



Universidad Veracruzana

Facultad de Estadística e Informática

Región Xalapa

Especialización en Métodos Estadísticos

Aspectos esenciales en el planteamiento y uso del sistema de familias de distribuciones de Johnson para transformar datos poblacionales a una distribución normal

Desarrollo estadístico
para obtener el diploma de Especialista en
Métodos Estadísticos

Presenta:

Jorge Luis Hernández Mortera

Director:

Dr. Ernesto Pedro Menéndez Acuña

Co-director:

M.C. Jesús Hernández Suárez

Junio de 2022

“Lis de Veracruz: Arte, Ciencia, Luz”



Universidad Veracruzana

Facultad de Estadística e Informática
Región Xalapa

Especialización en Métodos Estadísticos

Aspectos esenciales en el planteamiento y uso del
sistema de familias de distribuciones de Johnson
para transformar datos poblacionales a una
distribución normal

Reporte para obtener el diploma de Especialista en
Métodos Estadísticos

Presenta:
Jorge Luis Hernández Mortera

Director :
Dr. Ernesto Pedro Menéndez Acuña
Co-director :
M.C. Jesús Hernández Suárez

UNIVERSIDAD VERACRUZANA

Facultad de Estadística e Informática

El Comité Académico de la Especialización en Métodos Estadísticos y el director de este trabajo recepcional intitulado: Aspectos esenciales en el planteamiento y uso del sistema de familias de distribuciones de Johnson para transformar datos poblacionales a una distribución normal, autorizan la impresión y la constitución del jurado para la defensa

COMITÉ ACADÉMICO



Dra. Julia Aurora Montano Rivas

Coordinadora

Especialización en Métodos Estadísticos



Dr. Luis Gerardo Montané Jiménez

Director

Facultad de Estadística e Informática



M.C. Jesús Hernández Suárez

Vocal



M.C. María Yesenia Zavaleta

Sánchez

Vocal



Dr. Ernesto Pedro Menéndez Acuña

Director del Trabajo Recepcional

Índice

Introducción	4
Metodología	5
Descripción y análisis de datos a ser transformados	6
Comprobación de normalidad en el conjunto de datos	7
Cálculos para la selección de ecuación.	8
Comprobación de idoneidad de parámetros de ecuaciones y transformación de datos.....	11
Comprobación de normalidad de datos calculados.....	13
Cálculo de valor estadístico para selección de ecuación	16
Reporte y análisis de resultados	16
Conclusiones y recomendaciones	21
Referencias.....	22

Resumen

La estadística proporciona técnicas y enfoques que permiten el análisis de datos utilizados en la toma de decisiones en las organizaciones. Estas técnicas van desde aquellas contempladas en la estadística descriptiva hasta las consideradas en la estadística inferencial y son utilizadas para conocer y tomar decisiones oportunas y fiables. Cuando existe la necesidad de obtener y analizar datos constitutivos de una variable aleatoria cuantitativa continua, un buen número de las pruebas estadísticas inferenciales de estimación y pruebas de hipótesis empleadas se basan en considerar que se ha obtenido una muestra aleatoria de una distribución de probabilidad normal. Sin embargo, en ocasiones, esta consideración no es correcta y en otras, la consideración de que provengan de una distribución de probabilidad normal no es fácil de comprobar, por tratarse de muestras pequeñas. Para ambas situaciones, la estadística puede ayudar con el uso de técnicas que transformen los datos originales a una distribución de probabilidad normal en una primera instancia. Este trabajo presenta el proceso de transformación de Johnson que permite la transformación de datos de poblaciones no normales. Se presentan los aspectos conceptuales del proceso y la propuesta del uso de una hoja electrónica de cálculo a efecto de lograr una mayor comprensión que pueda ser utilizada en el aprendizaje de la técnica en los ámbitos académicos y de práctica estadística.

Palabras clave: Distribución Normal, percentiles, transformación Box Cox, transformación Johnson.

Introducción

Un supuesto muy utilizado, en particular en las técnicas de estadística, es que el comportamiento estadístico de los datos siga una distribución normal, de donde proviene el término principio de normalidad. Por ello, las técnicas recomiendan dentro de su metodología realizar pruebas a efecto de comprobar este supuesto de normalidad de los datos. No hacerlo, puede conducir a declarar conclusiones que pueden estar fuera de la situación real que se pretende medir que, en el caso de procesos de producción, pudieran bien sobreestimar o subestimar la calidad de los productos manufacturados y tener consecuencias negativas en la satisfacción del cliente reflejado en la disminución de ventas debido al rechazo de los productos que no cumplen las características de calidad.

Cuando el supuesto de normalidad en el comportamiento estadístico de la variable en estudio no se cumple, la estadística proporciona también técnicas que pueden utilizarse a efecto de seguir utilizando los modelos establecidos, transformando los datos con comportamiento no-normal a datos que si lo hagan. Así, técnicas tales como la transformación Box-Cox o la utilización del sistema de familias de distribución de Johnson pueden utilizarse para tal efecto. En la actualidad, algunos sistemas computacionales incluyen estas técnicas.

Johnson (1949) propone un método a efecto de poder usar datos que provienen de poblaciones con comportamiento no normal bajo el método de las así llamadas familias de distribuciones de Johnson. Para ello, determina que dada una variable aleatoria continua X , cuya distribución es desconocida y la cual se quiere aproximar, Johnson propone 3 transformaciones normalizadas teniendo la siguiente forma general:

$$Z = \gamma + \eta f\left(\frac{X - \epsilon}{\lambda}\right) \quad (1)$$

Donde $f(\cdot)$ denota la función de transformación, Z es una variable aleatoria normal estándar, γ y η son parámetros de forma, λ parámetro de escala y ϵ parámetro de localización. De manera general, se asume que $\eta > 0$ y $\lambda > 0$.

La primera transformación propuesta por Johnson define a la familia de distribuciones log normal, se denota por SL y está representada de la siguiente manera:

$$Z = \gamma + \eta \ln\left(\frac{X - \epsilon}{\lambda}\right), X > \epsilon \quad (2)$$

El sistema de familias de distribuciones acotadas SB se define por:

$$Z = \gamma + \eta \ln\left(\frac{X - \epsilon}{\lambda + \epsilon - X}\right), \quad \epsilon < X < \epsilon + \lambda \quad (3)$$

El sistema de familias de distribuciones no acotadas SU se define por:

$$Z = \gamma + \eta \ln\left[\left(\frac{X - \epsilon}{\lambda}\right) + \left\{\left(\frac{X - \epsilon}{\lambda}\right)^2 + 1\right\}^{\frac{1}{2}}\right], \quad -\infty < X < \infty \quad (4)$$

Metodología

Slifker y Shapiro (1980) desarrollaron una serie de criterios para ayudar a seleccionar una distribución Johnson a la transformación de datos. Chou et al. (1998) llevan más allá su método y proponen el enfoque más utilizado en la actualidad para calcular los parámetros de las ecuaciones basado en el cálculo de percentiles a partir de los datos originales. A partir de ello, se ha creído pertinente agrupar los pasos sugeridos en forma de diagrama de bloques a efecto de una mejor visualización general, que permita aplicarlo en cualquier caso que se explore la posibilidad de la transformación de los datos originales a fin de cumplir con los supuestos de normalidad de las técnicas estadísticas que así lo requieran. Este diagrama de bloques se muestra en la Figura 1, con base en este diagrama y de los datos originales se describe paso a paso para explicar el proceso de transformación.

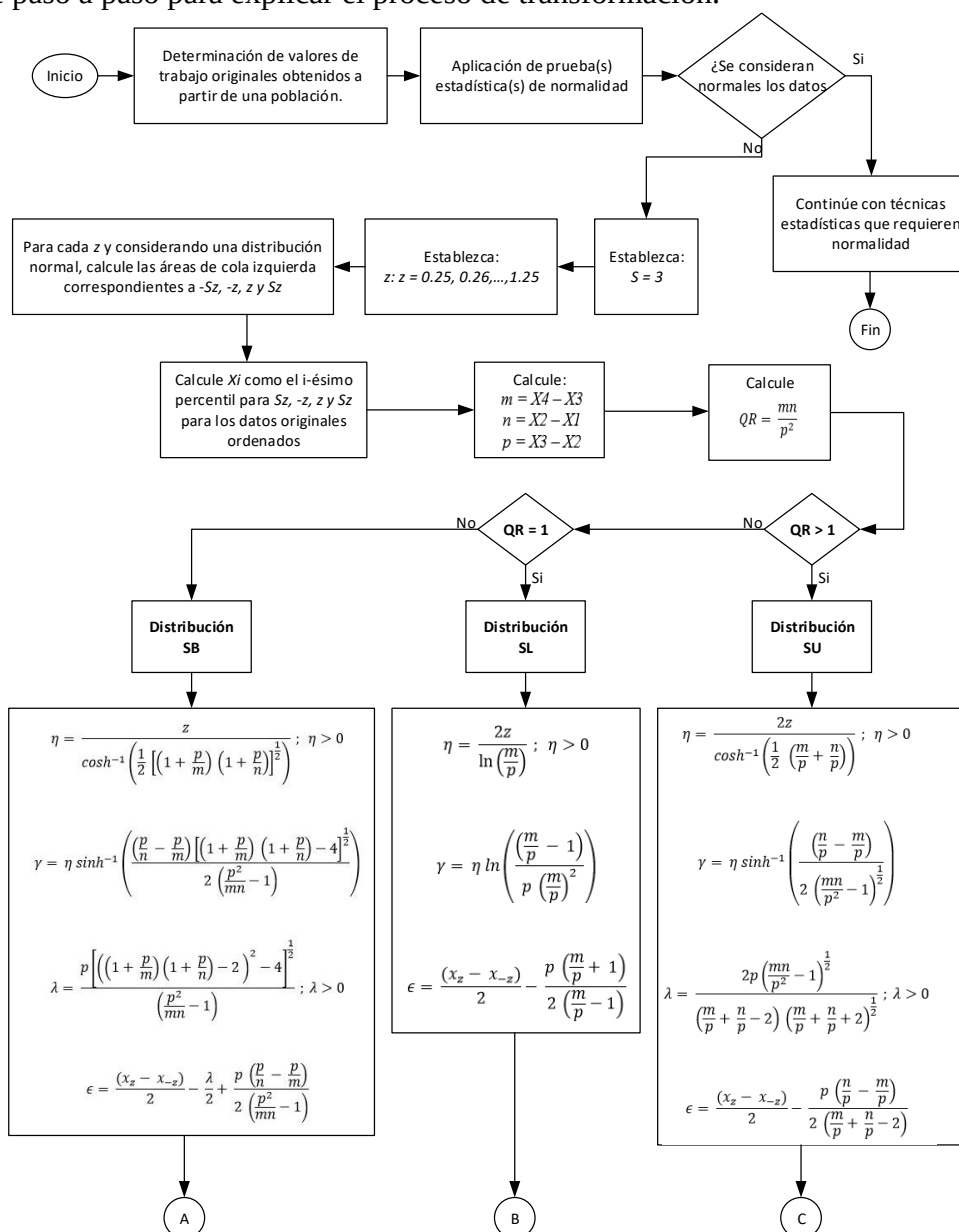


Figura 1. Diagrama de bloques de proceso de transformación de Johnson.
Fuente: elaboración propia

