Estadística Avanzada - A2

Leroy Deniz

2022-11-14

Contents

0 Importación de librerías

```
library(ggplot2)
library(dplyr)
```

1 Lectura del fichero

Se carga el contenido del fichero *gpa.csv* utilizando la función *read.csv* y se muestran sus primeros registros junto con los nombres de las columnas y sus tipos.

```
gpa <- read.csv("gpa.csv", sep = ",", dec = ".")
print(head(gpa))
```

```
##    sat tothrs hsize hsrank   hsperc colgpa athlete female white black gpaletter
## 1  920     43  0.10      4 40.00000   2.04    TRUE   TRUE FALSE FALSE         C
## 2 1170     18  9.40    191 20.31915   4.00   FALSE  FALSE  TRUE FALSE         A
## 3  810     14  1.19     42 35.29412   1.78    TRUE  FALSE  TRUE FALSE         C
## 4  940     40  5.71    252 44.13310   2.42   FALSE  FALSE  TRUE FALSE         C
## 5 1180     18  2.14     86 40.18692   2.61   FALSE  FALSE  TRUE FALSE         B
## 6  980    114  2.68     41 15.29851   3.03   FALSE   TRUE  TRUE FALSE         B
```

Se muestran más explícitamente los tipos de datos con los que R ha identificado cada columna.

```
t(t(sapply(gpa, class)))
```

```
##           [,1]
## sat       "integer"
## tothrs    "integer"
## hsize     "numeric"
## hsrank    "integer"
## hsperc    "numeric"
## colgpa    "numeric"
## athlete   "logical"
## female    "logical"
## white     "logical"
## black     "logical"
## gpaletter "character"
```

2 Estadística descriptiva y visualización

2.1 Análisis descriptivo

Se muestra a continuación información relativa al dataset con el que se va a trabajar,

```
summary(gpa)
```

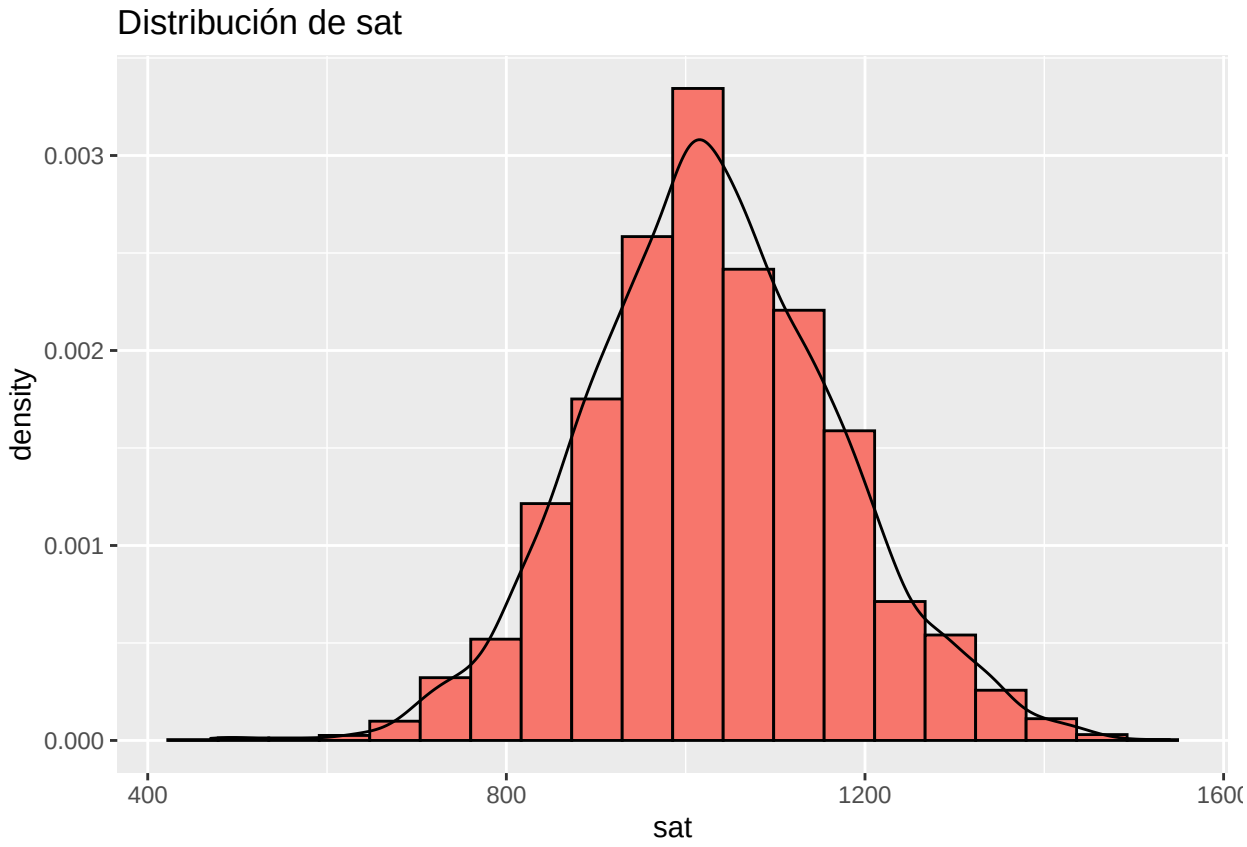
```
##       sat           tothrs           hsize          hsrank
##  Min.   : 470   Min.   :  6.00   Min.   :0.03   Min.   :  1.00
##  1st Qu.: 940   1st Qu.: 17.00   1st Qu.:1.65   1st Qu.: 11.00
##  Median :1030   Median : 47.00   Median :2.51   Median : 30.00
##  Mean   :1030   Mean   : 52.83   Mean   :2.80   Mean   : 52.83
##  3rd Qu.:1120   3rd Qu.: 80.00   3rd Qu.:3.68   3rd Qu.: 70.00
##  Max.   :1540   Max.   :137.00   Max.   :9.40   Max.   :634.00
##      hsperc            colgpa       athlete          female
##  Min.   : 0.1667   Min.   :0.000   Mode :logical   Mode :logical
##  1st Qu.: 6.4328   1st Qu.:2.210   FALSE:3943      FALSE:2277
##  Median :14.5833   Median :2.660   TRUE :194       TRUE :1860
##  Mean   :19.2371   Mean   :2.654
##  3rd Qu.:27.7108   3rd Qu.:3.120
##  Max.   :92.0000   Max.   :4.000
##    white           black          gpaletter
##  Mode :logical   Mode :logical   Length:4137
##  FALSE:308       FALSE:3908      Class :character
##  TRUE :3829      TRUE :229       Mode  :character
##
##
##
```

La cantidad de registros del dataset es de 4137 filas (observaciones) con 11 columnas (variables)

