MSEL
Modelling in Science Education and Learning

Modelling in Science Education and Learning
Volume 10 (2), 2017 doi: 10.4995/msel.2017.6561.
Instituto Universitario de Matemática Pura y Aplicada
Universitat Politècnica de València

Un Recorrido de Estudio e Investigación para el aprendizaje del concepto de variable aleatoria discreta mediante métodos de Monte Carlo

A Research and Study Course for learning the concept of discrete random variable using Monte Carlo methods

Vicente D. Estruch, Francisco J. Boigues, Anna Vidal
Universitat Politècnica de València
vdestruc@mat.upv.es, fraboipl@mat.upv.es, avidal@mat.upv.es

Abstract

El concepto de variable aleatoria es un constructo matemático que presenta cierta complejidad teórica. No obstante, el aprendizaje de dicho concepto puede facilitarse si se plantea como el final de un proceso secuencial de modelización de un suceso real. Más concretamente, para aprender el concepto de variable aleatoria discreta, la simulación de Monte Carlo puede ofrecer una herramienta sumamente útil puesto que en el proceso de modelización/simulación podremos abordar el concepto teórico de variable aleatoria, al tiempo que se observa a la variable aleatoria "en acción". Este trabajo expone un Recorrido de Estudio e Investigación (REI) basado en una serie de actividades relacionadas con variables aleatorias como entrenamiento e introducción de elementos de simulación, presentándose después la construcción de un modelo, que es la parte substancial de la actividad, generando una variable aleatoria y su función de probabilidad. Partiendo de una situación sencilla, relacionada con la reproducción y supervivencia de la camada de un roedor, con componentes aleatorios, se construye, paso a paso, el modelo que representa la situación planteada mediante una variable aleatoria "original". En las etapas intermedias de la construcción del modelo tienen un papel fundamental las distribuciones uniforme discreta y binomial. El recorrido de tales etapas permite reforzar el concepto de variable aleatoria al tiempo que se exploran las posibilidades que ofrecen los métodos de Monte Carlo para simular casos reales y se comprueba la sencillez que supone implementar dichos métodos mediante el lenguaje de programación de Matlab©.

The concept of random variable is a mathematical construct that presents some theoretical complexity. However, learning this concept can be facilitated if it is presented as the end of a sequential process of modeling of a real event. More specifically, to learn the concept of discrete random variable, the Monte Carlo simulation can provide an extremely useful tool because in the process of modeling / simulation one can approach the theoretical concept of random variable, while the random variable is observed "in action". This paper presents a Research and Study Course (RSC) based on series of activities related to random variables such as training and introduction of simulation elements, then the construction of the model is presented, which is the substantial part of the activity, generating a random variable and its probability function. Starting from a simple situation related to reproduction and survival of the litter of a rodent, with random components, step by step, the model that represents the real raised situation is built obtaining an "original" random variable. In the intermediate stages of the construction of the model have a fundamental role the uniform discrete and binomial distributions. The trajectory of these stages allows reinforcing the concept of random variable while exploring the possibilities offered by Monte Carlo methods to simulate real cases and the simplicity of implementing these methods by means of the Matlab© programming language.

Palabras clave: Variable aleatoria discreta, modelización, simulación, métodos de Monte Carlo, Microsoft Excel©, Matlab©.
Keywords: Discrete random variable, modeling, simulation, Monte Carlo methods, Microsoft Excel©, Matlab©.

1. Introducción

1.1. La Estadística en la enseñanza superior

La Estadística es considerada parte de las materias básicas en la formación que ha de proporcionar a los alumnos universitarios la adquisición de competencias para la investigación o el ejercicio profesional, fundamentalmente en ámbitos relacionados con las ciencias de la naturaleza e ingenierías. Dicha formación ha de incluir, obviamente, aspectos teóricos de la Estadística. No obstante, el aspecto formativo más importante en el aprendizaje de la Estadística en ciencias e ingenierías se centra, sobre todo, en el valor instrumental de la misma para abordar problemas prácticos reales, donde se trabaja con cantidades importantes de datos. Por ello, el uso de paquetes informáticos, de cálculo general o específicamente estadísticos, tiene un papel esencial.

Por otra parte, a la luz de investigaciones recientes, el aprendizaje efectivo de la Estadística en la universidad necesita sustentarse en la comprensión de los conceptos estadísticos básicos y en la construcción progresiva de representaciones mentales implícitas o explícitas. (Escalante Gómez, 2008), lo cual conlleva abordar con rigor las bases teóricas que sustentan las técnicas de análisis estadístico. En este punto aparecerán obstáculos diversos, donde destacaría la escasa formación en Estadística de los alumnos que acceden a la universidad; no siendo menos importantes los problemas relacionados del ámbito emocional, puesto que los déficits motivacionales y actitudinales constatados, así como la importancia concedida a la dimensión actitudinal en el desempeño competente, incluso han alentado el desarrollo de un área de investigación que aborda el dominio afectivo-actitudinal en los procesos de enseñanza-aprendizaje de la Estadística (Blanco, 2008).

1.2. Modelos probabilísticos y situaciones reales

Los elementos de las matemáticas necesarios para afrontar los problemas relacionados con resultados de experimentos aleatorios son los que llevan a clasificar los métodos probabilísticos en discretos y continuos (Brase y Brase, 2016, Peck, 2014). Un enfoque discreto se utiliza cuando el número de resultados experimentales es finito o infinito numerable, como por ejemplo el número de crías en una camada. Por otro lado, en un experimento en que se mide el tiempo que tarda una reacción química en completarse, es posible que los resultados puedan variar desde 0 a T segundos. En este caso los resultados posibles constituyen un conjunto infinito no numerable, el de los valores reales que forman el intervalo $[0, T]$, lo que nos lleva a un enfoque continuo.

En el enfoque discreto, es posible asignar probabilidades a los resultados individuales. En el enfoque continuo no será posible asignar una probabilidad no nula a cada resultado y sólo podremos asignar probabilidades a sucesos asociados a que la variable pertenezca a un subintervalo del intervalo total de los valores posibles o rango de la variable aleatoria. Siguiendo con el ejemplo de la reacción química, en ese caso podríamos plantear la probabilidad de que se complete la reacción en un tiempo t, $a \leq t \leq b$, con $[a, b] \subseteq [0, T]$.

El caso de rango discreto finito de una variable aleatoria es el más fácil de conceptualizar y de describir matemáticamente, lo cual facilita obtener ejemplos sencillos que permitan abordar mejor, en una primera etapa, el estudio formal de las variables aleatorias.

