Catedra 1 Análisis estadístico

Adán Marchena

1. Análisis descriptivo de los datos

Los datos utilizados provienen de una encuesta aplicada a estudiantes universitarios. El conjunto incluye 400 observaciones y 18 variables. A continuación, se realiza un análisis descriptivo siguiendo las instrucciones del enunciado.

1.1. Clasificación de variables

Visualización inicial

```
Rows: 400
Columns: 18
$ ID               <chr> "SB11201910010435", "SB11201910004475", "SB1120191001~
$ Sexo             <chr> "Femenino", "Masculino", "Masculino", "Masculino", "F~
$ Edad             <chr> "21.36", "21.07", "20.92", "18.41", "16.64", "16.02",~
$ Fuma             <chr> "No", "Si", "Si", "Si", "Si", "No", "Si", "Si", "Si",~
$ Estatura         <chr> "Alta", "Baja", "Alta", "Alta", "Alta", "Baja", "Baja~
$ Colegio          <chr> "Privado", "Privado", "Privado", "Privado", "Privado"~
$ Estrato          <dbl> 1, 2, 2, 2, 1, 2, 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 2, 1,~
$ Financiacion     <chr> "Beca", "Beca", "Beca", "Beca", "Beca", "Beca", "Beca~
$ Acumulado        <chr> "3.92", "3.96", "3.85", "3.69", "4.01", "3.80", "4.35~
$ Gastos           <chr> "48.9", "72.1", "85.2", "56.6", "64.6", "63.0", "40.8~
$ Ingreso          <chr> "0.61", "2.07", "0.84", "1.55", "2.32", "2.10", "0.69~
$ Clases           <chr> "Virtual", "Presencial", "Virtual", "Virtual", "Virtu~
$ Pandemia         <chr> "De acuerdo", "De acuerdo", "De acuerdo", "De acuerdo~
$ Clases_virtuales <chr> "Totalmente de acuerdo", "Ni de acuerdo, ni en desacu~
$ Estadistica      <chr> "Indeciso", "En desacuerdo", "En desacuerdo", "Totalm~
$ inseguridad      <chr> "De acuerdo", "De acuerdo", "Totalmente de acuerdo", ~
$ vida_cotidiana   <chr> "De acuerdo", "Totalmente en desacuerdo", "De acuerdo~
$ Puntaje          <dbl> 81, 78, 77, 70, 68, 65, 54, 50, 36, 35, 35, 34, 30, 2~
```

Descripción y clasificación de variables

Se realizó una inspección preliminar del conjunto de datos utilizando la función `str()` de R, lo cual permitió observar que algunas variables numéricas como `Edad`, `Ingreso`, `Gastos` y `Acumulado` estaban codificadas como texto (`character`). Estas fueron convertidas a tipo `numeric` para permitir su análisis estadístico.

A continuación, se presenta una clasificación general de las variables según su tipo estadístico:

Table 1: Clasificación y descripción de las variables

Variable	Descripción	Tipo
ID	Código o identificador del estudiante	Cualitativa (Nominal)
Sexo	Género del estudiante	Cualitativa (Nominal)
Edad	Edad del estudiante	Cuantitativa (Continua)
Fuma	¿Es fumador?	Cualitativa (Nominal)
Estatura	Altura del estudiante	Cualitativa (Ordinal)
Colegio	Tipo de colegio al que asiste	Cualitativa (Nominal)
Estrato	Nivel socioeconómico	Cualitativa (Ordinal)
Financiación	Tipo de financiación universitaria	Cualitativa (Nominal)
Acumulado	Promedio acumulado del semestre anterior	Cuantitativa (Continua)
Gastos	Promedio de gastos mensuales (en miles de $)	Cuantitativa (Continua)
Ingreso	Ingreso mensual del padre (en millones de $)	Cuantitativa (Continua)
Clases	¿Qué tipo de clase prefiere?	Cualitativa (Nominal)
Pandemia	¿Afectó económicamente la pandemia a su familia?	Cualitativa (Ordinal)
Clases_virtuales	Me preocupa no entender el contenido	Cualitativa (Ordinal)
Estadistica	Me gusta la estadística	Cualitativa (Ordinal)
inseguridad	Me siento inseguro al realizar ejecicios estadísticos	Cualitativa (Ordinal)
Vida_cotidiana	Utilizo la estadística en la vida cotidiana	Cualitativa (Ordinal)
Puntaje	Porcentaje de acierto en una determinada prueba	Cuantitativa (Discreta)

