Primer parcial

Daniel Gonzalez

2023-08-29

### Problema 1

[1.] En Colombia, durante un período histórico, se creía que la distribución de grupos sanguíneos era la siguiente: tipo A, 32%; tipo B, 24%; tipo AB, 5%; y tipo O, 39%. Se estimaba que en ese tiempo, el 3% de las personas pertenecientes al tipo O fue clasificado como del tipo A; el 90% de los del tipo A fue correctamente clasificado; el 5% de los del tipo B se clasificó como del tipo A, y el 8% de los del tipo AB fue clasificado incorrectamente como del tipo A. Un soldado herido fue llevado a la enfermería y se le clasificó como del tipo A. ¿Cuál es la probabilidad de que tal grupo sanguíneo sea ciertamente el suyo?

### **Solucion**

Utilizaremos el teorema de Bayes para calcular la probabilidad condicional:

Para calcular , utilizamos la ley de probabilidad total:

Finalmente, podemos calcular la probabilidad:

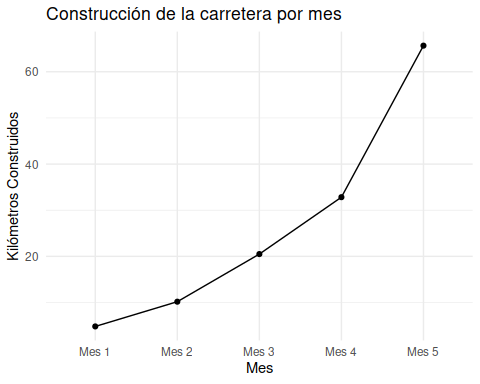
Por lo tanto, la probabilidad de que el grupo sanguíneo del soldado sea ciertamente Tipo A es aproximadamente del 99.65%.

### Problema 2

[2.] Elabore un gráfico para representar adecuadamente la siguiente información:

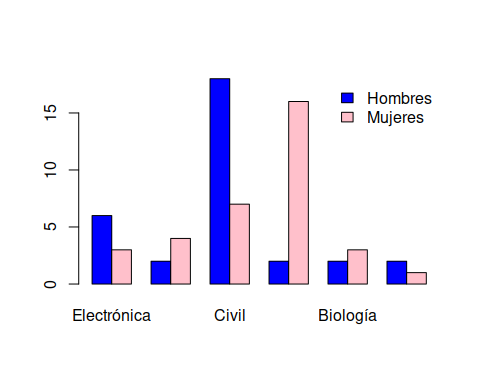
[a.] Durante 5 meses se construyen 134 kilómetros de carretera en la siguiente forma: En el primer mes, 3.60% del total del proyecto; en el segundo mes un 7.60% del total; en el tercer mes, el 15.3% del total; en el cuarto mes 24.5% del total y en último mes, el 49% restante.

# Instala y carga la biblioteca ggplot2 si aún no está instalada  
if (!requireNamespace("ggplot2", quietly = TRUE)) {  
 install.packages("ggplot2")  
}  
library(ggplot2)  
  
# Datos  
meses <- c("Mes 1", "Mes 2", "Mes 3", "Mes 4", "Mes 5")  
porcentaje\_construido <- c(3.60, 7.60, 15.3, 24.5, 49.0)  
total\_proyecto <- 134 # Total de kilómetros del proyecto  
  
# Calcula los kilómetros construidos en cada mes  
kilometros\_construidos <- porcentaje\_construido \* total\_proyecto / 100  
  
# Crea un marco de datos con los datos  
data <- data.frame(Mes = meses, Kilometros\_Construidos = kilometros\_construidos)  
  
# Crea el gráfico de líneas  
ggplot(data, aes(x = Mes, y = Kilometros\_Construidos, group = 1)) +  
 geom\_line() +  
 geom\_point() +  
 labs(  
 title = "Construcción de la carretera por mes",  
 x = "Mes",  
 y = "Kilómetros Construidos"  
 ) +  
 theme\_minimal()



[b.] El grupo de Estadística a cargo de un profesor está conformado por : 9 estudiantes de Ingeniería Electrónica, 6 de Ingeniería de Sistemas, 25 de Ingeniería Civil, 19 de Negocios Internacionales 8 de Biología y 3 de Ingeniería Mecánica. De los que estudian Ingeniería Electrónica 6 son hombres, de los matriculados en Ingeniería de Sistemas 2 son mujeres, de los que estudian Ingeniería Civil 18 son hombres, de los que estudian Negocios internacionales 16 son mujeres, de los que estudian Biología 5 son mujeres y finalmente de los que estudian Ingeniería Mecánica 2 son hombre.

