Simulación de variables aleatorias continuas y el teorema del límite central

Kendall Rodríguez Bustos[1] & Greivin Ramírez Arce[2]

Resumen

Se presenta una propuesta de simulación de variables aleatorias continuas y de ejercicios que involucran el desarrollo del teorema del límite central. Para esto se hará uso de Excel con programación básica en Visual Basic que permite desarrollar muestreo repetitivo que emula el comportamiento que siguen las distribuciones. La propuesta pretende que el estudiante universitario inicie desde el diseño de la distribución, haga el desarrollo repetitivo de experimentos, construya su representación gráfica y llegue hasta el cálculo de probabilidades desde el enfoque frecuencial. Las distribuciones continuas a simular son la uniforme, la exponencial, la normal y su relación de cada una de estas con el teorema del límite central.
El marco en que el que se sustenta la propuesta es a través del Conocimiento Tecnológico Pedagógico del Contenido (TPACK), presentado por Koehler y Mishra (2006) en el que interesa mostrar: a) las distintas representaciones que se pueden obtener con Excel, b) las técnicas pedagógicas que la programación aporta en forma constructiva para adquirir el concepto teorema del límite central, c) el conocimiento sobre qué hace fácil o difícil la comprensión del concepto y cómo la tecnología puede aportar al desarrollo del conocimiento.

Abstract

A proposal of simulation of continuous random variables and exercises involving the development of the Central Limit Theorem is presented. For this it is going to be used Excel with basic programming in Visual Basic, that allows to develop repetitive sampling which emulates the behavior that distributions follow. The proposal is intended for the university student to start from the layout design, make the development of repetitive experiments, build his/her plot and achieve to calculate probabilities from the frequency approach. The continuous distributions to be simulated are the uniform, the exponential, the normal and the relationship of each of these with the Central Limit Theorem.
The framework in which the proposal is based is through the Pedagogical Technological Content Knowledge (TPACK), presented by Koehler and Mishra (2006) which wants to show: a) the different representations that can be obtained with Excel, b) the pedagogical techniques that programming contributes constructively to acquire the Central Limit Theorem concept, c)the knowledge about what makes easy or difficult to understand the concept and how technology can contribute to the knowledge development.

Palabras claves: simulación, variables aleatorias continuas, teorema del límite central, conocimiento tecnológico pedagógico del contenido

Keywords: simulation, central limit theorem, technological pedagogical content knowledge

Modalidad: ponencia

El interés de la presente propuesta es incorporar la programación básica en Visual Basic, como complemento de Excel, para simular el comportamiento de variables aleatorias

[1]Instituto Tecnológico de Costa Rica, kendall2412@gmail.com
[2]Instituto Tecnológico de Costa Rica, gramirez@itcr.ac.cr

continuas tales como la uniforme, exponencial, normal y la relación de estas con el teorema del límite central.

Lo anterior es relevante considerando que los estudiantes universitarios ingresan con deficiencias en contenidos y habilidades de razonamiento estocástico, aunado a ello, los cursos que toman en su educación superior, aparte de ser pocos, muchas veces son desarrollados con metodologías clásicas reducidos a formalismos y a procesos rutinarios de aplicación de fórmulas; sin conocer el proceso de construcción de las distribuciones en los que carecen del desarrollo repetitivo de experimentos (Inzunsa, 2006; Ramírez, 2007).

Múltiples investigadores concuerdan que para modelar situaciones reales en las cuales hay presencia del azar, lo aleatorio o la incertidumbre, es importante el desarrollo del pensamiento o razonamiento probabilístico y buscar métodos que permitan de manera razonable emular estos sucesos o fenómenos aleatorios. (Alvarado; Batanero, 2008; Sánchez, 2009; Inzunsa; Guzmán, 2011; Jaimes; Yáñez, 2013; Burbano; Pinto; Valdivieso, 2015)

No obstante, los conocimientos matemáticos no son suficientes para que los profesores logren enseñar probabilidad de una manera correcta y fomentar en sus alumnos un adecuado razonamiento probabilístico; pues, desde el enfoque clásico no permite la debida comprensión de probabilidad ni desarrollar intuiciones adecuadas, y de esta forma no generan una construcción significativa en los estudiantes acerca de los conceptos asociados a experimentos aleatorios. (Batanero; Contreras; Díaz; Roa, 2012; Jaimes; Yáñez, 2013)

Jaimes y Yáñez (2013) consideran que muchos docentes todavía tienen preferencia por el enfoque clásico de la probabilidad; donde está más asociado a la visión determinista de las matemáticas y desconfían de la aproximación de los resultados que se obtienen al realizar o simular una serie de ensayos de un experimento aleatorio.

Por ello, se plantea el uso de la simulación computacional como un recurso didáctico que puede lograr despertar el interés e incrementar la motivación de los discentes para el aprendizaje de temas relacionadas con probabilidad. En particular, en el caso de la simulación de variables aleatorias continuas y el teorema del límite central, se aconseja realizarlo después de aprender los aspectos teóricos de los modelos de probabilidad continuos en la etapa de las operaciones formales. (Burbano; Pinto; Valdivieso; 2015).

Marco teórico

Interesa del Conocimiento Tecnológico Pedagógico del Contenido (TPACK, figura 1) según la propuesta realizada, las siguientes intersecciones:

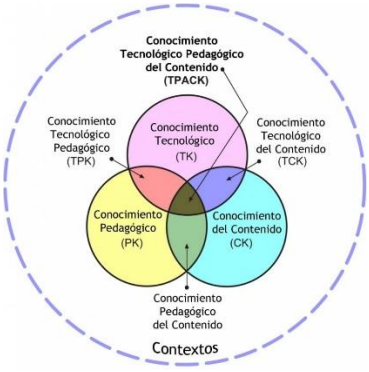


Figura 1. Modelo TPACK por Koehler y Mishra (2006)

Conocimiento pedagógico del contenido

1. *Conocimiento que permite comprender cómo se debe organizar y adaptar un contenido para ser enseñado.*

Inzunsa y Guzmán (2011) consideran que el éxito de un currículo de probabilidad para lograr fomentar el razonamiento probabilístico de los estudiantes depende, en gran medida, de la comprensión de los profesores acerca de la probabilidad, complementado con el conocimiento de las concepciones erróneas de los estudiantes y el uso de representaciones y herramientas.

Más aún, actualmente existe una tendencia por parte de los investigadores de la estocástica que es plantear propuestas del proceso de enseñanza – aprendizaje mediante el significado frecuencial de la probabilidad. De hecho, Batanero, Contreras y Gómez (2014) se centran en las siguientes características del significado frecuencial:

- Posibilidad de estimar una probabilidad teórica a partir de datos de frecuencias.
- Comprensión de las características de resultados aleatorios y de la convergencia.

