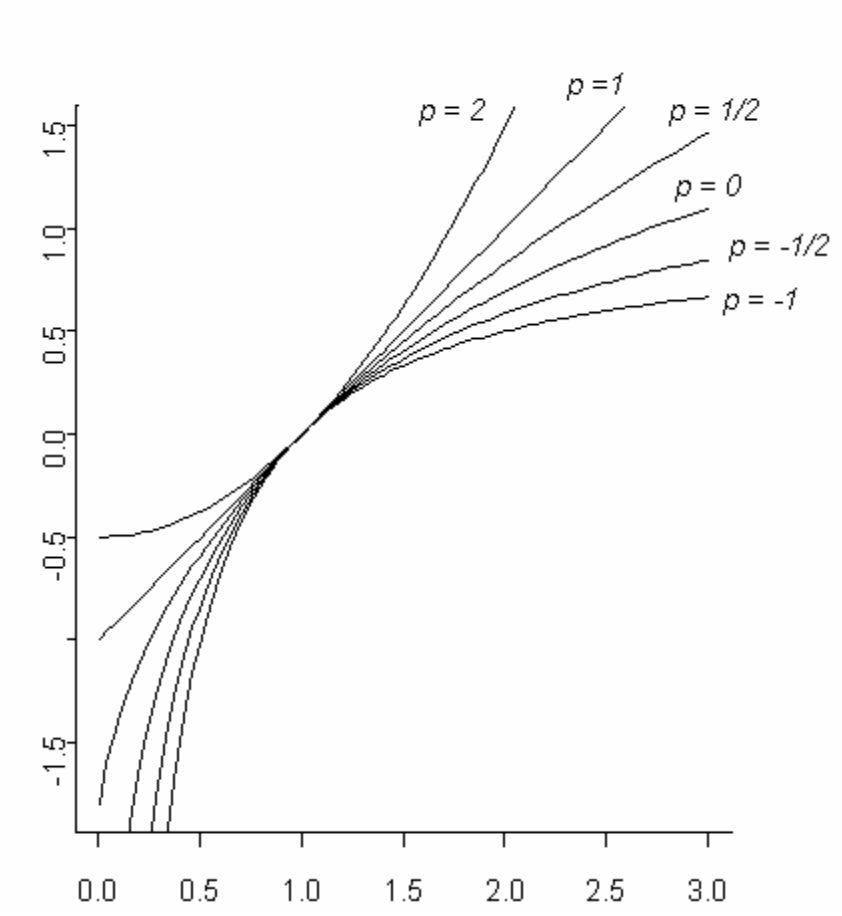
TRANSFORMACIONES APAREADAS

Hemos visto que la familia $T_p^*(x)$

$$T_p^*(x) = \begin{cases} \dfrac{x^p - 1}{p} & (p \neq 0) \\ \ln x & (p = 0) \end{cases}$$

constituye una familia de transformaciones apareadas en el punto $(x, T_p^*(x)) = (1,0)$ y las propiedades que resultan del apareamiento en ese punto.



Procedimiento para obtener transformaciones apareadas en el punto (x_0, x_0).

Sea *x* el dato original e *y* =*T*(*x*) la correspondiente transformación no lineal. Nos interesa hallar una nueva transformación

$$a + bT(x) = x$$

que esté apareada en *x* = x_0 , *e*s decir que deje fijo el punto x_0 :

$$T_{apareada\ en\ x_0}(x_0) = a + bT(x_0) = x_0$$

y

tenga derivada 1 en ese punto:

$$\left.\frac{dz}{dx}\right|_{x_0} = \left.\frac{d\left[a + bT(x)\right]}{dx}\right|_{x_0}$$

$$= b\left.\frac{dT(x)}{dx}\right|_{x_0} = 1.$$

O sea

$$b = \frac{1}{T'(x_0)}\ . \qquad (1)$$

Por lo tanto

$$a = x_0 - bT(x_0) = x_0 - \frac{T(x_0)}{T'(x_0)}. \qquad (2)$$

y la expresión apareada en *x* = x_0 resulta

$$z = x_0 + \frac{T(x) - T(x_0)}{T'(x_0)} \ . \qquad (3)$$

Para transformaciones de potencias $T(x) = x^p$, la expresión de una transformación apareada es sencilla. Como

$$T'(x_0) = p x_0^{p-1} \ ,$$

Tenemos

$$z = x_0 + \frac{x^p - x_0^p}{p x_0^{p-1}} \ .$$

El caso especial de $p = 0$, corresponde a la transformación logaritmo.

Las transformaciones logaritmo pueden realizarse

i) en base *e* (ln)

ó

ii) en base 10 ($\log_{10}$).

Consideraremos las dos opciones simultáneamente.

La transformación original es

$$i)\ T(x) = \ln(x)$$

ó

$$ii)\ T(x) = \log_{10}(x) = \log_{10}(e) \cdot \ln(x)$$

De manera que

$$i)\ T'(x) = \frac{1}{x}$$

ó

$$ii)\ T'(x) = \log_{10}(e)\frac{1}{x}$$

Recordemos que la expresión apareada en $x = x_0$ es

$$z = x_0 + \frac{T(x) - T(x_0)}{T'(x_0)} \quad . \qquad (3)$$

Por lo tanto, en el caso del logaritmo resulta

$$i)\, z = x_0 + x_0(\ln(x) - \ln(x_0))$$
ó
$$ii)\, z = x_0 + x_0 \frac{\log_{10}(x) - \log_{10}(x_0)}{\log_{10}(e)}$$

Equivalentemente

$$i)\, z = x_0(1 - \ln(x_0)) + x_0 \ln(x)$$
ó $$(4)$$
$$ii)\, z = x_0(1 - \frac{\log_{10}(x_0)}{\log_{10}(e)}) + x_0 \frac{\log_{10}(x)}{\log_{10}(e)}$$

Ejemplo: Consideremos los datos de la población de las ciudades más grandes de USA.

