# INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

DIVISIÓN ACADÉMICA DE CIENCIAS EXACTAS

DEPARTAMENTO ACADÉMICO DE ESTADÍSTICA

CUADERNO DE EJERCICIOS

ESTADÍSTICA I

EST - 10101

Agosto 2026

ITAM

**INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO**

División Académica de Ciencias Exactas

Departamento Académico de Estadística

Cuaderno de ejercicios

Estadística I (EST-10101)

# Contenido

Liga a las bases de datos: XXXXXXX

**Introducción**

En concordancia con la **Visión 5.0 del ITAM** y con su compromiso con una formación integral y de excelencia, el presente **Cuaderno de Ejercicios de Estadística I** busca contribuir al desarrollo del **pensamiento analítico y crítico, el razonamiento cuantitativo, la resolución de problemas y la toma de decisiones sustentada en evidencia**, competencias fundamentales para la formación académica y profesional de los estudiantes.

Este cuaderno surge de la experiencia docente acumulada y del propósito de ofrecer un recurso que acompañe y fortalezca el proceso de aprendizaje. Los ejercicios que lo integran han sido seleccionados, revisados y adaptados considerando los objetivos del curso y las principales dificultades que enfrentan los estudiantes en la comprensión y aplicación de los conceptos estadísticos.

A lo largo del cuaderno se incorporan problemas contextualizados en situaciones reales o plausibles, con el propósito de favorecer un aprendizaje que trascienda la aplicación mecánica de fórmulas y procedimientos. Se busca que el estudiante desarrolle la capacidad de **comprender una situación, identificar la información relevante, traducirla a un planteamiento estadístico, seleccionar las herramientas adecuadas e interpretar los resultados en el contexto del problema**.

Más que una colección de ejercicios, este cuaderno se concibe como una **herramienta de aprendizaje y formación** orientada a promover una comprensión rigurosa de la Estadística, no solo como un conjunto de métodos y técnicas, sino como una forma de razonar a partir de la información, enfrentar la incertidumbre y sustentar decisiones de manera crítica y fundamentada.

**Liliana De la Torre Desentis**

**liliana.delatorre@itam.mx**

# TEMA 1: Introducción

## 1.1 Uso de datos en la solución de problemas y toma de decisiones

1.1.1. Un partido político analiza la percepción de confianza en instituciones públicas por grupo de edad. Los datos muestran que los jóvenes confían menos que los adultos mayores.
¿Qué información adicional sería útil recopilar para entender las causas de esta diferencia?
¿Cómo podrían utilizarse los datos recopilados para diseñar estrategias orientadas a mejorar la confianza entre los jóvenes?

1.1.2. Un estudiante registra durante un semestre las horas de estudio dedicadas a cada materia, las horas de sueño y sus calificaciones obtenidas en los exámenes. ¿Qué decisiones podría tomar utilizando estos datos para mejorar su rendimiento académico y organizar mejor su tiempo?

1.1.3. El gobierno de un país analiza información sobre los tratados comerciales firmados durante los últimos 10 años, incluyendo exportaciones, importaciones y crecimiento del intercambio comercial con distintos países. ¿Cómo podrían utilizarse estos datos para decidir con qué países conviene priorizar nuevas negociaciones comerciales?

1.1.4. Un organismo internacional recopila información sobre incidentes diplomáticos ocurridos entre países vecinos durante los últimos 5 años, clasificándolos según el tipo de conflicto y su frecuencia. ¿Cómo podrían utilizarse estos datos para anticipar posibles riesgos y diseñar estrategias de prevención de conflictos?

1.1.5. Una pareja registra durante seis meses sus gastos mensuales en categorías como alimentos, transporte, entretenimiento, servicios y compras en línea. ¿Cómo podrían utilizar estos datos para identificar en qué áreas están gastando más dinero y tomar decisiones para mejorar su presupuesto?

1.1.6. Un gerente analiza las ventas mensuales de distintos productos en una cadena de tiendas para identificar cuáles tienen mayor demanda y cuáles presentan bajas ventas. ¿Qué decisiones relacionadas con producción, inventario y promociones podrían tomarse a partir de estos datos?

1.1.7. Una empresa estudia la información de los clientes que cancelaron sus contratos durante el último año, incluyendo edad, tipo de servicio contratado, tiempo como cliente y número de quejas registradas. ¿Cómo podrían utilizarse estos datos para diseñar estrategias que ayuden a reducir la pérdida de clientes?

1.1.8. Una persona quiere elegir la mejor ruta para ir diariamente de su casa a la universidad. Durante varias semanas registra datos sobre tiempo de traslado, costo, tráfico y puntualidad utilizando diferentes rutas y medios de transporte. ¿Cómo podría utilizar estos datos para decidir cuál es la opción de transporte más conveniente?

1.1.9. Una empresa de bebidas está considerando lanzar al mercado una nueva línea de té saborizado. ¿Qué datos debería recopilar sobre consumidores, competencia y hábitos de compra para decidir si el lanzamiento sería conveniente?

1.1.10. Una empresa prueba durante un mes dos programas de gestión, A y B, en distintos departamentos de la organización. Durante el periodo de prueba se registran indicadores como tiempo de uso, errores del sistema, productividad y satisfacción de los empleados. ¿Qué información debería analizarse para decidir cuál software conviene comprar?

1.1.11. El gobierno de la Ciudad de México implementa una nueva línea de Metrobús y afirma que los tiempos de traslado se redujeron en un 25%. Un centro de investigación recopila datos de geolocalización de teléfonos móviles para auditar esta afirmación. Explica cómo el uso de estos datos masivos (Big Data) transforma la evaluación de políticas públicas en comparación con el uso de encuestas tradicionales de opinión.

1.1.12. Una empresa transnacional de consultoría con sede en México detecta un incremento del 15% en la tasa de rotación de su personal de auditoría tras implementar un modelo 100% presencial postpandemia. ¿Qué datos cuantitativos y cualitativos deberían recolectar antes de decidir si regresan a un esquema híbrido?

## 1.2 Concepto de variación y aleatoriedad. Relevancia de la calidad de los datos.

1.2.1 Cinco alumnos lanzan un dado justo seis veces cada uno y registran el número obtenido en cada lanzamiento. Los resultados varían tanto entre lanzamientos como entre estudiantes.

a) Aunque las condiciones del experimento son las mismas, ¿por qué los resultados pueden ser diferentes en cada lanzamiento?

b) Si un estudiante lanzara el dado 100 veces, ¿esperarías que cada número apareciera exactamente el mismo número de veces? Explica tu respuesta.

c) Si el dado estuviera cargado de manera que el 6 tuviera una probabilidad mucho mayor de aparecer, ¿la variación observada podría atribuirse únicamente a la aleatoriedad?

1.2.2 Diez encuestadores aplican el mismo cuestionario sobre la confianza en el gobierno a ciudadanos en zonas similares. Aun así, los porcentajes de confianza obtenidos son ligeramente distintos entre encuestadores.

a) ¿Por qué los resultados pueden diferir, aunque el cuestionario y las condiciones sean prácticamente iguales?

b) ¿Las diferencias observadas corresponden a variación natural, variación aleatoria o ambas? Justifica tu respuesta.

c) Si uno de los encuestadores registra sistemáticamente niveles de confianza más altos que los demás, ¿seguiría siendo un caso de variación aleatoria? ¿Qué problema relacionado con la calidad de los datos podría existir?

1.2.3 Seis médicos miden la estatura de un mismo atleta utilizando instrumentos similares, pero obtienen resultados ligeramente distintos.

a) ¿Por qué pueden existir pequeñas diferencias en las mediciones, si se trata de la misma persona?

b) ¿A qué fuentes de variación podrían atribuirse estas diferencias: variación real de la característica medida, error aleatorio de medición o diferencias sistemáticas entre instrumentos u observadores? Explica tu respuesta.

c) ¿Qué acciones podrían tomarse para mejorar la precisión y calidad de las mediciones?

1.2.4 Cinco jugadores de un equipo realizan 20 tiros libres cada uno. El número de encestes es distinto entre los jugadores.

a) ¿Qué factores podrían explicar las diferencias observadas entre los jugadores? Distingue entre diferencias reales de habilidad y variación aleatoria debida al resultado de los 20 tiros.

b) Si quisieras estudiar la diferencia real de habilidad entre jugadores, ¿cómo diseñarías el experimento para reducir el efecto de la aleatoriedad?

c) ¿Por qué sería importante registrar correctamente todos los tiros realizados y encestados?

1.2.5 Cinco termómetros colocados en el mismo cuarto marcan temperaturas ligeramente distintas.

a) Aunque los termómetros se encuentren en el mismo cuarto, ¿qué factores podrían explicar las diferencias observadas? Distingue entre pequeñas variaciones reales de temperatura y diferencias debidas a los instrumentos o al proceso de medición.

b) ¿Las diferencias observadas podrían deberse a errores de medición o problemas en la calidad de los instrumentos? Explica tu respuesta.

c) ¿Cómo podrías determinar cuál de los termómetros es el más confiable?

1.2.6 Ocho embajadas de un mismo país registran el tiempo promedio que tardan en responder solicitudes rutinarias de ciudadanos. Los tiempos promedio difieren entre embajadas.

a) ¿Por qué pueden existir diferencias en los tiempos de respuesta, aunque el procedimiento oficial sea el mismo? Explica tu respuesta

b) ¿Las diferencias observadas representan variación natural, variación aleatoria o ambas? Explica tu respuesta.

c) Si quisieras detectar posibles errores o inconsistencias en los registros de alguna embajada, ¿qué indicadores revisarías para evaluar la calidad de los datos?

1.2.7 Durante una campaña electoral para la jefatura de gobierno, una casa encuestadora decide realizar su muestreo exclusivamente a través de anuncios pagados en Instagram y TikTok, obteniendo 50,000 respuestas en tres días.

a) ¿Por qué un volumen tan grande de datos no garantiza la representatividad?

b) ¿Cómo afecta el “sesgo de selección” a la calidad de los datos para predecir el resultado real de la población votante?

1.2.8 Un corporativo de supermercados nota que el precio de adquisición del aguacate Hass varía drásticamente semana a semana, impactando los márgenes de ganancia estimados. Al revisar los registros, descubren que algunos proveedores reportan precios en pesos mexicanos y otros en dólares, y que las fechas de registro no coinciden con las de entrega.

a) ¿Cómo afectan la calidad de los datos la falta de estandarización y los errores de registro?

b) Distingue entre la variación natural del precio del mercado debido a la oferta y demanda y la variación artificial introducida por la mala calidad de los datos contables.

## 1.3 Entendimiento de la variación empleando la probabilidad y la estadística

1.3.1 Un dado perfectamente balanceado se lanza 60 veces y el número “6” aparece en 8 ocasiones.

a) Calcula la proporción observada de resultados iguales a 6 y compárala con la probabilidad teórica $\frac{1}{6}$. ¿Cuál es la diferencia?

b) Explica por qué una diferencia entre el resultado observado y el valor teórico no implica necesariamente que el dado esté sesgado.

c) ¿Por qué este tipo de diferencias se consideran parte de la variación aleatoria esperada en experimentos probabilísticos?

d) Si el experimento se repitiera con 600 lanzamientos en lugar de 60, ¿esperarías que la proporción observada estuviera más cerca o más lejos del valor teórico? Explica tu respuesta.

1.3.2 Se mide el rendimiento de combustible (km por litro) de 15 autos idénticos y se obtiene un promedio de 14.8 km/l con una variación de ± 0.7 km/l.

a) Aunque los autos son iguales, ¿por qué el rendimiento puede variar entre ellos?

b) Clasifica las siguientes posibles causas de variación como principalmente sistemáticas o aleatorias: peso del conductor, estilo de manejo, viento, calidad del combustible.

c) Explica por qué no es posible esperar exactamente el mismo rendimiento en todos los autos.

d) ¿Qué nos indica este ejemplo sobre la importancia de analizar datos considerando variabilidad e incertidumbre?

1.3.3 En una encuesta electoral, el 52% de los entrevistados declara preferir al candidato A. El estudio reporta un margen de error de ± 3 puntos porcentuales.

a) ¿Por qué diferentes encuestas realizadas sobre la misma elección pueden producir resultados ligeramente distintos?

b) ¿Las diferencias entre encuestas se deben necesariamente a errores o fraude? Explica tu respuesta.

c) ¿Cómo ayuda el margen de error a entender la incertidumbre presente en los resultados de una encuesta?

d) ¿Qué condiciones deben cumplirse para confiar en que la variación observada proviene principalmente del azar y no de sesgos en la recolección de datos?

e) ¿Por qué la estadística permite hacer estimaciones útiles aun cuando no puede garantizar certeza absoluta sobre el resultado real de la elección?

1.3.4 Una empresa automotriz alemana planea abrir una planta en Nuevo León. Para proyectar sus costos operativos en los próximos 5 años, analiza la variación histórica del tipo de cambio peso/euro.

a) ¿Por qué la empresa no puede asumir un tipo de cambio fijo para sus proyecciones financieras?

b) Explica cómo la estadística y la probabilidad permiten a los administradores “cuantificar la incertidumbre” y crear escenarios (optimista, intermedio, pesimista) para blindar la inversión.

1.3.5 Un estudiante de Ciencia Política analiza el impacto de un programa social de becas en la reducción de la deserción escolar. Al observar los datos de 10 municipios, nota que en algunos la deserción bajó 8%, en otros se mantuvo igual y en uno aumentó 2%.

Utiliza los conceptos de variación para explicarle a un político (que argumenta que el programa es un fracaso total debido al municipio donde aumentó la deserción) por qué es un error juzgar una política pública basándose en un solo punto de datos aislado en lugar de analizar la distribución completa de la variación.

1.3.6 El departamento de control interno de una Fintech detecta que el monto promedio de las compras con tarjeta de crédito tiene cierta media y desviación estándar. De pronto, un usuario realiza 15 transacciones consecutivas de montos idénticos en un lapso de 10 minutos.

Explica cómo el entendimiento de la variación estadística ayuda a los sistemas automatizados a identificar comportamientos "atípicos" (outliers). ¿Cómo interviene la probabilidad para determinar si un patrón de transacciones es legítimo o un fraude potencial?

## 1.4 Poblaciones y muestras aleatorias

1.4.1 Un instituto de investigación desea conocer la opinión de los ciudadanos sobre una nueva política pública. Debido a limitaciones de tiempo y costo, encuesta a 1,000 personas seleccionadas aleatoriamente.

a) Identifica la población de estudio.

b) Identifica la muestra.

c) Explica por qué no es necesario encuestar a toda la población para obtener información útil sobre la opinión pública.

d) ¿Por qué los resultados obtenidos en la muestra pueden diferir ligeramente de la opinión real de toda la población?

e) ¿Qué acciones podrían ayudar a reducir la variabilidad y mejorar la representatividad de la muestra?

1.4.2 Una empresa desea estudiar cuántas horas al día pasan sus clientes en redes sociales para diseñar campañas de marketing digital. Para ello, selecciona aleatoriamente a 500 usuarios.

a) Identifica la población de estudio.

b) Identifica la muestra.

c) Si la muestra indica un promedio de 3.5 horas diarias, ¿puede concluirse que todas las personas de la población pasan exactamente ese mismo tiempo en redes sociales? Explica tu respuesta.

d) Si se seleccionara otra muestra distinta de 500 usuarios, ¿esperarías obtener exactamente el mismo promedio? ¿Por qué?

1.4.3 El jefe de gobierno de la CDMX desea conocer la opinión de los ciudadanos sobre el uso de transporte eléctrico. Para ello, se realiza una encuesta en línea a 2,000 personas.

a) Identifica la población de estudio.

b) Identifica la muestra.

c) Si el 70% de las personas encuestadas apoya la propuesta, ¿puede asegurarse que exactamente el 70% de toda la población también la apoya? Explica tu respuesta.

d) ¿Qué posibles sesgos podrían afectar la representatividad de una encuesta realizada únicamente por internet?

e) ¿Por qué es importante que la muestra sea representativa de la población?

1.4.4 La Secretaría de Relaciones Exteriores desea conocer la percepción sobre la seguridad turística en México entre los adultos residentes en Estados Unidos para diseñar una campaña de promoción.

a) Define la población objetivo.

b) Propón un método de muestreo que cumpla con el principio de que cada individuo tenga una probabilidad conocida y mayor a cero de ser elegido (muestra aleatoria).

c) Explica las barreras logísticas y de costos que harían imposible estudiar a toda la población.

1.4.5 Una firma internacional de auditoría debe revisar la validez de 500,000 facturas emitidas por un gigante tecnológico durante el año fiscal. Revisarlas todas tomaría meses y costaría millones de pesos.

a) Identifica la población objeto de estudio.

b) Explica detalladamente cómo una muestra aleatoria estadísticamente representativa permite a los auditores emitir un dictamen confiable sobre la totalidad de los estados financieros, reduciendo costos y tiempos de ejecución.

1.4.6 Tres alumnos del ITAM quieren lanzar una app de microcréditos para estudiantes universitarios de la Zona Metropolitana del Valle de México (ZMVM). Para estimar la demanda, deciden aplicar una encuesta únicamente a sus amigos y compañeros dentro del campus del ITAM.

a) Identifica la población objetivo real del negocio

b) Identifica la muestra que los alumnos seleccionaron.

c) Argumenta por qué esta muestra se considera "de conveniencia" y no "aleatoria", y cuáles son los riesgos para el plan de negocios si toman decisiones asumiendo que los datos obtenidos reflejan el comportamiento de toda la población universitaria de la ZMVM.

# TEMA 2: Análisis exploratorio de datos

## 2.1 Tipos de datos, clasificación y escalas de medición

2.1.1 Clasifica cada una de las siguientes variables por tipo de dato y escala de medición:

| | Tipo de dato | | | Escala de medición | | | |
|---|---|---|---|---|---|---|---|
| Variable | cualitativa | cuantitativa discreta | cuantitativa continua | nominal | ordinal | de intervalo | de razón |
| Tiempo diario de uso de TikTok (minutos) | | | | | | | |
| Nivel de batería del celular (%) | | | | | | | |
| Tipo de chatbot utilizado (IA generativa, asistente personal, servicio al cliente) | | | | | | | |
| Calidad del sueño medida por smartwatch (mala, regular, buena, excelente) | | | | | | | |
| Número de podcasts escuchados por semana | | | | | | | |
| Temperatura corporal según smartwatch (°C) | | | | | | | |
| Categoría del vehículo de Uber (Uber X, Comfort, Black) | | | | | | | |
| Velocidad de conexión a Internet (Mbps) | | | | | | | |
| Nivel de estrés detectado por anillos inteligentes (bajo, medio, alto) | | | | | | | |
| Emisiones de $CO_2$ de un viaje en avión (kg) | | | | | | | |
| Horas de estudio en plataformas como Coursera o Udemy por mes | | | | | | | |
| Género musical favorito en Spotify | | | | | | | |
| Número de pasos registrados por el celular en un día | | | | | | | |
| Puntaje de riesgo crediticio (0–850 puntos) | | | | | | | |
| Nivel de suscripción a Netflix (Básico, Estándar, Premium) | | | | | | | |
| Distancia recorrida en bicicleta eléctrica (km) | | | | | | | |
| Frecuencia cardiaca en reposo medida por smartwatch (latidos/min) | | | | | | | |
| Categoría de producto comprado en Amazon (electrónicos, hogar, moda, alimentos, etc.) | | | | | | | |
| Intensidad de ruido ambiental medida por el celular (dB) | | | | | | | |
| Calificación de una app en Google Play (1–5 estrellas) | | | | | | | |
| Porcentaje de avance en un curso online | | | | | | | |
| Cantidad de correos electrónicos recibidos al día | | | | | | | |
| Tiempo de carga de una batería de auto eléctrico (horas) | | | | | | | |
| Nivel de satisfacción con un servicio (muy bajo, bajo, medio, alto, muy alto) | | | | | | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Número de seguidores nuevos en Instagram por semana | | | | | | | |
| Kilogramos de residuos reciclados por mes | | | | | | | |
| pH del agua potable (escala de 0 a 14) | | | | | | | |
| Rol del usuario en una plataforma (estudiante, creador, administrador) | | | | | | | |
| Porcentaje de batería consumido por apps de IA | | | | | | | |
| Consumo energético del hogar según medidor inteligente (kWh/mes) | | | | | | | |
| Duración de videollamadas en Zoom (min) | | | | | | | |
| Nivel de severidad de errores en un software (bajo, medio, alto, crítico) | | | | | | | |
| Cantidad de “likes” recibidos en la última publicación | | | | | | | |
| Latencia de un servidor en videojuegos online (ms) | | | | | | | |
| Cantidad de vidrio/plástico en envases comprados en el supermercado (g) | | | | | | | |
| Categoría del nivel de riesgo sísmico en la zona donde vive (bajo, moderado, alto) | | | | | | | |
| Porcentaje de uso de CPU por aplicaciones activas | | | | | | | |
| Número de entregas realizadas por un repartidor de apps (UberEats / Rappi) | | | | | | | |
| Gasto mensual en servicios de streaming (MXN) | | | | | | | |
| Tiempo de espera para recibir un pedido de comida (minutos) | | | | | | | |

2.1.2 Clasifica cada una de las siguientes variables por tipo de dato y escala de medición:

| | | Tipo de dato | | | Escala de medición | | | |
|---|---|---|---|---|---|---|---|---|
| | Variable | cualitativa | cuantitativa discreta | cuantitativa continua | nominal | ordinal | de intervalo | de razón |
| Ciencia Política | Partido político con el que se identifica un votante | | | | | | | |
| | Nivel de confianza en las instituciones (1 = nada, 5 = totalmente) | | | | | | | |
| | Edad del candidato | | | | | | | |
| | Número de diputados en el Congreso por partido | | | | | | | |
| | Año de nacimiento de los senadores | | | | | | | |
| | Porcentaje de participación electoral en una elección | | | | | | | |
| R | País de origen del embajador | | | | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Índice de Desarrollo Humano (IDH) | | | | | | | |
| | Número de tratados internacionales firmados por un país | | | | | | | |
| | Rango de poder militar (alto, medio, bajo) | | | | | | | |
| | Número de conflictos fronterizos activos en un año | | | | | | | |
| Administración | Área funcional en la que trabaja un empleado (Finanzas, Marketing, Operaciones, RRHH) | | | | | | | |
| | Satisfacción laboral (escala Likert de 1 a 7) | | | | | | | |
| | Año de creación de las empresas | | | | | | | |
| | Salario mensual en pesos | | | | | | | |
| | Número de ventas cerradas en un mes | | | | | | | |
| | Años de antigüedad en la empresa | | | | | | | |
| | Puntaje de empleados en un test de liderazgo | | | | | | | |
| | Nivel de satisfacción del cliente (1 a 10) | | | | | | | |
| Contabilidad | Ingreso mensual de una empresa (en pesos) | | | | | | | |
| | Número de facturas emitidas por semana | | | | | | | |
| | Tipo de auditoría realizada (interna, externa, fiscal) | | | | | | | |
| | Método de depreciación utilizado por la empresa | | | | | | | |
| | Monto de cuentas por cobrar vencidas (en pesos) | | | | | | | |
| | Nivel de riesgo financiero asignado a una empresa (bajo, medio, alto) | | | | | | | |
| | Tiempo promedio de pago a proveedores (en días) | | | | | | | |
| | Número de empleados del departamento contable | | | | | | | |

## 2.2 Distribuciones de frecuencias para datos cualitativos y cuantitativos: diagramas de barras, diagramas circulares, histogramas, polígono de frecuencias y ojiva

2.2.1 Una universidad desea entender cómo sus estudiantes consumen contenido digital, con el fin de diseñar mejores estrategias de comunicación. Se aplicó una encuesta a 120 estudiantes sobre dos variables:

- Plataforma principal usada para informarse
- Minutos diarios dedicados a consumir noticias

Los resultados fueron los siguientes:

| ***Plataforma principal:*** | ***No. Alumnos*** |
|---|---|
| TikTok | 36 |
| X | 30 |
| Instagram | 26 |
| Foros / blogs especializados | 10 |
| Medios digitales formales (El País, The New York Times, BBC, etc.) | 18 |

| ***Minutos al día*** | ***Frecuencia*** |
|---|---|
| (0 – 10] | 8 |
| (10 – 20] | 15 |
| (20 – 30] | 25 |
| (30 – 40] | 32 |
| (40 – 50] | 20 |
| (50 – 60] | 12 |
| (60 – 70] | 8 |

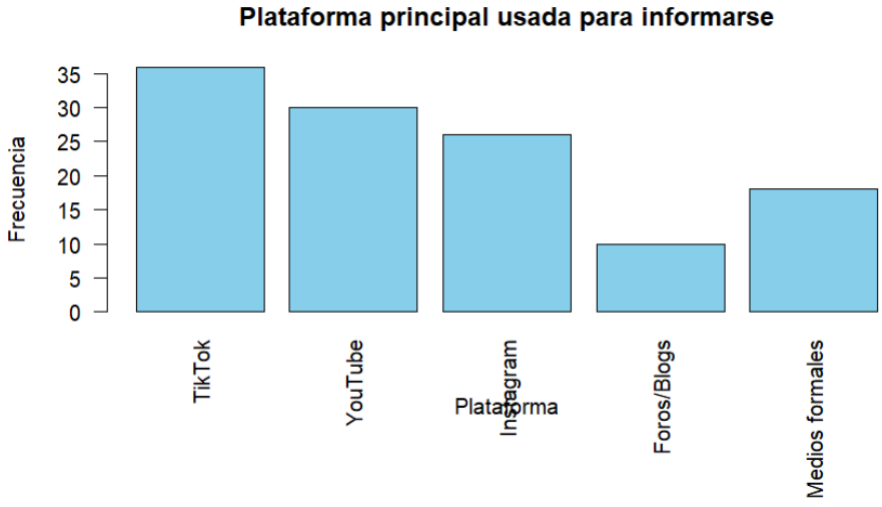
Para la variable “Plataforma principal usada para informarse” grafica:

- Diagrama de barras
- Diagrama circular

Para la variable “Minutos diarios dedicados a consumir noticias” grafica:

- Histograma
- Polígono de frecuencias
- Ojiva

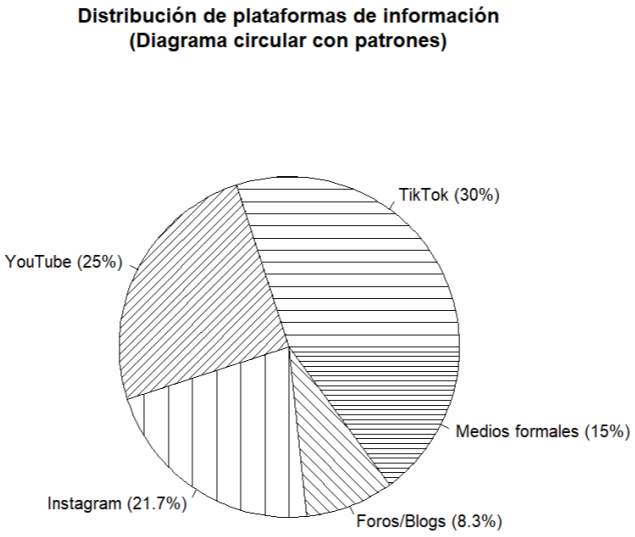
A partir de las gráficas obtenidas, responde las preguntas:



a) ¿Qué tan grande es la diferencia en número de estudiantes entre la plataforma más utilizada y la menos utilizada en relación con el tamaño total de la muestra?

b) ¿Cuántos estudiantes tendrían que cambiar de TikTok a Foros/Blogs para que ambas plataformas tuvieran la misma frecuencia?

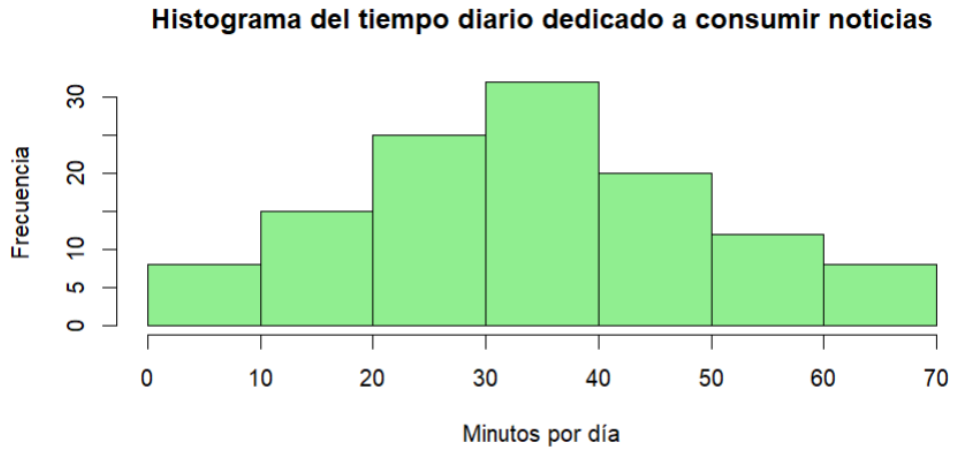
c) ¿Qué plataforma representa la mayor oportunidad de crecimiento para una campaña de comunicación institucional? Explica tu respuesta utilizando las frecuencias observadas.




a) Si la universidad solo pudiera difundir información en dos plataformas, ¿cuáles elegirías para alcanzar al mayor número de estudiantes? Justifica tu respuesta con los datos del gráfico.

b) Si la universidad decide no utilizar TikTok para difundir información, ¿qué porcentaje de estudiantes podría alcanzar utilizando únicamente las otras plataformas?

c) Supongamos que la universidad actualmente publica información únicamente en medios digitales formales. Con base en el gráfico, ¿consideras que esta estrategia es adecuada para alcanzar a la mayoría de los estudiantes? Explica tu respuesta.



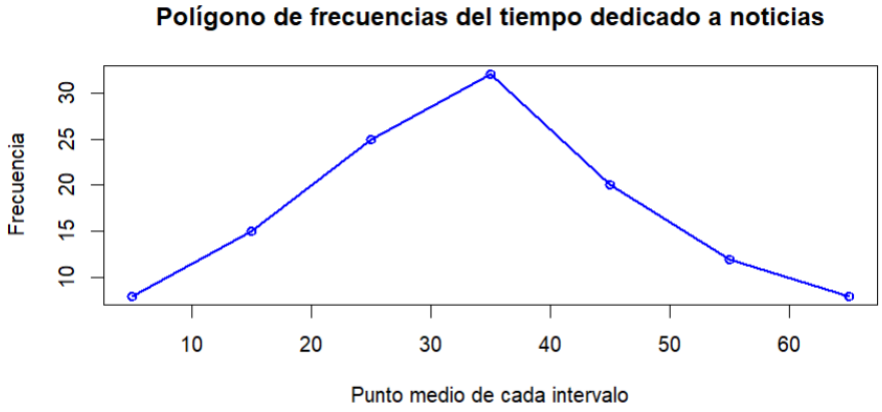

a) ¿En qué intervalo se concentra la mayor parte del tiempo de consumo?

b) ¿La distribución parece simétrica, sesgada a la derecha o a la izquierda? ¿Qué significa esta forma?

c) ¿Qué intervalo presenta un cambio brusco en la frecuencia?

d) ¿La variable es continua o discreta? ¿Por qué se usa histograma y no una gráfica de barras?

e) Si se agruparan los datos en intervalos de 15 minutos, ¿cambiaría la interpretación?



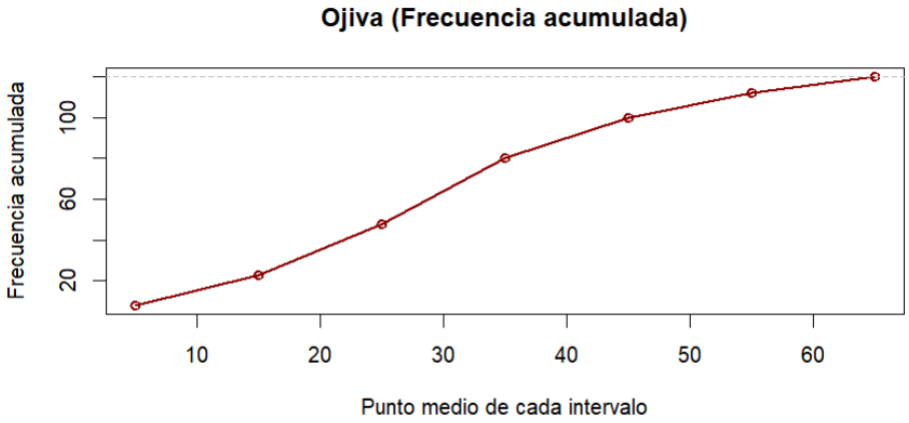

a) ¿En qué punto el polígono alcanza su máximo? ¿Qué significa este punto en el contexto del problema?

b) ¿Cuál es la forma general del polígono?

c) ¿En qué intervalo la frecuencia aumenta más rápidamente?

d) ¿Qué representa el área bajo el polígono?

e) ¿Qué patrón sugeriría un polígono con múltiples picos?




a) ¿Cuántos estudiantes consumen 30 minutos o menos de noticias al día?

b) ¿Cuál es el percentil 50 del tiempo dedicado a informarse?

c) ¿En qué intervalo ocurre el mayor incremento acumulado?

d) ¿Cuántos estudiantes consumen más de 50 minutos diarios?

e) ¿Cómo usarías la ojiva para identificar datos atípicos?

2.2.2 La Secretaría de Movilidad de la CDMX realiza un estudio para entender cómo se desplazan los habitantes después de los cambios recientes en infraestructura, nuevas líneas de transporte y el aumento del uso de vehículos compartidos. Se encuesta a 150 personas y se registran dos variables:

- Medio de transporte predominante
- Tiempo total de traslado diario (minutos)

Los resultados fueron los siguientes:

| ***Medio de transporte predominante*** | ***No. Personas*** |
|---|---|
| Metro | 42 |
| Metrobús | 30 |
| Automóvil particular | 28 |
| Aplicaciones de movilidad (Uber, Didi, Beat) | 20 |
| Bicicleta / Ecobici | 15 |
| Caminando | 15 |

| *Tiempo* | *Frecuencia* |
|---|---|
| 0–20 | 12 |
| 20–40 | 20 |
| 40–60 | 35 |
| 60–80 | 40 |
| 80–100 | 25 |
| 100–120 | 18 |

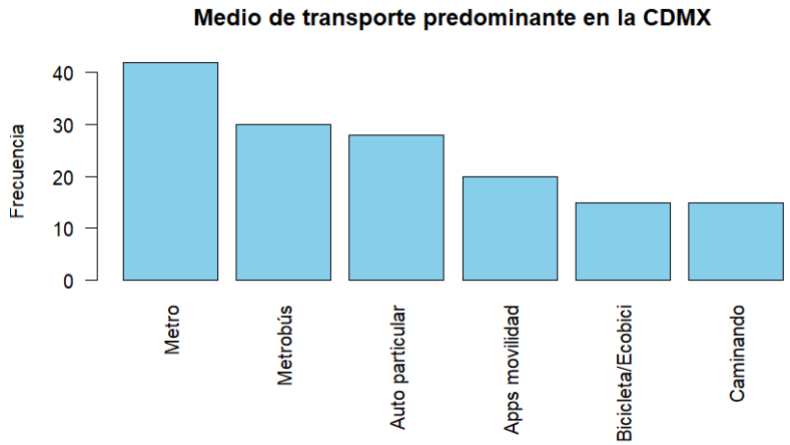
Para la variable “Medio de transporte predominante” grafica:

- Diagrama de barras
- Diagrama circular

Para la variable “Tiempo total de traslado diario” grafica:

- Histograma
- Polígono de frecuencias
- Ojiva

A partir de las gráficas obtenidas, responde las preguntas:



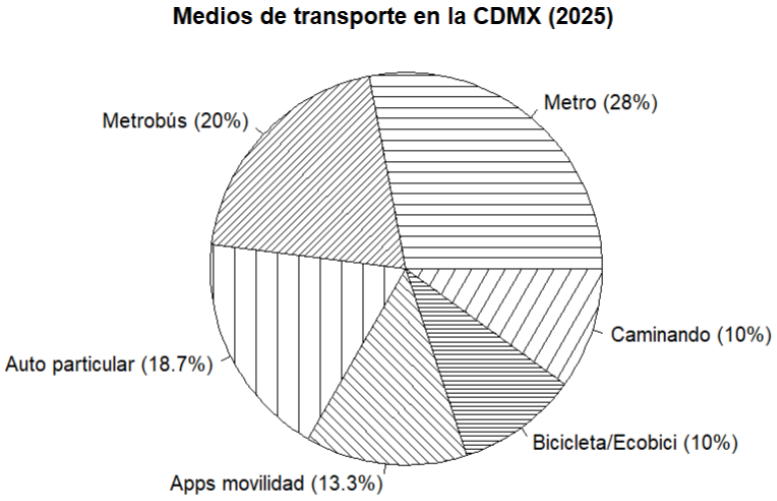

a) ¿Qué medio de transporte utilizan con mayor frecuencia las personas encuestadas?

b) ¿Cuántas personas más utilizan el Metro que el automóvil particular como principal medio de transporte? ¿Qué indica esta diferencia sobre los patrones de movilidad observados?

c) ¿Qué categoría representa la mayor oportunidad para promover políticas de movilidad sustentable? Justifica tu respuesta utilizando las frecuencias observadas.

d) ¿Consideras que un gráfico de barras apiladas sería una representación adecuada para estos datos? Explica por qué sí o por qué no.

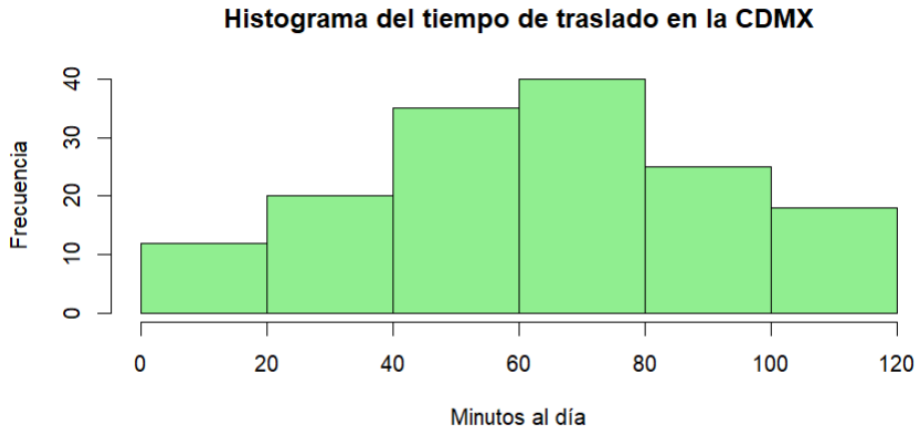
e) Las categorías con menor frecuencia son bicicleta/ecobici y caminando. ¿Qué podrían indicar estos resultados sobre los hábitos de movilidad de la población estudiada?




a) ¿Qué porcentaje de las personas encuestadas utiliza el Metro como principal medio de transporte?

b) ¿Qué medio de transporte concentra aproximadamente una quinta parte (20%) de los usuarios?

c) Si fueras responsable de la Secretaría de Movilidad (SEMOVI), ¿qué acciones o políticas podrías proponer a partir de la distribución observada de los medios de transporte? Justifica tu respuesta utilizando los porcentajes del gráfico.

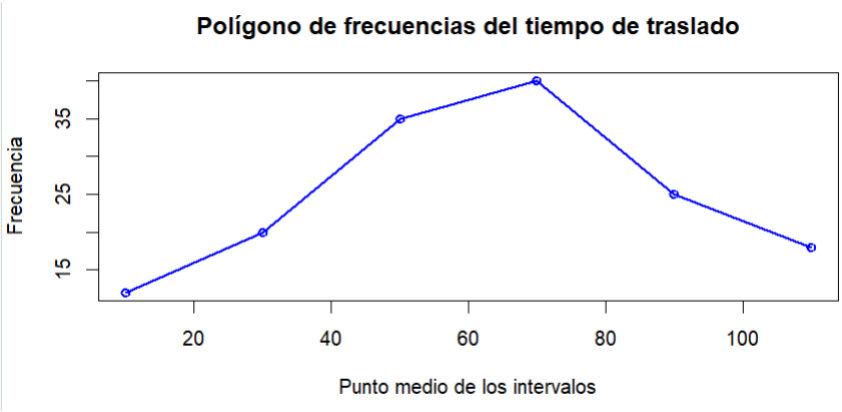



a) ¿En qué intervalo se concentran los tiempos de traslado más comunes?

b) Al observar la forma del histograma, ¿la distribución parece presentar sesgo hacia los tiempos de traslado más cortos o hacia los más largos? ¿Qué indica esto sobre los hábitos de movilidad de la población?

c) ¿Por qué un histograma es una representación más adecuada que una gráfica de barras para analizar el tiempo de traslado?

d) ¿La distribución observada sugiere que la mayoría de las personas experimenta tiempos de traslado moderados o extremos? Explica qué implicaciones podría tener esto para la planeación del transporte urbano.



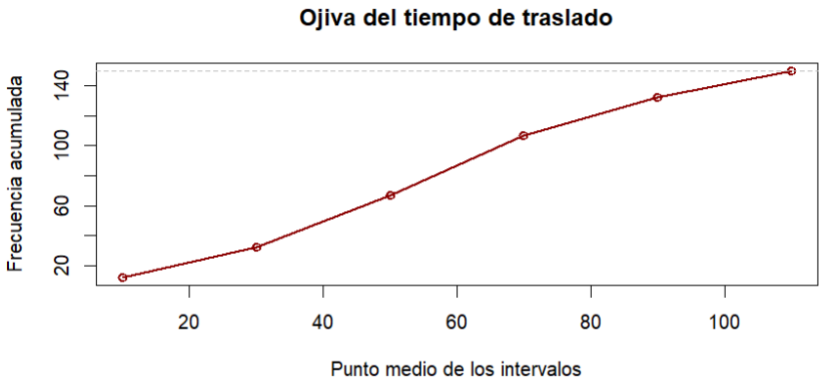

a) ¿En qué punto del polígono se observa la frecuencia más alta? ¿Qué indica este valor sobre los tiempos de traslado de la población estudiada?

b) ¿La distribución presenta un único pico principal (unimodal) o varios picos (multimodal)? ¿Qué sugiere esto sobre los patrones de tiempo de traslado de las personas?

c) ¿Entre qué intervalos de tiempo se observa el mayor incremento en la frecuencia? ¿Qué indica este cambio sobre la concentración de los datos?

d) ¿Qué representa, de manera general, el área bajo el polígono de frecuencias en relación con la población estudiada?

e) ¿Qué conclusiones sobre los hábitos de movilidad de la población pueden obtenerse a partir de la forma general del polígono de frecuencias?




a) Según la ojiva, ¿cuántas personas tienen un tiempo de traslado diario de 60 minutos o menos?

b) ¿Cuál es el tiempo de traslado por debajo del cual se encuentra aproximadamente el 75% de las personas?

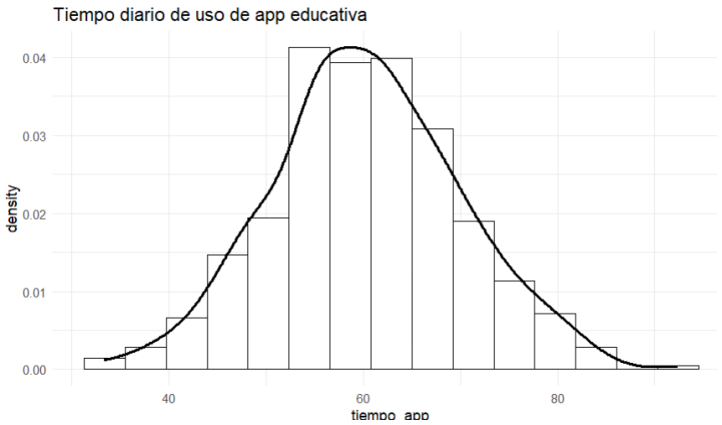
c) ¿En qué tramo de tiempos de traslado aumenta más rápidamente la frecuencia acumulada? ¿Qué indica esto sobre la concentración de los datos?

d) ¿Cuántas personas tienen tiempos de traslado superiores a 100 minutos diarios?

e) ¿Puede utilizarse la ojiva para identificar posibles tiempos de traslado inusualmente altos o bajos? Explica brevemente cómo.

## 2.3 Curvas de distribuciones de frecuencias (formas características)

2.3.1 La siguiente curva de frecuencias representa el tiempo diario (en minutos) que los alumnos de una escuela primaria dedican al uso de una aplicación educativa. Con base en la gráfica, responde las siguientes preguntas:



a) ¿La distribución parece centrada o presenta algún grado de asimetría? Explica tu respuesta a partir de la forma de la curva.

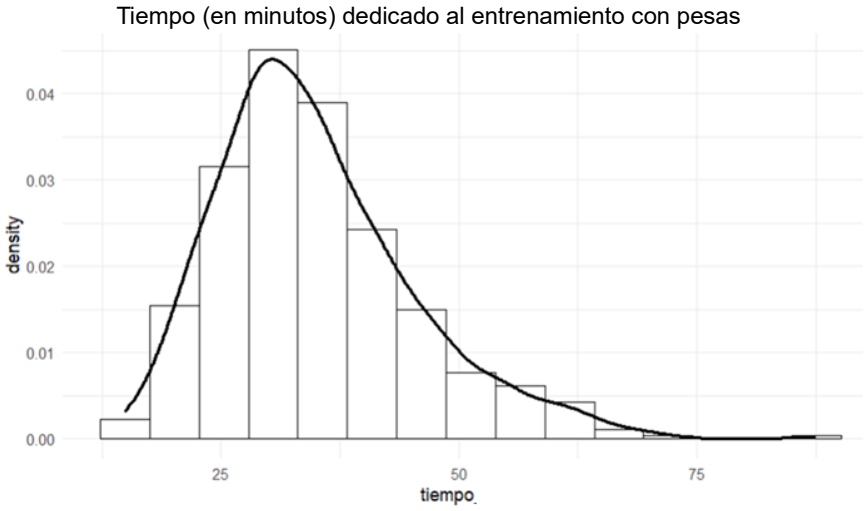
b) A partir de la forma de la distribución, ¿esperarías que la media y la mediana tuvieran valores similares? ¿Qué indica esto sobre los tiempos de uso de la aplicación?

c) ¿La gráfica muestra evidencia de tiempos de uso inusualmente bajos o altos? Explica cómo lo identificas.

d) ¿La forma de la curva sugiere que todos los alumnos tienen patrones de uso similares o que existen grupos con comportamientos diferentes? Justifica tu respuesta.

e) ¿En qué intervalo de tiempo se concentra la mayor cantidad de alumnos? ¿Qué indica esto sobre el tiempo de uso más común de la aplicación?

2.3.2 La distribución del tiempo (en minutos) que los usuarios de un gimnasio dedican al entrenamiento con pesas durante sus sesiones de ejercicio.




a) La distribución presenta una cola larga hacia la derecha. ¿Qué indica esta característica sobre los hábitos de entrenamiento de algunos usuarios del gimnasio?

b) ¿En qué rango de tiempos se observa la mayor concentración de usuarios? Explica tu respuesta con base en la gráfica.

c) Considerando la forma de la distribución, ¿esperarías que la media fuera mayor o menor que la mediana? Explica por qué.

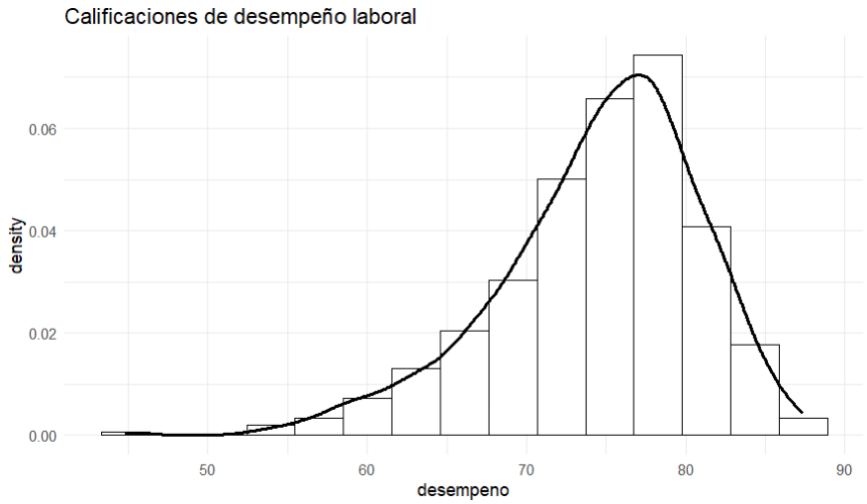
d) ¿Qué implicaciones podría tener esta distribución para la administración del gimnasio en aspectos como la disponibilidad de equipos, horarios o diseño de programas de entrenamiento?

e) ¿La gráfica sugiere la presencia de observaciones relativamente alejadas hacia valores altos? ¿Puede concluirse únicamente con esta gráfica que se trata de valores atípicos? Explica cómo llegas a esa conclusión.

2.3.3 Una empresa evalúa anualmente el desempeño de sus empleados mediante una calificación que va de 0 a 100 puntos. Con el propósito de analizar los resultados obtenidos, el departamento de Recursos Humanos construyó la siguiente curva de frecuencias para la variable:

$X$: Calificación de desempeño laboral (0–100 puntos).

Con base en la gráfica, responde las siguientes preguntas:




a) A partir de la distribución de las calificaciones, ¿podemos concluir que la mayoría de los empleados obtuvo un desempeño alto? Explica tu respuesta utilizando la gráfica.

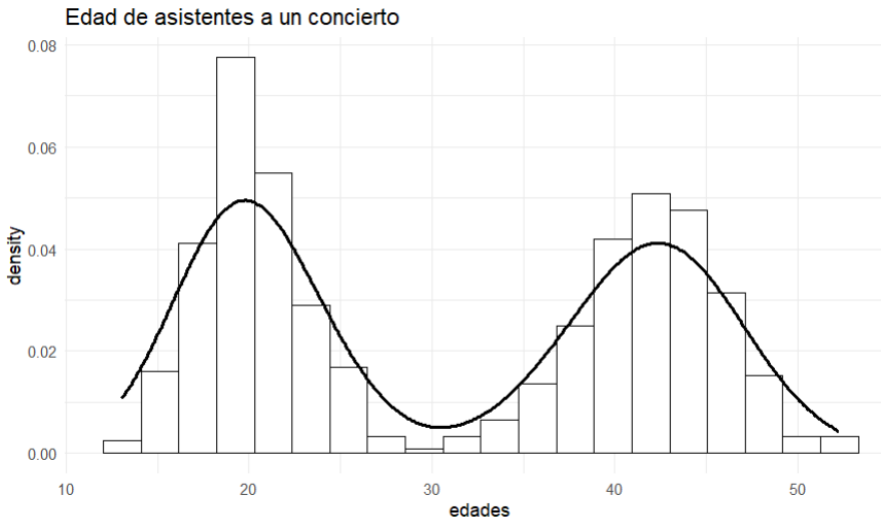
b) La distribución presenta una cola larga hacia la izquierda. ¿Qué sugiere esto sobre el desempeño de algunos empleados dentro de la empresa?

c) Con base en la forma de la distribución, ¿esperarías que la mediana fuera mayor o menor que la media? Explica qué significa esto en el contexto del desempeño laboral.

d) Con base en tu respuesta del inciso anterior, ¿cuál de las dos medidas consideras más adecuada para describir el desempeño típico de los empleados? Justifica tu respuesta.

e) Si la empresa modificara la forma de calcular el índice de desempeño laboral, ¿cómo podrían verse afectadas la distribución de las calificaciones y las conclusiones sobre el desempeño general de los empleados?

2.3.4 Se está analizando la edad de los asistentes a un concierto. La siguiente curva de frecuencias muestra la distribución de edades observada.




Con base en la gráfica, responde las siguientes preguntas:

a) ¿Cuáles son las edades aproximadas en las que se observan los dos picos principales de la distribución? ¿Qué representan estas edades dentro del contexto del concierto?

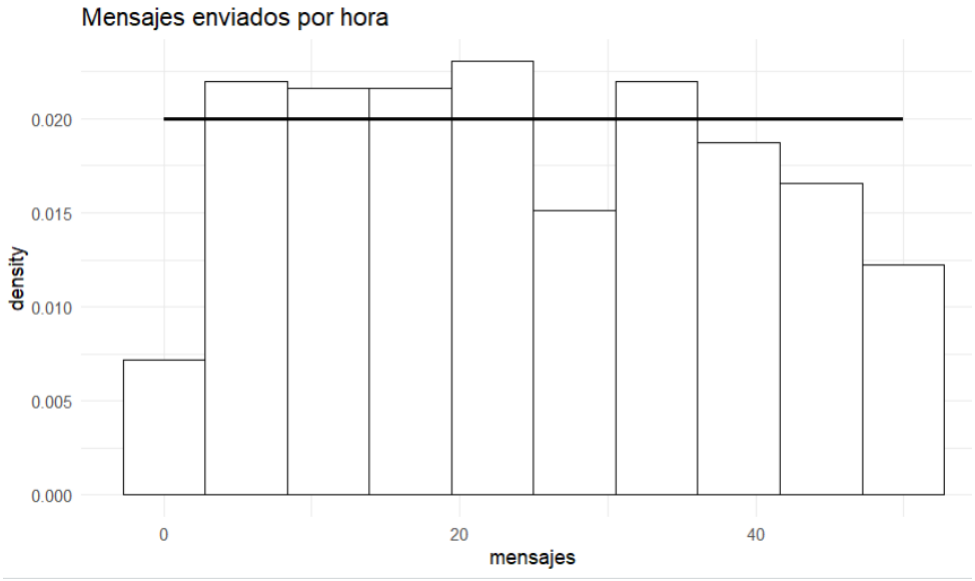
b) Al comparar la altura de los dos picos, ¿qué grupo parece tener una mayor cantidad de asistentes: los jóvenes o los adultos? Justifica tu respuesta utilizando la gráfica.

c) ¿Cuál de los dos grupos de asistentes presenta una mayor variabilidad en sus edades? Explica cómo lo identificas en la curva.

d) ¿Qué tan diferenciados parecen estar los dos grupos de asistentes en términos de edad? Fundamenta tu respuesta con base en la separación entre los picos de la distribución.

e) ¿La forma de la distribución sugiere que el concierto atrae a dos segmentos de público distintos? Explica tu respuesta utilizando la evidencia proporcionada por la gráfica.

2.3.5 Una empresa de telefonía analiza el número de mensajes enviados por hora durante el horario laboral. La siguiente curva de frecuencias muestra la distribución observada.




Con base en la gráfica, responde las siguientes preguntas:

a) ¿La distribución muestra algún valor o intervalo de mensajes enviados que concentre claramente más observaciones que los demás? Explica tu respuesta.

b) Si la distribución es aproximadamente uniforme, ¿qué indica esto sobre el comportamiento de envío de mensajes a lo largo del periodo analizado?

c) Aunque algunas barras son ligeramente más altas o más bajas que otras, ¿consideras que estas diferencias reflejan un patrón real o simplemente variación aleatoria? Justifica tu respuesta.

d) ¿Cuál es el rango aproximado de mensajes enviados por hora observado en la gráfica? ¿Qué información proporciona sobre los niveles mínimo y máximo de actividad?

e) ¿La distribución sugiere la existencia de picos de actividad, sesgos o concentraciones en ciertos valores, o más bien indica que los mensajes se envían de manera relativamente uniforme? Explica tu respuesta.

2.3.6 Una empresa que promueve el trabajo híbrido analiza la duración de las videollamadas realizadas por sus empleados. A partir de los datos observados, las reuniones pueden clasificarse aproximadamente en tres tipos:

- Reuniones cortas: alrededor de 10 minutos
- Reuniones medianas: alrededor de 30 minutos
- Reuniones largas: alrededor de 60 minutos

La siguiente curva de frecuencias muestra la distribución de las duraciones observadas.


Con base en la gráfica, responde las siguientes preguntas:

a) ¿Cuántos picos principales (modas) se observan en la distribución y a qué rangos de duración corresponden? ¿Qué representan estos picos dentro del contexto del problema?

b) ¿Cuál de los tres tipos de reuniones (cortas, medianas o largas) parece ser el más frecuente? Explica cómo lo identificas a partir de la altura de los picos de la distribución.

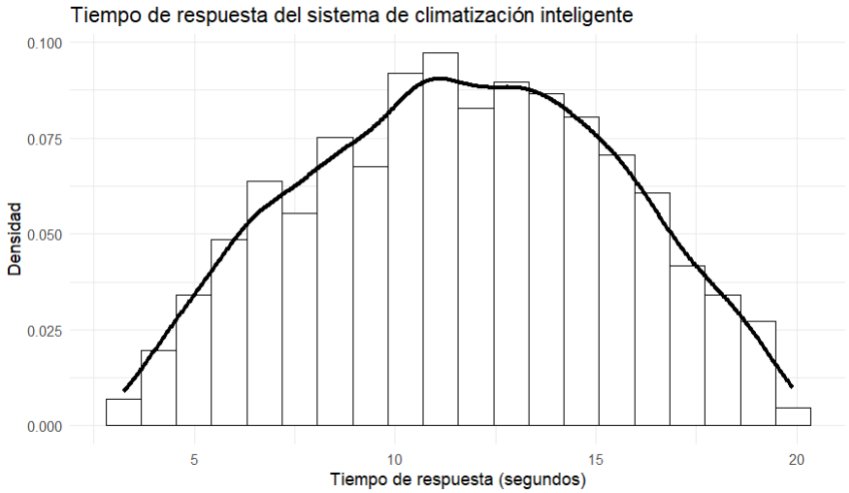
c) ¿Qué grupo de reuniones parece presentar mayor variabilidad en su duración? Explica cómo lo observas en la gráfica.

d) ¿La distribución muestra separaciones claras entre las reuniones cortas, medianas y largas, o existe traslape entre los grupos? ¿Qué podría explicar ese traslape?

e) ¿Qué implicaciones podría tener esta distribución para la gestión del tiempo y la productividad dentro de la empresa?

f) Si la empresa quisiera reducir interrupciones y mejorar la eficiencia operativa, ¿qué tipo de reuniones debería revisar o reorganizar primero? Justifica tu respuesta utilizando la información de la gráfica.

2.3.7 Muchas oficinas en CDMX están implementando un sistema de climatización inteligente que ajusta la temperatura según: número de personas en el espacio, hora del día, condiciones exteriores, $CO_2$ en interiores. La curva de frecuencias para el “tiempo de respuesta del sistema” (cuánto tarda en estabilizar la temperatura cuando detecta un cambio) es la siguiente:




a) ¿Qué indica la forma de la distribución sobre la variabilidad del tiempo de respuesta del sistema de climatización?

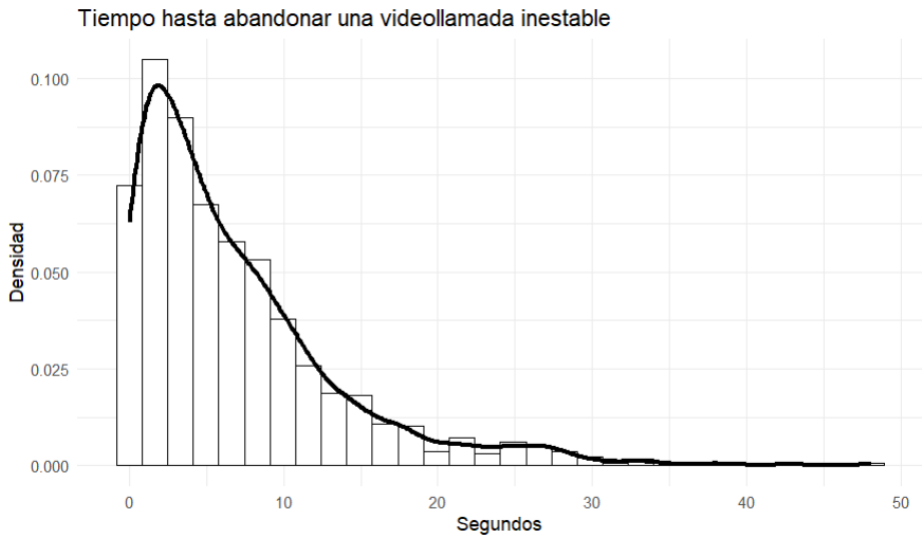
b) ¿Los tiempos de respuesta tienden a concentrarse alrededor de un valor típico o se encuentran ampliamente dispersos? Explica tu respuesta con base en la gráfica.

c) ¿Qué implicaciones podría tener para los administradores del edificio que los tiempos de respuesta presenten una alta variabilidad? Menciona una posible decisión que podría tomarse a partir de esta información.

d) ¿Consideras que la media es una medida adecuada para resumir el tiempo de respuesta del sistema? Justifica tu respuesta utilizando la forma de la distribución.

e) ¿Qué conclusión podrías obtener sobre la estabilidad y eficiencia del sistema a partir de la distribución observada de los tiempos de respuesta?

2.3.8 Se analiza el tiempo (en segundos) que tarda un usuario en abandonar una videollamada cuando experimenta problemas de conexión. La siguiente curva de frecuencias muestra la distribución observada.




Con base en la gráfica, responde las siguientes preguntas:

a) ¿La distribución indica que la mayoría de los usuarios abandona la videollamada rápidamente o que permanece conectada durante más tiempo? Justifica tu respuesta utilizando la forma de la curva.

b) La distribución presenta una cola hacia la derecha. ¿Qué indica esta característica sobre el comportamiento de algunos usuarios ante una conexión inestable?

c) ¿Los tiempos de abandono se concentran principalmente en unos pocos segundos o se distribuyen de manera uniforme a lo largo de todo el rango observado? Explica tu respuesta.

d) ¿Qué implicaciones podría tener esta distribución para el diseño y la mejora de la aplicación de videollamadas? Menciona una posible acción basada en los datos observados.

e) ¿La gráfica muestra evidencia de usuarios con una tolerancia inusualmente alta a los problemas de conexión? Explica cómo lo identificas a partir de la distribución.

## 2.4 Datos no agrupados

2.4.1 Los psicólogos están preocupados por la cantidad de tiempo que los adolescentes pasan en redes sociales. Con el fin de evaluar posibles cambios en sus hábitos digitales, se realizó un estudio con 20 adolescentes para medir el tiempo (en minutos por día) que dedican a TikTok después del lanzamiento de nuevas funciones de edición basadas en inteligencia artificial.

Los resultados obtenidos fueron los siguientes:

85, 92, 74, 60, 120, 110, 95, 88, 102, 76, 65, 70, 115, 130, 98, 90, 85, 72, 100, 108

Usando R:

a) Calcula las medidas de tendencia central: media, mediana y moda.

b) Calcula las medidas de dispersión: rango, rango intercuartílico (IQR), varianza, desviación estándar, desviación media absoluta (MAD) y coeficiente de variación.

c) Calcula el percentil 20, el cuartil 3 y el decil 2

Con base en los resultados obtenidos, responde las siguientes preguntas:

a) Los especialistas consideran que un uso superior a 60 minutos diarios puede estar asociado con mayores niveles de ansiedad. Con base en las medidas de tendencia central obtenidas, ¿qué se puede concluir sobre el tiempo de uso de TikTok en este grupo de adolescentes?

b) Si la mediana del tiempo de uso aumentó después de la incorporación de las nuevas funciones de inteligencia artificial, ¿qué indicaría este cambio sobre los hábitos de uso de los adolescentes?

c) Si no existe una moda claramente definida, ¿qué sugiere esto sobre la distribución de los tiempos de uso de TikTok entre los adolescentes estudiados?

d) ¿Qué información proporciona el rango acerca de la variabilidad en el tiempo de uso de TikTok dentro del grupo analizado?

e) ¿Qué indicaría un rango intercuartílico (IQR) pequeño acerca del comportamiento de la mayoría de los adolescentes?

f) Si la desviación estándar aumentara después de la actualización de la aplicación, ¿qué podría concluirse sobre los patrones de uso de TikTok entre los adolescentes?

g) ¿Cómo interpretarías un coeficiente de variación elevado en el contexto del tiempo diario dedicado a TikTok?¿Cómo interpretas un coeficiente de variación alto en el uso de la app?

h) Si la desviación media absoluta (MAD) es pequeña, ¿qué indica esto sobre la similitud de los tiempos de uso entre los adolescentes del estudio?

i) Interpreta el significado del percentil 20, el tercer cuartil (Q3) y el segundo decil (D2) en el contexto del tiempo diario dedicado a TikTok.

2.4.2 Ante el crecimiento del uso de herramientas de inteligencia artificial generativa (ChatGPT, Copilot, Gemini, etc.), un grupo de investigadores desea analizar cuánto tiempo dedican los estudiantes universitarios a estas plataformas para realizar tareas académicas.

Para ello, se realizó un estudio con 20 alumnos de licenciatura del ITAM durante una semana de exámenes parciales. Cada estudiante registró el tiempo diario promedio (en minutos) que utilizó asistentes de IA para estudiar, resolver ejercicios, redactar tareas o preparar exámenes.

 Los tiempos observados fueron:

45, 60, 30, 75, 120, 95, 80, 55, 90, 40, 65, 70, 110, 130, 85, 50, 60, 35, 100, 105

Usando R:

- Calcula las medidas de tendencia central: media, mediana y moda.
- Calcula las medidas de dispersión: rango, rango intercuartílico (IQR), varianza, desviación estándar, desviación media absoluta (MAD) y coeficiente de variación.
- Calcula el percentil 25, el cuartil 3 y el decil 8.

Con base en los resultados obtenidos, responde las siguientes preguntas:

a) La universidad considera que un uso superior a 90 minutos diarios de asistentes de IA podría indicar una fuerte dependencia de estas herramientas para realizar actividades académicas. ¿Qué porcentaje de estudiantes supera este umbral y qué conclusión puede obtenerse sobre el grupo estudiado?

b) Si la media del tiempo de uso es mayor que la mediana, ¿qué indica esto sobre la forma de la distribución de los datos y sobre el comportamiento de algunos estudiantes?

c) ¿Qué sugiere, en este contexto, que no exista una moda claramente definida o que existan varias modas en los tiempos de uso de asistentes de IA?

d) ¿Qué información proporciona el rango acerca de las diferencias en los hábitos de uso de herramientas de IA entre los estudiantes?

e) Si el rango intercuartílico (IQR) fuera pequeño, ¿qué indicaría sobre la variabilidad de los tiempos de uso de la mayoría de los estudiantes?

f) Si en un estudio realizado el siguiente semestre la desviación estándar aumentara, ¿qué podría concluirse acerca de los patrones de uso de asistentes de IA entre los alumnos?

g) ¿Cómo interpretarías un coeficiente de variación elevado en el contexto del tiempo diario dedicado al uso de herramientas de IA para fines académicos?

h) Si la desviación media absoluta (MAD) fuera pequeña, ¿qué indicaría sobre la similitud de los hábitos de uso entre los estudiantes del estudio?

i) ¿Cómo podrían utilizarse el percentil 25, el tercer cuartil (Q3) y el decil 8 para identificar distintos niveles de uso de asistentes de IA y apoyar la toma de decisiones académicas?

2.4.3 Como parte de una estrategia para reducir emisiones contaminantes y mejorar la movilidad urbana, el gobierno de la ciudad amplió la red de ciclovías y estableció nuevas zonas de baja emisión. Un grupo de investigación urbana desea analizar si estas políticas modificaron el uso de scooters eléctricos entre estudiantes universitarios.

Para ello, se seleccionó a 20 estudiantes y se registró el número de viajes semanales realizados en scooter eléctrico antes y después de la ampliación de ciclovías e implementación de la política.

 Se obtuvieron los siguientes datos:

| Estudiante | Antes | Después |
|---|---|---|
| 1 | 2 | 4 |
| 2 | 3 | 6 |
| 3 | 1 | 3 |
| 4 | 4 | 8 |
| 5 | 6 | 10 |
| 6 | 5 | 7 |
| 7 | 4 | 6 |
| 8 | 3 | 7 |
| 9 | 5 | 9 |
| 10 | 2 | 4 |
| 11 | 4 | 7 |
| 12 | 5 | 8 |
| 13 | 7 | 12 |
| 14 | 8 | 13 |
| 15 | 6 | 9 |
| 16 | 3 | 5 |
| 17 | 4 | 6 |
| 18 | 2 | 5 |
| 19 | 7 | 11 |
| 20 | 6 | 8 |

Utilizando R, realiza lo siguiente para los periodos “Antes” y “Después”:

- Calcula las medidas de tendencia central: media, mediana y moda.
- Calcula las medidas de dispersión: rango, rango intercuartílico (IQR), varianza, desviación estándar, desviación media absoluta (MAD) y coeficiente de variación.
- Calcula los percentiles 25, 50 y 75.
- Construye los diagramas de caja comparativos para ambos periodos.

