Distribuciones de probabilidad
LECCIÓN 3

En los ejemplos de la lección 1 como el caso de los dados y de la moneda, se tenía en cuenta que esas estuviesen balanceadas (justas) es decir, se consideraba para facilidad del cálculo que son elementos ideales sin imperfecciones físicas. Esto deriva en que la distribución de probabilidad para tales objetos es uniforme. Es decir, cada evento puede suceder un mismo número de veces. A este comportamiento se le llama distribución de probabilidad. Existen infinitas distribuciones de probabilidad. Sin embargo, en esta lección se dará estudio a las más importantes

Variables aleatorias y su distribución de probabilidad.

Para iniciar, la mejor forma de explicar la distribución de un a variable es emplear un registro numérico, por ejemplo, consideremos el evento de lanzar un dado y luego un segundo dado. En el primer experimento consideramos que la variable aleatoria X que registra el número de puntitos observados al lanzar del dado. X tomará uno y solo uno de los valores enteros entre el número 1 y el 6. Estos valores son igualmente probables, con lo que decimos que la probabilidad de observar algún valor i es:

$$p_i = P(X = i) = \frac{1}{6}, \; i = 1, 2, \ldots, 6$$

Hay dos funciones en estadística que se usan para describir la distribución de probabilidades. La función de distribución acumulativa (CDF) y la función de probabilidad de masa (PMF).

En el segundo evento (lanzar una segunda vez un dado) la variable aleatoria es X, pero esta vez registra la suma de los puntitos observados en los dos dados.
Se puede notar fácilmente que X solamente está definida para los números enteros desde el 2 hasta el 12. Pero estos valores no están igualmente distribuidos. Es decir, para el primer dado, cada combinación tiene 1/6 de probabilidad de verse, pero para lanzar un segundo dado cada combinación tiene 1/36 probabilidades de ocurrir.

Para evaluar mejor la situación, anotemos como un par ordenado (x1, x2) las entradas referentes al primer y segundo dado. Y anotemos en una tabla las posibles combinaciones a obtener:

	x2=1	x2=2	x2=3	x2=4	x2=5	x2=6
x1=1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
x1=2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
x1=3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
x1=4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
x1=5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
x1=6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

TIC

Generalmente, en el lanzamiento de dos dados, interesa saber el resultado como la suma de los dos números obtenidos, con lo cual se reduce la cantidad de números posibles a 12, como se señaló anteriormente. Ya que la probabilidad de todos los eventos suma uno, tenemos que:

$$\sum_{x=2}^{12} p(x) = 1$$

Para validar que la distribución no es uniforme, se puede evaluar en una tabla los posibles resultados, en donde F(x) es la función de distribución acumulativa (CDF), es decir, va sumando los valores y P(f) es la función de probabilidad de masa, es decir, cuanto suman los puntitos de los dos dados lanzados.

X	2	3	4	5	6	7	8	9	10	11	12
F(x)	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$
P(x)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Función de densidad de probabilidad

Para el caso de las variables continuas, la probabilidad de que determinado número salga se define como el número dividido entre infinito. Lo que hace que la probabilidad de obtener el número es cero. Por tal motivo no tiene sentido analizar las funciones de probabilidad como en el caso del ejercicio discreto. Por eso se empelan funciones de densidad de probabilidad. Esta función se escribe como la suma de la probabilidad de todos los elementos posibles. Estas sumas sobre elementos continuos se suelen escribir como integrales que describen la sumatoria de todos los posibles elementos en un rango dado por el límite inferior hasta el límite superior de la integral.

