2)
x<-append(x,x.1)
y<-append(y,y.1)}

}
else {cat('Distribuciones introducidas no validas', xdist,ydist↔
, 'las opciones son Normal/Weibull')}
xy<-cbind(x[-1],y[-1])
return(xy)}

```

Código 4.6: Generación de valores aleatorios de las variables X e Y a partir de las distribuciones de estas de la cópula y del valor de  $\theta$ .

De esta manera, se consiguen valores pseudoaleatorios  $(u_1, u_2) \in [0, 1] \times [0, 1]$  a partir, solamente, de la cópula a utilizar y el valor de tau en control. Esto se realiza mediante el código que se muestra en Código 4.7.

```

Rand.val.copula<-function(tau,C,n){
  theta<-theta.valor(tau,C)
  if(C=="Frank"){
    u1u2=Frank(theta,n)
  }
  else{
    if (C=="Clayton"){
      u1u2=Clayton(theta,n)}
    else{
      if (C=="Gumbel"){
        u1u2=Gumbel(theta,n)
      }
    }
  }
  else { cat('La cópula introcida:', C, ' no es ↔
correcta,
las opciones son Frank/Clayton/Gumbel')}
}
}
return(u1u2)
}

```

Código 4.7: Generación de valores aleatorios de una cópula.

Para calcular la medida de eficiencia del gráfico de control,  $ARL_1$ , se debe recurrir al uso de simulación, como ya se explicó en la sección 3.2.2. En este trabajo se va a recurrir al uso de simulación Montecarlo, debido a que no se puede modelizar analíticamente la distribución en el muestreo del estadístico  $\hat{\tau}$  cuando no se está bajo la hipótesis de independencia entre las dos variables X e Y. Esta simulación se va a realizar a partir de las cópulas, generando  $n$  valores aleatorios mediante la función creada en el Código 4.6.

#### 4. Experiencias computacionales

```
implementacio.final<-function(alpha,n,x,y,C,tau,y1,y2,x1,x2,NMAX)←
{
count=0
fuera.control=0
if (tau<0){
while (count<NMAX){
LCL<-qKendall(p=alpha,N=n,lower.tail = TRUE)
theta<-theta.valor(tau,C)
xy<-Generador.aleatorio(x,y,C,theta,y1,y2,x1,x2,n)
tau_cap=cor(xy[,1],xy[,2],method="kendall")
if (isTRUE(tau_cap<LCL )==TRUE){
fuera.control=fuera.control+1}
count=count+1}}
else{
while (count<NMAX){
UCL<-qKendall(p=1-alpha,N=n,lower.tail = TRUE)
theta<-theta.valor(tau,C)
xy<-Generador.aleatorio(x,y,C,theta,y1,y2,x1,x2,n)
tau_cap<-cor(xy[,1],xy[,2],method="kendall")
if (isTRUE(tau_cap>UCL)==TRUE){
fuera.control=fuera.control+1}
count=count+1}}
ARL1=NMAX/fuera.control
return(c(ARL1,theta))}
```

Código 4.8: Obtención del  $ARL_1$  mediante simulación Montecarlo en el caso del proceso fuera de control.

## 4.2. Análisis a realizar

En esta sección se va a detallar la batería de ejemplos representativos que se va a generar, además de qué métodos se van a utilizar para la realización del análisis del estudio. El objetivo de la creación de este banco de pruebas es evaluar el comportamiento del gráfico de control bajo el efecto de distintas condiciones iniciales que se consideran representativas y diferentes entre ellas.

Los casos concretos en que se va a centrar este trabajo son los siguientes:

- Observaciones bivariantes  $(X_1, Y_1), \dots, (X_n, Y_n)$  con tamaños muestrales  $n = 10$  y  $n = 30$ .
- Un nivel de significación  $\alpha = 0,0027$  que es el que se suele utilizar más en control estadístico.
- Las variables  $X$  e  $Y$  con las que se ha trabajado van a presentar distribución Weibull o distribución Normal con los siguientes coeficientes de variación  $\frac{\sigma}{\mu} \in \{0.1, 0.2, 0.5\}$ , es

decir, se estudiarán los casos en qué  $\mu = 10$  y  $\sigma \in \{1, 2, 5\}$ . Cuando se trate de una distribución Weibull, se deben calcular los parámetros de escala y forma que son  $\lambda$  y  $k$  respectivamente, a partir de  $\mu$  y  $\sigma$ , tal y como se va a explicar a continuación:

- El parámetro de escala  $\lambda$  se obtiene mediante el cálculo de la solución de la siguiente ecuación, para lo cual se han utilizado métodos numéricos.

$$\frac{\Gamma(\frac{2}{\lambda} + 1)}{\Gamma^2(\frac{1}{\lambda} + 1)} = \left(\frac{\sigma}{\mu}\right)^2 + 1$$

- El parámetro de forma  $k$  se obtiene a partir de los parámetros  $\lambda$  y  $\mu$  de la siguiente manera:

$$k = \frac{\mu}{\Gamma(\frac{1}{\lambda} + 1)}$$

Con que las distribuciones que se van a considerar en este trabajo son las expuestas en la Tabla 4.1

Normal( $\mu, \sigma$ )		Weibull( $\lambda, k$ )	
$\mu$	$\sigma$	$\lambda$	$k$
10	1	12.1534	10.4304
10	2	5.7974	10.7998
10	5	2.1013	11.2906

Cuadro 4.1.: Distribuciones de X e Y.

- Las cópulas Arquimedianas bivariantes que se van a utilizar en este trabajo van a ser tres y son las siguientes:
  - Cópula de Frank.
  - Cópula de Clayton.
  - Cópula de Gumbel.
- Por último, las  $\tau$  a detectar que presenta los siguientes valores:  $\{0,1, 0,3, 0,5, 0,7, 0,9, -0,1, -0,3, -0,5, -0,7, -0,9\}$

Una vez ya se han determinado los casos a estudiar, se procederá al análisis de los parámetros con el objetivo de conocer su influencia sobre la detección de los casos fuera de control, dado que no se puede apreciar a simple vista qué factores influyen. Con este objetivo, se ha realizado un análisis de la varianza (ANOVA). Es decir, el ANOVA va a analizar cómo afectan los distintos factores al  $ARL_1$ .

## 4.3. Resultados y discusión

### 4.3.1. Resultados de la experiencia numérica

Siguiendo los pasos detallados en el capítulo 3, aplicados a los casos particulares que se han mencionado en la sección anterior y haciendo uso de las funciones implementadas en R

#### 4. Experiencias computacionales

que se han ilustrado en el apartado 4.1.2, se van a obtener una serie de resultados que van a ser expuestos a lo largo de esta sección.

Tal y como se ha mencionado previamente, en este trabajo se van a estudiar los casos para  $n = 10$  y  $n = 30$  observaciones bivariantes  $(X_1, Y_1), \dots, (X_n, Y_n)$  con un  $\alpha = 0,0027$  que es el que se suele utilizar más en control estadístico. Por consiguiente, los límites con los que se trabajará a lo largo de este estudio son los que se muestran en la Tabla 4.2, estos límites han sido calculados mediante el Código 4.1.

	$n = 10$	$n = 30$
LCL	-0,6444444	-0,3517241
UCL	0,6444444	0,3517241

Cuadro 4.2.: Límites de control para los casos a estudiar obtenidos mediante el código 4.1.

Se debe tener en cuenta que el valor de  $\tau$  fuera de control a detectar toma los siguientes valores de  $\tau$ :  $\{0,1, 0,3, 0,5, 0,7, 0,9, -0,1, -0,3, -0,5, -0,7, -0,9\}$ . Mediante el Código 4.2 se consiguen calcular los valores presentados en la Tabla 4.3.

A modo de ejemplo, y para mostrar el efecto de elegir una u otra cópula para modelizar el comportamiento conjunto de las variables  $X$  e  $Y$ , se ha realizado una representación gráfica de la función densidad de cada una de las cópulas tomando distribuciones marginales  $N(10, 2)$  y  $W(5,7974, 10,7998)$  con una  $\tau$  a detectar de  $\tau = 0,3$ . Para realizar estos gráficos, presentados en las figuras 4.1, 4.2 y 4.3 se ha utilizado el software ‘Mathematica’.

Al observar las funciones de densidad expuestas en las figuras 4.1, 4.2 y 4.3, se puede notar que teniendo las mismas distribuciones marginales ( $N(10, 2)$  y  $W(5,7974, 10,7998)$ ) y fijando el parámetro  $\tau$  en los tres casos los gráficos son distintos; con esto, se muestra que la cópula elegida influirá en la distribución conjunta de  $X$  e  $Y$ . El código mediante el cual se han realizado estas gráficas se encuentra adjunto en A.3.

