APÉNDICE B

TRANSFORMACIÓN DE DATOS

Por lo general, la experiencia indica que para la mayoría de los datos biológicos el no cumplimiento de los supuestos requeridos en muchos de los métodos estadísticos paramétricos como son la aditividad, independencia, y normalidad de los datos no son de mayor importancia y dichos métodos pueden aplicarse sin mayores problemas. Sin embargo, hay ocasiones en las cuales no es posible obviar el incumplimiento de dichos supuestos. En estos casos se tienen dos alternativas. Una vía es recurrir a la estadística no paramétricas y usar métodos equivalentes como las pruebas de Mann-Whitney (= prueba de T), de Kruskall-Wallis (= Andeva de una vía), de Friedman (= Andeva de dos vías) o la correlación de Spearman (= correlación de Pearson). Este tipo de estadística no requiere de suposiciones previas acerca de la distribución de los datos. Sin embargo, cuando se cumplen los supuestos, aunque sea en forma aproximada, los métodos paramétricos son mucho más potentes que las pruebas no paramétricas en la verificación de diferencias significativas. La otra solución es la de transformar los datos de tal forma que los nuevos valores cumplan con los supuestos. Veamos como funciona la transformación con un grupo de datos artificiales. En la Tabla B1 se muestran tres grupos de datos no transformados. Los valores de los grupos B y C son el resultado de multiplicar los valores de A por un factor de 2 y 3 respectivamente.

Tabla B1. Datos ficticios no transformados

	Valores originales (x)		
	A	B	C
	4	8	12
	6	12	18
	7	14	21
	5	10	15
	3	6	9
	8	16	24
Promedio	5,5	11	16,5
Varianza	3,5	14	31,5

Bajo esta situación se violan dos de los supuestos básicos de las pruebas paramétricas, por un lado existe un efecto multiplicativo, el cual incumple la condición de aditividad. Por otro lado las varianzas de los grupos son muy diferentes. La aplicación de una transformación logarítmica puede resolver estos dos problemas (Tabla B2). El efecto multiplicativo se convierte en un efecto aditivo puesto que el $log\ xy$ es equivalente a $log\, x + log\, y$. Por otro lado las varianzas se hacen más parecidas. En el caso del ejemplo son exactamente iguales por tratarse de datos artificiales.

Tabla B2. Datos ficticios transformados

	Valores transformados (Log x)		
	A	B	C
	0,6021	0,9031	1,0792
	0,7782	1,0792	1,2553
	0,8451	1,1461	1,3222
	0,6990	1,0000	1,1761
	0,4771	0,7782	0,9542
	0,9031	1,2041	1,3802
Media	0,7174	1,0184	1,1945
Varianza	0,0252	0,0252	0,0252

El paso siguiente es efectuar el análisis paramétrico con los datos transformados. En este punto es importante reflexionar sobre el concepto de transformación, el cual es difícil de aceptar porque da la impresión que se quiere ver lo que se desea y no lo que es. Pero esta es una idea equivocada que posiblemente surge de nuestra costumbre de ver las relaciones cuantitativas en una escala lineal o aritmética. La transformación no es sino un cambio en la escala de observación, que da una perspectiva distinta y permite detectar relaciones que no se observaban en la escala original. Vamos a ilustrar esta idea con un ejemplo no matemático. Supóngase un viajero que desea trasladarse por primera vez desde un sitio A hacia otros dos sitios B y C. En la Figura B1 se muestra el trayecto de la carretera entre los tres puntos.

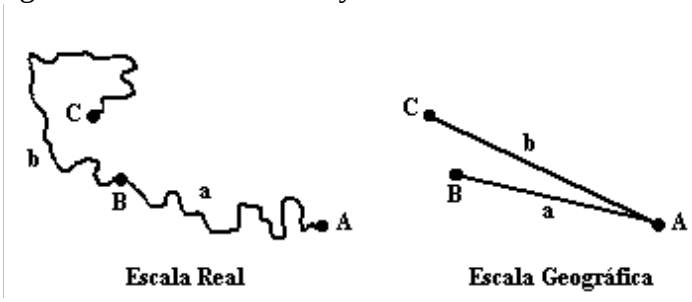

Figura B1

Para el viajero el punto C esta tan lejos del punto B como éste del punto A. Esto es verdad bajo la perspectiva de la situación que él debe resolver. El traslado entre los tres puntos se debe hacer a través de una carretera. Los accidentes del terreno determinan que las distancias a recorrer sean mayores que las distancia verdaderas. Si el viajero, se cambia a una escala de observación mayor, por ejemplo desde un avión o sobre un mapa, donde se observa simultáneamente la ubicación de los tres puntos, se dará cuenta que el punto C está mucho más cerca de B y A de lo que parecía al viajar por la carretera, tal y como se muestra en la Figura B1. Las dos situaciones son verdaderas y la conclusión que obtuvo de ambas dependió de la perspectiva o el nivel de la escala en la cual se colocó.

Los datos biológicos pueden producir situaciones caracterizadas por falta de aditividad, de heterogeneidad de las varianzas y no normalidad de los datos. Los tres tipos de problemas

pueden ser resueltos mediante algunas transformaciones estándar, las cuales veremos a continuación.

