ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

COMPARACIONES DE PERFILES DE SALUD DE INDIVIDUOS SANOS EN DIFERENTES SECTORES ECONÓMICOS

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

PROYECTO DE INVESTIGACIÓN

CRISTIAN RAÚL GUATEMAL AGUILAR
cristian.guatemal@epn.edu.ec

DIRECTOR: Ph.D. CARLOS ALBERTO ALMEIDA RODRIGUEZ
carlos.almeidar@epn.edu.ec

Quito, julio 2019

DECLARACIÓN

Yo CRISTIAN RAÚL GUATEMAL AGUILAR, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual, correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su reglamento y por la normatividad institucional vigente.

Cristian Raúl Guatemal Aguilar

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por CRISTIAN RAÚL GUATEMAL AGUILAR, bajo mi supervisión.

Ph.D. Carlos Alberto Almeida Rodríguez
Director del Proyecto

AGRADECIMIENTOS

Al finalizar el presente trabajo de investigación, es inexplicable el profundo sentimiento de felicidad y gratitud que llena mi ser. No existen palabras que puedan expresar cada una de las emociones que invaden mi alma y corazón.

Gracias a Dios por todas sus bendiciones, por la vida de mis padres, por mis amigos, familiares y cada una de las personas que de manera directa o indirecta han compartido un poco de su tiempo conmigo. Gracias Dios por permitirme culminar con éxito esta etapa de mi vida.

Es un verdadero placer utilizar este espacio y expresar mis más profundos agradecimientos, a cada una de las personas que me han brindado su cariño, confianza y amistad verdadera e incondicional durante todo tiempo compartido dentro y fueras de las aulas de clase.

En primer lugar agradezco a la Escuela Politécnica Nacional por haberme brindado la oportunidad de ser parte de ella, así como también a los diferentes docentes que me brindaron sus conocimientos, apoyo y consejo para seguir adelante día a día.

Mis más sinceros agradecimientos a mi director de tesis, el Ph.D. Carlos Almeida Rodríguez por su labor continua, sus consejos a nivel académico, sus motivaciones y por su tiempo infinito para guiarme con toda la paciencia del mundo durante el desarrollo de mi proyecto de titulación.

A mis padres por ser los principales promotores de mis sueños, gracias por su confianza, amor, por su sacrifico en todos estos años y sobre todo gracias por creer en mí. Gracias a mi padre Jorge por inculcarme desde pequeño el valor de esfuerzo, de la constancia y cada uno de los valores que hoy en día son parte de mí ser. Gracias a mi madre Rosita por su firmeza en los momentos más difíciles de mi vida, por no permitirme que me diera por vencido en las adversidades.

A Jhonny mi gran mejor amigo y hermano, quién ha sido mi mano derecha durante todo este tiempo, te agradezco por tu desinteresada ayuda, por echarme una mano cuando siempre lo necesitaba, gracias por tu amistad, por tu convicción y carisma, por tantas anécdotas compartidas dentro y fuera de la universidad.

A Priscilla mi gran amiga, gracias por cada uno de los consejos que me diste cuando sentía que el mundo me destrozaba, por tu infinita bondad y cariño; no sabes cuán agradecido estoy con Dios por cada risa, por cada lágrima y cada momento compartido contigo. El tiempo sigue pasando, y ahí estás, cerca de mí ofreciendo lo mejor de tu corazón, gracias por tus esfuerzos de mantener viva nuestra amistad y sobre todo gracias por tu ánimo en las fases finales de este proyecto.

A Paola, mi querida amiga. Tengo una gratitud enorme contigo amiga mía. Gracias por ser una persona incondicional, por tu cariño desinteresado y gran generosidad. Sin tu apoyo en las adversidades y dificultades que presentaba la universidad no hubiera sido posible culminar mi carrera.

A Maru. Quizás no hemos compartido grandes anécdotas por la limitación en nuestro tiempo, pero eso no quita que exista algún grato recuerdo de risas y ocurrencias. Gracias por tu consejo y palabras de ánimo cuando lo necesitaba, por tu forma de ser, por tu transparencia y cariño.

A Karen, mi loca amiga. El tiempo y la vida misma son un misterio, ¿recuerdas cuando nos conocimos?, pues ahora estamos juntos en las fases finales de nuestra carrera y es preciso decir gracias por los momentos de risas y más risas compartidos, por tus ocurrencias que hacían que nos olvidemos del estrés universitario. Amiga, gracias por tu forma de ser.

A Migue. Recuerdo la anécdota de la firma y no puedo creer como todo ha cambiado. Gracias por tu ayuda y amistad desinteresada de los últimos meses.

DEDICATORIA

A mis padres, les dedico este trabajo por su constante espíritu de lucha y superación, por su amor incondicional. Gracias por ser los ángeles que cuidaron y cuidan de mi.

A mi hermana adorada Arlet, tu sonrisa me ha cambiado la vida. No sabes la alegría infinita que me da de tenerte a mi lado.

Índice general

CAPÍTULOS	PÁGINA
Lista de Figuras	XI
Lista de Tablas	XV
Lista de Símbolos	XVI
Resumen	XVIII
Abstract	XIX
1. Introducción	**1**
1.1. Antecedentes	2
1.2. Justificación	3
1.3. Objetivos	6
1.3.1. Objetivo General	6
1.3.2. Objetivos Específicos	6
1.4. Hipótesis	7
1.5. Metodología	7
1.6. Estructura del Trabajo	8
2. Marco Teórico	**10**
2.1. Nociones Básicas	11
2.1.1. Prueba de Hipótesis	11
2.1.2. p-valor	13
2.1.3. Test Consistente	14
2.1.4. Corrección de Continuidad	15
2.1.5. Estadístico de Orden	16
2.2. Métodos No Paramétricos	17
2.2.1. Estadístico Rango-Orden (Estadístico de rango)	19

2.2.2. Ties (Empates) . . . 22
2.2.3. Problema de una Muestra o Muestras Relacionadas . . . 23
2.2.4. Sign Test (Prueba de Signo) . . . 23
2.2.5. Wilcoxon Signed-Rank Test (Test de Signo de Rango de Wilcoxon) . . . 28
2.2.6. Problema General de dos Muestras . . . 35
2.2.6.1. Modelo de Ubicación . . . 36
2.2.6.2. Modelo de Escala . . . 37
2.2.6.3. Modelo General Ubicación-Escala . . . 38
2.2.6.4. De Wald–Wolfowitz Runs Test . . . 39
2.2.6.5. Test U de Mann-Whitney . . . 42
2.2.7. Estadísticos Lineales de Rango y la Generalización al Problema de dos Muestras . . . 46
2.2.7.1. Propiedades del Estadístico Lineal de Rango . . . 48
2.2.8. Test de Rango Lineal para el Problema de Ubicación . . . 56
2.2.8.1. El Wilcoxon Rank-Sum Test (Test de Suma de Rangos de Wilcoxon) . . . 57
2.2.9. Otros Test de Rango Lineal para el Problema de Ubicación . . . 61
2.2.9.1. Terry–Hoeffding (Normal Scores) Test . . . 62
2.2.9.2. Van der Waerden Test . . . 63
2.2.10. Test de Rango Lineal para el Problema de Escala . . . 63
2.2.10.1. The Mood Test . . . 65
2.2.10.2. The Freund–Ansari–Bradley Test . . . 68
2.2.10.3. The Siegel-Tukey Test . . . 71
2.2.10.4. The Klotz Normal-Scores Test . . . 72
2.2.11. Análisis de Múltiples Muestras Independientes . . . 73
2.2.11.1. Test de la Mediana . . . 74
2.2.11.2. Test Anova Unidireccional de Kruskal-Wallis y Comparaciones Múltiples . . . 77
2.3. Eficiencia Relativa Asintótica . . . 83
2.4. Análisis de Componentes Principales (ACP) . . . 86
2.5. Construcción de Indicadores Compuestos . . . 91
2.5.1. Indicadores Compuestos (CI) . . . 92

3. Construcción de la Base de Datos y del Índice de Salud 102
3.1. Análisis y Depuración de Datos . . . 103
3.2. Construcción de los Índices de Salud . . . 105

4. Aplicación y Resultados **116**
4.1. Análisis Descriptivo 116
4.2. Test Anova 120
4.3. Test de Kruskal-Wallis 128

5. Conclusiones y Recomendaciones **141**
5.1. Conclusiones 141
5.2. Recomendaciones 143

Referencias Bibliográficas **150**

Anexo **151**

A. Anexo I: Depuración de Datos y ACP **152**
A.1. Variables codificadas y porcentaje de datos perdidos por variable . . 152
A.2. Variables que tienen a lo mucho el 10% de datos perdidos. 154
A.3. Variables utilizadas en el ACP. 154
A.4. Datos en relación al mejor individuo sano. 154
A.5. Componentes Principales y Proyección de los individuos 156

B. Anexo II: Aplicación y Resultados **157**
B.1. Test de Normalidad. 157
B.2. Accidentes de Trabajo por Actividad y Año. 160
B.3. Aplicación a una muestra que proviene (por medio de simulación) de una población normal. 161

C. Anexo III: Enlace Web **165**
C.1. Dirección web del aplicativo. 165

Índice de Figuras

2.1. Si $F_X(x)$ es la cdf de $\mathscr{X}$ e $F_Y(x)$ es la cdf de $\mathscr{Y}$ entonces a) Para $\theta < 0$ (Distribución Normal) y b) Para $\theta > 0$ (Distribución Exponencial) Tanto en el caso a) como en b), las dos poblaciones tienen la misma forma y variabilidad. La única diferencia está en que la cdf de la población $\mathscr{Y}$ está trasladada a la derecha ($\theta > 0$) o a la izquierda ($\theta < 0$) en comparación con la cdf de la población $\mathscr{X}$. Generalmente el parámetro de ubicación es igual a la diferencia entre cuantiles del mismo orden (particularmente medias o medianas). 36
2.2. Si $F_X(x)$ es la cdf de $\mathscr{X}$ e $F_Y(x)$ es la cdf de $\mathscr{Y}$, entonces a) Si $\theta > 1$ (Distribución Normal) y b) Si $\theta < 1$ (Distribución Exponencial) La cdf de la población $\mathscr{Y}$ es la misma que la cdf de la población $\mathscr{X}$ pero con una escala comprimida o expandida según $\theta > 1$ o $\theta < 1$, respectivamente. 37
2.3. Frontera de Rendimiento del DEA. 96

3.1. Porcentaje de datos perdidos por tipo de codificación. 104
3.2. Proporción de variabilidad explicada para cada componente (Índice ACP). 108
3.3. Gráfica de pesos de los dos primeros componentes (Índice ACP). . . 110
3.4. Proporción de variabilidad explicada para cada componente (Índice CI). 111
3.5. Gráfica de pesos de los dos primeros componentes (Índice CI). 113

4.1. Histograma y Función de Densidad para la primera muestra (Índice ACP). 117
4.2. Histograma y Función de Densidad para la segunda muestra (Índice ACP). 117
4.3. Histograma y Función de Densidad para la tercera muestra (Índice ACP). 118

4.4. Histograma y Función de Densidad para la primera muestra (Índice CI). 118
4.5. Histograma y Función de Densidad para la segunda muestra (Índice CI). 119
4.6. Histograma y Función de Densidad para la tercera muestra (Índice CI).119
4.7. Gráfica de Probabilidad Normal-Muestra 1 (Índice ACP). 120
4.8. Gráfica de Probabilidad Normal-Muestra 2 (Índice ACP). 121
4.9. Gráfica de Probabilidad Normal-Muestra 3 (Índice ACP). 121
4.10. Gráfica de Probabilidad Normal-Muestra 1 (Índice CI). 122
4.11. Gráfica de Probabilidad Normal-Muestra 2 (Índice CI). 122
4.12. Gráfica de Probabilidad Normal-Muestra 3 (Índice CI). 123
4.13. Diagrama de Caja y Bigote paras las muestras provenientes del Índice ACP. 125
4.14. Diagrama de Caja y Bigote paras las muestras provenientes del Índice CI. 125
4.15. Gráfica de Densidad de las muestras provenientes del Índice ACP. . 134
4.16. Gráfica de Densidad de las muestras provenientes del Índice CI. . . . 134

Índice de Tablas

2.1. Error tipo I y tipo II 12
2.2. Pruebas paramétricas y su contraparte no paramétrica 19
2.3. Regiones de rechazo unilaterales para el Sign Test 24
2.4. Regiones de rechazo para las alternativas de H_0 del Wilcoxon Signed-Rank Test. 33
2.5. Regiones de rechazo y p-valores aproximados para las alternativas de H_0 del Test T^+ 33
2.6. Regiones de rechazo para las alternativas de H_0 del Test Mann-Whitney 45
2.7. Regiones de rechazo y p-valores exactos para las alternativas de H_0 del Wilcoxon Rank-Sum Test 61
2.8. Regiones de rechazo y p-valores aproximados para las alternativas de H_0 del Wilcoxon Rank-Sum Test 61
2.9. Resumen de los test no paramétricos utilizados en el problema general que compara dos muestras independientes. 72
2.10. Valores de la $ARE(K_n, T_n^*)$, $ARE(T^+, T_n^*)$ y $ARE(K_n, T_n^+)$ 84
2.11. Valores de la $ARE(U, c_1)$ para cinco distribuciones de probabilidad . 86
2.12. Ventajas y Desventajas de los Indicadores Compuestos 92
2.13. Compatibilidad entre métodos de agregación y ponderación. 95
2.14. Tipo de restricciones "pie-share" 100

3.1. Número de variables por tipo de codificación. 104
3.2. Actividades económicas y número de observaciones. 105
3.3. Matriz de correlaciones de los datos depurados de la sección 3.1. . . 106
3.4. Matriz de correlaciones de los datos depurados de la sección 3.1 en relación al mejor individuo sano dado en (A.4). 107
3.5. Importancia de las Componentes Principales para el Índice ACP. . . 108
3.6. Pesos de los primeros dos componentes principales (índice ACP). . . 109

3.7. Proyección de los 20 primeros individuos en las dos primeras componentes principales (Índice ACP) 110
3.8. Importancia de las Componentes Principales para el Índice CI. 111
3.9. Pesos de los seis primeros componentes principales (índice CI) 112
3.10. Proyección de los 20 primeros individuos en las seis primeras componentes principales (Índice CI). 113
3.11. Transformación mediante Min-Max de la proyección de los 20 primeros individuos de las seis primeras componentes principales (Índice CI). 114
3.12. Primeros 1200 valores del indicador CI. 115

4.1. Resumen estadístico de las muestras (Índice ACP). 116
4.2. Resumen estadístico de las muestras (Índice CI). 116
4.3. Kolmogorov-Smirnov Test para la Muestra 1 (Índice ACP). 123
4.4. Kolmogorov-Smirnov Test para la Muestra 2 (Índice ACP). 123
4.5. Kolmogorov-Smirnov Test para la Muestra 3 (Índice ACP). 124
4.6. Kolmogorov-Smirnov Test para la Muestra 1 (Índice CI). 124
4.7. Kolmogorov-Smirnov Test para la Muestra 2 (Índice CI). 124
4.8. Kolmogorov-Smirnov Test para la Muestra 3 (Índice CI). 124
4.9. Levene's Test for Homogeneity of Variance (índice ACP) 126
4.10. Levene's Test for Homogeneity of Variance (índice CI) 126
4.11. Análisis de Varianza de un factor (Índice ACP). 127
4.12. Análisis de Varianza de un factor (Índice CI). 127
4.13. Comparaciones múltiples dadas por el proceso post hoc para el Test ANOVA. 128
4.14. Comparaciones múltiples dadas por el proceso post hoc para el Test ANOVA. 128
4.15. Test de Kruskal-Wallis (Índice ACP). 129
4.16. Test de Kruskal-Wallis (Índice CI). 129
4.17. Resumen estadísticos de las muestras dadas por el proceso post hoc (Índice ACP). 129
4.18. Comparaciones múltiples dadas por el proceso post hoc para el Test de Kruskal-Wallis (Índice ACP). 130
4.19. Resumen estadísticos de las muestras dadas por el proceso post hoc (Índice CI). 130
4.20. Comparaciones múltiples dadas por el proceso post hoc para el Test de Kruskal-Wallis (Índice CI). 130

4.21. Contraste post hoc dados por el Test U (Índice ACP). 131
4.22. Contraste post hoc dados por el Test U (Índice CI). 131
4.23. Análisis de Varianza de un factor aplicado a los rangos de las muestras provenientes del Índice ACP. 132
4.24. Análisis de Varianza de un factor aplicado a los rangos de las muestras provenientes del índice CI 132
4.25. Test F de Fisher: Muestra 1 y Muestra 2 (Índice ACP). 135
4.26. Test F de Fisher: Muestra 2 y Muestra 3 (Índice ACP). 135
4.27. Test F de Fisher: Muestra 1 y Muestra 3 (Índice ACP). 135
4.28. Test F de Fisher: Muestra 1 y Muestra 2 (Índice CI). 135
4.29. Test F de Fisher: Muestra 2 y Muestra 3 (Índice CI). 135
4.30. Test F de Fisher: Muestra 1 y Muestra 3 (Índice CI). 136
4.31. Mood Test: Muestra 1 y Muestra 2 (índice ACP) 136
4.32. Mood Test: Muestra 2 y Muestra 3 (índice ACP) 136
4.33. Mood Test: Muestra 1 y Muestra 3 (índice ACP) 136
4.34. Freund–Ansari–Bradley test Muestra 1 y Muestra 2 (índice ACP). . . 136
4.35. Freund–Ansari–Bradley Test Muestra 2 y Muestra 3 (índice ACP). . . 136
4.36. Freund–Ansari–Bradley Test Muestra 1 y Muestra 3 (índice ACP) . . 137
4.37. Siegel-Tukey Test Muestra 1 y Muestra 2 (índice ACP). 137
4.38. Siegel-Tukey Test Muestra 2 y Muestra 3 (índice ACP). 137
4.39. Siegel-Tukey Test Muestra 1 y Muestra 3 (índice ACP). 137
4.40. Klotz Normal-Scores Test Muestra 1 y Muestra 2 (índice ACP). 137
4.41. Klotz Normal-Scores Test Muestra 2 y Muestra 3 (índice ACP). 137
4.42. Klotz Normal-Scores Test Muestra 1 y Muestra 3 (índice ACP). 137
4.43. Mood Test: Muestra 1 y Muestra 2 (índice CI). 138
4.44. Mood Test: Muestra 2 y Muestra 3 (índice CI). 138
4.45. Mood Test: Muestra 1 y Muestra 3 (índice CI). 138
4.46. Freund–Ansari–Bradley test Muestra 1 y Muestra 2 (índice CI). . . . 138
4.47. Freund–Ansari–Bradley Test Muestra 2 y Muestra 3 (índice CI). . . . 138
4.48. Freund–Ansari–Bradley Test Muestra 1 y Muestra 3 (índice CI). . . . 138
4.49. Siegel-Tukey Test Muestra 1 y Muestra 2 (índice CI). 139
4.50. Siegel-Tukey Test Muestra 2 y Muestra 3 (índice CI). 139
4.51. Siegel-Tukey Test Muestra 1 y Muestra 3 (índice CI). 139
4.52. Klotz Normal-Scores Test Muestra 1 y Muestra 2 (índice CI). 139
4.53. Klotz Normal-Scores Test Muestra 2 y Muestra 3 (índice CI). 139
4.54. Klotz Normal-Scores Test Muestra 1 y Muestra 3 (índice CI). 139

A.1. Variables codificadas y porcentaje de datos perdidos (NA). 152

B.1. Kolmogorov-Smirnov Test para la Muestra 1 (índice ACP). 157
B.2. Kolmogorov-Smirnov Test para la Muestra 2 (índice ACP). 157
B.3. Kolmogorov-Smirnov Test para la Muestra 3 (índice ACP). 157
B.4. Lilliefors Test Test para la Muestra 1 (índice ACP). 157
B.5. Lilliefors Test Test para la Muestra 2 (índice ACP). 157
B.6. Lilliefors Test Test para la Muestra 3 (índice ACP). 158
B.7. Shapiro-Wilk normality Test para la Muestra 2 (índice ACP) 158
B.8. Shapiro-Wilk normality Test para la Muestra 3 (índice ACP) 158
B.9. Jarque-Bera Test for normality para la Muestra 1 (índice ACP). 158
B.10. Jarque-Bera Test for normality para la Muestra 2 (índice ACP). 158
B.11. Jarque-Bera Test for normality para la Muestra 3 (índice ACP). 158
B.12. Kolmogorov-Smirnov Test para la Muestra 1 (índice ACP). 159
B.13. Kolmogorov-Smirnov Test para la Muestra 2 (índice ACP). 159
B.14. Kolmogorov-Smirnov Test para la Muestra 3 (índice ACP). 159
B.15. Lilliefors Test Test para la Muestra 1 (índice ACP). 159
B.16. Lilliefors Test Test para la Muestra 2 (índice ACP). 159
B.17. Lilliefors Test Test para la Muestra 3 (índice ACP). 159
B.18. Shapiro-Wilk normality Test para la Muestra 2 (índice ACP) 159
B.19. Shapiro-Wilk normality Test para la Muestra 3 (índice ACP) 159
B.20. Jarque-Bera Test for normality para la Muestra 1 (índice ACP). 160
B.21. Jarque-Bera Test for normality para la Muestra 2 (índice ACP). 160
B.22. Jarque-Bera Test for normality para la Muestra 3 (índice ACP). 160
B.23. Accidentes de Trabajo por Actividad Económica para el periodo 2012-2018 160
B.24. Categorías Individuales de la CIIU por sección 161
B.25. Muestra 1 de tamaño 300 161
B.26. Muestra 2 de tamaño 300 162
B.27. Muestra 3 de tamaño 300 162
B.28. Kolmogorov-Smirnov Test para la Muestra 1 (población normal). . . 162
B.29. Kolmogorov-Smirnov Test para la Muestra 2 (población normal). . . 162
B.30. Kolmogorov-Smirnov Test para la Muestra 3 (población normal). . . 163
B.31. Levene's Test for Homogeneity of Variance (muestras normales). . . 163
B.32. Análisis de Varianza de un factor. 163
B.33. Test de Kruskal-Wallis (muestras normales). 163

Lista de Símbolos

Para una mejor comprensión del presente trabajo de titulación se detalla a continuación una tabla con las notaciones utilizadas.

Símbolo	**Significado**	**Término**
Ω	Espacio Muestral	
	Variable aleatoria	$v.a.$
$F_X(x)$	Función de distribución acumulada	cdf
$f_X(x)$	Función de densidad de probabilidad	pdf
$p_X(x)$	Función de masa de probabilidad	pmf
$E[g(X)]$	Valor esperado de alguna función $g(X)$	
$E[X^k]$	k-ésimo momento de una v.a. X	
$\mu(media)$	Primer momento central de una v.a. X sobre su media	$E(X)$
$\sigma^2(varianza)$	Segundo momento central de una v.a. X sobre su media	$Var(X)$
M_X	Mediana de la población $\mathscr{X}$	
M_Y	Mediana de la población $\mathscr{Y}$	
$cov(A,B)$	Covarianza entre la v.a. A y la v.a. B	
$corr(A,B)$	Correlación entre la v.a. A y la v.a. B	
$N(0,1)$	Distribución Normal Estándar	
$\xrightarrow[d]{}$	Convergencia en distribución	
H_0	Hipótesis nula	
H_1 o H_A	Hipótesis alternativa	
$\Phi(x)$	La cdf de la normal estándar	
x_p	p-ésimo cuantil de la distribución de una v.a.	

Símbolo	Significado	Término
$Q_X(p)$	Función cuantil de una v.a.	$F_X^{-1}(P)$

Si F_X es estrictamente creciente entonces $Q_X(p) = x_p$ tiene solución única. Sin embargo, como la cdf no puede ser creciente para todos los valores de x, se redefine a x_p como $x_p = inf\{x : F_X(x) \geq p\}$ con $0 < p < 1$ para obtener un valor único de la función cuantil incluso si X es una v.a. discreta.

$\mathscr{X}$	La población X
$\mathscr{Y}$	La población Y
X	La Muestra $X_1, X_2, \ldots, X_n$
Y	La Muestra $Y_1, Y_2, \ldots, Y_m$

Población continua es equivalente a decir que la población tiene función de distribución acumulada continua.

$X >^{ST} Y$	$\mathscr{X}$ es estocásticamente más grande que $\mathscr{Y}$.	
	Asymptotic Relative Efficiency	ARE
	Análisis de Componentes Principales	ACP
	Indicador Compuesto	CI
	Análisis Envolvente de Datos	DEA
	Benefit of the Doubt	BOD

Se utilizada la expresión Indicador ACP para hacer referencia al proceso de construcción de las muestras basadas en el procedimiento del ACP.

Se utilizada la expresión Indicador CI para hacer referencia al proceso de construcción de las muestras basadas en las ponderaciones del Benefit of the Doubt.

RESUMEN

El presente trabajo de investigación se desarrolla en el campo de la estadística no paramétrica. De las muchas opciones de análisis que presenta lo no paramétrico, optamos por realizar un estudio teórico-descriptivo de los estadísticos lineales de rango, junto con los test correspondientes que se desarrollaron a partir de estos; como una posibilidad para abordar el problema de la falta de normalidad de los datos. Se comienza con la presentación del Test de Signo y a medida que se avance en el documento se presentan test mucho más complejos; estos test están orientados a detectar diferencias, ya sea a nivel de ubicación o de escala dentro del problema general de comparar dos o k muestras independientes. Luego de un adecuado proceso de depuración de datos, el Análisis de Componentes Principales (ACP) y Benefit of the Doubt Approach (BOD) fueron considerados como herramientas en la construcción de los indicadores de salud de los individuos que se ubican dentro de las actividades económicas. Las proyecciones en los nuevos ejes de los grupos más sobresalientes son las variables observadas para el proceso de inferencia, divididos en los grupos de interés. A pesar de que las muestras no verifican el cumplimiento de las hipótesis del modelo lineal con errores normalmente distribuidos, se realiza el Test Anova de un factor con el fin de comparar los resultados obtenidos por éste, con los obtenidos por su contraparte no paramétrico, el Test de Kruskal-Wallis. Finalmente, se presenta un enlace web en donde el lector tiene la libertad de interactuar y realizar variaciones a todo el proceso práctico desarrollado en el presente trabajo de titulación.

ABSTRACT

The current research work is developed in the field of non-parametric statistics. Out of the many options for non-parametric analysis, we decided to put forward a theoretical-descriptive study of linear rank statistic along the corresponding tests developed from them as a possibility to approach the lack of normality of the data problem. It starts with the Sign Test, and as the document goes on, much more complex tests are presented; these test are aimed at detecting differences, both at location and at scale level in the general problem of comparing two or k independent samples. After an appropriate data cleaning process, the Principal Component Analysis (PCA) and Benefit of the Doubt Approach (BOD) were considered as tools for the construction of the individuals health indicators for the individuals involved in the economic activities. The projections in the ACP axes becomes the indicators wich are included in the inference process, divided into the interest groups. Despite the samples not verifying the fulfillment of the lineal model with normally distributed errors hypotheses, the Anova Test of a factor is performed, looking to compare the results obtained by it with the ones obtained by its nonparametric counterpart, the Kruskal-Wallis Test. Finally, a web link is presented, where the reader can interact and make variations to the whole practical process developed within this thesis project.

CAPÍTULO 1

Introducción

El objetivo de los procedimientos de inferencia estadística es usar los datos de una muestra para obtener información aunque incierta sobre el comportamiento de la población subyacente de la cual fue extraída. No obstante, los supuestos que se dan en el análisis de inferencia no siempre son verificables. A pesar de ello, los investigadores tanto experimentados como principiantes han preferido el uso técnicas estadísticas paramétricas o en su defecto han optado por la utilización de diversos procesos de transformación de datos, para así, como mencionan Conover e Iman (1981), ajustar los problemas del mundo real dentro del marco de la teoría estadística normal.

El contraste de potencia que brindan las técnicas paramétricas cuando se verifica el cumplimiento de las hipótesis de normalidad, hace que se deje de lado a lo no paramétrico. Sin embargo, generalmente no se puede encontrar evidencia estadística de que una población siga una distribución de probabilidad determinada y menos aún saber si esta distribución es o se aproxima a una normal. Por este motivo, las técnicas de estimación no paramétricas surgen como una alternativa más flexible a los modelos paramétricos.

A pesar de que la literatura nos proporciona una cantidad considerable de técnicas de inferencia no paramétrica, como por ejemplo: suavizamiento por Kernel, estimación de densidad o teoría del Minimax (Wasserman 2006), nos centramos de manera particular en los fundamentos teóricos de los estadísticos lineales de rango. Esta teoría nos brinda la posibilidad de utilizar la magnitud relativa a una observación solo para determinar su rango; y a partir de este punto trabajar con algún proceso adecuado para los rangos. Este enfoque genera una combinación de los métodos paramétricos con los no paramétricos. Un claro ejemplo de esta combinación,

llamada Transformación de Rango (Conover e Iman 1981), es el proceso de aplicar el Test Anova de un factor a los datos transformados por sus rangos.
Mencionado proceso también es conocido como el Test de Kruskal-Wallis (Montgomery 2004).

Ahora bien, bajo la incertidumbre del cumplimiento de los supuestos del modelo lineal con errores normalmente distribuidos, Montgomery (2004) recomienda que: "Cuando exista preocupación acerca del supuesto de normalidad o por el efecto de puntos atípicos o valores "absurdos", el análisis de varianza común se realice tanto a los datos originales como a los rangos" (p.118). Por lo tanto, para resolver un problema de carácter empírico es fundamental la elección de una prueba estadística adecuada, misma que dependerá del problema que se este tratando, así como también de la naturaleza propia de los datos.

En virtud de lo anterior, el presente trabajo de titulación que emplea transformaciones construidas a partir del procedimiento del Análisis de Componentes Principales y el enfoque Benefit of the Doubt, se convierte en una alternativa muy atractiva al inicio de una investigación para descubrir la estructura probabilística que gobierna a los datos; y va orientado no solo para los estudiantes de pregrado, sino de manera general para cualquier investigador experimentado o principiante que realice algún tipo de análisis de datos, en donde no se tenga certeza del cumplimiento de los supuesto de normalidad.

1.1. Antecedentes

Con base a la fundamentación teórica de las fuentes bibliográficas del presente estudio, algunos trabajos de investigación relacionados con las técnicas estadísticas no paramétricas son los siguientes:

- El trabajo de titulación de Hernando Gamarra, Álvaro Pérez y Ramiro Quiseno, previo a la obtención del título de Licenciado en Matemáticas, titulado **"Estadística no Paramétrica"** de la Universidad de Sucre. El objetivo del mencionado trabajo fue realizar un análisis descriptivo de los métodos estadísticos paramétricos y no paramétricos más relevantes desde el punto de vista de los autores, así como también dar instrucciones sobre la utilización del programa estadístico *Statistic System New Creation (NCSS)*. Todo esto encaminado a comparar los resultados obtenidos por los trabajos de titulación previamente

realizados por los estudiantes de Ingeniería, Biología, Zootecnia, Enfermería, entre otras carreras de la Universidad de Sucre, en los que se aplicaron métodos paramétricos sin tener la certeza del cumplimiento de los supuestos de normalidad; con los obtenidos por las pruebas estadísticas no paramétricas descritas por Gamarra, Pérez y Quiseno (2006).

- El trabajo de titulación de Andrés Rojas, previo a la obtención del título de Ingeniero en Estadística Informática, titulado **"Técnicas Estadísticas Paramétricas y No Paramétricas Equivalentes: Resultados Comparativos por Simulación"** de la Escuela Superior Politécnica del Litoral. El objetivo de este estudio se centra en realizar un análisis comparativo entre las técnicas estadísticas paramétricas y no paramétricas, enfocado principalmente en el valor plausible de dichas técnicas. La comparación se realiza por medio de simulaciones numéricas de muestras provenientes de diversas poblaciones, a fin de establecer por medio de resultados numéricos el comportamiento de técnicas equivalentes en igualdad de condiciones, violando y cumpliendo sus supuestos; para así determinar que pruebas se desempeñan mejor que otras en ciertos casos y el efecto de los supuestos teóricos en su robustez. En total se realizaron 100 simulaciones para cada caso indicado (Rojas 2003).

- La sección de Estadísticas No Paramétricas de la *American Statistical Association* en donde se incluye un fondo de dotación económico, el Premio Gottfried E. Noether cuyo nombre es en homenaje a Gottfried Emanuel Noether, un destacado académico en estadísticas no paramétricas. Dicho fondo tiene como fin fomentar la investigación en el campo de la estadística no paramétrica, y se da por parte de la esposa de Noether, Emiliana Noether y de su hija Mónica Noether; como un reconocimiento a distinguidos investigadores y docentes.

1.2. Justificación

Con frecuencia antes de realizar cualquier proceso de análisis de datos, la problemática que enfrentan tanto el investigador principiante como el experimentado, es el de decidir que prueba estadística es la más adecuada para analizar un conjunto de datos. Dado que la aplicación de la estadística en el análisis de datos es muy amplia abarcando campos como las ciencias exactas o las ciencias sociales, la elección de una prueba estadística apropiada es primordial si deseamos que los resultados alcanzados sean los adecuados.

Debido a la naturaleza inherente que presentan los datos, los instrumentos de medición con la que se obtiene y a medida que el tamaño de ésta se hace grande, los expertos sostienen que no es posible determinar la forma de la distribución poblacional de donde provienen los datos y que los supuestos del modelo general con errores normalmente distribuidos no llegan a cumplirse. En este contexto, Conover e Iman (1981) expresan que:

El gran problema que los estadísticos aplicados han enfrentado rigurosamente desde el inicio de la estadística paramétrica, es el de ajustar los problemas del mundo real dentro del marco de la teoría estadística normal, cuando muchos de los datos con los que tratan son claramente no normales. (p.124)

De esta situación surgen dos escuelas de pensamiento:

1. Transformación de datos.
2. Procedimientos de distribución libre (Estadística no paramétrica).

Molinero (2003) detalla que el primer enfoque es la solución más natural. La idea se basa en la modificación de los datos mediante alguna transformación matemática, siendo las más comunes las transformaciones logarítmicas, raíz cuadrada, arcsen, etc.

En cambio, una posibilidad de análisis de los procedimientos de distribución libre, llamados así porque no es necesario conocer la forma de la distribución probabilística de la población de la cual se extrajo la muestra y estudiada por Gibbons y Chakraborti (2011) al igual que Pratt y Gibbons (1981), nos permite basarnos en el criterio de los rangos de las observaciones. Si bien es cierto que los métodos no paramétricos son menos atractivos para los investigadores debido a que se llega a conclusiones débiles o más generales, ocasionadas por la flexibilidad en cuanto a las suposiciones que realiza su desarrollo teórico; son perfectamente válidos y de hecho son los más utilizados cuando el supuesto de normalidad no llega a cumplirse (Kvam y Vidakovic 2007).

La flexibilidad en los supuestos expuesta con anterioridad se relaciona con el hecho de considerar que las poblaciones son continuas (que tienen función de distribución acumulada continua), en algunos casos simétricas (función de distribución acumulada simétrica) y que se centran en cualquier medida de tendencia central. Esta medida de centralidad puede ser cualquier i-ésimo cuantil. De manera particular el cuantil 50 (la mediana), mismo que siempre existe a diferencia de la esperanza (media) en lo paramétrico que no necesariamente puede existir (ej. distribución de

Cauchy).

Lo expresado anteriormente genera estadísticos que pueden detectar diferencias entre dos o más poblaciones, ya sea a nivel de ubicación o de escala, con la característica de que los mismos llegan a ser casi tan eficientes como sus contrapartes paramétricos en términos de Eficiencia Relativa Asintótica (Gibbons y Chakraborti 2011). Por ejemplo, si se desea realizar una prueba a nivel de ubicación y bajo el cumplimiento de los supuestos del modelo general con errores normalmente distribuidos, el Test de Kruskal-Wallis tendrá como contraparte al Test Anova, obteniéndose resultados similares, pero con la ventaja de que las suposiciones de base o hipótesis sometidas a prueba son más flexibles. Así, en términos generales se puede considerar que aunque la potencia de las pruebas paramétricas es mayor que la que ofrecen sus similares no paramétricos, es conveniente comentar que el tamaño de una muestra es un requisito indispensable para aumentar la eficacia de un test, pues a medida que la muestra crece, la posibilidad de cometer un error de tipo II disminuye.

Algunos autores consideran que el verdadero comienzo de la estadística no paramétrica se da alrededor de 1936 con la presentación del llamado Test de Signo. No obstante, J. V. Bradley (1960), por ejemplo distingue la aparición de cuatro etapas, siendo en la cuarta, la llamada "sin parámetros" en donde se menciona que: "Los esfuerzos para identificar los parámetros de una población padre con el fin de poder especificar su ley de probabilidad fueron reemplazados en gran medida por intentos de determinar relaciones exactas, válidas para tamaños de muestra restringidos" (p.2); como los primeros indicios de lo no paramétrico.
Desde entonces, lo no paramétrico toma la forma de una disciplina separada de la estadística tradicional; y sus aplicaciones en la actualidad son diversas, que pueden ir desde el área de la rehabilitación (Martin 2011) hasta procesos de investigación en psicología (Moses 2011). A modo de ejemplo, el trabajo de investigación de Espitia (2014), cuyo propósito era identificar diferencias existentes en condiciones de empleo, condiciones de trabajo y salud mental laboral según la posición de la clase social para los trabajadores asalariados de Bogotá; es fácilmente aplicable a los procesos de distribución libre.

Sin embargo, debido a que los resultados de Espitia (2014) están condicionados a que las muestras cumplan con los supuestos de normalidad, es necesario realizar un estudio similar aplicable a nuestro país, pero sin limitarnos al conocimiento previo de que nuestras muestras cumplan o no con las hipótesis de normalidad, sino más

bien abordar el problema dentro una marco más general, en donde el conocimiento de la distribución de probabilidad de la población no sea un requisito previo.
Esto nos orienta a realizar comparaciones entre perfiles de salud de individuos sanos ecuatorianos por actividad económica, con el objetivo de ver si existe alguna diferencia a nivel de ubicación entre las poblaciones a la que pertenecen los individuos.

Para este fin, se desarrollan dos indicadores simples de salud que serán utilizados en la construcción de las variables, cuyas características son objeto de inferencia. El primer indicador se basa en el procedimiento del Análisis de Componentes Principales (indicador ACP) y el segundo es proporcionado por las ponderaciones del Benefit of the Doubt (indicador CI), mismas que son construidas a partir de las seis variables retenidas del ACP, en relación a los datos transformados debido al mejor individuo sano.

1.3. Objetivos

1.3.1. Objetivo General

- Comparar perfiles de salud entre grupos determinados por alguna característica socioeconómica mediante métodos no paramétricos.

1.3.2. Objetivos Específicos

- Analizar el desarrollo estadístico de los métodos no paramétricos basados en los rangos de las observaciones, para el caso de una, dos y k muestras independientes.
- Desarrollar indicadores simples de salud de los individuos dentro de las actividades económicas.
- Comprobar el cumplimiento o no de los supuestos del modelo general normal para cada muestra derivada de los grupos más sobresaliente de acuerdo al indicador establecido, y aplicar el proceso de inferencia estadística más adecuado.

1.4. Hipótesis

Mediante el diseño de indicadores representativos simples, que caracterizan el perfil de salud de los individuos sanos, determinar si existe diferencia en dichos perfiles usando métodos no paramétricos.

1.5. Metodología

El presente trabajo de titulación dispone de una base de datos que pertenece a una empresa de servicios médicos. Las observaciones corresponden a datos clínicos de diagnóstico general de individuos sanos que se encuentran en diferentes partes del sector formal ocupacional ecuatoriano.

Las fases que se llevarán a cabo se detallan a continuación.

Fase 1: Las aplicaciones de la teoría estadística para situaciones del mundo real se limita a la verificación del cumplimiento de supuestos fundamentales para el buen desenvolvimiento de la misma, pero cabe recalcar que un modelo matemático simplifica la complejidad inherente de las situaciones que ocurren en la vida diaria, logrando así solo resultados aproximados más no algo que sea verdaderamente cierto. De esta manera, a lo largo de la historia se han desarrollado teorías que tratan en la mayor parte posible de abarcar situaciones que se dan de manera cierta, pero considerando simplificaciones, si dichas situaciones lo permiten. Esto claramente es de gran ayuda para el investigador, pues le permite desarrollar nuevos fundamentos teóricos que se adapten a las nuevas necesidades o naturaleza de los datos.

En este sentido, el presente estudio inicia con la exhibición del desarrollo teórico de los métodos no paramétricos más sobresalientes a nivel de eficiencia relativa asintótica basados en los rangos de las observaciones. Es decir, dentro de la gran variedades de test no paramétricos utilizados para dar respuesta al problema general de comparar una, dos y k muestras independientes, se escogen aquellos que mejor se adaptan a los problemas ya mencionados, centrándonos principalmente en los test no paramétricos que son la contraparte directa a la prueba t y Anova de un factor que ofrece el campo paramétrico.

Fase 2: El proceso de depuración de los datos considerará solo a las variables que contengan al menos un cierto porcentaje de información completa (sin datos

perdidos o NA), y se hará la consideración que los individuos serán aquellos que pertenezcan a los grupos más sobresaliente (más numerosos).

La construcción del primer indicador de salud (indicador ACP) se basará en la reducción de la dimensión de las variables que se considerarán en el estudio mediante el proceso de Análisis de Componentes Principales (ACP). Esto con el fin de obtener nuevas variables (ejes) que recojan la mayor parte de información de las variables iniciales, en donde los individuos serán proyectados en estos nuevos ejes (Johnson y Wichern 2007), (Peña 2002).
El segundo indicador (indicador CI), desarrollado para confirmar o no los resultados obtenidos por el primero, se basa en el Análisis Envolvente de Datos aplicado por el Benefit of the Doubt Approach a las variables retenidas por el ACP (en relación a los datos sobre el mejor individuo sano).

Fase 3: Las variables observadas (muestras) con las que se va a realizar el proceso de inferencia serán las proyecciones de los individuos, dentro de sus nuevos ejes y dentro de los grupos más representativos (para el indicador del ACP). En cambio, para el segundo indicador, las muestras serán las ponderaciones dadas por el Benefit of the Doubt Approach dentro de los grupos más representativos.

Estas muestras serán sometidas a cada uno de los requerimientos que exige la teoría paramétrica, y ya sea que cumplan o no con los supuestos de normalidad se procederá con la aplicación del análisis de varianza (Anova) y los resultados de aquí se compararán con los obtenidos por su contraparte no paramétrico, el Test de Kruskal-Wallis.

1.6. Estructura del Trabajo

En el capítulo II se detalla toda la base teórica de la investigación. Se comienza con una breve presentación de la inferencia estadística paramétrica y en lo posterior el punto de análisis se centra en la estadística no paramétrica. Lo no paramétrico comienza con una pequeña explicación general de los procedimientos de distribución libre, y continua con el desarrollo teórico de los test más simples hasta sus versiones más generales. La culminación del capítulo se da con la descripción del Test Kruskal-Wallis, el proceso de comparaciones múltiples aplicado cuando la hipótesis de igualdad de poblaciones es rechazada y finalmente, con la exhibición de

los resultados concernientes a la Eficiencia Relativa Asintótica, así como también el procedimiento del Análisis de Componentes Principales y el Análisis Envolvente de Datos aplicado al Benefit of the Doubt Approach.

El capítulo III hace una descripción detallada del proceso de depuración de los datos, las variables utilizadas y las consideraciones planteadas en la obtención de cada una de las muestras objeto de estudio. Posteriormente se presenta el procedimiento utilizado en la construcción de los índices de salud y su relación con cada una de las muestras.

El capítulo IV muestra por separado los resultados de la aplicación del test paramétrico Anova y del test no paramétrico Kruskal-Wallis, a las muestras derivadas de los indicadores de salud. Todo encaminado a saber si la afirmación de que las k muestras provienen o no de poblaciones idénticas a un nivel de ubicación, es verdadera o falsa. Finalmente se realiza contrastes post hoc para identificar que pares de muestras difieren entre sí a nivel de ubicación, así como ver lo que sucede un sentido de variabilidad, mediante la aplicación de test no paramétricos para el problema de escala.

Adicionalmente, si el lector desea considerar variaciones en el proceso de depuración de datos, construcción de los índices de salud, muestras a considerar, tipo de método del proceso post hoc y test no paramétricos a nivel de escala, lo puede realizar ingresando al enlace web que se presenta tanto en el capítulo III como IV.

En el capítulo V se muestran las conclusiones y recomendaciones establecidas en base a los fundamentos teóricos y resultados alcanzados en el presente trabajo de investigación.

CAPÍTULO 2

Marco Teórico

El presente capítulo describe brevemente algunos de los conceptos usados más adelante. Por un lado, el tema paramétrico presenta definiciones básicas de por ejemplo, momentos de combinaciones lineales de variables aleatorias, prueba de hipótesis, p-valor, corrección de continuidad, estadístico de orden y test consistente. Todo con la finalidad de relacionar de mejor manera los aspectos teóricos-prácticos de los métodos no paramétricos.
Por otro lado, lo referente al campo no paramétrico comienza con una breve exhibición del por qué surge este nuevo tipo de estadística, su historia y el enfoque que le dan los investigadores especialistas del campo. A continuación se detalla el test no paramétrico más simple, Sign Test, continuando hasta sus versiones más generales; mismas que serán presentadas como las soluciones más adecuadas, desde la perspectiva no paramétrica basada en los rangos de las observaciones, para el problema general de comparar dos o k muestras independientes cuando el supuesto de normalidad no llega a cumplirse.
Finalmente, de darse el caso en el cual se rechace la hipótesis nula de poblaciones idénticas, el proceso de comparaciones múltiples nos permitirá saber que tipo de poblaciones son las que difieren entre si a nivel de ubicación. Adicionalmente, la ARE nos direccionará a conocer que test no paramétrico es más eficiente que otro no paramétrico, así como su eficiencia en relación al t Test o Anova Test.
La culminación del capítulo se dá con la presentación del procedimiento de Análisis de Componentes Principales y el proceso de construcción de indicadores compuestos mediante el Análisis Envolvente de Datos aplicado al Benefit of the Doubt Approach.

2.1. Nociones Básicas

En esta sección se describe, pero no de una manera tan detallada algunas definiciones fundamentales de probabilidad y estadística que están relacionadas de manera directa con el marco teórico del presente trabajo de titulación.

Definición 2.1.1 (Momentos de combinaciones lineales de variables aleatorias)**.** *Sean $X_1, X_2, \ldots, X_n$, n variables aleatorias y $a_1, a_2, \ldots, a_n$, n constantes reales arbitrarias, entonces*

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i) \tag{2.1}$$

$$var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 var(X_i) + 2\sum_{1\leq i<j\leq n}\sum a_i a_j cov(X_i, X_j) \tag{2.2}$$

$$cov\left(\sum_{i=1}^{n} a_i X_i, \sum_{i=1}^{n} b_i X_i\right) = \sum_{i=1}^{n} a_i b_i var(X_i) + \sum_{1\leq i<j\leq n}\sum (a_i b_j + a_j b_i) cov(X_i, X_j) \tag{2.3}$$

Teorema 2.1.1 (Teorema Central Límite)**.** *Sean $X_1, X_2, \ldots, X_n$, n variables aleatorias independientes idénticamente distribuidas (iid) con media μ y varianza $\sigma^2 < +\infty$, entonces*

$$\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu) = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow[d]{} N(0,1)$$

$$i.e.\ P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq x\right) \to \Phi(x)$$

donde $\bar{X}_n = \sum_{i=1}^{n} X_i$ y $\Phi(x)$ la cdf de la normal estándar.

2.1.1. Prueba de Hipótesis

Para Gibbons y Chakraborti (2011), una hipótesis estadística es: "Una afirmación o aseveración sobre la función de probabilidad de una o más variables aleatorias, o una declaración sobre las poblaciones de las que se extraen una o más muestras, por ejemplo, su forma o el valor de sus parámetros" (p.18).
Si la afirmación que se da especifica completamente a la población se llamará simple, caso contrario compuesta.

Para determinar si la declaración en cuestión es cierta o no, se realiza una prueba de hipótesis (regla de decisión), misma que considera una hipótesis nula, una

hipótesis alternativa y una adecuada región de rechazo.

Definición 2.1.2 (Hipótesis nula y alternativa)**.** *La hipótesis nula H_0 es la hipótesis bajo el test (prueba) y la hipótesis alternativa (unilateral o bilateral), H_1 o H_A, es la conclusión alcanzada si se rechaza la hipótesis nula.*

Definición 2.1.3 (Test de una hipótesis estadística)**.** *El test de una hipótesis estadística es una regla que permite tomar una decisión sobre si H_0 debe rechazarse o no en base al valor observado de un estadístico.*

La distribución de probabilidad del test estadístico bajo H_0 se conoce como distribución nula del test estadístico (Gibbons y Chakraborti 2011).

Definición 2.1.4 (Región crítica)**.** *La región crítica o región de rechazo R de un test es el subconjunto de valores asumidos por el test estadístico que, de acuerdo con el test conduce al rechazo de H_0.*

A los límites de R bajo un test se los conoce como valores críticos del test estadístico. En otras palabras, si W_N es un test estadístico que conduce al rechazo de H_0 si $W_N \geq W_\alpha$, entonces W_α es el valor crítico de R ($W_N \in R$ si $W_N \geq W_\alpha$).

Cada vez que se realiza una test para verificar la afirmación concerniente a la hipótesis nula, son posibles dos tipos de errores: Error tipo I y Error tipo II.

Definición 2.1.5 (Error Tipo I y Tipo II)**.** *Un error de tipo I es la acción de rechazar H_0 cuando la misma es verdadera. La probabilidad de este error se refiere al nivel de significancia del test y se denotará por α. Por otro lado, un error de tipo II aparece si se acepta H_0 cuando la misma es falsa. La probabilidad de cometer este tipo de error se denotará por β.*

Los tipos de errores y las decisiones correctas se muestran en la Tabla 2.1

Tabla 2.1: Error tipo I y tipo II

	No rechazamos H_0	Rechazamos H_0
H_0 es verdadera	✓	**Error tipo I (α)**
H_0 es falsa	**Error tipo II (β)**	✓

Las probabilidades de cometer estos errores para un test estadístico W_N en la prueba de hipótesis $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ son respectivamente

$$\alpha(\theta) = P_\theta(W_N \in R), \theta \in \Theta_0 \qquad y \qquad \beta(\theta) = P_\theta(W_N \notin R), \theta \in \Theta_1$$

para $\theta \in \Theta = \Theta_0 \cup \Theta_1$ y $\Theta_0 \cap \Theta_1 = \emptyset$; con Θ denotando al conjunto de parámetros.

Al supremo de $\alpha(\theta)$, para todo $\theta \in \Theta_0$ se lo conoce como el tamaño del test. En cambio, el nivel de significancia es una cota nominal preseleccionada para $\alpha(\theta)$, mis-

ma que no puede alcanzarse cuando la función de probabilidad relevante es discreta (lo cual suele ocurrir en pruebas de hipótesis no paramétricas). De darse el caso que la distribución de probabilidad sea discreta podría surgir cierta confusión respecto a estos términos. Por lo tanto, de aquí en adelante el símbolo α denotará el tamaño del test, el nivel de significancia o la probabilidad de cometer un error de tipo I, mismo que se acompañará del adjetivo "exacto" si $sup_{\theta\in\Theta_0}\alpha(\theta) = \alpha$ y de "conservador" si la probabilidad de cometer un error tipo I es a lo mucho α, es decir si $sup_{\theta\in\Theta_0}\alpha(\theta) \leq \alpha$

Definición 2.1.6 (Potencia de un test)**.** *La potencia de un test denotada por* $\Pi(\theta) = P_\theta(W_N \in R), \theta \in \Theta_1$ *es la probabilidad que el test estadístico conduzca al rechazo de* H_0.

En otras palabras, la potencia es la probabilidad de una decisión correcta calculada cuando H_0 es falsa, es decir cuando H_1 es verdadera. De esta manera

$$\Pi(\theta) = 1 - \beta(\theta) = P_\theta(W_N \in R), \theta \in \Theta_1$$

2.1.2. p-valor

Un criterio alternativo para verificar la validez de una prueba de hipótesis se realiza mediante el uso del p-valor, también llamado valor de probabilidad o probabilidad de significancia. Este valor se define como el nivel de significación más pequeño en el que la hipótesis nula sería rechazada usando el valor observado del test estadístico.

Por un lado, un p-valor pequeño se interpreta en un sentido que la muestra produjo un resultado que es bastante raro bajo el supuesto de la hipótesis nula. Esto conlleva al rechazo de la hipótesis nula.
Por otro lado, un p-valor grande indica que el resultado de la muestra es consistente con la hipótesis nula. En consecuencia, la hipótesis nula no se rechaza (Gibbons y Chakraborti 2011).

Si la distribución de probabilidad nula de un test estadístico se aproxima asintóticamente por medio de otra distribución, el p-valor encontrado a partir de esta nueva función de probabilidad se llamará p-valor asintótico o aproximado. Así, una adecuada decisión acerca de H_0 implica seleccionar el valor para α. Si este valor es mayor o igual al p-valor (o p-valor asintótico) se debe rechazar la hipótesis nula, caso contrario H_0 no se rechaza.
Los valores usuales para α son de 0.01, 0.05 o similares.

2.1.3. Test Consistente

"La consistencia es un criterio de prueba "bueno" para los métodos paramétricos y no paramétricos, y todos los procedimientos de prueba estándar comparten claramente esta propiedad" (Gibbons y Chakraborti 2011, p.22).

Definición 2.1.7 (Consistencia)**.** *Un test se llama consistente para una alternativa específica si la potencia del test cuando esa alternativa es cierta se acerca a* 1 *cuando* $n \to +\infty$.

No obstante, la consistencia de un test en ciertas situaciones se puede determinar en base de la convergencia en probabilidad del test estadístico al parámetro de interés. De esta manera, Gibbons y Chakraborti (2011) presentan un resultado útil para investigar la consistencia de un test.

A continuación se describe el resultado:

Dada una muestra aleatoria de tamaño n substraída de una población $F_X(x:\theta)$, $\theta \in \Theta$, T un test estadístico para la hipótesis nula $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ y $g(\theta)$ una función arbitraria de θ tal que

$$\begin{aligned} &g(\theta) = \theta_0 \quad si \quad \theta \in \Theta_0 \\ &\quad y \\ &g(\theta) \neq \theta_0 \quad si \quad \theta \in \Delta \quad para \quad \Delta \subset \Theta_1. \end{aligned} \tag{2.4}$$

Si para todo θ se tiene que

$$E(T) = g(\theta) \quad y \quad \lim_{n \to +\infty} Var(T) = 0 \ , \tag{2.5}$$

entonces el test de tamaño α y región de rechazo $T \in R$ para $|T - \theta| > c_\alpha$ es consistente para la subclase Δ.

De manera similar, si la subclase de alternativa unilateral es

$$\begin{aligned} &g(\theta) = \theta_0 \quad si \quad \theta \in \Theta_0 \\ &\quad y \\ &g(\theta) > \theta_0 \quad si \quad \theta \in \Delta \quad para \quad \Delta \subset \Theta_1 \ , \end{aligned} \tag{2.6}$$

entonces la consistencia del test de tamaño α tiene región de rechazo $T \in R$ para $T - \theta_0 > c'_\alpha$.

La demostración de este resultado citado por Gibbons y Chakraborti (2011) se

presenta en Frases (1957) y se deriva directamente de la desigualdad de Chebyshev.

2.1.4. Corrección de Continuidad

En la inferencia no paramétrica se da el caso que la distribución nula exacta de varios test estadísticos es discreta. A pesar de que la literatura presenta tablas de regiones de rechazo para muestras de tamaño pequeño, es necesaria una aproximación de la distribución nula cuando se abordan muestras de tamaño grande. Si tales aproximaciones asintóticas son continuas, la introducción de una corrección para la continuidad puede mejorar la aproximación. Esto tiene sentido pues se considera el valor del test estadístico discreto como el punto medio de un intervalo (Gibbons y Chakraborti 2011).

