

# Fontecha - TP 4 Datos categóricos y medidas de asociación

María Eugenia Fontecha

17 octubre, 2020

## Ejercicio 1

Use la base de datos “*Framingham.csv*”.

El estudio de Framingham del corazón es un trabajo que se puso en marcha en 1948, estudia a los habitantes de la ciudad de Framingham abalizando sus factores de riesgo y el desarrollo de enfermedad cardiovascular.

Esta base de datos cuenta con 4699 observaciones (personas) libres de enfermedad coronaria al inicio del estudio. Para cada sujeto se recolectó la siguiente información:

- Sbp: Presión arterial sistólica
- dbp: Presión arterial diastólica
- age: edad en años
- scl: colesterol en plasma (mg%)
- bmi: índice de masa corporal
- sex: sexo, 1 para hombres, 0 para mujeres
- month: mes del de año de evaluación basal
- Id: identificador
- followup: seguimiento en días
- chdfate: ocurrencia de la enfermedad coronaria en el seguimiento: 1, presencia de enfermedad coronaria; 0, ausencia.

```
setwd(path)
framingham = read.csv('framingham.csv')
summary(framingham)
```

```
##           X           sex           sbp           dbp
## Min.      : 1    Min.    :0.0000   Min.    : 80.0   Min.    : 40.00
## 1st Qu.:1176   1st Qu.:0.0000   1st Qu.:116.0   1st Qu.: 74.00
## Median :2350   Median :0.0000   Median :130.0   Median : 80.00
## Mean     :2350   Mean     :0.4361   Mean     :132.8   Mean     : 82.54
## 3rd Qu.:3524   3rd Qu.:1.0000   3rd Qu.:144.0   3rd Qu.: 90.00
## Max.     :4699   Max.     :1.0000   Max.     :270.0   Max.     :148.00
##
##           scl           chdfate           followup           age
## Min.      :115.0   Min.      :0.0000   Min.      : 18   Min.      :30.00
## 1st Qu.:197.0   1st Qu.:0.0000   1st Qu.: 5136   1st Qu.:39.00
## Median :225.0   Median :0.0000   Median : 8908   Median :45.00
## Mean     :228.3   Mean     :0.3135   Mean     : 8061   Mean     :46.04
## 3rd Qu.:255.0   3rd Qu.:1.0000   3rd Qu.:11648   3rd Qu.:53.00
```

```
## Max.      :568.0    Max.      :1.0000    Max.      :11688    Max.      :68.00
## NA's      :33
##      bmi          month          id
## Min.      :16.20    Min.      : 1.000    Min.      : 1
## 1st Qu.   :22.80    1st Qu.   : 3.000    1st Qu.   :1176
## Median    :25.20    Median    : 6.000    Median    :2350
## Mean      :25.63    Mean      : 6.369    Mean      :2350
## 3rd Qu.   :28.00    3rd Qu.   :10.000    3rd Qu.   :3524
## Max.      :57.60    Max.      :12.000    Max.      :4699
## NA's      :9
```

**1. ¿Cuál es la pregunta pico del estudio? ¿Qué tipo de estudio es? ¿Cuál es el la variable respuesta o Outcome?**

**P:** habitantes de Framingham sin enfermedad coronaria.

**I:** no hay intervención.

**C:** presencia o no de factores de riesgo. Los factores de riesgo son: Sbp, dbp, age, scl, bmi y sex.

**O:** desarrollo o no de la enfermedad coronaria.

Se trata de un estudio de cohortes propectiva, ya que se toma a un grupo de personas, habitantes de Framingham, y se las sigue en el tiempo para ver si desarrollan o no la enfermedad y si tenían presentes los factores de riesgo. La variable respuesta es el desarrollo o no de la enfermedad coronaria.