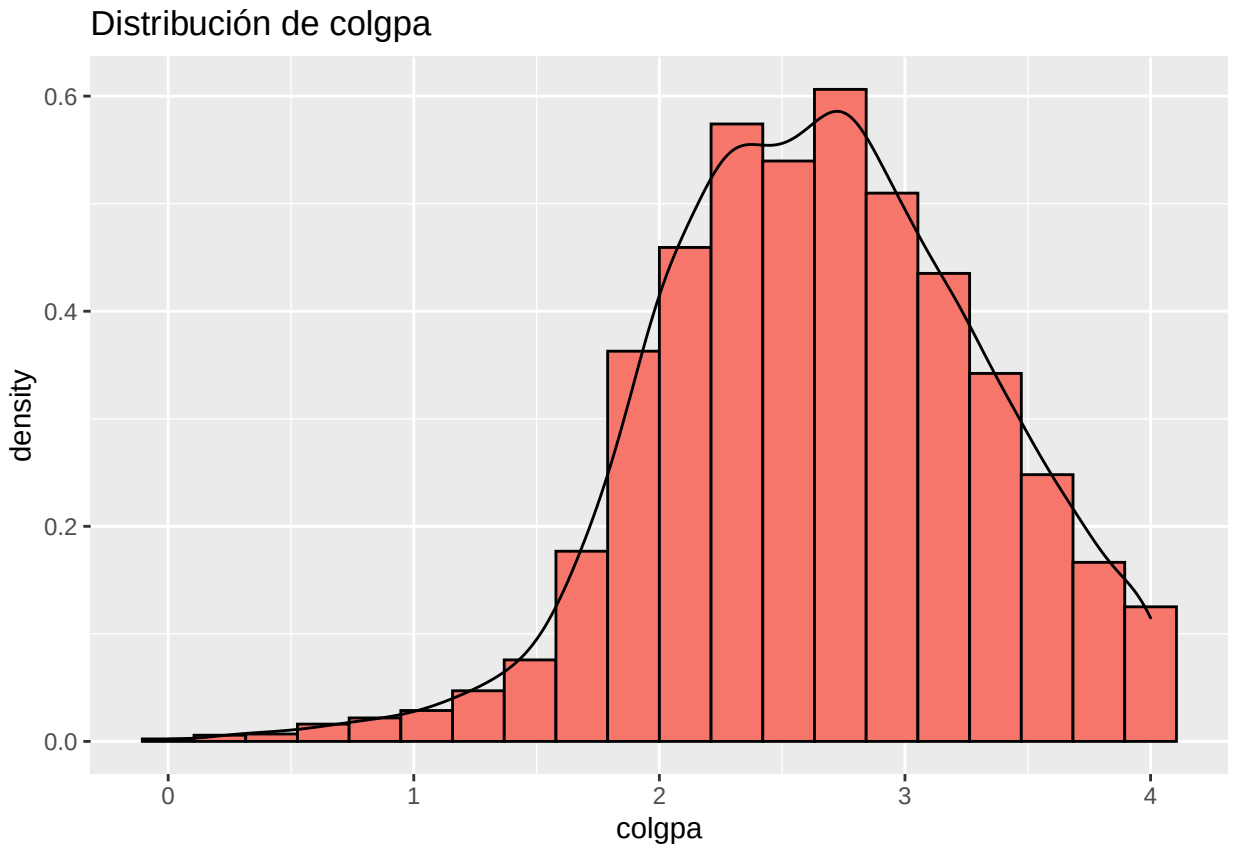
2.2 Visualización

2.2.1 Distribución de las variables *sat* y *colgpa*.

```
ggplot(gpa, aes(sat)) + geom_histogram(aes(y = ..density..), bins = 20, color = "black",
    fill = "#F7766C") + geom_density(aes(fill = sat), alpha = 0.2) + ggtitle("Distribución de sat")
```

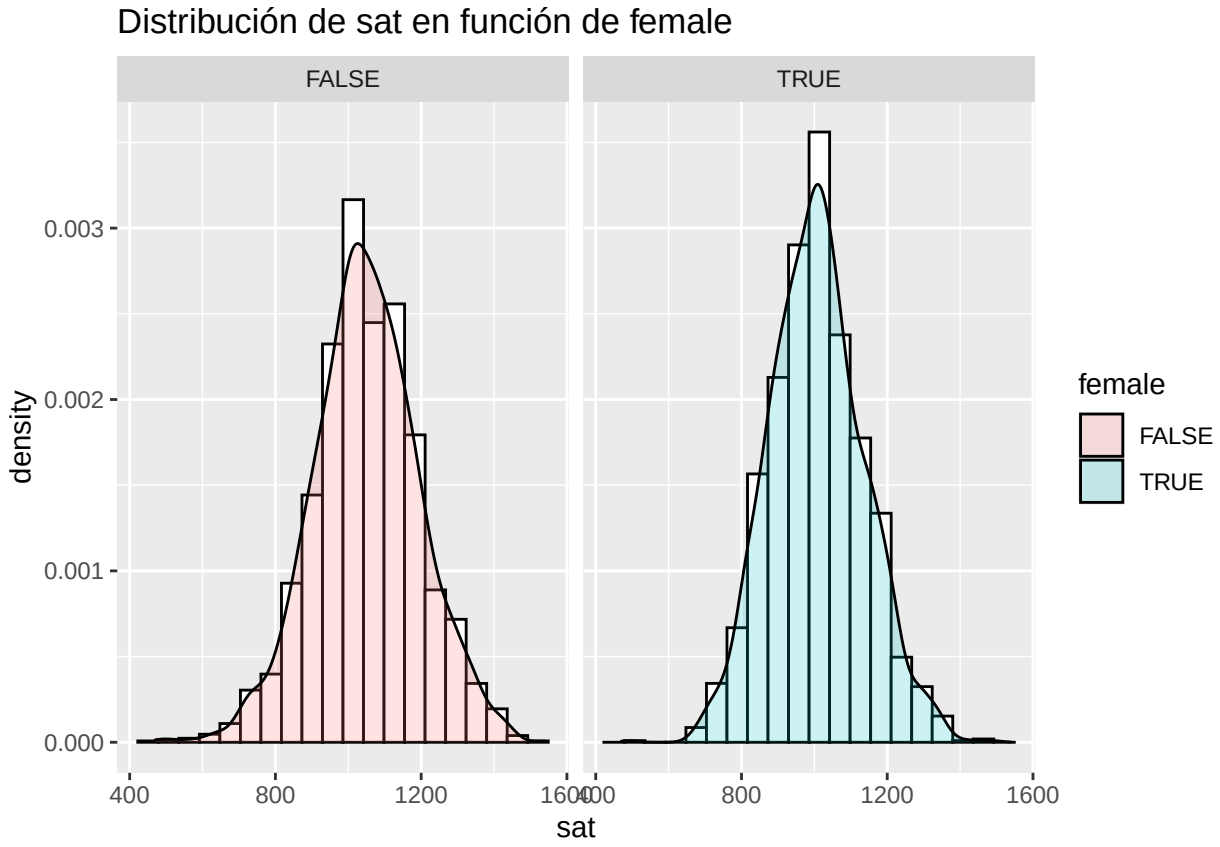


```
ggplot(gpa, aes(colgpa)) + geom_histogram(aes(y = ..density..), bins = 20, color = "black",
    fill = "#F7766C") + geom_density(aes(fill = colgpa), alpha = 0.2) + ggtitle("Distribución de colgpa")
```

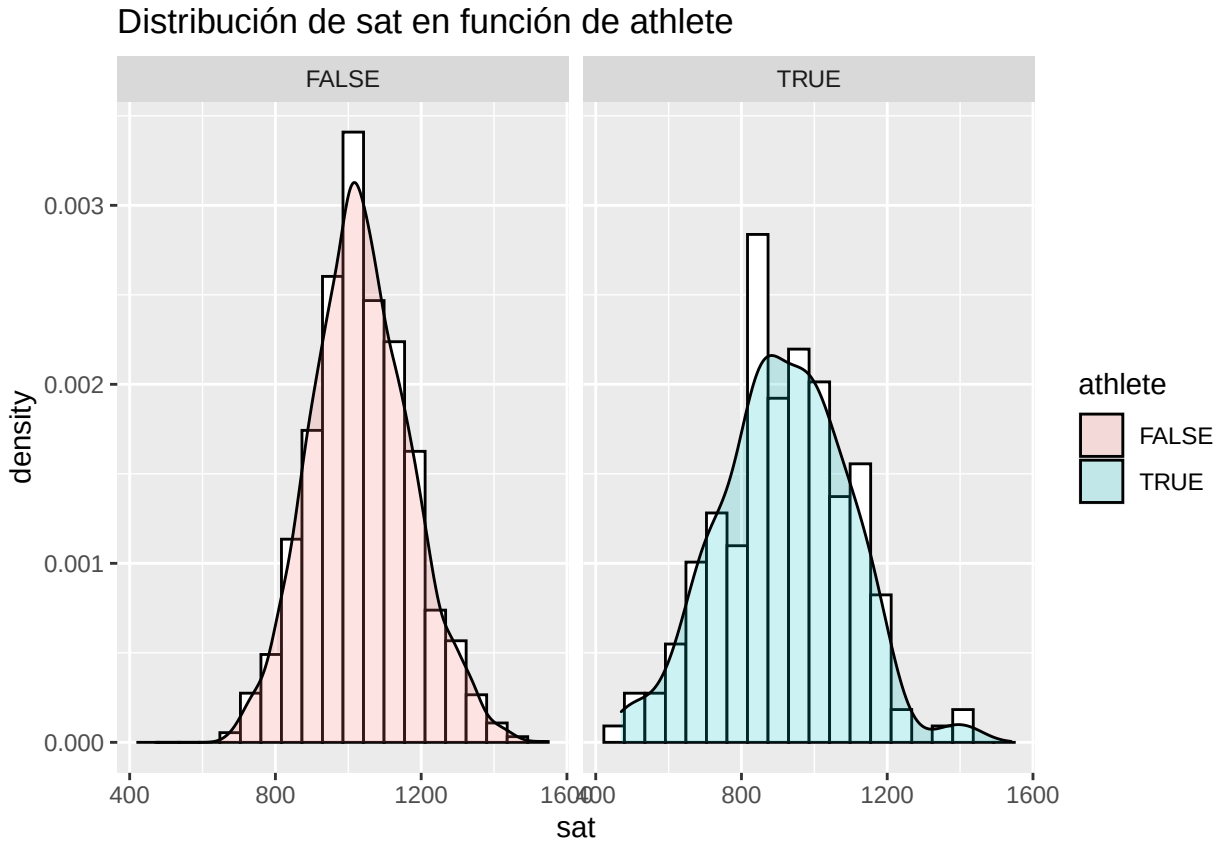


2.2.2 Distribución de la variable 'sat' con respecto a la variable género ('female'), la variable atleta ('athlete') y la raza ('white', 'black').