Dado un test estadístico T y considerando una alternativa bilateral que rechaza H_0 si $T \geq t_{\alpha/2}$ o $T \leq t'_{\alpha/2}$, entonces para tamaños de muestras grandes la distribución $\dfrac{T - E_{\theta_0}(T)}{\sigma_{\theta_0}(T)}$ es asintóticamente la distribución normal estándar bajo H_0 y las regiones de rechazo con corrección de continuidad se obtienen al resolver

$$\frac{t_{\alpha/2} - 0{,}5 - E_{\theta_0}(T)}{\sigma_{\theta_0}(T)} = z_{\alpha/2} \qquad y \qquad \frac{t'_{\alpha/2} + 0{,}5 - E_{\theta_0}(T)}{\sigma_{\theta_0}(T)} = -z_{\alpha/2} \tag{2.7}$$

donde $z_{\alpha/2}$ satisface que $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ para $\Phi(x)$ la cdf de la distribución normal estándar.

De esta manera, se rechazará H_0 si

$$T \geq E_{\theta_0}(T) + 0.5 + z_{\alpha/2}\sigma_{\theta_0}(T) \qquad \text{o} \qquad T \leq E_{\theta_0}(T) - 0.5 - z_{\alpha/2}\sigma_{\theta_0}(T).$$

De manera similar se pueden obtener las regiones de rechazo de tamaño α con corrección de continuidad para las alternativas unilaterales tanto de cola superior como inferior. Estas regiones respectivamente son

$$T \geq E_{\theta_0}(T) + 0.5 + z_{\alpha}\sigma_{\theta_0}(T) \qquad y \qquad T \leq E_{\theta_0}(T) - 0.5 - z_{\alpha}\sigma_{\theta_0}(T). \tag{2.8}$$

Ahora bien, si t_0 denota el valor observado del test estadístico T cuya distribución nula puede ser aproximada por una distribución continua, típicamente la normal para muestras de tamaño grande, entonces para la alternativa de cola superior el

p-valor con corrección de continuidad será

$$1 - \Phi\left(\frac{t_0 - 0.5 - E_{\theta_0}(T)}{\sigma_{\theta_0}(T)}\right) \tag{2.9}$$

y para la alternativa de cola inferior será

$$\Phi\left(\frac{t_0 + 0.5 - E_{\theta_0}(T)}{\sigma_{\theta_0}(T)}\right) \tag{2.10}$$

En el caso de la alternativa bilateral, que quiere decir que no hay dirección específica para calcular el p-valor, una idea es considerar al menor p-valor de las alternativas unilaterales; y si la distribución de probabilidad es simétrica tiene sentido duplicar este valor, no obstante en la inferencia no paramétrica se utiliza esta idea a pesar de que la distribución no sea simétrica (Gibbons y Chakraborti 2011).

2.1.5. Estadístico de Orden

Definición 2.1.8 (Estadístico de Orden). *Si $X_1, X_2, \ldots, X_n$ denotan una muestra aleatoria de alguna población con cdf F_X continua (la probabilidad de que dos o más v.a. tengan igual magnitud es cero) y supongamos que $X_{(1)}$ representa el valor más pequeño de $X_1, X_2, \ldots, X_n$; $X_{(2)}$ representa el segundo valor más pequeño de $X_1, X_2, \ldots, X_n$ y así sucesivamente; entonces $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ describe el único arreglo ordenado de la muestra original en forma creciente. La expresión $X_{(j)}$ denotará el j-ésimo estadístico de orden para $1 \leq j \leq n$*

Es fácil observar que algunas propiedades se derivan inmediatamente de la definición de estadístico de orden, por ejemplo:

1. $X_{(1)}$ y $X_{(n)}$ representan respectivamente el valor mínimo y máximo de la muestra.
2. El rango muestral (medida de dispersión) será $X_{(n)} - X_{(1)}$.
3. La mediana muestral (medida de ubicación) será el valor de $X_{((n+1)/2)}$ si n es impar, caso contrario será el promedio entre $X_{(n/2)}$ y $X_{(n/2+1)}$ para n par.

Para ver más a detalle la distribución de probabilidad conjunta, momentos exactos, momentos asintóticos entre otras propiedades de los estadísticos de orden puede ir a Gibbons y Chakraborti (2011) o Kvam y Vidakovic (2007).

2.2. Métodos No Paramétricos

En la actualidad todas las actividades, incluso las más cotidianas conllevan de manera directa o indirecta a la generación de información presentada a manera de datos, que dependiendo de las fuentes de recopilación y situaciones de interés tomarán un valor específico.

El adecuado tratamiento de estos datos nos puede dar una cierta noción, aunque algo incierta del comportamiento o tendencia que tiene la población subyacente de donde provienen los mismos. Y los procesos de inferencia, sujetos a estos tratamientos están limitados al cumplimiento de ciertos supuestos, pero, ¿qué hacer cuando dichos supuestos no llegan a cumplirse? En este contexto, el investigador debe optar por desarrollar o en su defecto buscar nuevas metodologías que se adapten a la información con la que se va a realizar el proceso de inferencia.

Por un lado, dentro de lo paramétrico, para obtener información de algún parámetro de una distribución poblacional, la idea comúnmente era ajustar los problemas del mundo real dentro del marco de la teoría estadística paramétrica bajo normalidad, ya sea por medio de procesos de transformación de datos o algún otro proceso que ajuste los datos a la distribución normal (Conover e Iman 1981). Sin embargo, debido a la naturaleza propia de los datos, el determinar a ciencia cierta cuál es la distribución que caracteriza a la población subyacente de la cual fueron extraídos suele no ser una tarea sencilla y menos aún saber si se ajustaban dentro de lo normal; así, el supuesto de normalidad en los procesos de inferencia ya no se considera un artículo de fe (J. V. Bradley 1960).
Por otro lado, en los procesos no paramétricos basados en los rangos de las observaciones vamos a ver que se pueden obtener contrapartes a las pruebas comunes t-Student y Anova de un factor, pero bajo una consideración más flexible, la continuidad de la distribución de probabilidad de la población. Adicionalmente, las hipótesis establecidas se centran en cualquier cuantil que no sea la media, debido a que esta no necesariamente existe (ej. Distribución de Cauchy). De manera particular se considerará a la mediana como medida de tendencia central, debido a que esta siempre existe para cualquier población ($Q_X(0.5) = F_X^{-1}(0.5) = M$ existe y es única) y por su robustez como medida de ubicación.

De todo lo anterior se presenta a la estadística no paramétrica, basada en los estadísticos de rango como una alternativa flexible al incumplimiento de los criterios del modelo general con errores normalmente distribuidos.

Jacob Wolfowitz, citado en Kvam y Vidakovic (2007) fue quién introdujo el término no paramétrico al decir que:

Nos referiremos a esta situación, donde una distribución está completamente determinada por el conocimiento de su conjunto de parámetros finitos, como el caso paramétrico, y denotaremos el caso opuesto, donde las formas funcionales de las distribuciones son desconocidas como el caso no paramétrico. (p.1)

De esta manera, la estadística no paramétrica se definió por lo que no es: "la estadística tradicional basada en la distribución conocida con parámetros desconocidos", pero Randles, Hettmansperger y Casella (2004) extienden la definición y sugieren que: "Las estadísticas no paramétricas pueden y deben definirse ampliamente para incluir toda la metodología que no utiliza un modelo basado en una sola familia paramétrica"(p.561). Así, al hablar de procedimientos no paramétricos se hace referencia a que no es necesario conocer a priori la distribución de probabilidad de los datos, razón por la cual a estos procedimientos también se los conoce como procesos de Distribución Libre.
Un aspecto importante que se debe tener en cuenta es no considerar a los métodos no paramétricos como la negación del caso paramétrico, sino más bien como una posibilidad para abordar el problema de la falta de normalidad de las observaciones; y si bien es cierto que estos métodos son menos poderosos que sus contrapartes paramétricos, bajo el supuesto de que se cumplan todos los requerimientos de normalidad los resultados obtenidos en ambos campos pueden ser ligeramente diferentes.

Las investigaciones desarrolladas por los matemáticos en lo relacionado a rangos y estadísticos de rango datan desde hace años, pero no fue sino hasta las décadas de 1940 y 1950 que la idea de test de rango ganó prominencia en la literatura estadística; y gran parte de esta popularidad como menciona Gómez (2010) se debe al trabajo de Erich Lehmann, quién en 1951 desarrolla test consistentes e insesgados no paramétricos, mismos que posteriormente serán publicados en el Annals of Mathematical Statistics. Así es como Lehmann y otros que desarrollaron propiedades teóricas de los métodos de rango hacen que los procesos de distribución libre reciban una considerable atención e interés por parte de la comunidad matemática.

Existen procedimientos de distribución libre para relacionar una muestra, dos muestras y la generalización a k muestras, en donde cada uno tiene respectivamente su contraparte paramétrico. Los detalles se presentan en la Tabla 2.2

Tabla 2.2: Pruebas paramétricas y su contraparte no paramétrica

Tipo de Problema [1]		Prueba paramétrica	Prueba no paramétrica
Muestras relacionadas	dos muestras	t-student	Sign Test Wilcoxon Signed-Ranks Test
	más de dos muestras	Anova	Friedman [2]
Muestras independientes	dos muestras	t-student	Test de la mediana Test U de Mann-Whitney
	más de dos muestras	Anova de un factor	Extensión del Test de la mediana Kruskal-Wallis

La clave según Kvam y Vidakovic (2007), para evaluar los datos en un marco no paramétrico es utilizar los rangos de las observaciones en lugar del valor real de sus magnitudes.

2.2.1. Estadístico Rango-Orden (Estadístico de rango)

Los datos se ordenan cuando se los organiza según algún criterio, como por ejemplo, del más pequeño al más grande o del mejor al peor. Si se considera el valor real de las observaciones solo para determinar su posición relativa dentro de la muestra común y luego de esto su magnitud es ignorada por completo en los procesos de inferencia, entonces un estadístico de rango-orden para una muestra aleatoria puede ser cualquier conjunto de constantes que indiquen algún orden de las observaciones.

De aquí en adelante el criterio de organización se basará en la ordenación de las observaciones de forma ascendente, lo cual se ejecuta mediante el estadístico de rango (o simplemente rango).

Definición 2.2.1 (Estadístico de Rango (Rango))**.** *El rango de la observación X_i de una muestra aleatoria de tamaño n, denotado por $r(X_i)$, es el número asignado a dicha observación según su orden en el arreglo (muestra) ordenado. La definición formal es la siguiente:*

$$r(X_i) = \sum_{j=1}^{n} \mathbb{1}(X_j \leq X_i) \qquad donde \qquad \mathbb{1}(X_j \leq X_i) = \begin{cases} 1 & si\ X_j \leq X_i \\ 0 & si\ X_j > X_i \end{cases} \tag{2.11}$$

[1] Existen más casos de test no paramétricos que depende del tipo de variable, pero solo se mencionan aquellos que son objeto de estudio. Para más detalle vea Gómez-Gómez, Danglot-Banck y Vega-Franco (2003).

[2] Solo se lo menciona debido a que es la contraparte del Test Anova para muestras pareadas, sin embargo, no es un caso de estudio del presente trabajo de titulación.

De manera similar, si $X_{(j)}$ representa el j-ésimo estadístico de orden (véase 2.1.8, pág. 16), entonces su rango correspondiente es j.

Debido a que trabajar con los rangos de las observaciones es algo simple y como en la práctica las mediciones reales a menudo son difíciles, costosas o incluso imposibles de obtener; el utilizar los rangos en un proceso de investigación hace que sea algo atractivo.

Gibbons y Chakraborti (2011) y Kvam y Vidakovic (2007) atribuyen que este estadístico es una variable aleatoria discreta y si la muestra aleatoria proviene de una población continua, entonces el estadístico seguirá la distribución uniforme discreta.

Propiedades del estadístico de rango

1. La función de probabilidad de masa (pmf) del estadístico es

$$f_{r(X)}(j) = P(r(X_i) = j) = \frac{1}{n} \qquad \text{para} \qquad 1 \leq j \leq n \tag{2.12}$$

2. La esperanza del estadístico es

$$E(r(X_i)) = \frac{n+1}{2} \qquad \text{para} \qquad 1 \leq j \leq n \tag{2.13}$$

3. La varianza del estadístico es

$$var(r(X_i)) = \frac{n^2-1}{12} \qquad \text{para} \qquad 1 \leq j \leq n \tag{2.14}$$

4. La covarianza del estadístico es

$$cov(r(X_i), r(X_j)) = -\frac{n+1}{12} \quad \text{para} \quad i \neq j \tag{2.15}$$

Demostración:
1.- Como $r(X_i)$ es una v.a. que puede tomar valores en $\{0,1,2,\ldots,n\}$, entonces cada uno de estos valores tiene igual probabilidad de ocurrir. Así,

$$P(r(X_i) = j) = \frac{1}{n} \qquad \text{para} \qquad 1 \leq j \leq n$$

2.- Si $Y_i = r(X_i)$, entonces

$$E(Y_i) = \sum_{j=1}^{n} j * f_Y(j) = \sum_{j=1}^{n} \frac{1}{n} * j = \frac{1}{n} \sum_{j=1}^{n} j = \frac{1}{\cancel{n}} * \frac{\cancel{n}(n+1)}{2} = \frac{n+1}{2}$$

3.- Si $Y_i = r(X_i)$ y de la definición de varianza para v.a. discretas tenemos que

$$var(Y_i) = E(Y_i^2) - (E(Y_i))^2 = \sum_{j=1}^{n} j^2 * f_Y(j) - \left(\frac{n+1}{2}\right)^2 = \frac{1}{n}\sum_{j=1}^{n} j^2 - \left(\frac{n+1}{2}\right)^2$$
$$= \frac{1}{\cancel{n}}\left(\frac{\cancel{n}(n+1)(2n+1)}{6}\right) - \left(\frac{n+1}{2}\right)^2 = (n+1)\left[\frac{n-1}{12}\right] = \frac{n^2-1}{12}$$

4.- Si $Y_i = r(X_i)$ e $Y_j = r(X_j)$ para $i \neq j$, entonces

$$cov(Y_i, Y_j) = E(Y_i Y_j) - E(Y_i)E(Y_j). \qquad (1)$$

Como (1) también es igual a

$$cov(Y_1, Y_2) = E(Y_1 Y_2) - E(Y_1)E(Y_2)\ , \qquad (2)$$

entonces el término $E(Y_1 Y_2)$ se puede expresar en términos de esperanza condicional. De esta manera, $E(Y_1 Y_2) = E(Y_1 E(Y_2|Y_1))$. Ahora bien, si Y_1 es fijo, Y_2 puede ser cualquiera de los $n-1$ valores presentes en el conjunto de los n primero enteros positivos, con probabilidad $\frac{1}{n-1}$. Esto implica que

$$E(Y_2|Y_1) = \frac{1}{n-1}\left[\frac{n(n+1)}{2} - Y_1\right]. \qquad (3)$$

Al reemplazar (3) en (2) tenemos

$$cov(Y_1, Y_2) = E(Y_1 E(Y_2|Y_1)) - E(Y_1)E(Y_2)$$
$$= \frac{1}{n-1}\left[\frac{n(n+1)}{2}\right]E(Y_1) - \frac{1}{n-1}E(Y_1^2) - E(Y_1)^2$$
$$= \frac{1}{n-1}\left[\frac{n(n+1)}{2}\left(\frac{n+1}{2}\right) - \frac{1}{n}\left(\frac{n(n+1)(2n+1)}{6}\right)\right] - \left(\frac{n+1}{2}\right)^2$$
$$= \frac{(3n+2)(n+1)}{12} - \left(\frac{n+1}{2}\right)^2 = -\frac{n+1}{12}$$

■

Ahora bien, si existe(n) alguna(s) observación(es) repetida(s) en la muestra aleatoria, entonces el rango ya no será bien definido. El problema radica en no poder determinar a ciencia cierta qué rango le corresponde a qué observación repetida. No obstante, para dar solución a este problema se presenta la teoría de ties.

2.2.2. Ties (Empates)

Dentro de un conjunto de datos existe la posibilidad de que una o varias observaciones tengan igual magnitud. Esto ocasiona varios problemas en la formulación de los test estadísticos no paramétricos, afectando de manera especial a su varianza.

Definición 2.2.2 (Ties)**.** *Si dos o más observaciones tienen la misma magnitud se dice que son valores ties.*

La consideración teórica que realizan los expertos para eliminar toda posibilidad de ocurrencia de valores ties es suponer que la cdf es continua, pero como recalcan Gibbons y Chakraborti (2011):

"En la práctica, los ties pueden ocurrir, ya sea porque la población es realmente discreta o debido a limitaciones prácticas en la precisión de la medición" (p.193).

En este sentido se debe determinar algún proceso que trate los valores ties. Los enfoques convencionales en la asignación de rangos para las observaciones repetidas pueden ser aleatorización y rangos medios.

Aleatorización

Este proceso divide en r grupos de números diferentes al arreglo ordenado (de menor a mayor) de tamaño N. Si denotamos por t_i a la frecuencia del grupo i, entonces para valores de frecuencia mayor o igual a 2, el grupo contendrá observaciones repetidas. En total existen $\prod_{i=1}^{r} t_i$ posibles asignaciones de los rangos a toda la muestra con valores ties y además $\sum_{i=1}^{r} ti = N$ (Gibbons y Chakraborti 2011).

La idea de aleatorización consiste en seleccionar por algún proceso aleatorio, una de todas las posibles asignaciones de los N primero enteros positivos (rangos) a las observaciones de la muestra.

Rangos Medios

La idea del método consiste en asignar a cada miembro de un grupo con observaciones ties, el promedio de los rangos que tendrían si las observaciones fueran diferentes. Este enfoque es el más atractivo, sin embargo, cada vez que se lo utilice es necesario incorporar una corrección para los ties, lo cual provoca que la varianza

del test estadístico se reduzca.

Existen otros tipos de procedimientos para observaciones ties [3] , sin embargo en lo que sigue se considerará el enfoque de los rangos medios.

2.2.3. Problema de una Muestra o Muestras Relacionadas

La expresión "problema de una muestra" quiere decir que vamos a realizar algún proceso de inferencia con datos que pertenecen a una sola muestra aleatoria.

Tanto para el caso de una muestra como muestras relacionadas se presenta el tradicional Sign Test (prueba de Signo) y el conocido Wilcoxon Signed-Rank Test.

2.2.4. Sign Test (Prueba de Signo)

Dentro de las aplicaciones del campo paramétrico, para el caso de una muestra, se presenta el test de teoría normal (para la varianza conocida) o el Test t-student (para varianza desconocida) (Gibbons y Chakraborti 2011), (Montgomery 2004); para probar la hipótesis

$$H_0 : \mu = \mu_0 \qquad \text{con} \qquad \mu_0 \qquad \text{un valor específico.}$$

La idea ahora es ver su análogo no paramétrico, el Sign Test.

La prueba de Signo o Sign Test, a veces llamada también prueba de signo binomial, es la más simple y la más antigua de las pruebas no paramétricas. Su versatilidad y facilidad de uso la hacen atractiva para los investigadores (Kraska-Miller 2014).

Sean $X_1, X_2, \ldots, X_N$ observaciones de una muestra aleatoria substraída de una población que tiene mediana desconocida M y distribución F_X continua y estrictamente creciente al menos en una vecindad de M (M existe y es única).

La prueba de hipótesis a realizar se refiere al valor de la mediana poblacional, es decir

$$H_0 : M = M_0 \qquad \text{con } M_0 \text{ un valor específico,} \tag{2.16}$$

[3] Para mayor detalle vea Gibbons y Chakraborti (2011).

contra cualquiera alternativa unilateral o bilateral.

El valor M_0 será el cuantil que divide el área bajo la curva de la pdf de F_X en dos partes iguales. Así (2.16) se puede reformular de la siguiente manera:

$$H_0 : \theta = P(X > M_0) = P(X < M_0) = 0.5 \ , \tag{2.17}$$

misma que es interpretado por Gibbons y Chakraborti (2011) de la siguiente manera: "Si los datos de la muestra son consistentes con el valor de la mediana hipotética, la mitad promedio de las observaciones de la muestra se encontrarán sobre M_0 y la otra mitad por debajo " (p.169).

A las observaciones que están por encima de M_0 se les asigna el signo $+$ (diferencias positivas, $X_i - M_0 > 0$) o el signo $-$ (diferencias negativas, $X_i - M_0 < 0$) caso contrario. El caso $X_i = M_0$ (también llamado tie (Pratt y Gibbons 1981)) es teóricamente imposible de ocurrir debido al supuesto de continuidad de la cdf (Kvam y Vidakovic 2007).

Si K denota a la variable aleatoria número de observaciones más grandes que M_0 (número total de signos +), entonces $K \sim Bi(N, 0.5)$ si H_0 es cierta. En consecuencia, el test no paramétrico basado en K es llamado Sign Test, y sus regiones de rechazo para la alternativa unilateral, ya sea de cola superior o inferior a un nivel de significancia α se presentan en la siguiente tabla

Tabla 2.3: Regiones de rechazo unilaterales para el Sign Test

Cola	Superior	Inferior
	[4] $H_1 : M > M_0$ o $P(X > M_0) > 0.5$	$H_1 : M < M_0$ o $P(X < M_0) < 0.5$
región de rechazo	$K \geq k_\alpha$ donde k_α es el entero más pequeño que satisface $P_\theta(K \geq k_\alpha), \theta \in \Theta_0 = \sum_{i=k_\alpha}^{N} \binom{N}{i} (0.5)^N \leq \alpha$	$K \leq k'_\alpha$ donde k'_α es el entero más grande que satisface $P_\theta(K \leq k'_\alpha), \theta \in \Theta_0 = \sum_{i=0}^{k'_\alpha} \binom{N}{i} (0.5)^N \leq \alpha$

De la Tabla 2.3, las regiones de rechazo para la alternativa bilateral

$$H_1 : M \neq M_0 \qquad \text{o} \qquad \theta = P(X > M_0) \neq 0.5 \tag{2.18}$$

se dan cuando $K > k_{\alpha/2}$ o $K < k'_{\alpha/2}$, siendo $k_{\alpha/2}$ y $k'_{\alpha/2}$ el entero más pequeño y

[4] Las regiones de rechazo para la alternativa unilateral (superior o inferior) son consultadas de Gibbons y Chakraborti (2011), Kvam y Vidakovic (2007)

grande que satisfacen

$$\sum_{i=k_{\alpha/2}}^{N} \binom{N}{i} (0.5)^N \leq \alpha/2 \qquad y \qquad \sum_{i=0}^{k'_{\alpha/2}} \binom{N}{i} (0.5)^N \leq \alpha/2 \qquad \text{respectivamente.} \tag{2.19}$$

Es importante notar que solo es necesario calcular o bien $k_{\alpha/2}$ o $k'_{\alpha/2}$, debido a que cuando la distribución binomial es simétrica, lo cual es verdadero para $\theta = 0.5$; se cumple la relación $k_{\alpha/2} = N - k'_{\alpha/2}$ (Pratt y Gibbons 1981).

Consistencia del Sign Test

La consistencia del Test de Signo para la alternativa de cola superior se basa en la definición (2.1.7). De este modo, la alternativa $H_1 : \theta > 1/2$ a un nivel de significancia α dado, se rechaza si $K > k_\alpha$ o equivalentemente si

$$P_\theta(K > k_\alpha), \theta \in \Theta_0 \leq \alpha \qquad \forall \theta \leq 1/2$$

Ahora bien,

$$\begin{aligned} P_{1/2}(K > k_\alpha) = \alpha &= P_{1/2}\left(\frac{K - E_{\theta_0}(K)}{\sigma_{\theta_0}(K)} > \frac{k_\alpha - E_{\theta_0}(K)}{\sigma_{\theta_0}(K)}\right) \text{[5]} \\ &= P_{1/2}\left(\frac{K - 0.5N}{0.5\sqrt{N}} > \frac{k_\alpha - 0.5N}{\sqrt{N}}\right) \to P_{1/2}(Z > z_\alpha), \end{aligned}$$

es decir $P_{1/2}(Z > z_\alpha) = \alpha$ para $z_\alpha = \dfrac{k_\alpha - 0.5N}{0.5\sqrt{N}}$.

De esta manera, la potencia del test bajo la alternativa de cola superior es la siguiente:

$$\begin{aligned} \Pi_N(\theta) &= P_\theta(K > k_\alpha) = P_\theta(K > 0.5N + 0.5\sqrt{N} z_\alpha) \\ &= P_\theta\left(\frac{K - N\theta}{\sqrt{N(\theta(1-\theta))}} > \frac{0.5N - N\theta + 0.5\sqrt{N} z_\alpha}{\sqrt{N(\theta(1-\theta))}}\right) \\ &\to P_\theta\left(Z > \frac{\frac{0.5N - N\theta}{\sqrt{N}} + 0.5 z_\alpha}{\sqrt{\theta(1-\theta)}}\right) = P_\theta\left(Z > \underbrace{\frac{\sqrt{N}(0.5-\theta) + 0.5 z_\alpha}{\sqrt{\theta(1-\theta)}}}_{\text{Bajo la alternativa, } 0.5 - \theta < 0}\right), \end{aligned}$$

[5] Como K sigue la distribución binomial $Bi(N, 0.5)$, entonces $E_{\theta_0}(K) = 0.5N$ y $\sigma_{\theta_0}(K) = \sqrt{N(0.5)(0.5)} = 0.5\sqrt{N}$

y si $N \to +\infty$ entonces $\Pi_N(\theta) \to 1$. Esto demuestra que el Sign Test es consistente para la alternativa de cola superior [6].

p-valor

Para la alternativa de cola superior $H_1 : M > M_0$ y k_0 el valor observado de la v.a. K, el p-valor es igual a

$$\sum_{i=k_0}^{N} \binom{N}{i} (0.5)^N \tag{2.20}$$

De manera análoga para la alternativa de cola inferior.

En cambio, para la alternativa bilateral se define K' como $min(K, N-K)$, lo cual implica que el p-valor sea igual a

$$2\sum_{i=0}^{K'} \binom{N}{i} (0.5)^N \tag{2.21}$$

Aproximación de la distribución nula del Test de Signo a la distribución normal

Como el estadístico K sigue una distribución $Bi(N,p)$, se pueden generar fácilmente tablas que contengan el valor exacto del Test de Signo para cualquier valor de N. Sin embargo, como la distribución binomial es simétrica para $p = 0.5$, entonces esta distribución se puede aproximar a la distribución normal; y debido a la naturaleza discreta de la binomial se debe usar la corrección de continuidad (Gibbons y Chakraborti 2011).

En este contexto, por ejemplo, si H_0 se rechaza en favor de la alternativa de cola superior $H_1 : M > M_0$ cuando $K \geq k_\alpha$, para

$$k_\alpha = 0.5N + 0.5 + 0.5\sqrt{N} z_\alpha \ , \qquad [7]$$

el correspondiente p-valor aproximado, dado de (2.9) es igual a

$$1 - \Phi\left(\frac{k_0 - 0.5 - 0.5N}{\sqrt{0.25N}}\right)$$

Un proceso similar se utiliza para obtener el p-valor de la alternativa unilateral de cola inferior, así como para la alternativa bilateral.

[6] Un proceso análogo determina la consistencia del test para los otros tipos de alternativas.

[7] Se sigue de que $K \sim Bi(N, 0.5)$ y de (2.8)

Problema de diferencias cero

Como se dijo previamente, el caso $X_i = M_0$ (o diferencia cero, $X_i - M_0 = 0$) es teóricamente imposible de ocurrir, pero como dicen Gibbons y Chakraborti (2011): "Las diferencias de cero pueden ocurrir y ocurren, ya sea porque el supuesto de continuidad es erróneo o debido a mediciones imprecisas" (p.171).

Una posible solución para estos casos es simplemente eliminar las observaciones que son iguales al valor hipotético M_0. Este proceso afecta al tamaño de muestra.
Una segunda idea radica en considerar a las diferencias cero como casos, tanto de signo + como de signo - (es decir, reparto equitativo para cada caso).
Finalmente, un tercer enfoque podría utilizar algún proceso aleatorio que determine el signo de estas diferencias.

Aplicación a muestras relacionadas (pareadas)

En situaciones prácticas frecuentemente las observaciones ocurren en pares. Si bien los pares entre sí pueden ser independientes, los miembros del par pueden estar relacionados de alguna manera. Una posibilidad, como mencionan Pratt y Gibbons (1981), se da al decir que: "Esta relación dentro de los pares puede estar presente debido a la naturaleza del problema, o puede ser impuesta artificialmente por el diseño, como cuando las unidades experimentales se emparejan de acuerdo con algún criterio" (p.104).

Desde esta perspectiva, los procesos de inferencia para muestras pareadas, por ejemplo, dan respuesta a la interrogante de si dos tratamientos son diferentes o si uno de ellos es "mejor" que el otro.

El proceso aplicado al caso de muestras relacionadas se presentan a continuación:

Sea $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ una muestra aleatoria de N pares y supongamos que sus diferencias $(X_1 - Y_1), \ldots, (X_N - Y_N)$ provienen de la población de diferencias con cdf continua y mediana desconocida M_D. Si K denota el número de diferencias positivas, $D_i > M_0$, donde $D_i = X_i - Y_i \ \ \forall i = 1, \ldots, N$ y M_0 un valor fijo, entonces el Sign Test se aplica sin mayor dificultad debido a que al considerar la población de diferencias se está trabajando con una sola muestra.

Un aspecto que se debe de enfatizar lo mencionan Kvam y Vidakovic (2007) al señalar que a diferencia del caso para dos poblaciones, en donde la hipótesis de igualdad de medias es equivalente a la hipótesis de que la diferencia de las medias de población es igual a cero; no ocurre lo mismo con respecto a las medianas. En

otras palabras, $M_D \neq M_X - M_Y$ para M_X e M_Y las medianas de las poblaciones $\mathscr{X}$ e $\mathscr{Y}$ respectivamente. Sin embargo, cuando las dos poblaciones de estudio son simétricas, $M_X = M_Y$ y la población de diferencias también es simétrica, entonces la media y la mediana coinciden y solo así se tendría una condición necesaria y suficiente para decir que la hipótesis de igualdad de medianas es equivalente a que la diferencia de las medianas sea cero (Gibbons y Chakraborti 2011).

Si bien el Test de Signo es perfectamente válido para realizar inferencias considerando muestras relacionadas, en lo que sigue se presenta un enfoque alternativo al test de muestras relacionadas, el Test de Signo de rango de Wilcoxon. Este test es la contraparte no paramétrica a la prueba t pareada.

2.2.5. Wilcoxon Signed-Rank Test (Test de Signo de Rango de Wilcoxon)

La construcción del estadístico K en el Sign Test solo consideraba el signo de la diferencia entre el valor de la observación y la mediana hipotética, pero ignora por completo la magnitud de estas diferencias. En virtud de esto, Frank Wilcoxon citado por Kvam y Vidakovic (2007) sugiere que: "Además del signo, el valor absoluto de estas diferencias también debería importar, y podría aumentar la eficiencia de la prueba de signo" (p.126).

Consecuentemente, suponiendo que la información de las diferencias esté disponible, que la distribución poblacional sea simétrica y continua, entonces Gibbons y Chakraborti (2011) mencionan que: " El estadístico de prueba de signo de rango de Wilcoxon proporciona una prueba de ubicación alternativa que se ve afectada por las magnitudes y los signos de estas diferencias" (p.195).

Así, en base a lo expuesto con anterioridad se procede a describir el Test de Signo de Rango de Wilcoxon.

Sea $X_1, X_2, \ldots, X_N$ una muestra aleatoria con N observaciones que provienen de una población con cdf F_X continua y simétrica cerca de una mediana desconocida M. Bajo la hipótesis nula

$$H_0 : M = M_0 \qquad \text{con } M_0 \text{ un valor específico,}$$

las diferencias $D_i = X_i - M_0$ son distribuidas simétricamente cerca de 0 (Pratt y Gibbons 1981), es decir

$$
\begin{aligned}
P(X_i \leq M_0 - x) &= P(X_i \geq M_0 + x) \qquad \forall x > 0 \\
P(X_i - M_0 \leq -x) &= P(X_i - M_0 \geq x) \\
F_D(D_i \leq -x) &= F_D(D_i \geq x) \\
F_D(D_i \leq -x) &= 1 - F_D(D_i \leq x)
\end{aligned}
$$

Si se calculan los rangos para $|D_i|$ y se conserva el signo original de las diferencias D_i, entonces los estadísticos T^+ y T^- se definirán como la suma de los rangos con signo positivo y negativo de las diferencias absolutas respectivamente. Adicionalmente se puede definir el estadístico T como $T = T^+ - T^-$.

Matemáticamente estos estadísticos se expresan como

$$
T^+ = \sum_{j=1}^{N} Z_i r(|D_i|) \qquad , \qquad T^- = \sum_{j=1}^{N} (1 - Z_i) r(|D_i|) \tag{2.22}
$$

y

$$
T = T^+ - T^- = 2\sum_{j=1}^{N} Z_i r(|D_i|) - \sum_{j=1}^{N} r(|D_i|) = 2\sum_{j=1}^{N} Z_i r(|D_i|) - \frac{N(N+1)}{2} \tag{2.23}
$$

con [8]

$$
Z_i = \begin{cases} 1 & \text{si } D_i > 0. \\ 0 & \text{si } D_i \leq 0. \end{cases} \tag{2.24}
$$

Debido a que estos tres estadísticos están relacionados de manera lineal son criterios equivalentes; razón por la cual, Pratt y Gibbons (1981) dicen que a cualquiera de estos tres estadísticos se los conoce como el estadístico de Signo de Rango de Wilcoxon (o Wilcoxon Signed-Rank Test) en honor a Frank Wilcoxon. Sin embargo, T^+ puede proporcionar una prueba de ubicación más eficiente para algunas distribuciones (Gibbons y Chakraborti 2011), motivo por el cual se lo considera en lo que sigue.

Considerando que H_0 sea verdadera, los momentos de primer y segundo orden de T^+ son

$$
\begin{aligned}
E_{\theta_0}(T^+) = E_{\theta_0}(T^-) &= \frac{N(N+1)}{4} \\
E_{\theta_0}(T) &= 0
\end{aligned} \tag{2.25}
$$

$$
\begin{aligned}
Var_{\theta_0}(T^+) = Var_{\theta_0}(T^-) &= \frac{N(N+1)(2N+1)}{24} \\
Var_{\theta_0}(T) &= 4Var_{\theta_0}(T^+)
\end{aligned} \tag{2.26}
$$

[8] La v.a. Z de (2.24) se distribuye como una Bernoulli con parámetro 0.5

Demostración:

1.- Para calcular el valor de la esperanza de T^+ nos ayudamos de (2.1). De esta manera,

$$E_{\theta_0}(T^+) = \sum_{j=1}^{N} E[r(|D_j|)E(Z_j)].$$

Pero como $r(|D_j|)$ es una permutación del conjunto $\{0,1,,2,\dots N\}$ y $Z_j \sim Be(0.5)$, entonces

$$E_{\theta_0}(T^+) = 0.5\sum_{k=1}^{N} r(|D_k|) = 0.5\left[\frac{N(N+1)}{2}\right] = \frac{N(N+1)}{4}$$

De manera análoga para $E_{\theta_0}(T^-)$.

En cambio para T, $E_{\theta_0}(T) = E(T^+ - T^-) = \cancel{\frac{N(N+1)}{4}} - \cancel{\frac{N(N+1)}{4}} = 0.$

2.- Para calcular el valor de la varianza para T^+ nos ayudamos de (2.2), así

$$\begin{aligned} Var_{\theta_0}(T^+) &= \sum_{j=1}^{N} Var(Z_j) r(|D_j|)^2 \\ &= \frac{1}{4}\left[\sum_{j=1}^{N} r(|D_j|)^2\right] = \frac{1}{4}\left[\frac{N(N+1)(2N+1)}{6}\right] = \frac{1}{24}\left[N(N+1)(2N+1)\right] \end{aligned}$$

De manera análoga para $Var(T^-|H_0)$.

En cambio para T se tiene que

$$\begin{aligned} Var_{\theta_0}(T) &= Var(T^+ - T^-) \\ &= Var\left(2\sum_{j=1}^{N} Z_i r(|D_i|) - \frac{N(N+1)}{2}\right) = 4Var\left(\sum_{j=1}^{N} Z_i r(|D_i|)\right) = 4Var_{\theta_0}(T^+) \end{aligned}$$

■

Un proceso equivalente para obtener la esperanza y la varianza de T^+ se presenta en Gibbons y Chakraborti (2011) y Pratt y Gibbons (1981). La idea es definir al estadístico de la suma de rangos con signo positivo como

$$T^+ = \sum_{1\le i\le j\le N} T_{ij} \qquad donde \qquad T_{ij} = \begin{cases} 1 & \text{si } D_i + D_j > 0 \\ 0 & \text{caso contrario} \end{cases} \tag{2.27}$$

y probabilidades p_1, p_2, p_3 y p_4 $\forall i,j,k$ distintas de la siguiente manera

$$p_1 = P(D_i > 0) \tag{2.28}$$
$$p_2 = P(D_i + D_j > 0) \tag{2.29}$$
$$p_3 = P(D_i > 0 \quad y \quad D_i + D_j > 0) \tag{2.30}$$
$$p_4 = P(D_i + D_j > 0 \quad y \quad D_i + D_k > 0) \tag{2.31}$$

De esta forma, los momentos de primer y segundo orden de la variable T_{ij} son

$$\begin{aligned}
E(T_{ii}) &= 1P(D_i + D_i > 0) + 0 = P(D_i > 0) = p_1 \qquad \forall i \\
Var(T_{ii}) &= 1P(D_i + D_i > 0) + 0 - (E(T_{ii}))^2 = p_1 - p_1^2 \qquad \forall i \\
E(T_{ij}) &= 1P(D_i + D_j > 0) + 0 = P(D_i + D_j > 0) = p_2 \qquad \forall i \neq j \\
Var(T_{ij}) &= 1P(D_i + D_j > 0) + 0 - (E(T_{ij}))^2 = p_2 - p_2^2 \qquad \forall i \neq j
\end{aligned}$$

en consecuencia, la media y la varianza de T^+ son respectivamente

$$E_{\theta_0}(T^+) = Np_1 + N(N-1)p_2/2 \tag{2.32}$$

$$\begin{aligned}
Var_{\theta_0}(T^+) = Np_1(1-p_1) + N(N-1)[p_2(1-p2)/2 + 2(p_3 - p_1p_2)] \\
+ N(N-1)(N-2)(p_4 - p_2^2)
\end{aligned} \tag{2.33}$$

Ahora bien, como las D_i se distribuyen simétricamente cerca de 0 y bajo el supuesto que H_0 es cierta, entonces $p_1 = 1/2$ (debe haber tantas diferencias positivas como negativas), $p_2 = 1/2$, $p_3 = 3/8$ y $p_4 = 1/3$ (Gibbons y Chakraborti 2011) y (Pratt y Gibbons 1981), mismos que al ser reemplazado en (2.32) y (2.33) dan respectivamente (2.25) y (2.26).

Consistencia del Wilcoxon Signed-Rank Test

Para demostrar la consistencia del Test de Wilcoxon, por ejemplo, para el caso de cola superior nos ayudamos de (2.6). De esta forma $g(p_2) = 1/2$ bajo la hipótesis nula y $g(p_2) > 1/2$ para la hipótesis alternativa.

Definimos el estadístico $T' = \dfrac{2T^+}{N^2}$ y se demostrará que es un estimador consistente [9] de p_2. Para ello, por un lado

[9] Pratt y Gibbons (1981) definen que θ_n es un estimador consistente de θ si $\lim_{n\to+\infty} E(\theta_n) = \theta$ y $\lim_{n\to+\infty} Var(\theta_n) = 0$

$$E(T') = E\left(\frac{2T^+}{N^2}\right) = \left(\frac{2}{N^2}\right) E(T^+)$$
$$= \left(\frac{2}{N^2}\right)\left[Np_1 + \frac{N(N-1)p_2}{2}\right] = \frac{2p_1}{N} + \frac{N-1}{N}p_2 = \frac{2p_1}{N} + \left[1 - \frac{1}{N}\right]p_2$$

y si $N \to +\infty$, entonces $E(T') \to p_2$.
Por otro lado,

$$Var(T') = \frac{4}{N^4}[Np_1(1-p_1) + N(N-1)[p_2(1-p2)/2 + 2(p_3 - p_1p_2)]$$
$$+ N(N-1)(N-2)(p_4 - p_2^2)]$$

y nuevamente si $N \to +\infty$, entonces $Var(T') \to 0$.

En consecuencia, T' es un estimador consistente de p_2 y de (2.5) se concluye que es un test consistente para la alternativa de cola superior con tamaño α y región de rechazo $T' \in R$ para $T' - p_2 > c_\alpha$.

Nótese que la consistencia se da ante cualquier distribución alternativa incluso no simétrica para la cual $p_2 > 1/2$ (Pratt y Gibbons 1981). Adicionalmente, si la mediana real de la población supera M_0, los datos de la muestra reflejarán esto al tener la mayoría de los rangos más grandes correspondientes a diferencias positivas (Gibbons y Chakraborti 2011).

Región de rechazo, p-valor y aproximación a la distribución normal del Test de Signo de Rango de Wilcoxon

Bajo la hipótesis nula $H_0 : M = M_0$, la distribución de probabilidad nula de T^+, misma que es simétrica cerca de su media e idénticamente distribuida como T^- se define como

$$P(T^+ = T) = \frac{u(t)}{2^N} \tag{2.34}$$

donde $u(t)$ son las formas posibles de asignar los N primeros enteros positivos a los signos + o -, de manera que la suma de estos números sea t.

La construcción de tablas con valores críticos de cola izquierda se basan en la variable aleatoria T, misma que puede ser dependiendo del caso T^+ o T^-, la que tenga su valor de suma más pequeño. Por ejemplo, en Gibbons y Chakraborti (2011) para $N \leq 15$ se exhibe la tabla H con valores críticos y niveles de significación exactos de la prueba de clasificación de Wilcoxon.
Por otro parte se pueden encontrar implementaciones computaciones de estas tablas

en el paquete estadístico Matlab o en R mediante la libreria MASS.

Si t_α es el número tal que $P(T < t_\alpha) = \alpha$, entonces las regiones de rechazo para las alternativas de H_0 de tamaño α son la siguientes

Tabla 2.4: Regiones de rechazo para las alternativas de H_0 del Wilcoxon Signed-Rank Test.

$T^- \leq t_\alpha$	para $H_1 : M > M_0$
$T^+ \leq t_\alpha$	para $H_1 : M < M_0$
$T^+ \leq t_{\alpha/2}$ o $T^- \leq t_{\alpha/2}$	para $H_1 : M \neq M_0$

Pero como en la práctica se trabaja con muestras de gran tamaño es necesario una distribución asintótica de T^+. Así, en concordancia a (2.25), (2.26) y de los resultados obtenidos para variables dependiente presentados en Hoeffding y Robbins (1948), o de manera particular el caso estudiado por Diananda (1955) , Orey (1958), Romano y Wolf (2000) cuando $f(n) = m$, se tiene que

$$Z = \frac{T^+ - E(T^+)}{Var(T^+)} = \frac{T^+ - \frac{N(N+1)}{4}}{\sqrt{\frac{1}{24}[N(N+1)(2N+1)]}} = \frac{4T^+ - N(N+1)}{\sqrt{\frac{2}{3}[N(N+1)(2N+1)]}} \tag{2.35}$$

sigue una distribución normal estándar a medida que $N \to +\infty$ (generalmente $N \geq 15$).

Si se utiliza la corrección de continuidad para mejorar la aproximación y el tamaño de la muestra es mayor a 15, entonces las regiones de rechazo apropiadas y los p-valores basados en la aproximación normal se presentan en la Tabla 2.5

Tabla 2.5: Regiones de rechazo y p-valores aproximados para las alternativas de H_0 del Test T^+

Alternativa	Región de Rechazo Aproximada	p-valor Aproximado
$M > M_0$	$T^+ \geq \frac{N(N+1)}{4} + 0.5 + z_\alpha\sqrt{\frac{N(N+1)(2N+1)}{24}}$	$1 - \Phi\left(\frac{t_0 - 0.5 - N(N+1)/4}{\sqrt{N(N+1)(2N+1)/24}}\right)$
$M < M_0$	$T^+ \leq \frac{N(N+1)}{4} - 0.5 - z_\alpha\sqrt{\frac{N(N+1)(2N+1)}{24}}$	$\Phi\left(\frac{t_0 + 0.5 - N(N+1)/4}{\sqrt{N(N+1)(2N+1)/24}}\right)$
$M \neq M_0$	ambos con $z_{\alpha/2}$	2*(el más pequeño de los anteriores)

Problema de diferencias ties para el Test de Signo de Rango de Wilcoxon

Como ya se mencionado anteriormente, los casos de diferencias ties o ceros son teóricamente imposibles de suceder bajo el supuesto de continuidad de la cdf de

la población. De existir estas diferencias se debe proceder como en la pág. 27, sin embargo, la distribución de probabilidad de T^+ no será la misma. En consecuencia se debe utilizar el proceso visto en (2.2.2) para abordar este problema.

Si el un número de ties es pequeño se produce un efecto leve y no es necesario hacer ninguna corrección (Gibbons y Chakraborti 2011), no obstante, para estimar (2.35) en presencia de un número considerable de valores ties es necesario corregir la varianza de (2.26), misma que bajo el enfoque de rango medios (véase 2.2.2, pág. 22) se transforma en

$$var(T^+) = \frac{1}{24}[N(N+1)(2N+1)] - \underbrace{\frac{\sum t(t^2-1)}{48}}_{\text{corrección para ties}} \quad [10], \tag{2.36}$$

donde el sumatorio se extiende sobre t, el número de diferencias ties para un rango determinado.

Aplicación a muestras relacionadas

La comparación entre dos tratamientos, según Wilcoxon (1945), generalmente corresponde a una de las siguientes dos categorías:

a) Tener varias repeticiones para cada uno de los dos tratamientos, que no están emparejados.

b) Tener una serie de comparaciones pareadas que nos llevan a una serie de diferencias, algunas de las cuales pueden ser positivas y otras negativas.

Desde este punto de vista, el Test de Wilcoxon se propuso en realidad para hacer inferencias sobre el valor de la mediana de la población de diferencias (en datos de muestras relacionadas (pareadas)), lo cual nos permite decir qué miembro del par es ”mejor ” que el otro.

Si en la metodología sobre la aplicación a muestras pareadas de la pág. 27 se considera adicionalmente que la población de diferencias es simétrica y que las diferencias son de la forma

$$D_i = X_i - Y_i - M_0 \qquad \forall i = 1, \ldots, N,$$

[10] La metodología de la introducción del factor de corrección para ties se presenta en Gibbons y Chakraborti (2011)

entonces el Test de Signo de Rango de Wilcoxon se aplica sin mayor limitación.

2.2.6. Problema General de dos Muestras

Como se vio anteriormente, el Test de Signo y el Test de Signo de Rango de Wilcoxon son perfectamente aplicables al problema de una muestra y muestras relacionadas. Ahora el interés radica en analizar los datos de dos muestras aleatorias que son mutuamente independientes (extraídas de manera aleatoria de sus respectivas poblaciones), no solo en sí mismas (los datos en cada muestras son independientes), sino también entre ellas (cada elemento de la primera muestra es independiente de cada elemento de la segunda muestra).

Nuestro universo consistirá de las poblaciones $\mathscr{X}$ e $\mathscr{Y}$ con cdf continuas F_X, F_Y respectivamente; y para $X_1, X_2 \ldots, X_m$ una muestra aleatoria de la población $\mathscr{X}$ e $Y_1, Y_2 \ldots, Y_n$ una muestra aleatoria de la población $\mathscr{Y}$, las cuales son extraídas de manera independiente de sus respectivas poblaciones; el problema de comparar dos muestras citado en Gibbons y Chakraborti (2011), bajo la consideración de que la única diferencia entre las poblaciones ocurre solo con respecto a algún parámetro (ej. mediana o varianza), desea determinar si las dos muestras son extraídas de poblaciones idénticas o equivalentemente

$$H_0 : F_X(x) = F_Y(x) \qquad \forall x \tag{2.37}$$

versus las alternativas de que las dos poblaciones difieren de alguna manera especificada.

Si se considera el campo paramétrico y bajo el supuesto de que las muestras provienen de poblaciones normales, los mejores test para probar la igualdad de medias e igualdad de varianzas son respectivamente la prueba t-student y la prueba F (Siegel 1956), (Wald y Wolfowitz 1939). Sin embargo, bajo la incertidumbre del cumplimiento de los supuestos que rigen el desarrollo teórico de los test antes mencionados, o si no se cuenta con información suficiente para juzgar su validez o si es apropiada una prueba de igualdad completamente general para distribuciones no especificadas; los procedimientos no paramétricos son los más recomendables.

En la práctica se hacen suposiciones sobre la forma de las poblaciones subyacentes, y con frecuencia la diferencia a nivel de "ubicación" (la diferencia es la misma independientemente del parámetro de ubicación que se elija, y es la cantidad

de cambio requerido para que las dos poblaciones sean idénticas Pratt y Gibbons (1981)) entre las dos poblaciones es de interés primordial. Por lo tanto, estudiar estas diferencias a través del llamado modelo de ubicación o modelo de cambio es muy importante.

2.2.6.1. Modelo de Ubicación

Bajo el supuesto que la única diferencia entre las poblaciones ocurre solo a nivel de ubicación, entonces el modelo propuesto es el siguiente

Definición 2.2.3. *Sean $X_1, X_2, \ldots, X_m$ y $Y_1, Y_2, \ldots, Y_n$ dos conjuntos de observaciones mutuamente independientes extraídos de las poblaciones $\mathscr{X}$ e $\mathscr{Y}$ respectivamente. Decimos que la población $\mathscr{Y}$ es la misma que la población $\mathscr{X}$ excepto posiblemente por un valor de desplazamiento desconocido θ (parámetro de ubicación) si*

$$F_Y(x) = P(Y \leq x) = P(X \leq x - \theta) = F_X(x - \theta) \qquad \forall x \ y \ \theta \neq 0. \tag{2.38}$$

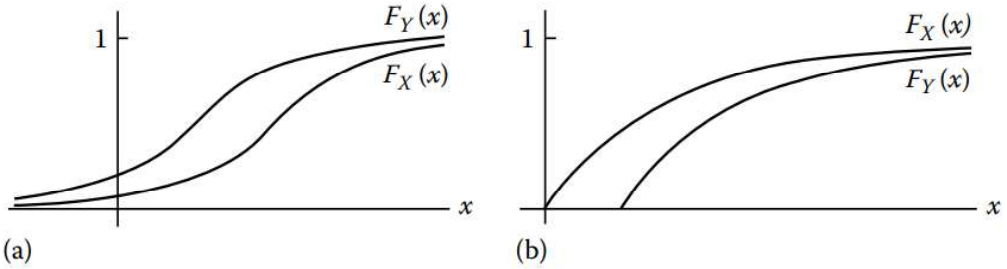
Esto significa que $\mathscr{Y}$ y $\mathscr{X} + \theta$ tienen la misma distribución de probabilidad o que la distribución de $\mathscr{X}$ es la misma que la de $\mathscr{Y} - \theta$.

Dependiendo del valor de θ tenemos los siguientes casos:

- Si $\theta = 0$, la cdf de $\mathscr{Y}$ es la misma cdf de $\mathscr{X}$.
- Si $\theta > 0$, la cdf de $\mathscr{Y}$ es la misma cdf de $\mathscr{X}$ pero trasladada a la derecha.
- Si $\theta < 0$, la cdf de $\mathscr{Y}$ es la misma cdf de $\mathscr{X}$ pero trasladada a la izquierda.

Gráficamente, (2.38) para cuando θ es mayor o menor a cero se da en la Figura 2.1.

Figura 2.1: Si $F_X(x)$ es la cdf de $\mathscr{X}$ e $F_Y(x)$ es la cdf de $\mathscr{Y}$ entonces
a) Para $\theta < 0$ (Distribución Normal) y b) Para $\theta > 0$ (Distribución Exponencial)
Tanto en el caso a) como en b), las dos poblaciones tienen la misma forma y variabilidad. La única diferencia está en que la cdf de la población $\mathscr{Y}$ está trasladada a la derecha ($\theta > 0$) o a la izquierda ($\theta < 0$) en comparación con la cdf de la población $\mathscr{X}$. Generalmente el parámetro de ubicación es igual a la diferencia entre cuantiles del mismo orden (particularmente medias o medianas).



Ahora bien, bajo el supuesto de desplazamiento las dos poblaciones tienen la mis-

ma forma y de manera particular sus varianzas, si estas existen, deben ser iguales (Pratt y Gibbons 1981). El valor de θ puede ser igual a la diferencia entre medias (si estas existen), entre medianas o de manera general a la diferencia entre cuantiles del mismo orden de las poblaciones objetas de estudio (Gibbons y Chakraborti 2011).

Otra suposición que se hace sobre la forma de las poblaciones subyacentes se da en cuanto al llamado modelo de escala. Este modelo asume que las poblaciones $\mathscr{X}$ e $\mathscr{Y}$ son las mismas excepto posiblemente por un factor de escala positivo $\theta > 0$.

La descripción de este modelo se presenta a continuación.

2.2.6.2. Modelo de Escala

De manera similar al modelo de ubicación, si consideramos que la única diferencia entre las poblaciones ocurre solo a nivel de escala, entonces la definición del modelo a utilizar es la siguiente:

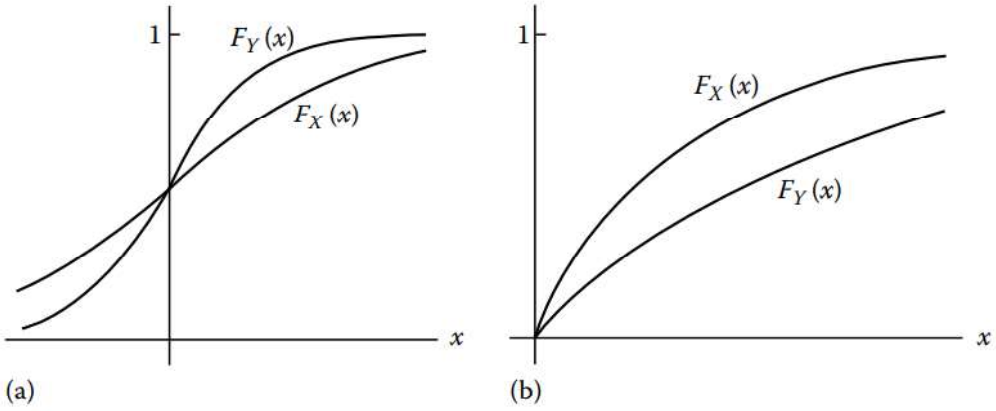
Definición 2.2.4. *Sean $X_1, X_2, \ldots, X_m$ y $Y_1, Y_2, \ldots, Y_n$ dos conjuntos de observaciones mutuamente independientes extraídas de las poblaciones $\mathscr{X}$ e $\mathscr{Y}$ respectivamente. Decimos que la población $\mathscr{Y}$ es la misma que la población $\mathscr{X}$ excepto posiblemente por un factor de escala desconocido θ (parámetro de escala) si*

$$F_Y(x) = P(Y \leq x) = P(X \leq \theta x) = F_X(\theta x) \qquad \forall x \ \ y \ \ \theta > 0 \ \ y \ \ \theta \neq 1 \tag{2.39}$$

Esto significa que $\mathscr{Y}$ e $\mathscr{X}$ $/\theta$ tienen la misma distribución de probabilidad.

Gráficamente, (2.39) para θ mayor y menor a 1 se da en la Figura 2.2.

Figura 2.2: Si $F_X(x)$ es la cdf de $\mathscr{X}$ e $F_Y(x)$ es la cdf de $\mathscr{Y}$, entonces
a) Si $\theta > 1$ (Distribución Normal) y b) Si $\theta < 1$ (Distribución Exponencial)
La cdf de la población $\mathscr{Y}$ es la misma que la cdf de la población $\mathscr{X}$ pero con una escala comprimida o expandida según $\theta > 1$ o $\theta < 1$, respectivamente.



Un contexto más general sobre la forma de las poblaciones subyacentes se obtiene al considerar el modelo de ubicación y modelo escala al mismo tiempo. Este proceso se describe a continuación.

