Documentación de R

El sitio de CRAN http://cran.r-project.org/ tiene contribuciones de los usuarios.

http://cran.r-project.org/doc/contrib/

W. J. Owen 2007. **The R Guide**
http://cran.r-project.org/doc/contrib/Owen-TheRGuide.pdf

R reference card by Tom Short es un resumen de 4 páginas muy útil.
http://cran.r-project.org/doc/contrib/Short-refcard.pdf

Algunas son en **castellano**.

http://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf

Emmanuel Paradis 2003. **R para Principiantes.**
http://cran.r-project.org/doc/contrib/rdebuts_es.pdf

MEDIDAS RESUMEN: Numéricas y Gráficas

Nos interesa resumir las características más importantes del conjunto de datos en una pequeña cantidad de números que sean fácilmente interpretables.

Para hacer esto bien, debemos limitar nuestros objetivos. Es equivocado esperar que un resumen estándar revele lo inusual. Si la distribución de los datos es distintiva es más fácil expresarlo con palabras.

En el diagrama tallo-hoja de los datos de la dureza de 30 incrustaciones de aluminio,
53.0 82.5 74.4 55.7 70.2 67.3 54.1 70.5 84.3 69.5 77.8 87.5 55.3 73.0 52.4 50.7 78.5 55.7 69.1 72.3 63.5 85.8 53.5 59.5 71.4 95.4 64.3 53.4 51.1 82.7,

se distinguen dos grupos y un punto aislado (esto no ocurre con frecuencia):

5 | 1123344
5 | 566
6 | 044
6 | 79
7 | 0011234
7 | 89
8 | 334
8 | 68
9 |
9 | 5

Intentar mostrar ocurrencias ocasionales de este tipo complicarían los procedimientos y confundirían las medidas resumen. Para la mayoría de los lotes nada de eso pasa.

Vemos que los datos se cortan más abruptamente hacia valores pequeños que hacia valores grandes (esto ocurre muy frecuentemente).

Los resúmenes pueden ser muy útiles pero no son los detalles. Generalmente los detalles agregan poco pero es importante estar preparados para las ocasiones en que sí agregan mucho.

Medidas Resumen

Los estadísticos clásicos para resumir un conjunto de datos de n observaciones, $x_1, x_2, \ldots, x_n$, utilizan solamente operaciones aritméticas simples, como suma, multiplicación, división y tal vez raíz cuadrada, en todos sus datos. Los más familiares son la *media muestral* dada por

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} \ ,$$

y la *varianza muestral*, frecuentemente dada por

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \ .$$

Es importante que las medidas resumen sean resistentes, es decir que varíen poco en presencia de un cambio arbitrario de una pequeña parte del lote.

La media muestral y la varianza muestral no tienen ese comportamiento. Un único dato aberrante puede producir un importante efecto adverso en ambas medidas.

En lo que sigue veremos una medida de posición, la mediana, y una medida de dispersión, distancia entre cuartos, basadas en los datos ordenados.

Primero debemos ordenar los datos y asignarles un rango.

Supongamos que ordenamos los datos, $x_1, x_2, \ldots, x_n$, en orden ascendente y obtenemos la muestra ordenada que indicamos por:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)};$$

esto es, $x_{(1)}$ es la observación más pequeña, $x_{(2)}$ la 2da. observación más pequeña , $x_{(i)}$ es la i-ésima observación más pequeña. Más formalmente $x_{(1)}$, $x_{(2)}$, $\ldots, x_{(n)}$ son llamados *estadísticos de orden*, y $x_{(i)}$ es el estadístico de orden i

Sobre la base del ordenamiento podemos definir el rango de una observación de dos maneras distintas.

Podemos contar desde el más pequeño hacia el más grande obteniendo el *rango ascendente*, ó desde el más grande hacia el más pequeño obteniendo el *rango descendente*.

$x_{(2)}$ tiene rango ascendente 2 y en general $x_{(i)}$ tiene rango ascendente *i*. Contando desde la observación mayor; $x_{(n)}$ tiene rango 1 (= $n + 1 - n$), $x_{(n-1)}$ tiene rango 2 (= $n + 1 - (n - 1)$) y $x_{(i)}$ tiene rango $n + 1 - i$.

