JUSTIFICACIÓN DEL GRÁFICO DE DISPERSIÓN-NIVEL

Suponemos

- las distribuciónes subyacentes a nuestros lotes tienen formas similares
- los lotes se encuentren a niveles diferentes
- los lotes tienen dispersiones diferentes

Si X_j es una variable aleatoria que representa una observación del lote *i*, el nivel de X_i depende del lote particular representado.

Eliminamos de la notación esta dependencia de X del nivel y utilizamos *X* para indicar una variable aleatoria correspondiente a cualquier lote.

Pero tenemos en cuenta que *X* en realidad representa varias variables con diferentes niveles y dispersiones.

Consideremos las transformaciones de potencia X^p , donde *p* es cualquier número real fijo.

Supongamos que para algún número real *p*, la distancia intercuartos (ó, equivalentemente, intercuartil) es constante entre todos los lotes a diferentes niveles de *X*.

Establecemos los supuestos y la notación:

Para X^p

$$\begin{aligned} \text{mediana} &= m && (m > 0) \\ \text{cuarto superior} &= m + d && (d > 0) \\ \text{cuarto inferior} &= m - c && (c > 0) \\ \text{distancia intercuartos} &= d + c && (\text{cte., indep. de } m) \end{aligned} \qquad (3)$$

Para X

$$\begin{aligned} \text{mediana} &= m^{1/p} \\ \text{cuarto superior} &= (m + d)^{1/p} \\ \text{cuarto inferior} &= (m - c)^{1/p} \\ \text{distancia intercuartos} &= (m + d)^{1/p} - (m - c)^{1/p} \end{aligned} \qquad (4)$$

Queremos hallar una expresión (aproximadamente) lineal entre

- el log de la distancia intercuartos y
- el log de la mediana

de los datos sin transformar, en la que aparezca explícitamente el valor *p* que hace que la correspondiente transformación de potencias estabilice la dispersión como en (3).

Debemos suponer que *c* es menor que *m*, en caso contrario el cuarto inferior de los datos crudos, que son positivos, sería negativo.

En realidad esperamos que *c* y *d* no sean mayores que *m* / 2. De esta manera es razonable expandir la expresión de la distancia intercuartos en función de *d* / *m* y *c* / *m.*

Por simplicidad, sea $q = 1 / p$, entonces si desarrollamos la función x^q en serie de Taylor alrededor del punto x = 1 y evaluamos la función en $1+ d / m$ y $1 - c / m$ obtenemos las expansiones que nos interesan:

$$(m+d)^q - (m-c)^q$$

$$= m^q \left[\left(1+\frac{d}{m}\right)^q - \left(1-\frac{c}{m}\right)^q \right]$$

$$= m^q \left[1 + q\frac{d}{m} + \frac{q(q-1)}{2}\left(\frac{d}{m}\right)^2 + \frac{q(q-1)(q-2)}{6}\left(\frac{d}{m}\right)^3 + \cdots \right.$$

$$\left. -1 + q\frac{c}{m} - \frac{q(q-1)}{2}\left(\frac{c}{m}\right)^2 + \frac{q(q-1)(q-2)}{6}\left(\frac{c}{m}\right)^3 + \cdots \right]$$

$$= m^{q-1} \left[q(d+c) + \frac{q(q-1)}{2}\frac{d^2 - c^2}{m} \right.$$

$$\left. + \frac{q(q-1)(q-2)}{6}\frac{d^3 - c^3}{m^2} + \cdots \right]$$

Luego

$$(m+d)^q - (m-c)^q = q(d+c)m^{q-1}\left[1 + \frac{(q-1)}{2}\frac{(d-c)}{m}\right.$$

$$\left. + \frac{(q-1)(q-2)}{6}\frac{(d^2 - dc + c^2)}{m^2} + \cdots \right] \qquad (5)$$

El término principal es $q(d+c)m^{q-1}$.

log(término ppal.) = log q +log (d + c) + (q -1)log m.

