Metodología de la Investigación: Tarea 2

Badillo Hernandez Christian Francisco Lopez Hernandez Lizbeth
Nuñez Rodas Luis Antonio Pichardo Jiménez Sealtiel
Santana Jaimes Luz María Concepción

Tabla de contenidos

Ejercicio 1 **2**
- Usando Proporciones . . . 2
  - Proporciones por programa . . . 2
  - Proporciones globales . . . 3
- Usando tasas . . . 3

Ejercicio 2 **4**

Ejercicio 3 **9**

Ejercicio 4 **10**
- *OR* para asma y alergia . . . 10
- Variable de interacción . . . 11
- Variable de confusión . . . 13
- Conclusión . . . 15

Ejercicio 5 **16**

Ejercicio 6 **19**

Ejercicio 7 **26**

Ejercicio 1

Para este problema se utiliza la base de datos *sexbias.csv* que contiene la información de 4526 sujetos que deseaban ingresar a los 6 (A,B,C,D,E,F) programas de posgrado más prestigiosos de la Universidad de California.

Tenemos en un principio la **hipótesis nula (H_0):** Las mujeres y hombres reciben el mismo trato, es decir, no hay preferencia hacia los hombres(misma tasa de aceptación). Nos lleva a la **Hipótesis Alternativa(H_α):** Los hombres están recibiendo un trato preferencial sobre las mujeres en la admisión a ciertos programas de posgrado.

Vamos a realizar el ejercicio de dos maneras, primero por medio de proporciones y finalmente por tasas.

Usando Proporciones

Empezamos colocando una tabla con el conteo de cada posgrado, señalando cuántas mujeres, hombres fueron aceptados o no en cada uno de los programas; se visualiza el total de candidatos a cada programa.

Tabla 1: Conteo General

department	H_Aceptados	H_No_Aceptados	M_Aceptadas	M_No_Aceptadas	Postulantes
A	512	313	89	19	933
B	353	207	17	8	585
C	120	205	202	391	918
D	138	279	131	244	792
E	53	138	94	299	584
F	22	351	24	317	714
Total	1198	1493	557	1278	4526

Proporciones por programa

Sea i el subíndice que indica el programa, entonces las propociones que representan las tasas de aceptación de mujeres y hombres están representadas por:

- $\hat{p}_{m_i} = \frac{x_{m_i}}{n_{m_i}}$
- $\hat{p}_{h_i} = \frac{x_{h_i}}{n_{h_i}}$

Donde x son los sujetos aceptados y n el total de postulantes.

Tabla 2: Proporción aceptación por programa

posgrado	proporción_m	proporción_h
A	0.8240741	0.6206061
B	0.6800000	0.6303571
C	0.3406408	0.3692308
D	0.3493333	0.3309353
E	0.2391858	0.2774869

posgrado	proporción_m	proporción_h
F	0.0703812	0.0589812

De esta tabla resaltamos como en el programa A, las mujeres tienen mayor proporción de aceptación.

Proporciones globales

Podemos combinar los datos por programas para obtener las proporciones globales:

- $\hat{p}_m = \frac{\sum_i x_{m_i}}{\sum_i n_{m_i}}$
- $\hat{p}_h = \frac{\sum_i x_{h_i}}{\sum_i n_{h_i}}$

Estas proporciones serían

```
[1] "Proporción global de mujeres aceptadas: 0.3035"
```

```
[1] "Proporción global de hombres aceptados: 0.4452"
```

Podríamos hacer una prueba de diferencia de proporciones de dos poblaciones

$$z = \frac{(\hat{p_m} - \hat{p_h}) - 0}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_m} + \frac{1}{n_h})}}$$

Donde la proporción combinada es $\hat{p} = (x_m + x_h)/(n_m + n_h)$

```
[1] "El valor de z= -9.606"
```

Para $\alpha = 0.05$ el valor crítico de la prueba de hipótesis es de $Z_{\alpha/2} = 1.96$.

Como:

$$|-9.606| > 1.96$$

Rechazamos la prueba de hipótesis (H_0), entonces se acepta la hipótesis nula (H_α).

Usando tasas

Creamos una tabla de tasas de incidencia:

Tabla 3: Tasas de incidencia

	Mujeres	Hombres	Totales
Aceptados	557	1198	1755
Candidatos	1835	2691	4526

Para comparar las tasas de aceptación entre los individuos Hombres y Mujeres, recurrimos a un análisis relativo, es decir, el cociente entre las dos tasas: tasa de hombres aceptados y tasa de mujeres aceptadas. Para ello se utiliza una función para calcular el RR y su intervalo de confianza; se llama `calcular_IC_RR` y se describe a detalle en el problema 7.

```
[1] "Este cociente usando la función fue de: 1.467"
[1] "Con un intervalo de confianza entre: 1.326, 1.622"
[1] "Estadístico Z: 7.468"
[1] "p-value: 8.14903700074865e-14"
```

Nuestra Hipótesis nula (H_0) es equivalente a pensar que el cociente entre estas tasas debería ser 1, lo cual no sucede; usando el nivel de significancia de 0.05 esta hipótesis resulta ser significativa con un p-value de orden 10^{-14} y un estadístico $z = 7.5$. Por tales motivos podemos rechazar la H_0 : misma tasa de aceptación a los programas de posgrado.

Ejercicio 2

La base de datos senic.csv corresponde a Study on the Efficacy of Nosocomial Infection Control (SENIC Project).Con esta base realice lo siguiente.

1. Ajuste un modelo de regresión lineal simple entre las variable INFRISK (Riesgo de infección. Como variable de respuesta) y LOS (Duración del periodo de estancia o tiempo de duración de la estancia. Como variable explicativa).
    - **INFRISK** : Probabilidad estimada promedio de adquirir una infección en el hospital (en porcentaje)
    - **LOS**: Duración promedio de la estadía de todos los pacientes en el hospital (en días)

```r
data_2 <- read.csv("senic.csv")

#Se crea una variable que mapee a todas las regiones

data_2 <- data_2 %>%
  mutate(N_REGION = case_when(
    REGION == 1 ~ "NE",
    REGION == 2 ~ "NC",
    REGION == 3 ~ "S",
    REGION == 4 ~ "W",

  ))

##Se aplica un modelo de regresión lineal simple que contiene a la variable INFRISK como re
#en función de la variable explicativa LOS :

modelo <- lm(INFRISK ~ LOS, data = data_2)
summary(modelo)
```

```
Call:
lm(formula = INFRISK ~ LOS, data = data_2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7781 -0.7083  0.1337  0.6790  2.5879

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.73223    0.56083   1.306    0.194
LOS          0.37510    0.05702   6.578 1.76e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.152 on 108 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.2861,    Adjusted R-squared:  0.2794
F-statistic: 43.27 on 1 and 108 DF,  p-value: 1.757e-09
```

Por lo que el modelo ajustado queda como:

$$y = \beta_0 + \beta_1 x = 0.732 + 0.375x$$

2. Interprete el parámetro estimado asociado a su única variable predictora.

El parámetro β_1 estimado es de 0.3751 y representa el cambio promedio en INFRISK por cada unidad adicional de LOS en términos del contexto de los datos, indica que por cada día adicional de estancia hospitalaria, el riesgo de infección aumenta en promedio 0.375 unidades; o en un 37.5% ,este parámetro, de acuerdo con el p-value < 0.05, es significativamente distinto de cero y se diría que sí hay un efecto positivo en la cantidad de días de estancia sobre el riesgo de infección.

3. ¿Tiene una interpretación lógica el intercepto del modelo ($\hat{\beta}_0$) en el contexto de los datos que lo generan? Si es así, interprete este parámetro estimado.

En este caso el valor estimado de $\beta_0 = 0.732$ y representaría el valor promedio de riesgo de infección cuando la duración de la estancia es 0. Es decir, el riesgo de infección cuando el paciente no estuvo hospitalizado lo cual solo podría significar el riesgo asociado al no estar en el hospital. Pero como los datos no toman en cuenta o no están diseñados para tener en cuenta a personas con estancia 0, la única interpretación del intercepto es como un ajuste matemático para generar la predicción dentro del rango posible de los datos.

