EXAMEN PROBABILIDAD Y SIMULACIÓN
FECHA: 19-12-2024
MARÍA PALLARES DIEZ

EJERCICIO 1

a) Probabilidad de que el tiempo de supervivencia supere 12 meses, 3 y 5 años en mujeres premenopáusicas y también en mujeres postmenopáusicas

Para mujeres premenopáusicas el tiempo de supervivencia calculado es el siguiente:

- Supere 12 meses: 0,76%, esto significa que aproximadamente el 75.6% de los pacientes premenopáusicos sobreviven más de 1 año.
- Supere 3 años: 0,22%, solo el 22% de los pacientes premenopáusicos sobreviven más de 3 años.
- Supere 5 años: 0,05%, apenas un 5% de los pacientes premenopáusicos consiguen sobrevivir más de 5 años a la enfermedad.

Para mujeres postmenopáusicas el tiempo de supervivencia calculado es el siguiente:

- Supere 12 meses: 0,85%, la gran mayoría de las pacientes (85%) superan el primer año de la enfermedad habiendo pasado también la menopausia.
- Supere 3 años: 0,27%, apenas más del 25% de las mujeres que hayan pasado la menopausia superan los 3 años tras el diagnostico.
- Supere 5 años: 0,044%, casi el 5% de las mujeres superan los 5 años desde que son diagnosticadas y han pasado la menopausia.

b) Utilizando el modelo probabilístico, que se ajusta a esta variable, calcula el tiempo de supervivencia medio (valor esperado) para cada grupo de mujeres.

El tiempo de supervivencia medio (valor esperado) para cada grupo de mujeres es el siguiente:

- Mujeres premenopáusicas 1,98 años (alrededor de 2 años o 1 año y 11 meses)
- Mujeres postmenopáusicas 2,32 años (alrededor de 2 años y 4 meses)

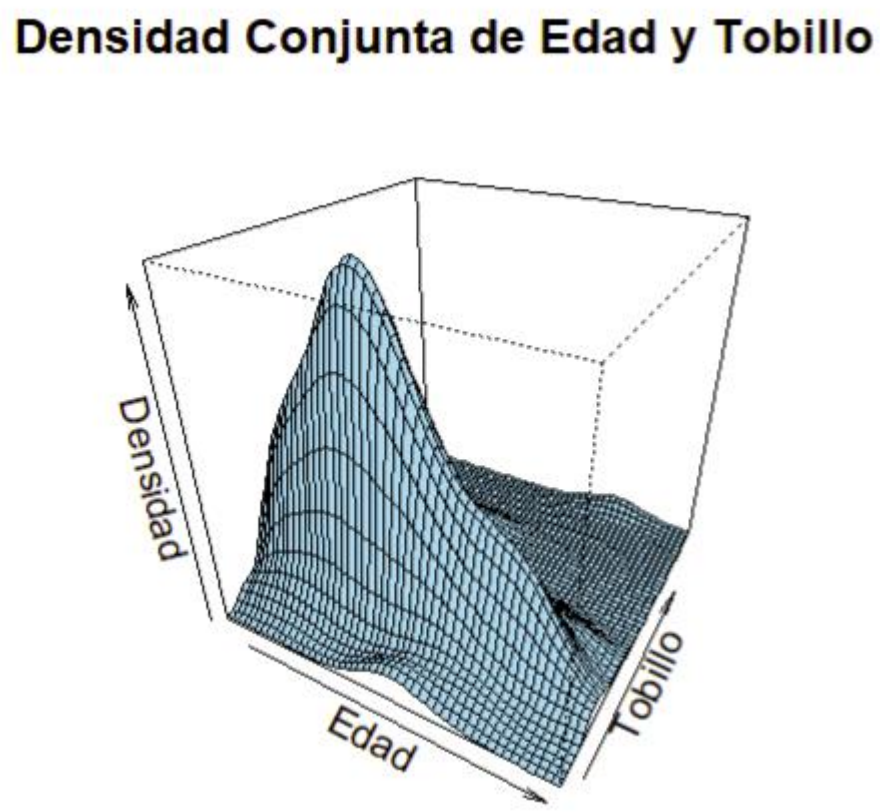
Con estos datos podemos concluir que las mujeres que han pasado la menopausia tienden a tener una esperanza de vida mayor que aquellas que aún no han pasado por esta tras ser diagnosticadas con cáncer de mama con ganglios positivos.

```R
# Cargar datos
datos <- read.table("breastcancer.txt", header = TRUE)
# Convertir tiempo a años para las comparaciones
datos$time_years <- datos$time / 365
# a) Probabilidad de supervivencia en diferentes intervalos
# Filtrar por estado menopáusico
pre <- subset(datos, menostat == "Pre")
post <- subset(datos, menostat == "Post")
# Calcular probabilidades
prob_pre <- c(
  mean(pre$time_years > 1),
  mean(pre$time_years > 3),
  mean(pre$time_years > 5)
)
prob_post <- c(
  mean(post$time_years > 1),
  mean(post$time_years > 3),
  mean(post$time_years > 5)
)
# b) Tiempo de supervivencia medio
mean_pre <- mean(pre$time_years)
mean_post <- mean(post$time_years)
```

