COMPARACIÓN DE LOTES MEDIANTE BOXPLOTS

EJEMPLO: LAS CIUDADES MÁS POBLADAS EN 16 PAÍSES

El *World Almanac* de 1967 lista 16 países que tienen 10 o más ciudades grandes; entre estas se han elegido las 10 más pobladas.

- El conjunto de datos da lugar a muchas preguntas respecto de las ciudades en estas 16 naciones.
- Cómo se comparan las medianas a través de las naciones?
- Son las menores ciudades más grandes de China mayores que las más grandes ciudades de algunos de los otros países?
- Tienen los países con ciudades más grandes una tendencia a tener mayor variabilidad entre las poblaciones?
- Cuánta asimetría tienen los diferentes conjuntos de datos?

(1) Sweden		(2) Netherlands		(3) Canada		(4) France	
Stockholm	7.87	Amsterdam	8.68	Montreal	11.91	Paris	28.11
Goteborg	4.22	Rotterdam	7.31	Toronto	6.72	Marseilles	7.83
Malmo	2.49	The Hague	6.02	Vancouver	3.84	Lyon	5.35
Norrkoping	0.94	Utrecht	2.64	Edmonton	2.81	Toulouse	3.30
Vasteras	0.89	Eindhoven	1.75	Hamilton	2.73	Nice	2.94
Uppsala	0.87	Haarlem	1.72	Ottawa	2.68	Bordeaux	2.54
Orebro	0.81	Groningen	1.51	Winnipeg	2.65	Nantes	2.46
Halsingborg	0.78	Tilburg	1.42	Calgary	2.49	Strasbourg	2.33
Linkoping	0.71	Enschede	1.31	Quebec	1.71	St. Etienne	2.03
Boras	0.69	Arnhem	1.29	London	1.69	Lille	1.99

(5) Mexico		(6) Argentina		(7) Spain		(8) England	
MexicoCity	31.18	BuenosAires	29.66	Madrid	25.99	London	79.86
Guadalajara	10.12	Rosario	7.61	Barcelona	16.96	Birmingham	11.02
Monterrey	8.06	Cordoba	6.35	Valencia	5.01		
Juarez	3.79	La Plata	4.10	Seville	4.74	Liverpool	7.22
Puebla	3.46	Avellaneda	3.80	Zaragoza	3.57	Manchester	6.38
Mexicali	2.91	Santa Fe	2.75	Bilboa	3.34	Leeds	5.09
Leon	2.71	Mar del Plata	2.70	Malaga	3.12	Sheffield	4.88
Torreon	2.17	Gral.SanMartin	2.69	Murcia	2.64	Bristol	4.30
Chihuahua	2.06	Tucuman	2.51	Cordoba	2.14	Coventry	3.30
SanLuisPotosi	1.86	Lanus	2.44	Palma	1.69	Nottingham	3.10
						Kingston	2.99

(9) Italy		(10) West Germany		(11) Brazil		(12) Soviet Union	
Rome	23.59	WestBerlin		Sao Paulo	49.81	Moscow	63.34
Milan	15.80	21.92		RiodeJaneiro	38.57	Leningrad	36.36
Naples	11.82	Hamburg		Recife	9.68	Kiev	13.32
Turin	11.14	18.56		BeloHorizonte	9.52	Baku	11.37
Genoa	7.84	Munich	11.42	Salvador	8.08	Tashkent	10.90
Palermo	5.90	Cologne	8.27	Porto Alegre	8.03	Gorky	10.84
Florence	4.54	Essen	7.28	Fortaleza	6.99	Kharkov	10.70
Bologna	4.44	Dusseldorf	7.02	Curitiba	5.02	Novosibirsk	10.27
Catania	3.61	Frankfurt	6.94	Belem	4.95	Kuibyshev	9.50
Venice	3.36	Dortmund	6.53	Niterol	2.78	Sverdlovsk	9.17
		Bremen	5.84				
		Hannover	5.66				

(13) Japan		(14) United States		(15) India		(16) China	
Tokyo	110.21	NewYork	77.81	Bombay	45.37	Shangai	69.00
Osaka	32.14	Chicago	35.50	Calcutta	30.03	Beijing	40.10
Nagoya	18.88	LosAngeles	24.79	Delhi	22.98	Hong Kong	36.92
Yokohama	16.39	Philadelphia	20.02	Hyderabad	20.62	Tianjin	32.20
Kyoto	13.37	Detroit	16.70	Madras	17.25	Shenyang	24.11
Kobe	11.95	Baltimore	9.39	Howrah	16.11	Wuhan	21.46
KitaKyush	10.70	Houston	9.38	Ahmedabad	11.49	Chongqing	21.21
		Cleveland	8.76	Kanpur	9.47	Canton	16.50
Kawasaki	7.89	Washington,DC	7.63	Bangalore	9.07	Xian	15.00
Fukuoka	7.71	St. Louis	7.50	Poona	7.21	Nanjing	11.13
Sapporo	7.04						