Utilizamos la ecuación ii) (4) para el ejemplo tomando

x_0 = 88, la mediana del lote y la transformación $\log_{10}$.

$$z = 88(1 - \frac{\log_{10}(88)}{0.4343}) + \frac{88}{0.4343}\log_{10}(x)$$
$$= -306.0 + 202.6\log_{10}(x)$$

Hemos elegido el punto de apareamiento igual a la mediana de las poblaciones, 88, es decir que

$$88 = a + b \log_{10} 88$$

obteniendo

$$z = 202.6 \log_{10} x - 306.0.$$

La transformación más simple

$$z = 200 \log10 x - 300$$

produce un resultado similar y es la utilizamos en el ejemplo.

Tabla 20

	Pobl.	$\log_{10}$(Pobl.)	200$\log_{10}$(Pobl.)-300
New York	778	2.891	278.196
Chicago	355	2.550	210.046
Los Angeles	248	2.394	178.890
Philadelphia	200	2.301	160.206
Detroit	167	2.223	144.543
Baltimore	94	1.973	94.626
Houston	94	1.973	94.626
Cleveland	88	1.944	88.897
Washington, DC	76	1.881	76.163
St. Louis	75	1.875	75.012
Milwakee	74	1.869	73.846
San Francisco	74	1.869	73.846
Boston	70	1.845	69.020
Dallas	68	1.833	66.502
New Orleans	63	1.799	59.868

La tabla 20 muestra los datos originales y los datos transformados.

La mayoría de los valores transformados han cambiado poco por la reexpresión

$$z = 200 \log_{10} x - 300.$$

esto es que *z* se parece a *x.*

Solamente las poblaciones de las ciudades más grandes--New York, Chicago, Los Angeles, Philadelphia, y Detroit han sido alteradas sustancialmente acercando sus valores a los de las demás.

Figura 22. Log de la población versus la población de ciudades de USA

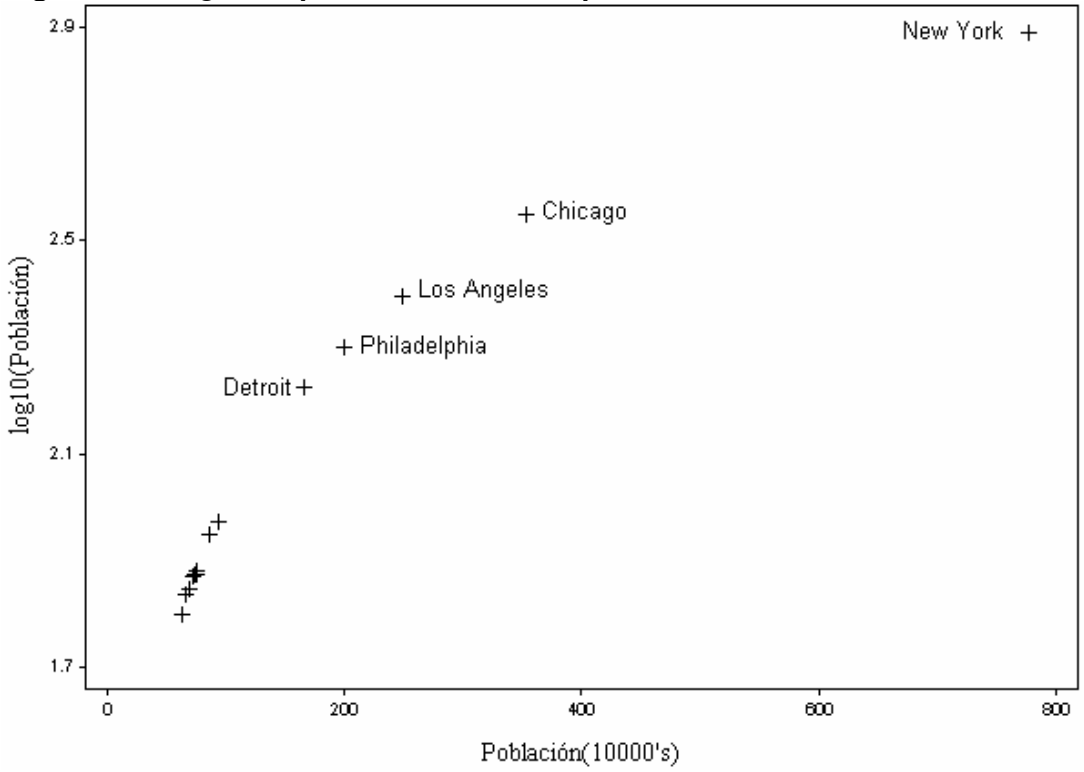
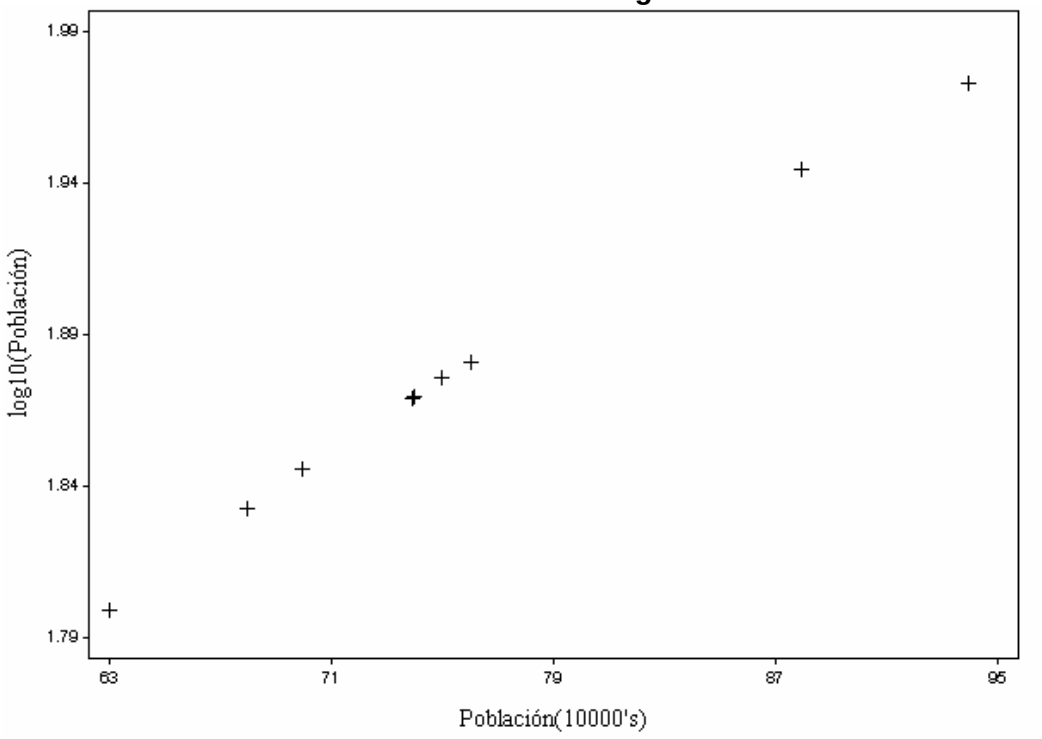