2.2.6.3. Modelo General Ubicación-Escala

Definición 2.2.5. *Sean $X_1, X_2, \ldots, X_m$ y $Y_1, Y_2, \ldots, Y_n$ dos conjuntos de observaciones mutuamente independientes extraídas de las poblaciones $\mathscr{X}$ e $\mathscr{Y}$ respectivamente. El modelo general satisface que*

$$P(Y - \mu_Y \leq x) = P(X - \mu_X \leq \theta x) \qquad \forall x \ y \ \theta > 0 \ , \ \theta \neq 1 \ , \tag{2.40}$$

para μ_X, μ_Y las medias respectivas de $\mathscr{X}$ e $\mathscr{Y}$

La expresión en (2.40) es equivalente a decir que $\mathscr{Y} - \mu_Y$ y $\dfrac{\mathscr{X} - \mu_X}{\theta}$ siguen la misma distribución de probabilidad. Un resultado similar se obtiene al reemplazar las medias μ_X, μ_Y por las medianas poblacionales respectivas.

Independientemente del tipo de modelo que se considere, el problema general de dos muestras es el más analizado dentro de lo no paramétrico. La prueba de hipótesis a realizar considera que las cdf de las poblaciones son continuas, que la hipótesis nula se formula como en (2.37) y que las alternativas pueden ser

- $H_A : F_X(x) \neq F_Y(x)$ para algún x (alternativa bilateral)
- $H_1 : F_X(x) \geq F_Y(x)$ para todo x
  $H_1 : F_X(x) > F_Y(x)$ para algún x (alternativas general de una cola, en este caso superior)

Una alternativa para el último caso que con frecuencia se utiliza se da cuando las observaciones de una población tienden a ser más grandes que las observaciones de la otra población. A esta propiedad Gibbons y Chakraborti (2011) la llaman "estocásticamente más grande" y Pratt y Gibbons (1981) "estocásticamente dominante".
Definición 2.2.6. *Sean $\mathscr{X}$, $\mathscr{Y}$ dos poblaciones con cdf continuas F_X y F_Y respectivamente. Si se cumple que $F_X > F_Y$, entonces $\mathscr{Y}$ será estocásticamente más grande que $\mathscr{X}$. Esto se denotará por $Y >^{ST} X$.*

De la definición anterior, si $\theta > 0$ ($\theta < 0$) en (2.38), entonces $Y >^{ST} X$ ($X >^{ST} Y$). Adicionalmente, si $\theta < 1$ ($\theta > 1$) en (2.39), entonces $Y >^{ST} X$ ($X >^{ST} Y$).

En resumen, cuando el interés de nuestro proceso de inferencia radique en ver si

existe diferencia entre las poblaciones a un nivel de ubicación, se debe el modelo de ubicación, mismo que da la llamada alternativa de ubicación

$$H_L : F_Y(x) = F_X(X - \theta) \qquad \forall x \text{ y algún } \theta \neq 0 \tag{2.41}$$

En cambio, si el interés es ver la existencia de diferencias a un nivel de escala entre las dos poblaciones, se debe usar la alternativa de escala

$$H_S : F_Y(x) = F_X(\theta x) \qquad \forall x \text{ y algún } \theta \neq 1 \tag{2.42}$$

A pesar de que la literatura proporciona una gran variedad de test para el problema general de dos muestras, solo los más conocidos dentro de lo no paramétrico serán mencionados.

2.2.6.4. De Wald–Wolfowitz Runs Test

Las pruebas univariadas tradicionales como el test de desviación máxima de Smirnov o Wald-Wolfowitz Run Test mencionadas en Friedman y Rafsky (1979) son perfectamente válidas para la alternativa bilateral del problema general de dos muestras. Sin embargo, solo el segundo test genera procesos de distribución libre y será el que se describa en lo que sigue.

A. Wald y J. Wolfowitz han desarrollado un método para probar la hipótesis nula de que dos muestras independientes provienen de la misma población contra la hipótesis alternativa de que los dos grupos difieren en cualquier aspecto (ya sea a nivel de ubicación o escala) en función de "runs".
El proceso desarrollado por los autores antes mencionados se describe a continuación.

Sean $X_1, X_2, \ldots, X_m$ e $Y_1, Y_2, \ldots, Y_n$ dos muestras aleatorias independientes extraídas de las poblaciones $\mathscr{X}$, $\mathscr{Y}$ respectivamente. Suponemos que $\mathscr{X}$ tiene cdf continua F_X y $\mathscr{Y}$ cdf continua F_Y. Agrupamos las dos muestras en un solo conjunto (arreglo) y ordenamos las $N = m + n$ observaciones en orden creciente, conservando la identidad de pertenencia de cada observación a sus respectivas muestras. Bajo el supuesto de continuidad de las poblaciones, los casos ties son teóricamente imposibles de suceder, lo que garantiza la existencia de un único arreglo ordenado (menor a mayor en este caso).

La hipótesis nula a ser probada es la misma que en (2.37), es decir

$$H_0 : F_X(x) = F_Y(x) \qquad \forall x. \tag{2.43}$$

Si (2.43) fuese verdadera, se esperaría que las observaciones estén bien mezcladas dentro del arreglo agrupado-ordenado de tamaño N.

El criterio de mezcla se basa en los llamados runs.

Definición 2.2.7 (Runs)**.** *Un run se define como la sucesión de uno o más tipos de símbolos (etiquetas) que son seguidos y precedidos por un símbolo diferente.*

Por ejemplo, si tengo dos muestras de tamaños $m = 4$ e $n = 5$ y si el arreglo ordenado es $XYYXXYXYY$, en donde la etiqueta X identifica a los elementos de la muestra de tamaño 4 e Y a los de la muestra de tamaño 5, entonces el total de runs que se tiene es 6.

Esto nos da la idea de que si las muestras provienen de poblaciones idénticas, el número de runs no debería ser demasiado pequeño, pues de serlo se estaría rechazando H_0, o como lo describen Gibbons y Chakraborti (2011): "Un patrón de arreglo con muy pocos runs sugeriría que este grupo de N observaciones no es un solo ejemplo aleatorio sino que está compuesto por dos muestras de dos poblaciones distinguibles" (p.231).

Por consiguiente, se define el estadístico R llamado de Wald-Wolfowitz Runs Test, como el número total de runs dentro de la muestra agrupada-ordenada para probar H_0, y de existir diferencias entre las cdf F_X y F_Y, el valor de R tenderá a ser pequeño (Gibbons y Chakraborti 2011),(Friedman y Rafsky 1979), (Wald y Wolfowitz 1939).

La utilidad del Test R es apropiada principalmente cuando la alternativa es general y bilateral. En otras palabras, cuando la alternativa es

$$H_A : F_X(x) \neq F_Y(x) \text{ para algún } x. \tag{2.44}$$

Sin embargo, el Test R generalmente funciona para la alternativa de cola inferior. De esta manera, la respectiva región de rechazo a un nivel de significancia α será

$$R \leq c_\alpha \ , \tag{2.45}$$

donde c_α es el entero más grande que satisface $P_\theta(R \leq c_\alpha), \theta \in \Theta_0 \leq \alpha$; y su correspondiente p-valor será igual a $P_\theta(R \leq R_0), \theta \in \Theta_0$, para R_0 el valor observado del estadístico R.

En base a la distribución nula exacta y k-ésimos momentos del Test R, presentados en Gibbons y Chakraborti (2011) o en Siegel (1956), la literatura proporciona tablas con valores críticos. Por ejemplo, para valores de $m, n \leq 20$ existe la tabla F_1 presentada en Siegel (1956) o para $m \leq n \leq 12$, la tabla D en Gibbons y Chakraborti (2011). No obstante, al trabajar con muestras de gran tamaño es indispensable utilizar, bajo la hipótesis nula, la aproximación de la distribución nula del Test R a la distribución normal $\left(z = \dfrac{R - E(R)}{Var(R)}\right)$. Esta aproximación la exhibe Siegel (1956) en la tabla A, misma que contiene el p-valor en función del valor observado R_0.

Finalmente, lo concerniente a la consistencia del Test R está fuera del alcance del presente estudio, pero se desarrolla en Wald y Wolfowitz (1939).

Problema de ties para el Wald-Wolfowitz Runs Test

Al considerar que las poblaciones son continuas los casos ties son teóricamente imposibles de suceder. De hecho, la presencia de valores ties dentro de cada muestra no afecta la formulación del Test R. El problema surge cuando estos valores se presentan entre muestras, es decir, si dos o más observaciones de las diferentes muestras tienen exactamente la misma magnitud.

La idea que trata al problema de los ties radica en romperlos de todas las formas posibles, y observar todos los posibles valores resultantes del Test R. Este número calculado por Gibbons y Chakraborti (2011) es igual a

$$\prod_{i=1}^{k} \binom{s_i + t_i}{s_i} \ , \tag{2.46}$$

en donde s_i es el número de observaciones de la una muestra que tienen igual magnitud con t_i observaciones de la otra muestra dentro del grupo de ties i. Esto $\forall i = 1, 2, \ldots, k$, siendo k el total de grupos ties entre las muestras.

Si todos los valores resultantes del Test R son significativos con respecto al valor previamente establecido para α, los ties no presentan un problema importante aunque aumentan el tedio de cálculo (Siegel 1956). Sin embargo, puede darse el caso que las posibles formas de romper los ties llevan a valores de R que pueden ser significativos y otros que no lo son, lo que provocaría que la decisión sea más difícil. En este caso Siegel (1956) recomienda que se determine la probabilidad de ocurrencia asociada con cada valor posible de R y se tome el promedio de los p-valores como la probabilidad obtenida para decidir si aceptar o rechazar H_0.

La utilidad principal del Wald-Wolfowitz Run Test se da en los análisis preliminares de los datos, es decir cuando todavía no se ha formulado ninguna forma particular de alternativa; y para el caso en el que se rechaza la hipótesis nula, se pueden realizar más estudios con otras pruebas en un intento de clasificar el tipo de diferencia entre las poblaciones (Gibbons y Chakraborti 2011).

2.2.6.5. Test U de Mann-Whitney

Para establecer si dos muestras aleatorias independientes provenían de una misma población, el Test Wald-Wolfowitz consideraba la idea de mezcla aleatoria basada en runs. Ahora para dar respuesta al mismo problema nos apoyamos en el Test U de Mann-Whitney, también llamado Wilcoxon Rank-Sum Test, Test Mann-Whitney, Test Mann Whitney-Wilcoxon o simplemente prueba U.

El Test U, al igual que los runs, analiza la disposición de las observaciones en el arreglo agrupado-ordenado, pero con la diferencia de que considera la relación entre la magnitud de las observaciones de la una muestra con la magnitud de las observaciones de la otra muestra. Desde esta perspectiva, Gibbons y Chakraborti (2011) dicen que: "Si la mayoría de las Y es mayor que la mayoría de las X, o viceversa, sería evidencia contra una mezcla aleatoria y por lo tanto tendería a desacreditar la hipótesis nula de distribuciones idénticas" (p.261).

Este enfoque hace del Test Mann-Whitney una prueba eficiente al tal punto que llega a ser la contraparte más habitual a la prueba t paramétrica, cuando el investigador desea evitar la tarea de verificar el cumplimiento de todos los supuestos de la prueba t, o cuando la medición en la investigación es más débil que la escala de intervalo (Siegel 1956).

La formulación matemática se describe a continuación.

Sean $X_1, X_2, \ldots, X_m$ e $Y_1, Y_2, \ldots, Y_n$ dos muestras aleatorias independientes extraídas de las poblaciones $\mathscr{X}$, $\mathscr{Y}$, mismas que se suponen tienen cdf continuas F_X y F_Y respectivamente. Agrupamos las dos muestras en un solo arreglo, ordenamos las $N = n + m$ observaciones en forma creciente y conservamos la identidad de pertenencia de los elementos a sus respectiva muestras.
Si se etiqueta a todos los elementos de la muestra $Y_1, Y_2, \ldots, Y_n$ por y y a los de la muestra $X_1, X_2, \ldots, X_m$ por x, entonces el estadístico U de Mann-Whitney se define como el número de veces que las y preceden a las x en el arreglo agrupado-

ordenado, o equivalentemente en símbolos

$$U = \sum_{i=1}^{m}\sum_{j=1}^{n} D_{ij} \quad donde \quad Dij = \begin{cases} 1 & \text{si } Y_j < X_i \\ 0 & \text{si } Y_j > X_i \ \ \forall i = 1,\ldots,n; j = 1,\ldots,m \end{cases} \tag{2.47}$$

La expresión en (2.47) provocará el rechazo de la hipótesis nula de poblaciones idénticas a favor de la alternativa H_1 que nos dice que $Y >^{ST} X$ o $F_Y(x) \leq F_X(x)$ con desigualdad estricta para algunos x; cuando los valores de U sean pequeños.

Consistencia del Test U

Como la hipótesis nula de poblaciones idénticas se basa en la idea de mezcla aleatoria, (2.37) se puede redefinir en función del criterio de que debe haber tantas observaciones de la muestra $X_1,\ldots,X_m$ menores a las observaciones de la muestra $Y_1,\ldots,Y_n$ como observaciones de la muestra $Y_1,\ldots,Y_n$ menores a las observaciones de la muestra $X_1,\ldots,X_m$. En otras palabras, H_0 se puede parametrizar como $H_0 : p = P(X > Y) = P(X < Y) = 1/2$. Del mismo modo, las alternativas unilaterales dependiendo del caso se parametrizan por $P(Y > X) > 1/2$ o $P(Y < X) < 1/2$.

A manera de ejemplo, una posible prueba de hipótesis a realizar sería

$$H_0 : p = 0.5$$
$$\text{vs}$$
$$H_1 : p < 0.5 \ .$$

Bajo el supuesto que H_0 sea cierta, las $n * m$ v.a. D_{ij} en (2.47) siguen una distribución Bernoulli con parámetro igual a $1/2$, pero de manera general, $D_{ij} \sim Be(p)$. En consecuencia, sus momentos de primer y segundo orden son iguales a

$$E(D_{ij}) = p \ , \qquad Var(D_{ij}) = p - p^2. \tag{2.48}$$

De lo anterior, la esperanza y varianza del Test de Mann-Whitney son

$$E(U) = mnp \ ^{[11]} \tag{2.49}$$

$$\begin{aligned} Var(U) &= mnp(1-p) + mn(n-1)(p_1 - p^2) + mn(m-1)(p_2 - p^2) \\ &= mn[p - p^2(N-1) + (n-1)p_1 + (m-1)p_2]. \ ^{[12]} \end{aligned} \tag{2.50}$$

[11] La esperanza es fácil de calcular, de (2.47) y (2.48) se tiene que $E(u) = \sum_{i=1}^{n}\sum_{j=1}^{m} E(D_{ij}) = mnp$

Ahora bien, si se define $U' = \dfrac{U}{mn}$, se calcula la $E(U')$, $Var(U')$ como

$$E(U') = E\left(\frac{U}{mn}\right) = \left(\frac{1}{mn}\right) E(u) = \left(\frac{1}{\cancel{mn}}\right)[\cancel{mn}p] = p$$

y

$$\begin{aligned} Var(U') &= Var\left(\frac{U}{mn}\right) = \frac{1}{(mn)^2} Var(U) \\ &= \frac{1}{(mn)^2}\Big(mn[p - p^2(N-1) + (n-1)p_1 + (m-1)p_2]\Big) \\ &= \frac{1}{(mn)}\Big([p - p^2(N-1) + (n-1)p_1 + (m-1)p_2]\Big) \end{aligned}$$

y a medida que $m, n \to +\infty$, entonces $Var(U') \to 0$. Esto implica que U' sea un estimador consistente de p y con la ayuda de (2.5), un test consistente para la alternativa de cola inferior de tamaño α y región de rechazo $U - mn/2 < k_\alpha$ (otra forma de demostración más avanzada de la consistencia se da en Mann y Whitney (1947)).

Región de rechazo, p-valor y aproximación a la distribución normal del Test U.

Bajo la hipótesis nula la distribución de probabilidad del Test U, misma que es simétrica cerca de su media ((2.49) para p=1/2) se define como

$$f_U(\mu) = P(U = \mu) = \frac{r_{m,n}(\mu)}{\binom{m+n}{m}} \ , \tag{2.51}$$

donde $r_{m,n}(\mu)$ es el número de arreglos distinguibles (muestra agrupada-ordenada) de tal manera que el número de veces que las y preceden a las x es exactamente μ.

Bajo lo expuesto anteriormente; gracias a la propiedad de simetría, solo se necesitan calcular los valores críticos de cola inferior para obtener las regiones de rechazo tanto de cola superior como bilateral. Así, si se define el estadístico U' como el número de veces que una x precede a una y en el arreglo agrupado, entonces

$$U' = \sum_{i=1}^{n}\sum_{j=1}^{m}(1 - D_{ij}). \tag{2.52}$$

Esto genera las regiones de rechazo a un nivel se significancia α presentadas en la Tabla 2.6.

[12] La demostración de (2.50) se da en Gibbons y Chakraborti (2011), en donde se definen los parámetros p_1, p_2 de manera explícita.

Tabla 2.6: Regiones de rechazo para las alternativas de H_0 del Test Mann-Whitney

Alternativa		Región de Rechazo
$p < 1/2$ ó	$F_Y(x) \leq F_X(x)$	$U \leq c_\alpha$
$p > 1/2$ ó	$F_Y(x) \geq F_X(x)$	$U' \leq c_\alpha$
$p \neq 1/2$ ó	$F_Y(x) \neq F_X(x)$	$U \leq c_{\alpha/2}$ ó $U' \leq c_{\alpha/2}$

Para el caso de muestras pequeñas, la literatura proporciona información acerca del p-valor, mismo que simplemente necesita el valor de m, n y U para ser calculado. Por ejemplo, para valores de $m, n \leq 8$, Siegel (1956) en la parte de anexos presenta la tabla J. Sin embargo, puede darse el caso de que el valor U sea tan grande que no aparezca en la tabla J, esta situación no quiere decir que la tabla este mal sino más bien que se está considerando el caso de precedencia de manera equivocada; por lo que de darse esta situación se debe tomar el valor $U = mn - U'$.
Finalmente para cuando $m, n \to +\infty$ se debe usar la aproximación de la distribución nula del Test U a la distribución normal. Este resultado requiere de los valores $p_1 = p_2 = 1/3$ y $p = 1/2$ presentados en Gibbons y Chakraborti (2011), mismos que al ser reemplazados en (2.49) y (2.50) nos dan respectivamente que

$$E(U) = \frac{mn}{2} \tag{2.53}$$

$$Var(U) = \frac{mn(N+1)}{12} \tag{2.54}$$

De este modo,

$$\frac{U - E(U)}{Var(U)} = \frac{U - mn/2}{\sqrt{mn(N+1)/12}} \xrightarrow[d]{} N(0,1) \ , \tag{2.55}$$

y en vista de que U solo puede asumir valores enteros es necesario usar la corrección de continuidad para mejorar la aproximación.

Problema de ties para el Test Mann-Whitney

Aunque teóricamente los ties no deben de existir por la suposición de continuidad de las distribuciones de probabilidad de las poblaciones $\mathscr{X}$ e $\mathscr{Y}$, en la práctica pueden ocurrir y se presentan como en la pág. 41.
Un enfoque conservador utiliza el mayor valor de U (o U') para romper los ties de todas las formas posibles (Gibbons y Chakraborti 2011).

No obstante, una definición opcional del Test de Mann-Whitney que permite la

presencia de ties es

$$U_T = \sum_{i=1}^{m}\sum_{j=1}^{n} D_{ij} \quad donde \quad Dij = \begin{cases} 1 & \text{si } Y_j < X_i \\ 0.5 & \text{si } Y_j = X_i \\ 0 & \text{si } Y_j > X_i \ \ \forall i = 1,\ldots,n; j = 1,\ldots,m \end{cases} \tag{2.56}$$

Si se especifica que $p^+ = P(X > Y)$ y $p^- = P(X < Y)$, entonces

$$\begin{aligned} E_\theta(U_T), \theta \in \Theta_0 &= mn(p^+ - p^-) \\ Var_\theta(U_T), \theta \in \Theta_0 &= \frac{mn(N+1)}{12}\left[1 - \frac{\sum t(t^2-1)}{N(N^2-1)}\right] \ , \end{aligned} \tag{2.57}$$

donde t denota la multiplicidad de los ties y la suma se extiende sobre todos los conjuntos de valores ties (Gibbons y Chakraborti 2011).

Así, U_T estandarizado se comportará asintóticamente como la distribución normal cuando $n, m \to +\infty$.

Más adelante se presenta el Test de Mann-Whitney basado en la suma de los rangos de las observaciones con etiqueta y que son menores a las observaciones con etiquetas x. El desarrollo teórico es similar, pero con ciertas consideraciones en cuanto al cálculo de los momentos de primer y segundo orden. Mayor detalle se presenta en Corder y Foreman (2009), Siegel (1956), Pratt y Gibbons (1981), Kvam y Vidakovic (2007), Randles (2012) y Mann y Whitney (1947).

2.2.7. Estadísticos Lineales de Rango y la Generalización al Problema de dos Muestras

Anteriormente se presentaron algunos métodos para abordar el problema general de dos muestras desarrollados a partir de la muestra agrupada-ordenada. Sin embargo existen otros procesos que proporcionan información de las posibles diferencias entre las poblaciones. Estos procesos dependen de los estadísticos de rango (véase 2.2.1, pág. 19) para muestras agrupadas (también llamadas combinadas).

Los test basados en rangos para el problema general de comparar dos muestras dependerán de ciertas variables indicadoras, mismas que en función de determinadas constantes generan el denominado estadístico lineal de rango. Estos estadísticos son esenciales en la teoría no paramétrica debido a que juegan un rol fundamental

en la resolución de algunos problemas de teoría general, como por ejemplo el establecimiento de pruebas asintóticas más potentes (Hálek 1968).

A continuación se presenta la definición de rango para el arreglo agrupado.

Definición 2.2.8 (Rango dentro de la muestra agrupada). *Sean $X_1, X_2, \ldots, X_m$ e $Y_1, Y_2, \ldots, Y_n$ dos muestras aleatorias independientes sustraídas de las poblaciones $\mathscr{X}$ e $\mathscr{Y}$ con cdf F_X y F_Y respectivamente.*
Bajo la hipótesis nula de igualdad de distribuciones $H_0 : F_X(x) = F_Y(x)\ \forall x$, se define el rango para cada observación dentro de la muestra agrupada como:

$$r_{XY}(X_i) = \sum_{k=1}^{m} S(X_i - X_k) + \sum_{k=1}^{n} S(X_i - Y_k) \tag{2.58}$$

$$r_{XY}(Y_i) = \sum_{k=1}^{m} S(Y_i - X_k) + \sum_{k=1}^{n} S(Y_i - Y_k) \tag{2.59}$$

donde

$$S(u) = \begin{cases} 0 & si\ u < 0 \\ 1 & si\ u \geq 0 \end{cases}, \tag{2.60}$$

A simple vista esta manera de calcular el rango resulta ser algo tediosa, lo que nos incentiva a plantearnos un enfoque más práctico. La idea es definir el arreglo agrupado-ordenado de tamaño $N = m + n$ como un vector
$\mathcal{Z} = (Z_1, Z_2, \ldots, Z_m, Z_{m+1}, \ldots, Z_N)$ de variables indicadoras, donde

$$Z_i = \begin{cases} 1 & \text{si la i-ésima observación es una X} \\ 0 & \text{si la i-ésima observación es una Y } \ \forall i = 1, \ldots, N \end{cases} \tag{2.61}$$

Claramente (2.61) no solo da la etiqueta de a que muestra pertenece cada observación, sino también el rango asociada a la misma dentro del arreglo agrupado-ordenado. Es decir, si la variable $Z_i = 1$ su rango asociado dentro de la muestra agrupada es i y pertenece a la muestra de las X.

Bajo esta consideración, toda función lineal respecto a las variables indicadoras Z_i expresada como

$$T_N(\mathcal{Z}) = \sum_{i=1}^{N} a_i Z_i \ \text{ con } \ a_i \ \text{ constante, llamada peso o score} \tag{2.62}$$

se denomina estadístico lineal de rango.

2.2.7.1. Propiedades del Estadístico Lineal de Rango

Teorema 2.2.1. *Bajo la hipótesis nula, $H_0 : F_X(x) = F_Y(x) = F(x)$ para todo x con F desconocida, se tiene que*

$$E(Z_i) = \frac{m}{N} \quad , \quad Var(Z_i) = \frac{mn}{N^2} \quad y \quad cov(Z_i, Z_j) = -\frac{mn}{N^2(N-1)} \tag{2.63}$$

Demostración:
De (2.61) tenemos que $Z_i \; \forall i = 1, \ldots, N$ sigue una distribución Bernoulli con pmf

$$f_{Z_i}(z_i) = \begin{cases} m/N & \text{si } z_i = 1 \ , \\ n/N & \text{si } z_i = 0 \ , \\ 0 & \text{caso contrario,} \end{cases}$$

de esta manera $E(Z_i) = m/N$ y $Var(Z_i) = (m/N)(n/N) = mn/N^2$.

Por otro lado para $i \neq j$,

$$cov(Z_i, Z_j) = E(Z_i Z_j) - E(Zi)E(Zj) = P(Z_i = 1 \cap Z_j = 1) - \left(\frac{m}{N} * \frac{m}{N}\right) = \frac{\binom{m}{2}}{\binom{N}{2}} - \left(\frac{m}{N}\right)^2$$

$$= \frac{m(m-1)\cancel{(m-2)!}}{\cancel{2!}\cancel{(m-2)!}} \Big/ \frac{N(N-1)\cancel{(N-2)!}}{\cancel{2!}\cancel{(N-2)!}} - \left(\frac{m}{N}\right)^2$$

$$= \frac{m(m-1)}{N(N-1)} - \left(\frac{m}{N}\right)^2 = -\frac{mn}{N^2(N-1)}$$

■

Teorema 2.2.2. *Bajo la hipótesis nula, $H_0 : F_X(x) = F_Y(x) = F(x)$ para todo x, entonces*

$$\begin{aligned} E(T_N) &= m \sum_{i=1}^{N} \frac{a_i}{N} \\ Var(T_N) &= \frac{mn}{N^2(N-1)} \left[N \sum_{i=1}^{N} a_i^2 - \left(\sum_{i=1}^{N} a_i \right)^2 \right] \end{aligned} \tag{2.64}$$

Demostración:

Como $T_N(\mathcal{Z}) = \sum_{i=1}^{N} a_i Z_i$ y en base a los resultados (2.1) , (2.63) tenemos que

$$\begin{aligned} E(T_N) = E\left(\sum_{i=1}^{N} a_i Z_i\right) &= \sum_{i=1}^{N} a_i E(Z_i) \\ &= \sum_{i=1}^{N} a_i \frac{m}{N} = m \sum_{i=1}^{N} \frac{a_i}{N} \end{aligned}$$

Ahora de (2.2) se tiene que

$$\begin{aligned} Var(T_N) &= \sum_{i=1}^{N} a_i^2 var(Z_i) + 2 \sum_{1 \le i < j \le N} \sum a_i a_j cov(Z_i, Z_j) \\ &= \frac{mn}{N^2} \sum_{i=1}^{N} a_i^2 - \frac{2mn}{N^2(N-1)} \sum_{1 \le i < j \le N} \sum a_i a_j \\ &= \frac{mn}{N^2} \left[\sum_{i=1}^{N} a_i^2 - \frac{2}{N-1} \sum_{1 \le i < j \le N} \sum a_i a_j \right] \\ &= \frac{mn}{N^2(N-1)} \left[N \sum_{i=1}^{N} a_i^2 - \sum_{i=1}^{N} a_i^2 - 2 \sum_{1 \le i < j \le N} \sum a_i a_j \right] \\ &= \frac{mn}{N^2(N-1)} \left[N \sum_{i=1}^{N} a_i^2 - \left(\sum_{i=1}^{N} a_i^2 + 2 \sum_{1 \le i < j \le N} \sum a_i a_j \right) \right] \\ &= \frac{mn}{N^2(N-1)} \left[N \sum_{i=1}^{N} a_i^2 - \left(\sum_{i=1}^{N} a_i^2 \right)^2 \right] \end{aligned}$$

■

Teorema 2.2.3. *Si* $B_N = \sum_{i=1}^{N} b_i Z_i$ *y* $T_N = \sum_{i=1}^{N} a_i Z_i$ *son dos estadísticos lineales de rango, bajo la hipótesis nula,* $H_0 : F_X(x) = F_Y(x) = F(x) \ \ \forall x$,

$$cov(B_N, T_N) = \frac{mn}{N^2(N-1)} \left(N \sum_{i=1}^{N} a_i b_i - \sum_{i=1}^{N} a_i \sum_{i=1}^{N} b_i \right) \tag{2.65}$$

Demostración:

Por definición de covarianza tenemos que

$$\begin{aligned}
cov(B_N, T_N) &= E(B_N * T_N) - E(B_N)E(T_N) \\
&= E\left(\sum_{i=1}^{N} b_i Z_i * \sum_{i=1}^{N} a_i Z_i\right) - \left(\frac{m}{N}\sum_{i=1}^{N} b_i\right)\left(\frac{m}{N}\sum_{i=1}^{N} a_i\right) \\
&= \sum_{i=1}^{N} a_i b_i E(Z_i)^2 + E(Z_i Z_j)\sum_{i=1}^{N} b_i \sum_{i=1}^{N} a_i - E(Z_i Z_j)\sum_{i=1}^{N} a_i b_i - \left(\frac{m}{N}\right)^2 \sum_{i=1}^{N} a_i \sum_{i=1}^{N} b_i \\
&= \frac{m}{N}\sum_{i=1}^{N} a_i b_i + \frac{m(m-1)}{N(N-1)}\sum_{i=1}^{N} b_i \sum_{i=1}^{N} a_i - \frac{m(m-1)}{N(N-1)}\sum_{i=1}^{N} a_i b_i - \left(\frac{m}{N}\right)^2 \sum_{i=1}^{N} a_i \sum_{i=1}^{N} b_i \\
&= \sum_{i=1}^{N} a_i b_i \left(\frac{m}{N} - \frac{m(m-1)}{N(N-1)}\right) + \sum_{i=1}^{N} a_i \sum_{i=1}^{N} b_i \left(\frac{m(m-1)}{N(N-1)} - \frac{m^2}{N^2}\right) \\
&= \left(\frac{m(N-m)}{N(N-1)}\right)\sum_{i=1}^{N} a_i b_i + \left(\frac{m(m-N)}{N^2(N-1)}\right)\sum_{i=1}^{N} a_i \sum_{i=1}^{N} b_i \ ,
\end{aligned}$$

y dado que $N = m + n$ entonces

$$\begin{aligned}
cov(B_N, T_N) &= \left(\frac{mn}{N(N-1)}\right)\sum_{i=1}^{N} a_i b_i - \left(\frac{mn}{N^2(N-1)}\right)\sum_{i=1}^{N} a_i \sum_{i=1}^{N} b_i \\
&= \frac{mn}{N^2(N-1)}\left[N\sum_{i=1}^{N} a_i b_i - \sum_{i=1}^{N} a_i \sum_{i=1}^{N} b_i\right].
\end{aligned}$$

■

Por otra parte, dado que el número total de vectores $\mathcal{Z}$ distinguibles (es decir, la disposición de los m unos y n ceros en el arreglo combinado) es igual a $\binom{m+n}{m}$; bajo H_0 la probabilidad de ocurrencia de un $\mathcal{Z}$ será $1/\binom{m+n}{m}$ (ed. todos los vectores $\mathcal{Z}$ distinguibles son igualmente probables de ocurrir). Esto implica que la distribución de probabilidad nula de cualquier estadístico lineal de rango sea igual a

$$P(T_N(\mathcal{Z}) = k) = \frac{t(k)}{\binom{m+n}{m}} \ , \tag{2.66}$$

donde $t(k)$ es el número de arreglos de las m X y n Y v.a. talque $T_N(Z) = k$.

Supongamos ahora que si para cada vector $\mathcal{Z}$ de m unos y n ceros existe un vector conjugado $\mathcal{Z}'$ de m ceros y n unos, de modo que si $T_N(\mathcal{Z}) = \mu + k$ y $T_N(\mathcal{Z}') = \mu - k$, entonces la distribución de $T_N(\mathcal{Z})$ es simétrica cerca de su media μ. Esto permite

establecer la condición de simetría para cualquier estadístico lineal de rango como

$$T_N(\mathcal{Z}) + T_N(\mathcal{Z}') = 2\mu. \tag{2.67}$$

Los teoremas siguientes establecen las condiciones que se deben cumplir para garantizar la simetría del estadístico lineal de rango.

Teorema 2.2.4. *La distribución nula de $T_N(\mathcal{Z})$ es simétrica cerca de su media $\mu = \frac{m}{N}\sum_{i=1}^{N} a_i$ si las constantes a_i satisfacen*

$$a_i + a_{N-i+1} = c \quad \textit{donde } c \textit{ es una constante } \forall i = 1,2,\ldots,N \tag{2.68}$$

Demostración:
Se define el vector conjugado de $\mathcal{Z}$, $\mathcal{Z}'$ como $\mathcal{Z}' = (Z'_1, Z'_2, \ldots, Z'_N)$, en donde $Z'_i = Z_{N-i+1}$ $\forall i = 1,2,\ldots,N$, luego

$$\begin{aligned} T_N(\mathcal{Z}) + T_N(\mathcal{Z}') &= \sum_{i=1}^{N} a_i Z_i + \sum_{i=1}^{N} a_i Z'_i = \sum_{i=1}^{N} a_i Z_i + \sum_{i=1}^{N} a_i Z_{N-i+1} \\ &= \sum_{i=1}^{N} a_i Z_i + \sum_{j=1}^{N} a_{N-j+1} Z_j = \sum_{i=1}^{N} (a_i + a_{N-i+1}) Z_i = c\sum_{i=1}^{N} Z_i = cm \end{aligned}$$

y como $E(T_N(\mathcal{Z}) + T_N(\mathcal{Z}')) = 2\mu = cm$ se sigue el resultado. ■

Teorema 2.2.5. *La distribución nula de $T_N(\mathcal{Z})$ es simétrica con respecto a su media para cualquier conjunto de constantes a_i, si $m = n = N/2$*

Demostración:
En este caso se define el vector conjugado de $\mathcal{Z}$, $\mathcal{Z}'$ como $Z'_i = 1 - Z_i$, así

$$T_N(\mathcal{Z}) + T_N(\mathcal{Z}') = \sum_{i=1}^{N} a_i Z_i + \sum_{i=1}^{N} a_i Z'_i = \sum_{i=1}^{N} a_i Z_i + \sum_{i=1}^{N} a_i (1 - Z_i) = \sum_{i=1}^{N} a_i \ ,$$

luego como $\mu = \frac{m}{N}\sum_{i=1}^{N} a_i$ y $N = m + n = 2m = 2n$, entonces

$$T_N(\mathcal{Z}) + T_N(\mathcal{Z}') = \frac{\mu N}{m} = 2\mu$$

■

Teorema 2.2.6. *La distribución nula de $T_N(\mathcal{Z})$ es simétrica cerca de su media si N es par y*

las constantes a_i satisfacen que

$$a_i = i \quad \forall i \leq N/2 \qquad y \qquad a_i = N - i + 1 \quad \forall i > N/2$$

Demostración:
Se definen las componentes del vector $\mathcal{Z}'$ por

$$\begin{aligned} Z'_i &= Z_{i+N/2} \qquad \forall i \leq N/2 \\ Z'_i &= Z_{i-N/2} \qquad \forall i > N/2. \end{aligned}$$

Si en

$$\begin{aligned} T_N(\mathcal{Z}) + T_N(\mathcal{Z}') &= \sum_{i=1}^{N} a_i Z_i + \sum_{i=1}^{N} a_i Z'_i \\ &= \underbrace{\sum_{i=1}^{N/2} i Z_i + \sum_{i=N/2+1}^{N/2} (N-i+1) Z_i}_{a} \\ &\quad + \sum_{i=1}^{N/2} i Z_{i+N/2} + \sum_{i=N/2+1}^{N/2} (N-i+1) Z_{i-N/2} \ , \end{aligned}$$

se hacen los cambio de índices $j = i + N/2$ y $k = i - N/2$, entonces

$$\begin{aligned} T_N(\mathcal{Z}) + T_N(\mathcal{Z}') &= a + \sum_{j=1+N/2}^{N} \left(j - \frac{N}{2} \right) Z_j + \sum_{k=1}^{N/2} \left(N - k - \frac{N}{2} + 1 \right) Z_k \\ &= \sum_{i=1}^{N/2} i Z_i + \sum_{i=N/2+1}^{N/2} (N-i+1) Z_i \\ &\quad + \sum_{j=1+N/2}^{N} \left(j - \frac{N}{2} \right) Z_j + \sum_{k=1}^{N/2} \left(N - k - \frac{N}{2} + 1 \right) Z_k \\ &= \sum_{i=1}^{N/2} \left(\frac{N}{2} - i + 1 + i \right) Z_i + \sum_{j=N/2+1}^{N} \left(j - \frac{N}{2} + N - j + 1 \right) Z_j \\ &= \left(\frac{N}{2} + 1 \right) \left[\sum_{i=1}^{N/2} Z_i + \sum_{j=N/2+1}^{N} Z_j \right] = \left(\frac{N}{2} + 1 \right) \left[\sum_{i=1}^{N} Z_i \right] \\ &= m \left(\frac{N}{2} + 1 \right) = m \left(\frac{N+2}{4} \right). \qquad (*) \end{aligned}$$

Por otro lado en vista que como

$$\begin{aligned}\sum_{i=1}^{N} a_i &= \sum_{i=1}^{N/2} i + \sum_{i=N/2+1}^{N} (N+1-i) = \sum_{i=1}^{N/2} i + \sum_{i=N/2+1}^{N} (N+1) - \sum_{i=N/2+1}^{N} i \\ &= \sum_{i=1}^{N/2} i + \sum_{i=N/2+1}^{N} (N+1) - \sum_{i=1}^{N} i + \sum_{i=1}^{N/2} i \\ &= 2\frac{\left(\frac{N}{2}\right)\left(\frac{N}{2}+1\right)}{2} + (N+1)\left(N - \frac{N}{2}\right) - \frac{N(N+1)}{2} \\ &= \frac{N^2+2N}{4}\ ,\end{aligned}$$

al ser igualado con $\sum_{i=1}^{N} a_i = \frac{N\mu}{m}$ (resultado de (2.64)) nos da que $\mu = \frac{N+2}{4}$, entonces al reemplazar este valor en (*) se obtiene que $T_N(\mathcal{Z}) + T_N(\mathcal{Z}') = 2u$. ■

También se presentan las siguientes propiedades que surgen como consecuencia de los teoremas mencionados anteriormente.

1. **Propiedad 1.** Para $T = \sum_{i=1}^{N} a_i Z_i$ y $T' = \sum_{i=1}^{N} a_i Z_{N-i+1}$, entonces

$$T = T' \quad si \quad a_i = a_{N-i+1}\ , \qquad \forall i = 1,2,\ldots,N.$$

2. **Propiedad 2.** Para $T = \sum_{i=1}^{N} a_i Z_i$ y $T' = \sum_{i=1}^{N} a_i (1 - Z_i)$, entonces

$$T + T' = \sum_{i=1}^{N} a_i.$$

3. **Propiedad 3.** Para $T = \sum_{i=1}^{N} a_i Z_i$ y $T' = \sum_{i=1}^{N} a_i (1 - Z_{N-i+1})$, entonces

$$T + T' = \sum_{i=1}^{N} a_i \quad si \quad a_i = a_{N-i+1}\ , \qquad \forall i = 1,2,\ldots,N.$$

Es preciso mencionar que para muestras de gran tamaño y considerando que m/n permanece constante, la distribución de probabilidad de un estadístico lineal de rango puede aproximarse a la distribución normal estándar (generalización del teorema central de límite) bajo ciertas condiciones de regularidad que están relacionadas

principalmente con la suavidad y tamaño de las constantes a_i (Gibbons y Chakraborti 2011). De hecho, en el paper de Chernoff y Savage (1957) se menciona que los autores Wald and Wolfowitz, Noether, Hoeffding, Lehmann, Madow y Dwass han dado condiciones suficientes para garantizar la normalidad asintótica de la distribución de probabilidad de cualquier estadístico lineal de rango.

No obstante, bajo el deseo de verificar la conjetura dada por Hodges y Lehmann (1956) y considerando supuestos que limitan el crecimiento de los pesos y proporcionan ciertas propiedades de suavidad, Chernoff y Savage (1957) extienden el resultado de normalidad asintótica dado por los autores mencionados anteriormente, con el fin de cubrir más situaciones. Estos supuestos son los numerales 1, 2, 3 y 4 descritos en el Teorema 2.2.7. Pero antes de mencionarlos es necesario exhibir ciertos aspectos que ayudan a tener una mejor interpretación de los mismos.

Sean $X_1, X_2, \ldots, X_m$, $Y_1, Y_2, \ldots, Y_n$ las observaciones ordenadas de una muestra aleatoria provenientes de las poblaciones con cdf continua $F(x)$, $G(x)$ respectivamente. Sea $N = m + n$, $\lambda_N = m/N$ y supongamos que todas las N desigualdades $0 < \lambda_0 \leq \lambda_N \leq 1 - \lambda_0 < 1$ se mantienen fijas para algunos $\lambda_0 \leq 1/2$.
Si para $F_m(x) = (\text{número de } X_i \leq x)/m$ y $G_n(x) = (\text{número de } Y_i \leq x)/n$ las funciones de distribución acumulativa de las muestras X e Y respectivamente, se define

$$H_N(x) = \lambda_N F_m(x) + (1 - \lambda_N) G_n(x)$$

como la función de distribución acumulativa de muestra combinada, entonces

$$H(x) = \lambda_N F(x) + (1 - \lambda_N) G(x)$$

será la función de distribución acumulativa poblacional combinada.

En este contexto, todo estadístico lineal de rango cuyos pesos son funciones de los rangos, puede ser representado equivalente como

$$T_N = m \int_{-\infty}^{\infty} J_N[H_N(x)] dF_m(x) \ ,$$

donde $J_N[H_N(x)]$ es una función que define los pesos a_i que entran en el estadístico lineal de rango.
A manera de ejemplo, si consideramos el caso más simple $J_N[H_N(x)] = H_N(x)$ y

$a_i = i/N$, entonces $J_N(i/N) = a_i$. Esto implica que

$$\begin{aligned} T_N &= m\int_{-\infty}^{\infty} H_N(x)dF_m(x) = \frac{m}{N}\int_{-\infty}^{\infty}[mF_m(x) + nG_n(x)]dF_m(x) \\ &= \frac{m}{N}\int_{-\infty}^{\infty}(\text{número de observaciones en la muestra combinada} \leq x) \\ &\times (1/m \text{ si } x \text{ es el valor de una v.a. } X \text{ y } 0 \text{ caso contrario}) \\ &= \frac{1}{N}\sum_{i=1}^{N} iZ_i = \sum_{i=1}^{N} a_i Z_i \end{aligned}$$

Ahora bien, si K es una constante genérica que puede depender de J_N pero no de $F_(x), G(x), m, n, N$, junto con que o_p o O_p son declaraciones que involucran uniformidad en $F(x), G(x)$ y λ_N en el intervalo $0 < \lambda_0 \leq \lambda_N \leq 1 - \lambda_0 < 1$ y además que I_N es el intervalo en el que $0 < H_N(x) < 1$ (Chernoff y Savage 1957), entonces el teorema siguiente proporciona la aproximación normal asintótica de cualquier estadístico lineal de rango.

Teorema 2.2.7. *Si*

1. $\lim_{N\to+\infty} J_N(H) = J(H)$ *existe para* $0 < H < 1$ *y no es una constante,*
2. $\int_{-I_N} [J_N(H_N) - J(H_N)]dF_m(x) = o_p(N^{-1/2})$,
3. $J_N(1) = o(\sqrt{N})$,
4. $|J^{(i)(H)}| = \left|\frac{d^i J}{dH^i}\right| \leq K[H(1-H)]^{-I-1/2+\delta}$ *para* $i = 0,1,2$ *y para algún* $\delta > 0$,

entonces, para F, G *y* λ_N *fijos,*

$$\lim_{N\to+\infty} P\left(\frac{T_N - \mu_N}{\sigma_N} \leq t\right) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}dx\,,$$

donde

$$\mu_N = \int_{-\infty}^{\infty} J[H(x)]dF(x)$$

y

$$\begin{aligned} N\sigma_N^2 = 2(1-\lambda_N)\Bigg\{ &\iint_{-\infty<x<y<\infty} G(x)[1-G(y)]J'[H(x)]J'[H(y)]dF(x)dF(y) \\ &+ \frac{1-\lambda_N}{\lambda_N}\iint_{-\infty<x<y<\infty} F(x)[1-F(y)]J'[H(x)]J'[H(y)]dG(x)dG(y)\Bigg\} \end{aligned}$$

siempre que $\sigma_N \neq 0$.

La demostración de este teorema está más allá del alcance del presente trabajo de titulación, sin embargo, si el lector desea verla a detalle puede revisar la demostración del Teorema 1 en la sección 3 del paper Chernoff y Savage (1957).

Para finalizar, como ya se ha dicho en varias ocasiones, la alternativa general de la hipótesis nula para el problema general de dos muestras es ver si las dos poblaciones no son idénticas. Aquí radica la importancia de los estadísticos lineales de rango, ya que con una elección adecuada de los coeficientes a_i, mismos que dependen del tipo de diferencia entre las poblaciones que uno espera detectar, se pueden realizar alternativas más complejas.

El tipo de situación más simple, mismo que es objeto de estudio, se da cuando uno espera que la diferencia entre las dos poblaciones solo se dé a nivel de ubicación o escala. Llamaremos a cada uno de estos casos como el problema de ubicación y el problema de escala para dos muestras respectivamente.

2.2.8. Test de Rango Lineal para el Problema de Ubicación

En el problema general de dos muestras que se describió con anterioridad, la idea consistía en determinar si las dos muestras extraídas de manera aleatoria y cada una independiente de la otra provenían de poblaciones idénticas. Ya sea que la diferencia ocurriese a nivel de ubicación o de escala, una de las poblaciones estaba afectada por el parámetro θ, mismo que dependiendo del caso podía tomar cierto valor específico.

En lo paramétrico y considerando que la diferencia ocurre en relación a algún parámetro, la prueba t-student para probar la igualdad de medias y la prueba F de Fisher para probar la igualdad de varianzas son los mejores test (con relación al marco de Neyman-Pearson) (Gibbons y Chakraborti 2011),(Montgomery 2004). Sin embargo, el cumplimiento de cada uno de los supuestos del modelo general con errores normalmente distribuidos en la práctica o debido a la naturaleza propia de los datos no llega a efectuarse (Conover e Iman 1981). Por esta razón, el conocimiento previo de la función de distribución poblacional es una limitación que presentan los procedimientos paramétricos. No obstante, el abordar un problema dentro de lo no paramétrico, no requiere especificación alguna de la distribución, lo cual resulta atractivo para el investigador al momento de realizar algún proceso de inferencia.

El trabajar con los rangos de las observaciones dentro de la muestra agrupada-ordenada proporciona información sobre el tamaño relativo de cualquier cuantil (de manera particular para las medianas poblacionales). Por ejemplo, al considerar (2.38) con $\theta < 0$ se tendría que la mediana de la población $\mathscr{Y}$ es menor a la mediana de la población $\mathscr{X}$ o que $X >^{ST} Y$.

El test basado en el estadístico lineal de rango para detectar diferencias a nivel de ubicación que se estudiará es el Wilcoxon Rank-Sum Test.

2.2.8.1. El Wilcoxon Rank-Sum Test (Test de Suma de Rangos de Wilcoxon)

El problema general de comparar dos muestras que considera el modelo de ubicación (problema de ubicación) descrito en (2.38) puede ser reformulado para cualquier alternativa unilateral. A manera de ejemplo, para el caso cola inferior se tendría que

$$\begin{aligned} H_0 : \theta = 0 \quad & (H_0 : F_Y(x) = F_X(x) \ \ \forall x) \\ & \quad vs \\ H_1 : \theta < 0 \quad & (H_L : F_X(x) \leq F_Y(x) \quad X >^{ST} Y). \end{aligned} \tag{2.69}$$

Si a cada observación dentro del arreglo agrupado de tamaño $N = m + n$ se le asigna su correspondiente rango, entonces, el decir que la mediana de la población $\mathscr{X}$ es más grande que la mediana de la población $\mathscr{Y}$ se verá reflejado en los rangos. En otras palabras, el suponer que lo afirmado es verdadero implica que los rangos de las observaciones de la muestra que provienen de la población $\mathscr{X}$ son más grandes que los que provienen de la otra población.

Basado en esta idea, Wilcoxon considera $a_i = i\,,$ para todo $i = 1,2,\ldots,N$ en (2.62) y propone el estadístico

$$W_N = \sum_{i=1}^{N} iZ_i, \tag{2.70}$$

llamado de Wilcoxon Ranks-Sum Test para la alternativa unilateral H_1.

La interpretación que se da a W_N no es otra que cosa que sumar todos los rangos correspondientes a las observaciones de la muestra que provienen de la población $\mathscr{X}$; y si H_1 es verdadera, entonces W_N debe ser grande.

Los posibles valores que puede tomar W_N para cuando los tamaños de muestra están en la relación $m \leq n$ son

$$\underbrace{\frac{m(m+1)}{2}}_{\text{valor mínimo}} \quad y \quad \underbrace{\frac{m}{2}(2N-m+1)}_{\text{valor máximo}} \ . \tag{2.71}$$

La demostración de estos valores se presenta a continuación.

Demostración:
El caso del valor mínimo se da cuando todas las observaciones de la muestra $X_1, X_2, \ldots, X_m$ son menores a la observación más pequeña de la muestra $Y_1, Y_2, \ldots, Y_n$. Así,

$$W_N = \sum_{i=1}^{N} iZ_i = \sum_{i=1}^{m} i = \frac{m(m+1)}{2}$$

En cambio para el valor máximo, todas las observaciones de la muestra que provienen de la población $\mathscr{X}$ son mayores a la observación más grande de la muestra que proviene de la población $\mathscr{Y}$. En este sentido

$$\begin{aligned} W_N = \sum_{i=1}^{N} iZ_i &= \sum_{i=N-m+1}^{N} i = \sum_{i=1}^{N} i - \sum_{i=1}^{N-m} i \\ &= \frac{N(N+1)}{2} - \frac{(N-m)(N-m+1)}{2} = \frac{m}{2}(2N-m+1) \end{aligned}$$

■

La esperanza y varianza de (2.70) se calculan a partir de (2.64). De esta forma, la esperanza será

$$E(W_N) = m\sum_{i=1}^{N} \frac{i}{N} = m\frac{\cancel{N}(N+1)}{2\cancel{N}} = \frac{m(N+1)}{2} \tag{2.72}$$

y la varianza será

$$\begin{aligned} Var(W_N) &= \frac{mn}{N^2(N-1)}\left[N\sum_{i=1}^{N} i^2 - \left(\sum_{i=1}^{N} i\right)^2\right] \\ &= \frac{mn}{N^2(N-1)}\left[N\left(\frac{N(N+1)(2N+1)}{6}\right) - \left(\frac{N(N+1)}{2}\right)^2\right] \\ &= \frac{mn}{N^2(N-1)}\left[\frac{1}{12}\left(N^2(N+1)(2(2N+1)-3(N+1)\right)\right] \\ &= \frac{1}{12}\left[\frac{mn}{\cancel{N^2}(N-1)}\right][\cancel{N^2}(N+1)][\cancel{N-1}] = \frac{mn}{12}(N+1) \ . \end{aligned} \tag{2.73}$$

Lo referente a la simetría de W_N se produce cuando el valor c de (2.2.4) es igual a $N+1$.

Demostración:
En base a la demostración del Teorema 2.2.4 y considerando que $a_i = i \ \ \forall i = 1,2,\ldots,N$ entonces

$$c = \frac{2}{N}\sum_{i=1}^{N} a_i = \frac{\cancel{2}}{\cancel{N}}\left(\frac{\cancel{N}(N+1)}{\cancel{2}}\right) = N+1$$

■

Para muestras pequeñas la distribución de probabilidad nula de W_N se sigue de (2.66). Es así que ya existen tablas que contienen esta probabilidad, por ejemplo, en Gibbons y Chakraborti (2011) se presenta la tabla J en la parte de los anexos, misma que funciona para $1 \leq m \leq n \leq 10$. Sin embargo, para muestras de gran tamaño nos ayudamos de los resultados obtenidos por los estadísticos lineales de rango. Así, la distribución de probabilidad de un estadístico lineal puede aproximarse a la distribución normal estándar sin mayor dificultad.

Problema de ties para el Wilcoxon Rank-Sum Test

En caso de existir observaciones repetidas entre las muestras, la varianza de W_N se verá afectada. Este hecho provoca que sea necesario usar la corrección para valores ties. De este modo (2.73) se reescribe como

$$\begin{aligned} Var(W_N) &= \frac{mn}{N^2(N-1)}\left[N\left(\sum_{i=1}^{N} i^2 - \underbrace{\frac{\sum t(t^2-1)}{12}}_{\text{corrección para ties}}\right) - \left(\sum_{i=1}^{N} i\right)^2\right] \\ &= \frac{mn}{N^2(N-1)}\left[N\left(\frac{N(N+1)(2N+1)}{6} - \frac{\sum t(t^2-1)}{12}\right) - \left(\frac{N(N+1)}{2}\right)^2\right] \\ &= \frac{mn}{N^2(N-1)}\left[N^2(N+1)\left(\frac{2N+1}{6} - \frac{N+1}{4}\right) - \frac{N}{12}\sum t(t^2-1)\right] \\ &= \frac{mn(N+1)}{12} - \frac{mn\sum t(t^2-1)}{12N(N-1)}, \end{aligned} \tag{2.74}$$

donde t es el número de observaciones ties para un rango arbitrario y el sumatorio se extiende sobre todo el conjunto de rangos con valores ties.

Equivalencia con el Test U de Mann-Whitney

Si comparamos la estructura del estadístico U con la estructura del Wilcoxon Rank-Sum Test se puede apreciar que existe cierta similitud en su formulación. Esto se debe a que U (U') se definió como el número de veces que las observaciones con etiqueta y (x) precedían a las observaciones de la muestra con etiqueta x (y) y el Test W_N como la suma de los rangos de las observaciones $X_1, X_2, \ldots, X_m$. En este sentido, si consideramos los Test U' y W_N, ambos definidos para la alternativa unilateral $H_1 : F_X(x) \leq F_Y(x)$ o $X >^{ST} Y$, provocarán el rechazo de H_0 a favor de H_1 para valores de U' y W_N pequeños.

Si recordamos la expresión (2.47), el $\sum_{j=1}^{n} D_{ij}\ \forall i = 1,2,\ldots,m$ representa el número de valores j para el cual $Y_j < X_i$. Pero este valor es equivalente al rango de X_i ($r(X_i)$) disminuido una cantidad t_i, misma que representa a la cantidad de observaciones $X_1, X_2, \ldots, X_m$ que son menores o iguales a X_i. En otras palabras, el Test U puede ser reescrito como

$$\begin{aligned} U &= \sum_{i=1}^{m}\sum_{j=1}^{n} D_{ij} = \sum_{i=1}^{m}[r(X_i) - t_i] = \sum_{i=1}^{m} r(X_i) - \sum_{i=1}^{m} t_i \\ &= \sum_{i=1}^{N} iZ_i - (1+2+3+\ldots+m) = W_N - \frac{m(m+1)}{2}\,, \end{aligned} \tag{2.75}$$

que simplemente es una función lineal respecto al estadístico W_N. Un proceso totalmente análogo se sigue para obtener que $U' = W'_N - \dfrac{n(n+1)}{2}$, donde W'_N es la suma de los rangos de las observaciones de la muestra $Y_1, Y_2, \ldots, Y_n$.

Existen otras expresiones del Test U que se desarrollan a partir de (2.75) y de que $U + U' = mn$. Estos resultados se muestran en Pratt y Gibbons (1981), Siegel (1956), Kraska-Miller (2014), Corder y Foreman (2009) y se expresan como

$$U = mn - W'_N + \frac{n(n+1)}{2} \tag{2.76}$$

$$U' = mn - W_N + \frac{m(m+1)}{2}. \tag{2.77}$$

En consideración a todo lo anterior, Gibbons y Chakraborti (2011) mencionan que estos dos test son equivalentes; y por lo tanto sus propiedades son las mismas, incluida la consistencia (también véase en Mann y Whitney (1947)).

