# Elección de los puntos de corte para convertir una variable cuantitativa en cualitativa

*Preparado por Luis M. Molinero (Alce Ingeniería)*
CorreoE: **bioestadistica@alceingenieria.net**

*Septiembre 2003*

**Artículo en formato PDF**

www.seh−lelha.org/stat1.htm

## Introducción

Existen diferentes motivos por los que variables que siendo originalmente de tipo cuantitativo, como el resultado de una prueba analítica, son convertidas en clasificaciones cualitativas de tipo ordinal o incluso en un resultado de tipo dicotómico (con sólo dos valores posibles). La razón principal suele ser el intento de simplificar la interpretación de la variable en cuestión, de tal manera que la clasificación en categorías facilite la toma de decisiones, por ejemplo a la hora de solicitar pruebas complementarias, o de considerar al paciente como candidato a una determinada terapia.

La conversión de una variable cuantitativa en cualitativa se denomina **categorización**.

La elección del número y valores de los puntos de corte puede efectuarse de acuerdo a criterios ya establecidos por trabajos anteriores, por razones teóricas basadas en la información clínica o fisiológica (esto es lo deseable), pero otras veces es el propio investigador quién tiene que decidir los puntos de corte que va a establecer. Cuando esto es así hay algunas cuestiones que conviene conocer y que vamos a comentar en este artículo.

## Métodos

Un procedimiento muy empleado para la elección de los puntos de corte se basa en escoger los valores de los cuartiles o de percentiles específicos de la distribución de los datos en nuestro estudio. Este método se suele utilizar para fijar intervalos de referencia de pruebas analíticas a partir de una muestra representativa de la población, eligiéndose dos percentiles centrados en torno a la mediana de la distribución, concretamente los valores 2.5 y 97.5, que definen por tanto un intervalo de referencia del 95 %. Pero hay otras posibilidades: otro punto de corte puede ser el considerar como valores elevados los que están por encima del tercer cuartil.

Si el objetivo de nuestro estudio es determinar un punto o puntos de corte para guiar la toma de decisiones, el cálculo de los percentiles de la distribución a partir de los valores de nuestra muestra da lugar a estimaciones sesgadas si el tamaño de muestra no es grande, y en general sus valores pueden variar en gran medida de una muestra a otra, por lo que es preferible calcularlos a partir de un modelo de distribución de probabilidad. Así si podemos suponer que los datos siguen una distribución normal, una vez calculada la media *m* y la desviación típica *s*, estimaremos por ejemplo el valor del tercer cuartil consultando en una tabla de probabilidad de la distribución normal el valor para el que $Pr(x \leq z)=0.75$, donde veremos que corresponde a 0.674, por lo que estimaremos dicho cuartil como m + 0.674 s.

Sin embargo lo más frecuente es que los resultados de las pruebas analíticas no se distribuyan según una normal, sino que tengan distribuciones asimétricas, concentradas en el lado izquierdo y con una cola larga al lado derecho, por lo que será preciso acudir a otro tipo de distribución o bien transformar los datos adecuadamente para que sea aproximadamente válida la utilización del distribución normal.

Una vez que se ha determinado que una variable cuantitativa se asocia de forma significativa con la variable resultado o con la presencia del suceso, lo primero que debemos hacer es examinar de forma gráfica dicha relación. Si el resultado es una variable cualitativa, por ejemplo dicotómica, la gráfica será en general poco

informativa salvo que haya una separación bastante definida.

En la figura 1 vemos un ejemplo de este tipo de gráfica que nos sugiere un punto de corte en torno al valor 35.