Con base en los resultados obtenidos, responde las siguientes preguntas:

a) ¿Las medidas de tendencia central (media, mediana y moda) sugieren que el número semanal de viajes en scooter aumentó después de la ampliación de ciclovías? Explica tu respuesta.

b) Al comparar las medidas de dispersión (rango, IQR, desviación estándar, MAD y coeficiente de variación), ¿la variabilidad en el número de viajes aumentó, disminuyó o permaneció similar después de la implementación de la política?

c) ¿En cuál de los dos periodos (“Antes” o “Después”) se observa una mayor diversidad en los hábitos de uso de scooters eléctricos entre los estudiantes? Justifica tu respuesta utilizando las medidas de dispersión.

d) Con base en los diagramas de caja y en los percentiles calculados, ¿la distribución del número de viajes parece haber cambiado después de la política? Describe los principales cambios observados.

e) Al comparar los valores individuales antes y después de la intervención, ¿qué grupo parece haber incrementado más su uso de scooters: los estudiantes con pocos viajes iniciales o aquellos que ya utilizaban frecuentemente este medio de transporte? Explica tu respuesta.

f) ¿Qué conclusiones sobre la efectividad de la ampliación de ciclovías pueden extraerse de los resultados obtenidos? ¿Qué acciones adicionales podría considerar la autoridad para fomentar la movilidad sustentable?

g) ¿Qué factores metodológicos o limitaciones del estudio podrían afectar la validez de las conclusiones obtenidas?

h) Además de la ampliación de ciclovías, ¿qué otras variables podrían haber influido en el cambio observado en el número de viajes semanales en scooter eléctrico?

2.4.4 **Datos poblacionales no agrupados.** La Secretaría del Medio Ambiente de la Ciudad de México monitorea diariamente la concentración de partículas suspendidas menores a 2.5 micrómetros (PM2.5), expresadas en microgramos por metro cúbico $(\mu g/m^3)$. Estas partículas son consideradas altamente dañinas para la salud debido a que pueden penetrar profundamente en el sistema respiratorio.

A continuación, se presentan las concentraciones promedio diarias de PM2.5 registradas durante el mes de enero en 3 alcaldías de la Ciudad de México (BJU, GAM, NEZ). Debido a que se cuenta con la información de todos los días del mes, los datos deben considerarse poblacionales.

| DIA | BJU | GAM | NEZ |
|---|---|---|---|
| 1 | 178 | 152 | 289 |
| 2 | 50 | 48 | 42 |

| | | | |
|---|---|---|---|
| 3 | 53 | 55 | 70 |
| 4 | 40 | 37 | 37 |
| 5 | 49 | 48 | 64 |
| 6 | 50 | 51 | 57 |
| 7 | 65 | 59 | 60 |
| 8 | 60 | 70 | 60 |
| 9 | 32 | 48 | 66 |
| 10 | 70 | 64 | 71 |
| 11 | 29 | 21 | 21 |
| 12 | 59 | 52 | 71 |
| 13 | 57 | 32 | 36 |
| 14 | 51 | 45 | 40 |
| 15 | 46 | 45 | 48 |
| 16 | 54 | 39 | 90 |
| 17 | 38 | 45 | 40 |
| 18 | 34 | 28 | 32 |
| 19 | 63 | 46 | 87 |
| 20 | 53 | 56 | 57 |
| 21 | 67 | 57 | 48 |
| 22 | 51 | 49 | 43 |
| 23 | 37 | 44 | 53 |
| 24 | 32 | 36 | 60 |
| 25 | 18 | 16 | 27 |
| 26 | 62 | 76 | 52 |
| 27 | 44 | 40 | 44 |
| 28 | 46 | 40 | 32 |
| 29 | 46 | 43 | 60 |
| 30 | 57 | 54 | 72 |
| 31 | 65 | 54 | 71 |

Para la alcaldía BJU:

a) Construye una tabla de frecuencias para las concentraciones de PM2.5 que incluya: frecuencia absoluta, frecuencia relativa, frecuencia acumulada, y frecuencia relativa acumulada.

b) Elabora el histograma y el diagrama de caja y brazos.

c) Calcula e interpreta las siguientes medidas de tendencia central: media, mediana y moda.

d) Calcula e interpreta las siguientes medidas de dispersión: rango, rango intercuartílico, varianza poblacional, desviación estándar poblacional, coeficiente de variación.

e) Calcula los siguientes cuantiles: $Q_1$, $Q_2$, $Q_3$, percentil 10, percentil 90.

Responde las siguientes preguntas:

g) ¿Qué características presenta la distribución de las concentraciones diarias de PM2.5 en la alcaldía BJU? Describe su forma (simetría, sesgo, concentración de datos y posibles valores extremos).

h) ¿El diagrama de caja sugiere la presencia de valores atípicos? En caso afirmativo, ¿qué podrían representar estos valores dentro del contexto de la contaminación atmosférica?

i) ¿Las concentraciones de PM2.5 se mantuvieron relativamente estables durante el mes o mostraron una variabilidad importante? Justifica tu respuesta utilizando medidas de dispersión como el rango, el IQR, la desviación estándar o el coeficiente de variación.

j) La Organización Mundial de la Salud (OMS) recomienda que la concentración promedio diaria de PM2.5 no exceda los 15 µg/m³. Compara esta recomendación con los resultados obtenidos para la alcaldía BJU y evalúa la calidad del aire observada durante el mes.

k) Si fueras asesor de salud pública, ¿qué recomendaciones harías a la población y a las autoridades con base en los niveles de contaminación observados? Fundamenta tu respuesta utilizando la información estadística obtenida.

l) ¿Consideras que la media describe adecuadamente el nivel típico de contaminación durante el mes o sería preferible utilizar la mediana? Justifica tu respuesta considerando la presencia o ausencia de valores extremos.

m) ¿Qué implicaciones tendría para la salud pública que varios días del mes presentaran concentraciones cercanas al percentil 90?

## 2.5 Datos agrupados

2.5.1 Un club deportivo que organiza carreras recreativas quiere analizar la frecuencia cardiaca promedio en reposo (latidos por minuto) de corredores amateurs para diseñar planes de entrenamiento personalizados.

Se midió la frecuencia cardiaca de 240 corredores antes de iniciar un programa de acondicionamiento físico.

 Los datos obtenidos son los siguientes:

| Frecuencia cardiaca | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 71 | 54 | 64 | 57 | 51 | 66 | 62 | 61 | 55 | 67 | 49 | 58 |
| 60 | 51 | 47 | 55 | 58 | 71 | 59 | 72 | 60 | 70 | 50 | 63 |
| 56 | 65 | 55 | 62 | 60 | 68 | 63 | 53 | 55 | 50 | 72 | 71 |
| 53 | 63 | 66 | 56 | 52 | 54 | 53 | 60 | 69 | 57 | 55 | 55 |
| 73 | 70 | 72 | 53 | 65 | 51 | 64 | 59 | 57 | 60 | 57 | 61 |
| 61 | 46 | 70 | 64 | 67 | 58 | 51 | 65 | 58 | 55 | 60 | 53 |
| 55 | 55 | 58 | 57 | 73 | 56 | 65 | 70 | 63 | 58 | 65 | 47 |
| 50 | 51 | 45 | 63 | 55 | 60 | 70 | 61 | 50 | 49 | 62 | 70 |
| 54 | 65 | 55 | 69 | 64 | 50 | 47 | 68 | 65 | 59 | 51 | 51 |
| 59 | 61 | 67 | 61 | 48 | 70 | 50 | 65 | 60 | 52 | 60 | 57 |
| 67 | 47 | 74 | 72 | 73 | 64 | 55 | 60 | 46 | 71 | 57 | 68 |
| 49 | 56 | 68 | 48 | 60 | 58 | 60 | 50 | 57 | 60 | 63 | 50 |
| 64 | 54 | 64 | 64 | 54 | 62 | 55 | 53 | 62 | 65 | 52 | 62 |
| 61 | 59 | 59 | 59 | 57 | 68 | 66 | 56 | 73 | 52 | 69 | 59 |
| 55 | 53 | 58 | 45 | 59 | 59 | 60 | 65 | 70 | 56 | 50 | 60 |
| 50 | 73 | 52 | 55 | 67 | 45 | 58 | 70 | 58 | 46 | 65 | 56 |
| 45 | 65 | 62 | 70 | 60 | 66 | 63 | 66 | 54 | 53 | 56 | 65 |
| 60 | 63 | 74 | 65 | 55 | 72 | 47 | 62 | 55 | 61 | 55 | 67 |
| 71 | 57 | 48 | 60 | 53 | 56 | 65 | 54 | 65 | 66 | 62 | 52 |
| 55 | 68 | 59 | 56 | 61 | 60 | 56 | 52 | 63 | 70 | 66 | 55 |

Utilizando R, obtén una tabla de distribución de frecuencias utilizando los siguientes intervalos:

| Frecuencia cardiaca | Frecuencia absoluta | Frecuencia relativa | Frecuencia acumulada absoluta | Frecuencia acumulada relativa |
|---|---|---|---|---|
| [45 – 50) | | | | |
| [50 – 55) | | | | |
| [55 – 60) | | | | |
| [60 – 65) | | | | |
| [65 – 70) | | | | |
| [70 – 75) | | | | |

A partir de la tabla de frecuencias:

- Calcula las medidas de tendencia central: media, mediana y moda.
- Calcula las medidas de dispersión: rango, rango intercuartílico (IQR), varianza, desviación estándar, desviación media absoluta (MAD) y coeficiente de variación.
- Construye un histograma, un polígono de frecuencias y una ojiva.
- Interpreta la forma general de la distribución.

Con base en los resultados obtenidos, responde las siguientes preguntas:

a) ¿En qué intervalo de frecuencia cardiaca se concentra la mayor cantidad de corredores?

b) ¿Cómo describirías el perfil cardiaco típico de los participantes del estudio utilizando las medidas de tendencia central?¿La distribución es simétrica o sesgada?

c) A partir del histograma y de las medidas calculadas, ¿la distribución parece simétrica o presenta sesgo? Explica tu respuesta.

d) ¿Qué información aporta la posición de la mediana sobre la distribución de las frecuencias cardiacas?

e) ¿Las medidas de dispersión sugieren que existe una alta heterogeneidad entre los corredores o que el grupo es relativamente homogéneo? Justifica tu respuesta.

f) ¿Cómo podrían utilizarse estos resultados para diseñar planes de entrenamiento diferenciados según el nivel de condición física de los corredores?

g) ¿Qué información individual se pierde al trabajar con datos agrupados en intervalos en lugar de utilizar los datos originales?

h) Al calcular medidas estadísticas a partir de datos agrupados, ¿qué supuestos se están realizando sobre los valores dentro de cada intervalo?

2.5.2 Una plataforma de streaming modificó su algoritmo de recomendaciones con el objetivo de priorizar contenido más personalizado para cada usuario. La empresa desea analizar si este cambio influyó en el número de horas semanales que los suscriptores dedican a la plataforma.

Para ello, se seleccionó una muestra de 200 suscriptores y se registró el tiempo de consumo semanal antes y después de la actualización del algoritmo.

 Los resultados obtenidos se agruparon en los siguientes intervalos:

Antes del cambio

| Horas/semana | Frecuencia |
|---|---|
| 0 – 5 | 18 |
| 5 – 10 | 42 |
| 10 – 15 | 60 |
| 15 – 20 | 48 |
| 20 – 25 | 22 |
| 25 – 30 | 10 |
| **Total** | **200** |

Después del cambio:

| Horas/semana | Frecuencia |
|---|---|
| 0 – 5 | 10 |
| 5 – 10 | 28 |
| 10 – 15 | 50 |
| 15 – 20 | 60 |
| 20 – 25 | 32 |
| 25 – 30 | 20 |
| **Total** | **200** |

Utilizando R, realiza lo siguiente para ambos periodos:

a) Calcula las medidas de tendencia central aproximadas: media, mediana y moda.

b) Calcula las medidas de dispersión: rango, rango intercuartílico (IQR), varianza, desviación estándar, desviación media absoluta (MAD) y coeficiente de variación.

c) Construye histogramas y polígonos de frecuencia comparativos.

Con base en los resultados obtenidos, responde las siguientes preguntas:

a) ¿En qué intervalo de frecuencia cardiaca se concentra el mayor número de corredores? ¿Qué indica esto sobre el nivel de condición física predominante en el grupo estudiado?

b) Utilizando la media, la mediana y la moda, ¿cómo describirías la frecuencia cardiaca típica de los corredores del estudio?

c) A partir del histograma, el polígono de frecuencias y la comparación entre la media y la mediana, ¿la distribución parece simétrica o presenta algún grado de sesgo? Explica tu respuesta.

d) ¿Qué información proporciona la mediana acerca de la distribución de las frecuencias cardiacas? Interpreta su significado en el contexto del estudio.

e) Considerando el rango, el IQR, la desviación estándar y el coeficiente de variación, ¿los corredores presentan frecuencias cardiacas relativamente homogéneas o existe una variabilidad importante entre ellos? Justifica tu respuesta.

f) ¿Cómo podrían utilizarse las medidas de tendencia central y dispersión para diseñar programas de entrenamiento adaptados a distintos perfiles de condición física?

g) ¿Qué información se pierde al agrupar las frecuencias cardiacas en intervalos en lugar de analizar los datos individuales? Proporciona un ejemplo.

h) Cuando se calculan medidas estadísticas a partir de datos agrupados, ¿qué suposiciones se hacen sobre los valores que pertenecen a cada intervalo? ¿Cómo podrían afectar estas suposiciones la precisión de los resultados?

i) ¿Qué porcentaje aproximado de corredores presenta una frecuencia cardiaca inferior a 60 latidos por minuto? ¿Qué podría indicar esto acerca de su condición física?

j) Si un entrenador quisiera identificar al 25% de los corredores con las frecuencias cardiacas más bajas en reposo, ¿qué cuantil utilizaría y cómo lo interpretaría?

2.5.3 El Servicio Sismológico realizó un análisis de la magnitud de los 50 sismos más recientes percibidos en la Ciudad de México, medidos de acuerdo con la escala de Richter.

 La siguiente tabla presenta la distribución de frecuencias de las magnitudes registradas:

| Magnitud (Escala Richter) | Frecuencia |
|---|---|
| (1.5 – 2.0] | 4 |
| (2.0 – 2.5] | 9 |
| (2.5 – 3.0] | 11 |
| (3.0 – 3.5] | 10 |
| (3.5 – 4.0] | 8 |
| (4.0 – 4.5] | 5 |
| (4.5 – 5.0] | 2 |
| (5.0 – 5.5] | 1 |
| **Total** | **50** |

a) Para cada intervalo de clase calcula: frecuencia relativa, frecuencia absoluta acumulada y su frecuencia acumulada relativa.

b) Construye el polígono de frecuencias.

c) Calcula las medidas descriptivas: media, mediana, clase modal, desviación estándar y el coeficiente de variación.

d) Interpreta la media, la mediana, la clase modal, la desviación estándar y el coeficiente de variación en el contexto de las magnitudes de los sismos registrados. ¿Qué información aporta cada medida sobre la actividad sísmica observada?

e) A partir del polígono de frecuencias y de las medidas de tendencia central calculadas, describe la forma general de la distribución. ¿La distribución parece simétrica, sesgada hacia magnitudes bajas o altas, o presenta algún otro patrón? Justifica tu respuesta.

f) Si se contara con una distribución similar para los sismos registrados en Los Ángeles, ¿qué herramientas estadísticas utilizarías para comparar ambas distribuciones? Explica qué aspectos compararías y qué información aportaría cada técnica.

g) ¿Qué porcentaje de los sismos registrados tuvo una magnitud menor a 3.5? ¿Qué indica este resultado sobre la intensidad de la actividad sísmica observada?

2.5.4 Un comprador desea elegir entre dos marcas de focos, A y B. Como ambas tienen el mismo precio, su decisión dependerá de cuál ofrezca una mayor duración y una mayor consistencia en su desempeño.

| Foco A | Foco B |
|---|---|
| 1104 | 1254 |
| 1160 | 1231 |
| 1202 | 1199 |
| 1265 | 1173 |
| 1203 | 1151 |

Con base en los datos observados, calcula las medidas estadísticas que consideres pertinentes y responde:

a) ¿Qué marca recomendarías comprar? Justifica tu respuesta utilizando medidas de tendencia central y dispersión.

b) ¿Cuál de las dos marcas muestra una duración más consistente entre sus focos?

c) Si el precio es el mismo, ¿qué marca recomendarías comprar y por qué?

2.5.5 **Datos poblacionales agrupados.** La Secretaría del Medio Ambiente de la Ciudad de México monitorea de manera continua la concentración de ozono $(O_3)$en el aire debido a sus efectos sobre la salud y el medio ambiente. El ozono troposférico puede provocar irritación respiratoria, disminución de la función pulmonar y aumento de enfermedades cardiovasculares, especialmente durante temporadas de alta radiación solar.

La siguiente tabla presenta una distribución agrupada de las concentraciones máximas diarias de ozono registradas durante un año en la Ciudad de México (datos poblacionales). Las concentraciones se expresan en partes por billón (ppb).

**Concentración de ozono (ppb)**

| Límite inferior | Límite superior | Frecuencia (días) |
|---|---|---|
| 20 | 30 | 18 |
| 30 | 40 | 42 |
| 40 | 50 | 67 |
| 50 | 60 | 84 |
| 60 | 70 | 73 |
| 70 | 80 | 42 |
| 80 | 90 | 21 |
| 90 | 100 | 12 |
| 100 | 110 | 5 |
| 110 | 120 | 1 |
| **Total** | | **365** |

Con base en la información anterior, realiza lo siguiente:

a) Para cada intervalo calcula frecuencia relativa, frecuencia absoluta acumulada y frecuencia relativa acumulada.

b) Construye el histograma y la ojiva.

c) Identifica el intervalo modal e interpreta su significado en el contexto de la calidad del aire.

d) Calcula e interpreta las siguientes medidas: media, mediana, moda, varianza, desviación estándar, coeficiente de variación.

e) Calcula los cuartiles $Q_1$, $Q_2$y $Q_3$ y los percentiles 10, 80 y 95.

f) Interpreta el significado del percentil 95 en términos de contaminación ambiental.

Responde las siguientes preguntas:

g) Con base en el histograma, la ojiva y las medidas de tendencia central calculadas, ¿cómo describirías la forma de la distribución de las concentraciones de ozono? ¿Parece aproximadamente simétrica o presenta algún grado de sesgo? Justifica tu respuesta.

h) ¿Las concentraciones de ozono se mantuvieron relativamente estables a lo largo del año o mostraron una variabilidad importante? Utiliza medidas como la desviación estándar y el coeficiente de variación para sustentar tu respuesta.

i) ¿Qué porcentaje de los días del año registró concentraciones de ozono superiores a 80 ppb? Interpreta este resultado en términos de calidad del aire.

j) Si las autoridades consideran preocupantes las concentraciones superiores a 90 ppb, ¿qué proporción de los días del año se ubicó en esta categoría? ¿Qué implicaciones tiene este resultado para la salud pública?

k) Considerando las medidas descriptivas, los cuantiles y las gráficas construidas, ¿qué conclusiones generales pueden obtenerse sobre la calidad del aire en la Ciudad de México durante el año analizado?

l) Si fueras asesor de la Comisión Ambiental de la Megalópolis (CAMe), ¿qué acciones recomendarías para reducir las concentraciones de ozono y disminuir la exposición de la población a este contaminante? Fundamenta tu respuesta utilizando los resultados obtenidos.

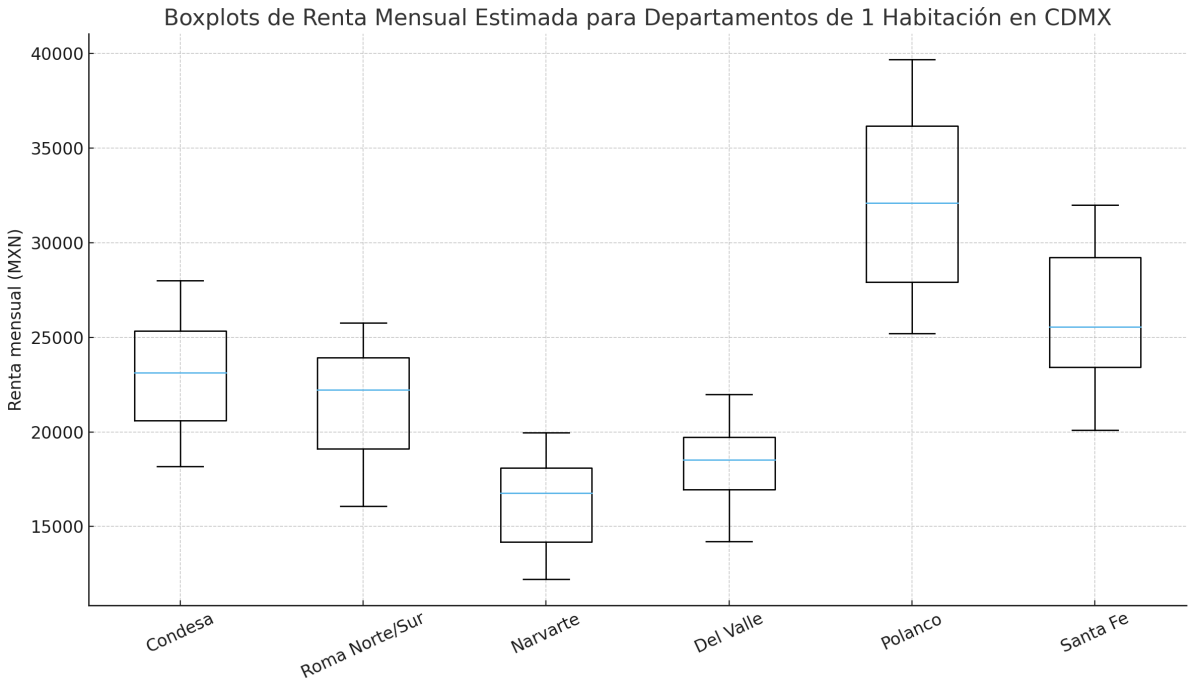
m) ¿Qué información proporciona el percentil 95 sobre los episodios de contaminación más severos del año? ¿Por qué este indicador puede ser útil para la planeación ambiental y la emisión de alertas de calidad del aire?

n) Si el objetivo de las autoridades fuera reducir los días con mayores niveles de contaminación, ¿sería más útil monitorear la media o el percentil 95? Justifica tu respuesta.

## 2.6 Diagramas de caja y brazos

2.6.1 Una joven profesionista de 24 años, recién egresada de licenciatura, consiguió su primer empleo en una empresa ubicada cerca del corredor Reforma–Juárez. Como desea independizarse sin comprometer su estabilidad financiera, comenzó a investigar el mercado de renta de departamentos de una habitación en distintas zonas de la Ciudad de México.

Después de recopilar información sobre precios de renta mensuales en varias colonias, decidió construir un boxplot comparativo para analizar la distribución de los precios y comparar el comportamiento del mercado inmobiliario entre zonas.




Con base en el boxplot obtenido, responde las siguientes preguntas:

a) ¿Qué zona parece tener las rentas más altas? Justifica tu respuesta utilizando la posición de la mediana.

b) ¿Qué zona presenta la mayor variabilidad en los precios de renta? Explica cómo lo identificas en el boxplot.

c) ¿Qué zona presenta la menor variabilidad en las rentas? ¿Qué podría indicar esto sobre la homogeneidad de los departamentos disponibles?

d) ¿Cuál es la diferencia entre la zona con la mediana más alta y la zona con la mediana más baja? ¿Qué indica esta diferencia sobre el mercado inmobiliario de la ciudad?

e) ¿En qué zonas el rango intercuartílico (IQR) es más amplio? ¿Qué significa esto respecto a la dispersión de las rentas?

f) ¿Se observan valores atípicos (outliers) en alguna zona? ¿Qué podrían representar estos departamentos dentro del mercado de renta?

g) Compara las colonias Condesa, Roma y Narvarte. ¿Qué diferencias observas en sus medianas y en la dispersión de las rentas?

h) ¿Existen zonas cuyos rangos de renta se traslapen? ¿Qué implica esto para una persona que está buscando departamento?

i) ¿Qué zona presenta la mayor diferencia entre la renta mínima y la renta máxima? ¿Qué podría indicar esto sobre la diversidad de inmuebles disponibles?

j) Si un departamento tiene una renta mensual de $30,000, ¿en cuáles zonas parecería una renta típica y en cuáles sería inusualmente alta o baja? Justifica tu respuesta utilizando el boxplot.

k) ¿Qué zonas muestran una distribución aproximadamente simétrica y cuáles presentan indicios de sesgo? Explica cómo lo identificas visualmente.

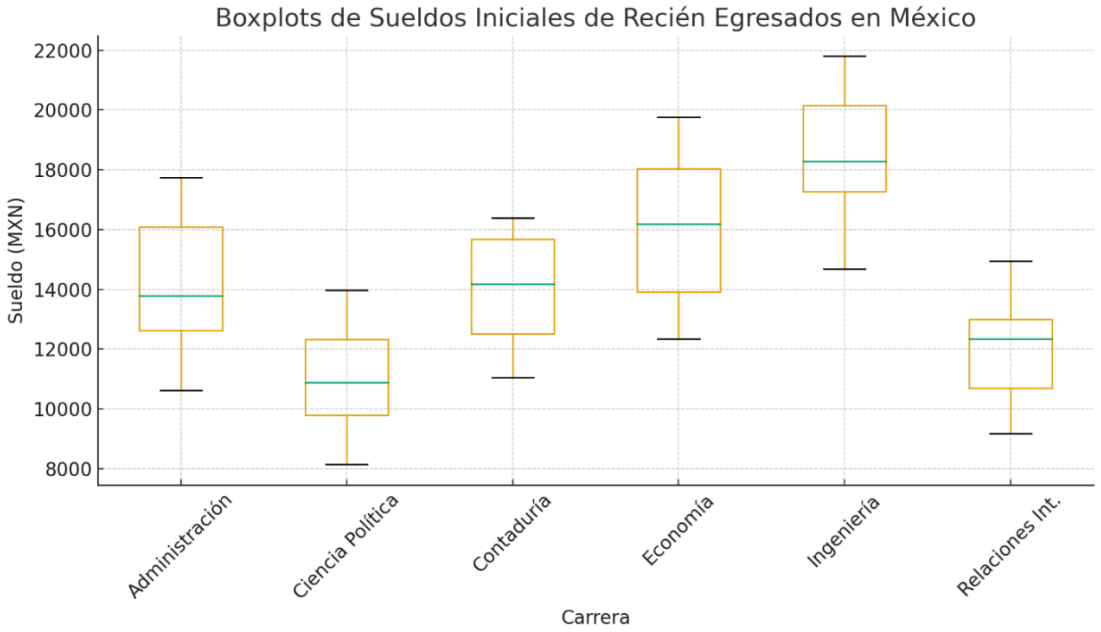
l) Si la joven profesionista busca una zona con rentas relativamente estables y predecibles, ¿cuál le recomendarías? Justifica tu respuesta utilizando el boxplot.

m) Si la joven está dispuesta a aceptar una mayor variabilidad en las rentas a cambio de encontrar oportunidades más económicas dentro de una zona, ¿qué colonia podría considerar? Explica tu respuesta.

2.6.2 Cada año, miles de jóvenes universitarios se incorporan al mercado laboral en México. Aunque muchos comienzan una etapa profesional similar, los sueldos iniciales pueden variar considerablemente dependiendo de factores como la carrera estudiada, el sector económico, la demanda laboral, las habilidades técnicas adquiridas y el tipo de empresa contratante.

Con el objetivo de orientar a futuros estudiantes y egresados, un centro universitario de orientación profesional realizó un estudio sobre los sueldos mensuales iniciales de recién egresados en seis áreas académicas distintas.

A partir de información obtenida de portales laborales, encuestas de egresados y estimaciones del mercado, se construyeron rangos aproximados de salarios iniciales y se elaboraron boxplots comparativos para analizar la distribución de los sueldos entre carreras.



Con base en los boxplots, responde las siguientes preguntas:

a) ¿Qué carrera parece ofrecer los salarios iniciales más altos? Justifica tu respuesta utilizando la posición de las medianas.

b) ¿Qué carrera presenta la mayor variabilidad salarial entre sus egresados? Explica cómo lo identificas en el boxplot.

c) ¿Qué carrera muestra la menor variabilidad salarial? ¿Qué podría indicar esto sobre la homogeneidad de las oportunidades laborales para sus egresados?

d) ¿En qué carreras la mediana parece representar adecuadamente el salario típico y en cuáles podría ser menos representativa debido a una alta dispersión o a la presencia de valores atípicos?

e) ¿Se observan valores atípicos (outliers) en alguna carrera? ¿Qué podrían representar estos casos dentro del mercado laboral?

f) ¿Qué carrera presenta el rango intercuartílico (IQR) más amplio? ¿Qué implica esto sobre la diversidad de salarios iniciales entre sus egresados?

g) Compara los salarios iniciales de Ciencia de Datos y Economía. ¿Qué diferencias observas en sus medianas y en la dispersión salarial?

h) ¿Existen carreras cuyos rangos salariales se traslapen? ¿Qué indica este traslape sobre las oportunidades económicas de los egresados?

i) Si un recién egresado recibe una oferta de $22,000 mensuales, ¿en cuáles carreras este salario parecería típico y en cuáles sería inusualmente alto o bajo? Justifica tu respuesta utilizando el boxplot.

j) ¿Qué carreras muestran evidencia de asimetría en la distribución salarial? Explica cómo puede identificarse esta característica en el boxplot.

k) Entre Administración y Contaduría, ¿cuál parece ofrecer un mayor potencial salarial máximo? ¿Qué evidencia proporciona el boxplot?

l) Si un estudiante busca maximizar su ingreso esperado al inicio de su carrera profesional, ¿qué carreras parecerían más atractivas según la información del boxplot?

m) Si un estudiante valora más la estabilidad salarial que la posibilidad de obtener ingresos excepcionalmente altos, ¿qué carrera podría resultarle más conveniente? Justifica tu respuesta.

n) ¿Qué conclusiones generales puede obtener un estudiante sobre la relación entre carrera universitaria y salario inicial a partir de estos boxplots?

## 2.7 El problema de comparación: análisis de subpoblaciones

2.7.1 Una universidad desea analizar si el uso de herramientas de inteligencia artificial (IA) está asociado con un mejor desempeño académico. Para ello, compara el promedio general de calificaciones entre estudiantes que utilizan IA y estudiantes que no la utilizan.

| Grupo | Promedio general |
|---|---|
| Usa IA | 8.4 |
| No usa IA | 8.1 |

Con base en estos resultados, parecería que el uso de IA mejora el desempeño académico.
Sin embargo, al desagregar la información por licenciatura, se obtienen los siguientes resultados:

| Licenciatura | Usa IA | No usa IA |
|---|---|---|
| Ingeniería | 8.0 | 8.3 |
| Sociales | 8.8 | 9.1 |

a) Al analizar los datos por licenciatura, ¿se mantiene la conclusión inicial sobre el efecto del uso de IA en el desempeño académico? Explica.

b) ¿Por qué es incorrecto comparar solo los promedios generales?

c) ¿Qué variable está actuando como factor de confusión en el análisis?

d) Si la universidad solo analizara los promedios generales, ¿qué decisión equivocada podría tomar?

2.7.2 Una universidad desea evaluar si las becas ayudan a reducir la deserción estudiantil. Para ello, compara la tasa de deserción entre alumnos con beca y alumnos sin beca.

| Grupo | % Deserción |
|---|---|
| Con beca | 12% |
| Sin beca | 15% |

Con base en estos resultados generales, parecería que las becas reducen la deserción escolar.
Sin embargo, al desagregar la información por nivel socioeconómico, se obtienen los siguientes resultados:

| Nivel | Con beca | Sin beca |
|---|---|---|
| Bajo | 20% | 18% |
| Medio | 10% | 8% |
| Alto | 5% | 4% |

a) Cuando se comparan los estudiantes dentro de cada licenciatura, ¿la evidencia sigue apoyando la conclusión de que el uso de IA mejora el desempeño académico? Explica tu respuesta.

b) ¿Por qué puede resultar engañoso basar el análisis únicamente en los promedios generales de ambos grupos?

c) ¿Qué variable está influyendo simultáneamente en el uso de IA y en el promedio académico, generando una posible interpretación incorrecta de los resultados?

d) Si la universidad utilizara únicamente los promedios generales para tomar decisiones, ¿qué conclusiones o políticas podría implementar de manera equivocada?

e) ¿Puede concluirse que el uso de IA causa un mejor desempeño académico? Explica qué información adicional sería necesaria para establecer una relación causal.

f) ¿Qué estrategia de análisis recomendarías para evaluar de manera más justa el efecto del uso de IA sobre el rendimiento académico?

## 2.8 El problema de la asociación

2.8.1 El director de recursos humanos de una empresa desea analizar si existe alguna relación entre el tiempo que tardan las secretarias en copiar un documento y el número de errores que cometen durante la tarea.

Para ello, evaluó a 20 secretarias registrando:

- el tiempo requerido para copiar un escrito (en segundos),
- y el número de errores cometidos.

 Los resultados obtenidos se muestran a continuación:

| Secretaria | Tiempo (seg) | Número de errores |
|---|---|---|
| 1 | 68 | 8 |
| 2 | 72 | 2 |
| 3 | 35 | 9 |
| 4 | 91 | 14 |
| 5 | 47 | 9 |

| | | |
|---|---|---|
| 6 | 52 | 13 |
| 7 | 75 | 12 |
| 8 | 63 | 3 |
| 9 | 55 | 0 |
| 10 | 65 | 14 |
| 11 | 84 | 0 |
| 12 | 45 | 14 |
| 13 | 58 | 14 |
| 14 | 61 | 12 |
| 15 | 69 | 2 |
| 16 | 22 | 2 |
| 17 | 46 | 5 |
| 18 | 55 | 5 |
| 19 | 66 | 13 |
| 20 | 71 | 2 |

Con base en la información anterior:

a) Construye un diagrama de dispersión que relacione el tiempo de copiado con el número de errores cometidos.

b) Con base en el diagrama de dispersión, ¿se observa alguna relación o tendencia entre el tiempo requerido para copiar el documento y el número de errores cometidos? Describe la dirección, la forma y la intensidad aparente de la relación.

c) Calcula el coeficiente de correlación de Pearson entre el tiempo de copiado y el número de errores. Interpreta su signo y magnitud en el contexto del problema.

d) ¿Se observan valores atípicos o casos inusuales en el diagrama de dispersión? ¿Cómo podrían influir en la correlación calculada?

e) ¿Puede concluirse que el tiempo de copiado causa cambios en el número de errores? Explica la diferencia entre correlación y causalidad en este contexto.

2.8.2 La siguiente tabla presenta información socioeconómica reciente para distintos países del continente americano. Las variables consideradas son:

- Esperanza de vida al nacer (años)
- PIB per cápita (USD)
- Tasa de alfabetización (%)
- Población urbana (%)

| País | Esperanza de vida | PIB per cápita (USD) | Alfabetización (%) | Población urbana (%) |
|---|---|---|---|---|
| Canadá | 82 | 53000 | 99 | 82 |
| Estados Unidos | 77 | 82000 | 99 | 83 |
| México | 75 | 14000 | 95 | 81 |
| Guatemala | 70 | 5600 | 82 | 53 |
| Costa Rica | 80 | 17000 | 98 | 82 |
| Cuba | 78 | 9500 | 99 | 78 |
| Brasil | 75 | 10000 | 94 | 88 |
| Chile | 81 | 17500 | 97 | 88 |
| Colombia | 77 | 7000 | 95 | 82 |
| Argentina | 77 | 13000 | 99 | 92 |
| Perú | 77 | 7500 | 94 | 80 |
| Haití | 64 | 1800 | 62 | 57 |
| República Dominicana | 74 | 10500 | 93 | 84 |
| Panamá | 79 | 19000 | 96 | 69 |
| Uruguay | 78 | 22000 | 99 | 95 |

a) Construye diagramas de dispersión para cada par de variables e identifica el tipo de relación existente:

- esperanza de vida y PIB per cápita
- tasa de alfabetización y esperanza de vida
- porcentaje de población urbana y PIB per cápita

b) ¿Qué pares de variables parecen presentar una asociación positiva? Describe la intensidad y dirección de la relación observada en cada caso.

c) ¿Se observan países que se aparten claramente de la tendencia general de los datos? Identifíquelos y explique por qué podrían considerarse observaciones atípicas.

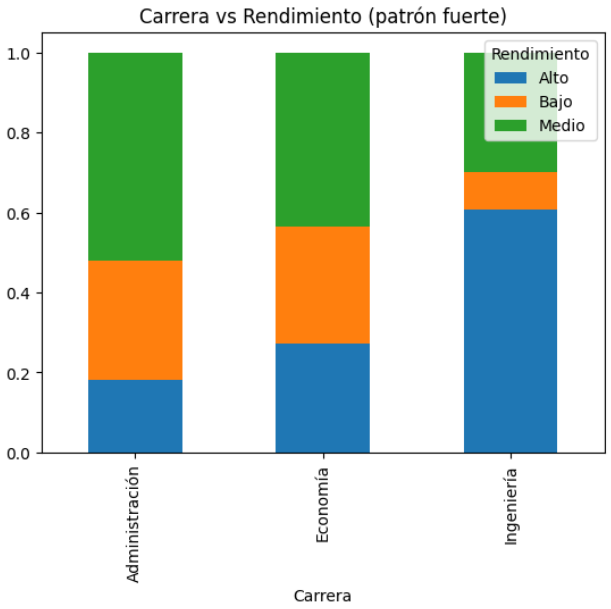
d) ¿Cuál de las variables analizadas parece estar más estrechamente relacionada con la esperanza de vida? Justifique su respuesta utilizando los diagramas de dispersión.

e) Compara la posición de Canadá y Estados Unidos respecto a los países de América Latina en términos de PIB per cápita, esperanza de vida, alfabetización y urbanización. ¿Qué diferencias destacan?

f) ¿La relación entre PIB per cápita y esperanza de vida parece lineal o muestra rendimientos decrecientes a niveles altos de ingreso? Explica tu respuesta.

g) ¿Es posible concluir que un mayor PIB per cápita causa una mayor esperanza de vida? Explica la diferencia entre asociación y causalidad en este contexto.

2.8.3 Una universidad realizó un estudio con estudiantes de tres carreras: Economía, Administración e Ingeniería. Para cada estudiante se registró la carrera a la que pertenece, su nivel de rendimiento académico (Alto, Medio o Bajo) y sus horas de estudio semanales.




Con base en la gráfica proporcionada, responde las siguientes preguntas:

a) ¿Qué carrera presenta la mayor proporción de estudiantes con rendimiento académico alto? ¿Qué podría indicar este resultado sobre el desempeño de sus estudiantes?

b) ¿En qué carrera se observa la mayor proporción de estudiantes con rendimiento académico bajo? ¿Qué implicaciones podría tener este resultado para la universidad?

c) ¿Alguna carrera muestra una distribución de niveles de rendimiento claramente diferente a las demás? Describe las principales diferencias observadas.

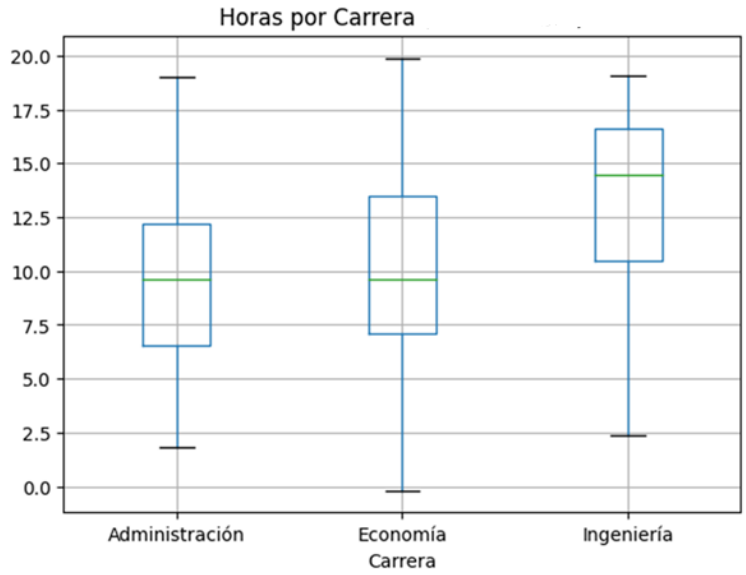
d) Al comparar las tres carreras, ¿qué nivel de rendimiento (alto, medio o bajo) parece ser el más frecuente? ¿Qué conclusión general puede obtenerse sobre el desempeño académico de los estudiantes?

e) Si la universidad tuviera recursos para implementar un programa de tutorías en una sola carrera, ¿cuál recomendarías priorizar? Justifica tu respuesta utilizando la información de la gráfica.

f) ¿Qué carrera presenta el mayor contraste entre estudiantes de rendimiento alto y bajo? ¿Qué podría indicar esto sobre la heterogeneidad académica de sus estudiantes?

## 2.9 Problema de comparación y problemas de asociación

2.9.1 Para el ejercicio anterior, se construyeron diagramas de caja y brazos para comparar la distribución de las horas de estudio semanales entre las distintas carreras universitarias.




Con base en los diagramas, responde lo siguiente:

Comparación entre carreras

a) ¿Qué carrera presenta la mediana más alta de horas de estudio semanales? ¿Qué indica este resultado sobre los hábitos de estudio de sus estudiantes?

b) ¿Cuál carrera muestra la mayor variabilidad en las horas de estudio? Explica cómo lo identificas utilizando el rango intercuartílico y la longitud de los brazos del diagrama.

c) ¿Se observan valores atípicos en alguna carrera? En caso afirmativo, ¿qué podrían representar estos estudiantes dentro de su grupo académico?

d) Considerando simultáneamente la mediana y la dispersión de las horas de estudio, ¿qué carrera parece presentar hábitos de estudio más consistentes? Justifica tu respuesta.

e) Si la universidad tuviera recursos para implementar un programa de fortalecimiento de hábitos de estudio en una sola carrera, ¿cuál recomendarías priorizar? Utiliza la información del boxplot para sustentar tu respuesta.

Asociación entre horas de estudio y rendimiento académico

a) ¿Qué nivel de rendimiento académico presenta la mediana más alta de horas de estudio semanales?

b) ¿Los diagramas sugieren una relación entre el número de horas de estudio y el nivel de rendimiento académico? Explica tu respuesta.

c) ¿Las distribuciones de horas de estudio para los distintos niveles de rendimiento están claramente separadas o presentan traslape? ¿Qué implica esto sobre la relación entre ambas variables?

d) ¿Las horas de estudio parecen ser un factor importante para explicar las diferencias en el rendimiento académico? Justifica tu respuesta utilizando la posición de las medianas y la dispersión observada.

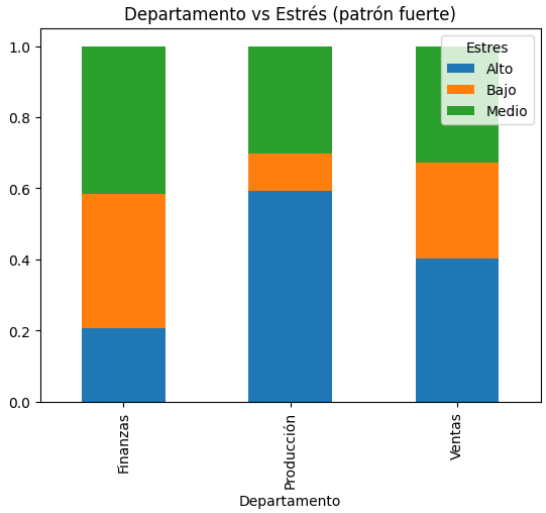
e) ¿Qué nivel de rendimiento presenta la mayor variabilidad en las horas de estudio? ¿Qué podría indicar esto acerca de las estrategias de estudio de los estudiantes de ese grupo?

f) ¿Existen estudiantes con pocas horas de estudio que logran un rendimiento alto, o estudiantes con muchas horas de estudio que obtienen un rendimiento bajo? ¿Qué sugiere esto sobre otros factores que podrían influir en el desempeño académico?

g) ¿Qué limitaciones tiene el uso de diagramas de caja para analizar la relación entre horas de estudio y rendimiento académico en comparación con un diagrama de dispersión?

2.9.2 Una empresa realizó un estudio sobre el bienestar laboral de sus empleados en tres departamentos: Ventas, Producción y Finanzas. Para cada empleado se registró:

- Departamento al que pertenece
- Nivel de estrés (Alto, Medio o Bajo)
- Número de días de ausencia al año



Con base en la información y las gráficas proporcionadas, responde las siguientes preguntas.

Asociación entre departamento y nivel de estrés

a) ¿Qué departamento presenta la mayor proporción de empleados con nivel de estrés alto? ¿Qué podría indicar este resultado sobre las condiciones de trabajo en dicho departamento?

b) ¿En qué departamento se observa la mayor proporción de empleados con nivel de estrés bajo? ¿Qué interpretación podría darse a este resultado?

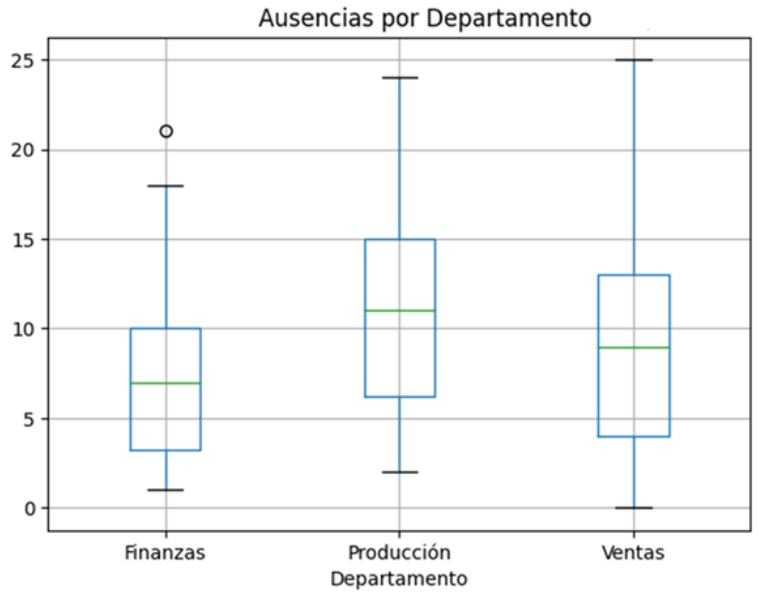
c) ¿Algún departamento presenta una distribución de niveles de estrés claramente diferente a la de los demás? Describe las principales diferencias observadas.

d) ¿Qué patrón general se observa en los niveles de estrés de los empleados al comparar los distintos departamentos?

e) Si la empresa tuviera recursos para implementar inicialmente un programa de bienestar laboral en un solo departamento, ¿cuál recomendarías priorizar? Justifica tu respuesta utilizando la información de la gráfica.

f) ¿Qué departamento parece presentar la combinación más favorable de bienestar laboral? Fundamenta tu respuesta utilizando la distribución de los niveles de estrés.

2.9.3 A continuación, se presentan los diagramas de caja y brazos para comparar el número de días de ausencia anual entre los distintos departamentos.




Con base en los diagramas, responde:

a) ¿Qué departamento presenta la mediana más alta de días de ausencia al año? ¿Qué indica este resultado sobre el ausentismo típico en ese departamento?

b) ¿Cuál departamento muestra la mayor variabilidad en los días de ausencia? Explica cómo lo identificas utilizando el rango intercuartílico y la longitud de los brazos del diagrama.

c) ¿Se observan valores atípicos en alguno de los departamentos? En caso afirmativo, ¿qué podrían representar estos casos dentro del contexto laboral?

d) Considerando simultáneamente el nivel típico de ausencias (mediana) y la dispersión de los datos, ¿qué departamento parece presentar el mejor desempeño en términos de asistencia? Justifica tu respuesta.

e) Si la empresa solo pudiera implementar un programa para reducir el ausentismo en un departamento, ¿cuál recomendarías priorizar? Sustenta tu respuesta utilizando la información del boxplot.

f) ¿La diferencia entre las medianas parece suficientemente grande para sugerir comportamientos distintos entre departamentos o las distribuciones son similares? Explica tu respuesta.

g) ¿Qué departamento parece presentar una distribución más asimétrica? ¿Cómo puede identificarse esta característica en el boxplot?

2.9.4 Posteriormente, se analizaron los días de ausencia según el nivel de estrés de los empleados.

a) ¿Qué nivel de estrés presenta la mediana más alta de días de ausencia al año? ¿Qué indica este resultado sobre el comportamiento típico de ese grupo?

b) ¿Los diagramas sugieren una relación entre el nivel de estrés y el número de días de ausencia? Explica la tendencia observada.

c) ¿Las distribuciones de días de ausencia para los distintos niveles de estrés están claramente separadas o presentan traslape? ¿Qué implica esto sobre la relación entre ambas variables?

d) Con base en la posición de las medianas y la dispersión observada en los diagramas, ¿el nivel de estrés parece ser un factor importante para explicar el ausentismo laboral? Justifica tu respuesta.

e) ¿Qué nivel de estrés presenta la mayor variabilidad en los días de ausencia? ¿Qué podría indicar esta variabilidad sobre el comportamiento de los empleados de ese grupo?

f) ¿Existen empleados con niveles de estrés similares, pero comportamientos muy diferentes en cuanto a ausentismo? ¿Cómo puede identificarse esta situación en los diagramas?

g) ¿Qué nivel de estrés parece mostrar la distribución más homogénea de días de ausencia? Explica cómo llegas a esa conclusión.

h) Si tuvieras que predecir el ausentismo de un empleado únicamente a partir de su nivel de estrés, ¿consideras que los diagramas proporcionan evidencia suficiente? Justifica tu respuesta.

## Verdadero o falso

Indica si cada una de las siguientes afirmaciones es verdadera o falsa. En todos los casos, justifica tu respuesta utilizando conceptos estadísticos apropiados.

1. La media aritmética de un conjunto de datos proporciona información sobre la ubicación central de los datos, no sobre sus valores extremos.
2. Una distribución de frecuencias es una tabla que organiza los datos en clases o categorías, las cuales deben ser exhaustivas y mutuamente excluyentes.
3. Un polígono de frecuencias relativas se construye utilizando las frecuencias relativas de cada clase y muestra la proporción de observaciones que pertenece a cada una de ellas.
4. Una muestra se considera representativa cuando refleja adecuadamente las características relevantes de la población de la que fue extraída.
5. Siempre es posible construir un histograma a partir de un polígono de frecuencias.
6. La moda de un conjunto de datos siempre corresponde a un único valor.
7. El coeficiente de variación es una medida relativa de dispersión que permite comparar la variabilidad entre conjuntos de datos.
8. Las variables cuantitativas discretas únicamente pueden tomar valores enteros.
9. La media, la mediana y la moda de un conjunto de datos pueden ser muy similares; sin embargo, nunca pueden coincidir exactamente.
10. Para calcular el promedio de datos expresados en porcentajes se utiliza la media aritmética.
11. Las medidas de tendencia central y de dispersión pueden calcularse tanto para variables cuantitativas como para variables cualitativas.
12. La mediana de un conjunto de datos siempre debe coincidir con al menos uno de los valores observados en dicho conjunto.

13. En una distribución fuertemente sesgada a la derecha, la media suele ser mayor que la mediana.
14. Dos conjuntos de datos pueden tener la misma media y desviación estándar, pero distribuciones completamente diferentes.
15. Si todos los valores de un conjunto de datos aumentan en una misma constante, la desviación estándar permanece sin cambio.
16. El rango intercuartílico no se ve afectado significativamente por valores atípicos extremos.
17. Una variable cualitativa ordinal puede resumirse mediante la mediana, pero no mediante la media aritmética.
18. Si el coeficiente de variación de un conjunto es mayor que 100%, entonces la desviación estándar es mayor que la media.
19. Una distribución bimodal necesariamente indica que existen dos poblaciones diferentes mezcladas en los datos.
20. Si una distribución es perfectamente simétrica, entonces la media y la mediana coinciden.
21. Una correlación cercana a cero implica necesariamente que no existe relación entre las variables.
22. El histograma y el diagrama de barras representan el mismo tipo de información estadística.
23. En una distribución de frecuencias agrupadas, la suma de las frecuencias relativas siempre debe ser igual a 1.
24. Un valor atípico puede modificar considerablemente la media sin afectar demasiado la mediana.
25. Si una variable tiene desviación estándar igual a cero, entonces todos los datos son idénticos.
26. En una muestra aleatoria simple, todos los elementos de la población tienen la misma probabilidad de ser seleccionados.
27. Una alta correlación entre dos variables implica necesariamente una relación de causalidad.
28. El percentil 90 representa el valor por debajo del cual se encuentra exactamente el 90% de las observaciones.
29. Una distribución uniforme presenta menor concentración de datos alrededor de la media que una distribución normal.

30. Si dos variables presentan asociación negativa, al aumentar una de ellas la otra tiende a disminuir.
31. En un boxplot, los valores atípicos siempre representan errores de captura o medición.
32. La suma de las desviaciones de todos los datos respecto a la media aritmética siempre es igual a cero.

## MINICASO: "El reto de recuperar la satisfacción del cliente"

La cadena de cafeterías Café Urbano, con presencia en distintas zonas de la ciudad, ha comenzado a recibir opiniones contradictorias en redes sociales y plataformas de reseñas. Mientras algunos clientes destacan la calidad del café y el ambiente de las sucursales, otros reportan largos tiempos de espera y una atención inconsistente.

La dirección de operaciones sospecha que el problema no afecta por igual a todas las sucursales, sino que podría estar concentrado en ciertos puntos de venta específicos. De ser así, una mala experiencia localizada podría estar dañando la percepción de toda la marca.

Con el objetivo de identificar posibles áreas de mejora, se realizó una encuesta piloto a 30 clientes seleccionados aleatoriamente en sus tres sucursales. A cada cliente se le solicitó la siguiente información:

1. La sucursal en la que fue atendido: norte, centro, sur
2. Su nivel de satisfacción con el servicio: alto, medio, bajo
3. El tiempo de espera, en minutos, para recibir su pedido.

El equipo de analítica consolidó la información en el archivo  "Cafeterías Café Urbano". La gerencia necesita evidencia clara para decidir en qué sucursal intervenir primero y qué tipo de problema es el más urgente.

| Cliente | Sucursal | Satisfacción | Tiempo_espera |
|---|---|---|---|
| 1 | Centro | Media | 6 |
| 2 | Norte | Alta | 4 |
| 3 | Sur | Baja | 10 |
| 4 | Centro | Alta | 3 |
| 5 | Sur | Media | 7 |
| 6 | Norte | Alta | 5 |
| 7 | Centro | Baja | 12 |
| 8 | Centro | Media | 6 |
| 9 | Sur | Alta | 4 |

| | | | |
|---|---|---|---|
| 10 | Norte | Media | 8 |
| 11 | Centro | Alta | 2 |
| 12 | Norte | Media | 7 |
| 13 | Sur | Alta | 5 |
| 14 | Centro | Media | 6 |
| 15 | Norte | Baja | 11 |
| 16 | Sur | Media | 9 |
| 17 | Centro | Alta | 3 |
| 18 | Norte | Alta | 4 |
| 19 | Sur | Media | 8 |
| 20 | Centro | Alta | 5 |
| 21 | Centro | Media | 7 |
| 22 | Norte | Media | 9 |
| 23 | Sur | Alta | 6 |
| 24 | Centro | Alta | 4 |
| 25 | Norte | Baja | 12 |
| 26 | Centro | Media | 8 |
| 27 | Sur | Media | 7 |
| 28 | Centro | Alta | 5 |
| 29 | Norte | Alta | 3 |
| 30 | Sur | Alta | 6 |

Utilizando R construye las siguientes tablas y gráficas:

- tablas de frecuencias para “Sucursal” y “Satisfacción”
- tabla de frecuencias con intervalos de 2 minutos para “Tiempo_espera”
- diagrama de barras para las variables “Sucursal” y “Satisfacción”
- diagrama circular para las variables “Sucursal” y “Satisfacción”
- histograma, polígono de frecuencias y ojiva para “Tiempo de espera”
- diagrama de caja y brazos (boxplot) para “Tiempo de espera por sucursal”

RESULTADOS OBTENIDOS CON R:

Tablas de frecuencia:

```
> # Tabla de frecuencia para Sucursal
> table(datos$Sucursal)

Centro  Norte    Sur
    12      9      9
```

```
> # Tabla de frecuencia para Satisfacción
> table(datos$Satisfaccion)

 Alta  Baja Media
   14     4    12

> # Tabla de frecuencia para Tiempo de espera (intervalos de 2 minutos)
> breaks <- seq(0, 14, by=2)
> table(cut(datos$Tiempo_espera_min, breaks=breaks, right=FALSE))

 (0,2]  (2,4]  (4,6]  (6,8] (8,10] (10,12]
     0      4      8      9      5       4
```

Con base en las tablas de frecuencias obtenidas, responde las siguientes preguntas:

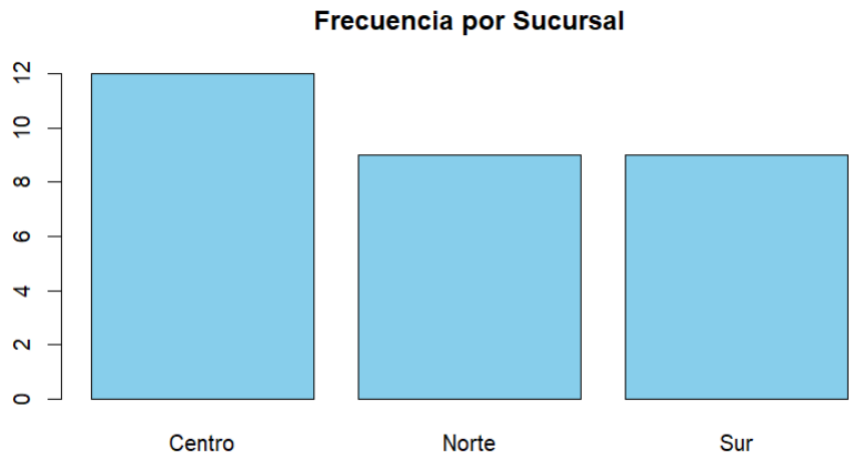
a) ¿Qué sucursal registra el mayor número de clientes encuestados?

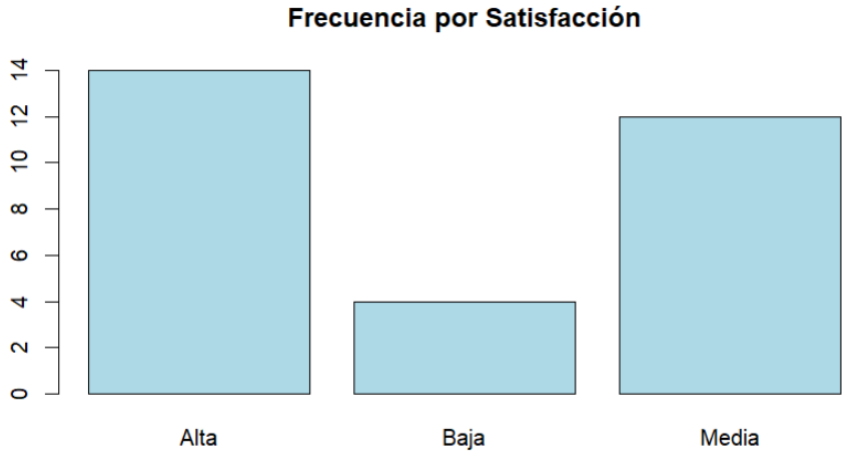
b) ¿Cuál es el nivel de satisfacción más frecuente entre los clientes?

c) ¿Cuántos clientes presentan un tiempo de espera menor a 6 minutos?

d) ¿Qué proporción de clientes tuvo un tiempo de espera entre 6 y 8 minutos?

e) ¿Qué porcentaje del total de clientes corresponde a cada sucursal?






Con base en los diagramas de barras elaborados, responde las siguientes preguntas:

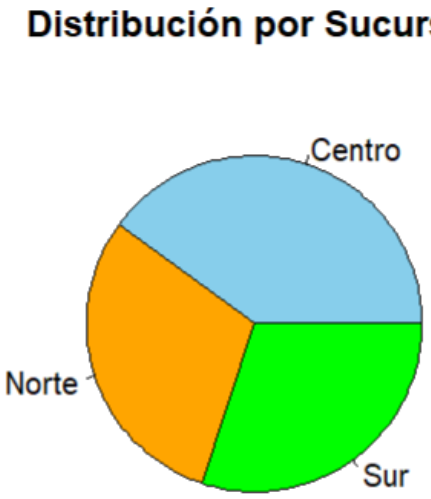
a) ¿Qué sucursal concentra el mayor número de clientes encuestados?

b) ¿Cuál es el nivel de satisfacción menos frecuente entre los clientes?

c) ¿Se observan diferencias importantes entre las sucursales en cuanto al número de clientes atendidos? Explica tu respuesta.

d) A simple vista, ¿se identifica algún patrón en los niveles de satisfacción de los clientes? Describe lo que observas.

e) ¿Existen categorías que presenten barras con la misma frecuencia? En caso afirmativo, ¿qué interpretación puede darse a este resultado?




Con base en los diagramas circulares (gráficas de pastel) elaborados, responde las siguientes preguntas:

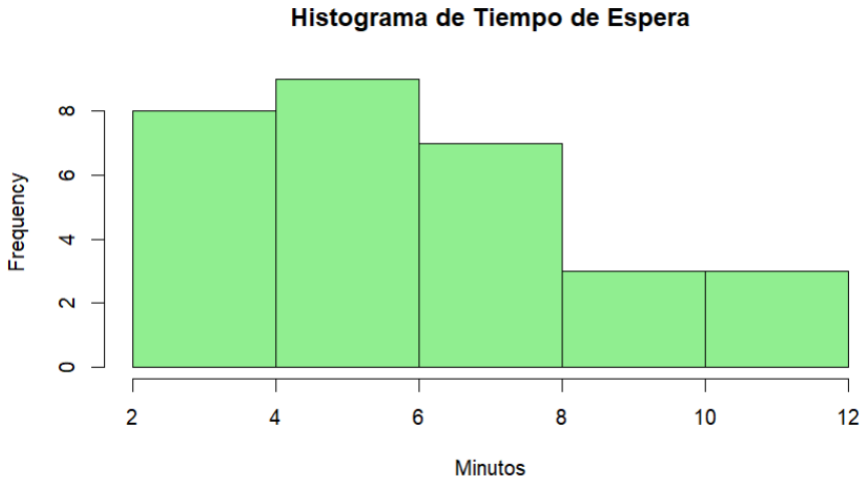
a) ¿Qué sucursal representa la mayor proporción de clientes encuestados?

b) ¿Qué porcentaje de clientes reportó un nivel de satisfacción “Alta”?

c) ¿Cuál es el segmento más pequeño en los diagramas y qué información proporciona sobre la distribución de los clientes?

d) Al comparar los diagramas circulares de sucursal y nivel de satisfacción, ¿qué variable parece presentar una distribución más homogénea? Explica tu respuesta.

e) ¿Es posible inferir alguna relación entre la sucursal y el nivel de satisfacción únicamente a partir de las gráficas de pastel? Justifica tu respuesta.




Con base en el histograma de la variable “Tiempo_espera”, responde las siguientes preguntas:

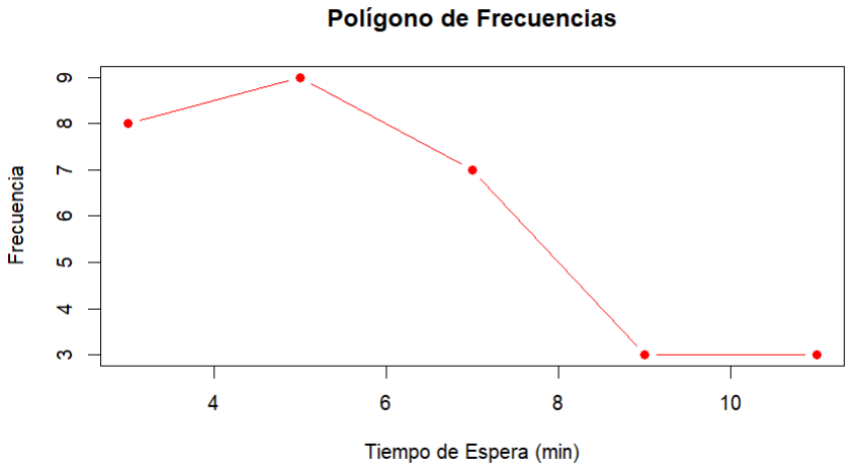
a) ¿En qué intervalo de tiempo de espera se concentra la mayor cantidad de clientes?

b) ¿Se observan tiempos de espera poco frecuentes o valores extremos? ¿Qué podrían indicar?

c) ¿La distribución del tiempo de espera parece ser simétrica o presenta algún tipo de sesgo? Explica tu respuesta.

d) ¿Cuál es el intervalo con mayor frecuencia y qué interpretación puede darse a este resultado?

e) Al comparar el histograma con las medidas de tendencia central (media y mediana), ¿qué información adicional puede obtenerse sobre la distribución de los tiempos de espera?




Con base en el polígono de frecuencias de la variable “Tiempo_espera”, responde las siguientes preguntas:

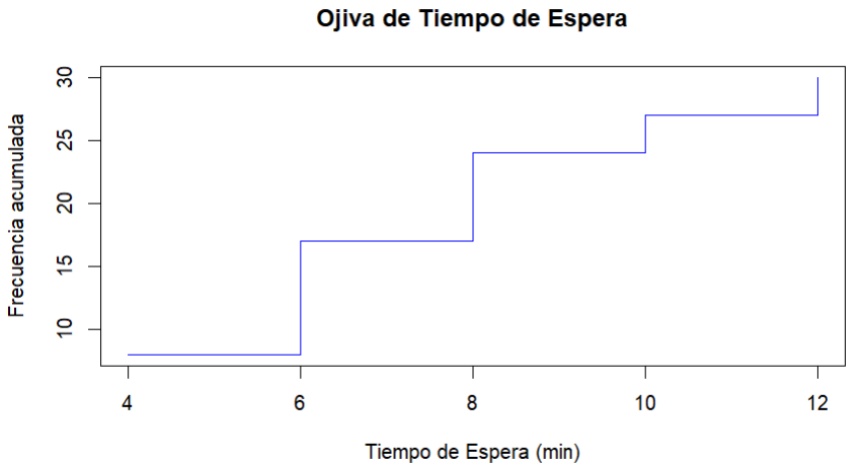
a) ¿Cuál es el punto más alto del polígono y qué representa en términos de frecuencia?

b) ¿Existen intervalos en los que la frecuencia disminuya de manera notable? Describe lo que observas.

c) ¿Cómo se compara la forma del polígono de frecuencias con la del histograma correspondiente?

d) ¿Qué tendencia general puede identificarse en la distribución de los tiempos de espera?

e) A partir del polígono, ¿se pueden identificar posibles valores extremos o intervalos poco frecuentes? Explica tu respuesta.




Con base en la ojiva de la variable “Tiempo_espera”, responde las siguientes preguntas:

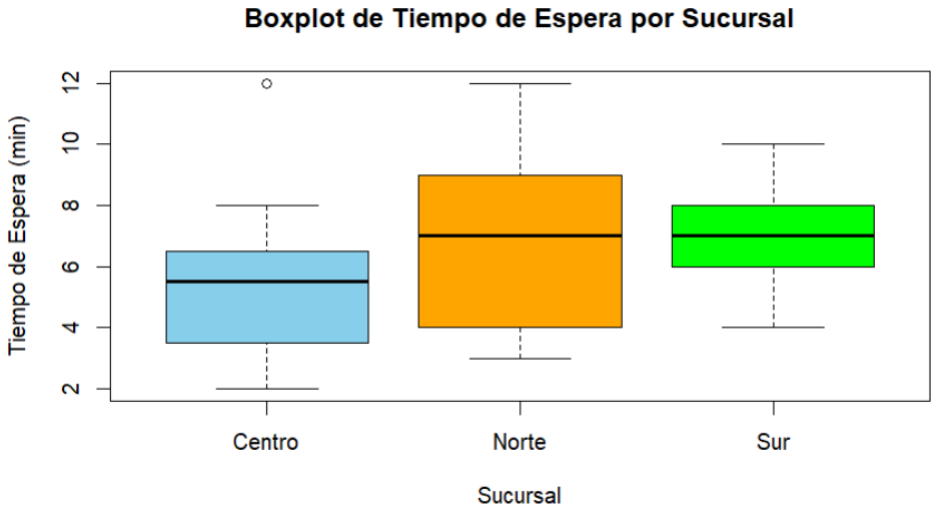
a) ¿Cuántos clientes presentan un tiempo de espera menor a 6 minutos?

b) ¿Qué proporción de clientes tiene un tiempo de espera inferior a 8 minutos?

c) ¿Cómo se comporta la frecuencia acumulada conforme aumenta el tiempo de espera? Describe la tendencia observada.

d) ¿En qué intervalo de tiempo se observa el mayor incremento en la frecuencia acumulada? ¿Qué indica esto sobre los datos?

e) ¿Qué porcentaje de clientes experimenta tiempos de espera mayores a 10 minutos?




Analiza el diagrama de caja y brazos (boxplot) correspondiente al tiempo de espera de los clientes en cada sucursal y responde las siguientes preguntas:

a) ¿Qué sucursal presenta la mediana más alta de tiempo de espera?

b) ¿Cuál sucursal muestra la mayor variabilidad en los tiempos de espera? Explica cómo lo identificas a partir del diagrama.

c) ¿Se observan valores atípicos en alguna sucursal? En caso afirmativo, ¿a qué sucursal pertenecen y qué podrían representar en términos operativos?

d) ¿Qué información proporciona el rango intercuartílico de cada sucursal sobre la dispersión de los tiempos de espera?

e) Al comparar las sucursales, ¿en cuál de ellas los clientes parecen experimentar menores tiempos de espera de manera general? Justifica tu respuesta utilizando elementos del diagrama.