2. *Forma en que representan y formulan los conceptos de la disciplina, técnicas pedagógicas, que hace que los conceptos sean fáciles o difíciles de aprender.*

Diversos investigadores recomiendan ampliamente hacer uso de simulación y experimentación en la enseñanza de la probabilidad, donde la idea de realizar experimentos es la observación de la estabilidad, que alrededor de la probabilidad empírica adquiere las frecuencias relativas asociado a un suceso aleatorio cuando se aumenta el número de repeticiones, tal como se plantea la Ley de los Grandes Números, y esto constituye el fundamento del enfoque frecuencial. (Inzunsa; Guzmán, 2011; Jaimes; Yáñez, 2013)

Además, el desarrollo de la simulación se sugiere primero, de manera física, con el fin de analizar el comportamiento de resultados aislados, más tarde, el proceso repetitivo de toma de muestras deberá permitir el surgimiento de patrones que lleven a deducciones e inferencias sobre el comportamiento de las distribuciones (Ramírez, 2013).

Alvarado y Batanero (2008, p.8) plantean esta necesidad de repetición con el concepto: "El teorema central del límite, uno de los fundamentos en estadística, estudia el comportamiento de la suma de variables aleatorias, cuando crece el número de sumandos, asegurando su convergencia hacia una distribución normal en condiciones muy generales"

Diversas investigaciones asociadas a la enseñanza del teorema del límite central consideran que hay una complejidad en la comprensión de su significado presentado en los libros de estadística aplicada a la ingeniería y se hallan una variedad de enfoques y aproximaciones. (Alvarado; Batanero, 2008).

Estas deficiencias en la comprensión del concepto de distribución y del teorema del límite central generan otras limitaciones; particularmente en la realización e interpretación en la estadística inferencial: intervalos de confianza y contrastes de hipótesis.

Es por esto, que el proceso repetitivo de toma de muestras a través de la generación de ciclos con la programación, debe buscar que el estudiante acorte la brecha entre el duro

proceso imaginativo que sugiere las hipótesis del teorema del límite central y la manipulación de los estimadores de cada muestra que se consiguen al imprimir sus resultados en la hoja de Excel desde Visual Basic.

3. *Estrategias, conocimientos previos, errores conceptuales y metodológicos más frecuentes de los estudiantes*

Estudios evidencian la existencia de concepciones erróneas y sesgos probabilísticos por parte de las personas adultas y estudiantes universitarios, de tal forma, que sin una preparación correcta sobre los futuros profesores y profesores en servicio pueden mostrar razonamientos y sesgos similares en estos estudios. (Inzunsa; Guzmán, 2011)

Sánchez (2009) considera que hay tres principales sesgos probabilísticos durante el proceso de aprendizaje de los estudiantes, esto son: el sesgo de equiprobabilidad, el sesgo de la atención y la representatividad; esto como producto de sus estudios acerca de los errores y dificultades de los alumnos en la resolución de problemas de probabilidad a nivel de secundaria en México.

De esta manera, una enseñanza basada en el uso de la simulación física como computacional, y la reflexión en pequeños grupos sobre estas dificultades pueden ayudar a superar estos sesgos. (Batanero; Contreras; Díaz; Roa, 2012; Batanero; Contreras; Gómez, 2014).

Alvarado y Batanero (2008) consideran que los elementos de significado relacionados con la distribución normal son necesarios para una debida comprensión del teorema de límite central, puesto que en éste es indispensable el uso de esta distribución. Además, en el caso de la enseñanza por simulación, mencionan que el uso de la tecnología por sí sola no es suficiente para la comprensión del teorema, sino que son las actividades de tipo constructiva favorecen el aprendizaje.

Así, en esta propuesta, se pretende que el estudiante vaya generando la representación gráfica instantánea que permite analizar de manera dinámica el comportamiento de la distribución, y su convergencia o no, al aumentar el número de muestras obtenidas mediante la programación básica.

Conocimiento tecnológico del contenido

1. *Involucra todas las formas en que la tecnología limita o facilita la representación, explicación o demostración de conceptos y métodos propios de la disciplina.*

Al realizar experimentos físicos ayuda a generar una mejor comprensión alrededor del experimento aleatorio en cuestiones como la identificación del espacio muestral y en la relación ordenada entre posibilidades a priori y resultados a posteriori. Sin embargo, dadas las pocas repeticiones que se realizan mediante la simulación física; entonces es difícil que los alumnos logren captar ciertas regularidades en el comportamiento del suceso aleatorio que permita dar algún significado a su experiencia y generar conceptos claro de la probabilidad de un evento aleatorio. De esta manera, ante las pocas repeticiones y a la toma de mayor número de muestras se facilita el uso de la tecnología; en particular, las simulaciones computacionales. (Jaimes; Yáñez, 2013)

En particular, en el caso del teorema del límite central, aunque una demostración matemática establece la veracidad del teorema, quizás esto no contribuya mucho a la idea

intuitiva del resultado. De esta forma, Alvarado y Batanero (2008) proponen a partir de una población sencilla realizar una simulación con lápiz y papel de la elección de la muestra aleatoria de distintos tamaños de la población y luego utilizar simulación utilizando tecnología para establecer una mejor comprensión del teorema.

Desde luego, el docente debe tener bien claro el alcance de las herramientas computacionales para la enseñanza de la estocástica, pues Arnaldos y Faura (2012) advierten el riesgo del uso de las simulaciones que supone presentarlos a los estudiantes del siguiente sentido: mostrar dónde se encuentran, qué proporcionan y cómo se usan; y dejar a su libertad el uso de las mismas como parte de los materiales a disponibilidad.

En el caso particular de la propuesta, el utilizar únicamente Excel (sin el complemento de Visual Basic), Fathom, Statistica o Geogebra hará que el proceso repetitivo de toma de muestras sea largo, el proceso de construcción de la distribución sea tedioso, la acumulación de los estimadores sea ineficiente, o simplemente se requiera de una programación elevada. Caso contrario a través de Visual Basic como complemento de Excel, que facilita todas las condiciones antes mencionadas.

2. *Qué tecnología son las mejores para enseñar un tema determinado y cómo utilizarlas de forma efectiva para abordarlo.*

En cierto sentido, la simulación es una representación que sustituye su experimento estocástico por otro y su empleo es de gran utilidad en la enseñanza de conceptos en el campo de las distribuciones en el muestreo. Además, se proponen ejercicios para generar muestras aleatorias de observaciones de distintas distribuciones de probabilidad usando los paquetes estadísticos como: Minitab, SAS o plantilla electrónica Excel. (Alvarado; Batanero, 2008)

Se concuerda en nuestra propuesta con Alvarado y Batanero (2008) en los que consideran que, para la enseñanza del teorema de límite central se puede utilizar una simulación gráfica en la computadora como una forma de ilustrar y argumentar esta proposición mediante el aumento progresivo del tamaño de la muestra. Tal como el caso de la aproximación de la distribución normal a la binomial para distintos valores de los parámetros n y p.

3. *De qué modo el contenido disciplinar es transformado por la aplicación de una tecnología.*

Jaimes y Yáñez (2013) afirman el uso correcto de una herramienta computacional, que permite la generación de diversos experimentos aleatorios en gran número de pruebas, puede ayudar a encontrar una aproximación y relacionarla con la probabilidad teórica y el espacio muestral para generar un concepto significativo.