Tabla 12: medidas resumen de las poblaciones de 10 ciudades mayores de 16 países

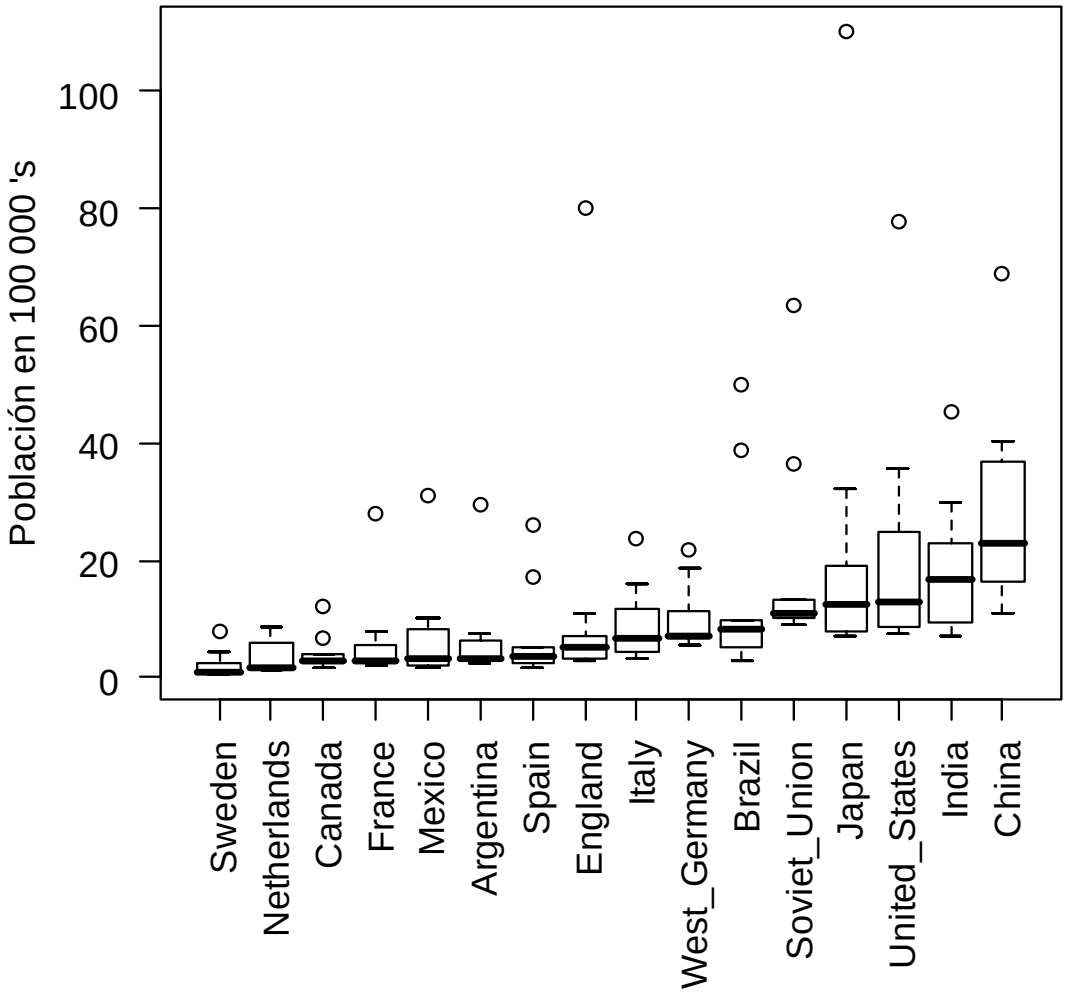
SWEDEN	NETHERLANDS	CANADA	FRANCE
Min. :0.6900 1st Qu.:0.7875 Median :0.8800 Mean :2.0270 3rd Qu.:2.1020 Max. :7.8700	Min. :1.290 1st Qu.:1.442 Median :1.735 Mean :3.365 3rd Qu.:5.175 Max. :8.680	Min. : 1.690 1st Qu.: 2.530 Median : 2.705 Mean : 3.923 3rd Qu.: 3.582 Max. :11.910	Min. : 1.990 1st Qu.: 2.362 Median : 2.740 Mean : 5.888 3rd Qu.: 4.837 Max. :28.110
MEXICO	ARGENTINA	SPAIN	ENGLAND
Min. : 1.860 1st Qu.: 2.305 Median : 3.185 Mean : 6.832 3rd Qu.: 6.992 Max. :31.180	Min. : 2.440 1st Qu.: 2.692 Median : 3.275 Mean : 6.461 3rd Qu.: 5.788 Max. :29.660	Min. : 1.690 1st Qu.: 2.760 Median : 3.455 Mean : 6.920 3rd Qu.: 4.942 Max. :25.990	Min. : 2.990 1st Qu.: 3.550 Median : 4.985 Mean :12.810 3rd Qu.: 7.010 Max. :79.860
ITALY	WGERMANY	BRAZIL	URSS
Min. : 3.360 1st Qu.: 4.465 Median : 6.870 Mean : 9.204 3rd Qu.:11.650 Max. :23.590	Min. :5.660 1st Qu.:6.632 Median :7.150 Mean :9.944 3rd Qu.:10.630 Max. :21.920	Min. : 2.780 1st Qu.: 5.512 Median : 8.055 Mean :14.340 3rd Qu.: 9.640 Max. :49.810	Min. : 9.17 1st Qu.:10.38 Median :10.87 Mean :18.58 3rd Qu.:12.83 Max. :63.34
JAPAN	USA	INDIA	CHINA
Min. : 7.040 1st Qu.: 8.592 Median : 12.660 Mean : 23.630 3rd Qu.: 18.260 Max. :110.200	Min. : 7.500 1st Qu.: 8.915 Median :13.040 Mean :21.750 3rd Qu.:23.600 Max. :77.810	Min. : 7.210 1st Qu.: 9.975 Median :16.680 Mean :18.960 3rd Qu.:22.390 Max. :45.370	Min. :11.13 1st Qu.:17.68 Median :22.78 Mean :28.76 3rd Qu.:35.74 Max. :69.00

Las respuestas a las preguntas planteadas pueden obtenerse a partir las medidas resumen dadas en la tabla 12

En la práctica, un gráfico con los boxplots en paralelo para los 16 grupos de datos hace que las respuestas a estas y otras preguntas similares aparezcan rápidamente .