### 4.3. Resultados y discusión

$\tau$	$\theta$		
	Frank	Clayton	Gumbel
0.1	0.91	0.22	1.11
0.3	2.92	0.86	1.43
0.5	5.74	2.00	2.00
0.7	11.41	4.67	3.33
0.9	38.28	18.00	10.00
-0.1	-0.91	-0.18	0.91
-0.3	-2.92	-0.46	0.77
-0.5	-5.74	-0.66	0.66
-0.7	-11.41	-0.82	0.59
-0.9	-38.28	-0.94	0.53

Cuadro 4.3.: Valor del parámetro  $\theta$  según el tipo de cópula y el valor de  $\tau$  a detectar fuera de control.

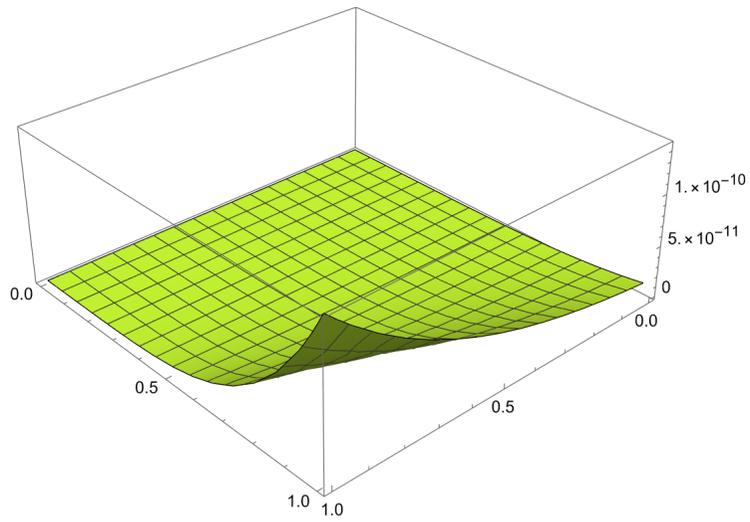


Figura 4.1.: Función de densidad utilizando la cópula de Frank.

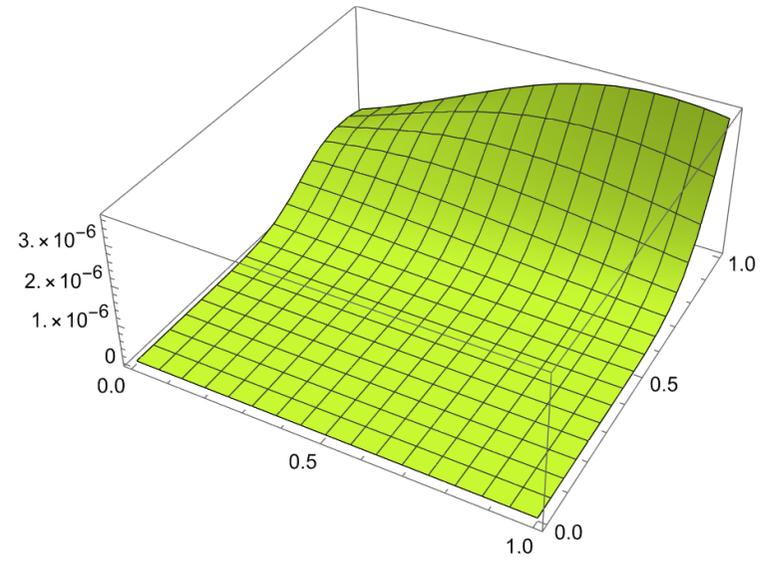


Figura 4.2.: Función de densidad utilizando la cópula de Clayton.

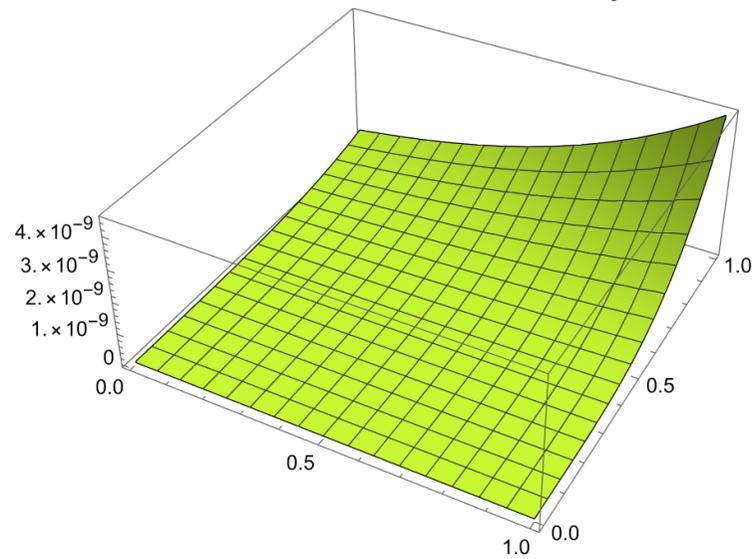


Figura 4.3.: Función de densidad utilizando la cópula de Gumbel.

A partir de los valores del parámetro  $\theta$ , expuestos en la tabla 4.3, sabiendo el valor de  $\tau$  que se desea detectar, se van a generar valores aleatorios de las variables  $X$  e  $Y$  y se ha calculado el coeficiente de correlación muestral  $\hat{\tau}$  de cada una de las muestras generadas para cada valor de  $\theta$  presente en la Tabla 4.3.

Seguidamente, se procede a realizar la simulación del proceso para obtener una estimación del parámetro  $ARL_1$ , haciendo uso del código 4.8. La simulación se realiza para cada uno de los casos que se han introducido en la sección 4.2. La tabla completa de los resultados obtenidos se encuentra disponible en el anexo A.2.

$\tau$	$\alpha$	$n$	Dist. XY	Desv	Cópula	$\theta$	$ARL_1$
0.1	0.0027	10	WN	d11	Gumbel	1.11111111	75.47169811
0.1	0.0027	10	WN	d12	Gumbel	1.11111111	75.52870091
...	...	...	...	...	...	...	...
-0.9	0.0027	30	NN	d55	Gumbel	0.526315789	1.008796707
-0.9	0.0027	30	WW	d55	Gumbel	0.526315789	1.008501669
0.1	0.0027	10	NN	d11	Frank	0.907355725	82.85004143
0.1	0.0027	10	WN	d15	Frank	0.907355725	83.82229673
...	...	...	...	...	...	...	...
-0.9	0.0027	30	NN	d55	Frank	-38.28121168	1
-0.9	0.0027	30	WW	d55	Frank	-38,28121168	1
0.1	0.0027	10	WN	d11	Clayton	0.22222222	113.2502831
0.1	0.0027	10	NN	d11	Clayton	0.22222222	108.1081081
...	...	...	...	...	...	...	...
-0.9	0.0027	30	NN	d55	Clayton	-0.947368421	65.87615283
-0.9	0.0027	30	WW	d55	Clayton	-0.947368421	63.33122229

Cuadro 4.4.: Extracto de la batería de ejemplos representativos que se puede ver completa en el anexo A.2.

Observando la tabla 4.4 no se puede apreciar con claridad el efecto de los distintos factores sobre el  $ARL_1$ . Por consiguiente, a continuación se llevará a cabo un análisis descriptivo y un análisis ANOVA sobre los resultados obtenidos con la finalidad de determinar el efecto de cada factor sobre la medida de desempeño del gráfico,  $ARL_1$ .

### 4.3.2. Análisis de los resultados obtenidos

Con la finalidad de validar el gráfico de control no paramétrico de forma óptima para el coeficiente de correlación de Kendall  $\tau$ , se ha realizado un estudio de los factores que pueden resultar influyentes y decisivos en el desempeño de los gráficos de control, que han sido introducidos previamente. Este estudio se llevará a cabo mediante un análisis descriptivo, y posteriormente un análisis de la varianza (ANOVA).

Se han obtenido diferentes valores de  $ARL_1$  mediante los diversos casos realizados para los distintos valores de los parámetros  $n$  (tamaño de muestra),  $\mu$  (media muestral),  $\sigma$  (desviación

#### 4. Experiencias computacionales

típica), distribución, cópula y  $\tau$ . Estos resultados son los ya mencionados en la sección 4.3.1 y que se encuentran disponibles el anexo A.2.

En primer lugar, se realiza un análisis descriptivo de dichos resultados mediante diagramas de caja y bigotes o de box-whisker, gráficos de cajas y bigotes que se definen como un método gráfico para mostrar la variación de un conjunto de datos. En la mayoría de los casos, un gráfico de cajas y bigotes puede proporcionar detalles adicionales al tiempo que permite mostrar múltiples conjuntos de datos en el mismo gráfico.

Dado que los gráficos de cajas y bigotes permiten comparar datos de diferentes categorías para tomar decisiones más fáciles y eficaces, en este trabajo se van a utilizar con la finalidad de detectar si hay diferencias entre los distintos niveles de cada factor sobre el  $ARL_1$ .

