Variable aleatoria

Tutoría - Parte I

{desafío} latam_

¿Cómo visualizamos estadísticamente eventos y la probabilidad de ocurrencia de los mismos?

{desafío} latam_

Autoaprendizaje

Recursos asincrónicos

- ¿Revisaste los recursos de la semana 3 (Guía y desafío)?
- ¿Tienes dudas sobre alguno de ellos?

{desafío} latam_

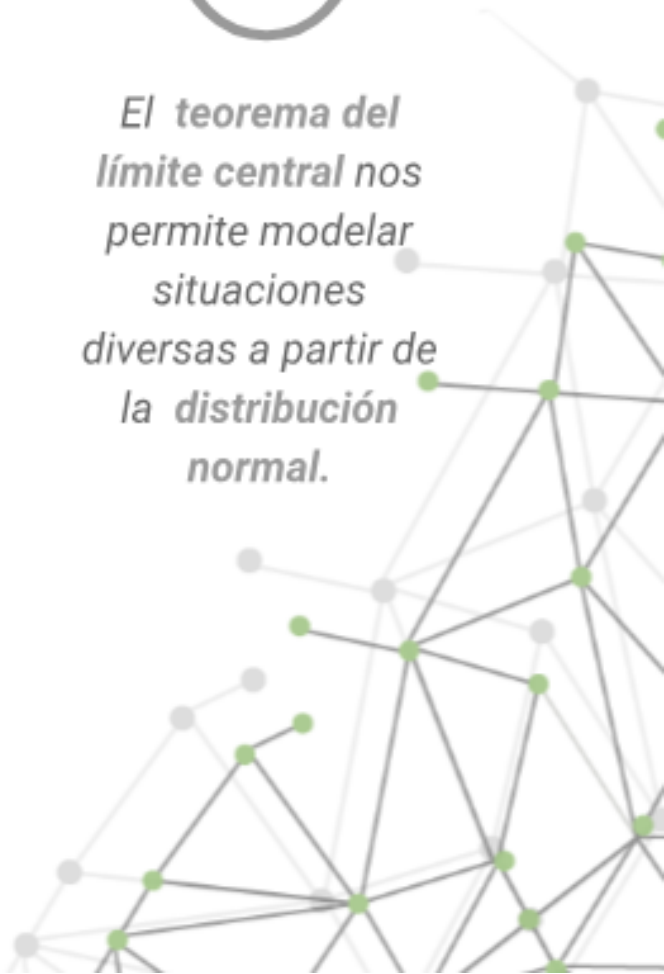
Ideas fuerza

A partir de un experimento u observación definimos ***variables aleatorias,*** *que dependen del aspecto que nos interesa analizar*

Las variables aleatorias pueden ser ***discretas o continuas,*** *presentando así distintas* ***distribuciones****.*

Las ***leyes de los grandes números*** *describen comportamientos de variables aleatorias al* ***aumentar*** *los tamaños de las* ***observaciones****.*

El ***teorema del límite central*** *nos permite modelar situaciones diversas a partir de la* ***distribución normal.***



{desafío} latam_

/* Distribución binomial*/

{desafío}
latam_

Distribución binomial

Definición

Se llama **experimento de Bernouilli** a uno que solo tiene dos resultados posibles, éxito (al que asociamos valor 1) con probabilidad **p**, o fracaso (al que asociamos valor cero) con probabilidad **1 - p**.

Si repetimos este experimento n veces, ¿cuál es la probabilidad de obtener k éxitos? Podemos calcular esta probabilidad mediante la siguiente fórmula

$$P(k) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

Distribución binomial: ejemplo

- Encuentra la media y la desviación estándar de una variable aleatoria que sigue una distribución binomial correspondiente a 50 pruebas, cada una con una probabilidad de éxito igual a 0.2.

En este caso tenemos $n = 50, p = 0{,}2$.

$$\mu = n \cdot p = 50 \cdot 0{,}2 = 10$$

Sabemos que $\sigma^2 = npq \implies \sigma = \sqrt{np(1-p)} = \sqrt{50 \cdot 0{,}2 \cdot 0{,}8} = 2{,}8284$

Distribución binomial: ejemplo

- Un biólogo examina ranas en busca de un rasgo genético que sospecha podría deberse a agua contaminada. Normalmente, el rasgo examinado se presenta en promedio en 1 de cada 8 ranas en la naturaleza. Si la frecuencia de este rasgo no ha cambiado y captura 12 ranas para examinar, ¿cuál es la probabilidad de que encuentre este rasgo en más de 4 ranas?