La figura 12 muestra este gráfico con los boxplots paralelos a si mismos y al eje de la población, ordenados de acuerdo a la mediana de las poblaciones de las 10 ciudades más grandes.

FIGURA 12



- Las ciudades más grandes de China tienden a ser más grandes que las de cualquier otra nación. todas las ciudades grandes de China son más grandes que todas las ciudades grandes de Suecia (Sweden) y Holanda (Netherlands).

- Comparamos la dispersión de estos 16 lotes mediante las longitudes de las cajas. Los datos de Canadá son los menos dispersos y los de China los más.

- La mayoría de los países presentan alguna asimetría en dirección a las grandes ciudades;

- Solamente India y Brasil tienen cajas que están sesgadas hacia la izquierda, pero ambos países tienen ciudades sustancialmente mayores que las representadas por las cajas.

- La mayor ciudad de todos los países, con la excepción de Holanda, es designada como un outlier; algunos tienen más de un outlier entre las 10 ciudades más pobladas.

Hemos detectado dos anormalidades en estos datos: **asimetría y outliers**.

Al haber ordenado los países en base a la mediana de los lotes podemos detectar otra característica:

tendencia de aumento en la dispersión a medida que aumenta el nivel.

Esta tendencia no es compatible con el supuesto de similar variabilidad entre lotes; cuando esto ocurre el análisis se simplifica.

Veremos transformaciones de los datos que permitan lograr homogeneidad de dispersiones y reducir la dependencia de éstas con el nivel.

En R

La función **`mar`** da los *márgenes* de los gráficos en *pulgadas* en el siguiente orden: margen inferior, margen izquierdo, margen superior y margen derecho respectivamente.

El valor por defecto es mar=c(5,4,4,2)+0.1.

Hemos obtenido el boxplot de las poblaciones de las 10 ciudades más pobladas mediante las siguientes instrucciones.

```
> par(las=2,cex=1,mar=c(7.2,4,2,2))
> boxplot(pobl16,
   ylab= "Población en 100 000 's" )
```

Para graficar los boxplots en orden creciente de las medianas

```
> orden.med <-
sort.list(sapply(pobl16,median))
> boxplot(pobl16[orden.med])
```

En este caso

```
sapply(pobl16,median)
```

es lo mismo que

```
apply(pobl16,2, median)
```

Para graficar en orden alfabético

Vemos que el data frame no tiene a los países con sus nombres en orden alfabético

```
   > names(pobl16)
 [1] "Sweden"        "Netherlands"
 [3] "Canada"        "France"
 [5] "Mexico"        "Argentina"
 [7] "Spain"         "England"
 [9] "Italy"         "West_Germany"
[11] "Brazil"        "Soviet_Union"
[13] "Japan"         "United_States"
[15] "India"         "China"
```

Primero hallamos los índices que nos dan el orden alfabético

```
> sort.list(names(pobl16))
 [1]  6 11  3 16  8  4 15  9 13  5  2 12  7  1
[15] 14 10
```

```
   > orden.alfab <- sort.list(names(pobl16))

   > names(pobl16[orden.alfab])
    [1] "Argentina"     "Brazil"
    [3] "Canada"        "China"
    [5] "England"       "France"
    [7] "India"         "Italy"
    [9] "Japan"         "Mexico"
   [11] "Netherlands"   "Soviet_Union"
   [13] "Spain"         "Sweden"
   [15] "United_States" "West_Germany"
```

NIVEL VERSUS DISPERSIÓN

Nos interesa hallar una ***transformación*** de los datos que reduzca o elimine el crecimiento, o decrecimiento, de la dispersión con el nivel.

Los datos reexpresados serán más adecuados tanto para exploración visual coma para aplicar técnicas usuales de comparación de grupos.

Por ejemplo el análisis de varianza de un factor es más simple y más efectivo cuando hay, exacta o aproximadamente, igualdad de varianzas entre grupos.

Transformaciones de potencia

Definimos la transformación de potencia con potencia (ó exponente) p como la transformación que reemplaza x por x^p.

Para p=0 utilizamos log x en vez de x^0.

Veremos que log x es el límite cuando p tiende a cero de $(x^p-1)/p$.

Definiremos un gráfico que nos permitirá encontrar la transformación adecuada.

Construcción de gráficos de dispersión versus nivel

Nos interesa eliminar la relación entre el nivel y la dispersión. Podemos suponer que la distancia intercuartos (cuartos! no cuartiles) es proporcional a una potencia de la mediana:

$$d_Q = cM^b, \qquad (1)$$

ó

$$\log d_Q = k + b \log M. \qquad (2)$$

Los logaritmos de las distancias intercuartos y el logaritmo de las medianas están relacionados linealmente.

El gráfico de dispersión versus nivel consiste en graficar los valores de log d_Q contra los valores de log M para todos los lotes y luego ajustar una recta al diagrama de dispersión obtenido.

Si b es la pendiente estimada entonces

$$\boldsymbol{p = 1 - b}$$

es un valor aproximado del exponente de una transformación de potencias de x para estabilizar la dispersión. Cuando $p = 0$ se utiliza el logaritmo.

Tabla 15: Logs de medianas - 5 (base 10) y distancias intercuartos para las mayores ciudades de 16 países.