4. Ajuste el mismo modelo controlando por región (REGION) (recuerde que esta variables es nominal con cuatro categorías). Interprete los parámetros estimados para las categorías de región. ¿Se puede considerar que esta variable es de confusión? Argumente su respuesta.

```r
#Se tranforma la variable región a factor de 4 niveles:
data_2$REGION <- as.factor(data_2$REGION)

#Se aplica el modelo de regresión tomando en cuenta las regiones:
modelo_x_region <- lm(INFRISK ~ LOS + N_REGION, data = data_2)

#Resumen del modelo.
summary(modelo_x_region)
```

```
Call:
lm(formula = INFRISK ~ LOS + N_REGION, data = data_2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7671 -0.7304  0.1280  0.6445  2.7546

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.37495    0.66185   0.567   0.5722
LOS          0.41392    0.06487   6.381 4.85e-09 ***
N_REGIONNE  -0.10421    0.31370  -0.332   0.7404
N_REGIONS   -0.25245    0.28254  -0.893   0.3736
N_REGIONW    0.64782    0.36684   1.766   0.0803 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.132 on 105 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.3295,    Adjusted R-squared:  0.304
F-statistic:  12.9 on 4 and 105 DF,  p-value: 1.404e-08
```

No sale NC porque es la que se usa de referencia para obtener las diferencias reportadas en las demás, es decir, comparadas con la region NC

LOS β_1: Su valor es de 0.41392, lo que indica que si se mantuviera constante la región (es decir, ajustando las diferencias que puedan existir entre regiones, eliminando dichas diferencias),cada día de estancia (LOS) incrementaría el riesgo de infección en 0.414 unidades. Y esto es significativo para un p-value < 0.05.

REGION NE: β_2: Su valor es de -0.10421, lo que indica que, en promedio, por cada unidad de LOS, el riesgo de infección es menor en -0.104 unidades respecto a lo que ocurre en la REGION1, del NC, pero su p-value es mayor a 0.05 por lo que el efecto no es significativo.

REGION S:β_3: Su valor es de -0.25245, lo que indica que, en promedio, por cada unidad de LOS, el riesgo de infección es menor en 0.25245 unidades respecto a lo que ocurre en la región de NC, pero su p-value es mayor a 0.05 por lo que el efecto tampoco es significativo.

REGION W: $\beta 4$: Su valor es de 0.64782, lo que indica que, en promedio, por cada unidad de LOS, el riesgo de infección es mayor en 0.64782 unidades respecto a lo que ocurre en la NC, pero su p-value es solo un poco mayor a 0.05 por lo que el efecto está cerca de ser significativo, pero no lo es.

¿Se puede considerar que la región es una variable es de confusión?

Para ello, se debe considerar al coeficiente β_1 estimado del modelo que no toma en cuenta las regiones y del que ponera a las regiones del β_1. Además de examinar si hay diferencias en los coeficientes de las regiones (respecto a la REGION1) y si estas son significativas

En el primer caso, se obtuvo un $\beta_1 = 0.37510$, en el segundo fue $\beta_1 = 0.41392$. Hay un cambio en la magnitud de los coeficientes, pero no pareciera ser muy grande el cambio. Además, el cambio en los coeficientes de las regiones no representa una diferencia estadísticamente significativa. Por lo que se concluye que la variable region no es una variable de confusión.

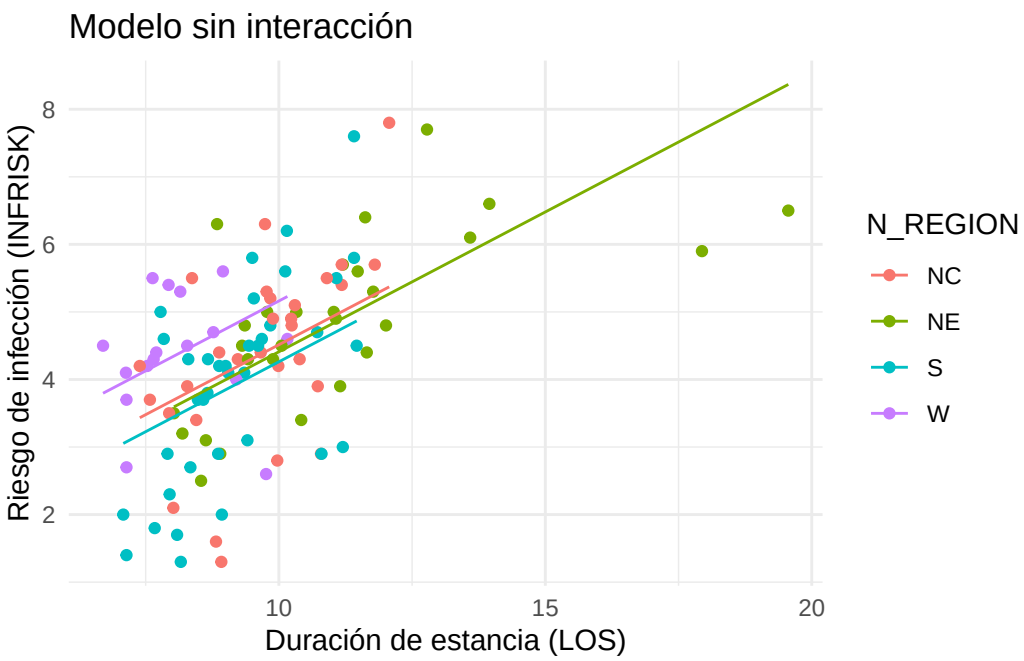
5. Adicione al modelo inicial la interacción (multiplicativa) entre las variable LOS y REGION. ¿Se puede afirmar a un nivel de significancia, $\alpha = 0.05$, que existe interacción entre estas variables? Y si así fuera, ¿qué significa en el contexto de esta investigación?

```r
##Se genera el modelo con interacción:

model_interaction <- lm(INFRISK ~ LOS * N_REGION, data = data_2)
summary(model_interaction)
```

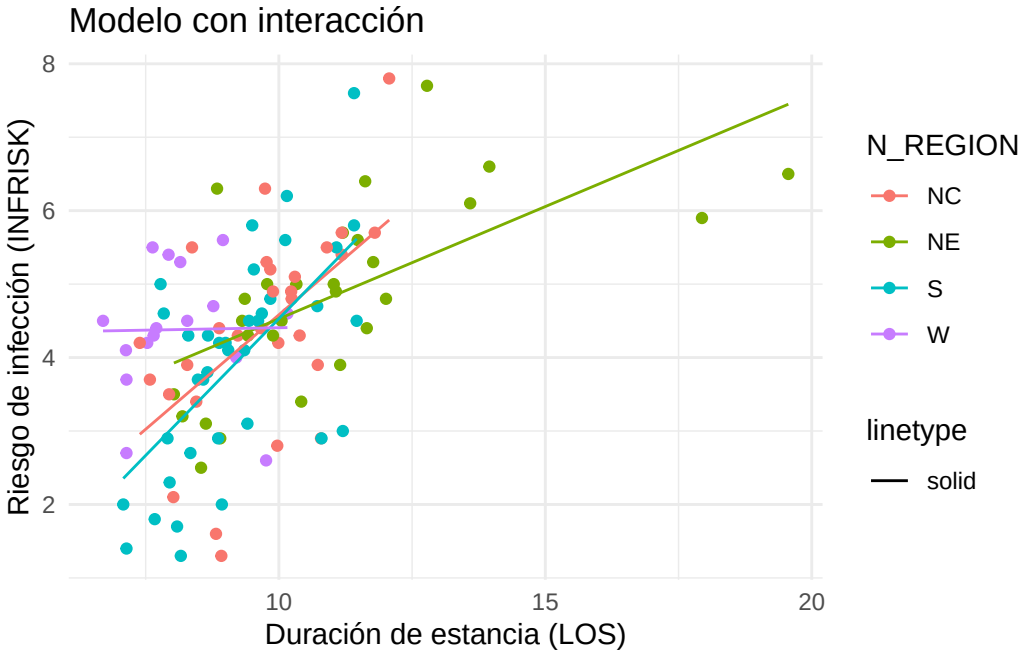
```
Call:
lm(formula = INFRISK ~ LOS * N_REGION, data = data_2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6092 -0.6593  0.1617  0.5130  2.3226

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -1.6510     1.6254  -1.016 0.312147
LOS              0.6233     0.1667   3.739 0.000305 ***
N_REGIONNE       3.1234     1.8574   1.682 0.095699 .
N_REGIONS       -1.2682     2.1323  -0.595 0.553300
N_REGIONW        5.9252     2.8175   2.103 0.037926 *
LOS:N_REGIONNE  -0.3178     0.1844  -1.723 0.087887 .
LOS:N_REGIONS    0.1215     0.2235   0.544 0.587823
LOS:N_REGIONW   -0.6102     0.3273  -1.864 0.065142 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.094 on 102 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.3917,    Adjusted R-squared:   0.35
F-statistic: 9.385 on 7 and 102 DF,  p-value: 6.13e-09
```

