INTERVALO DE CONFIANZA ROBUSTO PARA LA MEDIA

David Barrera

El mensaje principal de este trabajo es que los estimadores robustos son sumamente útiles cuando se identifica observaciones atípicas

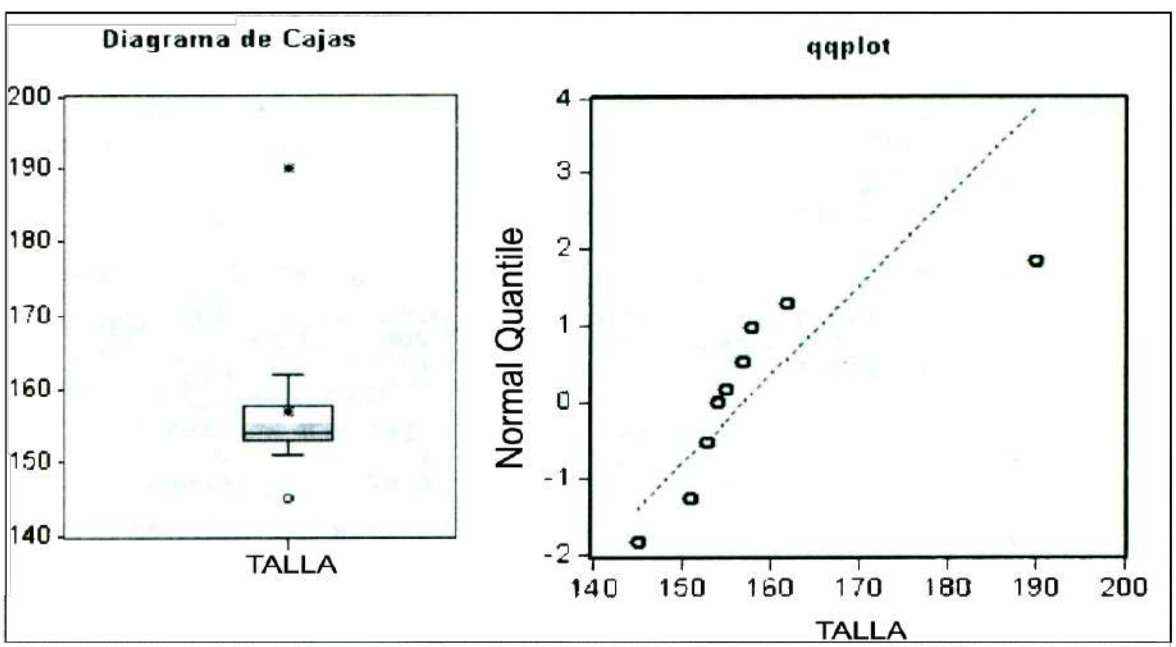
1. Introducción

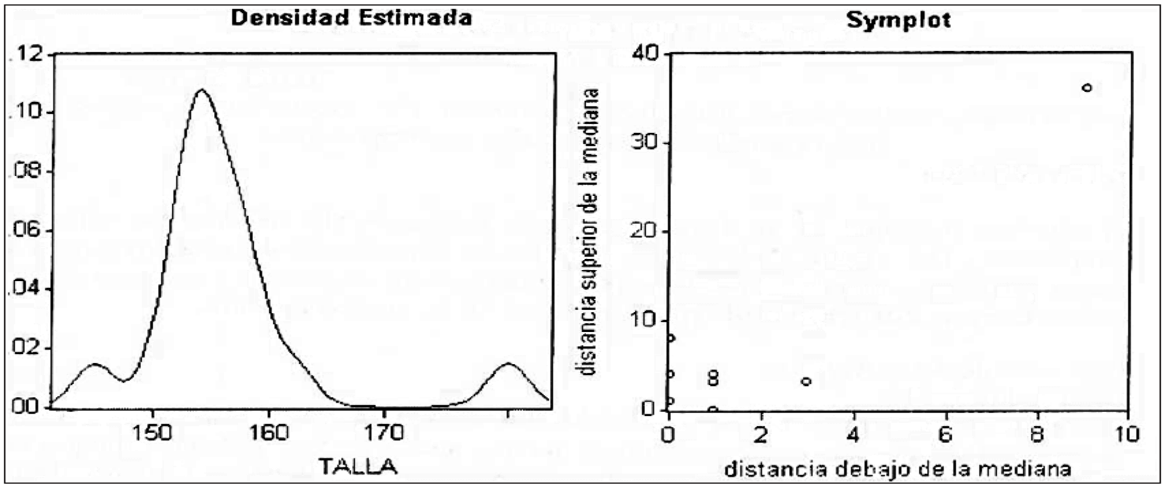
El objetivo principal de la Estadística es la obtención de información relevante (confiable) y útil a partir de los datos, por ello, es imprescindible que estos tengan la mayor precisión y fiabilidad posibles, por lo que uno de los procesos más importantes de esta Ciencia, está relacionado con el análisis y la depuración de datos.