Región de rechazo, p-valor y aproximación a la distribución normal del Wilcoxon Rank-Sum Test.

Las regiones de rechazo y p-valor exacto del Test de Wilcoxon Rank-Sum se presenta en la Tabla 2.7.

Tabla 2.7: Regiones de rechazo y p-valores exactos para las alternativas de H_0 del Wilcoxon Rank-Sum Test

Alternativa	Región de Rechazo	p-valor Exacto
$\theta < 0 \ \ (Y <^{ST} X)$	$W_N \geq w_\alpha$	$P(W_N \geq w_0)$
$\theta > 0 \ \ (Y >^{ST} X)$	$W_N \leq w'_\alpha$	$P(W_N \leq w_0)$
$\theta \neq 0$	$W_N \geq w_\alpha$ ó $W_N \leq w'_\alpha$	2*(el más pequeño de los anteriores)

Para arreglos de tamaño $N > 15$, las regiones de rechazo apropiadas y los p-valores aproximados basados en la aproximación normal con corrección de continuidad, para w_0 es el valor observado de W_N se exhiben en la Tabla 2.8.

Tabla 2.8: Regiones de rechazo y p-valores aproximados para las alternativas de H_0 del Wilcoxon Rank-Sum Test

Alternativa	Región de Rechazo Aproximada	p-valor Aproximado
$\theta < 0$	$W_N \geq \frac{m(N+1)}{2} + 0.5 + z_\alpha \sqrt{\frac{mn(N+1)}{12}}$	$1 - \Phi\left(\frac{w_0 - 0.5 - m(N+1)/2}{\sqrt{mn(N+1)/12}}\right)$
$\theta > 0$	$W_N \leq \frac{m(N+1)}{2} - 0.5 - z_\alpha \sqrt{\frac{mn(N+1)}{12}}$	$\Phi\left(\frac{w_0 + 0.5 - m(N+1)/2}{\sqrt{mn(N+1)/12}}\right)$
$\theta \neq 0$	ambos con $z_{\alpha/2}$	2*(el más pequeño de los anteriores)

2.2.9. Otros Test de Rango Lineal para el Problema de Ubicación

De manera general, cualquier conjunto de constantes a_i monótonas crecientes usadas en un estadístico lineal de rango proporcionarán una prueba consistente para el problema de ubicación (Gibbons y Chakraborti 2011). En este sentido, el Test de Suma de Rangos de Wilcoxon puede generalizarse para cualquier tipo de constantes que no son necesariamente rangos. A este nuevo tipo de constantes denotadas por c_k $\forall k = 1, 2, \ldots, N$ las llamaremos scores, en donde el score c_k se asigna a la observación X_i que tenga rango k dentro de la muestra agrupada-ordenada.

La suma de los scores para las observaciones de la muestra $X_1, X_2, \ldots, X_m$ en base a (2.61) y (2.70) se expresa como

$$\sum_{k=1}^{N} c_k Z_k \ , \tag{2.78}$$

lo cual es muy similar a la formulación de W_N. De hecho si se asigna a $c_k = k$ para todo $k = 1,2,\ldots,N$ se tiene exactamente W_N.

A continuación se presenta pero de manera no tan detallada, otros test para el problema de ubicación.

2.2.9.1. Terry–Hoeffding (Normal Scores) Test

La idea de los autores Gibbons y Chakraborti (2011), Pratt y Gibbons (1981), Klotz (1964) y Terry (1951) es asignar a c_k el valor esperado del i-ésimo estadístico de orden más pequeño correspondiente a una muestra de N variables aleatorias que tienen distribución normal estándar. De este manera, el estadístico que se obtiene será

$$c_1 = \sum_{i=1}^{N} E(\zeta_i) Z_i \tag{2.79}$$

mismo que es conocido como Fisher–Yates normal scores Test o c_1 Terry-Hoeffding Test.

Los valores para $E(\zeta_i)$, denominado "score normal" calculados mediante la aproximación de Bloms para N igual a 25,50,100,200 y 400, se dan en Harter (1961). Sin embargo en la actualidad estos valores se pueden obtener fácilmente de cualquier programa estadístico para un N arbitrario.

Además como $E(c_1) = 0$ y $Var(c_1) = \dfrac{mn}{N(N-1)} \sum_{i=1}^{N} E(\zeta_i)^2$ [13], la distribución de probabilidad nula de c_1 será simétrica cerca del origen; y para valores de $m \leq n \leq 10$, la tabla 1 presentada en Terry (1951) brinda información acerca de la distribución de probabilidad nula exacta de c_1.

Finalmente, Chernoff y Savage (1957) demuestran que este estadístico es asintóticamente normal y al menos tan eficiente como el t Test. Por lo tanto, Terry (1951) dice que

$$\frac{c_1}{\sqrt{Var(c_1)}} \xrightarrow[d]{} N(0,1) \text{ a medida que } N \to +\infty. \tag{2.80}$$

No obstante, una mejor aproximación asintótica de la distribución de probabilidad

[13] La demostración de la media se obtiene directamente de (2.2.4) y para la varianza en Terry (1951).

nula de c_1 para muestras grandes se da en Klotz (1964), misma expresa que

$$t = \frac{r(N-2)^{1/2}}{(1-r^2)^{1/2}} \quad donde \quad r = \frac{c_1}{(Var(c_1)(N-1))^{1/2}} \tag{2.81}$$

se distribuye aproximadamente como la t-student con $N-2$ grados de libertad.

2.2.9.2. Van der Waerden Test

Si el valor de $E(\zeta_i)$ en (2.79) se aproxima por $\Phi^{-1}\left(\frac{i}{N+1}\right)$, donde $\Phi(x)$ es la distribución normal estándar, entonces se tiene el Van der Waerden Test. Su formulación es la siguiente:

$$X_N = \sum_{i=1}^{N} \Phi^{-1}\left(\frac{i}{N+1}\right) Z_i. \tag{2.82}$$

En este caso los c_k se encuentran fácilmente en tablas de la distribución normal estándar acumulativa o en cualquier software estadístico. Por otro parte, de la similitud que existe con el c_1 Test se tiene que

$$\begin{aligned} E(X_N) &= 0\,, \\ Var(X_N) &= \frac{mn}{N(N-1)} \sum_{i=1}^{N} E\left(\Phi^{-1}\left(\frac{i}{N+1}\right)\right)^2 \end{aligned} \tag{2.83}$$

y que la distribución de probabilidad nula de X_N sea simétrica cerca de 0; además que el valor X_N estandarizado se aproximará a la distribución normal estándar.

Un resultado importante que se muestra en Gibbons y Chakraborti (2011) dice que el X_N Test es asintóticamente equivalente al c_1 Test. Por lo tanto, para muestras de tamaño grande el Test de Van de Waerden tendrá las mismas propiedades que el Test Terry-Hoeffding.

2.2.10. Test de Rango Lineal para el Problema de Escala

Como se vio anteriormente, los Test Wilcoxon Rank-Sum, Terry-Hoeffding o Van de Waerden eran sensibles para detectar diferencias entre dos poblaciones a nivel de ubicación. Pero si el interés radica en detectar diferencias de las poblaciones en un sentido de variabilidad o dispersión, entonces el modelo de escala es el más adecuado a usar.

La idea intuitiva es simplemente considerar que dos poblaciones son las mismas excepto posiblemente por un factor de escala desconocido. Es decir, que una población tiene una escala ya sea comprimida o expandida según el valor de θ en comparación con la otra (véase 2.42, pág. 39). A modo de ejemplo, si $\theta > 1$ se tendría que la población $\mathscr{Y}$ es la misma que la población $\mathscr{X}$, excepto que $\mathscr{Y}$ está comprimida con relación a $\mathscr{X}$ ($X >^{ST} Y$).

Al considerar el campo paramétrico y bajo el supuesto que las dos poblaciones son normales con medias desconocidas, la prueba de Fisher para probar la igualdad de varianzas es las más adecuada de usar dentro del marco de Neyman-Person (Gibbons y Chakraborti 2011), (Montgomery 2004). Pero valga la redundancia, en la práctica conocer previamente la forma de la distribución poblacional no es una tarea sencilla, mucho menos decir que la misma es normal. Por lo tanto, si hay razones para cuestionar los supuestos inherentes al modelo general con errores normalmente distribuidos y el interés radica en ver si existen diferencias entre dos poblaciones a nivel de escala, entonces es apropiado usar una prueba no paramétrica de dispersión.

En virtud de lo anterior y dado que la medianas son parámetros de ubicación habituales en los procedimientos de distribución libre, mismas que siempre existen, (2.40) se redefine para el caso no paramétrico como

$$\begin{aligned} P(Y - M_Y \leq x) &= P(X - M_X \leq \theta x) \\ F_{Y-M_Y}(x) &= F_{X-M_X}(\theta x) \qquad \forall x \text{ y } \theta > 0 \ , \ \theta \neq 1 \end{aligned} \tag{2.84}$$

donde M_X, M_Y son las medianas de las poblaciones $\mathscr{X}$,$\mathscr{Y}$ respectivamente.

No obstante, Gibbons y Chakraborti (2011) dicen que como en (2.84), las características respectivas de ubicación y dispersión están inextricablemente mezcladas en la muestra agrupada, las posibles diferencias de ubicación pueden enmascarar las diferencias de dispersión. Por lo tanto, si se conocen las medianas poblaciones M_X y M_Y se sugiere que las observaciones muestrales se ajusten por

$$X_i' = X_i - M_X \quad y \quad Y_i' = Y_i - M_Y \ \ \forall i = 1,2,\ldots,m \ \forall j = 1,2,\ldots,n \ , \tag{2.85}$$

para que las poblaciones $\mathscr{X}'$,$\mathscr{Y}'$ tengan medianas iguales a cero y sus respectivas observaciones en el arreglo agrupado especifiquen diferencias a nivel de escala que no estén afectadas por las diferencias de ubicación.

Si bien en la práctica no es posible saber a ciencia cierta el valor de las medianas

poblaciones, se puede considerar que $M_X = M_Y = M$, para M un valor desconocido. De esta forma, (2.84) se escribe como la alternativa general de escala de la siguiente manera

$$H_S : F_{Y-M}(x) = F_{X-M}(\theta x) \qquad \forall x \;\; y \;\; \theta > 0 \;\; , \;\; \theta \neq 1 \;\; , \tag{2.86}$$

misma que puede ser reformulada para cualquier alternativa unilateral. Por citar un caso, la alternativa de cola superior será

$$\begin{aligned} H_0 : \theta = 1 \quad & (H_0 : F_Y(x) = F_X(x) \;\; \forall x) \\ & \quad vs \\ H_1 : \theta > 1 \quad & (H_S : F_X(x) \leq F_Y(x) \;\; \text{o} \;\; X >^{ST} Y). \end{aligned} \tag{2.87}$$

En consecuencia, la elección adecuada de las constantes en el estadístico lineal de rango puede proporcionar información sobre la distribución relativa de las observaciones sobre su valor central común, y generar test sensibles para detectar diferencias a nivel de escala. De este modo, si la población $\mathscr{X}$ tiene una mayor dispersión que la población $\mathscr{Y}$, entonces las observaciones de la primera población se deben colocar aproximadamente de manera simétrica en ambos extremos de las observaciones de la segunda población. Esto condiciona a que las constantes a_i sean simétricas (pesos pequeños cerca de la mitad y grandes en los dos extremos, o viceversa).

Se han propuesto algunas pruebas basadas en los rangos de las observaciones para el problema de la escala, de entre ellas se presentan el Mood Test, Freund–Ansari–Bradley Test, Siegel-Tukey Test y Klotz Normal-Scores Test.

2.2.10.1. The Mood Test

En Mood (1954) se menciona que debido a los resultados tan exitosos obtenidos de los test desarrollados a partir de los estadísticos lineales de rango ante el problema de comparar dos muestras para detectar diferencias a nivel de ubicación, es natural investigar la eficiencia de un test basado en los rangos para detectar diferencias a nivel de dispersión. Y dado que la dispersión es una medida de propagación alrededor del valor central (media) de alguna población, entonces una primera idea sugiere tomar la desviación del rango de la i-ésima observación en el arreglo agrupado de tamaño N (m de la muestra X y n de la muestra Y) con respecto al rango medio, es decir la diferencia

$$r(XY_i) - \frac{N+1}{2} \;\; \text{, donde } XY \text{ representa el arreglo agrupado;} \tag{2.88}$$

para luego asignar este valor a las constantes a_i en (2.62). Sin embargo, el enfoque en (2.88) lleva implícito el problema descrito en el ejemplo siguiente

Ejemplo

Caso 1: $XYXYXY$ Caso 2: $XXYYYX$

$$\sum \left(r(XY_i) - \frac{N+1}{2} \right) Z_i = -1.5 \qquad \sum \left(r(XY_i) - \frac{N+1}{2} \right) Z_i = -1.5$$

El valor de la dispersión en ambos casos es el mismo, pero se aprecia claramente que en el caso 2, las observaciones de la muestra X están más dispersas que las observaciones de la muestra Y.

Este acontecimiento se corrobora con lo expresado por Gibbons y Chakraborti (2011) al decir que: "El hecho de que las desviaciones se dividan por igual entre números positivos y negativos presenta un problema al usar estas desviaciones reales como ponderaciones al construir estadísticas de rango lineal" (p.314).
Así, con la finalidad de dar solución a este problema, Mood citado en Gibbons y Chakraborti (2011) dice que se utilice los valores al cuadrado de estas desviaciones para dar igual peso a ambos extremos del valor central. De esta manera, el estadístico

$$M_N = \sum_{i=1}^{N} \underbrace{\left(i - \frac{N+1}{2} \right)^2}_{d_i} Z_i \,, \tag{2.89}$$

denominado Mood Test, donde Z_i se define de la misma manera que en (2.62), es sensible a detectar diferencias entre dos poblaciones a nivel de escala.

Por consiguiente, dado que los valores más grandes d_i se encuentran en los extremos del arreglos agrupado-ordenado, entonces valores grandes (pequeños) de M_N dan como resultado que las observaciones de la muestra X (Y) estén más dispersas. Por otro lado, para garantizar la simetría de las constantes d_i se considera que para el caso N impar, $d_{(N+1)/2} = 0$.

Bajo la hipótesis nula de (2.87) y del Teorema 2.2.2, tenemos que los momentos de primer y segundo orden de M_N son respectivamente

$$E(M_N) = \frac{m(N^2-1)}{122} \tag{2.90}$$

$$Var(M_N) = \frac{mn(N+1)(N^2-4)}{180} \quad \text{[14]} \tag{2.91}$$

Demostración:

Para la esperanza,

$$\begin{aligned} E(M_N) &= \frac{m}{N}\sum_{i=1}^{N}\left(i-\frac{N-1}{2}\right)^2 = \frac{m}{N}\left[\sum_{i=1}^{N} i^2 - (N+1)\sum_{i=1}^{N} i + \left(\frac{N+1}{2}\right)^2 \sum_{i=1}^{N} 1\right] \\ &= \frac{m}{\cancel{N}}\left[\frac{\cancel{N}(N+1)(2N+1)}{6} - \frac{\cancel{N}(N+1)^2}{2} + \frac{\cancel{N}(N+1)^2}{4}\right] \\ &= \frac{m}{12}\left[(N+1)\left(2(2N+1) - 6(N+1) + 3(N+1)\right)\right] = \frac{m}{12}(N^2-1) \end{aligned}$$

Para la varianza,

$$\begin{aligned} Var(M_N) &= \frac{mn}{N^2(N-1)}\left[N\sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)^4 - \left(\sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)^2\right)^2\right] \\ &= \frac{mn}{N^2(N-1)}\left[N\left(\sum_{i=1}^{N} i^4 - 4\left(\frac{N+1}{2}\right)\sum_{i=1}^{N} i^3 + 6\left(\frac{N+1}{2}\right)^2\sum_{i=1}^{N} i^2\right.\right. \\ &\left.\left. -4\left(\frac{N+1}{2}\right)^3\sum_{i=1}^{N} i + \left(\frac{N+1}{4}\right)^4\sum_{i=1}^{N} 1\right) - \left(\frac{N(N^2-1)}{12}\right)^2\right], \end{aligned} \tag{1}$$

si

$$\begin{aligned} \sum_{i=1}^{N} i^2 &= \frac{N(N+1)(2N+1)}{6} \\ \sum_{i=1}^{N} i^3 &= \left(\frac{N(N+1)}{2}\right)^2 \\ \sum_{i=1}^{N} i^4 &= \frac{N(N+1)(2N+1)(3N^2+3N-1)}{30} \end{aligned}$$

se reemplazan en (1), entonces

$$Var(M_N) = \frac{mn(N+1)(N^2-1)}{180}.$$

■

De la consideración que $m = n$ y del Teorema 2.2.5, entonces tenemos que la distribución de probabilidad nula de M_N es simétrica cerca de $E(M_N)/2$. Valores críticos exactos del Mood Test para muestras pequeñas (m=n=5 y m=n=10) se dan en Laubscher, Steffens y Lange (1968). Pero para muestras de tamaño grande es recomendable

[14] La demostración de estos momentos se presenta también en Mood (1954), Laubscher, Steffens y Lange (1968) y Gibbons y Chakraborti (2011).

utilizar la aproximación de la distribución M_N a la distribución normal, que dependiendo de si se utilice o no la corrección de continuidad, Laubscher, Steffens y Lange (1968) presentan tablas en las que comparan estos valores.

2.2.10.2. The Freund–Ansari–Bradley Test

Otra solución para evitar el problema inherente en (2.88), en el que las dispersiones positivas y negativas se equilibraban, es utilizar los valores absolutos de esas desviaciones. De este modo, el estadístico lineal de rango correspondiente es

$$A_N = \sum_{i=1}^{N} \left| i - \frac{N+1}{2} \right| z_i = (N+1) \sum_{i=1}^{N} \left| \frac{i}{N+1} - \frac{1}{2} \right| z_i \ , \tag{2.92}$$

llamado Freund-Ansari-Bradley Test.
Tablas con los valores críticos exactos del A_N Test se presentan en la tabla 1 de Wiel (s.f.).

Dependiendo de ciertas variaciones que se dé a (2.92), la literatura presenta alternativas de esta formulación, lo que ocasiona cierta confusión al momento de decir qué test pertenece a qué autor. Sin embargo, todos los test resultantes de estas variaciones son equivalentes.

A modo de ejemplo, una alternativa para A_N es realizar un proceso contrario al Mood Test, es decir asignar pesos más pequeños a los extremos de la muestra agrupada-ordenada. Esto da como resultado el test

$$\begin{aligned} F_N &= \sum_{i=1}^{N} \left(\frac{N+1}{2} - \left| i - \frac{N+1}{2} Z_i \right| \right) \\ &\text{o} \\ F_N &= \sum_{i=1}^{\lfloor (N+1)/2 \rfloor} i Z_i + \sum_{i=\lfloor (N+1)/2 \rfloor + 1}^{N} (N - i + 1) Z_i \ , \end{aligned} \tag{2.93}$$

que para cuando N es par las observaciones del centro tomarán el valor de $N/2$ y para N impar el valor de $(N+1)/2$.

La idea detrás de (2.93) dentro de la muestra agrupada-ordenada es dar el valor de 1 a la observación más pequeña y más grande, el valor de 2 a las segundas observaciones más pequeña y más grande; y así sucesivamente. En otras palabras, los valores pequeños (grandes) de F_N indican mayor dispersión para la muestra que

proviene de la población $\mathscr{X}$ ($\mathscr{Y}$) o de manera equivalente, valores pequeños (grandes) de F_N sugieren que $\theta > 1$ ($\theta < 1$) (Ansari y R. A. Bradley 1959).

De la equivalencia que existe con (2.92), Gibbons y Chakraborti (2011) establecen que

$$F_N = \frac{m(N+1)}{2} - A_N \ , \tag{2.94}$$

lo cual nos permite determinar que la esperanza F_N es igual a

$$E(F_N) = \underbrace{\frac{m}{4}[N+2]}_{\text{para } N \text{ par}} \qquad \text{o} \qquad E(F_N) = \underbrace{\frac{m}{4N}[N+1]^2}_{\text{para } N \text{ impar}} \tag{2.95}$$

Demostración:
En base a (2.93) y el Teorema 2.2.1 tenemos dos casos.
Para N par

$$E(F_N) = \sum_{i=1}^{\lfloor N/2 \rfloor} iE(Z_i) + \sum_{i=\lfloor N/2 \rfloor+1}^{N} (N-i+1)E(Z_i) = \frac{m}{N}\left[\sum_{i=1}^{N/2} i + \sum_{i=(N/2)+1}^{N} (N-i+1)\right] \ ,$$

haciendo el cambio de índices $j = N - i + 1$, entonces

$$E(F_N) = \frac{m}{N}\left[\sum_{i=1}^{N/2} i + \sum_{j=1}^{N/2} j\right] = \frac{2m}{N}\left[\sum_{i=1}^{N/2} i\right] = \frac{2\cancel{m}}{\cancel{N}}\left[\frac{\cancel{N}}{\cancel{2}}\left(\frac{N}{2}+1\right)\right] = \frac{m}{4}[N+2].$$

Para N impar

$$\begin{aligned} E(F_N) &= \frac{m}{N}\left[\sum_{i=1}^{\lfloor (N+1)/2 \rfloor} i + \sum_{i=\lfloor (N+1)/2 \rfloor+1}^{N} (N-i+1)\right] \\ &= \frac{m}{N}\left[\sum_{i=1}^{\lfloor (N+1)/2 \rfloor - 1} i + \frac{N+1}{2} + \sum_{i=\lfloor (N+1)/2 \rfloor+1}^{N} (N-i+1)\right] \ , \end{aligned}$$

haciendo el cambio de índices $j = N - i + 1$, entonces

$$\begin{aligned} E(F_N) &= \frac{m}{N}\left[\sum_{i=1}^{(N+1)/2-1} i + \frac{N+1}{2} + \sum_{j=1}^{(N-1)/2} j\right] = \frac{m}{N}\left[2\sum_{i=1}^{(N-1)/2} i + \frac{N+1}{2}\right] \\ &= \frac{m}{N}\left[\left(\frac{N-1}{2}\right)\left(\frac{N-1}{2}+1\right) + \frac{N+1}{2}\right] = \frac{m}{4N}(N+1)^2. \end{aligned}$$

■

Por otra parte, la varianza de F_N sea igual a

$$Var(F_N) = \underbrace{\frac{mn}{N-1}\left[\frac{N^2-4}{48}\right]}_{\text{para } N \text{ par}} \quad \text{o} \quad Var(F_N) = \underbrace{\frac{mn}{N^2}\left[\frac{N+1}{48}(N^2+3)\right]}_{\text{para } N \text{ impar}} \tag{2.96}$$

Demostración:

En base al Teorema 2.2.2 y de (2.94) tenemos que

$$\begin{aligned} Var(F_N) = Var(A_N) &= \frac{mn}{N^2(N-1)}\left[N\sum_{i=1}^{N}\left|i-\frac{N+1}{2}\right|^2 - \left(\sum_{i=1}^{N}\left|i-\frac{N+1}{2}\right|\right)^2\right] \\ &= \frac{mn}{N^2(N-1)}\left[N\sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)^2 - \left(\sum_{i=1}^{N}\left(i-\frac{N+1}{2}\right)\right)^2\right] \\ &= \frac{mn}{N^2(N-1)}\left[N\left(\frac{N(N^2-1)}{12}\right) - \left(\frac{N}{m}E(A_N)\right)^2\right] \\ &= \frac{mn}{N^2(N-1)}\left[\frac{N^2(N^2-1)}{12} - \left(\frac{N}{2}(N+1) - \frac{N}{m}E(F_N)\right)^2\right]. \end{aligned}$$

Si N es par, entonces

$$\begin{aligned} Var(F_N) &= \frac{mn}{N^2(N-1)}\left[\frac{N^2(N^2-1)}{12} - \left(\frac{N}{2}(N+1) - \frac{N}{\cancel{m}}\left(\frac{\cancel{m}}{4}(N+2)\right)\right)^2\right] \\ &= \frac{mn}{\cancel{N^2}(N-1)}\left[\frac{\cancel{N^2}(N^2-1)}{12} - \left(\frac{\cancel{N^2}}{4}\right)^2\right] = \frac{mn}{N-1}\left[\frac{N^2-4}{48}\right]. \end{aligned}$$

Si N es impar, entonces

$$\begin{aligned} Var(F_N) &= \frac{mn}{N^2(N-1)}\left[\frac{N^2(N^2-1)}{12} - \left(\frac{N}{2}(N+1) - \frac{\cancel{N}}{\cancel{m}}\left(\frac{\cancel{m}}{4\cancel{N}}(N+1)^2\right)\right)^2\right] \\ &= \frac{mn}{N^2(N-1)}\left[\frac{N^2(N^2-1)}{12} - \left(\frac{N^2-1}{4}\right)^2\right] \\ &= \frac{mn}{N^2(N-1)}\left[(N^2-1)\left(\frac{4N^2-3(N^2-1)}{48}\right)\right] = \frac{mn}{N^2}\left[\frac{N+1}{48}(N^2+3)\right]. \end{aligned}$$

■

Demostraciones alternativas se dan en Ansari y R. A. Bradley (1959).

La distribución de probabilidad nula del F_N Test para tamaños de muestras pe-

queños ($m,n \leq 10$) se presenta en la tabla 1 de Ansari y R. A. Bradley (1959). Sin embargo, para muestras de gran tamaño se debe utilizar la aproximación de la distribución de F_N a distribución normal, misma que si utiliza la corrección de continuidad dará resultados más ventajosos. Estos resultados se presentan en la tabla 2 de Ansari y R. A. Bradley (1959), la cual contiene la distribución de probabilidad nula exacta de F_N, así como también su aproximación mediante la distribución normal.

2.2.10.3. The Siegel-Tukey Test

Un enfoque alternativo para desarrollar un test que sea sensible a las diferencias a nivel escala se realiza mediante el reordenamiento de los rangos de las observaciones en el arreglo agrupado-ordenado. La idea es asignar el rango 1 a la observación que tenga la magnitud más pequeña en el arreglo, los rangos 3 y 2 a las dos observaciones más grandes respectivamente en el arreglo, los rangos 4 y 5 a las dos siguientes observaciones más pequeñas respectivamente; y así sucesivamente. Esta alternativa la proponen Siegel y Tukey (1960) mediante la siguiente formulación.

$$S_N = \sum_{i=1}^{N} a_i Z_i \quad [15] \qquad donde \qquad a_i = \begin{cases} 2i & \text{para i par, } 1 < i \leq N/2 \\ 2i-1 & \text{para i impar, } 1 \leq i \leq N/2 \\ 2(N-i)+2 & \text{para i par, } N/2 < i \leq N \\ 2(N-i)+1 & \text{para i par, } N/2 < i \leq N \end{cases} \tag{2.97}$$

El Test S_N en (2.97) asigna los rangos más bajos y más altos a los extremos y a la mitad del arreglo respectivamente. Esto ocasiona que se rechaza la hipótesis nula a favor de la alternativa para valores pequeños de S_N, debido a que esto indicaría que las observaciones de la muestra X tenderían a mostrar más variabilidad en comparación con las observaciones de la muestra Y.

En virtud que la distribución de probabilidad nula de S_N es la misma que la del W_N Test (véase 2.2.8.1, pág. 57); los momentos, tablas de distribución de probabilidad, aproximación normal y regiones de rechazo del S_N test son las mismas que el W_N Test. A pesar de esto, Siegel y Tukey (1960) presentan la tabla 2 que contiene valores exactos y aproximados de la distribución del Test S_N para tamaños de

[15] Si N es impar no se asigna rango a la observación que está en centro del arreglo, esto con el fin de que el rango más alto asignado sea par (Gibbons y Chakraborti 2011), (Siegel y Tukey 1960).

muestras $m \leq n \leq 20$. En cuanto a las regiones de rechazo exactas y aproximadas, simplemente en la Tabla 2.8 se debe cambiar $\theta < 1$, $\theta > 1$, $\theta \neq 1$ por los casos $\theta < 0$, $\theta > 0$, $\theta \neq 0$ respectivamente; S_N en lugar de W_N y s_0 por w_0, donde s_0 será valor observado de S_N.

2.2.10.4. The Klotz Normal-Scores Test

Si se utiliza el proceso visto en el Mood Test junto con el valor de las constantes del Van der Waerden Test, entonces se da el llamado score normal o Klotz Normal-Scores Test. La formulación es la siguiente:

$$K_N = \sum_{i=1}^{N} \left(\Phi^{-1} \left(\frac{i}{N+1} \right) \right)^2 Z_i. \tag{2.98}$$

El rechazo de H_0 se da para valores grandes de K_N, debido a que los pesos más grandes se encuentran en los extremos del arreglo agrupado-ordenado. En consecuencia, las observaciones que provienen de la muestra X tienen mayor variabilidad.

Por otro lado, los momentos de primer y segundo orden son respectivamente

$$\begin{aligned} E(K_N) &= \frac{m}{N} \sum_{i=1}^{N} \left(\Phi^{-1} \left(\frac{i}{N+1} \right) \right)^2 \\ Var(K_N) &= \frac{mn}{N^2(N-1)} \left[N \sum_{i=1}^{N} \left(\Phi^{-1} \left(\frac{i}{N+1} \right) \right)^4 - \left(\sum_{i=1}^{N} \left(\Phi^{-1} \left(\frac{i}{N+1} \right) \right)^2 \right)^2 \right], \end{aligned} \tag{2.99}$$

y tablas con valores críticos de (2.98) para $N \leq 20$ se presentan en Klotz (1960).

Para finalizar la sección 2.2.6, en la Tabla 2.9 se presenta un resumen de los test no paramétricos utilizados en el problema general que compara dos muestras independientes.

Tabla 2.9: Resumen de los test no paramétricos utilizados en el problema general que compara dos muestras independientes.

Tipo de Problema		Nombre del Estadístico	Símbolo
Problema General de dos Muestras Independientes	Problema de Ubicación	Wald-Wolfowitz Runs Test	R
		Wilcoxon Rank-Sum Test (o Test Mann-Whitney)	W_N (U)
		Terry-Hoeffding Test	c_1
		Van der Waerden Test	X_N
	Problema de Escala	Mood Test	M_N
		Freund-Ansari-Bradley Test	A_N
		Siegel-Tukey Test	S_N
		Klotz Normal-Scores Test	K_N

2.2.11. Análisis de Múltiples Muestras Independientes

Hasta el momento se ha realizado un análisis que compara las observaciones provenientes de dos muestras independientes (problema general que compara dos muestras). Sin embargo, con mucha frecuencia es deseable comparar las observaciones de k muestras independientes de manera simultánea. Este criterio nos lleva a realizar la generalización de lo visto en las secciones anteriores mediante el estudio del denominado problema de k ($k > 2$) muestras múltiples.

Supongamos que tenemos un conjunto de k muestras independientes (la independencia se da dentro y entre muestras) de tamaños $n_1, n_2, \ldots, n_k$ que provienen de k poblaciones con cdf continuas $F_1(x), F_2(x), \ldots, F_k(x)$ respectivamente. El interés radica en demostrar si las k muestras provienen de la misma población, es decir se desea probar la hipótesis nula

$$H_0 : F_1(x) = F_2(x) = \ldots = F_k(x) \quad \forall x \tag{2.100}$$

versus alguna alternativa que indique que las poblaciones difieran en algún sentido, pudiendo ser a nivel de ubicación o escala.

De manera particular nos centraremos en el estudio de las diferencias dadas a nivel de ubicación. Así, el modelo de ubicación para el problema de k muestras múltiples nos dice que las cdf para las k poblaciones serán continuas y respectivamente iguales a $F_1(x - \theta_1), F_2(x - \theta_2), \ldots, F_k(x - \theta_k)$, en donde θ_i denotará algún parámetro de ubicación, que en el caso no paramétrico es la mediana (siempre existe).

Esto nos permite formular H_0 y su alternativa en función de los θ_i de la siguiente manera:

$$\begin{aligned} &H_0 : \theta_1 = \theta_2 = \ldots = \theta_k \\ &\qquad vs \\ &H_1 : \theta_i \neq \theta_j \quad \text{para al menos un } i \neq j. \end{aligned} \tag{2.101}$$

Un caso particular de prueba de hipótesis para (2.101) se dá al considerar que las k muestras son independientes y sustraídas de poblaciones normales con la misma varianza (la diferencia entre las poblaciones solo puede ocurrir con respecto a su ubicación). En este contexto, la utilización del análisis de varianza de un factor (ANOVA) mediante el Test F es adecuado para resolver (2.101) (Montgomery 2004). Adicionalmente es preciso mencionar que las técnicas no paramétricas, para resol-

ver el problema de k muestras independientes, no requieren supuestos más allá de que las poblaciones sean continuas.

A continuación se presentan los test no paramétricos que son la contraparte más habitual al análisis de varianza de un factor.

2.2.11.1. Test de la Mediana

En principio se da un breve resumen del Test de la Mediana para el problema general que compara dos muestras independientes. En lo posterior se dará su generalización. Para mayor detalle del Test de la Mediana en el caso de 2 muestras vea Pratt y Gibbons (1981), Miller (1981), Gibbons y Chakraborti (2011), Kraska-Miller (2014) y Siegel (1956).

Test de la Mediana para el Problema General de dos Muestras

Sean $X_1, X_2, \ldots, X_m$ e $Y_1, Y_2, \ldots, Y_n$ dos muestras aleatorias independientes dentro y entre si; y sea δ un número arbitrario pero fijo. La idea es comparar la proporción de observaciones de la muestra $X_1, X_2, \ldots, X_m$ y de la muestra $Y_1, Y_2, \ldots, Y_n$ que son estrictamente menores a δ en el arreglo agrupado-ordenado.

Si denotamos por

$$\begin{array}{ll} U & \text{al número de observaciones de la muestra } X \text{ menores a } \delta \\ V & \text{al número de observaciones de la muestra } Y \text{ menores a } \delta \end{array},$$

entonces

$$\begin{array}{rcl} U & \sim & Bi(m, p_X) \\ V & \sim & Bi(n, p_Y) \end{array} \quad \text{donde} \quad \begin{array}{rcl} p_X & = & P(X < \delta) \\ p_Y & = & P(Y < \delta) \end{array},$$

con momentos de primer y segundo orden respectivamente iguales a

$$\begin{array}{rcl} E(U) & = & mp_X \\ Var(U) & = & mp_X(1-p_X) \end{array} \quad ; \quad \begin{array}{rcl} E(V) & = & np_Y \\ Var(V) & = & np_Y(1-p_Y). \end{array}$$

La elección de δ afecta la sensibilidad relacionada con el criterio del test. Por ejemplo, un valor δ demasiado grande o demasiado pequeño provocará que los valores de U e V adquieran valores muy poco confiables, esto nos lleva a establecer un criterio adecuado para su determinación. Así, de los posibles valores que puede adquirir δ, detallados en Gibbons y Chakraborti (2011) y Pratt y Gibbons (1981), se llega a la conclusión que δ deber ser justo la mediana del arreglo agrupado. Es decir, δ será

la mediana común de las dos poblaciones, y su valor será la observación que tenga rango $N/2$ para cuando N sea par o cualquier número que se encuentre entre las observaciones con rango $N/2$ y $(N+1)/2$ para N impar.

Por otra parte, si p denota la probabilidad que cualquier observación en el arreglo agrupado sea menor a δ y $T = U + V$, entonces la distribución de probabilidad de T se define como

$$f_T(t) = \binom{m+n}{t} p^t (1-p)^{m+n-t} \text{ para } t = 0,1,2,\ldots N\,, \tag{2.102}$$

y bajo el supuesto de poblaciones idénticas tendremos que $p = p_X = p_Y$. Esto ocasiona que la distribución condicional de U dado $T = t$ sea

$$f_{U|T}(u|t) = \frac{\binom{m}{u}\binom{n}{t-u}}{\binom{m+n}{t}} \text{ donde } u = max(0,t-n),1,2,\ldots,min(m,t)\ , \tag{2.103}$$

que no es otra cosa que la distribución de probabilidad hipergeométrica.

Ahora bien, dado que U/m es un estimador insesgado de p_X, entonces u/m debería ser cercano a $t/(m+n)$ bajo la hipótesis nula (Gibbons y Chakraborti 2011). De esta manera, si se considera que las dos poblaciones solo difieren a nivel de ubicación, la elección razonable de $t/(m+n)$ es 0.5, con lo cual $t = N/2$ si N es par o $t = (N-1)/2$ si N es impar.

De todo lo anterior, el test basado en el número de observaciones de la muestra X estrictamente menores a la mediana del arreglo agrupado se denominada Test de la Mediana o Mood-Median Test. Por consiguiente, si U es mucho más grande que $m/2$, entonces la mayoría de las observaciones de la muestra X son mucho más pequeñas que la mayoría de las observaciones de la muestra Y ($P(X < \delta) > P(Y < \delta)$) o de manera equivalente $X <^{ST} Y$.

Las regiones de rechazo así como también la aproximación de la distribución de probabilidad nula de U la distribución normal se dan en Gibbons y Chakraborti (2011).

En lo que sigue se presenta la generalización del Test de la Mediana.

Extensión de Test de la Mediana

La extensión del Test de la Mediana para k muestras independientes nos permi-

tirá determinar si las k muestras independientes de tamaño $n_1, n_2, \ldots, n_k$ provienen o no de la misma población; y como ya se mencionó anteriormente para el caso de 2 muestras, la elección adecuada del valor δ será la mediana (mediana común para las k muestras) del arreglado agrupado de tamaño $N = \sum_{i=1}^{k} n_i$.
Adicionalmente, como cada observación en el arreglo agrupado tiene igual probabilidad de caer antes o después del valor δ; y si la hipótesis nula de poblaciones idénticas es verdadera, entonces la mitad de las observaciones en cada una de las k muestras deben ser menores a este valor Gibbons y Chakraborti (2011).

Luego, en base a la noción presentada en el párrafo anterior, si se define la variable aleatoria U_i como el número de observaciones de la muestra i que son menores a la mediana común δ, t como el número total de observaciones que son menores a δ o equivalentemente en símbolos

$$t = \sum_{i=1}^{k} u_i = \begin{cases} N/2 & \text{si } N \text{ es par} \\ (N-1)/2 & \text{si } N \text{ es impar} \end{cases} \tag{2.104}$$

y a θ como la probabilidad que una observación en el arreglo agrupado sea menor a δ (t/N), entonces la distribución de probabilidad de las variables U_i, dado el valor de t, será la distribución de probabilidad hipergeométrica multivariada o

$$f(u_1, u_2, \ldots, u_k | t) = \frac{\binom{n_1}{u_1}\binom{n_2}{u_2}\cdots\binom{n_k}{u_k}}{\binom{N}{t}}. \tag{2.105}$$

En el caso que algún(os) U_i sea(n) demasiado(s) diferente(s) de su valor esperado, mismo que es igual a $n_i\theta$, se deberá rechazar el supuesto de poblaciones idénticas.

Las regiones de rechazo para la generalización del Test de la Mediana se realizan en base al test de bondad de ajuste *Chi-Cuadrado* (Gibbons y Chakraborti 2011). La idea es convertir a los datos en una tabla de $2xk$, donde el 2 representa el número de filas (1 fila para las frecuencias que están debajo de δ y la otra las frecuencias que son más grandes que δ) y k representa a cada una de las muestras.

Si denotamos por (i, j) a cada una de las $2 * k$ categorías con $i = 1,2$; $j = 1,2,\ldots,k$ y si f_{ij} e e_{ij} denotan la frecuencia observada y esperada (bajo la hipótesis nula) para

la categoría (i, j) respectivamente, entonces

$$f_{1j} = u_i \qquad e_{1j} = n_i\theta = n_i\left(\frac{t}{N}\right) \tag{2.106}$$

$$f_{2j} = n_i - u_i \qquad e_{2j} = n_i(1-\theta) = n_i\left(\frac{N-t}{N}\right). \tag{2.107}$$

En virtud de esto y del criterio propuesto por Pearson citado en Gibbons y Chakraborti (2011) para la prueba de bondad de ajuste de la Chi-Cuadrado, el estadístico resultante será

$$\begin{aligned} Q &= \sum_{j=1}^{k}\sum_{i=1}^{2}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} = \sum_{j=1}^{k}\left[\frac{(f_{1j}-e_{1j})^2}{e_{1j}} + \frac{(f_{2j}-e_{2j})^2}{e_{2j}}\right] \\ &= \sum_{j=1}^{k}\frac{\left(u_i - n_i\left(\frac{t}{N}\right)\right)^2}{n_i\left(\frac{t}{N}\right)} + \sum_{j=1}^{k}\frac{\left(\cancel{n_i} - u_i - \cancel{n_i} + \left(\frac{n_i t}{N}\right)\right)^2}{n_i - \left(\frac{n_i t}{N}\right)} \\ &= N\sum_{j=1}^{k}\frac{\left(u_i - \frac{n_i t}{N}\right)^2}{n_i}\left[\frac{1}{t} + \frac{1}{N-t}\right] = \frac{N}{t(N-t)}\sum_{j=1}^{k}\frac{(u_i - n_i t/N)^2}{n_i}\ , \end{aligned} \tag{2.108}$$

mismo que bajo H_0 sigue aproximadamente la distribución Chi-Cuadrada con $(K-1)(2-1) = (K-1)$ grados de libertad (Gibbons y Chakraborti 2011), (Siegel 1956). No obstante, si al valor de Q le multiplicamos por $(N-1)/N$ la aproximación mejora considerablemente, ocasionando que la región

$$Q \in R \ \ para \ \ \frac{N-1}{N} \geq \chi^2_{(k-1),\alpha} \tag{2.109}$$

rechace la hipótesis nula.

En lo referente al problema de ties, un enfoque es considerar a estas observaciones como menores a δ.

2.2.11.2. Test Anova Unidireccional de Kruskal-Wallis y Comparaciones Múltiples

Como vimos, la extensión del Test de la Mediana se basaba en la comparación relativa de cada una de las observaciones del arreglo agrupado respecto al valor de la mediana común. Ahora el enfoque es ir más allá al tratar de utilizar toda la información disponible de cada observación en el arreglo. Esto se logra mediante la

comparación en términos de los rangos de las observaciones (Gibbons y Chakraborti 2011).

El análisis de varianza unidireccional de Kruskal-Wallis por rangos, bajo el supuesto que los datos sean medidos al menos en escala ordinal y que las muestras sean extraídas de poblaciones con cdf continuas, es una prueba general que compara si k muestras independientes ($k > 2$) provienen de una población común.

La gran ventaja que nos brinda este análisis de varianza se refleja en el hecho que no es necesario conocer a priori la distribución de probabilidad de los datos; y si los datos verifican cada uno de los supuesto del modelo general con errores normalmente distribuidos, el Test de Kruskal-Wallis será tan efectivo como la prueba paramétrica Anova de un factor (Kraska-Miller 2014). Es por esta razón que al Test Kruskal-Wallis, que básicamente es la extensión del Wilcoxon Rank-Sum Test o también llamado Test Mann-Whitney se lo conoce como la contraparte no paramétrica a la prueba ANOVA.

Sin embargo, la prueba de Kruskal-Wallis solo nos brinda la información de si al menos una muestra difiere significativamente de las otras, pero no identifica en donde ocurren esas diferencias ni cuántas diferencias se producen (Corder y Foreman 2009). Así, para identificar las diferencias particulares entre pares de muestras se puede usar contrastes de muestras o pruebas post hoc, siendo la prueba de U de Mann-Whitney el método más útil para realizar estos contrastes entre pares de muestras específicos.

El procedimiento del Test de Kruskal-Wallis es la siguiente:

Bajo la hipótesis nula de poblaciones idénticas, para k muestras independientes de tamaños $n_1, n_2, \ldots, n_k$ se hace un seguimiento del rango que ha sido asignado a cada observación dentro arreglo agrupado de tamaño $\sum_{i=1}^{k} n_i$, así como la identidad de pertenencia de la observación a su respectiva muestra. En este contexto Gibbons y Chakraborti (2011), dicen que si los rangos adyacentes están bien distribuidos entre las k muestras, lo que sería cierto para una muestra aleatoria de una sola población; la suma total de los rangos, que no es otra cosa que la suma de los N primero enteros positivos se dividiría proporcionalmente de acuerdo al tamaño de la i-ésima muestra. Es decir, debido a que el rango esperado para cualquier observación es $\frac{N+1}{2}$ (véase 2.13, pág. 20), el valor esperado de la suma de los rangos asignados a

cada observación para la i-ésima muestra de tamaño n_i será

$$E(R_i) = n_i\left(\frac{N+1}{2}\right) \tag{2.110}$$

y su varianza será

$$Var(R_i) = \frac{n_i(N+1)(N-n_i)}{12}\ , \tag{2.111}$$

donde R_i denota la suma de los rangos de los elementos de la i-ésima muestra para $i = 1,2,\ldots,k$ (Kvam y Vidakovic 2007), (Gibbons y Chakraborti 2011).

Un estadístico razonable surge al considerar la suma de las variaciones de la suma de los rangos R_i con respecto a su valor esperado; y como ya se mencionó los inconvenientes que presenta esta idea en (2.88), lo indicado será considerar la suma del cuadro de estas desviaciones, es decir

$$S = \sum_{i=1}^{k}\left(R_i - n_i\left(\frac{N+1}{2}\right)\right)^2. \tag{2.112}$$

Valores muy grandes de S provocarán el rechazo de la hipótesis nula de poblaciones idénticas (Kvam y Vidakovic 2007).

Para determinar la distribución de probabilidad nula de S se deben clasificar a las observaciones en una tabla con k columnas (tabla de clasificación), en donde cada entrada de la i-ésima columna corresponde al rango asignado a cada una de las n_i observaciones de la muestra i, de tal forma que R_i será la suma de los valores presentes en la i-ésima columna. A pesar que Rijkoort (1952) citado en Gibbons y Chakraborti (2011) presenta tablas de la distribución de S para $k = 3,4$ y 5 con tamaños de muestras iguales y pequeños, la complejidad de esta clasificación hace que los cálculo sean tediosos incluso para muestras de tamaños pequeños (Kvam y Vidakovic 2007), en consecuencia un enfoque alternativo es necesario.

De lo anterior, y en vista que la aproximación normal estándar de los R_i implica que

$$\sum_{i=1}^{k}\left(\frac{(R_i - E(R_i))^2}{Var(R_i)}\right) \sim \chi^2_{k-1}\ , \tag{2.113}$$

Kruskal y Wallis (1952) y Kruskal (1952) proponen el estadístico

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{1}{n_i}\left(R_i - n_i\left(\frac{N+1}{2}\right)\right)^2\ , \tag{2.114}$$

que no es otro cosa que la suma ponderada de los cuadrados de las desviaciones con los recíprocos de los tamaños de las muestras respectivas utilizadas como pesos (Gibbons y Chakraborti 2011).
Valores muy grandes de H dan como resultado que la hipótesis nula en (2.101) sea rechazada.

Es claro ver que existe cierta similitud entre el Test S y el Test H. Esto lo ratifican Gibbons y Chakraborti (2011) al decir que los mismos son equivalente para tamaños de muestras iguales. En este sentido, como la determinación de la distribución de probabilidad nula de S resulta complicada de obtener, debido a la complejidad inherente en la clasificación de las observaciones, entonces la distribución de probabilidad de H tendrá el mismo rumbo.

A pesar de que existen tablas de probabilidad exacta de H, por ejemplo, la tabla K en Gibbons y Chakraborti (2011) para k=3 y $n_i \leq 3$ o para $k = 4$, $k = 5$ y $n_i \leq 4$ e $n_i \leq 3$ respectivamente presentadas en Iman et al. (1975) citado en libro anterior, es necesario realizar una aproximación razonable de la distribución H. Por lo tanto, bajo la hipótesis nula de poblaciones idénticas, cada uno de los rangos presentes en la i-ésima columna correspondiente a la tabla de clasificación, pueden ser considerados como v.a. del conjunto $1,2,\ldots,N$, donde $N = \sum_{i=1}^{k} n_i$. En otras palabras, cada i-ésima columna puede ser considera como una muestra aleatoria de tamaño n_i extraía sin reemplazo de la población de los primeros N enteros positivos, cuya media y varianza son respectivamente

$$\mu = \frac{N+1}{2} \qquad \text{y} \qquad \sigma^2 = \frac{N^2-1}{12}. \tag{2.115}$$

Por otro lado, como la esperanza y varianza del promedio de la suma de los rangos de la i-ésima columna son respectivamente

$$E(\overline{R_i}) = \frac{N+1}{2} \qquad \text{y} \qquad Var(\overline{R_i}) = \frac{1}{12n_i}[(N+1)(N-n_i]\,, \tag{2.116}$$

Demostración:
Con la ayuda de (2.110) tenemos que

$$E(\overline{R_i}) = E\left(\frac{R_i}{n_i}\right) = \frac{1}{n_i}E(R_i) = \frac{1}{\cancel{n_i}}\left[\cancel{n_i}\left(\frac{N+1}{2}\right)\right] = \mu$$

y de (2.111)

$$Var(\overline{R_i}) = \frac{1}{n_i^2} Var(R_i) = \frac{1}{\cancel{n_i^2}} \left[\frac{\cancel{n_i}(N+1)(N-n_i)}{12} \right] = \frac{1}{12n_i}[(N+1)(N-n_i)]$$

■

la estandarización $Z_i = \dfrac{\overline{R_i} - E(\overline{R_i})}{\sqrt{Var(\overline{R_i})}}$ se aproximará a la distribución normal estándar. Sin embargo, como los Z_i no son independientes debido a que

$$\begin{aligned} Cov(Z_i, Z_j) &= Cov\left(\frac{\overline{R_i} - E(\overline{R_i})}{\sqrt{Var(\overline{R_i})}}, \frac{\overline{R_j} - E(\overline{R_j})}{\sqrt{Var(\overline{R_j})}} \right) = \frac{1}{\sqrt{Var(\overline{R_i})}} \frac{1}{\sqrt{Var(\overline{R_j})}} Cov(\overline{R_i}, \overline{R_j}) \\ &= \frac{1}{\sqrt{Var(\overline{R_i})}} \frac{1}{\sqrt{Var(\overline{R_j})}} \left[-\frac{N+1}{12} \right] \neq 0\,, \end{aligned}$$

Kruskal (1952) mediante la aplicación del teorema de Wald-Wolfowitz para el caso especial que $L_N^{(i)} = \overline{R_i}$ y en base a (2.116) dice que

$$\begin{aligned} \sum_{i=1}^{k} \frac{N-n_i}{N} Z_i^2 &= \sum_{i=1}^{k} \left(\frac{\cancel{N-n_i}}{N} \right) \left(\frac{12n_i[\overline{R_i} - (N+1)/2]^2}{(N+1)\cancel{(N-n_i)}} \right) \\ &= \sum_{i=1}^{k} \frac{12n_i[\overline{R_i} - (N+1)/2]^2}{N(N+1)} \\ &= \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{1}{n_i} \left(R_i - n_i \left(\frac{N+1}{2} \right) \right)^2 \\ &= H \end{aligned} \tag{2.117}$$

es asintóticamente la distribución Chi-Cuadrado con $k-1$ grados de libertad (Gibbons y Chakraborti 2011), (Kruskal 1952). En consecuencia, la región de rechazo a un nivel de significancia α será para $H \geq \chi^2_{\alpha,K-1}$.

Problema con observaciones ties

Dada la suposición que las poblaciones son continuas, los casos ties son teóricamente imposibles de suceder. No obstante, el verdadero problema radica cuando los ties ocurren entre las muestras. Esta situación conlleva a realizar la corrección de la varianza mediante la llamada corrección para ties, que Gibbons y Chakraborti (2011), Kruskal y Wallis (1952), Siegel (1956), Corder y Foreman (2009) mencionan

que simplemente se debe dividir H para el factor de corrección

$$1 - \frac{\sum t(t^2-1)}{N(N^2-1)} \ .$$

Dicho de otra manera,

$$\frac{H}{1 - \frac{\sum t(t^2-1)}{N(N^2-1)}} \ , \tag{2.118}$$

donde t es el número de observaciones ties para un rango arbitrario y el sumatorio se extiende sobre todo el conjunto de rangos con valores ties.

Finalmente, la demostración de la consistencia del Test H está más allá del alcance del presente trabajo de titulación, pero si el lector desea analizarla con detalle puede revisar en Kruskal (1952).

Comparaciones Multiples

El estadístico de Kruskal-Wallis nos proporciona información de si hay o no diferencias entre las k poblaciones, pero no nos especifica explícitamente que poblaciones difieren entre sí. Para detectar las diferencias entre todo par de poblaciones vamos a realizar comparaciones dos a dos y decidir si las mismas provienen o no de poblaciones idénticas.

Si la hipótesis nula se rechaza, el proceso de comparaciones múltiples compara el valor de

$$Z_{ij} = \frac{|\overline{R_i} - \overline{R_j}|}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \qquad \text{con } \ 1 \leq i < j \leq k \tag{2.119}$$

con el cuantil superior $[\alpha/(k(k-1))]$ de la distribución normal estándar; y si el valor $Z_{ij} > Z_{\alpha/k(k-1)}$ los grupos muestrales i, j serán significativamente diferentes.

El valor de α, llamado tasa de error experimental o nivel de significancia global típicamente se escoge que sea al menos 0.20, debido a que se hace una gran cantidad de declaraciones. En consecuencia $1 - \alpha$ será la probabilidad que todas las declaraciones sean correctas (Gibbons y Chakraborti 2011).

Este proceso de comparaciones múltiples se debe a Dunn (1964).

2.3. Eficiencia Relativa Asintótica

Algo que comúnmente llama la atención del investigador es ver si dados dos test estadísticos que verifican que una declaración en cuestión es cierta o no, saber cuál es el más eficiente. En otras palabras, dados dos test estadísticos definidos para una misma hipótesis nula simple, la misma hipótesis alternativa, el mismo tipo de región de rechazo y el mismo nivel de significación; ¿cómo se puede comparar estos test para saber cual es el más eficiente? La respuesta a esta interrogante la da el concepto de eficiencia de Pitman descrito en Gibbons y Chakraborti (2011). Básicamente el enfoque es comparar dos pruebas y hacer que todos sus factores sean equivalentes, excepto su tamaño de muestra.

Dentro de lo paramétrico y dadas las consideraciones expuestas con anteriores, si tenemos un test A y un test B, donde el test B tiene su función potencia superior a la función potencia del test A, entonces se escoge el test B. Sin embargo, dado que en lo no paramétrico los cálculos exactos de la potencia son tediosos, junto con lo mencionado por Belmonte (2001) respecto a que la estructura de la distribución no es lo suficientemente rígida para obtener una teoría análoga a la que se usa en la teoría paramétrica, en cuanto a generar un test uniformemente más poderoso dentro del lema de Neyman-Pearson; se tendrá que una prueba no paramétrica rara vez es uniformemente eficiente comparada con otra.

Por esta razón, la solución al problema presentado en el párrafo anterior es definir un tipo de potencia de eficiencia límite. Esta definición se presenta a continuación.

Definición 2.3.1 (Eficiencia Relativa Asintótica (ARE)). *Sean A y B dos test consistentes para una hipótesis nula H_0, una hipótesis alternativa H_1 y una nivel de significancia α. La Eficiencia Relativa Asintótica (ARE por sus siglas en inglés) del test A en relación al test B es el valor límite de la relación n_b/n_a, donde n_a es el número de observaciones requeridas por el test A para que la potencia del test A sea igual a la potencia del test B basado en n_b observaciones, mientras simultáneamente $n_b \to +\infty$ y $H_1 \to H_0$.*

Teniendo en cuenta que la consistencia de los test hace que su potencia tienda a 1 para tamaños de muestras crecientes, se debe permitir que H_1 se aproxime a H_0. Esto con el fin de que la potencia de cada test se encuentre en el intervalo abierto $(\alpha, 1)$ para tamaños de muestra finitos y la relación límite generalmente sea un número distinto de 1 (Gibbons y Chakraborti 2011).

Más detalle de la teoría de la ARE se exhiben en Gibbons y Chakraborti (2011) y Mayorga (2004)

En lo que sigue se va a presentar los resultados de la ARE proporcionados al comparar ciertos test no paramétricos con los test paramétricos tradiciones, el t Test y ANOVA Test.