EJERCICIO 2

a) **Distribución conjunta de las variables X e Y, f X, Y (x, y)**

```
R
#a) Distribución conjunta de X (Edad) e Y (Tobillo)
library(MASS) kde_joint <- kde2d(datos$Edad, datos$Tobillo, n = 50)
#Visualización de la densidad conjunta
persp(kde_joint, theta = 30, phi = 30, col = "lightblue")
```



La gráfica que compartes representa la **densidad conjunta de las variables Edad (X) y Tobillo (*Y*)**. Este gráfico tridimensional muestra cómo se distribuyen conjuntamente las observaciones de estas dos variables.

Interpretación de la Gráfica

1. **Ejes:**
    a. El eje *XX*X representa la variable **Edad**.
    b. El eje *YY*Y representa la variable **Tobillo**.
    c. El eje *ZZ*Z muestra la **densidad conjunta estimada**, que indica qué tan probables son las combinaciones específicas de Edad y Tobillo.
2. **Regiones de Alta Densidad:**
    a. Las áreas donde la superficie es más alta corresponden a combinaciones de valores de Edad y Tobillo que ocurren con mayor frecuencia en los datos.

b. Estas regiones representan los puntos donde las dos variables están más correlacionadas.

3. **Regiones de Baja Densidad:**
   a. Las áreas más planas o cercanas al eje horizontal indican combinaciones de Edad y Tobillo que son menos comunes.
4. **Forma General:**
   a. La forma de la superficie indica la relación entre las dos variables:
      i. Si la densidad se concentra en una región específica, podría haber una correlación entre Edad y Tobillo.
      ii. Si la densidad está dispersa, las dos variables podrían ser más independientes entre sí.

b) E(Y |X)

```
R
edad <- datos$Edad
tobillo <- datos$Tobillo
# Calcular E(Y | X)
esperanza_condicional <- tapply(tobillo, edad, mean, na.rm = TRUE)
# Mostrar resultados
print(esperanza_condicional)
```

Con los datos y el código en R obtuvimos los resultados de E(Y|X), Esto representa la esperanza condicional de Y (Tobillo) dado X (Edad). Cada valor calculado muestra el promedio de Y para cada grupo de edades X.
Resultados Destacados:

- Para X=24, el promedio de Y es 24.8.
- Para X=45, el promedio de Y es 26.6.
- Para X=74, el promedio de Y es 24.6.
- Para X=81, el promedio de Y es 21.5.

Estos valores sugieren cómo varía el tobillo (Y) en función de la edad (X). Por ejemplo, en este caso, los valores promedio de Y no parecen seguir un patrón estrictamente creciente o decreciente, lo que podría indicar una relación no lineal entre las variables.

c) P(T < 30|Z < 175)

```
R
#Subconjunto donde Z < 175
peso <- datos$Peso biceps <- datos$Biceps subset_peso <- subset(datos, peso < 175)
#Calcular la probabilidad condicional
prob_t_condicional <- mean(subset_peso$Biceps < 30, na.rm = TRUE)
#Mostrar el resultado
print(prob_t_condicional)
```

Con los datos y el código en R obtuvimos los resultados de P(T < 30|Z < 175). Resultado: La probabilidad condicional calculada es 0.4483 (aproximadamente 44.8%).