Considerando ambos de estos rangos observamos que para cualquier dato

rango ascendente + rango descendente = $n+1$.

Definición: La *profundidad* de un dato en la muestra es el menor de los rangos ascendente y descendente.

Utilizaremos la noción de profundidad para obtener medidas resumen de los datos.

Medida de posición del centro de los datos

Definición: La *mediana*, M es el valor que deja la misma cantidad de los datos ordenados de cada lado, más formalmente, la proporción de datos menores que la mediana es menor o igual a ½ y la proporción de datos mayores que la mediana también es menor o igual a ½:

#(Datos $< M$) / $n \leq$ ½ y #(Datos > M) / n $\leq$ ½

Esta definición da un intervalo de valores posibles para la mediana. Con la siguiente forma de cálculo se obtiene el valor central de dicho intervalo.

La profundidad de la mediana es $p_M = \frac{n+1}{2}$.

Si p_M es entero, $n = 2k + 1$ (es impar), la mediana es el dato con profundidad

$\frac{n+1}{2} = k + 1$.

Si p_M es no es entero, $n = 2k$ (*es* par), la mediana tiene profundidad

$\frac{n+1}{2} = k + ½$ es decir que cae a mitad de camino entre $x_{(k)}$ y $x_{(k+1)}$.

O sea

$$M = \text{mediana} = \begin{cases} x_{(k+1)} & \text{si } n = 2k+1 \\ \dfrac{x_{(k)} + x_{(k+1)}}{2} & \text{si } n = 2k \end{cases} .$$

En el ejemplo de los puntos de fusión de las ceras de abeja $n = 59$. La mediana es el dato con profundidad $\frac{59+1}{2} = 30$ que podemos obtener inmediatamente a partir del diagrama Tallo-Hoja. La mediana se encuentra en la línea 8 del diagrama (corresponde al tallo que no tiene valor en la columna de profundidad).

La última hoja del tallo inmediato menor, 634, tiene profundidad 23, entonces la posición 30 se obtiene contado 7 hojas en el tallo 635; $M = 63.53$.

Los valores más sencillos de extraer del lote son los ***extremos***, los dos datos con **profundidad 1**, o sea el valor mayor y el valor menor del lote

Decimal point is 1 place to the left of the colon

PROF.	# hojas	TALLO	HOJAS
1	1	628	: 5
1	0	629	:
4	3	630	: 358
7	3	631	: 033
9	2	632	: 77
18	9	633	: 001446669
23	5	634	: 01335
	10	635	: 0000113668
26	7	636	: 0013689
19	2	637	: 88
17	6	638	: 334668
11	5	639	: 22223
6	0	640	:
6	1	641	: 2
5	3	642	: 147
2	0	643	:
2	2	644	: 02

Aquí, como para todos los resúmenes aparte de la mediana, **una profundidad** identifica ***dos* valores** de los datos, uno por debajo y otro por encima de la mediana.

A la mediana y los extremos agregamos otro **par de valores resumen**, los *cuartos*, definiendo

$$p_Q = \text{profundidad del cuarto} = \frac{[\text{profundidad de la mediana}] + 1}{2};$$

los corchetes en [x] indican, igual que antes, parte entera de x.
Esta definición simplifica la interpolación ya que la profundidad del cuarto sólo puede ser un entero ó un entero más ½

Hemos visto que si

n= 2*k* + 1 (es impar), la mediana es el dato con profundidad $\frac{n+1}{2} = k + 1$. Por lo tanto el cuarto tiene profundidad $\frac{k+1+1}{2} = \frac{\mathbf{k}}{\mathbf{2}} + \mathbf{1}$.

n = *2k* (*es* par), la mediana tiene profundidad $\frac{n+1}{2} = k + ½$. Por lo tanto, su parte entera es *k* y la profundidad del cuarto es $\frac{\boldsymbol{k+1}}{\boldsymbol{2}}$.

Cada cuarto está a mitad de camino entre la mediana y su correspondiente extremo, de manera que los cuartos *inferior* y *superior* encierran la mitad central de los datos.