1.2. Tabla de frecuencia de la variable Ingresos

Para el análisis, se seleccionó la variable `Ingreso`, ya que es de tipo cuantitativa continua y permite aplicar diversos métodos estadísticos. Dado que esta variable presenta valores dispersos,

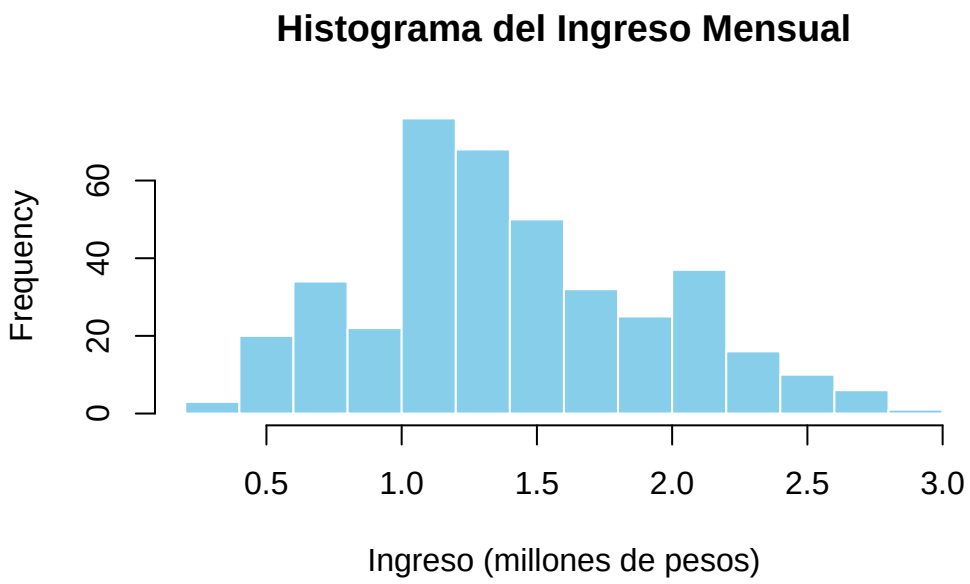
se agruparon los datos en intervalos para construir una tabla de frecuencias:

Table 2: Tabla de Frecuencia para la variable 'Ingreso

Clase	Marca	Frec_abs	Frec_rel	%	Frec_acum_abs	Frec_acum_rel
[0.358,0.609)	0.4835	23	0.0575	5.75	23	0.0575
[0.609,0.858)	0.7335	38	0.0950	9.50	61	0.1525
[0.858,1.11)	0.9840	48	0.1200	12.00	109	0.2725
[1.11,1.36)	1.2350	102	0.2550	25.50	211	0.5275
[1.36,1.6)	1.4800	62	0.1550	15.50	273	0.6825
[1.6,1.85)	1.7250	38	0.0950	9.50	311	0.7775
[1.85,2.1)	1.9750	38	0.0950	9.50	349	0.8725
[2.1,2.35)	2.2250	34	0.0850	8.50	383	0.9575
[2.35,2.6)	2.4750	10	0.0250	2.50	393	0.9825
[2.6,2.85)	2.7250	7	0.0175	1.75	400	1.0000

1.3. Gráficos

A continuación, se presentan dos gráficos representativos de la variable **Ingreso**, expresada en **millones de pesos**.



Histograma del ingreso mensual

El histograma permite observar la distribución de los ingresos de los estudiantes. Se aprecia una mayor concentración de observaciones entre $1.000.000 y $1.500.000, lo que concuerda con el valor de la mediana (1.32 millones). También se observan menos casos en los extremos, lo que sugiere una distribución ligeramente asimétrica hacia la derecha.