Sea X variable aleatoria que indica si una rana presenta o no el rasgo genético. En este caso X distribuye $Binomial(n, p)$ donde $n = 12$ y $p = 1/8$. Se busca

$$\begin{aligned} P(x > 4) &= 1 - (P(0) + P(1) + P(2) + P(3) + P(4)) \\ &= 1 - \left(\frac{12!}{0!(12-0)!} + \ldots + \frac{12!}{4!(12-4)!} \right) \\ &= \ldots \\ &= 0{,}0113 \end{aligned}$$

/* Distribución Normal */

Distribución Normal

Definición

La distribución normal es un patrón o forma de distribución de datos que se asemeja a una campana simétrica. Es comúnmente observada en muchos fenómenos de la vida real, como la altura de las personas o las puntuaciones en exámenes.

En esta distribución, la mayoría de los valores se encuentran cerca del valor promedio, y a medida que nos alejamos de él, la frecuencia de los valores disminuye gradualmente. Es una distribución muy utilizada en estadística para modelar y analizar datos.

Distribución Normal

¿Cómo la caracterizamos?

- Simetría
- Media, mediana y moda iguales
- Forma unimodal
- Tails o colas infinitas
- Parámetros de media y desviación estándar

{desafío}
latam_

Ejercicio:
La distribución normal en acción

{desafío}
latam_

La distribución normal en acción

Distribución normal con Python

Veremos cómo la distribución normal nos puede ayudar a modelar algunas situaciones, para lo que utilizaremos Python, Puedes abrir tu propio archivo de Jupyter Notebook para replicar los pasos que te mostrará tu profesor, con los que aprenderemos:

1. Modelamiento con la distribución normal
2. Análisis de los parámetros de la distribución normal
3. Aplicación de la distribución normal

Distribución normal

Probabilidades y estandarización

Pese a su gran utilidad, la función de distribución normal tiene una gran dificultad: no tiene primitiva, por lo que la integral que permite calcular el área bajo la curva no puede determinarse algebraicamente.

Para resolver este problema, se debe recurrir a valores de tabla que se encuentran calculados para **valores estandarizados,** es decir, una función de media igual a cero y desviación estándar igual a 1.

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

Distribución normal

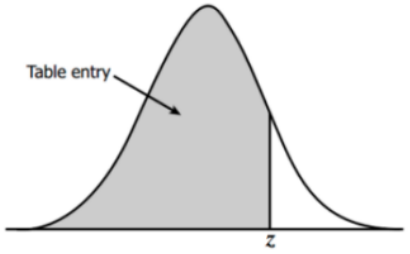
Probabilidades y estandarización

Así, por ejemplo, si la media de nuestro conjunto es μ, y su desviación estándar es σ, podemos **estandarizar** un valor cualquiera x de nuestro conjunto, restándole μ, y dividiendo el resultado por σ. Este valor estandarizado es el que puede ser buscado en una tabla.

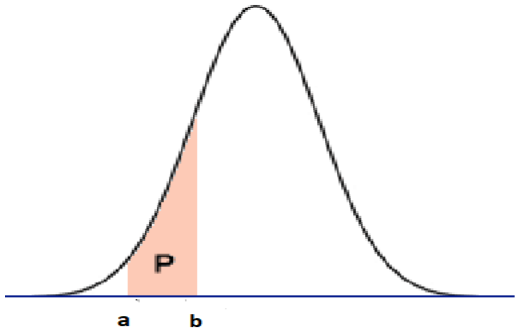
$$z = \frac{x - \mu}{\sigma}$$

Z score

$$z = \frac{x - \mu}{\sigma}$$



"Las entradas de la tabla representan el área bajo la curva de campana a la izquierda de z (también conocido como usar probabilidades directamente de la tabla)."



$$P(a \leq Z \leq b) = P(Z \leq b) - P(Z \leq a)$$

https://www.z-table.com/

Z score ejercicio

$$z = \frac{x - \mu}{\sigma}$$

Los puntajes de CI tienen una media de 100 y una desviación estándar de 16. Se dice que Albert Einstein tenía un CI de 160.

$\mu = 100,\ \sigma = 16 \qquad x = 160$

- ¿Cuál es la diferencia entre el CI de Einstein y la media?

$$x - \mu = 160 - 100 = 60$$

- ¿Cuántas desviaciones estándar son eso?