Utilizando R, calcula las siguientes medidas descriptivas para la variable “Tiempo_espera” considerando todos los clientes en conjunto:

- Medidas de tendencia central: media, mediana y moda
- Medidas de dispersión: rango, rango intercuartil, varianza, desviación estándar, desviación absoluta media y coeficiente de variación
- Medidas de posición: cuartiles 1, 2 y 3 y percentiles 10, 80 y 95

Posteriormente, repite el mismo análisis de manera separada para cada sucursal (Norte, Centro y Sur), con el fin de comparar el comportamiento de los tiempos de espera entre ellas.

RESULTADOS:

| Sucursal | Media | Mediana | Moda | Min | Max |
|---|---|---|---|---|---|
| TOTAL | 6.4 | 6 | 6 | 2 | 12 |
| Centro | 5.58 | 5.5 | 6 | 2 | 12 |
| Norte | 7 | 7 | 4 | 3 | 12 |
| Sur | 6.89 | 7 | 7 | 4 | 10 |

| Sucursal | Rango | Rango intercuartílico (IQR) | Varianza | Desv_Estándar | Desv_Abs_Media (MAD) | Coef_Variación (CV) |
|---|---|---|---|---|---|---|
| TOTAL | 10 | 3.75 | 7.08 | 2.66 | 2.12 | 0.416 |
| Centro | 10 | 2.5 | 7.17 | 2.68 | 1.92 | 0.48 |
| Norte | 9 | 5 | 10.5 | 3.24 | 2.67 | 0.463 |
| Sur | 6 | 2 | 3.61 | 1.9 | 1.46 | 0.276 |

| Sucursal | Q1 | Q2 | Q3 | P10 | P80 | P95 |
|---|---|---|---|---|---|---|
| TOTAL | 4.25 | 6 | 8 | 3 | 8.2 | 11.55 |
| Centro | 3.75 | 5.5 | 6.25 | 3 | 6.8 | 9.8 |
| Norte | 4 | 7 | 9 | 3.8 | 9.8 | 11.6 |
| Sur | 6 | 7 | 8 | 4.8 | 8.4 | 9.6 |

Con base en las medidas de tendencia central calculadas para la variable “Tiempo_espera” responde las siguientes preguntas:

a) ¿Cuál es el tiempo de espera promedio considerando a todos los clientes?

b) Al comparar las sucursales, ¿cuál presenta la media más alta de tiempo de espera y qué interpretación puede darse a este resultado?

c) Si existen valores extremos en los tiempos de espera, ¿la media representa adecuadamente el comportamiento general de los datos? Explica.

d) ¿Qué diferencias observas entre la media global y las medias calculadas para cada sucursal?

e) Si la empresa quisiera mejorar la eficiencia operativa, ¿qué sucursal debería atender prioritariamente con base en la media de los tiempos de espera?

f) ¿Cuál es el tiempo de espera que divide al conjunto de clientes en dos partes iguales?

g) Al comparar las sucursales, ¿cuál presenta la mediana más baja y qué indica esto sobre la experiencia de los clientes?

h) ¿Existen diferencias importantes entre la media y la mediana? En caso afirmativo, ¿qué podría indicar esto sobre la forma de la distribución de los tiempos de espera?

i) ¿Qué ventaja tiene utilizar la mediana en lugar de la media cuando existen valores atípicos?

j) ¿Cuál es el tiempo de espera más frecuente entre todos los clientes?

k) ¿Existe algún tiempo de espera que se repita con mucha frecuencia y que pueda sugerir un patrón en el servicio?

l) ¿Qué información aporta la moda que no puede observarse directamente a partir de la media o la mediana?

Con base en las medidas de dispersión calculadas para la variable “Tiempo_espera”, responde las siguientes preguntas.

a) ¿Cuál es la diferencia entre el tiempo de espera máximo y el mínimo considerando a todos los clientes?

b) Al comparar las sucursales, ¿cuál presenta el mayor rango y qué indica esto sobre la variabilidad de los tiempos de espera?

c) Si existen valores extremos, ¿el rango representa adecuadamente la dispersión general de los datos? Explica.

d) ¿Cómo cambiaría el rango si se eliminaran los valores atípicos?

e) ¿Qué sucursal parece tener la mayor dispersión en los tiempos de espera según el rango?

f) ¿Cuál es el intervalo que contiene el 50% central de los tiempos de espera de los clientes?

g) ¿Qué sucursal presenta el mayor rango intercuartílico y qué interpretación puede darse a este resultado?

h) Al comparar el IQR con el rango total, ¿qué se puede concluir sobre la distribución de los datos?

i) ¿Qué proporción de clientes se encuentra dentro del rango intercuartílico?

j) ¿Cómo puede ayudar el IQR a identificar la presencia de valores atípicos?

k) ¿Qué tan dispersos se encuentran los tiempos de espera con respecto a la media?

l) Al comparar las sucursales, ¿cuál muestra la mayor dispersión según la varianza y la desviación estándar? ¿Qué implica esto operativamente?

m) ¿Qué relación existe entre la varianza y la desviación estándar?

n) ¿De qué manera afectan los valores extremos a la varianza y a la desviación estándar?

o) ¿Qué sucursal parece mostrar mayor consistencia en sus tiempos de espera con base en estas medidas?

p) ¿Cuál es la desviación promedio de los tiempos de espera respecto a la media para el total de clientes?

q) ¿Qué sucursal presenta mayor consistencia en los tiempos de espera según la desviación absoluta media?

r) Al comparar la MAD con la desviación estándar, ¿qué diferencias observas en la interpretación de ambas medidas?

s) ¿Qué ventaja tiene utilizar la MAD en presencia de valores extremos?

t) ¿Cuál de las medidas analizadas parece describir mejor la dispersión típica de los tiempos de espera? Justifica tu respuesta.

u) ¿Qué tan grande es la dispersión de los tiempos de espera en relación con la media?

v) Al comparar las sucursales, ¿cuál presenta la mayor variabilidad relativa?

w) ¿Qué valor del coeficiente de variación indicaría que los tiempos de espera son más consistentes?

x) ¿Por qué es útil utilizar el coeficiente de variación para comparar sucursales con medias diferentes?

y) Si una sucursal presenta un coeficiente de variación muy alto, ¿qué implicaciones podría tener para la gestión operativa y la experiencia del cliente?

Con base en los cuantiles calculados para la variable “Tiempo_espera”, responde las siguientes preguntas:

a) ¿Cuáles son los tiempos de espera correspondientes a los cuartiles $Q_1$, $Q_2$y $Q_3$?

b) Al comparar las sucursales, ¿qué diferencias observas entre sus cuartiles?

c) ¿Qué porcentaje de clientes presenta tiempos de espera menores que $Q_1$?

d) ¿Qué utilidad tiene conocer $Q_3$para planear mejoras en el servicio y reducir tiempos de espera?

e) ¿Los cuartiles permiten detectar posibles valores extremos? Explica tu respuesta.

f) ¿Cuál es el tiempo de espera por debajo del cual se encuentra el 10% de los clientes?

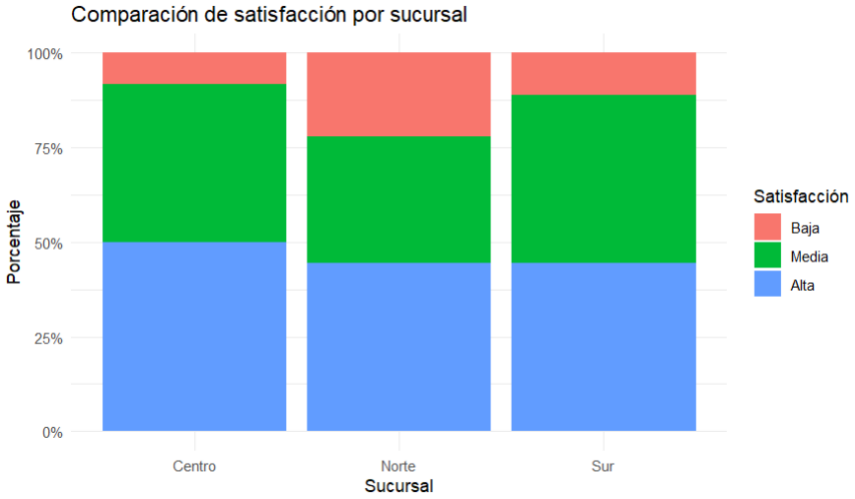
g) ¿Qué porcentaje de clientes presenta tiempos de espera inferiores a los percentiles $P_{80}$y $P_{95}$?

h) ¿Qué información aportan los percentiles altos sobre los clientes con mayores tiempos de espera?

i) Al comparar las sucursales, ¿cuál presenta los percentiles más altos y qué interpretación puede darse a este resultado?

j) ¿Cómo podrían utilizarse estos percentiles para establecer estándares de servicio o metas operativas en las sucursales?

Problemas de asociación y de comparación:




Con base en la gráfica que compara el nivel de satisfacción de los clientes entre las distintas sucursales, responde las siguientes preguntas:

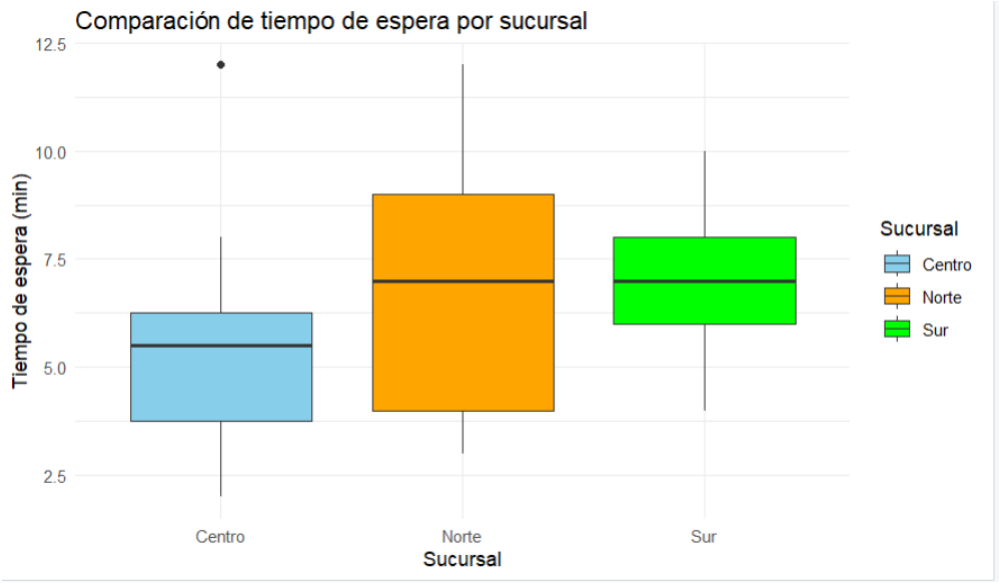
a) ¿Qué sucursal presenta la mayor proporción de clientes con nivel de satisfacción "Alto"?

b) ¿Cuál sucursal tiene el porcentaje más elevado de clientes con satisfacción "Baja"?

c) ¿Existe alguna sucursal cuya distribución de niveles de satisfacción sea claramente diferente a la de las demás? Explica tu respuesta.

d) ¿Qué patrón general observas en la satisfacción de los clientes entre las sucursales? ¿Predominan los clientes satisfechos, neutrales o insatisfechos?

e) Si la empresa tuviera que priorizar acciones de mejora en una sola sucursal, ¿cuál recomendarías y por qué, con base en la información de la gráfica?




Con base en el diagrama de caja y brazos del tiempo de espera por sucursal, responde las siguientes preguntas:

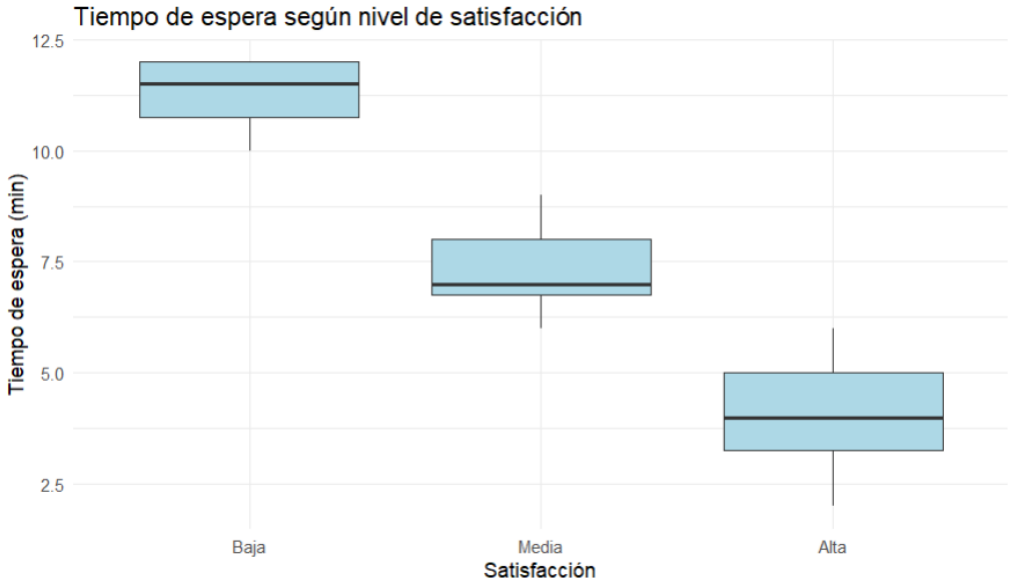
a) ¿Qué sucursal presenta la mediana más alta de tiempo de espera?

b) ¿Cuál sucursal muestra la mayor variabilidad en los tiempos de espera? Explica cómo lo identificas en el diagrama.

c) ¿Se observan valores atípicos en alguna sucursal? En caso afirmativo, ¿qué podrían sugerir sobre el funcionamiento del servicio en esa sucursal?

d) Considerando tanto la mediana como la dispersión de los datos, ¿qué sucursal parece mostrar el mejor desempeño en términos de tiempos de espera? Justifica tu respuesta.

e) Si la empresa solo pudiera intervenir en una sucursal para reducir los tiempos de espera, ¿cuál recomendarías priorizar y por qué?




Con base en los diagramas que relacionan el tiempo de espera con el nivel de satisfacción de los clientes, responde las siguientes preguntas:

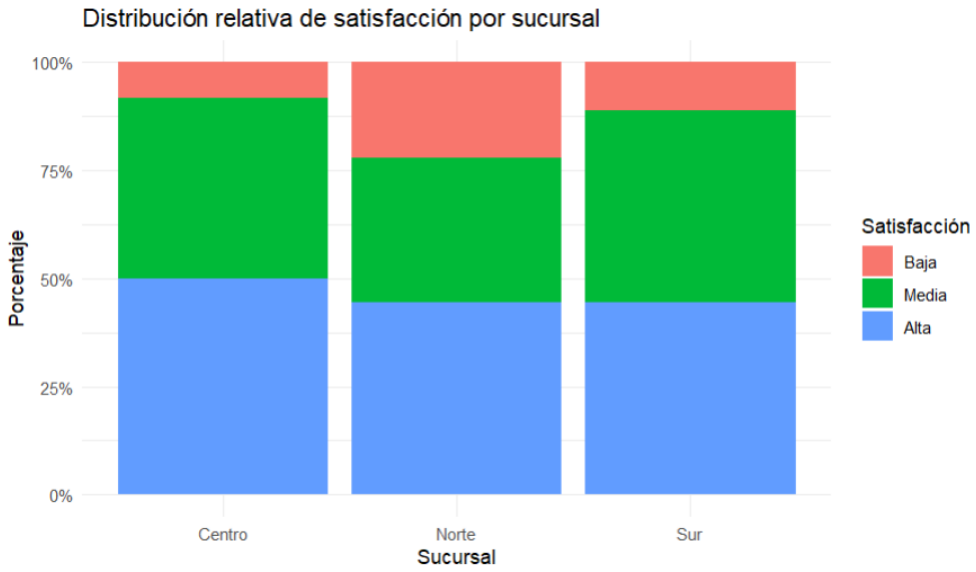
a) ¿Qué nivel de satisfacción presenta los mayores tiempos de espera?

b) ¿Se observa un patrón que sugiera que, a mayor tiempo de espera, menor es el nivel de satisfacción del cliente? Explica tu respuesta.

c) ¿Los distintos niveles de satisfacción se distinguen claramente por sus distribuciones de tiempo de espera, o existe traslape entre ellas?

d) Con base en la información de los diagramas, ¿el tiempo de espera parece influir de manera importante en la percepción del cliente sobre el servicio? Justifica tu respuesta.

e) ¿Qué nivel de satisfacción presenta la mayor variabilidad en los tiempos de espera y qué podría indicar esto sobre la experiencia de los clientes?




Con base en la gráfica de frecuencias relativas que compara el nivel de satisfacción entre sucursales, responde las siguientes preguntas:

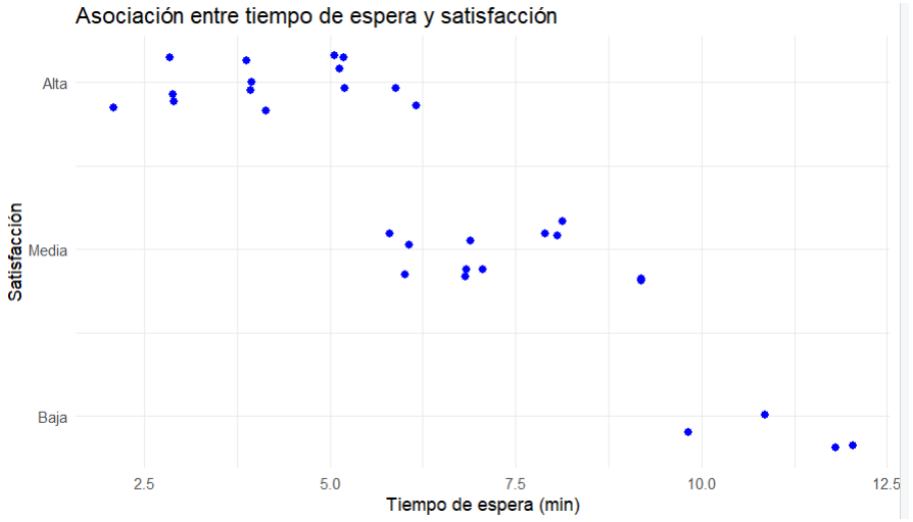
a) ¿Qué sucursal presenta la mayor proporción de clientes satisfechos dentro de su propia sucursal?

b) ¿Cuál sucursal muestra la mayor proporción relativa de clientes insatisfechos?

c) ¿Se observa una posible asociación entre la sucursal y el nivel de satisfacción de los clientes? Explica cómo se refleja esta relación en la gráfica.

d) ¿Qué sucursal parece tener la distribución más equilibrada entre los distintos niveles de satisfacción? Justifica tu respuesta.

e) Si la empresa tuviera que reconocer a una sucursal por la calidad de su servicio, ¿cuál elegirías con base en las frecuencias relativas observadas?




Con base en el diagrama de dispersión que relaciona el tiempo de espera con el nivel de satisfacción de los clientes, responde las siguientes preguntas:

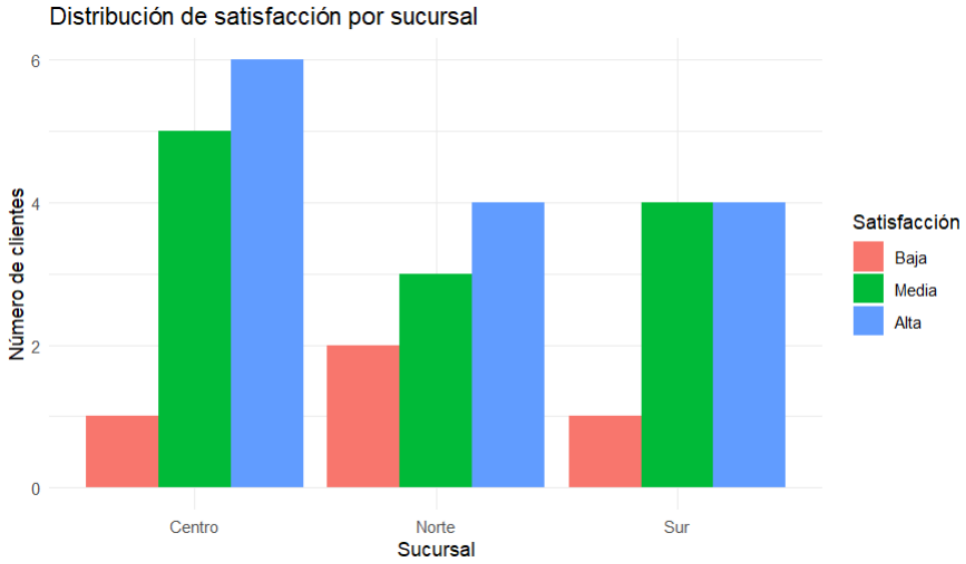
a) ¿El diagrama sugiere una relación positiva, negativa o inexistente entre el tiempo de espera y el nivel de satisfacción? Explica tu respuesta.

b) Si la relación entre ambas variables fuera fuerte, ¿cómo se observaría la tendencia en el diagrama?

c) ¿Cómo se representaría en el diagrama un patrón de independencia entre el tiempo de espera y la satisfacción de los clientes?

d) Si todos los clientes con alta satisfacción presentaran tiempos de espera bajos, ¿en qué zona del diagrama se concentrarían los puntos?

e) Si el diagrama mostrara una tendencia claramente decreciente, ¿qué decisiones operativas o de servicio podrían tomarse a partir de esa evidencia?




Con base en la gráfica que compara la distribución de los niveles de satisfacción entre sucursales, responde las siguientes preguntas:

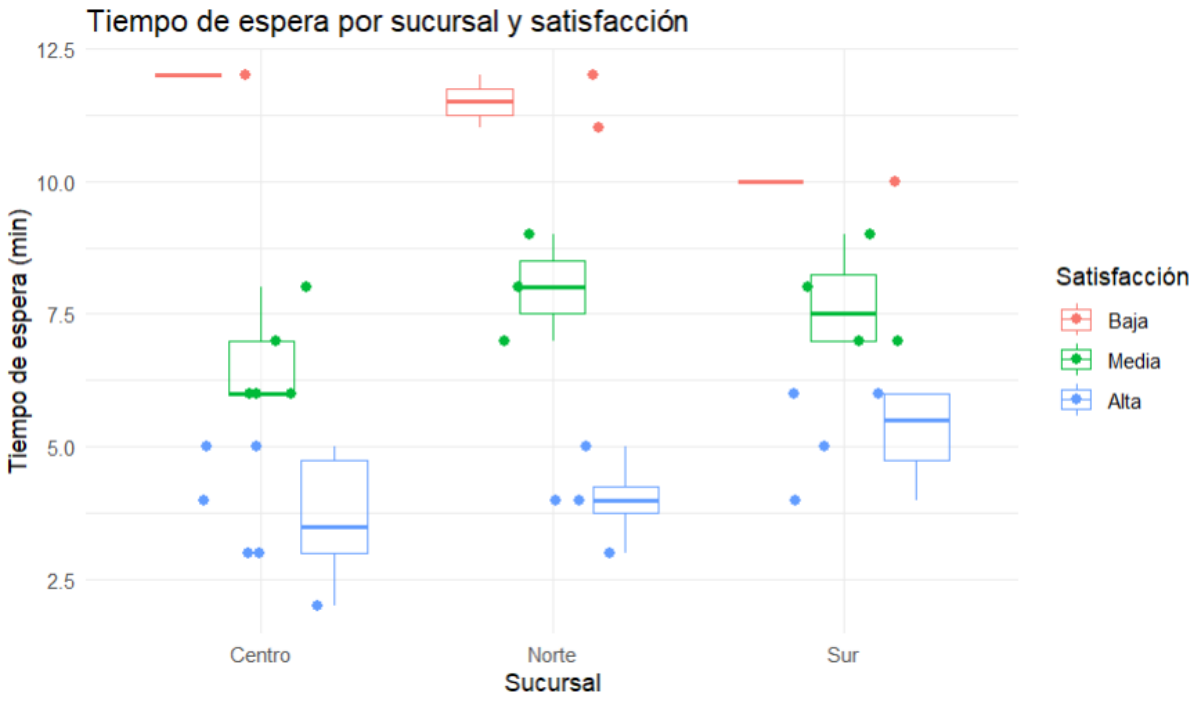
a) ¿Qué sucursal presenta los niveles de satisfacción más bajos entre sus clientes?

b) ¿Existen sucursales con distribuciones de satisfacción similares? En caso afirmativo, ¿cuáles y qué similitudes observas?

c) ¿Qué sucursal destaca por presentar una mayor cantidad o proporción de clientes con satisfacción Alta?

d) ¿Cómo cambia la percepción del servicio por parte de los clientes entre las distintas sucursales?

e) Con base en la gráfica, ¿parece existir una relación entre la sucursal y el nivel de satisfacción de los clientes? Justifica tu respuesta.




Con base en la gráfica de dispersión de los tiempos de espera por sucursal, responde las siguientes preguntas:

a) ¿Qué sucursal presenta la mayor concentración de tiempos de espera altos?

b) ¿Se observan agrupaciones de puntos que puedan sugerir problemas operativos específicos en alguna sucursal? Explica tu respuesta.

c) ¿El patrón observado en la gráfica de dispersión confirma o contradice lo mostrado en el diagrama de caja y brazos? Justifica tu respuesta.

d) ¿Qué sucursal presenta más observaciones muy por encima del límite superior mostrado en el boxplot y qué podría indicar esto?

e) A partir de la distribución de los puntos, ¿qué conclusiones puedes obtener sobre la consistencia del servicio en cada sucursal?

# TEMA 3: Probabilidad

## 3.1 Conjuntos y Conteo

3.1.1 **Espacio muestral.** Para cada uno de los siguientes experimentos aleatorios, construye el espacio muestral $S$, es decir, el conjunto de todos los resultados posibles del experimento.

a) Se lanzan tres monedas simultáneamente y se registra el resultado obtenido en cada una.

b) Se seleccionan cuatro personas al azar y se registra si el día de nacimiento de cada una corresponde a un número par (P) o impar (N).

c) Se extrae una bola al azar de cada una de las siguientes urnas y se registra el color obtenido en cada extracción:

- Urna 1: 3 bolas rojas (R), 2 blancas (B) y 1 azul (A)
- Urna 2: 2 bolas blancas (B) y 3 azules (A)

d) Se mezclan todas las bolas de ambas urnas del inciso anterior en una sola urna. Posteriormente, se extraen dos bolas al azar sin reemplazo y se registra el color de cada bola en el orden en que fueron extraídas.

e) Se lanza una moneda y un dado. Se registra el resultado de ambos.

f) Se lanzan dos dados y se registra el número obtenido en cada uno.

g) Una familia tiene tres hijos. Se registra el sexo de cada hijo en orden de nacimiento.

h) Un inversionista observa durante cuatro días consecutivos el comportamiento de una acción y registra si el precio sube (S) o baja (B) cada día.

i) Un banco evalúa las solicitudes de crédito de tres clientes consecutivos y registra si cada solicitud es aprobada (A) o rechazada (R).

j) Un centro de atención telefónica registra el resultado de cuatro llamadas consecutivas: resuelta (R) o no resuelta (N).

3.1.2 En una cafetería se observa el resultado de un cliente que solicita una bebida. El proceso puede terminar en uno de los siguientes seis resultados, igualmente probables:

- Resultado 1: Recibe la bebida correcta y a tiempo
- Resultado 2: Recibe la bebida correcta con retraso
- Resultado 3: Recibe una bebida incorrecta pero la acepta
- Resultado 4: Recibe una bebida incorrecta y solicita cambio
- Resultado 5: Su pedido se cancela por falta de ingredientes
- Resultado 6: Abandona la fila antes de recibir la bebida

Se observa el resultado de un cliente seleccionado al azar.

a) Define los siguientes eventos:

- el espacio muestral Ω
- A = {el cliente recibe la bebida correcta}
- B = {el cliente recibe alguna bebida}
- C = {existe algún problema en el servicio} (cualquier situación distinta de recibir la bebida correcta y a tiempo constituye un problema en el servicio)
- D = {el cliente recibe dos bebidas simultáneamente, resultado imposible para este modelo}
- E = {el cliente recibe la bebida correcta pero no de forma perfecta}
- $A \cup B$
- $A \cup C$
- $A \cap B$
- $A \cap C$
- $A \cup B \cup C$
- $A^c$
- $E^c$
- Valida $E \subset A$

b) Calcula la probabilidad de cada uno de los eventos anteriores

3.1.3 Se lanzan dos dados equilibrados de manera simultánea. Cada resultado se representa mediante un par ordenado $(x, y)$, donde:

- $x$ es el número de puntos obtenido en el dado 1
- $y$ es el número de puntos obtenido en el dado 2

Con base en este experimento:

a) Determina el espacio muestral $\Omega$.

b) Identifica los eventos simples que conforman el evento

$$A = \{\text{al menos uno de los dados muestra un } 1\}$$

c) Identifica los eventos simples que conforman el evento:

$$B = \{\text{la suma de los puntos obtenidos en ambos dados es igual a } 7\}$$

d) Identifica los eventos simples que conforman el evento

$C = \{$el número de puntos obtenido en el dado 1 es menor que el obtenido en el dado 2$\}$

e) Calcula las probabilidades $P(A)$, $P(B)$y $P(C)$.

f) Determina los eventos simples que integran:

$$A \cap B$$
$$A \cup B$$
$$A \cap C$$
$$B \cap C$$
$$A^c$$
$$C^c$$

g) Calcula las probabilidades correspondientes a los eventos anteriores.

3.1.4 Una empresa selecciona aleatoriamente dos candidatos para una entrevista final. Cada candidato proviene de una de las siguientes áreas académicas:

- Administración (A)
- Economía (E)
- Ingeniería (I)

Las tres áreas tienen la misma probabilidad de aparecer. Se registra el área académica del primer y del segundo candidato seleccionados.

a) Determina el espacio muestral de todas las parejas ordenadas posibles.

b) Identifica los eventos simples que pertenecen al evento

$$A = \{\text{al menos uno de los candidatos proviene de Administración}\}$$

c) Identifica los eventos simples que pertenecen al evento

$$B = \{\text{los candidatos provienen de áreas distintas}\}$$

d) Identifica los eventos simples que pertenecen al evento

$$C = \{\text{el primer candidato proviene de Economía}\}$$

e) Calcula $P(A)$, $P(B)$y $P(C)$.

## 3.2 Incertidumbre, fenómenos aleatorios y probabilidad

3.2.1 Analiza cada una de las siguientes situaciones y clasifícala como fenómeno determinístico, aleatorio o mixto. En cada caso, explica los elementos que sustentan tu clasificación.

a) El resultado de lanzar un dado equilibrado.

b) El Producto Interno Bruto (PIB) de México al cierre de este año.

c) La hora real de salida de un vuelo comercial programado para hoy.

d) El precio del dólar estadounidense mañana a las 12:00 horas.

e) El tipo de cambio peso-dólar al cierre de la jornada bursátil de mañana.

f) La inflación anual de México al cierre de este año.

g) El resultado de una encuesta electoral realizada a una muestra de votantes.

h) La temperatura registrada a las 12:00 horas del próximo 1 de mayo.

i) La fecha y hora exactas del próximo sismo de magnitud mayor a 7 en la Ciudad de México.

j) La cantidad de "me gusta" que recibirá una publicación en redes sociales durante la próxima semana.

k) El tiempo de espera de un paciente en el área de urgencias de un hospital.

l) La cantidad de lluvia que caerá mañana en tu ciudad.

m) Las ventas totales de una cafetería durante el día de mañana.

n) La temperatura máxima registrada durante el día de hoy.

o) El número de rayos que ocurrirán durante una tormenta eléctrica.

p) La hora del amanecer de mañana en la Ciudad de México.

q) El nivel medio del mar dentro de 100 años.

r) La caída de un rayo sobre un punto específico de una ciudad durante una tormenta.

s) El número de réplicas registradas después de un terremoto.

t) La calificación promedio que obtendrá un grupo en el examen final del curso.

u) El tiempo que transcurre desde que un cliente llega a un banco hasta que es atendido.

3.2.2 Indica si la afirmación es verdadera o falsa.

a) Si un fenómeno puede predecirse con mucha precisión, entonces es determinístico.

b) Un fenómeno aleatorio puede presentar patrones.

c) Un fenómeno mixto contiene componentes predecibles y componentes impredecibles.

d) Todo fenómeno económico es completamente aleatorio.

e) La probabilidad mide el grado de incertidumbre.

f) Que un fenómeno sea aleatorio no significa que todos sus resultados sean igualmente probables.

g) Una predicción puede ser muy precisa y aun así resultar incorrecta.

h) Conocer el resultado de un fenómeno aleatorio antes de que ocurra es imposible.

i) Si un fenómeno es completamente determinístico, entonces no existe incertidumbre sobre su resultado cuando se conocen todas las condiciones relevantes.

j) Si un meteorólogo pronostica un 80% de probabilidad de lluvia y no llueve, entonces el pronóstico fue incorrecto.

## 3.3 Enfoques de probabilidad: clásico o de Laplace, frecuentista y subjetivo

3.3.3 Lee cuidadosamente cada situación e identifica el enfoque de probabilidad que se está utilizando:

- Clásico o de Laplace
- Frecuentista
- Subjetivo

En cada caso, indica el enfoque utilizado y justifica brevemente tu respuesta.

a) Al lanzar un dado equilibrado de seis caras, ¿cuál es la probabilidad de obtener un número par?

b) Una fábrica produce 10,000 focos al día y, durante una jornada de producción, se observó que 1,230 fueron defectuosos. Con base en esta información, ¿cuál es la probabilidad de seleccionar al azar un foco defectuoso?

c) Un analista económico estima que existe un 70% de probabilidad de que ocurra una recesión el próximo año, considerando indicadores como la inflación, las tasas de interés y la situación geopolítica internacional.

d) Se selecciona una carta al azar de una baraja estándar de 52 cartas. ¿Cuál es la probabilidad de obtener un as?

e) Un meteorólogo estima una probabilidad del 40% de tormenta severa para mañana utilizando modelos meteorológicos, observaciones recientes y su experiencia profesional.

f) Se venden 200 boletos numerados para una rifa y cada participante adquiere un solo boleto. ¿Cuál es la probabilidad de ganar la rifa?

g) Durante una sesión de entrenamiento, un futbolista convirtió 72 de 100 tiros a portería. Utilizando esta información, se estima la probabilidad de que anote en su próximo intento.

h) Los fundadores de un startup consideran que existe un 60% de probabilidad de que la empresa continúe operando dentro de dos años, basándose en su experiencia, las condiciones del mercado y la competencia.

## 3.4 Axiomas de probabilidad

3.4.1 En una encuesta realizada a estudiantes universitarios sobre el uso de plataformas de streaming, se obtuvieron los siguientes resultados:

- La probabilidad de que un alumno tenga una suscripción a Netflix es $P(N) = 0.62$
- La probabilidad de que un alumno tenga una suscripción a Prime Video es $P(PV) = 0.41$
- La probabilidad de que un alumno tenga una suscripción al menos a una de las dos plataformas es $P(N \cup PV) = 0.78$

Con base en esta información:

a) Calcula la probabilidad de que un estudiante tenga ambas plataformas: $P(N \cap PV)$

b) Calcula la probabilidad de que un estudiante tenga Netflix pero no Prime Video: $P(N \cap P^c)$

c) Calcula la probabilidad de que un estudiante tenga Prime Video pero no Netflix: $P(P \cap N^c)$

d) Calcula la probabilidad de que un estudiante no tenga ninguna de las dos plataformas: $P((N \cup P)^c)$

e) Construye un diagrama de Venn y coloca en cada región la probabilidad correspondiente.

f) Verifica que la suma de las probabilidades de las cuatro regiones del diagrama de Venn sea igual a 1.

3.4.2 En una encuesta realizada a usuarios del sistema de transporte de una ciudad se obtuvo la siguiente información:

- La probabilidad de que un usuario utilice bicicleta es: $P(B) = 0.25$
- La probabilidad de que un usuario utilice transporte público es: $P(T) = 0.68$
- La probabilidad de que un usuario utilice ambos medios de transporte es: $P(B \cap T) = 0.12$

Con base en esta información:

a) Calcula la probabilidad de que un usuario utilice únicamente bicicleta: $P(B \cap T^c)$

b) Calcula la probabilidad de que un usuario utilice únicamente transporte público: $P(T \cap B^c)$

c) Calcula la probabilidad de que un usuario utilice al menos uno de los dos medios de transporte: $P(B \cup T)$

d) Verifica el axioma de Aditividad: $P(B \cup T) = P(B) + P(T) - P(B \cap T)$

e) Construye un diagrama de Venn y coloca la probabilidad correspondiente en cada una de sus regiones.

3.4.3 En un estudio sobre ciberseguridad realizado a empresas de distintos sectores se obtuvo la siguiente información:

- La probabilidad de que una empresa haya sufrido un ataque de phishing durante el último año es: $P(P) = 0.45$
- La probabilidad de que una empresa haya sufrido un ataque de malware durante el último año es: $P(M) = 0.30$
- La probabilidad de que una empresa no haya sufrido ninguno de estos dos tipos de ataques es: $P((P \cup M)^c) = 0.40$

Con base en esta información:

a) Calcula la probabilidad de que una empresa haya sufrido al menos uno de los dos tipos de ataques. $P(P \cup M)$

b) Calcula la probabilidad de que una empresa haya sufrido ambos tipos de ataques $P(P \cap M)$

c) Calcula la probabilidad de que una empresa haya sufrido únicamente phishing: $P(P \cap M^c)$

d) Calcula la probabilidad de que una empresa haya sufrido únicamente malware: $P(M \cap P^c)$

e) Verifica que la suma de las probabilidades correspondientes a:

- solo phishing
- solo malware
- ambos ataques
- ninguno de los ataques

sea igual a 1

Utilizando los resultados obtenidos, verifica las siguientes propiedades:

f) Axioma de no negatividad. Comprueba que todas las probabilidades calculadas son mayores o iguales a cero.

$$P(A) \geq 0, \text{ para cualquier evento } A$$

g) Axioma de normalización. Verifica que la probabilidad del espacio muestral es igual a 1.

$$P(\Omega) = 1$$

h) Regla del complemento. Comprueba que: $P(P \cup M) + P((P \cup M)^c) = 1$

i) Regla de adición para eventos no mutuamente excluyentes. Verifica que:

$$P(P \cup M) = P(P) + P(M) - P(P \cap M)$$

j) Determina si los eventos $P$ y $M$ son mutuamente excluyentes. Justifica tu respuesta.

k) Construye un diagrama de Venn y coloca la probabilidad correspondiente en cada región.

## 3.5 Probabilidad conjunta de eventos. Probabilidad marginal, condicional e independencia. Regla del producto

3.5.1 Una empresa de entrega de alimentos a domicilio (delivery) desea analizar la relación entre el uso de cupones promocionales y la frecuencia de compra de sus clientes.
A partir de una muestra representativa de usuarios, se estimaron las siguientes probabilidades conjuntas:

| | Pide 3 o más veces por semana | Pide menos de 3 veces por semana |
|---|---|---|
| Usa cupones | 0.18 | 0.22 |
| No usa cupones | 0.12 | 0.48 |

Sea:

- $C$: el cliente utiliza cupones promocionales.
- $F$: el cliente realiza 3 o más pedidos por semana.

Con base en la información proporcionada:

a) Calcula la probabilidad marginal de que un cliente utilice cupones P(C).

b) Calcula la probabilidad marginal de que un cliente realice 3 o más pedidos por semana P(F).

c) Calcula la probabilidad de que un cliente utilice cupones dado que realiza 3 o más pedidos por semana P(C|F).

d) Calcula la probabilidad de que un cliente realice 3 o más pedidos por semana dado que utiliza cupones P(F|C).

e) Determina si los eventos $C$ y $F$ son independientes P(C∩F)=P(C)P(F).

f) Interpreta el resultado obtenido en el inciso anterior en el contexto del problema.

g) Con base en los resultados obtenidos, ¿el uso de cupones parece estar asociado con una mayor frecuencia de pedidos? Justifica tu respuesta utilizando las probabilidades condicionales calculadas P(F|C) y P(F|$C^c$).

3.5.2 En una elección para presidente de la sociedad de alumnos, se preguntó si los estudiantes apoyan a la candidata X y si además están satisfechos con el desempeño actual del Consejo.

- $P(\text{apoya X}) = 0.54$
- $P(\text{satisfecho}) = 0.40$
- $P(\text{apoya X} \cap \text{satisfecho}) = 0.30$

a) Calcula la probabilidad de que un estudiante no apoye a X y no esté satisfecho.

b) ¿Cuál es la probabilidad de que apoye a X o esté satisfecho?

c) Obtén $P(\text{satisfecho} \mid \text{apoya X})$.

d) Calcula $P(\text{apoya X} \mid \text{satisfecho})$.

e) ¿Son independientes “apoyar a X” y “estar satisfecho?

3.5.3 Una plataforma quiere analizar si quienes ven series también ven documentales:

| | Ve documentales | No ve documentales |
|---|---|---|
| Ve series | 0.42 | 0.28 |
| No ve series | 0.10 | 0.20 |

a) Obtén la probabilidad marginal de ver series.

b) Obtén la probabilidad marginal de ver documentales.

c) Calcula $P(\text{ve documentales} \mid \text{ve series})$.

d) Usa la regla del producto para reconstruir $P(\text{ve series} \cap \text{ve documentales})$.

e) ¿Ver series y ver documentales son eventos independientes? Justifica tu respuesta.

3.5.4 Una encuesta registra:

| Sueño / Estrés | Bajo | Medio | Alto | Total |
|---|---|---|---|---|
| < 6 h | 0.05 | 0.06 | 0.07 | 0.18 |
| 6–8 h | 0.12 | 0.20 | 0.11 | 0.43 |
| > 8 h | 0.08 | 0.14 | 0.17 | 0.39 |
| Total | 0.25 | 0.40 | 0.35 | 1.00 |

a) Calcula la probabilidad de que una persona duerma > 8 h y tenga alto estrés.

b) Calcula la probabilidad marginal de tener alto estrés.

c) Calcula $P(\text{alto estrés} \mid < 6 \text{ h})$.

d) Calcula $P(>8 \text{ h} \mid \text{medio estrés})$.

e) ¿Dormir más horas reduce la probabilidad condicional de alto estrés? Interpreta.

3.5.5 Una empresa clasifica a los empleados según su nivel de satisfacción en el curso de capacitación (Alta / Media / Baja), modalidad (Remoto / Híbrido / Presencial) y edad (≤ 30 / > 30):

| Edad / Modalidad | Alta | Media | Baja | Total |
|---|---|---|---|---|
| ≤ 30 / Remoto | 0.06 | 0.04 | 0.03 | 0.13 |
| ≤ 30 / Híbrido | 0.07 | 0.05 | 0.03 | 0.15 |
| ≤ 30 / Presencial | 0.05 | 0.09 | 0.08 | 0.22 |
| > 30 / Remoto | 0.08 | 0.06 | 0.03 | 0.17 |
| > 30 / Híbrido | 0.12 | 0.07 | 0.05 | 0.24 |
| > 30 / Presencial | 0.05 | 0.03 | 0.01 | 0.09 |
| Total | 0.43 | 0.34 | 0.23 | 1.00 |

a) Calcula la probabilidad de un empleado mayor de 30 años con satisfacción alta.

b) Calcula la probabilidad marginal de modalidad remota.

c) Calcula $P(\text{satisfacción baja} \mid \text{presencial})$.

d) Calcula $P(\text{remoto} \mid \text{alta satisfacción})$.

e) ¿La satisfacción laboral y la modalidad de trabajo parecen independientes?

3.5.6 El INE publica semanalmente, en su página de Internet[1], los datos por rangos de edad, entidad de origen y sexo del Padrón Electoral y Lista Nominal. A partir de una muestra de esta información se obtiene la siguiente base de datos "Lista_Nominal_INE":

Lista_Nominal_INE.xlsx

Para realizar el análisis, clasifica la variable **Edad** en los siguientes grupos mutuamente excluyentes:

$$18,\ 19-24,\ 25-29,\ 30-34,\ 35-39,\ 40-44,$$
$$45-49,\ 50-54,\ 55-59,\ 60-64,\ 65 \text{ y mas.}$$

a) Construye una tabla de frecuencias conjuntas para las variables $X$ =sexo y $Y$ =grupo de edad. A partir de ella, construye la correspondiente tabla de probabilidades conjuntas.

b) Si seleccionamos a una persona al azar de la Lista Nominal, ¿cuál es la probabilidad de que sea mujer?

c) ¿Cuál es la probabilidad de que una persona elegida al azar pertenezca al grupo de edad de 18 años?

d) ¿Cuál es la probabilidad de seleccionar a una persona que sea hombre y tenga entre 25 y 29 años?

e) ¿Cuál es la probabilidad de que la persona elegida sea mujer y tenga 65 años o más?

f) Si sabemos que una persona pertenece al grupo de 30–34 años, ¿cuál es la probabilidad de que sea mujer?

g) Si sabemos que una persona es hombre, ¿cuál es la probabilidad de que esté en el grupo de 40–44 años?

[1] https://ine.mx/datos-rangos-edad-entidad-origen-y-sexo-del-padron-electoral-y-lista-nominal-2025/

h) Verifica si los eventos A = “ser mujer” y B = “pertenecer al grupo de 18 años” son independientes.

i) Verifica si los eventos A = “ser hombre” y B = “pertenecer al grupo 65 y más” son independientes.

j) Comprueba la regla del producto para P(mujer ∩ 25–29): P(mujer ∩ 25–29) = P(mujer)*P(25–29 | mujer).

k) Si un partido quisiera dirigir su campaña a mujeres de entre 19 y 24 años, ¿qué proporción representan en la Lista Nominal total?

## 3.6 Partición del espacio muestral. Regla de probabilidad total y Teorema de Bayes

3.6.1 Las empresas de tecnología están implementando programas de diversidad e inclusión. Se realizó una encuesta laboral que reveló datos sobre la distribución de género y posiciones de liderazgo:

- 40% de los empleados son mujeres
- 60% de los empleados son hombres
- Entre las mujeres, 30% ocupa puestos de liderazgo
- Entre los hombres, 20% ocupa puestos de liderazgo

a) ¿Cuál es la probabilidad de que un empleado ocupe un puesto de liderazgo?

b) ¿Cuál es la probabilidad de que un líder sea mujer?

c) Si la dirección de la empresa de tecnología lee el resultado del inciso b) y afirma: *'Dado que la probabilidad de que un líder sea mujer es alta (o baja, según el cálculo), ya hemos alcanzado la equidad de género en el liderazgo'*, evalúe críticamente esta afirmación utilizando los conceptos de probabilidad marginal, condicional y el tamaño de los grupos de origen."

d) Calcula la probabilidad de que un empleado no ocupe un puesto de liderazgo.

e) Dado que un empleado no ocupa un puesto de liderazgo, ¿cuál es la probabilidad de que sea hombre? Interpreta el resultado dentro del contexto del problema.

f) Compara la efectividad del liderazgo por género y determina si las mujeres o los hombres tienen mayor proporción de liderazgo relativo a su grupo.

g) Calcula la probabilidad de que un empleado no sea mujer y no esté en liderazgo.

3.6.2 Actualmente los países enfrentan una creciente presión las de calor, inundaciones y sequías extremas. Los gobiernos implementan diferentes estrategias y un factor clave es la adopción de energías renovables (solar, eólica, hidroeléctrica y geotérmica).

Según reportes recientes de la Agencia Internacional de Energía:

- 40% de los países han logrado alta adopción de energías renovables, cubriendo más del 50% de su consumo energético.
- 60% de los países aún tienen baja adopción, con menos del 50% de su energía proveniente de fuentes renovables.
- Entre los países con alta adopción, 70% logran reducir sus emisiones de $CO_2$ respecto al año anterior.
- Entre los países con baja adopción, solo 30% logran reducir emisiones.

a) Calcula la probabilidad de que un país seleccionado al azar logre reducir sus emisiones de $CO_2$ en 2025.

b) Si se sabe que un país logró reducir emisiones de $CO_2$, ¿cuál es la probabilidad de que tenga alta adopción de energías renovables?

c) Si un país no logró reducir emisiones, ¿cuál es la probabilidad de que tenga baja adopción de energías renovables?

d) Si un país tiene baja adopción de energías renovables, ¿cuál es la probabilidad de que no reduzca sus emisiones?

e) Si un país no redujo emisiones, ¿cuál es la probabilidad de que tenga alta adopción de energías renovables?

f) Un diplomático de un país con baja adopción de energías renovables argumenta que el 30% de los países de su bloque logran reducir emisiones sin invertir en renovables, por lo que la transición energética no es indispensable. Utiliza los resultados anteriores para refutar o apoyar estadísticamente este argumento desde una perspectiva global.

3.6.3 México continúa enfrentando desafíos en salud pública relacionados con enfermedades respiratorias estacionales, como la influenza. El gobierno ha implementado programas de vacunación masiva para reducir la morbilidad y mortalidad asociadas a la influenza. Los epidemiólogos están interesados en evaluar la efectividad de los programas de vacunación.
Según los datos más recientes:

- 60% de la población está completamente vacunada contra la influenza.
- Entre las personas vacunadas, la probabilidad de enfermar de influenza es de 5%, gracias a la protección que ofrece la vacuna.

- Entre las personas no vacunadas, la probabilidad de enfermar es mucho mayor: 20%, reflejando el riesgo de no recibir la vacuna.

a) Calcula la probabilidad de que un mexicano seleccionado al azar enferme de influenza.

b) Interpreta el resultado en términos de riesgo poblacional y efectividad de la vacunación.

c) Si un mexicano enferma de influenza, ¿cuál es la probabilidad de que no estuviera vacunado?

d) Si la cobertura de vacunación aumenta al 80%, ¿cómo cambia la probabilidad total de enfermar?

e) Manteniendo la cobertura original de vacunación del 60%, si surge una cepa de influenza más virulenta que aumenta la probabilidad de enfermar en vacunados a 10% y en no vacunados a 30%, ¿cómo cambia la probabilidad total?

f) Deseamos programar una función en R que calcule automáticamente la probabilidad total de enfermar para cualquier cobertura de vacunación $p$, probabilidad de enfermar entre vacunados $q_V$y probabilidad de enfermar entre no vacunados $q_N$. Expresa algebraicamente $P(E)$ como función de estos parámetros.y explica qué tipo de gráfico (de los vistos en el Tema 2) utilizarías para mostrar el impacto de la vacunación a las autoridades de salud.

3.6.4 Una app bancaria clasifica automáticamente las transacciones en tres categorías que forman una partición del espacio muestral:

- C1: Transacciones habituales (80%)
- C2: Transacciones inusuales (15%)
- C3: Transacciones de alto riesgo (5%)

El sistema de inteligencia artificial del banco analiza las transacciones y las marcas como "SOSPECHOSA" (S) bajo los siguientes parámetros de efectividad (probabilidades condicionales):

- La probabilidad de que una transacción habitual sea marcada como sospechosa es del 2% (error del algoritmo).
- La probabilidad de que una transacción inusual sea marcada como sospechosa es del 40%.
- La probabilidad de que una transacción de alto riesgo sea detectada y marcada como sospechosa es del 90%.

a) Calcula la probabilidad de que cualquier transacción al azar sea marcada como “SOSPECHOSA” por la app.

b) Si el departamento de prevención de fraudes recibe una alerta de transacción "SOSPECHOSA", ¿cuál es la probabilidad de que realmente provenga de la categoría de alto riesgo (C3)?

c) En el contexto bancario, un "Falso Positivo" ocurre cuando el sistema bloquea una transacción habitual (C1) asumiendo que es sospechosa, generando fricción con el cliente. Calcule la probabilidad de que una transacción marcada como sospechosa sea en realidad un falso positivo (P(C1 | S)). A la luz del resultado, ¿dirías que el algoritmo es eficiente para la operación diaria del banco? Justifica tu respuesta.

3.6.5 Durante una campaña electoral, una consultora política analiza la propagación de una noticia falsa (*Fake News*) de alto impacto en una red social. Los analistas políticos y de datos determinan que los usuarios de la plataforma en el país pueden dividirse en una partición de tres subpoblaciones mutuamente excluyentes según su comportamiento de consumo de medios:

- A1 (Consumidores Críticos): Usuarios que verifican activamente la información en múltiples fuentes independientes. Representan el 30% de la población activa en la red.
- A2 (Consumidores Pasivos): Usuarios que consumen información sin verificarla, pero no suelen interactuar intensamente con contenido político radical. Representan el 50% de la población.
- A3 (Consumidores Polarizados): Usuarios integrados en "cámaras de eco" que interactúan predominantemente con contenido alineado a su ideología radical. Representan el 20% de la población.

El algoritmo de recomendación de la plataforma etiqueta la publicación como "Sugerida para Compartir" (S) basándose en el nivel de interacción inicial de cada grupo. Los datos históricos de la consultora muestran las siguientes probabilidades condicionales de que el algoritmo potencie la difusión de la noticia falsa dentro de cada grupo:

- Para un usuario del grupo de *Consumidores Críticos*, la probabilidad de que el algoritmo logre que comparta la noticia es de apenas 0.05 (un 5%).
- Para un usuario del grupo de *Consumidores Pasivos*, la probabilidad de que la comparta es de 0.40 (un 40%).
- Para un usuario del grupo de *Consumidores Polarizados*, la probabilidad de que la comparta es de 0.85 (un 85%).

a) Calcula la probabilidad de que un usuario de esta red social, seleccionado completamente al azar, termine compartiendo la noticia falsa.

b) Si se detecta que un usuario específico ya compartió la noticia falsa en su perfil, ¿cuál es la probabilidad de que este usuario pertenezca al grupo de *Consumidores Polarizados* (A3)?

c) Compara la probabilidad marginal del grupo de *Consumidores Polarizados* (P(A3)) con la probabilidad condicional calculada en el inciso anterior (P(A3|S)). Desde la perspectiva de la Ciencia Política, ¿cómo explicas este cambio de probabilidad el fenómeno de la "radicalización digital" y la creación de burbujas informativas? Justifica tu respuesta utilizando conceptos estadísticos.

d) Simulación de Políticas Públicas (Análisis de Sensibilidad): Suponga que el Instituto Nacional Electoral (INE) implementa una campaña masiva de alfabetización digital que logra que la mitad de los *Consumidores Pasivos* (A2) migren al grupo de *Consumidores Críticos* (A1), modificando las proporciones de la partición a: A1 = 55%, A2 = 25%, A3 = 20%.

   i. Calcula la nueva probabilidad total de que un usuario comparta la noticia falsa.

   ii. A la luz de este resultado, ¿es esta política pública efectiva para frenar la desinformación si el grupo polarizado (A3) permanece intacto? Argumenta tu postura.

3.6.6 El Servicio de Administración Tributaria (SAT) ha implementado un nuevo modelo de Inteligencia Artificial basado en análisis de riesgos para optimizar sus procesos de fiscalización. El universo de contribuyentes (personas morales) de una región se puede estructurar en una partición de tres grupos mutuamente excluyentes según su comportamiento fiscal real:

- C1 (Empresas Formales Cumplidas): Contribuyentes que registran y pagan sus impuestos correctamente. Representan el 75% de las empresas.
- C2 (Empresas con Discrepancias Menores): Contribuyentes que cometen errores u omisiones involuntarias en sus declaraciones sin intención de fraude. Representan el 20% de las empresas.
- C3 (Empresas Fantasma / Evasoras): Organizaciones creadas con el único fin de simular operaciones y emitir facturas falsas para evadir impuestos. Representan el 5% de las empresas.

El algoritmo de IA del fisco analiza los Comprobantes Fiscales Digitales (CFDI) y las declaraciones en tiempo real. Cuando detecta patrones de riesgo, emite una "Alerta de Auditoría Profunda" (A). Los datos de calibración del sistema muestran las siguientes probabilidades condicionales de activar la alerta:

- Para una Empresa Formal Cumplida, la probabilidad de activar la alerta por un falso positivo del sistema es de apenas 0.02 (un 2%).
- Para una empresa con Discrepancias Menores, la probabilidad de activar la alerta es de 0.30 (30%).
- Para una Empresa Fantasma / Evasora, la probabilidad de que el sistema la detecte con éxito y active la alerta es de 0.90 (90%).

a) Calcula la probabilidad de que una empresa elegida al azar de la base de datos general sea seleccionada por el algoritmo para una "Alerta de Auditoría Profunda".

b) Si el SAT notifica una alerta de auditoría a una empresa, ¿cuál es la probabilidad de que se trate de una *Empresa Fantasma / Evasora* (C3)?

c) Si el algoritmo emite una alerta, ¿cuál es la probabilidad de que la empresa auditada resulte ser una *Empresa Formal Cumplida* (C1)?

a. Desde la perspectiva de la gestión de recursos del SAT y del impacto económico en las empresas legítimas, evalúa si este porcentaje de falsos positivos es aceptable o si el algoritmo requiere recalibración. Justifica tu postura utilizando los resultados obtenidos.

d) La autoridad fiscal decide endurecer los filtros del algoritmo para atrapar a más evasores, logrando que la probabilidad de detectar a las *Empresas Fantasma* (C3) suba del 90% al 98%. Sin embargo, esto provoca que la tasa de falsos positivos en las *Empresas Cumplidas* (C1) también aumente del 2% al 6%.

i. Calcula la nueva probabilidad de que una alerta corresponda realmente a una empresa fantasma (P(C3 | A)).

ii. A la luz del resultado anterior, ¿valió la pena endurecer el algoritmo? Explica el fenómeno estadístico de por qué la probabilidad de éxito de la auditoría varió de esa manera.

e) Un despacho contable afirma: *"Como el sistema tiene un 90% de efectividad para detectar empresas fantasma, si a mi cliente le llega una auditoría, es un "hecho casi seguro" (90%) que el SAT tiene pruebas de que es una empresa evasora"*. Con base en los conceptos de probabilidad condicional y los cálculos de los incisos anteriores, explica detalladamente el error conceptual en la afirmación del despacho.

# TEMA 4: Variables aleatorias y distribuciones de probabilidad

## 4.1 Concepto de variable aleatoria y soporte. Variables aleatorias discretas y continuas

4.1.1 Si se selecciona al azar una unidad (empresa/persona/país...) de la población correspondiente, indica cuáles de las siguientes características pueden definirse como variables aleatorias numéricas, Para las variables cualitativas, explica qué transformación sería necesaria.

a) Número de empleados en una empresa

b) Ciudad de nacimiento de una persona

c) Fecha de fundación de una empresa

d) Salario mensual de un empleado

e) Cantidad de productos vendidos por mes

f) Nombre del presidente de un país

g) Porcentaje de participación electoral en un distrito

h) Número de páginas de un libro

i) Número de reuniones realizadas en un mes

j) El PIB de un país

k) La cantidad de oficinas que tiene una empresa en un año determinado

l) El tipo de cambio dólar/peso el primer día del año

m) Número de protestas registradas en un año

n) Número de productos que fabrica una empresa

o) Edad de los legisladores

p) Número de clientes atendidos por un empleado

q) La cantidad de conferencias internacionales celebradas en un año

r) La política oficial de un gobierno sobre impuestos

s) Tiempo promedio de respuesta en atención al cliente

t) El día de la semana

u) Número de productos defectuosos en producción

v) El eslogan de una campaña publicitaria

w) Cantidad de contratos firmados en un trimestre

x) El nombre de los candidatos electorales para la próxima elección

y) Nivel de satisfacción de empleados en encuesta (0–10)

z) Número de seguidores de una empresa en redes sociales

4.1.2 Determina si las variables aleatorias son cualitativas, cuantitativas discretas o cuantitativas continuas:

a) Número de votos obtenidos por un candidato en una elección

b) Porcentaje de participación electoral por distrito

c) Número de partidos políticos registrados en un país

d) Cantidad de propuestas de ley aprobadas en un periodo legislativo

e) Índice de corrupción percibido (0–100)

f) Número de manifestaciones públicas en un año

g) Edad de los legisladores (años)

h) Número de países que firmaron un tratado internacional

i) Cantidad de escaños ocupados por un partido en el parlamento

j) Porcentaje de aprobación presidencial

k) Número de embajadas de un país en el extranjero

l) Cantidad de tratados comerciales firmados en un año

m) Volumen de importaciones de un país (millones de USD)

n) Número de conflictos armados activos en una región

o) Número de visitas oficiales entre jefes de estado

p) Porcentaje de población con pasaporte válido

q) Número de sanciones económicas impuestas a un país

r) Gasto militar como porcentaje del PIB

s) Número de refugiados recibidos por un país en un año

t) Tiempo promedio de respuesta a crisis internacionales (días)

u) Número de empleados en un departamento

v) Horas trabajadas por empleado a la semana

w) Número de reuniones realizadas en un mes

x) Porcentaje de proyectos completados a tiempo

y) Presupuesto asignado a un departamento (en miles de USD)

z) Número de quejas atendidas por servicio al cliente

aa) Número de horas de capacitación por empleado al año

bb) Porcentaje de satisfacción laboral medido en encuesta

cc) Cantidad de días de ausencia por enfermedad

dd) Tiempo promedio de resolución de problemas administrativos (horas)

## 4.2 Distribuciones de probabilidad discretas. Propiedades. Esperanza y varianza con sus propiedades

4.2.1 Una publicación permite reaccionar con 3 emojis: 👍, 😂 y 😮. Un usuario puede dejar exactamente 2 reacciones, pudiendo repetir emoji.

Sea la variable aleatoria $X =$ número de emojis distintos usados en las dos reacciones.

a) Enumera todos los posibles pares de reacciones.

b) ¿Cuántos arreglos tienen 1 solo emoji?

c) ¿Cuántos arreglos tienen 2 emojis distintos?

d) Construye la distribución de X.

e) Calcula el valor esperado de X.

f) Calcula la desviación estándar de X.

4.2.2 Una rifa asigna a cada participante un código QR formado por 3 dígitos del 0 al 9. Los dígitos se pueden repetir. Se define la variable aleatoria X: número de dígitos pares en el código QR.

a) Calcula el total de códigos posibles.

b) ¿Cuántos códigos tienen 0, 1, 2 y 3 dígitos pares?

c) Construye la distribución de probabilidad de $X$.

d) Calcula E(X).

e) Calcula Var(X).

4.2.3 En un videojuego, cada jugador crea su avatar eligiendo:

- 1 peinado entre 3 opciones (Peinado 1, Peinado 2, Peinado 3).

- 1 accesorio entre 4 opciones (Accesorio 1, Accesorio 2, Accesorio 3, Accesorio 4).

Supongamos que cada combinación de peinado y accesorio es igualmente probable.

Del análisis de los 100 avatares más populares, se identificó que las opciones favoritas son:

- Peinado popular: Peinado 2
- Accesorio popular: Accesorio 3

Definimos la variable aleatoria:

$X =$ número de coincidencias entre las elecciones del jugador y las opciones populares.

En donde;

- $X = 0$: No coincide en nada.
- $X = 1$: Coincide en peinado o en accesorio.
- $X = 2$: Coincide en ambas elecciones.

a) Enumera todas las combinaciones posibles de peinado y accesorio.

b) Para cada combinación, determina el valor de $X$.

c) Obtén la distribución de probabilidad de X.

d) ¿Cuál es la probabilidad de que un jugador elija al menos una opción popular?

e) Interpreta la moda de la distribución: qué indica sobre el comportamiento típico de los jugadores.

4.2.4 Una empresa tiene 6 proyectos de tecnología, 5 de finanzas y 3 de marketing. Se eligen al azar 4 proyectos para llevar al Consejo para su aprobación.

Sea X = número de proyectos de tecnología seleccionados.

a) Encuentre la distribución de probabilidad de X.

b) ¿Cuál es la probabilidad de que se seleccione por lo menos 1 proyecto de tecnología?

c) Si sabemos que se eligió al menos un proyecto de tecnología, ¿cuál es la probabilidad de que se hayan elegido exactamente 2?

d) Si la empresa quiere que las tres áreas —tecnología, finanzas y marketing— estén representadas entre los cuatro proyectos seleccionados, ¿cuál es la probabilidad de que no se cumpla esta condición?

e) Calcule E[X] e interpreta dentro del contexto del problema.

f) Calcula Var(X) e interpreta dentro del contexto del problema. (ej. más varianza = más incertidumbre en la representación política o en la asignación de becas).

4.2.5 En una encuesta a 20 ciudadanos, 13 confían en las instituciones democráticas y 7 no confían. Se seleccionan 5 personas al azar para una entrevista más detallada. Sea Z = número de personas que confían seleccionadas.

a) Obtén la distribución de probabilidad de Z.

b) ¿Cuál es la probabilidad de que no se seleccionen ciudadanos que confíen en las instituciones democráticas?

c) Si sabemos que se eligió al menos una persona que confía en las instituciones democráticas, ¿cuál es la probabilidad de que se hayan elegido exactamente 3?

d) Calcula E[Z] e interpreta dentro del contexto del problema.

e) Calcula Var(Z) e interpreta dentro del contexto del problema.

## 4.3 Distribuciones de probabilidad continuas. Propiedades. Esperanza y varianza con sus propiedades

4.3.1 Una empresa de análisis digital estudia el comportamiento de los usuarios al ver un video en una plataforma de contenido en línea.

La variable aleatoria $X$ representa la proporción del video que un usuario ve antes de abandonarlo, donde:

- $X = 0$: el usuario no ve el video.
- $X = 1$: el usuario ve el 100% del video.

Para modelar este comportamiento, los analistas proponen la siguiente función de densidad de probabilidad:

$$f(x) = \begin{cases} 2 - 2x & 0 < x < 1 \\ 0 & \text{en otro caso} \end{cases}$$

a) Grafica la función de densidad.

b) Verifica que $f(x)$ cumple las condiciones para ser una función de densidad de probabilidad.

c) Calcula las siguientes probabilidades:

i. $P(X < 0.2)$

ii. $P(0.3 < X < 0.5)$
iii. $P(X > 0.8)$

d) Determina el valor $k$ tal que $P(X < k) = 0.5$

e) Encuentra el valor $k$ tal que: $P(X > k) = 0.10$

f) Obtén la función de distribución acumulada $F(x)$.

g) Utiliza la función acumulada para calcular nuevamente $P(0.3 < X < 0.5)$

h) Calcula la esperanza matemática $E(X)$. Interprete el valor esperado en el contexto del tiempo de permanencia de los usuarios.

4.3.2 Un consultor externo está evaluando la eficiencia de la nueva plataforma digital del Servicio de Administración Tributaria (SAT) El tiempo $X$ (medido en horas) que le toma a un usuario completar con éxito un trámite complejo de declaración o registro es una variable aleatoria continua.

Debido a que el sistema cierra la sesión automáticamente al cabo de una hora para evitar saturación, la duración del proceso tiene la siguiente función de densidad de probabilidad:

$$f(x) = \begin{cases} x + cx^2 & 0 < x < 1 \\ 0 & \text{en otro caso} \end{cases}$$

Con base en esta información, resuelve los siguientes puntos para presentar tu reporte ejecutivo:

a) Determina el valor de la constante $c$ para que $f(x)$ cumpla con las condiciones de una función de densidad de probabilidad legítima.

b) Obtén la función de distribución acumulada para el tiempo de respuesta del trámite $F(x)$.

c) Si la meta de la institución es que los procesos sean ágiles, calcula la probabilidad de que un usuario logre terminar todo el trámite en menos de media hora.

d) Si un auditor detecta que un usuario ya lleva al menos 15 minutos intentando completar el trámite en la plataforma, encuentra la probabilidad de que este usuario requiera de al menos 30 minutos en total para finalizarlo.

e) Determina la probabilidad de que un usuario tarde en completar el trámite entre el 30% y el 60% del tiempo máximo permitido (es decir, entre 0.3 y 0.6 horas).

f) Encuentra el valor $k$ tal que $P(X < k) = 0.50$. Interpreta el resultado en términos del tiempo que le toma a la mitad de la población realizar el trámite.

g) Determina el valor $m$ tal que $P(X > k) = 0.20$. Interpreta el significado de este valor dentro del contexto del problema. ¿Qué representa este grupo de usuarios para el diseño de la plataforma?

h) Si un trámite presencial equivalente en ventanilla tarda exactamente 20 minutos (0.333 horas), calcula la probabilidad de que un usuario en la plataforma digital permanezca en el sistema al menos 15 minutos (0.25 horas).

i) Sabiendo que un usuario ya ha invertido 0.25 horas del tiempo máximo en la plataforma sin terminar, ¿es más probable que logre concluir el trámite antes o después de alcanzar el 0.75 horas? Justifica mediante el cálculo de probabilidades condicionales.

j) Calcula el valor esperado $E(X)$ e interprétalo dentro del contexto del problema.

k) Calcula la varianza $V(X)$ y la desviación estándar $\sigma_X$ del tiempo de finalización.

l) Con base en los resultados obtenidos en los incisos (f) y (g), ¿los usuarios de la plataforma tienden a resolver el trámite de manera ágil en los primeros minutos o la gran mayoría se concentra hacia el final del límite de tiempo?

m) La normatividad gubernamental considera que la plataforma digital es "altamente eficiente" si al menos el 60% de los usuarios logra completar su trámite en menos del 70% del tiempo límite (es decir, en menos de 0.7 horas). Utilizando el modelo propuesto, evalúa si este sistema cumple con dicho criterio de calidad y recomienda si debiera mantenerse, modificarse o reemplazarse.