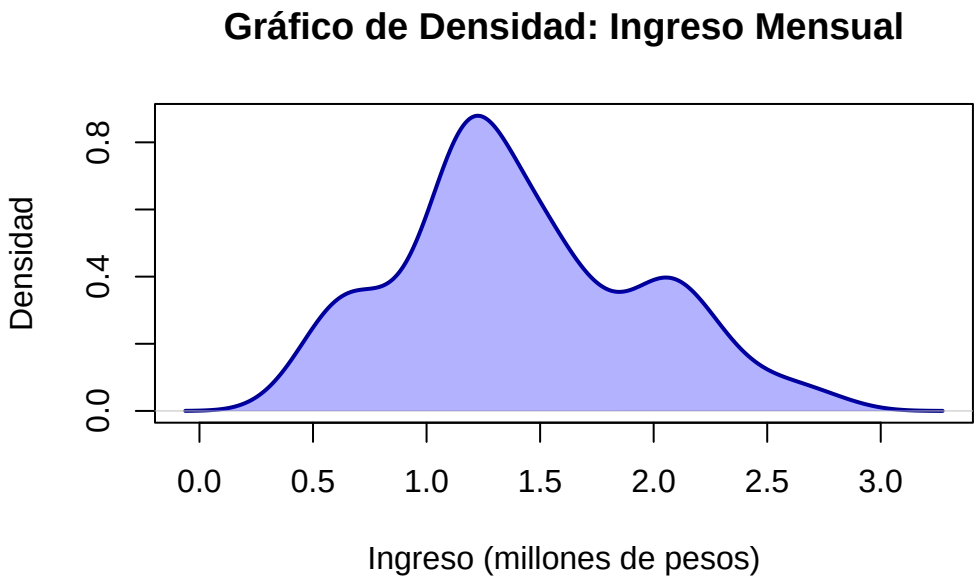


Gráfico de densidad del ingreso mensual

El gráfico de densidad ofrece una visualización suavizada de la distribución. Refuerza la impresión del histograma: la curva es ligeramente asimétrica hacia la derecha, con una leve cola en el extremo superior. Esto coincide con el valor positivo del coeficiente de asimetría (0.355).

1.4. Estadígrafo

Se analizan a continuación las principales medidas estadísticas descriptivas para la variable **Ingreso**, la cual representa el ingreso mensual de los estudiantes encuestados, expresado en **millones de pesos**.

Medidas de tendencia central

```
Media:  1.41
```

```
Mediana:  1.32
```

```
Moda:  1.16
```

- **Media:** 1.41 millones de pesos
  La media indica que, en promedio, los estudiantes reciben $1.410.000 mensuales.
- **Mediana:** 1.32 millones de pesos
  La mediana es menor a la media, lo que sugiere una ligera asimetría positiva en la distribución.
- **Moda:** 1.16 millones de pesos
  Es el ingreso más frecuente dentro de la muestra.

Medidas de dispersión

```
Rango:  0.36 - 2.85
```

```
Amplitud del rango:  2.49
```

```
Varianza:  0.29
```

```
Desviación estándar:  0.53
```

```
Coeficiente de variación:  37.87 %
```

- **Rango:** 0.36 a 2.85 millones de pesos
  Indica la diferencia entre el ingreso más bajo ($360.000) y el más alto ($2.850.000).
- **Amplitud del rango:** 2.49 millones de pesos
  Representa la extensión total de los ingresos observados.
- **Varianza:** 0.29
  Expresa la variabilidad de los ingresos respecto a la media.
- **Desviación estándar:** 0.53 millones de pesos
  En promedio, los ingresos se desvían $530.000 respecto de la media.
- **Coeficiente de variación:** 37.87%
  Indica un grado moderado de dispersión relativa. Un valor inferior al 50% sugiere una variabilidad aceptable respecto a la media.

