El histograma como un instrumento para la comprensión de las funciones de densidad de probabilidad

Behar Gutiérrez Roberto[1], Grima Cintas Pere[2]

[1]Universidad del Valle, Cali, Colombia
[2]UPC Barcelona, España

Resumen

Tradicionalmente los profesores de Estadística de nivel medio y superior hemos mirado la Estadística Descriptiva como una temática divorciada de la probabilidad y de la Inferencia. Cuando llega el momento de explicar el histograma, generalmente se construyen intervalos de igual tamaño y el eje de las ordenadas representa directamente la frecuencia relativa. Sin embargo, cuando trata la temática de las funciones de densidades en probabilidad, para calcular la probabilidad, que conceptualmente es el homólogo de la frecuencia relativa, si se mira como una extensión del concepto a la población entera, debe calcularse un área, ya no son las ordenadas las que proporcionan esta información.

La pregunta que surge es ¿Por qué si el concepto de probabilidad es una extensión de la frecuencia relativa a la población, en un caso se calcula un área y en el otro una altura? Esto parece conceptualmente incoherente. En el presente trabajo se plantea una estrategia para lograr coherencia, definiendo el histograma como un gráfico de la densidad empírica. Esto tiene una doble función, ganar potencial intuitivo para dar sentido real a la idea de densidad, logrando que la definición de variable aleatoria continua no suene artificial para los estudiantes y por otro lado resolver la mencionada incoherencia. En este trabajo se ilustra con un ejemplo la estrategia que se plantea.

Palabras clave: Histograma, función de densidad empírica, intervalos de clase. Función de densidad de probabilidad.

4.1. Introducción

En los cursos básicos de estadística, el capítulo que corresponde a Estadística Descriptiva, aparece como un tema aislado, que puede ir antes o después de la parte de probabilidad. En estas condiciones no se aprovechan algunos desarrollos de la Estadística Descriptiva que podrían ser usados como un puente intuitivo para la comprensión de resultados más abstractos de la teoría de la probabilidad. En este artículo se hará referencia específica al concepto de histograma, representación de la función empírica de densidad para dar sentido a la definición de variable aleatoria continua.

Una primera contradicción que podría enfrentar un estudiante, es que cuando aprendió su concepto de histograma, las ordenadas del gráfico representaban la frecuencia relativa, sin embargo en la extensión de la idea de histograma a la de densidad de probabilidad, se propone el cálculo del área bajo la curva para calcular la probabilidad y no las ordenadas. Esta fractura no tiene explicación alguna, convirtiéndose posiblemente en un obstáculo para el aprendizaje significativo de la función de densidad de probabilidad.

Si se quiere que la función de densidad de probabilidad sea una extensión de la idea de histograma, es conveniente que la definición de histograma se corresponda con el gráfico de función de densidad empírica. De esta manera se garantiza una continuidad en el concepto y se proporciona una base intuitiva para la comprensión de la definición de variable aleatoria continua, que es generalmente es matemática.

En J. M. Contreras, G. R. Cañadas, M. M. Gea y P. Arteaga (Eds.), Actas de las Jornadas Virtuales en Didáctica de la Estadística, Probabilidad y Combinatoria (pp. 229-235). Granada, Departamento de Didáctica de la Matemática de la Universidad de Granada, 2013.

Lee y Meletiou (2003) estudian algunos tipos de razonamientos erróneos al construir, interpretar y aplicar los histogramas en diferentes contextos de la vida real, sin embargo, no se refieren a la situación en la cual las áreas del histograma representan las frecuencias.

Wu (2004), típica algunos errores comunes relacionados con la interpretación y significado de algunos gráficos. Destaca la confusión entre gráficos parecidos pero de naturaleza distinta, en particular entre el histograma y gráfico de barras, pero no trata lo relativo al histograma como una representación de la función empírica de densidad, lo cual puede ser objeto de confusión, toda vez que esta no es observable de manera directa.

5. Definición de Histograma. (Función empírica de densidad)

Por comodidad, generalmente se toman los intervalos de clase del mismo ancho y se omite el concepto de densidad empírica, pues en caso de intervalos de igual ancho, la forma del histograma es idéntica, si se toma como ordenada la densidad o si se asume como la frecuencia relativa. El software de estadística, refuerza esta costumbre, pues por defecto hace gráficos de histograma con intervalos del mismo ancho.

Introduciendo el tema de la representación gráfica de los datos, usando intervalos de anchura desigual, se produce una ganancia conceptual importante, pues obliga a la representación del histograma como rectángulos que tienen como base el intervalo de clase y su área proporcional (o igual) a la frecuencia relativa.