Esto significa que, entre las personas cuyo peso (Z) es menor que 175, el 44.8%tiene un bíceps (T) menor a 30. Este valor nos proporciona una probabilidad condicional útil para entender la relación entre Z y T.

d) E(T|Z)

```
R
#Calcular E(T | Z)
esperanza_t_condicional <- tapply(biceps, peso, mean, na.rm = TRUE)
#Mostrar resultados
print(esperanza_t_condicional)
```

Con los datos y el código en R obtuvimos los resultados de E(T|Z) Esto representa la esperanza condicional de T (Bíceps) dadoZ (Peso). Cada valor indica el promedio de T para cada valor de Z.
Resultados Destacados:

- Para Z=125, el promedio de T es 29.3.
- Para Z=140.5, el promedio de T es 30.9.
- Para Z=174.5, el promedio de T es 34.6.
- Para Z=234.75, el promedio de T es 39.1.

Interpretación: Los resultados sugieren que el promedio de T (Bíceps) tiende a aumentar a medida que Z (Peso) aumenta. Esto refleja una relación positiva entre T y Z, lo cual tiene sentido fisiológico, ya que las personas con mayor peso podrían tener bíceps más grandes.

EJERCICIO 3

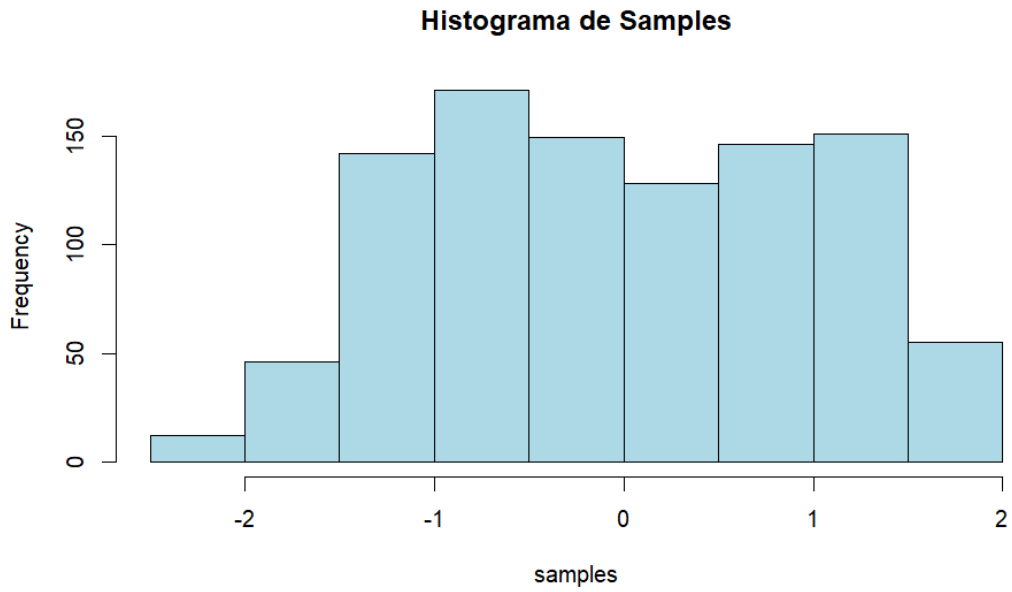
a) Si representamos con un histograma el vector samples, ¿qué conseguimos? Justifica tu respuesta.

Al representar un histograma del vector samples, podemos observar la distribución empírica de las muestras generadas por el algoritmo Metropolis-Hastings. Esto nos permite evaluar si las muestras convergen hacia la distribución objetivo definida por la función target (x).
¿Qué conseguimos?