Como el último sumando puede escribirse

$$(q\text{ -1})\log m = [(q\text{ - 1})/(1/p)]\log (m^{1/p}),$$

luego

- log(término ppal.) es lineal en
- log(mediana X) = log ($m^{1/p}$),

con pendiente

$$\frac{(q-1)}{1/\,p} = p\left(\frac{1}{p}-1\right)=1-p$$

Luego tendríamos que tomar

$$p \approx 1- \text{ pendiente}$$

Es evidente que el desarrollo precedente es aproximado.

En el caso particular de q=2 (p =1/2) y c = d el término principal es exacto:

Para q = 2 en la ecuación (5) resulta

$$(m+d)^2-(m-c)^2=2(d+c)m+(d^2-c^2)$$

Cuando c = d obtenemos

$$(m+d)^2-(m-c)^2=4dm$$

El término principal es exacto en este caso.

En el caso particular del logaritmo debemos rehacer los cálculos.

Para log X

$$\begin{aligned} \text{mediana} &= m \\ \text{cuarto superior} &= m + d \\ \text{cuarto inferior} &= m - c \\ \text{distancia intercuartos} &= d + c \end{aligned} \qquad (6)$$

Para X

$$\begin{aligned} \text{mediana} &= 10^{m} \\ \text{cuarto superior} &= 10^{m + d} \\ \text{cuarto inferior} &= 10^{m - c} \\ \text{distancia intercuartos} &= 10^{m}(10^{d} - 10^{c}) \end{aligned} \qquad (7)$$

Luego

$$\begin{aligned} \log(\text{distancia intercuartos}) &= m + \log(10^{d} - 10^{c}) \\ &= \log(\text{mediana de } X) + \log(10^{d} - 10^{c}) \end{aligned}$$

es lineal en log(mediana de *X*) con pendiente exactamente igual a 1, no hemos realizado ninguna aproximación.

Nos interesa saber cuán precisa es la aproximación cuando basamos nuestra transformación en

solamente en el término principal del desarrollo (5), o sea cuán cercana es la expresión

$$\frac{\text{distancia intercuartos}}{\text{término principal de (5)}} = \frac{(m+d)^q - (m-c)^q}{q(d+c)m^{q-1}}$$

de 1.

La expresión anterior puede escribirse

$$\frac{(1+d/m)^q - (1-c/m)^q}{q(d/m + c/m)}$$

Como *d* es la distancia desde la mediana de X^p al cuarto superior de X^p, la fracción *d* / *m* es la relación entre una medida de dispersión y la mediana para X^p.
Análogamente, *c* es la distancia desde la mediana de X^p al cuarto inferior de X^p, de manera que *c*/*m* tiene una interpretación similar. Para distribuciones con los cuartos relativamente cerca de la mediana, en la escala transformada, (esto es *d/m* y *c/m* no demasiado grandes) esperamos que el término principal del desarrollo (5) de una buena aproximación. Veamos que ocurre numéricamente.

La tabla 18 muestra los valores de esta expresión para varias combinaciones de (*d* / *m, c* / *m*). Para *d* / *m* = .2 el término principal está cerca de la distancia intercuartos. En algunas combinaciones de *d*/*m* = .4 el cociente es preocupan te. Para valores mayores de *d*/*m*, la aproximación es frecuentemente inadecuada.

Tabla 18. Valores de la relación:
[dist. intercuartos de X]/[término ppal de (5)]

		(*d* / *m*, *c* / *m*)					
q	*p*	(.2, .2)	(.2,.1)	(.4,.4)	(.4,.2)	(.5,.5)	(.8,.5)
1	1	1.00	1.00	1.00	1.00	1.00	1.00
2	1-2	1.00	0.95	1.00	1.10	1.00	1.15
3	1/3	1.01	1.11	1.05	1.24	1.08	1.46
4	1/4	1.04	1.23	1.16	1.43	1.25	2.01
---	0	1.00	1.00	1.00	1.00	1.00	1.00
-2	-1/2	1.04	0.92	1.19	0.89	1.78	1.42
-1	-1	1.08	0.90	1.42	0.88	1.33	1.11
-1/2	-2	1.03	0.94	1.11	0.91	1.20	1.03

En resumen, podremos estar conformes con la aplicabilidad del gráfico dispersión-nivel cuando, en los datos transformados, los cuartos estén relativamente cerca de la mediana. Esto es que deben estar alejados de la mediana a una pequeña fracción de la mediana. La elección de *p* a partir de la pendiente del gráfico del log(dist. intercuartos de X) contra log(mediana) es un procedimiento aproximado, que con frecuencia es muy bueno.