4.3.3 En el marco del Tratado entre México, Estados Unidos y Canadá (T-MEC), la velocidad de despacho en las aduanas del norte del país es un factor crítico de competitividad. Sea $Y$ la variable aleatoria continua que mide el tiempo (en días o fracciones de día) que tarda un cargamento de componentes tecnológicos de exportación en cruzar la aduana de Nuevo Laredo y ser liberado para su distribución.

Debido a la automatización de los procesos de revisión, el tiempo de liberación nunca excede de 1 día (24 horas), y se modela mediante la siguiente función de densidad de probabilidad:

$$f(y) = \begin{cases} ky^2 & para\ 0 < y < 1 \\ 0 & en\ otro\ caso \end{cases}$$

Con base en esta situación, resuelve los siguientes puntos para la auditoría operativa de la empresa:

a) Determina el valor de la constante $k$ de tal manera que la función $f(y)$ sea una función de densidad de probabilidad válida para el análisis logístico.

b) Realiza la gráfica de la función de densidad $f(y)$ utilizando el valor de $k$ obtenido. Explica visualmente si los cargamentos tienden a salir rápido o si la mayoría experimenta tiempos de espera cercanos al límite del día.

c) Obtén la función de distribución acumulada $F(y)$ del tiempo de liberación aduanal y realiza su gráfica correspondiente.

d) ¿Cuál es la probabilidad de que la liberación aduanal de un cargamento tome más de tres cuartos de día (es decir, más de 18 horas o $y > 0.75$)?

e) Obtén el primer cuartil ($Q_1$) del tiempo de despacho. Interpreta este resultado en términos de competitividad y tiempos mínimos esperados por los clientes más corporativos.

f) ¿Cuántos días (o qué fracción de día) se espera que tarde, en promedio, la liberación de un cargamento? Calcula también la varianza del proceso para medir el riesgo de desabasto en la cadena de suministro.

4.3.4 Con la apertura de una nueva ruta aérea directa orientada al turismo de lujo entre el Aeropuerto de París-Charles de Gaulle y el Aeropuerto Internacional de Oaxaca, el comité de administración de la aerolínea analiza la puntualidad de los vuelos. Sea $T$ la variable aleatoria que mide el tiempo de desviación (en horas) respecto a la hora programada de aterrizaje.

- Un valor negativo ($-1 < t < 0$) indica que el vuelo viene demorado o retrasado
- Un valor positivo ($0 \leq t < c$) indica que el vuelo viene adelantado

La política de optimización del espacio aéreo modela esta desviación mediante la siguiente función de densidad:

$$f(t) = \begin{cases} 1+t & -1 < t < 0 \\ 1 & 0 \leq t < c \\ 0 & en\ otro\ caso \end{cases}$$

a) Con base en esta información, resuelve el siguiente análisis de operaciones y rentabilidad:

b) Encuentra el valor de la constante $c$ que hace que esta función cumpla con las condiciones de una función de densidad de probabilidad.

c) ¿Cuál es la probabilidad de que un vuelo proveniente de París llegue con retraso?

d) Calcula la probabilidad de que un avión aterrice con un adelanto de más de 15 minutos.

e) La utilidad neta de la aerolínea por cada vuelo (en miles de dólares) está directamente ligada a la puntualidad debido al uso de combustible y costos de posicionamiento en puerta de embarque. Esta relación se describe mediante la función lineal:

$$U(t) = 1 + 0.5t$$

- Calcula la utilidad esperada por vuelo, $\mathrm{E[U(t)]}$, utilizando las propiedades del valor esperado.
- Calcula la varianza de la utilidad, $\mathrm{V[U(t)]}$, para medir la volatilidad de las ganancias de esta ruta internacional.

f) Encuentra la función de distribución acumulada $F(t)$ para todo el dominio de la variable.

g) Determina el valor del percentil 50 (la mediana) del tiempo de desviación. ¿El vuelo típico tiende a estar retrasado, adelantado o exactamente a tiempo? Justifica con tu resultado.

h) Las normativas de aviación internacional exigen que si un vuelo se retrasa más de 45 minutos (es decir, $t < -0.75$), la aerolínea debe otorgar un cupón de descuento para el próximo viaje. Si en un mes operan 40 vuelos en esta ruta, ¿cuántos vuelos se espera que tengan que indemnizar a los pasajeros?

4.3.5 Una Fintech que ofrece microcréditos para jóvenes universitarios está optimizando su canal de soporte técnico a través de videollamadas. Sea $X$ la variable aleatoria continua que mide la duración (en minutos) de una sesión de asesoría especializada para resolver dudas sobre contratos digitales.

Debido a que la mayoría de los problemas se resuelven rápido, pero unos pocos requieren mucho tiempo, la duración de las llamadas sigue una distribución exponencial con la siguiente función de densidad de probabilidad:

$$f(x) = \begin{cases} 0.1e^{-0.1x} & x > 0 \\ 0 & en\ cualquier\ otro\ caso \end{cases}$$

Con base en este modelo operativo, responde las siguientes preguntas de gestión:

a) Calcula el valor esperado $E(X)$ e interprétalo dentro del contexto del problema.

b) Encuentra la función de distribución acumulada $F(x)$ para todo el dominio de la variable.

c) Si un supervisor se conecta a la plataforma aleatoriamente justo cuando inicia una sesión de asesoría, utiliza $F(x)$ para calcular la probabilidad de que esta llamada se extienda por más de 10 minutos.

d) Encuentra el valor de la mediana del tiempo de atención (percentil 50) resolviendo F(m) = 0.5. Compara este resultado con el valor esperado obtenido en el inciso a) y explica a qué se debe la diferencia.

e) Si un asesor ya lleva 15 minutos atendiendo a un cliente con un problema complejo, calcula la probabilidad condicional de que la llamada dure más de 25 minutos en total (es decir, P(X > 25 | X > 15)).

f) Calcula la varianza $\mathrm{V}(\mathrm{X})$ y la desviación estándar $\sigma_X$ de la duración de las llamadas. ¿Qué nos dice una desviación estándar tan alta respecto a la predictibilidad del trabajo de los asesores?

g) La empresa estima que el costo operativo de mantener la plataforma en la nube es de $2 dólares fijos por llamada más un costo variable de $0.5 dólares por cada minuto que dure la sesión.

- Define la función de costo total por llamada $C(X)$
- Calcula el costo esperado por llamada y la varianza del costo total

h) La política de calidad de la Fintech exige que no más del 5% de los clientes experimenten llamadas de soporte que superen los 25 minutos (para evitar saturar las líneas). Con el modelo actual, ¿se está cumpliendo con esta política de calidad?

## 4.4 Distribuciones de probabilidad bivariadas discretas / 4.5 Distribuciones marginales de probabilidades. Probabilidad condicional de variables aleatorias / 4.6 Variables aleatorias independientes / 4.7 Covarianza y correlación

4.4.1. a) Explica el significado de la magnitud y el signo del coeficiente de correlación.
b) ¿En qué caso la correlación entre dos variables es igual a cero?

4.4.2 En una encuesta a jóvenes universitarios se registró:

- $X$: número de veces que una persona vio noticias políticas falsas en redes en una semana $x_i$ = {0, 1, 2}
- $Y$: número de veces que discutió temas políticos esa semana $y_i$ = {0, 1, 2}

La distribución conjunta es:

| X\Y | 0 | 1 | 2 | Total |
|---|---|---|---|---|
| 0 | 0.10 | 0.08 | 0.02 | 0.20 |
| 1 | 0.15 | 0.20 | 0.10 | 0.45 |
| 2 | 0.05 | 0.15 | 0.15 | 0.35 |
| **Total** | 0.30 | 0.43 | 0.27 | 1.00 |

a) Obtén las distribuciones marginales de $X$ y $Y$.

b) Calcula $PY = 2 \mid X = 1$.

c) Calcula $E(X)$, $E(Y)$, $E(XY)$.

d) Obtén $Cov(X, Y)$.

e) ¿Son independientes $X$ y $Y$? Justifica.

f) ¿Qué combinación $X$, $Y$ es la más frecuente? ¿Qué perfil describe?

g) ¿Existe evidencia estadística de que la exposición a noticias falsas incrementa la discusión política?

h) Si fueras asesor del INE, ¿en qué grupo enfocarías una campaña de alfabetización mediática?

i) ¿Son independientes $X$ y $Y$? Justifica tu respuesta.

j) Interpreta qué significa independencia en el contexto del problema.

k) ¿La independencia implicaría que ver noticias falsas no afecta la discusión política? Explica tu respuesta.

4.4.3 Una empresa de servicios profesionales ha comenzado a integrar herramientas de inteligencia artificial IA —como asistentes de escritura, análisis automatizado de datos y generación de reportes— con el objetivo de mejorar la productividad de sus empleados.

La dirección desea evaluar si existe una relación entre el nivel de adopción de IA y el desempeño mensual de los trabajadores. Para ello, se observó a 40 empleados durante un mes y se registró:

- $X$: número de herramientas de IA utilizadas regularmente
- $Y$: número de proyectos completados ese mes

Los resultados se muestran a continuación:

| Empleado | X | Y | Empleado | X | Y |
|---|---|---|---|---|---|
| 1 | 0 | 1 | 21 | 1 | 3 |
| 2 | 0 | 1 | 22 | 1 | 3 |
| 3 | 0 | 2 | 23 | 1 | 3 |
| 4 | 0 | 2 | 24 | 1 | 3 |
| 5 | 0 | 1 | 25 | 2 | 3 |
| 6 | 0 | 3 | 26 | 2 | 3 |
| 7 | 0 | 1 | 27 | 2 | 3 |
| 8 | 0 | 2 | 28 | 2 | 2 |
| 9 | 0 | 1 | 29 | 2 | 2 |
| 10 | 0 | 2 | 30 | 2 | 2 |
| 11 | 1 | 1 | 31 | 2 | 1 |
| 12 | 1 | 2 | 32 | 2 | 1 |
| 13 | 1 | 2 | 33 | 2 | 1 |
| 14 | 1 | 2 | 34 | 1 | 2 |
| 15 | 1 | 1 | 35 | 1 | 2 |
| 16 | 1 | 3 | 36 | 1 | 2 |
| 17 | 1 | 2 | 37 | 0 | 3 |
| 18 | 1 | 3 | 38 | 0 | 3 |
| 19 | 1 | 2 | 39 | 0 | 1 |
| 20 | 1 | 3 | 40 | 0 | 2 |

a) Construye la tabla de frecuencias conjuntas $n_{ij}$.

b) Obtén la tabla de probabilidades conjuntas $p_{ij}$.

c) Calcula las distribuciones marginales de $X$ y de $Y$.

d) Calcula $PY \geq 2 \mid X = 2$.

e) Calcula $E(X)$, $E(Y)$ y $Cov(X,Y)$.

f) ¿Qué nivel de uso de IA está asociado con mayor productividad esperada?

g) Si cada proyecto genera $20,000, estima el ingreso esperado por empleado.

h) ¿Conviene invertir en capacitación para aumentar el uso de IA? Justifica tu respuesta.

i) ¿Son independientes $X$ y $Y$?

4.4.4 Para envíos internacionales de una empresa exportadora:

- $X$: número de días de retraso en aduana, con $x_i = \{0, 1, 2\}$
- $Y$: número de reclamos del cliente, con $y_i = \{0, 1\}$

| X\Y | 0 | 1 | Total |
|---|---|---|---|
| 0 | 0.30 | 0.05 | 0.35 |
| 1 | 0.20 | 0.15 | 0.35 |
| 2 | 0.10 | 0.20 | 0.30 |
| **Total** | 0.60 | 0.40 | 1.00 |

a) Calcula $P(X \geq 1)$.

b) Calcula $P(Y = 1 \mid X = 2)$.

c) Encuentra $E(X)$ y $E(Y)$.

d) Calcula $Cov(X, Y)$. Interpreta el signo de la covarianza.

e) ¿Qué nivel de retraso representa mayor riesgo de reclamo?

f) ¿La empresa debería invertir en reducir retrasos de 1 a 0 días o de 2 a 1 días? Justifica tu respuesta.

g) Si se reducen en 50% los retrasos de 2 días, ¿cómo cambiaría el número esperado de reclamos?

4.4.5 Una empresa de comercio electrónico que opera en México ha detectado que, aunque sus ventas han crecido de manera sostenida, el volumen de devoluciones también ha aumentado, lo cual representa un costo logístico importante.

Para analizar este fenómeno, la empresa estudió el comportamiento de compra de un grupo de clientes durante un mes y registró las siguientes variables:

- $X$: número de productos comprados por un cliente en una sola orden
- $Y$: número de productos devueltos por ese mismo cliente

Con base en la información recolectada, se obtuvo la siguiente distribución conjunta de probabilidad:

| X \ Y | 0 | 1 | 2 | Total |
|---|---|---|---|---|
| 1 | 0.15 | 0.10 | 0.05 | 0.30 |
| 2 | 0.10 | 0.20 | 0.10 | 0.40 |
| 3 | 0.05 | 0.15 | 0.10 | 0.30 |
| **Total** | 0.30 | 0.45 | 0.25 | 1.00 |

a) Obtén las distribuciones marginales de $X$ y de $Y$.

b) Calcula la probabilidad de que un cliente devuelva al menos un producto, dado que compró tres productos.

c) Calcula el valor esperado de $X$ y de $Y$.

d) Calcula la covarianza entre $X$ y $Y$ e interpreta su signo.

e) ¿Qué tipo de cliente en términos de $X$ genera, en promedio, más devoluciones?

f) ¿Conviene a la empresa incentivar compras grandes, por ejemplo, con descuentos por volumen, aun cuando esto podría aumentar la tasa de devoluciones? Justifica.

g) Si el costo promedio por devolución es de $150, estima el costo esperado de devoluciones por cliente.

4.4.6. Se eligen aleatoriamente algunas familias de un área dada. Sea $X$ el número de carros en la familia y $Y$ el número de conductores con licencia en la familia. Las distribuciones marginales de $X$ y $Y$ son como sigue:

| X | 1 | 2 |
|---|---|---|
| P(X=x) | 0.60 | 0.40 |

| Y | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| P(Y=y) | 0.10 | 0.50 | 0.30 | 0.10 |

La probabilidad condicional de que haya un carro en una familia dado que hay tres conductores con licencia es 0.40. La probabilidad condicional de que haya dos conductores con licencia en una familia dado que hay dos carros es 0.30 y la probabilidad conjunta de que en una familia haya un automóvil y cuatro conductores con licencia es 0.03.

a) Encuentra la distribución conjunta de $X$ y $Y$.

b) Obtén el coeficiente de correlación de $X$ y $Y$ e interpreta el resultado dentro del contexto del problema.

c) ¿Son $X$ y $Y$ independientes? Justifica tu respuesta.

4.4.7. Una compañía de servicios turísticos debe determinar, con 48 horas de anticipación, cuántos guías contratar para atender las solicitudes de servicio que puedan recibirse de última hora para una fecha determinada.

Con el objetivo de mejorar la planeación del personal, el nuevo gerente analiza una muestra de los registros correspondientes a los últimos 10 años y obtiene la distribución conjunta de las variables aleatorias $X$ y $Y$, donde:

- $X$: número de solicitudes de última hora recibidas después de haber definido la cantidad de guías contratados para una fecha determinada
- $Y$: número de guías contratados por la empresa para atender los servicios de esa fecha

Supón que las frecuencias relativas observadas pueden utilizarse como probabilidades y que cada solicitud requiere exactamente un guía.

| Y \X | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **1** | 0.04 | 0.20 | 0.04 | 0.01 | 0.01 | 0.00 |
| **2** | 0.03 | 0.04 | 0.30 | 0.01 | 0.01 | 0.01 |
| **3** | 0.00 | 0.01 | 0.02 | 0.15 | 0.02 | 0.00 |
| **4** | 0.00 | 0.00 | 0.01 | 0.01 | 0.06 | 0.02 |

a) Obtén la distribución de probabilidades de la variable aleatoria $Z = X - Y$, es decir, del número de solicitudes no atendidas por falta de personal disponible.

a) Mediante la distribución de $Z$ comprueba que $E(Z) = E(X) - E(Y)$.

b) Calcula la covarianza entre $X$ y $Y$.

c) Calcula el coeficiente de correlación entre $X$ y $Y$. Interpreta su valor en relación con la efectividad del manejo de la empresa en los años en que se tomó la muestra.

d) ¿Son $X$ y $Y$ independientes? Justifica tu respuesta.

e) Obtén la varianza de $Z$.

4.4.8 Un cliente de una tienda compra una bebida ligera por semana, la cual puede ser embotellada o en lata. El vendedor registra el tipo de presentación elegida por el cliente durante cuatro semanas consecutivas.

Se considera que ocurre un cambio cuando, de una semana a la siguiente, el cliente compra una presentación diferente a la adquirida la semana anterior. Por ejemplo, si en una semana compra una bebida embotellada y en la siguiente compra una bebida en lata, se registra un cambio.

Define las variables aleatorias:

- $X$: número de cambios de presentación observados durante las cuatro semanas
- $Y$: número de bebidas embotelladas compradas durante las cuatro semanas

Supón que las dos presentaciones, embotellada y en lata, son igualmente probables en cada semana y que las elecciones realizadas en semanas distintas son independientes.

a) Determina el espacio muestral correspondiente a las posibles secuencias de compra durante las cuatro semanas.

b) Construye la distribución de probabilidad conjunta de $X$ y $Y$.

c) Obtén las distribuciones marginales de $X$ y de $Y$.

d) ¿Son $X$ y $Y$variables aleatorias independientes? Justifica numéricamente tu respuesta utilizando las probabilidades obtenidas.

4.4.9 La siguiente tabla presenta la distribución de probabilidad conjunta de las variables aleatorias $X$ y $Y$, así como sus respectivas distribuciones marginales, donde:

- $X$: número de años de estudio concluidos por el jefe de familia
- $Y$: estrato de ingreso del jefe de familia, definido de acuerdo con el número de veces que su ingreso representa el salario mínimo general

La distribución conjunta es:

| **X \ Y** | **1** | **2** | **3** | **4** | **5** | **P(X=x)** |
|---|---|---|---|---|---|---|
| **0** | 0.199 | 0.124 | 0.122 | 0.005 | 0.000 | **0.450** |
| **3** | 0.177 | 0.034 | 0.009 | 0.003 | 0.000 | **0.223** |
| **6** | 0.008 | 0.025 | 0.040 | 0.049 | 0.065 | **0.187** |
| **9** | 0.002 | 0.005 | 0.022 | 0.041 | 0.070 | **0.140** |
| **P(Y=y)** | **0.386** | **0.188** | **0.193** | **0.098** | **0.135** | **1.000** |

A partir de la información de la tabla, responde lo siguiente:

a) Determina si las variables aleatorias $X$ y $Y$son independientes. Justifica numéricamente tu respuesta e interpreta el resultado en el contexto del problema.

b) Calcula e interpreta las siguientes probabilidades:

i. $P(Y < 3)$

ii. $P(Y > 4)$

iii. $P(X < 6)$

iv. $P(Y = 5 \mid X = 6)$

v. $P(X = 3 \mid Y = 1)$

c) Calcula los valores esperados $E(X)$ y $E(Y)$ e interpreta los resultados dentro del contexto del problema.

d) Calcula las varianzas $Var(X)$ y $Var(Y)$.

e) Calcula el coeficiente de correlación entre $X$ y $Y$ e interpreta su resultado en el estrato de ingreso del jefe de familia.

4.4.10 Una agencia automotriz cuenta con 9 automóviles disponibles: 3 Volkswagen, 2 Ford y 4 Chevrolet. Se seleccionan al azar 2 automóviles, sin reemplazo, para realizar una inspección.

Sean las variables aleatorias:

- $X$: número de automóviles Volkswagen seleccionados
- $Y$: número de automóviles Ford seleccionados

a) Obtén la distribución de probabilidad conjunta de $X$ y $Y$.

b) Calcula las siguientes probabilidades y, en cada caso, describe con palabras el evento correspondiente:

i. $P(\mathrm{X} = 2, \mathrm{Y} = 0)$

ii. $P(X = 2 \mid Y = 1)$

iii. $P(X + Y < 1)$

iv. $P(Y = 0 \mid X \leq 1)$

c) Sea $Z$ el número de automóviles Chevrolet seleccionados. Expresa $Z$ como una función de $X$ y $Y$ y calcula $P(Z \leq 1)$.

d) Calcula el valor esperado y la desviación estándar del número de automóviles Chevrolet seleccionados.

4.4.11 Sean $X$ y $Y$ dos variables aleatorias discretas cuya distribución de probabilidad conjunta está dada por la siguiente tabla:

| **Y \X** | **-1** | **0** | **1** |
|---|---|---|---|
| **-1** | 1/8 | 1/8 | 1/8 |
| **0** | 1/8 | 0 | 1/8 |
| **1** | 1/8 | 1/8 | 1/8 |

A partir de esta distribución, responde lo siguiente:

a) Calcula $E(X)$, $E(Y)$ y $E(XY)$.

b) Verifica que $E(XY) = E(X)E(Y)$ y, a partir de este resultado, demuestra que la covarianza entre $X$ y $Y$es cero y, por lo tanto, que su coeficiente de correlación lineal satisface $\rho_{XY} = 0$.

c) Determina si $X$y $Y$ son independientes. Justifica numéricamente tu respuesta utilizando la distribución conjunta y las distribuciones marginales.

d) Con base en los resultados anteriores, explica brevemente por qué el hecho de que dos variables tengan correlación igual a cero no implica necesariamente que sean independientes.

4.4.12 Un restaurante de comida rápida analiza los hábitos de consumo de sus clientes durante una visita. Para ello, se selecciona al azar a un cliente y se definen las siguientes variables aleatorias:

- $X$: número de hamburguesas que consume el cliente durante su visita
- $Y$: número de refrescos que consume el cliente durante la misma visita

A partir de datos históricos, se obtiene la siguiente distribución de probabilidad conjunta de $X$ y $Y$:

| X \ Y | 0 | 1 | 2 |
|---|---|---|---|
| **0** | 0.10 | 0.03 | 0.00 |
| **1** | 0.00 | 0.50 | 0.17 |
| **2** | 0.00 | 0.20 | 0.00 |

Con base en esta información, responde:

a) Determina si $X$ y $Y$son variables aleatorias independientes. Justifica numéricamente tu respuesta.

b) Si se sabe que el cliente consumió al menos un refresco, ¿cuál es la probabilidad de que haya consumido menos de una hamburguesa? Interpreta el resultado dentro del contexto del problema.

c) Calcula el coeficiente de correlación entre el número de hamburguesas y el número de refrescos consumidos. Interpreta tanto el signo como la magnitud del coeficiente obtenido en términos de la relación lineal entre ambas variables.

4.4.13 Un comerciante analiza las ventas semanales, en kilogramos, de dos tipos de café. Sean:

- $X$: cantidad semanal vendida, en kg, de café tipo A
- $Y$: cantidad semanal vendida, en kg, de café tipo B

A partir de los registros históricos, se sabe que:

$$E(X) = 50.8,\ \sigma_X = 3$$
$$E(Y) = 51.2,\ \sigma_Y = 4.3$$

Además, el coeficiente de correlación entre las ventas de ambos tipos de café es:

$$\rho_{XY} = -0.7$$

Si $T = X + Y$ representa la cantidad total de café vendida semanalmente, determina:

a) El valor esperado de la venta total.

b) La varianza de la venta total

c) Interpreta brevemente el efecto que tiene la correlación negativa entre las ventas de ambos tipos de café sobre la variabilidad de las ventas totales.

4.4.14 Una empresa de tecnología recibe un lote de 10 teléfonos celulares reacondicionados, de los cuales 3 presentan algún defecto y 7 funcionan correctamente. Como parte de un proceso de control de calidad, se seleccionan al azar 2 celulares, uno después del otro y sin reemplazo, para realizar una inspección.

Sean las variables aleatorias:

- $X$: toma el valor 1 si el primer celular seleccionado es defectuoso y 0 si funciona correctamente.
- $Y$: toma el valor 1 si el segundo celular seleccionado es defectuoso y 0 si funciona correctamente.

a) Obtén la distribución de probabilidad conjunta de $X$ y $Y$.

b) Determina si $X$ y $Y$son variables aleatorias independientes. Justifica numéricamente tu respuesta.

c) Calcula el coeficiente de correlación lineal entre $X$ y $Y$. Interpreta el signo y la magnitud del resultado en el contexto del problema.

d) Sea $W$ el número total de celulares defectuosos encontrados en las dos selecciones. Expresa $W$ en términos de $X$ y $Y$ y calcula $E(W)$ y $Var(W)$.

4.4.15 Una caja contiene 6 monedas: 4 monedas de $10 y 2 monedas de $5. Se seleccionan al azar y sin reemplazo 3 monedas simultáneamente.

Sean las variables aleatorias:

- $X$: número de monedas de $10 seleccionadas
- $Y$: número de monedas de $5 seleccionadas

a) Obtén la distribución de probabilidad conjunta de $X$ y $Y$.

b) Calcula la probabilidad de seleccionar al menos 2 monedas de $10 y como máximo 2 monedas de $5.

c) Calcula el coeficiente de correlación entre $X$ y $Y$ e interpreta el resultado dentro del contexto del problema.

d) Sea $T$ la cantidad total de dinero, en pesos, obtenida con las tres monedas seleccionadas. Expresa $T$ como una función de $X$ y $Y$.

e) Calcula el valor esperado y la varianza de la cantidad total de dinero seleccionada.

4.4.16 Sean $X$y $Y$dos variables aleatorias discretas cuya función de probabilidad conjunta está dada por:

$P(\mathrm{X}=\mathrm{x},\mathrm{Y}=\mathrm{y})=c(x^2+y^2)$ para $x\in\{-1,0,1,3\}$ y $y\in\{-1,2,3\}$,

donde $c$ es una constante.

a) Determina el valor de $c$ para que la expresión anterior sea una función de probabilidad conjunta válida.

b) Construye la tabla de distribución de probabilidad conjunta de $X$ y $Y$.

c) Calcula la probabilidad de que $X$ sea igual a 0 y $Y$ sea menor o igual que 2.

d) Calcula la probabilidad de que $X$ sea menor que 0 y $Y$sea mayor que 2.

e) Calcula la probabilidad de que $X$ sea mayor que $2-Y$.

4.4.17 Sean $X$ y $Y$dos variables aleatorias discretas para las cuales se conoce la siguiente información:

$$E(X)=2,\ E(Y)=-1$$
$$Var(X)=4,\ Var(Y)=6$$
$$Cov(X,Y)=\frac{1}{2}$$

Para cada una de las siguientes variables aleatorias, calcula su valor esperado y su varianza:

a) $Z=X+Y$
Calcula $E(Z)$y $Var(Z)$.

b) $W=X-Y$
Calcula $E(W)$y $Var(W)$.

c) $U=2X+Y$
Calcula $E(U)$y $Var(U)$.

d) $T=3X-2Y+2$
Calcula $E(T)$y $Var(T)$.

4.4.18 Se lanzan simultáneamente tres dados equilibrados de seis caras. Se definen las siguientes variables aleatorias:

- $X$: número de dados en los que se obtiene 1

- $Y$: número de dados en los que se obtiene 2

a) Obtén la función de probabilidad conjunta de $X$ y $Y$ y represéntala mediante una tabla.

b) Si se sabe que en ninguno de los tres dados se obtuvo un 2, determina la función de probabilidad condicional de $X$, es decir: $P(X = x \mid Y = 0)$.

c) Determina si $X$ y $Y$son variables aleatorias independientes. Justifica numéricamente tu respuesta.

d) Calcula la covarianza y el coeficiente de correlación entre $X$ y $Y$.

e) Interpreta el signo del coeficiente de correlación en el contexto del problema.

f) Sea $Z$el número de dados cuyo resultado es distinto de 1 y de 2. Expresa $Z$ como una función de $X$ y $Y$ y calcula $E(Z)$ y $Var(Z)$.

4.4.19 Considera el caso de un repartidor de plataforma digital operando durante la tarde de un domingo; el objetivo es analizar su desempeño basándose en las incidencias reportadas en sus entregas. Sean:

$X$ = número de pedidos cancelados por el cliente al momento de la llegada
$Y$ = número de errores en el pedido reportados por el cliente (bebida faltante, artículo equivocado, etc.)

Su distribución de probabilidad conjunta está dada por la siguiente tabla:

| **Y \ X** | **0** | **1** | **2** |
|---|---|---|---|
| **0** | 1/60 | 2/60 | 3/60 |
| **1** | 2/60 | 4/60 | 6/60 |
| **2** | 3/60 | 6/60 | 9/60 |
| **3** | 4/60 | 8/60 | 12/60 |

a) Calcula las siguientes probabilidades:

i. $P(X \leq 1, Y \leq 1)$

i. $P(X + Y \leq 1)$

ii. $P(Y - X > 2)$

iii. Si se sabe que $Y = 2$, calcula la probabilidad de que $X > 0$.

b) Calcula el coeficiente de correlación entre $X$ y $Y$. Interpreta el resultado en términos de la dirección e intensidad de la relación lineal entre ambas variables.

c) Define la nueva variable aleatoria: $Z = Y - X$. Calcula su valor esperado y su varianza:

$$E(Z) = E(Y - X)$$
$$Var(Z) = Var(Y - X)$$

4.4.20 Estás a cargo de supervisar el uso de los cubículos de estudio durante la semana de exámenes finales. Sean:

$X$ = número de personas que solicitan extender el tiempo de uso más allá de lo permitido, donde $X \in \{1, 2, 3\}$

$Y$ = número de reportes de fallas técnicas en los cubículos (internet lento, fallas en los enchufes, etc.), donde $Y \in \{2, 3, 4\}$

La función de probabilidad conjunta de $X$ y $Y$está definida por:

$$f(x, y) = \begin{cases} cxy, & x = 1,2,3 \text{ y } y = 2,3,4, \\ 0, & \text{en cualquier otro caso,} \end{cases}$$

donde $c$ es una constante.

a) Determina el valor de $c$ para que $f(\mathrm{x}, \mathrm{y})$ sea una función de probabilidad conjunta válida.

b) Calcula la probabilidad de que $X$ sea mayor o igual que 2 y, simultáneamente, $Y$ sea menor o igual que 3.

c) Obtén la distribución marginal de $X$ y calcula su valor esperado y su varianza.

d) Obtén la distribución marginal de $Y$ y calcula su valor esperado y su varianza.

e) Determina si $X$ y $Y$ son variables aleatorias independientes. Justifica tu respuesta utilizando las distribuciones marginales obtenidas.

f) Si se sabe que $Y = 3$, calcula la probabilidad de que $X = 2$: $P(X = 2 \mid Y = 3)$.

g) Si se sabe que $Y \leq 3$, calcula la probabilidad de que $X \geq 2$: $P(X \geq 2 \mid Y \leq 3)$.

h) Calcula la covarianza y el coeficiente de correlación entre $X$ y $Y$. Interpreta el resultado obtenido en términos de la relación lineal entre ambas variables.

i) Determina si $X$ y $Y$ son variables aleatorias independientes. Justifica numéricamente tu respuesta utilizando la función de probabilidad conjunta y las distribuciones marginales obtenidas anteriormente.

j) Sea $Z$ el número de estudiantes franceses a quienes se otorgó la beca. Define la nueva variable aleatoria: $Z = 5X - 3Y$. Calcula su valor esperado y su varianza.

4.4.21 De un grupo de nueve estudiantes: cuatro mexicanos, dos españoles y tres franceses, se van a elegir a tres estudiantes para otorgarles una beca. Sea $\mathrm{X}$ el número de estudiantes mexicanos elegidos y $\mathrm{Y}$ el número de estudiantes españoles elegidos.

a) Construye la tabla de probabilidad conjunta de $\mathrm{X}$ y $\mathrm{Y}$.

b) Encuentra las distribuciones marginales de $\mathrm{X}$ y $\mathrm{Y}$.

c) Calcula el valor esperado y la varianza de $\mathrm{X}$ y $\mathrm{Y}$.

d) Obtén el coeficiente de correlación entre $\mathrm{X}$ y $\mathrm{Y}$.

e) Si $Z$ es el número de estudiantes franceses a los que se les otorgó la beca, y $Z = 3 - \mathrm{X} - \mathrm{Y}$. Obtén el valor esperado y la varianza de $Z$.

4.4.22 Sean $A$ y $B$ dos eventos tales que:

$$P(A) = 0.25,\ P(B \mid A) = 0.50,\ P(A \mid B) = 0.25$$

Se definen las siguientes variables aleatorias indicadoras:

$$X = \begin{cases} 1 & \text{si ocurre el evento } A \\ 0 & \text{si no ocurre el evento } A \end{cases}$$

y, de manera análoga,

$$Y = \begin{cases} 1 & \text{si ocurre el evento } B \\ 0 & \text{si no ocurre el evento } B \end{cases}$$

a) Construye la tabla de distribución de probabilidad conjunta de $X$ y $Y$.

b) Determina si cada una de las siguientes afirmaciones es verdadera o falsa. Justifica tu respuesta mediante los cálculos correspondientes.

i. Las variables aleatorias $X$y $Y$son independientes.

ii. $P(X^2 + Y^2 = 1) = 0.25$

iii. $P(XY = X^2Y^2) = 1$

4.4.23 Un estudio de mercado analiza la relación entre el número de televisores que hay en un hogar y el número de integrantes de la familia.

Sean las variables aleatorias:

- $X$: número de televisores que hay en el hogar
- $Y$: número de integrantes de la familia

A partir de la información recopilada en diversos estudios de mercado, se obtuvo la siguiente distribución de probabilidad conjunta de $X$y $Y$:

| X \ Y | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 0 | 0.07 | 0.05 | 0.03 | 0.02 |
| 1 | 0.11 | 0.08 | 0.09 | 0.10 |
| 2 | 0.03 | 0.07 | 0.09 | 0.10 |
| 3 | 0.00 | 0.00 | 0.05 | 0.11 |

a) Obtén las distribuciones marginales de $X$ y de $Y$.

b) Calcula e interpreta las siguientes probabilidades:

i. La probabilidad de seleccionar un hogar con 3 integrantes: $P(Y = 3)$.

ii. Si se sabe que el hogar tiene 3 integrantes, calcula la probabilidad de que tenga 2 televisores: $P(X = 2 \mid Y = 3)$.

c) Calcula el valor esperado del número de televisores y del número de integrantes por hogar: $E(X)$ y $E(Y)$. Interpreta ambos resultados en el contexto del problema.

d) Determina si $X$y $Y$ son variables aleatorias independientes. Justifica numéricamente tu respuesta utilizando las probabilidades conjuntas y marginales.

e) Calcula el coeficiente de correlación lineal entre el número de televisores y el número de integrantes del hogar. Interpreta el signo y la magnitud del resultado.

f) Sea

$$R = \frac{X}{Y}$$

la variable que representa el número de televisores por integrante del hogar. Calcula:

$$E(R) = E\left(\frac{X}{Y}\right)$$

e interpreta el resultado en el contexto del problema.

4.4.24 Un investigador desea analizar si existe alguna relación entre el desempeño académico de los estudiantes y el número de viajes al extranjero que realizaron durante el último año.

Para ello, recopiló información de 3,200 estudiantes y registró las siguientes variables:

- $X$: número de viajes al extranjero realizados durante el último año
- $Y$: número de materias reprobadas durante el mismo periodo

Los resultados obtenidos se presentan en la siguiente tabla de frecuencias conjuntas:

| **Y \ X** | **0** | **1** | **2** | **3** | **4** | **Total** |
|---|---|---|---|---|---|---|
| **0** | 187 | 300 | 150 | 225 | 338 | **1200** |
| **1** | 156 | 250 | 125 | 189 | 280 | **1000** |
| **2** | 125 | 200 | 100 | 150 | 225 | **800** |
| **3** | 32 | 50 | 25 | 36 | 57 | **200** |
| **Total** | **500** | **800** | **400** | **600** | **900** | **3200** |

a) Construye la tabla de distribución de probabilidad conjunta de $X$ y $Y$.

b) Obtén las distribuciones marginales de $X$ y de $Y$.

c) Calcula e interpreta las siguientes probabilidades:

i. La probabilidad de que un estudiante haya reprobado menos de 2 materias: $P(Y < 2)$.

ii. La probabilidad de que haya realizado menos de 2 viajes al extranjero: $P(X < 2)$.

iii. Si se sabe que un estudiante realizó 3 viajes al extranjero, ¿cuál es la probabilidad de que no haya reprobado ninguna materia?

iv. Si se sabe que un estudiante no reprobó ninguna materia, ¿cuál es la probabilidad de que no haya realizado ningún viaje al extranjero?

v. Si se sabe que un estudiante realizó 3 viajes al extranjero, ¿cuál es la probabilidad de que haya reprobado como máximo 1 materia?

d) Determina si $X$ y $Y$son variables aleatorias independientes. Justifica numéricamente tu respuesta.

e) Calcula el coeficiente de correlación lineal entre el número de viajes al extranjero y el número de materias reprobadas. Interpreta el signo y la magnitud del resultado.

f) A partir de los resultados obtenidos, ¿qué puede concluir el investigador acerca de la relación entre el número de viajes al extranjero y el número de materias reprobadas? ¿Los resultados permiten establecer una relación causal? Justifica tu respuesta.

4.4.25 Sea $X$ una variable aleatoria discreta con la siguiente distribución de probabilidad:

| **x** | **-2** | **-1** | **1** | **2** |
|---|---|---|---|---|
| P(X=x) | 1/4 | 1/4 | 1/4 | 1/4 |

Se define una nueva variable aleatoria $Y$ como una función de $X$:

$$Y = X^2$$

a) Determina los posibles valores de $Y$y construye la distribución de probabilidad conjunta de $X$y $Y$.

b) Calcula la covarianza y el coeficiente de correlación entre $X$y $Y$. Interpreta el resultado obtenido.

c) Determina si $X$ y $Y$son variables aleatorias independientes. Justifica numéricamente tu respuesta.

d) A partir de los resultados anteriores, explica por qué un coeficiente de correlación igual a cero no implica necesariamente independencia entre dos variables aleatorias.

## 4.8 Funciones lineales de variables aleatorias. Propiedades. Esperanza y varianza

4.8.1 El ingreso mensual nominal $X$en miles de pesos de un profesionista joven tiene:

$$E(X) = 18, Var(X) = 9$$

Debido a la inflación, su salario real se calcula como:

$$Y = 0.85X - 1$$

a) Calcula $E(Y)$.

b) Calcula $Var(Y)$y $SD(Y)$.

c) Interpreta económicamente $E(Y)$.

d) ¿Qué efecto tiene el factor 0.85 sobre la dispersión?

e) ¿Qué pasaría con la varianza si la inflación sube y el factor baja a 0.70?

4.8.2 Las ventas diarias $X$en miles de pesos de un vendedor cumplen:

$$E(X) = 40, Var(X) = 16$$

Su ingreso diario es:

$$Y = 0.05X + 2$$

a) Calcula $E(Y)$ y $Var(Y)$.

b) ¿Cuál es la desviación estándar del ingreso?

c) Interpreta el término $+2$ dentro del contexto del problema.

d) Si la comisión sube a 8%, ¿cómo cambian $E(Y)$ y $Var(Y)$?

e) ¿El ingreso es más riesgoso que las ventas?

4.8.3 El tipo de cambio diario $X$ pesos por dólar tiene:

$$E(X) = 17.5, \ \ Var(X) = 1$$

Una empresa importa mercancía por USD $50,000, por lo que el costo en pesos es:

$$Y = 50{,}000\, X$$

a) Calcula $E(Y)$ y $Var(Y)$.

b) Calcula $SD(Y)$.

c) Interpreta el riesgo cambiario.

d) ¿Qué pasaría con $Var(Y)$ si el tipo de cambio aumenta a 21.5?

# TEMA 5: Algunas distribuciones de probabilidad

Se usarán los códigos en R del Anexo 2: Códigos en R para distribuciones de probabilidad.

## 5.1 Distribuciones discretas

### DISTRIBUCIÓN UNIFORME DISCRETA

5.1.1 En un hackatón universitario sobre inteligencia artificial, los equipos deben presentar sus proyectos ante el jurado. El orden de presentación se asigna aleatoriamente entre los 8 equipos finalistas. Cada equipo tiene la misma probabilidad de quedar en cualquiera de las posiciones del 1 al 8.

a) ¿Cuál es la probabilidad de que el equipo “DeepCoders” presente en el primer turno?

b) ¿Cuál es la probabilidad de que presente en un turno par?

c) Si los organizadores seleccionan al azar un solo equipo para recibir comentarios del CEO invitado, ¿cuál es la probabilidad de que el equipo seleccionado sea DeepCoders?

d) ¿Cuál es el valor esperado del turno de presentación asignado a DeepCoders?

5.1.2 Una universidad recibió 10 laptops reacondicionadas, cada una de una marca diferente: Apple, Dell, HP, Lenovo, Asus, Acer, Samsung, Huawei, Microsoft y MSI.

Las laptops serán asignadas a 10 estudiantes beneficiarios mediante un sorteo, de manera que cada estudiante tenga la misma probabilidad de recibir cualquiera de las marcas disponibles.

Juan es uno de los estudiantes beneficiarios y recibirá una laptop seleccionada al azar.

a) ¿Cuál es la probabilidad de que a Juan le toque una laptop de una marca que él considera premium: Apple, Dell o Microsoft?

b) ¿Cuál es la probabilidad de que reciba una laptop cuyo nombre de marca comience con una letra comprendida entre A y H, inclusive?

c) Si cada laptop tiene un puntaje de desempeño igual a su posición alfabética Apple=1, Asus=2, Acer=3, …, MSI=10, ¿cuál es el valor esperado del puntaje asociado a la laptop de Juan?

d) ¿Cuál es la varianza de la distribución uniforme discreta correspondiente a estos puntajes?

**ENSAYOS BERNOULLI**

5.1.3 Una universidad está probando un sistema de autenticación facial para facilitar el acceso a plataformas académicas. Cada vez que un estudiante intenta desbloquear con reconocimiento facial, el sistema puede:

- Reconocer correctamente al usuario éxito = 1
- No reconocerlo y pedir contraseña fracaso = 0

Durante la fase piloto, se sabe que el sistema reconoce correctamente a un estudiante con probabilidad 0.87.

a) Si un estudiante realiza un intento, ¿cuál es la probabilidad de que el sistema falle?

b) ¿Cuál es el valor esperado y la varianza del ensayo Bernoulli que modela un intento de desbloqueo?

c) Si el sistema necesita mejorar, ¿qué interpretación le darías a que $p = 0.87$ sea "bueno" o "malo" dependiendo del contexto?

5.1.4 Un profesor utiliza un detector de plagio basado en inteligencia artificial para revisar tareas. Cada tarea que se analiza puede clasificarse como:

- Libre de plagio éxito = 1
- Con plagio significativo fracaso = 0

Según pruebas internas, el algoritmo identifica correctamente si una tarea está libre de plagio con probabilidad 0.93.

a) ¿Cuál es la probabilidad de que el algoritmo clasifique incorrectamente una tarea?

b) Modelando el proceso como un ensayo Bernoulli, ¿cuál es su valor esperado y su varianza?

c) ¿Qué implicaciones éticas tiene un $\mathrm{p} = 0.93$ en un curso con tareas de alto impacto en la calificación?

**DISTRIBUCIÓN BINOMIAL**

5.1.5 En una campaña digital, cada usuario puede reaccionar a un mensaje de 3 formas:

- "Like" con probabilidad 0.5
- "Compartir" con probabilidad 0.3
- "Ignorar" con probabilidad 0.2

Se seleccionan 8 personas al azar de manera independiente. Sea M = número de personas que dan “Like”.

a) Encuentra la distribución de probabilidad de M.

b) Calcula la probabilidad de que al menos 6 personas den “Like”

c) Calcula E(M) e interpreta dentro del contexto del problema.

d) Calcula Var(M) e interpreta dentro del contexto del problema.

5.1.6 Una empresa lanza un chatbot con IA para resolver dudas administrativas. Cada vez que un cliente hace una pregunta, el chatbot puede:

- Responder correctamente éxito con probabilidad p = 0.78
- Responder incorrectamente fracaso con probabilidad 1 – p.

En un día típico se registran 40 preguntas independientes entre sí.

a) ¿Cuál es la probabilidad de que el chatbot responda al menos 30 preguntas correctamente?

b) ¿Cuál es la probabilidad de que falle más de 10 veces?

c) ¿Cuál es el número esperado de respuestas correctas?

5.1.7 Una organización no gubernamental ONG en México está lanzando una campaña digital a través de redes sociales para promover una petición ciudadana en favor de una reforma ambiental. Con base en datos históricos de campañas previas, se sabe que la probabilidad de que un ciudadano joven decida firmar digitalmente la petición al ver la publicación es de exactamente p = 0.12.

Durante una prueba piloto, el algoritmo muestra la publicación de forma independiente a un grupo focalizado de 20 estudiantes universitarios.

Sea $X$ la variable aleatoria que representa el número total de estudiantes que deciden firmar la petición.

a) ¿Cuál es la probabilidad de obtener exactamente 15 firmas dentro del grupo de los 20 estudiantes?

b) ¿Cuál es la probabilidad de obtener al menos 15 firmas?

c) ¿Cuál es la probabilidad de tener más de 10 firmas?

d) ¿Cuál es el valor esperado media y la varianza de la distribución del número total de firmas obtenidas en esta muestra?

e) Calcula la probabilidad de que ninguno $X = 0$ de los 20 estudiantes decida firmar la petición. ¿Qué significaría operativamente este resultado para el equipo de comunicación de la ONG?

f) Determina la moda de esta distribución binomial el número de firmas que tiene la mayor probabilidad de ocurrir.

5.1.8 La Comisión Nacional Bancaria y de Valores (CNBV) exige a los bancos comerciales monitorear transacciones sospechosas mediante sistemas automatizados. En una sucursal bancaria de alta gama, se sabe que la probabilidad de que una alerta de transferencia internacional sea un "falso positivo" un cliente legítimo haciendo un movimiento inusual es de p = 0.85. Un auditor de cumplimiento normativo decide revisar detalladamente un lote aleatorio de 12 alertas emitidas durante el día.

Sea $X$ el número de alertas que resultan ser falsos positivos.

a) ¿Cuál es la probabilidad de que las 12 alertas revisadas sean falsos positivos?

b) Para optimizar el tiempo del equipo, se considera que la jornada es eficiente si al menos 10 de las 12 alertas resultan ser falsos positivos evitando investigaciones criminales innecesarias. Calcula esta probabilidad.

c) Determina el valor esperado y la desviación estándar de las alertas que sí requieren una investigación profunda por lavado de dinero, es decir, las que *no* son falsos positivos.

5.1.9 Tras una fuerte campaña de desinformación en redes sociales, la Organización Panamericana de la Salud OPS detecta que la tasa de rechazo a la vacuna de refuerzo contra el sarampión en una comunidad fronteriza es del 20% (p = 0.20). Una brigada médica internacional visita un centro comunitario y atiende de forma independiente a 25 madres y padres de familia.

a) ¿Cuál es el número esperado de padres de familia que rechazarán la vacuna en este grupo de 25? Calcula también su varianza.

b) Calcula la probabilidad de que más de 8 padres del grupo decidan no vacunar a sus hijos.

c) Toma de decisiones: La brigada médica cuenta únicamente con el presupuesto y el personal para aplicar talleres intensivos de concientización si el número de rechazos supera los 7 casos en la muestra. ¿Cuál es la probabilidad de que la

brigada se vea obligada a activar el protocolo de emergencia y gastar sus recursos en esa comunidad? ¿Consideras que el tamaño de la muestra es suficiente para capturar el riesgo real?

**DISTRIBUCIÓN GEOMÉTRICA**

5.1.10 Una persona usa su asistente virtual tipo Siri o Google Assistant. Cada vez que da una instrucción, el sistema la reconoce correctamente con probabilidad p = 0.85. Se repite el comando hasta que el reconocimiento sea correcto.

a) ¿Cuál es la probabilidad de que el asistente reconozca correctamente la instrucción en el primer intento?

b) ¿Cuál es la probabilidad de que se necesiten 3 intentos?

c) ¿Cuál es la probabilidad de que requiera más de 5 intentos?

5.1.11 En la plataforma académica CANVAS, a veces los archivos no cargan correctamente por saturación. Cada intento de subir un archivo tiene probabilidad p = 0.85 de completarse exitosamente. Los intentos continúan hasta que la carga se logra.

a) ¿Cuál es la probabilidad de que el archivo se suba en el segundo intento?

b) ¿Cuál es la probabilidad de que la carga se logre en 4 intentos o menos?

c) ¿Cuál es el número esperado de intentos para lograr la carga exitosa?

5.1.12 Un encuestador busca personas que respondan positivamente a la pregunta: *"¿Cree que el nuevo "Poder Judicial" es más efectivo?"*. Se sabe que la probabilidad de respuesta afirmativa es p = 0.4.
Sea $Y$ = número de personas entrevistadas hasta obtener la primera respuesta afirmativa.

a) Identifica la distribución de $Y$.

b) Calcula $P(Y = 3)$.

c) Calcula $E(Y)$. Interpreta el valor esperado dentro del contexto del problema.

5.1.13 Un alumno está intentando unirse a una videollamada de clase en Zoom. Cada intento de conexión se establece correctamente con probabilidad p = 0.72. Cada intento es independiente.

a) ¿Cuál es la probabilidad de que el alumno logre entrar en el tercer intento?

b) ¿Cuál es la probabilidad de que el alumno necesite al menos 4 intentos?

c) ¿Cuál es el valor esperado del número de intentos necesarios para conectarse?

**DISTRIBUCIÓN POISSON**

5.1.14 En una estación de bicicletas públicas cerca del campus, se observa que en promedio llegan $\lambda = 4.5$ bicicletas por hora. Suponiendo que las llegadas siguen un proceso Poisson:

a) ¿Cuál es la probabilidad de que en una hora lleguen exactamente 6 bicicletas?

b) ¿Cuál es la probabilidad de que lleguen al menos 5 bicicletas?

c) ¿Cuál es el número esperado de bicicletas que llegan en 2 horas?

5.1.15 Un empleado recibe en promedio $\lambda = 18$ notificaciones de mensajes por hora en sus grupos de trabajo y chats con compañeros.

a) ¿Cuál es la probabilidad de que reciba exactamente 20 notificaciones en una hora?

b) ¿Cuál es la probabilidad de recibir menos de 15?

c) ¿Cuántas notificaciones se esperan en 30 minutos?

5.1.16 Durante las videollamadas por celular, se observa que en promedio ocurren 3 fallas técnicas en llamadas de 15 minutos congelamientos, cortes breves de audio o pérdida momentánea de señal. Estas fallas se presentan de manera independiente y con una tasa constante, por lo que el número de interrupciones durante una llamada se modela con una distribución de Poisson con parámetro $\lambda = 3$.

a) ¿Cuál es la probabilidad de que durante una videollamada de 15 minutos no ocurra ninguna falla técnica?

b) ¿Cuál es la probabilidad de que la llamada presente exactamente 2 fallas?

c) ¿Cuál es la probabilidad de que ocurran 5 o más fallas en una llamada de media hora?

d) ¿Cuál es la probabilidad de que ocurran menos de 6 falladas en una llamada de una hora?

**DISTRIBUCIONES COMBINADAS**

5.1.17 Una plataforma de streaming reporta que:

- En promedio el servidor presenta 4 micro-fallas por hora.
- Cada micro-falla tiene una probabilidad p = 0.15 de provocar que un video se detenga.
- Cuando un video se detiene, el usuario vuelve a presionar “Play”. La probabilidad de que el video arranque correctamente es 0.8. El usuario vuelve a presionar “Play” hasta lograr de nuevo la conexión.

a) ¿Cuál es la probabilidad de que ocurran exactamente 5 micro-fallas en una hora?

b) Dadas esas 5 micro-fallas, ¿cuál es la probabilidad de que 2 detengan el video?

c) Si a un usuario el video se le detuvo, ¿cuál es la probabilidad de que necesite 3 intentos para que vuelva a reproducirse?

5.1.18 Una app de entregas en la ciudad analiza su funcionamiento:

- En promedio llegan 6 pedidos cada minuto.
- Cada pedido puede ser de uno de los siguientes tipos todos con igual probabilidad: comida, supermercado, farmacia, papelería
- Cada pedido tiene probabilidad 0.4 de ser asignado a un repartidor en motocicleta y probabilidad 0.6 de ser asignado a un repartidor en bicicleta. Para la variable indicadora de asignación, consideraremos ‘motocicleta’ como éxito.
- Si el pedido va en bicicleta, este podría no concretarse a tiempo.
- La probabilidad de entrega exitosa por intento es 0.7.
- Se intentará de nuevo hasta lograrlo.

a) ¿Cuál es la probabilidad de que lleguen exactamente 8 pedidos en un minuto?

b) Dado que llegó un pedido, ¿cuál es la probabilidad de que sea de farmacia?

c) Si llegaron 8 pedidos, ¿cuál es la probabilidad de que 3 sean asignados a moto?

d) Si un pedido va en bicicleta, ¿cuál es la probabilidad de que se entregue en el tercer intento?

5.1.19 Un sistema detecta intentos sospechosos de acceso:

- En promedio hay 9 intentos de acceso sospechoso por hora.
- Cada intento tiene una probabilidad de 0.25 de ser realmente malicioso
- Cuando un intento es malicioso, el firewall logra bloquearlo con probabilidad de 0.9. Si falla, vuelve a intentarlo hasta bloquearlo.

a) ¿Cuál es la probabilidad de que haya 12 intentos sospechosos en una hora?
b) Si hubo 12 intentos, ¿cuál es la probabilidad de que 3 sean maliciosos?
c) Para un intento malicioso, ¿cuál es la probabilidad de que el firewall necesite 3 intentos para bloquearlo?

5.1.20 Una universidad implementa una plataforma con IA para gestionar ejercicios y exámenes. Los alumnos envían dudas y tareas a esta plataforma.

- En promedio se reciben 12 solicitudes por hora.
- Cada consulta pertenece a uno de los 5 módulos del curso, con igual probabilidad: Estadística descriptiva, Probabilidad, Variables aleatorias, Distribuciones de probabilidad, Distribuciones bivariadas.
- Cada consulta enviada es respondida correctamente por la IA con probabilidad de 0.85.
- Si la IA no responde bien, el estudiante reenvía la misma duda. La probabilidad de respuesta correcta por reintento es 0.85 y se repite hasta obtenerla.

a) ¿Cuál es la probabilidad de recibir 15 consultas en una hora?
b) Si llega una consulta, ¿cuál es la probabilidad de que sea del módulo de “Probabilidad”?
c) Si llegaron 15 consultas, ¿cuál es la probabilidad de que 13 se respondan correctamente?
d) Si una consulta fue respondida incorrectamente, ¿cuál es la probabilidad de que el estudiante necesite 4 intentos para lograr una respuesta correcta?

## 5.2 Distribuciones continuas

### DISTRIBUCIÓN UNIFORME CONTINUA

5.2.1 Una plataforma gubernamental promete que el tiempo de respuesta para un trámite en línea se distribuye uniformemente entre 2 y 10 días.

Sea $X \sim U(2,10)$, donde $X$es el número de días que tarda un trámite en resolverse.

a) Escribe la función de densidad $fx$.

b) Calcula $P(X \leq 5)$.

c) Calcula $P(4 \leq X \leq 8)$.

d) Calcula $E(X)$y $Var(X)$.

e) Si un ciudadano espera más de 7 días, ¿qué proporción de trámites cumplen esa condición?

f) Interpreta por qué la distribución uniforme puede ser un modelo razonable o no en este contexto.

5.2.2 Durante una jornada cambiaria, el precio del dólar fluctúa entre $16.80 y $17.60 pesos. Supón que el precio intradía $X$se puede modelar como:

$$X \sim U(16.8, 17.6)$$

a) Escribe la función de densidad $fx$

b) Calcula $P(X < 17.0)$

c) Calcula $P(17.2 < X < 17.5)$

d) Calcula $E(X)$ y $Var(X)$.

e) ¿Cuál es la probabilidad de que el precio esté a más de $0.50 del valor esperado?

f) ¿Qué riesgos implica esta variabilidad para una empresa importadora?

**DISTRIBUCIÓN EXPONENCIAL**

5.2.3 En una plataforma de atención a clientes, el tiempo en horas entre dos quejas consecutivas se modela con una distribución exponencial con media de 4 horas.

$$X \sim \text{Exp}(\lambda), \qquad E(X) = \frac{1}{\lambda} = 4$$

a) Determina el valor de $\lambda$.

b) Calcula $P(X < 2)$.

c) ¿Qué tan común es esperar más de 6 horas entre quejas?

d) ¿El resultado sugiere que el sistema es estable o problemático?

e) Si la empresa quiere que la probabilidad de esperar más de 6 horas sea menor a 5%, ¿qué debería cambiar: el proceso o la tasa de llegada?

f) ¿Qué implicaciones tiene la propiedad de falta de memoria para la asignación de personal?

5.2.4 El tiempo en meses entre auditorías fiscales a empresas se modela como exponencial con media de 18 meses.

$$X \sim \text{Exp}(\lambda), \qquad E(X) = 18$$

a) Determina $\lambda$.

b) Calcula $P(X > 24)$.

c) Calcula $P(6 < X < 12)$.

d) ¿Qué tan riesgoso es para una empresa no ser auditada por varios años?

e) ¿Cómo cambiaría esta probabilidad si la tasa de auditorías aumenta?

5.2.5 El tiempo en años hasta que falla un servidor se distribuye exponencialmente con media de 5 años.

$$X \sim \text{Exp}(\lambda), E(X) = 5$$

a) Determina $\lambda$.

b) Calcula $P(X > 7)$.

c) Calcula $P(X \leq 3)$.

d) Explica por qué la distribución exponencial puede ser razonable para modelar fallas.

e) ¿Con qué probabilidad el servidor durará más que su vida media?

f) ¿Es conveniente reemplazar el servidor antes de los 5 años? Justifica tu respuesta.

**DISTRIBUCIÓN NORMAL**

5.2.6 Una empresa usa un algoritmo de IA para filtrar CVs. El puntaje de adecuación que asigna a los candidatos se distribuye como Normal con parámetros:

$$\mu = 72, \sigma = 8$$

a) ¿Cuál es la probabilidad de que un candidato tenga un puntaje mayor a 85?

b) ¿Cuál es la probabilidad de que un candidato tenga un puntaje entre 65 y 80?

c) ¿Qué proporción de candidatos queda por debajo de 60 puntos?

d) Si la empresa quiere entrevistar solo al 10% superior, ¿a partir de qué puntaje debe cortar?

e) ¿Qué significa estar en el percentil 90 en este contexto?

f) ¿Qué riesgos éticos implica usar este umbral automático?

5.2.7 En una ciudad, la concentración de microplásticos partículas por litro se distribuye como $N\sim(45, 36)$

a) Probabilidad de que un litro tenga más de 55 partículas.

b) Probabilidad de que esté entre 38 y 50 partículas.

c) ¿Qué porcentaje de muestras excede 60 partículas?

d) ¿Cuál es el valor que delimita el 95% superior?

e) ¿Qué significa que una muestra esté en el percentil 97?

f) Si el límite legal es 50, ¿qué proporción de la red incumple?

5.2.8 El tiempo diario en minutos que los universitarios pasan en redes sociales se distribuye como Normal con parámetros:

$$\mu = 195, \sigma = 40$$

a) Probabilidad de usar más de 4 horas diarias.

b) Probabilidad de usar entre 2 y 3 horas.

c) ¿Qué porcentaje usa menos de 120 minutos?

d) ¿A partir de cuántos minutos está el 20% más intenso?

e) ¿Qué representa estar por encima del percentil 80?

f) ¿Qué política universitaria podrías justificar con estos datos?

5.2.9 El tiempo de entrega en horas de pedidos exprés sigue una distribución Normal con parámetros:

$$\mu = 26, \sigma = 4$$

a) Probabilidad de que llegue en menos de 22 horas.

b) Probabilidad de que llegue entre 24 y 30 horas.

c) ¿Qué porcentaje tarda más de 35 horas?

d) ¿Cuál es el tiempo máximo que garantiza el 95% de entregas?

e) ¿Qué implica que un pedido esté en el percentil 90?

f) ¿Es realista prometer entregas en 24 horas?

**DISTRIBUCIÓN NORMAL BIVARIADA**

5.2.10 Una empresa multinacional ha implementado un sistema de inteligencia artificial que evalúa automáticamente a los candidatos que aplican a sus vacantes. El algoritmo asigna dos puntajes independientes del evaluador humano:

- $X$: puntaje de competencias técnicas
- $Y$: puntaje de habilidades socioemocionales

A partir de miles de evaluaciones pasadas, se ha observado que ambos puntajes se distribuyen de manera aproximadamente normal y presentan una correlación positiva, ya que los candidatos con mayor preparación técnica tienden también a desempeñarse mejor en entrevistas conductuales.

Se estima que:

$$\mu_X = 70, \mu_Y = 65, \sigma_X = 10, \sigma_Y = 8, \rho = 0.6$$

a) Calcula $P(X > 80, Y > 70)$

b) Calcula $P(X < 60, Y > 70)$

c) Calcula $P(X > 80)$, $P(Y > 70)$ y compáralos con el inciso (a).

d) Calcula $E(Y \mid X = 80)$

e) ¿Qué significa $\rho = 0.6$ en este contexto?

f) ¿Es razonable exigir ambos puntajes altos para contratar?

5.2.11 Durante los meses de verano, una empresa de energía monitorea la relación entre la temperatura diaria promedio en la ciudad y el consumo total de electricidad. En días más calurosos, el uso de aire acondicionado incrementa la demanda, generando una relación positiva entre ambas variables.

Definimos:

- $X$: temperatura diaria promedio °C
- $Y$: consumo diario de electricidad GWh

A partir de registros históricos, se ha determinado que el vector $(X, Y)$sigue una distribución normal bivariada con:

$$\mu_X = 28, \mu_Y = 420, \sigma_X = 4, \sigma_Y = 60, \rho = 0.75$$

a) Calcula $P(X > 32,\ Y > 480)$

b) Calcula $P(24 < X < 32,\ Y > 450)$

c) Calcula $E(Y \mid X = 35)$.

d) Calcula $Var(Y \mid X = 35)$.

e) ¿Qué indica una correlación alta positiva?

f) ¿Cómo usarías este modelo para planear infraestructura energética?

5.2.12 Un organismo internacional analiza datos de varios países para estudiar la relación entre el índice de percepción de corrupción y la tasa anual de crecimiento económico.

- $X$: índice de corrupción valores altos = más corrupción
- $Y$: crecimiento anual del PIB %

Los datos sugieren una relación negativa: los países con mayor corrupción tienden a crecer menos.

$$\mu_X = 55, \mu_Y = 2.8, \sigma_X = 10, \sigma_Y = 1.2, \rho = -0.65.$$

a) $P(X > 60,\ Y < 2)$

b) $P(40 < X < 60,\ Y > 3)$

c) Calcula $E(Y \mid X = 70)$.

d) ¿Qué signo tiene $Cov(X, Y)$?

e) ¿Qué significa una correlación negativa fuerte?

f) ¿Cómo usarías estos resultados para diseñar una política pública?

## 5.3 Relación entre la distribución Poisson y la distribución exponencial

5.3.1 En una app de servicios, llegan en promedio 3 quejas por hora.

a) ¿Cuál es la distribución del número de quejas en 2 horas?

b) ¿Cuál es la probabilidad de que no llegue ninguna queja en la próxima hora?

c) ¿Cuál es la probabilidad de que el tiempo de espera hasta la próxima queja sea mayor a 30 minutos?

d) Si ya han pasado 20 minutos sin quejas, ¿cuál es la probabilidad de esperar al menos 20 minutos más?

e) ¿Qué tan frecuente es una hora sin quejas? ¿Es un buen indicador de calidad del servicio?

f) Si la empresa quiere que la probabilidad de esperar más de 30 minutos sin quejas sea menor a 10%, ¿debería reducir la tasa de quejas? ¿A qué valor de $\lambda$?

g) ¿Qué implicaciones tiene la propiedad de falta de memoria para el diseño de turnos de atención?

h) ¿Qué factores externos podrían violar el supuesto de tasa constante?

5.3.2 En una sucursal llegan en promedio 10 clientes por hora.

a) ¿Cuál es la probabilidad de que lleguen exactamente 15 clientes en 1 hora?

b) ¿Cuál es la probabilidad de que lleguen menos de 5 clientes en 30 minutos?

c) ¿Cuál es la probabilidad de que el tiempo hasta el siguiente cliente sea mayor a 5 minutos?

d) Interpreta la propiedad de falta de memoria.

e) ¿Es probable que el banco esté subutilizado en horas pico?

f) ¿Cuántos clientes en promedio debería poder atender el personal en una hora para evitar saturación?

g) ¿Cómo usarías este modelo para planear el número de cajeros?

h) ¿En qué situaciones el proceso Poisson dejaría de ser válido?

5.3.3 Las auditorías a empresas ocurren con una tasa promedio de 2 por año.

a) ¿Cuál es la probabilidad de que en 3 años ocurra a lo más 2 auditorías?

b) ¿Cuál es la probabilidad de que no haya auditorías en el próximo año?

c) ¿Cuál es la distribución del tiempo hasta la próxima auditoría?

d) ¿Cuál es la probabilidad de que pasen más de 6 meses sin auditorías?

e) ¿Qué tan riesgoso es para una empresa no ser auditada en varios años?

f) ¿Cómo cambia la percepción de riesgo si la tasa de auditorías se duplica?

g) ¿Qué estrategias contables serían más prudentes con base en este modelo?

h) ¿Qué supuestos podrían fallar en contextos de reformas fiscales?

5.3.4 Un sistema eléctrico presenta en promedio 4 fallas por año.

a) ¿Cuál es la probabilidad de que ocurran exactamente 5 fallas en 1 año?

b) ¿Cuál es la probabilidad de que no ocurra ninguna falla en 3 meses?

c) ¿Cuál es la probabilidad de que el tiempo hasta la siguiente falla sea mayor a 2 meses?

d) Si ya transcurrió 1 mes sin fallas, ¿cuál es la probabilidad de esperar al menos 2 meses más?

e) ¿Qué tan confiable es el sistema con esta tasa de fallas?

f) ¿Cada cuánto debería programarse mantenimiento preventivo?

g) ¿Cómo afectaría a la varianza del número de fallas una mejora tecnológica?

h) ¿Qué eventos extremos podrían romper el modelo?

## 5.4 Aproximación de la distribución binomial y Poisson por medio de la distribución normal

5.4.1 Una encuesta a n = 400 ciudadanos muestra que la probabilidad de aprobar una reforma es p = 0.52.

Sea $X \sim Bin(400, 0.52)$.

a) Verifica si es válida la aproximación normal.

b) ¿Qué tan probable es observar al menos 220 personas a favor? Usa la corrección por continuidad.

c) Interpreta el resultado en términos políticos.

d) ¿Este resultado confirmaría que la mayoría apoya la reforma?

e) ¿Qué pasaría si la muestra fuera de solo 50 personas?

5.4.2 En una planta, la probabilidad de defecto es p = 0.04 y se inspeccionan n = 600 productos.

Sea $X \sim Bin(600,0.04)$.

a) Verifica si puede usarse la normal.

b) ¿Qué tan común es observar más de 30 defectos? Usa la corrección por continuidad.

c) Interpreta el resultado para control de calidad.

d) ¿Indica esto que el proceso está fuera de control?

e) ¿Qué riesgos implica una mala aproximación?

f) ¿Qué decisión tomarías si esta probabilidad supera el 10%?

g) ¿Cómo usarías esta información para definir estándares de calidad?

5.4.3 Una dependencia de la administración pública encargada de servicios urbanos limpieza, mantenimiento de vialidades y alumbrado ha registrado, durante los últimos años, un promedio de 50 accidentes laborales al año entre su personal operativo. Estos accidentes incluyen caídas, lesiones por maquinaria, descargas eléctricas leves y otros incidentes que requieren atención médica y generan ausentismo laboral.

La dirección desea evaluar si el número anual de accidentes que se presenta en un año determinado puede considerarse “normal” o si representa una señal de alerta que justifique reforzar los programas de seguridad y capacitación.

Se sabe que el número de accidentes por año se puede modelar como una variable aleatoria con distribución:

$$X \sim \text{Poisson}(50)$$

a) Verifica si es razonable usar la aproximación normal.

b) Aproxima $P(45 \leq X \leq 60)$ usando corrección por continuidad.

c) Interpreta el resultado en términos del funcionamiento habitual de la dependencia.

d) ¿El intervalo 45–60 accidentes representa un año “normal” para la dependencia?

e) ¿Qué tan probable es observar un año con más de 60 accidentes?

f) ¿Qué acciones preventivas se justifican si esa probabilidad no es despreciable?

g) ¿Cómo podrías evaluar si una nueva política de seguridad redujo efectivamente los accidentes?