FIGURA 23. Eliminando las ciudades grandes la relación es casi lineal



Las figuras 22 y 23 muestran que el logaritmo de la población es cercano a una transformación lineal de la población si excluimos las ciudades más grandes.

Ventajas de la reexpresión apareada.

Muchas de las técnicas que utilizamos en análisis de datos son transparentes a transformaciones lineales de la escala. Esto significa que si multiplicamos el dato *y* por una constante *d* y le sumamos un valor *c*, produciendo $z = c + dy$ los resultados cambian de $\underline{r}$ a $\underline{c + d\,r}$ ó $\underline{d\,r}$ ó se mantienen inalterados.

Por ejemplo: la mediana cambia a $c + dM$, la distancia intercuartos ó la distancia intercuartil es multiplicada por *d* y un test de *t* no cambia luego de realizar una transformación lineal.

En el ejemplo anterior vemos los beneficios.

- Primero: que los datos transformados tengan un aspecto muy similar, para la mayor parte, que los datos originales.

  Solamente los valores extremos cambian sustancialmente. Esto valdrá mientras que:

  (a) la transformación sea aproximadamente lineal en zona central del lote y

  (b) realicemos el apareamiento en un punto cercano al centro.

  Generalmente ya estamos acostumbrados a las magnitudes en la escala original y el apareamiento nos resulta más conveniente que modificar nuestra forma de pensar. Los datos transformados en el ejemplo tienen el comportamiento del logaritmo de la población pero el aspecto de los datos originales.

- Segundo: el apareamiento enfatiza los cambios debidos a la transformación. Solamente los datos de

Detroit y las ciudades más grandes fueron modificados sustancialmente, a mayor tamaño mayor cambio.

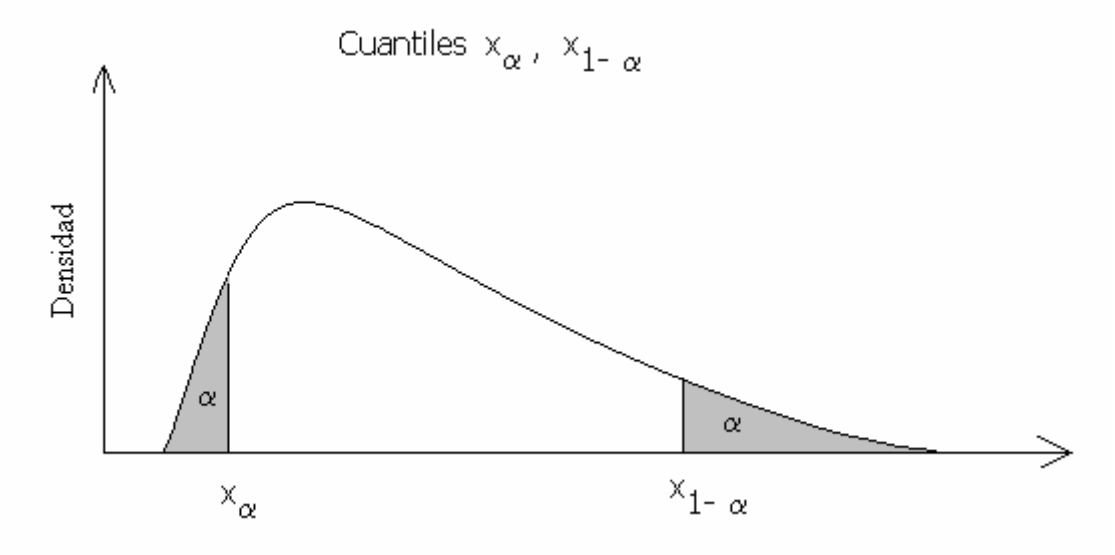
- Tercero: (que no se ve en este ejemplo) el apareamiento permite comparar los efectos de diferentes transformaciones. Veremos otro ejemplo de esta ventaja cuando veamos la transformación para simetría.

SIMETRIZACIÓN DE LOS DATOS

La simetría es una propiedad frecuentemente deseable en un conjunto de datos. Muchos estimadores de posición (posición del centro de los datos) tienen un mejor comportamiento y son más fáciles de interpretar cuando los datos son simétricos.

Como hicimos en el caso del gráfico de dispersión-nivel, utilizaremos un gráfico para elegir una transformación dentro de la familia de las transformaciones de potencias que simetrice el lote. Con ese objetivo presentamos los *valores letra*.

Los **valores letra** inferior y superior son los cuantiles que tienen potencias de ½ como valores de α .



La letra M corresponde a la mediana con $\alpha = ½$,
la letra F corresponde a los cuartos (fourths) con $\alpha = (½)^2$ y
la letra E corresponde a los octavos (eights) con $\alpha = (½)$.

Se sigue en el orden alfabético, observando que E precede inmediatamente a F, con D hacia atrás hasta A y dando la vuelta con Z, Y, X,...

Tabla 22

VALORES LETRA	α
M	1/2
F	$(1/2)^2$
E	$(1/2)^3$
D	$(1/2)^4$
C	$(1/2)^5$
B	$(1/2)^6$
A	$(1/2)^7$
Z	$(1/2)^8$
Y	$(1/2)^9$
X	$(1/2)^{10}$

La tabla 22 muestra la relación entre los valores letra y los valores α que en las distribuciones continuas corresponde al área que deja a izquierda bajo la curva de densidad el correspondiente cuantil.