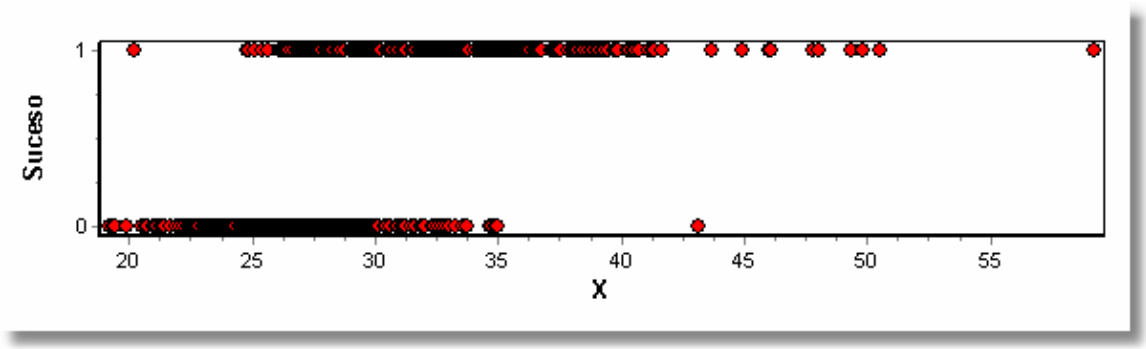


Figura 1

Sin embargo esta gráfica suele ser en general difícil de interpretar (por ejemplo como la de la figura 2) al distribuirse todos los puntos en dos valores del eje Y (suceso dicotómico).

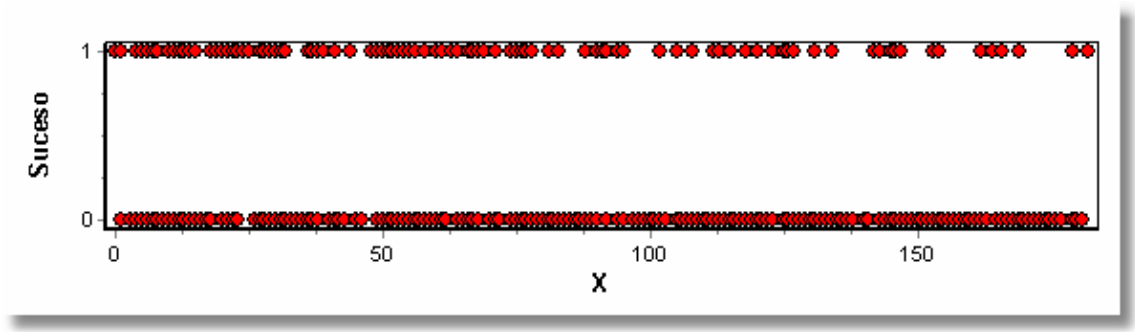


Figura 2

Una gráfica mucho más ilustrativa se obtiene dividiendo la variable X en intervalos iguales y calculando la proporción de sucesos para cada uno de esos intervalos, representando entonces dicha proporción frente al valor correspondiente al centro de cada intervalo, como en la figura 3.

El ejemplo concreto de la gráfica 3, nos sugiere un valor de corte para X en el entorno de 100.