6. Genere y muestre las gráficas de regresión para el caso del modelo que incluye a la región y el que incluye a la interacción.

```r
#Quitamos los NA
data_sin_na_2 <- data_2 %>%
  filter(!is.na(INFRISK))

# Crear un data frame con predicciones del modelo sin interacción, pero que toma en cuenta
#a las regiones
data_sin_na_2$predicted_no_interaction <- predict(modelo_x_region)

# Graficar
ggplot(data_sin_na_2, aes(x = LOS, y = INFRISK, color = N_REGION)) +
  geom_point() +
  geom_line(aes(y = predicted_no_interaction)) +
  labs(title = "Modelo sin interacción",
       x = "Duración de estancia (LOS)",
       y = "Riesgo de infección (INFRISK)") +
  theme_minimal()
```




```r
# Crear un data frame con predicciones del modelo con interacción
data_sin_na_2$predicted_interaction <- predict(model_interaction)

# Graficar
ggplot(data_sin_na_2, aes(x = LOS, y = INFRISK, color = N_REGION)) +
  geom_point() +
  geom_line(aes(y = predicted_interaction, linetype = "solid")) +
  labs(title = "Modelo con interacción",
       x = "Duración de estancia (LOS)",
       y = "Riesgo de infección (INFRISK)") +
  theme_minimal()
```



Ejercicio 3

- El riesgo de cáncer de mama está directamente relacionado con la edad de las mujeres a las que tuvieron su primer hijo. Cuando se evalúa la asociación entre el número total de hijos y el riesgo de cáncer de mama, ¿debe controlarse por edad de la madre a la que tuvieron el primer hijo? Argumente su respuesta. En esta cadena causal, ¿qué roles tienen las tres variables involucradas?

Variables de análisis: Número de hijos, edad de la madre al tener su primer hijo, y riesgo de cáncer de mama.

Pregunta de investigación: ¿Debe considerarse la edad de la madre al tener su primer hijo como una variable de control al estudiar la relación entre el número total de hijos y el riesgo de desarrollar cáncer de mama?

La pregunta de investigación anterior, se resume en determinar si la edad de la madre al tener al primer hijo actúa como una variable de confusión en la asociación de interés.

Identificar la presencia de una variable de confusión es crucial, ya que estas pueden distorsionar la estimación del efecto real de la variable de exposición (en este caso, el número total de hijos) sobre el resultado (riesgo de cáncer de mama).

Para evaluar la posibilidad de que la edad sea una variable de confusión, se hace referencia al artículo de la Secretaría de Salud (2009), titulado: “Factores de riesgo asociados con el cáncer de mama”, el cual proporciona evidencia relevante que fácilita argumentos en favor de considerar la edad como una variable de consufión. En este artículo, se menciona que las mujeres con mayor número de hijos presentan un menor riesgo de desarrollar cáncer de mama, atribuyendo este efecto protector a los cambios hormonales y estructurales que ocurren en la glándula mamaria durante el embarazo y la lactancia.

Asimismo, el artículo destaca que la edad al primer embarazo influye directamente en el riesgo de cáncer de mama. Las mujeres que tienen su primer hijo antes de los 30 años presentan un menor riesgo en comparación con aquellas que lo tienen después de esta edad o que no tienen hijos. Este fenómeno se relaciona con la diferenciación completa del tejido mamario inducida por un embarazo temprano, lo que reduce la susceptibilidad a transformaciones malignas.

Dado que la edad al primer hijo afecta tanto al número total de hijos (exposición) como al riesgo de cáncer de mama (resultado), debe considerarse como una variable de control para evitar sesgos en los análisis. No ajustar los resultados por esta variable podría llevar a interpretar incorrectamente la relación entre el número total de hijos y el riesgo de cáncer de mama.

En conclusión, la evidencia respalda la necesidad de incluir la edad de la madre al primer hijo como una variable de control al investigar esta asociación, con el fin de garantizar estimaciones precisas y evitar resultados erróneos.

Referencia.

Secretaría de Salud. (2009). Factores de riesgo asociados con el cáncer de mama. Salud Pública de México, 51(6), 422-429. Recuperado de https://www.scielo.org.mx/scielo.php?pid=S0036-36342009000800006&script=sci_arttext

Ejercicio 4

En la siguiente tabla se muestran los resultados de un estudio de casos y controles para estudiar el efecto del asma y el sexo sobre la alergia primaveral a las gramíneas. Se definen como casos a los individuos alérgicos y como controles a los no alérgicos.

Tabla 4: Alergias

	Hombre		Mujer	
	Asmáticos	**No Asmáticos**	**Asmáticos**	**No Asmáticos**
Casos	**19**	**8**	**10**	**24**
Controles	**51**	**84**	**31**	**89**

***OR* para asma y alergia**

- ¿Existe relación entre el hecho de ser asmático y la definición de caso y control (sin considerar la variable sexo). Utilizar la medida de riesgo adecuada, e interpretar los resultados obtenidos (Usar $\alpha = 0.05$).

Como no se tomará en cuenta el sexo, se deben sumar hombres y mujeres pero haciendo la distinción entre casos y controles; asmáticos y no asmáticos:

Tabla 5: Tabla de Asma y Alergia

Asma	Alergia (Casos)	No Alergia (Controles)	Total
Sí	29	82	111
No	32	173	205
Total	61	255	316

Como medida de riesgo, se utilizará la razón de momios (*OR*), para ello se necesita primero el cálculo de los momios en cada caso, estos se definen como:

$$\Omega_1 = \frac{\mathbb{P}[Y=1|X=1]}{\mathbb{P}[Y=2|X=1]} = \frac{\pi_1}{1-\pi_1} \qquad \Omega_2 = \frac{\mathbb{P}[Y=1|X=2]}{\mathbb{P}[Y=2|X=2]} = \frac{\pi_2}{1-\pi_2}$$

Donde Y es el desarrollo o no de la alergia; y X es el factor de exposición: tener o no asma. Entonces, el *OR* se calcula como el cociente del momio de tener alergia entre los que tienen asma y el momio de tener alergia entre los que no tienen asma:

$$\theta = \frac{\Omega_1}{\Omega_2}$$

Con intervalo de confianza igual a:

$$OR \in \left(\exp\left\{ \log(\hat{OR}) \pm Z_{(1-\alpha/2)} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \right\} \right)$$

Se genera una función para calcular el cociente de momios y su intervalo de confianza:

```
IC_OR <- function(X,alpha){

  a <- X[1]
  b <- X[3]
  c <- X[2]
  d <- X[4]

  OR <- (a*d)/(b*c)

  sd<-sqrt(1/a+1/b+1/c+1/d)
  IC<-exp(log(OR)+c(-1,1)*qnorm(1-alpha/2)*sd)

  return(list(OR = OR, IC = IC))
}
```

Se calcula el cociente de momios y su intervalo a un 95% de confianza:

```
resultado <- IC_OR(matrix_asma_alergia, alpha = 0.05)
OR_alergia <- resultado$OR
IC_alergia <- resultado$IC
```

El *OR* de tener alergia en los que tienen asma es 1.912 veces el momio de tener alergia en los que no tienen asma. Y el intervalo de confianza para el OR, al 95% de confianza, es de [1.084, 3.371]. Dado que el intervalo no contiene al valor nulo, se puede decir que existe una relación positiva débil entre tener asma y desarrollar alergia primaveral. O bien, que las personas que tienen asma desarrollan más la alergia primaveral que las personas que no tienen asma.