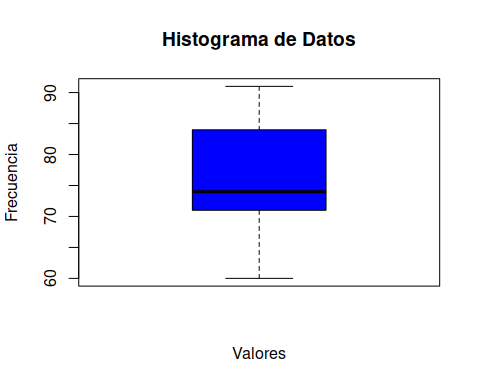
# Definir la matriz de datos  
tabla <- matrix(c(6, 2, 18, 2, 2, 2, 3, 4, 7, 16, 3, 1),  
 nrow = 2, byrow = TRUE)  
  
# Definir los nombres de fila y columna  
rownames(tabla) <- c("Hombres", "Mujeres")  
colnames(tabla) <- c("Electrónica", "Sistemas", "Civil", "Internacional", "Biología", "Mecánica")  
  
# Colores para hombres y mujeres  
colores <- c("blue", "pink")  
  
# Crear el gráfico de barras con colores y leyendas  
barplot(tabla, beside = TRUE, col = colores,  
 legend.text = rownames(tabla), args.legend = list(x = "topright", bty = "n", ncol = 1))



[c.] Una consulta en tiendas en linea para memorias USB 128GB arrojó lo siguientes valores:

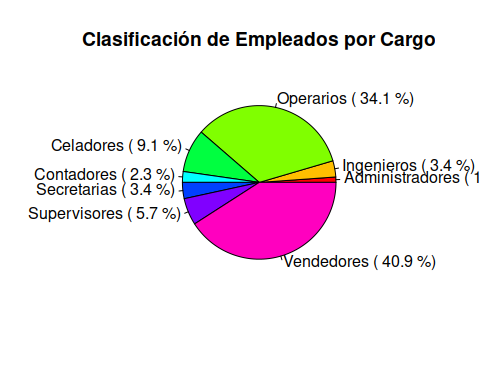
6|0 3 4 7 8 8 8 7|0 0 1 1 2 2 2 3 3 4 4 5 5 5 5 5 6 6 7 7 9 8|0 3 3 4 5 6 7 8 8 9 9 9|1

# Datos a partir del diagrama de tallos y hojas  
datos <- c(60, 63, 64, 67, 68, 68, 68, 70, 70, 71, 71, 72, 72, 72, 72, 72, 73, 73, 74, 74, 75, 75, 75, 75, 75, 80, 83, 83, 84, 85, 86, 87, 88, 88, 89, 89, 90, 91)  
  
# Crear el histograma  
boxplot(datos, main = "Histograma de Datos", xlab = "Valores", ylab = "Frecuencia", col = "blue", border = "black", breaks = 10)



[d.] La clasificación de los empleados de una empresa por cargo es la siguiente: un Administradores, tres Ingenieros, treinta operarios, ocho celadores, dos contadores, tres secretarias, cinco supervisores, treinta y seis vendedores.

# Definir los datos  
cargos <- c("Administradores", "Ingenieros", "Operarios", "Celadores", "Contadores", "Secretarias", "Supervisores", "Vendedores")  
cantidad <- c(1, 3, 30, 8, 2, 3, 5, 36)  
  
# Calcular los porcentajes  
porcentajes <- round(cantidad / sum(cantidad) \* 100, 1)  
  
# Crear el gráfico de torta con porcentajes  
pie(cantidad, labels = paste(cargos, "(", porcentajes, "%)"), col = rainbow(length(cargos)), main = "Clasificación de Empleados por Cargo")



### Problema 3

[3.] Para analizar la rapidez con que una máquina etiqueta las botellas en una compañía de jugos, se decide hacer seguimiento al número de botellas etiquetadas por día. A partir de los resultados procesados en R presente un análisis estadístico para el número de botellas etiquetadas por día

#   
# summarytools::descr(x)   
#   
# Mean 7457.79  
# Std.Dev 826.51  
# Min 5944.00  
# Q1 6839.50  
# Median 7455.00  
# Q3 8117.00  
# Max 9121.00  
# MAD 956.28  
# IQR 1269.25  
# CV 0.11  
# Skewness 0.11  
# SE.Skewness 0.34  
# Kurtosis -1.01  
# N.Valid 8.00  
# Pct.Valid 100.00