País	**log *M***	**log d_Q**
Sweden	-.06	.23
Netherlands	.24	.66
Canada	.43	.13
France	.44	.48
Mexico	.50	.77
Argentina	.51	.56
Spain	.54	.38
England	.70	.59
Italy	.84	.87
West Germany	.85	.69
Brazil	.91	.67
Soviet Union	1.04	.48
Japan	1.10	1.04
United States	1.12	1.20
India	1.22	1.13
China	1.36	1.31

EJEMPLO: GRÁFICO DE DISPERSIÓN VERSUS NIVEL PARA LAS MAYORES CIUDADES

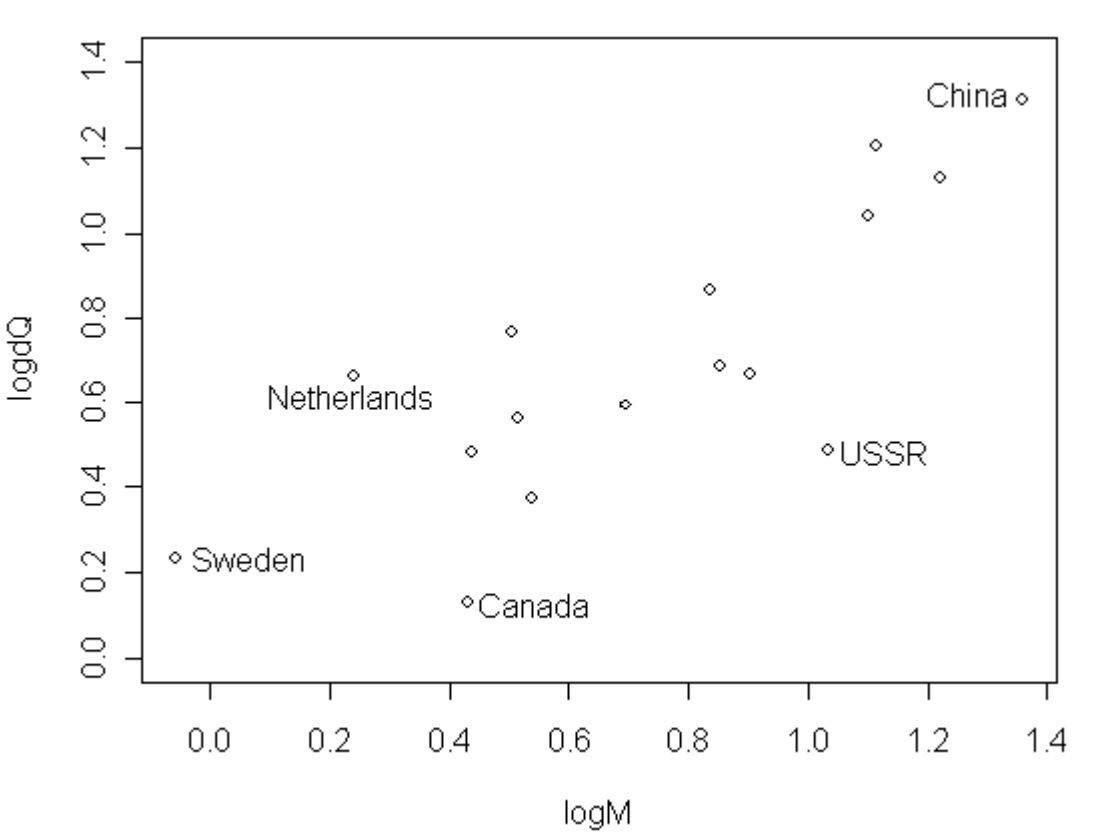
La tabla 15 muestra los logaritmos (en base 10) de las medianas y las distancias intercuartos para cada una de los 16 países.

La figura 15 muestra un gráfico de log d_Q contra log *M*, es decir un gráfico de dispersión versus nivel, para los 16 países. El log d_Q crece a medida que crece log *M*: en primera aproximación la relación parece casi lineal.

Ajustamos una recta a ojo a los puntos de la figura 15. Aunque dos personas no llegarán a la misma pendiente por este método, casi seguro trazarán una recta con pendiente entre ½ y 1, y probablemente más cerca de 1 (la recta de

regresión ajustada por cuadrados mínimos tiene pendiente 0.69).