Variable de interacción

- ¿Es la variable sexo modificadora del efecto? Justifique su respuesta.

Para poder detectar si la variable sexo es modificadora del efecto, también llamada una variable de interacción, se puede realizar análisis estratificado calculando el *OR* en los hombres y en las mujeres por separado y, al promediarlos, se debería tener un valor próximo al del *OR* sin estratificar.

Se genera la tabla de 2x2 para el estrato de los hombres:

Tabla 6: Tabla de Asma y Alergia en Hombres

Asma	Alergia (Casos)	No Alergia (Controles)	Total
Sí	19	51	70
No	8	84	92
Total	27	135	162

Se calcula el *OR* del estrato de los hombres:

```
resultado <- IC_OR(matrix_asma_alergia_h, alpha = 0.05)
OR_alergia_h <- resultado$OR
IC_alergia_h <- resultado$IC
```

El *OR* de tener alergia en los que tienen asma es 3.912 veces el momio de tener alergia en los que no tienen asma; en los hombres. Y el intervalo de confianza para el *OR*, al 95% de confianza, es de [1.596, 9.586]. Dado que el intervalo no contiene al valor nulo, se podría decir que existe una relación positiva moderada, en los hombres, entre tener asma y desarrollar alergia primaveral.

Ahora se calculará el *OR* para el estrato de las mujeres. Primero se genera la tabla de 2x2 para el estrato de las mujeres:

Tabla 7: Tabla de Asma y Alergia en Mujeres

Asma	Alergia (Casos)	No Alergia (Controles)	Total
Sí	10	31	41
No	24	89	113
Total	34	120	154

Se calcula el *OR* del estrato de las mujeres:

```
resultado <- IC_OR(matrix_asma_alergia_m, alpha = 0.05)
OR_alergia_m <- resultado$OR
IC_alergia_m <- resultado$IC
```

En las mujeres, el *OR* de tener alergia en los que tienen asma es 1.196 veces el momio de tener alergia en los que no tienen asma. Y el intervalo de confianza para el *OR*, al 95% de confianza, es de [0.515, 2.78]. Dado que el intervalo sí contiene al valor nulo, se podría decir que no existe una relación, en las mujeres, entre tener asma y desarrollar alergia primaveral.

Tomando en cuenta que en un caso sí hay relación y en otro no, ahora se realiza el promedio de los *OR* estratificados:

```
promedio_OR <- (OR_alergia_h + OR_alergia_m)/2
```

El promedio de los *OR* es de 2.554 y el *OR* sin estratificar era de 1.912, hay una pequeña diferencia entre ambos valores. Aunque la diferencia entre el *OR* de los hombres es 3.27 veces el de las mujeres.

Para confirmar si esta diferencia es lo suficientemente significativa, se recurre a la prueba de homogeneidad de Woolf (1955), que permite comparar el *OR* de los dos estratos (hombres y mujeres). En el caso en que el sexo sea una variable de interacción, los *OR* deberían diferir de manera significativa entre sí. Se toma como hipótesis nula que los grupos son homogéneos y como altera que no son homogéneos, tomando de referencia $\alpha = 0.05$.

Se aplica el estadístico de prueba:

```
alergias <- xtabs(freq ~ .,
                  cbind(expand.grid(risk=c("Asma","No Asma"),
                                    response=c("Alergia","No Alergia"),
                                    gender=c("Hombre","Mujer")),
                        freq=c(19,8,51,84,10,24,31,89))
)

WoolfTest(alergias)
```

```
    Woolf Test on Homogeneity of Odds Ratios (no 3-Way assoc.)

data:  alergias
X-squared = 3.5601, df = 1, p-value = 0.05918
```

Dado que el *p-value* resultó mayor a 0.05, se rechaza la hipótesis alterna de que los grupos no son homogéneos. Por lo tanto, la variable sexo no es una de interacción o que modifique el efecto de la relación entre el asma y el desarrollo de alergia. Es decir, el efecto del asma sobre la alergia en realidad es consistente para hombres y mujeres.

Variable de confusión

Se consideran los siguientes requisitos para poder considerar a una variable como de confusión. Se define al factor de exposición como la presencia de asma, la presencia del factor de riesgo o enfermedad es poseer alergia de primavera y al factor de confusión como el sexo de la persona.

1. **Asociación entre el factor de confusión y la exposición en los no enfermos**. Se debe encontrar una asociación entre el sexo y el asma en el grupo de personas no alérgicas (controles). Es decir, un *OR* diferente a 1 en las personas que no tienen alergia.

Se genera la tabla con los datos:

Tabla 8: Tabla de Sexo y Asma para Controles (No Alergia)

Sexo	Asma (Sí)	Asma (No)	Total
Hombre	51	84	135
Mujer	31	89	120
Total	82	173	255

Se calcula el *OR* para los controles:

```
resultado <- IC_OR(matrix_sexo_asma_controles, 0.05)

OR_sex_asm_control <- resultado$OR
IC_sex_asm_control <- resultado$IC
```

En los controles, el *OR* de tener asma en los hombres es 1.743 veces el momio de tener asma en las mujeres. Y el intervalo de confianza para el *OR*, al 95% de confianza, es de [1.019, 2.982]. Como el intervalo no contiene al uno, se podría decir que existe una asociación positiva entre el tener asma y ser hombre. O bien, que dentro de las personas que no tienen alergias, los hombres desarrollan más el asma que las mujeres. Para comprobar la fuerza de esta asociación, se realizará una prueba χ^2 para independencia.

Se calcula el estadístico de prueba utilizando la función que toma en cuenta la corrección por continuidad de Yates:

```
chisq.test(matrix_sexo_asma_controles)
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  matrix_sexo_asma_controles
X-squared = 3.6251, df = 1, p-value = 0.05691
```

El *p-value* obtenido con la función devuelve un valor de 0.05691, el cual es mayor al 0.05. Entonces se rechaza la hipótesis nula de que las variables no son independientes, es decir que no hay suficiente evidencia como para rechazar la independencia de las variables. Al igual que el resultado obtenido en el análisis de variables de interacción, se puede decir que el asma influye de la misma manera a hombres y mujeres que no tienen alergias. Por lo que, al no haber relación, el primer requisito no se cumple.

2. **Asociación entre el factor de confusión y la enfermedad en los grupos no expuestos**. Se debe evaluar si existe una relación entre el sexo y la presencia de alergia en el grupo de personas que no tienen asma (no expuestos).

Se genera la tabla con los datos:

Tabla 9: Tabla de Sexo y Alergia para No Expuestos (No Asma)

Sexo	Asma (Sí)	Asma (No)	Total
Hombre	8	84	92
Mujer	24	89	113
Total	32	173	205

Se calcula el *OR* para los no expuestos:

```
resultado <- IC_OR(matrix_sexo_asma_no_exp, 0.05)

OR_no_exp <- resultado$OR
IC_no_exp <- resultado$IC
```

En las personas que no tienen asma, el *OR* de tener alergia en las mujeres es 2.831 veces el momio de tener alergia en los hombres. Y el intervalo de confianza para el *OR*, al 95% de confianza, es de [1.206, 6.65]. Como el intervalo no contiene al uno, se podría decir que existe una asociación positiva entre el tener alergia y ser mujer. O bien, que dentro de las personas que no tienen asma, las mujeres desarrollan más alergias que los hombres. Para comprobar la fuerza de esta asociación, se realizará una prueba χ^2 para independecia.