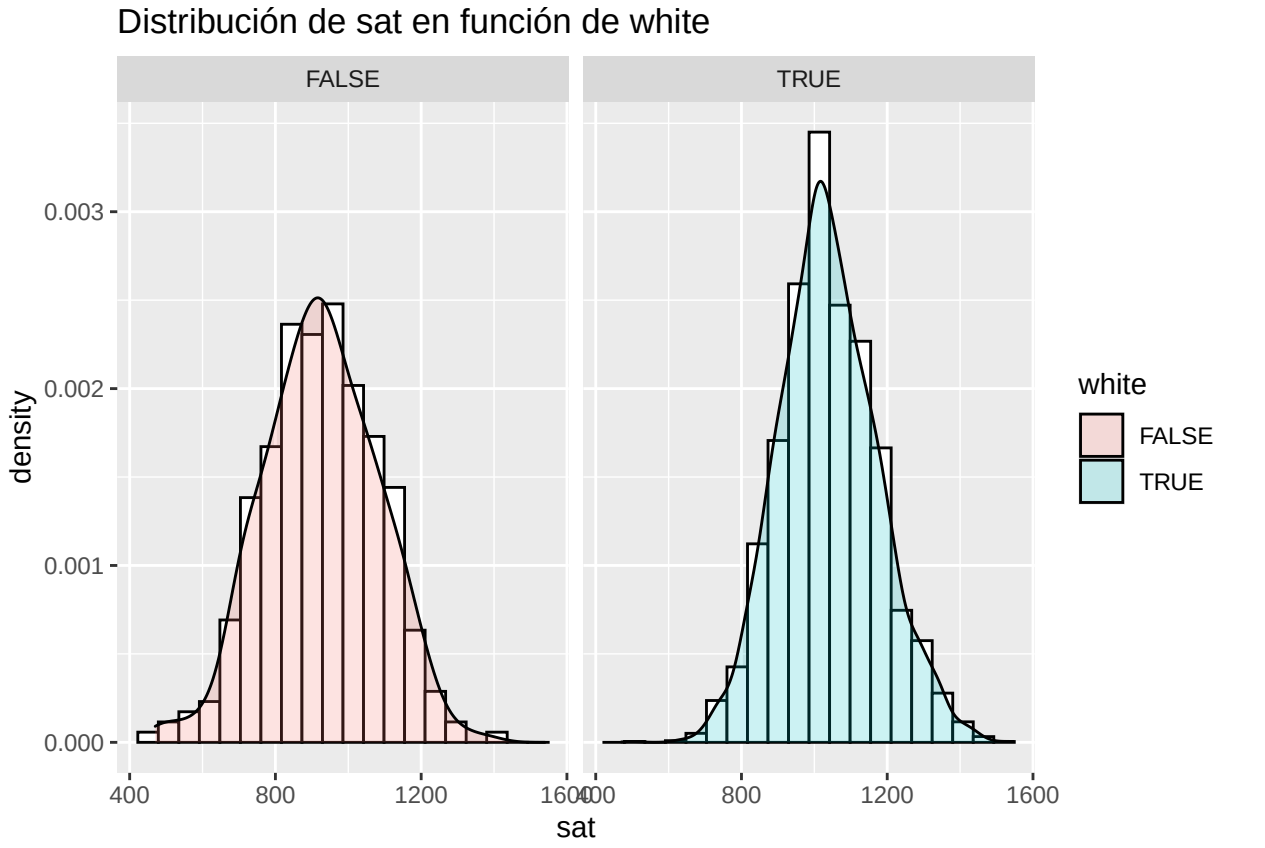
```
ggplot(gpa, aes(sat)) + geom_histogram(aes(y = ..density..), bins = 20, color = "black",
    fill = "white") + geom_density(aes(fill = female), alpha = 0.2) + facet_wrap(~female) +
    ggtitle("Distribución de sat en función de female")
```



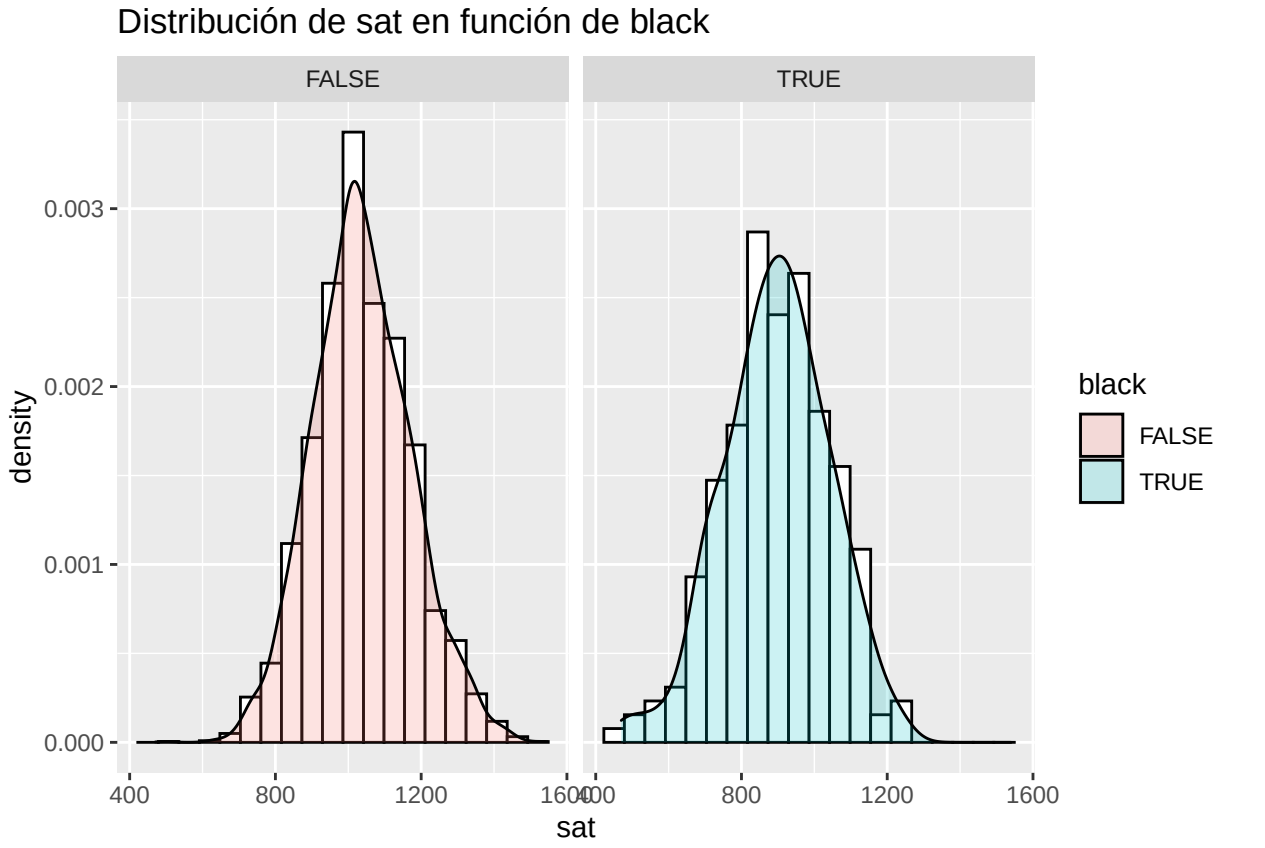
```
ggplot(gpa, aes(sat)) + geom_histogram(aes(y = ..density..), bins = 20, color = "black",
    fill = "white") + geom_density(aes(fill = athlete), alpha = 0.2) + facet_wrap(~athlete) +
    ggtitle("Distribución de sat en función de athlete")
```



```
ggplot(gpa, aes(sat)) + geom_histogram(aes(y = ..density..), bins = 20, color = "black",
    fill = "white") + geom_density(aes(fill = white), alpha = 0.2) + facet_wrap(~white) +
    ggtitle("Distribución de sat en función de white")
```