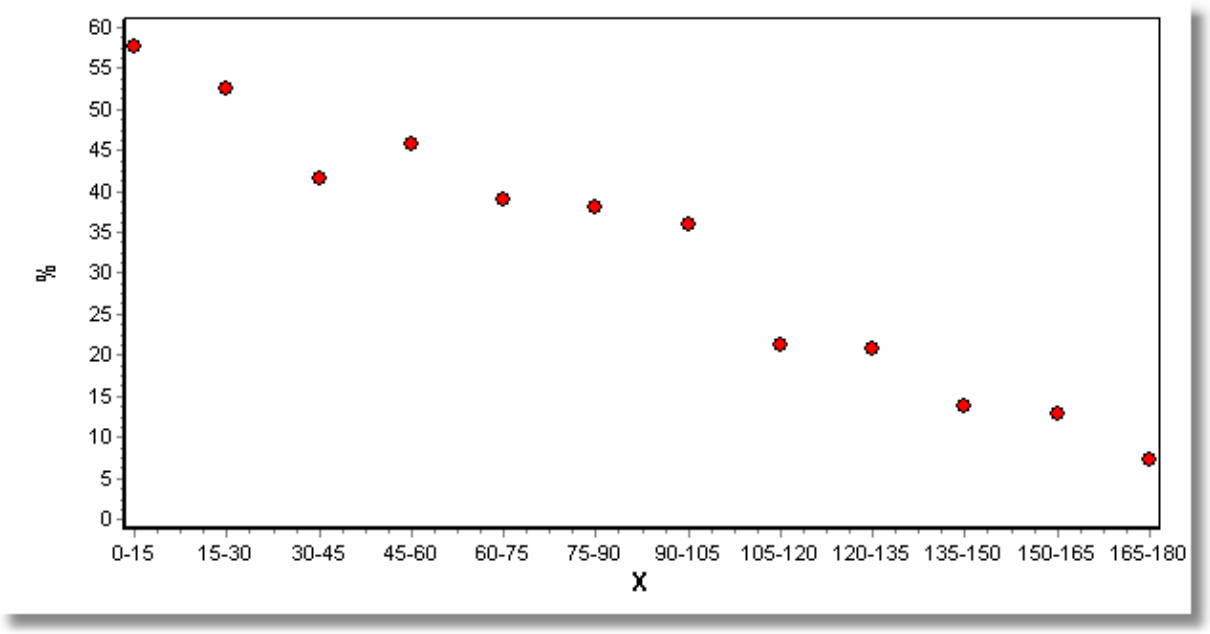


Figura 3

Por otro lado el inconveniente de este tipo de representación radica en que puede ser sensible a la amplitud del intervalo empleado, por lo que es una buena idea considerar diferentes amplitudes, y sobre todo verificar la frecuencia, número de datos, en cada intervalo, ya que si éste es muy pequeño la información será poco precisa.

Frente a la elección como punto de corte de un percentil, o a partir de la inspección visual de las gráficas, existe una alternativa sistemática que nos puede ayudar en la decisión. Consiste en determinar, para todos los valores de la variable *X* que se desea categorizar, el valor que mejor separa a los pacientes de acuerdo a la prueba de asociación del $chi^2$. De esta forma, si el resultado es dicotómico, para cada valor *C* observado en *X* se construye la siguiente tabla:

| | **X ≤ C** | **X > C** |
|---|---|---|
| **Suceso=NO** | $n_{11}$ | $n_{12}$ |
| **Suceso=SI** | $n_{21}$ | $n_{22}$ |

Se eligirá como punto de corte óptimo el valor de *C* para el que se obtiene el resultado de $chi^2$ más elevado, o lo que es lo mismo el que corresponde a un menor valor de probabilidad de la prueba. También se debe tener en cuenta los valores de riesgo relativo (o del odds ratio) en cada tabla.

Este procedimiento se denomina del ***"punto de corte óptimo"*** o también de *"valor mínimo de P"*.

En la figura 4 vemos el resultado de éste método aplicado a los datos que también se representaron en las figuras 2 y 3.

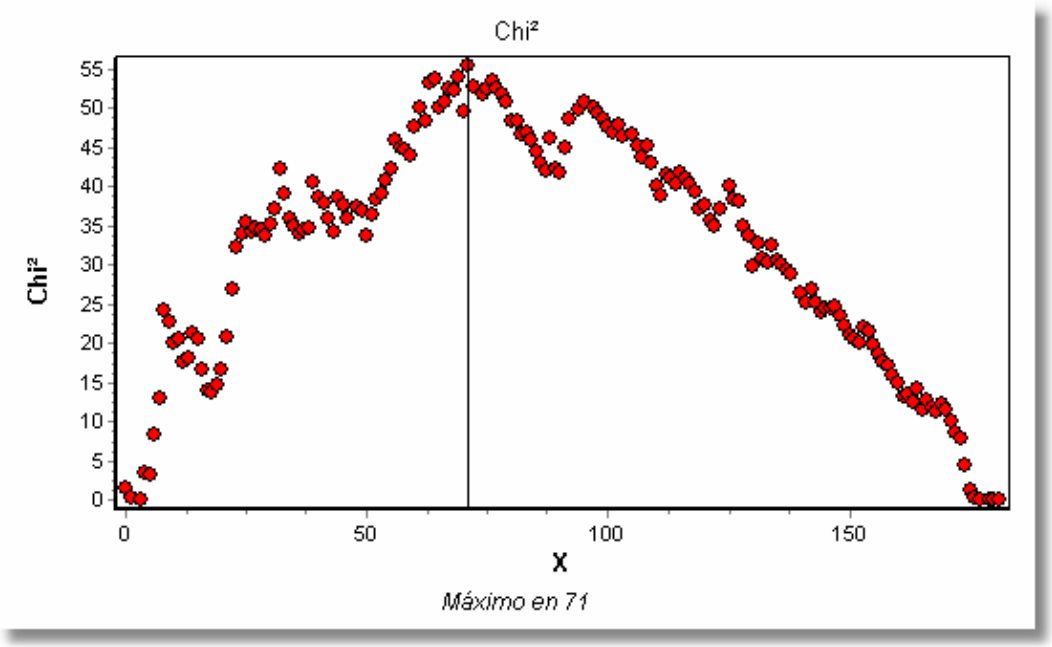


Figura 4

Para evaluar los posibles puntos de corte se recomienda no considerar los valores más extremos de la variable a ambos lados, excluyendo entonces entre el 5% o el 10% de ellos en cada extremo. Asimismo debido al aumento de la probabilidad de error de tipo I, que se produce al efectuar comparaciones múltiples, es también lógico utilizar alguna fórmula de ajuste para el valor de probabilidad mínimo obtenido, aunque un método como el de Bonferroni no es aquí adecuado, ya que las comparaciones no son independientes.