Transformación logarítmica.

Esta transformación se produce por la conversión de los datos originales en logaritmos, usualmente se utilizan logaritmos decimales. Cuando existen valores menores a 1, se puede usar el $log(x+1)$ para evitar trabajar con cantidades negativas. La transformación logarítmica ayuda a resolver situaciones de falta de aditividad e independencia de los datos, como hemos visto anteriormente. También es muy útil cuando existe dependencia de la varianza con respecto al valor de la media. Es decir, que a mayores valores de las medias le corresponden mayores varianzas. Considérese, por ejemplo, el caso del número de presas consumidas por un depredador. Esta relación varía desde cero (ninguna presa consumida), hasta valores extremadamente grandes cuando un solo depredador consume muchas presas, teóricamente este número puede ser infinitamente grande puesto que no hay límite para el número de presas consumidas. Si el registro de la relación presa/depredador se efectúa a lo largo del tiempo, es posible que en aquellos fechas con escasez de presas el numero promedio de esa relación sea bajo y también su varianza. Por el contrario, en las épocas con abundancias de presas, el promedio de la relación presa/depredador será grande y consecuentemente su varianza también será grande. Esta situación se puede clarificar graficando la desviación estándar vs el promedio.

En la Figura B2 se muestra el promedio y la desviación estándar del número de presas consumidas por individuo de la trucha Arco iris en ocho fechas diferentes, para valores no transformados (arriba) y valores transformados a logaritmos (abajo). En la parte superior del gráfico se observa que existe una relación aproximadamente lineal entre la desviación estándar y el promedio de presas consumidas por individuo. En la parte inferior se puede ver que la transformación a los logaritmos naturales eliminó la dependencia de la varianza con la media, es decir que las varianzas se hicieron homogéneas.

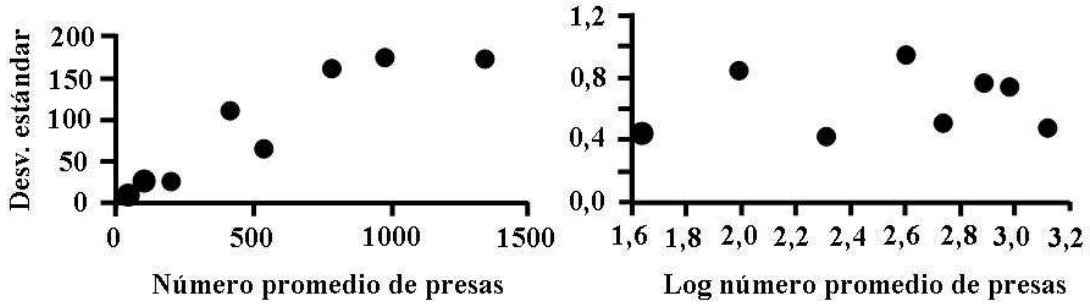

Figura B2. Dispersión del número de presas consumidas en una escala lineal (arriba) y una escala logarítmica (abajo)

Muchas veces no es la homogenización de las varianzas lo que determina una transformación logarítmica, sino la necesidad práctica de disminuir las diferencias de magnitud que pueden existir en un conjunto de datos. Por ejemplo, los cultivos de bacterias presentan un crecimiento exponencial, que es más apropiado representarlo en una escala logarítmica que en una escala aritmética, dada la naturaleza no lineal de este proceso.

Transformación raíz cuadrada

Los resultados de muchos experimentos se expresan como el número de veces que ocurre un resultado en un tiempo determinado o en un espacio dado. Por ejemplo: número de partículas desintegradas en una unidad de tiempo; número de electrones emitidos en una unidad de tiempo; número de glóbulos por campo; número de casos de una enfermedad en un año; número de anímales por unidad de área; número de bacterias por unidad de volumen y número de plantas por unidad de longitud. Usualmente, la distribución de este tipo de resultados se ajusta al modelo de Poison, por lo que sus varianzas y medias son muy similares. Esta falta de independencia de la varianza compromete los resultados de algunas prueba estadísticas paramétricas como el Andeva. Afortunadamente, la transformación de los datos en sus raíces cuadradas puede resolver este problema. En la Tabla B3 se expone un ejemplo, en el cual se aplicó la transformación raíz cuadrada.