**MINICASO: Monitoreo diario de e-scooters**

Una empresa de movilidad, que opera una flotilla de e-scooters, monitorea diariamente las fallas mecánicas frenos defectuosos, baterías dañadas, luces que no funcionan que puedan presentar. A partir de datos históricos se sabe que:

- la probabilidad de que una e-scooter presente alguna falla mecánica es $p = 0.1$
- los usuarios pueden enviar un reporte de fallas a través de la app. La plataforma recibe un promedio de 12 reportes al día por fallas mecánicas. Considera una jornada laboral de 9 horas.

Para cada una de las siguientes preguntas justifica el modelo probabilístico elegido y realiza los cálculos correspondientes:

a) Si se inspeccionan 15 e-scooters al azar, ¿cuál es la probabilidad de encontrar al menos 2 con fallas mecánicas?

b) ¿Cuál es la probabilidad de recibir exactamente 5 reportes en 2 horas por fallas mecánicas?

c) ¿Cuál es la probabilidad de que el intervalo entre dos reportes consecutivos por falla mecánica sea mayor a 3 horas?

d) Si inspeccionas e-scooters uno por uno cada inspección independiente con la misma probabilidad de detectar una falla mecánica, ¿cuál es la probabilidad de revisar exactamente 3 e-scooters en buen estado antes de encontrar el primer e-scooter con alguna falla mecánica?

e) Ahora se selecciona una muestra grande de 500 e-scooters para inspección. Usa la aproximación Normal para calcular la probabilidad de que entre 36 y 45 inclusive e-scooters tengan fallas mecánicas.

Para resolverlo, considera los siguientes valores de una Normal estandarizada. Usa el valor que más se aproxime:

- PZ < -2.162 = pnorm-2.162, mean=0, sd=1, lower.tail=TRUE = 0.0153
- PZ < -2.087 = pnorm-2.012, mean=0, sd=1, lower.tail=TRUE = 0.0184
- PZ < -0.745 = pnorm-0.745, mean=0, sd=1, lower.tail=TRUE = 0.2280
- PZ < -0.671 = pnorm-0.671, mean=0, sd=1, lower.tail=TRUE = 0.2512
- PZ < -0.311 = pnorm-0.311, mean=0, sd=1, lower.tail=TRUE = 0.3779
- PZ < -0.111 = pnorm0.111, mean=0, sd=1, lower.tail=TRUE = 0.4558

**MINICASO: Crisis viral en redes sociales de un candidato**

Durante una campaña electoral, un candidato enfrenta una crisis viral en redes sociales después de que se difunde un video polémico. El equipo de comunicación monitorea en tiempo real los comentarios publicados en X, Instagram y TikTok para estimar el impacto de la crisis y decidir su estrategia de respuesta.

A partir del análisis de datos históricos, se ha determinado que:

- El 20% de los comentarios publicados sobre el candidato son negativos.
- En promedio, se publican 24 comentarios negativos por día, y se supone que estos aparecen de manera independiente y aleatoria a lo largo del tiempo, a una tasa aproximadamente constante. Por lo tanto, para periodos cortos puede considerarse una tasa promedio de 1 comentario negativo por hora.
- El tiempo de respuesta del equipo de comunicación ante una publicación se encuentra entre 0 y 30 minutos, y cualquier tiempo dentro de este intervalo se considera igualmente probable.
- El porcentaje diario de sentimiento negativo, expresado en puntos porcentuales, puede modelarse aproximadamente mediante una distribución Normal con media de 45% y desviación estándar de 10 puntos porcentuales.
- En cada turno, uno de 6 community managers es seleccionado al azar para encargarse del monitoreo, y todos tienen la misma probabilidad de ser elegidos.

Supón, cuando corresponda, que los comentarios pueden considerarse independientes y que las condiciones descritas permanecen estables durante el periodo analizado.

En cada inciso:

1. Identifica y define la variable aleatoria.
2. Determina la distribución de probabilidad más adecuada e indica sus parámetros.

3. Justifica brevemente por qué esa distribución es apropiada.
4. Realiza el cálculo solicitado e interpreta brevemente el resultado en el contexto del problema.

Preguntas:

a) Se selecciona un solo comentario al azar y se desea registrar si es negativo o no. Define la variable aleatoria y especifica su distribución de probabilidad.

b) Se seleccionan 15 comentarios al azar. ¿Cuál es la probabilidad de que al menos 4 sean negativos?

c) Durante un intervalo de 1 hora, ¿cuál es la probabilidad de que se publiquen exactamente 3 comentarios negativos?

d) ¿Cuál es la probabilidad de que transcurran más de 45 minutos entre dos comentarios negativos consecutivos?

e) Los comentarios se revisan uno por uno. ¿Cuál es la probabilidad de que, antes de observar el primer comentario negativo, aparezcan exactamente 5 comentarios no negativos?

f) El tiempo de respuesta del equipo de comunicación ante una publicación puede tomar cualquier valor entre 0 y 30 minutos, con igual probabilidad. ¿Cuál es la probabilidad de que el equipo tarde más de 20 minutos en responder?

g) En cada turno se selecciona al azar a uno de 6 community managers, identificados con los números 1, 2, 3, 4, 5 y 6. Sea X el número que identifica al community manager seleccionado.

Determina la distribución de X y calcula: P(X=4), E(X), Var(X).

h) Sea Y el porcentaje diario de sentimiento negativo, medido en puntos porcentuales. Se sabe que: $Y \sim N(45,10^{2)}$
Calcula:

- La probabilidad de que el porcentaje diario de sentimiento negativo supere el 55%.
- La probabilidad de que se encuentre entre 35% y 50%.

i) En un día se publican 400 comentarios. Utiliza un modelo continuo como aproximación para estimar la probabilidad de que el número de comentarios negativos se encuentre entre 60 y 90, inclusive. Justifica la aproximación utilizada y realiza, si corresponde, la corrección por continuidad.

**Análisis y toma de decisiones**

j) ¿Qué riesgos estadísticos o estratégicos podrían surgir si las distribuciones elegidas no representan adecuadamente el comportamiento real de los comentarios en redes sociales?

k) A partir de los resultados obtenidos, menciona dos decisiones concretas de campaña que el equipo de comunicación podría tomar para gestionar la crisis.

l) ¿Qué evidencia empírica o análisis de datos serían necesarios para evaluar si las suposiciones utilizadas en los modelos anteriores son razonables?

## Anexo 1: Códigos básicos en R para Estadística descriptiva

```
# Supongamos que el dataframe se llama df
# y la variable de interés se llama variable

# Medidas de tendencia central
# media
meandf$variable, na.rm = TRUE

# mediana
mediandf$variable, na.rm = TRUE

#moda
as.numericnamessorttabledf$variable, decreasing = TRUE[1]


# Medidas de dispersión muestrales
# varianza
vardf$variable, na.rm = TRUE

# desv_std
sddf$variable, na.rm = TRUE

# rango
rangedf$variable, na.rm = TRUE

# desviación media con respecto a la mediana
meanabsdf$variable - mediandf$variable, na.rm = TRUE, na.rm = TRUE

#coeficiente de variación
desv_std / media


# Medidas de posición
# cuartil inferior
quantiledf$variable, 0.25, na.rm = TRUE

# cuartil superior
quantiledf$variable, 0.75, na.rm = TRUE

# amplitud intercuartílica
IQRdf$variable, na.rm = TRUE

# Percentil 10, 25, 50, 75 y 90
quantiledf$variable,
      probs = c0.10, 0.25, 0.50, 0.75, 0.90,
```

```
    na.rm = TRUE


# Medidas de asociación
# Covarianza
covarianza <- covdf_bi$variable_1, df_bi$variable_2

# Correlación de Pearson
cor_pearson <- cordf_bi$variable_1, df_bi$variable_2, method = "pearson"
```

# Anexo 2: Códigos en R para distribuciones de probabilidad

## Distribuciones Discretas

### BINOMIALn, p

PX = a dbinoma, size=n, prob=p
PX <= a pbinoma, size=n, prob=p, lower.tail=TRUE
PX < a pbinoma - 1, size=n, prob=p, lower.tail=TRUE
PX >= a pbinoma - 1, size=n, prob=p, lower.tail=FALSE
PX > a pbinoma, size=n, prob=p, lower.tail=FALSE
Pa <= X <= b pbinomb, size=n, prob=p, lower.tail=TRUE - pbinoma - 1, size=n, prob=p, lower.tail=TRUE
Pa < X <= b pbinomb, size=n, prob=p, lower.tail=TRUE - pbinoma, size=n, prob=p, lower.tail=TRUE
Pa <= X < b pbinomb - 1, size=n, prob=p, lower.tail=TRUE - pbinoma - 1, size=n, prob=p, lower.tail=TRUE
Pa < X < b pbinomb - 1, size=n, prob=p, lower.tail=TRUE - pbinoma, size=n, prob=p, lower.tail=TRUE

### POISSONlambda

PX = a dpoisa, lambda
PX <= a ppoisa, lambda, lower.tail=TRUE
PX < a ppoisa - 1, lambda, lower.tail=TRUE
PX >= a ppoisa - 1, lambda, lower.tail=FALSE
PX > a ppoisa, lambda, lower.tail=FALSE
Pa <= X <= b ppoisb, lambda, lower.tail=TRUE - ppoisa - 1, lambda, lower.tail=TRUE
Pa < X <= b ppoisb, lambda, lower.tail=TRUE - ppoisa, lambda, lower.tail=TRUE
Pa <= X < b ppoisb - 1, lambda, lower.tail=TRUE - ppoisa - 1, lambda, lower.tail=TRUE
Pa < X < b ppoisb - 1, lambda, lower.tail=TRUE - ppoisa, lambda, lower.tail=TRUE

### GEOMÉTRICAp

PX = a dgeoma - 1, prob=p
PX <= a pgeoma - 1, prob=p, lower.tail=TRUE
PX < a pgeoma - 2, prob=p, lower.tail=TRUE
PX >= a pgeoma - 2, prob=p, lower.tail=FALSE
PX > a pgeoma - 1, prob=p, lower.tail=FALSE
Pa <= X <= b pgeomb - 1, prob=p, lower.tail=TRUE - pgeoma - 2, prob=p, lower.tail=TRUE
Pa < X <= b pgeomb - 1, prob=p, lower.tail=TRUE - pgeoma - 1, prob=p, lower.tail=TRUE
Pa <= X < b pgeomb - 2, prob=p, lower.tail=TRUE - pgeoma - 2, prob=p, lower.tail=TRUE
Pa < X < b pgeomb - 2, prob=p, lower.tail=TRUE - pgeoma - 1, prob=p, lower.tail=TRUE

## Distribuciones Continuas

### UNIFORMEtheta1, theta2

PX <= a punifa, min=theta1, max=theta2, lower.tail=TRUE
PX < a punifa, min=theta1, max=theta2, lower.tail=TRUE
PX >= a punifa, min=theta1, max=theta2, lower.tail=FALSE
PX > a punifa, min=theta1, max=theta2, lower.tail=FALSE

Pa <= X <= b punifb, min=theta1, max=theta2, lower.tail=TRUE - punifa, min=theta1, max=theta2, lower.tail=TRUE

**EXPONENCIALtheta**

PX <= a pexpa, rate=1/theta, lower.tail=TRUE
PX < a pexpa, rate=1/theta, lower.tail=TRUE
PX >= a pexpa, rate=1/theta, lower.tail=FALSE
PX > a pexpa, rate=1/theta, lower.tail=FALSE
Pa <= X <= b pexpb, rate=1/theta, lower.tail=TRUE - pexpa, rate=1/theta, lower.tail=TRUE

**NORMALmu, sigma^2**

PX <= a pnorma, mean=mu, sd=sqrtsigma^2, lower.tail=TRUE
PX < a pnorma, mean=mu, sd=sqrtsigma^2, lower.tail=TRUE
PX >= a pnorma, mean=mu, sd=sqrtsigma^2, lower.tail=FALSE
PX > a pnorma, mean=mu, sd=sqrtsigma^2, lower.tail=FALSE
Pa <= X <= b pnormb, mean=mu, sd=sqrtsigma^2, lower.tail=TRUE - pnorma, mean=mu, sd=sqrtsigma^2, lower.tail=TRUE

**RESPUESTAS**

**IMPORTANTE:** En esta sección únicamente se incluye el resultado y/o una respuesta breve. No se incluye el análisis y/o la solución completa.

# Tema 1: Introducción

## Sección 1.1 Uso de datos en la solución de problemas y toma de decisiones

### 1.1.1 Confianza en instituciones públicas

¿Qué información adicional sería útil recopilar?

- Nivel educativo
- Ingreso económico
- Uso de redes sociales y medios de comunicación
- Participación política
- Experiencias previas con instituciones públicas
- Región geográfica de residencia
- Nivel de conocimiento sobre las instituciones

¿Pueden diseñarse estrategias para mejorar la confianza en los jóvenes?

Sí. Los datos permitirían identificar los factores asociados con la desconfianza y diseñar campañas de comunicación, programas de participación ciudadana o estrategias educativas dirigidas específicamente a los jóvenes.

### 1.1.2 Horas de estudio, sueño y calificaciones

¿Qué decisiones podría tomar el estudiante?

- Asignar más tiempo a las materias con menor desempeño
- Identificar si dormir más horas mejora las calificaciones
- Reducir actividades que consumen tiempo sin aportar al rendimiento académico
- Organizar un horario de estudio más equilibrado
- Detectar hábitos que favorecen mejores resultados

### 1.1.3 Tratados comerciales

¿Cómo podrían utilizarse estos datos?

- Identificar los países con mayor crecimiento en el comercio bilateral
- Detectar mercados con potencial de expansión
- Priorizar negociaciones con socios comerciales estratégicos

- Evaluar cuáles tratados han generado mayores beneficios económicos
- Dirigir recursos diplomáticos hacia relaciones comerciales más prometedoras

### 1.1.4 Incidentes diplomáticos

¿Cómo podrían utilizarse estos datos?

- Identificar patrones recurrentes de conflicto
- Detectar regiones con mayor riesgo diplomático
- Anticipar tensiones futuras entre determinados países
- Diseñar mecanismos de mediación y cooperación
- Fortalecer la vigilancia y prevención en áreas de mayor riesgo

### 1.1.5 Gastos familiares

¿Cómo podrían utilizar estos datos?

- Identificar las categorías con mayor gasto
- Detectar gastos innecesarios o impulsivos
- Establecer límites de gasto por categoría
- Elaborar un presupuesto mensual más eficiente
- Incrementar el ahorro mediante la reducción de gastos prescindibles

### 1.1.6 Ventas de productos

¿Qué decisiones pueden tomarse?

- Incrementar el inventario de productos con alta demanda
- Reducir o eliminar productos con bajas ventas
- Diseñar promociones para productos menos vendidos
- Ajustar la producción según la demanda observada
- Optimizar el espacio de almacenamiento y exhibición

### 1.1.7 Cancelación de contratos

¿Cómo podrían utilizarse estos datos?

- Identificar perfiles de clientes con mayor riesgo de cancelación
- Detectar problemas frecuentes en el servicio
- Diseñar programas de fidelización
- Mejorar la atención a clientes con múltiples quejas
- Implementar acciones preventivas para reducir la pérdida de clientes

### 1.1.8 Elección de ruta hacia la universidad

¿Cómo podría utilizar los datos?

- Comparar el tiempo promedio de traslado de cada ruta
- Evaluar los costos asociados a cada alternativa
- Analizar la frecuencia de retrasos
- Seleccionar la opción que ofrezca el mejor equilibrio entre costo, tiempo y puntualidad
- Elegir rutas alternativas para situaciones de tráfico intenso

### 1.1.9 Lanzamiento de una nueva línea de té saborizado

¿Qué datos debería recopilar?

Consumidores

- Preferencias de sabor
- Edad y perfil demográfico
- Frecuencia de consumo

Competencia

- Marcas existentes
- Precios
- Participación de mercado

Hábitos de compra

- Lugares de compra
- Frecuencia de adquisición
- Factores que influyen en la decisión de compra

Con esta información se podría estimar la demanda potencial y la viabilidad del producto.

### 1.1.10 Selección de software de gestión

¿Qué información debería analizarse?

- Productividad alcanzada con cada software
- Número de errores o fallas del sistema
- Tiempo requerido para realizar tareas
- Nivel de satisfacción de los usuarios

- Costos de implementación y mantenimiento
- Facilidad de uso y capacitación requerida

¿Cómo tomar la decisión?

Se debería seleccionar el software que proporcione mejores niveles de productividad y satisfacción, con menos errores y un costo razonable para la organización.

## 1.2 Concepto de variación y aleatoriedad. Relevancia de la calidad de los datos.

### 1.2.1 Lanzamiento de un dado

a) Aunque las condiciones del experimento sean las mismas, cada lanzamiento es un experimento aleatorio. Pequeñas diferencias en la fuerza, dirección o giro del dado producen resultados distintos que no pueden predecirse con certeza.

b) No. Aunque todos los números tienen la misma probabilidad de aparecer $1/6$, en 100 lanzamientos es poco probable que cada número aparezca exactamente el mismo número de veces. Sin embargo, se espera que las frecuencias observadas sean aproximadamente similares.

c) No. Si el dado estuviera cargado de manera que siempre cayera en 6, los resultados dejarían de reflejar la aleatoriedad propia de un dado equilibrado. En este caso existiría un sesgo causado por el instrumento utilizado.

### 1.2.2 Encuesta sobre confianza en el gobierno

a) Los resultados pueden diferir porque cada encuestador entrevista a personas distintas. Aun en zonas similares, las opiniones de los individuos varían naturalmente.

b) Corresponden principalmente a variación aleatoria, ya que las diferencias surgen por el proceso de selección de las personas encuestadas. También puede existir cierta variación natural debido a diferencias reales entre los ciudadanos.

c) No necesariamente. Si un encuestador obtiene sistemáticamente valores más altos que los demás, podría existir un sesgo de medición, un problema en la aplicación de la encuesta o un error en el registro de los datos.

### 1.2.3 Medición de la estatura

a) Pueden existir pequeñas diferencias debido a errores de medición, diferencias en la lectura de los instrumentos, postura del atleta o precisión limitada de los equipos.

b) Corresponden principalmente a variación aleatoria asociada al proceso de medición. La estatura real de la persona es prácticamente la misma.

c)

- Utilizar instrumentos calibrados
- Estandarizar el procedimiento de medición
- Capacitar a los observadores
- Realizar varias mediciones y utilizar el promedio
- Verificar periódicamente la precisión de los equipos

### 1.2.4 Tiros libres

a) Las diferencias observadas corresponden principalmente a variación natural, ya que los jugadores tienen distintos niveles de habilidad. También existe una componente aleatoria porque incluso un mismo jugador no siempre obtiene el mismo resultado.

b) Se podría aumentar el número de tiros realizados por cada jugador y mantener condiciones similares para todos. Esto reduce el impacto de la variación aleatoria y permite evaluar mejor la habilidad real.

c) Porque errores en el registro producirían conclusiones incorrectas sobre el desempeño de los jugadores y afectarían la validez del análisis.

### 1.2.5 Termómetros

a) Porque todos los termómetros están midiendo el mismo ambiente. La temperatura real del cuarto es prácticamente la misma para todos.

b) Sí. Las diferencias pueden deberse a errores de medición, falta de calibración, desgaste de los instrumentos o diferencias en su precisión.

c)

- Comparar las mediciones con un termómetro patrón
- Revisar la calibración de cada instrumento
- Analizar cuál presenta menor variabilidad en mediciones repetidas
- Verificar cuál produce resultados más cercanos al valor de referencia

### 1.2.6 Tiempos de respuesta en embajadas

a) Pueden existir diferencias debido a la cantidad de solicitudes recibidas, disponibilidad de personal, recursos tecnológicos, eficiencia administrativa o características locales de cada embajada.

b) Representan tanto variación natural como variación aleatoria.

- Variación natural: diferencias reales entre embajadas.
- Variación aleatoria: fluctuaciones diarias en la carga de trabajo o circunstancias particulares.

c) Se podrían revisar:

- Valores atípicos o inusuales
- Registros faltantes
- Consistencia entre fechas y tiempos reportados
- Promedios y variabilidad de los tiempos
- Cambios abruptos en los indicadores
- Diferencias excesivas respecto a otras embajadas similares

La presencia de anomalías podría indicar errores de captura, problemas de medición o inconsistencias en los procedimientos de registro.

## 1.3 Entendimiento de la variación empleando la probabilidad y la estadística

### 1.3.1 Lanzamiento de un dado

La probabilidad teórica de obtener un 6 es:

$$P(6) = \frac{1}{6} \approx 0.1667$$

La proporción observada es:

$$\hat{p} = \frac{8}{60} = 0.1333$$

a) Sí. La proporción observada es: $0.1333$, mientras que la probabilidad teórica es $0.1667$.

La diferencia es $0.1667 - 0.1333 = 0.0334$

Por lo tanto, la proporción observada es menor que la probabilidad teórica.

b) Porque en cualquier experimento aleatorio es normal que los resultados observados difieran del valor esperado teórico. Una diferencia moderada puede ocurrir simplemente por azar, especialmente cuando el número de observaciones es pequeño.

c) Porque cada lanzamiento es independiente y aleatorio. Aunque la probabilidad de obtener un 6 es constante, el número exacto de veces que aparece en una muestra finita fluctúa alrededor del valor esperado debido al azar.

d) Se esperaría que estuviera más cerca. De acuerdo con la Ley de los Grandes Números, al aumentar el número de lanzamientos, la frecuencia relativa tiende a aproximarse a la probabilidad teórica.

Por ello, con 600 lanzamientos la proporción observada probablemente sería más cercana a 0.1667 que con solo 60 lanzamientos.

### 1.3.2 Rendimiento de combustible

a) Porque existen múltiples factores que afectan el consumo de combustible:

- Diferencias en la conducción
- Condiciones del tráfico
- Condiciones climáticas
- Estado de las llantas
- Calidad del combustible
- Peso transportado

b) Clasifica las siguientes causas.

| Causa | Tipo de variación |
|---|---|
| Peso del conductor | Sistemática |
| Estilo de manejo | Sistemática |
| Viento | Aleatoria |
| Calidad del combustible | Principalmente sistemática |

c) Porque las condiciones reales de operación nunca son idénticas para todos los vehículos. Incluso pequeñas diferencias producen variación en el consumo observado.

d) Indica que una sola observación no describe completamente un fenómeno. Para comprender adecuadamente el comportamiento de los datos es necesario analizar tanto el promedio como la variabilidad observada.

### 1.3.3 Encuesta electoral

a) Porque cada encuesta selecciona una muestra diferente de votantes. Las personas entrevistadas pueden tener opiniones distintas, generando variaciones naturales entre muestras.

b) No. Las diferencias pueden surgir simplemente por variación muestral, que es una consecuencia natural de trabajar con muestras en lugar de toda la población.

c) El margen de error indica cuánto podría diferir el resultado de la encuesta respecto al valor real de la población debido al azar del muestreo.

En este caso: $52\% \pm 3\%$ sugiere que el apoyo real podría encontrarse aproximadamente entre: $49\%$ y $55\%$

d)

- La muestra debe ser representativa
- La selección de los participantes debe ser aleatoria
- El tamaño de muestra debe ser suficiente
- Las preguntas deben formularse de manera neutral
- No deben existir errores sistemáticos de medición o selección

e) Porque proporciona métodos para cuantificar la incertidumbre y evaluar qué tan confiables son las conclusiones obtenidas a partir de una muestra.

Aunque no ofrece certeza absoluta, permite tomar decisiones informadas basadas en evidencia y medir el grado de confianza asociado a las estimaciones.

## Sección 1.4 Poblaciones y muestras aleatorias

### 1.4.1 Opinión sobre una política pública

a) La población está formada por todos los ciudadanos sobre los cuales se desea conocer la opinión respecto a la nueva política pública.

b) La muestra está constituida por las 1,000 personas seleccionadas aleatoriamente y encuestadas.

c) Porque una muestra adecuadamente seleccionada puede reflejar las características de la población. Esto permite obtener estimaciones confiables con menor costo, tiempo y esfuerzo que un censo completo.

d) Porque la muestra representa solo una parte de la población. Las personas seleccionadas pueden no reflejar exactamente las opiniones de todos los ciudadanos, generando diferencias debidas al azar muestral.

e)

- Utilizar métodos de selección aleatoria
- Incrementar el tamaño de la muestra
- Asegurar la participación de distintos grupos demográficos
- Reducir la no respuesta
- Aplicar procedimientos de muestreo adecuados por ejemplo, estratificado

### 1.4.2 Uso de redes sociales

a) La población está formada por todos los clientes de la empresa.

b) La muestra está constituida por los 500 usuarios seleccionados aleatoriamente.

c) No. El valor de 3.5 horas representa un promedio. Algunas personas pueden utilizar redes sociales durante más tiempo y otras durante menos tiempo.

d) No. Cada muestra contiene individuos diferentes, por lo que es normal que los promedios varíen ligeramente debido a la variabilidad muestral.

### 1.4.3 Opinión sobre transporte eléctrico

a) La población está formada por todos los ciudadanos cuya opinión se desea conocer sobre el uso del transporte eléctrico.

b) La muestra está constituida por las 2,000 personas que respondieron la encuesta en línea.

c) No. El 70% es una estimación obtenida a partir de una muestra. El porcentaje real en toda la población puede ser ligeramente diferente debido al error muestral.

d)

- Exclusión de personas sin acceso a internet
- Menor participación de adultos mayores o grupos con baja conectividad
- Autoselección de los participantes
- Diferencias en el uso de plataformas digitales entre distintos sectores de la población

e) Porque las conclusiones obtenidas de la muestra se utilizan para inferir características de toda la población. Si la muestra no es representativa, las estimaciones pueden estar sesgadas y conducir a decisiones incorrectas.

# TEMA 2: Análisis exploratorio de datos

## 2.1 Tipos de datos, clasificación y escalas de medición

### 2.1.1 Clasificación por tipo de dato y escala de medición

| Variable | Tipo de dato | Escala de medición |
|---|---|---|
| Tiempo diario de uso de TikTok minutos | Cuantitativa continua | Razón |
| Nivel de batería del celular % | Cuantitativa continua | Razón |
| Tipo de chatbot utilizado | Cualitativa | Nominal |
| Calidad del sueño mala, regular, buena, excelente | Cualitativa | Ordinal |
| Número de podcasts escuchados por semana | Cuantitativa discreta | Razón |
| Temperatura corporal °C | Cuantitativa continua | Intervalo |
| Categoría del vehículo de Uber | Cualitativa | Nominal |
| Velocidad de conexión a Internet Mbps | Cuantitativa continua | Razón |
| Nivel de estrés bajo, medio, alto | Cualitativa | Ordinal |
| Emisiones de $CO_2$ de un viaje kg | Cuantitativa continua | Razón |
| Horas de estudio en plataformas online por mes | Cuantitativa continua | Razón |
| Género musical favorito en Spotify | Cualitativa | Nominal |
| Número de pasos registrados por día | Cuantitativa discreta | Razón |
| Puntaje de riesgo crediticio 0–850 | Cuantitativa discreta | Intervalo |
| Nivel de suscripción a Netflix | Cualitativa | Ordinal |
| Distancia recorrida en bicicleta eléctrica km | Cuantitativa continua | Razón |
| Frecuencia cardiaca en reposo latidos/min | Cuantitativa discreta | Razón |
| Categoría de producto comprado en Amazon | Cualitativa | Nominal |
| Intensidad de ruido ambiental dB | Cuantitativa continua | Intervalo |
| Calificación de una app 1–5 estrellas | Cualitativa | Ordinal |
| Porcentaje de avance en un curso online | Cuantitativa continua | Razón |
| Cantidad de correos electrónicos recibidos al día | Cuantitativa discreta | Razón |
| Tiempo de carga de una batería de auto eléctrico horas | Cuantitativa continua | Razón |
| Nivel de satisfacción con un servicio | Cualitativa | Ordinal |
| Número de seguidores nuevos en Instagram por semana | Cuantitativa discreta | Razón |
| Kilogramos de residuos reciclados por mes | Cuantitativa continua | Razón |
| pH del agua potable | Cuantitativa continua | Intervalo |
| Rol del usuario en una plataforma | Cualitativa | Nominal |
| Porcentaje de batería consumido por apps de IA | Cuantitativa continua | Razón |
| Consumo energético del hogar kWh/mes | Cuantitativa continua | Razón |
| Duración de videollamadas en Zoom min | Cuantitativa continua | Razón |
| Nivel de severidad de errores en software | Cualitativa | Ordinal |

| Variable | Tipo de dato | Escala de medición |
|---|---|---|
| Cantidad de "likes" recibidos | Cuantitativa discreta | Razón |
| Latencia de un servidor ms | Cuantitativa continua | Razón |
| Cantidad de vidrio/plástico en envases comprados g | Cuantitativa continua | Razón |
| Categoría de riesgo sísmico | Cualitativa | Ordinal |
| Porcentaje de uso de CPU | Cuantitativa continua | Razón |
| Número de entregas realizadas por un repartidor | Cuantitativa discreta | Razón |
| Gasto mensual en servicios de streaming MXN | Cuantitativa continua | Razón |
| Tiempo de espera para recibir un pedido min | Cuantitativa continua | Razón |

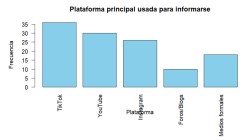
### 2.1.2 Clasificación por tipo de dato y escala de medición

| Variable | Tipo de dato | Escala de medición |
|---|---|---|
| Partido político con el que se identifica un votante | Cualitativa | Nominal |
| Nivel de confianza en las instituciones 1 = nada, 5 = totalmente | Cualitativa | Ordinal |
| Edad del candidato | Cuantitativa continua | Razón |
| Número de diputados en el Congreso por partido | Cuantitativa discreta | Razón |
| Año de nacimiento de los senadores | Cuantitativa discreta | Intervalo |
| Porcentaje de participación electoral en una elección | Cuantitativa continua | Razón |
| País de origen del embajador | Cualitativa | Nominal |
| Índice de Desarrollo Humano IDH | Cuantitativa continua | Intervalo |
| Número de tratados internacionales firmados por un país | Cuantitativa discreta | Razón |
| Rango de poder militar alto, medio, bajo | Cualitativa | Ordinal |
| Número de conflictos fronterizos activos en un año | Cuantitativa discreta | Razón |
| Área funcional en la que trabaja un empleado | Cualitativa | Nominal |
| Satisfacción laboral escala Likert de 1 a 7 | Cualitativa | Ordinal |
| Año de creación de las empresas | Cuantitativa discreta | Intervalo |
| Salario mensual en pesos | Cuantitativa continua | Razón |
| Número de ventas cerradas en un mes | Cuantitativa discreta | Razón |
| Años de antigüedad en la empresa | Cuantitativa continua | Razón |
| Puntaje de empleados en un test de liderazgo | Cuantitativa discreta | Intervalo |
| Nivel de satisfacción del cliente 1 a 10 | Cualitativa | Ordinal |
| Ingreso mensual de una empresa en pesos | Cuantitativa continua | Razón |
| Número de facturas emitidas por semana | Cuantitativa discreta | Razón |

| Variable | Tipo de dato | Escala de medición |
|---|---|---|
| Tipo de auditoría realizada interna, externa, fiscal | Cualitativa | Nominal |
| Método de depreciación utilizado por la empresa | Cualitativa | Nominal |
| Monto de cuentas por cobrar vencidas en pesos | Cuantitativa continua | Razón |
| Nivel de riesgo financiero asignado a una empresa bajo, medio, alto | Cualitativa | Ordinal |
| Tiempo promedio de pago a proveedores en días | Cuantitativa continua | Razón |
| Número de empleados del departamento contable | Cuantitativa discreta | Razón |

## 2.2 Distribuciones de frecuencias para datos cualitativos y cuantitativos: diagramas de barras, diagramas circulares, histogramas, polígono de frecuencias y ojiva

### 2.2.1

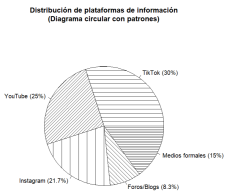



a) La diferencia es $36 - 10 = 26$ estudiantes, que equivale al $\frac{26}{120} \times 100 = 21.7\%$ de la muestra.

b) Sea $x$ el número de estudiantes que cambian:

$$36 - x = 10 + x$$
$$x = 13$$

c) Foros/Blogs especializados, porque es la plataforma con menor frecuencia 10 estudiantes. Esto sugiere que existe un mayor margen para aumentar su alcance mediante estrategias de promoción o contenido atractivo.




a) Se elegirían TikTok y YouTube, ya que son las dos plataformas con mayor número de usuarios. En conjunto alcanzan $36 + 30 = 66$ estudiantes, lo que representa 55%.

b)

$$\frac{30+26+10+18}{120}\times 100 = \frac{84}{120}\times 100 = 70\%$$

La universidad podría alcanzar al 70% de los estudiantes utilizando las demás plataformas.

c) No. Al utilizar únicamente medios digitales formales, la universidad estaría llegando solo al 15% de los estudiantes, por lo que sería recomendable complementar la comunicación con plataformas más populares como TikTok, X/YouTube e Instagram para lograr una mayor cobertura.

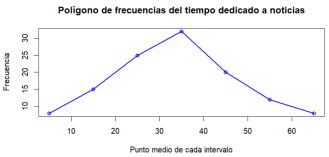

a) La mayor concentración de estudiantes dedica entre 30 y 40 minutos diarios al consumo de noticias.

b) La distribución parece aproximadamente simétrica. Esto sugiere que la mayoría de los estudiantes dedica una cantidad moderada de tiempo al consumo de noticias, mientras que pocos dedican muy poco o mucho tiempo.

c) El cambio más brusco se observa al pasar de 30–40 a 40–50 minutos, donde la frecuencia disminuye en 12 estudiantes.

d) La variable es continua. Se utiliza un histograma porque los datos están agrupados en intervalos numéricos continuos y las barras deben aparecer unidas para representar la continuidad de la variable. Una gráfica de barras se utiliza principalmente para variables cualitativas o cuantitativas discretas.

e) Sí, la interpretación podría cambiar porque los datos quedarían más resumidos. Se observaría con menor precisión la concentración de estudiantes en determinados rangos de tiempo, aunque la forma general de la distribución probablemente seguiría siendo similar.



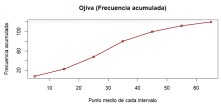
a) El punto donde se alcanza el máximo es en el intervalo de 30 a 40 minutos, que es el intervalo más frecuente; es decir, la mayor cantidad de estudiantes dedica entre 30 y 40 minutos diarios a consumir noticias.

b) Tiene una forma unimodal y aproximadamente simétrica, lo que sugiere que la mayoría de los estudiantes se concentra alrededor de un tiempo de consumo intermedio.

c) La frecuencia aumenta más rápidamente entre los intervalos 10–20 y 20–30 minutos.

d) El área bajo el polígono representa de manera aproximada la distribución total de las frecuencias, es decir, los 120 estudiantes incluidos en el estudio.

e) Un polígono con múltiples picos sugeriría una distribución multimodal, lo que podría indicar que existen diferentes grupos de estudiantes con hábitos de consumo de noticias distintos.




a) Aproximadamente 65 personas.

b) Aproximadamente 75 minutos: $P_{75}$=112.5. La frecuencia acumulada 112.5 se alcanza entre 70 y 90 minutos.

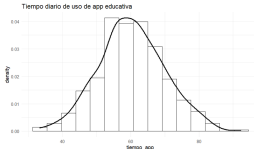
c) Entre 50 y 70 minutos, es donde la frecuencia acumulada aumenta más rápidamente.

d) Aproximadamente 20 personas: A los 90 minutos hay aproximadamente 130 personas acumuladas. 150−130=20.

e) Sí. Los posibles valores atípicos suelen identificarse en los extremos de la distribución, donde la ojiva se aplana y hay pocos datos acumulados. Sin embargo, para confirmarlos es preferible utilizar cuartiles y el rango intercuartílico.

## 2.3 Curvas de distribuciones de frecuencias formas características

### 2.3.1



a) Distribución aproximadamente simétrica alrededor de la media.

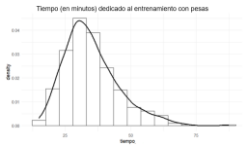
b) Sí, probablemente son muy cercanas. Debido a que la curva es casi simétrica.

c) No se observan valores extremos evidentes; las colas disminuyen de forma gradual.

d) Sugiere una sola población, ya que presenta un único pico distribución unimodal.

e) Aproximadamente entre 55 y 65 minutos de uso diario.

**2.3.2**




a) El sesgo a la derecha significa que algunos usuarios dedican mucho más tiempo al entrenamiento con pesas que la mayoría, generando valores altos.

b) Aproximadamente entre 25 y 40 minutos por sesión.

c) Sí. Los pocos usuarios que entrenan durante mucho tiempo elevan la media, haciendo que sea mayor que la mediana.

d) La programación de horarios, la disponibilidad de equipos y el diseño de planes de entrenamiento para distintos tipos de usuarios

e) Sí. Se observan algunos usuarios que dedican más de 70 u 80 minutos al entrenamiento, lo que podría considerarse un valor atípico alto.

**2.3.3**


a) Sí. La mayoría de las calificaciones se concentra entre aproximadamente 70 y 85 puntos, lo que indica un desempeño generalmente alto.

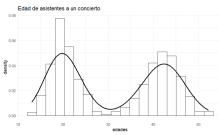
b) Significa que existe un pequeño grupo de empleados con calificaciones considerablemente más bajas que la mayoría.

c) Sí. La presencia de algunas calificaciones bajas reduce la media, por lo que la mediana resulta mayor. Esto indica que el empleado "típico" tiene un desempeño mejor que el sugerido por el promedio.

d) La mediana sería más representativa del desempeño típico de los empleados, porque es menos sensible a las pocas calificaciones muy bajas que aparecen en la cola izquierda de la distribución. Por ello, refleja mejor el desempeño de la mayoría de los empleados que la media, la cual puede verse afectada por esos valores extremos.

e) Dependiendo de cómo se modifique el índice, la distribución podría cambiar, afectando medidas como la media, la mediana y la dispersión de las calificaciones, así como la interpretación del desempeño general de los empleados

**2.3.4**




a) Los dos picos principales se observan aproximadamente en 20 años y 43 años. Representan las edades más frecuentes de dos grupos de asistentes al concierto.

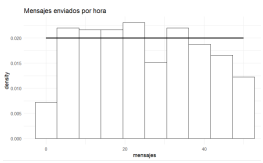
b) El grupo de jóvenes parece ser ligeramente más numeroso, ya que el primer pico alrededor de 20 años es más alto que el segundo.

c) El grupo de adultos presenta una mayor variabilidad, porque la curva alrededor del segundo pico está más extendida y abarca un rango de edades más amplio.

d) Los grupos están claramente diferenciados, ya que los picos se encuentran separados por aproximadamente 20 años y existe una zona de baja frecuencia entre ellos alrededor de los 30 años.

e) Sí. La distribución es bimodal, lo que sugiere que el concierto atrae a dos segmentos de público distintos: uno de jóvenes y otro de adultos.

**2.3.5**




a) No. Ningún intervalo concentra claramente más observaciones que los demás; las barras tienen alturas similares a lo largo de toda la distribución.

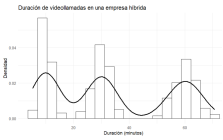
b) Indica que los mensajes se envían de manera relativamente constante durante el horario laboral, sin periodos de actividad especialmente alta o baja.

c) Probablemente corresponden a variación aleatoria. Las diferencias entre las barras son pequeñas y no muestran un patrón consistente.

d) El rango observado es aproximadamente de 0 a 55 mensajes por hora. Esto indica que la actividad mínima fue cercana a 0 mensajes y la máxima cercana a 55 mensajes por hora.

e) La distribución no muestra picos, sesgos ni concentraciones marcadas. En general, sugiere que los mensajes se envían de forma relativamente uniforme a lo largo del rango de valores observado.

**2.3.6**




a) Se observan tres picos principales alrededor de 10, 30 y 60 minutos. Estos picos representan los tres tipos de reuniones más frecuentes: cortas, medianas y largas.

b) Las reuniones cortas ≈10 minutos parecen ser las más frecuentes, ya que presentan el pico más alto de la distribución.

c) Las reuniones largas ≈60 minutos parecen mostrar mayor variabilidad, porque la curva alrededor de ese pico es más ancha y está más dispersa.

d) Los grupos están relativamente bien separados, aunque existe un ligero traslape entre las reuniones cortas y medianas, y entre las medianas y largas. Esto podría deberse a diferencias en los objetivos, participantes o duración real de las reuniones.

e) La distribución sugiere que la empresa utiliza distintos tipos de reuniones según sus necesidades. Esto puede favorecer la organización, pero también indica que una parte importante del tiempo laboral se dedica a reuniones largas.

f) La empresa debería revisar primero las reuniones largas, ya que consumen más tiempo y presentan una mayor variabilidad en duración. Reducirlas o hacerlas más eficientes podría generar un mayor impacto en la productividad.

**2.3.7**


a) La distribución muestra una variabilidad moderada. Los tiempos de respuesta se distribuyen en un rango relativamente amplio aproximadamente de 4 a 20 segundos, pero sin valores extremos muy alejados del resto.

b) Los tiempos de respuesta tienden a concentrarse alrededor de 11 a 13 segundos, donde se observa el punto más alto de la curva. Esto indica la existencia de un tiempo de respuesta típico.

c) Una alta variabilidad podría generar experiencias inconsistentes de confort para los usuarios del edificio. Los administradores podrían revisar la configuración o calibración del sistema para reducir las diferencias en los tiempos de respuesta.

d) Sí. La distribución es aproximadamente simétrica y unimodal, por lo que la media representa adecuadamente el tiempo de respuesta típico del sistema.

e) La distribución sugiere que el sistema es relativamente estable y predecible, ya que la mayoría de los tiempos de respuesta se concentran alrededor de un valor central y no se observan valores atípicos importantes ni una dispersión excesiva.

**2.3.8**


a) La distribución indica que la mayoría de los usuarios abandona la videollamada rápidamente, ya que las frecuencias más altas se concentran en los primeros segundos.

b) La cola hacia la derecha indica que existe un grupo pequeño de usuarios que tolera la mala conexión durante más tiempo antes de abandonar la llamada.

c) Los tiempos de abandono se concentran principalmente en los primeros segundos y no se distribuyen de manera uniforme. La frecuencia disminuye conforme aumenta el tiempo de permanencia.

d) La aplicación debería enfocarse en detectar y corregir problemas de conexión rápidamente, ya que la mayoría de los usuarios abandona la llamada en poco tiempo. Una posible acción sería mejorar la estabilidad de la conexión durante los primeros segundos de la videollamada.

e) Sí. La larga cola derecha muestra que algunos usuarios permanecen conectados durante 20, 30 o más segundos, lo que sugiere una tolerancia inusualmente alta a los problemas de conexión en comparación con la mayoría de los usuarios.

## 2.4 Datos no agrupados

### 2.4.1

En R:

```
# ==========================================
# Datos
# ==========================================

tiktok <- c(
  85, 92, 74, 60, 120, 110, 95, 88, 102, 76,
  65, 70, 115, 130, 98, 90, 85, 72, 100, 108
)

# ==========================================
# Tabla de frecuencias
# ==========================================

tabla <- table(tiktok)
modas <- names(tabla)[tabla == max(tabla)]

# ==========================================
# Resultados
# ==========================================

resultado <- data.frame(
  Media = mean(tiktok),
  Mediana = median(tiktok),
  Moda = as.numeric(modas),
  Rango = diff(range(tiktok)),
  IQR = IQR(tiktok),
  Varianza = var(tiktok),
  Desv_Estandar = sd(tiktok),
```

```
  MAD = mad(tiktok, constant = 1),
  CV = 100 * sd(tiktok) / mean(tiktok),
  P20 = quantile(tiktok, 0.20),
  Q3 = quantile(tiktok, 0.75),
  D2 = quantile(tiktok, 0.20)
)

resultado
```

```
> resultado
  Media  Mediana  Moda  Rango  IQR  Varianza  Desv_Estandar
  91.75  91       85    70     28   362.0921  19.02872

  MAD  CV        P20  Q3     D2
  16   20.73975  73.6 103.5  73.6
```

a) La media y la mediana son mayores a 60 minutos, por lo que el tiempo típico de uso supera el umbral asociado con posibles niveles elevados de ansiedad.

b) Un aumento de la mediana indicaría que el adolescente típico dedica más tiempo a TikTok que antes, sugiriendo un incremento general en el uso de la aplicación.

c) Si no existe una moda claramente definida, significa que no hay un tiempo de uso que se repita con suficiente frecuencia como para considerarse predominante en el grupo.

d) El rango muestra la diferencia entre el menor y el mayor tiempo de uso. Un rango amplio indica una gran variabilidad entre los adolescentes.

e) Un IQR pequeño indicaría que la mayoría de los adolescentes presenta tiempos de uso relativamente similares y concentrados alrededor de la mediana.

f) Un aumento de la desviación estándar sugeriría que los hábitos de uso se han vuelto más heterogéneos, con diferencias más marcadas entre los adolescentes.

g) Un coeficiente de variación alto indica que existe una gran variabilidad relativa en el tiempo de uso respecto al promedio del grupo.

h) Una MAD pequeña indica que los tiempos de uso suelen estar cerca de la media, por lo que los adolescentes presentan comportamientos relativamente similares.

i)

- Percentil 20 $P_{20}$: Aproximadamente el 20% de los adolescentes utiliza TikTok un tiempo igual o menor que este valor.
- Tercer cuartil $Q_3$: Aproximadamente el 75% de los adolescentes utiliza TikTok un tiempo igual o menor que este valor, y el 25% restante lo supera.
- Segundo decil $D_2$: Aproximadamente el 20% de los adolescentes utiliza TikTok un tiempo igual o menor que este valor. $D_2$ coincide con el percentil 20.

**2.4.2** En R:

```
# ==========================================
# Datos
# ==========================================
ia <- c(45, 60, 30, 75, 120, 95, 80, 55, 90, 40,
        65, 70, 110, 130, 85, 50, 60, 35, 100, 105)

# ==========================================
# Tabla de frecuencias
# ==========================================
tabla <- table(ia)
modas <- names(tabla)[tabla == max(tabla)]

# ==========================================
# Resultados
# ==========================================
resultado <- data.frame(
  Media = round(mean(ia), 2),
  Mediana = round(median(ia), 2),
  Moda = paste(modas, collapse = ", "),
  Rango = round(diff(range(ia)), 2),
  IQR = round(IQR(ia), 2),
  Varianza = round(var(ia), 2),
  Desv_Estandar = round(sd(ia), 2),
  MAD = round(mad(ia, constant = 1), 2),
  CV = round(100 * sd(ia) / mean(ia), 2),
  P25 = round(quantile(ia, 0.25), 2),
  Q3 = round(quantile(ia, 0.75), 2),
```

```
  D8 = round(quantile(ia, 0.80), 2)
)

# ==========================================
# Mostrar resultados
# ==========================================
print(resultado)
resultado
```

```
> resultado
  Media Mediana Moda Rango IQR Varianza Desv_Estandar
1    75    72.5   60   100  42   836.84        28.93

   MAD   CV   P25    Q3  D8
1 22.5 38.57 53.75 96.25 101
```

a) 6 de los 20 estudiantes 30% utilizan asistentes de IA más de 90 minutos al día. Esto sugiere que una proporción importante del grupo presenta un uso intensivo de estas herramientas.

b) Si la media es mayor que la mediana, la distribución presenta sesgo a la derecha. Esto indica que algunos estudiantes utilizan la IA durante tiempos considerablemente mayores que el resto.

c) La moda es 60 minutos. Esto significa que, dentro de la muestra, 60 minutos es el tiempo diario de uso observado con mayor frecuencia. Sin embargo, aparece únicamente en 2 de los 20 estudiantes, por lo que no representa una concentración muy marcada de los datos.

d) El rango muestra la diferencia entre el menor y el mayor tiempo de uso observado. Un rango amplio indica diferencias importantes en los hábitos de uso entre los estudiantes.

e) Un IQR pequeño indicaría que el 50% central de los estudiantes tiene tiempos de uso relativamente similares y concentrados alrededor de la mediana.

f) Un aumento de la desviación estándar sugeriría que los patrones de uso se han vuelto más heterogéneos, con diferencias más marcadas entre los estudiantes.

g) Un coeficiente de variación elevado indica una alta variabilidad relativa en los tiempos de uso respecto al promedio, es decir, comportamientos poco homogéneos.

h) Una MAD pequeña indicaría que los tiempos de uso suelen estar cerca de la media, reflejando hábitos de uso relativamente similares entre los estudiantes.

i)

- Percentil 25 $P_{25}$: El 25% de los estudiantes utiliza la IA un tiempo igual o menor a este valor.
- Tercer cuartil $Q_3$: El 75% de los estudiantes utiliza la IA un tiempo igual o menor a este valor; el 25% restante la utiliza más.
- Decil 8 $D_8$: El 80% de los estudiantes utiliza la IA un tiempo igual o menor a este valor; el 20% restante corresponde a los usuarios más intensivos.

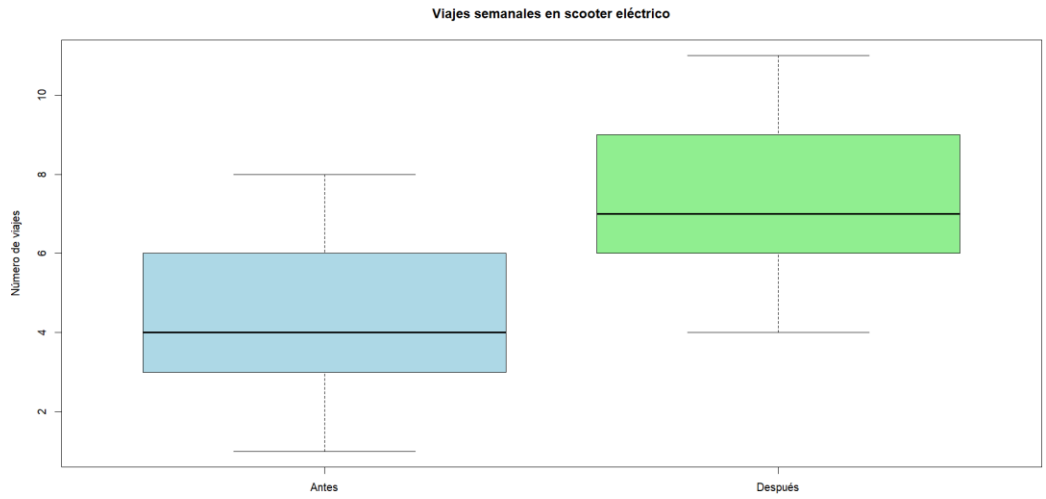
Estos cuantiles permiten segmentar a los estudiantes en niveles de uso bajo, medio y alto para diseñar estrategias académicas diferenciadas.

**2.4.3**

```
# ==========================================
# Datos
# ==========================================

antes <- c(
  2, 3, 1, 4, 6, 5, 4, 3, 5, 2,
  4, 5, 7, 8, 6, 3, 4, 2, 7, 6
)

despues <- c(
  5, 6, 4, 7, 9, 8, 7, 6, 8, 5,
  7, 8, 10, 11, 9, 6, 7, 5, 10, 9
)

# ==========================================
# Función para calcular estadísticas
# ==========================================

estadisticas <- function(x){

  tabla <- table(x)
  modas <- as.numeric(names(tabla)[tabla == max(tabla)])

  data.frame(
    Media = mean(x),
    Mediana = median(x),
    Moda = paste(modas, collapse = ", "),
    Rango = diff(range(x)),
    IQR = IQR(x),
    Varianza = var(x),
    Desv_Estandar = sd(x),
```

```
    MAD = mad(x, constant = 1),
    CV = 100 * sd(x) / mean(x),
    P25 = quantile(x, 0.25),
    P50 = quantile(x, 0.50),
    P75 = quantile(x, 0.75)
  )
}

# ============================================
# Resultados
# ============================================

estadisticas(antes)
estadisticas(despues)

# ============================================
# Boxplots comparativos
# ============================================

boxplot(
  antes, despues,
  names = c("Antes", "Después"),
  col = c("lightblue", "lightgreen"),
  main = "Viajes semanales en scooter eléctrico",
  ylab = "Número de viajes"
)
```

```
> # Resultados
> estadisticasantes
   Media   Mediana   Moda   Rango   IQR    Varianza    Desv_Estandar
   4.35    4         4      7       3      3.713158    1.926956 x
   MAD     CV          P25    P50    P75
   1.5     44.29783    3      4      6


> estadisticasdespues
   Media   Mediana   Moda   Rango   IQR   Varianza   Desv_Estandar
   7.35    7         7      7       3     3.713158   1.926956
   MAD     CV          P25    P50    P75
   1.5     26.21708    6      7      9
```




a) Sí. La media, la mediana y la moda aumentan después de la ampliación de ciclovías, lo que indica un mayor uso de scooters eléctricos.

b) La variabilidad permanece prácticamente igual. Las medidas de dispersión son muy similares en ambos periodos.

c) Ninguno de los periodos presenta una diversidad claramente mayor. Los hábitos de uso muestran una dispersión similar antes y después de la política.

d) La distribución conserva prácticamente la misma forma, pero se desplaza hacia valores más altos. Los percentiles y la mediana aumentan aproximadamente en la misma magnitud.

e) Ningún grupo incrementó más su uso que otro. Todos los estudiantes aumentaron el mismo número de viajes semanales.

f) Los resultados sugieren que la ampliación de ciclovías estuvo asociada con un aumento en el uso de scooters eléctricos. La autoridad podría complementar esta medida con infraestructura segura, estacionamientos para scooters y campañas de movilidad sustentable.

g) La muestra es pequeña, no existe un grupo de control y pueden existir factores externos no considerados que influyan en los resultados.

h) El clima, el costo del transporte público, la disponibilidad de scooters, la distancia recorrida, cambios en los hábitos de movilidad o campañas de promoción podrían haber influido en el aumento observado.

**2.4.4** Como los datos son poblacionales, debes utilizar en R:

- Varianza poblacional: varx*n-1/n
- Desviación estándar poblacional: sqrtvarianza_poblacional

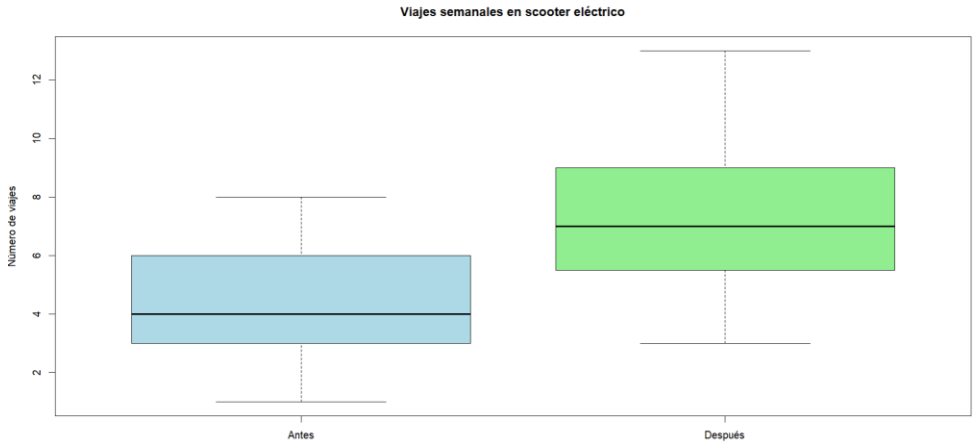
- MAD poblacional: promedio de las desviaciones absolutas respecto a la media.
- Coeficiente de variación poblacional: 100*sd_pob/media

En R:

```
# ==========================================
# Función para estadísticas poblacionales
# ==========================================

estadisticas_pob <- function(x){

  n <- length(x)

  tabla <- table(x)
  modas <- names(tabla)[tabla == max(tabla)]

  var_pob <- var(x) * (n - 1) / n

  sd_pob <- sqrt(var_pob)

  mad_pob <- mean(abs(x - mean(x)))

  cv_pob <- 100 * sd_pob / mean(x)

  data.frame(

    Media = mean(x),

    Mediana = median(x),

    Moda = paste(modas, collapse=", "),

    Rango = diff(range(x)),

    IQR = IQR(x),

    Varianza_Poblacional = var_pob,

    Desv_Estandar_Poblacional = sd_pob,

    MAD_Poblacional = mad_pob,

    CV_Poblacional = cv_pob,

    P25 = unname(quantile(x,0.25)),

    P50 = unname(quantile(x,0.50)),
```

```
    P75 = unname(quantile(x,0.75))

  )

 }
```

```
> estadisticas_pobantes
   Media   Mediana   Moda   Rango   IQR   Varianza_Poblacional
   4.35    4         4      7       3     3.5275
   Desv_Estandar_Poblacional   MAD_Poblacional   CV_Poblacional
   1.878164                    1.585             43.17618
   P25   P50   P75
   3     4     6
>
> estadisticas_pobdespues
   Media   Mediana   Moda    Rango  IQR    Varianza_Poblacional
   7.4     7         6,7,8   10     3.25   6.94
   Desv_Estandar_Poblacional  MAD_Poblacional    CV_Poblacional
   2.634388                   2.14               35.59984
   P25    P50   P75
   5.75   7     9
```




a) La distribución presenta sesgo positivo hacia la derecha debido a la presencia de un valor extremadamente alto 178 µg/m³. La mayoría de los días se concentra entre 30 y 70 µg/m³.

b) Sí. El valor de 178 µg/m³ aparece como un posible valor atípico. Podría corresponder a un episodio extraordinario de contaminación atmosférica o a condiciones meteorológicas desfavorables.

c) Mostraron una variabilidad importante, como lo evidencian el amplio rango y la presencia de valores extremos que incrementan la dispersión de los datos.

d) La calidad del aire fue deficiente durante prácticamente todo el mes, ya que incluso los valores más bajos observados superan la recomendación de la OMS.

e) A la población: reducir actividades físicas al aire libre durante episodios de alta contaminación.
A las autoridades: fortalecer programas de reducción de emisiones vehiculares e industriales y emitir alertas ambientales oportunas.

f) La mediana es más representativa, ya que la media se ve afectada por el valor extremo de 178 µg/m³.

g) Indicaría que hubo numerosos días con niveles de contaminación muy elevados, aumentando el riesgo de problemas respiratorios y cardiovasculares en la población, especialmente en grupos vulnerables.

## 2.5 Datos agrupados

### 2.5.1

En R:

```
# ==========================================
# Datos
# ==========================================

fc <- c(
71,54,64,57,51,66,62,61,55,67,49,58,60,51,47,55,58,71,59,72,
60,70,50,63,56,65,55,62,60,68,63,53,55,50,72,71,53,63,66,56,
52,54,53,60,69,57,55,55,73,70,72,53,65,51,64,59,57,60,57,61,
61,46,70,64,67,58,51,65,58,55,60,53,55,55,58,57,73,56,65,70,
63,58,65,47,50,51,45,63,55,60,70,61,50,49,62,70,54,65,55,69,
64,50,47,68,65,59,51,51,59,61,67,61,48,70,50,65,60,52,60,57,
67,47,74,72,73,64,55,60,46,71,57,68,49,56,68,48,60,58,60,50,
57,60,63,50,64,54,64,64,54,62,55,53,62,65,52,62,61,59,59,59,
57,68,66,56,73,52,69,59,55,53,58,45,59,59,60,65,70,56,50,60,
50,73,52,55,67,45,58,70,58,46,65,56,45,65,62,70,60,66,63,66,
54,53,56,65,60,63,74,65,55,72,47,62,55,61,55,67,71,57,48,60,
53,56,65,54,65,66,62,52,55,68,59,56,61,60,56,52,63,70,66,55
)

# ==========================================
# Tabla de frecuencias agrupadas
# ==========================================

intervalos <- cut(
  fc,
  breaks = c(45, 50, 55, 60, 65, 70, 75),
```

```
  right = FALSE
)

fa <- table(intervalos)
fr <- prop.table(fa)
Fa <- cumsum(fa)
Fr <- cumsum(fr)

tabla_frecuencias <- data.frame(
  Intervalo = names(fa),
  FA = as.vector(fa),
  FR = round(as.vector(fr), 4),
  FAA = as.vector(Fa),
  FRA = round(as.vector(Fr), 4)
)

print(tabla_frecuencias)

# ============================================
# Puntos medios
# ============================================

tabla_frecuencias$xi <- c(47.5, 52.5, 57.5, 62.5, 67.5, 72.5)

N <- sum(tabla_frecuencias$FA)

# ============================================
# Media
# ============================================

media <- sum(tabla_frecuencias$xi *
           tabla_frecuencias$FA) / N

# ============================================
# Mediana
# ============================================

L <- 55
F_ant <- 58
f <- 62
c <- 5

mediana <- L + ((N/2 - F_ant) / f) * c

# ============================================
# Moda
# ============================================
```

```
L <- 55
f1 <- 62
f0 <- 40
f2 <- 54
c <- 5

moda <- L +
  ((f1 - f0) /
   ((f1 - f0) + (f1 - f2))) * c

# ==========================================
# Medidas de dispersión
# ==========================================

# Rango

rango <- max(fc) - min(fc)

# Varianza

varianza <- sum(
  tabla_frecuencias$FA *
    (tabla_frecuencias$xi - media)^2
) / N

# Desviación estándar

desv_est <- sqrt(varianza)

# Coeficiente de variación

cv <- 100 * desv_est / media

# Desviación media absoluta

MAD <- sum(
  tabla_frecuencias$FA *
    abs(tabla_frecuencias$xi - media)
) / N

# ==========================================
# Cuartiles
# ==========================================

# Q1

L <- 50
F_ant <- 18
```

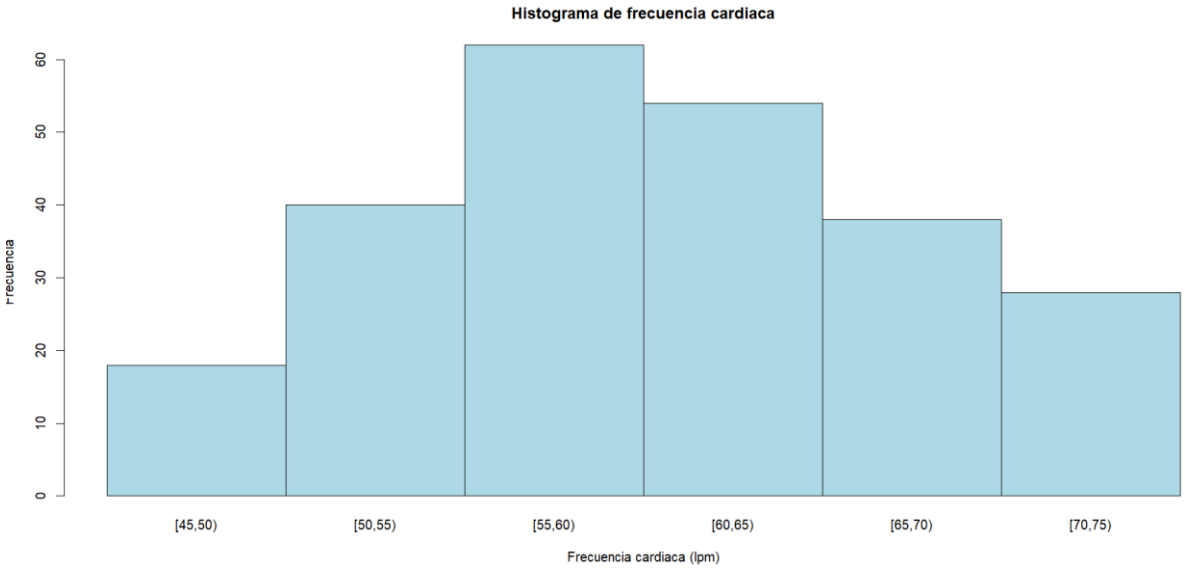
```
f <- 40
c <- 5

Q1 <- L + ((N/4 - F_ant) / f) * c

# Q3

L <- 65
F_ant <- 174
f <- 38
c <- 5

Q3 <- L + (((3 * N) / 4 - F_ant) / f) * c

# IQR

IQR <- Q3 - Q1

# ============================================
# Resumen
# ============================================

resultado <- data.frame(

  Media = round(media, 2),

  Mediana = round(mediana, 2),

  Moda = round(moda, 2),

  Rango = round(rango, 2),

  IQR = round(IQR, 2),

  Varianza = round(varianza, 2),

  Desv_Estandar = round(desv_est, 2),

  MAD = round(MAD, 2),

  CV = round(cv, 2)

)

print(resultado)

# ============================================
# Histograma (aproximado mediante barras)
```

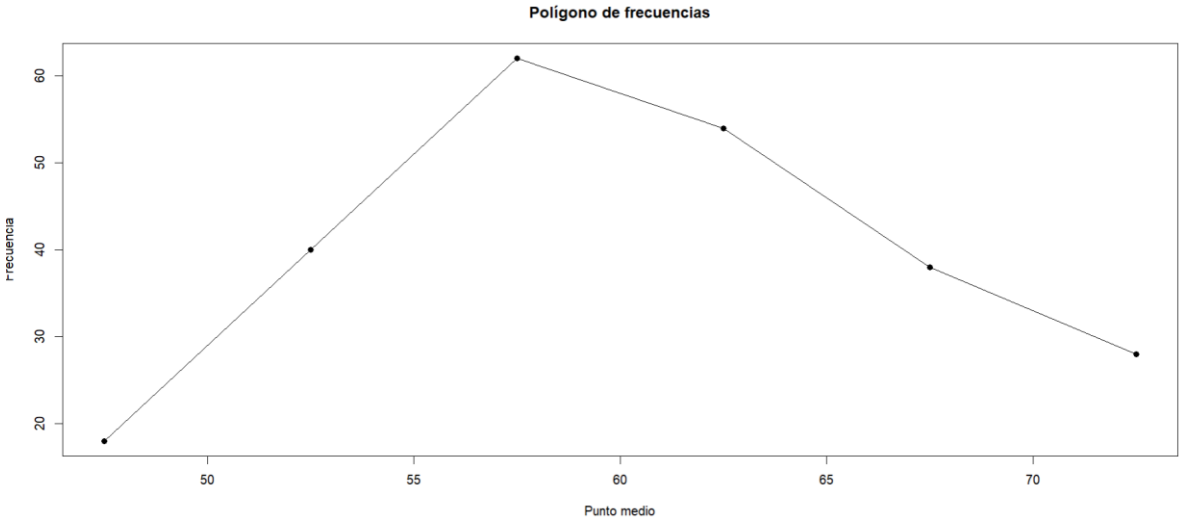
```
# ==========================================

barplot(
  tabla_frecuencias$FA,
  names.arg = tabla_frecuencias$Intervalo,
  space = 0,
  col = "lightblue",
  main = "Histograma de frecuencia cardiaca",
  xlab = "Frecuencia cardiaca (lpm)",
  ylab = "Frecuencia"
)

# ==========================================
# Polígono de frecuencias
# ==========================================

plot(
  tabla_frecuencias$xi,
  tabla_frecuencias$FA,
  type = "o",
  pch = 16,
  xlab = "Punto medio",
  ylab = "Frecuencia",
  main = "Polígono de frecuencias"
)

# ==========================================
# Ojiva
# ==========================================

plot(
  tabla_frecuencias$xi,
  tabla_frecuencias$FAA,
  type = "o",
  pch = 16,
  xlab = "Punto medio",
  ylab = "Frecuencia acumulada",
  main = "Ojiva"
)
```

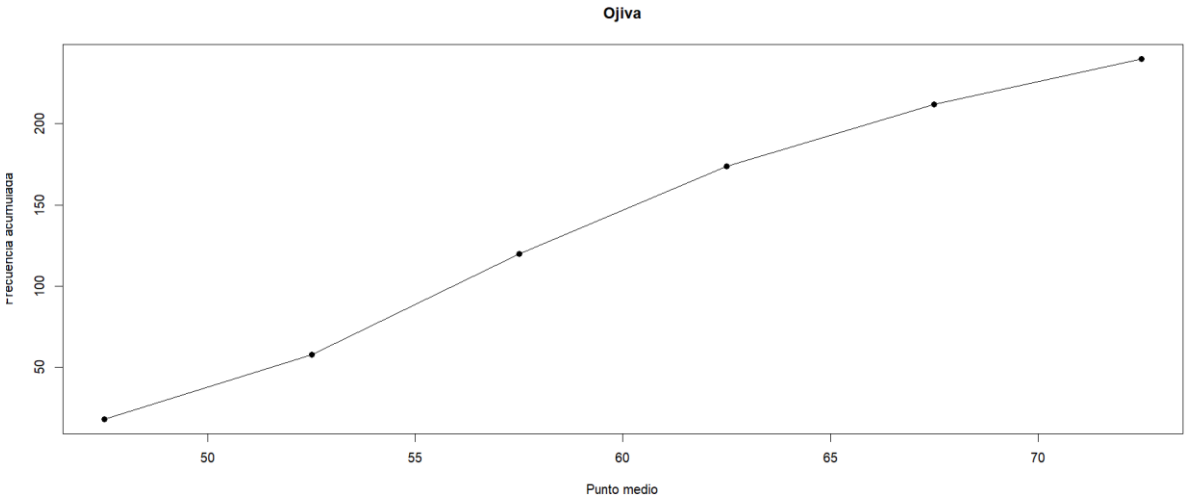
```
> tabla_frecuencias
  Intervalo Frecuencia_Absoluta Frecuencia_Relativa Frecuencia_Acumulada
1   [45,50                  18              0.0750                   18
2   [50,55                  40              0.1667                   58
3   [55,60                  62              0.2583                  120
4   [60,65                  54              0.2250                  174
5   [65,70                  38              0.1583                  212
6   [70,75                  28              0.1167                  240
```

```
  Frecuencia_Relativa_Acumulada
1                        0.0750
2                        0.2417
3                        0.5000
4                        0.7250
5                        0.8833
6                        1.0000
```

```
> resultado
  Media   Mediana   Moda    Rango   IQR     Varianza   Desv_Estandar
  60.38    60       58.67   30      10.54   51.11      7.15
  MAD     CV
  6.04    11.84
```






a) En el intervalo [55, 60 latidos por minuto, ya que presenta la frecuencia más alta 62 corredores.

b) El corredor típico tiene una frecuencia cardiaca en reposo cercana a 60 lpm, pues la media 60.38, la mediana 60 y la moda 58.67 son muy similares. Esto sugiere una distribución aproximadamente simétrica.

c) La distribución parece aproximadamente simétrica, ya que las medidas de tendencia central son muy cercanas entre sí y el histograma muestra una forma relativamente equilibrada alrededor del centro.

d) La mediana de 60 lpm indica que aproximadamente el 50% de los corredores tiene una frecuencia cardiaca menor o igual a 60 lpm y el otro 50% mayor o igual a ese valor.

e) El grupo parece relativamente homogéneo, ya que la desviación estándar 7.15 lpm y el coeficiente de variación 11.84% son bajos, indicando poca variabilidad respecto a la media.

f) Los corredores con frecuencias cardiacas más altas podrían requerir programas de acondicionamiento más progresivos, mientras que aquellos con frecuencias más bajas podrían realizar entrenamientos de mayor intensidad desde el inicio.

g) Se pierde el valor exacto de cada observación, así como detalles sobre posibles concentraciones, patrones particulares o valores atípicos dentro de cada intervalo.

h) Se supone que los datos están distribuidos de manera relativamente uniforme dentro de cada intervalo y que el punto medio de la clase representa adecuadamente a todas las observaciones de esa clase.

