Documentos educativos

Basic elements of probability IV: Random variables and probability distributions.

Elementos básicos de probabilidad IV: Variables aleatorias y distribuciones de probabilidad.

Juan Luis Soto Espinosa[1] : https://orcid.org/0000-0003-2600-9292
[1] *Facultad de Estudios Superiores Zaragoza*
Correo electrónico de contacto: soej@unam.mx
Palabras clave: Estadística, probabilidad, distribuciones, variables aleatorias.

Fecha de envío: 17-11-23
Fecha de aprobación:06-12-23

Como se vio en una entrega anterior de esta serie, existen eventos en los que pueden obtenerse diferentes resultados, cada uno de ellos con una determinada probabilidad de ocurrencia.

En un evento que puede tener varios resultados posibles, se dice que existe variabilidad en el resultado y cierto grado de incertidumbre en el evento, pues es imposible conocer con anticipación el resultado del evento.

Para estudiar este tipo de eventos desde un punto de vista estadístico, se recurre al cálculo de probabilidades, a través del cual es posible determinar la probabilidad de ocurrencia del resultado esperado.

Recordemos que el espacio muestral es el conjunto de todos los resultados posibles de un experimento. A este espacio se le representa con la letra griega Omega (W) y para enumerar los elementos que constituyen el espacio muestral se utilizan llaves {} y cada elemento del conjunto se incluye separado por comas de los demás. Si los resultados son eventos individuales que pueden ser contados, enumerados o listados, tendremos una variable aleatoria discreta.

Un ejemplo de espacio muestral lo constituyen los posibles resultados de tirar una moneda al aire y ver que cara queda expuesta al caer al piso. Este espacio tiene dos posibles resultados: Sol (cara en otros países) y águila (o cruz en otros países). Lo anterior se representa como:

X = Tirar una moneda y ver que cae

x = {águila, sol}

Si el conjunto de resultados tiene un número finito y enumerable de elementos, se conoce como espacio muestral discreto finito, mientras que, si presenta un número finito de resultados, pero resulta imposible enumerarlos se conoce como espacio muestral discreto infinito

Cuando se estudian estos eventos se utiliza lo que en estadística se conoce con una Variable Aleatoria.

Una variable aleatoria (identificada con la abreviatura v. a.), se define como una función matemática que permite asignar un valor numérico a la probabilidad de ocurrencia de un resultado posible dentro del espacio muestral de un experimento.

Los valores posibles de una variable aleatoria representan los posibles resultados de un experimento que aún no se realiza. También puede representar los posibles valores de una cantidad cuyo valor resultante es incierto. Una variable aleatoria puede tomar diferentes valores, cada uno de los cuales puede tener una frecuencia de ocurrencia particular y distinta a los demás.

De la afirmación anterior, tenemos que existen resultados que se presentan con mayor frecuencia que otros, por lo que la predicción de un resultado posible tiene una fuerte componente de azar.

Las variables aleatorias suelen adquirir valores reales, pero pueden adquirir valores aleatorios lógicos, funciones o cualquier tipo de elementos dentro de un espacio medible.

El conjunto Ω es una familia A de elementos que cumplen la condición:

A es subconjunto de P(Ω) ≠ ∅

Dónde:

Wv A (contiene al total)

$$\Omega \in A \implies A^c = \frac{\Omega}{A} \epsilon A$$

Al par (W, A) se le conoce como espacio medible o espacio de probabilidades, en función del contexto. A los elementos de la familia A se les conoce como conjuntos A-medibles (conjuntos medibles). En el contexto probabilístico, se les suele llamar sucesos (ver “Elementos básicos de probabilidad I: Experimentos, sucesos, puntos y espacio muestrales” de esta misma serie).

Desde un punto de vista matemático, una variable aleatoria (representada con una letra mayúscula) X es una función real definida dentro del espacio de probabilidad (Ω) asociado a un experimento aleatorio.

X: Ω→ ℝ

La variable aleatoria es la descripción numérica del resultado Anticipado de un experimento estadístico.