Más formalmente, un ¼ - ésimo cuantil es el número $\boldsymbol{x_{1/4}}$, tal que

#(Datos $< \boldsymbol{x_{1/4}}$) / $n \leq ¼$ y #(Datos > $\boldsymbol{x_{1/4}}$) / $n \leq 1 - ¼$

También en este caso, como en la mediana, se obtiene un intervalo de posibles valores para el ¼ - ésimo cuantil. Un valor posible es el cuarto inferior definido anteriormente.

En el ejemplo anterior teníamos p_M = 30, por lo tanto p_Q = $\frac{30+1}{2} = 15+1/2$. Los cuartos se obtienen promediando los valores en las profundidades 15 y 16. Para el cuarto inferior estos valores se encuentran en la fila 6 del diagrama Tallo-Hoja, que corresponde al tallo 633 y para el cuarto superior en la fila 11.

$$\text{CUARTO INFERIOR} = \frac{63.36 + 63.36}{2} = 63.36$$

$$\text{CUARTO SUPERIOR} = \frac{63.84 + 63.83}{2} = 63.835$$

Observación: en el diagrama Tallo-Hoja

para tallos menores que la mediana:

la profundidad se alcanza en la hoja con mayor valor;
la profundidad de cada hoja del tallo se obtiene contando de izquierda a derecha a partir de la profundidad del tallo inmediato inferior;

para tallos mayores que la mediana:

la profundidad se alcanza en la hoja con menor valor;
la profundidad de cada hoja del tallo se obtiene contando de derecha a izquierda partir de la profundidad del tallo inmediato superior.

Medidas de dispersión de los datos

Nos interesa describir cuán concentrados están los datos. Una medida simple y resistente es la *distancia intercuartos* (d_Q)

$d_Q =$ CUARTO SUPERIOR - CUARTO INFERIOR

que da el ancho de la mitad central del lote.

En el ejemplo d_Q = 63.835 - 63.36 = 0.475.

Obviamente el rango, es decir la diferencia entre los valores extremos, también refleja la dispersión pero valores sueltos afectan tanto el rango que su resistencia es despreciable.

La *distancia intercuartil*, d_C, o rango intercuartil, es muy similar a la distancia intercuartos pues está basada en los cuartiles que son casi lo mismo que los cuartos

$d_C =$ CUARTIL SUPERIOR - CUARTIL INFERIOR

$$\text{profundidad del cuartil} = \frac{n+1}{4}$$

En el ejemplo, la profundidad del cuartil es $\frac{59+1}{4} = 15$ y por lo tanto d_C= 63.84-63.36=0.48

Identificación de valores atípicos

Para examinar datos en busca de valores atípicos necesitamos una medida de dispersión que sea insensible a ellos.

Necesitamos una medida de dispersión que enfatice el comportamiento de la porción central de los datos. Esto lo hace la distancia intercuartos.

Definiremos *puntos de corte internos y externos* de manera que los datos que estén fuera de esos puntos serán identificados como *valores externos; moderados o severos* respectivamente.

Muchos de estos valores no serán realmente atípicos en el sentido que su comportamiento subyacente es el mismo que el de la mayoría de los datos.

$$\text{Valla Interna Inferior} = Q_I - 1.5\, d_Q$$

$$\text{Valla Interna Superior} = Q_S + 1.5\, d_Q$$

$$\text{Valla Externa Inferior} = Q_I - 3\, d_Q$$

$$\text{Valla Externa Superior} = Q_S + 3\, d_Q$$

Observación: Las vallas tienen la misma longitud desde sus respectivos cuartos.
La figura 8 muestra los cuartos y los puntos de corte

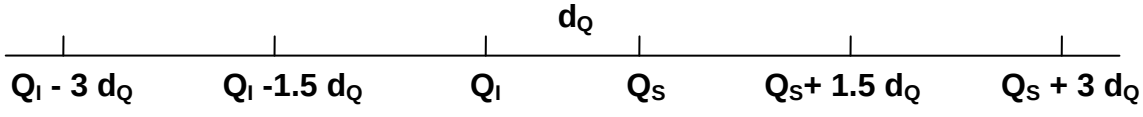


Figura 8

¿Qué chance tiene un dato de caer fuera de los puntos de corte, Q_I -1.5 d_Q y Q_S +1.5 d_Q , cuando los datos provienen de una distribución gaussiana?