**2.5.2**

```
# ==========================================
# Datos
# ==========================================

# Límites de clase
li <- c(0, 5, 10, 15, 20, 25)
ls <- c(5, 10, 15, 20, 25, 30)

# Puntos medios
xi <- (li + ls) / 2

# Frecuencias
antes <- c(18, 42, 60, 48, 22, 10)
despues <- c(10, 28, 50, 60, 32, 20)

# ==========================================
# Función para estadísticas agrupadas
# ==========================================

estadisticas_agrupadas <- function(fa, xi, li, amplitud){

  N <- sum(fa)

  #----------------------------------------
  # Media
  #----------------------------------------

  media <- sum(fa * xi) / N
```

```
#----------------------------------------
# Frecuencias acumuladas
#----------------------------------------

FAA <- cumsum(fa)

#----------------------------------------
# Mediana
#----------------------------------------

pos_med <- N / 2

clase_med <- which(FAA >= pos_med)[1]

L <- li[clase_med]

F_ant <- ifelse(clase_med == 1,
                0,
                FAA[clase_med - 1])

f_med <- fa[clase_med]

mediana <- L +
  ((pos_med - F_ant) / f_med) * amplitud

#----------------------------------------
# Moda
#----------------------------------------

clase_mod <- which.max(fa)

L <- li[clase_mod]

f1 <- fa[clase_mod]

f0 <- ifelse(clase_mod == 1,
             0,
             fa[clase_mod - 1])

f2 <- ifelse(clase_mod == length(fa),
             0,
             fa[clase_mod + 1])

moda <- L +
  ((f1 - f0) /
     ((f1 - f0) + (f1 - f2))) *
  amplitud
```

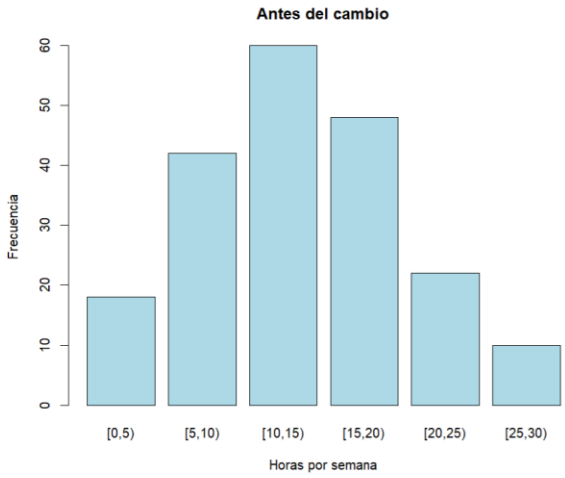
```
#----------------------------------------
# Varianza
#----------------------------------------

varianza <- sum(fa * (xi - media)^2) / N

#----------------------------------------
# Desviación estándar
#----------------------------------------

desv_est <- sqrt(varianza)

#----------------------------------------
# Desviación media absoluta
#----------------------------------------

MAD <- sum(fa * abs(xi - media)) / N

#----------------------------------------
# Coeficiente de variación
#----------------------------------------

CV <- 100 * desv_est / media

#----------------------------------------
# Función para cuantiles
#----------------------------------------

cuantil <- function(p){

  pos <- p * N

  clase <- which(FAA >= pos)[1]

  L <- li[clase]

  F_ant <- ifelse(clase == 1,
                  0,
                  FAA[clase - 1])

  f <- fa[clase]

  L + ((pos - F_ant) / f) * amplitud

}

Q1 <- cuantil(0.25)
```

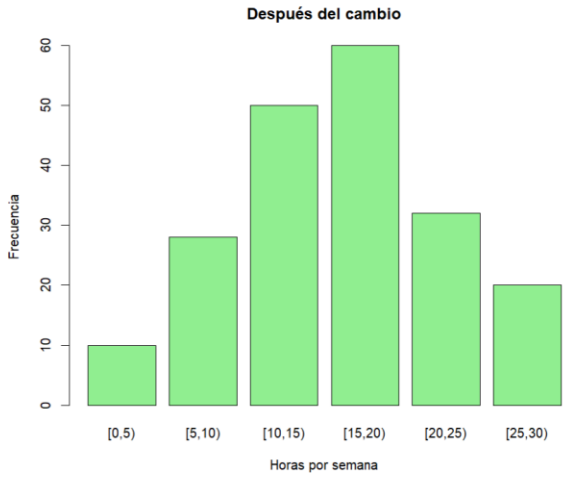
```
  Q3 <- cuantil(0.75)

  IQR <- Q3 - Q1

  data.frame(

    Media = round(media,2),

    Mediana = round(mediana,2),

    Moda = round(moda,2),

    Rango = max(ls) - min(li),

    IQR = round(IQR,2),

    Varianza = round(varianza,2),

    Desv_Estandar = round(desv_est,2),

    MAD = round(MAD,2),

    CV = round(CV,2)

  )

}

# ============================================
# Resultados
# ============================================

res_antes <- estadisticas_agrupadas(
  antes,
  xi,
  li,
  amplitud = 5
)

res_despues <- estadisticas_agrupadas(
  despues,
  xi,
  li,
  amplitud = 5
)

rbind(
```

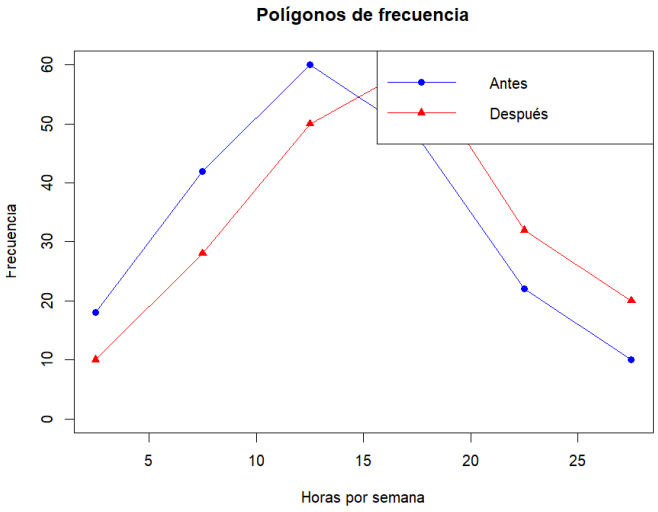
```
  Antes = res_antes,
  Despues = res_despues
)

# ==========================================
# Histogramas
# ==========================================

par(mfrow = c(1,2))

barplot(
  antes,
  names.arg = c("[0,5)", "[5,10)", "[10,15)",
                "[15,20)", "[20,25)", "[25,30)"),
  col = "lightblue",
  main = "Antes del cambio",
  xlab = "Horas por semana",
  ylab = "Frecuencia",
  space = 0
)

barplot(
  despues,
  names.arg = c("[0,5)", "[5,10)", "[10,15)",
                "[15,20)", "[20,25)", "[25,30)"),
  col = "lightgreen",
  main = "Después del cambio",
  xlab = "Horas por semana",
  ylab = "Frecuencia",
  space = 0
)

par(mfrow = c(1,1))

# ==========================================
# Polígonos de frecuencias
# ==========================================

plot(
  xi,
  antes,
  type = "o",
  pch = 16,
  ylim = c(0, max(c(antes, despues))),
  col = "blue",
  xlab = "Horas por semana",
  ylab = "Frecuencia",
  main = "Polígonos de frecuencia"
```

```
)

lines(
  xi,
  despues,
  type = "o",
  pch = 17,
  col = "red"
)

legend(
  "topright",
  legend = c("Antes", "Después"),
  col = c("blue", "red"),
  lty = 1,
  pch = c(16,17)
)
```

```
> resultados
         Media  Mediana  Moda   Rango  IQR    Varianza  Desv_Estandar
Antes     13.6  13.33    13.00  30     9.32   41.29     6.43
Despues   15.9  16.00    16.32  30     9.11   42.94     6.55

           MAD    CV
Antes     5.22   47.25
Despues   5.39   41.21
```






a) El mayor número de usuarios se concentra en:

Antes: intervalo [10,15 horas por semana 60 usuarios.

Después: intervalo [15,20 horas por semana 60 usuarios.

Esto indica que, tras la actualización, el consumo típico se desplazó hacia niveles más altos de uso.

b) Antes del cambio, el usuario típico consumía alrededor de 13 horas semanales media = 13.6, mediana = 13.33, moda = 13.00.

Después del cambio, el usuario típico consumía alrededor de 16 horas semanales media = 15.9, mediana = 16.0, moda = 16.32.

Las tres medidas aumentan, sugiriendo un incremento general en el tiempo de consumo.

c) La distribución parece aproximadamente simétrica en ambos periodos, ya que la media, la mediana y la moda son muy cercanas entre sí. No se observa un sesgo pronunciado.

d) La mediana indica el valor que divide a los usuarios en dos grupos iguales.

Antes: aproximadamente el 50% consumía menos de 13.3 horas semanales y el otro 50% más.

Después: aproximadamente el 50% consumía menos de 16 horas y el otro 50% más.

Esto muestra un desplazamiento hacia mayores niveles de consumo.

e) Los usuarios presentan una variabilidad moderada y bastante similar en ambos periodos.

Rango = 30 horas en ambos casos.

IQR ≈ 9 horas.

Desviación estándar ≈ 6.5 horas.

Sin embargo, el CV disminuye de 47.25% a 41.21%, lo que indica que el consumo es ligeramente más homogéneo después del cambio.

f) Las medidas de tendencia central permiten identificar el nivel típico de consumo, mientras que las de dispersión ayudan a distinguir usuarios de bajo, medio y alto consumo. Esto puede utilizarse para diseñar recomendaciones o promociones personalizadas para cada segmento.

g) Se pierde la información exacta de cada usuario.

Por ejemplo, dos usuarios que consumen 10.1 y 14.9 horas quedan en el mismo intervalo [10,15 aunque sus comportamientos sean distintos.

h) Se supone que los datos están distribuidos de manera relativamente uniforme dentro de cada intervalo y que el punto medio representa adecuadamente a todos los valores de la clase.

Estas aproximaciones pueden introducir pequeñas diferencias respecto a los resultados obtenidos con los datos originales.

i) La frecuencia acumulada hasta 15 horas es:

$$18 + 42 + 60 = 120$$

De un total de 200 usuarios:

$$\frac{120}{200} = 0.60$$

Por lo tanto, aproximadamente 60% de los usuarios consumía menos de 15 horas semanales antes del cambio.

Después del cambio:

$$10 + 28 + 50 = 88$$

$$\frac{88}{200} = 0.44$$

Aproximadamente 44% de los usuarios consume menos de 15 horas semanales, lo que confirma un aumento general del tiempo de uso.

j) Se utilizaría el primer cuartil $Q_1$, que corresponde al percentil 25.

Interpretación:

El 25% de los usuarios consume menos horas que $Q_1$.

El 75% consume más horas que $Q_1$.

Este cuantil permite identificar al grupo de usuarios de menor consumo y diseñar estrategias específicas para incrementar su interacción con la plataforma.

**2.5.3**
En R

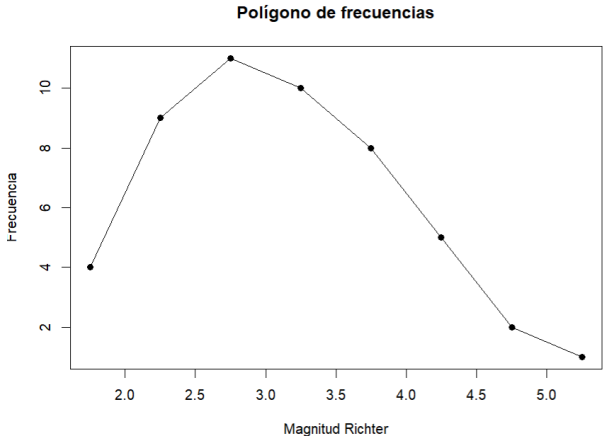
```
# ==========================================
# Tabla de frecuencias
# ==========================================

li <- c(1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0)
ls <- c(2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5)

fa <- c(4, 9, 11, 10, 8, 5, 2, 1)

N <- sum(fa)

# Frecuencia relativa
fr <- fa / N

# Frecuencia acumulada
Fa <- cumsum(fa)

# Frecuencia relativa acumulada
Fr <- cumsum(fr)

tabla <- data.frame(
  Intervalo = paste(li, ls, sep = " - "),
  FA = fa,
  FR = round(fr, 4),
  FAA = Fa,
  FRA = round(Fr, 4)
)

print(tabla)

# ==========================================
# Polígono de frecuencias
# ==========================================

# Puntos medios
xi <- (li + ls) / 2
```

```
plot(
  xi,
  fa,
  type = "o",
  pch = 16,
  xlab = "Magnitud Richter",
  ylab = "Frecuencia",
  main = "Polígono de frecuencias"
)

# ==========================================
# Medidas descriptivas
# ==========================================

# Media

media <- sum(xi * fa) / N

# Mediana

pos <- N / 2

clase <- which(Fa >= pos)[1]

L <- li[clase]

F_ant <- ifelse(clase == 1, 0, Fa[clase - 1])

f <- fa[clase]

c <- 0.5

mediana <- L + ((pos - F_ant) / f) * c

# Clase modal

clase_modal <- which.max(fa)

print(tabla[clase_modal, ])

# Moda

L <- li[clase_modal]

f1 <- fa[clase_modal]

f0 <- ifelse(clase_modal == 1, 0, fa[clase_modal - 1])
```

```
f2 <- ifelse(clase_modal == length(fa), 0, fa[clase_modal + 1])

moda <- L +
  ((f1 - f0) /
     ((f1 - f0) + (f1 - f2))) * c

# Varianza

varianza <- sum(fa * (xi - media)^2) / N

# Desviación estándar

desv <- sqrt(varianza)

# Coeficiente de variación

cv <- 100 * desv / media

# ============================================
# Resultados
# ============================================

resultado <- data.frame(

  Media = round(media, 3),

  Mediana = round(mediana, 3),

  Moda = round(moda, 3),

  Clase_Modal = paste(li[clase_modal],
                      ls[clase_modal],
                      sep = " - "),

  Desv_Estandar = round(desv, 3),

  Coeficiente_Variacion = round(cv, 2)

)

print(resultado)
```

a)
```
> tabla
     Intervalo FA   FR   FAA   FRA
1   1.5 - 2    4   0.08    4   0.08
2   2 - 2.5    9   0.18   13   0.26
3   2.5 - 3   11   0.22   24   0.48
4   3 - 3.5   10   0.20   34   0.68
5   3.5 - 4    8   0.16   42   0.84
6   4 - 4.5    5   0.10   47   0.94
7   4.5 - 5    2   0.04   49   0.98
8   5 - 5.5    1   0.02   50   1.00
```

b)




c)
```
> resultado
Media  Mediana  Moda   Clase_Modal  Desv_Estandar Coeficiente_Variacion
3.12   3.05     2.833  2.5 - 3.0    0.841                 26.97
```

d) Media = 3.12: La magnitud promedio de los sismos registrados fue de aproximadamente 3.12 grados Richter.

Mediana = 3.05: El 50% de los sismos tuvo una magnitud menor o igual a 3.05, mientras que el otro 50% presentó magnitudes superiores.

Clase modal = 2.5–3.0: Este fue el intervalo donde se concentró el mayor número de sismos, por lo que las magnitudes entre 2.5 y 3.0 fueron las más frecuentes.

Desviación estándar = 0.841: Las magnitudes suelen diferir del promedio en aproximadamente 0.84 grados Richter, lo que indica una dispersión moderada.

Coeficiente de variación = 26.97%: La variabilidad de las magnitudes representa cerca del 27% de la media, lo que sugiere una dispersión moderada de la actividad sísmica.

e) La distribución es unimodal, ya que presenta un único pico en el intervalo 2.5–3.0, que corresponde a la clase modal.

Se observa una cola más larga hacia las magnitudes altas 4.0 a 5.5, lo que sugiere un ligero sesgo positivo o hacia la derecha. La relación entre las medidas de tendencia central respalda esta conclusión:

Moda = 2.833 < Mediana = 3.05 < Media = 3.12

La diferencia entre estas medidas no es muy grande, por lo que el sesgo es moderado.

f) Se podrían utilizar:
- Media y mediana: para comparar la magnitud típica de los sismos en ambas ciudades.
- Desviación estándar y coeficiente de variación: para comparar la variabilidad de las magnitudes.
- Histogramas y polígonos de frecuencia: para comparar visualmente la forma de las distribuciones.
- Diagramas de caja: para identificar diferencias en dispersión, asimetría y posibles valores atípicos.

g) La frecuencia acumulada hasta el intervalo 3.0–3.5 es $4 + 9 + 11 + 10 = 34$

Por lo tanto:

$$\frac{34}{50} \times 100 = 68\%$$

El 68% de los sismos tuvo una magnitud menor a 3.5.
Interpretación: La mayoría de los eventos registrados fueron de baja a moderada intensidad, mientras que los sismos de magnitud elevada fueron relativamente poco frecuentes. Esto sugiere que la actividad sísmica observada estuvo dominada por movimientos de menor magnitud.

**2.5.4**

| **Medida** | **Foco A** | **Foco B** |
|---|---|---|
| Media | 1186.8 | 1201.6 |
| Mediana | 1202 | 1199 |
| Rango | 161 | 103 |
| Desv. estándar muestral | 58.34 | 42.59 |
| Coeficiente de variación % | 4.92 | 3.54 |

a) La marca B presenta una duración promedio mayor 1201.6 horas frente a 1186.8 horas y una menor variabilidad.

b) La marca B, porque tiene menor desviación estándar 42.59 horas y menor coeficiente de variación 3.54%.

c) La marca B, ya que combina una mayor duración promedio con una mayor consistencia entre focos.

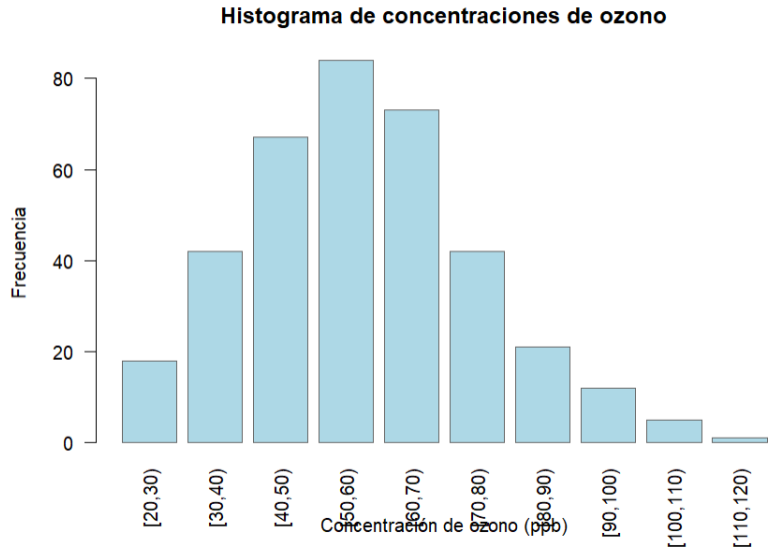
**2.5.5**

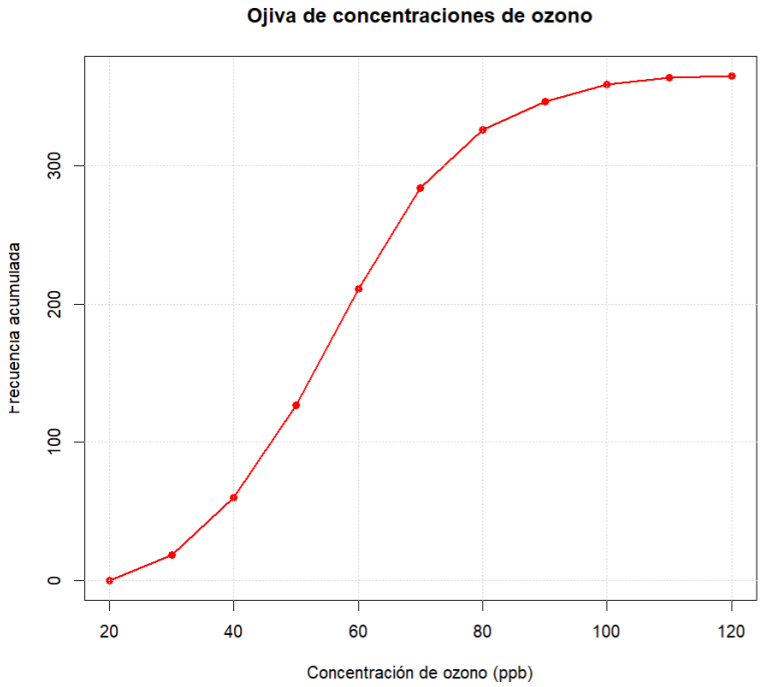
a)

| Intervalo | FA | FR | FAA | FRA |
|---|---|---|---|---|
| 20-30 | 18 | 0.0493 | 18 | 0.0493 |
| 30-40 | 42 | 0.1151 | 60 | 0.1644 |
| 40-50 | 67 | 0.1836 | 127 | 0.3479 |
| 50-60 | 84 | 0.2301 | 211 | 0.5781 |
| 60-70 | 73 | 0.2000 | 284 | 0.7781 |
| 70-80 | 42 | 0.1151 | 326 | 0.8932 |
| 80-90 | 21 | 0.0575 | 347 | 0.9507 |
| 90-100 | 12 | 0.0329 | 359 | 0.9836 |
| 100-110 | 5 | 0.0137 | 364 | 0.9973 |
| 110-120 | 1 | 0.0027 | 365 | 1.0000 |

```
< # ==========================================
# Histograma
# ==========================================

# Etiquetas de intervalos
intervalos <- c(
  "[20,30)", "[30,40)", "[40,50)", "[50,60)",
  "[60,70)", "[70,80)", "[80,90)", "[90,100)",
  "[100,110)", "[110,120)"
)
```

```
barplot(
  height = fa,
  names.arg = intervalos,
  col = "lightblue",
  border = "gray40",
  main = "Histograma de concentraciones de ozono",
  xlab = "Concentración de ozono (ppb)",
  ylab = "Frecuencia",
  las = 2,
  space = 0
)

# ========================================
# Ojiva
# ========================================

FAA <- c(0, cumsum(fa))
x <- c(20, ls)

plot(
  x,
  FAA,
  type = "o",
  pch = 16,
  lwd = 2,
  col = "red",
  main = "Ojiva de concentraciones de ozono",
  xlab = "Concentración de ozono (ppb)",
  ylab = "Frecuencia acumulada"
)

grid()
```




b) El intervalo modal es: $[50,60)$ porque contiene la mayor frecuencia 84 días.
Interpretación: La concentración máxima diaria de ozono observada con mayor frecuencia durante el año estuvo entre 50 y 60 ppb.

d) Códigos en R:

```
# ==========================================
# Datos
```

```
# ==========================================

li <- c(20, 30, 40, 50, 60, 70, 80, 90, 100, 110)
ls <- c(30, 40, 50, 60, 70, 80, 90, 100, 110, 120)

fa <- c(18, 42, 67, 84, 73, 42, 21, 12, 5, 1)

N <- sum(fa)

# Puntos medios
xi <- (li + ls) / 2

# ==========================================
# Medidas descriptivas
# ==========================================

# Media para datos agrupados
media <- sum(xi * fa) / N
media

# ==========================================
# Mediana
# ==========================================

FAA <- cumsum(fa)

pos <- N / 2

clase <- which(FAA >= pos)[1]

L <- li[clase]

F_ant <- ifelse(clase == 1, 0, FAA[clase - 1])

f <- fa[clase]

c <- ls[clase] - li[clase]

mediana <- L + ((pos - F_ant) / f) * c
mediana

# ==========================================
# Moda
# ==========================================

clase_modal <- which.max(fa)

L <- li[clase_modal]
```

```
f1 <- fa[clase_modal]

f0 <- ifelse(clase_modal == 1, 0, fa[clase_modal - 1])

f2 <- ifelse(clase_modal == length(fa), 0, fa[clase_modal + 1])

c <- ls[clase_modal] - li[clase_modal]

moda <- L +
  ((f1 - f0) /
     ((f1 - f0) + (f1 - f2))) * c

moda

# ==========================================
# Varianza poblacional
# ==========================================

varianza <- sum(fa * (xi - media)^2) / N
varianza

# ==========================================
# Desviación estándar
# ==========================================

desv_est <- sqrt(varianza)
desv_est

# ==========================================
# Coeficiente de variación
# ==========================================

cv <- 100 * desv_est / media
cv

# ==========================================
# Función para cuantiles agrupados
# ==========================================

cuantil_agrupado <- function(p){

  pos <- p * N

  clase <- which(FAA >= pos)[1]

  L <- li[clase]
```

```
  F_ant <- ifelse(clase == 1,
                  0,
                  FAA[clase - 1])

  f <- fa[clase]

  c <- ls[clase] - li[clase]

  Q <- L + ((pos - F_ant) / f) * c

  return(Q)

}

# ==========================================
# Cuartiles
# ==========================================

Q1 <- cuantil_agrupado(0.25)
Q2 <- cuantil_agrupado(0.50)
Q3 <- cuantil_agrupado(0.75)

Q1
Q2
Q3

# ==========================================
# Percentiles
# ==========================================

P10 <- cuantil_agrupado(0.10)
P80 <- cuantil_agrupado(0.80)
P95 <- cuantil_agrupado(0.95)

P10
P80
P95

# ==========================================
# Tabla resumen
# ==========================================

resultado <- data.frame(

  Media = round(media, 2),

  Mediana = round(mediana, 2),
```

```
  Moda = round(moda, 2),
  Varianza = round(varianza, 2),
  Desv_Estandar = round(desv_est, 2),
  CV = round(cv, 2),
  Q1 = round(Q1, 2),
  Q2 = round(Q2, 2),
  Q3 = round(Q3, 2),
  P10 = round(P10, 2),
  P80 = round(P80, 2),
  P95 = round(P95, 2)
 )
 print(resultado)
```

```
> resultado
  Media   Mediana   Moda   Varianza    Desv_Estandar    CV
  57.58   56.61     56.07  316.66      17.79            30.91

  Q1      Q2      Q3     P10    P80    P95
  44.66   56.61   68.6   34.4   71.9   89.88
```

Media: la concentración promedio fue de 57.58 ppb.
Mediana: el 50% de los días presentó concentraciones inferiores a 56.61 ppb.
Moda: la concentración más frecuente fue de 56.07 ppb.
Desviación estándar: las concentraciones suelen diferir del promedio en 17.79 ppb.
CV ≈ 30%: existe una variabilidad moderada.

e)
```
  Q1      Q2      Q3     P10    P80    P95
 44.66   56.61   68.6   34.4   71.9   89.88
```

f) El 95% de los días del año registró concentraciones máximas de ozono iguales o inferiores a 89.88 ppb, mientras que solo el 5% de los días presentó concentraciones superiores a este valor.

En términos ambientales, esto significa que el percentil 95 es un indicador de los niveles extremos de contaminación por ozono. Un valor cercano a 90 ppb sugiere que, aunque la mayoría de los días presentan concentraciones moderadas, existe un pequeño grupo de días con niveles significativamente más altos que pueden representar un riesgo importante para la salud pública.

La información del percentil 95 permite a las autoridades identificar la magnitud de los eventos más severos de contaminación y evaluar si son necesarias medidas extraordinarias, como contingencias ambientales o restricciones temporales a ciertas actividades contaminantes.

g) La distribución es unimodal y presenta un ligero sesgo positivo hacia la derecha. Esto se observa porque la media 57.58 es ligeramente mayor que la mediana 56.61, y la frecuencia disminuye gradualmente hacia las concentraciones más altas. Además, el histograma muestra una cola hacia los valores elevados de ozono.

h) Las concentraciones mostraron una variabilidad moderada. La desviación estándar es de aproximadamente 17.79 ppb y el coeficiente de variación es de 30.91%, lo que indica que las concentraciones fluctúan alrededor del promedio, aunque sin una dispersión excesiva.

i) Los días con concentraciones superiores a 80 ppb son: $21 + 12 + 5 + 1 = 39$ días

$$\frac{39}{365} \times 100 \approx 10.68\%$$

Aproximadamente 10.7% de los días presentaron concentraciones superiores a 80 ppb.
Interpretación: Aunque la mayoría de los días tuvo niveles moderados de ozono, alrededor de uno de cada diez días registró concentraciones relativamente altas.

j) Los días con concentraciones superiores a 90 ppb son: $12 + 5 + 1 = 18$ días

$$\frac{18}{365} \times 100 \approx 4.93\%$$

Respuesta: Aproximadamente 4.9% de los días presentaron niveles preocupantes de ozono.
Interpretación: Aunque estos episodios fueron poco frecuentes, pueden representar riesgos importantes para grupos vulnerables como niños, adultos mayores y personas con enfermedades respiratorias.

k) La calidad del aire fue generalmente moderada, con concentraciones de ozono concentradas entre 40 y 70 ppb. Sin embargo, existieron episodios de contaminación elevada, evidenciados por los valores altos del percentil 95 y por la presencia de días con concentraciones superiores a 80 y 90 ppb. Esto indica que,

aunque la mayoría de los días no presentó niveles extremos, sí ocurrieron eventos de contaminación relevantes.

l) Fortalecer programas de reducción de emisiones vehiculares.
- Promover el uso del transporte público y la movilidad sustentable.
- Incrementar la vigilancia de emisiones industriales.
- Emitir alertas preventivas durante condiciones meteorológicas favorables a la formación de ozono.
- Informar a la población vulnerable para limitar actividades al aire libre durante episodios de alta contaminación.
- Estas recomendaciones se justifican por la existencia de días con concentraciones elevadas que pueden afectar la salud pública.

m) El percentil 95 indica que el 95% de los días presentó concentraciones iguales o inferiores a 89.88 ppb, mientras que el 5% restante corresponde a los episodios más contaminados del año.
Este indicador es útil porque permite evaluar la intensidad de los eventos extremos, los cuales suelen tener un mayor impacto en la salud que los niveles promedio.

n) Sería más útil monitorear el percentil 95.
La media resume el comportamiento promedio de todo el año, pero puede ocultar episodios extremos. En cambio, el percentil 95 se enfoca específicamente en los días con mayores concentraciones de ozono, por lo que es más adecuado para evaluar riesgos ambientales, diseñar contingencias y medir el éxito de políticas dirigidas a reducir los eventos más severos de contaminación.

## 2.6 Diagramas de caja y brazos

2.6.1

a) Polanco presenta las rentas más altas, ya que tiene la mediana más elevada, cercana a $32,000 mensuales. Esto indica que la mitad de los departamentos en Polanco tienen rentas superiores a ese valor.

b) Polanco muestra la mayor variabilidad. Se identifica porque tiene el boxplot más alto IQR amplio y también una gran distancia entre los valores mínimo y máximo.

c) Del Valle parece presentar la menor variabilidad, ya que su caja es relativamente estrecha y los bigotes son más cortos que en otras zonas.
Esto sugiere que las rentas son más homogéneas y que existe menor diferencia entre departamentos.

d) Mediana más alta: Polanco ≈ $32,000
Mediana más baja: Narvarte ≈ $16,500

Diferencia:$32{,}000 - 16{,}500 = 15{,}500$
Esta diferencia refleja la fuerte segmentación del mercado inmobiliario de la ciudad y las diferencias en ubicación, servicios y prestigio de las zonas.

e) Los IQR más amplios se observan en Polanco y Santa Fe
Esto indica una mayor dispersión en los precios de renta y una oferta más diversa de departamentos.

f) No se observan puntos aislados fuera de los bigotes, por lo que no parecen existir valores atípicos visibles en ninguna zona.
Si existieran, podrían representar departamentos de lujo o inmuebles con características excepcionales.

g)

| Zona | Mediana aproximada |
|---|---|
| Condesa | $23,000 |
| Roma | $22,000 |
| Narvarte | $16,500 |

- Condesa y Roma presentan rentas similares.
- Narvarte es claramente más económica.
- La dispersión es mayor en Condesa y Roma que en Narvarte.

h) Sí. Por ejemplo:
- Condesa y Roma.
- Roma y Del Valle.
- Santa Fe y Polanco.

Esto implica que un mismo presupuesto puede permitir encontrar opciones en distintas zonas, aunque con características diferentes.

i) Polanco.
Su rango va aproximadamente de $25,000 a $40,000, lo que indica una gran diversidad de inmuebles y segmentos de mercado dentro de la misma zona.

j) Sería relativamente típico en Polanco y Santa Fe.
Sería inusualmente alto en Narvarte, Del Valle, Roma y Condesa
En estas últimas zonas, $30,000 se encuentra cerca o incluso por encima de los valores máximos observados.

k) Aproximadamente simétricas: Del Valle, Narvarte y Roma
Se observa porque la mediana está cerca del centro de la caja y los bigotes tienen longitudes similares.
Posible sesgo hacia rentas altas en Polanco y Santa Fe

Se aprecia porque la mitad superior de la distribución parece más extendida y los valores máximos están más alejados.

l) Del Valle.
Presenta una dispersión reducida y una mediana moderada, lo que sugiere un mercado más estable y predecible.

m) Roma Norte/Sur o Condesa.
Aunque muestran mayor dispersión que Del Valle o Narvarte, incluyen departamentos con rentas relativamente bajas dentro de zonas muy demandadas, lo que aumenta la posibilidad de encontrar oportunidades atractivas.

**2.6.2**

a) Ingeniería parece ofrecer los salarios iniciales más altos, ya que presenta la mediana más elevada, cercana a $18,000 mensuales. Esto indica que el egresado típico de Ingeniería recibe un salario superior al de las demás carreras analizadas.

b) Economía parece presentar la mayor variabilidad salarial. Esto se observa porque tiene una de las cajas IQR más amplias y una gran distancia entre los valores mínimo y máximo.
Esto sugiere que existen diferencias importantes entre los salarios iniciales de sus egresados.

c) Ciencia Política presenta la menor variabilidad. Su caja es relativamente estrecha y el rango total es menor que el de otras carreras.
Esto podría indicar que las oportunidades laborales para sus egresados son más homogéneas en términos salariales.

d) La mediana parece representar adecuadamente el salario típico en: Ciencia Política Relaciones Internacionales y Contaduría porque presentan una dispersión relativamente moderada.
Podría ser menos representativa en Economía e Ingeniería debido a su mayor dispersión salarial.

e) No se observan puntos fuera de los bigotes del boxplot, por lo que no hay valores atípicos visibles.
Si existieran, podrían representar egresados que consiguieron puestos excepcionalmente bien remunerados o empleos con condiciones muy distintas al promedio.

f) Economía presenta el IQR más amplio.

Esto implica una mayor diversidad en los salarios iniciales de sus egresados, posiblemente debido a la variedad de sectores y puestos en los que pueden emplearse.

g) Ingeniería presenta una mediana cercana a $18,000. Además: Ingeniería ofrece salarios iniciales más altos.
Economía presenta una mediana cercana a $16,000.
Ambas presentan una dispersión considerable.
Economía muestra una mayor heterogeneidad salarial.
Esto sugiere que, en promedio, los egresados de Ingeniería tienen mejores oportunidades salariales iniciales.

h) Sí. Se observa traslape entre varias carreras, por ejemplo:
- Administración y Contaduría
- Contaduría y Economía
- Relaciones Internacionales y Administración
- Economía e Ingeniería

Esto indica que algunos egresados de carreras distintas pueden obtener salarios iniciales similares, por lo que la carrera no es el único factor que determina el ingreso; también influyen la empresa, la experiencia, las habilidades y el sector laboral.

i) Un salario de $22,000 mensuales:
- Sería relativamente típico o cercano al extremo superior en Ingeniería, ya que se encuentra dentro de su rango observado.
- Sería inusualmente alto para Administración, Ciencia Política, Contaduría, Economía y Relaciones Internacionales, pues supera claramente sus medianas e incluso se acerca o rebasa los valores máximos observados en varias de ellas.

Esto sugiere que una oferta de $22,000 es especialmente atractiva para la mayoría de las carreras analizadas.

j) Las carreras que parecen mostrar mayor asimetría son: Economía, Ingeniería y Administración
Esto puede identificarse porque la mediana no se encuentra exactamente en el centro de la caja y los bigotes tienen longitudes diferentes. En particular, los bigotes superiores parecen más extensos, lo que sugiere una posible asimetría hacia salarios más altos.

k) Administración parece ofrecer un mayor potencial salarial máximo.
La evidencia es que el valor máximo observado en el boxplot de Administración alcanza aproximadamente $17,800, mientras que en Contaduría se sitúa alrededor de $16,300.

Esto sugiere que algunos egresados de Administración pueden acceder a puestos iniciales mejor remunerados.

l) Las carreras más atractivas serían: Ingeniería, Economía y Contaduría
Estas carreras presentan las medianas salariales más elevadas y, en general, los rangos superiores más altos.
Si el criterio principal es el ingreso inicial esperado, Ingeniería destaca claramente sobre las demás.

m) Ciencia Política parece ser la opción más conveniente.
Presenta una dispersión relativamente pequeña y un rango más reducido, lo que sugiere que los salarios iniciales son más homogéneos y predecibles.
También Relaciones Internacionales muestra una variabilidad moderada, aunque con una dispersión algo mayor.

n) La carrera elegida influye significativamente en el salario inicial, ya que las medianas difieren considerablemente entre áreas de estudio.
Ingeniería y Economía presentan los salarios iniciales más altos, mientras que Ciencia Política y Relaciones Internacionales muestran niveles salariales más bajos.
La dispersión salarial también varía entre carreras; algunas ofrecen ingresos más homogéneos y otras presentan mayores diferencias entre egresados.
Existen traslapes entre varias carreras, lo que indica que el salario no depende únicamente de la licenciatura, sino también de factores como habilidades, experiencia, empresa contratante, ubicación geográfica y sector económico.
Elegir una carrera implica considerar tanto el nivel salarial esperado como la estabilidad de los ingresos, ya que una mayor remuneración suele venir acompañada de una mayor variabilidad salarial.
En conjunto, los boxplots muestran que la elección de carrera puede influir de manera importante en las oportunidades económicas iniciales, aunque no determina completamente el éxito laboral de un egresado.

## 2.7 El problema de comparación: análisis de subpoblaciones

### 2.7.1

a) No. Cuando se analizan los datos por licenciatura, se observa que en ambas carreras los estudiantes que no utilizan IA tienen mejores promedios.
Por lo tanto, la conclusión inicial de que el uso de IA mejora el desempeño académico no se mantiene al controlar por licenciatura.

b) Porque los promedios generales mezclan estudiantes de distintas licenciaturas que pueden tener características académicas diferentes.

Al combinar grupos heterogéneos, se puede ocultar la verdadera relación entre las variables y llegar a conclusiones engañosas.

c) La variable licenciatura Ingeniería o Sociales.
Esta variable influye tanto en el promedio académico como en la composición de los grupos que usan o no usan IA, alterando la comparación global.

d) Podría concluir erróneamente que el uso de IA mejora el rendimiento académico y, con base en ello, promover o incentivar su uso de manera generalizada.
Sin embargo, el análisis por licenciatura muestra que dentro de cada carrera los estudiantes que usan IA obtienen promedios más bajos, por lo que la decisión estaría basada en una interpretación incorrecta de los datos agregados.
Conclusión
Este ejemplo muestra la importancia de analizar posibles variables de confusión antes de interpretar asociaciones estadísticas. Los resultados agregados pueden ser engañosos si no se consideran las características de los distintos subgrupos que conforman la población.

**2.7.2**

a) No. Dentro de cada nivel socioeconómico, los estudiantes con beca presentan una tasa de deserción ligeramente mayor.
Por lo tanto, la conclusión observada en los datos agregados no se mantiene al analizar los grupos por separado.

b) Porque los resultados generales mezclan estudiantes con distintos niveles socioeconómicos.
Si la distribución de becas no es uniforme entre los niveles, el promedio global puede ocultar lo que realmente ocurre dentro de cada grupo.

c) El nivel socioeconómico.
Esta variable afecta tanto la probabilidad de recibir una beca como la probabilidad de desertar, por lo que actúa como una variable de confusión.

d) Podría concluir que las becas están reduciendo la deserción y decidir ampliar el programa sin analizar otros factores.
Sin embargo, los datos desagregados sugieren que existen diferencias importantes asociadas al nivel socioeconómico que deben considerarse antes de evaluar la efectividad de las becas.

e) No.
Estos datos muestran una asociación, pero no permiten establecer causalidad.

Para demostrar causalidad sería necesario contar con información adicional, por ejemplo:

- Estudios experimentales
- Asignación aleatoria de becas
- Información sobre rendimiento académico, apoyo familiar, empleo, motivación y otros factores relevantes

f) Se recomienda:

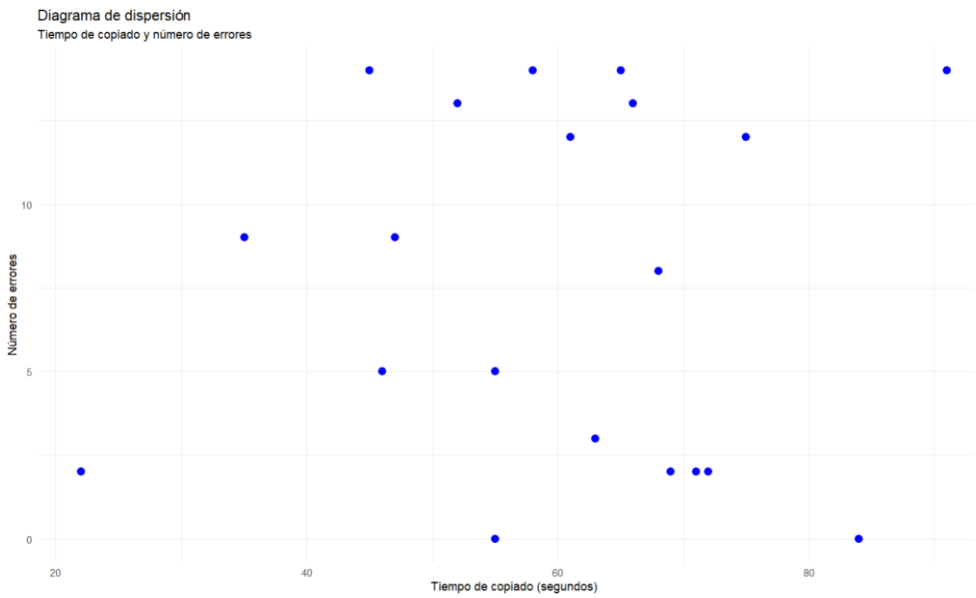
1. Analizar los datos por nivel socioeconómico y otros grupos relevantes.
2. Utilizar tasas ajustadas o modelos de regresión que controlen variables de confusión.
3. Comparar estudiantes con características similares
4. Evaluar la evolución de los estudiantes a lo largo del tiempo mediante estudios longitudinales.

De esta forma se obtendría una estimación más confiable del efecto real de las becas sobre la deserción escolar.

## 2.8 El problema de la asociación

### 2.8.1

a) Diagrama de dispersión




El diagrama de dispersión muestra una nube de puntos sin una tendencia clara.

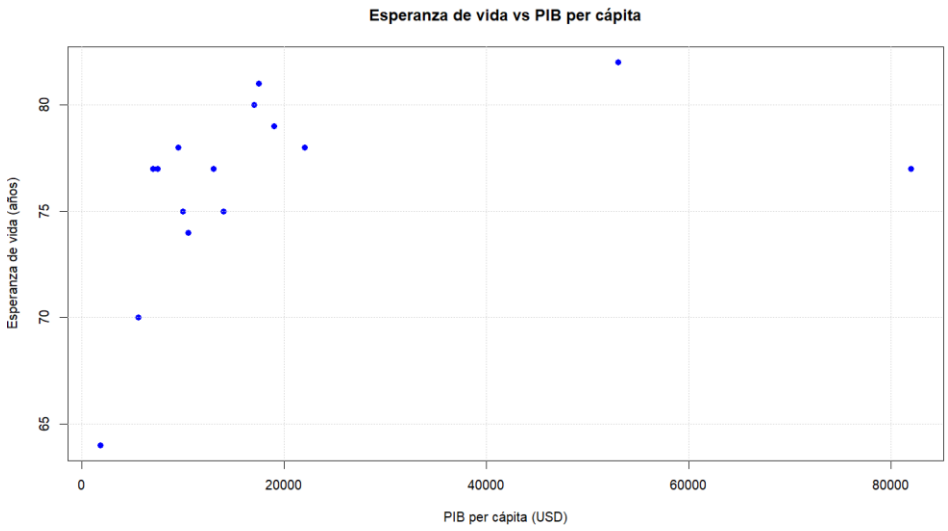
b) No se observa una relación evidente entre el tiempo de copiado y el número de errores. La asociación parece muy débil.

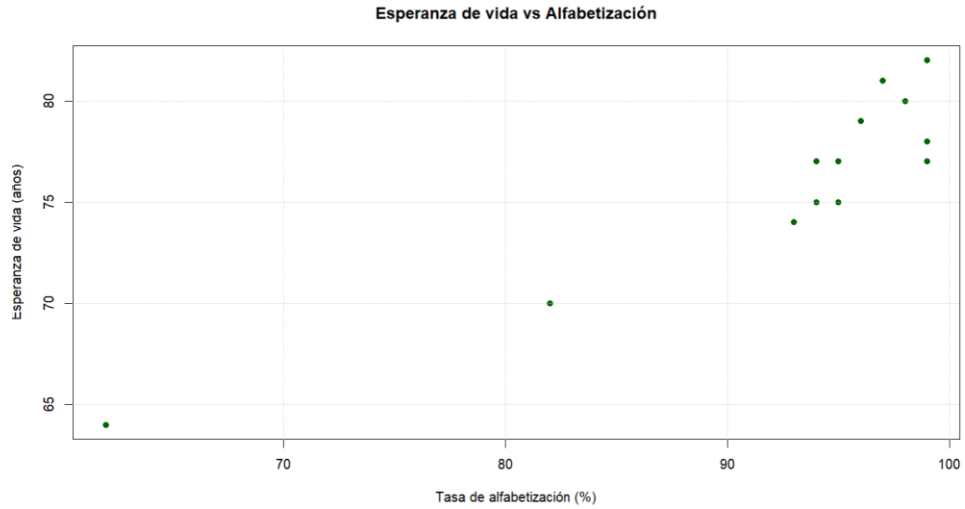
c) El coeficiente de correlación de Pearson es $r \approx 0.034$. Indica una correlación positiva prácticamente nula.
Interpretación: Existe una correlación lineal positiva prácticamente nula entre el tiempo de copiado y el número de errores.
Dentro del contexto del problema, el tiempo que tarda una secretaria en copiar un documento no parece estar relacionado con la cantidad de errores que comete.

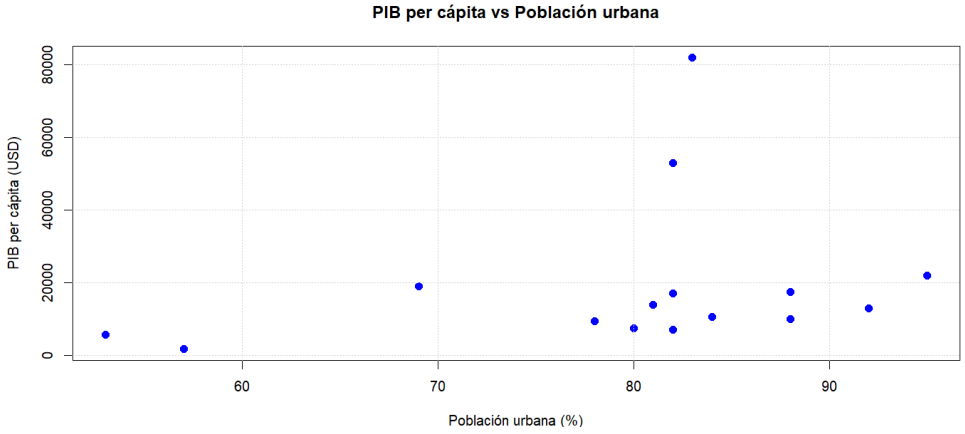
d) Se observan algunos casos inusuales, pero no parecen influir significativamente en la correlación general.

e) No. A partir de estos datos no puede concluirse que el tiempo de copiado cause cambios en el número de errores. La correlación únicamente mide el grado de asociación entre dos variables.

**2.8.2** a) Diagramas de dispersión






a)

- Esperanza de vida y PIB per cápita: asociación positiva moderada
- Alfabetización y esperanza de vida: asociación positiva fuerte
- Población urbana y PIB per cápita: asociación positiva moderada

b) Los tres pares muestran asociación positiva. La más fuerte parece ser entre alfabetización y esperanza de vida.

c) Haití, por sus bajos valores en todas las variables. También destacan Estados Unidos y Cuba por apartarse de algunas tendencias generales.

d) La tasa de alfabetización parece ser la variable más estrechamente relacionada con la esperanza de vida.

e) Canadá y Estados Unidos tienen PIB per cápita mucho más altos. Canadá también presenta una de las mayores esperanzas de vida. Ambos tienen alfabetización y urbanización elevadas.

f) La relación parece mostrar rendimientos decrecientes: a niveles altos de ingreso, grandes aumentos en PIB generan pequeñas mejoras en esperanza de vida.

g) No. La asociación no implica causalidad. Otros factores como educación, salud y condiciones de vida también pueden influir en la esperanza de vida.

**2.8.3**

a) Ingeniería. Presenta la mayor proporción de estudiantes con rendimiento alto, lo que sugiere un mejor desempeño académico general.

b) Administración. Tiene la mayor proporción de estudiantes con rendimiento bajo, por lo que podría requerir mayor apoyo académico.

c) Sí, Ingeniería. Tiene más estudiantes con rendimiento alto y menos con rendimiento bajo que las demás carreras.

d) Rendimiento medio. Es el nivel más frecuente en conjunto, indicando que la mayoría de los estudiantes tiene un desempeño intermedio.

e) Administración. Porque concentra más estudiantes con rendimiento bajo y menos con rendimiento alto.

f) Ingeniería. Presenta la mayor diferencia entre estudiantes de rendimiento alto y bajo, mostrando una clara predominancia del rendimiento alto.

## 2.9 Problema de comparación y problemas de asociación

**2.9.1**

a) Ingeniería. Presenta la mediana más alta de horas de estudio, lo que sugiere que sus estudiantes estudian más tiempo semanalmente.

b) Economía. Muestra la mayor variabilidad, pues presenta el rango total más amplio y una caja relativamente extensa.

c) No se observan valores atípicos evidentes en el diagrama.

d) Ingeniería. Tiene una mediana alta y una dispersión moderada, lo que sugiere hábitos de estudio más consistentes.

e) Administración. Sería la carrera prioritaria, ya que presenta la mediana más baja de horas de estudio.

Asociación entre horas de estudio y rendimiento académico

a) Rendimiento alto. Presenta la mediana más alta de horas de estudio.

b) Sí. A mayor número de horas de estudio, parece observarse un mejor rendimiento académico.

c) Presentan cierto traslape. Esto indica que las horas de estudio no explican completamente el rendimiento.

d) Sí, parecen importantes. Las medianas aumentan conforme mejora el nivel de rendimiento.

e) Rendimiento medio. Presenta la mayor variabilidad en las horas de estudio.

f) Sí. Esto sugiere que además de las horas de estudio influyen factores como habilidad, motivación, métodos de estudio o conocimientos previos.

g) El boxplot no muestra la relación individual entre variables. Un diagrama de dispersión permite observar mejor la asociación, tendencia y posibles patrones entre horas de estudio y rendimiento académico.

**2.9.2**

a) Producción. Presenta la mayor proporción de empleados con estrés alto, lo que podría indicar mayores exigencias o presión laboral.

b) Finanzas. Presenta la mayor proporción de empleados con estrés bajo, sugiriendo condiciones laborales relativamente más favorables.

c) Sí, Producción. Se distingue por concentrar una mayor proporción de empleados con estrés alto y una menor proporción con estrés bajo.

d) Se observa que los niveles de estrés varían entre departamentos. Producción muestra más estrés alto, mientras que Finanzas presenta más empleados con estrés medio y bajo.

e) Producción. Debería ser el departamento prioritario, ya que concentra la mayor proporción de empleados con estrés alto.

f) Finanzas. Parece presentar la situación más favorable, pues tiene menos empleados con estrés alto y más empleados con estrés bajo y medio.

**2.9.3**

a) Producción. Presenta la mediana más alta de días de ausencia, lo que indica un mayor nivel típico de ausentismo.

b) Ventas. Muestra la mayor variabilidad, ya que presenta el rango total más amplio y brazos más largos.

c) Sí, en Finanzas. Se observa un valor atípico alto, que podría corresponder a un empleado con circunstancias excepcionales de ausencia.

d) Finanzas. Presenta una mediana relativamente baja y menor dispersión, lo que sugiere mejor asistencia.

e) Producción. Debería priorizarse porque tiene la mayor mediana de ausencias.

f) Las diferencias existen, pero son moderadas. Las distribuciones presentan cierto traslape entre departamentos.

g) Ventas. Parece mostrar mayor asimetría positiva, identificable por un brazo superior más largo que el inferior y una mayor extensión hacia valores altos.

**2.9.4**

a) Estrés alto. Presenta la mediana más alta de días de ausencia, indicando un mayor ausentismo típico.

b) Sí. A medida que aumenta el nivel de estrés, tienden a aumentar los días de ausencia.

c) Las distribuciones presentan cierto traslape. Esto implica que el estrés influye en el ausentismo, pero no lo determina completamente.

d) Sí. Las medianas aumentan conforme aumenta el nivel de estrés, sugiriendo una relación positiva entre ambas variables.

e) Estrés alto. Presenta la mayor variabilidad en los días de ausencia, indicando comportamientos más diversos dentro del grupo.

f) Sí. Puede observarse por la amplitud de las cajas y los brazos, que muestran empleados con ausencias muy distintas dentro del mismo nivel de estrés.

g) Estrés bajo. Presenta la menor dispersión, por lo que su distribución parece más homogénea.

h) No completamente. Los diagramas sugieren una relación, pero el traslape entre distribuciones indica que también intervienen otros factores además del nivel de estrés.

## Verdadero o falso

| # | V/F | Justificación |
|---|---|---|
| 1 | **V** | La media es una medida de tendencia central; no describe directamente los valores máximo y mínimo. |
| 2 | **V** | Las clases deben cubrir todos los datos exhaustivas y cada observación debe pertenecer a una sola clase mutuamente excluyentes. |
| 3 | **V** | El polígono de frecuencias relativas utiliza frecuencias relativas y representa proporciones. |
| 4 | **V** | Una muestra representativa reproduce adecuadamente las características relevantes de la población. |
| 5 | **F** | No siempre es posible reconstruir exactamente el histograma a partir de un polígono si se desconoce la amplitud de las clases u otra información. |
| 6 | **F** | Un conjunto puede ser bimodal, multimodal o no tener moda. |
| 7 | **V** | El coeficiente de variación es una medida relativa porque compara la desviación estándar con la media. |
| 8 | **F** | Las variables discretas toman valores contables, que no necesariamente tienen que ser enteros por ejemplo, incrementos de 0.5. |

| # | V/F | Justificación |
|---|---|---|
| 9 | **F** | En distribuciones simétricas unimodales pueden coincidir exactamente. |
| 10 | **V** | Generalmente se utiliza la media aritmética para promediar porcentajes comparables. |
| 11 | **F** | Las medidas de dispersión y la media no pueden calcularse para variables cualitativas nominales. |
| 12 | **F** | Si hay un número par de observaciones, la mediana puede ser el promedio de dos valores y no coincidir con ningún dato observado. |
| 13 | **V** | En distribuciones sesgadas a la derecha, la cola derecha arrastra la media hacia valores mayores. |
| 14 | **V** | La media y la desviación estándar no determinan completamente la forma de una distribución. |
| 15 | **V** | Sumar una constante desplaza los datos, pero no modifica las distancias respecto a la media. |
| 16 | **V** | El IQR se basa en el 50% central de los datos y generalmente no se ve afectado por valores extremos outliers. |
| 17 | **V** | La mediana puede calcularse para variables cualitativas con escala ordinal y variables cuantitativas; la media solo puede calcularse para variables cuantitativas. |
| 18 | **V** | Si CV > 100%, entonces $s > \bar{x}$ |
| 19 | **F** | Una distribución bimodal puede surgir por diversas razones y no necesariamente por mezcla de poblaciones. |
| 20 | **V** | En una distribución perfectamente simétrica, media y mediana coinciden. |
| 21 | **F** | Una correlación cercana a cero solo indica ausencia de relación lineal; puede existir una relación no lineal. |
| 22 | **F** | El histograma se usa para variables cuantitativas continuas; el diagrama de barras para categorías o valores discretos. |
| 23 | **V** | Las frecuencias relativas representan proporciones y deben sumar 1 o 100%. |
| 24 | **V** | La media es sensible a valores extremos; la mediana es mucho más resistente. |
| 25 | **V** | Si la desviación estándar es cero, no existe variabilidad: todos los valores son iguales. |
| 26 | **V** | Esa es precisamente la definición de muestra aleatoria simple. |
| 27 | **F** | Correlación no implica causalidad. Puede haber variables de confusión. |
| 28 | **F** | El percentil 90 es un valor por debajo del cual se encuentra al menos el 90% de las observaciones; no necesariamente exactamente el 90%. |

| # | V/F | Justificación |
|---|---|---|
| 29 | **V** | La distribución uniforme reparte los datos de manera homogénea, mientras que la normal concentra más observaciones alrededor de la media. |
| 30 | **V** | En una asociación negativa, cuando una variable aumenta, la otra tiende a disminuir. |
| 31 | **F** | Un valor atípico puede ser una observación válida y no necesariamente un error. |
| 32 | **V** | Es una propiedad fundamental de la media aritmética: $\sum x_i - \bar{x} = 0$. |

# Tema III – Probabilidad

## 3.1 Conjuntos y Conteo

### 3.1.1 Espacio muestral

a) Sea C = cara y X = cruz

S = {(C, C, C), (C, C, X), (C, X, C), (C, X, X), (X, C, C), (X, C, X), (X, X, C), (X, X, X)}

b) S = {(P, P, P, P), (P, P, P, N), (P, P, N, P), (P, P, N, N), (P, N, P, P), (P, N, P, N), (P, N, N, P), (P, N, N, N), (N, P, P, P), (N, P, P, N), (N, P, N, P), (N, P, N, N), (N, N, P, P), (N, N, P, N), (N, N, N, P), (N, N, N, N)}

c) S = {(R, B), (R, A), (B, B), (B, A), (A, B), (A, A)}

d) S = {(R, R), (R, B), (R, A), (B, R), (B, B), (B, A), (A, R), (A, B), (A, A)}

Aunque la extracción es sin reemplazo, (R,R), (B,B) y (A,A) sí son posibles porque después de mezclar las urnas hay 3 bolas rojas, 4 blancas y 4 azules.

e) Sea C = cara y X = cruz

S = {(C, 1), (C, 2), (C, 3), (C, 4), (C, 5), (C, 6), (X, 1), (X, 2), (X, 3), (X, 4), (X, 5), (X, 6)}

f) S = {(i, j): i, j ∈ {1, 2, 3, 4, 5, 6}}

El espacio muestral contiene 36 resultados posibles

g) Sea M = mujer y H = hombre

S = {(M, M, M), (M, M, H), (M, H, M), (M, H, H), (H, M, M), (H, M, H), (H, H, M), (H, H, H)}

h) S = {(S, S, S, S), (S, S, S, B), (S, S, B, S), (S, S, B, B), (S, B, S, S), (S, B, S, B), (S, B, B, S), (S, B, B, B), (B, S, S, S), (B, S, S, B), (B, S, B, S), (B, S, B, B), (B, B, S, S), (B, B, S, B), (B, B, B, S), (B, B, B, B)}

i) S = {(A, A, A), (A, A, R), (A, R, A), (A, R, R), (R, A, A), (R, A, R), (R, R, A), (R, R, R)}

j) S = {(R, R, R, R), (R, R, R, N), (R, R, N, R), (R, R, N, N), (R, N, R, R), (R, N, R, N), (R, N, N, R), (R, N, N, N), (N, R, R, R), (N, R, R, N), (N, R, N, R), (N, R, N, N), (N, N, R, R), (N, N, R, N), (N, N, N, R), (N, N, N, N)}

**3.1.2** Denotemos los seis resultados posibles por 1, 2, 3, 4, 5 y 6.

a) Definición de los eventos

Ω = {1, 2, 3, 4, 5, 6}

A = {el cliente recibe la bebida correcta} = {1, 2}

B = {el cliente recibe alguna bebida} = {1, 2, 3, 4}

C = {existe algún problema en el servicio} = {2, 3, 4, 5, 6}

D = {el cliente recibe dos bebidas simultáneamente} = ∅

E = {el cliente recibe la bebida correcta pero no de forma perfecta} = {2}

$A \cup B = \{1, 2, 3, 4\}$

$A \cup C = \{1, 2, 3, 4, 5, 6\} = \Omega$

$A \cap B = \{1, 2\} = A$

$A \cap C = \{2\} = E$

$A \cup B \cup C = \{1, 2, 3, 4, 5, 6\} = \Omega$

$A^c = \{3, 4, 5, 6\}$

$E^c = \{1, 3, 4, 5, 6\}$

$E \subset A$, ya que todo resultado que pertenece a E también pertenece a A

b) Probabilidad de los eventos

Como los seis resultados son igualmente probables:

$P(\Omega) = 1$

$P(A) = 2/6 = 1/3$

$P(B) = 4/6 = 2/3$

$P(C) = 5/6$

$P(D) = 0$

$P(E) = 1/6$

$P(A \cup B) = 4/6 = 2/3$

$P(A \cup C) = 6/6 = 1$

$P(A \cap B) = 2/6 = 1/3$

$P(A \cap C) = 1/6$

$P(A \cup B \cup C) = 6/6 = 1$

$P(A^c) = 4/6 = 2/3$

$P(E^c) = 5/6$

3.1.3

a) Espacio muestral

$\Omega = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6),$
$(2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6),$
$(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6),$
$(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6),$
$(5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6),$
$(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)\}$

El espacio muestral contiene 36 resultados igualmente probables.

b) Evento A = {al menos uno de los dados muestra un 1}

$A = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1)\}$

c) Evento B = {la suma de los puntos obtenidos en ambos dados es igual a 7}

$B = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$

d) Evento C = {el número de puntos obtenido en el dado 1 es menor que el obtenido en el dado 2}

C = {(1, 2), (1, 3), (1, 4), (1, 5), (1, 6),
(2, 3), (2, 4), (2, 5), (2, 6),
(3, 4), (3, 5), (3, 6),
(4, 5), (4, 6),
(5, 6)}

e) Probabilidades de A, B y C

P(A) = 11/36

P(B) = 6/36 = 1/6

P(C) = 15/36 = 5/12

f) Intersecciones, uniones y complementos

$A \cap B$ = {(1, 6), (6, 1)}

$A \cup B$ = {(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 5), (3, 1), (3, 4), (4, 1),
(4, 3), (5, 1), (5, 2), (6, 1)}

$A \cap C$ = {(1, 2), (1, 3), (1, 4), (1, 5), (1, 6)}

$B \cap C$ = {(1, 6), (2, 5), (3, 4)}

$A^c$ = {(2, 2), (2, 3), (2, 4), (2, 5), (2, 6),
(3, 2), (3, 3), (3, 4), (3, 5), (3, 6),
(4, 2), (4, 3), (4, 4), (4, 5), (4, 6),
(5, 2), (5, 3), (5, 4), (5, 5), (5, 6),
(6, 2), (6, 3), (6, 4), (6, 5), (6, 6)}

$C^c$ = {(1, 1),
(2, 1), (2, 2),
(3, 1), (3, 2), (3, 3),
(4, 1), (4, 2), (4, 3), (4, 4),
(5, 1), (5, 2), (5, 3), (5, 4), (5, 5),
(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)}

g) Probabilidades

$P(A \cap B) = 2/36 = 1/18$

$P(A \cup B) = 15/36 = 5/12$

$P(A \cap C) = 5/36$

$P(B \cap C) = 3/36 = 1/12$

$P(A^c) = 25/36$

$P(C^c) = 21/36 = 7/12$

**3.1.4**

a) Espacio muestral

$\Omega = \{(A, A), (A, E), (A, I), (E, A), (E, E), (E, I), (I, A), (I, E), (I, I)\}$

El espacio muestral contiene 9 resultados igualmente probables.

b) Evento A = {al menos uno de los candidatos proviene de Administración}

$A = \{(A, A), (A, E), (A, I), (E, A), (I, A)\}$

c) Evento B = {los candidatos provienen de áreas distintas}

$B = \{(A, E), (A, I), (E, A), (E, I), (I, A), (I, E)\}$

d) Evento C = {el primer candidato proviene de Economía}

$C = \{(E, A), (E, E), (E, I)\}$

e) Probabilidades

$P(A) = 5/9$

$P(B) = 6/9 = 2/3$

$P(C) = 3/9 = 1/3$

**3.2 Incertidumbre, fenómenos aleatorios y probabilidad**

**3.2.1**

a) Aleatorio. Aunque se conocen los resultados posibles y sus probabilidades, no puede predecirse con certeza el resultado de un lanzamiento individual.

b) Mixto. Depende de factores económicos parcialmente predecibles, pero también de eventos inciertos e inesperados.

c) Mixto. Existe una hora programada y factores operativos predecibles, pero pueden presentarse retrasos por causas inciertas.

d) Mixto. El precio depende de factores económicos observables, pero también de noticias, decisiones y eventos impredecibles.

e) Mixto. El tipo de cambio responde a factores económicos y financieros identificables, junto con acontecimientos inciertos.

f) Mixto. La inflación presenta tendencias y factores económicos predecibles, pero también está sujeta a perturbaciones inesperadas.

g) Aleatorio. El resultado depende de los individuos seleccionados al azar y de sus respuestas.

h) Mixto. La temperatura sigue patrones meteorológicos y estacionales, pero su valor exacto depende de condiciones inciertas.

i) Aleatorio. No es posible predecir con certeza la fecha y hora exactas de un sismo futuro.

j) Aleatorio. Depende del comportamiento impredecible de los usuarios de la red social.

k) Mixto. Depende de factores predecibles, como la capacidad del hospital, y de factores inciertos, como el número y gravedad de los pacientes que llegan.

l) Mixto. La lluvia depende de condiciones meteorológicas observables, pero la cantidad exacta presenta incertidumbre.

m) Mixto. Las ventas presentan patrones asociados con el día, horario y demanda habitual, pero también dependen de factores impredecibles.

n) Mixto. La temperatura sigue patrones físicos y meteorológicos, pero su valor máximo exacto depende de condiciones variables.

o) Aleatorio. No puede determinarse con certeza el número exacto de rayos que ocurrirán durante una tormenta.

p) Determinístico. La hora del amanecer puede calcularse con gran precisión a partir de leyes astronómicas y de la ubicación geográfica.

q) Mixto. Existen tendencias y modelos físicos para proyectarlo, pero también intervienen factores futuros inciertos.

r) Aleatorio. No puede predecirse con certeza si un rayo caerá sobre un punto específico.

s) Aleatorio. Aunque existen patrones estadísticos en la ocurrencia de réplicas, no puede conocerse con certeza su número exacto.

t) Mixto. Depende de factores parcialmente predecibles, como el desempeño previo y la dificultad del examen, y de factores inciertos que afectan los resultados individuales.

u) Mixto. Depende de factores predecibles, como el número de empleados y el sistema de atención, y de factores aleatorios, como la llegada de otros clientes y la duración de cada servicio.