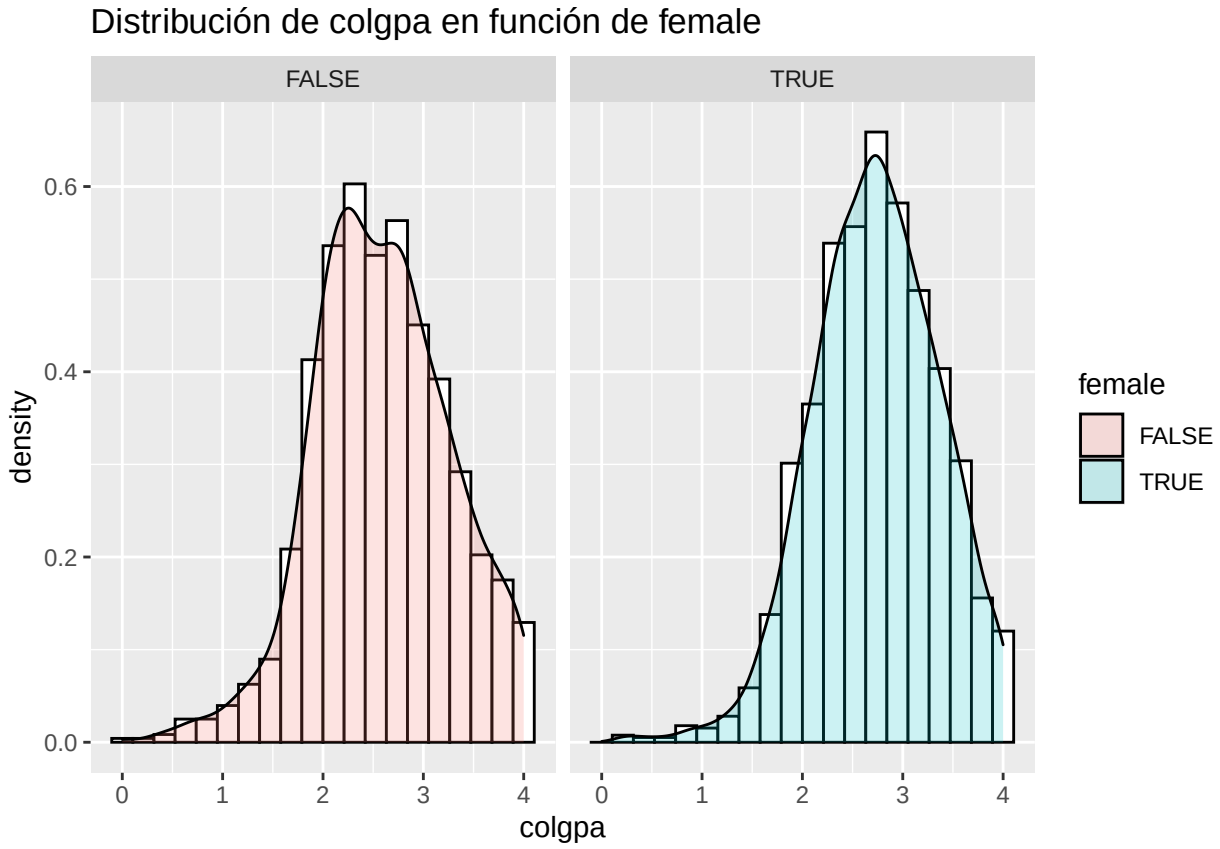


```
ggplot(gpa, aes(sat)) + geom_histogram(aes(y = ..density..), bins = 20, color = "black",
    fill = "white") + geom_density(aes(fill = black), alpha = 0.2) + facet_wrap(~black) +
    ggtitle("Distribución de sat en función de black")
```

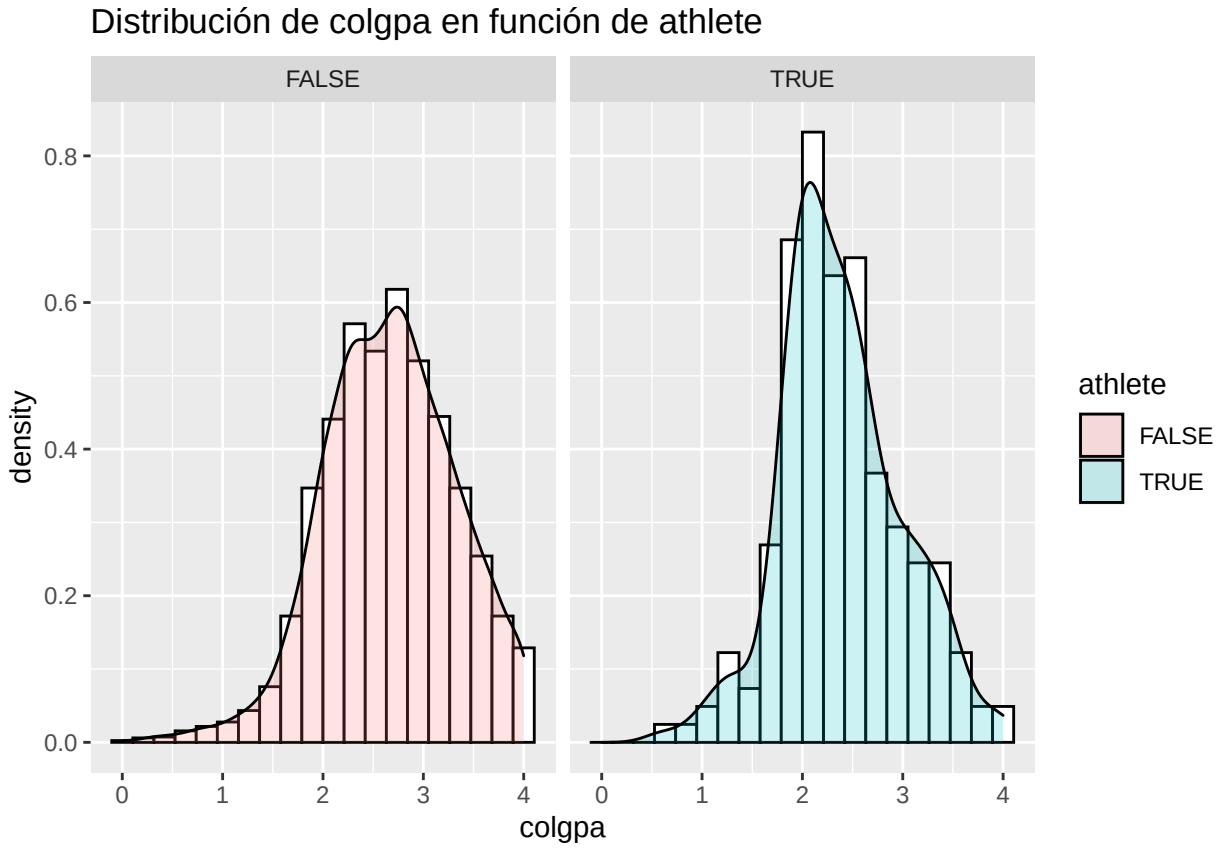


2.2.3 Distribución de la variable 'colgpa' con respecto a la variable género ('female'), la variable atleta ('athlete') y la raza ('white', 'black').