Tabla B3. Número de ninfas por hoja (datos originales y transformados) en dos fechas diferentes durante el desarrollo de cierto cultivo.

a. Datos originales		
N° ninfas (x)	Fecha 1	Fecha 2
0	13	0
1	27	1
2	28	5
3	18	9
4	9	13
5	4	16
6	1	18
7		14
8		10
9		7
10		4
11		2
12		1
13		1
$\sum fx$	199	606
$\sum fx^2$	581	4206
$\sum f$	100	101
Media	1,99	6,00
Varianza	1,87	5,70
Razón de Varianza	3,05***	
$F_{(0,05;\ 99/100)}$	1,87	

b. Datos transformados		
$N^{\circ}\ ninfas\ \left(\sqrt{x}\right)$	Fecha 1	Fecha 2
0,000	13	0
1,000	27	1
1,414	28	5
1,732	18	9
2,000	9	13
2,236	4	16
2,449	1	18
2,646		14
2,828		10
3,000		7
3,162		4
3,317		2
3,464		1
3,606		1
$\sum fx$	127,17	242,20
$\sum fx^2$	199	606
$\sum f$	100	101
Media	1,27	2,40
Varianza	0,38	0,25
Razón de Varianza	0,67 ns	
$F_{(0,05;\ 99/100)}$	1,87	

Los resultados de la Tabla B3a muestran que la segunda fecha tiene una varianza significativamente mayor que la primera. En la Tabla B3b se observa que una vez aplicada la transformación $\sqrt{x}$ no hay diferencias significativas entre las varianzas. La transformación homogenizó las varianzas de las dos muestras.

La $\sqrt{x}$ también puede usarse para transformar datos porcentuales, cuando el intervalo de variación se encuentra entre un 0 y un 20 por ciento. Sí el intervalo va de 80 a 100 por ciento, los porcentajes deberán restarse de 100 antes de la transformación.

Transformación Angular (Arco-seno).

Esta transformación se utiliza para valores expresados en porcentajes o proporciones. Este tipo de datos por lo general se distribuye siguiendo el modelo binomial. Como sabemos las distribuciones binomiales se caracterizan porque la varianza es función de la media.

$$Media = \mu = np \qquad Varianza = \sigma^2 = npq = \mu q$$

La Figura B3 muestra como las varianzas de distintas distribuciones binomial tienden a ser mayores para valores intermedios de las medias y son menores para valores pequeños o grandes de la media.

De la condición anterior se vislumbra que distribuciones de datos con medias diferentes pueden ser asimétricas y con varianzas diferentes. La transformación arco seno puede solucionar esta situación, puesto que al aplicarse alarga los extremos de la distribución y angosta la parte central.

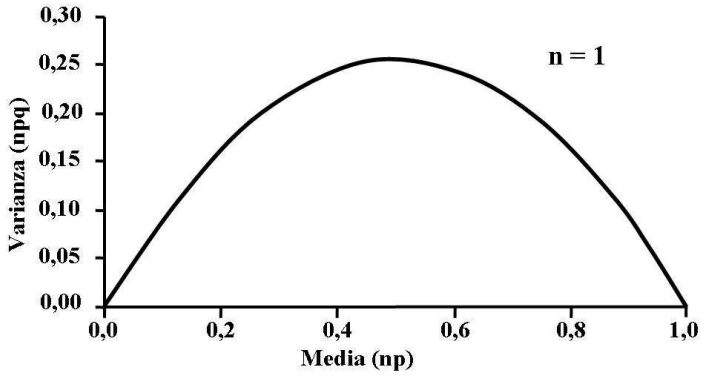

Figura B3. Distribución de la varianza para varias distribuciones binomial con diferentes medias.

Para transformar los datos se obtiene el arco seno (inverso del seno) de la raíz de la proporción ($ar\cos eno\sqrt{p}$), siendo p es el valor proporcional de los datos originales (los porcentajes deben dividirse entre 100). Las unidades de los valores transformados son grados o radianes.

En la Tabla B4 se muestra la distribución del número de truchas como un porcentaje del total de presas encontradas en los estómagos de 246 individuos, antes y después de aplicarles la transformación angular.

Tabla B4. Distribución de la proporción de presas en el estómago de 246 truchas, para los datos originales (% presas) y transformados en el *arcoseno* $\sqrt{(\%\ de\ \ presas)/100}$.

% presas	Grados	N° truchas
0	0,000	5
10	18,435	25
20	26,565	59
30	33,211	60
40	39,232	45
50	45,000	20
60	50,768	12
70	56,789	8
80	63,435	6
90	71,565	4
100	90,000	2

La aplicación de la transformación a los % de las presas aproxima la distribución a una normal, tal como lo muestra la Figura B4.

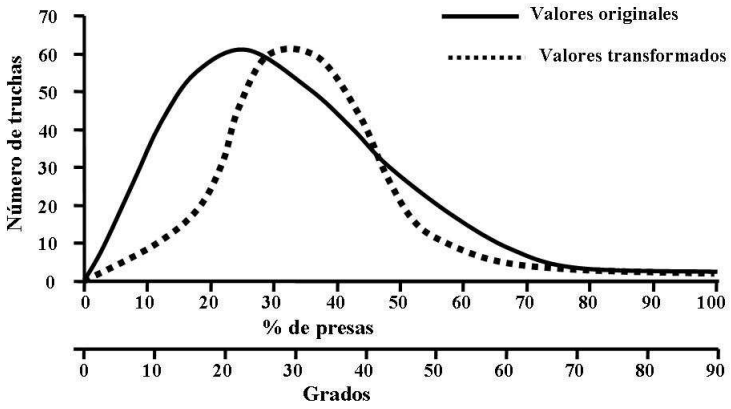

Figura B4. Distribución del número de presas (en % de presas ó en grados) en el contenido estomacal de 246 truchas.