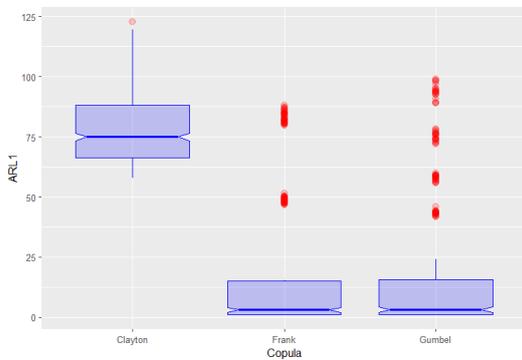


Figura 4.4.:  $ARL_1$  y cópulas.

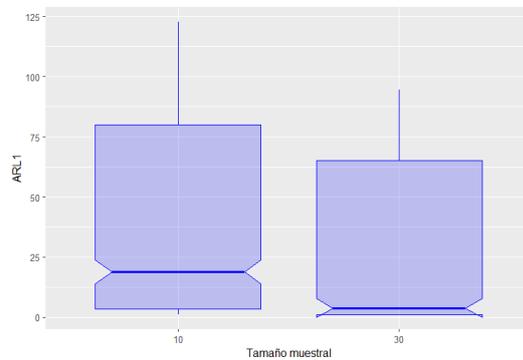


Figura 4.5.:  $ARL_1$  y tamaño muestral .

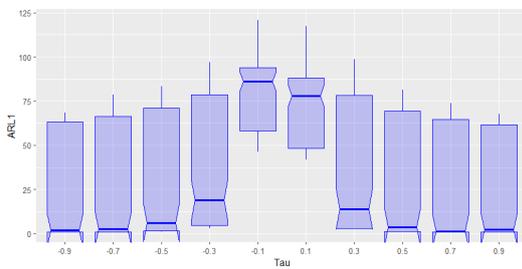


Figura 4.6.:  $ARL_1$  y  $\tau$  a detectar.

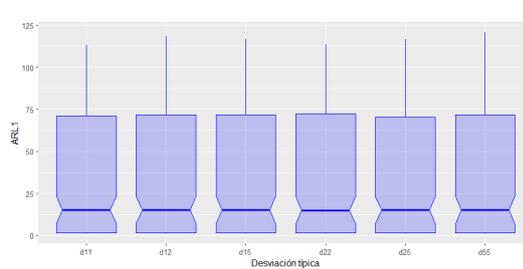


Figura 4.7.:  $ARL_1$  y desviación típica.

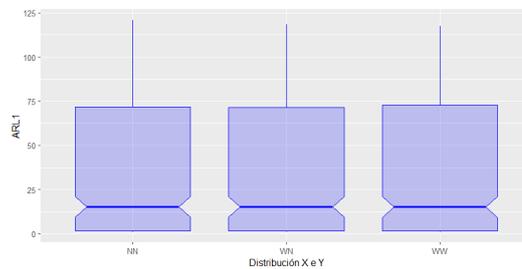


Figura 4.8.:  $ARL_1$  y Distribución X e Y.

Tal y como se puede observar en la figura 4.4, la cópula de Clayton presenta valores de

$ARL_1$  más elevados que en los casos de la cópula de Frank y de Gumbel. Por tanto, se puede deducir que existen diferencias relevantes entre el  $ARL_1$  calculado utilizando la cópula de Clayton y el  $ARL_1$  calculado por las otras dos cópulas. En concreto el diagrama de cajas múltiple sugiere que, por término medio (o mediano), cuesta más detectar una situación de falta de control ( $ARL_1$  superior) si la relación entre las variables  $X$  e  $Y$  sigue una cópula de Clayton, respecto a Frank o Gumbel, y suponiendo que el resto de condiciones sean idénticas

Al estudiar la figura 4.5, se observa que al analizar muestras de tamaño 10 el  $ARL_1$  presenta valores ligeramente más elevados que en el caso de analizar muestras de tamaño 30. Por consiguiente, se detectan con mayor facilidad los casos fuera de control con un tamaño de muestra  $n = 30$ . Este resultado era esperable debido a que cuanto mayor es el tamaño muestral, mayor es la potencia ( $\beta$ ) para detectar desviaciones del proceso (en este caso, para detectar correlación entre  $X$  e  $Y$ ).

Examinando al detalle la figura 4.6, se puede apreciar que a medida que aumenta (en valor absoluto) el valor de  $\tau$  a detectar fuera de control, la velocidad de detección será menor y, por tanto, el  $ARL_1$  presentará menores valores. Además, observando la amplitud de las cajas, el  $ARL_1$  presenta valores más homogéneos a la hora de detectar valores pequeños, es decir, presenta menor variación. Esto sugiere que la detección de una correlación pequeña entre  $X$  e  $Y$  cuesta bastante, en términos de  $ARL_1$ , con menor importancia del resto de factores que cuando la correlación  $\tau$  a detectar es mayor.

En cuanto al efecto del coeficiente de variación sobre el  $ARL_1$ , se procederá a observar cómo afectan los distintos valores de  $\sigma_x$  y  $\sigma_y$ , dado que la media ( $\mu$ ) permanece constante a lo largo de todo el estudio. Con este objetivo, se han observado dos gráficos de cajas múltiples en los que se puede apreciar que la desviación típica no afecta a la velocidad de detección de los valores fuera de control, tal y como se puede observar en la figura 4.8 (solo se ha adjuntado gráfico dado que para ambos actúan de igual manera).

A continuación, mediante el programa "Statgraphics", se van a obtener unos gráficos a partir de los cuales se estudiará la existencia de posibles interacciones de orden 2 entre los distintos factores, es decir, se estudiará si los distintos factores presentan efectos entre ellos. Nótese que se sigue tratando de un análisis puramente descriptivo.

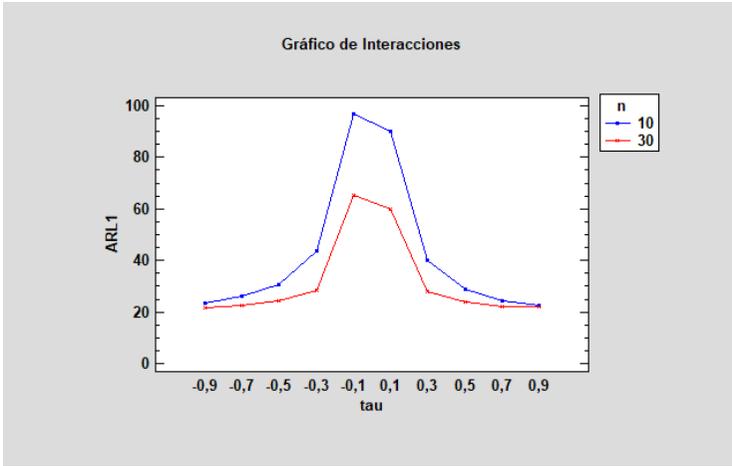


Figura 4.9.: Interacción  $\tau$  y  $n$ .

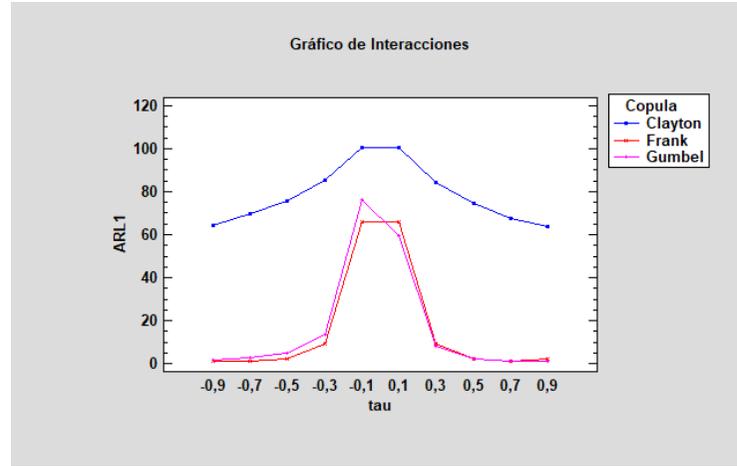


Figura 4.10.: Interacción  $\tau$  y cópula.

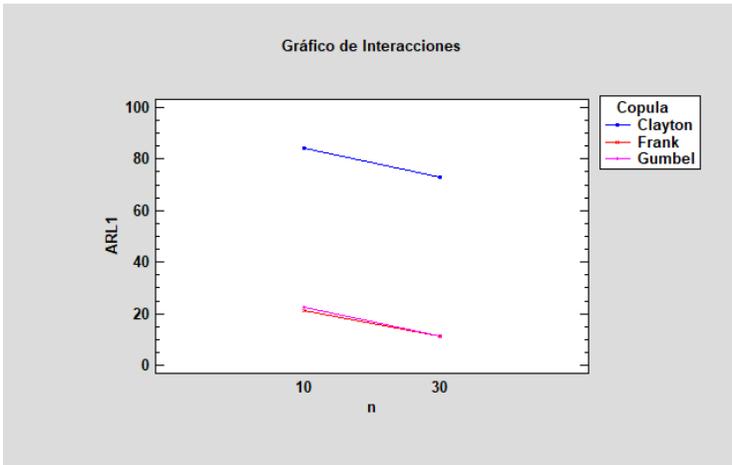


Figura 4.11.: Interacción  $n$  y cópula.