Salvo en el caso de la mediana, para cada letra tenemos *un par* de valores letra: el valor letra inferior y el valor letra superior.

Los extremos del lote se identifican por su profundidad: 1.

Profundidad de los valores letra

En lotes grandes puede ser útil resumir los datos con un poco más de detalle. Las cinco medidas resumen: mín, cuarto inferior, mediana, cuarto superior y máximo se pueden acomodar agregar dos valores más los octavos (eights). Se determinan por la siguiente profundidad:

$$\text{profundidad del octavo} = \frac{[\text{profundidad del cuarto}] + 1}{2}$$

en general, se define

$$\text{profundidad de un valor letra} = \frac{[\text{profundidad del anterior}] + 1}{2}$$

Resumen medio - mid summary

Una manera simple de chequear simetría consiste en definir una medida de tendencia central (resumen medio, midsummary) para cada par de valores letra. Cada resumen medio es el promedio de los dos valores letra correspondientes y se identifican por medF, medE, medD, medC, ..., etc.

Otro nombre, rango medio, es utilizado para el promedio de los extremos.

Una vez que hemos calculado los resúmenes medios para todos los pares de valores letra, podemos examinarlos para *hallar evidencia sobre una asimetría sistemática*.

Si la asimetría aparente es debida a uno o dos valores desviados, solamente serán afectados los valores letra más extremos. Por lo tanto el conjunto completo de resúmenes medios provee mayor resistencia.

En un lote

i) perfectamente simétrico: todos los resumen medios serían iguales a la mediana.

ii) asimétrico a derecha: los resumen medios aumentarían a medida que provinieran de valores letra correspondientes a α's más pequeños.

iii) asimétrico a izquierda: decrecerían.

EJEMPLO: Ingreso familiar

La tabla 23 muestra los valores letra correspondientes a los ingresos anuales de una muestra de 994 familias. Los cálculos para los resumen medios son inmediatos:

$$\text{medF} = \tfrac{1}{2}\,(2412 + 4944) = 3678$$

y así siguiendo.

En la escala original los valores de los meds aumentan sostenidamente a medida que descendemos en la

columna; la medY es 1.6 veces la mediana. Este comportamiento es una evidencia sólida de asimetría importante. Únicamente cuando pasamos de medY a la media de los extremos no observamos un incremento.

Tabla 23. Despliegue de los valores letra y los resúmenes medios para una muestra de ingresos familiares; en la escala original y en la escala logarítmica (base 10).

		Escala original			Escala logarítmica		
#	994	Ingresos familiares en dólares			Ingresos familiares		
M	497.5		3480.0			3.54	
F	249	2412.0	3678.0	4944.0	3.38	3.54	3.69
E	125	1788.0	4115.5	6443.0	3.25	3.53	3.81
D	63	1517.0	4400.5	7284.0	3.18	3.52	3.86
C	32	1248.0	4799.0	8350.0	3.10	3.51	3.92
B	16.5	963.5	4978.8	8994.0	2.98	3.47	3.95
A	8.5	727.5	5241.0	9754.5	2.86	3.43	3.99
Z	4.5	579.0	5394.5	10210.0	2.76	3.39	4.01
Y	2.5	345.0	5510.3	10675.5	2.54	3.28	4.03
	1	114.0	5494.0	10874.0	2.06	3.05	4.04

Vemos:

- En los datos originales, asimetría a derecha
- En los datos reexpresados, una tendencia a la asimetría a izquierda, como lo muestra la tendencia de decrecimiento monótono de los resumen medios.

Intentamos hallar otra transformación que logre una simetría mayor en este lote.

Cálculo de las profundidades de los valores letra

Ejemplo: Ingreso familiar

```
> nombres <- c( "M",
  rev(LETTERS[1:6]),rev(LETTERS))
> tot <- 994
> k <- floor(logb(tot, 2))
> prof <- vector(mode = "numeric", length = k)
> names(prof) <- nombres[1:k]
> n <- tot
```

Profundidad de los valores letra de acuerdo con la definición

([profundidad del valor letra anterior] + 1) / 2

```
>   for(i in 1:k) {
+ n <- (floor(n) + 1)/2
+ prof[i] <- n
+ }
>  prof
    M     F     E     D     C     B     A     Z     Y
497.5 249.0 125.0  63.0  32.0  16.5   8.5   4.5   2.5
```

Profundidad de los cuantiles de acuerdo con la definición α * (n+1) tal como se utiliza para los cuartiles

```
> profc <- round((tot + 1) * (1/2)^(1:k), 2)
```

Profundidad de los cuantiles de acuerdo con la definición de la función `quantile` de S: α * (n-1) +1

```
> profq <- round((tot - 1) * (1/2)^(1:k) +
  1, 2)
> rbind(prof, profc, profq)
```

```
          M      F      E     D     C     B    A    Z    Y
 prof 497.5 249.00 125.00 63.00 32.00 16.50 8.50 4.50 2.50
profc 497.5 248.75 124.38 62.19 31.09 15.55 7.77 3.89 1.94
profq 497.5 249.25 125.12 63.06 32.03 16.52 8.76 4.88 2.94
```

GRÁFICO DE TRANSFORMACIÓN PARA SIMETRÍA

Sea *M* la mediana de un lote. Representemos por x_I y x_S a varios valores letra inferior y superior, se grafica

$$\frac{(x_S - M)^2 + (M - x_I)^2}{4M} \qquad (5)$$

en el eje horizontal y

$$\frac{x_S + x_L}{2} - M \qquad (6)$$

en el eje vertical. Si el gráfico resultante es aproximadamente lineal, uno menos la pendiente es la potencia indicada de una transformación de la forma

$$T(x) = kx^p.$$

Como los otros gráficos de diagnóstico, el objetivo de este gráfico es obtener, en primera aproximación, una buena elección de la transformación.