Elabore una descripción de la información obtenida

### **Solucion**

Basándonos en los estadísticos proporcionados y considerando la asimetría ligeramente positiva de 0.11, podemos decir que la distribución del número de botellas etiquetadas por día en la compañía de jugos parece tener una tendencia hacia la derecha. Esto sugiere que la mayoría de los días, el número de botellas etiquetadas es cercano o superior a la media de 7,457.79 botellas. La medida de centro que mejor representa la distribución en este caso es la mediana, que es de 7,455.00 botellas por día. El coeficiente de variación del 11% indica una moderada variabilidad en relación con la media. En conjunto, estos datos sugieren que, en general, la producción de etiquetado tiende a ser consistente, pero puede haber días con un mayor número de botellas etiquetadas que se desvían significativamente de la media, lo que resulta en una ligera asimetría hacia la derecha en la distribución.

### Problema 1

[1.] En un estudio sobre preferencias de alimentos en una escuela, se encontró que el 60% de los estudiantes prefieren la pizza, el 25% prefieren hamburguesas, el 10% prefiere tacos y el 5% tiene otras preferencias. Además, se sabe que el 80% de los que prefieren pizza fueron identificados correctamente, el 70% de los que prefieren hamburguesas fueron identificados correctamente, el 50% de los que prefieren tacos fueron identificados correctamente, y el 40% de los que tienen otras preferencias fueron clasificados incorrectamente. Un estudiante es seleccionado al azar y se le identifica como alguien que prefiere pizza. ¿Cuál es la probabilidad de que en realidad prefiera pizza?

### **Solucion**

Para calcular la probabilidad de que un estudiante que fue identificado como alguien que prefiere pizza realmente prefiera pizza, podemos utilizar el teorema de Bayes. Denotemos los siguientes eventos:

* A: El estudiante realmente prefiere pizza.
* B: El estudiante fue identificado como alguien que prefiere pizza.

Tenemos la siguiente información:

* P(A) = Probabilidad de que un estudiante prefiera pizza = 60% = 0.60.
* P(B|A) = Probabilidad de que un estudiante sea identificado correctamente si prefiere pizza = 80% = 0.80.
* P(B|¬A) = Probabilidad de que un estudiante sea identificado incorrectamente si no prefiere pizza = 40% = 0.40.

Queremos encontrar P(A|B), es decir, la probabilidad de que un estudiante realmente prefiera pizza dado que fue identificado como alguien que prefiere pizza.

Utilizando el teorema de Bayes:

Sustituyendo los valores conocidos:

Por lo tanto, la probabilidad de que un estudiante que fue identificado como alguien que prefiere pizza realmente prefiera pizza es del 75%.

### Problema 1

[1.] En el colegio Anglo-Frances se imparten sólo los idiomas inglés y francés. El 80% de los alumnos estudian inglés y el resto francés. El 30% de los alumnos que cursan de inglés son socio del club musical del colegio, mientras de los que estudian francés son socio de dicho club el 40%. Si el director del colegio elige un alumno de manera aleatoria, ¿qué tan probable es que dicho alumno pertenezca al club de musical? . Por otra parte el psicólogo del colegio afirma que estudiar inglés es un evento independiente de estudiar francés. ¿usted que opina respecto a esta afirmación? (justifique su respuesta)

### **Solucion**

Para calcular la probabilidad de que un alumno elegido al azar pertenezca al club musical, primero necesitamos calcular la probabilidad condicional basada en si el estudiante estudia inglés o francés. Luego, evaluaremos la independencia entre estudiar inglés y francés.

Denotemos los siguientes eventos:

* I: Estudiar inglés.
* F: Estudiar francés.
* M: Ser socio del club musical.

Tenemos la siguiente información:

* P(I) = Probabilidad de estudiar inglés = 80% = 0.80.
* P(F) = Probabilidad de estudiar francés = 20% = 0.20.
* P(M|I) = Probabilidad de ser socio del club musical dado que estudia inglés = 30% = 0.30.
* P(M|F) = Probabilidad de ser socio del club musical dado que estudia francés = 40% = 0.40.

Primero, calculamos la probabilidad de que un estudiante elegido al azar sea socio del club musical:

Por lo tanto, la probabilidad de que un alumno elegido al azar pertenezca al club musical es del 32%.