**2. Que tipo de variables identifica.**

Las variables son:

- sex: cualitativa dicotómica
- Sbp: cuantitativa continua
- dbp: cuantitativa continua
- age: cuantitativa discreta
- scl: cuantitativa continua
- bmi: cuantitativa continua
- month: cualitativa categórica
- Id: cuantitativa discreta
- followup: cuantitativa discreta
- chdfate: cualitativa dicotómica

**3. Categorice la variable sbp en Hipertensión si o no ( si=spb > 140 y no=SBP<139)**

```
sbp <- framingham$sbp
framingham$sbp[sbp >= 140] <- 'Sí'
framingham$sbp[sbp < 140] <- 'No'

framingham$sbp = as.factor(framingham$sbp)

#Chequeo que se haya categorizado bien
str(framingham$sbp)
```

```
## Factor w/ 2 levels "No","Sí": 1 1 2 1 2 2 2 2 1 ...
```

4. Categorice la variable edad en Rango Etario 1=20 a 45 años, 2=46 a 60años, 3=60 a 75 años y 4=más de 75 años

```
age <- framingham$age
framingham$age[age >= 20 & age <= 45] <- 1
framingham$age[age >= 46 & age <= 60] <- 2
framingham$age[age >= 61 & age <= 75] <- 3
framingham$age[age > 75] <- 4

framingham$age = as.factor(framingham$age)

#Chequeo que se haya categorizado bien
str(framingham$age)
```

```
## Factor w/ 3 levels "1","2","3": 2 2 3 2 1 3 1 1 2 2 ...
```

Hay solo tres niveles porque el máximo de la variable *age* era 68, entonces no hay ningún dato en el último rango etario.

5. Categorice la variable scl en Hipercolesterolemia si o no ( si = scl>200 y no = scl <199)

```
scl <- framingham$scl
framingham$scl[scl >= 200] <- 'Sí'
framingham$scl[scl < 200] <- 'No'

framingham$scl = as.factor(framingham$scl)

#Chequeo que se haya categorizado bien
str(framingham$scl)
```

```
## Factor w/ 2 levels "No","Sí": 2 1 2 2 2 1 2 2 2 2 ...
```

6. Categorice la variable bmi en obesidad si o no ( si = bmi>30 y no bmi<30)

```
bmi <- framingham$bmi
framingham$bmi[bmi >= 30] <- 'Sí'
framingham$bmi[bmi < 30] <- 'No'

framingham$bmi = as.factor(framingham$bmi)

#Chequeo que se haya categorizado bien
str(framingham$bmi)
```

```
## Factor w/ 2 levels "No","Sí": 1 1 1 1 1 2 1 1 1 2 ...
```

7. Realice las tablas de frecuencia para cada una de las variables (Hipertensión, rango etario hipercolesterolemia, obesidad y sexo). Analice si existe asociación entre las mismas y el desarrollo de enfermedad coronaria. (platee el test de hipotesis para cada una, valide los supuestos y realice el test)

El Test de Chi Cuadrado compara las frecuencias observadas para una variable contra las esperadas suponiendo que la variable perteneciente al factor de riesgo y la variable correspondiente al evento son independientes. Si existe alguna de las frecuencias observadas que se aleja de su correspondiente frecuencia esperada, entonces se rechaza la hipótesis nula y se concluye que las variables no son independientes. El Test no da información sobre qué frecuencia se aleja de la esperada.

Los supuestos del test son:

- más del 80% de las celdas tienen una frecuencia esperada mayor o igual a 5 y ninguna frecuencia esperada menor a 1
- las frecuencias marginales de la tabla son conocidas.

## Hipertensión

```
tabla_sbp <- table(framingham$sbp, framingham$chdfate)
kable(tabla_sbp, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	0	1
No	2356	873
Sí	870	600

Las hipótesis del Test de Chi Cuadrado para la hipertensión son:

- $H_0$ : la hipertensión y el desarrollo de la enfermedad coronaria son variables independientes
- $H_1$ : la hipertensión y el desarrollo de la enfermedad coronaria no son variables independientes.

Se cumplen los supuestos del test ya que todas las celdas tiene una frecuencia mayor a 5, y bajo el supuesto de que las variables son independientes, las frecuencias esperadas son iguales a las observadas, y se conocen las frecuencias marginales.