El hecho que sea aproximado hace conveniente regraficar los pares

[log(mediana) , log(dist. intercuartos)]

para los datos transformados.

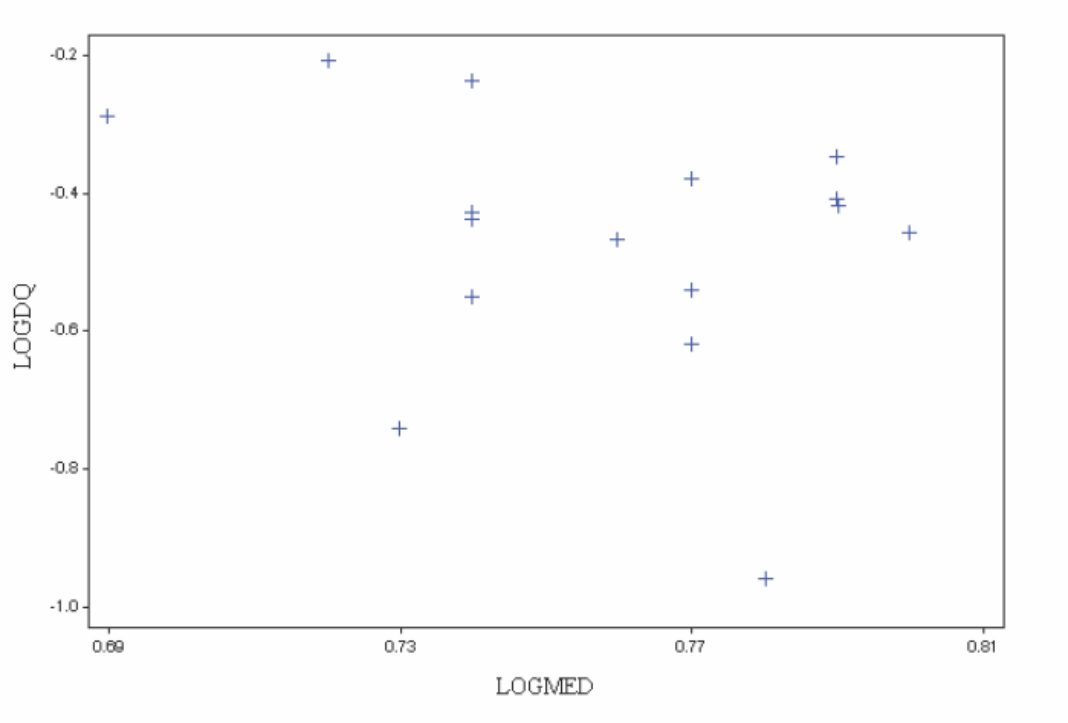
Tabla 19. Cálculos para el gráfico dispersión-versus-nivel para el log(población) de las 10 ciudades mayores en 16 países. (Se utilizaron logaritmos base 10.)

País	Mediana	d_Q	Log(*M*)	- Log(d_Q)
Sweden	4.94	.51	.69	.29
Netherlands	5.22	.62	.72	.21
Canada	5.43	.18	.73	.74
France	5.44	.36	.74	.44
Mexico	5.50	.57	.74	.24
Argentina	5.52	.37	.74	.43
Spain	5.54	.28	.74	.55
England	5.70	.34	.76	.47
Italy	5.84	.42	.77	.38
West Germany	5.85	.24	.77	.62
Brazil	5.91	.29	.77	.54
Soviet Union	6.04	.11	.78	.96
Japan	6.10	.39	.79	.41
United States	6.12	.45	.79	.35
India	6.22	.38	.79	.42
China	6.36	.35	.80	.46

Observación: Tukey recomienda repetir el gráfico de dispersión nivel para los datos transformados (figura 19). No debe confundirse este gráfico, con los gráficos de las figuras 16 y 17. En este gráfico se

toma logaritmo de la mediana y las distancias intercuartos con los datos ya transformados.

