SESION 4 R

***Sesión 4: Estadística descriptiva***

**Contenido**

Medidas de tendencia central y dispersión. Tablas bidimensionales en R.

**Actividad 4:**

Aplicaciones en R utilizando los datos de esta sesión.

**Estadística descriptiva con dplyr**

La función glimpse del paquete **dplyr** nos da un resumen de las variables de la base de datos.

glimpse(dt)

## Rows: 1,236

## Columns: 7

## $ bwt <int> 120, 113, 128, 123, 108, 136, 138, 132, 120, 143, 140, 144, …

## $ gestation <int> 284, 282, 279, NA, 282, 286, 244, 245, 289, 299, 351, 282, 2…

## $ parity <chr> "First born", "First born", "First born", "First born", "Fir…

## $ age <int> 27, 33, 28, 36, 23, 25, 33, 23, 25, 30, 27, 32, 23, 36, 30, …

## $ height <int> 62, 64, 64, 69, 67, 62, 62, 65, 62, 66, 68, 64, 63, 61, 63, …

## $ weight <int> 100, 135, 115, 190, 125, 93, 178, 140, 125, 136, 120, 124, 1…

## $ smoke <chr> "Not", "Not", "Yes", "Not", "Yes", "Not", "Not", "Not", "Not…

Vamos a extraer una sola variable con pull y luego a calcular la media, varianza y desviación estándar del peso de los bebés.

dt |> pull(bwt) |> mean()

## [1] 119.5769

dt |> pull(bwt) |> var()

## [1] 332.5682

dt |> pull(bwt) |> sd()

## [1] 18.23645

dt |> pull(bwt) |> quantile(probs=c(0.25, 0.40, 0.90))

## 25% 40% 90%

## 108.75 116.00 142.00

Vamos a calcular las mismas medidas de resumen anteriores pero con escritura secuencial usando el operador pipe |> y la función summarise.

dt |> summarise(media\_peso = mean(bwt),

varianza\_peso = var(bwt),

desvi\_peso = sd(bwt),

Q1 = quantile(bwt, probs=0.25),

D4 = quantile(bwt, probs=0.40),

P90 = quantile(bwt, probs=0.90))

## media\_peso varianza\_peso desvi\_peso Q1 D4 P90

## 1 119.5769 332.5682 18.23645 108.75 116 142

Vamos a ver cómo cambian las medidas anteriores cuando diferenciamos por el tipo de embarazo.

dt |> group\_by(parity) |>

summarise(media\_peso = mean(bwt),

varianza\_peso = var(bwt),

desvi\_peso = sd(bwt),

Q1 = quantile(bwt, probs=0.25),

D4 = quantile(bwt, probs=0.40),

P90 = quantile(bwt, probs=0.90),

n = n())

## # A tibble: 2 × 8

## parity media\_peso varianza\_peso desvi\_peso Q1 D4 P90 n

## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int>

## 1 First born 120. 345. 18.6 109 116 143 921

## 2 Unknown 118. 294. 17.1 107 114 139 315

Vamos a ver como cambiar las medidas anteriores cuando diferenciamos por fumar.

dt |> group\_by(smoke) |>

summarise(media\_peso = mean(bwt),

varianza\_peso = var(bwt),

desvi\_peso = sd(bwt),

Q1 = quantile(bwt, probs=0.25),

D4 = quantile(bwt, probs=0.40),

P90 = quantile(bwt, probs=0.90),

n = n())

## # A tibble: 3 × 8

## smoke media\_peso varianza\_peso desvi\_peso Q1 D4 P90 n

## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int>

## 1 Not 123. 303. 17.4 113 119 144 742

## 2 Yes 114. 328. 18.1 102 109. 138 484

## 3 <NA> 127. 476. 21.8 110. 122. 152. 10

Vamos a ver como cambiar las medidas anteriores cuando diferenciamos por el tipo de embarazo y fumar.

dt |> group\_by(parity, smoke) |>

summarise(media\_peso = mean(bwt),

varianza\_peso = var(bwt),

desvi\_peso = sd(bwt),

Q1 = quantile(bwt, probs=0.25),

D4 = quantile(bwt, probs=0.40),

P90 =quantile(bwt, probs=0.90),

n = n())

## `summarise()` has grouped

## output by 'parity'. You can

## override using the `.groups`

## argument.

## # A tibble: 5 × 9

## # Groups: parity [2]

## parity smoke media\_peso varianza\_peso desvi\_peso Q1 D4 P90 n

## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int>

## 1 First born Not 124. 308. 17.6 113 120 145 548

## 2 First born Yes 114. 345. 18.6 101 110 139. 363

## 3 First born <NA> 127. 476. 21.8 110. 122. 152. 10

## 4 Unknown Not 121. 284. 16.8 112 117 140 194

## 5 Unknown Yes 113. 275. 16.6 102 109 131 121

Calculando la correlación entre peso bebe y peso madre.

dt |> select(bwt, weight) |> cor()

## bwt weight

## bwt 1 NA

## weight NA 1

Como hay NA en alguna de las variables vamos a agregar use='complete.obs' para obtener la correlación.

dt |> select(bwt, weight) |> cor(use='complete.obs')

## bwt weight

## bwt 1.0000000 0.1540382

## weight 0.1540382 1.0000000

De la salida anterior vemos que la correlación lineal de Pearson entre ambas variables es 0.1541.