1.3. La probabilidad en el caso de variables aleatorias discretas

Consideremos un experimento aleatorio en que se pueden obtener los resultados numéricos del conjunto discreto de cardinal finito $X \in \{x_1, x_2, \ldots, x_n\}$. Entonces $P(X = x_1) + P(X = x_2) + \ldots + P(X = x_n) = 1$. Si el número de posibles resultados, $X \in \{x_1, x_2, \ldots, x_i, \ldots\}$, es infinito numerable, entonces $\sum_{i=1}^{\infty} P(X = x_i) = 1$; con lo que necesitamos recurrir al concepto de serie numérica y la función de probabilidad $f(x) = P(X = x)$ solo tendrá sentido si la serie de números no negativos dada por $\sum_{i=1}^{\infty} f(x_i)$ es convergente.

En el caso particular de que todos los resultados tengan la misma probabilidad de suceder, $P(X = x_i) = k > 0$, el número de resultados será necesariamente finito, ya que la serie $\sum_{i=1}^{\infty} k$, $k > 0$, es divergente. En el caso de un número finito de resultados numéricos del experimento, $X \in \{x_1, x_2, \ldots, x_n\}$, todos con la misma probabilidad, en base a la definición de probabilidad de Laplace, la probabilidad de que suceda cualquiera de ellos será $f(x_i) = P(X = x_i) = 1/n, i = 1, 2, \ldots, n$, con lo que sólo necesitamos un parámetro (número de elementos de X, n) para caracterizar completamente la distribución de probabilidad, y poder calcular cualquier medida estadística asociada.

Conocer con detalle, desde un punto de vista teórico, una distribución permite calcular, de forma inmediata y precisa, probabilidades de interés u otras características estadísticas de la distribución como, por ejemplo, el valor esperado, la desviación típica, percentiles teóricos, etc. Sin embargo, en muchos problemas prácticos no es fácil obtener una expresión analítica de la función de probabilidad asociada a la variable aleatoria que define un modelo, el cual representa un caso real. No obstante, si conocemos las características del fenómeno aleatorio, y no hay impedimento alguno para repetir el experimento un gran número de veces, podremos obtener la distribución de probabilidad de forma aproximada, o lo que es lo mismo una aproximación empírica de la distribución de probabilidad.

En los inicios de los estudios de probabilidad, era frecuente que los investigadores recurrieran a este tipo de cálculos empíricos, obviamente muy tediosos por el hecho de tener que repetir muchas veces el experimento aleatorio, pero sin embargo inevitables al afrontar ciertos procesos complejos. En esta situación, es muy útil tener la capacidad de simular los procedimientos, es decir, obtener una muestra de gran tamaño sin tener que realizar realmente el experimento. Los ordenadores permiten que esto sea posible y, además, rápido. Un ordenador puede realizar o repetir un cálculo miles o millones de veces. Lo único que se necesita es un medio (lenguaje de programación y la correspondiente rutina) para simular el proceso. La mayoría de los paquetes de software científico incorporan generadores de números pseudoaleatorios. En concreto la hoja de cálculo Excel de Microsoft y Matlab se manifiestan como herramientas útiles para afrontar la simulación de modelos mediante la generación de valores pseudoaleatorios.

En este trabajo, se presenta un Recorrido de Estudio e Investigación (REI) para la introducción del concepto de variable aleatoria discreta en la asignatura Instrumentos de Estadística y Simulación, de segundo curso de los estudios del Grado en Ciencias Ambientales en la Escuela Politécnica Superior de Gandia-Universitat Politècnica de València. Para ello se modelarán situaciones reales mediante modelos probabilísticos concatenados y se recurrirá a la simulación de Monte Carlo como medio para obtener información sobre las características de dichos modelos.

2. Objetivos

Los REI pueden funcionar localmente como mecanismos didácticos capaces de superar la tradicional desarticulación o atomización de las matemáticas que se aprenden en la universidad, asumiendo una función articuladora, la cual se deriva directamente de su capacidad para dar a la modelización matemática un papel más importante en los sistemas de enseñanza (Boigues et al, 2011).

El REI que se presenta en este trabajo, se materializa en una serie de actividades prácticas cuyo objetivo primario es que los alumnos adquieran las siguientes competencias:

- Distinguir entre modelo probabilístico y resultados de un experimento aleatorio.
- Modelar los resultados de situaciones con componentes aleatorias mediante variables aleatorias.
- Utilizar un asistente informático para obtener series de valores pseudoaleatorios.
- Estimar las características de una distribución de probabilidad mediante simulación de Monte Carlo.
- Desarrollar el problema y el proceso de modelización.
- Afrontar situaciones equivalentes y modelo general.

3. El problema y el proceso de modelización

3.1. Situaciones equivalentes y modelo general

Los modelos de probabilidad son aproximaciones simplificadas de la realidad. Como tales aproximaciones, deben contemplar las características más importantes del fenómeno aleatorio real para ser útiles como herramientas de predicción, teniendo en cuenta que, por otra parte, deben ser lo más simples posible ya que un modelo difícil de manejar no será útil para ser utilizado en la práctica.

Consideremos la siguiente situación real (R0):

R0: *Datos empíricos, obtenidos mediante observación directa, indican que cierto roedor se reproduce de forma que, en el momento del alumbramiento, en cada camada nos encontramos con* 1, 2, 3, 4, 5 *ó* 6 *crías vivas con la misma proporción; siendo la proporción de camadas nulas o la de camadas con un número de crías vivas superior a* 6 *prácticamente despreciable. Al cabo de un mes, se ha estimado que la tasa de supervivencia de las crías es de una camada es del* 85 %*. El objetivo es estimar cuántos individuos de una camada cualquiera sobreviven al cabo de un mes.*

La situación R0 es un ejemplo claro de sistema cuyo comportamiento tiene componentes aleatorias. Por una parte se tiene que en el alumbramiento tendremos un número determinado de crías, entre 1 y 6, con igual probabilidad y, por otra parte, de esas crías iniciales sobrevivirá un 85 %, o lo que es lo mismo, la probabilidad que una cría recién nacida esté viva al cabo de un mes es 0.85. Por lo tanto en R0 se distinguen dos etapas sucesivas en el tiempo. Llamaremos R1 a la situación que describe la primera etapa:

R1: *Se tiene que cierto roedor se reproduce de forma que, en el momento del alumbramiento, en cada camada nos encontramos con* 1, 2, 3, 4, 5 *ó* 6 *crías vivas con la misma proporción;*

siendo la proporción de camadas sin crías vivas o la de camadas con un número de crías vivas superior a 6 prácticamente despreciable.