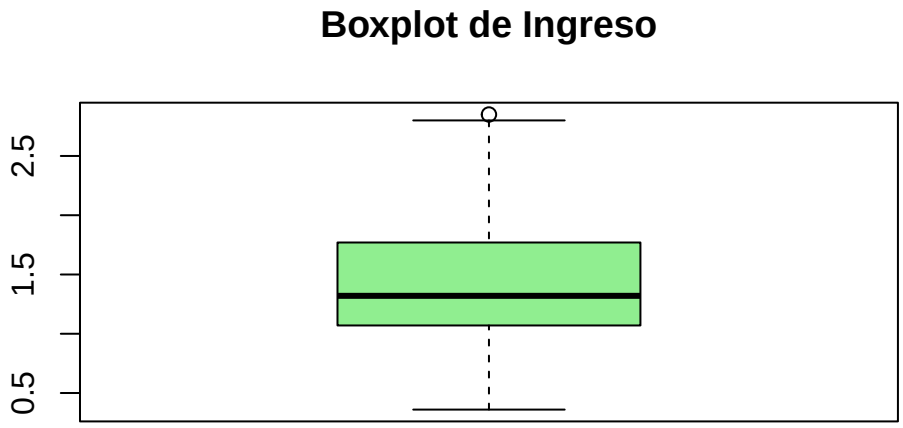
Medidas de forma

```
Asimetría:  0.355
```

```
Curtosis:  2.566
```

- **Asimetría (Skewness):** 0.355
  Positiva y cercana a cero, lo que indica que la distribución presenta una **ligera asimetría a la derecha**, es decir, existen algunos estudiantes con ingresos más altos que la mayoría.
- **Curtosis:** 2.566
  Inferior a 3, lo que indica una **distribución ligeramente platicúrtica** (menos concentrada en el centro y con colas más delgadas que una normal).

1.4. Posibles Outliers




```
Límite inferior:  0.03
```

```
Límite superior:  2.81
```

```
Cantidad de outliers:  1
```

```
Outliers:  2.85
```

A través del análisis gráfico mediante el boxplot y del cálculo del rango intercuartílico (IQR), se identificó un único valor atípico (outlier) en la variable *Ingreso*. Este valor sobrepasa el límite superior establecido por el método IQR, el cual considera como atípicos aquellos datos que se encuentran por encima de `Q3+1.5*IQR`. La presencia de este outlier fue claramente visible en el gráfico de caja (boxplot), lo que refuerza su clasificación como un valor extremo en comparación con la distribución general de los ingresos observados.

Aunque se trata de un único caso, su detección es relevante, ya que podría influir en algunas medidas estadísticas como la media y la varianza. Respecto a su tratamiento, la decisión de eliminarlo o conservarlo dependerá del objetivo del análisis. Si se busca describir el comportamiento general de la población sin la influencia de valores extremos, podría considerarse su exclusión. Sin embargo, si el ingreso elevado corresponde a una observación real y se desea mantener la representatividad del conjunto de datos, entonces sería apropiado conservarlo, dejando constancia de su presencia y posible impacto.

1.5. Conclusión

El análisis descriptivo de la variable *Ingreso*, expresada en millones de pesos, permitió obtener una caracterización clara y detallada del comportamiento de los datos dentro del conjunto observado.

Desde el punto de vista de las medidas de tendencia central, se observó que la **media** fue de 1.41 millones, la **mediana** de 1.32 millones y la **moda** de 1.16 millones. Esto sugiere una leve **asimetría positiva**, es decir, una distribución ligeramente sesgada hacia la derecha, lo que fue confirmado por el coeficiente de asimetría (0.355). En términos simples, la mayoría de los ingresos se concentran en valores más bajos, aunque hay algunos casos con ingresos más altos que elevan la media.

Respecto a la **dispersión**, se obtuvo una **desviación estándar** de 0.53 millones, lo que representa una **variabilidad moderada** en relación con la media. El **coeficiente de variación** fue de 37.87 %, indicando una dispersión significativa de los ingresos respecto a su promedio. La **curtosis** de 2.566 revela una distribución con colas ligeramente más delgadas que una distribución normal (platicúrtica), lo que sugiere que los valores extremos no son tan frecuentes, aunque sí se detectó **un outlier**.