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  matrix_sexo_asma_no_exp
X-squared = 5.1421, df = 1, p-value = 0.02335
```

El *p-value* obtenido con la función devuelve un valor de 0.02335, el cual es menor al 0.05. Por lo que se rechaza la hipótesis nula de que las variables son independientes. En este caso, se puede decir que existe una relación positiva, en las personas que no tienen asma, de desarrollar alergia y ser mujer. Por lo que, al haber asociación, el segundo requisito sí se cumple.

3. **El factor de confusión no debe ser un paso intermedio en la secuencia causal entre el tipo de tratamiento y su éxito**. En este caso, como el sexo es una característica definida o intrínseca de los individuos, no se puede considerar como un paso intermedio entre el asma y la alergia. Es decir, no es posible que el asma provoque el sexo de las personas y con ello un mayor o menor aumento del desarrollo de alergias. Por lo que este criterio se cumple.

Tomando de referencia que el valor del *OR* bruto es 1.912 con un intervalo de confianza de [1.084 , 3.371] se puede apreciar que el intervalo está muy cerca del valor nulo, donde se demostraría que no hay asociación entre el desarrollo de alergia primaveral y tener o no asma, independientemente del sexo. Al estratificar por sexo, el *OR* de los hombres es 3.912 con un intervalo de confianza de [1.596 , 9.586], el cual revelaría una posible asociación positiva moderada entre tener alergia primaveral y asma para los hombres. Por último, el *OR* de las mujeres es 1.196 con un intervalo de confianza de [0.515 , 2.78], el cual demuestra que en el caso de las mujeres no se encuentra asociación entre tener alergia primaveral y asma. Dentro de los criterios, se encontró que dos de los tres se cumplen para ser una variable de confusión y uno no lo cumple por tener un *p-value* de 0.05691, es decir un valor muy cercano a la zona de rechazo. Tomando toda esta información, se concluye que la variable sexo sí es una variable de confusión, la cual genera confusión al enmascarar el efecto, donde, de tener más información, se podría notar que el *OR* global sea cercano a 1 (sin asociación) pero que el *OR* de un estrato muestra asociación (el de los hombres) y el otro no.

Conclusión

Para toda la información obtenida, se puede decir que, al no estratificar por sexo, no hay asociación entre el desarrollo de alergia primaveral y tener asma. Al estratificar por sexo, solo los hombres muestran una asociación moderada entre entre el desarrollo de alergia y tener asma; esta relación no se encuentra en las mujeres. Por ello, la variable sexo se considera como una de confusión, pero no una de interacción; esto último debido a no tener suficiente evidencia estadística para indicar que los *OR* de cada estrato difieran de manera significativa entre ellos.

Ejercicio 5

La siguiente tabla corresponde a la clasificación cruzada de las variables: Enfermedad coronaria (CHD, por sus siglas en inglés), Fumar (Smoke) y bebedor de café (Coffee).

Tabla 10: CHD, Soke and Coffee

CHD	Smoke	Coffee	n
Sí	No	No	15
No	No	No	42
Sí	Sí	No	11
No	Sí	No	8
Sí	No	Sí	15
No	No	Sí	21
Sí	Sí	Sí	25
No	Sí	Sí	14

- Utilizando un modelo de regresión logística (*logit*) responda: ¿Están el tabaquismo y/o el café relacionados con el incremento del momio de CHD? (Use $\alpha = 0.05$).

Para aplicar el modelo de regresión logística, primero modificamos la tabla para que represente a cada observación en las filas y no como al conjunto de las frecuencias en cada grupo:

```
chd_expanded <- chd_smoke_coffee %>%
  uncount(n)

head(chd_expanded)
```

```
  CHD Smoke Coffee
1  Sí    No     No
2  Sí    No     No
3  Sí    No     No
4  Sí    No     No
5  Sí    No     No
6  Sí    No     No
```

Se convierte la variable dependiente (CHD) a una variable *dummy* de 0 y 1 y se aplica el modelo logístico con *Smoke* y *Coffee* como variables independientes; este se definiría como:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \cdot \text{Smoke} + \beta_2 \cdot \text{Coffee}$$

Donde:

π =probabilidad de que ocurra CHD

β_0 = Intercepto. El $log(OR)$ de tener CHD en los no fumadores y no bebedores de café

β_1 = Coeficiente asociado a fumar $(Smoke)$

β_2 = Coeficiente asociado al consumo de café $(Coffee)$

```
chd_expanded$CHD_bin <- ifelse(chd_expanded$CHD == "Sí", 1, 0)

modelo <- glm(CHD_bin ~ Smoke + Coffee, family = binomial, data = chd_expanded)
```

```
summary(modelo)
```

```
Call:
glm(formula = CHD_bin ~ Smoke + Coffee, family = binomial, data = chd_expanded)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9572     0.2703  -3.541 0.000398 ***
SmokeSí       1.1020     0.3610   3.053 0.002269 **
CoffeeSí      0.5270     0.3542   1.488 0.136798
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 206.93  on 150  degrees of freedom
Residual deviance: 191.74  on 148  degrees of freedom
AIC: 197.74

Number of Fisher Scoring iterations: 4
```

Se calcula e^{β_i} *con* $i = 1, 2$ para obtener cada *OR*

```
exp(cbind(OR = coef(modelo), confint(modelo)))
```

```
                    OR     2.5 %    97.5 %
(Intercept) 0.3839539 0.2215322 0.6424622
SmokeSí     3.0101153 1.4935502 6.1770819
CoffeeSí    1.6938032 0.8447583 3.4022183
```

El *OR* de tener CHD en los fumadores es 3.01 veces el *OR* de los no fumadores que tienen CHD; y con una confianza al 95% su intervalo no contiene al valor nulo, por lo que se dice que hay una asociación positiva entre tener CHD y ser fumador. O bien, al examinar el $ln(OR)$ se encuentra que el *p-value* es menor a 0.05, confirmando la asociación entre ser fumador y tener CHD.

El momio de tener CHD en los que sí beben café es 1.69 veces el momio de los que tienen CHD y no beben café; y su intervalo contiene al valor nulo, por lo que se dice que no hay una asociación entre tener CHD y ser bebedor de café, con una confianza de 95%. O bien, al examinar el $ln(OR)$ se encuentra que el *p-value* es mayor a $\alpha = 0.05$, confirmando que no hay asociación entre ser bebedor de café y tener CHD.

En conclusión, solo el ser fumador se encuentra asociado de manera positiva con tener enfermedad coronaria

- ¿Es la asociación de café y CHD mayor entre los fumadores? Es decir, ¿fumar es un efecto modificador de las asociaciones de café-CHD? (Use $\alpha = 0.05$)

Se genera el modelo que incluye a la interacción entre fumar y ser bebedor de café y se examina el coeficiente de la variable de interacción; este sería:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \cdot \text{Smoke} + \beta_2 \cdot \text{Coffee} + \beta_3 \cdot (\text{Smoke} \cdot \text{Coffee})$$

Donde:

$\beta_3 =$ Coeficiente asociado a la interacción $(Smoke \cdot Coffee)$

```r
# Modelo con interacción
modelo_con_interaccion <- glm(CHD_bin ~ Smoke * Coffee,
                              family = binomial,
                              data = chd_expanded)

summary(modelo_con_interaccion)
```

```
Call:
glm(formula = CHD_bin ~ Smoke * Coffee, family = binomial, data = chd_expanded)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.0296     0.3008  -3.423 0.000619 ***
SmokeSí            1.3481     0.5535   2.435 0.014873 *
CoffeeSí           0.6931     0.4525   1.532 0.125573
SmokeSí:CoffeeSí  -0.4318     0.7295  -0.592 0.553899
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 206.93  on 150  degrees of freedom
Residual deviance: 191.39  on 147  degrees of freedom
AIC: 199.39

Number of Fisher Scoring iterations: 4
```

Se calcula e^{β_i} *con* $i = 1, 2, 3$ para obtener cada *OR*

```r
exp(cbind(OR = coef(modelo_con_interaccion), confint(modelo_con_interaccion)))
```

```
                       OR     2.5 %     97.5 %
(Intercept)      0.3571429 0.1917723  0.6292934
SmokeSí          3.8500000 1.3186041 11.7863193
CoffeeSí         2.0000000 0.8242030  4.9042592
SmokeSí:CoffeeSí 0.6493506 0.1534047  2.7107380
```

Como se puede observar, el *p-value* del coeficiente asociado a la interacción entre las variables de ser bebedor de café y fumador es mayor al $\alpha = 0.05$, por lo que se puede concluir que no hay interacción significativa de los fumadores en la relación de ser bebedor de café y tener enfermedad coronaria. Si se aplica el exponencial y calculando el intervalo a un 95% de confianza, se observa que el intervalo $[0.1534, 2.7101]$ contiene al valor nulo, por lo que se confirma que, de manera significativa, no hay interacción de los fumadores en la relación de tomar café y tener enfermedad coronaria

Ejercicio 6

La base de datos Melanoma.csv, contiene información sobre un estudio de cohorte para estudiar la muerte por melanoma realizado en Dinamarca. La descripción de esta base es la siguiente:

Description: The Melanoma data frame has data on 205 patients in Denmark with malignant melanoma. This data frame contains the following columns: time: survival time in days, possibly censored. status: 1 died from melanoma, 2 alive, 3 dead from other causes. sex : 1 = male, 0 = female. age: age in years. year : of operation. thick- ness: tumour thickness in mm. ulcer : 1 = presence, 0 = absence. Con estos datos realice:

• Calcule la tasa de incidencia anual por 1000 de muerte por melanoma

Si definimos a D como el número total de ocurrencias y a T como el periodo de Ocurrencia, por lo visto en clase, podemos asumir que D tiene una distribución Poisson con una tasa de ocurrencia λ. De esta manera, el estimador máximo verosímil de λ queda definido como:

$$\left[\lambda = \frac{D}{T}\right]$$