De acuerdo con los valores que puede adquirir, una variable aleatoria puede ser:

Discretas: Si los valores que la variable puede adquirir son números enteros. (edad, Número de hijos, Número de accidentes).

Continuas: Si la variable puede adquirir cualquier valor de un número infinito de valores dentro de un intervalo. (estatura, peso, distancia, presión)

Variable aleatoria discreta

Consideremos una variable aleatoria discreta, con las siguientes características:

Sea el experimento aleatorio “Tirar un dado”, cuyo espacio muestral (S) está dado por:

S = {1, 2, 3, 4, 5, 6}

Donde cada posible resultado del experimento tiene una probabilidad de ocurrencia asociada. En este ejemplo, la probabilidad de obtener cualquier número del 1 al 6, es la misma para todos los casos, considerando que se trata de un dado normal. La probabilidad de obtener un número considerado dentro del espacio muestral es 1/6, ya que tenemos seis resultados posibles en el experimento, cada uno de los cuales se ubica en una de las seis caras del dado.

La variable aleatoria es una función que asigna un número real a cada resultado en el espacio muestral de un experimento aleatorio. Considerando el espacio muestral S, tendremos que la variable aleatoria X sería:

X = Resultado de tirar una vez un dado.

x = {1, 2, 3, 4, 5, 6}

Nótese que la variable aleatoria “Resultado de tirar una vez un dado” se identifica con una letra mayúscula, mientras que el espacio muestral asociado a la misma se representa con una letra minúscula. De esta forma, las variables aleatorias siempre se denotan con letras mayúsculas mientras que el espacio muestral asociado se denota con la misma letra, pero minúscula.

Es posible considerar variables aleatorias cuyo resultado no sea numérico, por ejemplo, sea la variable aleatoria “COVID-19 en trabajadores de una empresa” y denotémosla con la letra Y (recordemos que las variables aleatorias siempre se representan con mayúsculas).

Los posibles resultados de esta variable serían dos:

Trabajador con COVID-19 en una empresa.

Trabajador sin COVID-19 en una empresa.

En este caso, los resultados no son numéricos, sin embargo, es posible diferenciar uno de otro asignándoles un valor numérico:

0 = Trabajador sin COVID-19 en una empresa.

1 = Trabajador con COVID-19 en una empresa

De esta forma, el espacio muestral asociado a la variable:

Y = COVID-19 en trabajadores de una empresa

Sería:

y = {0,1}

En un tercer ejemplo, definamos una tercera variable aleatoria para un conjunto de participantes, donde Z estaría dada por “Alcaldía de procedencia”. Tenemos:

Z = Alcaldía de procedencia.

El espacio muestral estaría dado por:

z = {Coyoacán, Iztacalco, Iztapalapa, Tlalpan, Tláhuac, Xochimilco}

Si asignamos valores numéricos a cada una de las alcaldías, el espacio muestral puede representarse como:

z = {1,2,3,4,5,6}

Donde:

1 = Coyoacán

2 = Iztacalco

3 = Iztapalapa

4 = Tlalpan

5 = Tláhuac.

6 = Xochimilco

Si el conjunto de participantes incluye el mismo número de participantes de cada alcaldía, la probabilidad de ocurrencia sería la misma; pero si el número de participantes de cada alcaldía es diferente, la probabilidad de ocurrencia de cada valor es diferente.

Supongamos que hay 10 participantes de cada alcaldía. Para mostrar la forma en que la probabilidad de ocurrencia se distribuye en el espacio muestral, se construye una tabla como la siguiente (Tabla 1):

Tabla 1. Distribución de casos

Alcaldía	Participantes
Coyoacán	10
Iztacalco	10
Iztapalapa	10
Tlalpan	10
Tláhuac.	10
Xochimilco	10
Total	60