Observemos que la expresión (6) mide la distancia entre la mediana y un resumen medio, esto para cada valor letra. Estas diferencias deberían ser todas 0 en un lote perfectamente simétrico.

Cuando *p* es positivo *k g*eneralmente se elige igual a 1 y cuando *p* es negativo *k es* igual a -1. Cuando *p* = 0 se utiliza la transformación logarítmica.

Desarrollo.

Nos restringiremos a datos positivos. Esta condición se cumple automáticamente con cantidades o recuentos. Otros tipos de datos deben modificarse sumando una constante.

Indicamos por $x_1, x_2, \ldots, x_n$ al conjunto de datos de un lote. Buscamos una potencia, *p,* para la cual el lote transformado $x_1^p, x_2^p, \ldots, x_n^p$ es aproximadamente simétrico.

Si el lote transformado es simétrico, entonces para todos los valores letra tendremos que

$$x_S^p - M^p = M^p - x_I^p. \qquad (7)$$

Esto puede re escribirse como

$$\frac{x_S^p + x_I^p}{2} = M^p. \qquad (8)$$

Para *p* distinto de cero, un desarrollo de Taylor de orden 2 de x_S^p y x_L^p alrededor de *M* da:

$$x_S^p \approx M^p + pM^{p-1}(x_S - M) + \frac{p(p-1)}{2} M^{p-2}(x_S - M)^2$$

y

$$x_I^p \approx M^p + pM^{p-1}(x_I - M) + \frac{p(p-1)}{2} M^{p-2}(x_I - M)^2 \ .$$

Sustituyendo estas aproximaciones en la ecuación (8) obtenemos

$$\frac{1}{2}\left\{2M^p + pM^{p-1}(x_S + x_I - 2M)\right.$$

$$\left. + \frac{p(p-1)}{2} M^{p-2}\left[(x_S - M)^2 + (x_I - M)^2\right]\right\} \approx M^p.$$

Reagrupación y simplificación algebraica da:

$$M(x_S + x_I - 2M) + \frac{p-1}{2}\left[(x_S - M)^2 + (x_I - M)^2\right] \approx 0$$

$$\frac{x_S + x_I}{2} - M \approx (1-p)\frac{\left[(x_S - M)^2 + (M - x_I)^2\right]}{4M} \qquad (9)$$

La aproximación (9) sugiere una transformación cuyo objetivo es simetrizar un lote. El miembro izquierdo

representa la distancia del resumen medio a la mediana, para varios cuantiles.

Como hemos visto, esta diferencia es cero en un lote simétrico y resulta una recta horizontal. Luego la pendiente es 0 y *p* es 1. Por lo tanto ninguna transformación es necesaria ni sugerida por el gráfico.

Como estamos tratando datos positivos, *M* es positivo y la expresión

$$\frac{(x_S - M)^2 + (M - x_I)^2}{4M}$$

es no decreciente a medida que los pares de valores letra se alejan de la mediana hacia los extremos.

Luego si los meds aumentan, como lo hacen en un lote asimétrico a derecha, 1 - *p* será positivo y por lo tanto *p* <1. Tal transformación trae las colas hacia adentro y el lote transformado se vuelve más simétrico. Transformaremos con potencias bajas cuando los resumen medios aumentan desde la mediana hacia afuera.

EJEMPLO: Ingreso familiar- continuación.
Las coordenadas para construir el gráfico de transformación para simetría se presenta en la tabla 24.

La aproximación (9) nos permite estimar p para cada valor letra, estos estimadores (6ta. columna). Observamos que estos valores cambian a medida que cambiamos los valores letra; sin embargo son coherentes con la observación de que los datos presentan un alejamiento de la simetría que la transformación logaritmo no puede corregir.

Tabla 24. Cálculos para la construcción del gráfico para la transformación de simetría en el ejemplo de los ingresos familiares.

Valor letra	x_I	x_S	$\frac{x_S + x_L}{2} - M$	$\frac{(x_S - M)^2 + (M - x_I)^2}{4M}$
F	**2412.0**	**4944.0**	**198.0**	**235.9**
E	**1788.0**	**6443.0**	**635.5**	**836.4**
D	**1517.0**	**7284.0**	**920.5**	**1316.4**
C	**1248.0**	**8350.0**	**1319.0**	**2061.7**
B	**963.5**	**8994.0**	**1498.8**	**2639.2**
A	**727.5**	**9754.5**	**1761.0**	**3372.5**
Z	**579.0**	**10210.0**	**1914.5**	**3858.4**
Y	**345.0**	**10675.5**	**2030.3**	**4425.5**

El gráfico para obtener la transformación de simetría aparece en la figura 24. Hay muchas maneras de ajustarle una recta.

Queremos ajustar una recta que pase por el origen, con un método resistente, es decir con un método que no sea afectado fuertemente por unos pocos puntos que se desvíen considerablemente del patrón del resto de los puntos.

Una manera rápida de hacer esto es: considerar las ocho rectas que pasan por el origen y cada uno de los puntos y luego elegir la recta cuya pendiente sea la mediana de las ocho pendientes.

Los puntos que corresponden a valores letra internos indican una mayor asimetría que los puntos para valores letra más cercanos a las colas del lote.

Este enfoque da una pendiente de 0.60 y la recta resultante es la que se muestra en la figura 24.