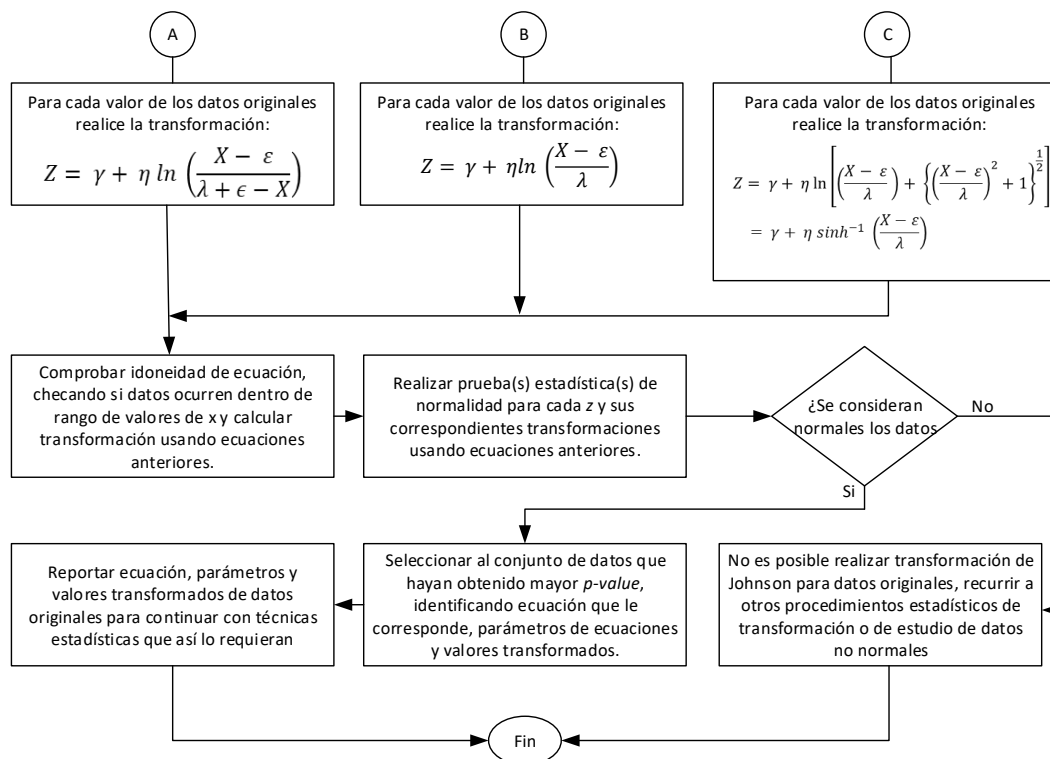


Figura 1 (cont.) Diagrama de bloques de proceso de transformación de Johnson.
Fuente: elaboración propia

Descripción y análisis de datos a ser transformados

Los datos a utilizar para describir el procedimiento de transformación se obtuvieron de una corrida de simulación a partir de una distribución no normal. Estos datos se muestran en la Tabla 1.

8.345	3.8754	2.8293	10.0682	3.0415	9.757	17.9816	8.8023	19.2514	0.6495
4.1845	3.1382	0.7861	2.6948	1.4105	6.5472	18.562	33.2815	5.7468	3.6426
6.4831	0.4288	4.0376	7.9375	2.8618	5.4818	23.0995	4.9505	14.9848	0.4954
15.562	10.0596	20.0732	16.086	7.9985	8.7752	8.2518	3.2451	3.2875	1.9043
2.8342	13.3593	13.6887	5.4666	4.1306	1.3208	14.0508	33.951	5.7251	12.5029
4.4225	4.1442	8.41	1.6227	1.8936	6.9811	3.2588	0.8391	2.1278	0.9814
4.1196	9.5129	4.3212	1.5754	6.5708	3.4719	3.1791	8.2204	7.4249	32.3065
6.3356	15.2977	2.9021	17.9454	1.6921	6.1445	5.4295	8.6401	1.2493	10.9183
4.9297	0.4265	1.8908	1.0595	9.108	0.3936	9.222	25.6693	4.4289	1.2951
2.7622	11.8491	15.1553	7.198	20.795	32.9194	8.7181	3.4745	27.9004	22.105

Un análisis estadístico inicial en RStudio (v2022.02.3+492) de este conjunto de datos puede observarse en la Figura 2, en cuya última columna se ilustra un histograma que muestra que no podría aseverarse el comportamiento normal de los datos.

```
> skim(datos)
-- Data Summary -----
Name      values
Number of rows      datos
Number of columns    100

Column type frequency:
numeric          1

Group variables    None

-- Variable type: numeric -----
# A tibble: 1 x 11
  skim_variable n_missing complete_rate mean    sd    p0    p25    p50    p75    p100 hist
*   <chr>          <int>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 Datos            0           1  8.41  7.98 0.394  2.89  5.74 10.3  34.0  ██████
```

Figura 2. Resumen estadístico de datos originales usando R.

Esta primera consideración puede verse más claramente en el histograma y la gráfica de la función de densidad obtenidos con el paquete *ggplot* versión 3.3.2 (Wickham, 2016) de R versión 3.6.3 (Core Team, 2020), mismo que se muestra en la Figura 3. Claramente se observa que el comportamiento presenta un sesgo a la derecha, alejado de un comportamiento normal.

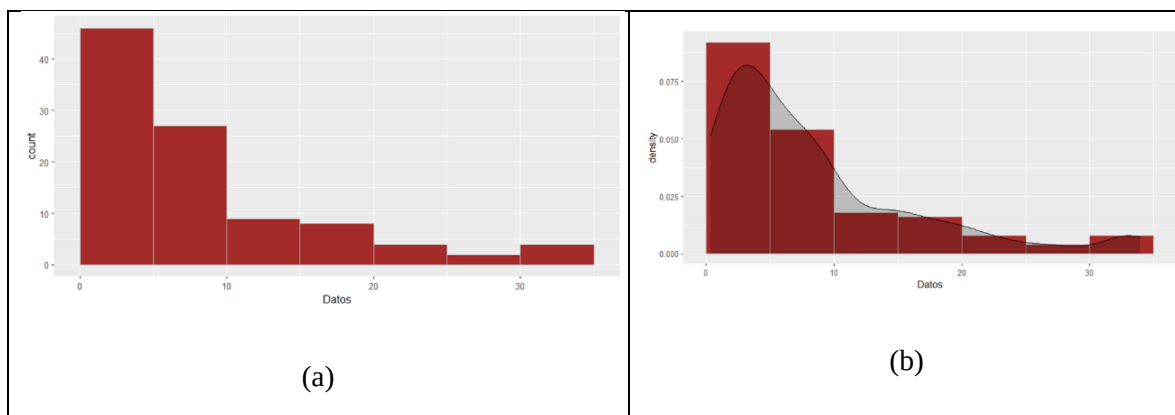


Figura 3. (a) Histograma y (b) función de densidad de datos originales usando R.

Comprobación de normalidad en el conjunto de datos

Para la comprobación inicial de la normalidad de los datos originales pueden utilizarse pruebas estadísticas. Para realizar esta comprobación, se proponen dos hipótesis, nula y alternativa como se muestra a continuación:

H_0 : los datos provienen de una distribución normal

H_a : los datos no provienen de una distribución normal

Para este trabajo en particular se utilizó la prueba de Anderson Darling (AD) propuesta en 1952 por Theodore Anderson y Donald Darling (Stephens, 1974). La prueba puede utilizarse como un procedimiento estadístico para determinar si un conjunto de datos proviene de una distribución de probabilidad dada. Se presenta una captura de pantalla del procedimiento en la hoja de cálculo propuesta en la Figura 4. Los datos originales se observan en la columna B, la columna C corresponde a la serie de números consecutivos de uno al total de datos en la muestra (en el ejemplo, 100 datos) a efecto de establecer la base de ordenación que se

muestra en la columna D. En las columnas H a K se aplica el procedimiento AD y los cálculos finales se resumen en las celdas F4 a G14.

El *p-value* calculado y mostrado en la celda G10 es el menor de las probabilidades calculadas en las celdas G11 a G14 acorde con el procedimiento AD. Este valor (1.48748E-13) es menor a un nivel de significancia de 0.05, por ejemplo. Por ello, se tiene evidencia estadística para rechazar la hipótesis nula planteada de que los datos provienen de una población con comportamiento estadístico de distribución normal.