```
#Test de Chi Cuadrado, con corrección de continuidad (porque la tabla es de 2x2)
chi_sbp <- chisq.test(tabla_sbp)
chi_sbp
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tabla_sbp
## X-squared = 88.492, df = 1, p-value < 2.2e-16
```

Como el p-valor da menor a 0.05 se rechaza la hipótesis nula, por lo que la hipertensión y la enfermedad coronaria no son independientes.

## Rango etario

```
tabla_age <- table(framingham$age, framingham$chdfate)
kable(tabla_age, row.names = TRUE, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	0	1
1	1780	600
2	1305	781
3	141	92

Las hipótesis del Test de Chi Cuadrado para el rango etario son:

- $H_0$ : el rango etario y el desarrollo de la enfermedad coronaria son variables independientes
- $H_1$ : el rango etario y el desarrollo de la enfermedad coronaria no son variables independientes.

Se cumplen los supuestos del test de Chi Cuadrado, ya que todas las frecuencias son mayores a 5.

```
chi_age <- chisq.test(tabla_age)
chi_age
```

```
##
## Pearson's Chi-squared test
##
## data:  tabla_age
## X-squared = 84.807, df = 2, p-value < 2.2e-16
```

Las variables rango etario y desarrollo de enfermedad coronaria no son independientes, ya que con un valor-P menor a 0.05 se rechaza la hipótesis nula.

## Hipercolesterolemia

```
tabla_scl <- table(framingham$scl, framingham$chdfate)
kable(tabla_scl, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	0	1
No	979	258
Sí	2221	1208

Las hipótesis del Test de Chi Cuadrado para la hipercolesterolemia son:

- $H_0$ : la hipercolesterolemia y el desarrollo de la enfermedad coronaria son variables independientes
- $H_1$ : la hipercolesterolemia y el desarrollo de la enfermedad coronaria no son variables independientes.

Realizo el test d Chi Cuadrado ya que se cumplen los supuestos:

```
chi_scl <- chisq.test(tabla_scl)
chi_scl
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tabla_scl
## X-squared = 86.477, df = 1, p-value < 2.2e-16
```

Como el test arrojó un p-valor menor a 0.05 se concluye que la hipercolesterolemia y la enfermedad coronaria no son variables independientes.

## Obesidad

```
tabla_bmi <- table(framingham$bmi, framingham$chdfate)
kable(tabla_bmi, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	0	1
No	2865	1224
Sí	353	248

Las hipótesis del Test de Chi Cuadrado para la obesidad son:

- $H_0$ : la obesidad y el desarrollo de la enfermedad coronaria son variables independientes
- $H_1$ : la obesidad y el desarrollo de la enfermedad coronaria no son variables independientes.

Se cumplen los supuestos del test de Chi Cuadrado ya que todas las celdas de la tabla son mayores a 5.

```
chi_bmi <- chisq.test(tabla_bmi)
chi_bmi
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tabla_bmi
## X-squared = 30.714, df = 1, p-value = 2.991e-08
```

La obesidad y el desarrollo de enfermedad coronaria no son independientes, ya que se obtuvo un p-valor menor a 0.05, por lo que se rechaza  $H_0$ .

## Sexo

```
tabla_sex <- table(framingham$sex, framingham$chdfate)
kable(tabla_sex, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	0	1
0	2000	650
1	1226	823

Las hipótesis del Test de Chi Cuadrado para el sexo son:

- $H_0$ : el sexo y el desarrollo de la enfermedad coronaria son variables independientes
- $H_1$ : el sexo y el desarrollo de la enfermedad coronaria no son variables independientes.

Los supuestos del test de Chi cuadrado también se cumplen en este caso.

```
chi_sex <- chisq.test(tabla_sex)
chi_sex
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tabla_sex
## X-squared = 130.58, df = 1, p-value < 2.2e-16
```

El sexo y la enfermedad coronaria no son variables independientes.

#### 8. Calcule los riesgos relativos para cada una de las variables con sus intervalos de confianza.

El riesgo relativo se calcula como:  $RR = \frac{\text{Probabilidad en los expuestos}}{\text{Probabilidad en los no expuestos}}$

El intervalo de confianza del riesgo relativo se calcula como:

- $\ln(L) = \ln RR - z_{\frac{\alpha}{2}} \cdot ES(\ln RR)$
- $\ln(U) = \ln RR + z_{\frac{\alpha}{2}} \cdot ES(\ln RR),$

donde  $ES(\ln RR) = \sqrt{\frac{1}{a} - \frac{1}{n_1} + \frac{1}{c} - \frac{1}{n_2}},$

siendo  $a$  la cantidad de expuestos con el evento,  $n_1$  el total de expuestos,  $c$  la cantidad de no expuestos con el evento y  $n_2$  el total de no expuestos.