Definiendo el histograma de esta manera sus ordenadas representan automáticamente la función empírica de densidad, generándose el enlace conceptual apropiado con la densidad de probabilidad de una variable aleatoria. Además la palabra "empírica" se asocia con muestral, y la densidad de probabilidad como su análogo poblacional. Ilustremos la situación con un ejemplo.

Ejemplo 1. En el sector de la industria metalmecánica, se toma una muestra al azar de 500 obreros y se determina la antigüedad en su trabajo. Por razones de índole administrativo, se quiere representar los datos por medio de un histograma que considere los siguientes intervalos de clase: 0-2 años, 2-3 años, 3-5 años, 5-10 años, 10-20 años. Después de contar el número de obreros que pertenecen a cada intervalo y expresarlo en porcentaje, se obtiene la Tabla 1. La frecuencia relativa se ha denotado por f_i

Tabla 1. Frecuencia relativa de la variable Antigüedad en el trabajo

i	Intervalo (Años de Antigüedad)	Frecuencia Relativa % (f_i)
1	(0-2]	10
2	(2-3]	5
3	(3-5]	40
4	(5-10]	40
5	(10-20]	5
Total		100

Ahora se procede a construir el histograma, como el gráfico de la función de densidad empírica. Note que en esta situación los intervalos son de diferente ancho (Ci). Se debe ahora construir un conjunto de rectángulos cuya base sea el intervalo de clase correspondiente y cuya área (Ai) represente la frecuencia relativa (fi) del intervalo respectivo. De esta manera, si un rectángulo asociado con un intervalo de clase tiene el doble de área que otro, es porque contiene el doble de datos. En nuestro ejemplo, si detallamos la frecuencia relativa en la Tabla 1, el área sobre el primer intervalo deberá ser el doble del área sobre el segundo. El área del rectángulo sobre el tercer intervalo deberá ser cuatro veces el área del primero. De esta manera la ordenada, es decir las alturas, digamos

f^*_i, del rectángulo construido sobre el i-ésimo intervalo, deberá ser tal que el área del rectángulo Ai coincida con su frecuencia fi, es decir que:

$$A_i = f_i = (base).(altura) = C_i.f_i^*$$

donde C_i es el ancho del intervalo. Así, despejando f^*_i, se obtiene la altura (ordenada eje vertical) que debe tener cada rectángulo: $f_i^* = {}^{f_i}\!/_{C_i}$.

Observe que se divide la frecuencia relativa entre el número de unidades que tenga el intervalo correspondiente, entonces las unidades de f^*_i son (% de datos por cada unidad de la variable en dicho intervalo). Veamos por ejemplo para el primer intervalo:

$$f_1 = 10\%\ C_1 = 2,$$

así que la altura del primer rectángulo es: $f_1^* = \frac{f_1}{C_1} = \frac{10\%}{2\ años} = 5\%/año$

Es intuitivamente claro, que si el primer intervalo tiene el 10% de los datos y estos datos están distribuidos en un intervalo que tiene una longitud de dos (2) unidades, pues en promedio hay 5% por cada unidad (f^*_i=5% /año=0.05/año).

El cuarto intervalo, (5; 10], por ejemplo, en sus 5 unidades (5 años) contiene 40% de los datos. Así que en promedio, hay 8% de los datos en cada unidad o lo que es lo mismo:

$$f_4^* = \frac{f_4}{C_4} = \frac{40\%}{5\ años} = 8\%/año \equiv 0,08/año$$

Es decir que las unidades del eje Y en el gráfico del histograma es %/unidad de intervalo, por eso se le conoce como densidad de frecuencia (f^*_i) y en este caso, para tomar en consideración que se calcula con base en los datos de una muestra, se le llama función empírica de densidad de frecuencia. En la siguiente tabla, se registra la densidad empírica de frecuencia para cada intervalo.

Tabla 2. Densidad empírica de frecuencia para la variable antigüedad

i	Intervalo (Años de Antigüedad)	Frecuencia Relativa f_i %	Densidad de Frecuencia f^*_i %/año
1	(0-2]	10	5
2	(2-3]	5	5
3	(3-5]	40	20
4	(5-10]	40	8
5	(10-20]	5	0,5
Total		100	

Si se realiza el gráfico de las densidades empíricas de frecuencias de la Tabla 2, se obtiene el histograma de la Figura 1.