La probabilidad de que un participante tomado al azar provenga de alguna de las alcaldías consideradas en el espacio muestral estaría dada por la fórmula para el cálculo de la probabilidad clásica:

$$P(x) = \frac{Casos\ favorables\ de\ x}{Casos\ totales}$$

De forma que el cálculo de la probabilidad de ocurrencia de, por ejemplo, Coyoacán, estaría dada por:

$$P(Coyoacán) = \frac{Participantes\ de\ Coyoacán}{Participantes\ totales}$$

$$P(Coyoacán) = \frac{10}{60} = \frac{1}{6} = 0.1666$$

Repitiendo la operación para cada uno de los elementos del espacio muestral, tenemos la siguiente tabla 2:

Tabla 2.		
Alcaldía	Participantes	Probabilidad
Coyoacán	10	0.1666
Iztacalco	10	0.1666
Iztapalapa	10	0.1666
Tlalpan	10	0.1666
Tláhuac.	10	0.1666
Xochimilco	10	0.1666
Total	60	1.0000

La tabla 2 presenta tres columnas, la primera muestra cada uno de los valores en el espacio muestral de la variable aleatoria definida. La segunda columna muestra el número de elementos que adquieren el valor definido en las líneas (eventos favorables), Esta columna es lo que conocemos como Distribución De Frecuencia de la variable aleatoria.

La tercera columna muestra la probabilidad de ocurrencia de cada valor, el cual se obtiene dividiendo la frecuencia del valor entre el total de resultados posibles; en este ejemplo la probabilidad es la misma dado que para cada valor posible de la variable se tiene el mismo número de eventos. Esta columna corresponde a la Distribución De Probabilidad de la variable aleatoria. Esta columna tiene las siguientes características.

Cada valor tiene una probabilidad asociada

Las probabilidades de ocurrencia de cada valor son mutuamente excluyentes, es decir, no es posible que un elemento adquiera dos valoras del espacio muestral.

La probabilidad asociada a cada elemento del espacio muestral está entre CERO y UNO.

La suma total de las probabilidades de los elementos del espacio muestral es igual a uno.

Supongamos ahora que tenemos la variable aleatoria “Tirar una vez dos dados de seis caras para ver el resultado”, la definición y el espacio muestral de esta variable estaría dado por:

X = Tirar una vez dos dados de seis caras para ver el resultado.

El espacio muestral estaría dado por:

$x = \{2,3,4,5,6,7,8,9,10,11,12\}$

Que son los posibles resultados que se pueden tener en el experimento, nótese que el número 1 no se incluye en el espacio muestral pues no hay forma de que al tirar dos dados tradicionales de seis caras se obtenga ese resultado, ya que el valor mínimo obtenible es dos. En otras palabras, la probabilidad de ocurrencia del valor uno en este experimento es igual a CERO, con lo que podemos afirmar que este resultado constituye un EVENTO IMPOSIBLE en este experimento.

Si se consideran todas las combinaciones que es posible obtener al tirar dos dados de seis caras tenemos lo siguiente (Tabla 3):

Tabla 3. Combinaciones						
	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

En la fila superior se presentan los valores que es posible obtener en el primer dado, mientras que en la columna de la izquierda se presentan los valores que es posible obtener en el segundo dado. En el resto de las celdas se presentan los valores que se obtienen sumando el número de puntos en el primer dado (columna) más el número de puntos que se obtienen en el segundo (fila). Esta tabla 3 muestra que tendremos 36 combinaciones posibles.

Construyendo la tabla de frecuencia tenemos:

Tabla 4	
Valor	Frecuencia
1	0
2	1
3	2
4	3
5	4
6	5
7	6
8	5
9	4
10	3
11	2
12	1

Procedemos a calcular la probabilidad de ocurrencia de un valor a través de la fórmula:

$$P(x) = \frac{Frecuencia\ de\ x}{Total\ de\ eventos}$$

O bien

$$P(x) = \frac{Eventos\ favorables}{eventos\ totales}$$

Para el caso del valor 2, se procede a realizar el cálculo de las probabilidades de cada valor:

$$P(2) = \frac{1}{36} = 0.0277$$

$$P(3) = \frac{2}{36} = 0.0555$$

$$P(4) = \frac{3}{36} = 0.0833$$

$$P(5) = \frac{4}{36} = 0.1111$$

$$P(6) = \frac{5}{36} = 0.1388$$

$$P(7) = \frac{6}{36} = 0.1666$$

$$P(8) = \frac{5}{36} = 0.1388$$

$$P(9) = \frac{4}{36} = 0.1111$$

$$P(10) = \frac{3}{36} = 0.0833$$

$$P(11) = \frac{2}{36} = 0.0555$$

$$P(12) = \frac{1}{36} = 0.0277$$

Vaciando la información en la tabla de distribución de frecuencia tenemos:

Tabla 5. Distribución de probabilidades		
Valor	Frecuencia	Probabilidad
1	0	0
2	1	0.0277
3	2	0.0555
4	3	0.0833
5	4	0.1111
6	5	0.1388
7	6	0.1666

Tanto en la Figura 1 como en la tabla 5 puede apreciarse que existen resultados que son más frecuentes que otros,

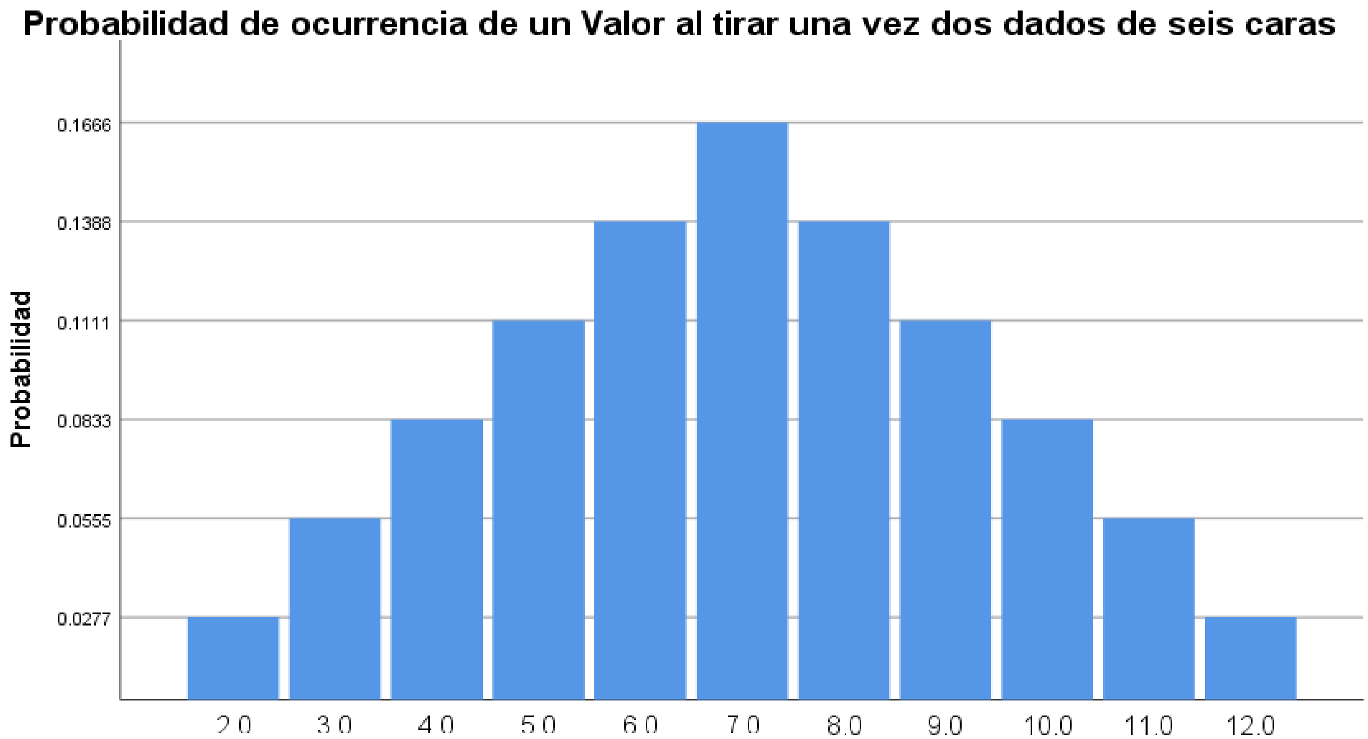


Figura 1. Probabilidad de ocurrencia

Tabla 5. Distribución de probabilidades		
Valor	Frecuencia	Probabilidad
8	5	0.1388
9	4	0.1111
10	3	0.0833