**14.5 Creando nuevas variables**

En esta sección vamos a mostrar cómo crear nuevas variables a partir de variables ya existentes en la base de datos.

Vamos a transformar la altura y el peso de la madre al [sistema internacional de unidades](https://es.wikipedia.org/wiki/Sistema_Internacional_de_Unidades).

dt |> mutate(height\_mt = height \* 0.0254,

weight\_kg = weight \* 0.453592) -> dt

head(dt, n=4)

## bwt gestation parity age height weight smoke height\_mt weight\_kg

## 1 120 284 First born 27 62 100 Not 1.5748 45.35920

## 2 113 282 First born 33 64 135 Not 1.6256 61.23492

## 3 128 279 First born 28 64 115 Yes 1.6256 52.16308

## 4 123 NA First born 36 69 190 Not 1.7526 86.18248

Ahora vamos a calcular el IMC con las nuevas variables así:

dt |> mutate(imc = weight\_kg / height\_mt^2) -> dt

dt |> head(n=4)

## bwt gestation parity age height weight smoke height\_mt weight\_kg imc

## 1 120 284 First born 27 62 100 Not 1.5748 45.35920 18.29004

## 2 113 282 First born 33 64 135 Not 1.6256 61.23492 23.17244

## 3 128 279 First born 28 64 115 Yes 1.6256 52.16308 19.73949

## 4 123 NA First born 36 69 190 Not 1.7526 86.18248 28.05778

De la salida anterior vemos la nueva base de datos con las tres nuevas variables creadas con mutate.

**14.6 Tablas de frecuencia**

Vamos a una la variable de clasificación según peso. Recien nacidos con pesos inferiores a 100 onzas son considerados livianos, aquellos con pesos entre 100 (inclusive) y 140 son aceptables y con pesos de 140 o más se consideran sanos.

dt |> mutate(clasi = cut(bwt,

breaks=c(-Inf, 100, 140, Inf),

labels=c("Liviano","Aceptable","Sano"),

include.lowest = TRUE)) -> dt

Tabla de frecuencia absoluta para clasificación.

dt |> pull(clasi) |>

table()

##

## Liviano Aceptable Sano

## 174 917 145

Tabla de frecuencia relativa para clasificación.

dt |> pull(clasi) |>

table() |>

prop.table() |>

round(digits=4)

##

## Liviano Aceptable Sano

## 0.1408 0.7419 0.1173

Tabla de frecuencia absoluta para fumar y clasificación.

dt |> select(smoke, clasi) |>

table() |>

addmargins()

## clasi

## smoke Liviano Aceptable Sano Sum

## Not 62 581 99 742

## Yes 111 331 42 484

## Sum 173 912 141 1226

Tabla de frecuencia relativa para fumar y clasificación.

dt |> select(smoke, clasi) |>

table() |>

prop.table()

## clasi

## smoke Liviano Aceptable Sano

## Not 0.05057096 0.47389886 0.08075041

## Yes 0.09053834 0.26998369 0.03425775

Tabla de frecuencia relativa por **columnas** para fumar y clasificación.

dt |> select(smoke, clasi) |>

table() |>

prop.table(margin=2)

## clasi

## smoke Liviano Aceptable Sano

## Not 0.3583815 0.6370614 0.7021277

## Yes 0.6416185 0.3629386 0.2978723

**14.7 Gráficos**

En esta sección se mostrará como crear algunos gráficos básicos. Para más detalles de como construir otros gráficos se recomienda consultar el libro [Gráficos con R](https://fhernanb.github.io/Graficos-con-R/).

Boxplot para el peso de los bebés.

boxplot(dt$bwt, las=1, col="darkorchid1",

ylab="Peso del bebe (onzas)")