	A	B	C	D	E	F	G	H	I	J	K
1											
2		Datos	Consecutivo	Datos Orig.		Prueba de Anderson					
3		Originales		ordenados		Darling		F1i	1-F1i	F2i	Si
4		8.345	1	0.3936		N=	100	0.15770043	0.84229957	0.00068861	-9.12788746
5		4.1845	2	0.4265		MediaMuestral	8.408683	0.1586957	0.8413043	0.0009181	-26.5019097
6		6.4831	3	0.4288		DesvEstMuestral	7.983517373	0.15876543	0.84123457	0.0010697	-43.4035126
7		15.562	4	0.4954		Suma (Si)	-10545.6643	0.16079334	0.83920666	0.00137946	-58.8958675
8		2.8342	5	0.6495		AD test statistic	5.456643321	0.16555007	0.83444993	0.0073133	-60.4488835
9		4.4225	6	0.7861		AD* test statistic	5.498795891	0.16984194	0.83015806	0.01530787	-65.4750281
10		4.1196	7	0.8391		P-value	1.48748E-13	0.1715262	0.8284738	0.03287359	-67.3153636
11		6.3356	8	0.9814		p1	1.48748E-13	0.17610086	0.82389914	0.04312	-73.2070025
12		4.9297	9	1.0595		p2	0	0.17864415	0.82135585	0.06039261	-76.997217
13		2.7622	10	1.2493		p3	0	0.18492069	0.81507931	0.07199744	-82.0601058
14		3.8754	11	1.2951		p4	0	0.18645555	0.81354445	0.08720998	-86.4989785
15		3.1382	12	1.3208				0.18732025	0.81267975	0.10172477	-91.0896587
16		0.4288	13	1.4105				0.1903577	0.8096423	0.11524742	-95.4881081

Figura 4. Pruebas de normalidad a datos originales usando la hoja de cálculo propuesta.

Adicionalmente, se realizaron cuatro pruebas de normalidad usando el paquete *normtest* versión 1.1 (Gavrilov, 2014) de RStudio (v2022.02.3+492) que realiza cuatro pruebas de normalidad que pueden observarse en Stephens (1974). Los resultados de estas pruebas se observan en la Figura 5 y puede apreciarse en cada una de ellas el *p-value* menor a un nivel de significancia de 0.05, deduciéndose que se rechaza la hipótesis nula de que los datos se adecuan a una distribución normal.

Anderson-Darling normality test data: Datos A = 5.4566, p-value = 1.487e-13	Cramer-von Mises normality test data: Datos W = 0.93549, p-value = 2.674e-09
Lilliefors (Kolmogorov-Smirnov) normality test data: Datos D = 0.17034, p-value = 1.449e-07	Shapiro-Francia normality test data: Datos W = 0.82494, p-value = 2.233e-08

Figura 5. Pruebas de normalidad a datos originales usando R.

Cálculos para la selección de ecuación.

Dado que el paso anterior ha demostrado que los datos originales no siguen una distribución normal se continua con el procedimiento de Chou et al. (1998) en el paso 2 que considera una serie de cálculos para determinar la ecuación de Johnson más adecuada.

En primer término, se usa un valor de *S* igual a 3, toda vez que de acuerdo con Chou (1998) permite realizar la discriminación entre las ecuaciones de las familias de Johnson. Bowman y Shenton (1988) también justifican el uso de *S* = 3 para el cálculo de los parámetros de las

ecuaciones de Johnson. Este valor de S permite particionar el eje del valor de z a efecto de calcular los percentiles como se mostrará más adelante.

Una vez determinado S , se definen los valores de z a ser utilizados en la transformación. Estos valores de z y los subsecuentes cálculos de p , m y n pueden usarse para obtener los estimadores de los parámetros de las ecuaciones de Johnson. En términos generales Slifker y Shapiro (1980) recomiendan valores de $z > 0$ toda vez que permite el cálculo de los puntos percentiles presentes en el procedimiento. Aunque la selección de estos valores puede ser arbitrarios, Slifker y Shapiro (1980) y Bowman y Shenton (1988) seleccionan cuatro desviaciones simétricas de la distribución normal, $-Sz$, $-z$, z y Sz , donde S se selecciona como se estableció en el párrafo anterior y z son constantes arbitrarias positivas. Chou et al. (1998) en su procedimiento, recomienda tomar valores de z que van de 0.25 a 1.25, con incrementos de 0.01, a efecto de tener 101 valores de z , siendo este el criterio más comúnmente usado en los sistemas computacionales. Con ello, se determinan los percentiles asociados a cada valor de Z (q_1 , q_2 , q_3 y q_4), calculándose el área correspondiente a la función de distribución de probabilidad (fdp) normal para estos percentiles en forma de porcentaje (X_1 , X_2 , X_3 y X_4). Estos valores forman la base de los cálculos de los parámetros m , n y p necesarios para la determinación del valor de QR que determinará la familia propuesta por Johnson.

Por ello, este paso puede ser resumido en los siguientes términos:

Para cada z en $S = \{z : z = 0.25, 0.26, \dots, 1.25\}$ calcule:

$$QR = \frac{mn}{p^2} ; \quad \text{donde: } m = (X_4 - X_3) ; n = (X_2 - X_1) ; p = (X_3 - X_2)$$

Si QR es menor que 1, puede usarse SB a los datos. Si QR es igual a uno, debe usarse SL . En tanto que si QR es menor que uno use SU para encontrar la transformación.

En la Figura 6 se muestra la captura de pantalla de la hoja de cálculo para los datos de ejemplo mostrándose los criterios citados.

	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1														
2	S =	3												
3														
4	Z	0.25	0.26	0.27	0.28	0.29	0.3	0.31	0.32	0.33	0.34	0.35	0.36	0.37
5	-SZ	-0.75	-0.78	-0.81	-0.84	-0.87	-0.9	-0.93	-0.96	-0.99	-1.02	-1.05	-1.08	-1.11
6	-Z	-0.25	-0.26	-0.27	-0.28	-0.29	-0.3	-0.31	-0.32	-0.33	-0.34	-0.35	-0.36	-0.37
7	Z	0.25	0.26	0.27	0.28	0.29	0.3	0.31	0.32	0.33	0.34	0.35	0.36	0.37
8	SZ	0.75	0.78	0.81	0.84	0.87	0.9	0.93	0.96	0.99	1.02	1.05	1.08	1.11
9														
10	q1	0.22662735	0.21769544	0.20897009	0.20045419	0.1921502	0.18406013	0.17618554	0.16852761	0.16108706	0.15386423	0.14685906	0.14007109	0.13349951
11	q2	0.40129367	0.39743189	0.39358013	0.38973875	0.38590812	0.38208858	0.37828048	0.37448417	0.37069998	0.36692826	0.36316935	0.35942357	0.35569125
12	q3	0.59870633	0.60256811	0.60641987	0.61026125	0.61409188	0.61791142	0.62171952	0.62551583	0.62930002	0.63307174	0.63683065	0.64057643	0.64430875
13	q4	0.77337265	0.78230456	0.79102991	0.79954581	0.8078498	0.81593987	0.82381446	0.83147239	0.83891294	0.84613577	0.85314094	0.85992891	0.86650049
14														
15	100*q1	22.6627352	21.7695438	20.8970088	20.0454193	19.2150202	18.4060125	17.6185542	16.8527607	16.108706	15.386423	14.6859056	14.007109	13.3499513
16	X1=Percentil	2.83143693	2.79922902	2.74117381	2.6689523	2.14076729	1.95390636	1.89833334	1.89271585	1.88039187	1.73830943	1.66010983	1.61641089	1.58563817
17														
18	100*q2	40.1293674	39.7431887	39.3580127	38.9738752	38.5908119	38.2088578	37.8280478	37.4484165	37.0699981	36.6928264	36.3169349	35.9423567	35.5691245
19	X2=Percentil	4.17354137	4.158134	4.14371628	4.13854426	4.13338669	4.12869446	4.12454744	4.12041326	4.09494245	4.06432365	4.03010077	3.96995175	3.91001888
20														
21	100*q3	59.8706326	60.2568113	60.6419873	61.0261248	61.4091881	61.7911422	62.1719522	62.5515835	62.9300019	63.3071736	63.6830651	64.0576433	64.4308755
22	X3=Percentil	7.9540875	7.97740884	8.00639242	8.09078011	8.17493186	8.22583945	8.23767731	8.24947852	8.27982541	8.31462629	8.34800524	8.37210935	8.39612684
23														
24	100*q4	77.3372648	78.2304562	79.1029912	79.9545807	80.7849798	81.5939875	82.3814458	83.1472393	83.891294	84.613577	85.3140944	85.992891	86.6500487
25	X4=Percentil	12.2177727	12.8866971	13.4620601	13.7448381	14.0425188	14.7774964	15.0798761	15.2002652	15.3115443	15.5005347	15.8035396	16.3332297	17.5429295
26														
27	m=X _U	4.26368516	4.90928826	5.45566764	5.65405801	5.86758691	6.551657	6.84219883	6.95078668	7.03171891	7.18590841	7.45553435	7.96112035	9.14680267
28	n=X _L	1.34210444	1.35890498	1.40254247	1.53164903	1.9926194	2.1747881	2.2262141	2.2276974	2.21455057	2.32601421	2.36999094	2.35354086	2.32438071
29	p=X _M	3.78054613	3.81927484	3.86267613	3.95223586	4.04154516	4.09714499	4.11312986	4.12906527	4.18488297	4.25030265	4.31790447	4.4021576	4.48610796
30														
31	QR	0.40037077	0.45734696	0.51284643	0.55441352	0.71579558	0.84880013	0.90036328	0.9082108	0.88916113	0.92523985	0.94771859	0.96686259	1.05642125
32														
33	Decisión	SB	SB	SB	SB	SB	SB	SB	SB	SB	SB	SB	SB	SU

Figura 6. Proceso de cálculo de QR

Se aprecia en la Figura 6 el establecimiento de $S = 3$ en la celda E2. Los valores de z de 0.25 a 1.25 se establecen en el rango de celdas E4 a DA4 con un incremento de 0.01. Para cada valor de z se establece $-Sz$, $-z$, z y Sz . Así, para $z = 0.25$ los valores calculados son $[-0.75, -0.25, 0.25$ y $0.75]$ respectivamente, mismos que se muestran en las celdas E5 a E8.

