Inteligencia Artificial

Act7: Laboratorio de repaso de Probabilidad y Estadística

Arturo Garza Rodríguez

February 2025

1. Tipos de datos y medidas de tendencia central

En una empresa se han recolectado los siguientes datos de 10 empleados:

Nombre	**Edad (años)**	**Área de trabajo**
Ana	25	Ventas
Luis	30	Administración
Marta	40	Producción
Carlos	35	Ventas
Elena	28	Recursos Humanos
Juan	50	Producción
SofÃa	45	Administración
Pedro	38	Ventas
Daniel	33	Producción
Laura	27	Recursos Humanos

1.1. Clasificación de variables

1.1.1. Variables cuantitativas

Las variables cuantitativas son aquellas que describen la cantidad de algo. Por lo tanto, decimos que:

'**Edad**' es la única variable cuantitativa.

1.1.2. Variables cualitativas

Las variables cualitativas se encargan de describir características no cuantificables. Es decir:

'**Nombre**' y '**Área de trabajo**' son variables cualitativas.

1.2. Medidas de tendencia central de Edad

Las medidas de tendencia central son valores que resumen un conjunto de datos indicando dónde se concentra la mayoría de los datos.

1.2.1. Media

Para calcular la media se utiliza la siguiente función:

$$\mu = \frac{\sum x_i}{n} \tag{1}$$

Donde μ es la media, x_i es cada valor de la muestra o población, y n es el tamaño de la muestra o población.

Aplicando esta fórmula al conjunto de datos respectivos a la *edad* de cada persona tenemos que:
De la fórmula 1

$$\mu = \frac{\sum x_i}{n} = \frac{25+30+40+35+28+50+45+38+33+27}{10} = 35{,}1 \tag{2}$$

Por lo tanto la edad media entre los empleados es de 35.1 años.

1.2.2. Mediana

Para obtener la mediana no existe una función definida, sino más bien, una serie de pasos a seguir:

1. Ordenar los datos de forma ascendente
2. Contar cuántos datos son:
   a) Si n es par, dividir $n/2$ y tomar los valores en las posiciones $n/2$ y $(n/2)+1$, y obtener su media.
   b) Si n es impar, sumar $n+1$ y luego dividir $(n+1)/2$ y el dato en la posición obtenida es la mediana.

Aplicando estos pasos a nuestros datos, obtenemos que, las edades ordenadas de forma ascendente son:

$$Edades = 25, 27, 28, 30, 33, 35, 38, 40, 45, 50$$

Como $n = 10$, entonces dividimos entre dos y tomamos los valores en las posiciones 5 y 6, y obtenemos su media. $x_5 = 33$, $x_6 = 35 \rightarrow m = \frac{(33+35)}{2} \rightarrow m = 34$.
Por lo tanto, la mediana entre las edades de los empleados es de 34 años.

1.2.3. Moda

La moda es el dato que más se repite en un conjunto de datos. Es más fácil realizar el conteo de frecuencias en un conjunto de datos ya ordenado, por lo que aprovecharemos que ordenamos las edades de forma ascendente para obtener la mediana para conseguir la moda.

$$Edades = 25, 27, 28, 30, 33, 35, 38, 40, 45, 50$$

Una vez que tenemos el conjunto de datos, procedemos a contar las frecuencias de cada dato.
en este caso, podemos observar que todos los datos aparecen únicamente una vez, por lo tanto, no hay una moda explícita, no existe una frecuencia que sobresalga de las demás. Por lo tanto se detemrina que es un conjutno de datos **amodal**.

1.3. Interpretación de resultados

Las medidas de tendencia central, como se mencionó al inicio, nos indican cómo se comporta, a grandes rasgos, un conjunto de datos. Como su nombre lo indica, reflejan el punto alrededor del cual tienden a agruparse los valores.

En este caso, observamos que la media es de 35.1 años, lo que sugiere que, en promedio, los empleados tienen alrededor de esa edad. La mediana es 34 años, lo que indica que la mitad de los empleados tiene menos de 34 años y la otra mitad más. Por otro lado, no existe una moda, lo que

implica que las edades están distribuidas de manera uniforme sin que ninguna edad se repita más que las demás.

Esto sugiere una distribución relativamente equilibrada, sin valores atípicos que afecten significativamente la tendencia central.