```
riesgo_relativo <- function(tabla, FR)
{
  z <- 1.96 #z de alpha/2 para un nivel de confianza del 95%
  a = tabla[2,2] #con factor de riesgo y evento
  n1 = sum(tabla[2,]) #total de individuos con factor de riesgo
  c = tabla[1,2] #sin factor de riesgo y con evento
  n2 = sum(tabla[1,]) #total de individuos sin factor de riesgo
```

```

RR = (a/n1)/(c/n2) #riesgo relativo
ES = sqrt(1/a - 1/n1 + 1/c - 1/n2)

L = exp(log(RR) - z*ES) #límite inferior del IC
U = exp(log(RR) + z*ES) #límite superior del IC

df <- data.frame(Lower= c(L) , Relative_Risk= c(RR),Upper= c(U),
                 row.names = c(FR))

tabla_RR <- kable(df, booktabs = T) %>%
  kable_styling(latex_options = "striped")

}

```

## Hipertensión

```

RR_sbp <- riesgo_relativo(tabla_sbp, 'Hipertensión')
RR_sbp

```

	Lower	Relative_Risk	Upper
Hipertensión	1.388519	1.50969	1.641435

La hipertensión aumenta en un 50% el riesgo de sufrir una enfermedad coronaria, con un intervalo de confianza al 95% entre 38.85% y 64.14%. Esta asociación es estadísticamente significativa, con un p-Valor del test de Chi Cuadrado menor a 0.05.

## Rango etario

Creo las variables dummy para los tres rangos etarios.

```

dummies_age <- dummy.data.frame(data = framingham, names = 'age')

```

```

tabla_age1 <- table(dummies_age$age1, dummies_age$chdfate)

```

```

tabla_age2 <- table(dummies_age$age2, dummies_age$chdfate)

```

```

tabla_age3 <- table(dummies_age$age3, dummies_age$chdfate)

```

```

RR_age1 <- riesgo_relativo(tabla_age1, 'Rango Etario 1')
RR_age1

```

	Lower	Relative_Risk	Upper
Rango Etario 1	0.6140003	0.6696699	0.730387

Tener de 20 a 45 años disminuye en un 33% las chances de desarrollar una enfermedad coronaria, con un intervalo de confianza al 95% entre 27% y 39%. Esta asociación es estadísticamente significativa ya que se obtuvo un p-valor menor a 0.05 en el test de Chi Cuadrado.



```
RR_age2 <- riesgo_relativo(tabla_age2, 'Rango Etario 2')
RR_age2
```

	Lower	Relative_Risk	Upper
Rango Etario 2	1.299049	1.413742	1.538561

El rango etario 2, que abarca desde los 46 hasta los 60 años, aumenta 1.41 veces las chances de desarrollar una enfermedad coronaria. El intervalo de confianza al 95% es de 1.29 a 1.54 y la asociación es estadísticamente significativa, con un p-valor menor a 0.05.

```
RR_age3 <- riesgo_relativo(tabla_age3, 'Rango Etario 3')
RR_age3
```

	Lower	Relative_Risk	Upper
Rango Etario 3	1.082789	1.2769	1.50581

Tener entre 60 y 75 años aumenta en un 28% el riesgo de desarrollar una enfermedad coronaria, con un intervalo de confianza entre 0.8% y 50%. La asociación es estadísticamente significativa, con un p-valor menor a 0.05.

## Hipercolesterolemia

```
RR_scl <- riesgo_relativo(tabla_scl, 'Hipercolesterolemia')
RR_scl
```

	Lower	Relative_Risk	Upper
Hipercolesterolemia	1.501585	1.689077	1.89998

La hipercolesterolemia aumenta las chances de desarrollar una enfermedad coronaria en un 69%, con un intervalo de confianza de 50% a 90%. Esta asociación es significativa, con un p-valor menor a 0.05.

## Obesidad