Figura 15: Gráfico de dispersión versus nivel, log d_Q contra log *M*



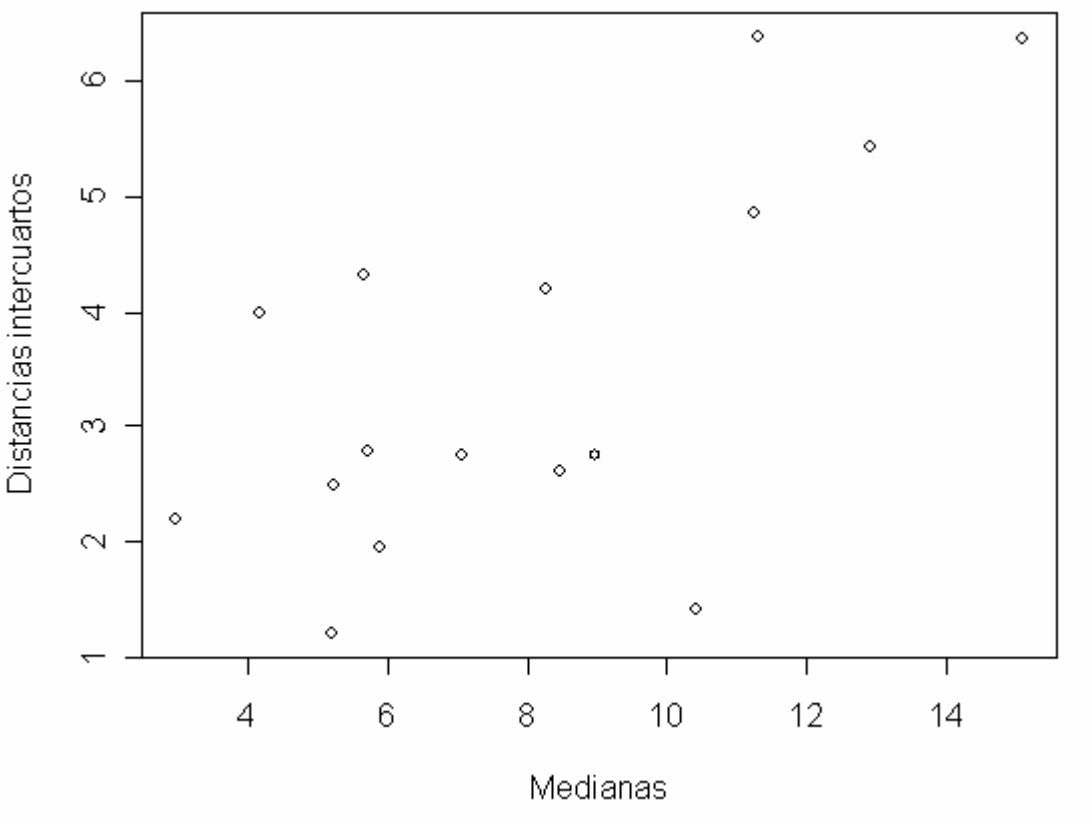
Para b = 1, p es cero y resulta la transformación logaritmo. Análogamente, si b = ½ lleva a p = ½ , la transformación raíz cuadrada.
A pesar que una potencia entre 0 y ½ puede ser mejor que alguno de estos dos para estabilizar la dispersión, por razones de simplicidad y de interpretabilidad, consideraremos las transformaciones logaritmo y raíz cuadrada.

Aplicamos cada una de las transformaciones.

Tabla 16. Medianas y distancias intercuartos para los datos de las 10 ciudades mayores de 16 países transformados por logaritmo base 10 y raíz cuadrada.

País	Logaritmo de la población		Raíz cuadrada de la población	
	Mediana (más 5)	d_Q	Mediana (por 10^2)	d_Q
Sweden	-.06	.51	2.97	2.19
Netherlands	.22	.62	4.17	3.98
Canada	.43	.18	5.21	1.20
France	.44	.36	5.24	2.49
Mexico	.50	.57	5.64	4.32
Argentina	.52	.37	5.72	2.78
Spain	.54	.28	5.88	2.12
England	.70	.34	7.07	2.75
Italy	.84	.42	8.29	4.21
West Germany	.85	.24	8.46	2.61
Brazil	.91	.29	8.98	2.75
Soviet Union	1.04	.11	10.43	1.41
Japan	1.10	.39	11.25	4.86
United States	1.12	.45	11.42	6.38
India	1.22	.38	12.92	5.42
China	1.36	.35	15.09	6.37

Figura 16. Distancia intercuartos versus mediana: datos transformados por la raíz cuadrada.