2. Medidas de dispersión

Dado el siguiente conjunto de datos correspondiente a las calificaciones de 8 estudiantes en un examen:

$$X = \{70, 85, 90, 95, 88, 92, 75, 80\}$$

2.1. Medidas de dispersión

Las medidas de dispersión nos indican qué tan dispersos están los datos en torno a la media de la muestra o población. Es decir, muestran cuánta variabilidad existe entre los datos y ayudan a comprender si estos se agrupan cercanamente a la media o si están más alejados

2.1.1. Varianza

La fórmula para obtener la varianza es la siguiente (considerando que esto es una muestra de todos los resultados de los estudiantes en la escuela):

$$S^2 = \frac{\sum(x_i - \mu)^2}{n-1} \tag{3}$$

Aplicando la ecuación 3 al conjunto de datos, primero obtenemos la media, la cual es $\mu = 84{,}375$. Ahora procedemos a calcular la varianza:

$$\begin{aligned} S^2 = &\frac{(70-84{,}375)^2 + (85-84{,}372)^2 + (90-84{,}375)^2 + (95-84{,}375)^2}{7} \\ &\frac{+(88-84{,}375)^2 + (92-84{,}375)^2 + (754-84{,}375)^2 + (80-84{,}375)^2}{7} \end{aligned} \tag{4}$$

No cupo en una sola línea, pero es la misma ecuación dividida en 2 partes. Por lo tanto, la varianza que encontramos es de: $S^2 \approx 75{,}7$

2.1.2. Desviación estándar

LA desviación estándar no es más que solamente la raíz cuadrada de la varianza:

$$\sigma = \sqrt{S^2} = \sqrt{\frac{\sum(x_i - \mu)^2}{n-1}} \tag{5}$$

Por lo tanto, $\sigma = \sqrt{75{,}7} \approx 8{,}7$

2.2. Interpretación de resultados

Las medidas de dispersión indican qué tan dispersos están los datos respecto a la media. En este caso, la varianza de aproximadamente 75.7 muestra una dispersión moderada de las calificaciones, lo que sugiere que las notas no están concentradas cercanamente alrededor de la media. Esto significa que existen diferencias significativas entre las calificaciones de los estudiantes.

Por otro lado, la desviación estándar, que es aproximadamente 8.7, indica que las calificaciones se desvían en promedio 8.7 puntos de la media de 84.375. Aunque no es una desviación extremadamente alta, muestra una variabilidad moderada, lo que sugiere que no todos los estudiantes tienen un rendimiento similar. Esto puede ser útil para identificar a los estudiantes cuyo rendimiento difiere mÃ¡s de la media.

3. Probabilidades y Teorema de Bayes

Una empresa de tecnología ha interpretado que el 60 % de sus empleados son programadores, y el 40 % son diseñadores. Se sabe que el 70 % de los programadores tienen conocimientos de inteligencia artificial (IA), mientras que solo el 30 % de los diseñadores tienen estos conocimientos.

Si se elige un empleado al azar y se sabe que tiene conocimientos de IA, ¿cuál es la probabilidad de que sea programador?

3.1. Teorema de Bayes

El Teorema de Bayes, o la fórmula de la probabilidad condicional nos dice cuál es la probabilidad de que suceda un evento dado que ya se conoce la ocurrencia de otro. Esto se expresa en la siguiente fórmula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{6}$$

3.2. Aplicación del Teorema

Se sabe que la probabilidad de que al seleccionaer un empleado al azar es de 60 %, y el 70 % de estos trabajadores tienen conocimientos sobre IA, mientras que del 40 % de trabajadores, es decir, los diseñadores, solo el 30 % tiene conocimientos sobre IA. Definamos los eventos:

$A \rightarrow Seleccionarunprogramador$
$B \rightarrow Seleccionarundiseñador$
$C \rightarrow TieneconocimientosdeIA$

El enunciado 'del 60 % solo el 70 %' se puede expresar como:

$$\frac{60}{100} \cdot \frac{70}{100} = \frac{42}{100} = 42\,\%$$

Mientras que el enunciado 'del 40 % solo el 30 %' así:

$$\frac{40}{100} \cdot \frac{30}{100} = \frac{12}{100} = 12\,\%$$

Ahora que ya tenemos las probabilidades de que ocurran los eventos de manera simultánea, esto es:
$P(A \cap C) = 42\,\%$
$P(B \cap C) = 12\,\%$
Y los eventos de manera independiente:
$P(A) = 60\,\%$
$P(B) = 40\,\%$
$P(C) = 54\,\%$

Como lo que queremos es saber la probabilidad de que al elegir un empleado al azar sea programador si se sabe que tiene conocimientos de IA; en la función 6 sustituimos los valores necesarios:

$$P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{42\,\%}{54\,\%} \tag{7}$$
$$= \frac{\frac{42}{100}}{\frac{54}{100}} = \frac{42}{54} = \frac{21}{27} \approx 0{,}78$$

Por lo tanto la probabilidad de que al seleccionar un empleado al azar sabiendo que posee conocimiento sobre IA, este empleado sea un programador es del **78 %**.