Comenzaremos, por simplicidad, con la distribución misma (correspondiente a una muestra muy muy grande) en vez de una muestra finita.

El cuarto corresponde a un área de 0.25

```
> qnorm(0.25)
[1] -0.6744898
```

de la cola de la distribución, de manera que los cuartos son μ - 0.6745σ y μ + 0.6745σ, dando una distancia intercuartos de 1.348σ.

Esto hace que las vallas internas estén en μ - 2.698σ y μ + 2.698σ, dando un área de .00349 en cada cola.

```
> 2*pnorm(-2.698)
[1] 0.006975744
```

Por lo tanto la proporción de la distribución que cae debajo de μ - 2.698σ ó encima de μ + 2.698σ es .00698.

En muestras finitas, la proporción promedio de observaciones que están más allá de las vallas internas es sustancialmente mayor que el valor poblacional. De un estudio de simulación, Hoaglin, Iglewicz y Tuckey (1981) obtuvieron la siguiente cota

superior de la fracción promedio de valores externos en muestras gaussianas con $n \geq 5$:

$$0.00698 + \frac{0.4}{n} \qquad (5)$$

Si nos interesa el la cantidad de datos externos que se espera tener, en promedio, debemos multiplicar la expresión (5) por n , obteniendo para un único lote

$$0.007n + 0.4,$$

y para varios lotes

0.007(cantidad total de observaciones) + 0.4 (cantidad de lotes).

Este resultado es útil siempre que recordemos que se aplica a distribuciones gaussianas y que las distribuciones con colas pesadas tendrán más valores externos por la derecha.

La distancia intercuartos permite obtener una estimación resistente de σ. Hemos visto que los cuartos en la distribución gaussiana son $\mu - 0.6745\sigma$ y $\mu + 0.6745\sigma$, dando d_Q(poblac. gaussiana) = 1.348σ. Por lo tanto, para un conjunto de datos, definimos

$$pseudosigma = \frac{d_Q}{1.349}.$$

Cuando los datos son gaussianos pseudosigma da un estimador de σ, y su valor será generalmente cercano al de *s.*

Si las dos estimaciones difieren sustancialmente deberíamos preferir pseudosigma y buscar las observaciones que han inflado a *s.*

BOXPLOTS

El boxplot es la representación gráfica de la mediana, los cuartos, los valores adyacentes y los valores externos moderados o severos. Permite extraer los siguientes aspectos del lote:

Posición del centro
Dispersión
Asimetría
Longitud de la cola
Puntos que yacen fuera del conjunto.

Este compacto diagrama es muy útil para comparar varios lotes de datos.

Para construir el diagrama nos falta definir los siguientes valores:

$$\text{VALOR ADYACENTE INFERIOR (VAI)} = \begin{cases} \text{valor más cercano, mayor o igual,} \\ \text{a la valla interna inferior} \end{cases}.$$

VALOR ADYACENTE SUPERIOR (VAS) = { valor má s cercano, menor o igual, a la valla interna superior.

Boxplot para un único lote

Introduciremos el boxplot analizando las poblaciones de las 15 ciudades más grandes de USA. Los datos se encuentran en un archivo texto que tiene 2 columnas la primera encabezada con “Ciudad” contiene los nombres de las ciudades, la segunda encabezada con “pobl” la cantidad de habitantes.

Ciudad	pobl
NewYork	778
Chicago	355
LosAngeles	248
Philadelphia	200
.....	

```
pobl <- read.table(file.choose(),
header=T, row.names="Ciudad")  #¿da
error?
pobl <- read.csv2(file.choose(),
header=T, row.names="Ciudad")
```

Podemos acceder a los valores de población de cada ciudad por su nombre

```
> pobl["New York",]
[1] 778
> pobl["Detroit",]
```

```
[1] 167
```