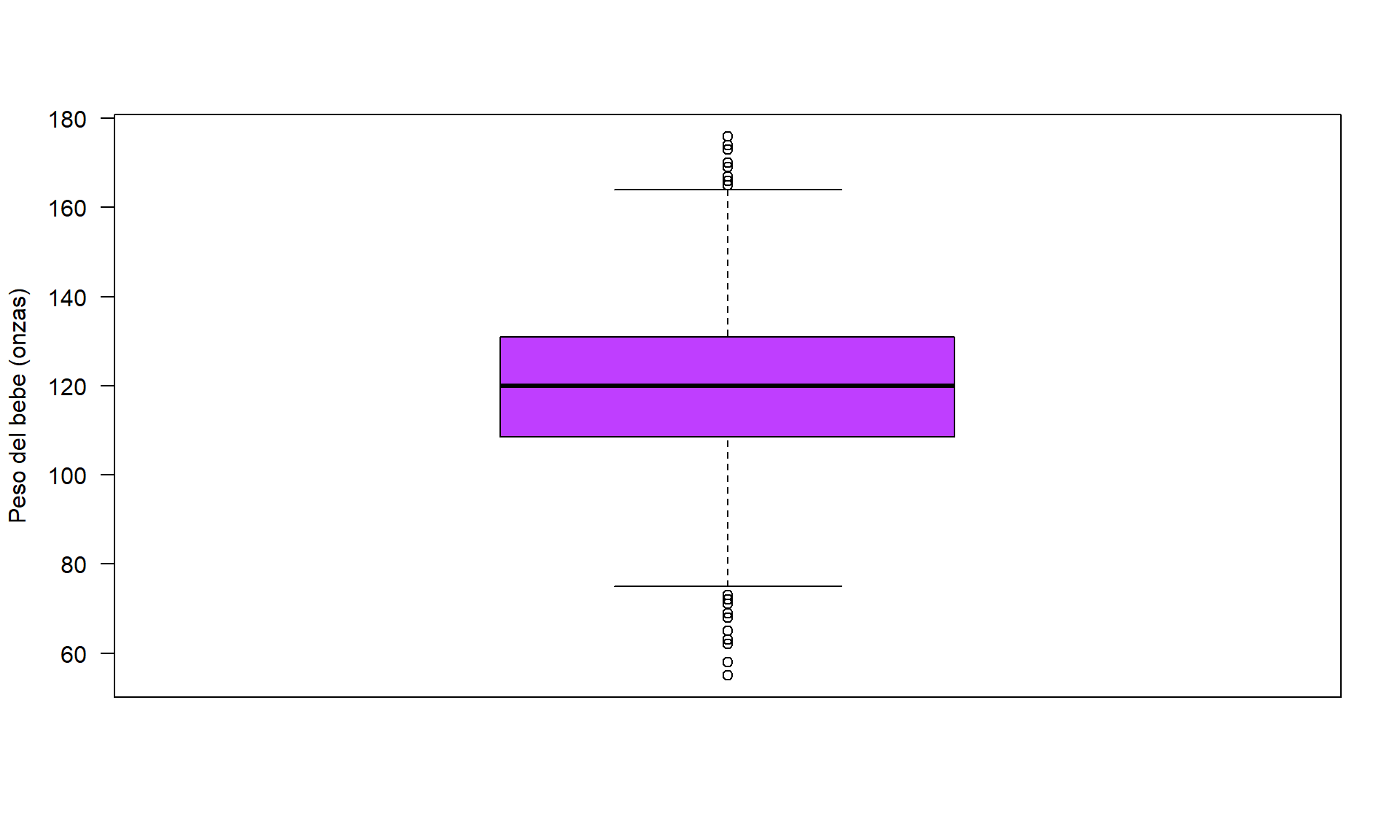


Figure 14.1: Boxplot para el peso de los bebés.

Boxplot horizontal para el peso de los bebés.

boxplot(dt$bwt, las=1, col="darkseagreen2",

horizontal=TRUE,

xlab="Peso del bebe (onzas)")