Figura 19. Gráfico de Dispersión -Nivel para los datos de las más grandes ciudades transformados por logaritmo



Se observa una leve tendencia hacia abajo: podría considerarse una transformación menos fuerte que el logaritmo.

Conclusiones:

- El gráfico de dispersión-nivel permite elegir una transformación de potencias para estabilizar la dispersión.
- Una dispersión estable puede ser esencial para continuar con el análisis (por ej. ANOVA)

- Los boxplots de los lotes de datos transformados permiten visualizar los efectos de la transformación y comparar las formas de los lotes en la nueva escala con los de los datos crudos**.**

Transformaciones.

Definición: Una *transformación* de un lote x_1, x_2, . . . ,x_n, es un función T que reemplaza cada x_i, por un nuevo valor $T(x_i)$, de manera que los nuevos valores del lote son $T(x_1)$, $T(x_2)$, . . . , $T(x_n)$.

En el estudio del gráfico dispersión--nivel hemos presentado una familia importante de transformaciones:

las *transformaciones de potencias.*

Utilizamos una de ellas, la *transformación logaritmo*, para re-expresar las poblaciones de las 10 ciudades mayores de 16 paises.

También presentamos la llamada escalera de Tukey, en la tabla 16, con algunas potencias. enteras o semienteras.

Nuestra definición de transformación es bastante general:

- Incluye la posibilidad de llevar los datos brutos a una constante, a pesar que ese procedimiento no es útil en ningún sentido.

- Incluye la categorización de los datos. Por ejemplo las ciudades podrían ser categorizadas en pequeñas, medianas y grandes de acuerdo con su población.

La categorización es un tipo útil de transformaciones pero no las estudiaremos.

Nos interesan transformaciones con las siguientes propiedades:

1. **Preserven el orden** de los datos en el lote:
   - funciones estrictamente crecientes los datos mayores en la escala original sigan siendo mayores en la nueva escala puede cambiar su distancia
   - preservan los cuartos (y los percentiles en general) salvo por pequeñas diferencias debido a interpolaciones ya que estos dependen de los estadísticos de orden.

2. **Sean continuas**:
   - puntos que están muy cerca en el lote bruto también estén muy cerca en el lote transformado, en forma relativa a la escala utilizada.

3. **Sean suaves:**
   - no tengan ángulos agudos, derivables en todos los órdenes.
   - las re-expresiones se puedan obtener fácilmente.

TRANSFORMACIONES DE POTENCIA

Definición: Las transformaciones de potencia tienen la forma

$$T_p(x) = \begin{cases} ax^p + b & (p \neq 0) \\ c \log x + d & (p = 0) \end{cases} \qquad (1)$$

donde *a, b, c, d* y *p* son números reales. Para que las condiciones 1 a 5 se satisfagan es necesario que $a > 0$ para $p > 0$ y que $a < 0$ para $p < 0$.

Los valores *a, b, c* y *d* son bastante arbitrarios son y en general elegidos por conveniencia.
Mientras que *p* interesa y es elegido para facilitar el análisis.

Tres situaciones usuales para las elecciones de las constantes *a, b, c* y *d.*

1. Cuando queremos re-expresar los lotes en una forma sencilla, podemos especializar las transformaciones (1) en

$$T_p(x) = \begin{cases} x^p & (p > 0) \\ \log x & (p = 0) \\ -x^p & (p < 0) \end{cases} \qquad (2)$$

2. Cuando queremos comparar las transformaciones entre sí y examinar sus propiedades matemáticas y geométricas especializamos en

$$T_p(x) = \begin{cases} \dfrac{x^p - 1}{p} & (p \neq 0) \\ \ln x & (p = 0) \end{cases} \qquad (3)$$

3. Cuando queremos re-expresar el lote de manera que los nuevos datos se parezcan a los originales en posición y dispersión, elegimos las constantes en un proceso que llamaremos *apareareamiento* (*matching*).