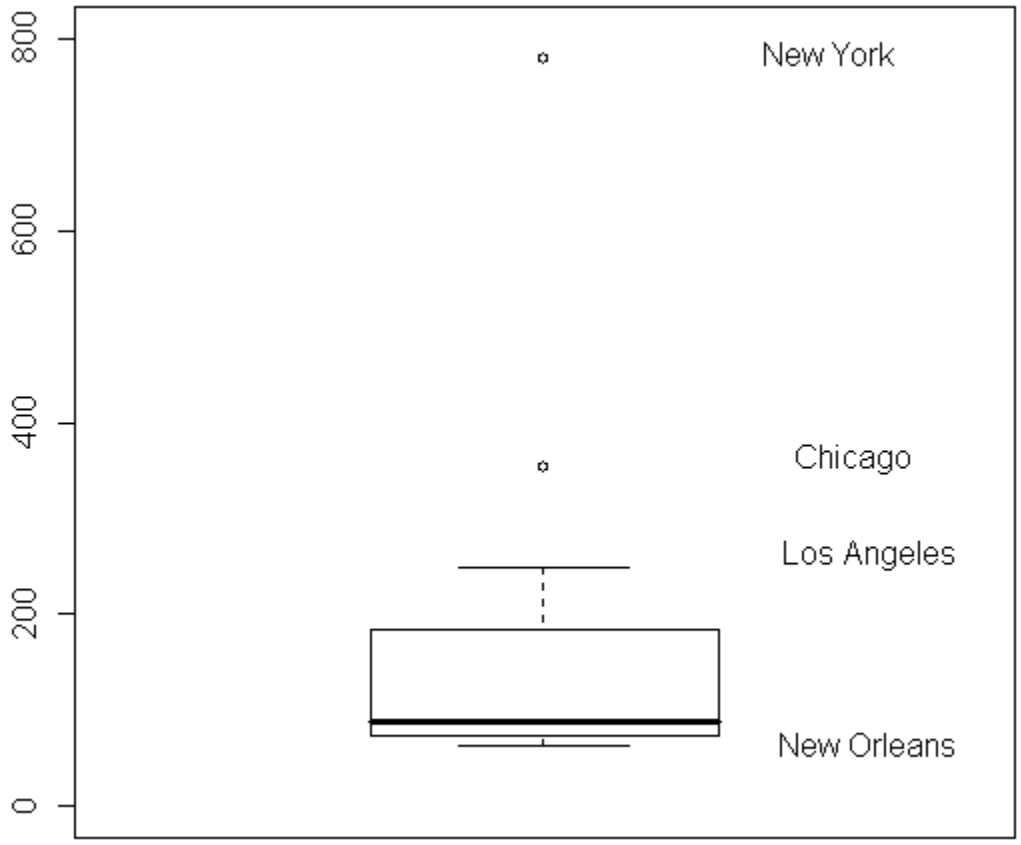


Figura 11. Boxplot de la población de las 15 ciudades más grandes de USA.

En **R**

```
boxplot(pobl$pobl, ylim=c(0,800))
  ciudades<- c("New York","Chicago","Los
Angeles","New Orleans")
 text(locator(1),ciudades[1])
 text(locator(1),ciudades[2])
 text(locator(1),ciudades[3])
 text(locator(1),ciudades[4])
```

Veamos como se construye el boxplot en el caso de las poblaciones de las 15 ciudades más grandes de USA.

Tabla 9

Ciudad	Población(10000)
New York	778
Chicago	355
Los Angeles	248
Philadelphia	200
Detroit	167
Houston	94
Baltimore	94
Cleveland	88
Washington DC	76
St. Louis	75
San Francisco	74
Milwakee	74
Boston	70
Dallas	68
New Orleans	63

El boxplot se construye dibujando

i) una caja cuyos extremos son los cuartos inferior (Q_I =74) y superior (Q_S= 183.5) y con una barra vertical en la mediana (88),

ii) una línea de cada extremo de la caja hasta el correspondiente valor adyacente (VAI = 63 VAS = 248),

iii) los valores que caen fuera de las vallas internas pero dentro de las externas son outliers moderados (Chicago),

iv) los valores que caen fuera de las vallas externas son outliers severos (New York).

Se puede ver,

- la *posición* del lote, indicada por la mediana, la barra que divide a la caja.

- la *dispersión,* longitud de la caja: distancia intercuartos.