11	2	0.0555
12	1	0.0277

Si se desea saber cuál es la probabilidad de ocurrencia del valor 5 en el experimento cuya variable aleatoria es "Tirar una vez dos dados de seis caras, basta leer el valor de la columna Probabilidad; en este ejemplo, la probabilidad de ocurrencia del valor 5 en él.

Lo anterior puede representarse a través de un histograma de frecuencia (Figura 1).:

y que por lo tanto presentan una mayor probabilidad de ocurrencia.

Después de un número elevado de repeticiones, los eventos empiezan a presentar patrones de comportamiento definidos por la probabilidad de ocurrencia de los valores que considera su espacio muestral, estos patrones pueden representarse de manera gráfica y ser expresados por una función matemática que define su comportamiento. A esta gráfica y expresión matemática se le conoce como

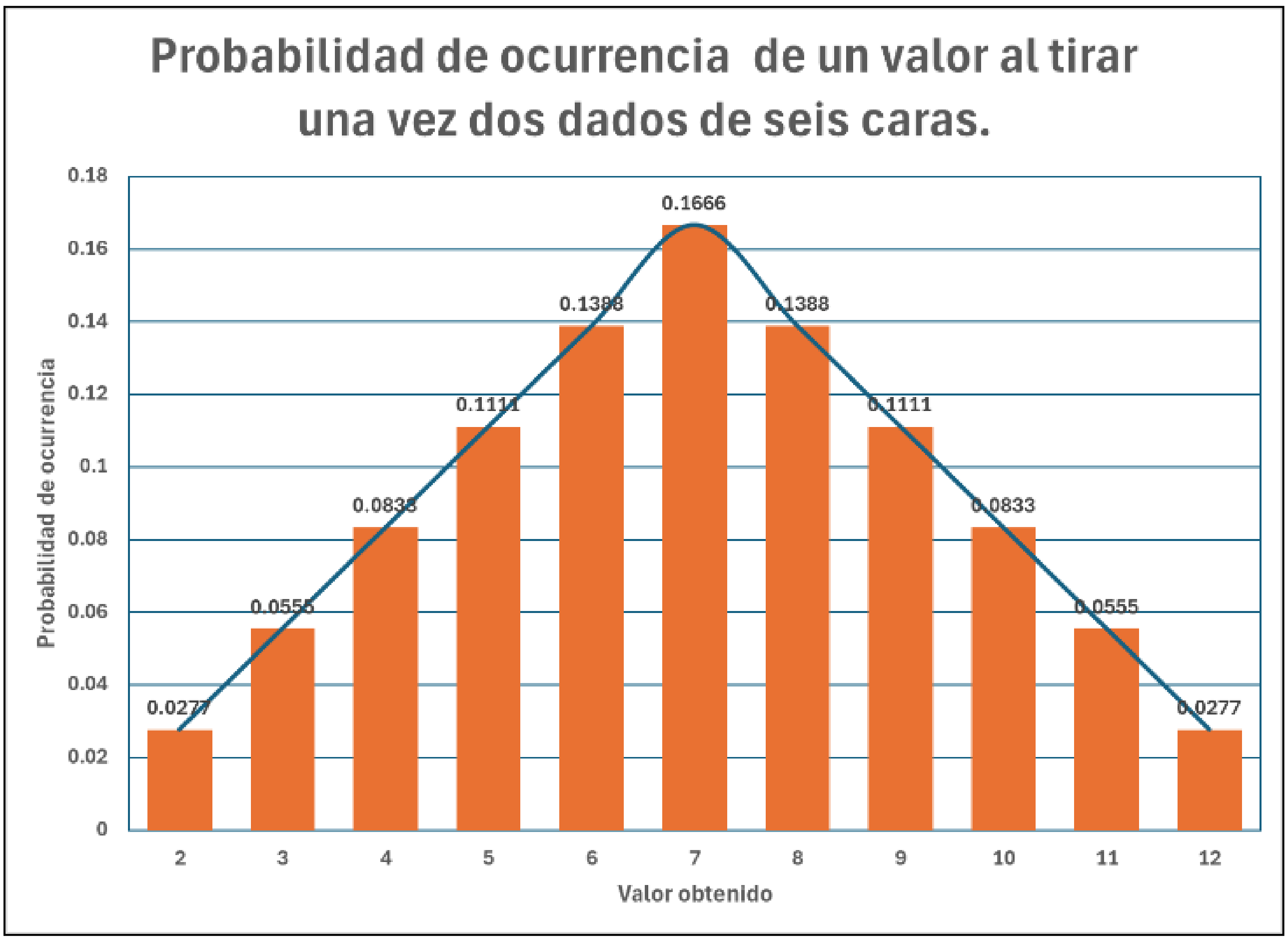
Función De Una Distribución.