Cuando *p* está fijo, cualquier eleción de las constantes *a* y *b* (ó *c* y *d*, si *p* = 0) representa una transformación lineal de cualquier otra:

$$\frac{A}{a}(ax^p + b) + (B - \frac{A}{a}b) = Ax^p + B.$$

Como una transformación lineal no constante simplemente cambia el origen y da un cambio uniforme de escala es que decimos que **la elección de las constantes no es esencial.**

Ventajas de la familia de transformaciones de potencias

- Cuando la familia está especificada de alguna de las tres formas anteriores se cumplen las cinco características anteriormente mencionadas.

- *Concavidad.* Cualquier transformación de potencias es cóncava hacia arriba ó cóncava hacia abajo dentro de su dominio de números positivos.

No hay puntos de inflección en los que la cavidad cambie.

- La transformación comprime la escala para los valores grandes más de lo que lo hace para valores chicos (ej. $T(x) = \log x$) ó a la inversa (ej. $T(x) = x^2$).
- No puede hacer cosas más complicadas como expandir la escala para valores grandes y chicos y comprimirla para los valores *intermedios*. Una transformación que se comporta de esa manera es

$$T(x) = (x-10)^3 ,$$

donde 10 es "*intermedio*" en el rango de los valores de x.

- *Flexibilidad.* En la práctica la flexibilidad que ofrece la familia es mayor de la necesaria. Muchas veces alcanza con restringirnos a los miembros de la escalera de potencias.

- *Unidad Geométrica*. Las transformaciones de potencias pueden ser consideradas como una familia de funciones, incluyendo al logaritmo.

Si consideramos de las curvas definidas en (1), simultáneamente para todo *p,* aquellas que

$$\begin{cases} T(1) = 0 & \text{(pasan por el punto (1,0))} \\ T'(1) = 1 & \text{(tienen pendiente 1)} \end{cases}$$

obtenemos la familia

$$T_p^*(x) = \begin{cases} \dfrac{x^p - 1}{p} & (p \neq 0) \\ \ln x & (p = 0) \end{cases} \qquad (3)$$

Como el análisis posterior no se altera mediante un cambio de posición y escala no perdemos generalidad al realizar una elección particular de las constantes (salvo que *a* = 0 ó *b* = 0).

El apareamiento realizado en el punto (1, 0) se evidencia en la figura 20 y la ausencia de éste se ve en la figura 21.

La figura 20 pone en evidencia las propiedades que tienen las transforma-ciones de potencia cuando están expresadas de la forma (3):

Propiedades de las curvas $T_p^*(x)$:

1. Todas las curvas son monótonas crecientes, de manera que para cada *p*, $T_p^*(x)$ preserva el orden de los datos transformados.

2. Las curvas comparten un punto (1, 0) para todo *p*.

3. Las curvas casi coinciden en los puntos cercanos al (1, 0); esto es que comparten una recta tangente común en ese punto.