```
RR_bmi <- riesgo_relativo(tabla_bmi, 'Obesidad')
RR_bmi
```

	Lower	Relative_Risk	Upper
Obesidad	1.239516	1.378519	1.533111

La obesidad aumenta 1.38 veces el riesgo de sufrir una enfermedad coronaria, con un intervalo de confianza al 95% entre 1.24 y 1.53. Esta asociación es estadísticamente significativa, con un p-Valor del test de Chi Cuadrado menor a 0.05.

## Sexo

```
RR_sex <- riesgo_relativo(tabla_sex, 'Sexo (masculino)')  
RR_sex
```

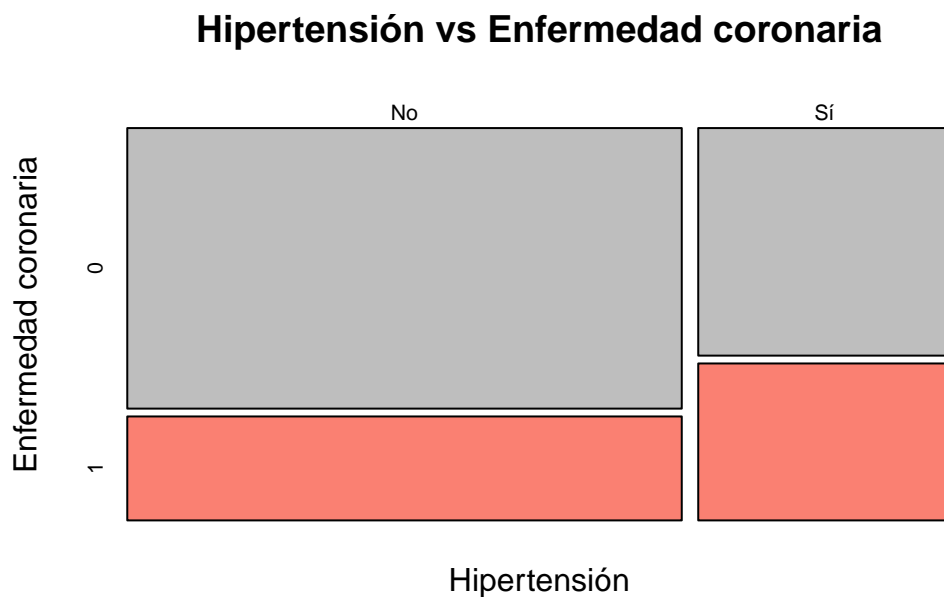
	Lower	Relative_Risk	Upper
Sexo (masculino)	1.503844	1.637534	1.783109

El sexo masculino aumenta el riesgo de desarrollar una enfermedad coronaria un 64%, con un intervalo de confianza al 95% entre 50.38% y 78.31%, y siendo la asociación estadísticamente significativa (p-valor menor a 0.05).

### 9. Realice un gráfico de mosaicos para cada comparación

## Hipertensión

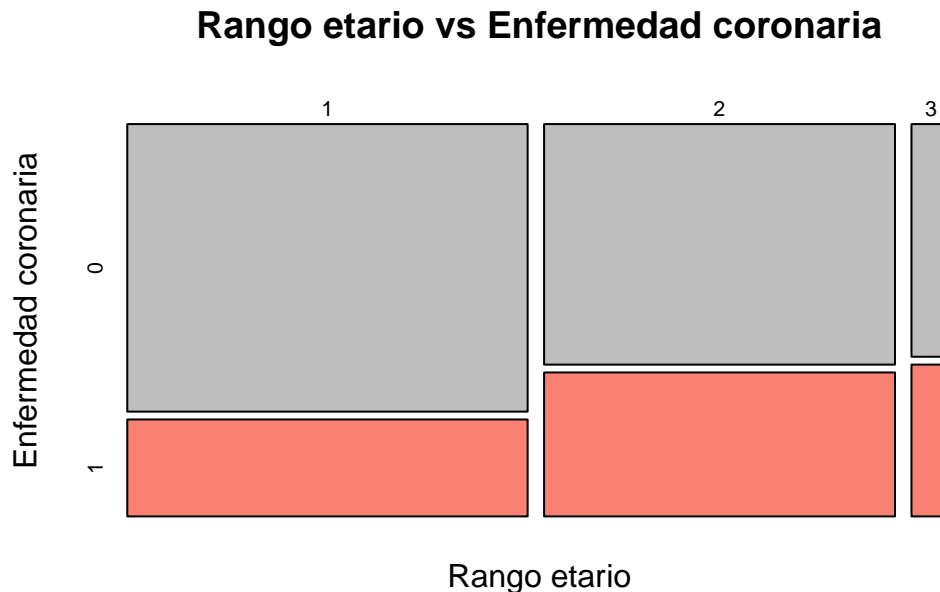
```
mosaicplot(tabla_sbp, main = 'Hipertensión vs Enfermedad coronaria', xlab = 'Hipertensión',  
            , ylab = 'Enfermedad coronaria', color = c('grey', 'salmon'))
```



Vemos que para el grupo con hipertensión la proporción de individuos con el evento es mayor que para el grupo sin hipertensión.

## Rango etario