Arnaldos y Faura (2012) mencionan las siguientes aplicaciones de las simulaciones usando tecnología en los aspectos de variables aleatorias y modelos de variables aleatorias; que respalda el trabajo realizado en este documento:

- Variables aleatorias: comprensión de los tipos de variables aleatorias, interpretación de sus principales características, funciones que describen su comportamiento aleatorio.

- Modelos de variables aleatorias: comprensión de la naturaleza y efecto de las variaciones paramétricas, mejorar la capacidad de distinción entre los tipos de modelos.

En la propuesta se utilizan los conceptos previos de biyectividad de funciones (aplicado a distribuciones según Sanabria, 2013), cálculo de inversas y el comportamiento de la función lineal; para generar a partir de un número aleatorio entre 0 y 1, las distribuciones de variables aleatorias continuas: uniforme, la exponencial y la normal. Además, a través de la generación de ciclos se transforma el problema en el proceso de muestreo repetitivo que cumple con las hipótesis del teorema del límite central.

Conocimiento tecnológico pedagógico

1. *Conocimiento de las características y el potencial de las múltiples tecnologías disponibles utilizadas en contextos de enseñanza aprendizaje. E inversamente, conocimiento sobre cómo la enseñanza y aprendizaje se modifican al utilizar la tecnología en particular.*

En el estudio realizado por Jaimes y Yáñez (2013) evidencian que la simulación computacional en probabilidad permite superar algunos de los sesgos o concepciones erróneas que los estudiantes poseen acerca de las secuencias aleatorias o sobre el valor de las probabilidades en experimentos compuestos; como el sesgo de los valores recientes, sesgo del desorden, sesgo de equiprobabilidad y la concepción de la variación constante de las frecuencias relativas.

Diversos investigadores mencionan que la promoción de las actuales tecnologías en la enseñanza de la matemática permite hacer simulaciones y a través de estas se construye un puente entre las ideas intuitivas del discente y los conceptos formales; permitiendo verbalizar sus pensamientos y apropiarse del razonamiento probabilístico. (Castro; Rodríguez, 2005; De Oliveira; Espasandin; 2011)

La programación básica en Visual Basic como complemento de Excel puede permitir que los estudiantes expresen sus interpretaciones, comparaciones y conjeturas (funciones psicológicas de nivel superior, según Feuerstein en Kozulin, 2000), al tener que construir ciclos anidados, condicionales, y actualización de variables.

Conocimiento pedagógico tecnológico del contenido

La triple intersección evidencia, según Mishra y Koehler (2006), que en nuestra propuesta el uso de simulación es fundamental debido a que:

> La tecnología puede jugar un papel esencial en la forma de representar, ilustrar, ejemplificar y demostrar las ideas y conceptos de una disciplina. Supone el desarrollo de una mente abierta y creativa para poder adaptar las herramientas que existen, que no siempre fueron creadas para fines educativos y reconfigurarlas. Siendo así, TPACK requiere de la comprensión de:
>
> - la representación de ideas utilizando la tecnología.
> - Técnicas pedagógicas que utilizan la tecnología en formas constructivas para enseñar un contenido.

- conocimiento sobre qué hace fácil o difícil la comprensión de un concepto y cómo la tecnología puede contribuir a compensar esas dificultades que enfrentan los alumnos.
- conocimiento de las ideas e hipótesis previas de los alumnos y sobre cómo la tecnología puede ser utilizada para construir conocimiento disciplinar.

La integración de la tecnología a la enseñanza de un contenido disciplinar requiere del desarrollo de una sensibilidad que atienda a la relación dinámica y transaccional entre componentes.

La programación de la simulación de estas actividades debería permitir a los estudiantes compartir su estrategia de simulación como una red de intercambio de salida de información. Aquí el profesor debe ser un guía que organice y dirija el conocimiento brindado por las redes, sabiendo destacar la información innecesaria en la programación realizada. Como utiliza Siemens (2004) el conectivismo:

> La integración de principios explorados por las teorías del caos, redes, complejidad y auto-organización. El aprendizaje es un proceso que ocurre al interior de ambientes difusos de elementos centrales cambiantes que no están por completo bajo control del individuo. El aprendizaje (definido como conocimiento aplicable) puede residir fuera de nosotros (al interior de una organización o una base de datos), está enfocada en conectar conjuntos de información especializada, y las conexiones que nos permiten aprender más tienen mayor importancia que nuestro estado actual de conocimiento.

Aspectos teóricos

Los siguientes conceptos son deseables que el estudiante conozca como contexto en el desarrollo de las actividades que se proponen:

Enfoque frecuencia de la probabilidad

Dado un evento asociado a una experiencia aleatoria, la *probabilidad frecuencial* es la frecuencia relativa observada con que ocurre el evento al repetirse la experiencia varias veces. Es decir, si A es un evento, entonces la probabilidad frecuencial corresponde a:

$$P(A) = \frac{núm\ de\ experimentos\ donde\ el\ evento\ A\ ocurre}{núm\ total\ de\ experimentos\ realizados}$$

Ley de los Grandes Números

Dado un experimento, y sea A un evento. Si el experimento se repite un número suficientemente grande de veces, entonces la probabilidad frecuencial de A será muy cercana al valor real de la probabilidad.

Variable aleatoria

Es una función de un espacio muestral Ω asociado a los números reales, es decir, una regla que asigna un único valor real a cada evento del espacio muestral.

Variable aleatoria continua

Se dice que *X* es una variable aleatoria continua si y sólo si su rango es continuo.

Distribución de probabilidad de una variable aleatoria continua

Distribución uniforme: Se dice que *X* sigue una distribución uniforma en el intervalo $[a, b]$ si su función de densidad está dada por:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & si \quad a \leq x \leq b \\ 0 & si \quad en\ otro\ caso \end{cases}$$

y se denota $X \sim U[a, b]$.

La función de distribución acumulada corresponde:

$$F_X(x) = \begin{cases} 0 & si \quad x > a \\ \frac{x-a}{b-a} & si \quad a \leq x \leq b \\ 1 & si \quad x > b \end{cases}$$

Por otro lado, si se quiere obtener valores aleatorios de una variable

$$Y \sim U[a, b]$$

Note que si v es un número aleatorio entre 0 y 1 que cumple que $v = F_Y(y)$, entonces

$$v = \frac{y-a}{b-a} \Longrightarrow y = v(a-b) + a$$

Distribución exponencial: Se dice que *X* sigue una distribución exponencial si su distribución de probabilidad está dada por:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & si \quad x \geq 0 \\ 0 & si \quad en\ otro\ caso \end{cases}$$

donde λ es una constante positiva. Se denota $X \sim Exp(\lambda)$.