4. Distribuciones de probabilidad

Suponga que el número de defectos en un lote de producción sigue una distribución de Poisson con media $\lambda = 3$ defectos por lote.

4.1. Distribución Poisson

La distribución de Poisson describe el número de eventos que ocurren en un intervalo de tiempo o espacio fijo, bajo la condición de que los eventos son independientes y ocurren a una tasa constante. Esto se expresa en la siguiente ecuación.

$$P(X = x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!} \tag{8}$$

4.2. Probabilidad de lote con exactamente 2 defectos

Esto es $P(x = 2)$, entonces sustituimos en 8:

$$P(x = 2) = \frac{3^2 \cdot e^{-3}}{2!} = \frac{9}{2e^3} \approx 0{,}224 \tag{9}$$

Por lo tanto, la probabilidad de que un lote contenga exactamente 2 defectos es de 22.4 %.

4.3. Probabilidad de lote con al menos 1 defecto

Esto es

$$P(x >= 1) = 1 - P(x < 1) = 1 - P(x = 1) - P(x = 0) \tag{10}$$

Como trabajamos con una variable discreta, esta solo toma valores enteros (o discretos) mayores o iguales a 0; también conocemos que la probabilidad de A evento oscila entre 0 y 1, por lo que cualquier probabilidad puede calcularse con la operación 'inversa' de $1-$*probabilidad de que no suceda el evento* A, por lo que se cumple lo anterior.
Entones, sustituimos en 10:

$$P(x >= 1) = 1 - P(x = 1) - P(x = 0) = 1 - \frac{3^1 \cdot e^{-3}}{1} - \frac{3^0 \cdot e^{-3}}{1} \tag{11}$$

$$\approx 1 - 0{,}15 - 0{,}05 \approx 0{,}8$$

Por lo tanto, la probabilidad de que en un lote, exista al menos un defecto, es del **80 %**.

5. Funciones de densidad y distribución acumulativa

Sea X una variable aleatoria con distribución normal de media $\mu = 50$ y deviación estándar $\sigma = 10$.
Entonces:

$$x \sim N(\mu = 50, \sigma^2 = 100)$$

Para calcular probabilidades de una distribución normal, la manera más sencilla es estandarizando la variable aleatoria, esto se hace mediante la siguiente fórmula:

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

y utilizar la distribución normal estándar con media 0 y varianza 1, cuyas probabilidades conocemos.

5.1. Probabilidad de $X < 45$

Esto es: $P(x < 45)$, y estandarizamos la variable x, llegando a: $P(\frac{x-50}{10} < \frac{45-50}{10}) = P(z < -0{,}5)$, dato que podemos obtener de forma directa.
Al ser una función simétrica $P(z < -0{,}5) = P(z > 0{,}5) = 0{,}3085$
Por lo tanto, la probabilidad de que $x < 45$ es del **30.85 %**.

5.2. Probabilidad de $40 <= X <= 60$

Sucede algo similar en este caso, solamente que ahora trabajamos con dos limites, inferior y superior. Estandarizamos la variable x, obteniendo: $P(\frac{40-50}{10} <= \frac{x-50}{10} <= \frac{60-50}{10} = P(-1 <= z <= 1)$.
Al igual que en las integrales, si conocemos de antemano que la función a tratar es simétrica en un punto, podemos partir la integral desde ese punto hasta el límite propuesto y multiplicar por 2 el resultado, aquí también se puede. La distribución normal estandarizada es simétrica en $x = 0$, y

como podemos ver, 0 es el punto medio entre -1 y 1, entones, podemos decir que: $P(-1 <= z <= 1) = 2 \cdot P(0 < z < 1)$, mismos valores que podemos obtener de forma directa:$2 \cdot (0{,}5 - 0{,}1587) = 2 \cdot 0{,}3413 = 0{,}6826$. Por lo tanto, la probabilidad de que $40 <= x <= 60$ es del **68.26 %**.

