Notas Estadísticas: Cuartiles y Percentiles

Carrera, Elena; Vaira, Stella M; Contini, Liliana E.

Dpto. de Matemática. Área Estadística. Facultad de Bioquímicas y Ciencias Biológicas. Universidad Nacional del Litoral. C.C. 242. Ciudad Universitaria. Paraje El Pozo S/Nro. (3000) Santa Fe. Argentina. e-mail: ecarrera@fbcb.unl.edu.ar

Resumen: Esta es la primera de una serie de notas sobre Estadística. Cuando se analizan mediciones de variables cuantitativas, algunas veces es necesario establecer conjuntos que contengan igual número de observaciones en cada uno. Los cuantiles son los valores de corte de esos grupos. Los cuartiles, deciles y percentiles son casos particulares de cuantiles. Si bien no son fáciles de calcular y puede hacerse de distintas manera, se propone una forma para ello.

Palabras claves: cuantiles, cuartiles, deciles, percentiles.

SUMMARY: Statistics notes: Quartiles and Percentiles. Carrera, Elena; Vaira, Stella M; Contini, Liliana E. This is the first of a series of notes on Statistics. When presenting analysing measurements of continuous variables, it is sometime necessary to place them in several groups, each of them showing an equal number of observations. So the quantiles are the cut off points that split the data. Quartiles, deciles and percentiles are particular occurrences of quantiles. Although quantiles are not easy to calculate, we propose one possible way in which this can be carried out.

Key words: quantiles, quartiles, deciles, percentiles.

Introducción

Si se quiere justificar el por qué de estas notas lo esencial es dar un concepto de Estadística, ésta *es el arte de hacer inferencias y extraer conclusiones a partir de datos imperfectos* [1]. La estadística aplicada aparece hoy fuertemente ligada a la economía, la ingeniería y cualquier otra área de la ciencia y de la técnica. Las Ciencias Biológicas y la Medicina no han sido ajenas a esto y para referirse a la aplicación de la estadística a sus áreas acuñaron los nombres de Bioestadística o Biometría. Ésta es una disciplina relacionada con el análisis estadístico de datos generados desde problemas biológicos, particularmente en el área de la salud y de la enfermedad; desarrolla y aplica métodos estadísticos y matemáticos, éstas teorías cuantitativas favorecen la aplicabilidad mencionada para proveer a un mejor entendimiento y solución de problemas en áreas como Biología, Medicina, Bioquímica, Epidemiología y Salud Ambiental. *Una de las principales razones de la existencia de los métodos estadísticos es la de hacer claras y simples todas las características interesantes o importantes que un conjunto de datos puede contener. Siendo los conceptos estadísticos, luego de su recolección, una guía de análisis que permite extraer la mayor información posible de los datos* [1].

Las Ciencias Biomédicas han incorporado el uso del método Estadístico desde su empírico nacimiento hasta el avanzado desarrollo de nuestros días. Hoy ya no se discute que los principios y métodos Estadísticos se necesitan no sólo para la comprensión, sino también para el ejercicio eficaz en cualquiera de las profesiones que actúan en el campo de la salud, ya que la variabilidad de los datos clínicos, biológicos y de laboratorio, tanto en individuos como en comunidades, hacen que la toma de decisiones

Recibido: 6-6-02
Aceptado: 24-10-02

vaya siempre acompañada de un grado de incertidumbre(2).

En años recientes el empleo de métodos estadísticos ha crecido enormemente. Las revistas donde se publican trabajos científicos exigen la justificación de las técnicas estadísticas empleadas(3, 4, 5). La cantidad de software disponible con el advenimiento de las computadoras personales lo ha hecho a un ritmo también creciente. Esto ha ocasionado, por un lado, el crecimiento y desarrollo sostenido de sofisticada metodología estadística para la aplicación por parte de los científicos y de los estadísticos aplicados, por el otro se ha detectado un uso indiscriminado y a veces equivocado de las técnicas estadísticas que se encuentran en los programas de computación(3, 6, 7, 8). Aparecen así con frecuencia algunos errores por el uso inapropiado de métodos que son adecuados para otro tipo de datos o para otro tipo de aplicaciones. Según Altman(9) (1994) *algunas personas piensan que todo lo que uno necesita para hacer estadística es una computadora y el software apropiado. Esta visión es errada cuando se analiza, pues, ciertamente ignora, la consideración esencial del estudio del diseño, los fundamentos con los cuales la investigación es construida.*