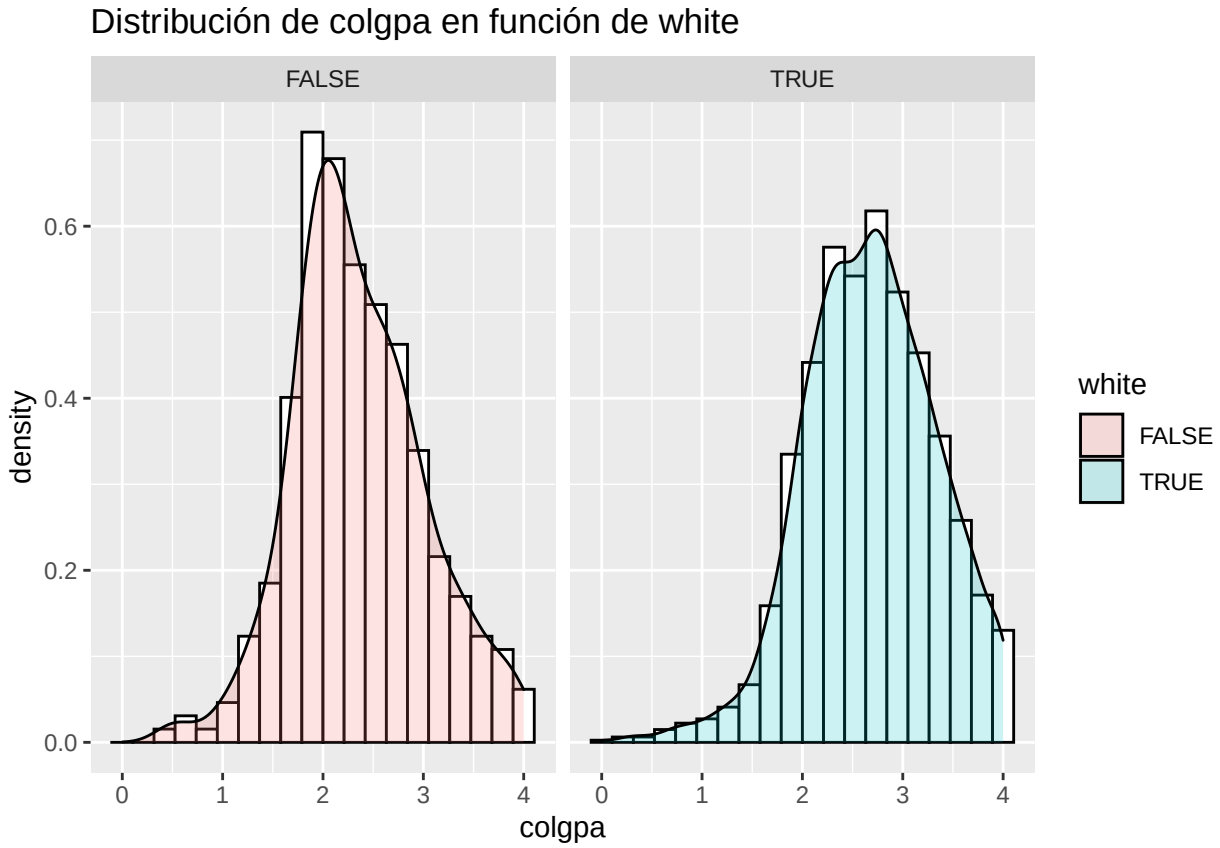
```
ggplot(gpa, aes(colgpa)) + geom_histogram(aes(y = ..density..), bins = 20, color = "black",
    fill = "white") + geom_density(aes(fill = female), alpha = 0.2) + facet_wrap(~female) +
    ggtitle("Distribución de colgpa en función de female")
```



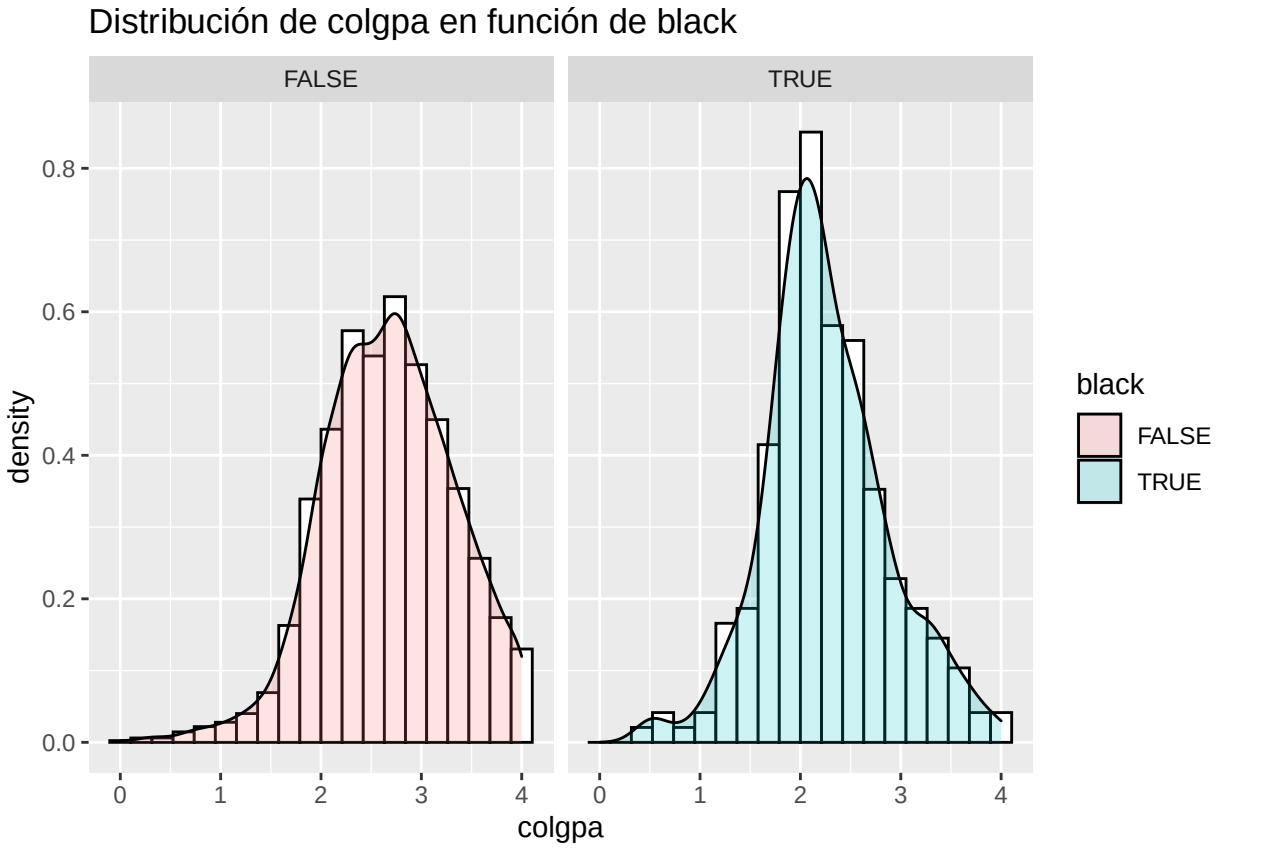
```
ggplot(gpa, aes(colgpa)) + geom_histogram(aes(y = ..density..), bins = 20, color = "black",
    fill = "white") + geom_density(aes(fill = athlete), alpha = 0.2) + facet_wrap(~athlete) +
    ggtitle("Distribución de colgpa en función de athlete")
```



```
ggplot(gpa, aes(colgpa)) + geom_histogram(aes(y = ..density..), bins = 20, color = "black",
    fill = "white") + geom_density(aes(fill = white), alpha = 0.2) + facet_wrap(~white) +
    ggtitle("Distribución de colgpa en función de white")
```



```
ggplot(gpa, aes(colgpa)) + geom_histogram(aes(y = ..density..), bins = 20, color = "black",
    fill = "white") + geom_density(aes(fill = black), alpha = 0.2) + facet_wrap(~black) +
    ggtitle("Distribución de colgpa en función de black")
```



2.2.4 Interpretación

A simple vista podríamos afirmar que la distribución de la variable sat con respecto a las cuatro variables con las que se estudia, presenta una distribución centrada, en algunos casos podría, previa verificación, incluso tener una distribución normal.