Y R2 a la siguiente etapa:

R2: *Dada una camada inicial de N crías, al cabo de un mes, se ha comprobado que sobreviven aproximadamente el 85 % de las mismas.*

Es obvio que, si asumimos la simplificación de considerar nula tanto la probabilidad de 0 crías como la de más de 6 crías vivas en la camada, la situación planteada en R1 es asimilable al experimento aleatorio de lanzar un dado perfecto y observar el valor que aparece en la cara superior, que llamaremos R3:

R3: *En el experimento aleatorio de lanzar un dado, en la cara superior puede aparecer cualquiera de los valores $\{1,2,3,4,5,6\}$, con la misma probabilidad.*

Diremos que las situaciones R1 y R3 son equivalentes en comportamiento probabilístico o lo que es lo mismo, responden al mismo modelo teórico probabilístico, que llamaremos M1, que se describe mediante una variable aleatoria discreta concreta:

M1: *Sea X una variable aleatoria que toma los valores $\{1,2,3,4,5,6\}$, con la misma probabilidad; dada por la función $f(i) = P(X = i) = 1/6$, $i = 1,2,3,4,5,6$. En este caso se dice que la variable aleatoria X sigue una distribución uniforme, o rectangular, discreta, en el intervalo de valores enteros $[1,6]_{\mathbb{Z}} = [1,6] \cap \mathbb{Z}$.*

El valor esperado (o media) y la varianza de la variable aleatoria X del modelo M1 son, respectivamente:

$$\mu_X = \sum_{i=1}^{6} i\frac{1}{6} = 3.5, \qquad \sigma_X^2 = \sum_{i=1}^{6}(i-3.5)^2\frac{1}{6} = 2.916.$$

El razonamiento anterior es un ejemplo de proceso de abstracción (modelización) consistente en el paso desde dos situaciones reales con comportamiento equivalente (R1 y R3) a un modelo estadístico general (M1), común para las dos situaciones.

Por otra parte la situación descrita en R2, se ajusta claramente a un modelo Binomial puesto que si se tiene que X es el número de crías vivas en una camada, al cabo de un mes el número de crías que habrán sobrevivido puede ser cualquier valor Y en el conjunto $\{0,1,\ldots,X\}$, siendo la probabilidad de que una cría de la camada sobreviva igual a 0.85.

Por lo tanto se tiene el modelo M2.

M2: *Fijado un valor entero positivo X, que puede tomar los valores $\{1,2,3,4,5,6\}$, sea Y la variable aleatoria que distribuye binomial con parámetros X y $p = 0.85$ —$B(X,p)$—. Por lo tanto Y puede tomar los valores $\{0,1,\ldots,X\}$ y la función de probabilidad de Y viene dada por*

$$f(k) = P(Y = k) = \binom{X}{k} \cdot 0.85^k \cdot (1-0.85)^{X-k}, \quad k = 0,1,2,\ldots,X.$$

A partir de los modelos M1 y M2 se deduce el modelo global M que representa la situación inicial R0.

M: *Sea X una variable aleatoria que se distribuye uniforme discreta en el intervalo $[1,6]_{\mathbb{Z}}$ y sea $Y = Y(X)$ la variable aleatoria que se distribuye binomial con parámetros X y p —$B(X,p)$—, en otras palabras Y es una variable aleatoria que se distribuye $B(X,p)$ donde X es una variable aleatoria que se distribuye uniforme discreta en el intervalo $[1,6]_{\mathbb{Z}}$. Por lo tanto Y puede tomar los valores $\{0,1,2,3,4,5,6\}$.*

Obtener la función de probabilidad para Y —$f_Y(k) = P(Y=k)$— es relativamente sencillo teniendo en cuenta el teorema de la probabilidad total:

$$\begin{aligned} P(Y=0) &= P(Y=0|X=1)P(X=1) + \ldots + P(Y=0|X=6)P(X=6) \\ &= \frac{1}{6}\sum_{i=1}^{6}\binom{i}{0}p^0(1-p)^{i-0} = \frac{1}{6}\sum_{i=1}^{6}(1-p)^i, \end{aligned}$$

$$\begin{aligned} P(Y=k) &= P(Y=k|X=k)P(X=k) + \ldots + P(Y=k|X=6)P(X=6) \\ &= \frac{1}{6}\sum_{i=k}^{6}\binom{i}{k}p^k(1-p)^{i-k} = \frac{1}{6}\sum_{i=1}^{6}p^k(1-p)^{i-k}, \quad k=1,\ldots,6, \end{aligned}$$

puesto que

$$P(X=i) = \frac{1}{6}, \quad i=1,\ldots,6,$$

$$P(Y=k|X=u) = \binom{u}{k}p^k(1-p)^{i-k}, \quad u=1,\ldots,6,\ k=0,\ldots 6, \quad u \geq k.$$

Por lo tanto, el modelo M viene dado por una variable aleatoria Y, que puede tomar los valores $\{0,1,2,3,4,5,6\}$, cuya función de probabilidad viene dada por las expresiones:

$$f_Y(x) = P(Y=x) = \begin{cases} \frac{1}{6}\sum_{i=1}^{6}(1-p)^i & x=0 \\ \frac{1}{6}\sum_{i=x}^{6}p^x(1-p)^{i-x} & x=1,\ldots,6 \end{cases}$$

En un planteamiento más general, si X se distribuye uniforme en el intervalo $[1,n]_{\mathbb{Z}}$, e Y se distribuye binomial $B(X,p)$, se tendrá que la función de probabilidad para Y vendrá dada por

$$f_Y(x) = P(Y=x) = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}(1-p)^i & x=0 \\ \frac{1}{n}\sum_{i=x}^{n}p^x(1-p)^{i-x} & x=1,\ldots,n \end{cases} \tag{1}$$

Por lo tanto, a partir del caso general, se observa que la distribución de probabilidad de la variable aleatoria Y queda totalmente determinada a partir de los parámetros n y p.