Teniendo en cuenta lo señalado anteriormente, las palabras de Mood(10) (1966) aparecen como siempre acertadas *el empleo del instrumental estadístico no es una simple cuestión de escoger la llave que mejor ajusta al perno, más bien se trata de elegir entre varias llaves, todas las cuales parecen adaptarse igual de bien, pero ninguna ajustar exactamente al mismo.* Numerosos estudios usan técnicas equivocadas, o bien la técnica correcta equivocadamente, hacen pobre interpretación de los resultados, reporte selectivo de los mismos y extracción de conclusiones injustificadas(9, 11).

Tema de interés permanente es distinguir las técnicas estadísticas para datos categóricos de aquellas que se utilizan para datos discretos y de las que se utilizan para datos continuos. Como analizar el tipo de variable para efectuar regresiones que no siempre son lineales, interpretar correctamente los coeficientes relacionados con la correlación, calcular tamaños muestrales y evitar la aparición de valores p sin la correspondiente identificación de la prueba estadística al que está asociado o según Abraira(11) (2001) valores p "huérfanos". Se debe también asegurar la aleatoriedad empleada en los intervalos de confianza (3, 12, 13, 14, 15) y tener en cuenta la representatividad de ciertas medidas tales como la media, la mediana, el desvío estándar y el error estándar para indicar cual resume mejor los datos. Hoy han aparecido con fuerza el empleo de los cuartiles, deciles y percentiles (16, 17, 18), una gran variedad de gráficos disponibles (1, 19, 20), técnicas inferenciales que se basan en el análisis de estos estadísticos como las pruebas no paramétricas para comparar medianas y pruebas de bondad de ajuste que se basan en el cálculo de cuantiles (1, 7, 17, 18, 21, 22, 23).

Medidas de posición

En algunos casos de medidas cuantitativas es necesario establecer puntos de corte, de tal manera que entre esos valores quede el mismo número de observaciones, surgen así los llamados *cuantiles* o *cuantilos*, que describen la posición de una observación relativa a las demás observaciones, son posibles valores de la variable. Así los más conocidos son los cuartiles, deciles (o decilos) y percentiles (o percentilos). Todos estos son *estadísticos de orden* porque surgen después que los datos se han ordenado en forma creciente. Cabe aclarar que estas medidas de posición se extienden a las variables numéricas en general, ya sean continuas o discretas, sin bien los deciles y percentiles son más utilizados para datos continuos.

Los cuartiles son los posibles valores de la variable que dividen el conjunto de datos en cuatro partes que contienen el mismo número de observaciones. El 25% del total está así contenido en cada grupo. La diferencia entre el máximo y el mínimo valor observado se denomina *rango* (19, 21, 22, 23). El rango ($x_{max} - x_{min}$) queda dividido por estos puntos que obviamente no deben necesariamente estar equidistantes (Figura 1). Se debe reiterar que el primer cuartil (Q_1), segundo cuartil (Q_2) y tercer cuartil (Q_3) no son los intervalos sino los valores posibles de la variable o puntos de corte (cut-off) que limitan superiormente cada grupo(16). La distancia entre el primer y tercer cuartil es denominado *rango intercuartílico* (Figura 1).