Pero antes de mostrar estos resultados es necesario detallar que, dados dos test estadísticos T_n y T_n^* para datos que consisten de n observaciones, se denotará la Eficiencia Relativa Asintótica de T_n en relación a T_n^* por $ARE(T_n, T_n^*)$.

La tabla siguiente muestra la ARE del Sign Test y del Wilcoxon Signed Rank Test con relación a t Test calculadas para cuatro distribuciones de probabilidad. Adicionalmente se exhibe la ARE del Sign Test en relación al Wilcoxon Signed Rank Test.

Tabla 2.10: Valores de la $ARE(K_n, T_n^*)$, $ARE(T^+, T_n^*)$ y $ARE(K_n, T_n^+)$

Distribución	$ARE(K_n, T_n^*)$	$ARE(T_n^+, T_n^*)$	$ARE(K_n, T_n^+)$
Uniforme	$1/3 \approx 0.33$	1	$1/3 \approx 0.33$
Normal	$2/3 \approx 0.67$	$3/\pi \approx 0.9549$	$2/\pi \approx 0.6366$
Logística	$3/4 = 0.75$	$\pi^2/9 \approx 1.0966$	$\pi^2/12 \approx 0.8225$
Doble Exponencial	$4/3 \approx 1.33$	$3/2 = 1.5$	2

T_n^* es el t Test
K_n es el Sign Test
T_n^+ es de Wilcoxon Signed Rank Test

Bajo el supuesto de distribuciones normales, la Tabla 2.10 nos dice que la ARE del Sign Test en relación al t Test es aproximadamente 0.67 y la del Wilcoxon Signed Rank Test en relación al t Test aproximadamente del 0.95. Estos valores de la ARE son menores que 1, lo cual es razonable, debido a que el t Test es la mejor prueba para la distribución normal.

Por un lado, si se observa la columna tres de la tabla anterior, claramente el Test de Signo de Rango de Wilcoxon es un fuerte competidor a la prueba t para muestras de gran tamaño. En cambio, la columna dos expresa que la prueba de Signo es mucho menos eficiente que la prueba t para distribuciones de cola ligeras, pero tiene un rendimiento superior para distribuciones de cola pesada, como por ejemplo en la doble exponencial.
De manera general, Gibbons y Chakraborti (2011) expresan que: "la ARE de la prueba de Rango con Signo de Wilcoxon siempre es al menos 0.864 para cualquier distribución simétrica continua, mientras que el límite inferior correspondiente para la prueba de Signo es solo 1/3"(p.492).

Por otro lado, en la última columna de la Tabla 2.10 se compara la eficiencia del

Sign Test en relacion al Wilcoxon Signed Rank Test. Dando como resultado que el Test de Signo es más eficiente solo para la distribución doble exponencial.

En resumen, la prueba de Signo de Rango de Wilcoxon es una alternativa muy viable a la popular prueba t de Student, debido a que se acerca mucho a su rendimiento en términos de la ARE, no obstante la prueba de Signo, quizás es una opción popular debido a su facilidad de uso más que a su rendimiento.

Ahora bien, al considerar la ARE para ciertos test que son una alternativa adecuada para el problema general que compara dos o k muestras, ya sea a nivel de ubicación o escala, los resultados son los siguientes:

Gibbons y Chakraborti (2011) demuestran que la ARE del Mann-Whitney Test (o equivalentemente de Wilcoxon Rank Sum Test) en relación al t Test viene dada por la ARE del Wilcoxon Signed Rank Test en relación al t Test. Adicionalmente, el valor de la ARE del Test de la Mediana relativa a la prueba de Mann Whitney es el mismo que la ARE de la prueba de Signo relativa a la prueba de Rango con Signo de Wilcoxon. Por consiguiente, tanto los valores de eficiencia presentados en la Tabla 2.10, así como la descripción realizada de la misma para estos casos; son los mismos, pero con la diferencia que para el caso del Test U no es necesario suponer que la distribución de probabilidad sea simétrica.

En lo referente al Terry–Hoeffding (Score Normal) Test, la literatura dice que bajo el cumplimiento de los supuestos que rigen en la formulación del t Test, la prueba de Terry es asintóticamente óptima en relación al t Test para distribuciones normales, es decir su ARE será de 1.

De manera general, la ARE del Test Score Normal en relación al t Test para otras familias de distribuciones continuas será mayor a 1 (Lehmann 2008).

De los dos párrafos anteriores se aprecia que tanto el Test U (o su equivalente de Wilcoxon Rank Sum Test) como el Terry–Hoeffding Test son competidores fuertes a la prueba t. Sin embargo, al comparar estos dos test se evidencia que el Test de Mann-Whitney es más eficiente frente a distribuciones de cola larga, y el c_1 Test es más eficiente para distribuciones de cola corta (Lehmann 2008).

Esto se puede evidenciar en la Tabla 2.11.

Tabla 2.11: Valores de la $ARE(U,c_1)$ para cinco distribuciones de probabilidad

	Uniforme	Normal	Logística	Doble Exponencial	Cauchy
$ARE(U,c_1)$	0	0.955	1.05	1.18	1.41

U es el Test Mann-Whitney
c_1 es el Terry–Hoeffding Test

Si se considera el problema general que compara dos muestras a nivel de escala, por un lado, la ARE del Mood Test en relación al Test $F_0 = S_1^2/S_2^2$ (el mejor test para probar la igualdad varianzas de dos poblaciones normales que difieren solo en la varianza Gibbons y Chakraborti (2011), Montgomery (2004)), donde S_1 y S_2 son las desviaciones estándar para la muestra 1 y 2 respectivamente; será $15/(2\pi^2) \approx 0.76$. Por otro lado, la ARE del Freund-Ansari-Bradley Test en relación al Test F_0 es igual a $6/\pi^2 \approx 0.6079$, igual a 0.60 en relación a la distribución uniforme continua e igual a 0.94 en relación a la doble exponencial. En última instancia, la ARE del Freund-Ansari-Bradley Test en relación al Mood Test es igual a 0.8.

Para el caso de múltiples muestras independientes ($k \geq 3$), la ARE de la extensión del Test de la Mediana en relación al Test ANOVA es de $2/\pi \approx 0.637$ y de $2/3$ en relación al Test de Kruskal-Wallis, ambos casos para la distribución normal. Finalmente, la ARE del Test de Kruskal Wallis en relación a cualquier distribución continua es de al menos el 0.864 y del 0.955 para la distribución normal (Gibbons y Chakraborti 2011). Este hecho nos indica que el Test de Kruskal- Wallis es la contraparte más eficiente al Test ANOVA.

2.4. Análisis de Componentes Principales (ACP)

La herramienta utilizada en la construcción de las observaciones objetas de inferencia se basará en las transformaciones proporcionadas por el procedimiento del Análisis de Componentes Principales (ACP).

El objetivo del ACP, dadas $X_1, X_2, \ldots, X_p$ variables (variables originales) es hallar p nuevas variables (ejes o componentes) $Y_1, Y_2, \ldots, Y_p$, cada una de las cuales es una combinación lineal de las variables X_j con $j = 1, 2, \ldots, p$; de tal manera que r ($r < p$) de ellas recojan la mayor parte de información posible concerniente a las variables originales [16]. Los nuevos ejes representarán las direcciones con variabilidad máxima y proporcionan una descripción más simple de la estructura de covarianza (Johnson y Wichern 2007). Adicionalmente, si consideremos que $\mathbf{X}$ es la matriz de datos que contiene las mediciones de los n individuos en p variables iniciales cuan-

titativas, entonces las componentes principales dependerán solamente de la matriz de covarianzas $\boldsymbol{\Sigma}$ (o matriz de correlaciones $\boldsymbol{\rho}$) de la matriz $\mathbb{X}$.

Algebraicamente, las componentes principales son combinaciones lineales particulares de las p variables originales. Geométricamente, esas combinaciones lineales representan un nuevo sistema de coordenadas obtenidas por rotación del sistema original con $X_1, X_2, \ldots, X_p$ como ejes de coordenadas.

La utilidad del ACP señalado por Peña (2002), por un lado permite representar óptimamente en un espacio de dimensión pequeña, observaciones de un espacio general p-dimensional. Es decir, sirve para identificar posibles variables "latentes" o no observadas, que están generando la variabilidad de los datos. Por otro lado, permite transformar las variables originales, en general correlacionadas, en nuevas variables no correlacionadas, facilitando la interpretación de los datos.

En lo que sigue se exhibe el procedimiento del ACP:

Si el vector aleatorio $\boldsymbol{\mathcal{X}'} = [X_1, X_2, \ldots, X_p]$ tiene matriz de covarianzas $\boldsymbol{\Sigma}$ con autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, entonces para las combinaciones lineales

$$\begin{aligned} Y_1 &= \boldsymbol{a_1'\mathcal{X}} = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\ Y_2 &= \boldsymbol{a_2'\mathcal{X}} = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\ &\vdots \qquad\qquad\qquad \vdots \\ Y_p &= \boldsymbol{a_p'\mathcal{X}} = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p \end{aligned} \tag{2.120}$$

tendremos que

$$Var(Y_j) = \boldsymbol{a_j'\Sigma a_j} \quad j = 1,2,\ldots,p \tag{2.121}$$

$$Cov(Y_j, Y_k) = \boldsymbol{a_j'\Sigma a_k} \quad j \neq k \text{ con valores en } 1,2,\ldots,p \quad . \tag{2.122}$$

En consecuencia, las componentes principales serán las combinaciones lineales no correlacionadas $Y_1, Y_2, \ldots, Y_p$ con las varianzas más grandes posibles. Así, si queremos que Y_1 recoja la mayor cantidad de información posible de los individuos, se debe exigir que las proyecciones de los individuos sobre este eje, en promedio sean lo más grandes posibles; junto con la restricción que a_1 en (2.120) sea el vector unitario. De manera análoga para el resto de componentes.

[16] Para mayor detalle vea Johnson y Wichern (2007), Peña (2002).

En virtud de lo expresado anteriormente, al definir como

$$\begin{aligned}
\text{Primera componente principal} = \text{combinación lineal } \boldsymbol{a_1'\mathcal{X}} \text{ que} \\
&\text{maximiza} \quad Var(\boldsymbol{a_1'\mathcal{X}}) \\
&\text{sujeta a:} \quad \boldsymbol{a_1'a_1} = 1 \ , \\
\text{Segunda componente principal} = \text{combinación lineal } \boldsymbol{a_2'\mathcal{X}} \text{ que} \\
&\text{maximiza} \quad Var(\boldsymbol{a_2'\mathcal{X}}) \\
&\text{sujeta a:} \quad \boldsymbol{a_2'a_2} = 1 \\
&\qquad\qquad Cov(\boldsymbol{a_1'\mathcal{X}}, \boldsymbol{a_2'\mathcal{X}}) = 0 \ , \\
\vdots \qquad\qquad &\qquad\qquad \vdots \\
\text{j-ésima componente principal} = \text{combinación lineal } \boldsymbol{a_j'\mathcal{X}} \text{ que} \\
&\text{maximiza} \quad Var(\boldsymbol{a_j'\mathcal{X}}) \\
&\text{sujeta a:} \quad \boldsymbol{a_j'a_j} = 1 \\
&\qquad\qquad Cov(\boldsymbol{a_j'\mathcal{X}}, \boldsymbol{a_k'\mathcal{X}}) = 0 \ \text{ para } k < j \ ,
\end{aligned}$$

el Teorema 2.4.1 nos da componentes principales no correlacionadas que tienen varianza igual a los valores propios de $\boldsymbol{\Sigma}$.

Teorema 2.4.1. *Si* $\boldsymbol{\Sigma}$ *es la matriz de covarianzas asociada a* $\boldsymbol{\mathcal{X}'}$*, que tiene los pares valor propio-vector propio* $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$ *para* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$*, entonces la j-ésima componente principal está dada por*

$$Y_j = \boldsymbol{e_j'\mathcal{X}} = e_{j1}X_1 + e_{j2}X_2 + \ldots + e_{jp}X_p \quad con \quad j = 1, 2, \ldots, p \tag{2.123}$$

y además

$$\begin{aligned}
Var(Y_j) &= \boldsymbol{e_j'\Sigma e_j} = \lambda_j \\
Cov(Y_j, Y_k) &= \boldsymbol{e_j'\Sigma e_k} \ \ para \ \ j \neq k
\end{aligned} \tag{2.124}$$

La demostración de este teorema es fácil de realizar y se encuentra detallada en Johnson y Wichern (2007).

Un resultado importante que se da del mismo teorema muestra que si $\sigma_1, \sigma_2, \ldots, \sigma_p$ son las varianzas de $X_1, X_2, \ldots, X_p$ respectivamente, entonces

$$\sigma_1 + \sigma_2 + \ldots + \sigma_p = \sum_{j=1}^{p} Var(X_j) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{j=1}^{p} Var(Y_j) \tag{2.125}$$

y la proporción de la varianza total debida a (explicada por) el j-ésimo componente principal será

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \text{ para } j = 1,2,\ldots,p \quad . \tag{2.126}$$

Un enfoque alternativo, mismo que es utilizado en el presente estudio, es obtener componentes principales para variables estandarizadas. La estandarización lleva todas las escalas de medida a una escala común de media 0 y varianza 1, para eliminar el problema de medición y variabilidad diferente presente en las variables originales. De esta manera, si μ_j, σ_j son la media y la varianza de la variable X_j respectivamente para $j = 1,2,\ldots,p$, entonces para las variables estandarizadas

$$\begin{aligned} Z_1 &= \frac{X_1 - \mu_1}{\sqrt{\sigma_1}} \\ Z_2 &= \frac{X_2 - \mu_2}{\sqrt{\sigma_2}} \\ \vdots \quad &\quad \vdots \\ Z_p &= \frac{X_p - \mu_p}{\sqrt{\sigma_p}} \end{aligned} \tag{2.127}$$

o en su forma matricial $\boldsymbol{\mathcal{Z}} = [Z_1, Z_2, \ldots, Z_p] = (\mathbf{V}^{1/2})^{-1}(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu})$, donde $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_p]$ y $\mathbf{V}$ es una matriz diagonal que tiene las desviaciones estándar de $X_1, X_2, \ldots, X_p$; se tiene que $E(\boldsymbol{\mathcal{Z}}) = 0$ y $Cov(\boldsymbol{\mathcal{Z}}) = (\mathbf{V}^{1/2})^{-1}\boldsymbol{\Sigma}(\mathbf{V}^{1/2})^{-1} = \boldsymbol{\rho}$.

En base a esto, los componentes principales de $\boldsymbol{\mathcal{Z}}$ se pueden obtener a partir de los vectores propios de la matriz de correlación $\boldsymbol{\rho}$ de $\boldsymbol{\mathcal{X}'}$ y todos los resultados dados por el Teorema 2.4.1 son igualmente aplicables para este tipo de variables.

Por consiguiente, la j-ésima componente principal será

$$Y_j = \boldsymbol{e_j'}\boldsymbol{\mathcal{Z}} = \boldsymbol{e_j'}(\mathbf{V}^{1/2})^{-1}(\boldsymbol{\mathcal{X}} - \boldsymbol{\mu}) \text{ para } j = 1,2,\ldots,p \ ,$$

$\sum_{j=1}^{p} Var(Y_j) = \sum_{j=1}^{p} Var(Z_j) = p$ y la proporción de varianza total debida al j-ésimo componente será $\frac{\lambda_j}{p}$.

Finalmente, la proyección del individuo i sobre el eje Y_j será la coordenada del individuo sobre el eje, esto es $Y_{ij} = \mathbf{X_{(i)}}e_j$ para $i = 1,2\ldots,n$. De manera general, la matriz de proyección de los individuos en los nuevos ejes será $\mathbf{Y} = \mathbf{XE}$, donde $\mathbf{E}$ es la matriz de vectores propios de $\boldsymbol{\rho}$. Mas propiedades del procedimiento del Análi-

sis de Componentes Principales se encuentran en Peña (2002) o Johnson y Wichern (2007).

Sin embargo, antes de aplicar el ACP se debe verificar si la correlación entre las variables analizadas es lo suficientemente grande como para justificar la factorización de la matriz de coeficientes de correlación. En otras palabras, se debe comprobar si la matriz de coeficientes de correlación no es significativamente distinta de la matriz identidad (las intercorrelaciones entre las variables son ceros). Esto se logra mediante la aplicación del Test de esfericidad de Bartlett, el cual consiste de la estimación *Chi-Cuadrado* a partir de una transformación del determinante de la matriz de correlaciones. En este sentido, si las variables no están intercorrelacionadas, entonces el test debe presentar un p-valor mayor a 0.05.

No obstante, algunos autores advierten que el Test de Bartlett tiende a ser estadísticamente significativo cuando el tamaño muestral crece, razón por la cual se lo deberá de utilizar únicamente cuando la razón n/k es menor que 5. Pero como nuestro estudio presenta un valor de n es grande, se debe calcular el Índice Kaiser-Meyer-Olkin (KMO) [17], mismo que sirve para comparar las magnitudes de los coeficientes de correlación general o simple con respecto a las magnitudes de los coeficientes de correlación parcial (Suárez 2007). Si la suma de los coeficientes de correlación parcial elevados al cuadrado entre todos los pares de variables es bajo en comparación con la suma de los coeficientes de correlación al cuadrado, entonces el índice estará próximo a uno y esto se considerará positivo e indicará que se puede continuar con el ACP. Pero si se obtienen valores bajos, entonces se indica que las correlaciones entre pares de variables no pueden ser explicadas por las otras variables y, por lo tanto, no es factible llevar a cabo el ACP. Los valores del índice KMO entre 0.5 y 1 indican que es apropiado aplicar el ACP a la matriz de datos bajo estudio.

Un enfoque alternativo toma en cuenta el determinante de la matriz de correlaciones. Si el valor del determinante es cercano a cero [18] existe evidencia que las variables tienen intercorrelaciones muy altas, lo que a su vez implica que sea factible continuar con el Análisis de Componentes Principales.

En conclusión, se debe de utilizar al menos un criterio de los antes mencionados para determinar si es adecuado el uso del ACP.

[17] Para mayor detalle vea Kaiser (1974).

[18] El determinante no debe ser igual a cero, pues en este caso los datos no serían válidos Suárez (2007).

2.5. Construcción de Indicadores Compuestos

Como se mencionó anteriormente, el ACP (más específicamente el análisis factorial) agrupa las variables individuales que son colineales para determinar un(os) nuevo(s) factor(es) (indicadores compuestos) que capture(n) tanto como sea posible la información común de las variables individuales. En este sentido, este(os) factor(es) ya no depende(n) de la dimensionalidad del conjunto de datos, sino más bien se basan en las dimensiones "estadísticas" de los datos (Nardo y col. 2008). No obstante, si la correlación es débil, entonces es poco probable que las variables compartan factores comunes.
Esta es una de las razones por las cuales es necesario la utilización de otro tipo de Indicadores Compuestos.

Organización para la Cooperación Económica y el Desarrollo (OECD por sus siglas en ingles)

La OECD es un foro único donde los gobiernos de 30 países (por ejemplo Estados Unidos, Suecia, España, Reino Unido, etc.) trabajan juntos para abordar los desafíos económicos, sociales y ambientales de la globalización. La idea de la organización es proporcionar un entorno en donde se pueda comparar experiencias de políticas, buscar respuestas a problemas comunes, identificar buenas prácticas y trabajar en conjunto para coordinar políticas nacionales e internacionales.

Dentro de las publicaciones de la OECD, que presentan los resultados de la recopilación estadísticas e investigaciones sobre los desafíos mencionados anteriormente; conjuntamente con la Unidad de Econometría y Estadística Aplicada del Centro de Investigación Conjunta (JRC) de la Comisión Europea en Ispra (Italia), se encuentra el manual **Handbook on Constructing Composite Indicators**. Este manual, que compara y clasifica el desempeño de los países en áreas tales como competitividad industrial, desarrollo sostenible, globalización y innovación; proporciona una guía en la construcción y el uso de indicadores compuestos, orientado a dar una mejor compresión acerca de la complejidad y las técnicas que se utilizan actualmente en su construcción; para así mejorar la calidad de los productos que se obtienen como resultado de su aplicación.

En términos generales, un **indicador (simple)** es una medida cuantitativa o cualitativa derivada de una serie de hechos observados que pueden revelar posiciones relativas (por ejemplo, de un país) en un área determinada (Nardo y col. 2008). A

modo de ejemplo, dentro del análisis de política se utilizan para la identificación de tendencias, dar un llamado de atención sobre temas particulares, establecimiento de prioridades de la política y la evaluación del rendimiento (desempeño) de los países. En cambio, la idea detrás de un **indicador compuesto**, que se forma cuando los indicadores individuales se compilan en un solo índice sobre la base de un modelo subyacente, es considerar conceptos multidimensionales que no pueden ser capturados por los indicadores simples. Algunos de estos conceptos, por ejemplo se relacionan con la competitividad, industrialización, sostenibilidad, integración en el mercado, sociedad basada en el conocimiento, etc.

2.5.1. Indicadores Compuestos (CI)

Los indicadores compuestos (que comparan el rendimiento de los individuos) son cada vez más reconocidos como una herramienta útil en el análisis de políticas y la comunicación pública, debido a que parece más fácil para el público en general interpretar este tipo de indicadores, que identificar tendencias comunes a través de muchos indicadores por separado (Rogge 2008), y también porque han demostrado ser útiles en la evaluación comparativa del desempeño de los países.

El objetivo principal de un CI es la comparación de un individuo en relación a otros individuos y/o con algún punto de referencia. Algunas de las principales ventajas y desventajas de los CI's se presentan en la Tabla 2.12.

Tabla 2.12: Ventajas y Desventajas de los Indicadores Compuestos

Ventajas	Desventajas
▪ Puede resumir realidades complejas y multidimensionales a fin de respaldar a quienes toman decisiones. ▪ Son más fáciles de interpretar que una series de muchos indicadores simples por separado. ▪ Puede evaluar el rendimiento de los países a lo largo del tiempo. ▪ Reduce el tamaño visible de un conjunto de indicadores sin perder la base fundamental de la información. ▪ Es posible incluir más información dentro del límite de tamaño existente. ▪ Considera los problemas del rendimiento y progreso del país dentro del tema político. ▪ Facilita la comunicación con el público en general (es decir, ciudadanos, medios de comunicación, etc.) y promueve la responsabilidad. ▪ Ayuda a construir/sustentar narrativas para un público no profesional y alfabetizado. ▪ Permite a los usuarios comparar dimensiones complejas de manera efectiva.	▪ Puede enviar mensajes de política engañosos si están mal construidos o mal interpretados. ▪ Puede generar políticas simplistas. ▪ Puede ser mal utilizado, por ejemplo, para apoyar una política engañosa cuando el proceso de construcción no es transparente y/o carece de principios estadísticos o conceptuales sólidos. ▪ La selección de indicadores y ponderaciones podría ser objeto de disputa política. ▪ Puede ocultar fallas series en algunas dimensiones, aumentando la dificultad de identificar acciones correctivas adecuadas, si el proceso de construcción no es transparente. ▪ Puede llevar a políticas inapropiadas si se ignoran las dimensiones de rendimiento que son difíciles de medir.

La forma más simple de representar un indicador compuesto es como un promedio ponderado de indicadores individuales, es decir

$$CI_c = \sum_{i=1}^{m} w_{c,i}.y_{c_i}^n \ , \tag{2.128}$$

donde CI_c será el índice compuesto para el individuo j, $y_{c_i}^n$ el valor (posiblemente normalizado) del individuo j en el indicador i $(i = 1,2,\ldots,m)$ y $w_{c,i}$ el peso asignado al indicador i.

De manera general, los pesos deben satisfacer con

$$0 \leq w_{c,i} \leq 1 \qquad y \qquad \sum_{i=1}^{m} w_{c,i} = 1 \ .$$

Un problema típico presente en los CI se relaciona con el hecho de que los sub indicadores que intervienen en su formulación se muestran en unidades de medición bastante diversas. Esto implica que sea necesario un proceso de normalización (para más información vea Nardo y col. (2008)). Sin embargo, esto no resuelve del todo al problema, debido a que la normalización oscurece el objetivo del indicador compuesto, pues ya no se está resumiendo a los datos originales, sino una transformación de los mismos.
A pesar de mantener un sistema de ponderación fijo, las clasificaciones eventuales aún dependerán de la opción de normalización que se utilice y por ende, los individuos con clasificaciones más bajas utilizarán esta dependencia para cuestionar el uso de indicadores compuestos. Eliminar el requisito de normalización de los datos eliminaría esta dependencia y, por lo tanto, una fuente importante de críticas.

Adicionalmente, un segundo problema relacionado con los indicadores compuestos tiene que ver con el esquema de ponderación utilizado para agregación de los sub indicadores. Lo ideal sería ponderar y combinar a los sub indicadores de manera que reflejen la estructura subyacente del fenómeno evaluado. No obstante, la ponderación proveniente de las partes interesadas se caracteriza a menudo por fuertes desacuerdos interindividuales.

Una posible solución sería utilizar la ponderación fija, que es un caso particular de ponderación, en virtud de su simplicidad. Si bien es simple, puede llegar a ser completamente engañosa puesto que se favorecería a unos individuos y se perjudicaría a otros. Por consiguiente, esta dependencia al esquema de ponderación nuevamente se podría utilizar para quebrantar la credibilidad de los indicadores compuestos.

En resumen, a pesar de su uso creciente, los indicadores compuestos siguen siendo un tema polémico. La falta de una metodología estándar, la indeseable dependencia de las clasificaciones de los individuos en la etapa preliminar de normalización y el desacuerdo entre los expertos/partes interesadas sobre el esquema de ponderación específico utilizado para agregar a los sub indicadores, entre otros, son ítems que se han utilizado para persuadir su credibilidad.

Las acciones fundamentales para superar estas limitaciones, por un lado, se apoyan considerablemente en el Data Envelopment Analysis (DEA) o Análisis Envolvente de Datos, a causa de que su atractivo de invariancia respecto a las unidades de medida implica que se puede omitir la etapa de normalización. Por otro lado, el problema informativo sobre el conjunto de ponderaciones "correctas" se resuelve al generar ponderaciones flexibles de Benefit of the Doubt Approach (BOD) para cada individuo evaluado. Ejemplos de la aplicación de estas acciones, citados en Rogge (2008), se dan en el contexto de la evaluación del desempeño de políticas, como las utilizadas en la evaluación del desempeño de los países con respecto a la privatización agregada, la proporción de un sistema de ponderaciones alternativo para el Índice de Desarrollo Humano o en la construcción de un indicador general para el desarrollo sostenible.

A causa de todo lo anterior, de las dos partes de teoría en que está divido el manual, solo se considera la parte II (A Toolbox for Constructors), centrándonos principalmente en el Data Envelopment Analysis (DEA) y el Benefit of the Doubt Approach (BOD) pertenecientes al paso Weighting and Aggregation, así como también en los fundamentos teóricos de Cherchye y col. (2016) y Rogge (2008). Sin embargo, no se analizan los aspectos técnicos/computacionales que implican su construcción.

Weighting and aggregation

Cuando los pesos se utilizan en un marco de referencia, generalmente tienen un efecto significativo en el indicador compuesto y en los rankings de los individuos. Existen varias técnicas de ponderación (véase la Tabla 2.13). Algunas provienen de los modelos estadísticos tradicionales como el ACP/Análisis Factorial (AF), DEA y modelos de componentes no observadas (UCM), modelos participativos como el proceso de asignación de presupuesto (BAP), procesos analíticos de jerarquía (AHP), etc.

Independientemente del método que se utilice, los pesos (que generalmente miden la importancia de la variable asociada) son sentencias de importancia y criterios bien aceptados dentro de los analistas. Sin embargo, puede darse el caso que otro

tipo de expertos deseen dar pesos a los componentes (variables) dependiendo de las prioridades políticas.

Tabla 2.13: Compatibilidad entre métodos de agregación y ponderación.

Método de Ponderación	Método de Agregación [19]		
	Lineal [20]	Geométrico [20]	Multicriterio
APC/AF	Si	Si	Si
BOD	Si	No	No
UCM	Si	No	No
BAP	Si	Si	Si
AHP	Si	Si	No

La ausencia de una forma "objetivo" para determinar los pesos y los métodos de agregación no necesariamente conllevan al rechazo de la validez de los indicadores compuestos, siempre que todo el proceso sea transparente (Nardo y col. 2008).

Data Envelopment Analysis (DEA) y Benefit of the Doubt Approach (BOD)

El Data Envelopment Analysis (DEA) o Análisis Envolvente de Datos desarrollo inicialmente por Charnes, Cooper and Rhodes (1978) citado por Rogge (2008) emplea herramientas de programación lineal para estimar una frontera de eficiencia que será utilizada como punto de referencia en la medición del rendimiento relativo de los individuos. Para ello requiere de la construcción de un punto de referencia (**la frontera**) y la medición de la distancia entre los individuos en un marco multidimensional.

Los supuestos concernientes al punto de referencia son los siguientes:

a) Pesos positivos (cuanto más alto sea el valor de un indicador individual, mejor será el individuo asociado).

b) La no discriminación de los individuos que son mejores en cualquier dimensión simple (indicador individual), clasificándolos por igual; y,

c) Una combinación lineal de los mejores rendimientos, es decir convexidad en la frontera

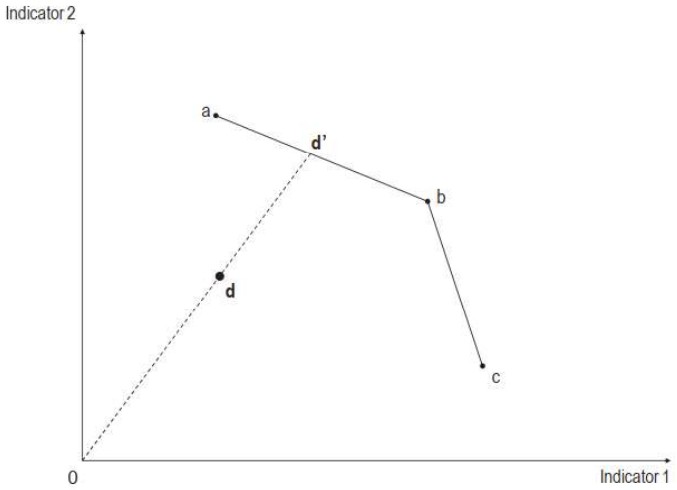
La distancia de cada individuo con respecto a la frontera se determina por la ubica-

[19] Los métodos de agregación se refieren al tipo de compensación entre los pesos. Es decir, un déficit en una dimensión (indicador) puede ser compensado por un excedente en otra

[20] Las agregaciones lineales son útiles cuando los indiciadores individuales tienen la misma unidad de medida, pues compensan proporcionalmente los pesos a los indicadores. En cambio, las agregaciones geométricas son adecuadas si se desea un cierto grado de no compensabilidad entre indicadores.

ción del individuo y su posición relativa a la frontera. A manera de ejemplo, para el caso simple de cuatro individuos y dos indicadores simples, la cuestión anterior se representa en la Figura 2.3

Figura 2.3: Frontera de Rendimiento del DEA.



La línea que conecta a los individuos a, b y c (clasificados como los de mejor desempeño) constituyen la frontera de rendimiento y el punto de referencia para el individuo d, mismo que se encuentra más allá de la frontera.

El indicador de rendimiento es la relación de la distancia entre el origen y el punto real observado sobre el origen y el punto proyectado en la frontera. Para nuestro ejemplo sería $0d/0d'$. Matemáticamente esta relación sería $(w_{1d}I_{1d} + w_{2d}I_{2d})/(w_{1d}I^*_{1d} + w_{2d}I^*_{2d})$, donde I^*_{id} es el valor fronterizo del indicador $i = 1,2$ e I_{id} su valor real.

Los individuos con el mejor rendimiento tendrán una puntuación de rendimiento igual a 1, mientras que lo de peor rendimiento menor a 1.

Actualmente, el alcance de la DEA ha crecido considerablemente en las últimas dos décadas, incluyendo las evaluaciones macro del rendimiento de la productividad de los países y varias aplicaciones en la construcción de CI's. Este último alcance ha sido llamado alternativamente como el método "Benefit of the Doubt Approach" introducido en el contexto de la evaluación del desempeño macroeconómico. La etiqueta de este método se deriva de uno de los principales puntos conceptuales de la DEA (la información sobre el esquema de ponderación apropiado para la evaluación comparativa del desempeño del individuo puede de hecho recuperarse de los propios datos del individuo Rogge (2008)) y se caracteriza por:

- La ponderación de este método está intrínsecamente relacionada con la idea de que incluso bajo una ponderación flexible, un individuo puede superar a

otro individuo en la muestra (ver (2.129) y (2.131)).

- Debido a la naturaleza flexible de los pesos (pesos adaptados a la elección de las unidades de medición), el problema de normalización de los CI puede ser evitado. Esta propiedad se conoce como unidad de invarianza.
- Si se dispone de información adicional (incluso aproximada) sobre los pesos apropiados, la misma se puede incorporar fácilmente en el ejercicio de evaluación.

Todas estas propiedades tienen como fin proporcionar, de alguna manera, un método práctico que siga la idea: "Si bien la posibilidad de llegar a un conjunto único de pesos es bastante improbable, esa singularidad no es realmente necesaria para emitir juicios acordados en muchas situaciones" expresada por Foster y Sen (1997, p. 206) citado en Rogge (2008).

A continuación se presenta la formulación del Benefit of the Doubt ($\boldsymbol{CI_c}$), definido como la relación entre el desempeño real de un individuo y su desempeño de referencia, paso a paso.

Paso 1: Idea de evaluación comparativa.

La puntuación de un individuo c está dada por:

$$CI_c = \frac{\text{rendimiento general real}}{\text{rendimiento general de referencia}} = \frac{\sum_{i=1}^{m} w_{c,i} y_{c,i}}{\sum_{i=1}^{m} w_{c,i} y_i^B} \tag{2.129}$$

donde y_i^B son los sub indicadores de referencia. Un valor del 100 % implica un rendimiento global similar al de los valores de referencia, un valor menor (más) que 1 se refiere a un peor (mejor) rendimiento.

Paso 2: Selección de un punto de referencia específico para el individuo.

Lo siguiente se relaciona con la identificación del rendimiento de referencia.
Se tomarán los puntos de referencia de la muestra observada. Es decir, el valor CI_c se manejará por comparación con otras observaciones existentes, en lugar de refe-

rencias externas. Esto genera la siguiente ecuación:

$$CI_c = \frac{\sum_{i=1}^{m} w_{c,i} y_{c,i}}{\underset{y_{j,i} \in \{\text{individuos estudiados}\}}{\text{maximizar}} \sum_{i=1}^{m} w_{c,i} y_{j,i}} \tag{2.130}$$

El denominador en (2.130) se obtiene al resolver un problema de maximización. La solución se alcanzará cuando un individuo que empleando ponderaciones $w_{c,i}$ obtiene la suma de ponderación máxima. En este sentido, el punto de referencia será endógeno (puede diferir de un individuo evaluado a otro).

Si por alguna razón, un individuo actúa como su propio punto de referencia (es decir, no existe otra observación de rendimiento superior) entonces el CI_c valdrá 1, lo que implica que se ha recuperado el valor máximo del indicador compuesto.

Paso 3: Selección de ponderaciones del Benefit of the Doubt específicos para cada individuo.

El problema de ponderación se maneja de forma separada para cada individuo y las ponderaciones específicas, asignadas a cada sub indicador se determinan de manera endógena. Dado que no se conoce las ponderaciones (políticas) "verdaderas" de un individuo, se supone que éstas se pueden inferir al observar sus fortalezas y debilidades relativas. Por lo tanto, a falta de información confiable, esto nos conduce a la idea del Benefit of the Doubt expresado en la siguiente ecuación:

$$CI_c = \underset{w_{c,i}}{\text{maximizar}} \ \frac{\sum_{i=1}^{m} w_{c,i} y_{c,i}}{\underset{y_{j,i} \in \{\text{individuos estudiados}\}}{\text{maximizar}} \sum_{i=1}^{m} w_{c,i} y_{j,i}} \tag{2.131}$$

Cualquier otro esquema de ponderación distinto empeoraría la posición del individuo evaluado con respecto a los demás individuos (Rogge 2008). Adicionalmente, dado que a cada individuo se le asigna el Benefit of the Doubt, no podrán afirmar que el rendimiento resultante se deba al esquema de ponderación (Cherchye y col. 2016).

A continuación se añaden dos restricciones más:

La **Restricción de Normalización** , $\sum_{i=1}^{m} w_{c,i} y_{j,i} \leq 1$, que indica que ningún otro individuo en el conjunto tiene un indicador compuesto resultante mayor que uno al

aplicar las ponderaciones óptimas para el individuo evaluado. Esta restricción destaca la idea evaluación comparativa (las ponderaciones más favorables para un individuo siempre se aplican a todas las observaciones.)
La **Restricción de Pesos no Negativos**, $w_{c,i} \geq 0$, que limita a que el indicador sea una función no decreciente de los sub indicadores. En consecuencia, $0 \leq CI_c \leq 1$; donde los valores más altos representarán un mejor rendimiento relativo a nivel global.

Considerando estas restricciones, el modelo será:

$$\begin{aligned} CI_c = \underset{w_{c,i}}{\text{maximizar}} \quad & \frac{\sum_{i=1}^{m} w_{c,i} y_{c,i}}{\underset{y_{j,i} \in \{\text{individuos estudiados}\}}{\text{maximizar}} \sum_{i=1}^{m} w_{c,i} y_{j,i}} \\ \text{sujeto a} \quad & \sum_{i=1}^{m} w_{c,i} y_{j,i} \leq 1 \quad \text{(n restricciones, una para cada individuo j)}, \\ & w_{c,i} \geq 0 \quad \text{(m restricciones, una para cada indicador i)}. \end{aligned} \tag{2.132}$$

Teniendo en cuenta el hecho de que por construcción, la observación de referencia alcanza el valor máximo del indicador compuesto de 1, el problema de maximización anterior se puede escribir en forma lineal (computacionalmente más fácil de manejar, por ejemplo, en solucionadores de Excel) de la siguiente manera:

$$\begin{aligned} CI_c = \underset{w_{c,i}}{\text{maximizar}} \quad & \sum_{i=1}^{m} w_{c,i} y_{c,i} \\ \text{sujeto a} \quad & \sum_{i=1}^{m} w_{c,i} y_{j,i} \leq 1 \quad \text{(n restricciones, una para cada individuo j)}, \\ & w_{c,i} \geq 0 \quad \text{(m restricciones, una para cada indicador i)}. \end{aligned} \tag{2.133}$$

Por otro lado, si se desea considerar referencias externas sobre la opinión de los expertos y su utilidad para resolver el problema de que si se cambia del tipo de transformación de los datos, (2.132) dé resultados diferentes, entonces es necesario incorporar restricciones sobre sub indicadores compartidos.

Restricciones de Sub indicadores Compartidos

El modelo propuesto hasta ahora (ecuación (2.132)) permite libremente estimar las ponderaciones para maximizar la puntuación de eficiencia relativa del individuo evaluado. Sin embargo, esta flexibilidad del Benefit of the Doubt podría enfrentar el riesgo de basar el rendimiento global en un subconjunto pequeño de todos los sub

indicadores, contradiciendo así las opiniones de los expertos sobre los pesos.

Por este motivo, para una mayor aceptación y credibilidad de los CI, se debe considerar la incorporación de la opinión de los expertos, cuando los mismos no están de acuerdo con su valor. Afortunadamente, los modelos DEA tiene la facilidad de agregar restricciones adicionales al problema básico definido en (2.132); y con toda garantía, este nuevo escenario del Benefit of the Doubt será más poderoso. Evidentemente, la naturaleza de estas restricciones (que consideran el criterio de los expertos) puede variar. En lo que sigue presentaremos brevemente algunas de ellas.

Con referencia a la ecuación (2.133), un sub indicador compartido, que es completamente independiente de las unidades de medición, es el producto de $y_{c,i}$ con $w_{c,i}$. Este enfoque alternativo permite que la ecuación (2.133) pueda ser reinterpretada como la suma de $i = 1,2,\ldots,m$ sub indicadores compartidos y no esté influenciado por las unidades de medida de los sub indicadores (Cherchye y col. 2016). Desde otra perspectiva, estos m términos pueden también ser considerados como "pie-share". En conjunto constituyen el volumen total del CI_c y el i-esimo término representará la porción del volumen del i-esimo sub indicador.
Así, todas la restricciones de "pie-share" se agregan al problema de maximización al añadir restricciones adicionales.

La interpretación del "pie-share" y las restricciones de sub indicadores compartidos permiten una representación fácil y natural de la información previa sobre la importancia de los componente del CI_c.

Estas restricciones se exhiben en la Tabla 2.14.

Tabla 2.14: Tipo de restricciones "pie-share"

Restricciones absolutas sobre sub indicadores compartidos $\alpha_i \leq w_{j,i}y_{j,i} \leq \beta_i$
Restricciones ordinarias sobre sub indicadores compartidos $w_{j,6}y_{j,6} \leq w_{j,5}y_{j,5} \leq w_{j,2}y_{j,2} \leq w_{j,3}y_{j,3} \leq w_{j,1}y_{j,1} \leq w_{j,7}y_{j,7} \leq w_{j,4}y_{j,4} \leq w_{j,8}y_{j,8}$
Restricciones relativas sobre sub indicadores compartidos $\alpha_i \leq \frac{w_{j,i}y_{j,i}}{w_{j,k}y_{j,k}} \leq \beta_i$
Restricciones proporcionales sobre sub indicadores compartidos $\alpha_i \leq \frac{w_i y_{j,i}}{\sum_{i=1}^{m} w_i y_{j,i}} \leq \beta_i$
Restricciones pertenecientes a categorías compartidas $\alpha \leq \frac{\sum_{i \in S_a} w_{j,i}y_{j,i}}{\sum_{i=1}^{m} w_{j,i}y_{j,i}} \leq \beta$

Para más detalles de cada una de estas nuevas restricciones puede ir a Cherchye

y col. (2016).

Finalmente, para más información referente al manual o a temas relacionados con indicadores compuestos puede ir a http://composite-indicators.jrc.ec.europa.eu/

CAPÍTULO 3

Construcción de la Base de Datos y del Índice de Salud

En este capítulo se presenta, en primera instancia, un análisis descriptivo de las variables existentes en la base de datos original. Esto con finalidad de establecer un adecuado criterio de depuración de datos. El proceso de depuración de los datos considerará aquellos campos que tengan al menos un cierto porcentaje de información (sin NA o datos perdidos) completa, para posteriormente utilizarlos en la construcción de los índices de salud, así como en el análisis de inferencia tanto paramétrico como no paramétrico.

En la construcción de los indicadores (índices) de salud, como primera aproximación (indicador ACP) se basará en la aplicación del ACP a los datos depurados, debido que se intenta tener la mejor información posible sin tener que considerar datos perdidos, porque considerarlos sale fuera del alcance del presente trabajo de titulación. Además se establecerá que las variables objetas de estudio serán las proyecciones de los individuos en el primer eje proporcionado por el ACP para los grupos más sobresalientes (más numerosos).

Por otro lado, el segundo indicador (indicador CI) se obtiene como resultado de la aplicación del Benefit of the Doubt Approach a las proyecciones (datos transformados de las proyecciones dadas por el ACP) de los individuos de las componentes principales cuyos valores propios son mayores que 1, mismas que son resultado de la aplicación del ACP a los datos depurados pero con un cambio de escala en el sentido del mejor individuo sano (vea (A.4) de la parte de anexos). En este caso, se hará la consideración que la variable objeta de estudio será las ponderaciones dadas por el Benefit of the Doubt Approach para los grupos más sobresalientes.

3.1. Análisis y Depuración de Datos

La base de datos a utilizar pertenece a una empresa de servicios médicos, misma que tiene 106 variables (campos) y 23335 registros en la variable que contiene la mayor cantidad de información. Las observaciones corresponden a datos clínicos de diagnóstico general de individuos sanos que se encuentran en diferentes sectores de las actividades económicas durante el periodo 2015-2017. Por ejemplo, de la Clasificación Internacional Industrial Uniforme (CIIU) adaptada a la economía del Ecuador tenemos que

- A - Agricultura, Ganadería, Silvicultura y Pesca.
- B - Explotación de Minas y Canteras.
- C - Industrias Manufactureras.
- D - Suministro de Electricidad, Gas, Vapor y Aire Acondicionado.
- E - Distribución de Agua; Alcantarillado, Gestión de Desechos y Actividades de Saneamiento.
- F - Construcción.

son algunas de las actividades (para más detalle vea la Tabla 3.2) en las que registró al menos un dato de diagnóstico clínico por parte de los individuos.

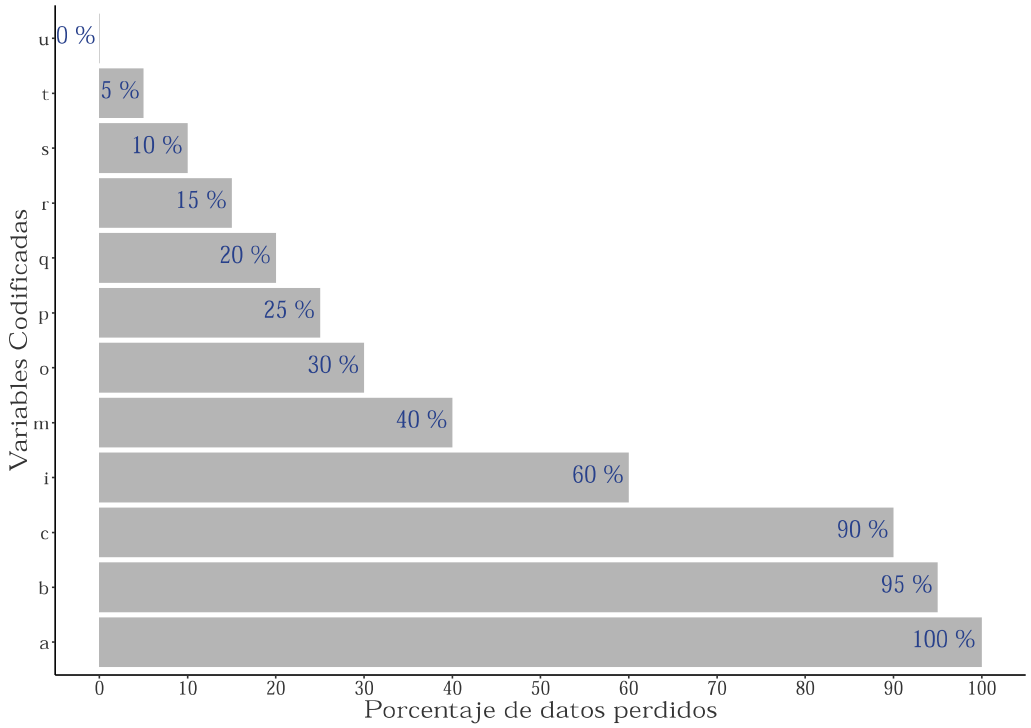
Se eliminan las variables *Basofilos, Eosinofilos, Eritrocitos, Linfocitos, Monocitos, Leucocitos, Neutrofilos* debido a inconsistencias presentadas en relación a su valor. De esta manera, los datos depurados tendrán 99 campos.

A los campos resultantes se los agrupa (codifica) de acuerdo a cierto porcentaje de datos perdidos presente en cada uno de ellos, esto con la finalidad de establecer el máximo porcentaje de información incompleta. A modo de ejemplo, la codificación "a" representará a las variables que tengan entre el 95 y 100 por ciento de datos perdidos. Algunos de los miembros pertenecientes a este grupo son *Calcio Serico Total, Hierro Serico, Potasio(k), Sodio(NA), Transferrina, Glucosa Postprandial, Insulina Basal, etc* [21].

El porcentaje de datos perdidos por codificación se presenta en la siguiente figura.

[21] La codificación de todas las variables así como el número total y porcentaje de datos perdidos por variable se presenta en (A.1) de la parte de anexos.

Figura 3.1: Porcentaje de datos perdidos por tipo de codificación.



De la Figura 3.1 es fácil ver que las variables con el 0% de datos perdidos corresponden a los miembros pertenecientes a la codificación "u", por el contrario, los campos con la mayor cantidad de información incompleta corresponde a los integrantes de la codificación "a".

El número de variables por tipo de codificación se presenta en la Tabla 3.1

Tabla 3.1: Número de variables por tipo de codificación.

Codificación	Nro. de variables	% de NA's
a	50	(95%,100%]
b	5	(90%,95%]
c	3	(85%,90%]
i	3	(55%,60%]
m	1	(35%,40%]
o	2	(25%,30%]
p	4	(20%,25%]
q	1	(15%,20%]
r	2	(10%,15%]
s	6	(5%,10%]
t	11	(0%,5%]
u	11	0%

En base a todo esto, el proceso de depuración considerará solo a las variables que tengan a lo mucho el 10% de datos perdidos. Esto implica que se va a trabajar con 28 campos [22]. Para estos campos se eliminan los registros faltantes a nivel de fila, ocasionando que posteriormente el número de observaciones presentes en cada uno

de ellos sea 18049.

La desagregación de este número por actividades económicas respecto a la variable *CIIU4.01* se exhibe en la Tabla 3.2.

Tabla 3.2: Actividades económicas y número de observaciones.

Actividad Económica	Nro. de observaciones
B - EXPLOTACIÓN DE MINAS Y CANTERAS.	7415
K - ACTIVIDADES FINANCIERAS Y DE SEGUROS.	2296
M - ACTIVIDADES PROFESIONALES, CIENTÍFICAS Y TÉCNICAS.	2145
F - CONSTRUCCIÓN.	1760
G - COMERCIO AL POR MAYOR Y AL POR MENOR; REPARACIÓN DE VEHÍCULOS AUTOMOTORES Y MOTOCICLETAS.	1512
N - ACTIVIDADES DE SERVICIOS ADMINISTRATIVOS Y DE APOYO.	709
J - INFORMACIÓN Y COMUNICACIÓN.	586
Q - ACTIVIDADES DE ATENCIÓN DE LA SALUD HUMANA Y DE ASISTENCIA SOCIAL.	548
C - INDUSTRIAS MANUFACTURERAS.	434
H - TRANSPORTE Y ALMACENAMIENTO.	309
D - SUMINISTRO DE ELECTRICIDAD, GAS, VAPOR Y AIRE ACONDICIONADO.	96
U-ACTIVIDADES DE ORGANIZACIONES Y ÓRGANOS EXTRATERRITORIALES.	62
E - DISTRIBUCIÓN DE AGUA; ALCANTARILLADO, GESTIÓN DE DESECHOS Y ACTIVIDADES DE SANEAMIENTO.	48
P - ENSEÑANZA.	37
S - OTRAS ACTIVIDADES DE SERVICIOS.	37
A - AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA.	36
O-ADMINISTRACIÓN PÚBLICA Y DEFENSA; PLANES DE SEGURIDAD SOCIAL DE AFILIACIÓN OBLIGATORIA.	11
L - ACTIVIDADES INMOBILIARIAS.	8

En virtud de la Tabla 3.2, los registros por filas con los que se va a realizar el estudio serán los pertenecientes a los tres grupos más numerosos, es decir a las actividades *B - Explotación de Minas y Canteras, K - Actividades Financieras y de Seguros y M - Actividades Profesionales, Científicas y Técnicas*. Esto ocasiona que nuestros datos depurados tengan 28 variables con 11856 observaciones por variable.

Si el lector piensa que el proceso de depuración debe considerar a las variables que tengan un porcentaje distinto del 10% de datos perdidos, puede cambiar este porcentaje ingresando a la pestaña ***Depuración de datos –Índice de salud–Aplicación*** del enlace web https://cristian-guatemal-work.shinyapps.io/TesisCG/

3.2. Construcción de los Índices de Salud

Por facilidad, en la construcción de los dos indicadores de salud solo se considerará a variables cuantitativas (solo se utilizarán 14 de las 28 variables resultantes del proceso de depuración [23], debido a que el resto de variables son por ejemplo cédu-

[22] Las variables que cumplen la condición de a lo mucho el 10% de datos perdidos se presenta en (A.2) de la parte de anexos

[23] Estas variables se presentan en (A.3) de la parte de anexos.

la de identificación, fecha de admisión, tipo de hábito, fecha de nacimiento, género, etc; mismas que no aportan mucho en nuestro estudio.), y se basarán en la reducción de dimensiones mediante el proceso del Análisis de Componentes Principales (indicador ACP) (véase 2.4, pág. 86) y el índice multidimensional del DEA aplicado al Benefit of the Doubt (indicador CI)(véase 2.5.1, pág. 92).
Pero antes, como se dijo en (2.4), previo a utilizar el ACP se debe verificar si la correlación entre las variables analizadas es lo suficientemente grande como para justificar la factorización de la matriz de coeficientes de correlación.

En este contexto, si aplicamos el Test de esfericidad de Bartlett expresado como

$$-[n-1-(2k+5)/6]ln|R| \sim \chi^2_{(k^2-k)/2} \tag{3.1}$$

donde k es el número de variables, n el tamaño de la muestra y R la matriz de correlaciones; para el caso de estudio del indicador ACP con

$$n = 11856$$
$$k = 14$$

y R (obtenida de los datos depurados de la sección 3.1) presentada en la Tabla 3.3

Tabla 3.3: Matriz de correlaciones de los datos depurados de la sección 3.1.

	peso	estt	imc	prs_	pr_2	CONC	HGB	VPM	PLAQ	VCM	HCM.	HEMA	ANCH	GLUC
peso	1.00	0.01	0.81	0.35	0.32	0.00	0.32	0.09	-0.05	-0.09	-0.01	0.02	0.02	0.21
estt	0.01	1.00	0.00	0.02	0.01	0.00	0.00	0.00	-0.01	-0.01	-0.01	0.00	0.00	0.00
imc	0.81	0.00	1.00	0.32	0.29	0.00	0.16	0.05	0.02	-0.07	-0.03	0.01	0.05	0.20
prs_	0.35	0.02	0.32	1.00	0.62	0.00	0.11	-0.01	-0.04	0.01	-0.02	0.01	0.01	0.13
pr_2	0.32	0.01	0.29	0.62	1.00	0.00	0.15	0.05	-0.04	-0.01	0.01	0.02	-0.02	0.13
CONC	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	-0.01	0.00	0.00	0.26	0.00
HGB	0.32	0.00	0.16	0.11	0.15	0.00	1.00	-0.01	-0.07	0.13	0.32	0.09	-0.12	0.13
VPM	0.09	0.00	0.05	-0.01	0.05	0.00	-0.01	1.00	-0.39	-0.14	-0.01	-0.01	-0.01	0.06
PLAQ	-0.05	-0.01	0.02	-0.04	-0.04	0.00	-0.07	-0.39	1.00	-0.10	-0.15	0.01	0.08	-0.03
VCM	-0.09	-0.01	-0.07	0.01	-0.01	-0.01	0.13	-0.14	-0.10	1.00	0.72	0.02	-0.14	-0.08
HCM,	-0.01	-0.01	-0.03	-0.02	0.01	0.00	0.32	-0.01	-0.15	0.72	1.00	0.02	-0.28	-0.01
HEMA	0.02	0.00	0.01	0.01	0.02	0.00	0.09	-0.01	0.01	0.02	0.02	1.00	-0.01	0.01
ANCH	0.02	0.00	0.05	0.01	-0.02	0.26	-0.12	-0.01	0.08	-0.14	-0.28	-0.01	1.00	0.00
GLUC	0.21	0.00	0.20	0.13	0.13	0.00	0.13	0.06	-0.03	-0.08	-0.01	0.01	0.00	1.00

entonces (3.1) adquiere el valor de 337.258 con un correspondiente p-valor menor a 2.22e-16.
De manera similar para el indicador CI, si aplicamos el test de esfericidad para

$$n = 11856$$
$$k = 12$$

y R (obtenida de los datos depurados de la sección 3.1 en relación al mejor individuo sano considerando (A.4) de la parte de anexos) presentada en la Tabla 3.4

Tabla 3.4: Matriz de correlaciones de los datos depurados de la sección 3.1 en relación al mejor individuo sano dado en (A.4).

	imc	prs_	pr_2	conc	hgb	vpm	plaq	vcm	hcm	hema	anch	gluc
imc	1.00	-0.15	-0.18	-0.00	-0.03	-0.06	0.01	-0.02	0.01	0.00	-0.01	0.10
prs_	-0.15	1.00	0.45	-0.00	0.01	0.01	0.02	0.02	0.00	-0.00	0.01	-0.00
pr_2	-0.18	0.45	1.00	0.00	0.03	0.05	-0.00	0.02	0.03	-0.01	0.02	-0.02
conc	-0.00	-0.00	0.00	1.00	-0.01	0.00	-0.01	-0.01	0.00	-0.00	0.33	-0.00
hgb	-0.03	0.01	0.03	-0.01	1.00	0.01	0.04	0.11	0.19	0.04	0.08	0.01
vpm	-0.06	0.01	0.05	0.00	0.01	1.00	-0.09	0.11	0.03	0.01	-0.03	0.01
plaq	0.01	0.02	-0.00	-0.01	0.04	-0.09	1.00	0.03	0.04	0.01	0.06	0.03
vcm	-0.02	0.02	0.02	-0.01	0.11	0.11	0.03	1.00	0.36	0.01	0.08	0.01
hcm	0.01	0.00	0.03	0.00	0.19	0.03	0.04	0.36	1.00	-0.00	0.18	0.02
hema	0.00	-0.00	-0.01	-0.00	0.04	0.01	0.01	0.01	-0.00	1.00	-0.01	0.00
anch	-0.01	0.01	0.02	0.33	0.08	-0.03	0.06	0.08	0.18	-0.01	1.00	-0.01
gluc	0.10	-0.00	-0.02	-0.00	0.01	0.01	0.03	0.01	0.02	0.00	-0.01	1.00

entonces (3.1) adquiere el valor de 253.25 con un correspondiente p-valor menor a 2.22e-16.