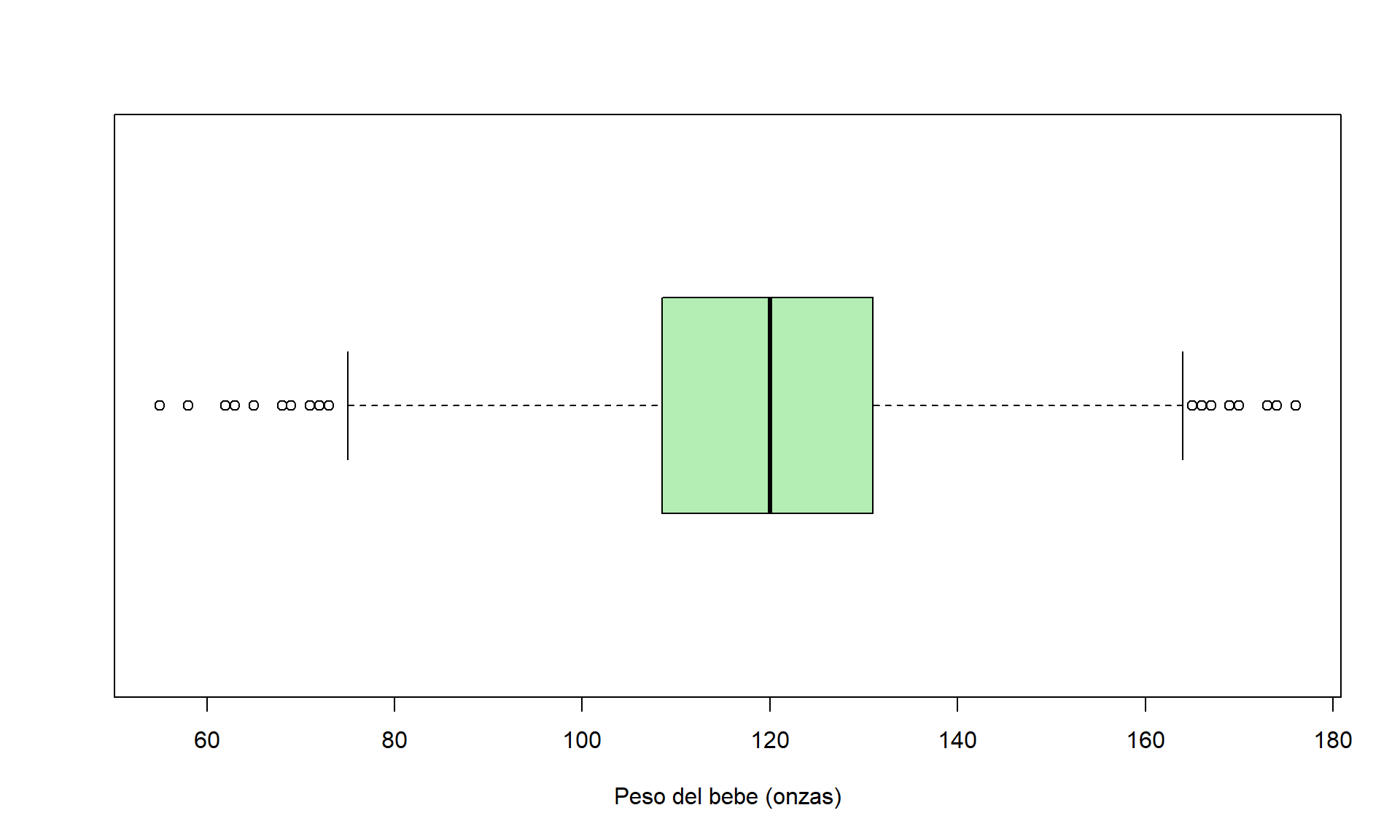


Figure 14.2: Boxplot horizontal para el peso de los bebés.

Boxplot para el peso de los bebés diferenciando por la variable smoke.

boxplot(dt$bwt ~ dt$smoke, las=1)

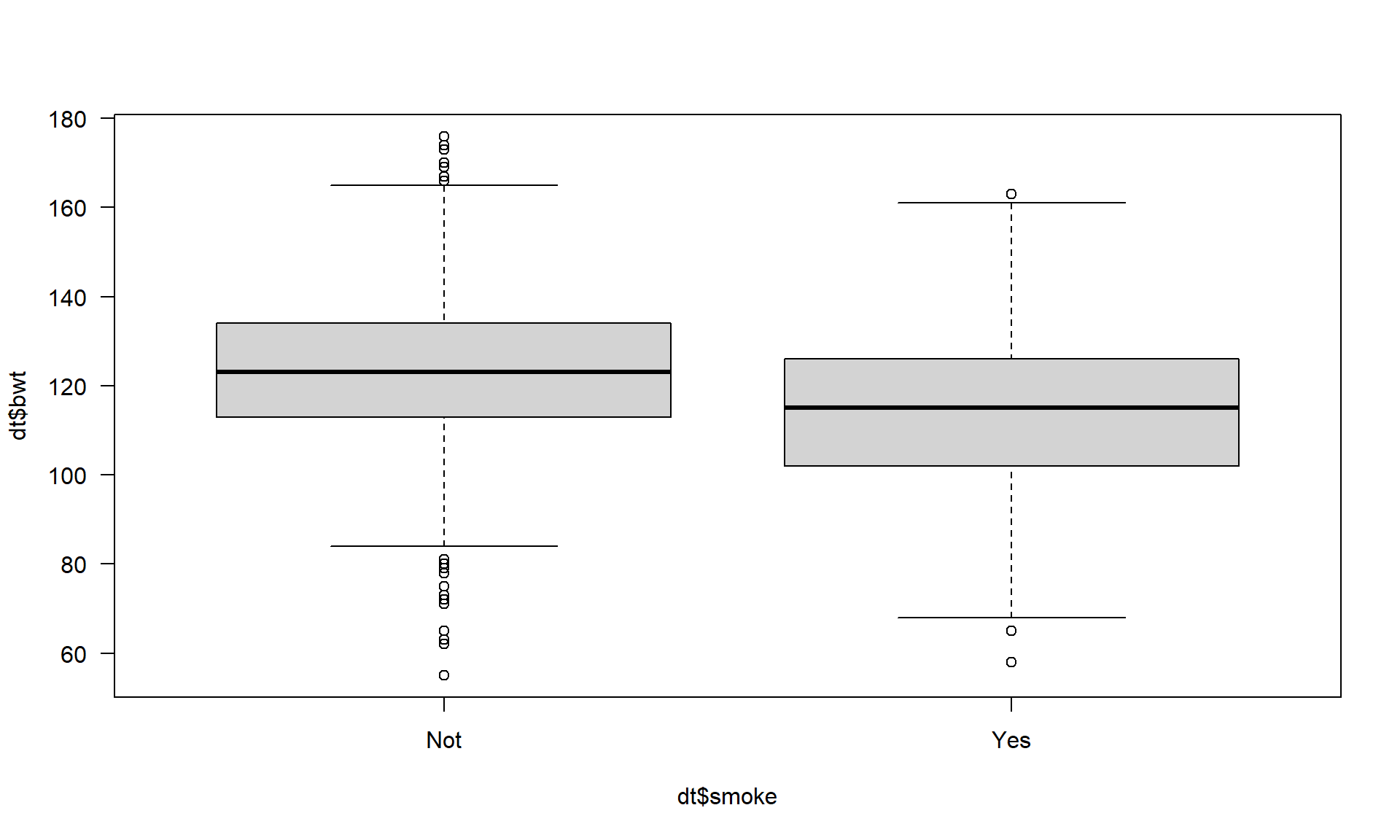


Figure 14.3: Boxplot para el peso diferenciando por fumar.