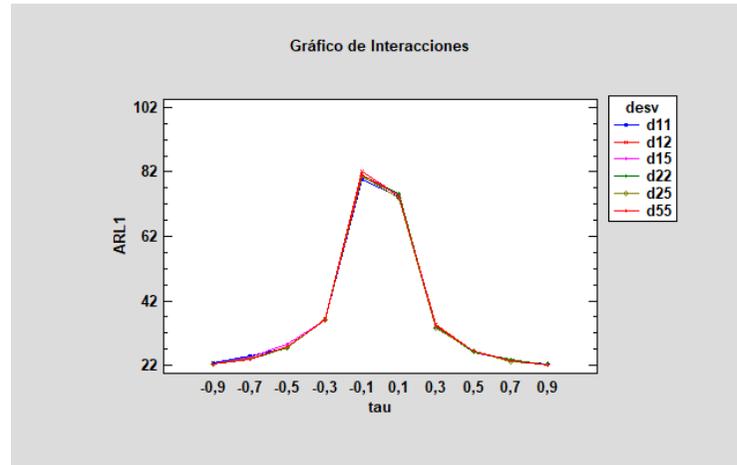


Figura 4.12.: Interacción  $\tau$  y desviación  $X$  e  $Y$ .

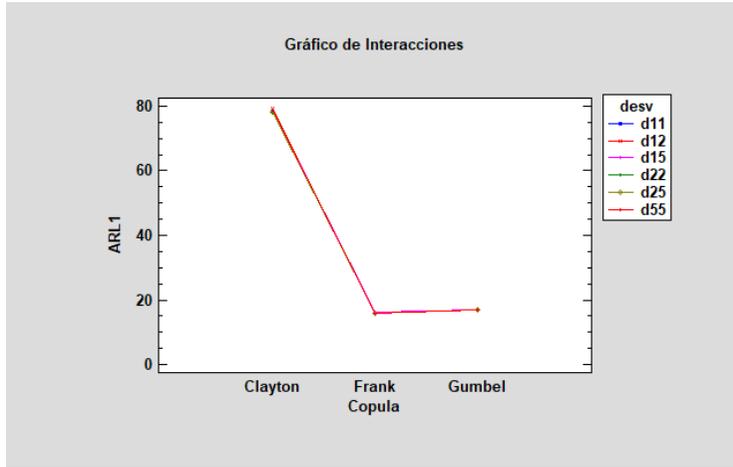


Figura 4.13.: Interacción cópula y desviación.

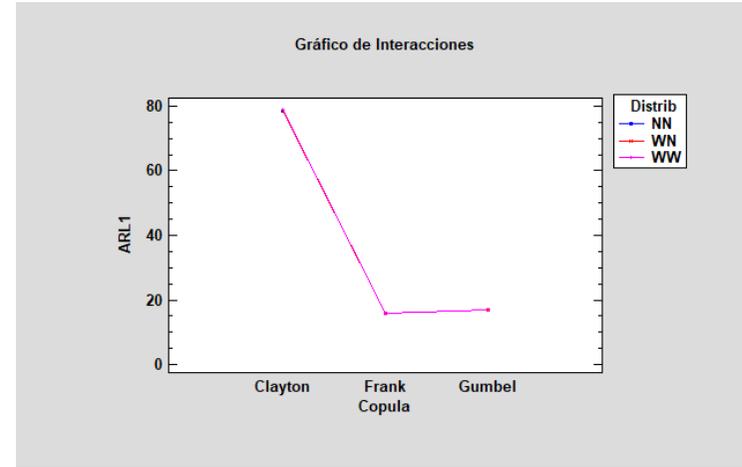


Figura 4.14.: Interacción cópula y distribución.

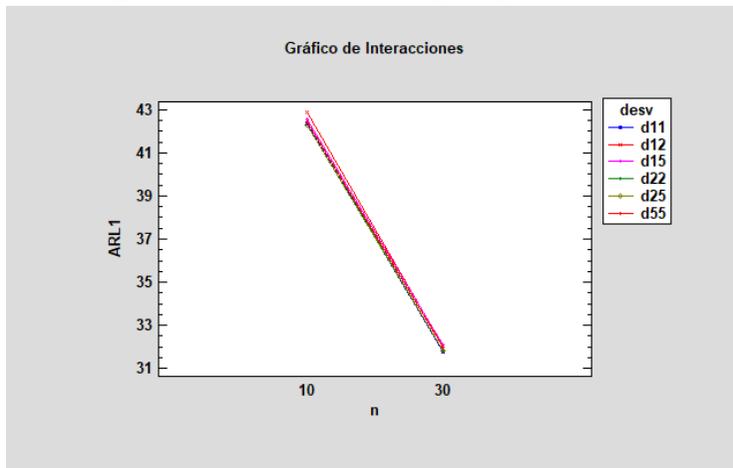


Figura 4.15.: Interacción  $n$  y desviación.

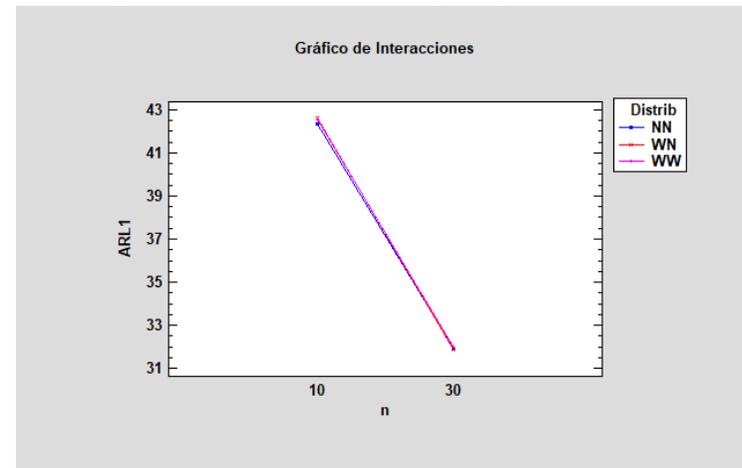


Figura 4.16.: Interacción  $n$  y distribución.

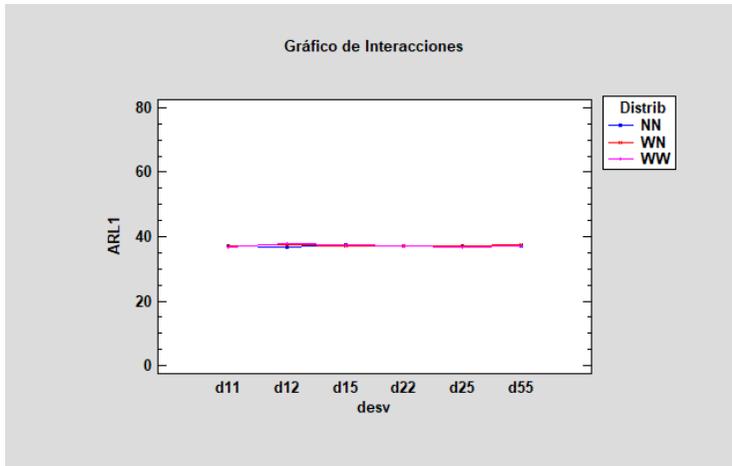


Figura 4.17.: Interacción desviación y distribución.

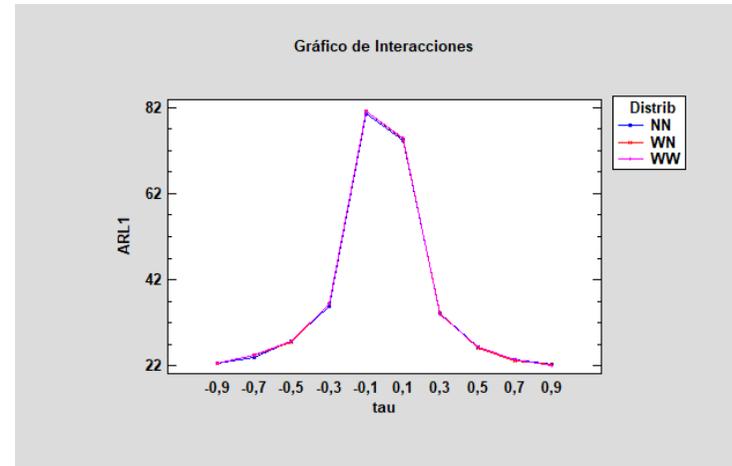


Figura 4.18.: Interacción  $\tau$  y distribución.