La función de distribución acumulada corresponde:

$$F_X(x) = \begin{cases} 0 & si \quad x < 0 \\ 1 - e^{-\lambda x} & si \quad x \geq 0 \end{cases}$$

Si se quiere obtener valores aleatorios de una variable $X \sim Exp(\lambda)$, note que si v es un número aleatorio entre 0 y 1 que cumple $v = F_X(x)$, entonces

$$v = 1 - e^{-\lambda x} \Longrightarrow x = \frac{-\ln(1-v)}{\lambda} = \frac{-\ln w}{\lambda}$$

con $w = 1 - v \in]0,1]$.

Distribución normal: Se dice que *X* sigue una distribución normal si su función de probabilidad está dada por

$$f_X(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

Donde μ y σ son constantes. Se denota $X \sim N(\mu, \sigma^2)$.

Para obtener valores aleatorios de una variable $X \sim N(\mu, \sigma^2)$, se presenta el problema de que no se cuenta con una fórmula explícita de F_X. Sin embargo, Excel tiene predeterminada de forma numérica la función inversa de F_X mediante el comando =DISTR.NORM.INV(v;μ;σ).

Distribución normal estándar: Se dice que *Z* sigue una distribución normal estándar si Z sigue una distribución normal con media cero y varianza uno, es decir, $Z \sim N(0,1)$.

Teorema del Límite Central: Sean $X_1, X_2, X_3, \ldots, X_n$ variables aleatorias mutuamente independientes que siguen una misma distribución, tales que

$$E(X_i) = \mu \quad \text{y} \quad Var(X_i) = \sigma^2, \text{ para } i = 1,2,3, \ldots, n$$

- Considere la variable suma

$$S_n = X_1 + X_2 + X_3 + \cdots + X_n$$

Se tiene que:

- $E(S_n) = n\mu$
- $Var(S_n) = n\sigma^2$
- Cuando $n \longrightarrow \infty$ se tiene que $\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \sim N(0,1)$

- Considere la variable aleatoria promedio

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

Se tiene que:

- $E(\bar{X}) = \mu$
- $Var(\bar{X}) = \frac{\sigma^2}{n}$
- Cuando $n \longrightarrow \infty$ se tiene que $\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$

Actividades

Actividad 1

1. *Una batería funciona en un tiempo exponencial con promedio de 5 horas. En un lote de 20 baterías, determine la probabilidad de que entre 7 y 12 tarden más de 6 horas.*

Simulación computacional en Excel

Considere la variable *X* como la duración de una batería en horas. Dado que *X* sigue una distribución exponencial con media $E(X) = 5$, donde $E(X) = \frac{1}{\lambda}$, entonces se tiene que $\lambda = \frac{1}{5}$. De esta forma

$$X \sim Exp\left(\frac{1}{5}\right)$$

Luego, sabiendo que la función acumulada de la distribución exponencial asociada a la variable aleatoria continua *X* está dada por:

$$F_X(x) = 1 - e^{-\lambda x} \Longrightarrow F_X(x) = 1 - e^{-\frac{x}{5}}, \text{ con } x \geq 0$$

Observe que si v es un número aleatorio entre 0 y 1 que satisface que $v = F_X(x)$ sea una función biyectiva, entonces

$$v = 1 - e^{-\frac{x}{5}} \Longrightarrow x = -5\ ln(1 - v) \text{ (sugerido por Sanabria, 2013)}$$

$$\Longrightarrow x = -5\ ln\, w, \text{ donde } w = 1 - v$$

De esta manera, para generar valores aleatorios de la variable *X* se utiliza la fórmula:

$$x = -5\ ln\, w, \text{ con } w \in\ ]0,1]$$

Esta transformación del conocimiento disciplinar, hace que el conocimiento pedagógico del estudiante se vea alterado a partir de la construcción de la distribución.

Con los datos anteriores, se plantea la simulación computacional:

1. Se define la variable "Valor aleatorio" en la celda A1. En la siguiente celda A2 se escribe el comando: =ALEATORIO() para generar valores aleatorios en el intervalo [0, 1]. Dado que se desean realizar 1000 valores aleatorios entonces se arrastra la fórmula de la celda A2 hasta la celda A1001.
2. Se defina la variable "*X*" en la celda B1 que corresponde a un valor aleatorio de la variable *X* distribuida de manera exponencial. Para ello, en la celda B2 se escribe el siguiente comando: $= -5 * LN(A2)$ relacionado con la formula $x = -5\ ln\, w$, donde la variable *w* corresponde a los valores aleatorios de la columna anterior en la hoja de cálculo.

   Luego se arrastra dicha fórmula hasta la celda B1001 para obtener 1000 valores aleatorios de la variable *X*; considerando este valor como una buena cantidad de experimentos realizados para mostrar la tendencia de la distribución (según Inzunsa; Guzmán, 2011).

   Por ejemplo, en la siguiente imagen en la celda B2 significa que la primera batería tuvo una duración superior a 6 horas (aprox 7.6126 horas).
3. Como primera parte de esta simulación interesa determinar la probabilidad de que una batería tarde más de 6 horas, es decir, $P(X > 6)$. De esta forma, se debe contabilizar los valores de la segunda columna que satisfacen la condición de que $X > 6$ y luego dividir dicha suma entre 1000 para hallar la probabilidad.

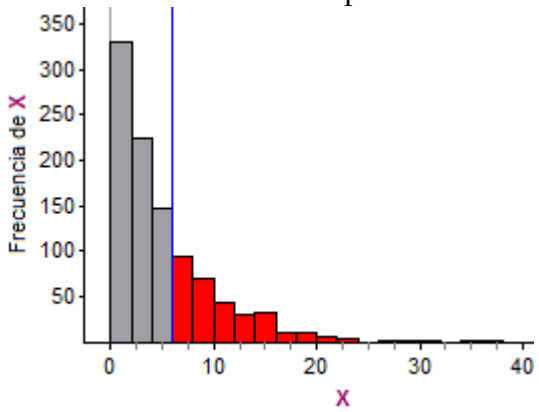
   Se define la variable "*Probabilidad de que tarde más de 6 horas*" en alguna celda cualquiera de la columna D y en la siguiente celda se escribe el comando: $= CONTAR.SI(B2{:}B1001;"> 6")/1000$

	A	B	C	D
1	**Valor aleatorio**	**X**		
2	0.218161504	7.612598243		
3	0.440567445	4.098458671		Probabilidad de que tarde más de 6 horas
4	0.965041132	0.177922775		0.300
1000	0.353359429	5.201347643		
1001	0.430509012	4.213935104		

Figura 2. Simulación computacional en Excel: Batería

La actividad permite variar el conocimiento tecnológico a partir de la construcción gráfica de la distribución, así como señalar la probabilidad solicitada; logrando una interpretación visual del cálculo solicitado.

Gráfico 1. Distribución exponencial del tiempo de duración de la batería.



4. Ahora bien, se desea hallar la probabilidad de que en un lote de 20 baterías, entre 7 y 12 baterías tarden más de 6 horas.

   Para ello, se plantea la siguiente estrategia: se define la variable "Núm Baterias i", con $i \in \{1, 2, \ldots, 100\}$ que corresponde a los 100 experimentos a realizar. Donde a partir de la cuarta columna (D1) hasta la centésima tercera columna (CY) se escribe en la primera celda: Núm Bacterías i (Ver figura 3).