```
mosaicplot(tabla_age, main = 'Rango etario vs Enfermedad coronaria', xlab = 'Rango etario',  
           , ylab = 'Enfermedad coronaria', color = c('grey', 'salmon'))
```

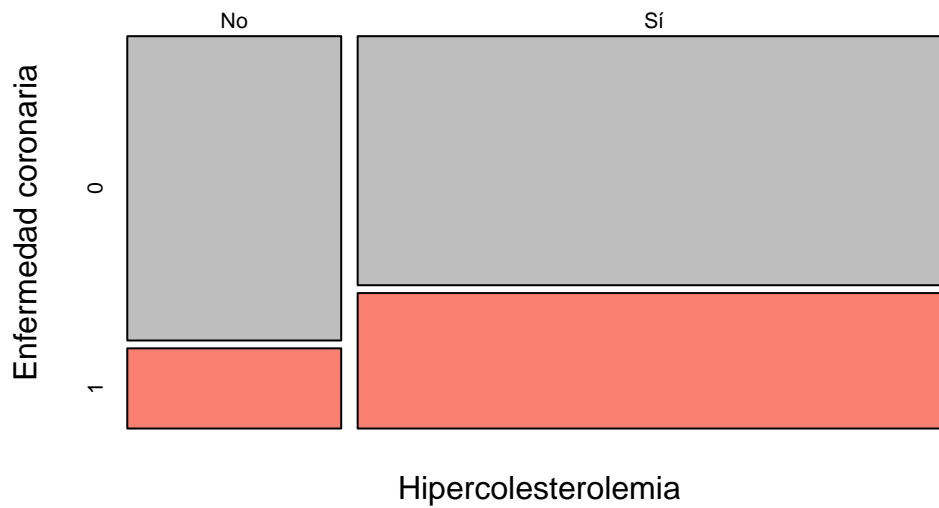


Para el caso de la edad, se ve que el grupo con mayor proporción de presencia del evento es el rango etario 3, que no se condice con los resultados de riesgo relativo que obtuve anteriormente, donde el rango etario 2 presentaba un riesgo mayor. Puede tener que ver con el hecho de que el grupo etario 3 tiene una menor cantidad de individuos que los demás rangos, entonces cuando se crean las variables dummies, los rangos de mayor cantidad de individuos tendrán mayor peso. Por ejemplo, en el caso de la variable dummy 2, el rango etario 3 queda mezclado con el 1, que tiene una mayor proporción de individuos sin el evento y cuya cantidad de individuos es mayor. Entonces, el efecto que tiene el rango etario 3 sobre el evento se ve camuflado por el rango etario 1. En cambio, para el caso de la variable dummy 3, al mezclarse el rango etario 1 con el 2, los efectos se ven más compensados, por tener una cantidad de individuos similar. Por esta razón, el riesgo relativo del rango etario 3 dio menor que el del rango etario 2.

## Hipercolesterolemia

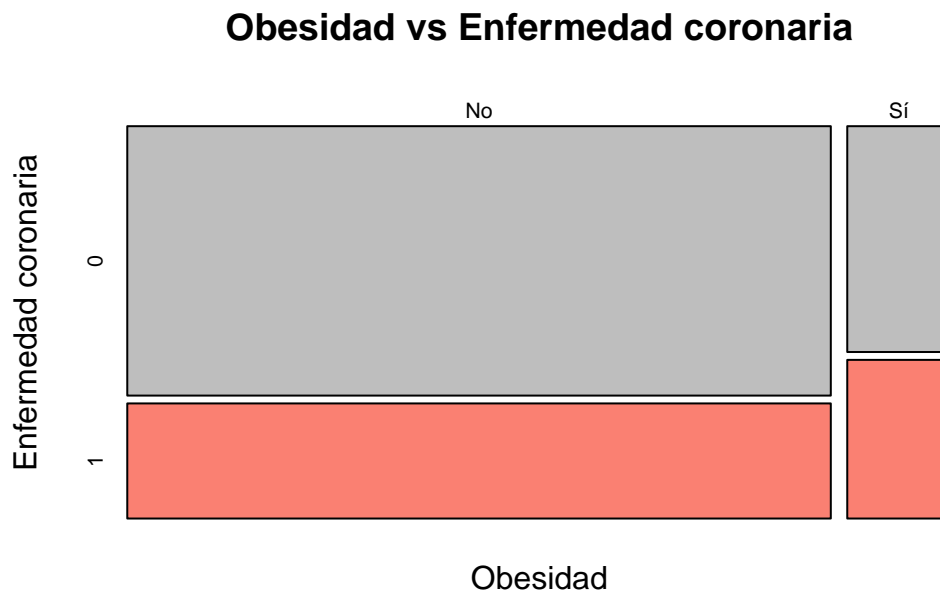
```
mosaicplot(tabla_scl, main = 'Hipercolesterolemia vs Enfermedad coronaria',  
           xlab = 'Hipercolesterolemia', ylab = 'Enfermedad coronaria',  
           color = c('grey', 'salmon'))
```

## Hipercolesterolemia vs Enfermedad coronaria



La proporción de eventos es mayor en el grupo de individuos con hipercolesterolemia que en el grupo sin hipercolesterolemia.