Este **valor atípico** fue identificado mediante el boxplot y el método del rango intercuartílico (IQR), evidenciando un ingreso que excede el límite superior. Aunque es solo un caso, su existencia podría influir en algunas medidas estadísticas, por lo que se recomienda considerar el contexto del análisis para decidir si debe ser excluido o no.

En síntesis, la variable *Ingreso* presenta una distribución relativamente simétrica con ligera asimetría positiva, una variabilidad importante y una única observación atípica, lo que entrega información valiosa para futuras etapas del análisis estadístico.

2. Análisis de Probabilidades y Muestreo estratificado

2.1. Selección de muestra aleatoria simple

Se seleccionó una muestra aleatoria simple de tamaño $\mathbf{n = 150}$ de la base de datos original utilizando una semilla fija (`set.seed(123)`) para garantizar la reproducibilidad de los resultados. El análisis se centra en la variable **Ingreso**, expresada en millones de pesos.

Se calcularon los estadísticos descriptivos para esta variable

- Media del ingreso: 1.43 millones de pesos
- Desviación estándar: 0.54 millones de pesos

2.2. Cálculo de probabilidades

a) Probabilidad de que el ingreso sea a lo más de $1.500.000

Asumiendo que la variable **Ingreso** sigue una distribución normal, se estimó la probabilidad de que el ingreso de un estudiante universitario sea a lo más de $1.500.000 (es decir, 1.5 millones de pesos):

```
[1] 0.5501717
```

Resultado: 0.5502

Es decir, hay aproximadamente un 55.02% de probabilidad de que un estudiante tenga ingresos menores o iguales a $1.500.000.

b) Probabilidad de que el ingreso sea de más de $1.000.000

```
[1] 0.7879306
```

Resultado: 0.7879

Lo que corresponde a una probabilidad del 78.79% de que el ingreso de un estudiante supere el millón de pesos.

2.3. Muestreo estratificado por nivel socioeconómico

Se realizó un **muestreo estratificado** considerando la variable `Estrato`, que se divide en tres niveles: **1 = bajo**, **2 = medio**, y **3 = alto**. Se seleccionaron **50 observaciones aleatorias por cada estrato**, totalizando nuevamente **150 observaciones**.

Los resultados de media y varianza por estrato se presentan en la siguiente tabla:

Table 3: Media y varianza de ingreso por estrato socioeconómico

Estrato	media	varianza
1	1.37	0.31
2	1.45	0.23
3	1.43	0.26

Comparación con la muestra completa

A modo de comparación, en la muestra aleatoria simple completa (n=150), se obtuvo:

- Media general: 1.43
- Varianza general: 0.29

Esto permite observar si existen diferencias significativas entre los ingresos según el nivel socioeconómico de los estudiantes.

2.4. Conclusión del análisis de probabilidades y muestreo estratificado

A partir de una muestra aleatoria simple de 150 estudiantes, se analizaron las probabilidades asociadas a los ingresos mensuales (en millones de pesos), asumiendo una distribución normal. Los resultados muestran que existe una probabilidad cercana al **52.02%** de que el ingreso mensual sea igual o inferior a **$1.500.000**, y una probabilidad de aproximadamente **78.79%** de que dicho ingreso supere el **millón de pesos**. Estos valores permiten entender la concentración de los ingresos dentro del rango observado.

Posteriormente, al aplicar un **muestreo estratificado** según el nivel socioeconómico de los estudiantes, se obtuvieron 50 observaciones por estrato (bajo, medio y alto). Los resultados revelan diferencias en las medias y varianzas de ingreso entre los distintos estratos, lo que sugiere que el nivel socioeconómico influye en los ingresos mensuales reportados por los estudiantes. En particular, los estudiantes pertenecientes al **estrato alto** tienden a registrar mayores ingresos promedio, en comparación con los estratos bajo y medio.

Esta comparación entre la muestra general y los grupos estratificados refuerza la importancia de considerar el contexto socioeconómico al analizar variables financieras como el ingreso, permitiendo un análisis más detallado y representativo de la población estudiantil.