   En la siguiente celda correspondiente a cada columna se escribe el comando: =ALEATORIO(), que asigna un valor aleatorio entre 0 y 1. Ahora, dado que en el enunciado del problema nos indica que se cuenta con un lote de 20 baterías, entonces en cada experimento se debe contar con una muestra de 20 baterías.
   Por lo que, en la cuarta columna se arrastra la celda D1 hasta la celda D21. De igual forma, se realiza con las demás 99 columnas restantes. Es decir, en total tendrán 2000 datos aleatorios que corresponde a 20 baterías por experimento (en este caso, se realizan 100 experimentos).

5. De esta manera se realizan 20 extracciones de baterías en cada experimento. Se considera un experimento exitoso, si se extraen entre 8 y 11 baterías que duran más de 6 horas. Así, en la celda D22 se escribe el siguiente comando:

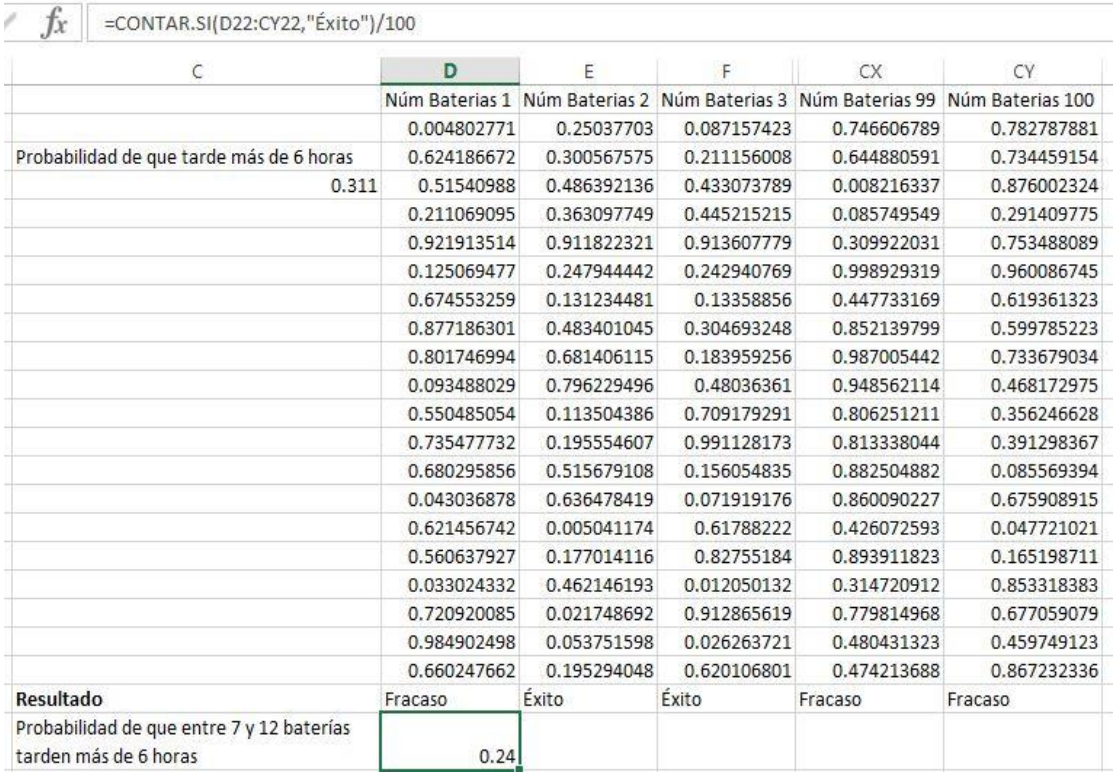
```
=SI(Y(CONTAR.SI(D1:D20;"<"&$C$4)<=11;CONTAR.SI(D1:D20;"<"&
$C$4)>=8)=VERDADERO;"Éxito";"Fracaso")
```

donde retorna como posibles resultados: Éxito o Fracaso y además el valor de la celda C4 corresponde a la probabilidad de que una batería tarde más de 6 horas.

Observe que si se considera la variable Y como el número de baterías de un lote de 20 que duren más de 6 horas, entonces lo solicitado en este problema es $P(7 < y < 12)$ que es equivalente a calcular $P(8 \leq y \leq 11)$

Luego, esto mismo se realiza con las demás columnas y así en cada experimento obtener un resultado (éxito o fracaso) que nos permita distinguir aquellos procesos donde cumplen o no las condiciones del problema dado. Se desea contabilizar la cantidad de éxitos que resultan de los 100 experimentos realizados y luego hacer el cálculo de la probabilidad respectiva.

Para esto se define la variable "Probabilidad de que entre 7 y 12 baterías tarden más de 6 horas" y en una celda posterior se escribe el comando: $= CONTAR.SI\ (D22{:}CY22;"Exito")/100$ que retorna la probabilidad buscada.



fx =CONTAR.SI(D22:CY22,"Éxito")/100

C	D	E	F	CX	CY
	Núm Baterias 1	Núm Baterias 2	Núm Baterias 3	Núm Baterias 99	Núm Baterias 100
	0.004802771	0.25037703	0.087157423	0.746606789	0.782787881
Probabilidad de que tarde más de 6 horas	0.624186672	0.300567575	0.211156008	0.644880591	0.734459154
0.311	0.51540988	0.486392136	0.433073789	0.008216337	0.876002324
	0.211069095	0.363097749	0.445215215	0.085749549	0.291409775
	0.921913514	0.911822321	0.913607779	0.309922031	0.753488089
	0.125069477	0.247944442	0.242940769	0.998929319	0.960086745
	0.674553259	0.131234481	0.13358856	0.447733169	0.619361323
	0.877186301	0.483401045	0.304693248	0.852139799	0.599785223
	0.801746994	0.681406115	0.183959256	0.987005442	0.733679034
	0.093488029	0.796229496	0.48036361	0.948562114	0.468172975
	0.550485054	0.113504386	0.709179291	0.806251211	0.356246628
	0.735477732	0.195554607	0.991128173	0.813338044	0.391298367
	0.680295856	0.515679108	0.156054835	0.882504882	0.085569394
	0.043036878	0.636478419	0.071919176	0.860090227	0.675908915
	0.621456742	0.005041174	0.61788222	0.426072593	0.047721021
	0.560637927	0.177014116	0.82755184	0.893911823	0.165198711
	0.033024332	0.462146193	0.012050132	0.314720912	0.853318383
	0.720920085	0.021748692	0.912865619	0.779814968	0.677059079
	0.984902498	0.053751598	0.026263721	0.480431323	0.459749123
	0.660247662	0.195294048	0.620106801	0.474213688	0.867232336
Resultado	Fracaso	Éxito	Éxito	Fracaso	Fracaso
Probabilidad de que entre 7 y 12 baterías tarden más de 6 horas	0.24				

Figura 3. Simulación computacional en Excel: Batería

Con este proceso repetitivo de experimentos y el dinamismo de Excel, el estudiante debe ser capaz de analizar la variabilidad de los datos obtenidos. Además, al comparan resultados entre pares, debe despertar la curiosidad por ver que cada uno obtendrá resultados diferentes pero bastante aproximados entre ellos.