A continuación, se calculan los percentiles correspondientes a estos valores tomando a la distribución normal a efecto de calcular el área como puede observarse en la Figura 1. En la celda E₁₀ de la Figura 2 (a menos que se indique lo contrario, los cálculos siguientes estarán basados en ella) se establece el cálculo usando la función de Excel: $[=DISTR.NORM.ESTAND.N(E5,1)]$. El valor resultante (0.22662735) correspondería al área bajo la curva de la distribución normal estándar para z de $-\infty$ a -0.75 . De manera similar, 0.40129367 sería el área de $-\infty$ a -0.25 , 0.59870633 el área de $-\infty$ a 0.25 y 0.77337265 el área de $-\infty$ a 0.75 . Las celdas E15, E18, E21 y E24 reflejan estos valores en puntos porcentuales a efecto de encontrar los percentiles que corresponden a los datos originales ordenados. Estos valores se han denominado $X1$, $X2$, $X3$ y $X4$ en la hoja de cálculo y son los que forman la base de los parámetros m , n y p que permitirán el cálculo de los parámetros de las ecuaciones de Johnson. Así entonces, $X1$ en la celda E16 se calcula en Excel a través del cálculo: $[=PERCENTIL(\$D\$53:\$D\$152,E10)]$, donde en el rango de celdas $[\$D\$53:\$D\$152]$ se encuentran los datos ordenados originales como puede observarse en la Figura 2. En tanto, en la celda E10 se encuentra el percentil correspondiente a -0.75 ($-Sz$). La interpretación del valor calculado, 2.83143693, equivaldría a establecer que debajo de ese valor se esperaría encontrar el 22.62% de los valores totales de los datos. De manera similar se calcularían e interpretarían los valores de $X2$, $X3$ y $X4$ que se encuentran en las celdas E19, E22 y E25 respectivamente, encontrándose que a la izquierda de 4.17354137 se encuentra el 40.13% de los datos; a la izquierda de 7.9540875 se posiciona el 59.87% de los datos y a la izquierda de 12.2177727 se ubicaría el 77.34%.

En los mismos términos se interpretan los resultados para el conjunto de valores de $\{z : z = 0.25, 0.26, \dots, 1.25\}$.

Se procede a continuación a calcular los valores de m , n y p de acuerdo a lo establecido en el procedimiento, mismas que se muestran a continuación:

$$m = X4 - X3$$

$$n = X2 - X1$$

$$p = X3 - X2$$

Para el caso que se ha estado describiendo, el cálculo de estos valores está dado por:

$$m = X4 - X3 = 12.2177727 - 7.9540875 = 4.26368516$$

$$n = X2 - X1 = 4.17354137 - 2.83143693 = 1.34210444$$

$$p = X3 - X2 = 7.9540875 - 4.17354137 = 3.78054613$$

Estos valores se muestran en las celdas E27, E28 y E29 respectivamente.

Una vez calculados estos valores, se puede calcular ahora el valor QR . Así, en el primer caso, el QR equivaldría a:

$$QR = \frac{mn}{p^2} = \frac{(4.26368516)(1.34210444)}{(3.78054613)^2} = 0.40037077$$

Este valor se encuentra en la celda E31 de la hoja de cálculo propuesta. El resto de valores de QR para cada valor z se calcula de la misma manera.

Como se estableció en el capítulo anterior, si QR es menor que 1, puede usarse SB a los datos. Si QR es igual a uno, debe usarse SL . En tanto que, si QR es mayor que uno, use SU para encontrar la transformación. Así el valor de QR calculado de 0.40037 correspondería a una ecuación de la familia SB . Es de hacer notar que cuando z toma el valor 0.37, el mismo procedimiento calcula un valor de QR de 1.05642125, el cual, de acuerdo con el criterio establecido, correspondería a una ecuación de la familia SU , misma que puede observarse en la celda Q33.

Cabe hacer notar que, dada la naturaleza del cálculo de la transformación, los valores de las X calculadas son variables aleatorias, así que difícilmente se encontrará que el cociente QR sea exactamente igual a uno, por lo que es poco probable usar el modelo SL . Ante ello, Slifker y Shapiro (1980) proponen que se pueda establecer un intervalo alrededor de uno para que se pueda usar SL , por ejemplo, para valores de QR entre 0.95 y 1.05.

Para el caso que se está considerando, se observa en la Figura 7 el proceso de selección a partir del cálculo de QR . Así, para el valor de QR de 0.4004 que se muestra en la celda E31 de la Figura 7, la selección es de la familia SB por lo que se procede a calcular los parámetros correspondientes basados en las fórmulas 1.12, 1.13, 1.14 y 1.15 y que también se observa en el procedimiento descrito en la Figura 3. Los valores de los parámetros de esta familia se muestran en las celdas E47 a E50. Nótese que la selección de la ecuación de la familia SB , excluye a las ecuaciones de las familias SL y SU por lo que se ha colocado en las celdas E35 y E47 la leyenda de que no existe modelo para ese valor de z en particular aun cuando se calcule los valores de los parámetros para esas ecuaciones.

	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
31	QR	0.40037077	0.45734696	0.51284643	0.55441352	0.71579558	0.84880013	0.90036328	0.9082108	0.88916113	0.92523985	0.94771859	0.96686259	1.05642125
32														
33	Decisión	SB	SB	SB	SB	SB	SB	SB	SB	SB	SB	SB	SB	SU
34														
35	Familia SU	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	Existe
36	$\eta=$	0.64382935	0.70512777	0.76713615	0.80777453	0.87652968	0.97362572	1.03645014	1.07770813	1.10539763	1.15202716	1.20411333	1.27327158	1.43181834
37	$\gamma=$	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-0.11677768
38	$\lambda=$	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	1.7922084
39	$\varepsilon=$	8.88822903	11.015349	15.1023573	17.4485057	41.2864591	-10.6718298	-5.09126935	-4.4098652	-5.3124473	-4.02295814	-3.03957141	-2.00108647	0.02931934
40														
41	Familia SB	Existe	Existe	Existe	Existe	Existe	Existe	Existe	Existe	Existe	Existe	Existe	Existe	No Existe
42	$\eta=$	0.51089326	0.51307674	0.51690129	0.52129606	0.49599895	0.49362863	0.50410933	0.51985348	0.53911846	0.54984374	0.56374533	0.57993002	0.591035
43	$\gamma=$	0.62591642	0.64129348	0.67595396	0.74477672	1.0349905	1.27883913	1.42874098	1.47906374	1.45040889	1.64125587	1.74608818	1.75233382	1.90839647
44	$\lambda=$	12.1198645	13.9571168	15.9915637	17.5049814	24.2969289	41.2870763	59.3585363	64.5391334	56.2721703	78.8654716	111.603937	183.460364	-106.349274
45	$\varepsilon=$	2.44003821	2.36053287	2.23928015	1.98799518	0.82291378	0.00751367	-0.3340813	-0.35670629	-0.21820005	-0.74001986	-0.97666258	-0.94813049	-1.13197122
46														
47	Familia SL	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe
48	$\eta=$	4.15747274	2.07114871	1.56388125	1.56384366	1.55572565	1.27815235	1.21825444	1.22886971	1.27179283	1.29491313	1.28161372	1.21524125	1.03871475
49	$\gamma=$	0.71015229	-0.64831662	-0.84587292	-0.83206575	-0.82178606	-0.87921426	-0.86061601	-0.83028717	-0.80443269	-0.7847193	-0.77355221	-0.77571666	-0.78030118
50	$\varepsilon=$	-25.4091004	-9.2241418	-5.22247738	-5.03995198	-4.81169209	-2.71038268	-2.07457668	-1.92170697	-2.05688569	-2.08945654	-1.91205931	-1.4751717	-0.40804225

Figura 7. Proceso de cálculo de idoneidad y selección de familia de ecuaciones

Comprobación de idoneidad de parámetros de ecuaciones y transformación de datos.

En el punto anterior se ha determinado, con base en QR , la familia de ecuaciones de Johnson que pueden ser más convenientes de transformar para los datos originales. Ello no significa que los parámetros de la ecuación a ser transformada puedan ser convenientes en términos numéricos. Aunque QR indique la posibilidad de usar una ecuación en particular, el cálculo de los parámetros del modelo pudiera no ser posible dado el origen de las fórmulas y devenir en resultados indeterminados. Ello porque se utilizan funciones trascendentes o de potencia que requieren de criterios a cumplir para realizar el cálculo. Por ejemplo, en la función trascendente logarítmica, solo es posible realizar el cálculo cuando el argumento del algoritmo es positivo. Así entonces, para ese caso particular de z , no sería posible calcular

ecuación alguna. En la Figura 8 se pueden observar estas situaciones. En las celdas verticales a partir de E53, Excel previene del cálculo numérico realizado por lo que, aunque se utilice el modelo SB, el cálculo implica violación de alguna regla numérica. Ante ello, Excel coloca la leyenda de cálculo erróneo por lo que, para ese valor de z, no se puede realizar transformación alguna de los datos originales. Caso contrario ocurre cuando el valor de z es igual a 0.3 que se muestra en la columna J en la que ya es posible realizar el cálculo de la transformación al cumplirse los criterios numéricos de las fórmulas utilizadas a partir de un modelo SB.

La transformación se realiza utilizando las fórmulas correspondientes a cada modelo como se describió anteriormente y que puede observarse en el procedimiento de la Figura 3.