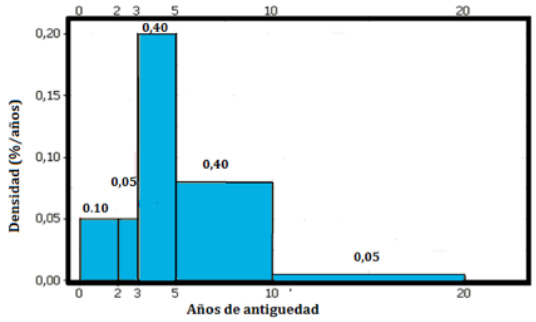

Figura 1. Histograma. Función empírica de densidad de frecuencia

Sobre cada rectángulo se ha colocado su área, es decir la frecuencia relativa. La ordenada correspondiente representa la densidad.De esta manera la estimación de un porcentaje relacionado con evento de la variable antigüedad, se convierte en el cálculo de un área, tal como ocurrirá más tarde, cuando se trate el tema de variables aleatorias continuas.

Así por ejemplo si se está interesado en estimar el porcentaje de obreros con antigüedad menor o igual a 4 años, digamos P(X≤4), bastará calcular el área del histograma comprendida entre cero (0) y cuatro (4), como se muestra en la Figura 2

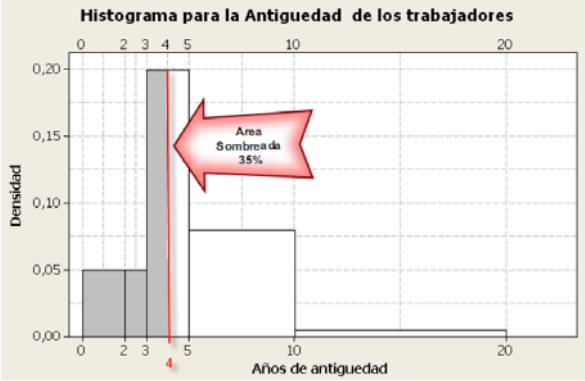

Figura 2. Área oscura del gráfico representa P(X≤4)

Observe que el área sombreada se calcula sumando por un lado las áreas de los primeros rectángulos (10%+5%) y por otro lado la parte del tercer rectángulo comprendida entre 3 y 4, como se conoce su densidad, que es 20% , y se requiere un año, Así que el porcentaje de trabajadores con antigüedad de 4 años o menos se estima en:

$$P(X\leq 4) = 10\% + 5\% + 20\% \,.(1 \text{ año}) = 35\%$$

Análogamente, si se desea estimar el porcentaje de obreros con antigüedad entre 4 y 7,5 años, es decir $P(4\leq X\leq 7{,}5)$. La respuesta será calcular el área del histograma entre dichos valores, como se muestra en la Figura 3.

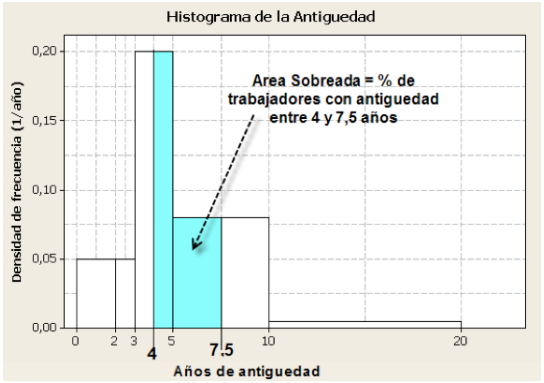

Figura 3. Representación de P(4≤X≤7.5). Área sombreada.

Haciendo el cálculo, usando el concepto de densidad, se obtiene:

$$P(4 \leq X \leq 7{,}5) = f_3^* * (5-4) + f_4^* * (7{,}5-5) = 20\%/\,año * (1año) + 8\%/\,año * (2{,}5años) = 40\%$$

Después de éste recorrido, abordemos la definición de variable aleatoria continua.

6. Variable aleatoria. Definición (Función de densidad de probabilidad)

Se dice que X es una variable aleatoria continua si existe una función f(x), llamada función densidad de probabilidad (fdp) de X, que satisface las siguientes condiciones:

a. $f(x) \geq 0 \quad \forall x \in \Re$;

Es razonable que no tome valores negativos, si se asocia con la función empírica de densidad de frecuencia.

b. $\int_{-\infty}^{+\infty} f(x).dx = 1$

Ya hemos dicho antes que el área del histograma y ahora el área bajo la función de densidad, debe ser 100%.

c. Para cualquier a, b se tiene que $P(a \leq X \leq b) = \int_{a}^{b} f(x).dx$

El área atrapada entre los valores a y b es justamente el porcentaje de datos de la población que cumple con esas especificaciones, análogamente a lo observado en el histograma. Mirado como la experiencia aleatoria de sacar al azar un valor de X, esta área puede interpretarse como probabilidad.