Esto provoca que en ambos caso se rechace la hipótesis nula de esfericidad y se asegure la existencia de variables correlacionadas.

Por otro lado, debido a que nuestro valor de n es grande, se debe calcular el Índice Kaiser-Meyer-Olkin. El resultado del cálculo de este índice, para la matriz de datos en relación al indicador ACP es 0.5680601 y para la matriz de datos en relación al indicador CI es 0.5417473; y, como estos valores están entre 0.5 y 1 entonces es apropiado continuar con la aplicación del ACP (Suárez 2007). Adicionalmente, como el valor del determinante de la matriz de correlaciones, relacionada al indicador ACP es igual a 0.03784842 y para la matriz relacionada al indicador CI es 0.5198626, entonces se sigue favoreciendo la utilización del ACP.

Así, en virtud de todo lo anterior, es adecuado el uso del ACP en la construcción de los índices de salud. Los resultados se presentan a continuación.

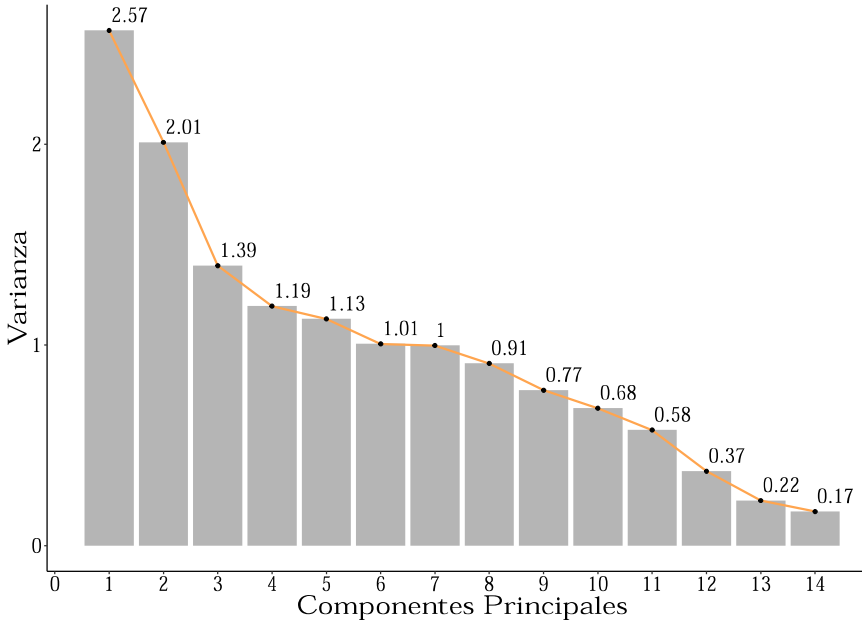
Índice basado en el ACP.

Debido a que las variables consideradas en el presente estudio tienen escalas y medidas diferentes, es necesario estandarizarlas como en (2.127). Luego, en base a la Tabla 3.5, la Figura 3.2 y los criterios de selección del número de componentes presentados en Peña (2002), tenemos que el número de componentes que resumen la mayor parte de la información de las variables iniciales son las seis primeras.
A pesar de ello, solo se va a considerar la primera y segunda componente, junto a sus respectivas proyecciones [24].

Tabla 3.5: Importancia de las Componentes Principales para el Índice ACP.

	Desviación Estándar	Proporción de Varianza	Proporción de Varianza Acumulada
CP1	1.6019563	0.18330457	0.1833046
CP2	1.4173973	0.14350107	0.3268056
CP3	1.1809095	0.09961052	0.4264162
CP4	1.0922154	0.08520961	0.5116258
CP5	1.0626048	0.08065206	0.5922778
CP6	1.0025204	0.07178908	0.6640669
CP7	0.9983660	0.07119533	0.7352622
CP8	0.9526591	0.06482566	0.8000879
CP9	0.8798075	0.05529009	0.8553780
CP10	0.8273134	0.04888910	0.9042671
CP11	0.7588364	0.04113090	0.9453980
CP12	0.6088488	0.02647835	0.9718763
CP13	0.4734753	0.01601278	0.9878891
CP14	0.4117673	0.01211088	1.0000000

Figura 3.2: Proporción de variabilidad explicada para cada componente (Índice ACP).



De la Tabla 3.6, que presenta los pesos de los dos primeros componentes principales, el primer componente principal, que acumula el 18.33 % de información tiene por ecuación

$$Y_1 = 0.5245peso + 0.0127estt + 0.4928imc + 0.4219prs\_ + 0.4129pr_2 + \ldots$$
$$\ldots - 0.0057anch + 0.2298gluc \ ,$$

[24] Las componentes principales y la proyección de las primeras veinte observaciones en todas las componentes principales se presenta en (A.5) de la parte de anexos.

Tabla 3.6: Pesos de los primeros dos componentes principales (índice ACP).

	PC1	PC2
peso	0.5245	0.0467
estt	0.0127	0.0112
imc	0.4928	0.0862
prs_	0.4219	0.0206
pr_2	0.4129	-0.0027
conc	-0.0028	0.0945
hgb	0.2538	-0.2961
vpm	0.0867	0.0297
plaq	-0.0728	0.1817
vcm	-0.0375	-0.5795
hcm,	0.0250	-0.6360
hema	0.0357	-0.0537
anch	-0.0057	0.3383
gluc	0.2298	0.0411

y acoge como variables principales a peso, imc, prs_ y prs_2.

En cambio, para la segunda componente principal, que agrupa el 14.35% de información tenemos que las variables que contribuyen en gran medida a su variación son hcm,vcm y anch.

La expresión matemática de este eje es la siguiente:

$$Y_2 = 0.0467peso + 0.0112estt + 0.0862imc + 0.0206prs\_ - 0.0027pr_2 + \ldots \\ \ldots + 0.3383anch + 0.0411gluc.$$

Gráficamente, la contribución de las variables a la formación de los dos primeros componentes se presenta en la Figura 3.3.

Para los individuos en el plano (Y_1, Y_2), por ejemplo, se tendrá que los individuos con valores altos en las variables peso, imc, prs_, prs_2, conc y anch se ubicarán en el primer cuadrante. Las personas que tienen valores altos para las variables peso, imc, prs_ y prs_2 pero bajo vcm y hcm se ubicarán en el cuarto cuadrante; y así sucesivamente.

Las primeras veinte nuevas coordenadas (proyecciones) de los individuos que se obtienen al reemplazar las observaciones de las variables originales en los dos primeros componentes, se presentan en la Tabla 3.7.

Figura 3.3: Gráfica de pesos de los dos primeros componentes (Índice ACP).

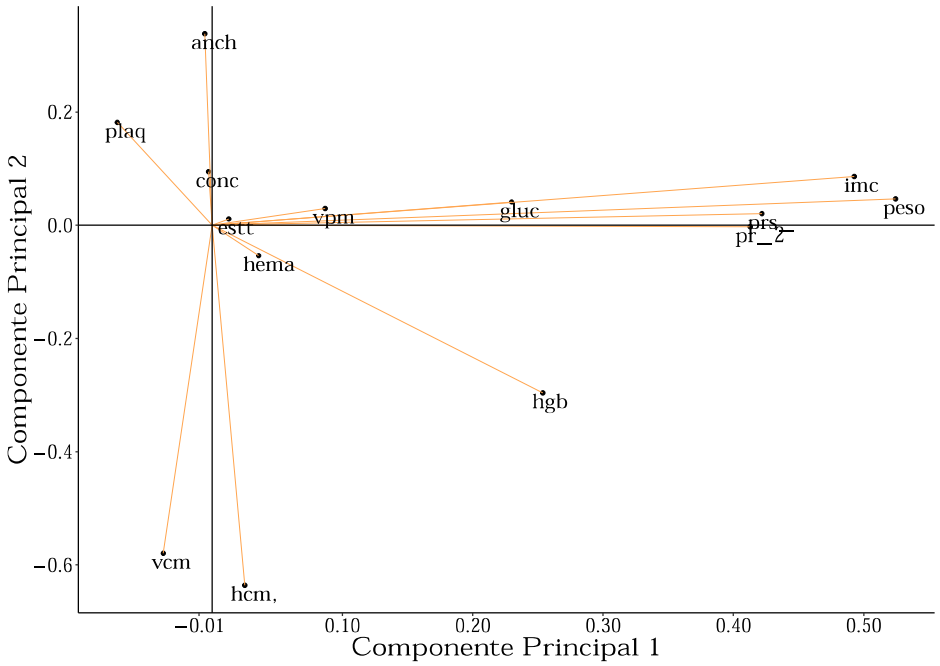


Tabla 3.7: Proyección de los 20 primeros individuos en las dos primeras componentes principales (Índice ACP)

	PC1	PC2
1	0.314965	-1.281236
2	-1.038018	-1.756577
3	-3.284438	-0.141684
4	0.083076	-1.749775
5	1.581764	-1.260650
6	1.241496	-1.615989
7	0.852427	1.044862
8	0.888750	0.570441
9	0.550141	0.662910
10	-1.210098	0.277280
11	-0.817526	-1.019584
12	-4.147063	-0.423389
13	0.363694	0.099121
14	-4.708464	1.654549
15	-1.320563	0.012210
16	0.193574	-1.094196
17	-2.115753	0.787109
18	0.322749	-0.940947
19	-0.625968	-0.210200
20	-0.430472	-0.323347
⋮	⋮	⋮

Basándonos en estas coordenadas y los grupos presentes en la Tabla 3.2, se establece que la primera muestra corresponderá a las proyecciones en la primera componente principal de los individuos pertenecientes a la actividad *B - Explotación de Minas y Canteras*, la segunda muestra a las proyecciones en la primera componente principal de los individuos pertenecientes a la actividad *K - Actividades Financieras*

y de Seguros y finalmente, la tercera muestra abarca las proyecciones en la primera componente principal de los individuos pertenecientes a la actividad *M - Actividades Profesionales, Científicas y Técnicas*.

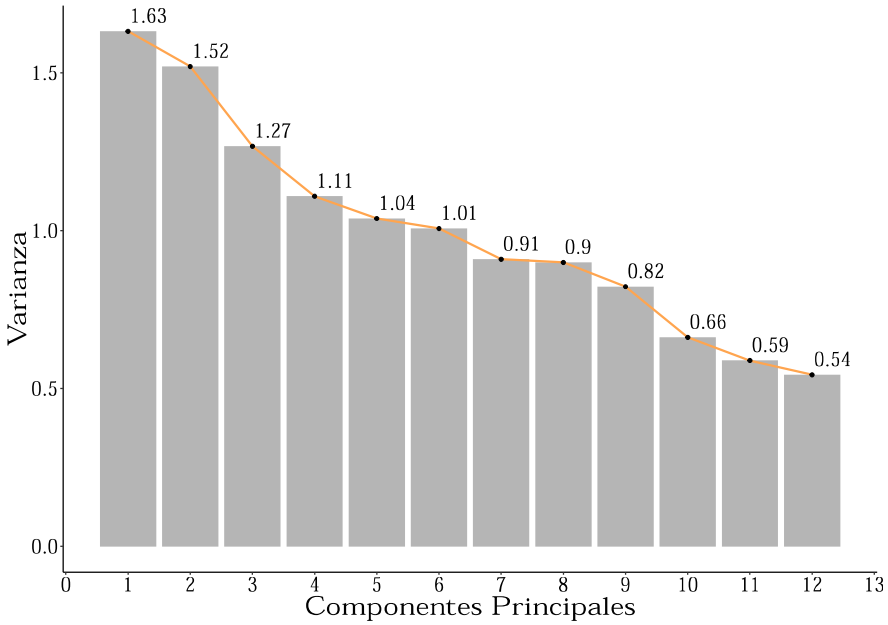
Índice basado en el Benefit of the Doubt Approach.

Bajo la misma metodología en lo relacionado a la elección del número de componentes principales aplicado en la construcción del indicador ACP; y en relación a la Tabla 3.8 y la Figura 3.4, que son obtenidas de la aplicación del ACP a los datos depurados que consideran el criterio del mejor individuo sano (vea (A.4) de la parte de anexos) tenemos que el número de componentes que resumen la mayor parte de la información de las variables son las seis primeras.

Tabla 3.8: Importancia de las Componentes Principales para el Índice CI.

	Desviación Estándar	Proporción de Varianza	Proporción de Varianza Acumulada
CP1	1.2774232	0.1359842	0.1359842
CP2	1.2329045	0.1266711	0.2626553
CP3	1.1258463	0.1056275	0.3682828
CP4	1.0531453	0.0924262	0.4607090
CP5	1.0191106	0.0865489	0.5472579
CP6	1.0035081	0.0839190	0.6311769
CP7	0.9537018	0.0757956	0.7069725
CP8	0.9484615	0.0749649	0.7819375
CP9	0.9069847	0.0685518	0.8504892
CP10	0.8137461	0.0551819	0.9056711
CP11	0.7671873	0.0490480	0.9547192
CP12	0.7371365	0.0452808	1.0000000

Figura 3.4: Proporción de variabilidad explicada para cada componente (Índice CI).



De la Tabla 3.9, que presenta los pesos de los primeros seis componentes principales para el índice CI, el primer componente principal, que acumula el 13.60% de información tiene por ecuación

$$Y_1 = -0.2414imc + 0.3716prs\_ + 0.4072pr_2 + \ldots$$
$$\ldots + 0.3441anch - 0.0301gluc \ ,$$

y acoge como variables principales a vcm, hcm, hgb y anch.

Tabla 3.9: Pesos de los seis primeros componentes principales (índice CI)

	PC1	PC2	PC3	PC4	PC5	PC6
imc	-0.2414	0.3269	-0.0193	0.1739	-0.3842	-0.0279
prs_	0.3716	-0.5030	0.0501	0.1765	-0.1859	-0.0327
pr_2	0.4072	-0.5007	0.0356	0.0949	-0.1508	-0.0194
conc	0.1748	0.1756	0.6545	-0.2280	-0.1138	-0.0994
hgb	0.2954	0.1867	-0.1859	0.1707	0.2183	-0.1719
vpm	0.1274	-0.0417	-0.2461	-0.5928	-0.2565	-0.1433
plaq	0.0866	0.0987	0.0724	0.6370	0.1542	0.0567
vcm	0.4129	0.2913	-0.3478	-0.1042	-0.0300	0.1055
hcm	0.4544	0.3610	-0.2452	0.0323	0.0095	0.1270
hema	0.0095	0.0325	-0.0576	0.0442	0.2049	-0.9316
anch	0.3441	0.2888	0.5263	-0.0472	-0.0139	-0.0103
gluc	-0.0301	0.1127	-0.0761	0.2743	-0.7757	-0.1982

Para la segunda componente principal, que agrupa el 12.67% de información tenemos que las variables que contribuyen en gran medida a su variación son icm, gluc, prs_ y pr_2. La expresión matemática de este eje es la siguiente:

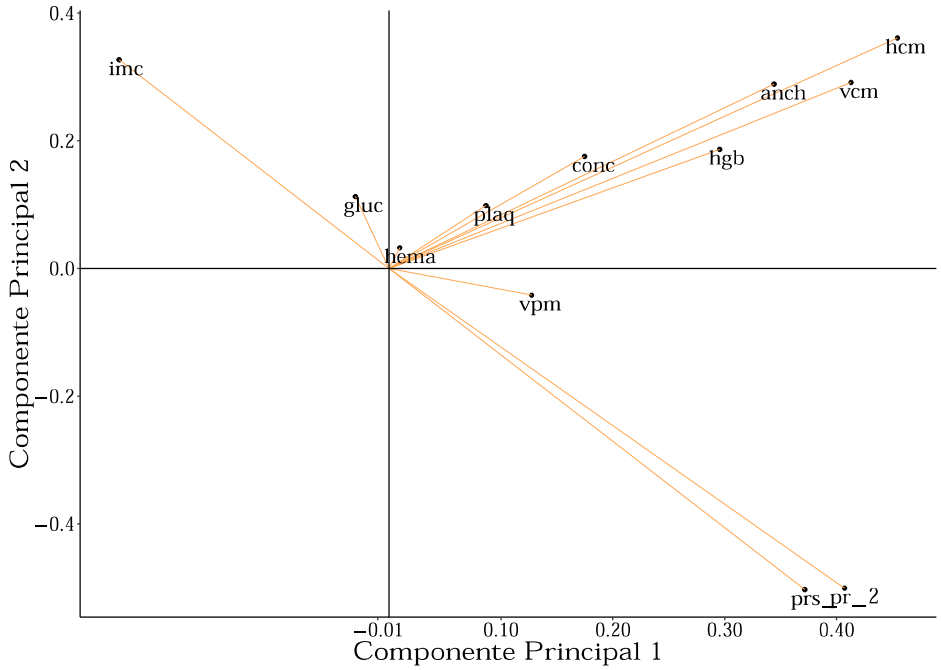
$$Y_2 = 0.3269imc - 0.5030prs\_ - 0.5007pr_2 + \ldots$$
$$\ldots + 0.2888anch + 0.1127gluc.$$

De manera similar para el resto de componentes.

Gráficamente, la contribución de las variables a la formación de los dos primeros componentes se presenta en la Figura 3.5.

Para los individuos en el plano (Y_1,Y_2), por ejemplo, se tendrá que los individuos con valores altos en las variables hcm, vcm, anch y hgb se ubicarán en el primer cuadrante. Las personas que tienen valores altos en las variables imc y gluc se ubicarán en el segundo cuadrante; y así en adelante.

Figura 3.5: Gráfica de pesos de los dos primeros componentes (Índice CI).



Las primeras veinte nuevas coordenadas (proyecciones) de los individuos que se obtienen al reemplazar las observaciones de las variables originales en los seis primeros componentes, se presentan en la Tabla 3.10.

Tabla 3.10: Proyección de los 20 primeros individuos en las seis primeras componentes principales (Índice CI).

	PC1	PC2	PC3	PC4	PC5	PC6
1	0.827925	-0.870675	-0.572182	-0.656463	-0.955230	-0.081755
2	-0.241928	1.135372	-0.975677	-0.257579	-0.030649	-0.055580
3	-1.645010	0.833140	-0.291989	0.069063	-1.066197	0.062212
4	-0.213055	0.627646	0.249890	-0.574887	0.554421	0.253881
5	0.745995	-1.518025	-0.478134	-0.985846	-0.411090	0.270433
6	0.543119	-0.656765	-0.923508	-1.259569	-0.521884	0.340794
7	0.862237	-1.013417	0.133169	0.135879	-0.889349	0.114148
8	-0.000522	0.429540	-0.762921	-1.212473	0.632388	0.483543
9	-1.112860	-0.104372	0.731649	0.766231	0.990806	0.587888
10	-1.495947	-0.697264	0.793447	0.129203	-0.312502	0.032277
11	-1.055040	0.974963	-0.677853	-2.029558	-1.129440	-0.574441
12	1.532457	-3.856045	0.385149	1.009719	-1.547592	-0.504473
13	-0.145303	-0.185042	1.830684	1.480372	0.948973	-0.103018
14	-0.316616	1.549896	-0.214065	1.138030	-4.619508	-1.232387
15	-1.240301	0.667427	-0.428708	-0.380526	-0.596161	-0.368368
16	-0.997685	0.746667	-0.540974	-1.262403	0.697527	0.415196
17	-1.179693	1.100586	-0.022000	1.671221	-2.455485	-0.705520
18	-0.926220	-0.063737	0.472635	0.267901	0.724645	0.050949
19	-1.731567	-0.678080	0.864053	0.181005	-0.128758	0.165505
20	-1.487828	0.685474	0.296188	0.876689	0.958213	0.079252
⋮	⋮	⋮	⋮	⋮	⋮	⋮

A estas nuevas coordenadas se las transforma mediante el método **Min-Max** (para

más detalle vea Nardo y col. (2008)). El resultado para las primeras 20 observaciones se presenta en la Tabla 3.11.

Tabla 3.11: Transformación mediante Min-Max de la proyección de los 20 primeros individuos de las seis primeras componentes principales (Índice CI).

	PC1	PC2	PC3	PC4	PC5	PC6
1	0.175463	0.225324	0.062832	0.722834	0.391970	0.975109
2	0.146151	0.277379	0.058675	0.733571	0.416184	0.975361
3	0.107710	0.269536	0.065719	0.742364	0.389064	0.976498
4	0.146942	0.264204	0.071302	0.725030	0.431507	0.978347
5	0.173218	0.208526	0.063801	0.713968	0.406221	0.978507
6	0.167660	0.230875	0.059212	0.706600	0.403319	0.979186
7	0.176403	0.221620	0.070099	0.744162	0.393695	0.976999
8	0.152765	0.259063	0.060867	0.707868	0.433548	0.980563
9	0.122290	0.245209	0.076265	0.761130	0.442935	0.981570
10	0.111794	0.229824	0.076902	0.743983	0.408803	0.976209
11	0.123874	0.273217	0.061743	0.685874	0.387408	0.970355
12	0.194765	0.147856	0.072695	0.767684	0.376456	0.971030
13	0.148799	0.243115	0.087588	0.780353	0.441840	0.974903
14	0.144105	0.288136	0.066522	0.771138	0.296005	0.964006
15	0.118798	0.265236	0.064310	0.730262	0.401374	0.972343
16	0.125445	0.267293	0.063153	0.706524	0.435254	0.979904
17	0.120459	0.276476	0.068500	0.785490	0.352679	0.969090
18	0.127403	0.246263	0.073597	0.747716	0.435965	0.976389
19	0.105339	0.230322	0.077629	0.745377	0.413615	0.977494
20	0.112017	0.265705	0.071779	0.764103	0.442082	0.976662
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Cada una de estas nuevas seis componentes, con sus respectivas observaciones transformadas (Tabla 3.11) van a conformar el indicador compuesto de Benefit of the Doubt Approach, **CI**. Los primeros 1200 valores de este indicador se muestran en la Tabla 3.12. Basándonos en estos valores transformados y los grupos presentes en la Tabla 3.2, se establece que la primera muestra corresponderá a los valores del indicador CI pertenecientes a la actividad *B - Explotación de Minas y Canteras*, la segunda muestra a los valores del indicador CI pertenecientes a la actividad *K - Actividades Financieras y de Seguros* y finalmente, la tercera muestra abarcará a los valores del indicador CI pertenecientes a la actividad *M - Actividades Profesionales, Científicas y Técnicas*.

Adicionalmente, para el caso del índice ACP, el lector tiene la libertad de escoger las proyecciones en la segunda, tercera o cualquier componente principal, así como el escoger cualquier actividad económica para no necesariamente condicionarse a las tres que se presentan en este estudio. Para el índice CI, se permite escoger cualquier número de componentes principales (para aplicar luego el CI) y cualquier actividad económica. Esto lo puede hacer ingresando a la subpestaña ***Muestras y Gráfica de Probabilidad Normal*** de la pestaña ***Depuración de datos –Índice de salud–Aplicación*** en el enlace web https://cristian-guatemal-work.shinyapps.io/TesisCG/.

Tabla 3.12: Primeros 1200 valores del indicador CI.

0.977752	0.969305	0.974378	0.986508	0.974288	0.981890	0.978777	0.974726	0.987579	0.979339	0.982099	0.990469	0.982317	0.989072	0.985589
0.977956	0.983186	0.980447	0.981896	0.981869	0.977973	0.973764	0.985856	0.990385	0.981627	0.979918	0.984658	0.980239	0.985147	0.981644
0.980113	0.982586	0.986350	0.986958	0.980400	0.972858	0.977602	0.981794	0.983859	0.981360	0.977432	0.978822	0.979860	0.983775	0.977248
0.981905	0.978929	0.992688	0.973872	0.979859	0.985446	0.979117	0.990653	0.982009	0.979262	0.991053	0.975331	0.982203	0.975861	0.974757
0.981238	0.990533	0.986797	0.974589	0.985160	0.980876	1.000000	0.986633	0.979053	0.977908	0.977953	0.979431	0.992641	0.978963	0.975341
0.981433	0.987332	0.984103	0.984387	0.981037	0.978940	0.984172	0.977728	0.975613	0.988458	0.980952	0.986380	0.984875	0.980024	0.973499
0.981047	0.981278	0.978806	0.982792	0.983009	0.978094	0.977671	0.980252	0.977965	0.977580	0.977845	0.986909	0.981453	0.979728	0.982072
0.983592	0.979673	0.974153	0.978808	0.982423	0.974921	0.975993	0.981198	0.975336	0.973283	0.975084	0.983697	0.985126	0.981591	0.974412
0.989319	0.981532	0.981541	0.985539	0.972495	0.990643	0.984496	0.977308	0.975413	0.977010	0.972273	0.995451	0.988840	0.983318	0.976554
0.980830	0.986549	0.971062	0.980052	0.981031	0.984335	0.989985	0.982891	0.978226	0.976588	0.985116	0.984708	0.981214	0.986550	0.984195
0.972903	0.980079	0.971832	0.979087	0.981820	0.984754	0.985388	0.977405	0.977166	0.984365	0.980163	0.982769	0.974414	0.984514	0.987467
0.976909	0.978991	0.975572	0.973805	0.982582	0.974104	0.984401	0.990407	0.987951	0.975655	0.985683	0.982432	0.979967	0.979069	0.977407
0.985161	0.979891	0.985929	0.979808	0.980132	0.975263	0.987012	0.983648	0.978359	0.980359	0.977420	0.977155	0.980980	0.982280	0.987786
0.969905	0.987003	0.980745	0.978505	0.978957	0.984236	0.982699	0.983453	0.976529	0.979028	0.984935	0.978244	0.985198	0.984776	0.981986
0.975339	0.978781	0.980475	0.983162	0.981780	0.985202	0.982232	0.984219	0.981863	0.977254	0.981173	0.976111	0.986638	0.981758	0.981196
0.983768	0.983951	0.978070	0.975019	0.981536	0.979587	0.985027	0.982492	0.981289	0.978833	0.976740	0.978706	0.977636	0.985099	0.981218
0.975742	0.981506	0.982436	0.982984	0.982607	0.978904	0.980365	0.983181	0.979149	0.983736	0.973076	0.979117	0.979816	0.976246	0.984388
0.982255	0.976049	0.981068	0.978199	0.978445	0.990889	0.983460	0.980681	0.977657	0.983569	0.985626	0.983637	0.984253	0.971544	0.981450
0.982196	0.986674	0.988202	0.979431	0.981909	0.981328	0.976626	0.982178	0.980029	0.972165	0.978919	0.985725	0.981323	0.979644	0.981443
0.984818	0.980039	0.984200	0.983721	0.975417	0.975957	0.982249	0.983355	0.983567	0.985513	0.981014	0.979352	0.978571	0.983923	0.976659
0.977581	0.981312	0.981987	0.983255	0.977410	0.981038	0.984405	0.979563	0.984735	0.991885	0.976000	0.987645	0.977601	0.979945	0.976909
0.979777	0.983778	0.976542	0.974697	0.981027	0.980087	0.980026	0.979946	0.977391	0.978358	0.982563	0.982778	0.982852	0.987559	0.981805
0.978678	0.978727	0.977338	0.979410	0.980089	0.979476	0.981472	0.978311	0.988065	0.982153	0.983241	0.990908	0.983285	0.983923	0.980591
0.981867	0.975947	0.981479	0.983769	0.966071	0.978020	0.992208	0.980063	0.973492	0.981141	0.982863	0.987952	0.979947	0.982429	0.979694
0.976706	0.984227	0.987726	0.979301	0.959488	0.984673	0.983471	0.986746	0.980342	0.975945	0.979999	0.977876	0.981225	0.981775	0.981256
0.978317	0.982321	0.982546	0.980906	0.968607	0.982963	0.980833	0.984840	0.990917	0.980514	0.972981	0.978416	0.978719	0.977894	0.977635
0.986292	0.980555	0.979597	0.979950	0.979023	0.972051	0.980127	0.983461	0.986384	0.982590	0.979032	0.979789	0.982335	0.982393	0.979140
0.985210	0.977461	0.993077	0.977133	0.976726	0.981182	0.977685	0.976698	0.982512	0.980041	0.977658	0.979896	0.987086	0.979629	0.980335
0.986684	0.979730	0.989503	0.982044	0.976532	0.973935	0.977588	0.977773	0.979769	0.987126	0.977470	0.979243	0.987390	0.977227	0.976492
0.980631	0.977827	0.978134	0.979746	0.978993	0.981092	0.978357	0.974995	0.978439	0.986883	0.977457	0.961555	0.979950	0.983014	0.976737
0.979097	0.976296	0.977330	0.981339	0.978391	0.964879	0.977639	0.973966	0.978024	0.973399	0.987156	0.983133	0.976337	0.984172	0.978298
0.982109	0.980368	0.980076	0.981080	0.975646	0.980682	0.984396	0.975382	0.976745	0.981837	0.979293	0.982040	0.987277	0.980887	0.974780
0.986664	0.979715	0.981564	0.975165	0.982963	0.977132	0.985277	0.980987	0.982360	0.977148	0.984270	0.985795	0.982637	0.977841	0.982811
0.975719	0.988775	0.977936	0.977593	0.981853	0.976626	0.978948	0.981449	0.980144	0.979205	0.983304	0.981014	0.976733	0.984178	0.988334
0.985579	0.984341	0.978621	0.980539	0.977998	0.976030	0.976503	0.976573	0.976898	0.984118	0.977561	0.976058	0.980247	0.982741	0.979883
0.983376	0.975985	0.987424	0.981416	0.975797	0.976398	0.981572	0.975457	0.985523	0.981873	0.977412	0.978064	0.984721	0.982627	0.980259
0.985114	0.978809	0.978468	0.980229	0.980601	0.973177	0.977236	0.979496	0.983473	0.976967	0.983105	0.986647	0.983500	0.981185	0.981493
0.986107	0.973521	0.975134	0.982253	0.979080	0.980029	0.973485	0.977832	0.976268	0.976980	0.977863	0.980446	0.982772	0.982048	0.981956
0.979630	0.982481	0.979401	0.976918	0.977910	0.975445	0.983244	0.981997	0.978923	0.975660	0.979121	0.980558	0.980164	0.978812	0.983636
0.979515	0.976771	0.975998	0.984095	0.980603	0.981817	0.980301	0.981280	0.980187	0.975911	0.981734	0.988033	0.980059	0.980943	0.994770
0.978886	0.975618	0.981877	0.982714	0.989373	0.980037	0.983879	0.988185	0.987005	0.975563	0.972908	0.976010	0.978689	0.977586	0.975209
0.981015	0.980846	0.978698	0.987410	0.977520	0.976836	0.986416	0.972782	0.988894	0.977283	0.965868	0.982226	0.981446	0.979164	0.984555
0.979105	0.985446	0.985488	0.977336	0.981429	0.979583	0.979325	0.981900	0.985004	0.979828	0.979998	0.979713	0.978064	0.978768	0.979829
0.975521	0.981919	0.986181	0.984516	0.984468	0.976671	0.984678	0.984680	0.984802	0.979834	0.974346	0.979994	0.979053	0.976158	0.978746
0.979768	0.984335	0.982293	0.979023	0.987845	0.974234	0.976645	0.979654	0.985469	0.974812	0.978249	0.979731	0.985392	0.980512	0.976298
0.982857	0.982903	0.982873	0.978839	0.976490	0.976063	0.974634	0.985186	0.973415	0.980303	0.980046	0.976443	0.985076	0.986369	0.985294
0.987807	0.979289	0.979674	0.988469	0.984872	0.975551	0.986348	0.980445	0.976189	0.983000	0.979593	0.977576	0.986086	0.986309	0.991492
0.987122	0.977089	0.987033	0.976921	0.988838	0.981899	0.986619	0.978668	0.984599	0.979987	0.982570	0.975731	0.984509	0.986083	0.983211
0.985474	0.982925	0.985609	0.981650	0.975725	0.981547	0.981696	0.984679	0.983822	0.982493	0.978183	0.985142	0.978028	0.984789	0.985196
0.980001	0.980047	0.988615	0.980868	0.981989	0.986259	0.978504	0.980147	0.983683	0.976733	0.977059	0.980112	0.984306	0.983766	0.985404
0.981147	0.985250	0.983858	0.983537	0.982357	0.983391	0.975906	0.977845	0.978941	0.976479	0.979064	0.983153	0.988118	0.982629	0.978003
0.978083	0.984296	0.982456	0.980974	0.986398	0.986782	0.980668	0.987439	0.974230	0.981088	0.973050	0.979977	0.975265	0.980164	0.980718
0.978354	0.982413	0.980701	0.979502	0.980542	0.985427	0.979091	0.976910	0.986783	0.980831	0.978909	0.977530	0.980287	0.983447	0.988387
0.976116	0.985661	0.977572	0.980666	0.977748	0.979938	0.981937	0.982194	0.981307	0.977896	0.980485	0.980329	0.976717	0.981183	0.983687
0.973329	0.982442	0.982061	0.979522	0.977938	0.991706	0.977858	0.981885	0.979807	0.980882	0.979994	0.981095	0.985974	0.981678	0.985671
0.980829	0.978448	0.979467	0.989365	0.982998	0.985708	0.980796	0.980266	0.982925	0.972989	0.980288	0.982791	0.982698	0.985757	0.980006
0.982045	0.980682	0.981083	0.978126	0.984976	0.980724	0.978228	0.981356	0.983411	0.982690	0.982261	0.977176	0.981908	0.982111	0.984510
0.976445	0.980993	0.989154	0.976701	0.978360	0.983897	0.983228	0.976557	0.983988	0.977738	0.980175	0.979436	0.984255	0.977565	0.980438
0.974697	0.988806	0.978898	0.979273	0.987503	0.980122	0.981139	0.981639	0.980743	0.984448	0.974397	0.984237	0.976070	0.977999	0.978738
0.978668	0.980018	0.980355	0.989121	0.985824	0.981286	0.984155	0.971849	0.979234	0.981922	0.977962	0.980996	0.984860	0.992353	0.984832
0.978704	0.981400	0.976974	0.979697	0.975788	0.979070	0.978998	0.979058	0.985230	0.984321	0.977140	0.978745	0.981642	0.974007	0.975272
0.987264	0.978041	0.978033	0.977353	0.974883	0.976319	0.983387	0.975553	0.981341	0.984195	0.980893	0.984850	0.974317	0.981197	0.976749
0.984219	0.980168	0.979667	0.983568	0.984445	0.981976	0.983209	0.976881	0.981186	0.980092	0.982983	0.986983	0.978651	0.975312	0.979185
0.980762	0.982328	0.983286	0.982369	0.976500	0.978213	0.981994	0.983631	0.981897	0.978945	0.987506	0.987879	0.977462	0.981830	0.977821
0.982593	0.983892	0.984735	0.978250	0.981955	0.978150	0.986523	0.981761	0.982404	0.978694	0.984950	0.986864	0.984835	0.978619	0.976211
0.982522	0.976173	0.979057	0.976060	0.970549	0.980565	0.977982	0.980966	0.972695	0.979499	0.977152	0.982863	0.982613	0.980516	0.985303
0.979721	0.976922	0.973840	0.979385	0.980963	0.980286	0.986066	0.986241	0.982401	0.987313	0.979480	0.986101	0.976847	0.985904	0.980724
0.980893	0.983697	0.982932	0.977284	0.983036	0.975405	0.983732	0.975409	0.982407	0.986033	0.981177	0.980817	0.983564	0.987009	0.987157
0.982340	0.979292	0.988624	0.977677	0.978006	0.984724	0.973530	0.987339	0.980943	0.977466	0.983680	0.980455	0.976958	0.981263	0.987104
0.982853	0.983493	0.986310	0.984059	0.975389	0.977911	0.985507	0.981714	0.990561	0.976437	0.985464	0.975442	0.980625	0.980526	0.986262
0.974509	0.981308	0.979512	0.975724	0.966435	0.982588	0.976459	0.979923	0.982343	0.977993	0.977875	0.974499	0.979634	0.978447	0.983674
0.987473	0.984208	0.974544	0.983472	0.975804	0.978209	0.976504	0.976291	0.977674	0.983867	0.983861	0.974904	0.985720	0.973850	0.979952
0.978961	0.979527	0.985313	0.981783	0.980291	0.979812	0.981726	0.981041	0.975301	0.980647	0.988654	0.978149	0.982416	0.984768	0.978351
0.979517	0.984435	0.986483	0.984857	0.980486	0.976946	0.980874	0.977542	0.988079	0.981161	0.983513	0.984959	0.981331	0.979991	0.982774
0.982476	0.991637	0.978856	0.983523	0.976766	0.983791	0.979925	0.974211	0.986667	0.977507	0.980625	0.981484	0.976927	0.975529	0.975851
0.990159	0.983562	0.985104	0.980102	0.981187	0.984829	0.982992	0.979726	0.978883	0.986053	0.983974	0.983752	0.977231	0.978217	0.978847
0.975988	0.984026	0.983786	0.980375	0.981911	0.980825	0.983910	0.979698	0.978391	0.985854	0.981053	0.979459	0.982646	0.977976	0.979229
0.986799	0.986419	0.978159	0.984566	0.980522	0.989473	0.976666	0.977351	0.977391	0.985971	0.981516	0.986639	0.976961	0.979515	0.981813
0.977092	0.985221	0.975183	0.987427	0.975922	0.986008	0.979539	0.982822	0.988914	0.984860	0.979420	0.982854	0.979193	0.984708	0.977544
0.977715	0.976929	0.990863	0.975503	0.979169	0.981960	0.976710	0.981304	0.985792	0.987025	0.978319	0.980756	0.979915	0.980929	0.979755
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

CAPÍTULO 4

Aplicación y Resultados

En este capítulo se presentan los resultados obtenidos de la aplicación del Test ANOVA, así como los del Test de Kruskal-Wallis; ambos aplicados a las muestras provenientes de los dos índices. Se comprobará efectivamente que para estas muestras, ambos test dan el mismo resultado. Como instancia final se realizará el proceso de comparaciones múltiples que permitirá identificar qué pares de muestras específicos a nivel de ubicación difieren entre sí, así como los resultados de la aplicación de test no paramétricos para el problema de escala.

4.1. Análisis Descriptivo

En esta sección se exhibe un resumen de las principales características que presentan las muestras utilizadas en el proceso de inferencia. Se enfatiza en las medidas de tendencia central, medidas de variabilidad y medidas de forma. Los valores de estas medidas se pueden apreciar, dependiendo del índice, en las siguientes tablas:

Tabla 4.1: Resumen estadístico de las muestras (Índice ACP).

	Nro. Observaciones	Media	Desviación Estándar	Mediana	Min	Max	Rango	Sesgo	kurtosis
Muestra 1	7415	-0.1795	1.5658	-0.1654	-7.9336	5.8546	13.7882	-0.2962	0.7856
Muestra 2	2296	0.5059	1.5912	0.6354	-6.4894	6.9228	13.4122	-0.3849	0.2465
Muestra 3	2145	0.0789	1.6154	0.1357	-10.8290	5.5268	16.3559	-0.5202	1.5833

Tabla 4.2: Resumen estadístico de las muestras (Índice CI).

	Nro. Observaciones	Media	Desviación Estándar	Mediana	Min	Max	Rango	Sesgo	kurtosis
Muestra 1	7415	0.9810	0.0042	0.9808	0.9494	1.0000	0.0506	-0.1894	2.7475
Muestra 2	2296	0.9803	0.0046	0.9798	0.9540	1.0000	0.0460	0.6206	2.0753
Muestra 3	2145	0.9804	0.0041	0.9801	0.9582	1.0000	0.0418	0.6050	2.9224

A continuación se presenta para cada muestra el gráfico de la distribución de frecuencias, el ajuste por parte de su función de densidad y la distribución normal que seguiría si efectivamente la muestra fuera extraída de una población normal.

Para las muestras provenientes del índice ACP.

Figura 4.1: Histograma y Función de Densidad para la primera muestra (Índice ACP).

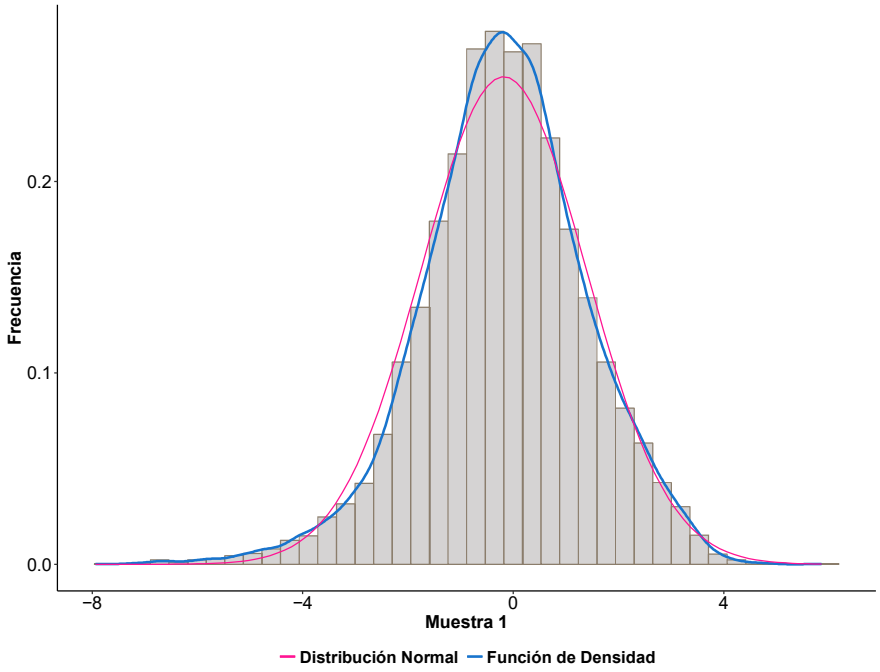


Figura 4.2: Histograma y Función de Densidad para la segunda muestra (Índice ACP).

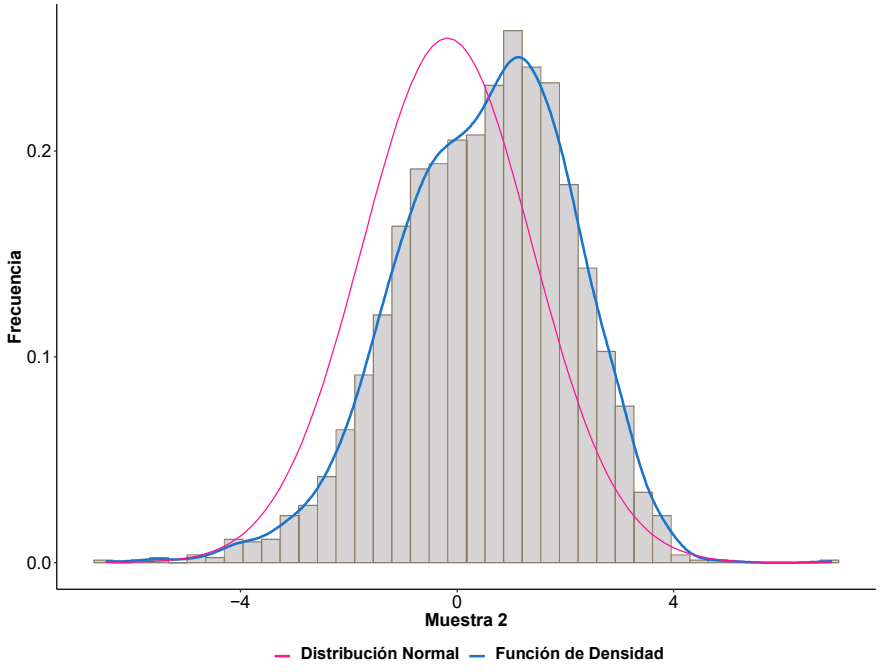
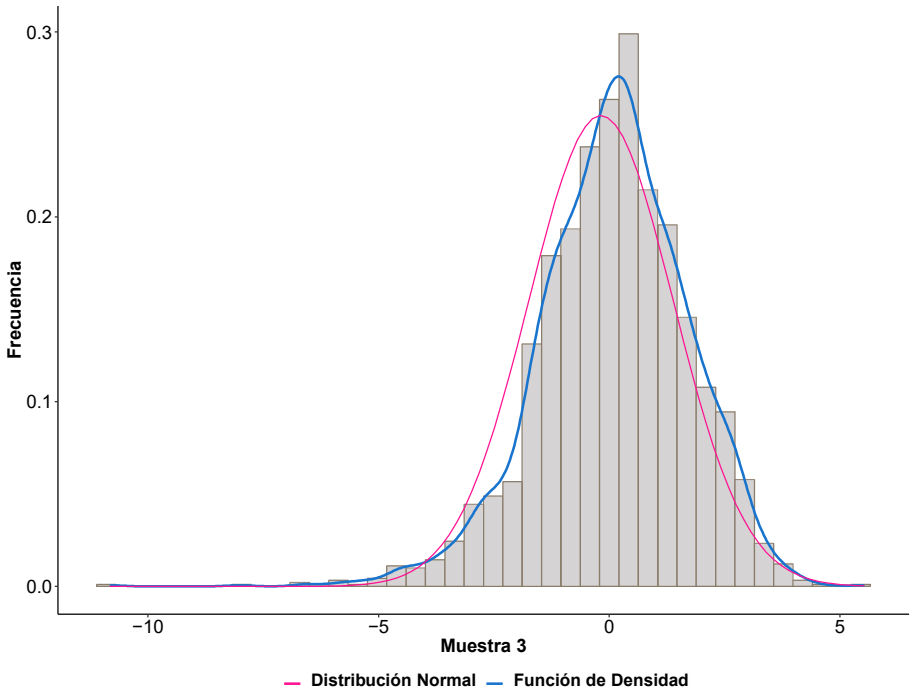


Figura 4.3: Histograma y Función de Densidad para la tercera muestra (Índice ACP).



Para las muestras provenientes del índice CI.

Figura 4.4: Histograma y Función de Densidad para la primera muestra (Índice CI).

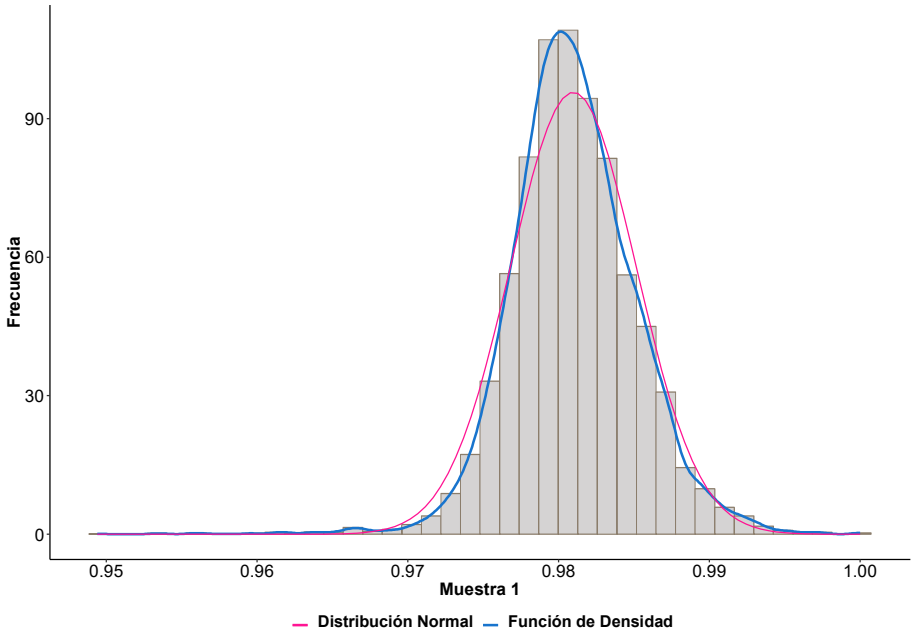


Figura 4.5: Histograma y Función de Densidad para la segunda muestra (Índice CI).

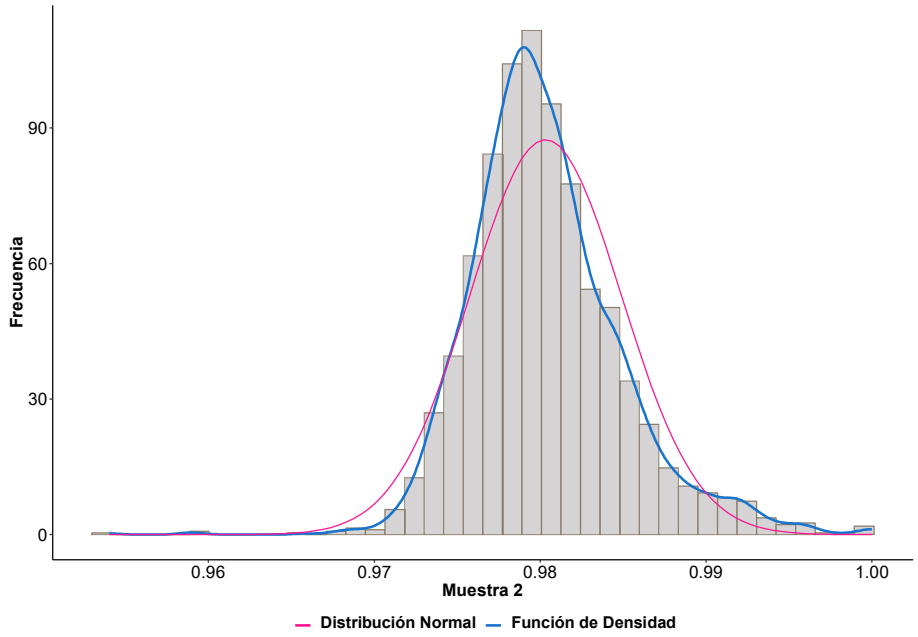
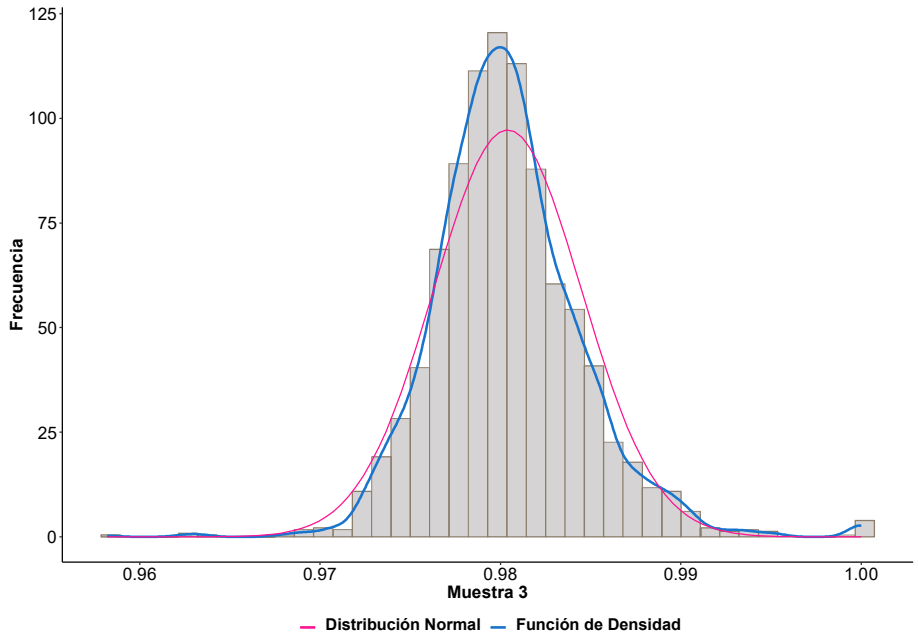


Figura 4.6: Histograma y Función de Densidad para la tercera muestra (Índice CI).



La pregunta que inmediatamente surge es de si las muestras cumplen con los requisitos que exigen las pruebas paramétricas. De manera particular en la aplicación de Análisis de Varianza de un solo factor (ANOVA).

4.2. Test Anova

Para la correcta utilización del Test ANOVA es necesario verificar el cumplimiento de ciertos supuestos fundamentales. Estos supuestos establecen que las muestras hayan sido extraídas de poblaciones independientes que pueden describirse como con una distribución normal, que las varianzas de las poblaciones sean iguales y que las observaciones sean v.a.

El supuesto de independencia es crítico, pero como menciona Montgomery (2004), basta que el orden de las observaciones sea aleatorio para satisfacer este supuesto. Ahora bien, para verificar el cumplimiento de los supuestos de normalidad e igualdad de varianzas se utiliza el siguiente procedimiento:

Supuesto de normalidad.

Para determinar si los datos muestrales se ajustan a una distribución normal, en primera instancia podemos basarnos de un examen visual a través de la gráfica de probabilidad normal. Esta gráfica para cada una de las muestras se presenta a continuación:

Para las muestras provenientes del índice ACP.

Figura 4.7: Gráfica de Probabilidad Normal-Muestra 1 (Índice ACP).

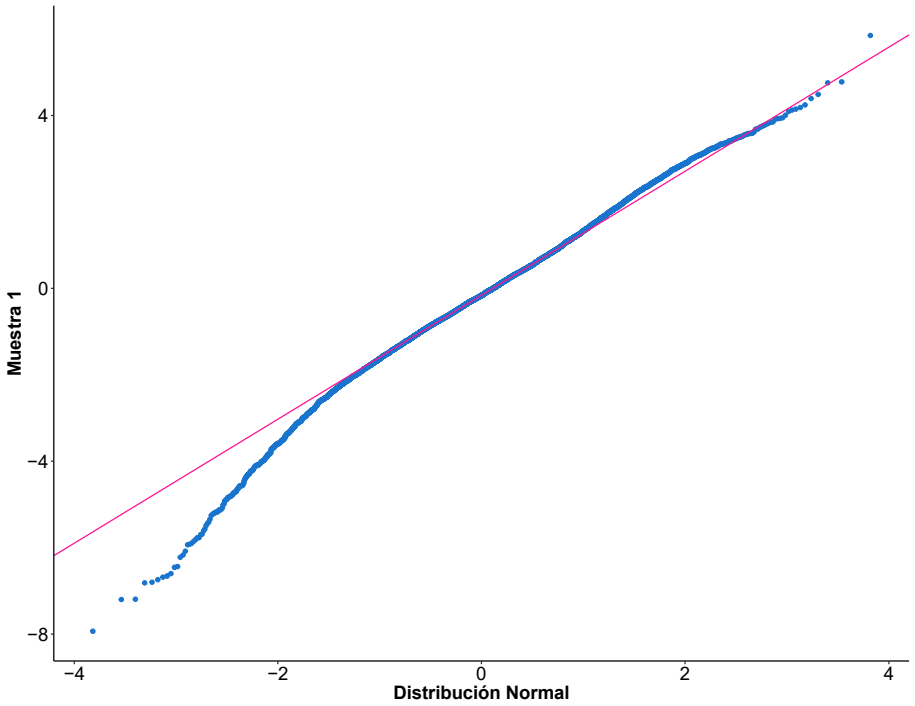


Figura 4.8: Gráfica de Probabilidad Normal-Muestra 2 (Índice ACP).