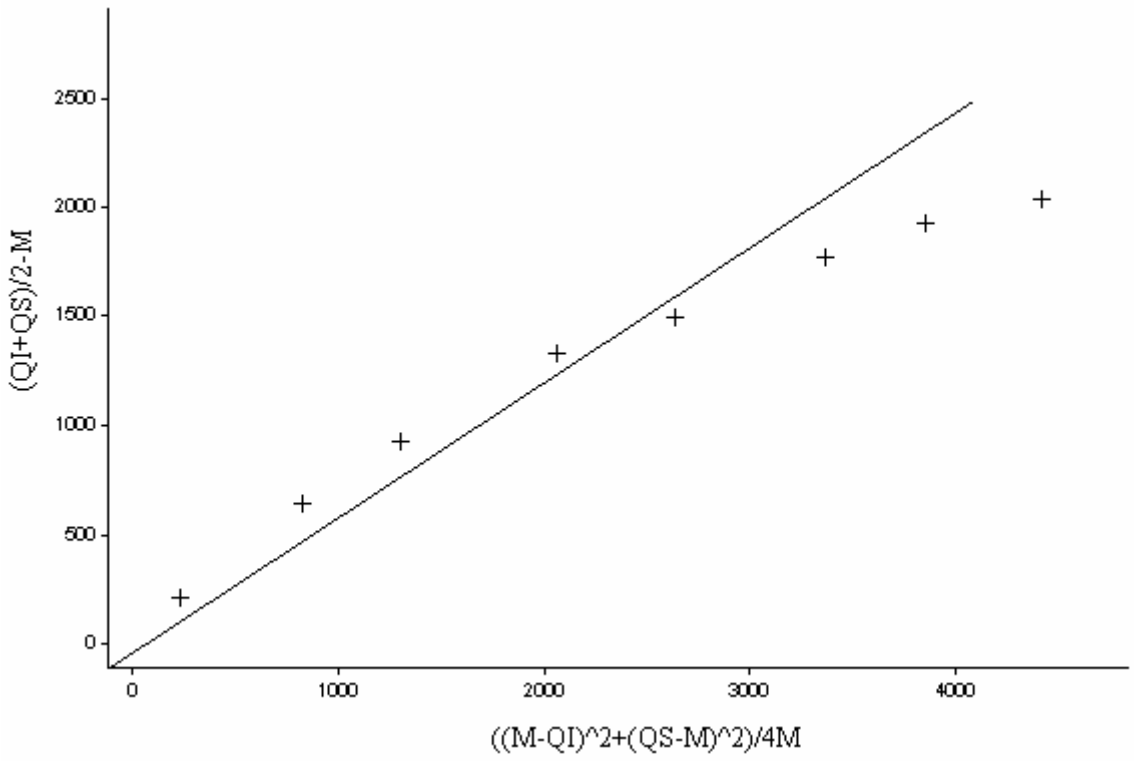
	$\frac{x_S + x_L}{2} - M$	/	$\frac{(x_S - M)^2 + (M - x_I)^2}{4M}$		Pendientes	Estimador de p
F	**198.0**	**/**	**235.9**	**=**	**0.84**	**0.16**
E	**635.5**	**/**	**836.4**	**=**	**0.76**	**0.24**
D	**920.5**	**/**	**1316.4**	**=**	**0.70**	**0.30**
C	**1319.0**	**/**	**2061.7**	**=**	**0.64**	**0.36**
B	**1498.8**	**/**	**2639.2**	**=**	**0.57**	**0.43**
A	**1761.0**	**/**	**3372.5**	**=**	**0.52**	**0.48**
Z	**1914.5**	**/**	**3858.4**	**=**	**0.50**	**0.50**
Y	**2030.3**	**/**	**4425.5**	**=**	**0.45**	**0.54**

La relación

potencia = 1- pendiente

indica una potencia de 0.40 para la reexpresión con una transformación de potencia (Observar que es equivalente tomar la mediana de las potencias estimadas).

Figura 24. Gráfico para simetría en el ejemplo de los ingresos familiares.



En la práctica, no usamos frecuentemente potencias tales como 0.40, redondeamos el exponente a la potencia semientera más próxima;
aquí $p = ½$ es la más cercana. Si el lote resultante no fuese lo suficientemente simétrico podríamos probar con la transformación logaritmo (correspondiente a $p = 0$) ó quizás con la raíz cuarta (correspondiente a $p = ¼$).

Tabla 25. Despliegue de valores letra junto con los resumen medios para las escalas de raíz cuadrada y raíz cuarta en el ejemplo de los ingresos familiares.

	Escala raíz cuadrada Ingresos familiares			Escala raíz cuarta Ingresos familiares		
# 994						
M 497.5		58.1			7.68	
F 249	49.1	59.7	70.3	7.01	7.70	8.39
E 125	42.3	61.3	80.3	6.50	7.73	8.96
D 63	38.9	62.1	85.3	6.24	7.74	9.24
C 32	35.3	63.4	91.4	5.94	7.75	9.56
B 16.5	31.0	62.9	94.8	5.57	7.65	9.74
A 8.5	27.0	62.9	98.8	5.19	7.57	9.94
Z 4.5	24.1	62.6	101.0	4.91	7.48	10.1
Y 2.5	18.6	60.9	103.3	4.31	7.24	10.2
1	10.7	57.5	104.3	3.27	6.74	10.2

Comparación de la asimetría del lote original y de tres lotes transformados

- Los datos crudos tienen una tendencia monótona fuerte de crecimiento en los meds hasta el anteúltimo valor letra. Esta tendencia indica asimetría a derecha en la distribución del ingreso.

- El logaritmo del ingreso muestra una tendencia monótona decreciente en los resumen medios indicando que éste tiene una distribución asimétrica a izquierda. A pesar de que este patrón es persistente, es leve para los primeros valores letra y más acentuado a medida que avanzamos hacia las colas. Si nuestro interés principal es la simetría en la parte principal del lote y no en los extremos,

podríamos estar satisfechos con la transformación logarítmica. Más aún, la experiencia en economía ha mostrado que las distribuciones de los ingresos pueden frecuentemente ser modeladas por la distribución log normal (es decir que la distribución del ln de los ingresos es aproximadamente normal). Si esto fuera cierto en este caso, la aplicación logarítmica daría "gaussianidad" y por lo tanto simetría.

- La transformación raíz cuadrada produce resumen medios que son monótonos crecientes hasta la letra *C* y monótonos decrecientes desde allí hasta los extremos. Todos los resúmenes son mayores que la mediana salvo el último, de manera que podemos pensar que este lote conserva un poco de la asimetría a derecha presente en los datos originales.

- La transformación raíz cuarta también produce resumen medios que son monótonos crecientes hasta la letra *C* y monótonos decrecientes desde allí hasta los extremos, pero en este caso están bien balanceados ya que cuatro son mayores que la mediana y cinco son menores.