Solución teórica

Dado que *X* es la duración de una batería en horas, donde $E(X) = \frac{1}{\lambda} = 5 \Longrightarrow \lambda = \frac{1}{5}$, entonces

$$X \sim Exp\left(\frac{1}{5}\right)$$

De esta forma:

$$\begin{aligned} P(X > 6) &= 1 - P(X \leq 6) \\ &= 1 - F_X(6) \\ &= 1 - \left(1 - e^{\frac{-1}{5}\cdot 6}\right) \\ &= e^{\frac{-6}{5}} \approx 0.3012 \end{aligned}$$

Considere la variable *Y* como el número de baterías que tardan más de seis horas de las 20 baterías, donde *Y* sigue una distribución binomial:

$$Y \sim B\left(20, e^{\frac{-6}{5}}\right)$$

De esta forma:

$$\begin{aligned} P(7 < Y < 12) &= P(8 \leq Y \leq 11) \\ &= \sum_{k=8}^{11} C(20,k) \cdot \left(e^{\frac{-6}{5}}\right)^k \left(1 - e^{\frac{-6}{5}}\right)^{20-k} \\ &= 0.226606 \end{aligned}$$

Por lo tanto, la probabilidad de que entre 7 y 12 baterías duren más de 6 horas es 0.226606.

Actividad 2

2. *El tiempo que tarda un empacador en empacar una caja de bananos sigue una distribución exponencial con media 90 segundos. La empresa bananera "Costa Rican Bananas" ha decidido despedir a aquellos que en una inspección sorpresa tarden más de 2 minutos.*
    a. *Determine la probabilidad de que un empacador sea despedido.*
    b. *La empresa "Costa Rican Bananas" cuenta con diversas empacadoras en todo el país, y cada una contiene 35 empleados empacadores. La empresa ha decidido además cerrar empacadoras, en las que el promedio por empacar una caja por empacadora sea mayor a 100 segundos. Determine la probabilidad de que una empacadora sea cerrada.*

Simulación computacional en Excel

Parte (a)

Considere la variable *X* como el tiempo que tarda un empacador en empacar una caja de bananos en segundos. Dado que *X* sigue una distribución exponencial con media

$E(X) = \frac{1}{\lambda} = 90 \Longrightarrow \lambda = \frac{1}{90}$. De esta forma: $X \sim Exp\left(\frac{1}{90}\right)$

Con los datos anteriores se puede plantear una posible simulación computacional de la siguiente manera:

1. Se define la variable "Valor aleatorio" en la celda A1. En la siguiente celda A2 se escribe el comando: =ALEATORIO() para generar valores aleatorios en el intervalo [0, 1]. Dado que se desean realizar 1000 valores aleatorios entonces se arrastra la celda A2 hasta la celda A1001.
2. Se define la variable "*X*" en la celda B1 correspondiente a la variable *X* distribuida de manera exponencial. Para ello se generan valores aleatorios mediante la siguiente formula:

   $$x = -90\ ln\ w, \text{ con } w \in]0,1]$$

   Para este caso, la variable *w* corresponde a los valores aleatorios de la columna anterior en la hoja de cálculo. Así, en la celda B2 se escribe el comando:
   $= -90 * LN(A2)$ Luego, se arrastra dicha fórmula hasta la celda B1001 para obtener 1000 valores aleatorios asociado a la variable *X*.
3. En este caso se debe determinar la probabilidad de que $X > 120$ (pues 2 minutos es equivalente a 120 segundos). De esta forma se debe contabilizar los valores generados de la segunda columna que satisfacen la condición de que $X > 120$ y luego dividir esa suma entre 1000 (cantidad de experimentos realizados) para hallar la probabilidad solicitada.
   De esta manera se define la variable "Probabilidad de ser despedido" en alguna celda de la columna C y en la siguiente celda se escribe el comando:
   $= CONTAR.SI(B2:B1001," > 120")/1000$

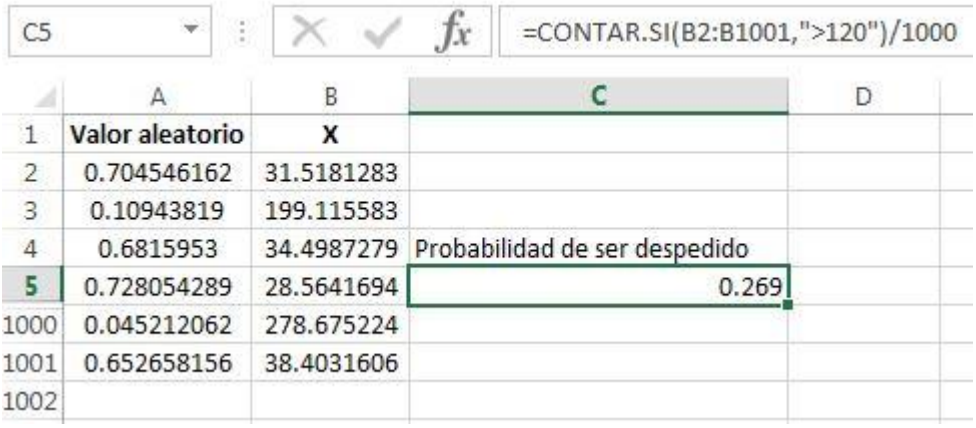


Figura 4. Simulación computacional: Costa Rica Bananas (Parte (a))

Parte (b)

Teníamos que X es el tiempo por empacador al empacar una caja de bananos por segundos; de esta manera, se define la variable $\bar{X}$ como el tiempo promedio que dura una empacadora (35 empleados) en empacar una caja en segundos. Es decir:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{34} + X_{35}}{35}$$

Como $X \sim Exp\left(\frac{1}{90}\right)$ entonces se deduce que la media $E(X) = \frac{1}{\lambda} = 90$ y la desviación estándar $\sigma = \frac{1}{\lambda} = 90$. De esta forma, para generar X_i valores aleatorios asociados a la variable X con $i = \{1,2,3, \ldots ,34,35\}$ se utiliza la siguiente fórmula: $x = -90\ ln\ w$, con $w \in\ ]0,1]$.

Se puede realizar una posible simulación computacional de la siguiente manera:

El programa principal se llama ***PromedioExponencial***, en ella se escriben las funciones que se utilizará en la respectiva simulación.