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
2			S =	3												
3																
4			Z	0.25	0.26	0.27	0.28	0.29	0.3	0.31	0.32	0.33	0.34	0.35	0.36	0.37
5			-SZ	-0.75	-0.78	-0.81	-0.84	-0.87	-0.9	-0.93	-0.96	-0.99	-1.02	-1.05	-1.08	-1.11
6			-Z	-0.25	-0.26	-0.27	-0.28	-0.29	-0.3	-0.31	-0.32	-0.33	-0.34	-0.35	-0.36	-0.37
7			Z	0.25	0.26	0.27	0.28	0.29	0.3	0.31	0.32	0.33	0.34	0.35	0.36	0.37
8			SZ	0.75	0.78	0.81	0.84	0.87	0.9	0.93	0.96	0.99	1.02	1.05	1.08	1.11
31			QR	0.40037077	0.45734696	0.51284643	0.55441352	0.71579558	0.84880013	0.90036328	0.9082108	0.88916113	0.92523985	0.94771859	0.96686259	1.05642125
32																
33			Decisión	SB	SB	SB	SB	SB	SB	SB	SB	SB	SB	SB	SB	SU
34																
35			Familia SU	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	Existe
36			η =	0.64382935	0.70512777	0.76713615	0.80777453	0.87652968	0.97362572	1.03645014	1.07770813	1.10539763	1.15202716	1.20411333	1.27327158	1.43181834
37			γ =	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-0.1167777
38			λ =	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	1.7922084
39			ϵ =	8.88822903	11.015349	15.1023573	17.4485057	41.2864591	-10.67183	-5.0912693	-4.4098652	-5.3124473	-4.0229581	-3.0395714	-2.0010865	0.02931934
40																
41			Familia SB	Existe	Existe	Existe	Existe	Existe	Existe	Existe	Existe	Existe	Existe	Existe	Existe	No Existe
42			η =	0.51089326	0.51307674	0.51690129	0.52129606	0.49599895	0.49362863	0.50410933	0.51985348	0.53911846	0.54984374	0.56374533	0.57993002	0.591035
43			γ =	0.62591642	0.64129348	0.67595396	0.74477672	1.0349905	1.27883913	1.42874098	1.47906374	1.45040889	1.64125587	1.74608818	1.75233382	1.90839647
44			λ =	12.1198645	13.9571168	15.9915637	17.5049814	24.2969289	41.2870763	59.3585363	64.5391334	56.2721703	78.8654716	111.603937	183.460364	-106.34927
45			ϵ =	2.44003821	2.36053287	2.23928015	1.98799518	0.82291378	0.00751367	-0.3340813	-0.3567063	-0.2182	-0.7400199	-0.9766626	-0.9481305	-1.1319712
46																
47			Familia SL	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe	No existe
48			η =	4.15747274	2.07114871	1.56388125	1.56384366	1.55572565	1.27815235	1.21825444	1.22886971	1.27179283	1.29491313	1.28161372	1.21524125	1.03871475
49			γ =	0.71015229	-0.6483166	-0.8458729	-0.8320658	-0.8217861	-0.8792143	-0.860616	-0.8302872	-0.8044327	-0.7847193	-0.7735522	-0.7757167	-0.7803012
50			ϵ =	-25.4091	-9.2241418	-5.2224774	-5.039952	-4.8116921	-2.7103827	-2.0745767	-1.921707	-2.0568857	-2.0894565	-1.9120593	-1.4751717	-0.4080422
51																
52	Originales	Consecutivo	Ordenados													
53	8.345	1	0.3936	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-1.0228764	-0.7838721	-0.8305682	-0.9813486	-0.6834006	-0.7274008	-1.0955275	6.87929036
54	4.1845	2	0.4265	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-0.9821116	-0.7612975	-0.8079907	-0.952791	-0.6674374	-0.713857	-1.081374	6.2914686
55	6.4831	3	0.4288	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-0.7993815	-0.7597556	-0.8064476	-0.9508488	-0.666338	-0.7129219	-1.0803972	6.25396953
56	15.562	4	0.4954	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-0.9061266	-0.7169893	-0.7635971	-0.8973821	-0.6353962	-0.6864808	-1.052792	5.32738438
57	2.8342	5	0.6495	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-0.7687652	-0.6297602	-0.6759219	-0.7904744	-0.569671	-0.6295658	-0.9934786	3.93947565
58	4.4225	6	0.7861	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-0.6718761	-0.563022	-0.6086257	-0.7103245	-0.5171417	-0.5833939	-0.9454641	3.17534583
59	4.1196	7	0.8391	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-2.5923881	-0.6387218	-0.5392588	-0.584624	-0.6820816	-0.4979935	-0.5664218	2.94681588
60	6.3356	8	0.9814	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-1.4578473	-0.5590091	-0.4803122	-0.5250053	-0.6126161	-0.4495387	-0.5231561	-0.8829545
61	4.9297	9	1.0595	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-1.2575204	-0.5199729	-0.4505593	-0.4948731	-0.577847	-0.424585	-0.5007055	-0.8596942
62	2.7622	10	1.2493	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-0.9614325	-0.4357602	-0.3845384	-0.4279263	-0.5013159	-0.3680947	-0.4494857	-0.8067002
63	3.8754	11	1.2951	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-0.9098742	-0.4173169	-0.3697639	-0.4129301	-0.4842968	-0.3552515	-0.437768	-0.7945906
64	3.1382	12	1.3208	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-0.8830518	-0.4072439	-0.3616493	-0.4046917	-0.4749648	-0.3481675	-0.4312938	-0.7879021
65	0.4288	13	1.4105	#NUM!	#NUM!	#NUM!	#NUM!	#NUM!	-0.7990159	-0.3735207	-0.3342555	-0.3768693	-0.4435399	-0.3240985	-0.4092394	-0.7651295

Figura 8. Proceso de comprobación de idoneidad de parámetros de ecuaciones y transformación de datos

Así, por ejemplo, cuando el valor de z es igual a 0.3 en la celda J4 y dado que QR (0.8488) en la celda J31 indica el uso de una ecuación del tipo SB que se muestra en la celda J33. Con ello ya se puede realizar la transformación sobre los datos originales. En primer término, se requiere ordenar los datos e indizarlos. En las celdas B53 a B152 se muestran los datos originales, en tanto, en las celdas C53 a C152 se despliega la indización de los valores. En las celdas D53 a D152 se ha realizado la ordenación ascendente de los datos originales. En particular, es posible observar en la captura que el dato original de la celda B65 (0.4265), se ha ordenado y se encuentra en la celda D54.

En el rango de celdas que va de E53 a DA152 se presenta la transformación de los datos originales a partir del modelo marcado por QR y el uso de las ecuaciones correspondientes a cada modelo y que se muestran en el procedimiento de la Figura 3. Dentro de este rango de celdas, es posible observar que no se puede realizar la transformación, por la índole del

cálculo numérico, en particular en las celdas correspondientes a las columnas E, F, G, H y J. A partir de la celda J53 a DA152 ya es posible realizar todos los cálculos usando las ecuaciones referidas en el procedimiento. En la Figura 9 se describe el cálculo realizado por Excel usando el modelo *SB* con z igual a 0.3, en tanto que en la Figura 10 se muestra el cálculo cuando se usa el modelo *SU* para z igual a 0.37. No se muestra cálculo para modelo *SL* toda vez que en ningún valor de z el valor de QR se acerca a uno, sin embargo, el cálculo de la transformación se haría en términos similares con la ecuación correspondiente.

S =	3		
Z	0.3		
QR	0.84880013		
Decisión	SB		

Familia SB	Existe
$\eta=$	0.49362863
$\gamma=$	1.27883913
$\lambda=$	41.2870763
$\varepsilon=$	0.00751367

Indización	Ordenados	Transformación
1	0.3936	-1.022876422

$$Z = \gamma + \eta \ln \left(\frac{X - \varepsilon}{\lambda + \varepsilon - X} \right) = 1.2788 + 0.4936 \ln \left(\frac{0.3936 - 0.0075}{41.287 + 0.0075 - 0.3936} \right) = -1.0229$$

Figura 9. Ejemplo de proceso de transformación para familia SB

S =	3		
Z	0.37		
QR	1.05642125		
Decisión	SU		

Familia SU	Existe
$\eta=$	1.43181834
$\gamma=$	-0.1167777
$\lambda=$	1.7922084
$\varepsilon=$	0.02931934

Indización	Ordenados	Transformación
1	0.3936	6.879290364

$$Z = \gamma + \eta \ln \left[\left(\frac{X - \varepsilon}{\lambda} \right) + \left\{ \left(\frac{X - \varepsilon}{\lambda} \right)^2 + 1 \right\}^{\frac{1}{2}} \right]$$

$$= -0.1167 + 1.4318 \ln \left[\left(\frac{0.3936 - 0.0293}{1.7922} \right) + \left\{ \left(\frac{0.3936 - 0.0293}{1.7922} \right)^2 + 1 \right\}^{\frac{1}{2}} \right] = 6.8793$$

Figura 10. Ejemplo de proceso de transformación para familia SU

Comprobación de normalidad de datos calculados

Una vez calculados los valores transformados mostrados en la hoja de cálculo de Excel en las celdas E53 a DA152 para cada valor de z y que se muestran en las celdas E4 a E104 se procede a realizar la prueba de normalidad para cada conjunto de datos. Para ello, se utilizará la prueba de Anderson Darling que, de acuerdo con Stephens (1974), es una prueba más sensible que otras, como la de Kolmogorov Smirnov. Aunque es posible usar sistemas computacionales para el cálculo de idoneidad de normalidad (R, Minitab, SAS, etc.), se ha realizado dentro de la hoja de Excel propuesta la prueba referida a efecto de tener el esquema completo de transformación de Johnson.

En la Figura 11 se muestra el resumen del procedimiento. Se presentan cuatro casos de cálculo para Z igual a 0.25 por ser el primer valor considerado, para z igual a 0.3 por ser el primer valor para el cual es posible calcular todas las transformaciones usando el modelo *SB*, para z igual a 0.37, primer valor para el que se define el cálculo completo bajo el modelo *SU* y para z igual a 0.74 que es el mejor valor para realizar la transformación como se describirá en el siguiente apartado. En las celdas J162 a J172 se muestra el así llamado *P-value* que es resultado de la selección del mayor de los cuatro valores de probabilidad que el procedimiento de Anderson-Darling sugiere, mismos que se muestran en las celdas E163 a DA163. Estos valores se obtienen a través de un método iterativo y que requieren cálculos adicionales, mismos que se realizan en las celdas E168 a DA571, parte de los cuales se muestran en la Figura 12 correspondiente a cada valor de z referido anteriormente.