Del estudio de estas gráficas de interacción se obtienen las siguientes conclusiones:

- Tras el estudio de la gráfica de interacción 4.9, se deduce que la relación entre tau ( $\tau$ ) y el valor de  $ARL_1$  depende del tamaño muestral. Las líneas no paralelas en la gráfica de interacción indican los efectos de interacción entre  $\tau$  y el tamaño muestral, dado que se observa que a medida que aumenta el valor de  $\tau$  a detectar la distancia entre ambas líneas disminuye hasta solaparse y, por tanto, el valor de  $ARL_1$  presenta un decrecimiento respecto el valor esperado. Por consiguiente, esta interacción provoca que se den valores menores de  $ARL_1$  cuando se trata de valores de la  $\tau$  a detectar elevados y un tamaño muestral de  $n = 10$ .
- Tras el estudio de la gráfica de interacción 4.10, se deduce que la relación entre tau ( $\tau$ ) y el valor de  $ARL_1$  depende de la cópula utilizada. Tal y como era de esperar tras el estudio del diagrama de box-whisker no hay diferencias significativas entre la cópula Observando esta gráfica, se puede apreciar que la línea correspondiente a la cópula de Clayton no es totalmente paralela a la de las otras dos gráficas, se observa que para los valores de  $\tau \in 0.1, -0.1$  presenta valores de  $ARL_1$  más bajos de lo esperado.
- Al observar el gráfico de interacción 4.11, se observan líneas paralelas y un solapamiento entre las líneas asociadas a las cópulas Gumbel y Frank. Esto ya era esperado al observar el diagrama de box-whisker 4.4 que nos indicaba que no había diferencia entre ellas, pero sí con la cópula de Clayton. Por tanto, esto indica que la relación entre el tamaño muestral y el valor de  $ARL_1$  no depende de la cópula que se esté utilizando.
- Tras el estudio del resto de gráficos de interacción, se aprecia que se comportan tal y como era de esperar tras observar los diagramas de box-whisker expuestos previamente, por ende, estas interacciones no presentan efectos significativos sobre el valor del  $ARL_1$ .

Por lo tanto, de las gráficas de interacción se deduce que las interacciones que presentan efectos significativos sobre el  $ARL_1$  son:

- La interacción entre la  $\tau$  a detectar y el tamaño muestral.
- La interacción entre la  $\tau$  a detectar y la cópula utilizada.

En segundo lugar, se va a interpretar los resultados obtenidos mediante la realización de un análisis de la varianza (ANOVA) con el fin de analizar cómo afecta cada factor al  $ARL_1$  y ver si los supuestos que se han realizado mediante el análisis descriptivo son ciertos.

### **Análisis de la varianza (ANOVA) paramétrico**

Además de realizar un análisis descriptivo, que ya ha permitido extraer algunas conclusiones claras, efectuaremos a continuación un análisis de la varianza (ANOVA), con el fin de cuantificar la importancia relativa de las diferentes fuentes de variabilidad identificadas mediante el análisis gráfico.

Hay que tener en cuenta que los p-valores que se obtengan del análisis solo serán completamente válidos en el caso de que se cumplan las hipótesis de esta técnica. No obstante, como es sabido, dado que Montgomery (2008) [11] afirma que si se viola hipótesis de homocedasticidad

#### 4. Experiencias computacionales

de las varianzas, la prueba  $F$  solamente resulta afectada ligeramente en el modelo balanceado, es decir, cuando se tiene el mismo tamaño de la muestra en todos los tratamientos con efectos fijos. Dado que el modelo que se pretende estudiar en este trabajo es balanceado y se trata de modelos con efectos fijos, la falta de homocedasticidad no producirá alteraciones significativas. En el caso de los modelos de efectos aleatorios, las varianzas desiguales pueden perturbar significativamente las inferencias sobre los componentes de la varianza, incluso si se utilizan diseños equilibrados. Por tanto, se procede a realizar también un análisis de la varianza ANOVA, los resultados del cual se muestran en la Tabla 4.5. Además, en todo caso, estamos más interesados en utilizar la descomposición de la suma de cuadrados y en comparar los cuadrados medios, para así poder averiguar qué factores producen diferencias en el  $ARL_1$  más importantes que otros.

Fuente	Suma de Cuadrados	Gl	Cuadrado Medio	Razón-F	Valor-P
<b>EFFECTOS PRINCIPALES</b>					
A:tau	468034	9	52003,8	12155,07	0
B:n	30295,6	1	30295,6	7081,11	0
C:Copula	927117	2	463559	108349,49	0
D:desv	23,0868	5	4,61735	1,08	0,3702
E:Distrib	4,75839	2	2,37919	0,56	0,5736
<b>INTERACCIONES</b>					
AB	32794	9	3643,78	851,68	0
AC	56140,1	18	3118,89	728,99	0
AD	124,525	45	2,76721	0,65	0,966
AE	30,141	18	1,6745	0,39	0,9894
BC	85,0217	2	42,5109	9,94	0,0001
BD	4,6586	5	0,93172	0,22	0,955
BE	5,91335	2	2,95668	0,69	0,5013
CD	32,9253	10	3,29253	0,77	0,6584
CE	1,39245	4	0,348113	0,08	0,9881
DE	40,6303	10	4,06303	0,95	0,4864
RESIDUOS	4008,83	937	4,27836		
TOTAL (CORREGIDO)	1,52E+06	1079			

Cuadro 4.5.: Análisis de Varianza para  $ARL_1$  - Suma de Cuadrados Tipo III.

En el cuadro 4.5 donde se observa que tanto los efectos principales como las interacciones de orden 2, de las variables  $\tau$ , cópula y  $n$  son significativas (no se han buscado interacciones de mayor orden que 2).

Para validar los resultados obtenidos mediante el análisis de la varianza, se procederá a estudiar si se satisfacen los supuestos necesarios:

#### **Falta de Normalidad:**

Para observar si el modelo no presenta falta de normalidad, primero se debe realizar un estudio del gráfico de la probabilidad normal de los residuos.

En las figuras 4.19 y 4.20 no se aprecia claramente la falta de normalidad, por lo tanto, se procederá a hacer un análisis sobre los residuos del modelo y observar si siguen una distribución normal. Para ello, se van a estudiar el sesgo estandarizado y la curtosis estandarizada.

### 4.3. Resultados y discusión

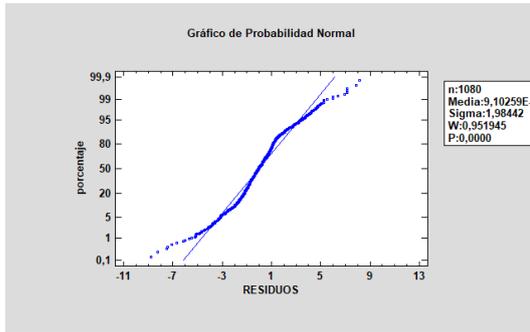


Figura 4.19.: Probabilidad Normal.

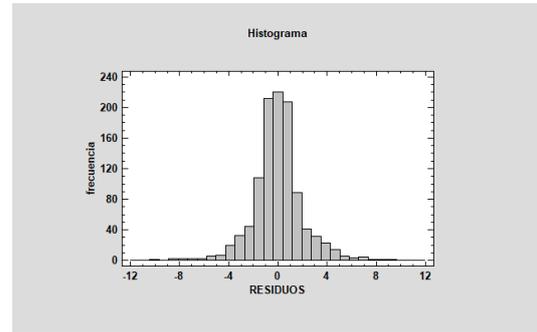


Figura 4.20.: Histograma de los residuos.

Sesgo Estandarizado	0,593885
Curtosis Estandarizada	22,2799

Los valores de estos estadísticos fuera del rango  $[-2,2]$  indican desviaciones significativas de la normalidad, lo que significaría falta de normalidad y, por tanto, invalidaría el modelo.

#### **Heterocedasticidad:**

Para detectar si el modelo elaborado presenta heterocedasticidad, es decir, la varianza de las perturbaciones del modelo no es constante. Por consiguiente, teniendo en cuenta que la estimación de la varianza residual del modelo es  $s^2 = CMR = \frac{\sum_j e_j^2}{N-I-1}$ , se ha realizado un análisis de los residuos al cuadrado respecto de las variables explicativas.

Fuente	Suma de Cuadrados	Gl	Cuadrado Medio	Razón-F	Valor-P
EFFECTOS PRINCIPALES					
A:n	359,872	1	359,872	6,01	0,0142
B:Copula	8521,81	2	4260,91	71,22	0
C:tau	16151,5	9	1794,61	30	0
D:desv	99,4318	5	19,8864	0,33	0,8935
E:Distrib	53,1904	2	26,5952	0,44	0,6413
RESIDUOS	63419,6	1060	59,8298		
TOTAL (CORREGIDO)	88605,4	1079			

Cuadro 4.6.: Análisis de Varianza para  $RESIDUOS^2$  - Suma de Cuadrados Tipo III.

Al estudiar el Cuadro 4.6, se detecta que tres variables explicativas ( $\tau$ , cópulas y  $n$ ) tienen un efecto significativo sobre  $RESIDUOS^2$  y, por tanto, sobre la varianza de los residuos. Entonces, el modelo elaborado presenta problemas de heterocedasticidad.