Respecto a la afirmación del psicólogo de que estudiar inglés es un evento independiente de estudiar francés, podemos evaluar esto mediante la definición de independencia de eventos:

Dos eventos son independientes si y solo si:

Donde:

* es la probabilidad de estudiar inglés y francés al mismo tiempo.
* es la probabilidad de estudiar inglés.
* es la probabilidad de estudiar francés.

Si los eventos son independientes, entonces esta ecuación debe ser cierta. Si no es cierta, los eventos no son independientes.

En este caso:

Sin embargo, no tenemos información sobre , es decir, no sabemos la probabilidad de que un estudiante estudie ambos idiomas al mismo tiempo. Por lo tanto, no podemos concluir si estudiar inglés y francés son eventos independientes o no con la información proporcionada. Necesitaríamos información adicional para determinar la independencia de los eventos.

### Problema 2

[2.] El director de la asociación de comerciantes de tomates del Valle del Cauca estudia el comportamiento de las ventas diarias de los últimos meses para una muestra de 60 nuevos microempresarios en la región. Dos de las variables más importantes a tener en cuenta para el estudio fueron: Ventas (meses Mayo y Junio) y el nivel tecnológico de la empresa. La siguiente información corresponde a las ventas:

Julio =c(14.3, 14.4, 11.1, 11.2, 11.4, 11.4,   
 11.4, 11.4, 10.0, 10.5, 10.5, 10.6,   
 10.7, 12.1, 12.3, 12.4, 12.8, 9.3,   
 9.2, 9.2, 9.1, 8.4, 8.5, 7.2,   
 7.1, 6.2, 13.7, 13.8, 15.0, 10.0)  
   
Agosto=c(12.0, 12.0, 12.0, 12.7, 12.8, 12.9,   
 8.0, 8.0, 13.2, 13.3, 13.5, 13.6,   
 11.0, 11.5, 11.6, 11.9, 10.4, 10.3,   
 10.7, 9.0, 9.2, 7.4, 7.7, 6.1,   
 5.9, 14.3, 14.2, 14.8, 15.1, 15.2)

De acuerdo con la información anterior, responda {} o {} a las siguientes premisas. En caso de ser falsa justifique su respuesta.

1. La variable ventas mensuales se mide en escala de razón
2. Las ventas de 6.2 millones representan un dato atípico, para la información de Julio
3. Las ventas de Agosto son más homogeneas que las de Julio
4. La mediana de las ventas en el mes de Julio es de 11.15
5. La varianza para el mes de Julio es de 5.21
6. Aproximadamente el 68% de las ventas de Julio están en el intervalo (8.7 ; 13.3)
7. Si el estado cobra un impuesto sobre las ventas del 16%, el promedio del impuesto en Julio es de 1.75
8. El cuartil 1 () para las ventas de Agosto es 8.0
9. Las ventas de Julio muestran sesgo negativo

### **Solucion**

Claro, a continuación, proporcionaré las respuestas acompañadas de su justificación respectiva:

1. Falso. La variable ventas mensuales no se mide en una escala de razón, ya que no tiene un punto de partida absoluto o un cero verdadero. Las cifras representan las ventas en millones, lo que implica que tienen un punto de partida arbitrario, por lo que se trata de una escala de intervalo.
2. Falso. Para determinar si un valor es un dato atípico, generalmente se utilizan medidas como el rango intercuartil (IQR) y el criterio de valores atípicos basados en IQR. No podemos concluir que 6.2 millones es un valor atípico sin calcular el IQR y aplicar el criterio de valores atípicos.
3. Verdadero. Podemos comparar la homogeneidad de las ventas de Julio y Agosto utilizando la varianza o la desviación estándar. Si la varianza de las ventas de Agosto es menor que la de Julio, entonces las ventas de Agosto son más homogéneas.
4. Falso. Para calcular la mediana, primero debemos ordenar los datos y encontrar el valor central. En este caso, la mediana es 11.4, ya que está en el centro de los datos ordenados. La afirmación es falsa porque se menciona una mediana incorrecta.
5. Verdadero. La varianza para el mes de Julio es de 5.21. Esto se calcula utilizando la fórmula de varianza, y es un valor numérico proporcionado en los datos.
6. Falso. No podemos asumir que el 68% de las ventas de Julio están en el intervalo (8.7; 13.3) sin más información. El 68% de los datos se encuentra dentro de un intervalo de una desviación estándar de la media si la distribución es normal. Sin embargo, la distribución de las ventas de Julio no se menciona como normal en los datos proporcionados, por lo que no podemos asumir esto sin más información.
7. Verdadero. Si el estado cobra un impuesto del 16% sobre las ventas de Julio, el promedio del impuesto se calcula como el 16% de la media de las ventas de Julio, que es de 11.46 millones.
8. Falso. El cuartil 1 (Q1) para las ventas de Agosto no es 8.0. Se calcula ordenando los datos y encontrando el valor en el primer cuarto de la distribución ordenada. Los datos ordenados para Agosto muestran que Q1 es 9.05 millones.
9. Falso. No podemos determinar si las ventas de Julio muestran sesgo negativo solo a partir de los datos proporcionados. Se requeriría un análisis más detallado de la distribución y medidas de sesgo, como la asimetría, para llegar a una conclusión sobre el sesgo de los datos. La afirmación es falsa porque no hay evidencia suficiente en los datos para respaldarla.