	D	E	J	Q	BB
2	S =	3			
3					
4	Z	0.25	0.3	0.37	0.74
31	QR	0.40037077	0.84880013	1.056421253	0.60249159
32					
33	Decisión	SB	SB	SU	SB
35	Familia SU	No existe	No existe	Existe	No existe
36	$\eta=$	0.64382935	0.97362572	1.431818337	2.87116934
37	$\gamma=$	#¡NUM!	#¡NUM!	-0.116777679	#¡NUM!
38	$\lambda=$	#¡NUM!	#¡NUM!	1.792208395	#¡NUM!
39	$\varepsilon=$	8.88822903	-10.67183	0.029319338	-9.2121433
40					
41	Familia SB	Existe	Existe	No Existe	Existe
42	$\eta=$	0.51089326	0.49362863	0.591034997	1.47714989
43	$\gamma=$	0.62591642	1.27883913	1.908396469	1.50041229
44	$\lambda=$	12.1198645	41.2870763	-106.3492738	62.517398
45	$\varepsilon=$	2.44003821	0.00751367	-1.131971218	-0.2565277
46					
47	Familia SL	No existe	No existe	No existe	No existe
48	$\eta=$	4.15747274	1.27815235	1.03871475	1.77562138
49	$\gamma=$	0.71015229	-0.8792143	-0.780301178	-0.5957368
50	$\varepsilon=$	-25.4091	-2.7103827	-0.408042249	-4.2282098
51					
52	Ordenados				
53	0.3936	#¡NUM!	-1.02288	6.87929	-5.22886
54	0.4265	#¡NUM!	-0.98211	6.29147	-5.15515
55	0.4288	#¡NUM!	-0.97938	6.25397	-5.15013
155	Anderson Darling				
156	N=	61	100	100	100
157	MediaMuestral	NO	0.3968184	0.5751331	-1.7946053
158	DesvEstadMuestral	NO	0.66392247	1.389138495	1.65461874
159	Suma	#¡NUM!	-10028.483	-15396.76482	-10018.099
160	AD test statistic	NO	0.28482958	53.96764816	0.18099415
161	AD* test statistic	NO	0.28702989	54.38454824	0.18239233
162	P-value	NO	0.62179933	<0.0005	0.91210894
163	p1	0	0	0	0
164	p2	0	0	0	0
165	p3	0	0.62179933	0	0
166	p4	0	0	0	0.91210894

Figura 11. Resumen de cálculo de indicador de normalidad de Anderson-Darling

Una interpretación de los resultados que se muestran en la Figura 11 indica que para el valor de z igual a 0.25, el valor de *QR* es de 0.4004 que da origen a un modelo *SB*, los parámetros

de las ecuaciones se muestran en celdas E42 a E45 y se realiza la transformación de los primeros tres valores originales en celdas E53 a E55 pero Excel previene del cálculo numérico que no puede realizarse. Hay 61 datos para los que se realiza la transformación (celda E156) y dado ello, no se calcula la media muestral, desviación estándar muestral ni los cálculos para la obtención de los parámetros de desempeño de la prueba Anderson-Darling (celdas E159 a E161). En celda E162 no es posible calcular P-value a partir de valores p que se muestran en celdas E163 a E166. En resumen, para z igual a 0.25 no puede realizarse la transformación de Johnson a los datos originales.

D	E	J	Q	BB
S =	3			
Z	0.25	0.3	0.37	0.74
Fii	#NUM!	0.01624419	0.999997163	0.01896749
	#NUM!	0.01890346	0.999980643	0.02112707
	#NUM!	0.01909405	0.999978247	0.02128142
	#NUM!	0.0248521	0.999688047	0.02593476
	#NUM!	0.03957758	0.992280084	0.03790324
	#NUM!	0.05373458	0.969384229	0.04965569
	#NUM!	0.0594121	0.956117566	0.05444859
	#NUM!	0.07498058	0.912103488	0.0678382
	#NUM!	0.08365998	0.884930429	0.07545456
	#NUM!	0.10491563	0.817316225	0.09454461
	#NUM!	0.11005132	0.801415477	0.09924526
	#NUM!	0.11293376	0.792630489	0.10189612
	#NUM!	0.12296666	0.762885125	0.1112084

(a)

D	E	J	Q	BB
S =	3			
Z	0.25	0.3	0.37	0.74
F2i	#NUM!	0.00681811	2.83713E-06	0.01537303
	#NUM!	0.0084926	1.9357E-05	0.01692587
	#NUM!	0.00948731	2.17528E-05	0.01781806
	#NUM!	0.01131732	0.000311953	0.01941644
	#NUM!	0.03047239	0.007719916	0.034866
	#NUM!	0.04481486	0.030615771	0.04611143
	#NUM!	0.06607142	0.043882434	0.06300772
	#NUM!	0.07587067	0.087896512	0.07094752
	#NUM!	0.09030755	0.115069571	0.08283989
	#NUM!	0.09906988	0.182683775	0.09017709
	#NUM!	0.10981262	0.198584523	0.09929056
	#NUM!	0.11949713	0.207369511	0.10761986
	#NUM!	0.12815939	0.237114875	0.11515856

(c)

D	E	J	Q	BB
S =	3			
Z	0.25	0.3	0.37	0.74
1-F1i	#NUM!	0.98375581	2.83713E-06	0.98103251
	#NUM!	0.98109654	1.9357E-05	0.97887293
	#NUM!	0.98090595	2.17528E-05	0.97871858
	#NUM!	0.9751479	0.000311953	0.97406524
	#NUM!	0.96042242	0.007719916	0.96209676
	#NUM!	0.94626542	0.030615771	0.95034431
	#NUM!	0.9405879	0.043882434	0.94555141
	#NUM!	0.92501942	0.087896512	0.9321618
	#NUM!	0.91634002	0.115069571	0.92454544
	#NUM!	0.89508437	0.182683775	0.90545539
	#NUM!	0.88994868	0.198584523	0.90075474
	#NUM!	0.88706624	0.207369511	0.89810388
	#NUM!	0.87703334	0.237114875	0.8887916

(b)

D	E	J	Q	BB
S =	3			
Z	0.25	0.3	0.37	0.74
Si	#NUM!	-9.108193	-12.7727218	-8.140169
	#NUM!	-26.210913	-32.55742155	-23.808338
	#NUM!	-43.080895	-53.6789506	-39.387317
	#NUM!	-57.233636	-56.51078416	-53.156645
	#NUM!	-60.483842	-43.84531519	-59.660657
	#NUM!	-66.318056	-38.6906766	-66.872703
	#NUM!	-72.023594	-41.22450233	-73.774952
	#NUM!	-13.623723	-294.8375655	-58.093457
	#NUM!	-7.0537602	-297.9248699	-53.651049
	#NUM!	-5.6182026	-301.0121742	-51.015585
	#NUM!	-5.4398158	-304.0994786	-45.056336
	#NUM!	-4.620587	-307.1867829	-42.041878
Suma	#NUM!	-10028.483	-15396.76482	-39.175872

(d)

Figura 12. Proceso de cálculo de cálculo iterativo de indicador de normalidad de Anderson-Darling

De manera similar, es posible realizar la interpretación para el resto de valores de z. En la misma Figura 11 se han presentado la captura de pantalla para el primer valor de z (0.3) donde el modelo SB se define por completo, para z igual a 0.37 donde se define por vez primera el modelo SU y para z igual a 0.74 que corresponde al mejor valor de P-value. Cuando z es 0.3, el P-value calculado es de 0.6218 que comparado, por ejemplo, con un nivel de significancia de 0.05 no sería posible rechazar la hipótesis nula de que los datos provienen de una población con distribución normal. De manera similar se haría la conclusión con el valor de z igual a 0.74.

Sin embargo, en el caso de z igual a 0.37, el P-value calculado es menor a 0.0005 que a un nivel de significancia de 0.05, se rechazaría la hipótesis de que los datos provienen de una población con distribución normal.

Cálculo de valor estadístico para selección de ecuación

Ya se ha dejado entrever en la sección anterior cual es el mejor valor de *P-value*. En la Figura 13 puede observarse el reporte que se realiza en la hoja de cálculo a través de la búsqueda en el rango de celdas de E162 a DA162 del mejor valor de *P-value*. Así, este valor es el correspondiente a 0.9122 que se obtiene en la celda BB162, mismo que se reporta en la celda C162 para un valor de *z* de 0.74 que puede encontrarse en la celda BB4 y reportado en la celda C163. A lo largo del rango de celdas donde se calcula los valores de *P-value* es posible observar un buen número de valores que cumplen con el criterio de ser mayores a un nivel de significancia de 0.05.

	B	C	D	E	J	Q	BB
2			S =	3			
3							
4			Z	0.25	0.3	0.37	0.74
155			Anderson Darling				
156			N=	61	100	100	100
157			MediaMuestral	NO	0.3968184	0.5751331	-1.7946053
158			DesvEstMuestral	NO	0.66392247	1.389138495	1.65461874
159			Suma	#¡NUM!	-10028.483	-15396.76482	-10018.099
160			AD test statistic	NO	0.28482958	53.96764816	0.18099415
161			AD* test statistic	NO	0.28702989	54.38454824	0.18239233
162	P-value(max)=	0.91210893	P-value	NO	0.62179933	<0.0005	0.91210894
163	Z óptimo =	0.74	p1	0	0	0	0
164			p2	0	0	0	0
165			p3	0	0.62179933	0	0
166			p4	0	0	0	0.91210894

Figura 13. Proceso de obtención de valores óptimos a partir de la prueba Anderson-Darling

Sin embargo, desde un punto de vista estadístico se sugiere seleccionar el mayor valor de *P-value*. Así entonces, ya se han determinado estos valores y se ha hecho de manera exitosa. A manera de resumen y tratando de tener una visualización global del procedimiento seguido, se presenta en la Figura 14 la captura del rango de celdas más relevantes en el cálculo, enmarcándose en color rojo los pasos seguidos en el procedimiento planteado en la Figura 3.

Reporte y análisis de resultados

Una vez que se ha comprobado la normalidad de los datos transformados y obtenido el mejor valor de *P-value*, resta reportar los datos transformados junto con la ecuación de transformación seleccionada. Toda vez que el mejor valor de *P-value* es 0.9121 que corresponde a un valor de *Z* de 0.74, los valores que se calculan en las celdas BB53 a BB152 cuyos cuatro primeros valores pueden observarse en las celdas BB53 a BB56 de la Figura 14. En la Tabla 2 se ha conjuntado los valores transformados siguiendo un orden vertical. En tanto, en la Tabla 3 se ha conjuntado los valores originales una vez que se han ordenado y que se muestran en las celdas D53 a D152, junto con los valores transformados y que deben interpretarse por pares de valores a efecto de poder compararse. Desde luego, puede darse el caso de que los datos obtenidos de la población original no puedan transformarse. Ello se puede observar cuando el *P-value* obtenido sea menor al valor de significancia de la prueba, en cuyo caso, el analista pudiera recurrir a otros enfoques estadísticos cuya descripción

quedan fuera de este trabajo. En caso de que se cumpla la prueba de normalidad, los valores transformados pueden usarse en las aplicaciones estadísticas que requieran datos a partir de una población con distribución normal y realizar las conclusiones y recomendaciones pertinentes con la certeza de que se ha cumplido ya con ese criterio específico.

