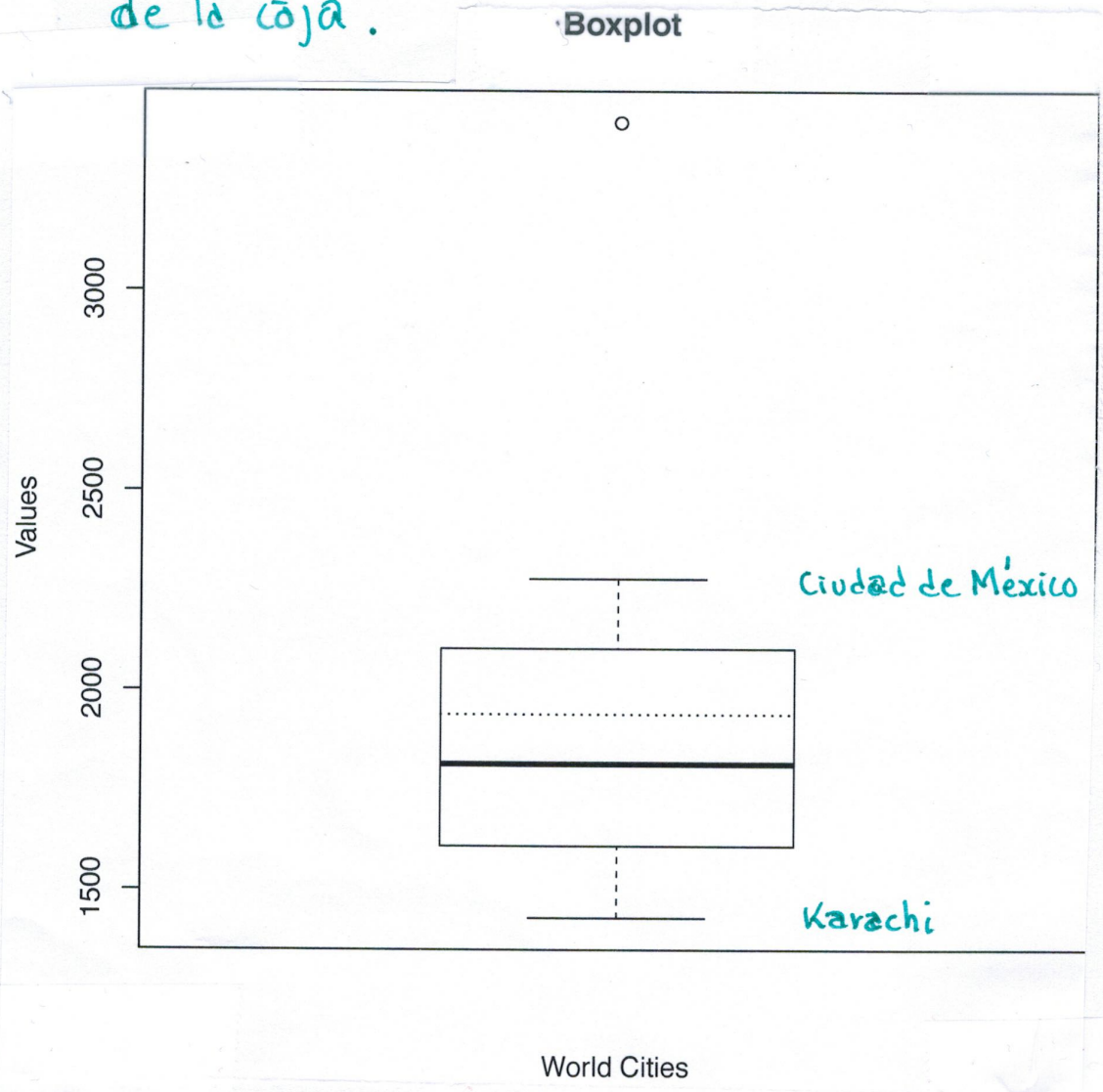
$x^* = 2280$ (Ciudad de México)

$x_* = 1430$ (Karachi)

Entonces los bigotes se dibujan desde el borde superior de la caja, hasta el valor $x^*$ (ciudad de México), en la parte de arriba de la caja y, desde el borde inferior de la caja hasta el valor $x_*$ (Karachi), en la parte de abajo de la caja.