La variable *colgpa* por el contrario, presenta una desviación hacia la derecha en todos sus distribuciones, lo que podría suponer un desbalanceo de los casos y, por tanto, un sesgo más elevado en comparación con los demás

3 Intervalo de confianza de la media poblacional de la variable *sat* y *colgpa*

3.1 Supuestos

Como se puede apreciar en las gráficas anteriores, la variable *sat* podría llegar a tener una distribución normal, no así *colgpa*, pero de todas formas podemos valernos del Teorema de Límite Central (TLC) que nos asegura que la media de una muestra de tamaño superior se comporta como una distribución normal.

En este caso, como la muestra supera los cuatro mil casos y el TLC establece que esta regla se cumple para muestras superiores a 30, estamos en una situación cómoda para poder apoyarnos en él.

3.2 Función de cálculo del intervalo de confianza

```r
# Normal
IC_norm <- function(x, NC) {
    n <- length(x)
    alpha <- 1/(NC/100)
    SE <- sd(x)/sqrt(n)

    z <- qnorm(alpha/2, lower.tail = TRUE)
    z_SE <- z * SE
    Low <- mean(x) - z_SE
    Up <- mean(x) + z_SE

    return(c(Low, Up))
}

# Student T
IC_t <- function(x, NC) {
    n <- length(x)
    alpha <- 1/(NC/100)
    SE <- sd(x)/sqrt(n)

    t <- qt(alpha/2, df = n - 1, lower.tail = TRUE)
    t_SE <- t * SE
    Low <- mean(x) - t_SE
    Up <- mean(x) + t_SE

    return(c(Low, Up))
}

# Por defecto toma el valor para utilizar Nomal
IC <- function(x, NC, kind = "normal") {
    if (kind == "t") {
        return(IC_t(x, NC))
    } else {
        return(IC_norm(x, NC))
    }
}
```

3.3 Intervalo de confianza de la variable sat

```
sat_IC_90 <- IC(gpa$sat, 90)
sat_IC_95 <- IC(gpa$sat, 95)
```

El intervalo de confianza de sat al 90% es [1030.02836, 1030.6339557] y al 95% es [1030.1880887, 1030.474227]

3.4 Intervalo de confianza de la variable colgpa

```
colgpa_IC_90 <- IC(gpa$colgpa, 90)
colgpa_IC_95 <- IC(gpa$colgpa, 95)
```

El intervalo de confianza de colgpa al 90% es [2.6534559, 2.6548061] y al 95% es [2.6534559, 2.6548061]

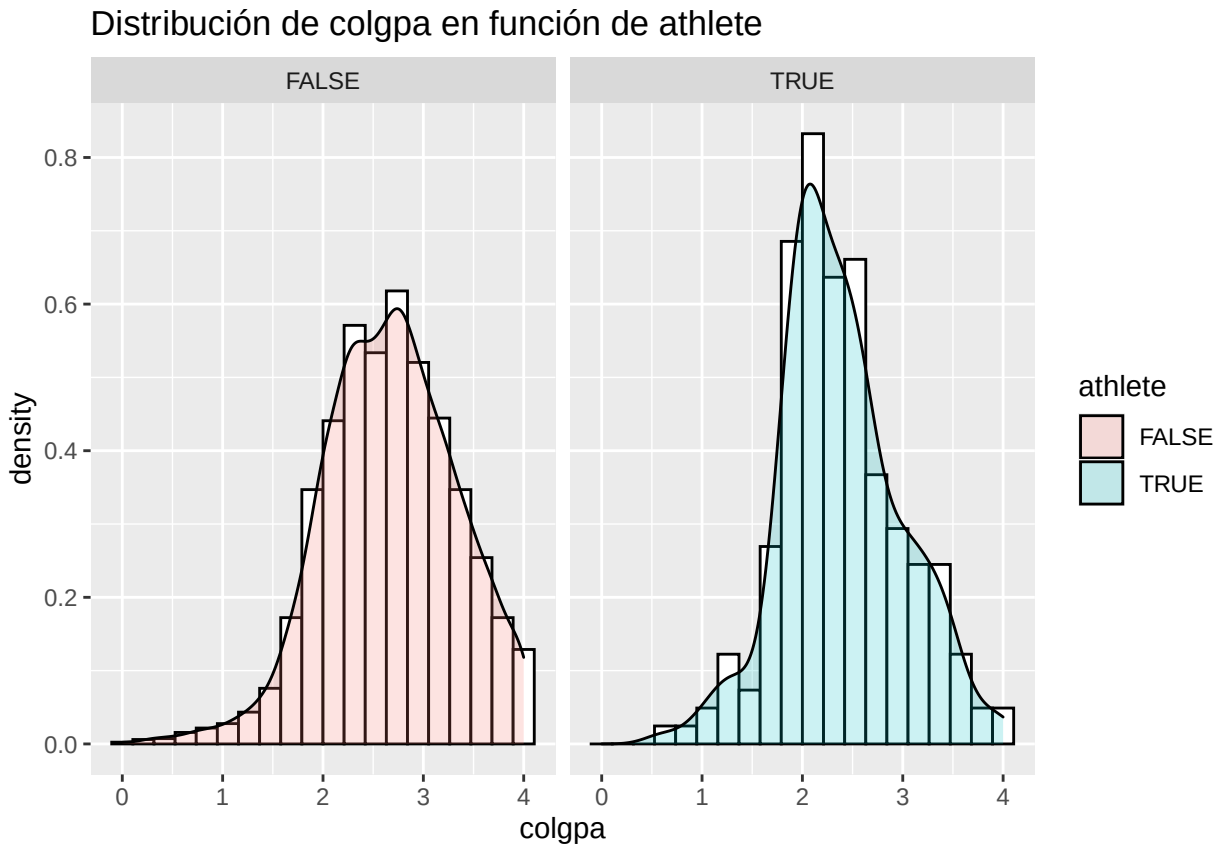
3.5 Interpretación

El intervalo de confianza puede interpretarse como un margen de seguridad en los datos obtenidos con un nivel de seguridad. Esto quiere decir que, de realizar el muestreo con un número grande de muestras, sabemos que hay una probabilidad fija de que el valor de la media de toda la población esté en ese intervalo. De esto se desprende que con un nivel de confianza de un 90%, la media estará en el intervalo [2.6527022, 2.6555599], asimismo, con un valor de confianza de 95%, la media estará en [2.6534559, 2.6548061].

4 ¿Ser atleta influye en la nota?

4.1 Análisis visual

```
ggplot(gpa, aes(colgpa)) + geom_histogram(aes(y = ..density..), bins = 20, color = "black",
    fill = "white") + geom_density(aes(fill = athlete), alpha = 0.2) + facet_wrap(~athlete) +
    ggtitle("Distribución de colgpa en función de athlete")
```



4.2 Función para el contraste de medias

```r
contraste_medias <- function(s1, s2, alt, CL) {

    # Cálculo de Medias
    mean1 <- mean(s1)
    mean2 <- mean(s2)

    # Cálculo del tamaño de la muestra
    n1 <- length(s1)
    n2 <- length(s2)

    # Cálculo de la desviación estándar
    sd1 <- sd(s1)
    sd2 <- sd(s2)

    # Cálculo del nivel de significancia
    alpha <- (1 - CL/100)

    # Cálculo de los grados de libertad (Apartado 5.2.2 de la teoría)
    denominador <- ((sd1^2/n1)^2/(n1 - 1) + (sd2^2/n2)^2/(n2 - 1))
    df <- ((sd1^2/n1 + sd2^2/n2)^2)/denominador

    # Cálculo del valor t (z según la distribución normal estandarizada)
    sb <- sqrt(sd1^2/n1 + sd2^2/n2)
    t <- (mean1 - mean2)/sb

    # Evaluación de la condición =
    if (alt == "bilateral") {
        t_critical <- qt(alpha/2, df, lower.tail = FALSE)
        p_value <- pt(abs(t), df, lower.tail = FALSE) * 2

        # Evaluación de la condición <
    } else if (alt == "<") {
        t_critical <- qt(alpha, df, lower.tail = TRUE)
        p_value <- pt(t, df, lower.tail = TRUE)

        # Evaluación de la condición > (alt == '>')
    } else {
        t_critical <- qt(alpha, df, lower.tail = FALSE)
        p_value <- pt(t, df, lower.tail = FALSE)
    }

    # Definición del vector resultado
    vector_data <- c(mean1, mean2, t, t_critical, p_value, alpha, df)
    names(vector_data) <- c("mean1", "mean2", "t", "t_critical", "p_value", "alpha",
        "df")
    return(vector_data)
}
```

4.3 Pregunta de investigación

¿Es igual la nota *colgpa* de los atletas y los no-atletas?