Dado que el modelo elaborado no satisface los supuestos de normalidad y de heterocedasticidad los p-valores presentados en el Cuadro 4.5 no pueden ser considerados fiables. Sin embargo, sigue siendo válida la información proporcionada por la columna suma de cuadrados para evaluar la contribución de cada factor en este estudio; teniendo en cuenta que cuanto mayor sea la suma de cuadrados asociados a cada factor, más diferentes serán las medias de estos.

#### 4. Experiencias computacionales

Observando el Cuadro 4.5, se confirman las suposiciones realizadas al observar los gráficos 4.4,4.5,4.6 y 4.8, dado que las variables con mayor valor en la suma de cuadrados respecto a la suma de cuadrados son las variables  $\tau$ ,  $n$  y cópula. En cambio, la suma de cuadrados de la variable desv presenta un valor ínfimo en comparación al resto.

#### Comparaciones 'Post-Hoc'

Tal y como se ha visto previamente, el modelo realizado no satisface los supuestos de normalidad y de homocedasticidad (igualdad de las varianzas). Por consiguiente, se hace uso del test de Games-Howell [12], debido a que se tiene el problema de heterocedasticidad, lo que hace que la prueba de Games-Howell sea un candidato ideal para realizar análisis 'post-hoc'. Como la prueba también es no paramétrica, en este caso, no es necesario el supuesto de normalidad.

La prueba de Games-Howell se utiliza para comparar todas las combinaciones posibles de diferencias de grupo cuando se viola el supuesto de homogeneidad de varianzas. Esta prueba post hoc proporciona intervalos de confianza para las diferencias entre las medias de los grupos y muestra si las diferencias son estadísticamente significativas.

La prueba se basa en la corrección de grados de libertad de Welch y utiliza la distribución de rangos de Tukey para calcular los p-valores. La prueba compara la diferencia entre cada par de medias con un ajuste adecuado para las pruebas múltiples. Por lo tanto, no es necesario aplicar correcciones adicionales del p-valor.

Estas pruebas 'Post-Hoc', que se van a mostrar a continuación, se van a ejecutar mediante el programa Jamovi [13] [14] en su versión online.

En las tablas 4.7,4.9 y 4.8 se aclara cómo afecta cada factor al  $ARL_1$ , es decir, cuál es la rapidez del gráfico para detectar que el proceso se ha salido de control. Así, para el factor  $n$  tamaño de muestral, se halla que, cuanto mayor sea el tamaño muestral, menor  $ARL_1$  se tiene y, por lo tanto, mayor eficiencia presenta el gráfico de control. Y esto verifica lo que ya se había intuido al observar los diagramas de Box-Whisker. Por otro lado, teniendo en cuenta que la variable cópula influye sobre el desempeño del gráfico, sobre el valor de  $ARL_1$ . Con el objetivo de analizar el efecto de esta variable se estudia la tabla 4.7, en la que se muestra la prueba de Games-Howell para el factor cópulas, y en este se aprecia que hay diferencias significativas entre el  $ARL_1$  obtenido mediante la cópula de Clayton y las de Frank y Gumbel. Sin embargo, entre las cópulas de Frank y la de Gumbel no se aprecian diferencias significativas.

Dado que mediante el análisis ANOVA se ha podido notar, que el valor  $\tau$  a detectar influye en el valor de  $ARL_1$ , es decir, en el desempeño del gráfico en la tabla 4.9 se muestra el resultado del test de Games-Howell con el objetivo de ver en qué valores de la  $\tau$  a detectar se aprecian diferencias significativas en el valor de  $ARL_1$ . Tal y como se ha podido intuir mediante los diagramas de Box-Whisker, los valores de  $\tau$  a detectar que presentan diferencias

### 4.3. Resultados y discusión

Prueba Post-Hoc de Games-Howell – ARL1

		Gumbel	Frank	Clayton
Gumbel	Diferencia de medias	—	0.980	-61.7***
	valor t	—	0.492	-38.0
	gl	—	718	558
	valor p	—	0.875	< .001
Frank	Diferencia de medias		—	-62.6***
	valor t		—	-39.2
	gl		—	566
	valor p		—	< .001
Clayton	Diferencia de medias			—
	valor t			—
	gl			—
	valor p			—

Nota. \* p < .05, \*\* p < .01, \*\*\* p < .001

Cuadro 4.7.: Prueba Post-Hoc de Games-Howell ARL1 y Copulas.

significativas respecto el resto son  $-0,1$  y  $0,1$ .

Prueba Post-Hoc de Games-Howell – ARL1

		10	30
10	Diferencia de medias	—	10.6***
	valor p	—	< .001
30	Diferencia de medias		—
	valor p		—

Nota. \* p < .05, \*\* p < .01, \*\*\* p < .001

Cuadro 4.8.: Prueba Post-Hoc de Games-Howell ARL1 y tamaño muestral.

Prueba Post-Hoc de Games-Howell – ARL1

		-0.9	-0.7	-0.5	-0.3	-0.1	0.1	0.3	0.5	0.7	0.9
-0.9	Diferencia de medias	—	-1.78	-5.15	-13.67	-58.4***	-52.18***	-11.63	-3.72	-0.859	0.277
	valor p	—	1.000	0.975	0.080	< .001	< .001	0.253	0.998	1.000	1.000
-0.7	Diferencia de medias		—	-3.37	-11.89	-56.6***	-50.40***	-9.85	-1.94	0.921	2.057
	valor p		—	0.999	0.239	< .001	< .001	0.530	1.000	1.000	1.000
-0.5	Diferencia de medias			—	-8.52	-53.2***	-47.02***	-6.48	1.43	4.293	5.430
	valor p			—	0.747	< .001	< .001	0.944	1.000	0.994	0.963
-0.3	Diferencia de medias				—	-44.7***	-38.51***	2.04	9.95	12.811	13.947
	valor p				—	< .001	< .001	1.000	0.541	0.150	0.062
-0.1	Diferencia de medias					—	6.21	46.75***	54.66***	57.523***	58.660***
	valor p					—	0.608	< .001	< .001	< .001	< .001
0.1	Diferencia de medias						—	40.54***	48.45***	51.317***	52.453***
	valor p						—	< .001	< .001	< .001	< .001
0.3	Diferencia de medias							—	7.91	10.775	11.911
	valor p							—	0.829	0.390	0.211
0.5	Diferencia de medias								—	2.863	4.000
	valor p								—	1.000	0.996
0.7	Diferencia de medias									—	1.136
	valor p									—	1.000
0.9	Diferencia de medias										—
	valor p										—

Nota. \* p < .05, \*\* p < .01, \*\*\* p < .001

Cuadro 4.9.: Prueba Post-Hoc de Games-Howell ARL1 y tau.

## 5. Ejemplo práctico

En esta sección se va a ilustrar la utilidad de estos gráficos mediante un ejemplo práctico. Para el cual se va a hacer uso de los resultados que se presentan en la tabla 4.4.

Una empresa fabrica una pieza cuya característica de calidad principal es el diámetro del vástago, esta característica de calidad debe ser independiente de la longitud de la pieza.

- $X$  := Diámetro del vástago.
- $Y$  := Longitud de la pieza.

Cuando el proceso de fabricación está bajo control, la longitud de esta pieza se distribuye según una normal de media 9.5 cm y desviación típica 0.22,  $N(9.5, 0.22)$ cm y el diámetro del vástago se distribuye según una normal  $N(2, 0.5)$ cm. El tamaño de muestreo será  $n = 20$  piezas. Para el diseño del gráfico se considerará una probabilidad de error de tipo I=  $0,0027 = \alpha$ , y se va a utilizar la cópula de Gumbel.

Teniendo en cuenta que los límites de control solamente dependen del valor de  $\alpha$  y el tamaño muestral, ya se pueden determinar, debido a que ya es sabido que se está trabajando con un  $\alpha = 0,0027$  y un tamaño de muestra  $n = 20$ . Los límites de control son los siguientes (de acuerdo a las ecuaciones (3.1) y (3.2)):

$$F(LCL) = P(\hat{\tau} \leq LCL \mid \tau = 0) = \alpha \implies LCL = F^{-1}(\alpha) = -0,4421053$$

para el gráfico unilateral diseñado para detectar correlaciones negativas, y

$$F(UCL) = P(\hat{\tau} \leq UCL \mid \tau = 0) = 1 - \alpha \implies UCL = F^{-1}(1 - \alpha) = 0,4421053$$

para el gráfico unilateral diseñado para detectar correlaciones positivas.

### 5.1. Proceso bajo control

El primer paso consiste en estudiar el comportamiento del gráfico de control cuando el proceso de fabricación se encuentra bajo control, es decir, cuando el diámetro y la longitud son independientes ( $\tau = 0$ ).

Como ya se han establecido los límites de control, se va a proceder a la monitorización de 10000 pruebas independientes, en los gráficos unilaterales 5.1 y 5.2 se muestran las primeras 200 muestras.