4. Las curvas forman una familia paramétrica continua.

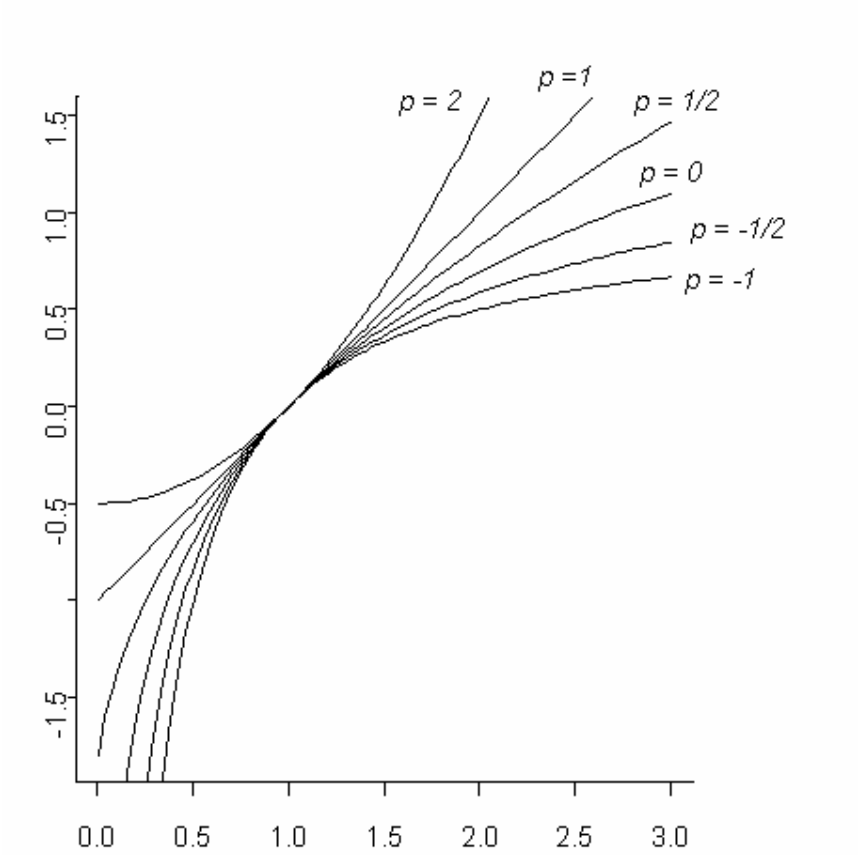


Figura 20. Gráficos de funciones de potencias apareadas, $T_p^*(x)$, **para valores de *p* elegidos.**

Observaciones:

- Las propiedades 2. y 3. dan el *apareamiento* de las transformaciones en el punto (1, 0).
- Las curvas están ordenadas por los valores de *p*, mayores valores de *p* dan curvas que están por encima de aquellas con menores valores de *p*.
- Si se agregara la curva con $p = ¾$ en la figura 20, ésta quedaría entre las curvas con $p = ½$ y $p = 1$.

- La transformación logaritmo calza entre las transformaciones de potencia $T^*_{1/2}(x) = 2\sqrt{x} - 2$ y $T^*_{-1/2}(x) = 2 - \frac{2}{\sqrt{x}}$.
- Para valores pequeños y positivos de λ , el gráfico de $T^*_0(x) = \ln x$ caería entre los gráficos de $T^*_\lambda(x)$ y de $T^*_{-\lambda}(x)$ para todo *x*.

Esta última observación da una justificación geométrica de nuestro uso de la transformación logaritmo como la transformación de potencias adecuada cuando *p* = 0.

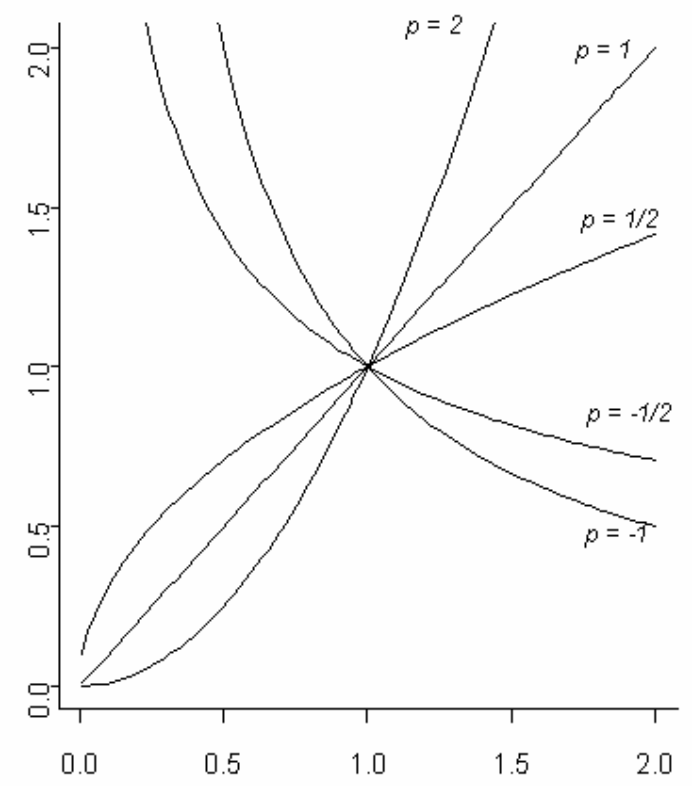


Figura 21. Gráficos de funciones de potencias, $T_p(x) = x^p$, para valores de *p* elegidos.