3. Problemas de probabilidad

a) Estrategias de marketing

Enunciado resumen:

> **Dos estrategias A y B tienen éxito en un 65% y 50% respectivamente. Actúan de forma independiente. ¿Cuál es la probabilidad de que un individuo compre debido a alguna de las dos?**

Desarrollo:

Queremos la probabilidad de que ocurra al menos una de las dos ventas.

Sea:

$$P(A) = 0.65$$

$$P(B) = 0.50$$

Como son eventos independientes:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) = P(A) x P(B) = 0.65 x 0.50 = 0.325$$

$$P(A \cup B) = 0.65 + 0.50 - 0.325 = 0.825$$

Resultado:

Considerando que las estrategias A y B actúan de forma independiente y tienen tasas de éxito del 65% y 50% respectivamente, se aplicó la fórmula de la probabilidad de la unión de dos eventos:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B))$$

Esto arrojó una probabilidad de **82.5%** de que un individuo realice una compra por alguna de las estrategias de marketing.

b) Probabilidad condicional y teorema de Bayes

Enunciado resumen:

> Si el reservorio es aprobado, hay un 90% de que construyan la granja. Si no es aprobado, hay un 2%. La probabilidad de aprobación es 60%.
>
> 1. ¿Cuál es la probabilidad de que construyan la granja?
> 2. Si se construyó, ¿cuál es la probabilidad de que el reservorio haya sido aprobado?

Desarrollo

Llamemos:

A: El reservorio es aprobado

¬A: No es aprobado

G: Se construye la granja

Datos:

$$P(G|A) = 0.90$$

$$P(G \mid \neg A) = 0.02$$

$$P(A) = 0.60, P(\neg A) = 0.40$$

1. Probabilidad de construir la granja:

$$P(G) = P(G \mid A) \cdot P(A) + P(G \mid \neg A) \cdot P(\neg A) = (0.90)(0.60) + (0.02)(0.04) = 0.54 + 0.008 = 0.548$$

Resultado:

Aplicando el teorema de probabilidad total, se determinó que la probabilidad de que la empresa construya la granja es de **54.8%**. Luego, utilizando el teorema de Bayes, se estimó que, dado que la granja fue efectivamente construida, existe una probabilidad del **98.45%** de que el reservorio haya sido aprobado.

c) Taller: Probabilidad condicional

Enunciado resumen:

> Por la mañana y tarde llegan vehículos con problemas eléctricos, mecánicos y de chapa. ¿Cuál es la probabilidad de que, si un vehículo llega por la tarde, sea por chapa?

Datos en tabla:

Problema	Mañana	Tarde	Total
Eléctrico	3	2	5
Mecánico	8	3	11
Chapa	3	1	4
Total	14	6	20

Desarrollo

Queremos:

$$P(Chapa|Tarde) = \frac{N°deautosporlatardeychapa}{Totalautosporlatarde} = \frac{1}{6} \approx 0.9845$$

Resultado:

Se construyó una tabla de contingencia considerando el tipo de problema y la jornada (mañana o tarde). Aplicando probabilidad condicional, se determinó que la probabilidad de que un vehículo asista por problemas de chapa, **dado que** lo hace en la tarde, es de aproximadamente **16.7%**.

4. Resolución de Problemas de Distribuciones de Probabilidad

a) Probabilidad y valor esperado con distribución binomial

Se sabe que un 15% de las dueñas de casa encuestadas en la comuna de Lo Espejo ha sido víctima de alguna estafa telefónica. Sea X la variable aleatoria que representa el número de víctimas de estafa telefónica en un grupo de 14 dueñas de casa encuestadas.

Entonces,

$$X \ B(n = 14, p = 0.15)$$

es decir, sigue una distribución binomial.