- Validación de convergencia: Verificamos si las muestras generadas siguen correctamente la distribución objetivo.
- Forma de la distribución: La forma del histograma debe reflejar la distribución objetivo. En este caso, la función target(x) define una distribución no estándar.
- Exploración del espacio: Evaluamos si el algoritmo explora adecuadamente la región de mayor densidad de la distribución objetivo.

```
R
samples <- metropolis_hastings(1000, 1, 0)
```

```
hist(samples, main = "Histograma de Samples", col = "lightblue")
```



b) Detalla qué aporta la línea 12 ¿Podríamos proponer una línea 12 con otra distribución de probabilidad? Justifica tu respuesta.

La línea 12 genera una propuesta para el siguiente valor de la cadena de Markov utilizando una distribución normal centrada en el valor actual *x_current* con desviación estándar *proposal_sd*.

- Exploración del espacio: Define los saltos que el algoritmo puede realizar en cada iteración.
- Control del tamaño de los pasos: El parámetro *proposal_sd* determina cuánto puede alejarse la propuesta del valor actual:

  -Un valor pequeño de *proposal_sd* genera pasos cortos (mayor aceptación, pero menos exploración).

  -Un valor grande de *proposal_sd* genera pasos largos (mejor exploración, pero menor aceptación).

¿Podemos usar otra distribución? Sí, podemos usar otras distribuciones para generar propuestas, dependiendo del problema. Como, por ejemplo:

- Distribución uniforme: Permite explorar un rango fijo alrededor de x_current:

  R
  ```
  x_proposed <- runif(1, min = x_current - delta, max = x_current + delta)
  ```

- Distribución exponencial: Útil para distribuciones con colas largas:

  R
  ```
  x_proposed <- rexp(1, rate = lambda) + x_current
  ```

La elección de la distribución propuesta depende de la forma de la distribución objetivo (target). Mientras sea simétrica y permita explorar el espacio de parámetros adecuadamente, cualquier distribución puede ser utilizada.

c) Detalla qué aporta el valor de Alpha, en la línea 15.

El valor de Alpha controla la probabilidad de aceptar la propuesta *x_proposed*, basada en la densidad de la distribución objetivo (target).

- Aceptación incondicional de mejores propuestas:

Si target(x_proposed) > target(x_current), entonces Alpha = 1 y se acepta siempre.

- Aceptación condicional de peores propuestas:

Si target(x_proposed) < target(x_current), entonces Alpha es menor que 1 y se acepta con una probabilidad igual a Alpha.

Este enfoque asegura que el algoritmo favorezca regiones de alta densidad de la distribución objetivo, pero también permite aceptar propuestas de menor densidad, lo que evita quedar atrapado en máximos locales.

d) Mediante un ejemplo, que tú misma/o propongas, analiza cómo afecta el valor de sd, de la línea 12, en el valor de Alpha, en la tasa de aceptación.

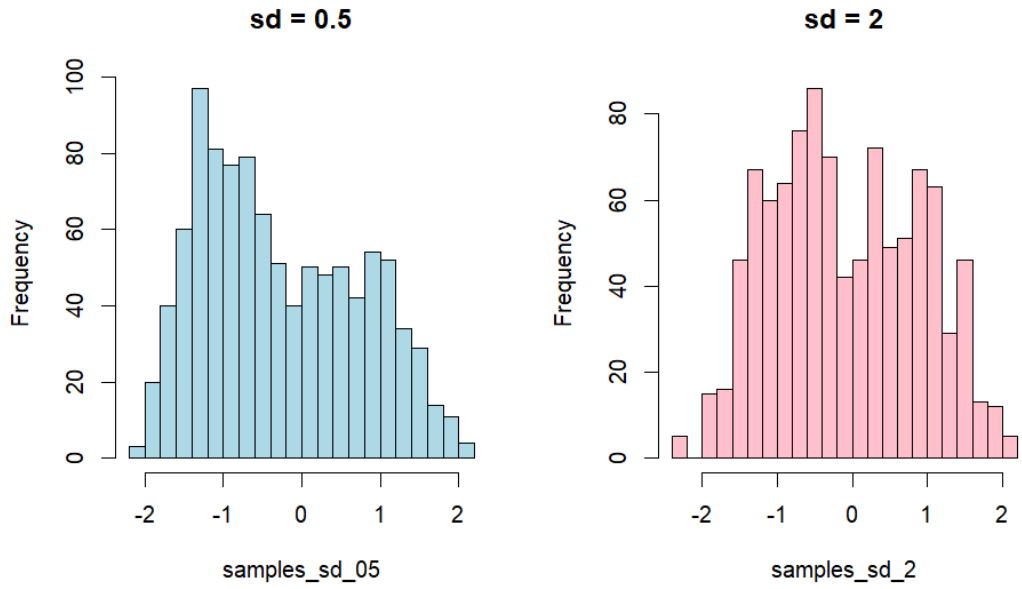
Probaremos con diferentes valores de *proposal_sd* para analizar cómo afecta al valor de Alpha y a la tasa de aceptación.