4.4 Hipótesis nula y la alternativa

Hipótesis nula:

$$H_0 : \mu_{colgpa\_athletes} = \mu_{colgpa\_non\_athletes}$$

Hipótesis alternativa:

$$H_1 : \mu_{colgpa\_athletes} \neq \mu_{colgpa\_non\_athletes}$$

4.5 Justificación del test a aplicar

La razón por la cual se está utilizando este test específico es porque tenemos dos muestras que no abarcan toda la población y por lo tanto no conocemos las varianzas reales, y además, tenemos dos muestras que son independientes (atletas y no atletas, casos disjuntos, misma columna en los datos). De esta forma, se utiliza contraste de medias.

4.6 Cálculo

```
x <- gpa[gpa$athlete == TRUE, ]$colgpa
y <- gpa[gpa$athlete == FALSE, ]$colgpa
datos <- contraste_medias(x, y, "bilateral", 95)
print(t(t(datos)))
```

```
##                    [,1]
## mean1       2.382732e+00
## mean2       2.667484e+00
## t          -6.459086e+00
## t_critical  1.970969e+00
## p_value     6.828951e-10
## alpha       5.000000e-02
## df          2.167579e+02
```

4.7 Interpretación del test

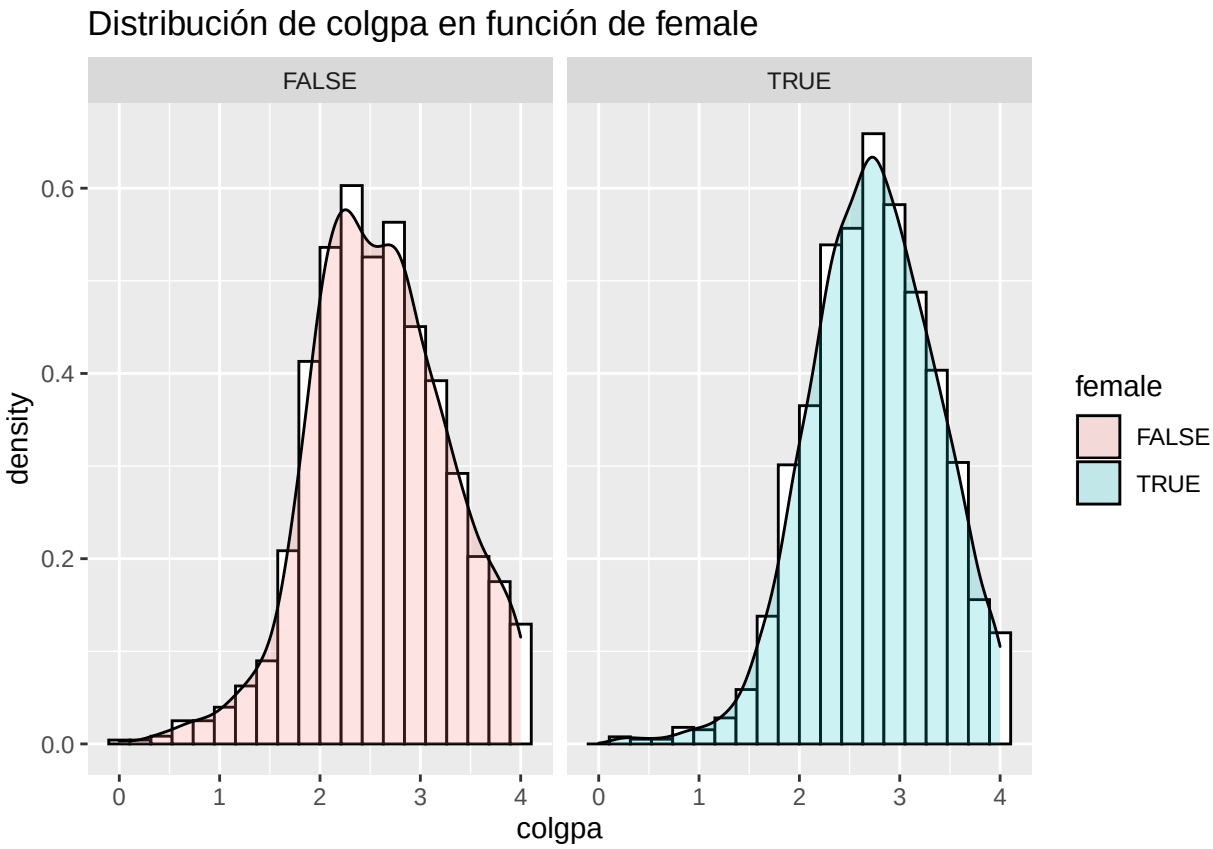
Puesto que p es significativamente menor que *alpha*, estamos en condiciones de rechazar la hipótesis nula y, por el contrario, decantarnos a favor de la hipótesis alternativa.

Respondiendo a la pregunta de investigación, la respuesta es no, las notas entre atletas y no atletas no son iguales.

5 ¿Las mujeres tienen mejor nota que los hombres?

5.1 Análisis visual

```
ggplot(gpa, aes(colgpa)) + geom_histogram(aes(y = ..density..), bins = 20, color = "black",
    fill = "white") + geom_density(aes(fill = female), alpha = 0.2) + facet_wrap(~female) +
    ggtitle("Distribución de colgpa en función de female")
```



5.2 Función

La función está definida en el punto 4.2.

5.3 Pregunta de investigación

¿La nota de las mujeres es mayor que la de los hombres?

5.4 Hipótesis nula y la alternativa

Hipótesis nula:

$$H_0 : \mu_{colgpa\_female} \geq \mu_{colgpa\_non\_female}$$

Hipótesis alternativa:

$$H_1 : \mu_{colgpa\_female} < \mu_{colgpa\_non\_female}$$

5.5 Justificación del test a aplicar

De la misma forma que la pregunta anterior, se aplica un test de contraste de medias para el rechazo o no de la hipótesis nula, puesto que tenemos dos conjuntos de variables independientes y con varianzas reales desconocidas.

5.6 Cálculo

Se generan las dos series según la condición requerida.

```
x <- gpa[gpa$female == TRUE, ]$colgpa  # mujeresbres
```

Para un nivel de confianza de un 95%:

```
datos <- contraste_medias(x, y, ">", 95)
print(t(t(datos)))
```

```
##                    [,1]
## mean1      2.733016e+00
## mean2      2.667484e+00
## t          3.686064e+00
## t_critical 1.645250e+00
## p_value    1.154434e-04
## alpha      5.000000e-02
## df         3.849466e+03
```

Para un nivel de confianza de un 90%:

```
datos <- contraste_medias(x, y, ">", 90)
print(t(t(datos)))
```

```
##                    [,1]
## mean1      2.733016e+00
## mean2      2.667484e+00
## t          3.686064e+00
## t_critical 1.281772e+00
## p_value    1.154434e-04
## alpha      1.000000e-01
## df         3.849466e+03
```

5.7 Interpretación del test

Como puede apreciarse en las tablas inmediatamente anteriores, en ambos casos el *p_value* es significativamente menor que el *alpha*, por lo que podemos rechazar la hipótesis nula en favor de la alternativa para ambos niveles de confianza.