Editando el boxplot para obtener una mejor apariencia.

boxplot(dt$bwt ~ dt$smoke, las=1,

col=c("dodgerblue", "hotpink2"),

main='Babies weight given smoke status',

xlab='Does the mother smoke?',

ylab='Weight (ounces)')

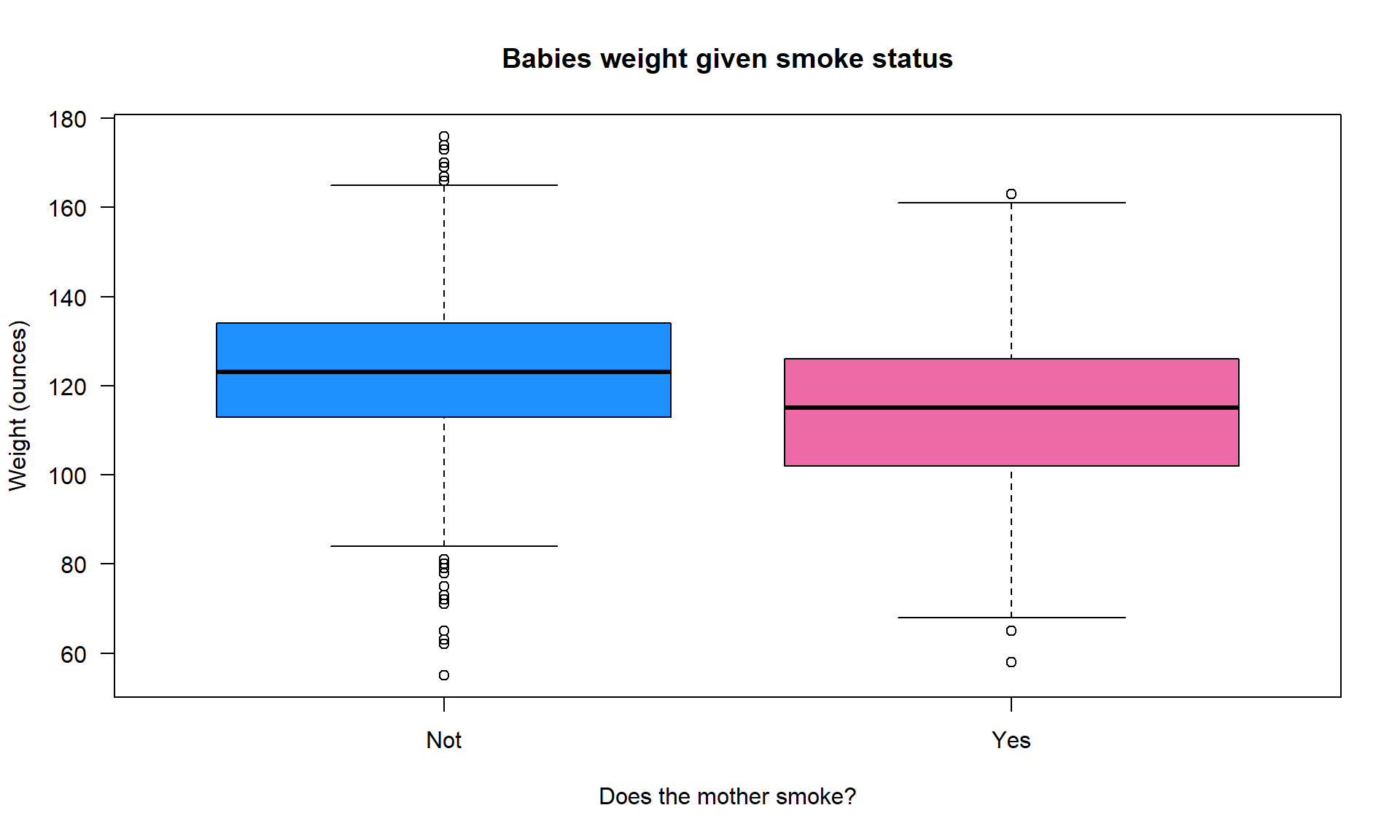


Figure 14.4: Boxplot mejorado para el peso diferenciando por fumar.

¿Como cambiar Not y Yes de smoke para colocar el gráfico en español?

dt$fuma <- ifelse(dt$smoke == 'Not', 'No', 'Si')

Boxplot en español

boxplot(dt$bwt ~ dt$fuma, las=1,

xlab='¿Fuma la madre?',

ylab='Peso del bebe (onzas)', las=1)