**3.2.2**

a) Falsa. Una alta precisión en la predicción no implica que el fenómeno sea determinístico.

b) Verdadera. Un fenómeno aleatorio puede presentar regularidades y patrones probabilísticos.

c) Verdadera. Un fenómeno mixto combina componentes predecibles y componentes inciertos.

d) Falsa. Los fenómenos económicos pueden contener componentes sistemáticos y aleatorios.

e) Verdadera. La probabilidad permite cuantificar el grado de incertidumbre asociado con un evento.

f) Verdadera. En un fenómeno aleatorio, los distintos resultados pueden tener probabilidades diferentes.

g) Verdadera. Una predicción muy precisa no garantiza que el resultado observado coincida con ella.

h) Verdadera. Si el resultado pudiera conocerse con certeza antes de ocurrir, el fenómeno no sería aleatorio.

i) Verdadera. En un fenómeno completamente determinístico, el resultado queda determinado cuando se conocen todas las condiciones relevantes.

j) Falsa. Una probabilidad de lluvia de 80% implica que existe un 20% de probabilidad de que no llueva; un solo resultado no permite concluir que el pronóstico fue incorrecto.

### 3.3 Enfoques de probabilidad: clásico o de Laplace, frecuentista y subjetivo

**3.3.1**

a) Clásico o de Laplace. Los seis resultados son igualmente probables y tres de ellos son números pares.

P(número par) = 3/6 = 1/2

b) Frecuentista. La probabilidad se estima a partir de la frecuencia relativa observada en la producción.

P(defectuoso) = 1,230/10,000 = 0.123

c) Subjetivo. La probabilidad se basa en el juicio del analista, considerando la información disponible y su experiencia.

d) Clásico o de Laplace. Las 52 cartas son igualmente probables y cuatro de ellas son ases.

P(as) = 4/52 = 1/13

e) Subjetivo. La probabilidad se obtiene a partir de modelos, observaciones y el juicio profesional del meteorólogo.

f) Clásico o de Laplace. Los 200 boletos tienen la misma probabilidad de resultar ganadores.

P(ganar) = 1/200 = 0.005

g) Frecuentista. La probabilidad se estima a partir de la frecuencia relativa de tiros anotados.

P(anotar) = 72/100 = 0.72

h) Subjetivo. La probabilidad se basa en la experiencia de los fundadores y en su evaluación de las condiciones del mercado y la competencia.

### 3.4 Axiomas de probabilidad

**3.4.1** Datos:

- P(N) = 0.62
- P(PV) = 0.41
- P(N ∪ PV) = 0.78

a) P(N ∩ PV) = 0.25

b) $P(N \cap PV^c) = 0.37$

c) $P(PV \cap N^c) = 0.16$

d) $P(N \cup PV^c) = 0.22$

e) Regiones del diagrama de Venn:

Solo Netflix: 0.37

Netflix y Prime Video: 0.25

Solo Prime Video: 0.16

Ninguna de las dos plataformas: 0.22

f) Verificación: 0.37 + 0.25 + 0.16 + 0.22 = 1

**3.4.2** Datos:

- $P(B) = 0.25$
- $P(T) = 0.68$
- $P(B \cap T) = 0.12$

a) $P(B \cap T^c) = 0.13$

b) $P(T \cap B^c) = 0.56$

c) $P(B \cup T) = 0.81$

d) Verificación de la regla de adición:

$P(B \cup T) = P(B) + P(T) – P(B \cap T)$

$0.81 = 0.81$ ✓

e) Regiones del diagrama de Venn:

Solo bicicleta: 0.13

Bicicleta y transporte público: 0.12

Solo transporte público: 0.56

Ninguno de los dos medios de transporte: $1 - 0.81 = 0.19$

**3.4.3** Datos:

- $P(P) = 0.45$
- $P(M) = 0.30$

- $P((P \cup M)^c) = 0.40$

a) $P(P \cup M) = 0.60$

b) $P(P \cap M) = 0.15$

c) $P(P \cap M^c) = 0.30$

d) $P(M \cap P^c) = 0.15$

e) Verificación:

Solo phishing + solo malware + ambos ataques + ninguno

$0.30 + 0.15 + 0.15 + 0.40 = 1$

Verificación de propiedades

f) Axioma de no negatividad

Todas las probabilidades calculadas son mayores o iguales a cero:

g) Axioma de normalización

$P(\Omega) = 1$ ✓

h) Regla del complemento

$P(P \cup M) + P(P \cup M)^c = 1$ ✓

i) Regla de adición para eventos no mutuamente excluyentes

$P(P \cup M) = P(P) + P(M) – P(P \cap M)$

$0.60 = 0.45 + 0.30 − 0.15$

$0.60 = 0.60$ ✓

j) Los eventos P y M no son mutuamente excluyentes, ya que:

$P(P \cap M) = 0.15 \neq 0$

Por lo tanto, una empresa puede haber sufrido ambos tipos de ataques.

k) Regiones del diagrama de Venn:

Solo phishing: 0.30

Phishing y malware: 0.15

Solo malware: 0.15

Ninguno de los dos ataques: 0.40

## 3.5 Probabilidad conjunta de eventos. Probabilidad marginal, condicional e independencia. Regla del producto

**3.5.1**

a) $P(C) = 0.40$

b) $P(F) = 0.30$

c) $P(C \mid F) = 0.60$

d) $P(F \mid C) = 0.45$

e) Como $P(C \cap F) \neq P(C)P(F)$, los eventos C y F no son independientes.

f) El uso de cupones y la frecuencia de compra están asociados. Conocer si un cliente utiliza cupones proporciona información sobre la probabilidad de que realice 3 o más pedidos por semana.

g) Como $P(F \mid C) = 0.45 > P(F \mid C^c) = 0.20$, el uso de cupones parece estar asociado con una mayor frecuencia de pedidos. Entre quienes usan cupones, 45% realiza 3 o más pedidos por semana, mientras que entre quienes no los usan, solo 20% lo hace.

**3.5.2** Sean:

X = {el estudiante apoya a la candidata X}

S = {el estudiante está satisfecho}

a) $P(X^c \cap S^c) = 0.36$

b) $P(X \cup S) = 0.64$

c) $P(S \mid X) = 0.5556$

d) $P(X \mid S) = 0.75$

e) Como $P(X \cap S) \neq P(X)P(S)$ los eventos no son independientes.

**3.5.3** Sean:

S = {ve series}

D = {ve documentales}

a) $P(S) = 0.70$

b) P(D) = 0.52

c) P(D | S) = 0.60

d) P(S ∩ D) = 0.42

e) Como P(S ∩ D) ≠ P(S)P(D), los eventos no son independientes.

**3.5.4**

a) P(> 8 h ∩ alto estrés) = 0.17

b) P(alto estrés) = 0.35

c) P(alto estrés | < 6 h) = 0.3889

d) P(> 8 h | estrés medio) = 0.35

e) P(alto estrés | < 6 h) = 0.07 / 0.18 = 0.3889

P(alto estrés | 6–8 h) = 0.11 / 0.43 = 0.2558

P(alto estrés | > 8 h) = 0.17 / 0.39 = 0.4359

No se observa que dormir más horas reduzca sistemáticamente la probabilidad condicional de alto estrés. La probabilidad disminuye al pasar de menos de 6 horas a entre 6 y 8 horas, pero aumenta para quienes duermen más de 8 horas.

**3.5.5**

a) P(> 30 ∩ alta) = 0.25

b) P(remoto) = 0.30

c) P(baja | presencial) = 0.2903

d) P(remoto | alta) = 0.3256

e) La satisfacción y la modalidad no son independientes.

**3.5.6** a) Tabla de frecuencias conjuntas

| Sexo / Edad | 18 | 19–24 | 25–29 | 30–34 | 35–39 | 40–44 | 45–49 | 50–54 | 55–59 | 60–64 | 65 y más | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mujer | 3 | 7 | 7 | 6 | 5 | 5 | 4 | 4 | 3 | 3 | 5 | **52** |
| Hombre | 2 | 6 | 6 | 6 | 5 | 5 | 4 | 4 | 3 | 3 | 4 | **48** |
| **Total** | **5** | **13** | **13** | **12** | **10** | **10** | **8** | **8** | **6** | **6** | **9** | **100** |

Tabla de probabilidades conjuntas

| Sexo / Edad | 18 | 19–24 | 25–29 | 30–34 | 35–39 | 40–44 | 45–49 | 50–54 | 55–59 | 60–64 | 65 y más | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mujer | 0.03 | 0.07 | 0.07 | 0.06 | 0.05 | 0.05 | 0.04 | 0.04 | 0.03 | 0.03 | 0.05 | **0.52** |
| Hombre | 0.02 | 0.06 | 0.06 | 0.06 | 0.05 | 0.05 | 0.04 | 0.04 | 0.03 | 0.03 | 0.04 | **0.48** |
| **Total** | **0.05** | **0.13** | **0.13** | **0.12** | **0.10** | **0.10** | **0.08** | **0.08** | **0.06** | **0.06** | **0.09** | **1.00** |

b) $P(M) = \frac{52}{100} = 0.52$

c) $P(18) = \frac{5}{100} = 0.05$

d) $P(H \cap 25\text{–}29) = \frac{6}{100} = 0.06$

e) $P(M \cap 65+) = \frac{5}{100} = 0.05$

f) $P(M \mid 30\text{–}34) = \frac{P(M \cap 30\text{–}34)}{P(30\text{–}34)} = \frac{6}{12} = 0.50$

g) $P(40\text{–}44 \mid H) = \frac{P(H \cap 40\text{–}44)}{P(H)} = \frac{5}{48} \approx 0.1042$

h) Para que sean independientes debe cumplirse:

$$P(M \cap 18) = P(M)P(18)$$

$$0.03 \neq 0.026$$

Los eventos “Ser mujer” y “tener 18 años” no son independientes.

i) ¿"Ser hombre" y "tener 65 años o más" son independientes?

$$P(H \cap 65+) = \frac{4}{100} = 0.04$$

Por otro lado:

$$P(H)P(65+) = (0.48)(0.09) = 0.0432$$

Como:

$$0.04 \neq 0.0432$$

Los eventos "Ser hombre" y "tener 65 años o más" no son independientes.

j) Regla del producto. Queremos comprobar:

$$P(M \cap 25\text{–}29) = P(M)P(25\text{–}29 \mid M)$$

De la tabla:

$$P(M) = \frac{52}{100} \text{ y } P(25\text{–}29 \mid M) = \frac{7}{52}$$

Entonces:

$$P(M)P(25\text{–}29 \mid M) = \frac{52}{100}\left(\frac{7}{52}\right) = \frac{7}{100} = 0.07$$

Por otro lado:

$$P(M \cap 25\text{–}29) = \frac{7}{100} = 0.07$$

Por lo tanto:

$$P(M \cap 25\text{–}29) = P(M)P(25\text{–}29 \mid M) = 0.07$$

Se verifica la regla del producto.

k) $P(M \cap 19\text{–}24) = \frac{7}{100} = 0.07$

**3.6 Partición del espacio muestral. Regla de probabilidad total y Teorema de Bayes**

**3.6.1** Sean:

M = {el empleado es mujer}

H = {el empleado es hombre}

L = {el empleado ocupa un puesto de liderazgo}

Datos:

- P(M) = 0.40
- P(H) = 0.60
- P(L | M) = 0.30
- P(L | H) = 0.20

a) Por la regla de probabilidad total:

P(L) = P(L | M)P(M) + P(L | H)P(H) = 0.24

b) Por el Teorema de Bayes:

P(M | L) = P(L | M)P(M) / P(L) = 0.50

c) La afirmación no puede justificarse únicamente con P(M | L) = 0.50. Esta probabilidad indica que 50% de las personas en puestos de liderazgo son mujeres, pero no muestra por sí sola las oportunidades relativas de liderazgo de cada género.

Las mujeres representan 40% de los empleados, pero 50% de los líderes. Además:

P(L | M) = 0.30

P(L | H) = 0.20

Por lo tanto, la proporción de mujeres que ocupa puestos de liderazgo es mayor que la proporción correspondiente entre los hombres. Para evaluar la equidad deben analizarse conjuntamente la composición de la población, la composición del grupo de líderes y las probabilidades condicionales de alcanzar un puesto de liderazgo.

d) $P(L^c) = 1 - P(L) = 0.76$

e) $P(H \mid L^c) = 0.6316$

f) P(L | M) = 0.30

P(L | H) = 0.20

Las mujeres tienen una mayor proporción de liderazgo dentro de su grupo, ya que 30% de las mujeres ocupa puestos de liderazgo, frente a 20% de los hombres.

g) $P(M^c \cap L^c) = P(H \cap L^c) = P(L^c \mid H)P(H) = 0.48$

**3.6.2** Sean:

A = {alta adopción de energías renovables}

B = {baja adopción de energías renovables}

R = {el país reduce sus emisiones de $CO_2$}

Datos:

- P(A) = 0.40
- P(B) = 0.60
- P(R | A) = 0.70
- P(R | B) = 0.30

a) Por la regla de probabilidad total:

$P(R) = P(R \mid A)P(A) + P(R \mid B)P(B) = 0.46$

b) Por el Teorema de Bayes:

$P(A \mid R) = P(R \mid A)P(A) / P(R) = 0.6087$

c) $P(B \mid R^c) = P(R^c \mid B)P(B) / P(R^c) = 0.7778$

d) $P(R^c \mid B) = 1 - P(R \mid B) = 1 - 0.30 = 0.70$

e) $P(A \mid R^c) = P(R^c \mid A)P(A) / P(R^c) = 0.2222$

f) El argumento del diplomático no está respaldado por los datos. Aunque el 30% de los países con baja adopción logra reducir sus emisiones P(R | B) = 0.30, entre los países con alta adopción la proporción es P(R | A) = 0.70

Por lo tanto, la probabilidad de reducir emisiones es más del doble entre los países con alta adopción. Además, entre los países que no redujeron sus emisiones, 77.78% pertenece al grupo de baja adopción.

Los datos no demuestran que las energías renovables sean la única causa posible de la reducción de emisiones, pero sí muestran una asociación clara entre una mayor adopción de energías renovables y una mayor probabilidad de reducirlas.

**3.6.3** Sean:

V = {la persona está vacunada}

E = {la persona enferma de influenza}

Datos:

- P(V)=0.60

- $P(V^c) = 0.40$
- $P(E \mid V) = 0.05$
- $P(E \mid V^c) = 0.20$

a) Por la regla de probabilidad total:

$P(E) = P(E \mid V)P(V) + P(E \mid V^c)P(V^c) = 0.11$

b) La probabilidad total de enfermar en la población es 11%. La vacunación reduce el riesgo individual de enfermar de 20% entre los no vacunados a 5% entre los vacunados.

c) Por el Teorema de Bayes:

$P(V^c \mid E) = P(E \mid V^c)P(V^c) / P(E) = 0.7273$

d) Si la cobertura aumenta a 80%:

$P(V) = 0.80$

$P(V^c) = 0.20$

$P(E) = (0.05)(0.80) + (0.20)(0.20) = 0.08$

La probabilidad total de enfermar disminuye de 11% a 8%.

e) Con la cepa más virulenta:

$P(E \mid V) = 0.10$

$P(E \mid V^c) = 0.30$

Manteniendo la cobertura original de vacunación de 60%:

$P(E) = 0.18$

La probabilidad total de enfermar aumenta de 11% a 18%.

f) Sean:

$p$ = proporción de la población vacunada

$rV$ = probabilidad de enfermar estando vacunado

$rNV$ = probabilidad de enfermar sin estar vacunado

Entonces:

$P(E) = p(rV) + (1 - p)(rNV)$

Por lo tanto, la función general es:

$f(p) = p(rV) + (1 - p)(rNV)$

Para este problema:

$f(p) = 0.05p + 0.20(1 - p) = 0.20 - 0.15p$

Utilizaría un diagrama de dispersión con línea, colocando la cobertura de vacunación p en el eje horizontal y la probabilidad total de enfermar en el eje vertical. El gráfico mostraría cómo cambia el riesgo poblacional a medida que aumenta la cobertura de vacunación.

**3.6.4** Datos:

- $P(C_1) = 0.80$
- $P(C_2) = 0.15$
- $P(C_3) = 0.05$
- $P(S \mid C_1) = 0.02$
- $P(S \mid C_2) = 0.40$
- $P(S \mid C_3) = 0.90$

a) Por la regla de probabilidad total:

$P(S) = P(S \mid C_1)P(C_1) + P(S \mid C_2)P(C_2) + P(S \mid C_3)P(C_3) = 0.121$

b) Por el Teorema de Bayes:

$P(C_3 \mid S) = P(S \mid C_3)P(C_3) / P(S) = 0.3719$

c) $P(C_1 \mid S) = P(S \mid C_1)P(C_1) / P(S) = 0.1322$

El algoritmo detecta 90% de las transacciones de alto riesgo, pero aproximadamente una de cada ocho alertas corresponde a una transacción habitual. Su eficiencia operativa debe evaluarse considerando el costo de estos falsos positivos frente al beneficio de detectar operaciones de alto riesgo.

**3.6.5** Datos:

- $P(A_1) = 0.30$
- $P(A_2) = 0.50$
- $P(A_3) = 0.20$
- $P(S \mid A_1) = 0.05$
- $P(S \mid A_2) = 0.40$
- $P(S \mid A_3) = 0.85$

a) Por la regla de probabilidad total:

$P(S) = P(S \mid A_1)P(A_1) + P(S \mid A_2)P(A_2) + P(S \mid A_3)P(A_3) = 0.385$

b) Por el Teorema de Bayes:

$P(A_3 \mid S) = P(S \mid A_3)P(A_3) / P(S) = 0.4416$

c) $P(A_3) = 0.20$

$P(A_3 \mid S) = 0.4416$

Aunque los consumidores polarizados representan solo 20% de la población, constituyen aproximadamente 44.16% de quienes comparten la noticia falsa.

Estadísticamente, observar que una persona compartió la noticia modifica la probabilidad de que pertenezca al grupo polarizado, debido a que $P(S \mid A_3) = 0.85$ es mucho mayor que las probabilidades de compartir en los otros grupos.

d) i. Con las nuevas proporciones:

- $P(A_1) = 0.55$
- $P(A_2) = 0.25$
- $P(A_3) = 0.20$

La nueva probabilidad total es:

$P(S) = 0.2975$

La probabilidad de compartir la noticia falsa disminuye de 38.5% a 29.75%.

d) ii. La política es efectiva porque reduce la probabilidad total de compartir la noticia falsa en $0.385 - 0.2975 = 0.0875$, es decir, 8.75 puntos porcentuales.

Sin embargo, su efectividad es limitada mientras el grupo polarizado permanezca intacto. Este grupo representa solo 20% de la población, pero aporta $(0.20)(0.85) = 0.17$ a la probabilidad total de compartir. Por lo tanto, después de la política, los consumidores polarizados siguen siendo responsables de una parte importante de la difusión.

**3.6.6** Datos:

- $P(C_1) = 0.75$
- $P(C_2) = 0.20$
- $P(C_3) = 0.05$
- $P(A \mid C_1) = 0.02$
- $P(A \mid C_2) = 0.30$
- $P(A \mid C_3) = 0.90$

a) Por la regla de probabilidad total:

$P(A) = P(A \mid C_1)P(C_1) + P(A \mid C_2)P(C_2) + P(A \mid C_3)P(C_3) = 0.120$

b) Por el Teorema de Bayes:

$P(C_3 \mid A) = P(A \mid C_3)P(C_3) / P(A) = (0.90)(0.05) / 0.12 = 0.375$

c) $P(C_1 \mid A) = P(A \mid C_1)P(C_1) / P(A) = (0.02)(0.75) / 0.12 = 0.125$

Desde la perspectiva de la gestión de recursos, 12.5% de las auditorías generadas por el sistema se dirigiría a empresas cumplidas. La aceptación de este porcentaje depende del costo de una auditoría innecesaria y del beneficio de detectar empresas evasoras. El resultado sugiere que existe margen para recalibrar el algoritmo y reducir la fricción sobre las empresas legítimas.

d) i. Con los nuevos parámetros:

- $P(A \mid C_1) = 0.06$
- $P(A \mid C_2) = 0.30$
- $P(A \mid C_3) = 0.98$

Primero se calcula la nueva probabilidad total de alerta:

$P(A) = 0.154$

Por lo tanto, $P(C_3 \mid A) = 0.3182$

d) ii. No parece haber valido la pena si el objetivo es aumentar la precisión de las alertas. Aunque la probabilidad de detectar una empresa fantasma aumentó de 90% a 98%, la probabilidad de que una alerta corresponda realmente a una empresa fantasma disminuyó:

Antes: $P(C_3 \mid A) = 0.3750$

Después: $P(C_3 \mid A) = 0.3182$

Esto ocurre porque las empresas cumplidas representan 75% de la población. Al aumentar su tasa de falsos positivos de 2% a 6%, se genera un número mucho mayor de alertas provenientes de este grupo numeroso. El incremento de falsos positivos supera el beneficio obtenido por el aumento en la detección de empresas fantasma.

e) La afirmación del despacho confunde dos probabilidades condicionales diferentes: $P(A \mid C_3) = 0.90$ y $P(C_3 \mid A) = 0.375$

La primera indica que, si una empresa es fantasma, existe 90% de probabilidad de que el sistema genere una alerta.

La segunda responde a la pregunta relevante para una empresa que ya recibió una alerta: ¿cuál es la probabilidad de que sea una empresa fantasma?

Esta probabilidad es $P(C_3 \mid A) = 37.5\%$ y no 90%.

# TEMA 4: Variables aleatorias y distribuciones de probabilidad

## 4.1 Concepto de variable aleatoria y soporte. Variables aleatorias discretas y continuas

### 4.1.1

| Inciso | ¿Es una variable aleatoria numérica? | Explicación |
|---|---|---|
| a) Número de empleados en una empresa | Sí | Es una variable cuantitativa discreta |
| b) Ciudad de nacimiento de una persona | No | Es cualitativa nominal. Puede transformarse mediante variables indicadoras (una por ciudad) o asignando códigos numéricos (solo como identificadores) |
| c) Fecha de fundación de una empresa | Sí | Puede expresarse numéricamente (por ejemplo, año de fundación o número de días desde una fecha de referencia) |
| d) Salario mensual de un empleado | Sí | Variable cuantitativa continua |
| e) Cantidad de productos vendidos por mes | Sí | Variable cuantitativa discreta |
| f) Nombre del presidente de un país | No | Es cualitativa nominal. Puede transformarse mediante variables indicadoras o códigos numéricos |
| g) Porcentaje de participación electoral en un distrito | Sí | Variable cuantitativa continua |
| h) Número de páginas de un libro | Sí | Variable cuantitativa discreta |
| i) Número de reuniones realizadas en un mes | Sí | Variable cuantitativa discreta |
| j) PIB de un país | Sí | Variable cuantitativa continua |
| k) Cantidad de oficinas que tiene una empresa en un año determinado | Sí | Variable cuantitativa discreta |
| l) Tipo de cambio dólar/peso el primer día del año | Sí | Variable cuantitativa continua |
| m) Número de protestas registradas en un año | Sí | Variable cuantitativa discreta |

| Inciso | ¿Es una variable aleatoria numérica? | Explicación |
|---|---|---|
| n) Número de productos que fabrica una empresa | Sí | Variable cuantitativa discreta |
| o) Edad de los legisladores | Sí | Variable cuantitativa continua (o discreta si se registra en años completos) |
| p) Número de clientes atendidos por un empleado | Sí | Variable cuantitativa discreta |
| q) Cantidad de conferencias internacionales celebradas en un año | Sí | Variable cuantitativa discreta |
| r) Política oficial de un gobierno sobre impuestos | No | Es cualitativa. Puede transformarse en categorías (por ejemplo: expansiva, neutral, restrictiva) y luego codificarse mediante variables indicadoras |
| s) Tiempo promedio de respuesta en atención al cliente | Sí | Variable cuantitativa continua |
| t) Día de la semana | No | Es cualitativa nominal. Puede transformarse mediante variables indicadoras (lunes, martes, …) o códigos numéricos (1–7, únicamente como etiquetas) |
| u) Número de productos defectuosos en producción | Sí | Variable cuantitativa discreta |
| v) Eslogan de una campaña publicitaria | No | Es cualitativa nominal. Puede transformarse mediante variables indicadoras o códigos numéricos |
| w) Cantidad de contratos firmados en un trimestre | Sí | Variable cuantitativa discreta |
| x) Nombre de los candidatos electorales para la próxima elección | No | Es cualitativa nominal. Puede transformarse mediante variables indicadoras o códigos numéricos |
| y) Nivel de satisfacción de empleados en encuesta (0–10) | Sí | Puede tratarse como variable cuantitativa discreta (escala de 0 a 10) |
| z) Número de seguidores de una empresa en redes sociales | Sí | Variable cuantitativa discreta |

**4.1.2**

| | Variable | Tipo |
|---|---|---|
| a) | Número de votos obtenidos por un candidato en una elección | Cuantitativa discreta |
| b) | Porcentaje de participación electoral por distrito | Cuantitativa continua |
| c) | Número de partidos políticos registrados en un país | Cuantitativa discreta |
| d) | Cantidad de propuestas de ley aprobadas en un periodo legislativo | Cuantitativa discreta |
| e) | Índice de corrupción percibido 0–100 | Cuantitativa continua |
| f) | Número de manifestaciones públicas en un año | Cuantitativa discreta |
| g) | Edad de los legisladores años | Cuantitativa continua |
| h) | Número de países que firmaron un tratado internacional | Cuantitativa discreta |
| i) | Cantidad de escaños ocupados por un partido en el parlamento | Cuantitativa discreta |
| j) | Porcentaje de aprobación presidencial | Cuantitativa continua |
| k) | Número de embajadas de un país en el extranjero | Cuantitativa discreta |
| l) | Cantidad de tratados comerciales firmados en un año | Cuantitativa discreta |
| m) | Volumen de importaciones de un país millones de USD | Cuantitativa continua |
| n) | Número de conflictos armados activos en una región | Cuantitativa discreta |
| o) | Número de visitas oficiales entre jefes de Estado | Cuantitativa discreta |
| p) | Porcentaje de población con pasaporte válido | Cuantitativa continua |
| q) | Número de sanciones económicas impuestas a un país | Cuantitativa discreta |
| r) | Gasto militar como porcentaje del PIB | Cuantitativa continua |
| s) | Número de refugiados recibidos por un país en un año | Cuantitativa discreta |
| t) | Tiempo promedio de respuesta a crisis internacionales días | Cuantitativa continua |
| u) | Número de empleados en un departamento | Cuantitativa discreta |
| v) | Horas trabajadas por empleado a la semana | Cuantitativa continua |
| w) | Número de reuniones realizadas en un mes | Cuantitativa discreta |
| x) | Porcentaje de proyectos completados a tiempo | Cuantitativa continua |

| | Variable | Tipo |
|---|---|---|
| y) | Presupuesto asignado a un departamento miles de USD | Cuantitativa continua |
| z) | Número de quejas atendidas por servicio al cliente | Cuantitativa discreta |
| aa) | Número de horas de capacitación por empleado al año | Cuantitativa continua |
| bb) | Porcentaje de satisfacción laboral medido en encuesta | Cuantitativa continua |
| cc) | Cantidad de días de ausencia por enfermedad | Cuantitativa discreta |
| dd) | Tiempo promedio de resolución de problemas administrativos horas | Cuantitativa continua |

**4.1.3** Para que f(x) sea una función de probabilidad debe cumplir:

1. $f(x) \geq 0$ para todos los valores posibles de x
2. $\sum f(x) = 1$

a) $f(x) = x$, para $x = 0.01, 0.04, 0.09, 0.16$

- No negatividad: Todos los valores de f(x) son mayores o iguales a cero. ✓
- Suma de probabilidades: $\sum f(x) = 0.01 + 0.04 + 0.09 + 0.16 = 0.30$

  Como $\sum f(x) = 0.30 \neq 1$, f(x) no es una función de probabilidad.

b) $f(x) = x/5$, para $x = 1, 2, 3$

- No negatividad:

  $f(1) = 1/5 = 0.20$

  $f(2) = 2/5 = 0.40$

  $f(3) = 3/5 = 0.60$

  Todos los valores son mayores o iguales a cero. ✓
- Suma de probabilidades: $\sum f(x) = 1/5 + 2/5 + 3/5 = 6/5 = 1.20$

  Como $\sum f(x) = 1.20 \neq 1$, f(x) no es una función de probabilidad.

c) $f(x) = (x - 1)/2$, para $x = 0, 1, 2$

- No negatividad:

  f(0) = (0 − 1)/2 = −0.50

  f(1) = (1 − 1)/2 = 0

  f(2) = (2 − 1)/2 = 0.50

  Como f(0) = −0.50 < 0, no se cumple la condición de no negatividad.

- Suma de probabilidades: $\sum f(x) = -0.50 + 0 + 0.50 = 0 \neq 1$, por lo tanto, f(x) no es una función de probabilidad.

d) $f(x) = x - x^2 + 0.1$, para x = 0.1, 0.2, 0.3, 0.4, 0.5

- No negatividad:

  $f(0.1) = 0.1 - (0.1)^2 + 0.1 = 0.19$

  $f(0.2) = 0.2 - (0.2)^2 + 0.1 = 0.26$

  $f(0.3) = 0.3 - (0.3)^2 + 0.1 = 0.31$

  $f(0.4) = 0.4 - (0.4)^2 + 0.1 = 0.34$

  $f(0.5) = 0.5 - (0.5)^2 + 0.1 = 0.35$

  Todos los valores son mayores o iguales a cero. ✓

- Suma de probabilidades: $\sum f(x) = 0.19 + 0.26 + 0.31 + 0.34 + 0.35 = 1.45$

  Como $\sum f(x) = 1.45 \neq 1$, f(x) no es una función de probabilidad.

e) $f(x) = (3/4)/[x!(3 - x)!]$, para x = 0, 1, 2, 3

- No negatividad:

  f(0) = (3/4)/[0!3!] = (3/4)/6 = 1/8

  f(1) = (3/4)/[1!2!] = (3/4)/2 = 3/8

  f(2) = (3/4)/[2!1!] = (3/4)/2 = 3/8

  f(3) = (3/4)/[3!0!] = (3/4)/6 = 1/8

  Todos los valores son mayores o iguales a cero. ✓

- Suma de probabilidades: $\sum f(x) = 1/8 + 3/8 + 3/8 + 1/8 = 8/8 = 1$ ✓

  Por lo tanto, f(x) sí es una función de probabilidad.

## 4.2 Distribuciones de probabilidad discretas. Propiedades. Esperanza y varianza con sus propiedades

**4.2.1** a) Los emojis disponibles son:

$$\{👍, 😂, 😮\}$$

Como el usuario deja exactamente 2 reacciones y puede repetir emoji, los posibles pares ordenados son:

Par de reacciones

(👍, 👍)

(👍, 😂)

(👍, 😮)

(😂, 👍)

(😂, 😂)

(😂, 😮)

(😮, 👍)

(😮, 😂)

(😮, 😮)

Total de arreglos: $3^2 = 9$

b) Número de arreglos con 1 solo emoji = 3

$$(👍,👍),\ (😂,😂),\ (😮,😮)$$

c) Número de arreglos con 2 emojis diferentes = 6

$$(👍,😂),(😂,👍),(👍,😮),(😮,👍),(😂,😮),(😮,😂)$$

d) Distribución de $X =$ número de emojis distintos utilizados

| $X$ | Frecuencia | Probabilidad |
|---|---|---|
| 1 | 3 | $P(X=1)=\frac{3}{9}=\frac{1}{3}$ |
| 2 | 6 | $P(X=2)=\frac{6}{9}=\frac{2}{3}$ |

e) Valor esperado

$$E(X)=\sum xP(X=x)=1.667$$

f) Desviación estándar

$$E(X^2)=\sum x^2P(X=x)=3$$

$$Var(X)=E(X^2)-[E(X)]^2=3-\left(\frac{5}{3}\right)^2\approx 0.222$$

$$\sigma\approx 0.471$$

**4.2.2** Sea X = número de dígitos pares en el código QR

$S_X$ = {0, 1, 2, 3}

a) Total de códigos posibles: 10 × 10 × 10 = 1,000

b) Número de códigos para cada valor de X:

Para X = 0, los tres dígitos deben ser impares: 5 × 5 × 5 = 125 códigos

Para X = 1, se elige la posición del único dígito par de 3 formas: Escribir $\binom{3}{1}5^1 5^2 = 375$ códigos

Para X = 2, se elige la posición del único dígito impar de 3 formas: Escribir $\binom{3}{1}5^1 5^2 = 375$ códigos

Para X = 3, los tres dígitos deben ser pares: 5 × 5 × 5 = 125 códigos

c) Distribución de probabilidad de X:

| X | P(X = x) |
|---|---|
| 0 | 125/1,000 = 0.125 |
| 1 | 375/1,000 = 0.375 |

2 375/1,000 = 0.375

3 125/1,000 = 0.125

d) Valor esperado:

$E(X) = \Sigma x P(X = x) = (0)(0.125) + (1)(0.375) + (2)(0.375) + (3)(0.125) = 1.5$

e) Varianza:

$E(X^2) = \Sigma x^2 P(X = x) = (0^2)(0.125) + (1^2)(0.375) + (2^2)(0.375) + (3^2)(0.125) = 3$

$Var(X) = E(X^2) - [E(X)]^2 = 3 - (1.5)^2 = 0.75$

**4.2.3** Sea X = número de coincidencias entre las elecciones del jugador y las opciones populares.

La opción de peinado popular es el Peinado 2 (P2) y la opción de accesorio popular es el Accesorio 3 (A3).

a) y b) Las 12 combinaciones posibles y el valor correspondiente de X son:

(P1, A1) X = 0

(P1, A2) X = 0

(P1, A3) X = 1

(P1, A4) X = 0

(P2, A1) X = 1

(P2, A2) X = 1

(P2, A3) X = 2

(P2, A4) X = 1

(P3, A1) X = 0

(P3, A2) X = 0

(P3, A3) X = 1

(P3, A4) X = 0

Por lo tanto:

X = 0: 6 combinaciones

X = 1: 5 combinaciones

X = 2: 1 combinación

c) Suponiendo que las 12 combinaciones son igualmente probables, la distribución de probabilidad de X es:

| X | P(X = x) |
|---|---|
| 0 | 6/12 = 0.5000 |
| 1 | 5/12 = 0.4167 |
| 2 | 1/12 = 0.0833 |

d) $P(X \geq 1) = P(X = 1) + P(X = 2) = 0.50$

e) Moda = 0

El resultado más frecuente es no coincidir con ninguna de las dos opciones populares. De las 12 combinaciones posibles, 6 no incluyen ni el peinado popular, ni el accesorio popular.

**4.2.4** Sea X = número de proyectos de tecnología seleccionados para xi = {0, 1, 2, 3, 4}

El número total de selecciones posibles es C(14, 4) = 1,001

a) La distribución de probabilidad de X es:

$$P(X = x) = \frac{\binom{6}{x}\binom{8}{4-x}}{\binom{14}{4}}$$

Distribución de probabilidad:

| X | P(X = x) |
|---|---|
| 0 | 0.0699 |
| 1 | 0.3357 |
| 2 | 0.4196 |
| 3 | 0.1598 |
| 4 | 0.0150 |

b) $P(X \geq 1) = 1 - P(X = 0) = 0.9301$

c) $P(X = 2 \mid X \geq 1) = P(X = 2) / P(X \geq 1) = 0.4511$

d) P(no están representadas las tres áreas) = 0.5055

e) E(X) = n(K/N) = 1.7143

f) Var(X) = n(K/N)(1 − K/N)((N − n)/(N − 1)) = 0.7535

**4.2.5** a) La distribución de probabilidad de Z es:

$$P(Z = z) = \frac{\binom{13}{z}\binom{7}{15 - z}}{\binom{20}{5}}$$

Para los valores $z_i$ = {0, 1, 2, 3, 4, 5}

La tabla de la distribución de probabilidad:

| Z | P(Z = z) |
|---|---|
| 0 | 0.0014 |
| 1 | 0.0293 |
| 2 | 0.1761 |
| 3 | 0.3874 |
| 4 | 0.3228 |
| 5 | 0.0830 |

b) P(Z = 0) = 0.0014

c) P(Z = 3 | Z ≥ 1) = P(Z = 3) / P(Z ≥ 1) = 0.3879

d) E(Z) = n(K/N) = 3.25

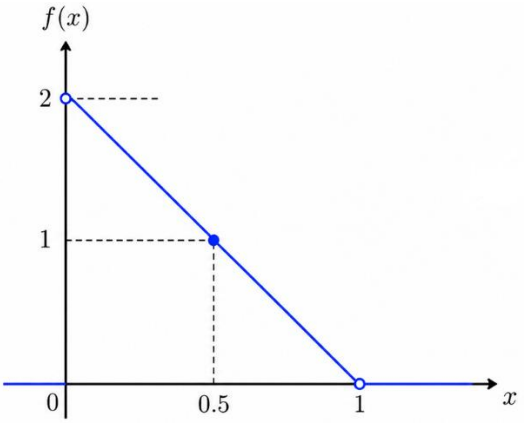
e) Var(Z) = n(K/N)(1 − K/N)((N − n)/(N − 1)) = 0.8980

## 4.3 Distribuciones de probabilidad continuas. Propiedades. Esperanza y varianza con sus propiedades

**4.3.1** La función de densidad es:

$$f(x) = \begin{cases} 2 - 2x, & \text{para } 0 < x < 1 \\ 0, & \text{en otro caso} \end{cases}$$

a) Gráfica de f(x):




b) Para que f(x) sea una función de densidad debe cumplir dos condiciones.

- No negatividad:

  Para $0 < x < 1$: $f(x) = 2 - 2x \geq 0$ ✓

- Área total igual a 1:

  $\int_0^1 (2 - 2x)\, dx = [2x - x^2]_0^1 = (2 - 1) - 0 = 1$ ✓

Por lo tanto, f(x) sí es una función de densidad de probabilidad.

c) $P(X < 0.2) = \int_0^{0.2} (2 - 2x)\, dx = 0.36$

d) $P(0.3 < X < 0.5) = 0.24$

e) $P(X > 0.8) = 0.04$

f) Se busca k tal que $P(X < k) = 0.50$

Entonces:

$\int_0^k (2 - 2x)\, dx = 0.50$

$k = 0.2929$

f) Se busca k tal que $P(X > k) = 0.10$

Entonces:

$P(X < k) = 0.90$

$k = 0.6838$

g) La función de distribución acumulada es:

$$F(x) = \begin{cases} 0, & \text{para } x \leq 0 \\ 2x - x^2, & \text{para } 0 < x < 1 \\ 1, & \text{para } x \geq 1 \end{cases}$$

h) $P(0.3 < X < 0.5) = F(0.5) - F(0.3) = 0.24$

i) $E(X) = \int_0^1 x\, f(x)\, dx = \int_0^1 x(2 - 2x)\, dx = 0.3333$

**4.3.2** La función de densidad es:

$$f(x) = \begin{cases} x + cx^2 & \text{para } 0 < x < 1 \\ 0 & \text{en otro caso} \end{cases}$$

a) Para que fx sea una función de densidad debe cumplirse:

$\int_0^1 x + cx^2\, dx = 1$

Por lo tanto, c = 3/2

b) La función de distribución acumulada es:

$$F(x) = \begin{cases} 0 & \text{para } x \leq 0 \\ \dfrac{x^2 + x^3}{2} & \text{para } 0 < x < 1 \\ 1 & \text{para } x \geq 1 \end{cases}$$

c) $P(X < 0.5) = F(0.5) = [0.5^2 + 0.5^3]/2 = 0.1875$

d) $P(X \geq 0.5 \mid X \geq 0.25) = [1 - F(0.5)] / [1 - F(0.25)] = 0.8125/0.9609375 = 0.8455$

e) $P(0.3 < X < 0.6) = F(0.6) - F(0.3) = [0.6^2 + 0.6^3]/2 - [0.3^2 + 0.3^3]/2 = 0.2295$

f) k = 0.7549 horas = 45.29 minutos

h) m = 0.9142 horas = 54.85 minutos

i) $P(X \geq 0.25) = 0.9609$

j) La probabilidad de terminar antes de 0.75 horas, dado que ya ha superado 0.25 horas, es:

$P(0.25 < X < 0.75 \mid X > 0.25) = 0.4715$

La probabilidad de terminar después de 0.75 horas es:

$P(X > 0.75 \mid X > 0.25) = 0.5285$

Como 0.5285 > 0.4715 es ligeramente más probable que el usuario termine después de alcanzar el 75% del tiempo límite, es decir, después de 45 minutos.

k) E(X) = 0.7083 horas = 42.5 minutos

l) $E(X^2) = 0.55$

Var(X) = 0.0483 horas²

$\sigma_X$ = 0.2197 horas = 13.18 minutos

m) Mediana = 0.7549 horas = 45.3 minutos

E(X) = 0.7083 horas = 42.5 minutos

Además, 20% de los usuarios tarda más de 54.9 minutos.

El modelo asigna mayor densidad a tiempos cercanos al límite superior de una hora. La mediana de aproximadamente 45.3 minutos y la media de 42.5 minutos indican que una proporción importante de los trámites requiere una parte considerable del tiempo máximo permitido.

n) Para que la plataforma sea considerada altamente eficiente debe cumplirse:

$P(X < 0.7) \geq 0.60$

$P(X < 0.7) = 0.4165$

Por lo tanto, 41.65% < 60%

**4.3.3** La función de densidad es:

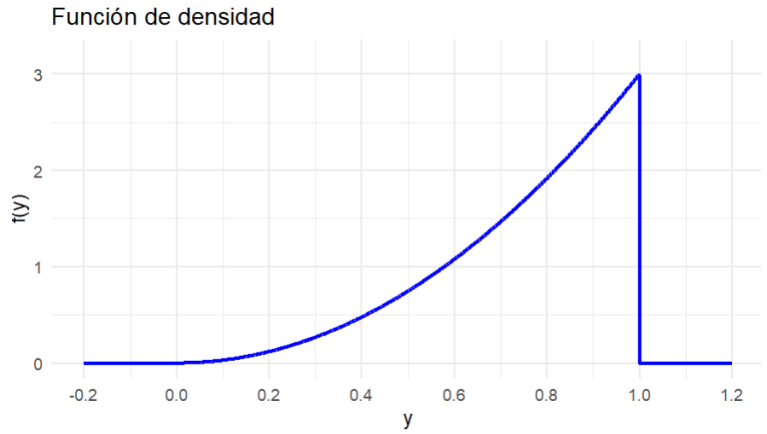
$$f(y) = \begin{cases} ky^2 & \text{para } 0 < y < 1 \\ 0 & \text{en otro caso} \end{cases}$$

a) Para que fy sea una función de densidad debe cumplirse $\int_0^1 ky^2\, dy = 1$

Por lo tanto, k = 3

Además, fy ≥ 0 para todos los valores de y, por lo que se cumplen las condiciones de una función de densidad de probabilidad.

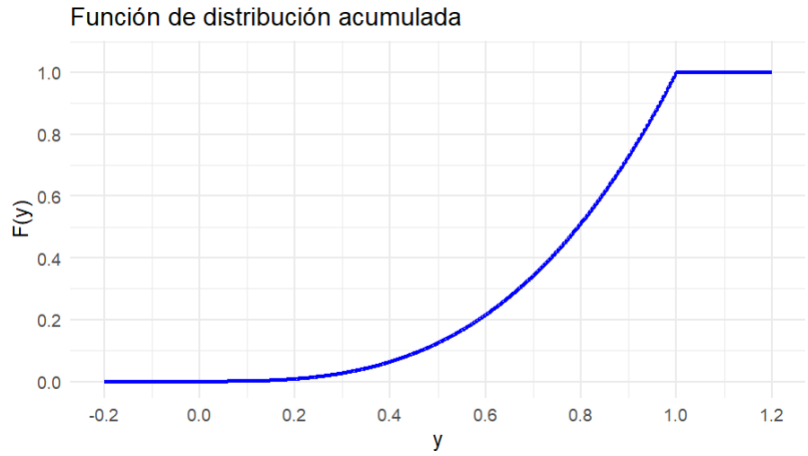
b) Gráfica de la función de densidad es:




Por lo tanto, los cargamentos no tienden a ser liberados rápidamente. El modelo indica que una mayor proporción de los tiempos de despacho se concentra cerca del límite máximo de 24 horas.

c) La función de distribución acumulada es:

$$Fy = \begin{cases} 0 & \text{para } y \leq 0 \\ y^3 & \text{para } 0 < y < 1 \\ 1 & \text{para } y \geq 1 \end{cases}$$




d) P(Y > 0.75) = 0.578125

e) $Q_1$ = 0.6300 días = 15.12 horas

f) E(Y) = ¾ = 0.75 días = 18 horas

E(Y²) = 3/5 = 0.60

Var(Y) = 3/80 = 0.0375 días²

**4.3.4** La función de densidad es:

$$f(t) = \begin{cases} 1+t & para -1 < t < 0 \\ 1 & para\ 0 \le t < c \\ 0 & en\ otro\ caso \end{cases}$$

a) Para que f(t) sea una función de densidad debe cumplirse $\int_{-1}^{0} (1 + t)\, dt + \int_{0}^{c} 1\, dt = 1$

Para ello, c = 1/2

b) Un vuelo llega con retraso cuando −1 < t < 0

Por lo tanto, PT < 0 = $\int_{-1}^{0} (1 + t)\, dt = 0.50$

c) P(T > 0.25) = $\int_{0.25}^{0.5} 1\, dt = 0.25$

d) La utilidad está dada por U(T) = 1 + 0.5T

E[U(T)] = 0.9792 miles de dólares = 979.17 dólares

e) Var[U(T)] = 71/2304 = 0.0308 (miles de dólares)²

f) La función de distribución acumulada es:

$$Ft = \begin{cases} 0 & \text{para } t \le -1 \\ t + \frac{t^2}{2} + \frac{1}{2} & \text{para } -1 < t < 0 \\ \frac{1}{2} + t & \text{para } 0 \le t < \frac{1}{2} \\ 1 & \text{para } t \ge \frac{1}{2} \end{cases}$$

g) El percentil 50 o mediana es el valor m que cumple F(m) = 0.50, por lo tanto, la mediana = 0 horas

El 50% de los vuelos llega a tiempo o adelantado y el 50% llega con retraso.

La mediana del tiempo de desviación es $m = 0$ horas. Esto significa que el 50% de los vuelos presenta una desviación menor que 0, es decir, llega con retraso, mientras que el 50% presenta una desviación mayor que 0, es decir, llega adelantado. La mediana igual a cero indica que el horario programado constituye el punto central de la distribución en términos probabilísticos; no significa que los vuelos tengan una probabilidad positiva de llegar exactamente a la hora programada.

h) P(T < −0.75) = 0.03125

Si se operan 40 vuelos en un mes, el número esperado de vuelos que requerirán indemnización es 40(0.03125) = 1.25

En promedio, se requerirán aproximadamente 1.25 indemnizaciones por cada 40 vuelos operados.

**4.3.5** a) $E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx = \int_{0}^{\infty} 0.1x\, e^{-0.1x}\, dx = 10$

b) Función de distribución acumulada:

$$F(x) = \begin{cases} 0 & x \leq 0 \\ 1 - e^{-0.1x} & x > 0 \end{cases}$$

c) P(X > 10) = 0.3679

d) La mediana m satisface F(m) = 0.50, m = 6.9315 minutos.

La mediana es aproximadamente 6.93 minutos, menor que la media de 10 minutos. Esta diferencia se debe a que la distribución exponencial es asimétrica a la derecha: la mayoría de las llamadas tiene duraciones relativamente cortas, mientras que unas pocas llamadas de larga duración elevan el valor de la media.

e) P(X > 25 | X > 15) = P(X > 25 ∩ X > 15) / P(X > 15) = 0.3679

Por la propiedad de falta de memoria de la distribución exponencial: P(X>25|X>15) = P(X>10) = 0.3679

Dado que una llamada ya ha durado 15 minutos, la probabilidad de que dure al menos 10 minutos adicionales es 36.79%. Debido a la propiedad sin memoria de la distribución exponencial, esta probabilidad es la misma que la de que una llamada recién iniciada dure más de 10 minutos.

f) Var(X) = 1/(0.1)² = 100 minutos² y $\sigma_X$ = 10 minutos

Una desviación estándar de 10 minutos, igual al valor de la media, indica una variabilidad considerable en la duración de las llamadas. Por tanto, aunque la duración promedio sea de 10 minutos, existe una dispersión importante y resulta difícil predecir con precisión cuánto durará una sesión individual.

g) La función de costo total es C(X) = 2 + 0.5X, donde C(X) está expresado en dólares.

E[C(X)] = E(2 + 0.5X) = 7 dólares

Var[C(X)] = Var(2 + 0.5X) = 25 dólares²

h) La política de calidad exige que $P(X > 25) \leq 0.05$

$P(X > 25) = 0.0821$

Como 8.21% > 5%, la Fintech no cumple con la política de calidad establecida. De acuerdo con el modelo, aproximadamente 8.21% de las llamadas supera los 25 minutos, cuando el máximo permitido es 5%.

**4.4 Distribuciones de probabilidad bivariadas discretas / 4.5 Distribuciones marginales de probabilidades. Probabilidad condicional de variables aleatorias / 4.6 Variables aleatorias independientes / 4.7 Covarianza y correlación**

**4.4.1** a) El signo del coeficiente de correlación indica la dirección de la relación lineal: positivo si las variables tienden a aumentar juntas y negativo si una tiende a disminuir cuando la otra aumenta. La magnitud indica la intensidad de la relación lineal: cuanto más cercano esté $|\rho|$ a 1, más fuerte es la relación; cuanto más cercano a 0, más débil es.

b) La correlación es igual a cero cuando $Cov(X,Y) = 0$. Por lo tanto, correlación cero no implica necesariamente independencia.

Esto indica que no existe relación lineal entre las variables, pueden existir otros tipos de dependencia no lineal.

**4.4.2** a) Distribución marginal de X:

| X | P(X=x) |
|---|---|
| 0 | 0.20 |
| 1 | 0.45 |
| 2 | 0.35 |

Distribución marginal de Y:

| Y | P(Y=y) |
|---|---|
| 0 | 0.30 |
| 1 | 0.43 |
| 2 | 0.27 |

b) $P(Y = 2 \mid X = 1) = 0.2222$

c) $E(X) = 1.15,\ E(Y) = 0.97,\ E(XY) = 1.30$

d) $\mathrm{Cov(X,Y)} = 0.1845$

Como $\mathrm{Cov}(X,Y) > 0$, existe una asociación lineal positiva: valores altos de exposición a noticias falsas tienden a presentarse junto con valores altos de discusión política.
Sin embargo, la covarianza positiva no demuestra que las noticias falsas causen una mayor discusión política.

e) $X$ y $Y$ no son independientes.

f) Buscamos la probabilidad conjunta más alta de la tabla, que corresponde a:

$$P(\mathrm{X} = 1, \mathrm{Y} = 1) = 0.20$$

Por lo tanto (X, Y) = (1, 1) es la combinación más frecuente y corresponde a jóvenes que durante la semana estuvieron expuestos una vez a noticias políticas falsas y que discutieron temas políticos una vez.

g) $\mathrm{Cov}(X,Y) = 0.1845 > 0$ sugiere una asociación positiva entre exposición a noticias falsas y discusión política.

Sin embargo, no podemos afirmar que la exposición “incrementa” o causa la discusión política. Podrían existir otras variables, por ejemplo, el interés previo por la política: una persona muy interesada en política podría consumir más contenido político, encontrarse con más noticias falsas y, al mismo tiempo, discutir más sobre política.

h) Un grupo especialmente relevante sería $X = 2,\ Y = 2$ porque representa a quienes tienen alta exposición a noticias falsas y alta participación en discusiones políticas. Esto podría aumentar la posibilidad de que información falsa sea compartida, discutida o amplificada en sus redes sociales.

Sin embargo, desde una estrategia preventiva también sería razonable dirigir la campaña a todos los individuos con $X = 2$, ya que $P(X = 2) = 0.35$, que equivale al 35% de las personas con mayor exposición, independientemente de cuánto discutan política.

i) Si $X$ y $Y$ fueran independientes, significaría que conocer cuántas veces una persona estuvo expuesta a noticias políticas falsas no proporcionaría información adicional sobre cuántas veces discutió temas políticos. Conocer el nivel de exposición a

noticias falsas sí cambia nuestra información probabilística sobre el nivel de discusión política.

j) No necesariamente. Independencia estadística no implica causalidad.

**4.4.3** a) Tabla de frecuencias conjuntas $n_{ij}$:

| X\Y | 1 | 2 | 3 | Total |
|---|---|---|---|---|
| 0 | 6 | 5 | 3 | **14** |
| 1 | 2 | 8 | 7 | **17** |
| 2 | 3 | 3 | 3 | **9** |
| **Total** | **11** | **16** | **13** | **40** |

b) Tabla de probabilidades conjuntas $p_{ij}$:

| X\Y | 1 | 2 | 3 | Total |
|---|---|---|---|---|
| 0 | 0.150 | 0.125 | 0.075 | **0.350** |
| 1 | 0.050 | 0.200 | 0.175 | **0.425** |
| 2 | 0.075 | 0.075 | 0.075 | **0.225** |
| **Total** | **0.275** | **0.400** | **0.325** | **1.000** |

c) Distribución marginal de X:

| X | P(X=x) |
|---|---|
| 0 | 0.350 |
| 1 | 0.425 |
| 2 | 0.225 |

Distribución marginal de Y:

| Y | P(Y=y) |
|---|---|
| 0 | 0.275 |

| Y | P(Y=y) |
|---|---|
| 1 | 0.400 |
| 2 | 0.325 |

d) $P(Y \geq 2 \mid X = 2) = \frac{6}{9} = 0.6667$

e) $E(X) = 0.875, \quad E(Y) = 2.05, \quad \text{Cov}(X, Y) = 0.08125$

La covarianza positiva indica una ligera asociación positiva entre uso de IA y productividad.

f) El mayor número esperado de proyectos corresponde a $X = 1$:

$$E(Y \mid X = 0) = 1.786,\ E(Y \mid X = 1) = 2.294,\ E(Y \mid X = 2) = 2.000$$

Por lo tanto, usar una herramienta de IA está asociado con la mayor productividad esperada.

g) Ingreso esperado por empleado:

$$E(\text{Ingreso}) = 20{,}000E(Y) = \$41{,}000$$

h) No puede concluirse que convenga aumentar indiscriminadamente el uso de IA. Los datos muestran que $X = 1$ tiene mayor productividad promedio que $X = 2$; además, la asociación observada no implica causalidad. Se requeriría evaluar el efecto de la capacitación y su costo.

i) X y Y no son independientes.

**4.4.4** a) $P(X \geq 1) = P(X = 1) + P(X = 2) = 0.35 + 0.30 = 0.65$

b) $P(Y = 1 \mid X = 2) = 0.6667$

c) $E(X) = 0.95$

$E(Y) = 0.40$

d) $E(XY) = 0.55$

$\text{Cov}\,X, Y = 0.17$

e) La covarianza es positiva, lo que indica que mayores retrasos están asociados con un mayor número de reclamos.

f ) El mayor riesgo corresponde a 2 días de retraso:

$$P(Y = 1 \mid X = 0) = 0.1429$$

$$P(Y = 1 \mid X = 1) = 0.4286$$

$$\boxed{P(Y = 1 \mid X = 2) = 0.6667}$$

g) Conviene priorizar la reducción de retrasos de 2 a 1 día, pues la probabilidad de reclamo disminuiría de $66.67\%$ a $42.86\%$, una reducción de aproximadamente 23.81 puntos porcentuales, mayor que al pasar de 1 a 0 días $(28.57$puntos porcentuales si se compara 42.86% con 14.29%; de hecho, esta reducción es mayor.

Por lo tanto, si el criterio es exclusivamente la mayor reducción en probabilidad de reclamo por envío, convendría reducir de 1 a 0 días.

h) Si se interpreta que la mitad de los envíos con retraso de 2 días pasan a tener 1 día de retraso, el número esperado de reclamos pasa de: $E(Y) = 0.40$ a aproximadamente $E(Y_{\text{nuevo}}) = 0.3643$

Es decir, los reclamos esperados disminuirían en aproximadamente 0.0357 por envío, equivalente a una reducción aproximada del 8.93%.

**4.4.5** a) Distribuciones marginales:

$$P_X(1) = 0.30,\ P_X(2) = 0.40,\ P_X(3) = 0.30$$

$$P_Y(0) = 0.30,\ P_Y(1) = 0.45,\ P_Y(2) = 0.25$$

b) $P(Y \geq 1 \mid X = 3) = 0.8333$

c) $E(X) = 2$

$E(Y) = 0.95$

d) $E(XY) = 2.10$

$\text{Cov X, Y} = 2.10 - (2)(0.95) = 0.20$

La covarianza es positiva, lo que significa que un mayor número de productos comprados está asociado con un mayor número de devoluciones.

e) Calculando las devoluciones esperadas según el tamaño de la compra:

$$E(Y \mid X = 1) = 0.6667$$

$$E(Y \mid X = 2) = 1.0000$$
$$E(Y \mid X = 3) = 1.1667$$

Por lo tanto, los clientes que compran 3 productos generan, en promedio, más devoluciones.

f) No puede determinarse únicamente con estos datos si conviene incentivar compras grandes. Aunque las órdenes mayores están asociadas con más devoluciones, habría que comparar el ingreso adicional generado por las ventas con el costo adicional de las devoluciones.

g) Si cada devolución cuesta $150:

$$E(C) = 150E(Y) = 142.50$$

4.4.6 a) Distribución conjunta de $X$ y $Y$

| X\Y | 1 | 2 | 3 | 4 | P(X=x) |
|---|---|---|---|---|---|
| **1** | 0.07 | 0.38 | 0.12 | 0.03 | **0.60** |
| **2** | 0.03 | 0.12 | 0.18 | 0.07 | **0.40** |
| **P(Y=y)** | **0.10** | **0.50** | **0.30** | **0.10** | **1.00** |

b) Coeficiente de correlación

$$E(X) = 1.40,\ E(Y) = 2.40$$
$$E(XY) = 3.49$$

$$Cov(X,Y) = E(XY) - E(X)E(Y) = 3.49 - (1.40)(2.40) = 0.13$$

$$Var(X) = 0.24,\ Var(Y) = 0.64$$

$$\rho_{XY} = \frac{0.13}{\sqrt{(0.24)(0.64)}} \approx 0.332$$

Existe una relación lineal positiva débil a moderada entre el número de automóviles y el número de conductores con licencia: en general, las familias con más automóviles tienden a tener más conductores con licencia, aunque la asociación no es fuerte.

c) Independencia

$X$ y $Y$ no son independientes. Por ejemplo:

$$P(\mathrm{X}=1,\mathrm{Y}=1)=0.07$$

Y si fueran independientes:

$$P(X=1)P(Y=1)=(0.60)(0.10)=0.06$$

Como $0.07 \neq 0.06$, se concluye que $X$ y $Y$ no son independientes.

**4.4.7** a) Distribución de $Z=X-Y$

| **z** | **-4** | **-3** | **-2** | **-1** | **0** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|---|---|---|---|---|
| P(Z=z) | 0.00 | 0.00 | 0.05 | 0.11 | 0.71 | 0.09 | 0.02 | 0.02 | 0.00 |

Los valores negativos indican que hubo más guías que solicitudes, $Z=0$ indica que el número de guías coincidió con el número de solicitudes y los valores positivos representan solicitudes que no pudieron ser atendidas por falta de guías.

b) Comprobación de $E(Z)=E(X)-E(Y)$

$$E(X)-E(Y)=2.08-2.10=-0.02$$

Utilizando directamente la distribución de $Z$:

$$E(Z)=\sum_{z} z\,P(Z=z)=-0.02$$

Por lo tanto: $E(Z)=E(X)-E(Y)=-0.02$ ✓

d) $Cov(X,Y)=E(XY)-E(X)E(Y)=0.802$

La covarianza positiva indica que, en general, cuando aumenta el número de solicitudes de última hora, también tiende a ser mayor el número de guías contratados.

d) $\rho_{XY}=\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}=\frac{0.802}{\sqrt{(1.3736)(0.89)}}\approx 0.7254$

Existe una relación lineal positiva relativamente fuerte entre el número de solicitudes y el número de guías contratados. Esto sugiere que la empresa ha logrado anticipar razonablemente la demanda: en fechas con mayor número de solicitudes suele

disponer de más guías. Sin embargo, la relación no es perfecta, por lo que todavía existe riesgo de exceso o falta de personal.

e) $X$ y $Y$ no son independientes

Por ejemplo:

$$P(X = 1, Y = 1) = 0.20$$

$$P(X = 1)P(Y = 1) = (0.25)(0.30) = 0.075$$

Como $0.20 \neq 0.075$, se concluye que $X$ y $Y$ no son independientes.

f) $Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y) = 0.6596$

Por lo tanto, la desviación estándar sería:

$$\sigma_Z = \sqrt{0.6596} \approx 0.8122$$

**4.4.8** Denotemos:

- $E$: bebida embotellada
- $L$: bebida en lata

Como en cada una de las cuatro semanas hay dos posibles elecciones, existen $2^4 = 16$ secuencias posibles, todas con la misma probabilidad:

$$P(\text{cada secuencia}) = \frac{1}{16}$$

a) Espacio muestral: $S = \{LEEE, LEEL, LELE, LELL, LLEE, LLEL, LLLE, LLLL\}$

Por lo tanto $| \Omega | = 16$

b) Distribución de probabilidad conjunta de $X$ y $Y$

| X\Y | 0 | 1 | 2 | 3 | 4 | P(X=x) |
|---|---|---|---|---|---|---|
| **0** | 1/16 | 0 | 0 | 0 | 1/16 | **2/16** |
| **1** | 0 | 2/16 | 2/16 | 2/16 | 0 | **6/16** |
| **2** | 0 | 2/16 | 2/16 | 2/16 | 0 | **6/16** |
| **3** | 0 | 0 | 2/16 | 0 | 0 | **2/16** |
| **P(Y=y)** | **1/16** | **4/16** | **6/16** | **4/16** | **1/16** | **1** |

c) Distribuciones marginales

La distribución marginal de $X$es:

| $x$ | **0** | **1** | **2** | **3** |
|---|---|---|---|---|
| $P(X=x)$ | $1/8$ | $3/8$ | $3/8$ | $1/8$ |

Es decir $X \sim Binomial(3, 0.5)$

La distribución marginal de $Y$ es:

| $y$ | **0** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|---|
| $P(Y=y)$ | $1/16$ | $4/16$ | $6/16$ | $4/16$ | $1/16$ |

Por lo tanto $Y \sim Binomial(4, 0{,}5)$

d) $X$ y $Y$ no son independientes.

Por ejemplo $P(X=0, Y=0) = \frac{1}{16}$

Pero $P(X=0)P(Y=0) = \frac{2}{16}\left(\frac{1}{16}\right) = \frac{1}{128}$

Como $\frac{1}{16} \neq \frac{1}{128}$, se concluye que $X$ y $Y$ no son independientes.

Conocer el número de cambios de presentación proporciona información sobre el número posible de bebidas embotelladas. Por ejemplo, si $X = 0$, el cliente compró la misma presentación las cuatro semanas, por lo que necesariamente $Y = 0$ o $Y = 4$.

**4.4.9** a) $X$ y $Y$no son independientes. Por ejemplo:

$$P(\mathrm{X}=0, \mathrm{Y}=1) = 0.199$$

$$P(X=0)P(Y=1) = (0.450)(0.386) = 0.1737$$

Como $0.199 \neq 0.1737$ se concluye que $X$ y $Y$ no son independientes, es decir, existe una relación entre los años de escolaridad y el estrato de ingreso del jefe de familia.

b) i. $P(Y<3) = P(Y=1) + P(Y=2) = 0.574$

ii. Como $Y$ toma valores de 1 a 5, $P(Y>4) = P(Y=5) = 0.13$

iii. $P(X<6) = P(X=0) + P(X=3) = 0.67$

iv. $P(Y = 5 \mid X = 6) = \frac{P(X, = 6Y = 5)}{P(X=6)} \approx 0.3476$

v. $P(X = 3 \mid Y = 1) = \frac{P(X=3,Y=1)}{P(Y=1)} \approx 0.4585$

c) $E(X) = \sum_x x\, P(X = x) = 3.051$

$E(Y) = 2.408$

d) $Var(X) \approx 10.7704$

$Var(Y) \approx 2.0195$

e) $\rho_{XY} \approx 0.6491$

4.4.10 La probabilidad conjunta puede obtenerse mediante:

$$P(\mathrm{X} = \mathrm{x}, \mathrm{Y} = \mathrm{y}) = \frac{\binom{3}{x}\binom{2}{y}\binom{4}{2-x-y}}{\binom{9}{2}}$$

para $x + y \leq 2$.

a) Distribución de probabilidad conjunta de $X$ y $Y$

| $X\backslash Y$ | **0** | **1** | **2** | $P(X = x)$ |
|---|---|---|---|---|
| **0** | $6/36$ | $8/36$ | $1/36$ | $15/36$ |
| **1** | $12/36$ | $6/36$ | $0$ | $18/36$ |
| **2** | $3/36$ | $0$ | $0$ | $3/36$ |
| $P(Y = y)$ | $21/36$ | $14/36$ | $1/36$ | **1** |

b) Probabilidades

i. $P(\mathrm{X} = 2, \mathrm{Y} = 0) = 0.0833$

ii. $P(X = 2 \mid Y = 1) = 0$

iii. $P(X + Y < 1) = 0.1667$

iv. $P(Y = 0 \mid X \leq 1) \approx 0.5455$

c) Como siempre se seleccionan exactamente 2 automóviles $= 2 - X - Y$

$$P(Z \leq 1) = 1 - P(Z = 2) \approx 0.8333$$

d) Valor esperado y desviación estándar de $Z$

El número de Chevrolet seleccionados sigue una distribución hipergeométrica:

$$Z \sim H(N = 9, K = 4, n = 2)$$

Por lo tanto, $E(Z) = n\frac{K}{N} = \approx 0.8889$

$$Var(Z) = n\frac{K}{N}\left(1 - \frac{K}{N}\right)\frac{N-n}{N-1} = \frac{35}{81}$$

$$\sigma_Z \approx 0.6573$$

**4.4.11** a) $E(X) = 0, E(Y) = 0, E(XY) = 0$

b) $Cov(X,Y) = E(XY) - E(X)E(Y) = 0, Var(X) = Var(Y) = \frac{3}{4}, \rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = 0$

c) X y Y no son independientes. Por ejemplo:

$$P(\mathrm{X} = 0, \mathrm{Y} = 0) = 0$$

y $P(X = 0)P(Y = 0) = \left(\frac{1}{4}\right)\left(\frac{1}{4}\right) = \frac{1}{16}$

Como $0 \neq \frac{1}{16}$ se concluye que $X$ y $Y$ no son independientes.

d) Conclusión: Una correlación igual a cero indica únicamente ausencia de relación lineal entre las variables, pero puede existir otro tipo de dependencia.

$$\rho_{XY} = 0 \ \Rightarrow / \ X \text{ y } Y \text{ son independientes}$$

**4.4.12** a) X y Y no son independientes.Por ejemplo:

$$P(\mathrm{X} = 0, \mathrm{Y} = 0) = 0.10$$

$$P(X = 0)P(Y = 0)(0.13)(0.10) = 0.013$$

Como $0.10 \neq 0.013$ se concluye que $X$ y $Y$ no son independientes.

El número de hamburguesas consumidas y el número de refrescos consumidos están relacionados.

b) $P(X < 1 \mid Y \geq 1) = \frac{0.03}{0.90} \approx 0.0333$

Entre los clientes que consumen al menos un refresco, aproximadamente el 3.33% no consume ninguna hamburguesa.

c) Coeficiente de correlación

$E(X) = 1.07,\ E(Y) = 1.07, E(XY) = 1.64$

$Cov(\mathrm{X}, \mathrm{Y}) = 0.4951$

$Var(X) = 0.2051, Var(Y) = 0.2051$

$\rho_{XY} \approx 0.4637$

En general, los clientes que consumen más hamburguesas tienden también a consumir más refrescos, aunque la relación no es perfecta.

**4.4.13** Sea $T = X + Y$

a) $E(T) = 102$ kg

b) $Var(T) = 9.43$ kg$^2$; $\sigma_T \approx 3.07$ kg

c) $\rho_{XY} = -0.7$ significa que las ventas de ambos tipos de café tienden a moverse en direcciones opuestas: cuando aumentan las ventas de un tipo, las del otro tienden a disminuir.