Estos cuartiles permiten la realización de un gráfico muy difundido en la actualidad llamado gráfico de "caja y bigotes" o Box and Whisker Plot (1, 19, 20) (Figura 2). Un clásico diagrama de cajas contiene 5 medidas resumen de los datos: mínimo, primer cuartil, segundo cuartil, tercer cuartil y máximo (1, 18, 19, 20). En otros, según el comportamiento de la

variable, además de estos cinco números se identifican los valores extremos o inusuales (outliers) (1, 17, 18, 19, 20). Intuitivamente estos valores son aquellos que se encuentran suficientemente lejos de los restantes (Figura 2-b). Esta lejanía se establece tomando como extensión máxima del "bigote" una vez y media el rango intercuartílico, por encima de Q_3 y por debajo de Q_1. Por lo cual se puede suponer que hubo errores en la toma de datos o bien que el individuo tiene características que lo hace pertenecer a otra población, como consecuencia de esto debe analizarse especialmente ese dato y esto posiblemente abra nuevas investigaciones(18).

La mediana es el segundo cuartil, su uso es muy común en el análisis de los tiempos de sobrevida y distribuciones de parásitos entre otros (6, 7, 16).

Cuando el número total de datos se divide en grupos que contengan el 10% de las observaciones o el 1% de las mismas, con lo cual el total de observaciones es dividida en 10 o 100 partes iguales, los puntos de corte que quedan determinados son los llamados deciles y percentiles respectivamente. A estos últimos también se los llama centilos o percentilos y fueron inventados por William Porter(24). Resulta evidente que el primer cuartil es el percentil 25 y que el tercer cuartil es el percentil 75. Es aceptado que los centiles se numeran de acuerdo al porcentaje de individuos que están debajo de ellos(7, 17, 18, 20, 21). Así el percentil 75, deja debajo de él el 75% de las observaciones y el 25% de las mismas están por encima de él. Es importante no olvidar la relación de los cuantiles con los valores críticos de las distribuciones teóricas que se utilizan en intervalos de confianza y pruebas de hipótesis. Al construir un intervalo del 95% de confianza para la media de una población normal, en el cálculo interviene $z_{0,025}$ y $z_{0,975}$ que no son más que los percentiles 2,5 y 97,5 respectivamente de una distribución normal estándar (1, 6, 7, 17, 18, 20, 21, 22, 23).

Los centiles son vastamente utilizados en medicina, psicología y educación y en la actualidad su empleo ha aumentado debido a la importancia adquirida por el control de calidad, ya sea en los laboratorios o en la industria.

Cálculo de percentiles

Existen variadas formas de calcular los percentiles de un conjunto de datos obtenidos en una experiencia o sea de datos muestrales. Todas ellas dan resultados similares, si bien la expresión matemática que permite calcular su posición presenta algunas variantes con fórmulas más o menos complicadas y que son las utilizadas por los programas de computación especiales para estadísticas. También el criterio para elegir el valor del percentil correspondiente difiere en los distintos autores consultados. Se propone a continuación una manera de efectuar este cálculo que es compartida por varios autores relacionados con la Biometría (1, 6, 7,16).

Una vez ordenadas las n observaciones en forma creciente, para calcular el percentil k que se puede simbolizar P_k, se procede calculando su posición a la que se llamará i, donde

$$i = k\frac{n+1}{100}$$

Este valor puede ser entero o no. Si es un número entero habrá que buscar la observación que ocupa esa posición, así el percentil k será el valor de la variable que tiene la posición i. Si no es entero, situación más frecuente, el i se hallará comprendido entre dos enteros consecutivos ($j < i < j+1$) entonces el P_k será el promedio entre $x_{(j)}$ y $x_{(j+1)}$ o bien se puede hallar por interpolación. Por ejemplo se tienen 53 observaciones y se desea calcular el percentil 33, se hace la cuenta

$$33\frac{53+1}{100} = 17{,}82$$

por lo tanto el P_{33} es un valor posible de la variable entre aquellos que ocupan la posición 17 y 18, ¿cómo encontrar ese valor? Por interpolación o sea, si el valor de la posición 17 es 12,2 y el de la posición 18 es 16,3 debemos hacer: $(16{,}3 - 12{,}2)*0{,}82 + 12{,}2 \cong 15{,}6$ éste será el valor buscado de P_{33}. Una alternativa más simple y muy utilizada es promediando los valores 16,3 y 12,2, que dará aproximadamente 14,3. Algunos autores en el caso de medidas discretas aconsejan redondear las cifras obtenidas por estos cálculos al entero más próximo(16, 17, 25).