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

Lo correcto en este caso es determinar la probabilidad de que una variable aleatoria esté en un rango determinado [a, b]. Esta probabilidad está dada por la ecuación:

$$\int_{a}^{b} f(x)dx = P[a \leq X \leq b]$$

Por ejemplo, se quiere calcular la probabilidad de que una persona al azar escogida en la ciudad sea un milenial (nacido entre 1980 y 2000 según la definición de Wikipedia) En este caso la respuesta se encuentra acotada por un rango [1980, 2000] con lo cual es posible determinar la función de densidad de probabilidad que conteste a esa pregunta. Veremos a continuación algunas distribuciones de probabilidad en diferentes intervalos:

Distribución normal o gaussiana

La distribución Gaussiana, también conocida como distribución normal, es un concepto estadístico que describe cómo se distribuyen los datos alrededor de un valor promedio. Su forma es similar a una «campana» simétrica que muestra cómo los valores se agrupan alrededor de un punto central. En esta distribución, la mayoría de los datos se concentran cerca del valor medio, y a medida que nos alejamos del valor medio, la probabilidad de que un dato pertenezca a un rango disminuye gradualmente. Su importancia radica en que gran cantidad de fenómenos naturales y artificiales se ajustan a este patrón. Por ejemplo, la altura de las personas, los puntajes en pruebas estandarizadas, el error en cualquier medición y muchas variables asociadas con fenómenos naturales. Por esto es muy conocida y se suelen explicar muchos conceptos estadísticos basados en las características de una distribución gaussiana. Esta distribución se puede modelar con dos parámetros que son la media y la desviación estándar. De la distribución gaussiana se deriva la regla empírica, que significa que aproximadamente el 68% de los datos se encuentran contenidos dentro de una desviación estándar del valor medio y el 95% de los datos se encuentra en dos desviaciones estándar. También implica que en tres desviaciones estándar alrededor de la media se encuentra el 99.7% de los datos. La función que modela la forma de una distribución gaussiana se modela con la ecuación:

$$f(x) = a \times e^{-(x-b)^2/2c^2}$$

$$a = \frac{1}{c\sqrt{2\pi}}$$

En esta ecuación a, b y c son constantes reales mayores que –1. El valor a simboliza el punto más alto de la campana, b es la posición del centro de la campana y c es la deviación estándar que modela el ancho de la campana. Usualmente al valor de a se le asigna:

También se suele encontrar en la literatura que los valores de b se le llama también con la letra griega μ (valor de la media) y a c se le asigna la letra griega (valor de la desviación estándar). En la figura 1 se muestran algunas curvas para diferentes valores de media y de desviación estándar.

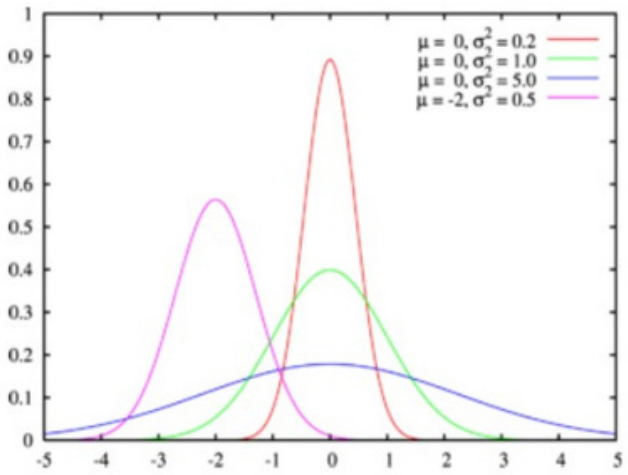

Figura 1: Curvas gaussianas

Distribución rectangular

Es una familia de distribuciones de probabilidad en donde todos los valores de un rango determinado tienen una probabilidad uniforme de aparición. El dominio de estas funciones está definido por los parámetros a y b que son su valor mínimo y su máximo respectivamente. La función de densidad está definida por:

$$f(x) = \frac{1}{b-a}; para\ x \in [a,\ b]$$

Es importante resaltar que solo está definida para el intervalo entre los valores a y b (figura 2).