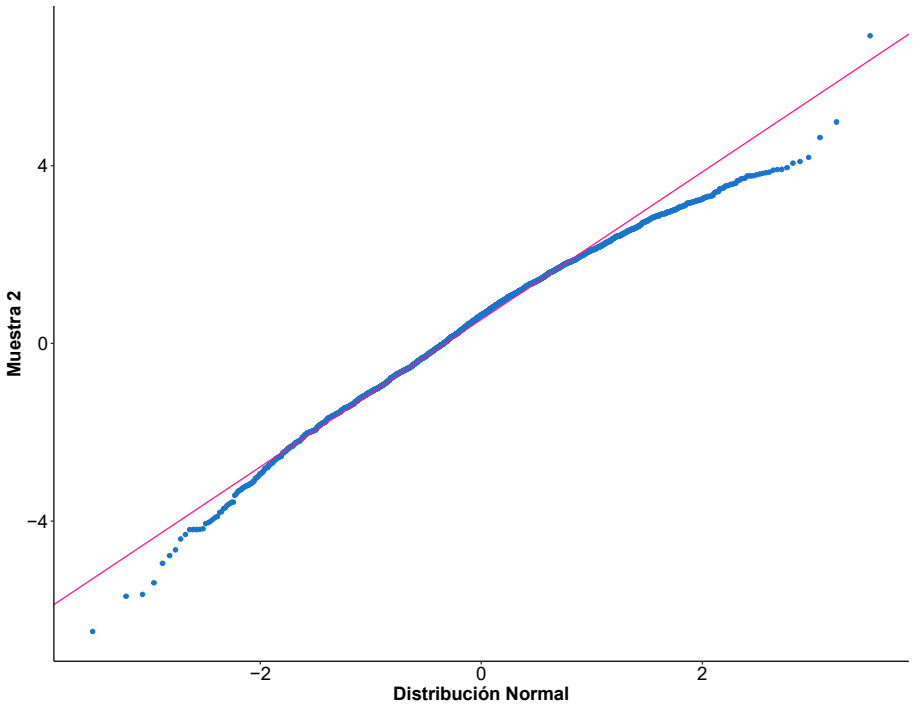
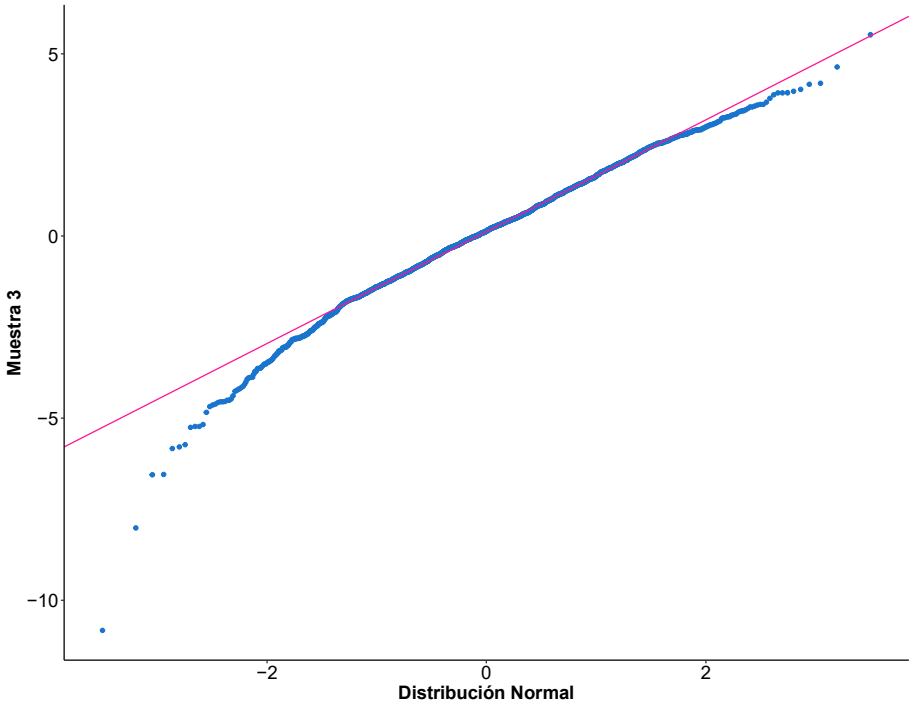


Figura 4.9: Gráfica de Probabilidad Normal-Muestra 3 (Índice ACP).



Para las muestras provenientes del índice CI.

Figura 4.10: Gráfica de Probabilidad Normal-Muestra 1 (Índice CI).

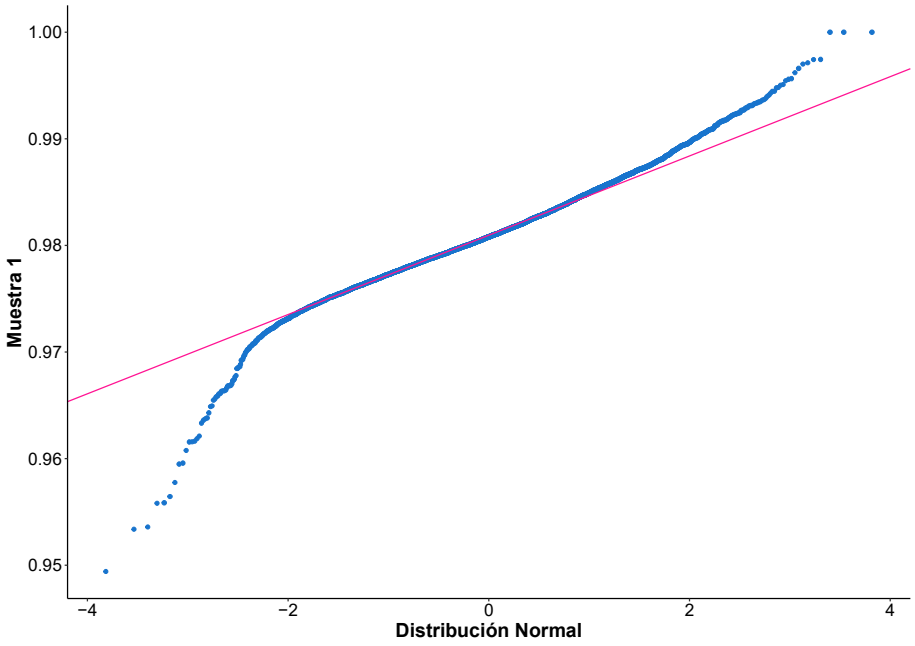


Figura 4.11: Gráfica de Probabilidad Normal-Muestra 2 (Índice CI).

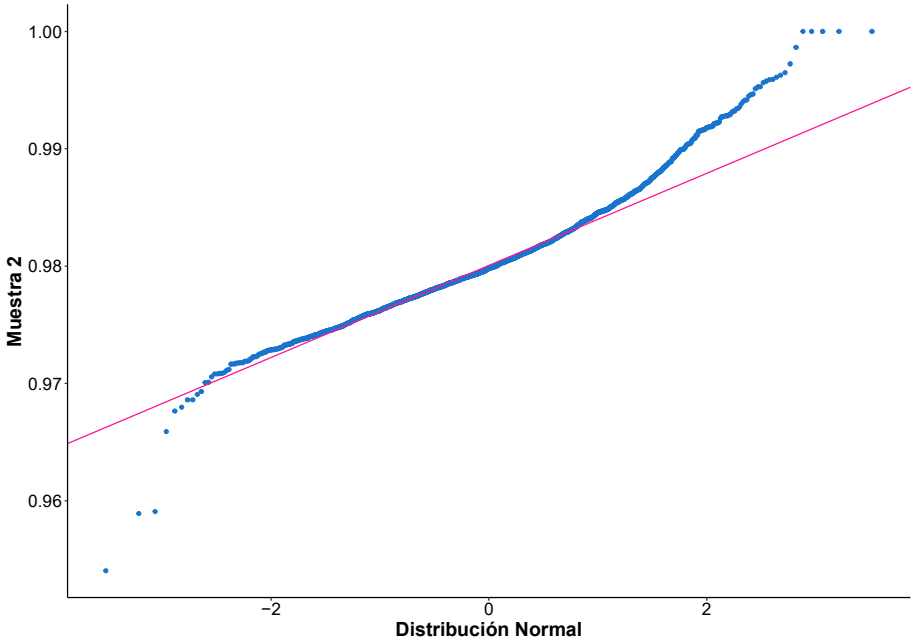
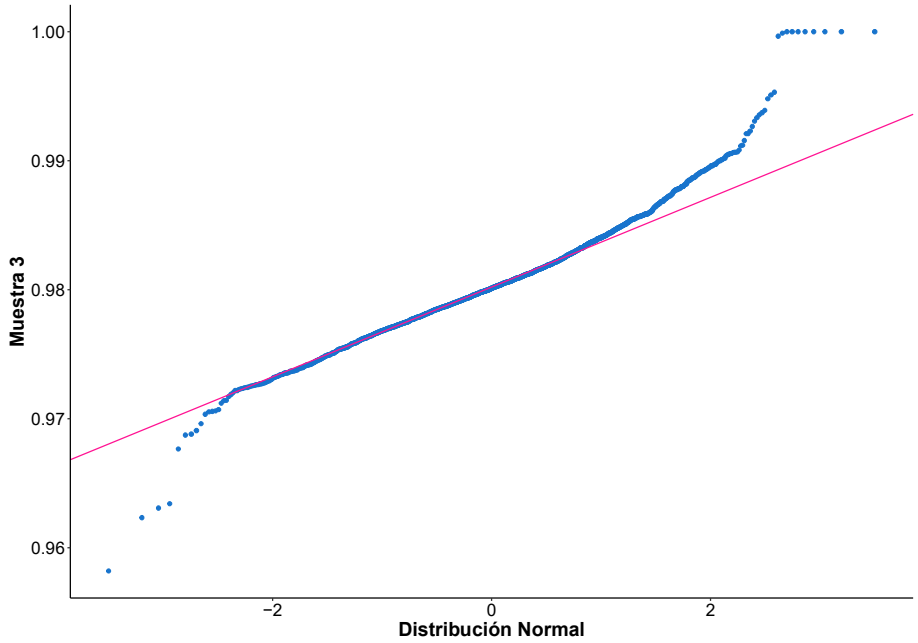


Figura 4.12: Gráfica de Probabilidad Normal-Muestra 3 (Índice CI).



Si la distribución hipotética de normalidad describe de manera adecuada a los datos, los puntos graficados estarán aproximadamente sobre la línea recta; si los puntos graficados muestran una desviación significativa de la recta, el modelo hipotético no es apropiado (Montgomery 2004). Así, de la Figura 4.7, Figura 4.8, Figura 4.9, Figura 4.10, Figura 4.11 y Figura 4.12 se sospecha que el supuesto de normalidad de los datos no es razonable.

Por otro lado, el resultado anterior se puede ratificar mediante el empleo de test que contrasten la normalidad de los datos. Para este fin se emplean 4 test y sus resultados hacen que se rechace el supuesto que las muestras provienen de una población normalmente distribuida. A modo de ejemplo, en las Tablas 4.3, 4.4 y 4.5 se presentan los resultados de aplicar el Test Kolmogorov-Smirnov a las muestras del índice ACP y en las Tablas 4.6, 4.7 y 4.8 a las muestras del índice CI.

El resultado de la aplicación de los otros tres test se puede apreciar en (B.1) de la parte de anexos.

Para las muestras provenientes del índice ACP.

Tabla 4.3: Kolmogorov-Smirnov Test para la Muestra 1 (Índice ACP).

	D
Test statistic	0.02470489
p value	0.0002344967
Alternative hypothesis	two-sided

One-sample Kolmogorov-Smirnov Test: Muestra 1

Tabla 4.4: Kolmogorov-Smirnov Test para la Muestra 2 (Índice ACP).

	D
Test statistic	0.03751539
p value	0.003120842
Alternative hypothesis	two-sided

One-sample Kolmogorov-Smirnov Test: Muestra 2

Tabla 4.5: Kolmogorov-Smirnov Test para la Muestra 3 (Índice ACP).

	D
Test statistic	0.03285506
p value	0.01949268
Alternative hypothesis	two-sided

One-sample Kolmogorov-Smirnov Test: Muestra 3

Para las muestras provenientes del índice CI.

Tabla 4.6: Kolmogorov-Smirnov Test para la Muestra 1 (Índice CI).

	D
Test statistic	0.03080939
p value	1.539941e-06
Alternative hypothesis	two-sided

One-sample Kolmogorov-Smirnov Test: Muestra 1

Tabla 4.7: Kolmogorov-Smirnov Test para la Muestra 2 (Índice CI).

	D
Test statistic	0.06503447
p value	7.349484e-09
Alternative hypothesis	two-sided

One-sample Kolmogorov-Smirnov Test: Muestra 2

Tabla 4.8: Kolmogorov-Smirnov Test para la Muestra 3 (Índice CI).

	D
Test statistic	0.05773625
p value	1.231279e-06
Alternative hypothesis	two-sided

One-sample Kolmogorov-Smirnov Test: Muestra 3

Supuesto de Homocedasticidad.

Siempre es buena idea representar gráficamente un conjunto de datos y de acuerdo al fin que exista, la presentación visual puede describir varias características importantes al mismo tiempo, tales como la dispersión y simetría. En nuestro caso, la representación gráfica de una serie de datos numéricos a través de sus cuartiles es necesaria. Esto motiva la utilización del Diagrama de Caja y Bigote.

La visualización de este diagrama, para el índice ACP se exhibe en la Figura 4.13 y para el índice CI en la Figura 4.14.

Figura 4.13: Diagrama de Caja y Bigote paras las muestras provenientes del Índice ACP.

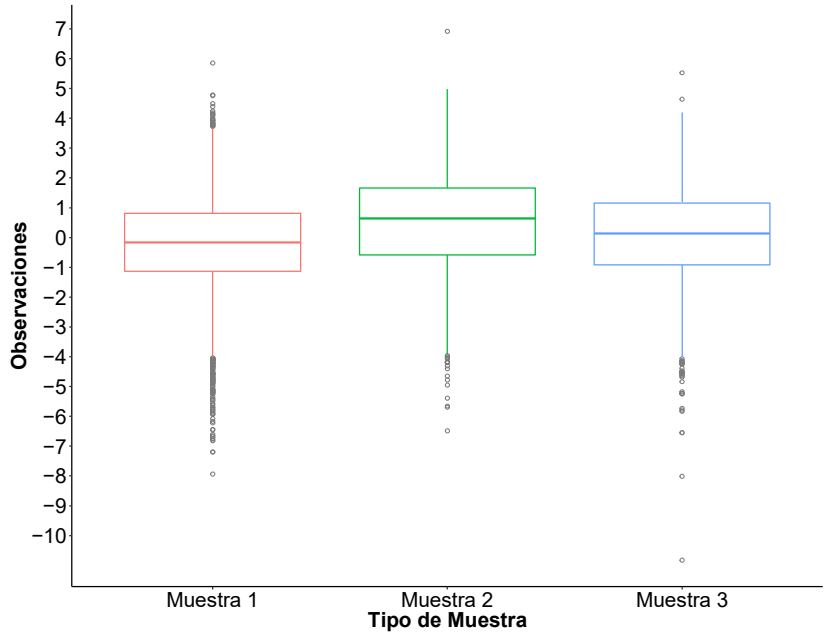
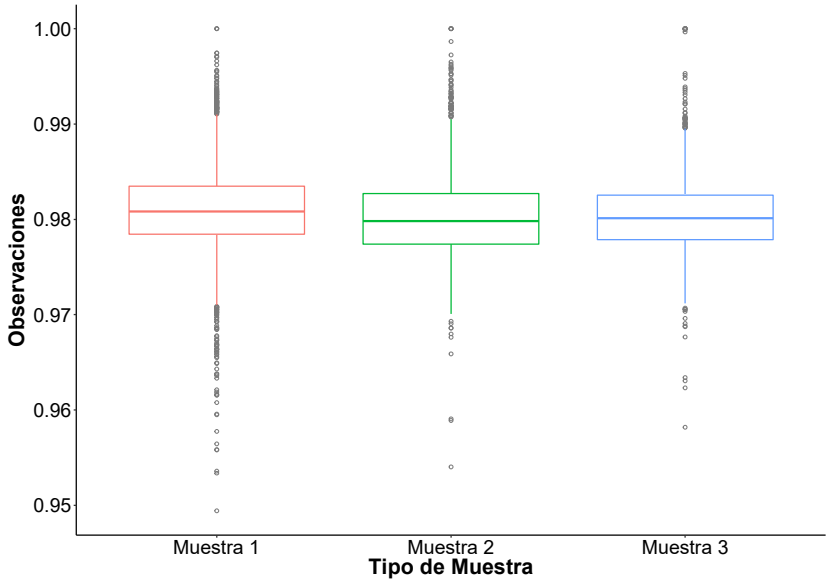


Figura 4.14: Diagrama de Caja y Bigote paras las muestras provenientes del Índice CI.



Tanto en la Figura 4.13 como en la Figura 4.14, se observa que cada muestra produce distribuciones razonablemente simétricas con una variabilidad similar. Pero también se puede apreciar una alta presencia de datos atípicos. Esto conlleva a verificar si las varianzas poblacionales de las que se extraen las diferentes muestras son iguales o no, mediante la utilización del Test de Levene. Este test prueba la hipótesis nula de que las varianzas de las poblaciones son iguales.

El resultado de aplicar el test anterior, para las muestras provenientes del índice ACP es

Tabla 4.9: Levene's Test for Homogeneity of Variance (índice ACP)

	W
Df	2
F value	4.5132
Pr(>F)	0.01098

y para las muestras provenientes del índice CI es

Tabla 4.10: Levene's Test for Homogeneity of Variance (índice CI)

	1
Df	2
F value	10.335
Pr(>F)	0.00003

Dado que el p-valor para ambos casos es menor a 0.05 existe una diferencia estadísticamente significativa entre las desviaciones estándar de las poblaciones a un nivel de confianza del 95.0 %. En otras palabras, se rechaza la hipótesis nula de igualdad de varianzas.

En conclusión, en vista que se rechaza tanto el supuesto de normalidad como el de igualdad de varianzas, si se aplicase el análisis de varianza a estas muestras, el resultado sería **poco confiable**. A pesar de ello se realizará el ANOVA con el fin de comparar los resultados obtenidos por éste, con los resultados dados por el test no paramétrico de Kruskal-Wallis.

El Análisis de Varianza de un factor o también conocido como Análisis de Varianza unidireccional (ANOVA), bajo el cumplimiento de los supuestos mencionados al comienzo de esta sección nos permite comparar tres o más medias poblacionales, para determinar si éstas pueden ser iguales ($H_0 : \mu_1 = \mu_2 = \ldots = \mu_n$).

Para Triola (2009):

El término unidireccional se utiliza porque los datos de la muestra se separan en grupos según alguna característica. En lugar de referirse al objetivo principal de probar la igualdad de medias, el término análisis de varianza se refiere al método que utilizamos, que se basa en un análisis de las varianzas de las muestras. (p. 628)

Así pues, el método ANOVA hace uso de la prueba F, llamada así en honor a Ronald Fisher, uno de los pioneros de la estadística actual. Con ella se compara

varias medias poblacionales en forma simultánea [25].

La aplicación del Análisis de Varianza a la muestra 1, muestra 2 y muestra 3 provenientes del índice ACP se presenta en la Tabla 4.11.

Tabla 4.11: Análisis de Varianza de un factor (Índice ACP).

Fuente	Grados de Libertad	Suma de Cuadrados	Cuadrado Medio	valor F	Pr(>F)
Entre Grupos	2	839.7463	419.8731	168.2285	9.057803e-73
Error(dentro de los grupos)	11853	29 583.3130	2.4959		

y para las muestras provenientes del índice CI en la Tabla 4.12.

Tabla 4.12: Análisis de Varianza de un factor (Índice CI).

Fuente	Grados de Libertad	Suma de Cuadrados	Cuadrado Medio	valor F	Pr(>F)
Entre Grupos	2	0.001064	0.000532	29.644481	1.44e-13
Error(dentro de los grupos)	11853	0.212666	0.000018		

De la Tabla 4.11 y Tabla 4.12 se concluye que existe diferencia estadísticamente significativa entre las medias de las tres muestras (tanto para las muestras provenientes del índice ACP como las muestras provenientes del índice CI) a un nivel de confianza del 95.0 %. Dicho de otra manera, se rechaza la hipótesis nula de igualdad de medias.

Si se desease saber que medias son significativamente diferentes entre si, se debe realizar el contraste post hoc. Este proceso básicamente compara dos a dos las medias de todos los tratamientos presentes en el ANOVA mediante algún test que compare dos medias poblaciones de manera simultánea. La prueba habitual es el conocido t-Test. Sin embargo, cuando se haga el proceso de comparaciones múltiples en lo no paramétrico se utilizará (2.119) y el Test U.

Los resultados obtenidos del proceso post hoc que utiliza el Test t junto con los métodos *Bonferroni y Holm* mencionados en R. Joaquin (2016) [26], para las muestras provenientes del índice ACP son los siguientes:

[25] Para más detalle del Test ANOVA vea Triola (2009), Montgomery (2004), Lind, Marchal y Wathen (2012)

[26] Más detalle de los métodos de contraste post hoc para el Test ANOVA se presentan en R. Joaquin (2016).

Tabla 4.13: Comparaciones múltiples dadas por el proceso post hoc para el Test ANOVA.

Método: Bonferroni

	Muestra 1	Muestra 2
Muestra 2	2.89315873024235e-72	
Muestra 3	7.97069215263998e-11	7.7952306273836e-19

Método: Holm

	Muestra 1	Muestra 2
Muestra 2	2.89315873024235e-72	
Muestra 3	2.65689738421333e-11	5.19682041825573e-19

y para las muestras proveniente del índice CI es

Tabla 4.14: Comparaciones múltiples dadas por el proceso post hoc para el Test ANOVA.

Método: Bonferroni

	Muestra 1	Muestra 2
Muestra 2	4.04725114416336e-10	
Muestra 3	6.33040952987612e-08	1

Método: Holm

	Muestra 1	Muestra 2
Muestra 2	4.04725114416336e-10	
Muestra 3	4.22027301991741e-08	0.59303126911544

De esta manera, la Tabla 4.13 nos dice que los tres pares de muestras son significativamente diferentes entre sí. Sin embargo, el resultado de la Tabla 4.14 para las muestras 2 y 3 nos dice que ambas muestras provienen de la misma población a nivel de ubicación, tanto con el método de *Bonferroni* como *Holm*.

No obstante, dado que no se cumplen los supuestos fundamentales de los modelos paramétricos bajo normalidad, no es recomendable la utilización del Test Anova para probar la hipótesis nula de poblaciones idénticas a nivel de ubicación. Lo propio a realizar en este punto es ver qué sucede desde el punto de vista no paramétrico.

4.3. Test de Kruskal-Wallis

Como se pudo constatar, la correcta utilización del test paramétrico ANOVA requiere verificar el cumplimiento de ciertos supuestos fundamentales para garantizar que el resultado sea el adecuado. Pero dar seguimiento a cada uno de estos supuestos no suele ser una tarea sencilla.

Desde este punto de vista, el Test Anova unidireccional de Kruskal-Wallis nos da una salida rápida cuando no se está del todo seguro de si nuestros datos satisfacen los requisitos que exige la teoría paramétrica, con la ventaja que solo es necesario que las muestras hayan sido extraídas de poblaciones que tengan su cdf continua.

Si aplicamos el Test de Kruskal-Wallis, a las muestras provenientes del índice ACP, el resultado es

Tabla 4.15: Test de Kruskal-Wallis (Índice ACP).

	Kruskal-Wallis chi-squared
Statistic	351.3006
Df	2
p.value	5.200481e-77

Kruskal-Wallis rank sum Test

y para las muestras provenientes del índice CI es

Tabla 4.16: Test de Kruskal-Wallis (Índice CI).

	Kruskal-Wallis chi-squared
Statistic	115.4536
Df	2
p.value	8.502987e-26

Kruskal-Wallis rank sum Test

En consecuencia, los p-valores de la Tabla 4.15 y Tabla 4.16 conducen al rechazo de la hipótesis nula que las muestras provienen de poblaciones idénticas.

Por otro lado, si se desea saber que poblaciones (dos a dos) son las que difieren entre sí, es necesario realizar el proceso de comparaciones múltiples. Los resultados son los siguientes:

Usando el comando $kruskal(data, tratamiento, group = FALSE, alpha = 0.2)$ del paquete estadístico *R*, para los métodos *Bonferroni y Holm*, se tienen las siguientes tablas:

Para las muestras provenientes del índice ACP.

Tabla 4.17: Resumen estadísticos de las muestras dadas por el proceso post hoc (Índice ACP).

	datos	Rango Promedio	std	n	Min	Max	Q25	Q50	Q75
Muestra 1	-0.18	5525.85	1.57	7415	-7.93	5.85	-1.13	-0.17	0.81
Muestra 2	0.51	7037.36	1.59	2296	-6.49	6.92	-0.58	0.64	1.66
Muestra 3	0.08	6133.49	1.62	2145	-10.83	5.53	-0.91	0.14	1.16

Tabla 4.18: Comparaciones múltiples dadas por el proceso post hoc para el Test de Kruskal-Wallis (Índice ACP).

Método: Bonferroni

	Difference	pvalue	Signif.
Muestra 1 - Muestra 2	-1511.51	0.00	***
Muestra 1 - Muestra 3	-607.64	0.00	***
Muestra 2 - Muestra 3	903.87	0.00	***

Método: Holm

	Difference	pvalue	Signif.
Muestra 1 - Muestra 2	-1511.51	0.00	***
Muestra 1 - Muestra 3	-607.64	0.00	***
Muestra 2 - Muestra 3	903.87	0.00	***

Para las muestras provenientes del índice CI.

Tabla 4.19: Resumen estadísticos de las muestras dadas por el proceso post hoc (Índice CI).

	datos	Rango Promedio	std	n	Min	Max	Q25	Q50	Q75
Muestra 1	0.98098	6187.14963	0.00417	7415	0.94941	1.00000	0.97845	0.98081	0.98347
Muestra 2	0.98033	5419.63567	0.00456	2296	0.95403	1.00000	0.97741	0.97982	0.98271
Muestra 3	0.98039	5579.06667	0.00410	2145	0.95819	1.00000	0.97786	0.98011	0.98255

Tabla 4.20: Comparaciones múltiples dadas por el proceso post hoc para el Test de Kruskal-Wallis (Índice CI).

Método: Bonferroni

	Difference	pvalue	Signif.
Muestra 1 - Muestra 2	767.513958	0.000000	***
Muestra 1 - Muestra 3	608.082962	0.000000	***
Muestra 2 - Muestra 3	-159.430996	0.357300	

Método: Holm

	Difference	pvalue	Signif.
Muestra 1 - Muestra 2	767.513958	0.000000	***
Muestra 1 - Muestra 3	608.082962	0.000000	***
Muestra 2 - Muestra 3	-159.430996	0.119100	

Usando el Test de Mann-Whitney para los métodos de *Bonferroni y Holm* [27], tenemos las siguientes tablas:

Para las muestras provenientes del índice ACP.

[27] Más detalle de los métodos de contraste post hoc para el Test de Kruskal-Wallis se presentan en R. A. Joaquin (2016).

Tabla 4.21: Contraste post hoc dados por el Test U (Índice ACP).

Método: Bonferroni		
	Muestra.1	Muestra.2
Muestra 2	2.73506072995587e-75	
Muestra 3	7.71886390220646e-13	7.53849724582797e-19
Método: Holm		
	Muestra.1	Muestra.2
Muestra 2	2.73506072995587e-75	
Muestra 3	2.57295463406882e-13	5.02566483055198e-19

Para las muestras provenientes del índice CI.

Tabla 4.22: Contraste post hoc dados por el Test U (Índice CI).

Método: Bonferroni		
	Muestra.1	Muestra.2
Muestra 2	4.99861750737219e-20	
Muestra 3	5.58032804103125e-13	0.19461868831691
Método: Holm		
	Muestra.1	Muestra.2
Muestra 2	4.99861750737219e-20	
Muestra 3	3.72021869402083e-13	0.0648728961056367

Por otro lado, usando (2.119) junto con la Tabla 4.17, Tabla 4.19 y $\alpha = 0.2$, se tiene que:

Para las muestras provenientes del índice ACP.

$$Z_{12} = 18.49$$
$$Z_{13} = 7.2413 \qquad y \qquad Z_{\alpha/k(k-1)} = 1.833915.$$
$$Z_{23} = 8.7942$$

En este caso, dado que Z_{ij} para todo $i, j = 1,2,3$ con $i \neq j$ es mayor a 1.833915 y debidos a los valores exhibidos en las Tablas 4.18 y 4.21; entonces existen diferencias significativas entre toda par de medianas. Por lo tanto, todo par de muestras conducen al rechazo de la hipótesis nula de poblaciones idénticas.

Para las muestras provenientes del índice CI.

$$Z_{12} = 9.389 \qquad Z_{13} = 7.246 \qquad Z_{23} = 1.551 \qquad y \qquad Z_{\alpha/k(k-1)} = 1.833915.$$

En este caso, dado que Z_{23} no es mayor a 1.833915 y debidos a los valores exhibidos en las Tablas 4.20 y 4.22; no existe diferencia significativa entre las medianas de las muestras 2 y muestra 3. Por lo tanto, solo los pares de muestras 1 y 2 junto con las muestras 1 y 3 conducen al rechazo de la hipótesis nula de poblaciones idénticas.

En conclusión, Kruskal-Wallis aparentemente nos proporciona los mismos resultados que los obtenidos por el procedimiento de Análisis de Varianza Anova, pero con la ventaja de que solo es necesario suponer que las muestras hayan sido extraídas de poblaciones que tengan cdf continua. Sin embargo, estos resultados no son comparables por la misma razón de que los supuesto del Test Anova no se cumplen, y por ende sus resultados no son confiables.

No obstante, una alternativa en la cual los resultados del Test Kruskal-Wallis son comparables con el Test Anova, es aplicar este último a los datos transformados por sus rangos. Los resultados son los siguientes:

Para los rangos de las muestras provenientes del índice ACP.

Tabla 4.23: Análisis de Varianza de un factor aplicado a los rangos de las muestras provenientes del Índice ACP.

Fuente	Grados de Libertad	Suma de Cuadrados	Cuadrado Medio	valor F	Pr(>F)
Entre Grupos	2	4115386785.29	2057693392.65	180.98	3.767078e-78
Error(dentro de los grupos)	11853	134762571394.71	11369490.54		

Para los rangos de las muestras provenientes del Índice CI.

Tabla 4.24: Análisis de Varianza de un factor aplicado a los rangos de las muestras provenientes del índice CI

Fuente	Grados de Libertad	Suma de Cuadrados	Cuadrado Medio	valor F	Pr(>F)
Entre Grupos	2	1352506011.3083	676253005.6542	58.2847	6.470611e-26
Error(dentro de los grupos)	11853	137525451828.6917	11602585.9975		

Bajo esta situación, cuando se aplica el Test Anova tanto a las muestras originales como a sus datos transformados por los rangos; y los resultado son diferentes, se debe dar preferencia a los resultados proporcionados por los rangos, debido a que es menos posible que estos sean distorsionados por una condición de no normalidad o por la presencia de observaciones inusuales (Montgomery 2004).

Así, en base a los resultados de la Tabla 4.23, 4.24 y los proporcionados por las Tablas 4.15 y 4.16 se ratifica el hecho de rechazar la hipótesis nula de poblaciones idénticas.

Para ilustrar otra situación en donde si es posible comparar los resultados de estos dos test, es necesario que las muestras cumplan los requisitos que exige el campo paramétrico. Esta situación se presenta en (B.3) de la parte de anexos.

En esta parte se simulan 3 muestras aleatorias de tamaño 300 que provienen de una población normal con media 0 y varianza 1 (vea las Tablas B.25, B.26 y B.27). Las muestras evidentemente verifican el supuesto que provienen de una población normal (vea las Tablas B.28, B.29 y B.30) y el supuesto que la varianzas de las poblaciones son iguales (vea la Tabla B.31).

Los resultados obtenidos tanto de la aplicación del Test Anova (vea la Tabla B.32 de la parte de anexos) como los del test no paramétrico de Kuskal-Wallis (vea la Tabla B.33 de la parte de anexos) conducen a no rechazar la hipótesis nula de poblaciones idénticas. Y dado que el Test de Kruskal-Wallis asume que la única diferencia entre las poblaciones ocurre a nivel de ubicación, entonces los resultados obtenidos por este, serán los mismos y de hecho comparables con los obtenidos por el Test Anova.

Por consiguiente, este ejemplo muestra la ventaja y facilidad de utilizar el Test Kruskal-Wallis para evitar la tediosa tarea de verificar si las muestras cumplan o no con los requisitos que exigen los modelos paramétricos bajo normalidad.

Adicionalmente, a pesar de que algunas muestras del estudio presentan suficiente evidencia estadística para decir que no provienen de la misma población a nivel de ubicación, resulta interesante ver qué sucede a nivel de escala.

De acuerdo con esto y dependiendo del índice, al observar la Figura 4.15 y Figura 4.16 se podría intuir si las poblaciones tienen o no la misma variabilidad.

No obstante, para dar respuesta a lo anterior de manera adecuada se utilizarán los Test F de Fisher, dentro de lo paramétrico; y, Mood Test, Freund Ansari Bradley Test, Siegel Tukey Test y Klotz Normal Scores Test, en lo no paramétrico.

Los resultados son los siguientes:

Para el caso paramétrico, la prueba habitual para probar la igualdad de varianzas, bajo el supuesto que las dos poblaciones son distribuciones normales con medias desconocidas, es la prueba F de Fisher (prueba F de igualdad de varianzas) [28].

Figura 4.15: Gráfica de Densidad de las muestras provenientes del Índice ACP.

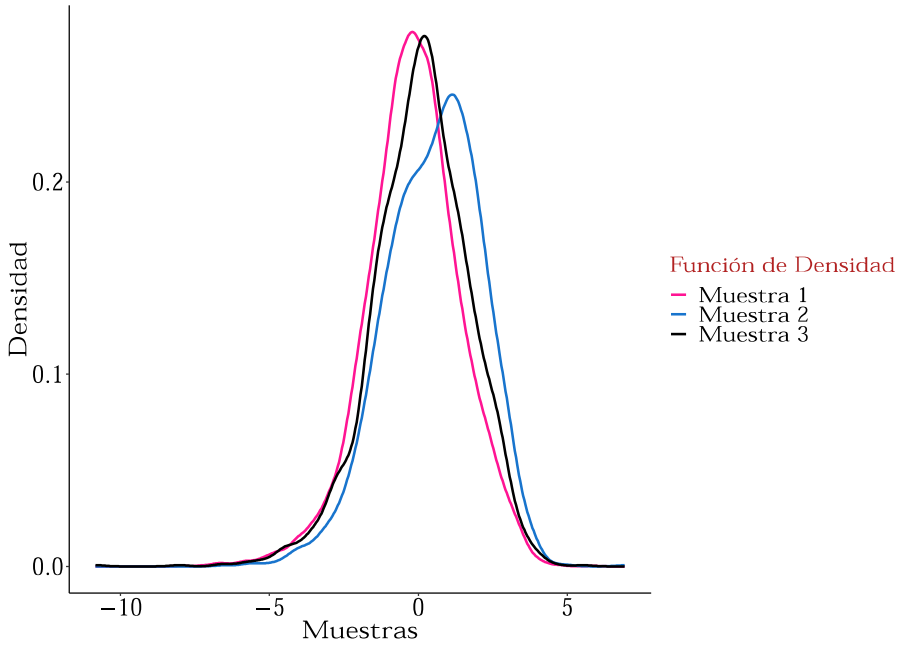
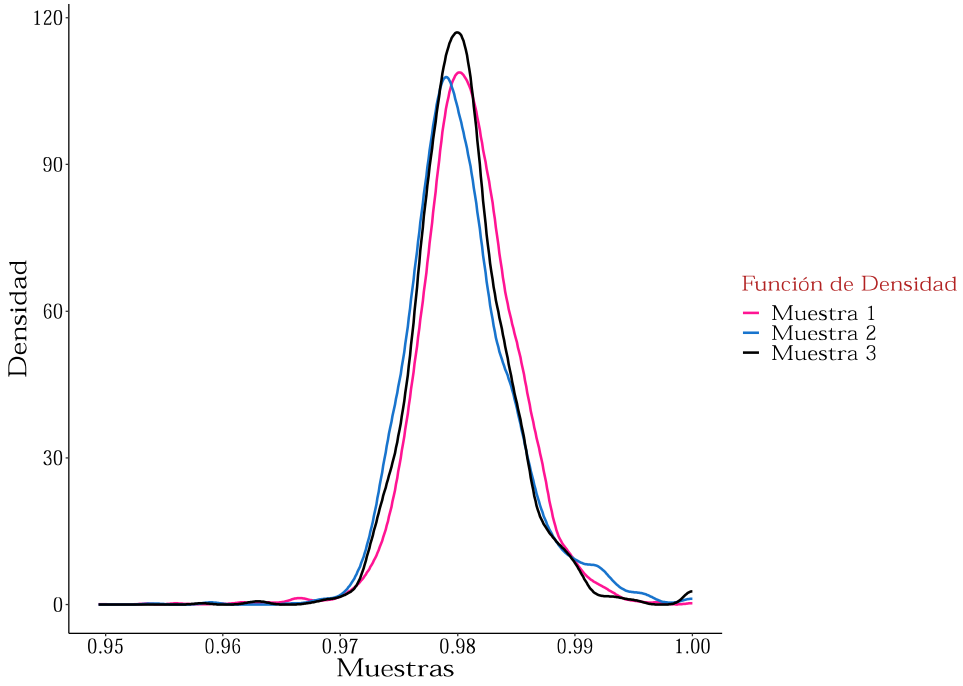


Figura 4.16: Gráfica de Densidad de las muestras provenientes del Índice CI.



Esta prueba no requiere ninguna suposición con respecto a la ubicación de las dos poblaciones normales. Lo cual implica que las magnitudes de las varianzas mues-

[28] Más detalle acerca de la prueba F de Fisher vea en Montgomery (2004).

trales sean directamente comparables, ya que cada una de ellas se computa como medidas de desviación alrededor de las medias muestrales respectivas (Gibbons y Chakraborti 2011).

De este modo, los resultados de aplicar la prueba F son:

Para las muestras provenientes del índice ACP.

Tabla 4.25: Test F de Fisher: Muestra 1 y Muestra 2 (Índice ACP).

	F
Test statistic	0.9684002
df numerador	7414
df denominador	2295
p value	0.337529
alternative	two.sided

F test to compare two variances: Muestra 1 y Muestra 2

Tabla 4.26: Test F de Fisher: Muestra 2 y Muestra 3 (Índice ACP).

	F
Test statistic	0.9703084
df numerador	2295
df denominador	2144
p value	0.477642
alternative	two.sided

F test to compare two variances: Muestra 2 y Muestra 3

Tabla 4.27: Test F de Fisher: Muestra 1 y Muestra 3 (Índice ACP).

	F
Test statistic	0.9396468
df numerador	7414
df denominador	2144
p value	0.07005081
alternative	two.sided

F test to compare two variances: Muestra 1 y Muestra 3

Como los p-valores en cada caso son mayores al 0.05, entonces aparentemente no existe suficiente evidencia estadística para rechazar la hipótesis nula de igualdad de varianzas, según el caso planteado.

Para las muestras provenientes del índice CI.

Tabla 4.28: Test F de Fisher: Muestra 1 y Muestra 2 (Índice CI).

	F
Test statistic	0.8338801
df numerador	7414
df denominador	2295
p value	4.561641e-08
alternative	two.sided

F test to compare two variances: Muestra 1 y Muestra 2

Tabla 4.29: Test F de Fisher: Muestra 2 y Muestra 3 (Índice CI).

	F
Test statistic	1.237193
df numerador	2295
df denominador	2144
p value	5.761284e-07
alternative	two.sided

F test to compare two variances: Muestra 2 y Muestra 3

Tabla 4.30: Test F de Fisher: Muestra 1 y Muestra 3 (Índice CI).

	F
Test statistic	1.03167
df numerador	7414
df denominador	2144
p value	0.3734206
alternative	two.sided

F test to compare two variances: Muestra 1 y Muestra 3

Como el p-valor para las muestras 1 y 3 es mayor al 0.05, entonces aparentemente no existe suficiente evidencia estadística para rechazar la hipótesis nula de igualdad de varianzas para estas muestras.

No obstante, como la prueba F no es robusta respecto al supuesto de normalidad y dado que anteriormente se probó que las muestras no provienen de una población normalmente distribuida entonces los resultados presentados en las Tablas 4.25, 4.26, 4.27, 4.28, 4.29 y 4.30 **no son confiables**. Por lo tanto es apropiado usar test no paramétricos de dispersión. Los resultados son los siguientes:

Para las muestras provenientes del índice ACP.

Tabla 4.31: Mood Test: Muestra 1 y Muestra 2 (índice ACP)

	M_N
z value	-3.623982
p value	0.0002901019
alternative	two.sided

Mood two-sample test of scale: Muestra 1 y Muestra 2

Tabla 4.32: Mood Test: Muestra 2 y Muestra 3 (índice ACP)

	M_N
z value	1.500558
p value	0.1334699
alternative	two.sided

Mood two-sample test of scale: Muestra 2 y Muestra 3

Tabla 4.33: Mood Test: Muestra 1 y Muestra 3 (índice ACP)

	M_N
z value	-1.420312
p value	0.155517
alternative	two.sided

Mood two-sample test of scale: Muestra 1 y Muestra 3

Tabla 4.34: Freund–Ansari–Bradley test Muestra 1 y Muestra 2 (índice ACP).

	A_N
Test statistic	18269545
p value	6.817556e-06
alternative	two.sided

Ansari-Bradley Test: Muestra 1 y Muestra 2

Tabla 4.35: Freund–Ansari–Bradley Test Muestra 2 y Muestra 3 (índice ACP).

	A_N
Test statistic	2498340
p value	0.01497213
alternative	two.sided

Ansari-Bradley test: Muestra 2 y Muestra 3

Tabla 4.36: Freund–Ansari–Bradley Test
Muestra 1 y Muestra 3 (índice ACP)

	A_N
Test statistic	17795902
p value	0.2113872
alternative	two.sided

Ansari-Bradley test: Muestra 1 y Muestra 3

Tabla 4.37: Siegel-Tukey Test
Muestra 1 y Muestra 2 (índice ACP).

	S_N
suma rank x	36535387
suma rank y	10621229
p value	5e-05
alternative	two.sided

Siegel-Tukey-test for equal variability: Muestra 1 y Muestra 2

Tabla 4.38: Siegel-Tukey Test
Muestra 2 y Muestra 3 (índice ACP).

	S_N
suma rank x	4995518
suma rank y	4867943
p value	0.0153
alternative	two.sided

Siegel-Tukey-test for equal variability: Muestra 2 y Muestra 3

Tabla 4.39: Siegel-Tukey Test
Muestra 1 y Muestra 3 (índice ACP).

	S_N
suma rank x	35588106
suma rank y	10113475
p value	0.21175
alternative	two.sided

Siegel-Tukey-test for equal variability: Muestra 1 y Muestra 3

Tabla 4.40: Klotz Normal-Scores Test
Muestra 1 y Muestra 2 (índice ACP).

	K_N
z value	-0.5205162
p value	0.6027039
alternative	two.sided

Two-Sample Klotz Test: Muestra 1 y Muestra 2

Tabla 4.41: Klotz Normal-Scores Test
Muestra 2 y Muestra 3 (índice ACP).

	K_N
z value	-0.6933497
p value	0.4880901
alternative	two.sided

Two-Sample Klotz Test: Muestra 2 y Muestra 3

Tabla 4.42: Klotz Normal-Scores Test
Muestra 1 y Muestra 3 (índice ACP).

	K_N
z value	-0.8721366
p value	0.3831339
alternative	two.sided

Two-Sample Klotz Test: Muestra 1 y Muestra 3

Como se puede apreciar en las Tablas 4.31, 4.34, 4.37 existe suficiente evidencia estadística para rechazar el supuesto que la muestra 1 y muestra 2 tienen la misma

variabilidad. Para las muestras 2 y 3, dependiendo de test que se utilice, se rechaza (Freund–Ansari–Bradley Test y Siegel-Tukey Test) y no se rechaza (Mood Test y Klotz Normal-Scores Test) el supuesto de igualdad de varianzas. Para las muestras 1 y 3, todos los test muestran que existe suficiente evidencia estadística para no rechazar el supuesto de igualdad de varianzas.

Para muestras provenientes del índice CI.

Tabla 4.43: Mood Test: Muestra 1 y Muestra 2 (índice CI).

	M_N
z value	-2.813113
p value	0.004906436
alternative	two.sided

Mood two-sample test of scale: Muestra 1 y Muestra 2

Tabla 4.44: Mood Test: Muestra 2 y Muestra 3 (índice CI).

	M_N
z value	3.923185
p value	8.738601e-05
alternative	two.sided

Mood two-sample test of scale: Muestra 2 y Muestra 3

Tabla 4.45: Mood Test: Muestra 1 y Muestra 3 (índice CI).

	M_N
z value	2.252641
p value	0.02428179
alternative	two.sided

Mood two-sample test of scale: Muestra 1 y Muestra 3

Tabla 4.46: Freund–Ansari–Bradley test Muestra 1 y Muestra 2 (índice CI).

	A_N
Test statistic	18134685
p value	0.02769858
alternative	two.sided

Ansari-Bradley Test: Muestra 1 y Muestra 2

Tabla 4.47: Freund–Ansari–Bradley Test Muestra 2 y Muestra 3 (índice CI).

	A_N
Test statistic	2470963
p value	0.0002028544
alternative	two.sided

Ansari-Bradley test: Muestra 2 y Muestra 3

Tabla 4.48: Freund–Ansari–Bradley Test Muestra 1 y Muestra 3 (índice CI).

	A_N
Test statistic	17583045
p value	0.01134321
alternative	two.sided

Ansari-Bradley test: Muestra 1 y Muestra 3

Tabla 4.49: Siegel-Tukey Test
Muestra 1 y Muestra 2 (índice CI).

	S_N
suma rank x	36265665
suma rank y	10890951
p value	0.0279
alternative	two.sided

Siegel-Tukey-test for equal variability: Muestra 1 y Muestra 2

Tabla 4.50: Siegel-Tukey Test
Muestra 2 y Muestra 3 (índice CI).

	S_N
suma rank x	4940823
suma rank y	4922638
p value	0.00015
alternative	two.sided

Siegel-Tukey-test for equal variability: Muestra 2 y Muestra 3

Tabla 4.51: Siegel-Tukey Test
Muestra 1 y Muestra 3 (índice CI).

	S_N
suma rank x	35162411
suma rank y	10539170
p value	0.0119
alternative	two.sided

Siegel-Tukey-test for equal variability: Muestra 1 y Muestra 3

Tabla 4.52: Klotz Normal-Scores Test
Muestra 1 y Muestra 2 (índice CI).

	K_N
z value	-6.066959
p value	1.303551e-09
alternative	two.sided

Two-Sample Klotz Test: Muestra 1 y Muestra 2

Tabla 4.53: Klotz Normal-Scores Test
Muestra 2 y Muestra 3 (índice CI).

	K_N
z value	4.196458
p value	2.711222e-05
alternative	two.sided

Two-Sample Klotz Test: Muestra 2 y Muestra 3

Tabla 4.54: Klotz Normal-Scores Test
Muestra 1 y Muestra 3 (índice CI).

	K_N
z value	0.07974379
p value	0.936441
alternative	two.sided

Two-Sample Klotz Test: Muestra 1 y Muestra 3

Como se puede apreciar, en la gran mayoría de las tablas (Tabla 4.43 a la Tabla 4.53) existe suficiente evidencia estadística para rechazar el supuesto que las muestras tienen la misma variabilidad. La excepción se dá para las muestras 1 y 3 con el test Klotz Normal-Scores (vea la Tabla 4.54).

Finalmente, en adición a lo expuesto en los capítulo III y IV, el lector está en la libertad de visualizar y realizar variaciones en lo referente al criterio de depuración de datos, muestras consideradas, verificación de los supuestos de normalidad para cada muestra, construcción del índice de salud derivado del ACP no condicionado al primer componente principal o para el caso del índice de salud basado en

las ponderaciones dadas por el Benefit of the Doubt no limitarse a las seis primeras componentes principales, resultados de la aplicación del Test Anova y Test de Kruskal-Wallis, proceso de comparaciones múltiples (tanto para el Test ANOVA como para el Test de Kruskal-Wallis) y resultados de la aplicación del test paramétrico F de Fisher y los test Mood Test y Freund–Ansari–Bradley Test dentro de lo no paramétrico para el problema de escala. Esto se puede realizar ingresando en el enlace web https://cristian-guatemal-work.shinyapps.io/TesisCG/.

En este mismo enlace (subpestaña ***Anova vs Kruskal-Wallis bajo el cumplimiento de normalidad*** de la pestaña ***Depuración de datos - -índices de Salud- -Aplicación***) se presenta también un claro ejemplo de que si las muestras cumplen con los requisitos que exige la teoría paramétrica bajo normalidad, los resultados alcanzados tanto por el Test Anova como del Test Kruskal-Wallis serán los mismos, y de hecho perfectamente comparables.

CAPÍTULO 5

Conclusiones y Recomendaciones

5.1. Conclusiones

- En vista de que los resultados proporcionados por ambos indicadores de salud tienden a mostrar suficiente evidencia estadística para rechazar el supuesto que las muestras consideradas en el estudio provienen de la misma población, un factor implícito que podría desencadenar esta diferencia a nivel de ubicación sería los hábitos de vida que tienen los trabajadores según la actividad económica que realicen. Es decir, dependiendo de la actividad existe cierta probabilidad de que un trabajador expuesto a esos hábitos pueda presentar lesiones o enfermedades severas que le dejen como secuela a corto o largo plazo una incapacidad temporal, incapacidad permanente, incapacidad total, incapacidad absoluta o incluso la muerte. En este sentido, a pesar que el individuo está aparentemente sano, sus diagnósticos clínicos estarán condicionados a sus hábitos de vida, lo que implica que el individuo según su actividad económica tienda a mostrar perfiles de salud distintos a los demás. Este hecho podría reflejar las diferencias existentes a nivel de ubicación, mismas que evitan que las muestras según la actividades económica sean necesariamente sustraídas de una misma población. Sin embargo, es posible que de alguna manera ciertas actividades en su mayor parte posible estén expuestas a los mismos hábitos de vida que otras, y de darse el caso de considerarlas, no necesariamente se rechazaría el supuesto que las muestras consideradas como las observaciones de los individuos pertenecientes a esas actividades provengan de la misma población. Un caso particular se ve reflejado en los resultados proporcionados por el enlace web al considerar las observaciones provenientes de las activida-

des económicas *P - ENSEÑANZA*, *S - OTRAS ACTIVIDADES DE SERVICIOS* y *O-ADMINISTRACIÓN PÚBLICA Y DEFENSA; PLANES DE SEGURIDAD SOCIAL DE AFILIACIÓN OBLIGATORIA*.

Adicionalmente, otro indicio que podría explicar las diferencias existentes entre poblaciones a nivel de ubicación, sería el número de accidentes de trabajos registrados por actividad económica. Estos valores presentados en la Tabla B.2, muestran la incidencia de los individuos a registrar un accidente de trabajo por actividad económica. En virtud de esto, las secciones de la CIIU con el mayor número de accidentes podrían estar expuestas a mayores factores de riesgo, y por ende, los individuos en éstas actividad tendrían a mostrar características en sus perfiles de salud que impliquen diferencias del resto de individuos de las otras actividades. Por ejemplo, las actividades *K - ACTIVIDADES FINANCIERAS Y DE SEGUROS*, *N - ACTIVIDADES DE SERVICIOS ADMINISTRATIVOS Y DE APOYO* y *L - ACTIVIDADES INMOBILIARIAS* en la Tabla B.2 presentan ciertas similitud en sus valores año tras año. Esto podría ser señal que sus perfiles de salud sean similares a nivel de ubicación. De hecho, con la ayuda del enlace web se verifica que efectivamente existe suficiente evidencia estadística para no rechazar el supuesto que estas muestras (observaciones de los individuos en la primera componente principal según la actividad) provienen de la misma población a nivel de ubicación.

En consecuencia, este estudio de alguna forma refleja que existen diferencias entre las poblaciones a nivel de ubicación. Sin embargo, las causas que generan dichas diferencias necesitan ser analizadas con más detalle, pues podría darse el caso de que los diferentes hábitos de vida a los que está expuesto un individuo, según su actividad económica sea un determinante de esas diferencias y por ende se tienda a tener perfiles de salud diferentes.

- Para el indicador del ACP, pese a que el número de componentes principales que explican la mayor parte de la información concerniente a las variables originales es grande (seis para ser exactos), y de ellas se escogen las proyecciones de los individuos en la primera componente principal pertenecientes a las actividades económicas más numerosas; no es limitación en la ejecución del presente estudio, debido a que el objetivo del mismo es la comparación entre test paramétricos y no paramétricos para una muestra arbitraria.
- En nuestro caso, el indicador CI (que generalmente considera conceptos multidimensionales que no pueden ser capturados por los indicadores simples para

dar una mejor compresión acerca de la complejidad inherente de los desafíos económicos, sociales y ambientales de la globalización) compara a los distintos individuos respecto al punto de referencia "individuo enfermo" utilizando las componentes principales que abarcan la mayor variabilidad de la información. Desde esta perspectiva, valores de las ponderaciones del Benefit of the Doubt (algunos registros en la Tabla 3.12) cercanos a 1 muestran evidencia que el individuo está enfermo (y de hecho la gran mayoría muestra este padecimiento). En este sentido, la comparación se la estaría realizando entre poblaciones "enfermas". A pesar de ello, se sigue mostrando evidencia para rechazar el supuesto de poblaciones idénticas.
Esto nos conlleva a concluir que las diferencias entre las poblaciones a nivel de ubicación ya no se deben a las muestras derivadas de los indicadores de salud, sino más bien por las características propias de la población (actividades económicas), y esto podría estar relacionado con lo expuesto en la primera conclusión.

- En un sentido de Eficiencia Relativa Asintótica, la prueba U de Mann-Whitney (o Wilcoxon Sum Rank Test) es la contraparte más eficiente a la prueba paramétrica t-Student tradicional y el test de Kruskal-Wallis al test paramétrico Anova. Así, como consecuencia directa de los resultados de la ARE vista en (2.3), para que la potencia de los test no paramétrico antes mencionados sea igual a la potencia de los test paramétrico, igualmente antes mencionados, se requiere que el tamaño de muestra de estos últimos sea 0.95 veces el tamaño de muestra de los primeros.

5.2. Recomendaciones

- A pesar de que en desarrollo del presente trabajo de titulación se considera el término no paramétrico como sinónimo de distribución libre, es necesario recalcar que para algunos autores estos términos son diferentes. Esto se corrobora con lo expresado por Bradley (1968) citado en Martínez (2004), al decir que: "Un test no paramétrico no hace hipótesis sobre el valor de los parámetros en una cierta distribución, mientras que un test de distribución libre no hace hipótesis sobre la forma precisa de la distribución".
- El proceso de depuración de datos puede considerar a las variables que tenga otra u otras cotas máximas de datos perdidos, y luego seguir la misma meto-

dología propuesta como en el caso cuando las variables tenían a lo mucho el 10 % de datos perdidos.