Cada punto que aparece representado en la figura 5.1 y en la figura 5.2 representa un valor del coeficiente de correlación de Kendall calculado para una muestra de 20 piezas, es decir, de 20 valores de  $X$  e  $Y$ . Este gráfico se realiza bajo el supuesto de que el proceso de fabricación se encuentra bajo control, es decir, que  $X$  e  $Y$  son independientes y, por tanto  $\tau = 0$ . Bajo

## 5. Ejemplo práctico

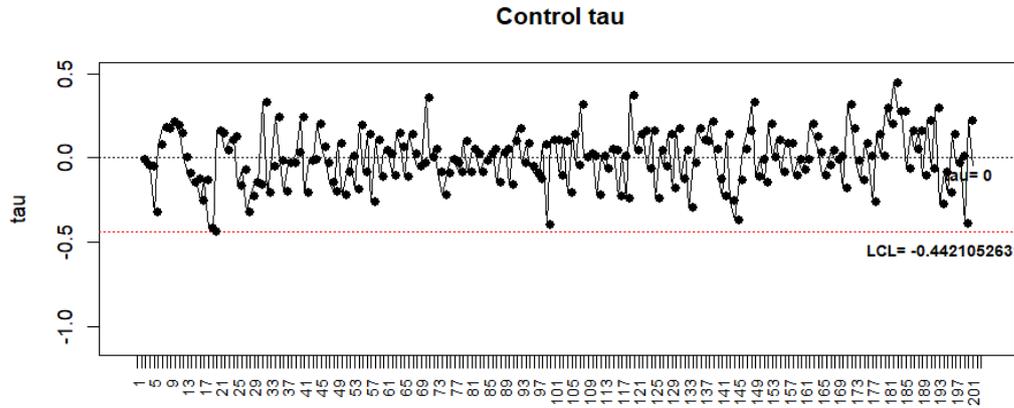


Figura 5.1.: Gráfico de control para la detección de correlaciones negativas. Situación bajo control ( $\tau = 0$ ). Primeras 200 muestras.

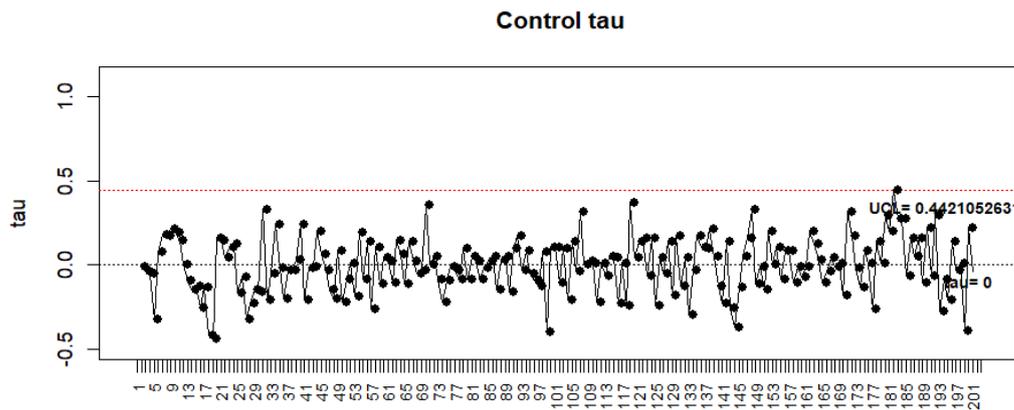


Figura 5.2.: Gráfico de control para la detección de correlaciones positivas. Situación bajo control ( $\tau = 0$ ). Primeras 200 muestras.

este supuesto, se puede apreciar que casi en su totalidad los puntos se encuentran dentro de los límites de control establecidos.

A continuación, se va a calcular el  $ARL_0$  esperado y el observado de cada uno de estos gráficos cuando el proceso se encuentra bajo control:

$$ARL_0 \text{ esperado} := \frac{1}{\alpha} = \frac{1}{0,0027} = 370,3704$$

$$ARL_0 \text{ observado Gráfico } UCL := \frac{10000}{27} = 370,3704$$

$$ARL_0 \text{ observado Gráfico } LCL := \frac{10000}{13} = 769,2308$$

Ambos son muy elevados, por lo tanto, tal y como se mencionó al inicio de este trabajo, es lo deseado debido a que este hecho provoca que hay pocas falsas alarmas.

## 5.2. Proceso fuera de control

### 5.2.1. Coeficiente de correlación a detectar positivo

A continuación, se procede a estudiar el comportamiento del gráfico de control cuando el proceso de fabricación se encuentra fuera control, es decir, cuando el diámetro y la longitud no son independientes ( $\tau \neq 0$ ). En este caso particular, se decide que el valor a detectar de la medida de correlación  $\tau$  como fuera de control va a ser  $\tau = 0,6$ . Por tanto, se va a proceder a la monitorización de 10000 muestras con un coeficiente de correlación de Kendall  $\tau = 0,6$ , en el gráfico 5.3 se van a ilustrar las 200 primeras muestras para mejor visibilidad.

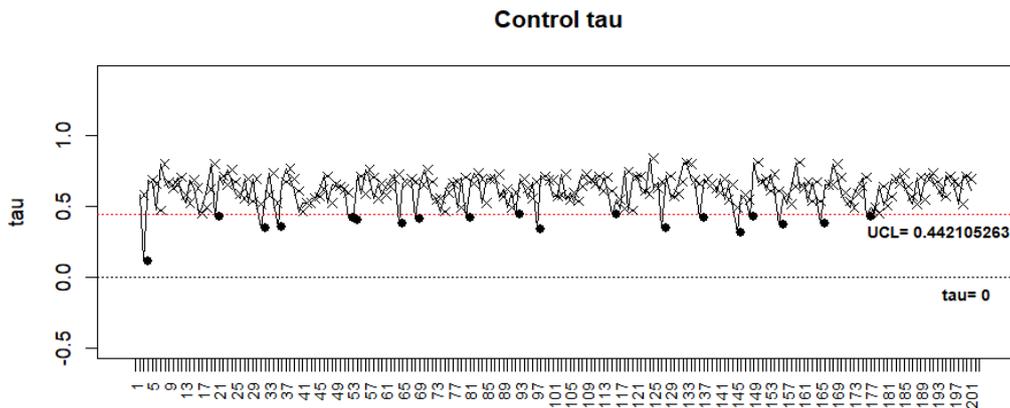


Figura 5.3.: Gráfico de control proceso fuera de control ( $\tau = 0,6$ ).

Al observar la figura 5.3, se puede apreciar gráficamente que gran parte de los puntos (representados por cruces) se encuentran fuera de la banda delimitada por el límite de control superior. Ahora, la medida de desempeño o velocidad de este gráfico para detectar este proceso fuera de control nos indica que el gráfico de control logra detectar la correlación entre el diámetro y la longitud con tan solo una prueba, debido a que el  $ARL_1 \text{ observado} = 1,105217$ . Mientras que el  $ARL_1$  esperado es 1,104606, que ha sido obtenido mediante simulación haciendo uso del código 4.8 expuesto en la sección 4.1.2.

Por consiguiente, tanto el  $ARL_1$  observado como el esperado determinan que solamente es necesaria una prueba para detectar que este proceso está fuera de control.

## 5. Ejemplo práctico

### 5.2.2. Coeficiente de correlación a detectar negativo

En este apartado se pretende estudiar el comportamiento del gráfico de control cuando se pretende detectar el caso en el que el diámetro y la longitud se encuentran correlacionadas negativamente. Para ello, se procede a tomar como valor a detectar de la medida de correlación  $\tau$  como fuera de control  $\tau = 0,6$ . Por tanto, se va a proceder a la monitorización de 10000 muestras con un coeficiente de correlación de Kendall  $\tau = -0,6$ , en el gráfico 5.4 se van a ilustrar las 200 primeras muestras para mejor visibilidad.

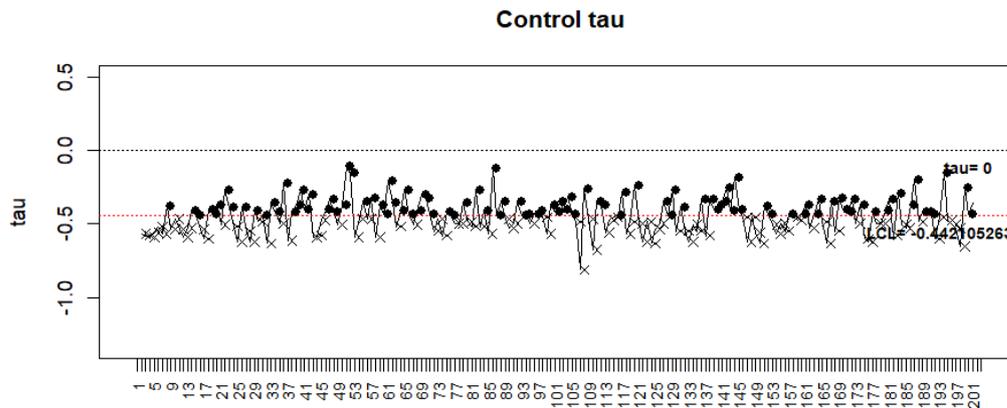


Figura 5.4.: Gráfico de control proceso fuera de control ( $\tau = -0,6$ ).