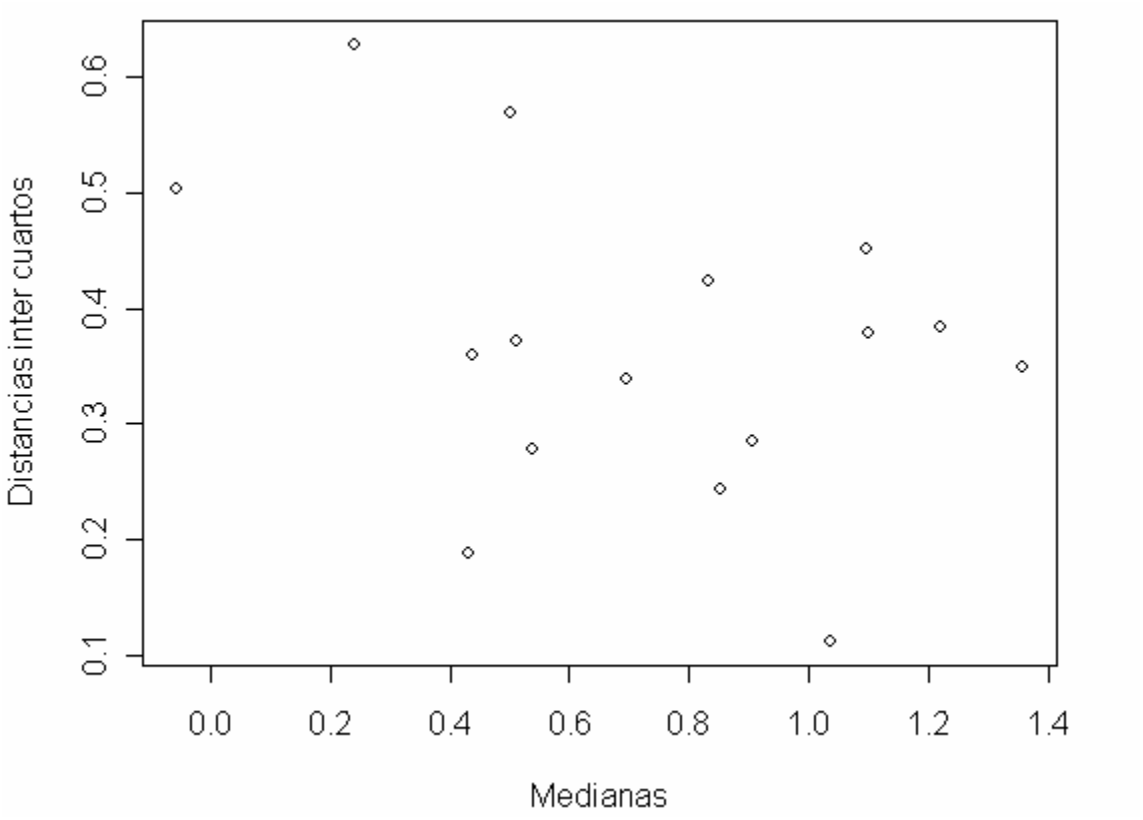


Las magnitudes relativas de las **pendientes** de los gráficos de las figuras 16 y 17 nos permiten elegir entre las dos transformaciones.

Las distancias intercuartos de los logaritmos decrecen levemente con el nivel mientras las de la raíz cuadrada aumentan con el nivel.

Deberíamos elegir el logaritmo.

Figura 17. Distancia intercuartos versus nivel: datos transformados por logaritmo.



Tomando la decisión

¿Cómo podemos tomar la decisión entre la transformación logaritmo, raíz cuadrada o alguna otra transformación de potencia con *p* entre 0 y 1?

Idealmente, una transformación no sólo iguala dispersiones sino también tiene una explicación temática.

Por ejemplo, en demografía, un modelo muy utilizado supone que las poblaciones tienden a crecer exponencialmente. Si esto es así, el logaritmo de la población crecerá aproximadamente de manera lineal.

Las ventajas del crecimiento lineal, tales como la facilidad de detectar apartamientos del ajuste y la

conveniencia en la interpolación sugieren al logaritmo como una transformación adecuada para poblaciones humanas.

También la raíz cúbica, $p = 1/3$ es una transformación que algunas veces tiene significado físico.

Tabla 17. Transformaciones de potencia más usadas[a]

Transformación	Potencia	Pendiente del gráfico dispersión - nivel
Cúbica	3	-2
Cuadrada	2	-1
No cambio	1	0
Raíz cuadrada	1/2	1/2
Logaritmo	0	1
Inversa de la raíz cuadrada	-1/2	3/2
Inversa	-1	2

[a]Corresponden a los miembros principales de la "escalera de potencias" de Tukey.

Para las ciudades más grandes elegimos la transformación logaritmo por que se trata de datos poblacionales. Este modelo teórico simple que favorece al logaritmo, más que una fuerte evidencia en los datos, es la base para tomar esa decisión.

Reanálisis en la escala logarítmica.

Las transformaciones de potencia son monótonas para valores positivos, luego:

los estadísticos de orden de los datos transformados serán los estadísticos de orden originales transformados

(salvo por los efectos del redondeo e interpolación).

Para obtener los boxplots es necesario recalcular la distancia intercuartos y los puntos de corte para los outliers.

Puede ocurrir que algunos datos que originalmente eran outliers del lado alto dejen de serlo y aparezcan algunos del lado bajo. Esto último es poco probable cuando tomamos los 10 valores mayores.

En la figura 18 observamos que las cajas son más similares en longitud y que la desigualdad remanente no parece estar muy relacionada con el nivel (aunque quizás se hayan reducido un poco de más las dispersiones de los niveles más altos, recordemos que la pendiente de la recta ajustada era 0.7).

En la nueva escala muchos outliers han sido llevados hacia adentro. De los 19 originales, 8 ya no son outliers y los demás se han movido hacia los puntos de corte superior.

Los nuevos boxplots son más fáciles de mirar y los países están desplegados a un nivel de detalle similar.