```
mosaicplot(tabla_bmi, main = 'Obesidad vs Enfermedad coronaria', xlab = 'Obesidad',  
           ylab = 'Enfermedad coronaria', color = c('grey', 'salmon'))
```

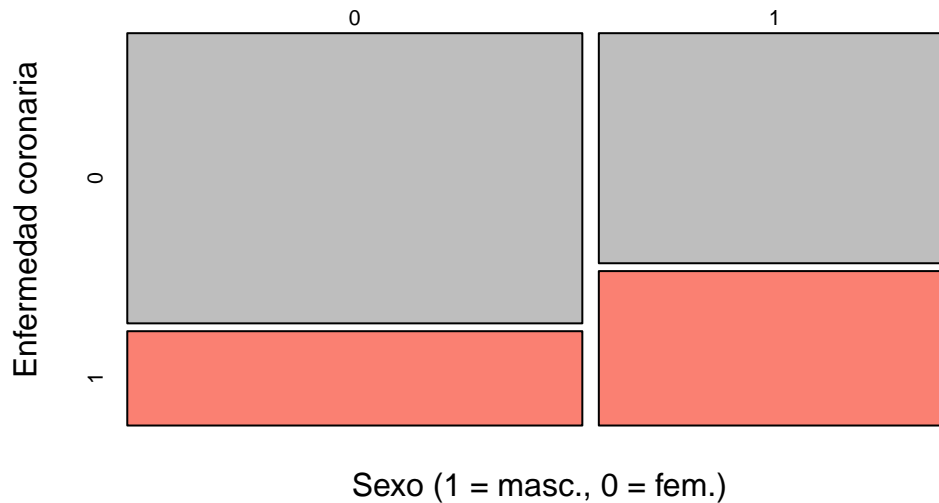


En el grupo con obesidad se ve una mayor proporción de individuos que desarrollaron enfermedad coronaria que en el grupo sin obesidad.

**Sexo**