	B	C	D	E	J	Q	BB
1							
2			S =	3			
3							
4			Z	0.25	0.3	0.37	0.74
31			QR	0.40037077	0.84880013	1.056421253	0.60249159
32							
33			Decisión	SB	SB	SU	SB
34							
35			Familia SU	No existe	No existe	Existe	No existe
36			$\eta=$	0.64382935	0.97362572	1.431818337	2.87116934
37			$\gamma=$	#¡NUM!	#¡NUM!	-0.116777679	#¡NUM!
38			$\lambda=$	#¡NUM!	#¡NUM!	1.792208395	#¡NUM!
39			$\varepsilon=$	8.88822903	-10.67183	0.029319338	-9.2121433
40							
41			Familia SB	Existe	Existe	No Existe	Existe
42			$\eta=$	0.51089326	0.49362863	0.591034997	1.47714989
43			$\gamma=$	0.62591642	1.27883913	1.908396469	1.50041229
44			$\lambda=$	12.1198645	41.2870763	-106.3492738	62.517398
45			$\varepsilon=$	2.44003821	0.00751367	-1.131971218	-0.2565277
46							
47			Familia SL	No existe	No existe	No existe	No existe
48			$\eta=$	4.15747274	1.27815235	1.03871475	1.77562138
49			$\gamma=$	0.71015229	-0.8792143	-0.780301178	-0.5957368
50			$\varepsilon=$	-25.4091	-2.7103827	-0.408042249	-4.2282098
51							
52	Originales	Indización	Ordenados				
53	8.345	1	0.3936	#¡NUM!	-1.02288	6.87929	-5.22886
54	4.1845	2	0.4265	#¡NUM!	-0.98211	6.29147	-5.15515
55	6.4831	3	0.4288	#¡NUM!	-0.97938	6.25397	-5.15013
56	15.562	4	0.4954	#¡NUM!	-0.90613	5.32738	-5.01154
155			Anderson Darling				
156			N=	61	100	100	100
157			MediaMuestral	NO	0.3968184	0.5751331	-1.7946053
158			DesvEstadMuestral	NO	0.66392247	1.389138495	1.65461874
159			Suma	#¡NUM!	-10028.483	-15396.76482	-10018.099
160			AD test statistic	NO	0.28482958	53.96764816	0.18099415
161			AD* test statistic	NO	0.28702989	54.38454824	0.18239233
162	P-value(max)=	0.91210893	P-value	NO	0.62179933	<0.0005	0.91210894
163	Z óptimo =	0.74	p1	0	0	0	0

Figura 14. Resumen de procedimiento de Johnson

Si el procedimiento estadístico correspondiente implica el cálculo de algún parámetro de desempeño, se realiza la transformación inversa para la obtención del valor en términos de la variable original, haciendo el despeje en la ecuación de transformación.

Tabla 2. Datos transformados por procedimiento de Johnson									
-5.22886	-3.92222	-2.93817	-2.6658	-2.29715	-1.81162	-1.28107	-1.04323	-0.29402	0.42203
-5.15515	-3.89733	-2.90315	-2.65307	-2.24889	-1.70643	-1.23584	-0.99046	-0.17171	0.49906
-5.15013	-3.81345	-2.86901	-2.57352	-2.21397	-1.65794	-1.22952	-0.94715	-0.14994	0.63565
-5.01154	-3.67011	-2.86654	-2.57242	-2.21179	-1.62135	-1.21087	-0.89464	-0.13188	0.73697
-4.73247	-3.6313	-2.85272	-2.5031	-2.04894	-1.60567	-1.19797	-0.89316	-0.09866	0.99142
-4.52175	-3.57604	-2.83275	-2.41156	-2.04249	-1.59993	-1.15294	-0.75203	-0.03385	1.20629
-4.44724	-3.42776	-2.76548	-2.35057	-1.9001	-1.5028	-1.13789	-0.60682	0.18603	1.62376
-4.26344	-3.42576	-2.72038	-2.32055	-1.88953	-1.45338	-1.12695	-0.50984	0.19018	1.68184
-4.17118	-3.41817	-2.70167	-2.31656	-1.88522	-1.40299	-1.12178	-0.38824	0.25594	1.71622
-3.96758	-3.2673	-2.67191	-2.31164	-1.81753	-1.29369	-1.06428	-0.34294	0.33258	1.77994

Tabla 3. Resumen de datos originales y datos transformados usando procedimiento de Johnson.									
Original	Transf.	Original	Transf.	Original	Transf.	Original	Transf.	Original	Transf.
0.3936	-5.22886	2.6948	-2.93817	4.1845	-2.29715	7.9985	-1.28107	14.0508	-0.29402
0.4265	-5.15515	2.7622	-2.90315	4.3212	-2.24889	8.2204	-1.23584	14.9848	-0.17171
0.4288	-5.15013	2.8293	-2.86901	4.4225	-2.21397	8.2518	-1.22952	15.1553	-0.14994
0.4954	-5.01154	2.8342	-2.86654	4.4289	-2.21179	8.345	-1.21087	15.2977	-0.13188
0.6495	-4.73247	2.8618	-2.85272	4.9297	-2.04894	8.41	-1.19797	15.562	-0.09866
0.7861	-4.52175	2.9021	-2.83275	4.9505	-2.04249	8.6401	-1.15294	16.086	-0.03385
0.8391	-4.44724	3.0415	-2.76548	5.4295	-1.9001	8.7181	-1.13789	17.9454	0.18603
0.9814	-4.26344	3.1382	-2.72038	5.4666	-1.88953	8.7752	-1.12695	17.9816	0.19018
1.0595	-4.17118	3.1791	-2.70167	5.4818	-1.88522	8.8023	-1.12178	18.562	0.25594
1.2493	-3.96758	3.2451	-2.67191	5.7251	-1.81753	9.108	-1.06428	19.2514	0.33258
1.2951	-3.92222	3.2588	-2.6658	5.7468	-1.81162	9.222	-1.04323	20.0732	0.42203
1.3208	-3.89733	3.2875	-2.65307	6.1445	-1.70643	9.5129	-0.99046	20.795	0.49906
1.4105	-3.81345	3.4719	-2.57352	6.3356	-1.65794	9.757	-0.94715	22.105	0.63565
1.5754	-3.67011	3.4745	-2.57242	6.4831	-1.62135	10.0596	-0.89464	23.0995	0.73697
1.6227	-3.6313	3.6426	-2.5031	6.5472	-1.60567	10.0682	-0.89316	25.6693	0.99142
1.6921	-3.57604	3.8754	-2.41156	6.5708	-1.59993	10.9183	-0.75203	27.9004	1.20629
1.8908	-3.42776	4.0376	-2.35057	6.9811	-1.5028	11.8491	-0.60682	32.3065	1.62376
1.8936	-3.42576	4.1196	-2.32055	7.198	-1.45338	12.5029	-0.50984	32.9194	1.68184
1.9043	-3.41817	4.1306	-2.31656	7.4249	-1.40299	13.3593	-0.38824	33.2815	1.71622
2.1278	-3.2673	4.1442	-2.31164	7.9375	-1.29369	13.6887	-0.34294	33.951	1.77994

La ecuación de Johnson que realiza la transformación de los datos originales se describe en función de los parámetros calculados con acuerdo al modelo *SB* que se ha determinado. Los valores de estos parámetros se encuentran en las celdas BB42 a BB45, mismo que pueden observarse en la Figura 15, donde también se puede apreciar el mejor valor de *Z* (0.74), el valor *QR* (0.6025), la decisión con base en *QR* (*SB*) y los parámetros descritos. En la celda BB162 se despliega el mejor valor de *P-value* (0.9121), en la celda D53 el primer valor original una vez ordenado y en la celda BB53 el valor transformado (-5.22886) para el valor original 0.3936.

La ecuación de transformación con base en los criterios y cálculos realizados utilizando el modelo *SB* está dada por lo siguiente:

$$Z = \gamma + \eta \ln \left(\frac{X - \varepsilon}{\lambda + \varepsilon - X} \right) \quad (5)$$

	D	BB
4	Z	0.74
31	QR	0.60249159
32		
33	Decisión	SB
41	Familia SB	Existe
42	η=	1.47714989
43	γ=	1.50041229
44	λ=	62.517398
45	ε=	-0.2565277
52	Ordenados	
53	0.3936	-5.22886
162	P-value	0.91210894

Figura 15. Resumen del modelo usado en la transformación.

Que en términos de los parámetros calculados y realizando operaciones aritméticas se reduciría a lo siguiente:

$$Z = 1.5004 + 1.4771 \ln \left(\frac{X - (-0.2565)}{62.5174 + 0.2565 - X} \right) \quad (6)$$

$$Z = 1.5004 + 1.4771 \ln \left(\frac{X + 0.2565}{62.7739 - X} \right) \quad (7)$$

Así, por ejemplo, para el primer dato original correspondiente a 0.3936 la transformación ocurriría en los siguientes términos:

$$Z = 1.5004 + 1.4771 \ln \left(\frac{0.3936 + 0.2565}{62.7739 - 0.3936} \right) = -5.2289 \quad (8)$$

Este valor se muestra en la celda BB162. Como se mencionó anteriormente, la relación completa de datos transformados se puede encontrar en el rango de celdas BB162 a BB171 de la hoja de Excel propuesta.

Finalmente, se ha considerado pertinente presentar, de manera similar a como se realizó con los datos originales, la determinación de las pruebas de normalidad usando programas y librerías de RStudio (v2022.02.3+492) . La Figura 16 muestra la salida de cuatro pruebas

realizadas a los datos transformados, pudiéndose observar en particular la igualdad de la determinación del valor *P-value* para la prueba de Anderson-Darling. Igualmente se presenta en la Figura 17(a) el histograma de los datos transformados, en tanto que la Figura 17(b) presenta el mismo histograma junto con la función de densidad aproximada a los datos. Ambos esquemas son producto del mismo programa de RStudio (v2022.02.3+492) anteriormente referenciado. Puede notarse en las gráficas una gran aproximación a un comportamiento estadístico de una distribución normal, lo cual concuerda con los elementos encontrados de manera analítica y que se han presentado en este trabajo.