```
# Muertes por melanoma
D<- datos[datos$status == 1, ]
# Tiempo total anualizado
T <- sum(datos$time)/365
# Tasa de incidencia anual por 1000 (estimador máximo verosímil)
lam <- (nrow(D) / T) * 1000
cat("La muerte anual por melanoma es de", lam, "personas de cada mil.")
```

```
La muerte anual por melanoma es de 47.14224 personas de cada mil.
```

- Realice los 4 tipos intervalos de confianza vistos en clase (la programación debe ser de ustedes, no la del script) ($\alpha = 0.05$)

```
# Intervalos de confianza por aproximación normal
error_estandar <- sqrt(nrow(D)) / T
z <- qnorm(0.975)  # Para un nivel de confianza del 95%
ic_normal <- c(lam - z * error_estandar * 1000,
               lam + z * error_estandar * 1000)
ic_normal=round(ic_normal,2)
# Resultados
cat("IC Aproximación Normal:\n")
```

```
IC Aproximación Normal:
```

```
cat("Límite inferior:", ic_normal[1], "\n")
```

```
Límite inferior: 34.9
```

```
cat("Límite superior:", ic_normal[2], "\n")
```

```
Límite superior: 59.38
```

```
# Intervalos de confianza por conteos bajos
casos <- sum(datos$status == 1)
tiempo <- sum(datos$time) / 365  # Tiempo total de observación en años
k <- 1000
tasa <- casos / tiempo
alpha <- 0.05

# Cuantil de la distribución normal estándar
z <- qnorm(1 - alpha / 2)
zi <- (z / 3) * sqrt(1 / (casos + 0.5))

# Límite inferior del intervalo
li <- ((casos + 0.5) / tiempo) * ((1 - (1 / (9 * (casos + 0.5))) - zi)^3) *

# Límite superior del intervalo
ls <- ((casos + 0.5) / tiempo) * ((1 - (1 / (9 * (casos + 0.5))) + zi)^3) *

# Redondeo
li <- round(li, 2)
ls <- round(ls, 2)

# Resultados
cat("IC Aproximación Conteos Bajos:\n")
```

```
IC Aproximación Conteos Bajos:
```

```
cat("Límite inferior :", li, "\n")
```

```
Límite inferior : 36.06
```

```
cat("Límite superior :", ls, "\n")
```

```
Límite superior : 60.61
```

```
# Intervalos de confianza Gamma
casos <- sum(datos$status == 1)
tiempo <- sum(datos$time) / 365
k <- 1000
alpha <- 0.05
tasa <- casos / tiempo

# Cuantiles de la distribución Gamma
g_lower <- qgamma(alpha / 2, casos)  # Límite inferior
g_upper <- qgamma(1 - alpha / 2, casos + 1)  # Límite superior
```

```r
# Cálculo de los límites
li <- (g_lower / tiempo) * k  # Límite inferior escalado
ls <- (g_upper / tiempo) * k  # Límite superior escalado

# Redondeo
li <- round(li, 2)
ls <- round(ls, 2)

cat("IC Método Gamma:\n")
```

```
IC Método Gamma:
```

```r
cat("Límite inferior :", li, "\n")
```

```
Límite inferior : 35.71
```

```r
cat("Límite superior :", ls, "\n")
```

```
Límite superior : 61.08
```

```r
# Intervalos de confianza exacto por Cassela y Berger
casos <- sum(datos$status == 1)
tiempo <- sum(datos$time) / 365
k <- 1000
alpha <- 0.05

# Función para calcular el límite inferior
calcular_LI <- function(LI, Y) {
  1 - ppois(Y, LI) + dpois(Y, LI) - alpha / 2
}

# Función para calcular el límite superior
calcular_LS <- function(LS, Y) {
  ppois(Y, LS) - alpha / 2
}

# Calcular el límite inferior
root_LI <- uniroot(calcular_LI, interval = c(0, casos * 10), Y = casos, exte
li <- root_LI$root / tiempo  # Convertir el límite inferior a tasa

# Calcular el límite superior
root_LS <- uniroot(calcular_LS, interval = c(0, casos * 10), Y = casos, exte
ls <- root_LS$root / tiempo  # Convertir el límite superior a tasa

li <- round(li, 2) * k
ls <- round(ls, 2) * k
```

```r
# Imprimir los resultados
cat(" IC Método Exacto (Casella y Berger):\n")
```

```
 IC Método Exacto (Casella y Berger):
```

```r
cat("Límite inferior :", li, "\n")
```

```
Límite inferior : 40
```

```r
cat("Límite superior :", ls, "\n")
```

```
Límite superior : 60
```

- ¿Es estadísticamente distinta esta tasa de incidencia entre hombres y mujeres?. Argumente su respuesta. ($\alpha = 0.05$)

```r
# Filtrar datos para hombres
hombres <- datos[datos$sex == 1, ]
hm <- sum(hombres$status == 1)  # Muertes por melanoma en hombres
th <- sum(hombres$time) / 365  # Tiempo total de observación para hombres en

# Filtrar datos para mujeres
mujeres <- datos[datos$sex == 0, ]
mm <- sum(mujeres$status == 1)  # Muertes por melanoma en mujeres
tm <- sum(mujeres$time) / 365  # Tiempo total de observación para mujeres en

alpha <- 0.05
z <- qnorm(1 - alpha / 2)
tasa_hombres <- hm / th
error_hombres <- sqrt(hm) / th
li_hombres <- (tasa_hombres - z * error_hombres) * 1000
ls_hombres <- (tasa_hombres + z * error_hombres) * 1000
tasa_hombres <- tasa_hombres * 1000
tasa_mujeres <- mm / tm
error_mujeres <- sqrt(mm) / tm
li_mujeres <- (tasa_mujeres - z * error_mujeres) * 1000
ls_mujeres <- (tasa_mujeres + z * error_mujeres) * 1000
tasa_mujeres <- tasa_mujeres * 1000

# Redondeo
tasa_hombres <- round(tasa_hombres, 2)
li_hombres <- round(li_hombres, 2)
ls_hombres <- round(ls_hombres, 2)

tasa_mujeres <- round(tasa_mujeres, 2)
li_mujeres <- round(li_mujeres, 2)
```

```r
ls_mujeres <- round(ls_mujeres, 2)

#  Resultados
resultados <- data.frame(
  Grupo = c("Hombres", "Mujeres"),
  Tasa = c(tasa_hombres, tasa_mujeres),
  LI = c(li_hombres, li_mujeres),
  LS = c(ls_hombres, ls_mujeres)
)
print(resultados)
```

```
   Grupo  Tasa    LI    LS
1 Hombres 68.86 43.80 93.93
2 Mujeres 35.53 22.37 48.70
```

```r
# comparación por medio del riesgo relativo
RR <- tasa_hombres/tasa_mujeres
# Intervalo de confianza
se_logRR <- sqrt(1/hm+1/mm)
#Intervalo de confianza del log(RR)
alpha<-0.05
z<-qnorm(1-alpha/2)
LS<-exp(log(RR)-(z*se_logRR))
LI<-exp(log(RR)+(z*se_logRR))
#IRR con intervalo de confianza del 95%
RR_IC<-data.frame(RR=RR,LS=LS,LI=LI)
print(RR_IC)
```

```
      RR      LS       LI
1 1.93808 1.15305 3.257582
```

```r
# prueba de hipótesis
#Donde H0: log(rr)=0 vs. H1: log(rr)/=0
Z <- (log(RR)-0)/se_logRR
p_val <- 2*(1-pnorm(Z))
hp<-data.frame("Z"=Z,"p-Value"=p_val)
print(hp)
```

```
        Z    p.Value
1 2.497471 0.01250827
```

```r
# Se observa que el P-value es menor a 0.05 (nivel de significancia) por lo
```

- Calcule la tasa de incidencia anual por 1000 de muerte dentro de este estudio.