5.3. Comprbación de resultados

Para comprobar que los resultados obtenidos son correctos, los compararemos con la función de probabilidad acumulada la cual se representa como $\Phi(k)$, donde k es el valor estandarizado. Como ya tenemos los valores estandarizados, pasaremos directo a la verificación.

5.3.1. Verificación de $X < 45$

$$P(x < 45) = P(-\infty < x < 45) = P(-\infty < z < -0{,}5) = \Phi(-0{,}5) \approx 0{,}3085$$

5.3.2. Verificación de $40 <= X <= 60$

$$P(40 < x < 60) = P(-1 < z < 1) = \Phi(1) - \Phi(-1) = 1 - 0{,}1587 - 0{,}1587 = 0{,}6826$$

6. Probabilidad condicional

Un dado justo de seis caras se lanza dos veces.

6.1. Espacio muestral

El espacio muestral está definido como:

$$\begin{gathered} S = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2), \\ (2,3),(2,4),(2,5),(2,6),(3,1),(3,2),(3,3),(3,4),(3,5), \\ (3,6),(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),(5,1),(5,2),(5,3), \\ (5,4),(5,5),(5,6),(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\} \end{gathered}$$

Definimos eventos:

$$A \rightarrow \text{Primer dado es impar}$$

$$B \rightarrow \text{Segundo dado es par}$$

Y probabilidades:

$$P(A) = \frac{1}{2}$$

$$P(B) = \frac{1}{2}$$

$$(P(A \cap B) = P(A) \cdot P(B) = \frac{1}{4}$$

6.2. ¿Cuál es la probabilidad de obtener un número par en el segundo lanzamiento, par dado que el primero salió un impar?

La probabilidad de que el segundo número sea par si el primero ya es impar está dada por:

$$P(B|A) = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2} \tag{12}$$

Por lo tanto, la probabilidad de que el segundo dado sea par, dado que el primero es impar es del **50 %**.

6.3. Interpretación de resultados

Es lógico que el resultado sea de 50 %, ya que, en un dado de 6 lados, hay 3 números pares, y 3 impares, es decir, 0.5 de probabilidad de obtener número par o impar (equiprobables), que al considerar el primero ya como impar, podemos decir que ese evento no nos interesa mucho, pues ya es algo conocido, mientras que del segundo dado, sí consideramos la probabilidad de que sea par, por lo tanto, obtenemos $1 \cdot \frac{1}{2} = \frac{1}{2}$. **Como los eventos son independientes, conocer que el primer dado es impar no afecta la probabilidad de que el segundo sea par.**

7. Distribución binomial

Un examen de opción múltiple tiene 5 preguntas, cada una con 4 posibles respuestas, de las cuales solo una es correcta. Un estudiante responde al azar.

En una distribución binomial consideramos una variable aleatoria discreta x como la cantidad de éxitos en las n repeticiones del experimento, mientras que tenemos θ como la probabilidad de éxito. En este caso, $n = 5, \theta = 0{,}25$. La función de distribución binomial es la siguiente:

$$f(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \tag{13}$$

7.1. Probabilidad de que el estudiante acerte exactamente 3 respuestas

Para este caso definimos:

$$n = 5$$

$$x = 3$$

$$\theta = 0{,}25$$

Sustituyendo los valores en 13, obtenemos:

$$f(3) = \binom{5}{3} \cdot (0{,}25)^3 \cdot (0{,}75)^2 = 0{,}0879$$

Por lo tanto, la probabilidad de que el estudiante acierte exactamente 3 de las 5 preguntas es del **8.79 %**.

7.2. Probabilidad de la menos un acierto por el estudiante

Análogamente al caso anterior, ahora tenemos el caso en que acierte al menos una pregunta, es decir: $P(x >= 1)$. Como $x \in [0, 5]$, porque son 5 preguntas, y así como puede acertar todas, puede errar todas. En lugar de calcular la probabilidad para cada caso en que $x >= 1$, mejor calculamos la probabilidad 'inversa', que nos dice que:

$$P(x >= 1) = 1 - P(x < 1), \text{ como trabamos con variables discretas} \rightarrow 1 - P(x = 0)$$

Calculando para $x = 0$ en 13:

$$f(0) = \binom{5}{0} \cdot (0{,}25)^0 \cdot (0{,}75)^5 = 0{,}2373$$

Y obtenemos $P(x >= 1)$ como:

$$P(x >= 1) = 1 - P(x = 0) = 1 - 0{,}2373 = 0{,}7627$$

Por lo tanto, la probabilidad de que el estudiante tenga al menos un acierto es del **76.27 %**.