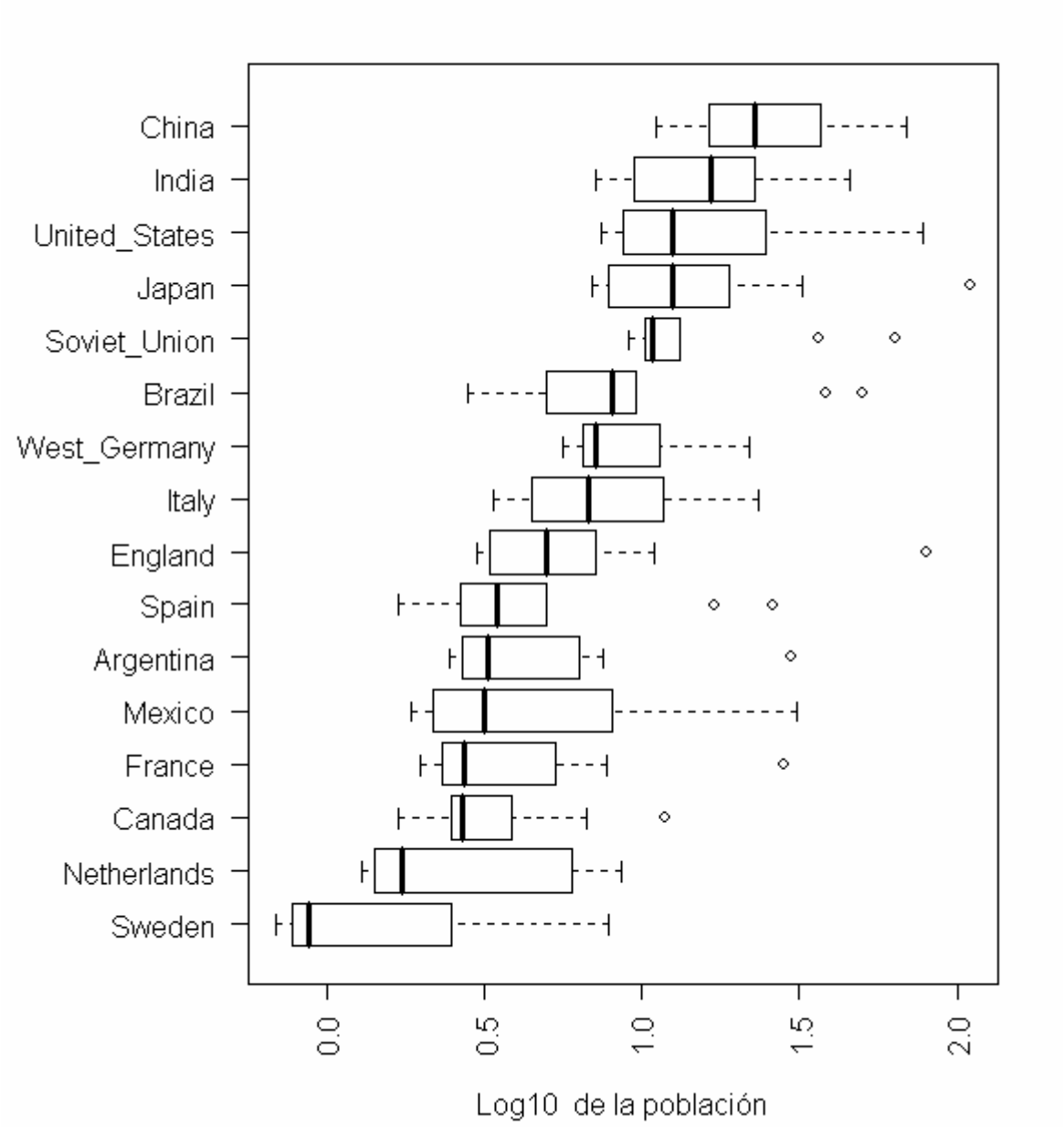
En la escala original los valores de Suecia (Sweden), Holanda (Netherlands) y Canadá son más difíciles de leer del gráfico que los de India y China.

En los gráficos con escala logarítmica los detalles aparecen similarmente bien para todos los países.

Tabla 18: medidas resumen del logaritmo de las poblaciones de las 10 ciudades mayores de 16 países

SWEDEN	NETHERLANDS	CANADA	FRANCE
Min. :-0.1612	Min. :0.1106	Min. :0.2279	Min. :0.2989
1st Qu.:-0.1038	1st Qu.:0.1590	1st Qu.:0.4030	1st Qu.:0.3733
Median :-0.0555	Median :0.2393	Median :0.4321	Median :0.4366
Mean : 0.1270	Mean :0.4041	Mean :0.5081	Mean :0.5827
3rd Qu.: 0.2904	3rd Qu.:0.6901	3rd Qu.:0.5504	3rd Qu.:0.6759
Max. : 0.8960	Max. :0.9385	Max. :1.0760	Max. :1.4490
MEXICO	**ARGENTINA**	**SPAIN**	**ENGLAND**
Min. :0.2695	Min. :0.3874	Min. :0.2279	Min. :0.4757
1st Qu.:0.3606	1st Qu.:0.4302	1st Qu.:0.4397	1st Qu.:0.5473
Median :0.5015	Median :0.5096	Median :0.5382	Median :0.6976
Mean :0.6340	Mean :0.6436	Mean :0.6570	Mean :0.8122
3rd Qu.:0.8244	3rd Qu.:0.7553	3rd Qu.:0.6938	3rd Qu.:0.8451
Max. :1.4940	Max. :1.4720	Max. :1.4150	Max. :1.9020
ITALY	**WGERMANY**	**BRAZIL**	**URSS**
Min. :0.5263	Min. :0.7528	Min. :0.4440	Min. :0.9624
1st Qu.:0.6498	1st Qu.:0.8215	1st Qu.:0.7366	1st Qu.:1.0160
Median :0.8326	Median :0.8542	Median :0.9061	Median :1.0360
Mean :0.8744	Mean :0.9469	Mean :0.9744	Mean :1.1600
3rd Qu.:1.0660	3rd Qu.:1.0230	3rd Qu.:0.9841	3rd Qu.:1.1070
Max. :1.3730	Max. :1.3410	Max. :1.6970	Max. :1.8020
JAPAN	**USA**	**INDIA**	**CHINA**
Min. :0.8476	Min. :0.8751	Min. :0.8579	Min. :1.046
1st Qu.:0.9302	1st Qu.:0.9499	1st Qu.:0.9973	1st Qu.:1.245
Median :1.1020	Median :1.0980	Median :1.2220	Median :1.357
Mean :1.1900	Mean :1.2000	Mean :1.2110	Mean :1.400
3rd Qu.:1.2610	3rd Qu.:1.3710	3rd Qu.:1.3500	3rd Qu.:1.552
Max. :2.0420	Max. :1.8910	Max. :1.6570	Max. :1.839

Figura 18. Boxplots de los logaritmos de las poblaciones de las 10 ciudades mayores en 16 países.