```r
muertes_totales <- sum(datos$status == 1 | datos$status == 3)  # Muertes por
tiempo_total <- sum(datos$time) / 365

# Tasa de incidencia total
tasa_total <- (muertes_totales / tiempo_total) * 1000
cat("Tasa de incidencia total por 1000 personas al año:", tasa_total, "\n")
```

```
Tasa de incidencia total por 1000 personas al año: 58.72103
```

- ¿Es estadísticamente distinta esta tasa de incidencia de la de muerte solo por melanoma?. Argumente su respuesta. ($\alpha = 0.05$)

```r
muertes_melanoma <- sum(datos$status == 1)
muertes_totales <- sum(datos$status == 1 | datos$status == 3)
tiempo_total <- sum(datos$time) / 365

tasa_melanoma <- muertes_melanoma / tiempo_total
tasa_total <- muertes_totales / tiempo_total

# Varianzas
var_melanoma <- muertes_melanoma / (tiempo_total^2)
var_total <- muertes_totales / (tiempo_total^2)

# Estadístico Z
z <- (tasa_melanoma - tasa_total) / sqrt(var_melanoma + var_total)

# Cuantil crítico para alpha = 0.05
z_critico <- qnorm(1 - 0.05 / 2)

# Resultados
cat("Tasa por melanoma (por 1000 personas-año):", round(tasa_melanoma * 1000
```

```
Tasa por melanoma (por 1000 personas-año): 47.14
```

```r
cat("Tasa total (por 1000 personas al año):", round(tasa_total * 1000, 2), "
```

```
Tasa total (por 1000 personas al año): 58.72
```

```r
cat("Estadístico Z:", round(z, 2), "\n")
```

```
Estadístico Z: -1.24
```

```r
cat("Cuantil :", round(z_critico, 2), "\n")
```

```
Cuantil : 1.96
```

```r
if (abs(z) > z_critico) {
  cat("Conclusión: Las tasas son estadísticamente distintas (alpha = 0.05).\
} else {
  cat("No hay evidencia suficiente para afirmar que las tasas son distintas
}
```

```
No hay evidencia suficiente para afirmar que las tasas son distintas (alpha =
```

Ejercicio 7

- ¿Existe diferencia significativa en las tasas de muerte al tiempo de inicio en el empleo (`start`)?. Responda esta pregunta construyendo las tasas por nivel de la covariable y después comparándolas como en el ejercicio previo.

Tabla 11: Tabla de tasas de incidencia

	pre1925	post1925	Total
Casos	161.0	115.0	276.0
Tiempo-Persona	96819.9	28170.8	124990.7

Suponiendo que nuestro nivel de exposición aumento con el tiempo, se tomará como exposición al periodo posterior a 1925 y como no expuesto al periodo anterior a 1925.

Definimos una función para estimar la razón de tasas de incidencia (IRR).

```
## Función para calcular el RR y su intervalo de confianza
calcular_IC_RR <- function(E1, T1, E2, T2, nivel_confianza = 0.95) {
  # Calcular el RR
  RR <- (E1 / T1) / (E2 / T2)

  # Logaritmo natural del RR
  ln_RR <- log(RR)

  # Error estándar del log(RR)
  SE_ln_RR <- sqrt(1 / E1 + 1 / E2)

  # Valor crítico de Z para el nivel de confianza especificado
  Z <- qnorm(1 - (1 - nivel_confianza) / 2)

  # Intervalo de confianza en escala log
  lower_ln <- ln_RR - Z * SE_ln_RR
  upper_ln <- ln_RR + Z * SE_ln_RR

  # Convertir a escala original
  IC_lower <- exp(lower_ln)
  IC_upper <- exp(upper_ln)

  # Resultado
  list(
    RR = RR,
    IC = c(IC_lower, IC_upper),
    SE = SE_ln_RR,
    nivel_sig = nivel_confianza)
 }
```

Estimamos el IRR y hacemos la prueba de hipótesis.

```
rr.stats <- calcular_IC_RR(tasas.start.year$cases[2],
                tasas.start.year$personyear[2],
                tasas.start.year$cases[1],
                tasas.start.year$personyear[1])

Z.test <- log(rr.stats$RR) / rr.stats$SE

p.value.rr.rate <- 2*(1-pnorm(Z.test))
```

Las tasas de incidencia 1000 persona-año es de 1.6629 para el periodo del inicio de trabajo antes de 1925 y 4.0822 para el periodo posterior a 1925. La razón de las tasas de incidencia es de 2.4549, con un intervalo de confianza al 95% de [1.9325, 3.1186].

Usando un nivel de significancia del 0.05, la prueba de hipótesis ($H_0 : \log RR = 0$) resulta ser significativa con un *p-value* = 1.898481e-13 y un estadístico Z de 7.3558. Por lo cual podemos rechazar la hipótesis nula de que ambas tasas de incidencia son iguales, al ser de dos colas solo se puede afirmar que son distintas sin indicar su dirección.

2. Utilizando el modelo de regresión Poisson, utilizando como *offset* la variable `personyrs`, responda la misma pregunta del inciso anterior. Este ejercicio es para que comprueben que sale lo mismo con los dos procedimientos y aprendan a usar este modelo.

El modelo de regresión Poisson es:

$$\begin{aligned} \log \text{rate} &= X_i^T \beta \\ &= \log \text{personyrs} + \beta_0 + \beta_1 \cdot \text{start} \end{aligned}$$

Donde:

rate $= \frac{\text{respdeath}}{\text{personyrs}} =$ es la tasa de incidencia de muerte respiratoria.

$\log$ personyrs $=$ Variable *offset* que ajusta por el tiempo de exposición.

$\beta_0 =$ Corresponde al $\log$ de la tasa esperada en el nivel de referencia (`start` = 1).

$\beta_1 =$ Coeficiente asociado a la razón de la tasa de incidencia $I\hat{R}R$ para el tiempo de inicio en el empleo (`start`).

Se escribe el modelo en `R`.

```
poisson.rate <- glm(
    formula = respdeath ~ start,
    family = "poisson",
    offset =  log(personyrs),
    data = df.resp.deaths)
```

Tabla 12

Como se observa en la tabla 12, el coeficiente estimado por el modelo de regresión corresponde al riesgo

relativo estimado anteriormente, además de que su intervalos de confianza son idénticos. Además su estimación es significativa estadísticamente a un nivel de $\alpha = 0.05$ ($Z = 7.356,$ p-value $= 1.9e - 13$).

Se puede probar la suposición de que la varianza y la media de los datos son iguales, que subyace al modelo de regresión Poisson, con la siguiente instrucción de R.

```
epiDisplay::poisgof(poisson.rate)
```

```
$results
[1] "Goodness-of-fit test for Poisson assumption"

$chisq
[1] 326.0066

$df
[1] 112

$p.value
[1] 8.963121e-23
```

Dado que nuestra hipótesis nula es que este supuesto se cumple, podemos ver que el modelo de Poisson no es el más indicado, por lo cual se realizará con el modelo quasipoison.

```
qpoisson.rate <- glm(
    formula = respdeath ~ start,
    family = "quasipoisson",
    offset =  log(personyrs),
    data = df.resp.deaths)
```

Tabla 13: Estimadores de Regressión QuasiPoisson (start).

Variable	IRR[1]	95% CI[1]	p-value
(Intercept)	0.00068	0.00033, 0.00134	<0.001
start	2.45492	1.52958, 3.89271	<0.001

[1] IRR = Razón de tasas de incidencia, CI = Intervalo de confianza

Se observa en la tabla 13 que la IRR no cambia a diferencia de la amplitud del intervalo de confianza que se amplia tomando en consideración que existe sobre dispersión o subdispersión, para verificar este hecho se utiliza el mismo modelo estimado.