La Figura 4 ilustra el comportamiento de la distribución de la variable aleatoria Y que representa al modelo M, para el caso concreto $n=6$ y $p=0.85$:

Valor x	Probabilidad $f_Y(x) = P(Y = x)$
0	0.0294
1	0.1961
2	0.1958
3	0.1937
4	0.1816
5	0.1405
6	0.0629

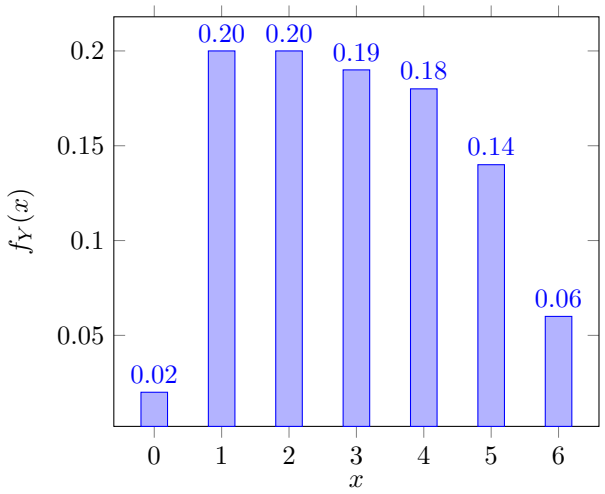

Figura 1: a) Tabla con los valores de la función de probabilidad correspondientes a los distintos valores que puede tomar la variable aleatoria Y que define el modelo M. b) Gráfico de barras de las probabilidades teóricas correspondientes.

Obtener una expresión genérica para el valor esperado de Y, μ_Y, y para la varianza de Y, σ_Y^2, exige cierto trabajo de cálculo aritmético, complicado para los estudiantes. Se puede probar que, para la variable aleatoria Y con parámetros n y p,

$$\mu_Y = \sum_y y f_Y(y) = \frac{1+n}{2}p, \tag{2}$$

$$\sigma_Y^2 = \sum_y (y-\mu_Y)^2 = \frac{1+n}{2}\big((1-p) + \frac{1+2n}{3}p\big) - \big(\frac{1+n}{2}p\big)^2. \tag{3}$$

Para el caso particular $n = 6$ y $p = 0.85$, se obtiene que $\mu_Y = 2.975$ y $\sigma_Y^2 \approx 2.554$. Como se puede observar en la Figura 4 (derecha), la distribución de la variable aleatoria Y tiene un comportamiento singular en el que destaca cierta asimetría observable a simple vista.

3.2. Análisis teórico frente a la simulación por ordenador

En muchas ocasiones, en el proceso de modelización de un caso realista mediante variables aleatorias, se tropieza con el obstáculo de la dificultad de obtener la función de probabilidad precisa, con lo que también será difícil obtener expresiones analíticas para las medidas estadísticas que caracterizan la distribución de la variable aleatoria. Es en este contexto en el que resulta indicado recurrir a los métodos de Monte Carlo, los cuales pueden proporcionar una muestra extensa de posibles valores de la variable aleatoria, cuya distribución empírica supondrá una aproximación a la distribución teórica.

Utilizar, de forma efectiva, métodos de Monte Carlo va asociado a tener que recurrir a un asistente de cálculo capaz de generar valores pseudo aleatórios. Por ejemplo, Matlab, puede generar un valor, x, en el intervalo $(0, 1)$ mediante la sentencia `x=rand(1,1)`. El número x es elegido, teóricamente, al azar por lo que se situará en cualquier parte del intervalo $(0, 1)$. Para simular el lanzamiento de una moneda, puesto que la probabilidad de que el valor generado x pertenezca al intervalo $(0, 1/2]$ será 0.5, entonces si $x \in (0, 1/2]$, contabilizaremos cara y si $x \in (1/2, 1)$, lo cual tiene también una probabilidad 0.5, contabilizaremos cruz. Para obtener muchos valores pseudo-aleatorios, basta con ejecutar en Matlab `x=rand(U,V)` para obtener

una matriz con U filas y V columnas, cada uno de cuyos elementos se supone seleccionado aleatoriamente y con las mismas posibilidades de aparecer en el intervalo real $(0, 1)$. La hoja de cálculo Excel también permite generar dichos valores mediante la orden `=aleatorio()`.

El lanzamiento de un dado (R3), y por lo tanto la aproximación al problema de las crías en una camada (R1), también puede hacerse mediante una rutina y un ordenador. En el caso del dado, se procedería del siguiente modo:

Algoritmo

(1) Generamos un valor aleatorio $x \in (0, 1)$.

(2) Si $x \in (0, 1/6]$, contabilizamos 1, si $x \in (1/6, 2/6]$, contabilizamos 2,
si $x \in (2/6, 3/6]$, contabilizamos 3, si $x \in (3/6, 4/6]$, contabilizamos 4,
si $x \in (4/6, 5/6]$, contabilizamos 5 y, por último,
si $x \in (5/6, 1)$, contabilizamos 6.

No obstante, Matlab también dispone de la orden `unidrnd(k,U,V)` que nos proporciona, directamente, una matriz $U \times V$ de valores pseudoaleatorios que se distribuyen uniforme discreta en el intervalo $[1, k]_{\mathbb{Z}}$.

La simulación por ordenador de fenómenos aleatorios es una herramienta indispensable en la investigación científica moderna. Los métodos de Monte Carlo también pueden utilizarse para una mejor comprensión de los problemas probabilísticos.

3.3. Valores aleatorios y pseudoaleatorios

Es importante distinguir entre generar valores aleatorios y valores pseudoaleatorios que, supuestamente, siguen cierta distribución, Estos últimos son realmente los que nosotros podemos obtener para las simulaciones que se llevan a cabo.

> *En [COLABORADORES DE WIKIPEDIA] se define generador de números aleatorios como: "...un dispositivo informático o físico diseñado para producir secuencias de números sin un orden aparente". Por otra parte, "Un generador pseudoaleatorio de números (GPAN) es un algoritmo que produce una sucesión de números que es una muy buena aproximación a un conjunto aleatorio de números. La sucesión no es exactamente aleatoria en el sentido de que queda completamente determinada por un conjunto relativamente pequeño de valores iniciales, llamados el estado del GPAN. ..., los números pseudoaleatorios son importantes en la práctica para simulaciones (por ejemplo, de sistemas físicos mediante el método de Montecarlo)?" (Wiki 1).*

Los paquetes de cálculo realmente proporcionan sucesiones de valores pseudoaleatorios, y una forma de verificar si dichos valores se ajustan al modelo teórico lo proporciona, por ejemplo, el estadístico Chi-cuadrado, que se calcula a partir de comparar frecuencias absolutas observadas y valores teóricos esperados de las frecuencias absolutas. En el caso de la simulación de los resultados del lanzamiento de un dado, para verificar si N valores obtenidos mediante un generador se ajustan al modelo teórico, elaboraremos una tabla con los posibles valores de la variable aleatoria y, para cada valor, la frecuencia absoluta en la simulación, la probabilidad teórica, la frecuencia absoluta esperada y el sumando correspondiente para el cálculo del estadístico Chi-cuadrado (Tabla 1).