Resultados esperados:

- Propuestas pequeñas (sd = 0.5):
  - Alta tasa de aceptación porque las propuestas están cerca del valor actual.
  - Sin embargo, la exploración del espacio es más lenta, lo que puede llevar a una convergencia más lenta.
- Propuestas grandes (sd = 2):
  - Menor tasa de aceptación porque las propuestas lejanas tienen menor densidad objetivo.
  - Mayor exploración del espacio, lo que puede ser útil para distribuciones multimodales.

Resultados obtenidos:

- Propuestas pequeñas (sd = 0.5): 0,87 tasa de aceptación.
- Propuestas grandes (sd = 2): 0,532 tasa de aceptación.



Como conclusión vemos que los valores esperados y obtenidos concuerdan, evidenciando que el valor de *proposal_sd* afecta el equilibrio entre tasa de aceptación y exploración del espacio de soluciones. Un valor adecuado depende de la forma de la distribución objetivo y del objetivo del análisis (convergencia rápida vs. exploración exhaustiva).

```
R

# Caso 1: sd = 0.5 (propuestas pequeñas)

samples_sd_05 <- metropolis_hastings(1000, proposal_sd = 0.5, x_init = 0)

acceptance_rate_sd_05 <- length(unique(samples_sd_05)) / length(samples_sd_05)

# Caso 2: sd = 2 (propuestas grandes)

samples_sd_2 <- metropolis_hastings(1000, proposal_sd = 2, x_init = 0)

acceptance_rate_sd_2 <- length(unique(samples_sd_2)) / length(samples_sd_2)

# Comparar histogramas

par(mfrow = c(1, 2))

hist(samples_sd_05, main = "sd = 0.5", col = "lightblue", breaks = 30)

hist(samples_sd_2, main = "sd = 2", col = "pink", breaks = 30)

# Tasa de aceptación

acceptance_rate_sd_05

acceptance_rate_sd_2
```

EJERCICIO 4

Datos:
P(E)= 0,01 <- Enfermo
P(Positivo|E) = 0,95 <- Sensibilidad: dar positivo estando enfermo
P(Negativo|NE) = 0,9 <- Especificidad: dar negativo no estando enfermo
P(NE) = 1- P(E) = 0,99 <-No enfermo

Resolución:
Queremos saber cuál es la probabilidad de que si eligen una persona al azar y le hacen una prueba y sale positivo que esta persona realmente tenga la enfermedad, es decir P(E|Positivo). Para ello podemos utilizar el Teorema de Bayes:

$$P(E|\text{Positivo}) = \frac{P(\text{Positivo}|E) \cdot P(E)}{P(\text{Positivo})}$$

Nos falta el dato P(Positivo), por ello vamos a calcularlo:

P(Positivo)= P(Positivo|E) x P(E) + (1-P(Negativo|NE) x P(NE)
P(Positivo)= 0,95 x 0,01 + (1-0,9 x 0,99)
P(Positivo)= 0,1085

Con este dato ya podemos sustituir el valor en el Teorema de Bayes y resolver:

$$P(E|\text{Positivo}) = \frac{0.95 \cdot 0.01}{0.1085}$$

$$P(E|\text{Positivo}) = \frac{0.0095}{0.1085} \approx 0.0876$$

Solución:
La probabilidad de que la persona tenga realmente la enfermedad habiéndole hecho una prueba al azar y que salga positivo es de un 8,76%.