$$\frac{x - \mu}{\sigma} = \frac{160 - 100}{16} = 60/16 = 3{,}75$$

- Convierte el puntaje de CI de Einstein a un puntaje z

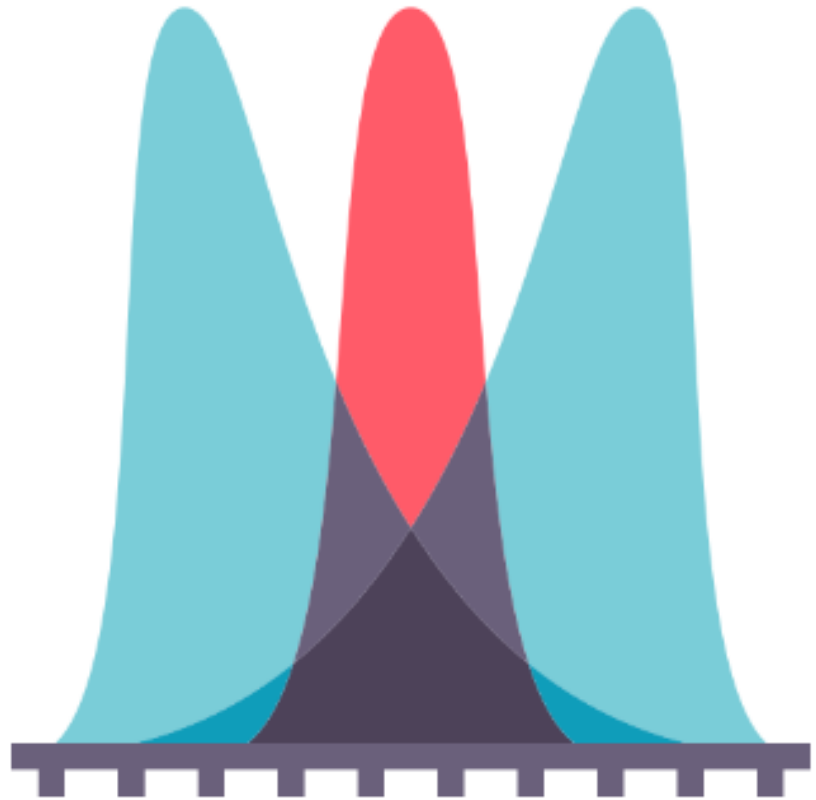
- Si consideramos que los puntajes de CI "normales" son aquellos que convierten puntajes z entre -2 y 2, ¿el CI de Einstein es normal o anormal?

Distribución normal

Probabilidades y estandarización

Un aspecto muy importante de la distribución normal es que nos permite establecer cuál es la probabilidad de que un valor dado se encuentre a una distancia dada de la media, en términos de la desviación estándar. Así, sabemos que si un conjunto de datos se distribuye en forma normal, entonces:

- Aproximadamente, un 68% de los datos se encuentran entre **x - μ** y **x + μ**
- Aproximadamente, un 95% de los datos se encuentran entre **x - 2μ** y **x + 2μ**
- Aproximadamente, un 99,7% de los datos se encuentran entre **x - 3μ** y **x + 3μ**



{desafío} latam_

/* Ley de los grandes números */

Ley de los grandes números

Ley Débil y Ley Fuerte

Ley débil

la media muestral se acerca a la media poblacional cuando el tamaño de la muestra aumenta

Ley fuerte

la media muestral tiende, con probabilidad 1, a la media poblacional en la medida que el tamaño n de la muestra tiende a infinito

Ley de los grandes números: ejemplo

Por ejemplo, una sola tirada de un dado de seis caras produce uno de los números 1, 2, 3, 4, 5 ó 6, cada uno con la misma probabilidad.
Por tanto, el valor esperado del promedio de las tiradas es:

$$\frac{1+2+3+4+5+6}{6} = 3.5$$

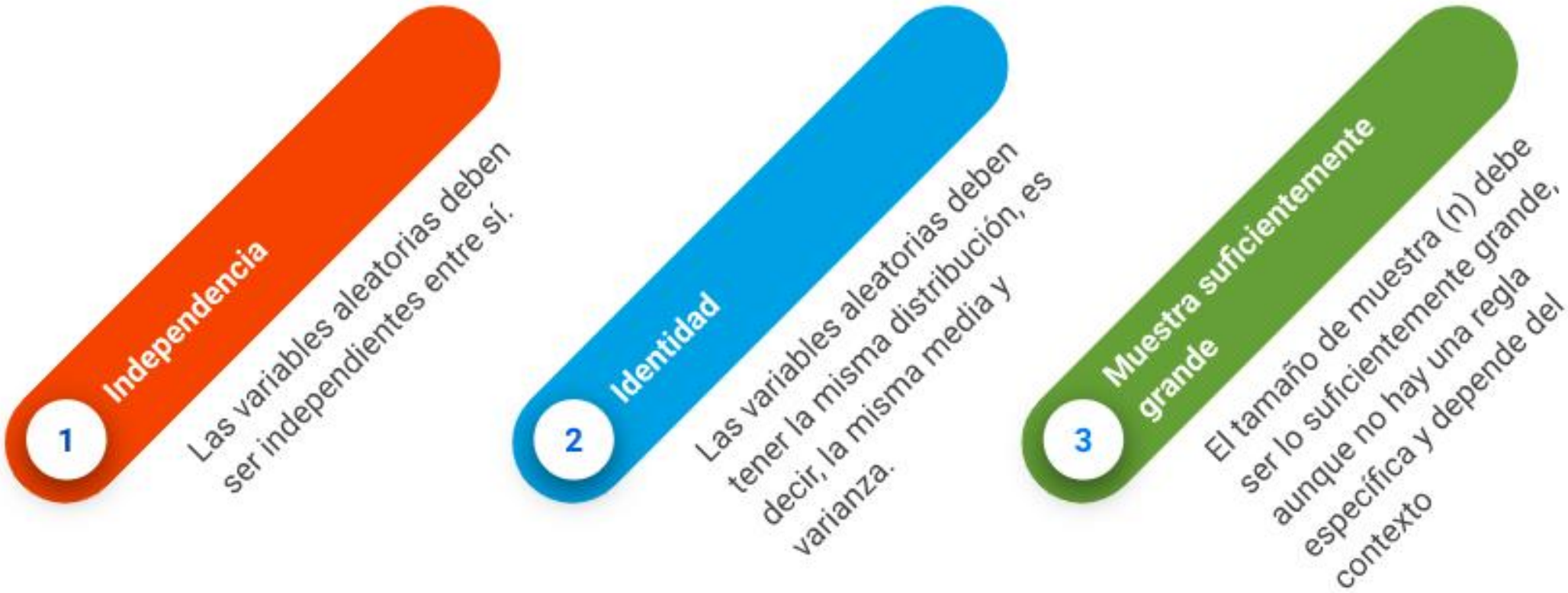
De acuerdo con la ley de los grandes números, si se lanza una gran cantidad de dados de seis caras, el promedio de sus valores (a veces llamado media muestral) se aproximará a 3,5, y la precisión aumentará a medida que se lancen más dados.

/* Teorema del límite central*/

{desafío}
latam_

Teorema del límite central

Condiciones



Teorema del límite central

Interpretación y consecuencias

A medida que aumenta el tamaño de la muestra, la distribución de la suma o media se aproxima cada vez más a una distribución normal.

+

La distribución normal puede ser utilizada para estimar probabilidades o intervalos de confianza relacionados con la suma o media de una muestra grande.

→

Por ejemplo, permite realizar inferencias sobre la media de una población utilizando muestras grandes, ya que se puede confiar en que la distribución de las medias muestrales se aproxima a una distribución normal.

Teorema del límite central: ejercicio

Una encuesta encontró que la familia estadounidense genera un promedio de 17.2 libras de basura de vidrio cada año. Supongamos que la desviación estándar de la distribución es de 2.5 libras.

- Encuentra la probabilidad de que la **media** de una muestra de 55 familias esté entre 17 y 18 libras.

Para la distribución de medias muestrales $\mu = 17{,}2$, mientras que $\sigma = \frac{2{,}5}{\sqrt{55}} = 0{,}3371$. Se busca $P(17 < X < 18)$, entonces primero estandarizamos observando que $Z_{17} = \frac{17-17{,}2}{0{,}3371} = -0{,}5933$ y $Z_{18} = \frac{18-17{,}2}{0{,}3371} = 2{,}373$ entonces $P(-0{,}5933 < Z < 2{,}373) = P(Z < 2{,}373) - P(Z \leq -0{,}5933) = 0{,}7146$.

- ¿Por qué se puede aplicar el teorema del límite central?

Estamos considerando una distribución de medias muestrales, por lo que se aplica el Teorema del Límite Central. (Además, como $55 > 30$, podemos aproximar esta distribución de medias muestrales como una distribución normal).

Ejercicio
Verifiquemos con Python

{desafío}
latam_

Ley de los grandes números y límite central

Verificando con Python

Verificaremos con Python los resultados anteriores, para lo que deberás seguir la presentación que te mostrará tu profesor. Puedes abrir tu propio archivo de Jupyter Notebook para replicar los pasos, con los que observaremos:

1. Una aplicación de la ley débil de los grandes números
2. Una aplicación de la ley fuerte de los grandes números
3. Una aplicación del teorema del límite central

Distribución binomial y ley de los grandes números

Esperanza y desviación teórica

Se llama **esperanza** al "promedio teórico" en un experimento. Para una variable aleatoria X con distribución binomial tenemos que

$$E(X) = n \cdot p$$

$$\sigma(X) = \sqrt{n \cdot p \cdot (1-p)}$$

Desafío

"Estadística descriptiva y probabilidades (parte II)"

- ¿Leíste el desafío de esta semana? ¿Comprendes bien lo que se solicita en cada caso?
- ¿Hay contenidos que necesitas repasar antes de comenzar este desafío?
- ¿Necesitas algún ejemplo o indicación para alguna pregunta o requerimiento específico?

Próxima sesión...

- *Crea visualizaciones que permiten identificar la distribución de variables de un dataset*