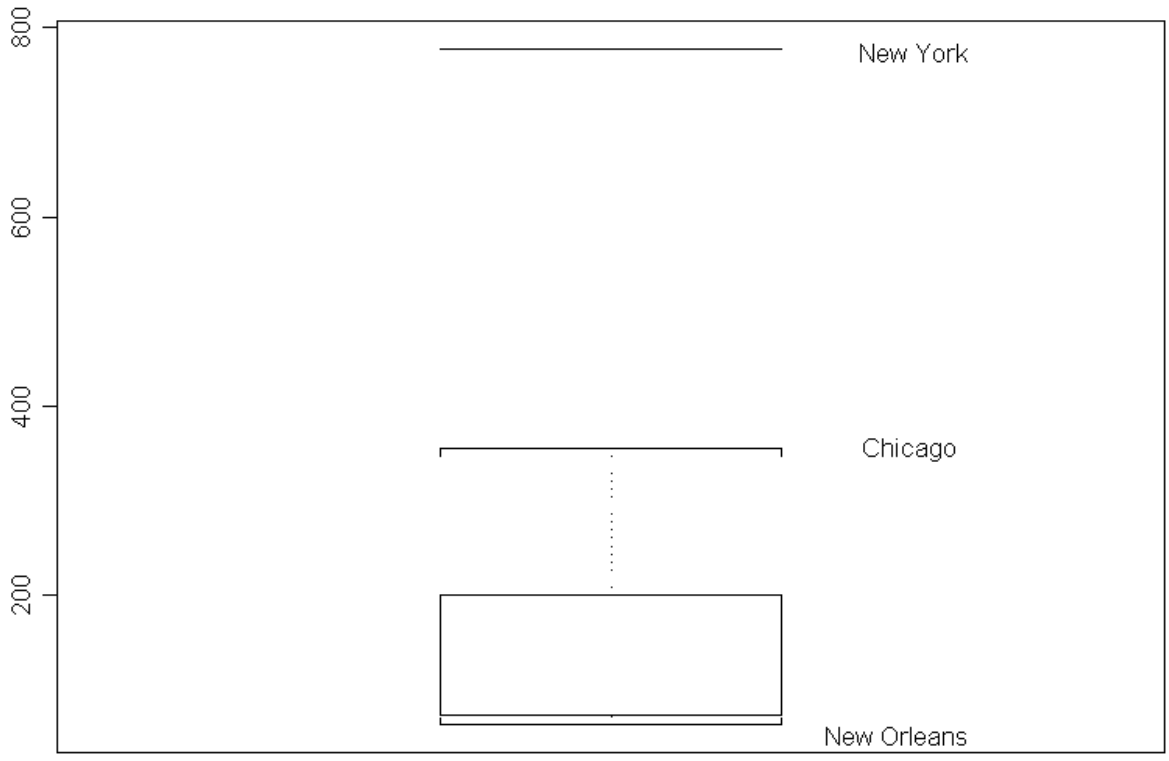
- la *asimetría;* posición relativa de la mediana, al cuarto inferior y al cuarto superior. Vemos asimetría positiva (ó a derecha), una situación común en datos positivos no acotados.

- la *longitud de las colas* mediante las líneas que se extienden desde New Orleans hasta Los Angeles

- los *valores atípicos* (Chicago y New York).

El mensaje del boxplot para la población de las 15 ciudades más grandes de USA es fuerte: el lote es pesadamente asimétrico y hay dos puntos atípicos.

En S-PLUS, Chicago no aparece como valor atípico (outlier).




```
> boxplot(pobl,boxcol=-1)
```

```
> ciudades<- c("New York","Chicago","New Orleans")
> text(locator(1),ciudades[1])
> text(locator(1),ciudades[2])
> text(locator(1),ciudades[3])
```

```
> summary(pobl)
 Min. 1st Qu. Median  Mean 3rd Qu. Max.
   63      74     88 168.3   183.5  778
```

```
> aaa<- boxplot(pobl,plot=F)
```

```
> aaa
$stats:
     [,1]
[1,]  355
[2,]  200
[3,]   88
[4,]   74
[5,]   63
```

El cuartil superior no coincide con el calculado anteriormente (183.5), sí lo hacen el inferior, la mediana, el máximo (sin el outlier) y el mínimo. Los cuartiles resultan de elegirlos en la profundidad (n+1)*0.25.