```
mosaicplot(tabla_sex, main = 'Sexo vs Enfermedad coronaria',
           xlab = 'Sexo (1 = masc., 0 = fem.)',
           ylab = 'Enfermedad coronaria', color = c('grey', 'salmon'))
```

## Sexo vs Enfermedad coronaria



El sexo masculino tiene mayor proporción de individuos con enfermedad coronaria que el grupo femenino.

### 10. Analice los residuos y residuos estandarizados.

Los residuos nos dicen qué celda se aparta de la independencia, y cuanto mayor sea el valor del residuo, mayor es la probabilidad de que la asociación sea significativa.

## Hipertensión

Residuos

```
kable(chi_sbp$observed - chi_sbp$expected, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	0	1
No	139.1977	-139.1977
Sí	-139.1977	139.1977

Residuos estandarizados

```
kable(chi_sbp$residuals, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	0	1
No	2.956437	-4.375215
Sí	-4.381713	6.484474

La celda que más se aleja de la independencia es la correspondiente a hipertensión y enfermedad coronaria ya que es la que tiene el residuo estandarizado mayor. Esto significa que hay una asociación entre la hipertensión y la enfermedad coronaria.

## Rango etario

Residuos

```
kable(chi_age$observed - chi_age$expected, row.names = TRUE, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	0	1
1	146.06086	-146.06086
2	-127.09960	127.09960
3	-18.96127	18.96127

Residuos estandarizados

```
kable(chi_age$residuals, row.names = TRUE, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	0	1
1	3.613399	-5.347450
2	-3.358594	4.970366
3	-1.499201	2.218660

El residuo más alto corresponde al rango etario 2 y presencia del evento, por lo que hay una asociación entre las edades de 46 a 60 y la enfermedad coronaria. La celda rango etario 1-evento tiene el residuo más negativo, y el de mayor el módulo, por lo que hay una asociación entre este rango etario y la enfermedad coronaria, siendo el rango etario 1 un factor de prevención.

## Hipercolesterolemia

Residuos

```
kable(chi_scl$observed - chi_scl$expected, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	0	1
No	130.6502	-130.6502
Sí	-130.6502	130.6502

Residuos estandarizados

```
kable(chi_scl$residuals, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	0	1
No	4.485622	-6.627210
Sí	-2.694163	3.980447

Las celdas que más se alejan de la independencia son las de ausencia de hipercolesterolemia. Como el residuo es negativo para el caso del evento, la ausencia de hipercolesterolemia es un factor de prevención.

## Obesidad

Residuos

```
kable(chi_bmi$observed - chi_bmi$expected, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	0	1
No	59.37058	-59.37058
Sí	-59.37058	59.37058

Residuos estandarizados

```
kable(chi_bmi$residuals, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	0	1
No	1.120872	-1.657277
Sí	-2.923664	4.322814

La celda de obesidad y presencia del evento es la que más se aleja de la independencia ya que tiene el mayor residuo estandarizado, por lo que hay una asociación entre la obesidad y la enfermedad coronaria.

## Sexo

Residuos

```
kable(chi_sex$observed - chi_sex$expected, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	0	1
0	180.698	-180.698
1	-180.698	180.698

Residuos estandarizados



```
kable(chi_sex$residuals, booktabs = T) %>%
  kable_styling(latex_options = "striped")
```

	0	1
0	4.236439	-6.269484
1	-4.817846	7.129905

El mayor residuo estandarizado se da en la celda sexo masculino-enfermedad coronaria, por lo que hay una asociación entre estas variables, siendo el sexo masculino un factor de riesgo.

#### 11. ¿Cómo interpreta los resultados para cada variable?

### Hipertensión

La hipertensión es un factor de riesgo para la enfermedad coronaria.

### Rango etario

Los rangos etarios 2 y 3 son factores de riesgo para la enfermedad coronaria, mientras que el 1 es un factor de prevención o protección. La mayoría de los resultados indicarían que el rango etario 2 tiene mayor riesgo de desarrollar la enfermedad, pero pienso que esto tiene que ver con un desequilibrio en la cantidad de individuos del rango etario 3 vs. los de los rangos 1 y 2.

```
summary(framingham$age)
```

```
##      1      2      3
## 2380 2086  233
```

### Hipercolesterolemia

La hipercolesterolemia es un factor de riesgo para la enfermedad coronaria.

### Obesidad

La obesidad es un factor de riesgo para la enfermedad coronaria.

### Sexo

El sexo masculino es un factor de riesgo para la enfermedad coronaria.