## Warning in (function (z, notch = FALSE, width = NULL, varwidth = FALSE, :

## Duplicated argument las = 1 is disregarded

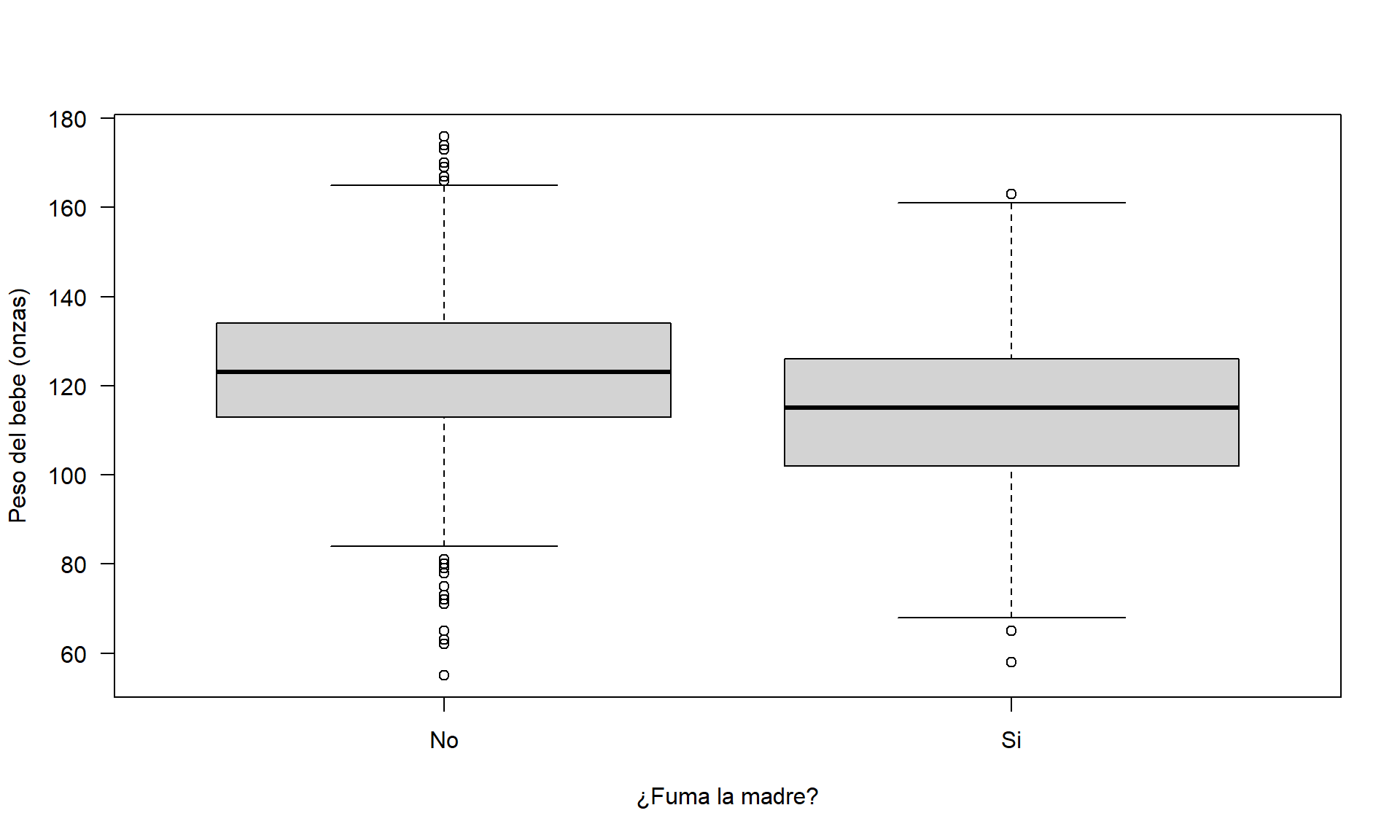


Figure 14.5: Boxplot para el peso de los bebés en español.

¿Cual es la relación entre peso del bebe y de la madre?

plot(x=dt$weight, y=dt$bwt,

xlab='Peso de la madre (libras)',

ylab='Peso del bebe (onzas)', las=1)