Resultado	Frecuencia absoluta en la simulación (N_i)	Probabilidad Teórica (P_i)	Frecuencia absoluta esperada ($N \cdot P_i$)	$C_i = \frac{(N_i - N/6)^2}{N/6}$
1	N_1	1/6	$N/6$	C_1
2	N_2	1/6	$N/6$	C_2
3	N_3	1/6	$N/6$	C_3
4	N_4	1/6	$N/6$	C_4
5	N_5	1/6	$N/6$	C_5
6	N_6	1/6	$N/6$	C_6

Tabla 1: Formato de tabla para el análisis de resultados de una simulación donde se generan N valores pseudo-aleatorios distribuidos uniforme discreta $[1, 6]_{\mathbb{Z}}$.

Y calculamos el valor del estadístico χ^2_c:

$$\chi^2_c = \sum_{i=1}^{6} C_i = \sum_{i=1}^{6} \frac{(N_i - N \cdot P_i)^2}{N \cdot P_i} = \sum_{i=1}^{6} \frac{(N_i - N/6)^2}{N/6} = \frac{6}{N} \sum_{i=1}^{6} \left(N_i - \frac{N}{6}\right)^2.$$

La expresión de χ^2_c cuantifica en qué medida se desvían las frecuencias absolutas obtenidas de la simulación respecto de las frecuencias absolutas esperadas si se asume como cierta la distribución teórica. El valor χ^2_c puede ser utilizado para contrastar la hipótesis de que la simulación se ajusta al modelo teórico frente a la alternativa de que no se ajusta, con un nivel de confianza prefijado. El estadístico muestral χ^2_c , en el caso de distribuciones discretas finitas, se distribuye Chi-cuadrado con $n-1$ grados de libertad, siendo n el número de valores distintos que puede tomar la variable aleatoria. En el caso de una distribución uniforme discreta en el intervalo $[1, 6]_{\mathbb{Z}}$, se tiene que $n = 6$. Por lo tanto los grados de libertad serán $n - 1 = 5$.

Al comparar el valor obtenido para el estadístico, χ^2_c , con el valor teórico $\chi^2_{5,0.05} = 11.070$, correspondiente a una variable aleatoria χ^2_c distribuida Chi-cuadrado con 5 grados de libertad de forma que se cumple que $P(\chi^2 \geq \chi^2_{5,0.05}) = 0.05$; si sucede que $\chi^2_c \leq 11.070$ podemos aceptar, o no podemos rechazar, a un nivel de confianza del 95 %, que la simulación ha proporcionado una muestra que se ajusta al modelo teórico. En caso contrario ($\chi^2_c > 11.070$) no podemos aceptar, a un nivel de confianza del 95 %, que la simulación haya proporcionado una muestra que se ajuste al modelo teórico.

4. Un recorrido de estudio e investigación (REI) para aprender el concepto de variable aleatoria discreta y su relación con la modelización de procesos aleatorios

El REI que se presenta, en este trabajo busca introducir el concepto de variable aleatoria discreta, desde la perspectiva de la modelización, en la asignatura Instrumentos de Estadística y Simulación, de segundo curso de los estudios del Grado en Ciencias Ambientales en la Escuela Politécnica Superior de Gandia-Universitat Politècnica de València. El esquema del REI se muestra en la Figura **2**.

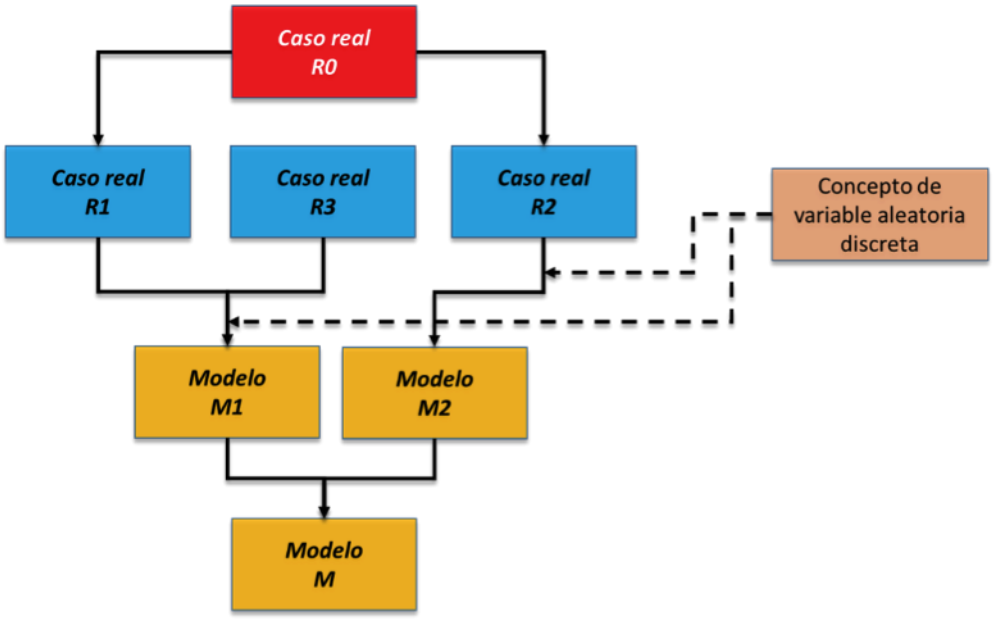

Figura 2: Esquema del desarrollo del REI para el estudio de un modelo basado en una variable aleatoria discreta.