1. Se define la variable *i* correspondiente a un valor aleatorio *X* distribuida de manera exponencial. A dicha variable le corresponde el siguiente comando: $-90 * Log(Rnd())$, donde $Rnd()$ es un comando especial de Visual Basic para generar valores aleatorios en el intervalo [0, 1].
2. Se define la variable *s* que correspondiente a la suma de *n* valores X_i que sigue una distribución normal, donde $i = \{1,2, \ldots, n\}$. Para determinar el promedio de cada muestra se realiza la instrucción $s/35$.
3. Se define la variable *fila* que corresponde a un contador del primer ciclo que controla la cantidad de valores a ejecutar en la hoja de cálculo. Para este caso se van generar 1000 valores aleatorios desde la celda A2 hasta la celda A1001 correspondientes a 1000 experimentos de esta simulación.
4. Se define la variable *n* asociada al tamaño de la muestra de cada experimento, en este caso *n* = 35. Este contador del segundo ciclo nos permite sumar *n* valores donde su resultado se guardará en la variable *s*.
5. Se desea hallar la probabilidad de que el promedio tarde más de 100 segundos, es decir, $P(\bar{X} > 100)$ entonces se define la variable *c* que se le asocia el comando: CountIf(Range("B1:B1001"),">100") que permite contabilizar la cantidad de valores que satisfacen la condición buscada.
6. El comando Cells(i,j) nos permite imprimir datos en la hoja de cálculo en la fila *i* y la columna *j*.

Ahora bien, el código completo del programa diseñado en Visual Basic directamente en las macros de Excel se muestra a continuación:

```
Option Explicit

Sub PromedioExponencial()

Dim i As Single
Dim s As Single
Dim fila As Integer
Dim n As Integer
Dim c As Single
s = 0
fila = 2

Do While fila <= 1001
    n = 1
    s = 0
    Do While n <= 35
        i = -90 * Log(Rnd())
        s = s + i
        n = n + 1
    Loop

    Cells(fila, 2) = s / 35
    fila = fila + 1
Loop

Cells(1, 1) = "Promedio Valores"
Cells(1, 2) = "Probabilidad"
c = Application.WorksheetFunction.CountIf(Range("A2:A1001"),">100")
Cells(2, 2) = c / 1000

End Sub
```

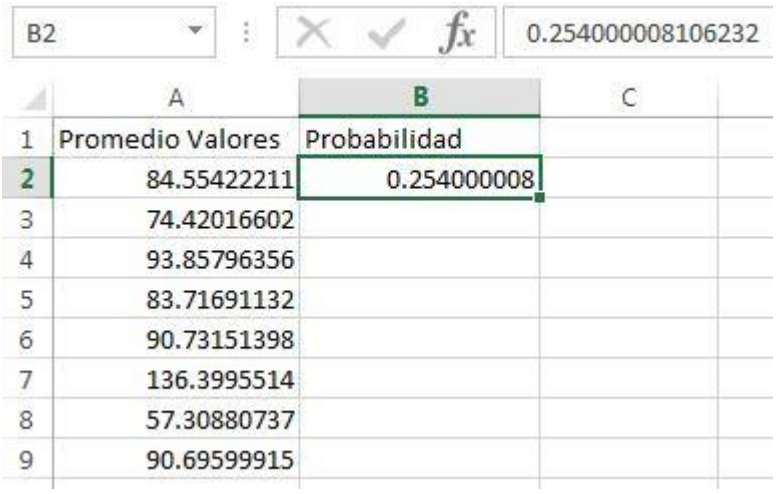
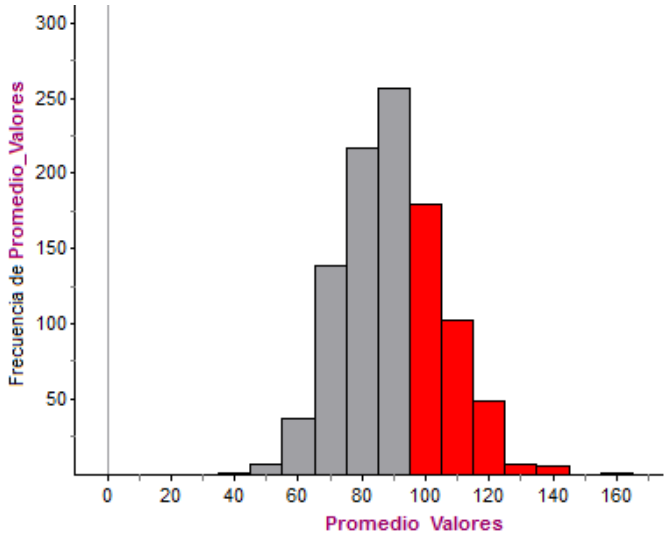


Figura 5. Simulación computacional: Costa Rica Bananas (Parte (b))

La representación gráfica de la distribución de los promedios muestrales permite al estudiante analizar de los 1000 experimentos, en cuántos de ellos la duración promedio fue superior a 100 segundos. Además, de observar la tendencia hacia la distribución normal.

Gráfico 2. Distribución de los tiempos promedios que tarda una empacadora en empacar una caja de bananos (Distribución de los promedios muestrales)



Solución teórica

a) Considere X la duración de un empacador para empacar una caja de bananos en segundos, donde $E(X) = \frac{1}{\lambda} = 90 \Longrightarrow \lambda = \frac{1}{90}$, entonces:

$$X \sim Exp\left(\frac{1}{90}\right)$$

De esta forma:

$$\begin{aligned} P(X > 120) &= 1 - P(X \leq 120) \\ &= 1 - F_X(120) \\ &= 1 - \left(1 - e^{\frac{-1}{90} \cdot 120}\right) \\ &= e^{\frac{-4}{3}} \\ &= 0.2636 \end{aligned}$$

Por lo tanto, hay una probabilidad de 0.2636 de que un empacador sea despedido.

b) Se define la variable $\bar{X}$ como el promedio de la duración de una empacadora (35 empacadores) en empacar una caja.

Como $n = 35 \geq 30$, entonces aplicando el Teorema del Límite Central:

$$\bar{X} \backsim \underbrace{N\left(90, \frac{8100}{35}\right)}_{\left(\mu, \frac{\sigma^2}{n}\right)}$$

Pues si $X \sim Exp(\lambda)$, entonces $\sigma^2 = \frac{1}{\lambda^2} = \frac{1}{(1/90)^2} = 8100$

De esta forma:

$$\begin{aligned} P(\bar{X} > 100) &= 1 - P(\bar{X} < 100) \\ &= 1 - P\left(Z < \frac{100-90}{\sqrt{\frac{8100}{35}}}\right) \\ &= 1 - P(Z < 0.66) \\ &= 1 - \phi(0.66) \\ &= 1 - 0.7454 \\ &= 0.2546 \end{aligned}$$

Por lo tanto, hay una probabilidad aproximada de 0.2546 de que una empacadora sea cerrada.

Actividad 3

3. *Jorge elaboró un software para resolver integrales no triviales, el cual tarda en promedio 6 segundos resolviendo una integral no trivial, con una desviación estándar de 0.75 segundos. Si el programa resolvió una lista de 30 integrales no triviales, determine la probabilidad de que el tiempo total en resolverlas sea inferior a 195 segundos.*

Simulación computacional en Excel

Para este caso *X* es el tiempo en segundos que tarda el software para resolver integrales no triviales, donde $\mu = 6$ y $\sigma = 0.75$. Además, se define S_{30} como el tiempo total que tarda el programa en resolver 30 integrales no triviales, es decir:

$$S_{30} = X_1 + X_2 + \cdots + X_{29} + X_{30}$$

Dado que *X* sigue una distribución desconocida con media μ y desviación estándar σ conocidos, entonces a dicha variable no se le puede asociar una función de distribución de variable continua, pues no se sabe con certeza la manera en que se comportan los datos aleatorios asociado a *X*.