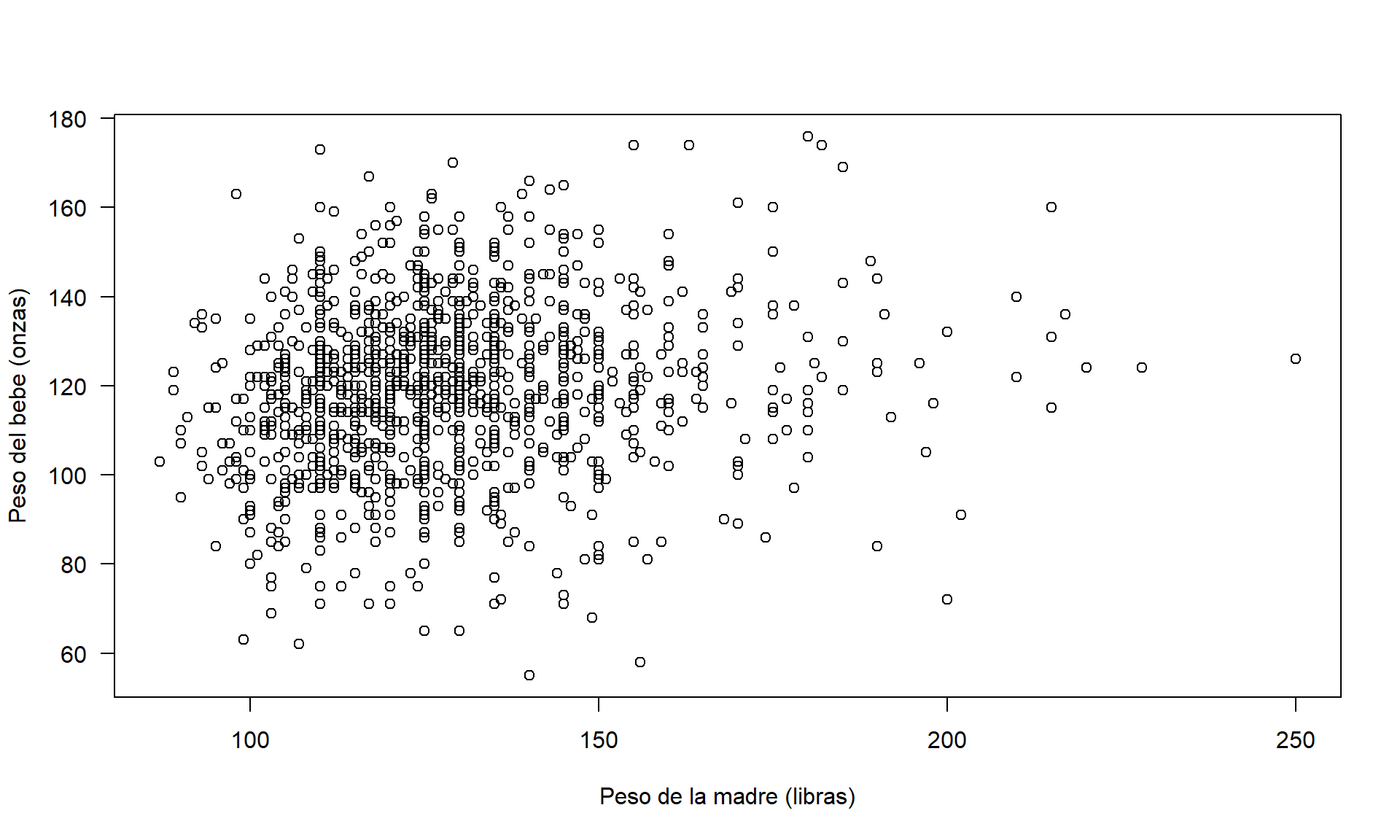


Figure 14.6: Diagrama de dispersión para peso del bebé versus peso de la madre.

Diagrama de dispersión entre pesos diferenciando por fumar

dt2 <- na.omit(dt)

plot(x=dt2$weight, y=dt2$bwt, pch=20, las=1,

xlab='Peso de la madre (libras)',

ylab='Peso del bebe (onzas)',

col=as.factor(dt2$smoke))

legend('bottomright', legend=c('No', 'Si'),

pch=20, col=1:2)