La distribución de probabilidad de una variable aleatoria describe la forma en la que la probabilidad se distribuye entre las diferentes modalidades de una variable aleatoria.

En el caso de las variables discretas, la distribución de probabilidad está definida por una función conocida como función de masa de probabilidad o, simplemente, como función de probabilidad, que regularmente se denota como f(x).

La función de densidad de probabilidad indica la probabilidad de cada valor (modalidad) en la variable aleatoria. Para que una función pueda ser considerada de densidad de probabilidad para una variable aleatoria, deben cumplirse dos condiciones, implícitas en una distribución de probabilidad (Figura 2):

Figura 2. Probabilidad de ocurrencia



F(x) debe tener un valor positivo para cada una de las modalidades de la variable aleatoria, y

La suma de todas las modalidades de la variable aleatoria debe ser igual a uno.

Variable aleatoria continua

En contraste con las variables aleatorias discretas (que tienen un número finito de valores que pueden ser expresados a través de números enteros), las variables aleatorias continuas pueden adquirir cualquier valor dentro de un intervalo que pertenezca a los números reales o que se encuentre dentro de un conjunto de intervalos.

Debido a que existe un número infinito de valores dentro de un intervalo determinado, la variable aleatoria continua no toma un valor determinado como en el caso de las v.a.d., en este caso, la probabilidad de ocurrencia de una variable aleatoria continua se encuentra dentro de un intervalo de valores determinado.

Mientras que en las V.A.D. existe la función de masa de probabilidad para determinar la posibilidad de ocurrencia de un valor, en el caso de las V.A.C. existe la Función de Densidad de Probabilidad, que también se denota como f(x). Para una variable continua, la función de densidad de probabilidad proporciona el valor de la función para cualquier valor que adquiera la variable x, sin embargo, no calcula directamente la probabilidad de que la variable aleatorias continua x adquiera un valor específico.

A diferencia del cálculo de la probabilidad de una variable aleatoria discreta, cuya probabilidad puede calcularse de forma directa, en el caso de las variables aleatorias continuas, la probabilidad de ocurrencia de un valor está determinado por el área bajo la curva dada por f(x) de la distribución de probabilidad en un intervalo determinado. El valor correspondiente a la probabilidad en un punto dado se calcula a través de la integral de f(x) en ese intervalo, Esta operación proporciona la probabilidad que la variable tendrá si toma un valor dentro de ese intervalo.

De manera análoga al caso de la función de masa de probabilidad, la función densidad de probabilidad debe cumplir con dos condiciones:

F(x) debe considerar valores positivos para cada valor de la variable aleatoria continua, y

El resultado de la integral sobre todos, los valores de la variable aleatoria deben ser igual a uno.