De esta manera para calcular la mediana, solamente tenemos que observar que es el percentil 50, el primer cuartil el percentil 25 y el tercer cuartil el percentil 75. Si los datos no son numerosos, fácilmente se calcula la mediana como el valor central si

el número de observaciones es impar, si éste es par, la mediana será el promedio de las dos observaciones centrales. Esto siempre sobre los datos ordenados previamente en forma creciente.

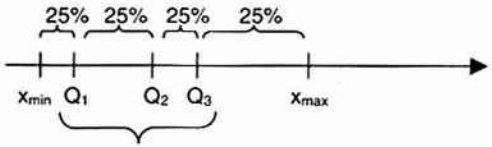
No todo es $\bar{x}$ y s

Cuando se trabaja con datos continuos, en algunas investigaciones o aplicaciones, como ya se vio, es necesario agruparlos en clases que contengan igual cantidad de observaciones para esto se crearon los cuantiles.

Es común considerar como medidas resúmenes del comportamiento de las variables a la media ($\bar{x}$) y el desvío estándar (s) como representantes de su centro y su dispersión respectivamente. Esto es óptimo cuando los datos experimentales siguen una distribución simétrica (Figura 3-a). Si los datos tienen una distribución asimétrica (Figura 3-b y c), la $\bar{x}$ no sigue siendo la medida óptima para resumir centralmente los datos por su sensibilidad a valores extremos. Dado que el desvío estándar se calcula en función de la media, esta s tampoco es representativa como medida de dispersión(1, 6, 7, 25). En este caso se utiliza la mediana como medida de tendencia central ya que representa con mayor realismo el dato «promedio». El rango intercuartílico es utilizado como una medida de dispersión cuando se trabaja con estos cuantiles(17) Contiene el 50% de las observaciones que se hallan en el medio y no se encuentra influenciado por valores extremos (Figura 1). La longitud del rectángulo o caja del box and whisker plot es precisamente ese rango (Figura 2). También se puede usar la mediana y dos percentiles, por ejemplo el 25 y el 75 (16). Pueden también ser construidos intervalos de confianza para cualquier cuantil (10, 17, 18, 21, 22).

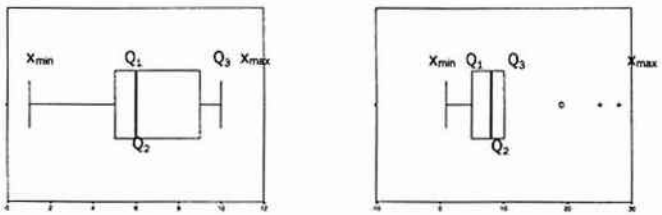
Podemos concluir que además de su importancia como medidas que agrupan entre sí igual cantidad de elementos, los cuantiles tienen un destacado papel como medidas resúmenes cuando los datos tienen una distribución no simétrica.

Figura 1: Puntos de Corte, entre el valor mínimo y el valor máximo, dado por los Cuartiles.



Rango intercuartílico: $RI = Q_3 - Q_1$

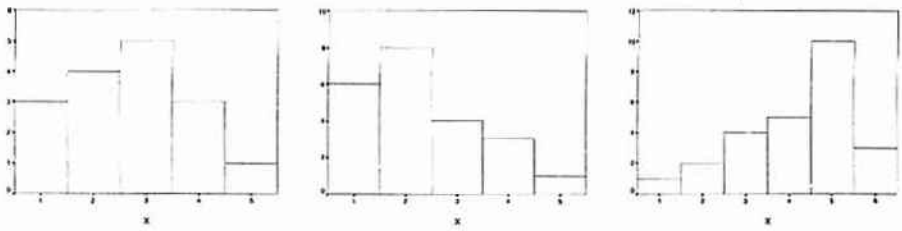
Figura 2: Diagrama de Caja (Box and Whisker Plot) y los cinco números resumen.



a) b)

x_{min}: mínimo, Q_1: primer cuartil, Q_2: segundo cuartil, Q_3:tercer cuartil, x_{max}: máximo. ° y *: valores extremos

Figura 3: Histogramas para una variable aleatoria continua



a) b) c)

Bibliografía

1- Siegel, A. (1988). *Statistics en data Analysis*. John Willey & Sons. New York. 1-137, 260-281.