**4.4.14** a) Distribución de probabilidad conjunta de $X$ y $Y$

| $X \backslash Y$ | **0** | **1** | $P(X = x)$ |
|---|---|---|---|
| **0** | $7/10 \cdot 6/9 = 7/15$ | $7/10 \cdot 3/9 = 7/30$ | $7/10$ |
| **1** | $3/10 \cdot 7/9 = 7/30$ | $3/10 \cdot 2/9 = 1/15$ | $3/10$ |
| $P(Y = y)$ | $7/10$ | $3/10$ | **1** |

En decimales:

| $X\backslash Y$ | **0** | **1** | Total |
|---|---|---|---|
| **0** | 0.4667 | 0.2333 | 0.7000 |
| **1** | 0.2333 | 0.0667 | 0.3000 |
| Total | 0.7000 | 0.3000 | 1.0000 |

b) $X$ y $Y$ no son independientes. Por ejemplo:

$$P(X' = 1Y = 1) = \frac{1}{15} \approx 0.0667$$

$$P(X = 1)P(Y = 1) = \left(\frac{3}{10}\right)\left(\frac{3}{10}\right) = 0.09$$

Como $0.0667 \neq 0.09$ se concluye que $X$ y $Y$ no son independientes. Esto se debe a que la selección es sin reemplazo: el resultado de la primera selección modifica la composición del lote y, por lo tanto, la probabilidad de que el segundo celular sea defectuoso.

c) Coeficiente de correlación

$$Cov(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{15} - (0.3)(0.3) \approx -0.0233$$

$$Var(X) = Var(Y) = (0.3)(0.7) = 0.21$$

$$\rho_{XY} = \frac{-0.0233}{\sqrt{(0.21)(0.21)}} \approx -0.1111$$

Existe una relación lineal negativa débil entre los resultados de ambas selecciones. Al seleccionar sin reemplazo, el resultado de la primera extracción modifica la composición del lote para la segunda extracción. Si el primer celular resulta defectuoso, disminuye el número de defectuosos disponibles, reduciendo la probabilidad de volver a seleccionar otro defectuoso. Esta dependencia explica que la correlación sea negativa.

d) Como $X$ indica si el primer celular es defectuoso y $Y$si el segundo es defectuoso: $W = X + Y$

$$E(W) = 0.6$$
$$Var(W) \approx 0.3733$$

**4.4.15** a) Distribución de probabilidad conjunta de $X$ y $Y$

Como siempre se seleccionan exactamente tres monedas, las variables satisfacen necesariamente $X + Y = 3$.

$$P(\mathrm{X} = \mathrm{x}, \mathrm{Y} = \mathrm{y}) = \frac{\binom{4}{x}\binom{2}{y}}{\binom{6}{3}},\ x + y = 3$$

Los únicos pares posibles son $(\mathrm{X}, \mathrm{Y}) = (1,2),\ (2,1),\ (3,0)$

| $X\backslash Y$ | **0** | **1** | **2** | $P(X = x)$ |
|---|---|---|---|---|
| **1** | 0 | 0 | $4/20$ | $4/20$ |
| **2** | 0 | $12/20$ | 0 | $12/20$ |
| **3** | $4/20$ | 0 | 0 | $4/20$ |
| $P(Y = y)$ | $4/20$ | $12/20$ | $4/20$ | **1** |

En decimales:

| $X\backslash Y$ | **0** | **1** | **2** |
|---|---|---|---|
| **1** | 0 | 0 | 0.20 |
| **2** | 0 | 0.60 | 0 |
| **3** | 0.20 | 0 | 0 |

b) $P(X \geq 2, Y \leq 2) = 0.80$

c) $\rho_{XY} = -1$

d) Si $T$ representa el valor total, en pesos, de las tres monedas seleccionadas:

$$T = 10X + 5Y$$

Como $Y = 3 - X$, también puede escribirse como $T = 5X + 15$

e) $E(T) = 25$ pesos; $Var(T) = 10$ pesos$^2$

**4.4.16** a) $c = \frac{1}{89}$

b) Tabla de distribución de probabilidad conjunta

$$P(\mathrm{X} = \mathrm{x}, \mathrm{Y} = \mathrm{y}) = \frac{x^2 + y^2}{89}$$

| $Y\backslash X$ | **-1** | **0** | **1** | **3** | $P(Y = y)$ |
|---|---|---|---|---|---|
| **-1** | $2/89$ | $1/89$ | $2/89$ | $10/89$ | $15/89$ |
| **2** | $5/89$ | $4/89$ | $5/89$ | $13/89$ | $27/89$ |
| **3** | $10/89$ | $9/89$ | $10/89$ | $18/89$ | $47/89$ |
| $P(X = x)$ | $17/89$ | $14/89$ | $17/89$ | $41/89$ | **1** |

c) $P(\mathrm{X} = 0, Y \leq 2) = \frac{5}{89} \approx 0.0562$

d) $P(\mathrm{X} < 0, \mathrm{Y} > 2) = \frac{10}{89} \approx 0.1124$

e) $P(X > 2 - Y) = P(1,2) + P(3,2) + P(0,3) + P(1,3) + P(3,3) = \frac{55}{89} \approx 0.6180$

**4.4.17** $E(X) = 2$, $E(Y) = -1$, $Var(X) = 4$, $Var(Y) = 6$, $Cov(X, Y) = \frac{1}{2}$

a) $Z = X + Y$

$E(Z) = 1;\ Var(Z) = 11$

b) $W = X - Y$

$E(W) = 3;\ Var(W) = 9$

c) $U = 2X + Y$

$E(U) = 3;\ Var(U) = 24$

e) $T = 3X - 2Y + 2$

$E(T) = 10; Var(T) = 54$

**4.4.18** $P(1) = \frac{1}{6},\ P(2) = \frac{1}{6},\ P(\text{otro}) = \frac{4}{6}$

Por lo tanto, para $x, y \geq 0$ y $x + y \leq 3$:

$$P(\mathrm{X} = \mathrm{x}, \mathrm{Y} = \mathrm{y}) = \frac{3!}{x!\, y!\, (3 - x - y)!} \left(\frac{1}{6}\right)^{x} \left(\frac{1}{6}\right)^{y} \left(\frac{4}{6}\right)^{3-x-y}$$

a) Distribución conjunta de $X$y $Y$

| $X \backslash Y$ | 0 | 1 | 2 | 3 | $P(X = x)$ |
|---|---|---|---|---|---|
| 0 | 64/216 | 48/216 | 12/216 | 1/216 | 125/216 |
| 1 | 48/216 | 24/216 | 3/216 | 0 | 75/216 |
| 2 | 12/216 | 3/216 | 0 | 0 | 15/216 |
| 3 | 1/216 | 0 | 0 | 0 | 1/216 |
| $P(Y = y)$ | 125/216 | 75/216 | 15/216 | 1/216 | 1 |

b) Distribución condicional de $X$dado $Y = 0$

$$P(X = x \mid Y = 0) = \frac{P(\mathrm{X} = \mathrm{x}, \mathrm{Y} = \mathrm{y})}{P(Y = 0)}$$

| $x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $P(X = x \mid Y = 0)$ | 64/125 | 48/125 | 12/125 | 1/125 |

Equivalentemente: $X \mid (Y = 0) \sim Binomial(3, \frac{1}{5})$

Al saber que no apareció ningún 2, cada dado solo puede ser un 1 o alguno de los valores $3,4,5,6$; por ello, condicionalmente:

$$P(1 \mid \text{no es } 2) = \frac{1}{5}$$

c) $X$ y $Y$ no son independientes. Por ejemplo:

$$P(\mathrm{X} = 3, Y = 1) = 0$$
$$P(X = 3)P(Y = 1) = \frac{1}{216}\frac{75}{216}$$

d) $X \sim Binomial\left(3, \frac{1}{6}\right)$, $Y \sim Binomial\left(3, \frac{1}{6}\right)$

$$E(X) = E(Y) = \frac{1}{2}$$
$$Var(X) = Var(Y) = 3\left(\frac{1}{6}\right)\left(\frac{5}{6}\right) = \frac{5}{12}$$
$$Cov(X,Y) = E(XY) - E(X)E(Y) = -\frac{1}{12} \approx -0.0833$$
$$\rho_{XY} = \frac{-1/12}{\sqrt{(5/12)(5/12)}} = -\frac{1}{5} = -0.20$$

Interpretación: existe una relación lineal negativa débil entre el número de unos y el número de doses. Cada dado sólo puede clasificarse como 1, 2 u "otro". Por ello, cuando aumenta el número de unos necesariamente disminuye el número de posiciones disponibles para obtener doses, originando una asociación negativa.

e) Como cada dado debe clasificarse como 1, 2 u otro valor: $Z = 3 - X - Y$

$$Z \sim Binomial\left(3, \frac{2}{3}\right)$$
$$E(Z) = 2$$
$$Var(Z) = \frac{2}{3} \approx 0.6667$$

**4.4.19** a) Probabilidades

i. $P(\mathrm{X} \leq 1Y \leq 1) = \frac{9}{60} = 0.15$

ii. $P(X + Y \leq 1) = \frac{5}{60} \approx 0.0833$

iii. $P(Y - X > 2) = \frac{4}{60} \approx 0.0667$

iv. $P(X > 0 \mid Y = 2) = \frac{5}{6} \approx 0.8333$

b) Coeficiente de correlación

$$E(X) = \frac{4}{3},\ E(Y) = 2$$

$$Var(X) = \frac{5}{9},\ Var(Y) = 1$$

$$E(XY) = \frac{8}{3}$$

$$Cov(\mathrm{X}, \mathrm{Y}) = 0$$

$$\rho_{XY} = 0$$

c) $Z = Y - X$

$$E(Z) = \frac{2}{3} \approx 0.6667$$

$$Var(Z) = \frac{14}{9} \approx 1.5556$$

**4.4.20** a) $c = \frac{1}{54}$

b) $P(\mathrm{X} \geq 2, Y \leq 3) = \frac{25}{54} \approx 0.4630$

c) Distribución marginal de $X$

$$f_X(x) = \sum_y f\,(\mathrm{x}, \mathrm{y}) = \frac{x}{6}$$

| $x$ | 1 | 2 | 3 |
|---|---|---|---|
| $P(X = x)$ | 1/6 | 1/3 | 1/2 |

$$E(X) = \frac{7}{3} \approx 2.3333$$

$$E(X^2) = 6$$

$$Var(X) = \frac{5}{9} \approx 0.5556$$

d) Distribución marginal de Y

$$f_Y(y) = \sum_x f(x,y) = \frac{y}{9}$$

| $y$ | 2 | 3 | 4 |
|---|---|---|---|
| $P(Y=y)$ | $2/9$ | $1/3$ | $4/9$ |

$$E(Y) = \frac{29}{9} \approx 3.2222$$
$$E(Y^2) = 11$$
$$Var(Y) = \frac{50}{81} \approx 0.6173$$

e) $X$ y $Y$ son independientes

$$f_X(x) f_Y(y) = \left(\frac{x}{6}\right)\left(\frac{y}{9}\right) = \frac{xy}{54} = f(x,y)$$

f) $P(X = 2 \mid Y = 3) = \frac{1}{3} \approx 0.3333$

g) $P(X \geq 2 \mid Y \leq 3) = \frac{5}{6} \approx 0.8333$

h) $Cov(X,Y) = 0;\ \rho_{XY} = 0$

En este caso, además de estar no correlacionadas, las variables son independientes.

i) $X$ y $Y$ son independientes

j) $Z = 5X - 3Y$

$$E(Z) = 2$$
$$Var(Z) = \frac{175}{9} \approx 19.4444$$

**4.4.21** a) Distribución de probabilidad conjunta de $X$y $Y$

La función de probabilidad conjunta de $X$ y $Y$ es:

$$P(X, = xY = y) = \frac{\binom{4}{x}\binom{2}{y}\binom{3}{3-x-y}}{\binom{9}{3}}$$

Tabla de probabilidades conjuntas:

| $X\backslash Y$ | 0 | 1 | 2 | $P(X=x)$ |
|---|---|---|---|---|
| 0 | $1/84$ | $6/84$ | $3/84$ | $10/84$ |
| 1 | $12/84$ | $24/84$ | $4/84$ | $40/84$ |
| 2 | $18/84$ | $12/84$ | 0 | $30/84$ |
| 3 | $4/84$ | 0 | 0 | $4/84$ |
| $P(Y=y)$ | $35/84$ | $42/84$ | $7/84$ | 1 |

b) Distribuciones marginales

Para $X$:

| $x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $P(X=x)$ | $10/84$ | $40/84$ | $30/84$ | $4/84$ |

Para $Y$:

| $y$ | *0* | *1* | *2* |
|---|---|---|---|
| $P(Y=y)$ | $35/84$ | $42/84$ | $7/84$ |

c) $E(X) = \frac{4}{3} \approx 1.3333;\ Var(X) = \frac{5}{9} \approx 0.5556$

$E(Y) = \frac{2}{3} \approx 0.6667;\ Var(Y) = \frac{7}{18} \approx 0.3889$

d) $E(XY) = \frac{3}{4};\ Cov(X, Y) = -\frac{5}{36} \approx -0.1389;\ \rho_{XY} = -\frac{1}{\sqrt{14}} \approx -0.2673$

Existe una relación lineal negativa débil entre el número de estudiantes mexicanos y españoles seleccionados. Al haber únicamente tres becas disponibles, seleccionar más estudiantes de una nacionalidad reduce las oportunidades disponibles para seleccionar estudiantes de las otras nacionalidades.

e) $Z = 3 - X - Y$; $E(Z) = 1$; $Var(Z) = \frac{1}{2} = 0.5$; $Cov(X, Y) = -\frac{2}{9}$

$$\rho_{XY} = \frac{-2/9}{\sqrt{(5/9)(7/18)}} = -\frac{4}{\sqrt{70}} \approx -0.4781$$

**4.4.22** Se tiene: $P(A) = 0.25$, $P(B \mid A) = 0.50$, $P(A \mid B) = 0.25$

$$P(A \cap B) = P(B \mid A)P(A) = 0.125$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = 0.25$$

$$P(B) = \frac{0.125}{0.25} = 0.50$$

a) Distribución conjunta de $X$y $Y$

La tabla conjunta es:

| $X\backslash Y$ | 0 | 1 | $P(X = x)$ |
|---|---|---|---|
| 0 | 0.375 | 0.375 | 0.750 |
| 1 | 0.125 | 0.125 | 0.250 |
| $P(Y = y)$ | 0.500 | 0.500 | 1.000 |

b) Verdadero o falso

i. “Las variables aleatorias $X$ y $Y$son independientes.”

Verdadero. $X$ y $Y$ son independientes.

ii. $P(X^2 + Y^2 = 1) = 0.25$

Falso. $P(X^2 + Y^2 = 1) = 0.50$

iii. $P(XY = X^2Y^2) = 1$ Verdadero

**4.4.23** a) Distribuciones marginales de $X$y $Y$

Distribución marginal del número de televisores:

| $x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $P(X = x)$ | 0.17 | 0.38 | 0.29 | 0.16 |

Distribución marginal del número de integrantes:

| $y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $P(Y = y)$ | 0.21 | 0.20 | 0.26 | 0.33 |

b) Probabilidades

i. $P(Y = 3) = 0.26$

ii. $P(X = 2 \mid Y = 3) \approx 0.3462$

c) $E(X) = 1.44;\ E(Y) = 2.71$

d) $X$ y $Y$ no son independientes. Por ejemplo:

$$P(\mathrm{X} = 0, Y = 1) = 0.07$$

$$P(X = 0)P(Y = 1) = 0.0357$$

Como: $0.07 \neq 0.0357$ se concluye que $X$ y $Y$ no son independientes

e) $E(XY) = 4.39;\ Cov(X, Y) = 0.4876$

$Var(X) = 0.9064;\ Var(Y) = 1.2859$

$\rho_{XY} \approx 0.4516$

Existe una relación lineal positiva moderada entre el número de integrantes y el número de televisores del hogar. Un coeficiente positivo indica que, en promedio, los hogares con más integrantes tienden también a tener un mayor número de televisores, aunque esta asociación no implica una relación causal.

f) Número de televisores por integrante. Se define:

$$R = \frac{X}{Y}$$

$$E(R) = 0.5775$$

$0.58$ televisores por integrante

**4.4.24** Como se estudiaron $3{,}200$estudiantes, cada probabilidad conjunta se obtiene mediante:

$$P(\mathrm{X} = x, Y = y) = \frac{n_{xy}}{3200}$$

a) Distribución de probabilidad conjunta

| $Y\backslash X$ | 0 | 1 | 2 | 3 | 4 | $P(Y=y)$ |
|---|---|---|---|---|---|---|
| 0 | 0.05844 | 0.09375 | 0.04688 | 0.07031 | 0.10563 | 0.37500 |
| 1 | 0.04875 | 0.07813 | 0.03906 | 0.05906 | 0.08750 | 0.31250 |
| 2 | 0.03906 | 0.06250 | 0.03125 | 0.04688 | 0.07031 | 0.25000 |
| 3 | 0.01000 | 0.01563 | 0.00781 | 0.01125 | 0.01781 | 0.06250 |
| $P(X=x)$ | 0.15625 | 0.25000 | 0.12500 | 0.18750 | 0.28125 | 1.00000 |

b) Distribuciones marginales

Para $X$ = número de viajes:

| $x$ | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $P(X=x)$ | 0.15625 | 0.25000 | 0.12500 | 0.18750 | 0.28125 |

Para $Y$ = número de materias reprobadas:

| $y$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $P(Y=y)$ | 0.3750 | 0.3125 | 0.2500 | 0.0625 |

c) Probabilidades

i. $P(Y<2)=0.6875$

ii. $P(X<2)=0.40625$

iii. $P(Y=0 \mid X=3)=0.375$

iv. $P(X=0 \mid Y=0)\approx 0.1558$

v. $P(Y\leq 1 \mid X=3)=0.69$

d) $X$ y $Y$ no son estrictamente independientes. Por ejemplo:

$$P(\mathrm{X}=0, Y=0)=\frac{187}{3200}=0.05844$$

$$P(X=0)P(Y=0)=(0.15625)(0.375)=0.05859$$

Como $0.05844\neq 0.05859$ se concluye que $X$ y $Y$ no son estrictamente independientes.

d) $E(X)=2.1875,\ E(Y)=1,\ E(XY)=2.18594$

$$\rho_{XY} = \frac{-0.00156}{\sqrt{(2.15234)(0.875)}} \rho_{XY} \approx -0.00114$$

e) Conclusión

Los resultados muestran que el número de viajes al extranjero y el número de materias reprobadas presentan una asociación prácticamente nula en estos datos. Aunque las variables no cumplen exactamente la condición matemática de independencia, las diferencias son mínimas y el coeficiente de correlación es prácticamente cero:

$$\rho_{XY} \approx -0.00114$$

Por lo tanto, no hay evidencia descriptiva de una relación lineal relevante entre ambas variables.

**4.4.25** a) Valores de $Y$y distribución conjunta de $X$ y $Y$

La distribución conjunta es:

| $Y\backslash X$ | $-2$ | $-1$ | $1$ | $2$ | $P(Y=y)$ |
|---|---|---|---|---|---|
| 1 | 0 | $1/4$ | $1/4$ | 0 | $1/2$ |
| 4 | $1/4$ | 0 | 0 | $1/4$ | $1/2$ |
| $P(X=x)$ | $1/4$ | $1/4$ | $1/4$ | $1/4$ | 1 |

b) Covarianza y coeficiente de correlación

$$E(X) = 0;\ E(Y) = \frac{5}{2};\ E(XY) = E(X^3) = 0;\ Cov(X,Y) = 0;\ \rho_{XY} = 0$$

c) $X$ y $Y$ no son independientes. Por ejemplo:

$$P(\mathrm{X} = 2, Y = 1) = 0$$

$$P(X = 2)P(Y = 1) = \left(\frac{1}{4}\right)\left(\frac{1}{2}\right) = \frac{1}{8}$$

Como $0 \neq \frac{1}{8}$ se concluye que $X$ y $Y$ no son independientes.

d) Correlación cero e independencia. Este ejercicio demuestra que:

$$\rho_{XY} = 0 \;\Rightarrow\; X \text{ y } Y \text{ son independientes}$$

Un coeficiente de correlación igual a cero indica únicamente que no existe relación lineal entre las variables; no descarta la existencia de una relación no lineal.

En este caso $\rho_{XY} = 0$, pero existe una dependencia exacta $Y = X^2$.

### 4.8 Funciones lineales de variables aleatorias. Propiedades. Esperanza y varianza

**4.8.1** $Y = 0.85X - 1,\ E(X) = 18,\ Var(X) = 9$

a) $E(Y) = 0.85(18) - 1 = 14.3$

$Var(Y) = (0.85)^2(9) = 6.5025$

$SD(Y) = 0.85(3) = 2.55$

b) El ingreso real esperado es de 14.3 mil pesos mensuales, después de ajustar el ingreso nominal por el efecto de la inflación considerado en el modelo.

c) El factor $0.85$ reduce la dispersión: la desviación estándar se multiplica por $0.85$ y la varianza por $0.85^2 = 0.7225$

d) Si el factor baja a $0.70$, $Var(Y) = (0.70)^2(9) = 4.41$. Por lo tanto, la varianza disminuiría de $6.5025$ a $4.41$.

**4.8.2** $E(X) = 40,\ Var(X) = 16,\ Y = 0.05X + 2$

a) $E(Y) = 0.05(40) + 2 = 4;\ Var(Y) = (0.05)^2(16) = 0.04$

b) $SD(Y) = \sqrt{0.04} = 0.20$

c) El término $+2$ representa un ingreso fijo diario de 2 mil pesos, independiente del nivel de ventas.

d) Si la comisión aumenta al $8\%$:

$$Y = 0.08X + 2$$

$$E(Y) = 0.08(40) + 2 = 5.2$$

$$Var(Y) = (0.08)^2(16) = 0.1024$$

La comisión mayor aumenta tanto el ingreso esperado como su variabilidad.

e) No. En términos absolutos, el ingreso tiene menor dispersión que las ventas:

$$SD(X) = 4,\ SD(Y) = 0.20$$

El ingreso varía menos debido a que solo una fracción de las ventas $(5\%)$ se incorpora como comisión.

**4.8.3** $E(X) = 17.5,\ Var(X) = 1,\ Y = 50{,}000X$

a) $E(Y) = 50{,}000(17.5) = \$875{,}000$
$Var(Y) = (50{,}000)^2(1) = 2{,}500{,}000{,}000$

b) $SD(Y) = 50{,}000\sqrt{1} = \$50{,}000$

c) El riesgo cambiario implica que las fluctuaciones del tipo de cambio generan variabilidad en el costo de importación. Una desviación estándar de 1 peso por dólar se traduce en una desviación estándar de $50,000 en el costo.

d Si únicamente aumenta el valor esperado del tipo de cambio de $17.5$ a $21.5$, pero se mantiene:

$$Var(X) = 1$$

entonces: $Var(Y) = 2{,}500{,}000{,}000$ no cambia. La varianza depende de la dispersión del tipo de cambio, no de su nivel promedio. Lo que sí aumentaría sería el costo esperado:

$$E(Y) = 50{,}000(21.5) = \$1{,}075{,}000$$

## 5.1 Distribuciones discretas

**5.1.1**

a) $P(X = 1) = \frac{1}{8} = 0.125$

b) ($X$ sea par) $= \frac{4}{8} = 0.5$

c) $P(\text{DeepCoders}) = \frac{1}{8} = 0.125$

d) $E(X) = \frac{1+8}{2} = 4.5$

**5.1.2**

*a)* $P(\text{premium}) = \frac{3}{10} = 0.30$

*b)* $P(A \text{ a } H) = \frac{6}{10} = 0.60$

*c)* $E(X) = \frac{1+10}{2} = 5.5$

*d)* $Var(X) = \frac{10^2-1}{12} = \frac{99}{12} = 8.25$

**5.1.3**

a) $P(X = 0) = 1 - 0.87 = 0.13$

b) $E(X) = \mathrm{p} = 0.87$
$Var(X) = \mathrm{p}(1 - \mathrm{p}) = (0.87)(0.13) = 0.1131$

Un $87\%$ de éxito puede parecer alto, pero implica que el sistema falla en el 13% de los intentos. Si se utiliza con mucha frecuencia o en procesos críticos, esta tasa de error podría ser demasiado elevada; Por lo tanto, considerar $\mathrm{p} = 0.87$ como "bueno" depende del nivel de confiabilidad requerido y del costo de los errores.

**5.1.4**

a) $P(X = 0) = 1 - 0.93 = 0.07$

b) $E(X) = p = 0.93$

$Var(X) = p(1 - p) = (0.93)(0.07) = 0.0651$

c) Un 7% de errores puede afectar injustamente a algunos estudiantes. En evaluaciones de alto impacto, el detector de IA no debería ser el único criterio de decisión; sus resultados deben complementarse con una revisión por parte del profesor para garantizar un proceso justo.

**5.1.5** $M$ =número de personas que dan "Like"

a) La función de probabilidad es:

$$(M = m) = \binom{8}{m}(0.5)^m(0.5)^{8-m}, \quad m = 0,1,\dots,8$$

b) $P(M \geq 6) = P(M = 6) + P(M = 7) + P(M = 8) = \binom{8}{6}(0.5)^8 + \binom{8}{7}(0.5)^8 + \binom{8}{8}(0.5)^8 = 0.1445$

c) $E(M) = np = 8(0.5) = 4$

Se espera que, en promedio, 4 de cada 8 personas den “Like” al mensaje.

d) $Var(M) = np(1-p) = 8(0.5)(0.5) = 2$

Indica la variabilidad del número de “Likes” alrededor de su valor esperado de 4 entre distintos grupos de 8 personas.

**5.1.6**

a) $P(X \geq 30) = \sum_{x=30}^{40} \binom{40}{x} (0.78)^x (0.22)^{40-x} = 0.7483$

b) Sea $\boldsymbol{F}$ = número de respuestas incorrectas:

$$F \sim Binomial(40, 0.22)$$

$$P(F > 10) = P(F \geq 11) = 0.2517$$

c) $E(X) = np = 31.2$

Se espera que el chatbot responda correctamente, en promedio, 31.2 de las 40 preguntas diarias.

**5.1.7** Sea $X$ = número de estudiantes que firman la petición:

$$X \sim Binomial(20, 0.12)$$

a) $P(X = 15) = \binom{20}{15}(0.12)^{15}(0.88)^5 \approx 1.26 \times 10^{-10}$b

b) $P(X \geq 15) = \sum_{x=15}^{20} \binom{20}{x} (0.12)^x (0.88)^{20-x} \approx 1.32 \times 10^{-10}$

c) $P(X > 10) = P(X \geq 11) \approx 0.00000439$

d) $E(X) = np = 20(0.12) = 2.4$

$Var(X) = np(1-p) = 20(0.12)(0.88) = 2.112$

e) $P(X = 0) = (0.88)^{20} = 0.0776$

Existe aproximadamente un 0.0776 de probabilidad de no obtener ninguna firma. Operativamente, esto indicaría que obtener cero firmas no sería un resultado excepcional y podría ocurrir aun si la tasa histórica de firma del $12\%$ se mantiene.

f ) Para una distribución binomial, la moda es:

$$(n+1)p = 21(0.12) = 2.52$$

$$\text{Moda}(X) = 2$$

Esto significa que el resultado individual más probable es de 2 firmas.

**5.1.8** Sea $X$ =número de alertas que resultan ser falsos positivos:

$$X \sim Binomial(12, 0.85)$$

a) $P(X = 12) = (0.85)^{12} = 0.1422$

b) $P(X \geq 10) = \sum_{x=10}^{12} \binom{12}{x} (0.85)^x (0.15)^{12-x} = 0.7358$

Existe aproximadamente un 0.7358 de probabilidad de que al menos 10 de las 12 alertas sean falsos positivos.

c) Sea $Y =$ número de alertas que sí requieren investigación profunda. Entonces:

$$Y \sim Binomial(12, 0.15)$$

$$E(Y) = 12(0.15) = 1.8$$

$$SD(Y) = \sqrt{12(0.15)(0.85)} = 1.237$$

En promedio, se espera que 1.8 de cada 12 alertas requieran una investigación profunda, con una desviación estándar de aproximadamente 1.24 alertas.

**5.1.9** Sea $X =$ número de padres de familia que rechazan la vacuna:

$$X \sim Binomial(25, 0.20)$$

a) $E(X) = np = 25(0.20) = 5$

$Var(X) = np(1-p) = 25(0.20)(0.80) = 4$

b) $P(X > 8) = P(X \geq 9) = \sum_{x=9}^{25} \binom{25}{x} (0.20)^x (0.80)^{25-x} = 0.0468$

c) El protocolo se activa si los rechazos superan 7 casos:

$$P(X > 7) = P(X \geq 8) = 0.1091$$

Existe aproximadamente un 10.91% de probabilidad de que la brigada tenga que activar el protocolo.

Una muestra de 25 personas permite una evaluación inicial, pero puede ser insuficiente para estimar con precisión el riesgo real de toda la comunidad. Sería recomendable utilizar una muestra más grande y representativa antes de tomar decisiones de asignación de recursos a gran escala.

**5.1.10** Sea $X =$ número de intentos necesarios hasta que el asistente reconoce correctamente la instrucción:

$$X \sim Geométrica(0.85)$$

con:

$$P(X = x) = (0.15)^{x-1}(0.85),\ x = 1,2,\dots$$

a) $P(X = 1) = 0.85$

b) Para necesitar exactamente 3 intentos, debe fallar los dos primeros y acertar el tercero:

$$P(X = 3) = (0.15)^2(0.85) = 0.019125$$

c) Requerir más de 5 intentos significa que los primeros 5 intentos fallan:

$$P(X > 5) = (0.15)^5 = 0.00007594$$

**5.1.11** Sea $X =$ número de intentos necesarios hasta lograr subir correctamente el archivo:

$$X \sim Geométrica(0.85)$$

con:

$$P(X = x) = (0.15)^{x-1}(0.85),\ x = 1,2,\dots$$

a) $P(X = 2) = (0.15)(0.85) = 0.1275$

b) $P(X \leq 4) = 1 - P(X > 4) = 1 - (0.15)^4 = 0.99949375$

c) $E(X) = \frac{1}{p} = \frac{1}{0.85} = 1.1765$

Se requieren en promedio aproximadamente 1.18 intentos para lograr una carga exitosa.

**5.1.12** Sea $Y =$número de personas entrevistadas hasta obtener la primera respuesta afirmativa.

a) $Y \sim Geométrica(0.4\,)$ con:

$$P(Y = y) = (0.6)^{y-1}(0.4),\ y = 1,2,\dots$$

c) $P(Y = 3) = (0.6)^2(0.4) = 0.144$

d) $E(Y) = \frac{1}{p} = \frac{1}{0.4} = 2.5$

El encuestador necesita entrevistar a 2.5 personas en promedio, para obtener la primera respuesta afirmativa.

**5.1.13** Sea $X =$ número de intentos necesarios hasta lograr conectarse:

$$X \sim Geométrica(0.72)$$

con $q = 1 - p = 0.28$.

a) Para conectarse en el tercer intento debe fallar los dos primeros:

$$P(X = 3) = (0.28)^2(0.72) = 0.05645$$

b) Necesitar al menos 4 intentos significa fallar los primeros 3:

$$P(X \geq 4) = (0.28)^3 = 0.02195$$

c) $E(X) = \frac{1}{0.72} = 1.3889$

Se requieren aproximadamente 1.39 intentos, en promedio, para lograr conectarse.

**5.1.14** Sea $X =$ número de bicicletas que llegan en una hora:

$$X \sim Poisson(4.5)$$

a) $P(X = 6) = \frac{e^{-4.5}(4.5)^6}{6!} = 0.1281$

b) $P(X \geq 5) = 1 - P(X \leq 4) = 1 - \sum_{x=0}^{4} \frac{e^{-4.5}(4.5)^x}{x!} = 0.4679$

c) Para 2 horas, el parámetro es: $\lambda_{2h} = 2(4.5) = 9$

d) $E(X) = 9$

En un periodo de 2 horas se espera que lleguen, en promedio, 9 bicicletas.

**5.1.15** Sea $X =$ número de notificaciones recibidas en una hora:

$$X \sim Poisson(18)$$

a) $P(X = 20) = \frac{e^{-18}(18)^{20}}{20!} = 0.0798$

b) Recibir menos de 15 significa $X < 15$, es decir:

$$P(X < 15) = PX \leq 14 \sum_{x=0}^{14} \frac{e^{-18}(18)^{x}}{x!} = 0.2081$$

c) Para 30 minutos:

$$\lambda_{30\ \text{min}} = 18\left(\frac{30}{60}\right) = 9$$

Por lo tanto, se esperan 9 notificaciones en un periodo de 30 minutos.

**5.1.16** Para una llamada de 15 minutos:

$$X \sim Poisson(3)$$

a) $P(X = 0) = e^{-3} = 0.0498$

b) $P(X = 2) = \frac{e^{-3}3^{2}}{2!} = 0.2240$

c) Para media hora, el parámetro cambia proporcionalmente: $\lambda_{30} = 3(2) = 6$
Entonces $X \sim Poisson(6)$:

$$P(X \geq 5) = 1 - P(X \leq 4) = 0.7149$$

d) Para una hora: $\lambda_{60} = 3(4) = 12$
Entonces $X \sim Poisson(12)$:

$$P(X < 6) = P(X \leq 5) = 0.0203$$

**5.1.17** a) Sea $X =$ número de micro-fallas en una hora:

$$X \sim Poisson(4)$$
$$P(X = 5) = \frac{e^{-4}4^{5}}{5!} = 0.1563$$

b) Dado que ocurrieron 5 micro-fallas, sea $\boldsymbol{Y}$ =número que detienen el video:

$$Y \sim Binomial5, 0.15$$
$$P(Y = 2) = \binom{5}{2}(0.15)^{2}(0.85)^{3} = 0.1382$$

c) Sea $Z =$ número de intentos hasta que el video vuelve a reproducirse:

$$Z \sim Geométrica(0.8)$$

Para necesitar exactamente 3 intentos, deben fallar los dos primeros y funcionar el tercero:

$$P(Z = 3) = (0.2)^2(0.8) = 0.032$$

**5.1.18** a) Sea $X =$ número de pedidos que llegan en un minuto:

$$X \sim Poisson(6)$$
$$P(X = 8) = \frac{e^{-6}6^8}{8!} = 0.1033$$

b) Hay 4 tipos de pedidos con igual probabilidad:

$$P(\text{Farmacia}) = \frac{1}{4} = 0.25$$

c) Dado que llegaron 8 pedidos, sea $\boldsymbol{Y}$ =número asignado a motocicleta:

$$Y \sim Binomial8, 0.4$$
$$P(Y = 3) = \binom{8}{3}(0.4)^3(0.6)^5 = 0.2787$$

d) Sea $Z =$ número de intentos hasta lograr la entrega exitosa en bicicleta:

$$Z \sim Geométrica(0.7)$$

Para entregarse en el tercer intento, deben fallar los dos primeros y tener éxito el tercero:

$$P(Z = 3) = (0.3)^2(0.7) = 0.063$$

**5.1.19** Este ejercicio combina Poisson, Binomial y Geométrica.

a) Sea $X$ =número de intentos sospechosos en una hora:

$$X \sim Poisson(9)$$
$$P(X = 12) = \frac{e^{-9}9^{12}}{12!} = 0.0728$$

b) Dado que hubo 12 intentos, sea $Y$ =número de intentos realmente maliciosos:

$$Y \sim Binomial(12, 0.25)$$
$$P(Y = 3) = \binom{12}{3}(0.25)^3(0.75)^9 = 0.2581$$

c) Sea $Z$ =número de intentos hasta que el firewall logra bloquear el acceso:

$$Z \sim Geometrica(0.9)$$

Para necesitar exactamente 3 intentos, debe fallar los dos primeros y bloquearlo en el tercero:

$$P(Z = 3) = (0.1)^2(0.9) = 0.009$$

Por lo tanto, la probabilidad es 0.9%.

**5.1.20** a) Sea $X$ = número de consultas recibidas en una hora:

$$X \sim Poisson(12)$$
$$P(X = 15) = \frac{e^{-12}12^{15}}{15!} = 0.0724$$

b) $P(\text{Probabilidad}) = \frac{1}{5} = 0.20$

c) Dadas 15 consultas, sea $Y$ =número respondido correctamente en el primer intento:

$$Y \sim Binomial(15, 0.85)$$
$$P(Y = 13) = \binom{15}{13}(0.85)^{13}(0.15)^2 = 0.2303$$

d) Como ya se sabe que la primera respuesta fue incorrecta, para necesitar 4 intentos en total deben fallar también los intentos 2 y 3, y acertar en el 4:

$$P(T = 4 \mid T > 1) = (0.15)^2(0.85) = 0.019125$$

## 5.2 Distribuciones continuas

**5.2.1** Sea:

$$X \sim U(2, 10)$$

a) La función de densidad es:

$$f(x) = \begin{cases} \frac{1}{8}, & 2 \le x \le 10 \\ 0, & \text{en otro caso} \end{cases}$$

b) $P(X \le 5) = \frac{5-2}{10-2} = 0.375$

c) $P(4 \leq X \leq 8) = \frac{8-4}{10-2} = 0.50$

d) $E(X) = \frac{2+10}{2} = 6$ días

$Var(X) = \frac{(10-2)^2}{12} = \frac{16}{3} \approx 5.3333$ días$^2$

e) $P(X > 7) = \frac{10-7}{10-2} = 0.375$

Por lo tanto, el 37.5% de los trámites tarda más de 7 días.

f) La distribución uniforme es razonable si cualquier tiempo entre 2 y 10 días tiene aproximadamente la misma posibilidad de ocurrir. Puede no ser adecuada si los tiempos se concentran alrededor de ciertos días o si existen retrasos que hacen algunos tiempos más frecuentes que otros.

**5.2.2** Sea $X \sim U16.8, 17.6$

a) La función de densidad es:

$$f(x) = \begin{cases} \frac{1}{17.6-16.8} = 1.25 & 16.8 \leq x \leq 17.6 \\ 0 & \text{en otro caso} \end{cases}$$

b) $P(X < 17) = \frac{17-16.8}{17.6-16.8} = 0.25$

c) $P(17.2 < X < 17.5) = \frac{17.5-17.2}{17.6-16.8} = 0.375$

d) $E(X) = \frac{16.8+17.6}{2} = 17.2$

$Var(X) = \frac{(17.6-16.8)^2}{12} = 0.0533$

e) Como $E(X) = 17.2$, estar a más de \$0.50 implica:

$$| X - 17.2 | > 0.50$$

es decir, $X < 16.7$ o $X > 17.7$. Ambos valores están fuera del intervalo $[16.8, 17.6]$, por lo que:

$$\boxed{P(| X - 17.2 | > 0.50) = 0}$$

f ) La variabilidad cambiaria genera incertidumbre en el costo de las importaciones. Una depreciación del peso aumenta el costo en pesos de las compras en dólares y puede reducir los márgenes de ganancia.

**5.2.3** Sea $X \sim Exp(\lambda)$, $E(X) = 4$

a $\frac{1}{\lambda} = 4 \Rightarrow \ = 0.25$ quejas/hora

b) $P(X < 2) = 1 - e^{-0.25(2)} = 0.3935$

c) $P(X > 6) = e^{-0.25(6)} = 0.2231$

d) Este resultado, por sí solo, no permite concluir si el sistema es estable o problemático. Una menor frecuencia de quejas podría ser favorable, pero se necesita información adicional sobre el volumen de clientes y las causas de las quejas.

e) Si se desea:

$$P(X > 6) < 0.05$$

entonces:

$$e^{-6\lambda} < 0.05$$

lo que requiere:

$$\lambda > 0.4993$$

Es decir, tendría que aumentar la tasa de llegada de quejas, lo cual contradice un objetivo usual de mejora del servicio. Por ello, convendría reformular el criterio: si la empresa busca reducir las quejas, debería disminuir $\boldsymbol{\lambda}$, lo que haría más probable esperar más de 6 horas entre ellas.

- aumentar $\lambda$ implica más quejas por unidad de tiempo
- disminuir $\lambda$ implica menos quejas por unidad de tiempo

f) Por la propiedad de falta de memoria, el tiempo que ya ha transcurrido sin recibir una queja no modifica la distribución del tiempo adicional de espera. Por lo tanto, la asignación de personal no debería basarse únicamente en cuánto tiempo ha pasado desde la última queja.

**5.2.4** Sea $X \sim Exp(\lambda)$, $E(X) = 18$

a) $\frac{1}{\lambda} = 18 \Rightarrow \lambda = \frac{1}{18} \approx 0.0556$ auditorías/mes

b) $P(X > 24) = e^{-(1/18)(24)} = 0.2636$

c) $P(6 < X < 12) = P(X > 6) - P(X > 12) = e^{-6/18} - e^{-12/18} = 0.2031$

d) Existe una probabilidad cada vez menor de permanecer varios años sin auditoría. Por ejemplo, la probabilidad de pasar más de 2 años sin auditoría es de 26.36%; esto no elimina el riesgo de ser auditado posteriormente.

e) Si aumenta la tasa de auditorías $\lambda$, disminuye el tiempo promedio entre auditorías y $P(X > t) = e^{-\lambda t}$ disminuye. Por lo tanto, sería menos probable que una empresa permanezca largos periodos sin ser auditada.

**5.2.5** Sea $X \sim Exp(\lambda)$, $E(X) = 5$

a) $\frac{1}{\lambda} = 5 \Rightarrow \lambda = 0.20$ fallas/año

b) $P(X > 7) = e^{-0.2(7)} = 0.2466$

c) $P(X \leq 3) = 1 - e^{-0.2(3)} = 0.4512$

d) La distribución exponencial puede ser razonable si las fallas ocurren aleatoriamente, de manera independiente y con una tasa aproximadamente constante en el tiempo.

e) $P(X > 5) = e^{-0.2(5)} = e^{-1} = 0.3679$

f) No necesariamente. Reemplazarlo antes de 5 años depende del costo de una falla frente al costo de reemplazo preventivo. Además, bajo el modelo exponencial, el servidor tiene una tasa de falla constante, por lo que llegar a los 5 años no implica que el riesgo instantáneo de falla aumente por su antigüedad.

**5.2.6** Sea:

$$X \sim N72, 64$$

a) $P(X > 85) = P\left(Z > \frac{85-72}{8}\right) = P(Z > 1.625) \approx 0.0521$

b) $P(65 < X < 80) = P(-0.875 < Z < 1) \approx 0.6508$

c) $P(X < 60) = P(Z < -1.5) = 0.0668$

d) Para seleccionar al 10% superior, se requiere el percentil 90:

$$z_{0.90} \approx 1.2816$$
$$x = 72 + 1.2816(8) \approx 82.25$$

El punto de corte debe ser aproximadamente 82.25 puntos.

e) Estar en el percentil 90 significa tener un puntaje igual o superior al de aproximadamente el 90% de los candidatos; solo alrededor del 10% obtiene un puntaje mayor.

f) Un umbral automático puede reproducir sesgos presentes en los datos o en el algoritmo y excluir candidatos adecuados. Por ello, conviene complementar la selección automática con revisión humana, auditorías de sesgo y criterios transparentes.

**5.2.7** Sea:

$$X \sim N45, 36$$

a) $P(X > 55) = P\left(Z > \frac{55-45}{6}\right) = P(Z > 1.6667) \approx 0.0478$

b) $P(38 < X < 50) = P(-1.1667 < Z < 0.8333) \approx 0.6763$

c) $P(X > 60) = P(Z > 2.5) = 0.0062$

d) Si se busca el valor que delimita el 95% superior, corresponde al percentil 5:

$$z_{0.05} = -1.6449$$
$$x = 45 - 1.6449(6) \approx 35.13$$

Por lo tanto, aproximadamente el 95% de las muestras tiene más de 35.13 partículas por litro.

Si la intención era encontrar el valor por debajo del cual se encuentra el 95% de las muestras percentil 95, entonces sería $54.87$.

e) Estar en el percentil 97 significa que la muestra tiene una concentración mayor que aproximadamente el 97% de las muestras; solo el 3% presenta concentraciones superiores.

f) Si el límite legal es 50:

$$P(X > 50) = P\left(Z > \frac{50 - 45}{6}\right) = P(Z > 0.8333) \approx 0.2023$$

Aproximadamente 20.23% de la red excede el límite establecido.

**5.2.8** Suponiendo que:

$$X \sim N196, 40^2$$

a) Más de 4 horas equivale a más de 240 minutos:

$$P(X > 240) = P\left(Z > \frac{240 - 195}{40}\right) = P(Z > 1.125) \approx 0.1303$$

b) Entre 2 y 3 horas equivale a entre 120 y 180 minutos:

$$P(120 < X < 180) = P(-1.875 < Z < -0.375) \approx 0.3237$$

c) $P(X < 120) = P(Z < -1.875) = 0.0304$

d) El 20% con mayor uso se encuentra por encima del percentil 80:

$$z_{0.80} \approx 0.8416$$
$$x = 195 + 0.8416(40) \approx 228.66 \text{ minutos}$$

Es decir, aproximadamente 229 minutos diarios 3 h 49 min.

e) Estar por encima del percentil 80 significa pertenecer al 20% de los universitarios que más tiempo dedica a redes sociales.

f) Los datos podrían justificar programas de bienestar digital y uso responsable de redes, especialmente dirigidos a estudiantes con tiempos de uso elevados. Sin embargo, estos datos por sí solos no demuestran que el uso intensivo cause problemas académicos o de bienestar, por lo que no justificarían medidas restrictivas sin evidencia adicional.

**5.2.9** Si

$$X \sim N26, 16$$

a) $P(X < 22) = P\left(Z < \frac{22-26}{4}\right) = P(Z < -1) = 0.1587$

b) $P(24 < X < 30) = P(-0.5 < Z < 1) = 0.5328$

c) $P(X > 35) = P(Z > 2.25) = 0.0122$

d) Para garantizar que el 95% de las entregas llegue antes de cierto tiempo, calculamos el percentil 95:

$$x_{0.95} = 26 + 1.6449(4) \approx 32.58 \text{ horas}$$

Aproximadamente, el 95% de los pedidos se entrega en 32.6 horas o menos.

e) Un pedido en el percentil 90 tarda más que aproximadamente el 90% de los pedidos; solo el 10% tarda más.

f) No parece realista como promesa general. La probabilidad de entregar en 24 horas o menos es:

$$P(X \leq 24) = P(Z \leq -0.5) = 0.3085$$

Solo aproximadamente 30.85% de los pedidos cumpliría la promesa de 24 horas; cerca del 69.15% no la cumpliría.

**5.2.10** Suponiendo una distribución normal bivariada:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2 \left[ \begin{pmatrix} 70 \\ 65 \end{pmatrix}, \begin{pmatrix} 100 & 48 \\ 48 & 64 \end{pmatrix} \right]$$

donde:

$$CovX, Y = \rho \sigma_X \sigma_Y = (0.6)(10)(8) = 48$$

a) $PX > 80, Y > 70 = 0.1004$

Aproximadamente 10.04% de los candidatos supera simultáneamente ambos puntajes.

b) $P\mathrm{X} < 60, \mathrm{Y} > 70 = 0.0052$

Aproximadamente 0.52% tiene un puntaje técnico bajo pero un puntaje socioemocional alto.

c) Individualmente:

$$P(X > 80) = 0.1587$$
$$P(Y > 70) = 0.2660$$

mientras que:

$$P(X > 80, Y > 70) = 0.1004$$

La probabilidad de cumplir ambos criterios simultáneamente es menor que la de cumplir cada criterio por separado.

d Para una normal bivariada:

$$E(Y \mid X = x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

Por lo tanto:

$$E(Y \mid X = 80) = 65 + (0.6)\frac{8}{10}(80 - 70) = 69.8$$

Un candidato con puntaje técnico de 80 tiene un puntaje socioemocional esperado de 69.8 puntos.

e) $\rho = 0.6$ indica una relación lineal positiva moderada-fuerte: candidatos con mayores puntajes técnicos tienden a presentar también mayores puntajes socioemocionales, aunque la relación no es perfecta.

f) Exigir ambos puntajes altos puede ser razonable si ambas competencias son esenciales para el puesto, pero es un criterio restrictivo: solo alrededor del 10% cumpliría $X > 80$ y $Y > 70$. Además, una decisión de contratación no debería depender exclusivamente de puntajes generados por IA; conviene considerar otros criterios y revisar posibles sesgos del sistema.

**5.2.11** Sea:

$$(X, Y) \sim N_2,\ \mu_X = 28,\ \mu_Y = 420,\ \sigma_X = 4,\ \sigma_Y = 60,\ \rho = 0.75$$

a) Estandarizando $P(\mathrm{X} > 32, \mathrm{Y} > 780) = P(Z_X > 1, Z_Y > 1) \approx 0.0838$

b) $P(24 < X < 32, Y > 450) = P(-1 < Z < 1, Y > 0.5) \approx 0.2037$

c) Para una normal bivariada:

$$E(Y \mid X = x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

$$E(Y \mid X = 35) = 420 + (0.75)\frac{60}{4}(35 - 28) = 498.75 \text{ GWh}$$

En un día con temperatura promedio de $35°C$, el consumo esperado es de 498.75 GWh.

d) $Var(Y \mid X = x) = \sigma_Y^2(1 - \rho^2) = 60^2(1 - 0.75^2) = 1575$ GWh$^2$
La varianza condicional no depende del valor específico de $\boldsymbol{X}$.

e) Una correlación de $\rho = 0.75$ indica una relación lineal positiva fuerte: a mayores temperaturas, el consumo eléctrico tiende a ser mayor.

e) El modelo permite pronosticar la demanda eléctrica según la temperatura esperada, identificar periodos de alta demanda y planear capacidad de generación, reservas e infraestructura para reducir el riesgo de saturación o apagones.

**5.2.12** Suponiendo que X,Y sigue una normal bivariada con:

$$\mu_X = 55,\ \mu_Y = 2.8,\ \sigma_X = 10,\ \sigma_Y = 1.2,\ \rho = -0.65$$

a) $P(\mathrm{X} > 60, \mathrm{Y} < 2) = P(Z_X > 0.5, Z_Y < -0.6667) \approx 0.1627$

b) $P(40 < \mathrm{X} < 60, \mathrm{Y} > 3) = \mathrm{P} - 1.5 < Z_X < 0.5, Z_Y > 0.1667 \approx 0.3058$

c) Para la normal bivariada:

$$E(Y \mid X = x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

$$E(Y \mid X = 70) = 2.8 + (-0.65)\frac{1.2}{10}(70 - 55) = 0.0163$$

Para un país con índice de corrupción de 70, el crecimiento esperado del PIB es aproximadamente 1.63% anual.

d) $CovX, Y = \rho\sigma_X\sigma_Y = (-0.65)(10)(1.2) = -7.8$

e) Una correlación de $\rho = -0.65$ indica una relación lineal negativa relativamente fuerte: países con mayores niveles de corrupción tienden a presentar menores tasas de crecimiento económico.

f) Los resultados podrían apoyar políticas de transparencia, rendición de cuentas y fortalecimiento institucional. Sin embargo, la correlación no demuestra causalidad, por lo que deben considerarse otros factores económicos, políticos y sociales antes de diseñar una política pública.

### 5.3 Relación entre la distribución Poisson y la distribución exponencial

5.3.1 Sea $N(t)$el número de quejas en $t$horas. Se supone un proceso de Poisson con tasa:

$$\lambda = 3 \text{ quejas/hora}$$

y el tiempo $T$ entre quejas sigue:

$$T \sim Exp(3)$$

a) En 2 horas: $N(2) \sim Poisson(6)$

b) $P(N(1) = 0) = e^{-3} = 0.0498$

c) 30 minutos $= 0.5$horas:

$$P(T > 0.5) = e^{-3(0.5)} = 0.2231$$

d) Por la propiedad de falta de memoria:

$$P(T > 40/60 \mid T > 20/60) = P(T > 20/60) = e^{-3(1/3)} = 0.3679$$

e) Una hora completa sin quejas ocurre con probabilidad: $0.0498$

Es poco frecuente. Sin embargo, una sola hora sin quejas no es evidencia suficiente de una mejora en la calidad; podría deberse al azar.

f) Se desea:

$$P(T > 0.5) < 0.10$$

Como:

$$e^{-0.5\lambda} < 0.10$$

se requiere:

$$\lambda > 4.6052 \text{ quejas/hora}$$

Por lo tanto, habría que aumentar, no reducir, la tasa de quejas, lo cual no es deseable.

Si el objetivo empresarial es reducir las quejas, sería más lógico plantear que la probabilidad de pasar más de 30 minutos sin quejas sea alta, no menor al 10%.

g) La falta de memoria implica que el tiempo transcurrido desde la última queja no modifica la distribución del tiempo hasta la siguiente. Por ello, los turnos no deberían planearse únicamente con base en cuánto tiempo ha pasado desde la última queja.

h) La tasa podría variar por hora del día, número de usuarios conectados, fallas del sistema, promociones, actualizaciones de la app o días de alta demanda. En esos casos, el supuesto de una tasa $\boldsymbol{\lambda}$constante podría no ser adecuado.

**5.3.2** Sea $N(t)$ el número de clientes que llegan en $\boldsymbol{t}$horas. Se supone un proceso de Poisson con:

$$\lambda = 10 \text{ clientes/hora}$$

y el tiempo $T$entre llegadas sigue:

$$T \sim Exp(10)$$

a) $P(N(1) = 15) = \frac{e^{-10}10^{15}}{15!} = 0.0347$

b) En 30 minutos:

$$\lambda_{30} = 10(0.5) = 5$$
$$P(N(0.5) < 5) = P(N(0.5) \leq 4) = 0.4405$$

c) 5 minutos $= \frac{\mathbf{1}}{\mathbf{12}}$ de hora:

$$P\left(T > \frac{1}{12}\right) = e^{-10(1/12)} = 0.4346$$

d) La propiedad de falta de memoria significa que el tiempo que ya ha transcurrido sin que llegue un cliente no modifica la distribución del tiempo adicional hasta la próxima llegada.

e) No puede determinarse solo con la tasa de llegadas. La subutilización depende también de la capacidad y velocidad de atención de los cajeros.

f) El personal debe poder atender más de 10 clientes por hora para que la capacidad promedio supere la tasa de llegadas y reducir el riesgo de saturación.

g) El modelo permite estimar la demanda esperada y la probabilidad de distintos volúmenes de clientes, ayudando a determinar cuántos cajeros se requieren para mantener tiempos de espera aceptables.

h) El proceso de Poisson podría no ser válido si la tasa de llegada cambia según la hora, los clientes llegan en grupos, las llegadas no son independientes o existen eventos especiales, quincenas o promociones que modifican el flujo habitual.

**5.3.3** Sea $N(t)$ el número de auditorías en $\boldsymbol{t}$ años. Se supone un proceso de Poisson con:

$$\lambda = 2 \text{ auditorías/año}$$

y el tiempo $T$ hasta la próxima auditoría sigue una distribución exponencial.

a) En 3 años:

$$N(3) \sim Poisson(6)$$

$$P(N(3) \leq 2) = \sum_{x=0}^{2} \frac{e^{-6}6^{x}}{x!} = 0.0620$$

b) $P(N(1) = 0) = e^{-2} = 0.1353$

c) $T \sim Exp(2)$
El tiempo esperado hasta la próxima auditoría es:

$$E(T) = \frac{1}{2} \text{ año} = 6 \text{ meses}$$

d) 6 meses $= 0.5$años:

$$P(T > 0.5) = e^{-2(0.5)} = 0.3679$$

e) La probabilidad de permanecer varios años sin auditoría disminuye rápidamente. Por ejemplo:

$$P(T > 3) = e^{-6} = 0.0025$$

Solo aproximadamente 0.25% pasaría más de 3 años sin auditoría bajo este modelo.

f) Si la tasa se duplica a $\lambda = 4$, el tiempo promedio entre auditorías disminuye de 6 a 3 meses, aumentando la probabilidad de ser auditado en periodos cortos.

g) La empresa debería mantener cumplimiento fiscal permanente, registros contables actualizados y documentación adecuada, sin depender de la expectativa de que una auditoría tarde en ocurrir.

h) En reformas fiscales, la tasa de auditorías podría cambiar con el tiempo, concentrarse en ciertos sectores o depender del perfil de riesgo de cada empresa. Esto violaría los supuestos de tasa constante e independencia del proceso de Poisson.

**5.3.4** Sea $N(t)$ el número de fallas en $t$años. Se supone un proceso de Poisson con:

$$\lambda = 4 \text{ fallas/año}$$

y el tiempo $T$entre fallas sigue una distribución exponencial.

a) $P(N(1) = 5) = \frac{e^{-4}4^5}{5!} = 0.1563$

b) 3 meses $= 0.25$años:

$$P(N(0.25) = 0) = e^{-4(0.25)} = 0.3679$$

c) 2 meses $= \frac{1}{6}$ de año:

$$P\left(T > \frac{1}{6}\right) = e^{-4(1/6)} = 0.5134$$

d) Por la propiedad de falta de memoria:

$$P(T > 3 \text{ meses} \mid T > 1 \text{ mes}) = P(T > 2 \text{ meses}) = 0.5134$$

e) El sistema presenta en promedio una falla cada 3 meses. Su confiabilidad debe evaluarse considerando también la duración, gravedad y costo de las fallas; la tasa por sí sola no determina si el sistema es suficientemente confiable.

f) El modelo no determina por sí solo la frecuencia óptima de mantenimiento. Como referencia, el tiempo promedio entre fallas es:

$$E(T) = \frac{1}{4} \text{ año} = 3 \text{ meses}$$

El mantenimiento podría programarse antes de ese intervalo, considerando costos y evidencia técnica.

g) En una distribución Poisson $Var(N) = E(N) = \lambda$

h) Si una mejora tecnológica reduce la tasa de fallas $\boldsymbol{\lambda}$, también disminuyen tanto el número esperado como su varianza.

i) Eventos como huracanes, olas de calor, terremotos, sobrecargas masivas o fallas en cascada podrían generar múltiples fallas dependientes y alterar la tasa constante, rompiendo los supuestos del modelo de Poisson.

## 5.4 Aproximación de la distribución binomial y Poisson por medio de la distribución normal

**5.4.1** Sea $X \sim Binomial(400, 0.52)$

a) Aproximación normal

$$np = 400(0.52) = 208 > 5$$
$$n(1-p) = 400(0.48) = 192 > 5$$

Por lo tanto, la aproximación normal es adecuada:

$$X \approx N208, 99.84$$
$$\sigma = \sqrt{99.84} \approx 9.992$$

b) $P(X \geq 220)$

Con corrección por continuidad:

$$P(X \geq 219.5) = \mathrm{P}\left(Z \geq \frac{219.5 - 208}{9.992}\right) = \mathrm{PZ} \geq 1.151 \approx 0.125$$

c) Aunque el apoyo esperado es del 52%, observar 220 o más personas a favor —es decir, 55% o más de la muestra— es relativamente poco frecuente: ocurre con una probabilidad aproximada del 12.5% bajo este modelo.

d) No necesariamente. Observar una mayoría en la muestra no basta para confirmar estadísticamente que la mayoría de toda la población apoya la reforma. Se requeriría realizar inferencia estadística sobre la proporción poblacional.

e) Si $n = 50$:

$$np = 26,\ n(1-p) = 24$$

La aproximación normal seguiría siendo válida, pero habría mayor variabilidad relativa y menor precisión que con una muestra de 400 personas.

**5.4.2** Sea:

$$X \sim Binomial(600, 0.04)$$

a) Aproximación normal

$$np = 600(0.04) = 24 > 5$$
$$n(1-p) = 600(0.96) = 576 > 5$$

Por lo tanto, sí puede utilizarse la aproximación normal:

$$X \approx N(24, 23.04)$$

$$\sigma = \sqrt{23.04} = 4.8$$

$b)\ P(X > 30) = P(X \geq 31)$

Con corrección por continuidad:

$$P(X > 30.5) = \mathrm{P}\left(Z > \frac{30.5 - 24}{4.8}\right) = \mathrm{PZ} > 1.3542 \approx 0.0879$$

e) Observar más de 30 defectos es relativamente poco frecuente si la tasa histórica de defectos realmente es del 4%. Esto podría justificar revisar el proceso de producción.

f) No necesariamente. Un resultado aislado con más de 30 defectos no demuestra por sí solo que el proceso esté fuera de control; se requiere analizar su comportamiento a lo largo del tiempo mediante herramientas de control estadístico.

g) Una mala aproximación puede producir probabilidades inexactas y llevar a falsas alarmas o a no detectar problemas reales en el proceso.

h) Si esta probabilidad superara el 10%, observar más de 30 defectos sería relativamente común bajo el proceso actual, por lo que 30 defectos podría ser un límite demasiado estricto. Conviene revisar el estándar o, preferentemente, implementar acciones para reducir la tasa de defectos si ese nivel no es aceptable.

i) Esta información permite establecer límites de tolerancia basados en probabilidades, definir umbrales de alerta y detectar cuándo el número de defectos es suficientemente inusual como para investigar el proceso.

**5.4.3** Sea $X \sim Poisson(50)$

Como $\lambda$ es grande, utilizamos:

$$X \approx N(50, 50),\ \sigma = \sqrt{50} \approx 7.071$$

a) Aproximación normal

Como:

$$\lambda = 50 > 10$$

la distribución Poisson es aproximadamente simétrica y la aproximación normal es razonable.

b) $P(45 \leq X \leq 60)$

Con corrección por continuidad:

$$P(44.5 < X < 60.5) = \mathrm{P}\left(\frac{44.5-50}{\sqrt{50}} < Z < \frac{60.5-50}{\sqrt{50}}\right) = \mathrm{P} - 0.778 < Z < 1.485 \approx 0.713$$

c) Aproximadamente el 71% de los años se esperaría observar entre 45 y 60 accidentes si se mantiene la tasa histórica de 50 accidentes anuales.

d) Sí. El intervalo de 45 a 60 accidentes puede considerarse relativamente habitual bajo el modelo histórico, ya que contiene aproximadamente el 71% de los posibles resultados.

e) $P(X > 60) = P(X \geq 61)$
Con corrección por continuidad:

$$P(X > 60.5) = \mathrm{P}\left(Z > \frac{60.5-50}{\sqrt{50}}\right) = \mathrm{PZ} > 1.485 \approx 0.0688$$

f) Aunque superar 60 accidentes es poco frecuente, la probabilidad no es nula. Se justifican acciones como reforzar la capacitación, revisar protocolos, identificar áreas de mayor riesgo y mejorar las medidas preventivas.

g) Se puede comparar la tasa de accidentes antes y después de implementar la política, considerando también la cantidad de trabajadores u horas trabajadas. Una reducción estadísticamente significativa permitiría evaluar si la política fue efectiva.

**Minicaso: Monitoreo diario de e-scooters**

La tasa de reportes por hora es:

$$\lambda = \frac{12}{9} = \frac{4}{3} = 1.3333 \text{ reportes/hora}$$

a) Al menos 2 e-scooters con fallas entre 15 inspeccionados
Se usa una Binomial, porque se inspecciona un número fijo de e-scooters, cada uno puede presentar falla o no, con $p = 0.1$:

$$X \sim Bin(15, 01)$$

$$P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - (0.9)^{15} - 15(0.1)(0.9)^{14} = 0.4510$$

b) Exactamente 5 reportes en 2 horas
Se usa una Poisson, porque se cuenta el número de reportes en un intervalo de tiempo.
Para 2 horas:

$$\lambda_2 = \frac{4}{3}(2) = \frac{8}{3} = 2.6667$$

$$X \sim Poisson(2.6667)$$

$$P(X = 5) = \frac{e^{-2.6667}(2.6667)^5}{5!} \approx 0.0781$$

c) Más de 3 horas entre dos reportes consecutivos
Se usa una Exponencial, porque se estudia el tiempo entre eventos de un proceso Poisson:

$$T \sim Exp\left(\frac{4}{3}\right)$$

$$P(T > 3) = e^{-(4/3)(3)} = e^{-4} = 0.0183$$

d) Exactamente 3 e-scooters en buen estado antes de encontrar la primera con falla
Se usa una Geométrica, con éxito = encontrar una e-scooter con falla, $p = 0.1$.
Para encontrar exactamente 3 en buen estado antes de la primera con falla, la secuencia debe ser:

$$B, B, B, F$$

Por lo tanto:

$$PX = 3 = (0.9)^3(0.1) = 0.0729$$

e) Entre 36 y 45 e-scooters con fallas, de una muestra de 500
Originalmente:

$$X \sim Bin(500, 0.1)$$

La aproximación Normal es válida porque:

$$np = 50,\ n(1-p) = 450$$

Ambos son suficientemente grandes.
Entonces:

$$X \approx N(50, 45)$$
$$\mu = 50,\ \sigma = \sqrt{45} \approx 6.708$$

Se pide:

$$P(36 \leq X \leq 45)$$

Aplicando corrección por continuidad:

$$P(35.5 < X < 45.5) = \mathrm{P}\left(\frac{35.5-50}{6.708} < Z < \frac{45.5-50}{6.708}\right) = \mathrm{P} - 2.162 < Z < -0.671$$
$$= 0.2512 - 0.0153 \approx 0.2359$$

**Minicaso: Crisis viral en redes sociales**

a) Un solo comentario: negativo o no
Definimos:

$$X = \begin{cases} 1 & \text{si el comentario es negativo} \\ 0 & \text{si no es negativo} \end{cases}$$

Por lo tanto:

$$X \sim Bernoulli(0.20)$$
$$P(X = 1) = 0.20,\ P(X = 0) = 0.80$$

b) Al menos 4 negativos entre 15 comentarios

$$X \sim Bin15, 0.20$$
$$P(X \geq 4) = 1 - P(X \leq 3) \approx 0.3518$$

c) Exactamente 3 comentarios negativos en una hora
Se usa una Poisson, porque contamos eventos aleatorios en un intervalo de tiempo con tasa constante:

$$X \sim Poisson(1)$$
$$P(X = 3) = \frac{e^{-1}(1)^3}{3!} = 0.0613$$

d) Más de 45 minutos entre dos comentarios negativos
Se usa una Exponencial, porque modela el tiempo entre eventos de un proceso Poisson:

$$T \sim Exp(1)$$

45 minutos $= 0.75$horas:

$$P(T > 0.75) = e^{-1(0.75)} = 0.4724$$

e) Primer comentario negativo después de exactamente 5 no negativos
Se usa una Geométrica, porque se revisan comentarios hasta encontrar el primer negativo.

La secuencia debe ser:

$$NN,\ NN,\ NN,\ NN,\ NN,\ N$$

Por lo tanto:

$$PX = 5 = (0.80)^5(0.20) = 0.0655$$

f) Tiempo de respuesta mayor a 20 minutos
Suponiendo que cualquier tiempo entre 0 y 30 minutos es igualmente probable:

$$T \sim U(0, 30)$$

Se usa una Uniforme continua:

$$P(T > 20) = \frac{30 - 20}{30 - 0} = 0.3333$$

g) Asignación entre 6 community managers
Sea $X$ =número del community manager asignado:

$$X \in \{1, 2, 3, 4, 5, 6\}$$

Se usa una Uniforme discreta, porque todos tienen la misma probabilidad:

$$P(X = x) = \frac{1}{6}$$

Por lo tanto:

$$P(X = 4) = \frac{1}{6} = 0.1667$$

$$E(X) = \frac{1 + 6}{2} = 3.5$$

$$Var(X) = \frac{6^2 - 1}{12} = \frac{35}{12} \approx 2.9167$$

Nota: el valor esperado $3.5$ es el promedio numérico de las etiquetas asignadas a los managers; no representa un "manager promedio" con significado operativo.

h) Porcentaje diario de sentimiento negativo
Suponemos:

$$Y \sim N45{,}100$$

Para superar 55%:

$$P(Y > 55) = P\left(Z > \frac{55 - 45}{10}\right) = P(Z > 1) = 0.1587$$

Para estar entre 35% y 50%:

$$P(35 < Y < 50) = P(-1 < Z < 0.5) = 0.5328$$

i) Entre 60 y 90 comentarios negativos de un total de 400
Originalmente:

$$X \sim Bin400, 0.20$$

Como:

$$np = 80,\ n(1-p) = 320$$

se puede utilizar la aproximación Normal:

$$X \approx N80, 64$$
$$\sigma = \sqrt{64} = 8$$

Con corrección por continuidad:

$$P(60 \leq X \leq 90) \approx P(59.5 < Y < 90.5) = \mathrm{P}\left(\frac{59.5-80}{8} < Z < \frac{90.5-80}{8}\right)$$
$$= \mathrm{P} - 2.5625 < Z < 1.3125 \approx 0.9001$$

j) Riesgos si el modelo no refleja el proceso real
Si el modelo no representa adecuadamente el proceso real, las probabilidades estimadas pueden ser incorrectas y conducir a decisiones equivocadas. Por ejemplo, podría existir dependencia entre comentarios, cambios en la tasa de publicación o mayor variabilidad que la supuesta.

k) Decisiones de campaña
Los resultados podrían utilizarse para reforzar el equipo de comunicación en periodos de alta negatividad, activar protocolos de respuesta rápida, ajustar mensajes y detectar incrementos inusuales del sentimiento negativo. Las decisiones no deberían basarse únicamente en el volumen de comentarios, sino también en su alcance, origen y contenido.

l) Se requieren datos históricos sobre frecuencia, clasificación y tiempos de publicación de los comentarios. Debe verificarse, entre otros aspectos, la estabilidad de la proporción de comentarios negativos, la constancia de la tasa de publicación, la independencia de los eventos y el ajuste de las distribuciones propuestas.

**Historial de versiones**

| Versión | Fecha | Revisión | Cambios |
|---|---|---|---|
| **V0** | Agosto 2026 | Elaboración por Liliana De la Torre Desentis | Versión original |
| | | | |
| | | | |