Distribuciones de probabilidad más comunes.

Dependiendo del tipo de variable que se considera en el estudio y de los valores que pueda adquirir, se pueden presentar diferentes tipos de distribuciones de probabilidad. Entre las más importantes podemos mencionar:

Para el caso de Variables Aleatorias Discretas:

Distribución Bernoulli: En esta distribución, la V.A.D. puede adquirir dos posibles valores, por lo que se dice que es una distribución que se aplica a variables dicotómicas (Dos posibles valores/ modalidades). Por ejemplo, lanzar una moneda al aire.

Distribución Binomial: Se considera una generalización de Bernoulli, que puede aplicarse a eventos dicotómicos compuestos, por ejemplo: tirar varias monedas al aire.

Distribución Poisson: Se considera una generalización de la distribución Binomial, se caracteriza por la existencia de infinitos eventos de probabilidad muy baja.

Para el caso de variables aleatorias continuas tenemos:

Distribución uniforme: En esta distribución, todos los valores que puede adquirir la variable aleatoria tienen la misma probabilidad.

Distribución Exponencial: Es una distribución utilizada para modelar tiempos medios de espera para la ocurrencia de determinados eventos.

Distribución normal o Gaussiana: Es la distribución de probabilidad más usada, es un modelo teórico capaz de determinar satisfactoriamente la probabilidad de ocurrencia del valor de una variable aleatoria en una situación real. La gráfica de su función de densidad tiene forma de campana y es simétrica respecto a la media de la distribución.

Distribución T: Es una distribución que surge de la necesidad de estimar la media de una población normalmente distribuida cuando el tamaño de muestra es pequeño y la desviación estándar desconocida.

Distribución Chi cuadrada: Es una distribución de probabilidad que se identifica con el símbolo χ^2, define una distribución teórica de valores de una población cuya forma depende de los grados de libertad de las variables de estudio.

Las distribuciones de probabilidad, en el área de la salud, permiten el estudio de las relaciones entre los valores (modalidades) de una variable aleatoria y su probabilidad de ocurrencia. Durante el análisis de datos provenientes de una población, se construyen tablas, gráficas o modelos matemáticos que faciliten el entendimiento del comportamiento de una variable, sea posible estimar la frecuencia de ocurrencia de un evento.

Las distribuciones de probabilidad permiten hacer afirmaciones acerca de la variable aleatoria de estudio, determinar su ocurrencia, tomar decisiones acerca de un determinado evento y hacer predicciones acerca de un suceso con cierto grado de confianza estadística.

Próxima entrega:

Distribuciones de probabilidad I: Distribución normal.

Referencias:

Anderson, D. R., Sweeney, D., & Williams, T. A. (1999). Estadística para la Administración y Economía. México DF, México: International Thompson Editores.

Departamento de Didáctica de la Matemática. (2011). Estadística con proyectos. (C. Batanero, & C. Díaz, Eds.) Granada, España: Facultad de Ciencias de la Educación, Universidad de Granada.

García Pérez, A. (2008). Estadística aplicada: conceptos básicos (2a edición ed.). Madrid, España: Educación permanente / Universidad Nacional de Educación a Distancia.

Kazmier, L. J., Díaz Mata, A., & Eslava Gómez, G. (1991). Estadística Aplicada a Administración y Economía. Naucalpan, Estado de México, Atlacomulco, México: McGraw Hill.

Miettinen, O. S. (2011). Epidemiological research: terms and concepts. Springer Science & Business Media.

Pérez López, C. (1999). Control estadístico de la calidad. +- Madrid, España: Alfa Omega.

Wackerly, D. D., Mendenhall III, W., & Scheaffer, R. (2010). Estadística Matemática con aplicaciones. México, D.F., México: Cengage Learning Editores, S.A.

Declaración de conflicto de intereses

Los autores de este artículo expresan que no tuvieron ningún conflicto de intereses durante la preparación de este documento ni para su publicación.