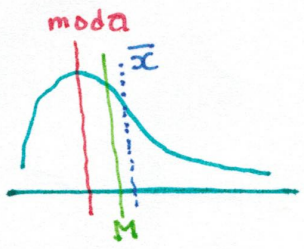
El diagrama de caja muestra que los datos de tamaños poblacionales de las ciudades tienen un sesgo[1] hacia arriba (hay falta de simetría en la distribución de X = tamaño poblacional).

La mitad superior (los datos arriba de la mediana) de la muestra tiene una dispersión mayor que la mitad inferior. Hay un dato aberrante marcado con caracter "O" y este corresponde a la ciudad de Tokyo. Debido a la sobredispersión en la mitad de los datos que están arriba de M la media, que no es una medida de tendencia central robusta, está desplazada hacia arriba y no coincide con la mediana.

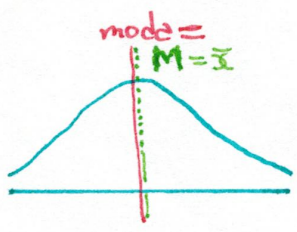
---

(1) Si $\bar{x} \neq M$ podemos decir que tenemos evidencia muestral que nos sugiere la hipótesis de que la distribución esta sesgada (la distribución de X).

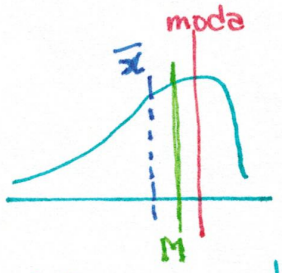
En el caso de distribuciones unimodales, en general pero no siempre, se tiene




(a) sesgo positivo




(b) Simetría




(c) Sesgo negativo

Ejercicio: Para los datos correspondientes al rendimiento, millas por galón de combustible, de los automóviles provenientes de Japón, Norte América y Europa, encontrar en cada caso, las estadísticas:

$x_*, x^*, M, F_L, F_U, b_L, b_U, \overline{x}$

Haga un análisis de estos datos usando los diagramas de cajas, desarrolle sus conclusiones.[1]

1 Las conclusiones son parte importante del ejercicio.

Ejemplo: Datos de billetes del banco de Suiza

$n = 200$ observaciones $\mathbf{x}_1, \ldots, \mathbf{x}_{200}$,

$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{i6})$ ; $i = 1, 2, \ldots, 200$.

Las observaciones corresponden a $\mathbf{X} = (X_1, X_2, \ldots, X_6)$ donde

$X_1$ = largo del billete

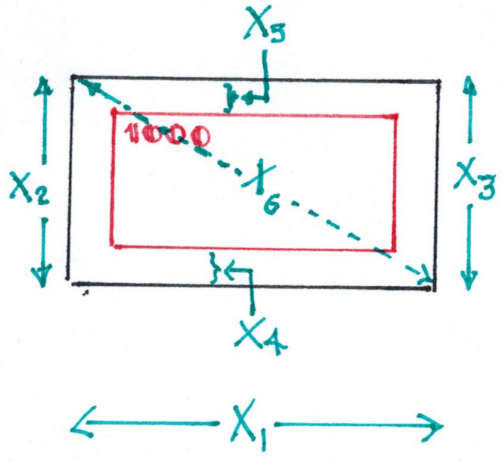
$X_2$ = Ancho del billete (izquierda)

$X_3$ = Ancho del billete (derecha)

$X_4$ = Distancia de la Figura en el billete al borde inferior del billete.