Entrando un poco más en el detalle, el REI seguiría los pasos que se enumeran a continuación:

1. Se expone la situación real R0.
2. Se distinguen las dos partes que conforman R0: R1 y R2.
3. Se constata, asumiendo ciertas simplificaciones, que la situación real R1 es análoga a la situación real R3. Esta constatación ilustra lo que sucede muchas veces en la realidad: es necesario simplificar como paso previo a modelar.
4. Teniendo en cuenta la definición de variable aleatoria discreta, se evidencia que las situaciones reales R1 y R3 son dos concreciones del modelo teórico M1 (distribución uniforme discreta en el intervalo $[1,6]_{\mathbb{Z}}$). Se constata que R3 corresponde a un modelo Binomial (M2) A partir de M1 y M2, teniendo en cuenta el teorema de la probabilidad total, se construye el modelo M. Se utiliza el programa Matlab para generar 1000 valores pseudoaleatorios distribuidos uniforme discreta en el intervalo $[1,6]_{\mathbb{Z}}$ (según el modelo M1).
5. Se comprueba el nivel de ajuste de los valores obtenidos al modelo teórico M1.
6. Se utiliza otra rutina de Matlab para obtener 1000 valores pseudoaleatorios distribuidos según la distribución correspondiente al modelo M.
7. Se analizan y valoran los resultados de la simulación del modelo M.

5. Desarrollo del aprendizaje

La actividad propuesta se desarrolla en una sesión de 2 horas de clase de prácticas en aula informática. El trabajo se realiza por etapas, y no se pasa de una etapa a otra si no se ha completado la anterior.

Etapa 1.-Se expone como elemento de discusión a todo el grupo el caso real R1.

No resulta difícil que los alumnos lleguen a la conclusión de que el caso R1 se comporta de manera análoga al experimento aleatorio de lanzar un dado sobre una superficie plana y registrar el valor que aparece en la cara superior, que hemos llamado R3.

Etapa 2.-Reconocer que R1 y R3 son manifestaciones diferentes de un mismo modelo aleatorio.

Teniendo en cuenta la definición de variable aleatoria discreta, que se habrá estudiado previamente en clase de teoría, se discute en grupo la formalización de los casos R1 y R3 mediante una variable aleatoria que se distribuye uniforme discreta en el intervalo $[1,6]_{\mathbb{Z}}$. Esta etapa puede abrirse a la creatividad discutiendo ejemplos de otros procesos aleatorios equivalentes, que pueden ser representados por un modelo aleatorio único, basado en una variable aleatoria discreta.

Etapa 3.-Simulación de una distribución uniforme discreta en el intervalo $[1,6]_{\mathbb{Z}}$.

Para generar 1000 valores pseudoaleatorios según una distribución uniforme discreta en el intervalo $[1,6]_{\mathbb{Z}}$ utilizaremos Matlab, concretamente el script `uniforme.m`:

```
%uniforme.m (genera 1000 valores uniforme discreta [1,6]_Z)
A=unidrnd(6,1000,1);
tabulate(A)
media=mean(A)
varianza=var(A)
table=tabulate(A);
```

Que proporciona la media aritmética la varianza muestral, y la tabla de frecuencias absolutas (recuentos) y relativas (en porcentaje) de los valores generados que constituyen la salida del programa. Realizamos un ensayo obteniendo:

```
media =
            3.4750
varianza =
            2.8903
>>   Value    Count   Percent
                1         171        17.10%
                2         164        16.40%
                3         165        16.50%
                4         180        18.00%
                5         159        15.90%
                6         161        16.10%
```

En este caso el valor obtenido para $\chi^2_c = 1.784 < \chi^2_{5,0.05} = 11.070$ indica un buen ajuste al modelo teórico. El gráfico de barras de frecuencias relativas correspondiente es el expuesto en la Figura **3**.

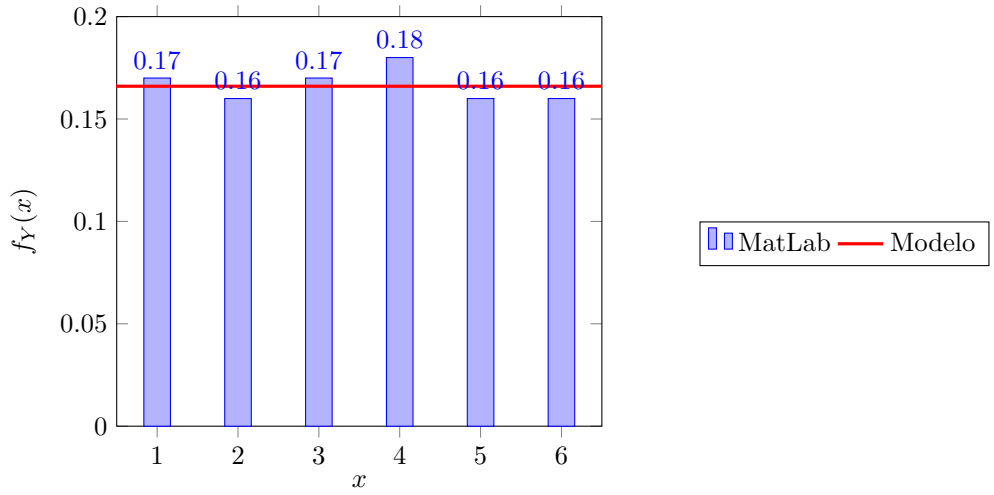

Figura 3: Gráfico de barras de frecuencias relativas correspondiente a 1000 valores que siguen una distribución uniforme discreta en el intervalo $[1,6]_{\mathbb{Z}}$.

Los resultados se resumen en la Tabla 2:

	Modelo Teórico	Simulación Matlab
1	$1/6 = 0.16$	0.17
2	0.16	0.16
3	0.16	0.17
4	0.16	0.18
5	0.16	0.16
6	0.16	0.16
Valor esperado (μ)	3.5	3.48
Varianza (σ^2)	2.916	2.89
χ^2_c	0	1.784

Tabla 2: Valores de probabilidad correspondientes al modelo teórico y resultados obtenidos mediante simulación con Matlab al considerar una muestra de 1000 realizaciones para una variable aleatoria uniforme discreta en el intervalo $[1,6]_{\mathbb{Z}}$.

Etapa 4.-La simulación de un modelo binomial $B(n,p)$.

¿Por qué puede ser interesante tener una herramienta para simular una variable aleatoria uniforme discreta en el intervalo $[1,6]_{\mathbb{Z}}$?, o lo que es lo mismo, ¿Por qué sería importante simular el resultado obtenido al lanzar un dado? La cuestión puede tener múltiples respuestas cuya discusión puede dar lugar a una actividad muy interesante para ser desarrollada en una clase presencial o en un foro, por ejemplo. No obstante, en términos generales, en modelización, normalmente se actúa por pasos y la simulación de un dado mediante un ordenador permite obtener valores supuestamente aleatorios sin necesidad de efectuar el lanzamiento físico del dado. A partir de aquí, partiendo de la situación real R1≡ R3, que es la más simple, se procede secuencialmente, haciendo más complejo el sistema hasta abarcar el problema real planteado inicialmente.