- Para mejorar o en su defecto crear un nuevo indicador de salud de los individuos dentro de las actividades económicas, se puede considerar el criterio y/o recomendaciones de un especialista en el área de salud (aplicado a las ponderaciones dadas tanto del indicador ACP como a las del indicador CI). Esto con el fin de si fuera el caso, considerar nuevas variables, pesos y ejes dados por el procedimiento del ACP aplicado a los datos depurados (indicador ACP) o a su vez cambiar las ponderaciones dadas por el Benefit of the Doubt mediante la añadidura de nuevas restricciones (como las de la Tabla 2.14) que consideren la opinión de los expertos.
- El proceso de inferencia se puede extender a las proyecciones de los individuos en la segunda, tercera o p-ésima componente principal, esto con el fin de luego de aplicar del Test Anova y Test de Kruskal-Wallis, comparar los resultados obtenidos por cada uno de estos y ver si se sigue rechazando el supuesto que las muestra provienen de la misma población.
- Seguir la sugerencia dada por Montgomery (2004): " Cuando exista preocupación acerca del supuesto de normalidad o por el efecto de puntos atípicos o valores "absurdos", se recomienda que el análisis de varianza común se realice tanto a los datos originales como a los rangos" (p.118). Y luego comparar sus resultados para tomar una adecuada de decisión.
- Si el investigador en cuestión desea realizar una comparación entre poblaciones, para las cuales asume que estas son iguales en todos los sentidos excepto posiblemente a nivel de ubicación o de escala, y desee evitar el trabajo tedioso de verificar el cumplimiento de cada uno de los supuestos de normalidad, entonces con toda confianza puede utilizar el Test de Mann-Whitney para comparar dos muestras independientes o el Test de Kruskal-Wallis para comparar k muestras independientes.
- Se debe realizar un análisis más profundo de las posibles causas que hacen que las muestras en cuestión rechacen el supuesto de poblaciones idénticas. Adicionalmente, el mismo procedimiento práctico expuesto en este trabajo podría ser aplicado a una muestra de datos más extensa, como por ejemplo, los registros de diagnóstico clínico de las bases de salud del Instituto Ecuatoriano de Seguridad Social.

Referencias Bibliográficas

Conover, W. J. y R. L. Iman (1981). «Rank Transformations as a Bridge between Parametric and Nonparametric Statistics». En: *The American Statistician* 35, págs. 124-129. URL: https://doi.org/10.1080/00031305.1981.10479327.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Science+Business Media, Inc.

Montgomery, D. C. (2004). *DISEÑO Y ANÁLISIS DE EXPERIMENTOS*. 2.ª ed. Limusa S.A.

Gamarra, H., A. Pérez y R. Quiseno (2006). *Estadística No Paramétrica*. URL: http://repositorio.unisucre.edu.co/handle/001/146.

Rojas, A. (2003). *Técnicas Estadísticas Paramétricas y No Paramétricas Equivalentes: Resultados Comparativos Por Simulación*. URL: http://www.iuma.ulpgc.es/~nunez/mastertecnologiastelecomunicacion/RecursosGenerales/TesisEstadisticaParametricayNoParametrica.pdf.

Molinero, L. M. (2003). *¿Y si los datos no siguen una distribución normal?... Bondad de ajuste a una normal. Transformaciones. Pruebas no paramétricas*. URL: https://www.alceingenieria.net/bioestadistica/noparame.pdf.

Gibbons, J. D. y S. Chakraborti (2011). *Nonparametric Statistical Inference*. 5.ª ed. Taylor y Francis Group, LLC.

Pratt, J. W. y J. D. Gibbons (1981). *Concepts of Nonparametic Theory*. 1.ª ed. Springer-Verlag New York, Inc.

Kvam, P. H. y B. Vidakovic (2007). *Nonparametric Statistics with Applications to Science and Engineering*. 1.ª ed. John Wiley & Sons, Inc.

Bradley, J. V. (1960). *Distribution-Free Statistical Tests*. URL: http://www.dtic.mil/dtic/tr/fulltext/u2/249268.pdf.

Martin, P. (2011). «Aplicación de la estadística no paramétrica en el área de rehabilitación.» En: *Rehabil. integral* 6, págs. 93-99. URL: https://www.rehabilitacionintegral.cl/wp-content/files_mf/6sanmart%C3%ADn99.pdf.

Moses, L. E. (2011). «Non-parametric statistics for psychological research». En: *American Psychological Association* 49, págs. 122-143. URL: http://psycnet.apa.org/doiLanding?doi=10.1037%2Fh0056813.

Espitia, B. M. (2014). *DIFERENCIAS EN CONDICIONES DE EMPLEO, CONDICIONES DE TRABAJO Y EN SALUD MENTAL LABORAL, SEGÚN LA POSICIÓN DE CLASE SOCIAL DE TRABAJADORES ASALARIADOS DE BOGOTÁ, 2013*. URL: http://bdigital.unal.edu.co/49675/1/333689642015.pdf.

Johnson, R. A. y D. W. Wichern (2007). *Applied Multivariate Statistical Analysis*. 6.ª ed. Pearson Educación, Inc.

Peña, D. (2002). *Análisis de datos multivariantes*.

Randles, R. H., T. P. Hettmansperger y G. Casella (2004). «Introduction to the Special Issue: Nonparametric Statistics.» En: *Institute of Mathematical Statistics*. 19, pág. 561. URL: https://projecteuclid.org/download/pdfview_1/euclid.ss/1113832719.

Gómez, M.A. (2010). *ERICH LEHMANN (1917-2009) OBITUARIO*. URL: http://www.mat.ucm.es/~villegas/AnLehmanSEIO.pdf.

Gómez-Gómez, M., C. Danglot-Banck y L. Vega-Franco (2003). «Sinopsis de pruebas estadísticas no paramétricas. Cuándo usarlas». En: *Revista Mexicana de Pediatría* 70, págs. 91-97. URL: http://www.medigraphic.com/pdfs/pediat/sp-2003/sp032i.pdf.

Kraska-Miller, M. (2014). *NONPARAMETRIC STATISTICS FOR SOCIAL AND BEHAVIORAL SCIENCES*. 1.ª ed. Taylor & Francis Group, LLC.

Hoeffding, W. y H. Robbins (1948). «The central limit theorem for dependent random variables.» En: *Duke Mathematical Journal*. 15, págs. 773-780. URL: https://projecteuclid.org/euclid.dmj/1077475030.

Diananda, P. (1955). «The central limit theorem for m-dependent variables .» En: *Mathematical Proceedings of the Cambridge Philosophical Society*. 15, págs. 92-95. URL: https://doi.org/10.1016/S0167-7152(99)00146-7.

Orey, S. (1958). «A central limit theorem for m-dependent random variables.» En: *Duke Mathematical Journal*. 25, págs. 543-546. URL: https://projecteuclid.org/euclid.dmj/1077468187.

Romano, J. y M. Wolf (2000). «A more general central limit theorem for m-dependent random variables with unbounded m.» En: *Statistics & Probability Letters*. 47, págs. 115-124. URL: https://doi.org/10.1016/S0167-7152(99)00146-7.

Wilcoxon, Frank (1945). «Individual Comparisons by Ranking Methods». En: *International Biometric Society.* 1, págs. 80-83. URL: http://links.jstor.org/sici?sici=0099-4987%28194512%291%3A6%3C80%3AICBRM%3E2.0.CO%3B2-P.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. McGRAW-HILL.

Wald, A. y J. Wolfowitz (1939). «On a test whether two samples are from the same population». En: *The Annals of Mathematical Statistics.* Págs. 147-162. URL: www.jstor.org.

Friedman, J. H. y L. C. Rafsky (1979). «MULTIVARIATE GENERALIZATIONS OF THE WALD-WOLFOWITZ AND SMIRNOV TWO-SAMPLE TESTS.» En: *The Annals of Statistics.* 7, págs. 697-717. URL: http://www.jstor.org/stable/2958919.

Mann, H. B. y D. R. Whitney (1947). «On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other.» En: *The Annals of Mathematical Statistics.* 8, págs. 50-60. URL: http://www.jstor.org/stable/2236101.

Corder, G. W. y D. I. Foreman (2009). *Nonparametric Statistics for Non-Statisticians*. 1.ª ed. John Wiley & Sons.

Randles, R. H. (2012). «On Neutral Responses (Zeros) in the Sign Test and Ties in the Wilcoxon–Mann–Whitney Test.» En: *The American Statistician.* Págs. 96-101. URL: http://dx.doi.org/10.1198/000313001750358554.

Hálek, J. (1968). «ASYMPTOTIC NORMALITY OF SIMPLE LINEAR RANK STATISTIC UNDER ALTERNATIVES.» En: *The Annals of Mathematical Statistics.* 39, págs. 325-346. URL: www.jstor.org.

Chernoff, H. y R. Savage (1957). «ASYMPTOTIC NORMALITY AND EFFICIENCY OF CERTAIN NONPARAMETRIC TEST STATISTICS.» En: *The Annals of Mathematical Statistics.* 39, págs. 972-994. URL: www.jstor.org.

Hodges, J. y E. Lehmann (1956). «The Efficiency of Some Nonparametric Competitors of the *t*-Test.» En: *The Annals of Mathematical Statistics.* 27, págs. 324-336. URL: https://www.jstor.org/stable/2236996?seq=1#page_scan_tab_contents.

Klotz, J. H. (1964). «ON THE NORMAL SCORES TWO-SAMPLE RANK TEST.» En: *Journal of the American Statistical Association.* Págs. 652-664. URL: http://dx.doi.org/10.1080/01621459.1964.10480720.

Terry, M. E. (1951). «SOME RANK ORDER TESTS WHICH ARE MOST POWERFUL AGAINST SPECIFIC PARAMETRIC ALTERNATIVES.» En: *The Annals of Mateha-matical Statistics.* Págs. 346-366. URL: www.jstor.org.

Harter, H. Leon (1961). «Expected Values of Normal Order Statistics.» En: *Biometrika Trust.* Págs. 151-165. URL: http://www.jstor.org/stable/2333139.

Mood, A. M. (1954). «ON THE ASYMPTOTIC EFFCIENCY OF CERTAIN NON-PARAMETRIC TWO-SAMPLE TESTS.» En: *The Annals of Mathematical Statistics.* Págs. 514-522. URL: www.jstor.org.

Laubscher, N. F., F. E. Steffens y E. M. de Lange (1968). «Exact Critical Values for Mood's Distribution-Free Test Statistic for Dispersion and Its Normal Approximation.» En: *American Statistical Association and American Society for Quality.* 10, págs. 514-522. URL: http://www.jstor.org/stable/1267104.

Wiel, M. A. van de. *The probability generating funvtion of the Freund-Ansari-Bradley statistic.* URL: http://www.mat.ufrgs.br/~viali/estatistica/mat2282/material/textos/Ansari_Bradley.pdf.

Ansari, A. R. y R. A. Bradley (1959). «RANK-SUM TESTS FOR DISPERSIONS.» En: *The Annals of Mathematical Statisics.* 10, págs. 1174-1189. URL: www.jstor.org.

Siegel, S. y J. W. Tukey (1960). «A Nonparametric Sum of Ranks Procedure for Relative Spread in Unpaired Samples.» En: *Journal of the American Statistical Association.* Págs. 429-445. URL: http://dx.doi.org/10.1080/01621459.1960.10482073.

Klotz, J. H. (1960). «NONPARAMETRIC TESTS FOR SCALE.» En: *The Annals of Mathematical Statistics.* Págs. 498-512. URL: www.jstor.org.

Miller, R. G. (1981). *Simultaneous Statistical Inference*. 2.ª ed. Springer-Verlag.

Kruskal, W. H. y W. A. Wallis (1952). «Use of Ranks in One-Criterion Variance Analysis.» En: *Journal of the American Statistical Association.* 47, págs. 583-621. URL: http://www.jstor.org/stable/2280779.

Kruskal, W. H. (1952). «A Nonparametric test for the Several Sample Problem.» En: *The Annals of Mathematical Statistics.* 23, págs. 525-540. URL: http://www.jstor.org/stable/2236578.

Dunn, O. (1964). «Multiple Comparisons Using Rank Sums.» En: *American Society for Quality.* 6, págs. 241-252. URL: http://www.jstor.org/stable/1266041.

Belmonte, R. (2001). *EFICIENCIA RELATIVA ASINTÓTICA.* URL: http://www.revistasbolivianas.org.bo/scielo.php?pid=S9876-67892001000100004&script=sci_arttext.

Mayorga, J. (2004). *Inferencia Estadística*. 1.ª ed. Universidad Nacional de Colombia.

Lehmann, E. (2008). «Parametric versus nonparametrics: two alternative methodologies». En: *Journal of Nonparametric Statistics.* 21, págs. 337-445. URL: https://core.ac.uk/download/pdf/81538486.pdf.

Suárez, O. (2007). *APLICACIÓN DEL ANÁLISIS FACTORIAL A LA INVESTIGACIÓN DE MERCADOS. CASO DE ESTUDIO*. URL: https://dialnet.unirioja.es/descarga/articulo/4804281.pdf.

Kaiser, H. F. (1974). «AN INDEX OF FACTORIAL SIMPLICITY*.» En: *PSYCHOMETRIK*. 39, págs. 31-36. URL: https://jaltcue.org/files/articles/Kaiser1974_an_index_of_factorial_simplicity.pdf.

Nardo, M., M. Saisana, A. Hoffmann A. Saltelli S. Tarantola y E. Giovannini (2008). *Handbook on Constructing Composite Indicators.* 5.ª ed. OECD, the Econometrics y Applied Statistics Unit of the Joint Research Centre (JRC) of the European Commission.

Rogge, N. (2008). *'Benefit of the doubt'composite indicators Deliverable 5.3*. URL: https://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/KEI-WP5-D5.3.pdf.

Cherchye, L., W. Moesen, N. Rogge y T. Puyenbroeck (2016). «AN INTRODUCTION TO "BENEFIT OF THE DOUBT" COMPOSITE INDICATORS.» En: *Social Indicators Research.* 82, págs. 111-145. URL: https://link.springer.com/article/10.1007/s11205-006-9029-7.

Triola, M. F. (2009). *Nonparametric Statistics for Non-Statisticians*. 11.ª ed. Addison-Wesley.

Lind, D., W. Marchal y S. Wathen (2012). *Estadística aplicada a los negocios y la economía*. 15.ª ed. McGraw-Hill.

Joaquin, R. (2016). *ANOVA análisis de varianza para comparar múltiples medias*. URL: https://rpubs.com/Joaquin_AR/219148.

Joaquin, R. A. (2016). *Kruskal-Wallis test. Alternativa no paramétrica al ANOVA independiente*. URL: https://rpubs.com/Joaquin_AR/219504.

Martínez, E. (2004). *Métodos No Paramétricos I.* URL: http://www.dm.uba.ar/materias/optativas/metodos_no_parametricos/2004/2/NoparI01.pdf.

MSP-Ecuador (2010). *NORMAS Y PROTOCOLOS para la alimentación y nutrición en la atención integral a personas que viven con VIH/SIDA*. URL: http://www.coalicionecuatoriana.org/web/pdfs/Normas_Alimentacion_para_PVV.pdf.

infosalus.com (2019). *Hipertensión y presión arterial*. URL: https://www.infosalus.com/enfermedades/cardiologia/hipertension/que-es-hipertension-69.html.

Sáenz, K., L. Narváez y M. Cruz (2008). *Valores de referencia hematológicos en población altoandina ecuatoriana*. URL: https://www.researchgate.net/publication/288891490_Valores_de_referencia_hematologicos_en_poblacion_altoandina_ecuatoriana.

tuotromedico (2018). *RDW en la sangre normal*. URL: https://www.tuotromedico.com/parametros/RDW-en-sangre-normal.htm.

Yánez, M. (2016). *Frecuencia de obesidad exógena y sobrepeso en estudiantes de los terceros años de bachillerato en contabilidad y administración, administración sistemas del colegio simón bolívar en el período abril-junio de 2016*. URL: http://www.dspace.uce.edu.ec/bitstream/25000/8084/1/T-UCE-0006-010.pdf.

Anexos

ANEXO A

Anexo I: Depuración de Datos y ACP

A.1. Variables codificadas y porcentaje de datos perdidos por variable

Tabla A.1: Variables codificadas y porcentaje de datos perdidos (NA).

Variable	NA's	% de NA's	Codificación
CALCIO.SERICO.TOTAL	23334	100.00	a
HIERRO.SERICO	23334	100.00	a
POTASIO..K.	23334	100.00	a
SODIO..NA.	23334	100.00	a
TRANSFERRINA	23334	100.00	a
GLUCOSA.POSTPRANDIAL	23334	100.00	a
INSULINA.BASAL	23334	100.00	a
APOLIPOPROTEINA.A1	23334	100.00	a
ASTO.CUANTITATIVO	23334	100.00	a
CROMO.EN.SANGRE	23334	100.00	a
MAGNESIO.EN.SANGRE	23334	100.00	a
ANTICUERPOS.ANTI.CHLAMYDIA.TRACHOMATIS.IGG	23334	100.00	a
ANTICUERPOS.ANTI.CHLAMYDIA.TRACHOMATIS.IGM	23334	100.00	a
ANTICUERPOS.ANTI.HERPES.II.IGG	23334	100.00	a
ANTICUERPOS.ANTI.HERPES.II.IGM	23334	100.00	a
COLESTEROL.VLDL	23333	99.99	a
PLOMO.EN.SANGRE	23332	99.99	a
HIV.1.2.ANTICUERPOS...ANTIGENO.P24	23330	99.98	a
LIPIDOS.TOTALES	23329	99.97	a
RETICULOCITOS	23328	99.97	a
COLINESTERASA.ACETIL.ERITROCITARIA..SANGRE.TOTAL.	23327	99.97	a
TIEMPO.DE.PROTROMBINA..TP.	23324	99.95	a
FT3.LIBRE	23322	99.94	a
ALFA.FETO.PROTEINA..AFP.	23321	99.94	a
CA.125	23321	99.94	a
CA.72.4	23321	99.94	a
HOMOCISTEINA	23321	99.94	a
T3.TOTAL	23321	99.94	a
FT4.LIBRE	23320	99.94	a
CK.MB..MASA.	23320	99.94	a
CPK	23319	99.93	a
T4.TOTAL	23318	99.93	a
PCR.CUANTITATIVO	23312	99.90	a
AMILASA	23312	99.90	a
APOLIPOPROTEINA.B	23302	99.86	a
GLOBULINA	23301	99.85	a
LDH	23298	99.84	a
VELOCIDAD.DE.SEDIMENTACION.1.HORA	23296	99.83	a
ALBUMINA	23295	99.83	a
HEPATITIS.B..ANTICUERPOS.ANTI.HBS.AG	23291	99.81	a

continúa en la página siguiente...

...continúa de la página anterior

PROTEINAS.TOTALES	23284	99.78	a
HEMOGLOBINA.GLICADA..HBA1C.	23260	99.68	a
X..PSA.TOTAL.PSA.LIBRE	23162	99.26	a
PSA.LIBRE	23142	99.17	a
BUN	23110	99.04	a
BILIRRUBINA.TOTAL	23033	98.71	a
HEPATITIS.B..ANTICUERPOS.ANTI.HBS	23018	98.64	a
CEA..ANTIGENO.CARCINOEMBRIONARIO.	22996	98.55	a
COLESTEROL.LDL..CUANTIFICADO.	22872	98.02	a
FOSFATASA.ALCALINA	22573	96.73	a
BILIRRUBINA.DIRECTA	22146	94.90	b
BILIRRUBINA.INDIRECTA	22146	94.90	b
TIEMPO.DE.TROMBOPLASTINA.PARCIAL..TTP.	22114	94.77	b
TSH	21999	94.27	b
PSA.TOTAL..ANTIGENO.PROSTATICO.ESPECIFICO.	21433	91.85	b
DENSIDAD	20986	89.93	c
PH	20986	89.93	c
GAMMA.GLUTAMIL.TRANSPEPTIDASA..GGT.	20234	86.71	c
UREA	13252	56.79	i
TGO...AST	13220	56.65	i
TGP...ALT	13217	56.64	i
ACIDO.URICO	8364	35.84	m
COLESTEROL.LDL..CALCULADO.	6934	29.72	o
COLESTEROL.HDL	6390	27.38	o
TEMPERATURA	5183	22.21	p
RESPIRACION	4799	20.57	p
TRIGLICERIDOS	4734	20.29	p
COLESTEROL.TOTAL	4680	20.06	p
SATURACION_OXIGENO	3915	16.78	q
CREATININA	2608	11.18	r
FRECUENCIA_CARDIACA	2337	10.01	r
GLUCOSA.BASAL	2106	9.03	s
IMC	1999	8.57	s
ESTATURA	1286	5.51	s
PRESION_ARTERIAL2	1286	5.51	s
PESO	1283	5.50	s
PRESION_ARTERIAL	1278	5.48	s
TIPO_SANGRE	988	4.23	t
HEMOGLOBINA.CORPUSCULAR.MEDIA	715	3.06	t
CIIU4.0_1	352	1.51	t
CIIU4.0_DESAGREGADO	352	1.51	t
VOLUMEN.PLAQUETARIO.MEDIO	254	1.09	t
CONCENTRACION.CORPUSCULAR.MEDIA.HEMOGLOBINA	252	1.08	t
HEMOGLOBINA	252	1.08	t
PLAQUETAS	251	1.08	t
VOLUMEN.CORPUSCULAR.MEDIO	251	1.08	t
HEMATOCRITO	251	1.08	t
ANCHO.DE.DISTRIBUCION.G.R.	251	1.08	t
IDENTIFICACION	0	0.00	u
FECHA_ADMISION	0	0.00	u
N	0	0.00	u
HABITO.A	0	0.00	u
HABITO.D	0	0.00	u
HABITO.S	0	0.00	u
HABITO.T	0	0.00	u
FECHA_NACIMIENTO	0	0.00	u

continúa en la página siguiente...

...continúa de la página anterior

GENERO	0	0.00	u
EDUCACION	0	0.00	u
EMPRESA	0	0.00	u

A.2. Variables que tienen a lo mucho el 10% de datos perdidos.

IDENTIFICACION	PESO
FECHA_ADMISION	ESTATURA
N	IMC
HABITO.A	PRESION_ARTERIAL
HABITO.D	PRESION_ARTERIAL2
HABITO.S	CONCENTRACION.CORPUSCULAR.MEDIA.HEMOGLOBINA
HABITO.T	HEMOGLOBINA
FECHA_NACIMIENTO	VOLUMEN.PLAQUETARIO.MEDIO
GENERO	PLAQUETAS
TIPO_SANGRE	VOLUMEN.CORPUSCULAR.MEDIO
EDUCACION	HEMOGLOBINA.CORPUSCULAR.MEDIA
EMPRESA	HEMATOCRITO
CIIU4.0_1	ANCHO.DE.DISTRIBUCION.G.R.
CIIU4.0_DESAGREGADO	GLUCOSA.BASAL

A.3. Variables utilizadas en el ACP.

Se considera solo a estas variables debido a que son cuantitativas y no categóricas.

Variable	Abreviatura
PESO	peso
ESTATURA	estt
IMC	imc
PRESION_ARTERIAL	prs_
PRESION_ARTERIAL2	pr_2
CONCENTRACION.CORPUSCULAR.MEDIA.HEMOGLOBINA	conc
HEMOGLOBINA	hgb
VOLUMEN.PLAQUETARIO.MEDIO	vpm
PLAQUETAS	plaq
VOLUMEN.CORPUSCULAR.MEDIO	vcm
HEMOGLOBINA.CORPUSCULAR.MEDIA	hcm,
HEMATOCRITO	hema
ANCHO.DE.DISTRIBUCION.G.R.	anch
GLUCOSA.BASAL	gluc

A.4. Datos en relación al mejor individuo sano.

La matriz de datos relacionada al mejor individuo sano corresponde a una transformación de las variables presentes en los datos depurados (véase 3.1, pág. 103). El procedimiento realizado fue:

- Como el Índice de Masa Corporal (imc) es una función del peso y la altura, solo se considera a esta variable en el análisis (siempre que el peso y la altura estén presentes).
- Imc.- A cada observación se le resta el promedio entre 18.5 y 24.99. Más información vea en MSP-Ecuador (2010).
- Presión Arterial.- A cada observación se le resta la presión arterial Sistólica óptima (120). Más información vea en infosalus.com (2019).
- Presión Arterial 2.- A cada observación se le resta la presión arterial Diastólica óptima (80). Más información vea en infosalus.com (2019).
- Concentración Corpuscular Media Hemoglobina.- A cada observación se le resta el promedio entre 34.64 (para hombre) y 34.06 (para mujeres). Más información vea en Sáenz, Narváez y Cruz (2008).
- Hemoglobina.- A cada observación se le resta el promedio entre 16.70 (para hombre) y 14.50 (para mujeres). Más información vea en Sáenz, Narváez y Cruz (2008).
- Volumen Plaquetario Medio.- A cada observación se le resta el promedio entre 10.54 (para hombres) y 10.50 (para mujeres). Más información vea Sáenz, Narváez y Cruz (2008).
- Plaquetas.- A cada observación se le resta el promedio entre 256.63 (para hombre) y 284.03 (para mujeres). Más información vea en Sáenz, Narváez y Cruz (2008).
- Volumen Corpuscular Medio.- A cada observación se le resta el promedio entre 88.04 (para hombre) y 88.37 (para mujeres). Más información vea en Sáenz, Narváez y Cruz (2008).
- Hemoglobina Corpuscular Media.- A cada observación se le resta el promedio entre 30.52 (para hombre) y 30.20 (para mujeres). Más información vea en Sáenz, Narváez y Cruz (2008).
- Hematócrito.- A cada observación se le resta el promedio entre 48.03 (para hombre) y 42.60 (para mujeres). Más información vea en Sáenz, Narváez y Cruz (2008).
- Ancho de Distribución G.R.- A cada observación se le resta el promedio entre 10.6 (para hombre) y 14.7 (para mujeres). Más información vea en tuotromedi-

co (2018).

- Glucosa Basal.- A cada observación se le resta el promedio entre 80 y 100 (niveles en ayuno). Más información vea en Yánez (2016).

Este enfoque nos permite establecer que el mejor individuo sano será aquel que tenga valores muy cercanos a cero en estas variables. En el mejor de los casos, el individuo totalmente sano será aquel que tenga todos estos valores iguales a cero.

A.5. Componentes Principales y Proyección de los individuos

Peso de los componentes principales

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
peso	0.5245	0.0467	-0.0142	0.0149	-0.3066	-0.1269	0.1034	-0.2589	-0.0529	0.0482	0.0661	0.0001	-0.0138	0.7244
estt	0.0127	0.0112	0.0124	-0.0213	0.1037	0.4565	0.8826	0.0064	-0.0136	-0.0047	-0.0245	0.0104	0.0006	-0.0075
imc	0.4928	0.0862	-0.0669	0.0087	-0.2841	-0.1733	0.1090	-0.2917	-0.2703	0.1286	0.0304	0.0664	0.0099	-0.6651
prs_	0.4219	0.0206	-0.1154	-0.0355	0.5283	0.0741	-0.0965	0.0619	0.0116	-0.0301	0.0703	-0.7033	0.0845	-0.0377
pr_2	0.4129	-0.0027	-0.0760	-0.0484	0.5337	0.1002	-0.1318	0.0845	0.1102	0.0019	-0.1633	0.6792	-0.0403	0.0152
conc	-0.0028	0.0945	-0.0584	0.7622	0.0365	-0.0032	0.0223	0.0872	0.2255	0.5860	0.0217	-0.0205	-0.0453	-0.0034
hgb	0.2538	-0.2961	-0.0414	0.0759	-0.3280	0.1200	-0.0140	0.1239	0.7232	-0.3444	-0.0283	-0.0625	-0.1725	-0.1636
vpm	0.0867	0.0297	0.7079	0.0706	0.0120	0.0177	-0.0372	-0.0820	-0.0326	0.0107	-0.6657	-0.1387	-0.1150	-0.0127
plaq	-0.0728	0.1817	-0.6509	-0.1327	-0.1333	-0.0251	0.0162	0.0509	0.0200	0.0898	-0.6903	-0.1046	-0.0359	0.0438
vcm	-0.0375	-0.5795	-0.1611	0.2023	0.1209	-0.0887	0.0410	-0.0585	-0.3699	-0.0941	-0.0402	-0.0383	-0.6467	0.0366
hcm,	0.0250	-0.6360	-0.0470	0.1548	-0.0180	-0.0739	0.0446	0.0192	-0.1156	0.0009	-0.1910	0.0341	0.7136	0.0212
hema	0.0357	-0.0537	-0.0572	0.0485	-0.2073	0.8366	-0.4072	-0.1756	-0.2085	0.0802	0.0182	-0.0035	0.0191	0.0078
anch	-0.0057	0.3383	-0.0907	0.5645	0.0278	-0.0019	-0.0003	-0.0278	-0.2067	-0.7026	-0.0512	0.0473	0.1257	0.0092
gluc	0.2298	0.0411	0.0806	-0.0199	-0.2550	0.0364	-0.0046	0.8770	-0.3161	0.0326	0.0237	0.0077	-0.0336	0.0174

Proyecciones de los 20 primeros individuos en las componentes principales

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
1	0.314965	-1.281236	-0.587206	0.706806	-1.625510	-0.494755	0.436678	-0.282985	-0.721161	-0.411115	0.603540	-0.016496	0.018889	-0.532318
2	-1.038018	-1.756577	-0.221568	0.562079	-0.398178	-0.079117	0.045350	0.308589	-0.024351	-0.443711	-0.405522	0.642753	0.124410	-0.392826
3	-3.284438	-0.141684	-0.214251	0.146072	-0.579016	-0.848883	0.508412	-1.120036	-1.621539	0.556296	0.736836	-0.327942	-0.216498	0.526665
4	0.083076	-1.749775	0.121153	-0.239805	0.508750	0.027595	-0.087116	0.052657	0.361050	0.430199	0.064833	0.606346	-0.202108	-0.148502
5	1.581764	-1.260650	-0.288860	0.415610	-0.781134	-0.378040	0.268590	0.111831	-0.797565	0.004780	-0.191663	-0.145843	0.663060	-0.179298
6	1.241496	-1.615989	-0.601508	0.495659	-0.049662	-0.363144	0.181350	-0.369905	-0.768224	-0.117940	-0.214077	0.508784	-0.241878	-0.149201
7	0.852427	1.044862	0.274954	0.249112	-1.243456	-0.114689	0.189016	0.207626	-0.272815	-0.471948	1.515466	0.015126	-0.759211	-0.391598
8	0.888750	0.570441	-0.161351	-0.134141	1.082305	0.196329	-0.242085	0.136744	-0.120708	-0.093943	0.029259	0.557134	-1.392437	0.223417
9	0.550141	0.662910	1.613277	0.188586	0.957855	0.344835	-0.316197	-0.098100	0.567308	-0.398500	-0.396105	0.549956	0.265823	0.195177
10	-1.210098	0.277280	-0.674475	0.105807	1.225463	0.152621	-0.310231	0.625621	-0.211651	-0.306663	-0.878281	-1.023031	0.387916	-0.417589
11	-0.817526	-1.019584	-1.515565	-0.330274	0.102556	0.098394	-0.089037	1.341387	0.005577	0.167055	0.811924	0.251944	-0.404485	-0.400481
12	-4.147063	-0.423389	-0.601625	-0.115217	3.765797	0.748109	-0.988000	1.137237	0.744420	-0.397705	-0.865236	0.720986	0.379332	-0.857005
13	0.363694	0.099121	1.572050	1.363734	-0.182195	0.405177	-0.219544	0.686315	0.374850	-1.845468	-0.390842	-0.125590	0.210023	-0.237764
14	-4.708464	1.654549	-0.169634	0.236624	-1.269360	0.156565	-0.051848	3.240637	-1.057894	-0.590333	0.041835	0.405114	-0.294376	0.054772
15	-1.320563	0.012210	0.039639	-0.021384	-0.027707	0.353581	-0.216353	1.467858	0.521276	-0.277809	-0.173350	-0.066866	1.105296	-0.055349
16	0.193574	-1.094196	-0.263259	0.419507	1.323357	-0.005014	-0.187275	0.164317	-0.786285	-0.233870	-0.365454	0.080298	-0.431307	-0.050834
17	-2.115753	0.787109	0.829658	0.161205	0.347992	-0.148071	-0.072598	2.436741	-2.243908	0.197738	0.356309	-0.107311	-0.044585	0.090487
18	0.322749	-0.940947	0.661439	-0.104050	0.754187	0.279158	-0.218749	-0.009176	0.677896	0.023588	0.862897	0.722332	-0.110643	-0.049386
19	-0.625968	-0.210199	1.695179	0.013017	-0.492184	-0.117434	0.069263	0.172928	-0.285559	0.344905	-1.451762	-0.537649	0.199341	0.325390
20	-0.430472	-0.323347	-1.071172	-0.267355	0.371179	0.162191	-0.140513	-0.289677	0.718176	0.064170	-1.773310	-0.228927	-0.306801	0.242099
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

ANEXO B

Anexo II: Aplicación y Resultados

B.1. Test de Normalidad.

Para las muestras provenientes del índice ACP

Kolmogorov-Smirnov Test

Tabla B.1: Kolmogorov-Smirnov Test para la Muestra 1 (índice ACP).

	D
Test statistic	0.02470489
p value	0.0002344967
Alternative hypothesis	two-sided

One-sample Kolmogorov-Smirnov Test: Muestra 1

Tabla B.2: Kolmogorov-Smirnov Test para la Muestra 2 (índice ACP).

	D
Test statistic	0.03751539
p value	0.003120842
Alternative hypothesis	two-sided

One-sample Kolmogorov-Smirnov Test: Muestra 2

Tabla B.3: Kolmogorov-Smirnov Test para la Muestra 3 (índice ACP).

	D
Test statistic	0.03285506
p value	0.01949268
Alternative hypothesis	two-sided

One-sample Kolmogorov-Smirnov Test: Muestra 3

Lilliefors Test

Tabla B.4: Lilliefors Test Test para la Muestra 1 (índice ACP).

	D
Test statistic	0.025
p value	0

Lilliefors (Kolmogorov-Smirnov) normality test: Muestra 1

Tabla B.5: Lilliefors Test Test para la Muestra 2 (índice ACP).

	D
Test statistic	0.038
p value	0.00000

Lilliefors (Kolmogorov-Smirnov) normality test: Muestra 2

Tabla B.6: Lilliefors Test Test para la Muestra 3 (índice ACP).

	D
Test statistic	0.033
p value	0.00001

Lilliefors (Kolmogorov-Smirnov) normality test: Muestra 3

Shapiro–Wilk Test

Para la muestra 1 no es posible utilizarlo debido a que el tamaño de muestra es mayor a 5 000

Tabla B.7: Shapiro-Wilk normality Test para la Muestra 2 (índice ACP)

	W
Test statistic	0.990
p value	0

Shapiro-Wilk normality test: Muestra 2

Tabla B.8: Shapiro-Wilk normality Test para la Muestra 3 (índice ACP)

	W
Test statistic	0.984
p value	0

Shapiro-Wilk normality test: Muestra 3

Jarque–Bera Test

Tabla B.9: Jarque-Bera Test for normality para la Muestra 1 (índice ACP).

	JB
Test statistic	299.660
p value	0

Jarque-Bera test for normality: Muestra 1

Tabla B.10: Jarque-Bera Test for normality para la Muestra 2 (índice ACP).

	JB
Test statistic	62.720
p value	0

Jarque-Bera test for normality: Muestra2

Tabla B.11: Jarque-Bera Test for normality para la Muestra 3 (índice ACP).

	JB
Test statistic	322.121
p value	0

Jarque-Bera test for normality: Muestra 3

Para las muestras provenientes del índice CI

Kolmogorov-Smirnov Test

Tabla B.12: Kolmogorov-Smirnov Test para la Muestra 1 (índice ACP).

	D
Test statistic	0.03080939
p value	1.539941e-06
Alternative hypothesis	two-sided

One-sample Kolmogorov-Smirnov Test: Muestra 1

Tabla B.13: Kolmogorov-Smirnov Test para la Muestra 2 (índice ACP).

	D
Test statistic	0.06503447
p value	7.349484e-09
Alternative hypothesis	two-sided

One-sample Kolmogorov-Smirnov Test: Muestra 2

Tabla B.14: Kolmogorov-Smirnov Test para la Muestra 3 (índice ACP).

	D
Test statistic	0.05773625
p value	1.231279e-06
Alternative hypothesis	two-sided

One-sample Kolmogorov-Smirnov Test: Muestra 3

Lilliefors Test

Tabla B.15: Lilliefors Test Test para la Muestra 1 (índice ACP).

	D
Test statistic	0.031
p value	0

Lilliefors (Kolmogorov-Smirnov) normality test: Muestra 1

Tabla B.16: Lilliefors Test Test para la Muestra 2 (índice ACP).

	D
Test statistic	0.065
p value	0

Lilliefors (Kolmogorov-Smirnov) normality test: Muestra 2

Tabla B.17: Lilliefors Test Test para la Muestra 3 (índice ACP).

	D
Test statistic	0.058
p value	0

Lilliefors (Kolmogorov-Smirnov) normality test: Muestra 3

Shapiro–Wilk Test

Para la muestra 1 no es posible utilizarlo debido a que el tamaño de muestra es mayor a 5 000

Tabla B.18: Shapiro-Wilk normality Test para la Muestra 2 (índice ACP)

	W
Test statistic	0.963
p value	0

Shapiro-Wilk normality test: Muestra 2

Tabla B.19: Shapiro-Wilk normality Test para la Muestra 3 (índice ACP)

	W
Test statistic	0.964
p value	0

Shapiro-Wilk normality test: Muestra 3

Jarque–Bera Test

Tabla B.20: Jarque-Bera Test for normality para la Muestra 1 (índice ACP).

	JB
Test statistic	2,379.227
p value	0

Jarque-Bera test for normality: Muestra 1

Tabla B.21: Jarque-Bera Test for normality para la Muestra 2 (índice ACP).

	JB
Test statistic	561.366
p value	0

Jarque-Bera test for normality: Muestra2

Tabla B.22: Jarque-Bera Test for normality para la Muestra 3 (índice ACP).

	JB
Test statistic	897.200
p value	0

Jarque-Bera test for normality: Muestra 3

B.2. Accidentes de Trabajo por Actividad y Año.

Según el REGLAMENTO DEL SEGURO GENERAL DE RIESGOS DEL TRABAJO de Resolución Nro. C.D. 513, del Instituto Ecuatoriano de Seguridad Social, un accidente de trabajo es todo suceso imprevisto y repentino que sobrevenga por causa, consecuencia o con ocasión del trabajo originado por la actividad laboral relacionada con el puesto de trabajo, que ocasione en el afiliado lesión corporal o perturbación funcional, una incapacidad, o la muerte inmediata o posterior.

Tabla de Accidentes de Trabajo por Actividad Económica.

Los datos que se presentan en la Tabla B.23 son querys realizados a la base Accidentes de Trabajo proporcionada por el Seguro General de Riesgos del Trabajo, del Instituto Ecuatoriano de Seguridad Social.

Tabla B.23: Accidentes de Trabajo por Actividad Económica para el periodo 2012-2018

Año	R	I	Q	T	N	K	L	M	O	A	G	F	P	B	C	J	S	E	D	H
2012	9	243			127	285	146	22	76	119	361	141	2		34	37		37	560	167
2013	41	355		2	213	416	357	44	156	234	668	309	5	6	46	96	4	67	870	206
2014	51	1075		9	1201	1729	1396	187	712	1333	2829	1655	30	377	253	272	32	440	4049	504
2015	45	1157	2	9	1457	1998	1806	201	930	2151	3052	1835	6	412	308	259	36	448	3997	536
2016	71	1100		4	1531	1756	1894	250	773	2544	2647	1116	3	366	336	262	42	475	3443	367
2017	60	835	9	4	1300	1240	1401	192	480	2145	1993	719	21	321	232	220	57	359	2673	285
2018	58	917	30	8	1529	1205	1482	297	512	1967	2046	588	19	350	201	258	46	338	2891	324

Categorías Individuales de la CIIU

Tabla B.24: Categorías Individuales de la CIIU por sección

Descripción de la Actividad Económica	Sección
AGRICULTURA, GANADERÍA, SILVICULTURA Y PESCA	A
EXPLOTACIÓN DE MINAS Y CANTERAS	B
INDUSTRIAS MANUFACTURERAS	C
SUMINISTRO DE ELECTRICIDAD, GAS, VALOR Y AIRE ACONDICIONADO	D
SUMINISTRO DE AGUA	E
CONSTRUCCIÓN	F
COMERCIO AL POR MAYOR Y AL POR MENOR	G
TRANSPORTE Y ALIMENTO	H
ACTIVIDADES DE ALOJAMIENTO	I
INFORMACIÓN Y COMUNICACIONES	J
ACTIVIDADES FINANCIERAS Y DE SEGUROS	K
ACTIVIDADES INMOBILIARIAS	L
ACTIVIDADES PROFESIONALES, CIENTÍFICAS Y TÉCNICAS	M
ACTIVIDADES DE SERVICIOS ADMINISTRATIVOS	N
ADMINISTRACIÓN PÚBLICA	O
ENSEÑANZA	P
ACTIVIDADES DE ATENCIÓN DE LA SALUD HUMANA	Q
ACTIVIDADES ARTÍSTICAS	R
OTRAS ACTIVIDADES DE SERVICIOS	S
ACTIVIDADES DE LOS HOGARES	T

B.3. Aplicación a una muestra que proviene (por medio de simulación) de una población normal.

Consideremos tres muestras de tamaño 300 que provienen de una población normal con media cero y varianza 1.

Tabla B.25: Muestra 1 de tamaño 300

0.99531	-0.26784	0.12458	-0.58586	1.15350	-1.45920	-1.27196	0.64727	2.66333	-0.40390	1.80960	-0.73589	0.22758	-2.72427	-0.74550
-0.04178	-0.54110	0.24662	-0.90613	-0.95255	0.97708	0.44302	-0.19145	1.27355	-0.24970	2.89974	0.75035	-0.45132	-1.60610	-0.33317
1.55326	-1.36555	0.06844	-0.55398	0.33708	-1.10962	1.10987	-1.39598	-0.21628	0.65146	-0.65848	-0.44929	-0.63796	2.05027	0.37531
-0.14024	-0.21756	-1.49613	0.20430	1.06470	-2.05587	2.12831	0.37288	1.13304	-1.13193	1.29304	-1.81978	1.04697	0.84567	0.00075
-0.45207	0.86605	1.32371	-1.18295	0.22328	-0.35254	-1.57260	0.71300	1.62472	0.70108	-1.60321	-0.02172	-1.14474	-0.70750	0.41322
0.35045	2.01851	0.05004	0.86807	0.38263	-0.28348	0.08178	-0.97838	1.78199	2.44390	-1.74165	1.19733	-0.76338	-2.31614	-0.31002
0.58289	-0.95601	0.06349	-0.11312	0.86045	0.23931	-0.24955	-0.07387	-1.78797	1.41859	-1.42945	0.15344	-1.02876	0.82653	0.14811
0.92675	-1.02515	-0.44181	0.19007	0.72708	-1.44245	-1.26563	1.37120	-0.31482	-1.52296	1.23828	-0.22841	-0.43271	-1.67031	0.95985
-0.70056	-0.61981	0.51471	-0.36495	-1.20119	-0.38969	0.84291	1.10280	1.89989	-0.93410	0.04897	0.11055	1.09161	1.14058	-0.11626
0.06435	-0.26061	-0.45468	0.20054	-0.73153	-0.69405	-0.60145	-0.40834	0.77978	0.23458	0.70574	-0.22231	-0.44167	-0.45683	0.65719
-0.26736	0.65554	0.37665	-0.82399	0.94222	0.01297	1.50754	-0.10377	0.39867	0.31207	1.10245	-0.93377	1.03141	-0.20773	1.30616
0.42539	-0.32339	-1.52971	-0.52286	0.12932	-0.03483	-0.35995	0.22627	1.19744	0.90046	-0.66230	0.40300	-0.97485	0.15651	1.12066
0.64676	0.00336	-1.27666	0.34583	-2.10999	2.45746	0.88865	-1.18811	1.36594	-0.22852	0.28497	0.11321	-1.31952	-1.27678	1.70558
-0.62505	-0.23346	-0.36829	0.31583	-0.13878	-0.30483	0.05473	-0.34273	2.01892	0.63595	-0.10029	-0.12340	-0.65535	-1.27632	-0.06479
-0.08499	0.74600	1.79280	0.88787	0.01901	0.47388	0.49447	0.51112	-0.38768	0.20240	-0.53952	1.40390	-1.06815	-1.44652	2.09472
-1.74495	-1.02942	0.52401	0.34375	-0.10429	1.92342	-0.63026	0.27291	-0.85535	0.33858	0.04124	-1.13850	-0.69165	-1.09821	1.24222
-0.13443	1.53364	-2.00134	-1.69883	0.82597	-1.43940	1.17707	-0.14263	-0.99535	-0.04630	-0.22402	-1.30689	-0.77156	-0.19351	0.31215
1.96040	-0.30975	1.19303	-0.24544	0.28752	-2.15237	-0.08762	0.59721	-0.70024	-2.30309	0.81165	-0.20292	0.87956	-0.73940	0.31217
-0.82562	0.75099	0.51813	0.63265	0.55843	1.20719	-0.89412	0.52525	0.63189	0.89970	1.25323	-1.55274	0.10179	-0.84832	-0.65564
0.04746	-0.81321	-1.16311	0.23719	0.45682	0.77744	-0.35763	-0.92061	-1.87530	-0.05290	0.58804	0.33577	-1.52560	-1.20933	1.01952

Tabla B.26: Muestra 2 de tamaño 300

-0.47358	-0.47917	-0.67022	1.27680	2.01988	-0.89308	0.02221	-0.99491	0.23559	1.38793	2.09644	-0.97752	-0.76450	0.15781	-0.63989
-0.82760	0.68283	1.37292	-1.03632	1.25666	-0.10685	0.01357	-0.65204	-0.72062	-0.94822	-1.23562	-0.50816	1.46453	0.16713	-0.50497
-0.31109	-1.76171	0.33977	1.13619	0.04015	-0.57939	0.92911	1.39565	0.30617	-1.23610	0.63109	1.16465	1.46616	0.99900	0.35695
-0.20438	-1.03129	-0.94598	1.74695	-0.75577	-0.96698	-0.29158	0.53656	1.63011	0.65343	-0.97963	2.02884	0.16158	-0.07757	1.23431
0.54252	-0.51202	-0.34416	-1.22731	0.33461	1.20666	-0.88092	-0.62225	1.61107	0.14974	0.30210	0.58811	-0.48344	1.17548	-0.15018
0.34833	2.21774	-0.01933	-1.21111	-1.02225	-0.09246	1.08251	-0.07909	2.30363	-0.30539	0.51085	0.26777	-0.72015	-1.86376	0.65533
-1.85896	-0.51321	0.04876	0.23426	-1.71417	0.47841	-0.76554	-2.27031	-0.40151	-1.04212	0.00041	0.16472	0.40715	0.49821	-1.67839
-0.76129	-2.22651	1.28194	1.02649	-0.49088	0.01327	-0.30296	1.55095	0.86377	1.52516	1.50534	-0.45119	-1.24926	0.76416	-0.46584
-0.39801	-0.98070	1.50252	-0.88866	0.15450	1.16318	-1.78558	0.79283	-0.05199	-1.69581	-1.71301	-1.00858	0.71652	-3.11816	-0.56206
1.49430	-1.28850	2.20480	0.32831	0.16989	-0.56300	0.98420	-0.42289	0.45718	1.64638	0.40611	0.60177	0.75007	1.22161	-2.22561
1.91097	0.31154	-1.11417	1.08538	-0.74587	1.04637	-0.51029	0.19318	0.04993	1.55818	0.38895	1.28427	-0.51511	1.37234	0.69209
1.86754	-1.10517	-0.19339	-1.30153	-2.62025	-1.02690	0.29308	0.70134	0.65317	-0.07902	-0.51016	-1.38820	-1.06694	0.18709	-1.27450
0.69364	-0.74973	-0.35941	-2.00056	-0.52307	-0.67909	-1.58354	0.09386	0.52572	0.01019	-0.47562	0.41584	-0.56811	-0.29465	-0.85546
0.50005	1.48100	0.36733	-1.48356	-0.91038	0.23917	1.71832	0.56348	0.59684	0.95198	-0.79939	-0.59620	-0.13184	0.51703	0.09340
0.36024	0.21167	0.77817	1.53647	-0.66223	-0.32895	0.42094	-1.47715	-0.50248	-0.81606	-0.78050	0.87265	0.69361	-0.38199	1.02394
-0.08985	2.48061	-0.31215	0.96155	-0.33226	0.03263	-0.81800	1.39994	-1.64171	0.06293	-0.29553	0.08088	0.79904	-0.76409	0.13806
1.23352	0.01749	0.36117	0.38844	1.37880	-0.67686	-0.59658	1.06283	2.14739	-1.17036	-0.14896	-0.98540	-0.09094	0.10558	-0.40228
1.89225	0.71675	-0.30327	0.12341	-0.31890	-0.24678	-1.04984	1.96675	-0.78540	-0.93179	-0.25087	-2.11115	0.25079	-0.28813	-0.63951
0.77055	0.41662	0.33753	-0.78195	-0.58018	-0.39522	-1.00196	-1.38591	0.62318	-1.52853	0.37194	0.56275	0.31100	0.42802	-0.72464
0.67128	1.23380	0.72504	-0.39130	-0.87875	-0.07948	1.93495	-1.35531	0.03384	-0.38362	1.96388	-1.20670	-0.18575	1.13885	-0.79396

Tabla B.27: Muestra 3 de tamaño 300

2.05889	-0.72775	-1.02770	-1.14645	-0.30368	0.57685	0.69264	-0.32481	-0.91476	0.17626	-0.60564	-0.24753	0.93682	-0.95921	-0.65415
0.85502	1.05618	-0.09321	-2.23015	-1.19303	-0.81052	-0.19159	-0.51276	-0.74894	-1.11789	-0.35780	-0.37141	1.88660	0.54245	-0.72691
1.72965	-0.91796	1.11495	0.01134	-0.56282	0.31412	0.75749	-1.65166	1.62342	0.46523	0.70096	-0.74428	0.38520	0.06413	2.41086
0.46153	-0.65750	-0.94841	0.65226	-0.20803	-0.76685	-0.75325	-2.37581	-1.27938	-0.35199	3.02533	0.23008	0.78003	-0.14120	0.81711
-0.04262	-0.51469	-1.01211	-0.34202	-1.90538	0.45488	0.68228	0.37264	0.62708	-0.54532	0.70337	-0.43291	-0.19967	1.47369	1.43636
1.56791	-0.48615	1.62962	-0.52061	0.29273	0.89955	1.35448	0.73735	0.77307	1.38668	0.43728	0.40857	0.64088	0.96021	-0.76238
-1.74067	2.32777	-0.21107	0.55497	1.75959	1.08076	-0.62941	0.69355	0.07334	0.01150	-1.76472	-0.58309	0.32837	-0.77056	-0.24136
-1.18443	0.48632	-0.47935	-0.71755	0.70283	-0.68484	-1.00336	-0.53046	-2.92039	-0.75027	1.18305	-0.63337	1.03915	-1.70970	0.27602
-0.68074	-0.59804	0.34196	-0.70562	-0.06660	1.03696	-0.29816	0.74384	-1.39124	0.31695	1.60949	0.15174	-1.41449	1.13890	0.17085
-1.04106	1.53833	-0.31192	-1.06462	2.06539	0.52205	-0.01210	-1.59568	0.48619	0.00022	-0.03785	0.79652	-0.56071	0.31240	1.55513
0.15981	-2.04112	-0.32834	0.47763	1.84705	-1.35551	2.17232	-0.92201	-0.97845	0.09482	-0.29346	1.04022	1.39331	1.03570	2.85132
1.71595	0.47861	1.14658	1.47262	0.40215	-0.22012	-1.25447	0.04959	-0.57308	1.03301	0.24804	1.06634	-1.04837	-0.74452	-1.21539
-0.27357	0.57413	-1.30159	2.22573	-1.93308	-0.82901	-0.16664	1.39837	-0.81319	-1.72318	-1.16934	-0.31767	-0.99952	0.62583	-0.94237
-0.71491	-1.01617	-2.17173	0.57537	0.51266	-1.26436	2.09734	-0.11646	2.57296	0.36022	0.28623	0.49398	-0.74789	1.46748	0.53827
0.74388	0.56335	0.00951	-0.64941	0.77643	0.39279	-1.27982	0.25757	0.91417	-0.86137	-1.25284	-0.27421	-0.29135	0.65075	0.14471
2.91068	-1.11163	-0.39097	0.54236	1.19215	0.41289	0.13782	-1.70710	1.14525	-0.38016	-1.37107	-1.53979	1.11457	-1.79139	-0.46420
0.96838	-0.16761	2.29985	1.50888	-0.20958	-0.89708	1.23079	-0.47479	-0.14600	-1.70518	-0.25791	1.13399	-1.38354	0.26723	-0.46151
0.65086	-0.89520	1.21576	1.78280	-1.07033	0.63238	-1.01560	-0.01609	-0.21557	0.87008	-0.37872	0.40424	2.32919	1.04598	1.34720
0.79359	0.11881	0.66242	0.38441	-0.03013	1.11250	0.37259	-0.35965	-0.11351	-1.08999	-0.35778	-0.91314	-1.36631	0.61251	-1.14998
0.96700	-0.22230	0.71364	-0.01580	-1.24631	-0.69637	-0.99669	-0.24169	-1.47741	0.76065	0.50739	-0.35957	0.86076	0.17154	0.69828

Cada muestra, usando el test Kolmogorov-Smirnov, no rechaza la hipótesis nula de que las muestras provienen de poblaciones normales. Esto se puede apreciar en las Tablas B.28, B.29 y B.30.

Tabla B.28: Kolmogorov-Smirnov Test para la Muestra 1 (población normal).

	D
Test statistic	0.03899136
p value	0.7517802
Alternative hypothesis	two-sided

One-sample Kolmogorov-Smirnov Test: Muestra 1

Tabla B.29: Kolmogorov-Smirnov Test para la Muestra 2 (población normal).

	D
Test statistic	0.04935654
p value	0.457936
Alternative hypothesis	two-sided

One-sample Kolmogorov-Smirnov Test: Muestra 2

Tabla B.30: Kolmogorov-Smirnov Test para la Muestra 3 (población normal).

	D
Test statistic	0.04419858
p value	0.6010767
Alternative hypothesis	two-sided

One-sample Kolmogorov-Smirnov Test: Muestra 3

Adicionalmente, al realizar la prueba de igualdad de varianzas poblacionales, tenemos:

Tabla B.31: Levene's Test for Homogeneity of Variance (muestras normales).

	1
Df	2
F value	0.716
Pr(>F)	0.489

Como el p-valor de la Tabla B.31 es mayor a 0.05, entonces no se puede rechazar la hipótesis nula de igualdad de varianzas.

De todo lo anterior, se verifica el cumplimiento de los supuestos fundamentales del Test Anova. Su aplicación a estas muestras nos da el siguiente resultado:

Tabla B.32: Análisis de Varianza de un factor.

Fuente	Grados de Libertad	Suma de Cuadrados	Cuadrado Medio	valor F	Pr(>F)
Entre Grupos	2	0.75773	0.37886	0.35800	0.69917
Error(dentro de los grupos)	897	949.27018	1.05827		

Ahora bien, veremos que pasa desde el punto de vista no paramétrico mediante la aplicación del Test Kruskal-Wallis. Esto se puede apreciar en la Tabla B.33

Tabla B.33: Test de Kruskal-Wallis (muestras normales).

	Kruskal-Wallis chi-squared
Statistic	0.1948647
Df	2
p.value	0.9071637

Kruskal-Wallis rank sum Test

En este sentido, dado que el Test Anova si verifica el cumplimiento de los supuestos fundamentales de su formación, entonces el Test Anova y el Test de Kruskal-Wallis son comparables.

Se aprecia claramente que ambos test no rechazan la hipótesis nula de que las muestras provienen de la misma población a nivel de ubicación. Sin embargo, este ejemplo muestra la utilidad del test Kruskal- Wallis, pues solo requiere del cumplimiento del supuesto que las muestras hayan sido extraídas de poblaciones que tienen su cdf continua; y, evita la tarea de verificar los supuestos necesarios para la aplicación test paramétrico Anova.

ANEXO C

Anexo III: Enlace Web

C.1. Dirección web del aplicativo.

En el enlace web que se presenta más adelante, el lector podrá interactuar con todo el proceso práctico del presente trabajo de titulación. Podrá no solo ver los resultados ya expuestos en este estudio, sino también realizar variaciones a cada criterio considerado. Por ejemplo, no necesariamente puede limitarse al 10% como límite máximo de datos perdidos presentes en las variable, o considerar que las muestras sean las proyecciones de los individuos en la primera componente principal para las actividades económicas más numerosas. El enlace ofrece la libertad de manipular cada criterio según se considere adecuado.
Adicionalmente, dentro del proceso de comparaciones múltiples, la elección no se limita al método de *holm* y *bonferroni*, todo lo contrario, se ofrecen ocho diferentes métodos.

Enlace
https://cristian-guatemal-work.shinyapps.io/TesisCG/