Detección de datos Atípicos
Pruebas Informales
Los gráficos pueden ser útiles, para identificar datos atípicos de una variable en un momento dado. Pasar por alto estas observaciones puede ser letal para las conclusiones luego del análisis estadístico. Las causas pueden ser variadas: Uno puede deberse a un accidente en el momento de la trascripción de los datos, y otro que realmente exista estas observaciones.

Como ilustración, vamos a considerar los siguientes datos numéricos, relativos a las tallas de 15 alumnos de un colegio, las unidades están centímetros

Alumno	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Talla	153	145	154	154	162	155	158	154	157	151	158	157	153	190	153





Los cuatro gráficos claramente muestran, la presencia de un valor atípico, es decir éste es posible que los estadísticos clásicos muestren un sesgo fuerte devido a éste valor.

Pruebas Formales

En la ventana de comandos del Stata, ejecutar:

```
File  Edit  Search  Tools

program drop _all
preserve
grubbs talla, gen(gr_talla ) iter(1000) level(95)
list talla if gr_talla==1
iqr talla
```

Luego de pulsar RUN, tendremos:

```
      talla
 1.     190

 iqr talla

   mean=  156.9         std.dev.=   9.91      (n= 15)
 median=    154  pseudo std.dev.=  3.706    (IQR=       5)
10 trim=  155.1
                                    low          high
                                  -------------------
                  inner fences      145.5        165.5
                # mild outliers     1            0
                % mild outliers     6.67%        0.00%

                  outer fences        138          173
              # severe outliers     0            1
              % severe outliers     0.00%        6.67%
```

El Estadístico iqr, afirma que efectivamente existe la presencia de un valor atípico severo en el extremo derecho, y la prueba de Grubbs confirma con una seguridad del 95%. Es decir con la prueba formal queda confirmado lo que se sospechaba con la prueba informal.

Algunos estadísticos básicos:

	Estadísticos	
	Con atípicos	Sin atípico
Mean	156.93	154.57
Median	154.00	154.00
Maximum	190.00	162.00
Minimum	145.00	145.00
Std. Dev.	9.91	3.96
Skewness	2.59	-0.54
Kurtosis	9.67	4.05

Los Estimadores Puntuales no difieren mucho(o no parece muy significativo), empero si se observa la desviación estándar, en ambos casos difiere bastante. ¿Entonces que pasaría si deseamos estimar en un intervalo de confianza el promedio poblacional, es decir tomando en cuenta la precisión del estimador?

Para estimar en un intervalo de confianza, vamos a asumir alegremente que provienen de una población normal. Usando un nivel de confianza del 95%, tendremos:

	A	B	C
1		Con Atípicos	Sin Atípicos
2	Lim. Inf.	151.445	152.287
3	Lim. Sup.	162.421	156.855
4	Longitud	10.976	4.568

Tomando en cuenta la precisión del estimador, la diferencia de las longitudes si es significativa. A simple vista, si pasamos por alto éste valor atípico la estimación tendría la el error estándar de éste sería mayor.

¿Cómo obtener Un Intervalo de Confianza Robusto, a sabiendas de que existe un valor atípico significativo?

Uno de los estimadores robustos más viejos es la mediana, vamos a utilizar en ésta ocasión como estimador de la media a la Mediana. No olvide que estamos asumiendo que la población es normal.

Inferencia con la Mediana Muestral:

$$\textbf{Sea } U_n = n - L_n \textbf{, donde } L_n = \left[\frac{n}{2}\right] - \left[\sqrt{\frac{n}{4}}\right] \textbf{ y la desviación estándar de la mediana}$$

$$SE(MED(X)) = 0.5 * (Y_{(U_n)} - Y_{(L_{n+1})})$$

$$\textbf{Donde } p = U_n - L_n - 1\textbf{. Con un intervalo de confianza para la mediana es}$$

$$MED(X) \mp t_{p,1-\frac{\alpha}{2}}\ SE(MED(X))$$

observación, con el único

	A	B	C	D	E	F	G
1	Intervalo de Confianza Robusto para la Media						
2	Obs.	Mediana	E.E	Lim. Inf.	Lim. Sup.	Nivel Conf.	Longitud
3	15	154.0000	0.5000	151.8487	156.1513	0.950	4.3027

supuesto de qué existe valores atípicos. Si comparamos con el intervalo de confianza clásico, la longitud del intervalo de confianza robusto es menor. La diferencia es significativa, por tanto éste último es resistente a los valores atípicos.

Nota.- Es necesario puntualizar ésta técnica es válido cuando la distribución es aproximadamente normal, y se detecta la presencia de valores atípicos. Con éste pequeño trabajo, es posible extender el tema también a pruebas de hipótesis robustas, regresión robusta, componentes principales robusto, etc. Es decir que existe un gran desafío para el estadístico.

Bibliografía:

- Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972), *Robust Estimares of Location*, Princeton University Press, Princeton, NJ.
- Olive, D.J. (2005), "Applied Robust Statistics" Southern Illinois University.

Software:
Eviews, Stata

David Barrera Ojeda
Email: david_barrera__ojeda@yahoo.es

Índice