$X_5$ = Distancia de la Figura en el billete al borde superior del billete

$X_6$ = Longitud de la Diagonal del billete



Objetivo: Estudiar cómo estas mediciones podrían usarse para determinar si un billete es genuino ó si es falso.

Preguntarse como determinar si un billete es verdadero o falso usando estas seis mediciones esta relacionado con una de las características de interés que se mencionaron al inicio del tema del análisis exploratorio de datos:

¿Hay componentes de $X$ que indican la existencia de subgrupos ó aglomeraciones en los datos?

Ver página 2 →

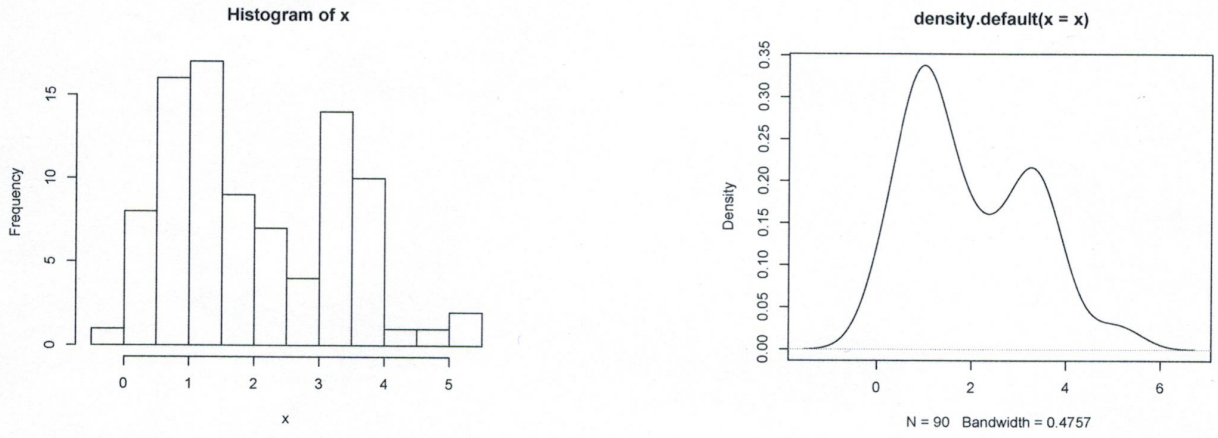
Dicho de otra forma, nos interesaría saber si una de las seis mediciones $X_1, .., X_6$ nos permite ver dos subgrupos dentro de los datos: los verdaderos y los falsos.

En este caso, partir directamente de los diagramas de cajas posiblemente no sea una idea muy útil, estos diagramas son representaciones unidimensionales de la distribución de una variable aleatoria (cualquiera de $X_1, X_2, .., X_6$). La forma

en que el diagrama de caja esta construido no permite al usuario de esta representación determinar si la distribución de $X_i$ tiene más de una moda. Por ejemplo, si simulamos una muestra en donde hay bimodalidad

```
> x ← c(rnorm(50,1,0.5), rnorm(40,3,1))
> hist(x)
> plot(density(x))
```

Figura A datos simulados




En la realidad, podemos tener una muestra como esta, es decir una muestra en donde al parecer

hay dos subpoblaciones que juntas forman la población total. Pensemos por un momento que la muestra no fue simulada, que los datos fueron observados con la naturaleza que muestran las figuras, entonces al percatarnos de que hay dos modas, la idea de que hay dos tipos de individuos en la población es inmediata. A continuación tendríamos que investigar si las observaciones de uno de estos dos grupos provienen de individuos con una característica particular, por ejemplo si hablamos de los billetes, hay que averiguar si podemos "clasificar" los individuos de un grupo como billetes falsos y a los individuos del otro grupo como billetes verdaderos. Como ya se mencionó arriba los diagramas de caja no nos ayudan a determinar si hay dos (ó más) grupos: La Figura B muestra el diagrama de caja correspondiente a los datos simulados en la Figura A