```
summary(qpoisson.rate)$dispersion
```

```
[1] 3.77482
```

Lo que nos indica una sobre dispersión de 3.77, por ende es más adecuado el uso de un modelo de quasipoisson o uno binomial negativo.

- Utilizando regresión Poisson pruebe si hay diferencias significativas en las tasas muerte por nivel de arsénico.

El modelo de regresión Poisson es:

$$\begin{aligned} \log \mu_i &= X_i^T \beta \\ &= \log \text{personyrs} + \beta_0 + \beta_{1l} \cdot \text{arsenic} \end{aligned}$$

Donde:

rate $= \frac{\text{respdeath}}{\text{personyrs}} =$ es la tasa de incidencia de muerte respiratoria.

$\log$ personyrs $=$ Variable *offset* que ajusta por el tiempo de exposición.

$\beta_0 =$ Corresponde al $\log$ de la tasa esperada en el nivel de referencia (`arsenic` = 1)

$\beta_{1l} =$ Coeficiente asociado a la razón de la tasa de incidencia $I\hat{R}R$ para los distintos niveles de exposición ($l = 2, 3, 4$) al arsénico (`arsenic`).

Se escribe el modelo en `R`.

```
poisson.rate.arsenic <- glm(
    formula = respdeath ~ arsenic,
    family = "poisson",
    offset =  log(personyrs),
    data = df.resp.deaths)
```

Tabla 14: Estimadores de Regressión Poisson (arsenic).

Variable	IRR[1]	95% CI[1]	p-value
(Intercept)	0.00161	0.00137, 0.00188	<0.001
arsenic			
1	—	—	
2	2.08395	1.51816, 2.81843	<0.001
3	2.03434	1.33190, 2.99317	<0.001
4	4.08421	2.86695, 5.69117	<0.001

[1] IRR = Razón de tasas de incidencia, CI = Intervalo de confianza

Los coeficientes estimados son significativos (veáse tabla 14), indicando que el riesgo de muerte por enfermedad respiratoria es dos veces mayor para un nivel de exposición 2 (1-4 años) y 3 (5-14 años) con respecto al nivel de exposición 1 (<1 año) y es cuatro veces mayor cuando el tiempo de exposición al arsénico es mayor a 15 años (nivel 4). El nivel de riesgo esperado cuando el nivel de exposición es el más bajo es prácticamente nulo, por tanto el modelo indica que este nivel de exposición no afecta la tasa de muerte por enfermedad respiratoria.

Igualmente se comprueba la suposición del modelo de regresión Poisson.

```
epiDisplay::poisgof(poisson.rate.arsenic)
```

```
$results
[1] "Goodness-of-fit test for Poisson assumption"

$chisq
[1] 311.1098

$df
[1] 110

$p.value
[1] 4.214491e-21
```

Se vuelve a obtener que el modelo no es el más apropiado. Al ajustar un modelo quasipoisson, se obtiene que el parámetro de dispersión es de 3.31293 (menor que la del modelo para `start`), indicando una sobre dispersión. Igualmente se obtienen los mismos estimadores de los coeficientes del modelo con un intervalo de confianza al 95% más amplio (veáse tabla 15) como consecuencia, el coeficiente estimado para el nivel de exposición 3 al arsénico resulta no ser significativo.

Tabla 15: Estimadores de Regressión QuasiPoisson (arsenic).

Variable	IRR[1]	95% CI[1]	p-value
(Intercept)	0.00161	0.00119, 0.00213	<0.001
arsenic			
1	—	—	
2	2.08395	1.15455, 3.58003	0.012
3	2.03434	0.90845, 4.01967	0.061
4	4.08421	2.10103, 7.37576	<0.001

[1] IRR = Razón de tasas de incidencia, CI = Intervalo de confianza

4. Utilizando regresión Poisson verifique si existe interacción entre el tiempo de inicio en el empleo y los niveles de exposición de arsénico.

El modelo de regresión Poisson es:

$$\begin{aligned}\log \mu_i &= X_i^T \beta \\ &= \log \text{personyrs}_i + \beta_0 + \beta_1 \cdot \text{start} + \beta_{2l} \cdot \text{arsenic} + \beta_{3l} \cdot \text{start} * \text{arsenic}\end{aligned}$$

Donde:

rate $= \frac{\text{respdeath}}{\text{personyrs}}$ = es la tasa de incidencia de muerte respiratoria.

$\log \text{personyrs}_i$ = Variable *offset* que ajusta por el tiempo de exposición.

β_0 = Corresponde al $\log$ de la tasa esperada en el nivel de referencia (`arsenic` = 1).

$\beta_1 =$ Coeficiente asociado a la razón de la tasa de incidencia $I\hat{R}R$ para el tiempo de inicio en el empleo (`start`).

$\beta_{2l} =$ Coeficiente asociado a la razón de la tasa de incidencia $I\hat{R}R$ para los distintos niveles de exposición ($l = 2, 3, 4$) al arsénico (`arsenic`).

$\beta_{3l} =$ Coeficiente asociado a la razón de la tasa de incidencia $I\hat{R}R$ para los distintos niveles de interacción entre el tiempo de inicio en el empleo (`start`) y el nivel de exposición al arsénico ($l = 2, 3, 4$).

Estimamos el modelo de interacción con `R`.

```
poisson.rate.inter <- glm(
    formula = respdeath ~ start*arsenic,
    family = "poisson",
    offset =  log(personyrs),
    data = df.resp.deaths)
```

Tabla 16: Estimadores de Regressión Poisson (Interacción: arsenic * start).

Variable	IRR[1]	95% CI[1]	p-value
(Intercept)	0.00067	0.00041, 0.00108	<0.001
start	2.00270	1.41941, 2.79100	<0.001
arsenic			
1	—	—	
2	1.81760	0.71373, 4.58525	0.2
3	1.90768	0.51954, 6.37911	0.3
4	1.72964	0.32878, 6.92450	0.5
start * arsenic			
start * 2	1.14016	0.57694, 2.19058	0.7
start * 3	0.98853	0.43239, 2.23949	>0.9
start * 4	1.37609	0.62033, 3.39683	0.5

[1] IRR = Razón de tasas de incidencia, CI = Intervalo de confianza

Al ajustar el modelo se obtiene que los estimadores de la interacción no son significativos a un nivel de $\alpha = 0.05$, por lo cual no hay evidencia estadística para hablar de una posible interacción entre el tiempo de inicio del trabajo y el nivel de exposición al arsénico. Solo un efecto principal es significativo (`start`) indicando que la razón de la tasa de incidencia es dos veces mayor cuando se empieza el empleo en años posteriores a 1925 controlando por el nivel de exposición al arsénico. Esta exposición resulta no ser significativa cuando se controla por el periodo de inicio del trabajo.

Se procede a corroborar la aplicabilidad del modelo.

```
epiDisplay::poisgof(poisson.rate.inter)
```

```
$results
[1] "Goodness-of-fit test for Poisson assumption"

$chisq
[1] 277.981

$df
[1] 106

$p.value
[1] 2.32494e-17
```

Tabla 17: Estimadores de Regressión QuasiPoisson (Interacción: arsenic * start).

Variable	IRR[1]	95% CI[1]	p-value
(Intercept)	0.00067	0.00029, 0.00151	<0.001
start	2.00270	1.09554, 3.52872	0.021
arsenic			
1	—	—	
2	1.81760	0.35778, 8.98415	0.5
3	1.90768	0.18526, 14.8249	0.6
4	1.72964	0.07751, 16.8197	0.7
start * arsenic			
start * 2	1.14016	0.34261, 3.48143	0.8
start * 3	0.98853	0.23114, 4.10759	>0.9
start * 4	1.37609	0.36327, 7.23870	0.7

[1] IRR = Razón de tasas de incidencia, CI = Intervalo de confianza

El modelo no es el más adecuado. La dispersión estimada es de 2.9637428 indicando la presencia de sobre dispersión (se estimo usando un modelo quasipoisson con la interacción incluida). Los parámetros estimados se muestran en la tabla 17. A excepción de la amplitud de los intervalos de confianza, las estimaciones e interpretaciones de este modelo no son distintas al anterior.