Ejemplo 2. El histograma de una cierta característica continua X, es el que muestra sombreado en la Figura 4. Se pretende ajustar una función empírica densidad continua y suena razonable la que aparece formando un triángulo equilátero. Encuentre la definición de dicha función de densidad de probabilidad estimada, f(x).

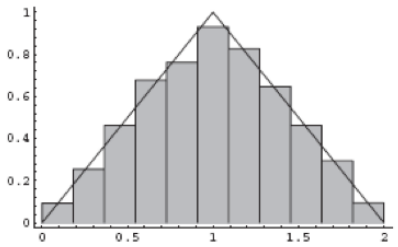

Figura 4. El gráfico sombreado es un histograma y las líneas una aproximación a una densidad empírica continua.

En primer lugar se observa que el rango de valores que puede tomar la variable aleatoria X son los puntos en el intervalo que va de cero (0) a dos (2). Es decir que:

$$\Omega_X = \{x \in \Re / 0 \prec x \leq 2\}$$

El rango o recorrido de la variable aleatoria X. algunas veces se denota por $\Re_X$

¿Cual deberá ser la ecuación que defina las dos rectas que conforman el triángulo equilátero y que definen la función de densidad de probabilidad estimada?Pues como el área debe ser igual a la unidad, esto significa que la altura h del triángulo, debe ser tal que el área valga 1.

$$Area = 1 = \frac{base * altura}{2} = \frac{2 * h}{2} = 1$$

De donde se deduce que la altura h=1. Por lo tanto la ecuación de la recta de pendiente positiva es f(x)=x. la ecuación de la recta con pendiente negativa será: f(x)=2 –x, así pues:

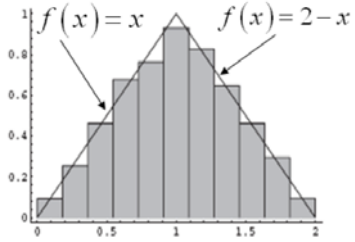

Figura 5. Función de densidad empírica ajustada

$$f(x)=\begin{cases} x & 0<x\leq 1 \\ 2-x & 1<x\leq 2 \end{cases}$$

Si se produce una realización de la variable aleatoria X, estime el porcentaje de veces en el que dicho valor resulta entre 0,5 y 1,5?

$$P(0,5\leq X\leq 1,5)=\int_{0,5}^{1,5} f(x).dx \; P(0,5\leq X\leq 1,5)=\int_{0,5}^{1,0} x.dx+\int_{1,0}^{1,5}(2-x).dx=$$

$$P(0,5\leq X\leq 1,5)=\int_{0,5}^{1,0} x.dx+\int_{1,0}^{1,5}(2-x).dx = P(0,5\leq X\leq 1,5)=\frac{x^2}{2}\Big|_{0,5}^{1,0}+\left(2x-\frac{x^2}{2}\right)\Big|_{1,0}^{1,5}=$$

$$P(0,5\leq X\leq 1,5)=\frac{3}{4}$$

Observe que el área, en este caso, se hubiera podido calcular como el área de dos trapecios, con base mayor la altura del triángulo.

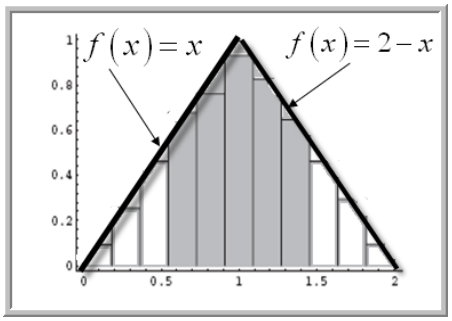

Figura 6. Representación de P(4≤X≤7.5).

7. Conclusión

La definición de variable aleatoria continua, es muy poco intuitiva e introduce la función de densidad de probabilidad de manera muy artificial. Desarrollar la idea de función

empírica de densidad, al momento de tratar la representación gráfica de variables de tipo continuo, a través de una definición apropiada de histograma, para una situación de intervalos de clase desiguales, en la cual las áreas y no las alturas representen la frecuencia relativa, hace que la definición y los procesos operativos con variables aleatorias sean más naturales y con una buena componente intuitiva.

Referencias

Lee y Meletiou (2003). Some difficulties of learning histograms in introductory statistics. Trabajo presentado en el *Joint Statistical Meetings Section on Statistical Education.* Online: http://www.statlit.org/PDF/2003LeeASA.pdfNadaraya, E.A. (1964).

Wu, Y. (2004). Singapore secondary school students´understanding of statisticalgraphs. Trabajo presentado en el *10th International Congress on Mathematics Education.*