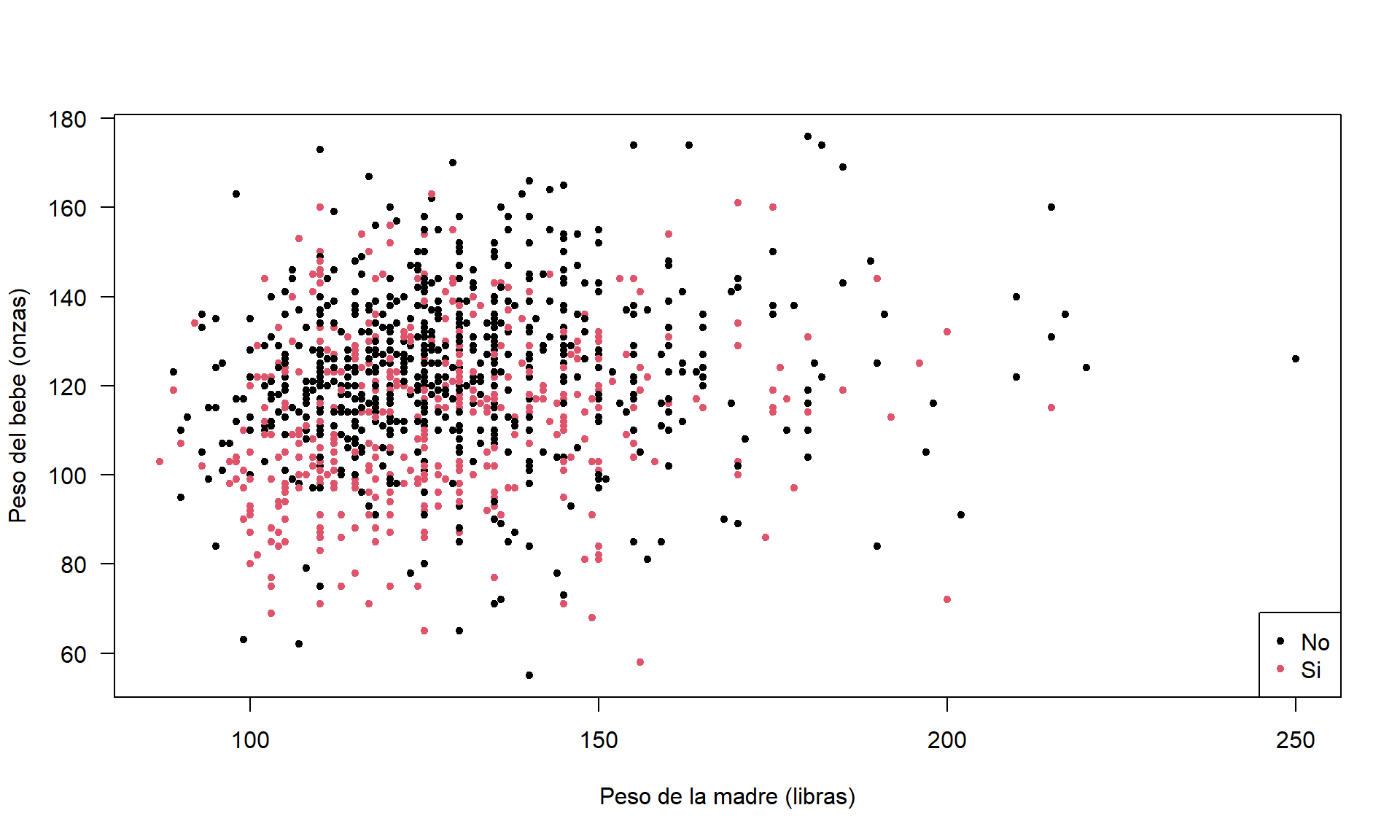


Figure 14.7: Diagrama de dispersión para peso del bebé versus peso de la madre diferenciando por fumar.

Para ver la relación entre el peso de la madre y la edad de la madre.

plot(dt$age, dt$weight, pch=20, col="tomato", las=1,

xlab="Edad de la madre",

ylab="Peso de la madre")

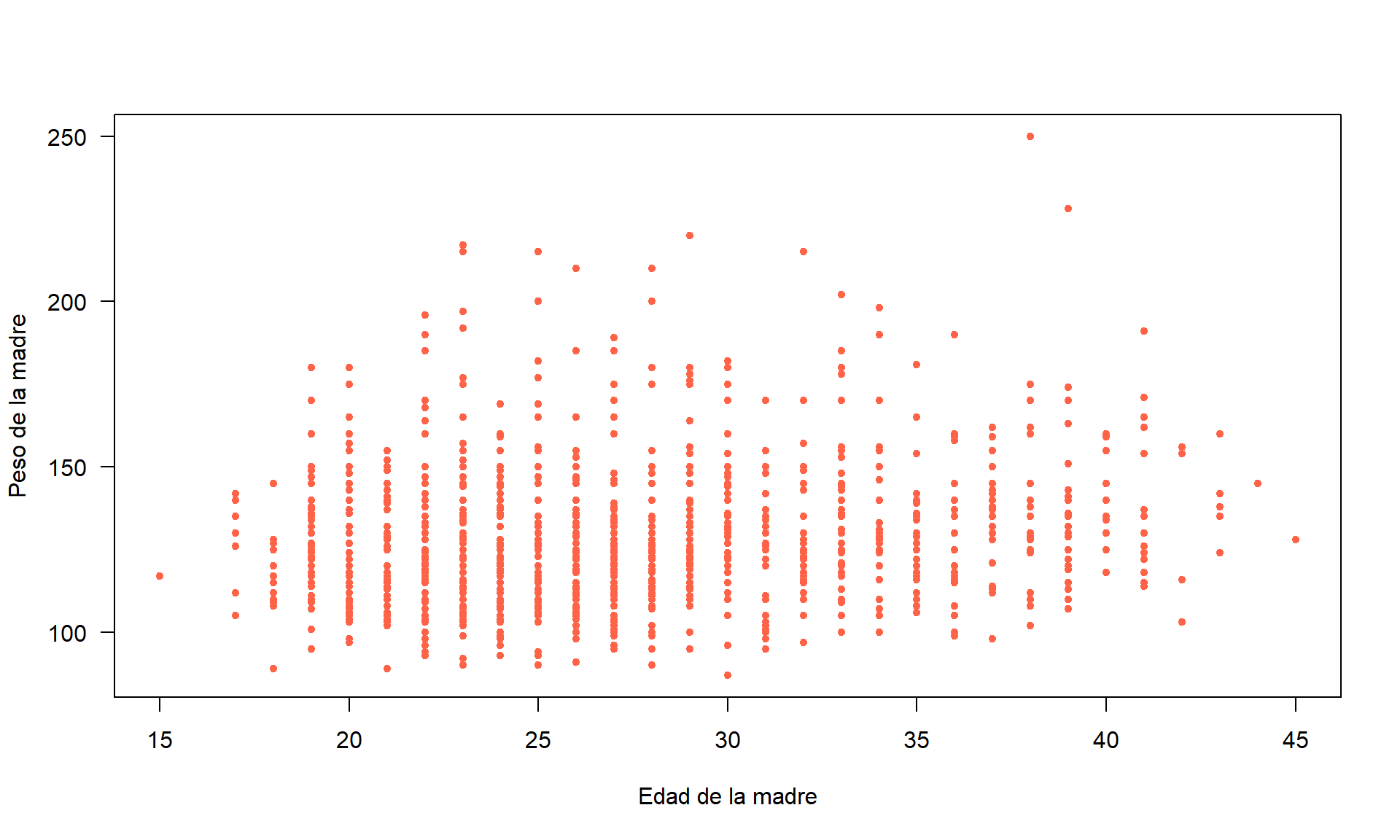


Figure 14.8: Diagrama de dispersión del peso de la madre versus su edad.