Anderson-Darling normality test data: Datos A = 0.18151, p-value = 0.9112	Cramer-von Mises normality test data: Datos W = 0.024362, p-value = 0.9166
Pearson chi-square normality test data: Datos P = 10.76, p-value = 0.3765	Shapiro-Francia normality test data: Datos W = 0.99289, p-value = 0.8049

Figura 16. Resumen de pruebas de normalidad utilizando R

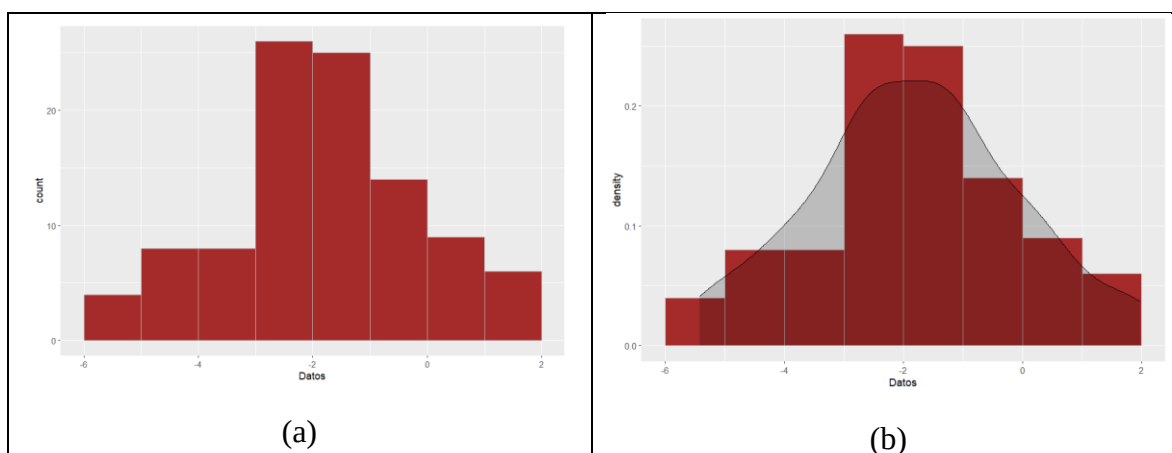


Figura 17. (a) Histograma de datos transformados y (b) histograma con función de densidad aproximada

Conclusiones y recomendaciones

Desde sus orígenes, la estadística ha proporcionado técnicas y procedimientos que se han utilizado para la recolección, ordenamiento, representación y análisis de un conjunto de datos de manera que pueda utilizarse para la toma de decisiones en una organización. El uso de estas técnicas, en particular aquellas relacionadas con el análisis, basan su actuación en el cálculo de indicadores de desempeño que están acotados por el cumplimiento de condiciones y supuestos a partir de los cuales han sido obtenidos sus modelos matemáticos y procedimientos de cálculo.

Un supuesto que requieren algunas técnicas de la estadística consiste en que los datos cumplan con provenir de una población cuyo comportamiento estadístico sea el de una distribución normal. La misma estadística ha aportado técnicas que comprueban con un grado de confianza si se cumple ese supuesto o no, previniendo la existencia de conclusiones erróneas en caso de no cumplir ese supuesto.

También, la estadística aporta enfoques de solución en caso de no cumplir con el supuesto de normalidad de los datos. Un camino es el de realizar un proceso de transformación de datos no normales a datos que cumplan con el criterio de normalidad y, una vez, transformados, pueda continuar el uso de las técnicas inferenciales.

El método de transformación de Johnson es uno de los métodos que más se recomienda ante el caso de incumplimiento del principio de normalidad. Propuesto por Johnson (1949), utiliza un procedimiento iterativo para tratar de encontrar una ecuación que mejor represente a los datos originales ante el criterio de normalidad.

Para la realización de las operaciones utilizadas se ha usado una hoja de cálculo muy popular en la actualidad como lo es Excel, dado que permite introducir los cálculos paso por paso y entender así el procedimiento de solución de tal manera que, una vez entendido el proceso, se pueda replicar el mismo cuantas veces sea necesario ante diferentes tipos de datos.

Básicamente, la hoja de cálculo propuesta realiza tres funciones: (1) usa el procedimiento iterativo de Johnson para obtener la mejor, de entre alternativas de ecuación de transformación; (2) realiza el proceso de cumplimiento del supuesto de normalidad a través de pruebas estadísticas; y (3) reporta la ecuación de transformación, los datos transformados y sus indicadores de desempeño.

Los autores de este trabajo están convencidos de que un practicante de las técnicas de estadística puede entender cabalmente el proceso de transformación que se ha presentado en este trabajo siguiendo la explicación que se ha realizado y tomar ventaja de ello ante la resolución de casos reales.

El trabajo ha cumplido los objetivos originalmente trazados. Sin embargo, aún puede extenderse incorporando otro u otros elementos que mejoren su desempeño. Se identifican dos en particular: (1) la posibilidad de hacer un análisis comparativo con otros sistemas computacionales estadísticos que ya incluyen este procedimiento de Johnson, tales como Minitab o el mismo R y; (2) la utilización de R para realizar los cálculos señalados, llevándolo incluso a proponer un paquete que realice la transformación de una manera amigable con el usuario.

Referencias

- Bowman, K. O. & Shenton L. R. (1988). *Solutions to Johnson's S_B y S_U* , Communications in Statistics - Simulation and Computation, 17:2, pp. 343-348
- Box G. E. P.; Cox D. R. (1964), *An Analysis of Transformations*. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 26, No. 2. pp. 211-252.
- Chou, Y., Polansky, A. M. & Mason R. L. (1998) *Transforming Non-Normal Data to Normality in Statistical Process Control*, Journal of Quality Technology, 30:2,133-141.
- Chou Y., Turner, S. Henson, D. Meyer & K. S., Chen (1994). *On using percentiles to fit data by a Johnson distribution*, Communications in Statistics - Simulation and Computation, 23:2, 341-354.
- Coss Bu, R., (1990) “*Simulación: un enfoque práctico*” Edit. Limusa. México, 1990.
- Gabrilov, I. y Pusev, R. “*Package ‘normtest’*”. Available at: <https://cran.r-project.org/web/packages/normtest/normtest.pdf> (consultado el 26 de mayo de 2022)
- George, F. and Ramachandran, K. M. (2011). *Estimation of Parameters of Johnson's System of Distributions*. Journal of Modern Applied Statistical Methods: Vol. 10: Iss. 2, Article 9.
- Gutiérrez P. H. y de la Vara S. R. (2009) *Control Estadístico de Calidad y Seis Sigma*. 2ª. Ed, Edit. McGraw Hill. México.
- Hans J.H. (2001) *An algorithm to determine the parameters of SU-curves in the Johnson system of probability distributions by moment matching*, Journal of Statistical Computation and Simulation, 70:4.
- Johnson, N. L. (1949). *Systems of frequency curves generated by methods of translation*. Biometrika, 36, 149-176.
- Lagos, I., Vargas, J. (2003) *Sistema de familias de distribuciones de Johnson, una alternativa para el manejo de datos no normales en cartas de control*. Revista Colombiana de Estadística, vol. 26, núm. 1, junio, pp. 25-40. Universidad Nacional de Colombia. Bogotá, Colombia
- Mohammed A. Shayib (1989) “*The Procedure for Selection of Transformations from the Johnson System*”, Communications in Statistics - Simulation and Computation, 18:4, 1457-1464
- Osborne, Jason (2002) “*Notes on the use of data transformations*,” Practical Assessment, Research, and Evaluation: Vol. 8, Article 6. DOI: <https://doi.org/10.7275/4vng-5608> Available at: <https://scholarworks.umass.edu/pare/vol8/iss1/6> (consultado el 3 de febrero de 2022)
- Osborne, Jason (2010) “*Improving your data transformations: Applying the Box-Cox transformation*,” Practical Assessment, Research, and Evaluation: Vol. 15, Article 12. DOI: <https://doi.org/10.7275/qbpc-gk17> Available at: <https://scholarworks.umass.edu/pare/vol15/iss1/12> (consultado el 2 de febrero de 2022)
- Pyzde, T. (1995) *Why Normal Distributions Aren't [All That Normal]*, Quality Engineering, 7:4, 769-777.
- R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
- Ryan, T. P. (2011) *Statistical Methods for Quality Improvement* 3rd ed. Edit. Wiley.
- Shapiro S. S. and Wilk M. B. (1965). *An Analysis of Variance Test for Normality (Complete Samples)*. Biometrika, Vol. 52, No. 3/4, pp. 591-611
- Simonato, J.G. (2011). *The Performance of Johnson Distribution for Computing Value at Risk and Expected Shortfall*. Social Science Electronic Publishing, 38.
- Slifker, J. F. and Shapiro, S. S. (1980). *The Johnson System: Selection and Parameter Estimation*. (A. S. Association, Ed.) Technometrics, 22, 239-246
- Stephanie Glen. “*Box Cox Transformation: Definition, Examples*” From StatisticsHowTo.com: Elementary Statistics for the rest of

us! <https://www.statisticshowto.com/box-cox-transformation/> (consultado el 23 de enero de 2022)

Stephens, M. A. “*EDF Statistics for Goodness of Fit and Some Comparisons*”. Journal of the American Statistical Association. Vol. 69, No. 347 (Sep., 1974), pp. 730-737.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

D'Agostino, Ralph B., Belanger, Albert and D'Agostino, Ralph B. Jr. (1990) *A Suggestion for Using Powerful and Informative Tests of Normality*. The American Statistician, Vol. 44, No. 4, pp. 316-32.

Referencias web:

<https://r-coder.com/transformation-box-cox-r/>, consultado el 7 de febrero de 2022.

https://rpubs.com/dvillasanao/Pruebas_de_Normalidad, consultado el 7 de febrero de 2022.

“Lis de Veracruz: Arte, Ciencia, Luz”

www.uv.mx