Para simular el modelo M2 simplemente necesitamos un generador de valores pseudoaleatorios con distribución binomial. Matlab proporciona la sentencia `binornd(n,p,U,V)` que pro-

porciona una matriz de valores pseudoaleatorios con distribución binomial $B(n, p)$, de tamaño $U \times V$.

Para realizar una prueba, se ejecuta por ejemplo `binornd(6,0.85,10,10)`. Obteniéndose una matriz 10×10 con 100 valores entre 0 y 6 que se supone responden a una distribución `binomial B(6,0.85)`.

```
ans =
     5     5     6     6     6     5     6     5     6     5
     6     6     6     5     6     5     4     5     4     5
     6     5     4     5     6     5     5     5     6     5
     5     5     5     4     5     6     5     5     3     6
     5     5     5     5     5     5     6     6     6     6
     6     4     5     6     6     4     6     5     4     4
     6     4     4     5     6     6     5     5     4     6
     5     4     6     6     5     3     5     5     5     4
     6     4     6     5     5     5     5     6     5     5
     5     5     5     4     5     6     6     6     4     4
```

Etapa 5.-La situación inicial R0 y el modelo global M

Aunque el modelo M corresponde a una variable aleatoria que tiene una función de probabilidad teórica y unos valores de media y varianza que se pueden obtener por cálculo aritmético —ver (4), (5) y (6)—, obtendremos, mediante simulación, valores aproximados para la función de probabilidad y para la media, μ_Y, y varianza, σ_Y^2.

Para obtener una muestra de 10000 valores de la variable Y, con parámetros $n = 6$ y $p = 0.85$, se utiliza la rutina `crias.m` de Matlab, que se proporciona a los alumnos, la cual sirve para un modelo más general con valores n y p cualesquiera (dentro de los admisibles pues n debe ser un número entero positivo y p un valor real, $0 \leq p \leq 1$). La rutina `crias.m` proporciona como salida un número predeterminado de valores pseudoaleatorios de la variable aleatoria Y, una tabla con la distribución de dichos valores, la media muestral, la varianza muestral y el valor muestral del estadístico Chi cuadrado, que posteriormente podemos comparar con el valor teórico para determinar si aceptamos o rechazamos que la muestra obtenida se ajusta a la distribución teórica:

```
%rutina crias.m
n=input('parametro n= ') %se introduce el valor n
p=input('parametro p= ') %se introduce la probabilidad de supervivencia p
repe=input('numero de simulaciones= ') %se introduce el tamaño de la
                                       muestra a generar
A=unidrnd(n,repe,1); %se genera una muestra de valores distribuidos
                                       uniforme continua en %(0,1)
Y=[];
for i=1:repe %se genera una muestra para la variable Y
y=binornd(A(i),p,1,1);
Y=[Y,y];
end
tabulate(Y) %se genera una tabla con las frecuencias absoluta y relativa
                 (en %) para los  valores de Y generados
media=mean(Y) %se calcula la media de la muestra
```

```
varianza=var(Y) %se calcula la varianza de la muestra
table=tabulate(Y);
bar(0:6,table(:,3)) %se genera un gráfico de barras para la muestra
                de valores de Y
xlabel('Resultados')
ylabel('frecuencia %')
Chi=sum(((table(:,2)-[0.0294;0.1961;0.1958;0.1937;0.1816;0.1405;0.0629]
        *repe).^2)./([0.0294;0.1961;0.1958;? 0.1937;0.1816;0.1405;0.0629]
        *repe)) %cálculo de Chi-cuadrado para la muestra obtenida (Y)
```

A partir de los comentarios del script (texto situado tras el símbolo %) es sencillo saber qué es lo que éste ejecuta en cada momento. Un resultado posible tras ejecutar el script `crias.m` con $n = 6$, $p = 0.85$ y `repe=10000` simulaciones es:

```
parametro n= 6
n =
     6
parametro p= 0.85
p =
    0.8500
numero de simulaciones= 10000
repe =
       10000
  Value    Count   Percent
      0      335      3.35%
      1     1946     19.46%
      2     1880     18.80%
      3     1943     19.43%
      4     1852     18.52%
      5     1417     14.17%
      6      627      6.27%
media =
    2.9790
varianza =
    2.5840
Chi=
   9.7808
```

Y el gráfico de barras de la Figura 4.

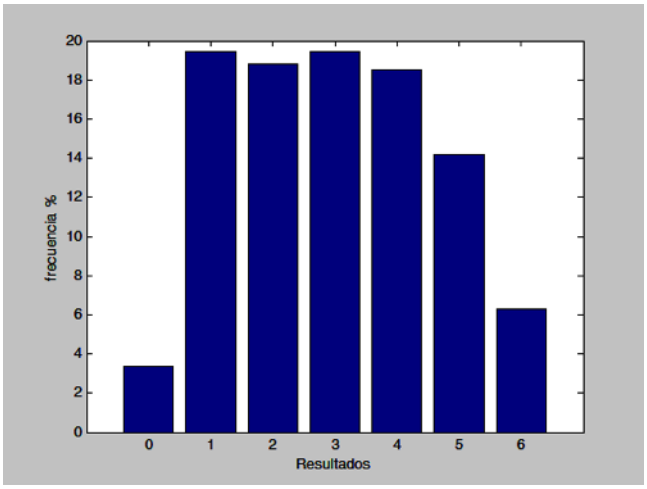

Figura 4: Gráfico de barras correspondiente a los valores de Y obtenidos de la simulación del modelo M realizada con Matlab.

La Figura 4 obtenida reproduce aproximadamente la forma de la Figura 4 (derecha) con lo que, a simple vista, la simulación se asemeja al modelo teórico. A los alumnos, en este punto, se les proporcionan las probabilidades teóricas —Figura 4 (izquierda)— y los valores teóricos de la media y la varianza de Y. En el ejemplo que nos ocupa, se construiría la Tabla 3 que facilita la comparación entre las probabilidades teóricas y los valores empíricos obtenidos de la simulación

	Modelo Teórico	Simulación Matlab
0	0.0294	0.0350
1	0.1961	0.1946
2	0.1958	0.1880
3	0.1937	0.1943
4	0.1816	01852
5	0.1405	01417
6	0.0629	00627
Valor esperado (μ)	2.975	2.9790
Varianza (σ^2)	2.554	2.5840
χ^2_c	$--$	9.7808

Tabla 3: Probabilidades, valores esperados y varianza, para el modelo teórico y para la muestra obtenida mediante simulación, y valor del estadístico χ^2_c correspondiente a la muestra.