Al observar la figura 5.4, se puede observar gráficamente que gran parte de los puntos (representados por cruces) se encuentran fuera de la banda delimitada por el límite de control inferior. Ahora, la medida de desempeño o velocidad de este gráfico para detectar este proceso fuera de control nos indica que el gráfico de control logra detectar la correlación entre el diámetro y la longitud con tan dos pruebas, debido a que el  $ARL_1$  observado = 1,921968. Mientras que el  $ARL_1$  esperado es 1,94742, que ha sido obtenido mediante simulación haciendo uso del código 4.8 expuesto en la sección 4.1.2.

Por consiguiente, tanto el  $ARL_1$  observado como el esperado determinan que solamente son necesarias dos pruebas para detectar que este proceso está fuera de control.

## 6. Conclusiones y trabajo futuro

En este trabajo se ha llevado a cabo el diseño e implementación de un gráfico de control para monitorización de tipo Shewhart del coeficiente de correlación de Kendall, es decir, se ha diseñado e implementado un gráfico de control que detecte cuándo dos variables dejan de ser independientes para este coeficiente.

Adicionalmente, se ha realizado un diseño de experimentos para observar qué variables son las que más efecto presentan sobre la medida de desempeño  $ARL_1$ . Para ello, en primer lugar, se diseñan y ejecutan experiencias computacionales mediante el uso del lenguaje R, basadas en simulación Montecarlo, con el objetivo de evaluar la medida de desempeño del gráfico ( $ARL_1$ ) bajo diferentes niveles de los distintos factores. Los resultados obtenidos han sido analizados mediante ANOVA y pruebas 'post-hoc' de Games-Howell.

En último lugar, se ha realizado un ejemplo práctico para ilustrar la utilidad de este trabajo. Las principales conclusiones que pueden extraerse del trabajo realizado son las siguientes:

- Los objetivos de este trabajo han sido cumplidos, dado que se ha conseguido diseñar e implementar un gráfico de control para el coeficiente de correlación de Kendall. Además de haberlo evaluado y validado a partir de experiencias numéricas que han permitido saber qué factores afectan a la medida de desempeño. Es más, en primer lugar, se ha introducido la teoría y los conceptos necesarios para una buena comprensión del trabajo.
- Se ha proporcionado un método para poder calcular la medida de desempeño del gráfico de control, tanto en el caso en el que el proceso está bajo control ( $\tau = 0$ ) como en el caso en el que el gráfico está fuera de control ( $\tau \neq 0$ ), haciendo uso de simulación y de las cópulas.
- Se ha proporcionado una forma de calcular los límites de control sin necesidad de conocer la función de distribución del coeficiente de correlación de Kendall,  $\tau$ , dado que ya se ha visto que el cálculo de esta presenta gran dificultad.
- Se ha realizado un estudio del comportamiento del  $ARL_1$  (medida de desempeño del gráfico de control) con el objetivo de conocer qué factores presentan una influencia sobre este. De este estudio se deduce que los factores que presentan un cierto efecto sobre el valor del  $ARL_1$  son:
  - Las cópulas, provocando un aumento sobre el  $ARL_1$  cuando se utiliza la cópula de Clayton.
  - El tamaño muestral; a medida que aumenta el tamaño muestral, disminuye el valor del  $ARL_1$ .
  - El coeficiente de correlación a detectar  $\tau$ , provocando que el  $ARL_1$  disminuya con el aumento de estos factores.

## 6. Conclusiones y trabajo futuro

Además de la influencia de estos factores sobre el  $ARL_1$ , también tienen influencia sobre esta medida de desempeño la interacción entre la cópula utilizada y el coeficiente de correlación a detectar. Y de igual manera, la interacción entre el tamaño muestral y el coeficiente de correlación a detectar.

### 6.1. Trabajo futuro

Tras la investigación realizada a lo largo de este trabajo y las conclusiones obtenidas, quedan aún ciertos aspectos por investigar en el futuro. Por consiguiente, este trabajo se podría ampliar a partir de las siguientes propuestas en investigaciones futuras:

- En primer lugar, se deberían ampliar los casos de estudio que han sido expuestos en el apartado 4.2 explorando diferentes distribuciones de  $X$  e  $Y$  con diferentes medias y desviaciones típicas, utilizando distintos tamaños muestrales y hacer uso de más cópulas. De la misma manera, también resultaría interesante introducir diferentes probabilidades de error tipo I ( $\alpha$ ); todo esto con el objetivo de descubrir su efecto sobre la medida de desempeño del gráfico.
- En este trabajo solamente se ha diseñado e implementado un gráfico de control clásico o tipo Shewhart. Por ello, en líneas futuras resultaría interesante investigar en la construcción de gráficos de control con memoria, como por ejemplo, los gráficos CUMSUM o EWMA. La ventaja de estos gráficos de control sería que los gráficos proporcionarían valores de  $ARL_1$  más pequeños que el gráfico de Shewhart y, en consecuencia, los gráficos obtenidos serían más eficientes. Este hecho es debido a que los valores calculados y graficados por estos gráficos no se basan solamente en cada muestra extraída, sino en la acumulación de toda la información disponible hasta ese momento.

## Bibliografía

- [1] Maite Santurtún, Ana Santurtún, and María T. Zarrabeitia. ¿afecta el medio a los suicidios que se cometen en españa? análisis descriptivo del patrón temporoespacial. *Revista de Psiquiatría y Salud Mental*, 11(4):192–198, October 2018.
- [2] David Badia Folguera. *Metodología de los mapas de concordancia para la estratificación de variables cuantitativas: Aplicación a la asignatura de medidas electrónicas*. PhD thesis, Departamento de Comunicaciones Teoría del Señal. Ingeniería y Arquitectura La Salle, Septiembre 2012.
- [3] Douglas C. Montgomery. *Introduction to statistical quality control*. Wiley, New York, NY [u.a.], 3. ed edition, 1997.
- [4] Wolfe D.A. Hollander, M. and E. Chicken.
- [5] Rivest L.-P. Genest, C. Statistical inference procedures for bivariate archimedean copulas. *J. Amer. Statist. Assoc.* 88, page 1034–1043, 1993.
- [6] 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.
- [7] D.J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures(5th Edition)*. Chapman and Hall/CRC, 2011.
- [8] Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric Statistical Methods*. Wiley Series in Probability and Statistics. Wiley-Blackwell, Hoboken, NJ, 3 edition, November 2013.
- [9] Dennis O Wackerly, William Mendenhall, and Richard L Scheaffer. *Estadística Matemática con Aplicaciones*. Cengage Learning Editores S.A. de C.V, Mexico City, Mexico, 7 edition, August 2009.
- [10] B. Wheeler. *SuppDists: Supplementary distributions*. distributions. R package version 1.8., 2009.
- [11] Douglas C. Montgomery. *Design and Analysis of Experiments*. Design and Analysis of Experiments. John Wiley Sons, Inc., New York., 2008.
- [12] Paul A. Games and John F. Howell. Pairwise multiple comparison procedures with unequal n’s and/or variances: A monte carlo study. *Journal of Educational Statistics*, 1(2):113–125, 1976.
- [13] R Core Team (2021). *R:A Language and environment for statistical computing(version 4.1)[Computer software]*. Retrieved from <https://cran.rproject.org>. (RpackagesretrievedfromMRANSanpshot2022-01-01.

## *Bibliografía*

- [14] The jamovi project (2022). *jamovi.(Version 2.3)[Computer Software]*. Retrieved from <https://www.jamovi.org>.

## **A. Anexos**

### **A.1. Código R**

El código de R que se ha utilizado para obtener todos los resultados numéricos de este trabajo se encuentra adjunto en el siguiente enlace:

<https://drive.google.com/file/d/1HCIWu53d52tKcmTVr7cm-nLuIZlqFHtU/view?usp=sharing>

### **A.2. Archivo Excel con los resultados numéricos**

Los resultados numéricos que se han obtenido al ejecutar el código presentado en el apartado 4.1.2, en el cual se han ido almacenando cada uno de los distintos casos en un archivo Excel al que se puede acceder en mediante el siguiente enlace: <https://docs.google.com/spreadsheets/d/1RPQzyvYm0KoYGSfrJIItld7tMVdWc12uS/edit?usp=sharing&oid=115656105769555660&rtpof=true&sd=true>

### **A.3. Código Mathematica**

Para realizar los gráficos de las funciones de densidad para las diferentes cópulas se ha hecho uso del software Mathematica. El código que se ha realizado para ello se encuentra adjunto en el siguiente enlace:

<https://drive.google.com/file/d/1s11NYv-GSa1gN15p0CBJdYWsQj3sSyks/view?usp=sharing>