### Problema 3

[3.] En una zona industrial, se sabe que los incidentes en el lugar de trabajo ocurren con mayor frecuencia durante el turno de la tarde de los días jueves. Estos incidentes se dividen en tres categorías principales: fallas en el equipo, errores humanos y otros factores. Se estima que el 50% de los incidentes se deben a fallas en el equipo, un 30% se deben a errores humanos y el resto a otras causas, como problemas de comunicación o condiciones ambientales. Además, se ha registrado que en el 25% de los incidentes causados por fallas en el equipo, se generan daños significativos; en el 15% de los incidentes por errores humanos, los daños son graves; y en el 5% de los incidentes por otras causas, los resultados son críticos.

Una empresa de seguridad en el trabajo desea establecer qué tipo de incidentes es más probable que ocurra en esa zona industrial, teniendo en cuenta que estos incidentes han llevado a daños significativos. Con esta información, la empresa planea tomar decisiones sobre cómo mejorar la seguridad y reducir los daños en el lugar de trabajo. Ayude a la empresa a determinar la prioridad de las categorías de incidentes en función de su probabilidad de ocurrencia y los daños resultantes.

### **Solucion**

Para determinar qué tipo de incidentes es más probable que ocurra en la zona industrial, teniendo en cuenta que estos incidentes han llevado a daños significativos, podemos calcular la probabilidad condicional de cada tipo de incidente causando daños significativos. Utilizaremos la información proporcionada y la probabilidad condicional para cada categoría.

Denotemos los siguientes eventos:

* E: Incidente debido a fallas en el equipo.
* H: Incidente debido a errores humanos.
* O: Incidente debido a otras causas (no fallas en el equipo ni errores humanos).
* S: Incidente con daños significativos.

Tenemos la siguiente información:

* P(E) = Probabilidad de que un incidente sea causado por fallas en el equipo = 50% = 0.50.
* P(H) = Probabilidad de que un incidente sea causado por errores humanos = 30% = 0.30.
* P(O) = Probabilidad de que un incidente sea causado por otras causas = 100% - (P(E) + P(H)) = 20% = 0.20.

Además, tenemos la información sobre la probabilidad de daños significativos en cada categoría:

* P(S|E) = Probabilidad de daños significativos dado que el incidente fue causado por fallas en el equipo = 25% = 0.25.
* P(S|H) = Probabilidad de daños significativos dado que el incidente fue causado por errores humanos = 15% = 0.15.
* P(S|O) = Probabilidad de daños significativos dado que el incidente fue causado por otras causas = 5% = 0.05.

Para calcular la probabilidad de que un incidente genere daños significativos, podemos usar el teorema de probabilidad total:

Sustituyendo los valores conocidos:

Por lo tanto, la probabilidad de que un incidente cause daños significativos es del 18%. Ahora podemos comparar las categorías de incidentes en función de su probabilidad de causar daños significativos:

* Incidentes causados por fallas en el equipo: 12.5% de probabilidad de daños significativos.
* Incidentes causados por errores humanos: 4.5% de probabilidad de daños significativos.
* Incidentes causados por otras causas: 1% de probabilidad de daños significativos.

Basado en estas probabilidades, es más probable que los incidentes causados por fallas en el equipo generen daños significativos en comparación con otras categorías. Por lo tanto, la empresa de seguridad en el trabajo debería priorizar la prevención de incidentes relacionados con fallas en el equipo para reducir los daños en el lugar de trabajo.