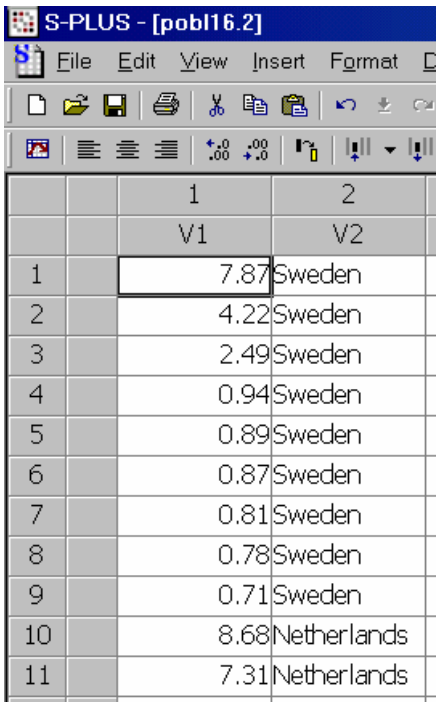
Observación: en R, cuando el argumento **`horizontal`** de la función **`boxplot`** toma el valor **`TRUE`** obtenemos boxplots horizontales al eje x.

En S-plus

Cambiamos la presentación de los datos de pobl16 y los guardamos en pobl16.2

```
menuStackColumns(pobl16.2, target.col.spec="<END>", pobl16,
    source.col.spec="<ALL>", rep.source.col.spec=NULL,
    rep.target.col.spec="<END>", group.col.p=T,
    group.col.name="<END>",
    show.p=.Options$show.data.in.view)
```

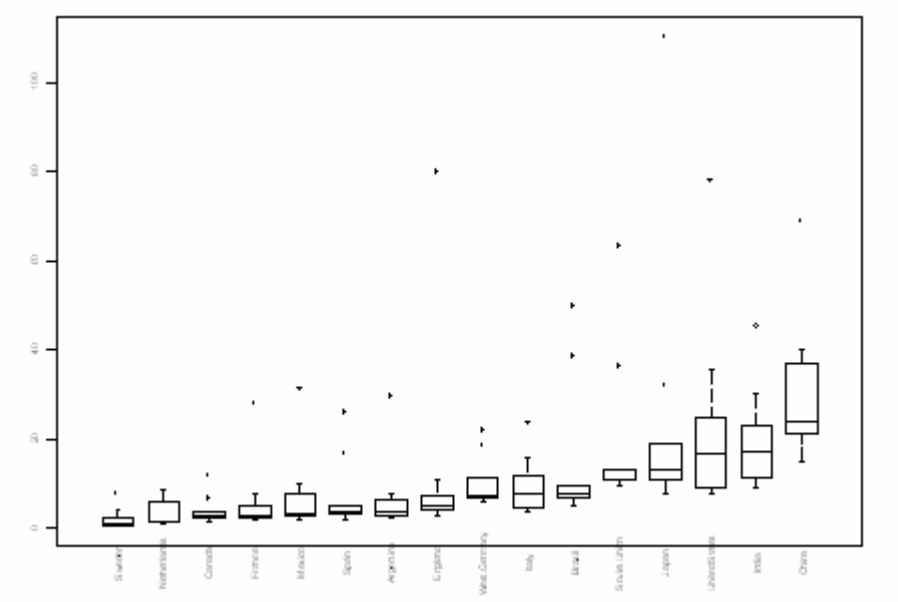
Los datos aparecen en dos columnas,



		1	2
		V1	V2
1		7.87	Sweden
2		4.22	Sweden
3		2.49	Sweden
4		0.94	Sweden
5		0.89	Sweden
6		0.87	Sweden
7		0.81	Sweden
8		0.78	Sweden
9		0.71	Sweden
10		8.68	Netherlands
11		7.31	Netherlands

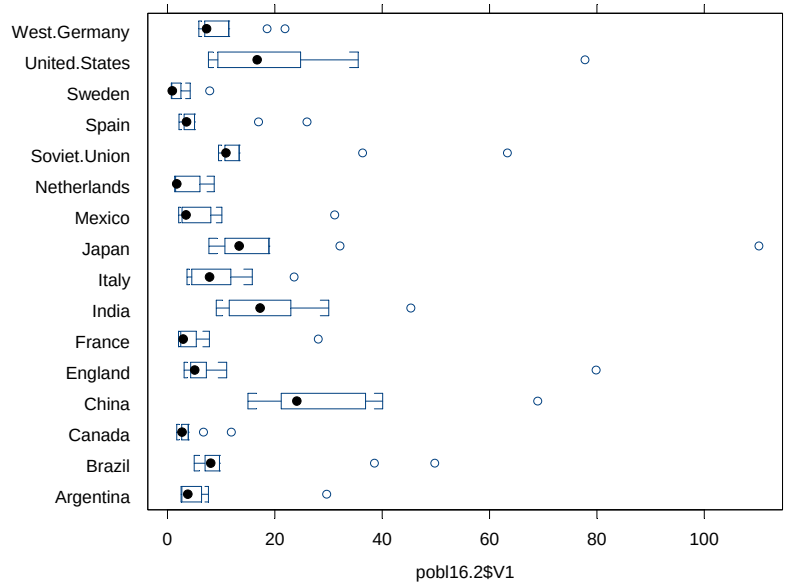
V1: valores de la variable población,
V2: nombre del país

```
boxplot(split(pobl16.2$V1,
 pobl16.2$V2)[sort.list(sapply((split(
 pobl16.2$V1, pobl16.2$V2)), median))], srt =
 90, cex = 0.5, style.bxp = "old")
```



Usando la función **`bwplot`** obtenemos boxplots paralelos al eje horizontal

```
bwplot(pobl16.2$V2 ~ pobl16.2$V1)
```



Mire el help. Además pruebe con

```
plot.factor(pobl16.2$V2,pobl16.2$V1)
```