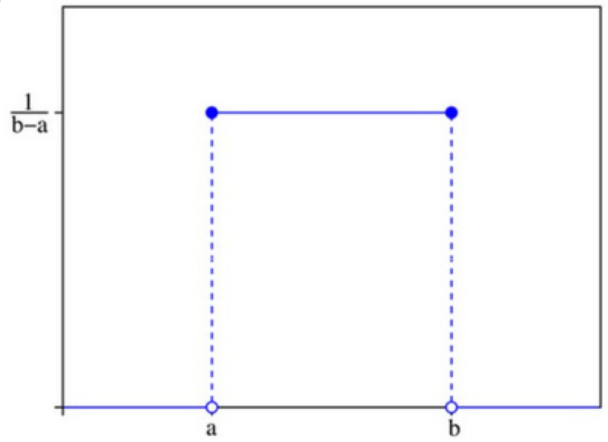

Figura 2: representación de la función uniforme de probabilidad.

Para este tipo de distribuciones la esperanza (valor medio de la variable aleatoria) está dado por:

$$E[x] = \frac{a+b}{2}$$

Y la varianza está dada por:

$$\frac{(b-a)^2}{12}$$

Distribución Pearson (X^2)

La distribución de Pearson o (2) chi cuadrado es la suma del cuadrado de k variables aleatorias independientes con distribución normal estándar. Esta función se emplea en inferencia estadística para las pruebas de hipótesis y la construcción de los intervalos de confianza. En la figura 3 se muestra la densidad de probabilidad de esta función. La media está dada por el valor k y la varianza es 2k

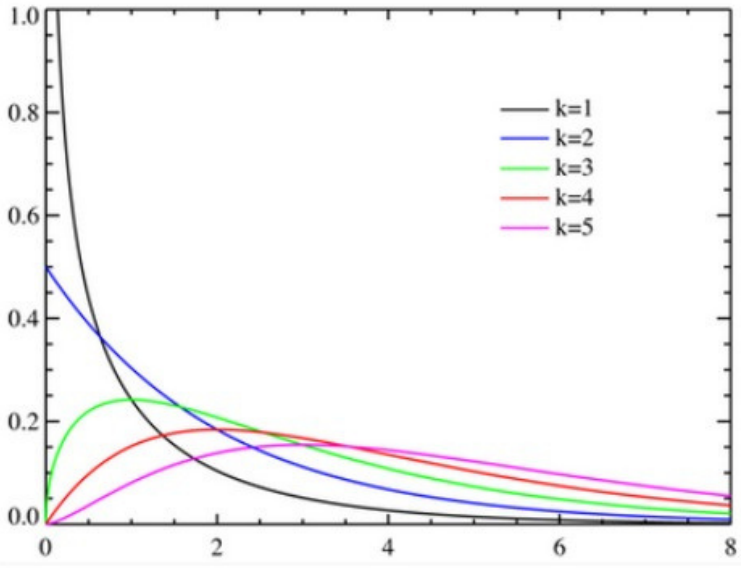

Figura 3: Densidad de probabilidad χ^2

Distribución T Student

Representa la distribución de probabilidad cuando se quiere estimar la media de una población normalmente distribuida cuando el tamaño de la muestra es pequeño y la desviación estándar poblacional es desconocida. Su media es cero, y su varianza es

$$\frac{v}{v-2}$$

Para v > 2 y es indefinida para otros valores.

En la figura 4 se muestra la curva de la densidad de probabilidad

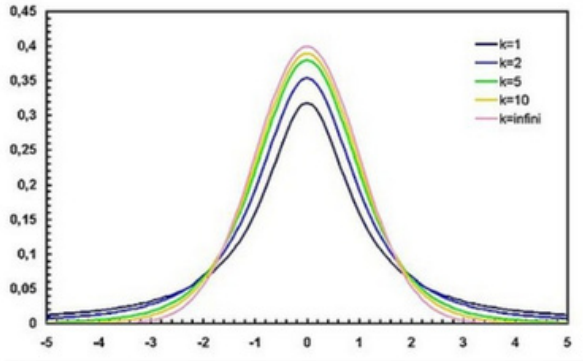

Figura 4: densidad de probabilidad T student.