2- Leiva, M.; Carrera, E.; Bottai, H.; Contini, L.; Vaira, S. (1999). Educación Estadística en la formación de grado y posgrado en las ciencias biomédicas. Actas IV Congreso Latinoamericano de Sociedades de Estadística. Mendoza, República Argentina.

3- Glantz, S. A.(1994). "Todo está en los cifras". JACC (J. Am. College Cardiology) - **3**, 6: 308-310.

4- International Committee of Medical Journal Editors. (1997). Uniform Requirements for Manuscripts Submitted to Biomedical Journals. New English Journal of Medicine. **336**, 4:309-316.

5- Reglamento de Publicaciones. Archivos Argentinos de Pediatría.

6- Dawson-Saunders, B; Trapp, R. G. (1999). *Bioestadística Médica*. Manual Moderno. México. 2ª ed., 3ª reimpresión. 1-7, 25-46, 223-246.

7- Altman, D. G. (1997). *Practical Statistics for Medical Research*. Chapman & Hall. London. 2da ed, 7ma reimpresión. 1-45, 107,170, 365-394.

8- Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* John Willey & Sons. New York.. 1- 8

9- Altman, D. (1994). The scandal of poor medical research. BMJ 308: 283 - 284.

10- Mood, A. (1966). *Introducción a la Teoría de la Estadística.* Editorial Aguilar. 6.

11- Abraira, V. (2001). El uso de la estadística en la investigación médica. Brotes **2**: 55-58

12- Henderson, R. (1993). Chemistry with Confidence: Should Clinical Chemistry Require Confidence Intervals for Analytical and Other Data? Clin. Chem. **39**, 6: 929-935.

13- Harris,E. (1993). On P value and Confidence Intervals (Why Can't We P with More Confidence? Clin. Chem. **39**, 6: 927-928.

14- Schervish, M.J. (1996). P values: What they are and What are not. The American Statistician. **50**, 3: 203-206.

15- Carrera, E; Contini, L; Vaira, S. (1999). Algo de Estadistica de todos los días: el valor P o P-value. FABICIB,1999. **3**: 163 – 166.

16- Altman, D. G.; Bland, J. M. (1994). Statistics Notes: Quartiles, Quintiles, centiles and other quantiles. BMJ **309**: 996-997

17- Casella, G. y Berger, R. (2002). *Statistical Inference.* Duxbury Advanced Series, 2nd Edition. Australia. 226-231.

18- Bartoszynski, R. Y Niewiadomska - Bugaj, M. (1996). *Probability and Statistical Inference.* John Wiley & Sons. 17-18, 344-400.

19- Tukey, J. (1977). *Exploratory Data Analysis.* Addison – Wesley. Massachusetts. 27-55.

20- Grimm, L. G. (1993), *Statistical applications for the behavioral Sciences.* John Willey & Sons. (New York). 24 - 20.

21- Bickel, P.J., Doksum, K.A. (1977). *Mathematical Statistics. Basic Ideas and Selected Topics.* Prentice-Hall, Inc. 17-18, 344-400.

22- Lindgren, B. W. (1976). Statistical Theory. Macmillan Publishing Co., Inc. New York. 3rd ed. 217-221.

23- Mood, A.M., Graybill, F.A., Boes, D.C. (1974). *Introduction to the Theory of Statistics.* McGraw-Hill Series in Probabilty and Statistics. 3rd ed.73-77, 508- 514.

24- Sociedad Argentina de Pediatría. Comité Nacional de Crecimiento y Desarrollo. (2001). Guías para la evaluación del crecimiento. 2da ed.

25- Mendenhall, W.; Sincich, T. (1997). Probabilidad y Estadistica para Ingenieria y Ciencias. Prentice Hall Iberoamericana, S. A. 4ta Edición. México. 39 - 62.