- **Probabilidad de que al menos 2 hayan sido estafadas:**

$$P(X >= 2) = 1 - P(X = 0) - P(X = 1)$$

$$P(X = 0) = (140)(0.15)0(0.85)14 \approx 0.1106$$

$$P(X = 1) = (141)(0.15)1(0.85)13 \approx 0.2723$$

$$P(X >= 2) = 1 - 0.1106 - 0.2723 = 0.6171$$

- **Valor esperado:**

$$E(X) = n \cdot p = 14 \cdot 0.15 = 2.1$$

Conclusión: La probabilidad de que al menos 2 dueñas de casa hayan sido estafadas es aproximadamente **0.6171**, y se espera que en promedio **2.1 mujeres** hayan sido víctimas.

b) Probabilidad con distribución de Poisson

La llegada de clientes a una caja de supermercado ocurre a una tasa promedio de 4 personas por minuto. En un intervalo de 2 minutos, la tasa esperada es:

$$\lambda = 4 \cdot 2 = 8$$

Sea:

$$X \sim Poisson(\lambda = 8)$$

Buscamos:

$$P(X >= 1) = 1 = P(X = 0) = 1 - \frac{e^8 X 8^0}{0!} = 1 - e - 8 \approx 1 - 0.000335 \approx 0.9997$$

Conclusión: La probabilidad de que al menos una persona llegue a la cola en 2 minutos es **0.9997**, prácticamente segura.

c) Probabilidad con distribución uniforme

La concentración de un contaminante se distribuye uniformemente en el intervalo [0,20]. Se considera tóxica a partir de los **8 millones**.

- **Probabilidad de toxicidad**:

$$P(X >= 8) = \frac{20-8}{20-0} = \frac{12}{20} = 0.6$$

- **Probabilidad exacta en 10 millones**:

  En una distribución continua uniforme, la probabilidad de obtener un valor exacto es **0**:

$$P(X = 10) = 0$$

  Conclusión: Hay un **60%** de probabilidad de que una muestra supere los niveles tóxicos, y la probabilidad de que la concentración sea exactamente de 10 millones es **0**.

5. Conclusión General

A lo largo de este informe se abordaron diversos aspectos del análisis estadístico aplicado al conjunto de datos *Estudiantes.xlsx*, así como problemas adicionales que requirieron el uso de distribuciones de probabilidad.

En primer lugar, se realizó un análisis descriptivo completo de la variable **Ingreso**, el cual permitió caracterizarla mediante medidas de tendencia central, dispersión y forma. Se observó una ligera asimetría positiva, una curtosis inferior a 3 (indicando menor concentración que una distribución normal), y se identificó un único **outlier por exceso** que fue visualizado mediante boxplot. El ingreso promedio se ubicó en torno a **$1.410.000**, con una mediana de **$1.320.000**, lo cual revela una distribución levemente sesgada hacia la derecha.

Posteriormente, se desarrolló un ejercicio de inferencia basado en una muestra aleatoria de 150 estudiantes. Se calcularon probabilidades bajo el supuesto de normalidad, y se compararon con los resultados obtenidos en un **muestreo estratificado por nivel socioeconómico (estrato)**. Las medias y varianzas por estrato arrojaron diferencias leves respecto a los parámetros de la muestra completa, sugiriendo que el estrato puede incidir, aunque no de manera drástica, en el nivel de ingreso.

En la tercera parte se resolvieron problemas de **probabilidad clásica y condicional**, incluyendo el uso de **reglas del complemento, la independencia, el teorema de Bayes** y

tablas de contingencia. Estas situaciones permitieron aplicar el enfoque teórico en contextos como el marketing, la planificación de proyectos y el análisis de decisiones.

Finalmente, se resolvieron ejercicios vinculados a **distribuciones discretas y continuas**. Se aplicaron correctamente los modelos **binomial, Poisson y uniforme**, permitiendo modelar fenómenos reales como estafas telefónicas, llegada de clientes y concentración de contaminantes. Estas herramientas fueron fundamentales para estimar probabilidades y valores esperados en distintos escenarios.

En conjunto, este trabajo permitió fortalecer la comprensión y aplicación de herramientas estadísticas tanto descriptivas como inferenciales, reafirmando su utilidad en la toma de decisiones basada en datos.