Altman et al proponen una fórmula de corrección muy sencilla para el caso de que se excluya el 5 % de los valores más extremos de X a ambos lados (percentiles 5 y 95), y otra para cuando se excluye el 10 % (percentiles 10 y 90). Son las siguientes

| | |
|---|---|
| Excluyendo el 5 % | $p = -3.13\ p_{min}(1 + 1.65\ \ln(p_{min}))$ |
| Excluyendo el 10 % | $p = -1.63\ p_{min}(1 + 2.35\ \ln(p_{min}))$ |

donde $p_{min}$ es el valor de probabilidad mínimo obtenido y p es el valor corregido.

Cuando en la aparición del suceso interviene también el tiempo y tenemos observaciones incompletas ("censored"), en las que se aplicarán modelos de análisis de supervivencia, se utilizará en lugar de una prueba de $chi^2$ la prueba logrank para comparación de curvas de supervivencia.

Cuando la variable que se convierte en cualitativa va a intervenir en un modelo de regresión múltiple multivariante, por ejemplo una regresión logística o un modelo de supervivencia de Cox, parece lógico que la búsqueda sistemática del punto de corte óptimo lo tenga en cuenta y se elija de tal forma que proporcione la separación óptima de acuerdo al modelo multivariante, en lugar de utilizar solo el criterio univariante, lo cual complica bastante la metodología.

## Inconvenientes de la conversión de una variable cuantitativa a cualitativa

Aunque resulta atractiva la utilización de un método sistemático para la elección de los puntos de corte como el anteriormente descrito, la categorización de una variable cuantitativa supone siempre una pérdida importante de información, y si además los puntos de corte se eligen en base a la información proporcionada por los propios datos del estudio puede dar lugar a que las conclusiones sean menos extrapolables a otras situaciones; por ello en los modelos de regresión siempre es preferible utilizar las variables cuantitativas como tales, y no convertidas a cualitativas ya que, además de no perder eficiencia, nos permite calcular la

modificación en el riesgo para cambios diferentes en la magnitud del factor pronóstico, lo que es aún más importante si los puntos de corte se van a elegir a partir de los mismos datos, ya que en este caso probablemente cambiarán de un estudio a otro. Si se construye un modelo de regresión con las mismas variables en otro estudio, los puntos de corte, elegidos según percentiles o según el criterio del mínimo valor de p, serán diferentes y dificultarán la comparación de los modelos.

Solo está justificada la utilización de puntos de corte cuando se quiere construir un modelo de decisión y se desea proporcionar una regla de actuación sencilla, y aun así en este caso el concepto de punto de corte óptimo puede venir influido por otros criterios diferentes de los aquí expuestos, como pueden ser la sensibilidad y especificidad del sistema de clasificación y los costes (entendidos en un sentido amplio) asociados a una decisión errónea en cualquiera de los sentidos.

Cuando se convierte una variable cuantitativa en cualitativa la utilización de sólo dos categorías debe estar bien justificada y de entrada no hay que rechazar la posibilidad de definir más de dos, aunque conlleve aumentar la complejidad en el modelo de regresión, dado que implica introducir variables internas (dummy) para modelar el efecto de la variable cualitativa.

## Referencias


- Mazumdar M, Glassman JR. **Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision−making about cancer treatments.** Statistics in Medicine 2000 Volume 19, Issue 1 , Pages 113 − 132. [Abstract]

- Buettner P, Garbe C, Guggenmoos−Holzmann I. **Problems in defining cutoff points of continuous prognostic factors: example of tumor thickness in primary cutaneous melanoma**. J Clin Epidemiol 1997; 50: 1201−1210 [Abstract]

- Altman DG, Lausen B, Sauerbrei W, Schumacher M. **Dangers of using "optimal" cutpoints in the evaluation of prognostic factors**. J Natl Cancer Inst 1994; 86: 829−835 [Medline]



Indice de artículos

Principio de la página