En este caso hay que comparar $\chi^2_c = 9.7808$ con $\chi^2_{7-1,0.05} = \chi^2_{6,0.05} = 12.592$. Se tiene que $9.7808 < 12.592$, con lo que, a un nivel de confianza del 95 % podríamos aceptar que los datos obtenidos se ajustan a la distribución teórica.

En los casos analizados se conocía la distribución teórica, que hemos podido comparar con los resultados de la simulación. Pero en muchos casos reales se desconocerá la distribución teórica, o el cálculo de la misma puede que sea complejo, y una forma eficiente de tener información de la distribución de la variable aleatoria que describe el modelo será la distribución empírica que se deduce de la simulación de Monte Carlo. Podremos considerar válida dicha distribución si al repetir la experiencia muchas veces, observamos cierta robustez en los resultados (frecuencias relativas, medias y varianzas parecidas al realizar varias simulaciones).

5.1. La experiencia en el aula

Durante el curso 2015-16 se han desarrollado sólo las etapas 1, 2 y 3, en clases prácticas en aula informática. Los alumnos han trabajado en grupos de dos personas y han tenido que cumplimentar una ficha donde, tras constatar la equivalencia entre las realidades R1 y R3, secuencialmente, tenían que trasladar los resultados de la experimentación con un dado físico, con un generador de valores aleatorios distribuidos según un uniforme discreta en el intervalo $[1,6]_{\mathbb{Z}}$ basado en una fórmula de Excel y, por último, han hecho ensayos con el generador de valores pseudoaleatorios distribuidos según una uniforme discreta en el intervalo $[1,6]_{\mathbb{Z}}$ de Matlab (`unidrnd(6)`). Por lo tanto, se ha experimentado con tres formas de simular un dado presuntamente perfecto, y los resultados, evaluados mediante el estadístico Chi-cuadrado, fueron satisfactorios en todos los casos: Todos los equipos obtuvieron valores aceptables para las muestra proveniente de experimentar con un dado físico, con Excel y con Matlab, con un nivel de confianza del 95 %.

Como primer aspecto a considerar, podemos destacar el interés mostrado por los alumnos en el desarrollo de la actividad. Lejos de centrarse únicamente en el registro de los resultados, los alumnos se han implicado activamente en el trabajo de simulación, por una parte analizando el script de Matlab y el fichero Excel que se les proporcionó, y por otra haciendo diversos ensayos para observar distintas simulaciones, constatando la utilidad y potencialidad de los métodos de Monte Carlo. Cabe destacar la influencia, claramente positiva, que ha supuesto la realización de la actividad manipulativa de obtener valores de la distribución utilizando un dado físico. No obstante ha sido la constatación de la potencia de los programas informáticos a la hora de simular procesos, generando miles de valores pseudoaleatorios en pocos segundos, lo que más ha removido la curiosidad de los alumnos.

Han sido precisamente las preguntas de los alumnos al realizar la experiencia descrita anteriormente las que han motivado el diseño del REI presentado en este trabajo, lo cual supone completar la experiencia, pero esta vez siguiendo el proceso de modelización descrito en el punto 5 que, siendo sencillo, proporciona al final una distribución no estándar muy interesante por sus características. Como se ha podido constatar, es relativamente fácil describir dicha distribución mediante su correspondiente función de probabilidad teórica, pero el obtener expresiones explícitas para la media y de la varianza teóricas presenta dificultades debido a la complejidad del cálculo aritmético. Este tipo de dificultades son las que justifican que el científico o ingeniero conozca y sea competente en el uso de los métodos de Monte Carlo.

6. Conclusiones

La comprensión del concepto formal de variable aleatoria no supone un proceso mental sencillo para los alumnos. El hecho de manejar una variable que sabemos qué valores puede tomar, pero que no se pueden predecir con exactitud en una realización va asociado a establecer un constructo cognitivo que haga corresponder la frecuencia para los diversos resultados posibles, al repetir muchas veces un experimento, con la probabilidad teórica de que suceda cada uno de dichos resultados. En un enfoque del aprendizaje por competencias, es muy importante que el alumno sepa trasladar el suceso real con componentes aleatorias al terreno formal de las variables aleatorias, tanto en el caso discreto como continuo. Los resultados de la experiencia desarrollada en clase (etapas 1,2,3) indican que, los alumnos son capaces de establecer la relación entre una realidad con componentes aleatorios discretos y un modelo estadístico discreto (variable aleatoria discreta). El aprendizaje efectivo del concepto de variable aleatoria discreta, a partir de la modelización y la simulación de Monte Carlo, proporciona al alumno un apren-

dizaje efectivo del concepto de variable aleatoria y la adquisición de una serie de competencias específicas en base a las cuales será capaz de afrontar problemas más complejos.

Referencias

Blanco Blanco A. (2008).
Una revisión crítica de la investigación sobre las actitudes de los estudiantes universitarios hacia la estadística.
Revista Complutense de Educación, 19(2), 311–330.

Boigues F. J., Estruch V. D., Roig B., Vidal A. (2011).
Un modelo de transmisión de plagas para la enseñanza del álgebra lineal en el contexto de estudios en ciencias ambientales.
Modelling in Science Education and Learning, 4, 5–117.

Brase C.H., Brase C.P. (2016).
Understanding Basic Statistics. Metric Version.
Seventh Edition CENGAGE Learning.

Colaboradores De Wikipedia.
Generador de números aleatorios.
Wikipedia, La enciclopedia libre.
https://es.wikipedia.org/w/index.php?title=Generador_de_n%C3%BAmeros_aleatorios&oldid=88483665

Colaboradores De Wikipedia.
Generador de números pseudoaleatorios.
Wikipedia, La enciclopedia libre.
https://es.wikipedia.org/w/index.php?title=Generador_de_n%C3%BAmeros_pseudoaleatorios&oldid=88579925

Escalante Gómez E. (2008).
Actitudes de alumnos de posgrado hacia la estadística aplicada a la investigación.
Encuentro 2010/ Año XLII, n° 85, 27–38.

Kay S.M. (2006).
Intuitive probability and random processes using MATLAB.
Springer. New York.

Peck, R. (2014).
Statistics. Learning from Data. Preliminary edition.
Brooks/Cole, CENGAGE Learning, Boston.

Modelling in Science Education and Learning
http://polipapers.upv.es/index.php/MSEL