No hay diferencias prácticas en utilizar las transformaciones de las figuras 20 ó 21.

Pero si nos interesa comparar varias alternativas es ventajoso eliminar al máximo las fuentes de confusión.

Dificultades de las transformaciones $y = x^p$

Si $x > 0$ y $p < 0$, $y = x^p$ es decreciente y por lo tanto invierte el orden.

Esto puede solucionarse tomando las transformaciones presentadas en las ecuaciones (2).

Las diferencias en las pendientes en el punto (1, 1) hacen que las transformaciones tengan, incluso para valores arbitrariamente cercanos a 1, un comportamiento muy diferente.

Comparemos $y = x$ e $y = x^2$ en el punto (1, 1):
Tienen pendientes: 1 y 2 respectivamente
$x = 1.001$ es transformado a $y = x = 1.001$
$y = x^2 = 1.002001$

El punto 1.001 es transformado a una distancia poco más del doble de $y = 1$ por $y = x^2$ que por $y = x$. Este "estiramiento" por $y = x^2$ es aún más intenso para todos valores de x mayores que 1.

Los miembros de la familia $y = x^p$ no son comparables en escala.

Justificación analítica de las propiedades de las curvas $T_p^*(x)$:

Para cada *p,* incluído *p* =0, $T_p^*(1)=0$, de manera que se satisface 2.

La derivada de $T_p^*(x)$ es $T_p^{*\prime}(x)=x^{p-1}$ para todos los valores de *p*, incluyendo al cero. Como $T_p^{*\prime}$ existe y es positiva $\forall\, x>0$, $T_p^*(x)$ es una función continua y monótona creciente de x para todo *p*. Luego la propiedad 1. se satisface.

La pendiente de la recta tangente a cada curva en x =1 es $T_p^{*\prime}(1)=1$. Luego las curvas comparten una recta tangente común.

Las *derivadas* $T_p^{*\prime}(x)$también son *continuas* en todos los valores positivos de x y en particular para los valores de x cercanos a 1.

Como $T_p^{*\prime}(x)$ está cerca de 1 cuando x está cerca de 1, las curvas tienen casi la misma pendiente en valores de x cercanos a 1. Luego casi coinciden como la propiedad 3. lo requiere.

La cuarta propiedad, posiblemente la principal razón para la unidad de la familia de potencias es más sutil. A medida que *p* cambia su valor el cambio resultante de un miembro de la familia a otro ocurre en forma suave y continua: para cada x fijo $T_p^*(x)$ es una función continua de *p* y también los son sus derivadas primera, segunda, ... etc.

i) Cuando $p \neq 0$, $T_p^*(x)$ es el cociente de dos funciones continuas de *p*, y su cociente también es continuo para $p \neq 0$.

ii) Cuando *p* = 0 el cociente es una indeterminación del tipo 0/0 que podemos resolver utilizando la regla de L'Hospital. Recordemos que

x^p es igual a $e^{p \ln x}$ y su derivada respecto de *p* es $e^{p \ln x} . \ln x$, por lo tanto

$$\lim_{p \to 0} T_p^*(x) = \lim_{p \to 0} \frac{e^{p \ln x} - 1}{p}$$

$$= \lim_{p \to 0} \frac{e^{p \ln x} . \ln x}{1}$$

$$= \ln x$$

Como el resultado muestra que

$$\lim_{p \to 0} T_p^*(x) = T_0^*(x)$$

resulta que $T_p^*(x)$ es continua en *p* =0 para cada x positivo. Decimos que $T_p^*(x)$ es *una familia de funciones que está indexada en forma continua por el parámetro p.*

Un argumento similar muestra que la primera, segunda, ... derivadas $\{T_p^{*\prime}(x)\}$,$\{T_p^{*\prime\prime}(x)\}$,... también están indexadas continuamente por el parámetro *p*.