Resistencia del Boxplot

La caja rectangular del boxplot está construida con medidas resistentes. Hasta un 25% de los datos pueden hacerse arbitrariamente grandes (valores salvajes) sin perturbar demasiado a la mediana, a los cuartos y a la caja en el boxplot.

Las colas del boxplot están determinadas por los datos extremos **dentro** de los puntos de corte para los outliers.

De esta manera no están perturbadas por grandes cambios en los valores de los outliers. Los puntos de corte también pueden resistir grandes perturbaciones del 25% de los datos pues están basados solamente en los cuartos.

Un gráfico similar podría construirse en base a la media muestral y el desvío estándar muestral tal gráfico carecería de resistencia, inclusive ante la presencia de un único dato salvaje.

Interpretación del Boxplot.

Los puntos de corte para identificar outliers (Q_I -1.5 d_Q, Q_S +1.5 d_Q) son un poco arbitrarios, pero la experiencia muestra que esta definición sirve para identificar puntos que requieren atención especial.

Mostramos que, para una población gaussiana, el .7% de la población está fuera de los puntos de corte. Es decir que dentro de los puntos de corte se encuentra el 99.3% de la población.

Consideremos los valores poblacionales de la mediana, los cuartos y los puntos de corte para los valores atípicos en varias distribuciones conocidas.

Las tablas 10 y 11 muestran estos valores y las probabilidades más allá de los puntos de corte para distribuciones simétricas y asimétricas respectiv.

Tabla 10. Valores poblacionales de la mediana, los cuartos, cortes para outliers para varias distribuciones simétricas.

Distr.	Mediana	Cuarto Superior	Cortes para outliers	% de outliers	Valor de 1.96σ	% Fuera de $\mu \pm 1.96\sigma$
$U(-1,1)$	0	0.500	±2.000	0	1.132	0
$N(0,1)$	0	0.674	±2.698	0.70	1.960	5.00
t_{20}	0	0.687	±2.748	1.24	2.066	5.20
t_{10}	0	0.700	±2.800	1.88	2.191	5.32
t_5	0	0.727	±2.908	3.35	2.530	5.25
t_1	0	1.000	±4.000	19.59	------	-----

Para la distribución uniforme (con colas extremadamente cortas), todos los valores están dentro de los puntos de corte para outliers.

Para la distribución gaussiana, el 0.7% de los valores son considerados outliers.

Elegimos las distribuciones *t* con distintos grados de libertad para representar distribuciones simétricas con colas pesadas. A medida que las colas se vuelven más pesadas el porcentaje de valores fuera de los puntos de corte es mayor y por lo tanto tenemos una mayor probabilidad de observar outliers.

Esto nos permite juzgar si nuestros datos presentan colas más pesadas que las gaussianas a través de la

cantidad de valores que caen fuera de los puntos de corte.

Tabla 11. Valores poblacionales de la mediana, los cuartos, cortes para outliers para varias distribuciones asimétricas.

Distr.	Mediana	Cuartos Superior e Inferior	Cortes para outliers	% de outliers	Valor de 1.96σ	% Fuera de $\mu \pm 1.96\sigma$
χ^2_1	0.45	0.102 1.323	-1.750[f] 3.155	7.58	-1.772 3.772	5.22
χ^2_5	4.35	2.675 6.626	-3.252[f] 12.552	2.80	-1.198 11.198	4.78
χ^2_{20}	19.34	15.452 23.828	2.888 36.392	1.39	7.604 32.396	4.53

[f]Para distribuciones asimétricas uno de los puntos de corte generalmente cae fuera del rango de los valores posibles.

Las distribuciones chi-cuadrado son ejemplos de **distribuciones asimétricas**; desde la más asimétrica, χ^2_1, hasta las **un poco más simétricas** χ^2_5 y χ^2_{20}.

Observamos una característica que frecuentemente ocurre en situaciones asimétricas (ocurrió en nuestro ejemplo de las ciudades de USA):

el menor de los puntos de corte está por debajo del menor dato posible,

la probabilidad de un outlier en este lado es 0, obtenemos una indicación de la asimetría no sólo por

la posición de la mediana y los cuartos sino también por la posición de los outliers a un lado de la caja.