8. Regla de Laplace

Una urna contiene 5 bolas rojas y 7 bolas azules. Se extrae una bola al azar.

8.1. Probabilidad de que la bola extraída sea roja

La cantidad de bolas rojas es 5, mientras que el total de bolas en la urna es de 12, por lo tanto, la probabilidad de que la bola extraída sea roja es de:

$$P(roja) = \frac{5}{12} \approx 0{,}4167$$

Es decir **41.67 %**.

8.2. Si se extraen dos bolas sin reemplazo, ¿cuál es la probabilidad de que ambas sean azules?

Para este caso, hay que considerar que son eventos dependientes, ya que uno influye en el otro. La ocurrencia de sacar una bola sin reemplazo afecta directamente en la probabilidad del siguiente evento.

$$P(azul_1) = \frac{7}{12} = 0{,}5833$$

$$P(azul_2) = \frac{6}{11} = 0{,}5454$$

Para la extracción de la segunda bola se redujo en 1 la cantidad total ya que se extrajo una bola anteriormente, y como fue azul, también se reduce en 1 la cantidad de azules en la urna. Es decir:

$$P(azul_{1,2}) = 0{,}5833 \cdot 0{,}5454 \approx 0{,}3181$$

Por lo tanto, la probabilidad de sacar una bola azul y no reemplazarla y luego volver a sacar una bola azul de la urna es del **31.81 %**.

9. Esperanza matemática

Suponga que una persona juega una lotería donde el premio es de 1000 dólares con una probabilidad de 0.01, y el costo de cada boleto es 10 dólares.

9.1. Esperanza matemática de ganancia del jugador

La esperanza matemática de una variable aleatoria X es el valor promedio que se espera obtener al realizar un experimento de manera repetida, ponderado por las probabilidades de cada posible resultado.
En este caso, los dos posibles resultados es que gane o no gane la lotería, cada uno con distintas cantidades de ganancia o périda.

1. Ganar la lotería = 1000 dólares - 10 dólares (ganancia menos precio del boleto) = 990 dólares
2. Perder la lotería = -10 dólares (costo del boleto)

Y las probabilidades asociadas son:

$$P(G) = 0{,}01$$
$$P(P) = 0{,}99$$

Definimos la esperanza matemática como la suma de las ganancias multiplicadas por sus respectivas probabilidades:

$$E(G) = (990 \cdot 0{,}01) + (-10 \cdot 0{,}99) = 9{,}9 - 9{,}9 = 0 \tag{14}$$

Por lo tanto, la esperanza de ganancia para el usuario es de **0 dólares**.

9.2. Interpretación de resultados

Esto significa que, en promedio, el jugador no ganará ni perderá dinero a largo plazo. Si el costo de cada boleto es 10 dólares, la esperanza matemática de la ganancia es 0. Esto indica que, si juegas muchas veces, ganarás en promedio lo mismo que gastaste en los boletos, lo que muestra que, a largo plazo, el juego es "justo"para el jugador en términos de ganancia neta.

10. Ley de grandes números

Un experimento consiste en lanzar una moneda justa 1000 veces y calcular la frecuencia relativa de obtener cara.
La frecuencia relativa de obtener cara es simplemente el número de veces que se obtiene cara en los 1000 lanzamientos dividido por 1000, es decir, es una variable aleatoria que depende de cuántas veces se obtiene cara en los lanzamientos.

10.1. ¿Cuál es el valor esperado de la frecuencia relativa de obtener cara?

Conociendo que la moneda solamente tiene dos caras, y al ser justas estos eventos se vuelven equipribables, el valor esperado es de **50 %**.

10.2. ¿Cómo se relaciona esto con la Ley de los Grandes Números?

A medida que el número de lanzamientos aumenta, la frecuencia relativa de obtener cara se acercará cada vez más a 0.5. Esta es una manifestación directa de la Ley de los Grandes Números, que predice que los resultados experimentales se alinearán con la probabilidad teórica conforme se repiten los experimentos.