Ninguna de las transformaciones de potencia puede simetrizar el lote completamente debido a la mayor asimetría cerca del centro del lote y menor asimetría hacia los extremos. Esto se puede ver nítidamente en el gráfico de transformación para simetría de la figura 24 si consideramos la pendiente de los primeros cuatro puntos por separado de la pendiente de los últimos cinco. De los primeros cuatro tendríamos una pendiente claramente mayor que de los cinco últimos.

La elección de la transformación es en parte una cuestión de juicio personal, especialmente cuando interesa la simetría. varios factores pueden ayudarnos para realizar la elección en el ejemplo de los ingresos familiares.

1. Si interesa la simetría en la parte principal de los datos pero la asimetría en las colas es relativamente menos importante, probablemente preferiríamos la transformación logaritmo. Los economistas pueden favorecerla por su facilidad en la interpretación.

2. Si es importante obtener simetría en las colas , podríamos preferir la transformación raíz cuadrada como un forma fácil de reexpresar los datos.

3. Si queremos que la reexpresión balancee el grado de dispersión de la parte principal de la distribución contra la más modesta de los extremos, podríamos elegir la raíz cuarta.

Cuando nos interesa comparar varias transformaciones de un mismo conjunto de datos es mejor utilizar **transformaciones apareadas.**

Tabla 26. Datos de ingresos familiares y tres transformaciones apareadas. Despliegue de valores letra y resumen medios.

	Datos originales			Log		
M		3480.0			3480.0	
F	2412.0	3678.0	4944.0	2201.0	3449.8	4698.6
E	1788.0	4115.5	6443.0	1159.2	3389.7	5620.1
D	1517.0	4400.5	7284.0	587.3	3317.2	6047.1
C	1248.0	4799.0	8350.0	-92.0	3215.2	6522.4
B	963.5	4978.8	8994.0	-992.3	2894.3	6780.9
A	727.5	5241.0	9754.5	-1970.1	2546.7	7063.4
Z	579.0	5394.5	10210.0	-2764.6	2228.8	7222.2
Y	345.0	5510.3	10675.5	-4566.3	1405.5	7377.3
1	114.0	5494.0	10874.0	-8419.8	-489.2	7441.5

Tabla 26 Continuación. Datos de ingresos familiares y tres transformaciones apareadas. Despliegue de valores letra y resumen medios.

	Raíz cuadrada			Raíz cuarta		
M		3480.0			3480.0	
F	2314.6	3565.3	4816.0	2260.4	3508.4	4756.4
E	1509.1	3749.8	5990.4	1344.6	3570.5	5796.5
D	1115.5	3852.5	6589.5	870.2	3586.2	6302.2
C	688.2	3994.7	7301.2	331.5	3607.6	6883.8
B	182.5	3945.8	7709.2	-343.2	3432.7	7208.5
A	-297.5	3937.6	8172.6	-1028.0	3271.2	7570.3
Z	-640.8	3900.4	8441.6	-1550.2	3113.4	7777.0
Y	-1288.3	3711.0	8710.3	-2629.5	2675.8	7981.2
1	-2220.0	3301.6	8823.1	-4518.2	1774.0	8066.2

Recordando que las transformaciones apareadas en x0 se obtienen con

$$z = x_0 + \frac{x^p - x_0^p}{p x_0^{p-1}} \text{ si p } \neq 0$$

$$i)\, z = x_0 + x_0(\ln(x) - \ln(x_0))$$

ó si p = 0

$$ii)\, z = x_0 + x_0 \frac{\log_{10}(x) - \log_{10}(x_0)}{\log_{10}(e)}$$

Las transformaciones apareadas en la mediana de los datos originales son:

$$\text{"log"} = 3480 + \frac{\log_{10}(\text{ingreso}) - 3.542}{0.0001248},$$

$$\text{"raíz cuadrada"} = 3480 + \frac{(\text{ingreso})^{1/2} - 58.99}{0.008476},$$

$$\text{"raíz cuarta"} = 3480 + \frac{(\text{ingreso})^{1/4} - 7.681}{0.0005518},$$

- Para cada transformación, la mediana del lote transformado es 3480 y los valores cercanos a la mediana son modificados lo menos posible.

- Los valores transformados son una transformación lineal de logaritmos y raíces, pueden (y lo hacen) tomar valores negativos.

- Cada transformación acerca los valores más altos hacia el centro y aleja los valores más bajos hacia afuera algunas veces fuera de lo que podría se la cota natural del cero.

Para comparar directamente la simetría de los lotes transformados, restamos la mediana, 3480, a cada una de los resumen medios.
En un lote perfectamente simétrico estas diferencias son cero.

Tabla 27. Resumen medios menos la mediana para los datos del ingreso familiar y tres reexpresiones apareadas.

	Datos Originales	Log	Raíz Cuarta	Raíz Cuadrada
F	198	-30	28	85
E	636	-90	91	270
D	921	-163	106	373
C	1319	-265	128	515
B	1499	-586	-47	466
A	1761	-933	-209	458
Z	1915	-1251	-367	420
Y	2030	-2074	-804	231

Para comparar el efecto de las diferentes transformaciones la figura 25 muestra el gráfico de los *resumen medios menos la mediana* en las escalas *log, raíz cuadrada y raíz cuarta* dados en la tabla 27 contra los valores de

$$\frac{(x_S - M)^2 + (M - x_I)^2}{4M}$$

en la escala original.