Sin embargo, para generar valores aleatorios en esta simulación computacional, se considera el siguiente resultado empírico relacionado a la probabilidad de una distribución normal:

$$P(\mu - 4\sigma \leq X \leq \mu + 4\sigma) \approx 0.99$$

De esta forma, se van a generar los valores aleatorios mediante la relación anterior, dado que se pueden obtener casi todos los datos aleatorios en cada experimento. Para este caso, en particular como $\mu = 6$ y $\sigma = 0.75$ entonces $X \in [3,9]$.

Se debe aclarar que la probabilidad frecuencial que se obtenga de esta simulación corresponde en realidad a una aproximación conservadora, donde en ocasiones es bastante cercana a la probabilidad teórica, esto pues, al no saber con certeza la distribución de la variable *X* entonces no es posible hallar la probabilidad esperada de manera computacional.

Dicha simulación se realiza mediante las macros de Visual Basic en Excel. Debido a que para cada experimento se deben tomar 30 muestras y obtener su respectivo tiempo

total. Por lo que a la hora de realizar 1000 experimentos se obtendrán un total de 30000 datos aleatorios que implicará es una simulación muy engorrosa por la cantidad de datos.

Por lo que se propone utilizar programación básica, específicamente por medio de ciclos iterativos para realizar cálculos de manera eficiente. De esta forma, la simulación de este problema se realiza directamente en la hoja de programación de Visual Basic y se ejecuta el programa mostrando los datos en la hoja de cálculo de Excel.

El programa principal se llama ***SumaAproxNormales***, en él se escriben las funciones que se utilizarán en la respectiva simulación.

1. Se define la variable *i* correspondiente a un valor aleatorio *X* en el intervalo [3, 9]. A dicha variable le corresponde el siguiente comando: $6 * Rnd() + 3$, donde $Rnd()$ es un comando especial de Visual Basic para generar valores aleatorios en el intervalo [0, 1].

   Observe que la variable i está asociada a la función $y = (b - a)x + a$, donde $x \in [0,1]$, $y \in [a, b]$ con $a = 3$ y $b = 9$, o sea, $y = 6x + 3$.

2. Se define la variable *S* que correspondiente a la suma de *n* valores X_i que sigue una distribución normal, donde $i = \{1,2,,\dots,n\}$.
3. Se define la variable *fila* que corresponde a un contador del primer ciclo que controla la cantidad de valores a ejecutar en la hoja de cálculo de Excel. En este caso se van generar 1000 valores aleatorios desde la celda A2 hasta la celda A1001.
4. Se define la variable *n* asociada al tamaño de la muestra de cada experimento, en este caso *n* = 30. Este contador del segundo ciclo nos permite sumar *n* valores donde su resultado se guardará en la variable *s*.
5. En este caso se desea hallar $P(S < 195)$ entonces se define la variable *c* que se le asocia el comando: CountIf(Range("B1:B1001"),"<195") que permite contabilizar la cantidad de valores que satisfacen la condición buscada.
6. El comando Cells(i,j) nos permite imprimir datos en la hoja de cálculo en la fila *i* y la columna *j*.

   Ahora bien, el código completo del programa diseñado en Visual Basic directamente en las macros de Excel se muestra a continuación:

```
Sub SumaAproxNormales()

Dim i As Single
Dim s As Single
Dim fila As Integer
Dim n As Integer
Dim c As Single
s = 0
fila = 2
Do While fila <= 1001
    n = 1
    s = 0
    Do While n <= 30
        i = 6 * Rnd() + 3
        s = s + i
        n = n + 1
    Loop
    Cells(fila, 1) = s
    fila = fila + 1
Loop
Cells(1, 1) = "Suma Valores"
Cells(1, 2) = "Probabilidad"
c = Application.WorksheetFunction.CountIf(Range("A2:A1001"), "<195")
Cells(2, 2) = c / 1000
End Sub
```

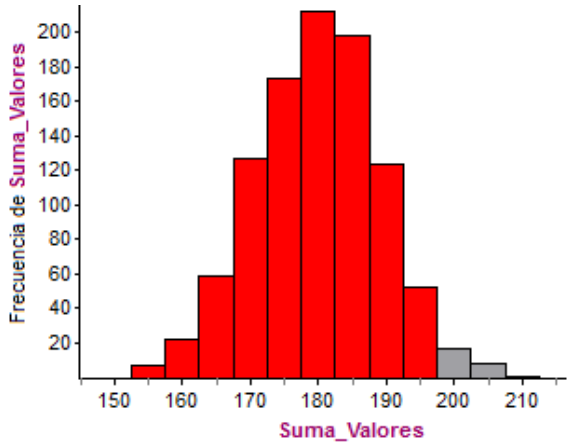
B2 | 0.9549999

	A	B	C
1	Suma Valores	Probabilidad	
2	175.131546	0.954999983	
3	172.0077515		
4	173.9867096		
5	196.7796173		
6	191.5630188		
7	167.6452179		
8	187.6789856		
9	178.0463257		

Figura 6. Simulación computacional: Integrales

La gráfica de las sumas muestrales se presenta a continuación:

Gráfico 3. Distribución de los tiempos totales que tarda el software en resolver 30 integrales (Distribución de las sumas muestrales)



Solución teórica

Sean X el tiempo en segundos que tarda el software para resolver integrales no triviales.

$S = X_1 + X_2 + \cdots + X_{29} + X_{30}$ el tiempo total en segundos que tarda el programa en resolver 30 integrales no triviales,

Como $n = 30 \geq 30$, entonces aplicando el Teorema del Límite Central:

$$S \backsim N\underbrace{(180, 16.875)}_{(n\mu, n\sigma^2)}$$

De esta forma:

$$\begin{aligned} P(S < 195) &= P\left(Z < \frac{195 - 180}{\sqrt{16.875}}\right) \\ &= P(Z < 3.65) \\ &= \phi(3.65) \\ &= 0.9999 \end{aligned}$$

Por lo tanto, la probabilidad aproximada de que en total tarde menos de 195 segundos en resolver las 30 integrales no triviales es 0.9999.

Bibliografía

Alvarado H., Batanero C. (2008). Significado del teorema de central del límite en textos universitarios de probabilidad y estadística. *Estudios Pedagógicos*. 34(2), 10-26. Universidad Austral de Chile, Valdivia, Chile.

Arnaldos F., Faura U. (2012). Aprendizaje de los fundamentos de la probabilidad apoyado en las TICs. *@tic revista d'innovació educativa*. 9, 131-137. Universitat de València, Valencia, España.