¡OJO! En el eje horizontal
se toma SIEMPRE la escala original

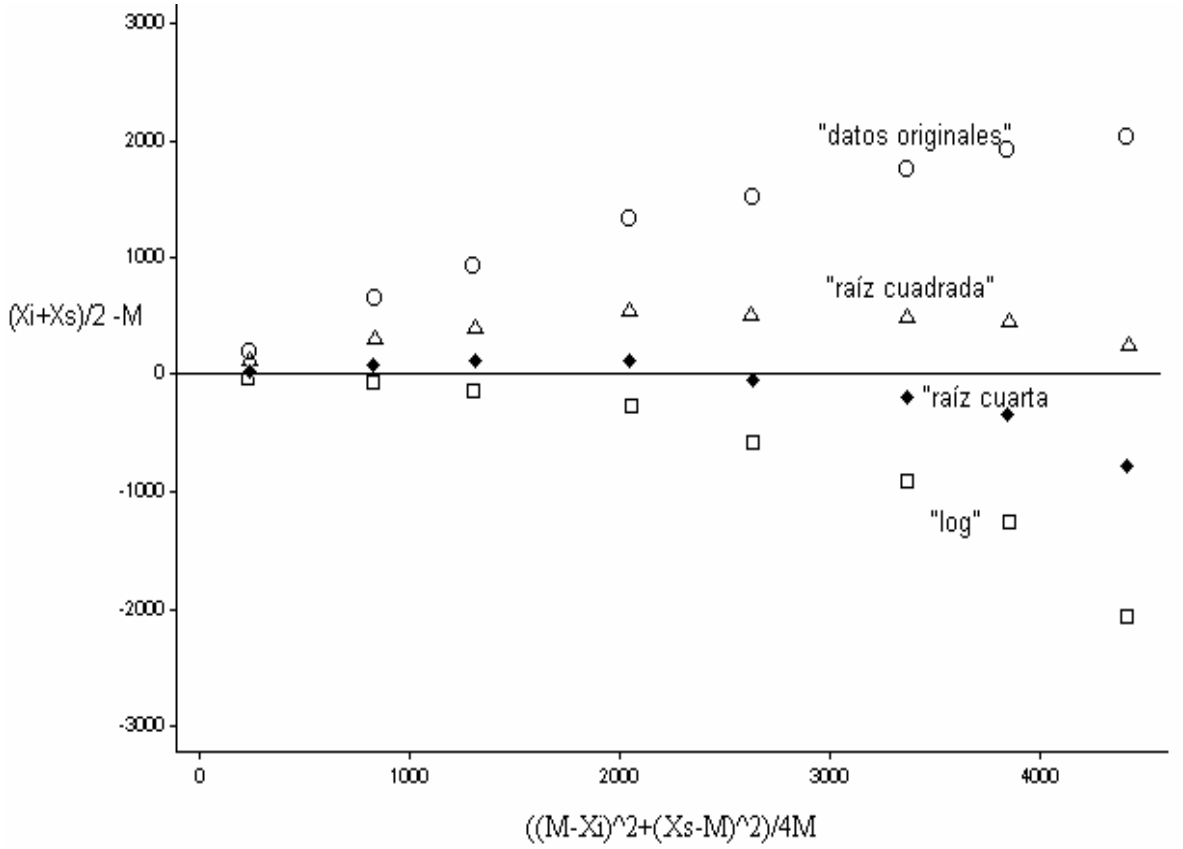


Figura 25. Medidas resumen para los datos del ingreso familiar reexpresados por transformaciones apareadas

- Los resumen medios correspondientes a la raíz cuarta se encuentran entre los del log y la raíz cuadrada. Esto se debe a la monotonía de las funciones de potencia apareadas respecto de la potencia (0< ¼ < ½), Tukey dice: "esto corresponde a la posición relativa de las transformaciones en la escalera de potencias".

- La transformación raíz cuarta esté probablemente cerca de la transformación óptima. Una transformación con *p* entre .25 y.5, quizás .40 como fue estimado antes, podría disminuir los residuos para los valores letra extremos pero a costa de aumentar los residuos positivos para los valores letra más centrales.

La escala común provista por las transformaciones apareadas facilita ambas observaciones.

APAREANDO VALORES ESPERADOS

La mediana de un lote transformado es igual a la transformación de la mediana del lote (cuando n es impar y la transformación es monótona) pero la media muestral no satisface una relación similar. Esto es, en general:

$$\frac{\sum T(x_i)}{n} \neq T(\bar{x})$$

para la v.a. X, o sea si E indica el valor esperado:

$$E[T(X)] \neq T(E[X])$$

El apareamiento no puede eliminar estas diferencias pero las puede reducir. Consideremos la transformación T apareada en x_0:

$$Z = a + b\,T(X)$$

con

$$a = x_0 - \frac{T(x_0)}{T'(x_0)}$$

$$b = \frac{1}{T'(x_0)}.$$

Si aproximamos $T(X)$ por un desarrollo de Taylor alrededor de x_0, un valor cercano a la media de X, hasta el término cuadrático y tomamos esperanza tendremos:

$$E(Z) \approx E(X) + \frac{1}{2}\text{Var}(X)\frac{T''(x_0)}{T'(x_0)} \qquad (10)$$

Si Z es una transformación apareada, de X en un punto cercano a la media de X, entonces la media de Z es aproximadamente igual a la media de X más un término que depende de la dispersión de X y la curvatura de la transformación.

Para las transformaciones de potencia incluido el logaritmo

$$\frac{T''(x_0)}{T'(x_0)} = \frac{p-1}{x_0}$$

Esta identidad simplifica aún más la expresión (10).

Hemos visto las siguientes ventajas de las transformaciones apareadas:

- Los datos reexpresados involucran números de tamaño familiar, algunos de los cuales quedan casi inalterados.
- Podemos ver el efecto de la transformación sobre cada uno de los puntos.
- Al comparar diferentes reexpresiones para un único lote, el apareamiento facilita la comparación al poner los datos en una escala común.

- El apareamiento en el valor esperado de una variable aleatoria permite que el valor esperado de la variable transformada sea aproximadamente el valor esperado de la variable sin transformar. El grado de aproximación depende de la dispersión de la variable en escala original y la curvatura de la reexpresión.

Observación final:

Una regla para decidir si una reexpresión cambiará la forma de la distribución de los datos está basada en el cociente siguiente:

valor mayor / valor menor

La transformación podría ser útil si este cociente es grande, digamos, mayor que 20. Pero inútil, si es menor que 2.

Para el ejemplo de las grandes ciudades los cocientes varían de 4 a 27 para los 16 lotes y para el conjunto total es de aproximadamente 160. Los datos transformados fueron beneficiosos en este ejemplo.