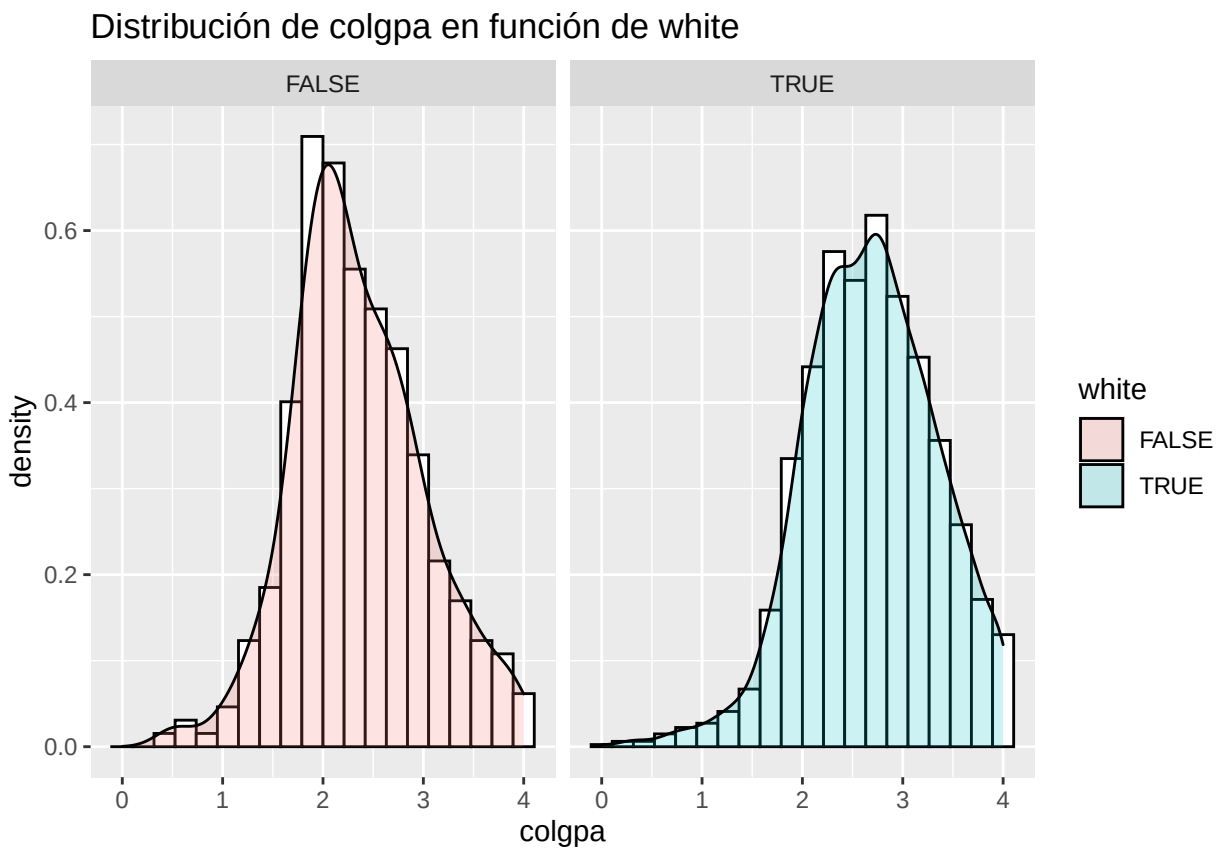
Respondiendo a la pregunta de investigación, la media de la nota de las mujeres es menor igual, a la de los hombres.

6 ¿Hay diferencias en la nota según la raza?

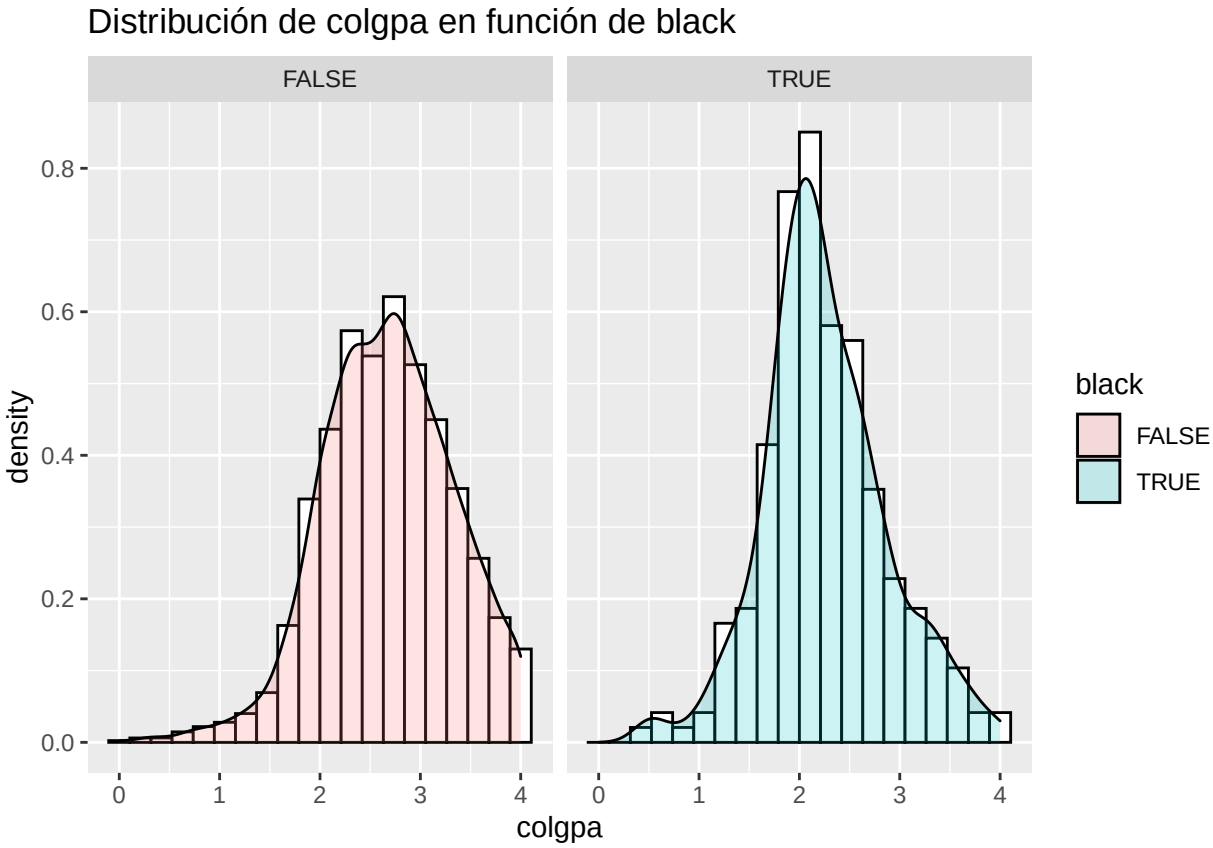
6.1 Análisis visual

Para este caso se deben analizar las dos gráficas de la izquierda, cuando la raza es blanca o negra, ignorando los demás datos que, en este estudio, no serían relevantes.

```
ggplot(gpa, aes(colgpa)) + geom_histogram(aes(y = ..density..), bins = 20, color = "black",
    fill = "white") + geom_density(aes(fill = white), alpha = 0.2) + facet_wrap(~white) +
    ggtitle("Distribución de colgpa en función de white")
```



```
ggplot(gpa, aes(colgpa)) + geom_histogram(aes(y = ..density..), bins = 20, color = "black",
    fill = "white") + geom_density(aes(fill = black), alpha = 0.2) + facet_wrap(~black) +
    ggtitle("Distribución de colgpa en función de black")
```



6.2 Función

La función está definida en el punto 4.2.

6.3 Pregunta de investigación

¿Son iguales las notas en ambas razas?

6.4 Hipótesis nula y la alternativa

Hipótesis nula:

$$H_0 : \mu_{colgpa\_white} = \mu_{colgpa\_black}$$

Hipótesis alternativa:

$$H_1 : \mu_{colgpa\_white} \neq \mu_{colgpa\_black}$$

6.5 Justificación del test a aplicar

De la misma forma que la pregunta anterior, se aplica un test de contraste de medias para el rechazo o no de la hipótesis nula, puesto que lo que se busca es verificar un valor de significancia.

6.6 Cálculo

```
x <- gpa[gpa$white == TRUE, ]$colgpa  # white
y <- gpa[gpa$black == TRUE, ]$colgpa  # black
```

Para un nivel de confianza de un 95%:

```
datos <- contraste_medias(x, y, "bilateral", 95)
print(t(t(datos)))
```

```
##                    [,1]
## mean1      2.678360e+00
## mean2      2.255546e+00
## t          1.007982e+01
## t_critical 1.969141e+00
## p_value    2.214163e-20
## alpha      5.000000e-02
## df         2.597000e+02
```

Como el *p_value* es significativamente menor que *alpha*, podemos rechazar de plano la hipótesis nula y decantarnos por la hipótesis alternativa; es decir, que las notas no son iguales. En tal caso, deberíamos ir un poco más allá e intentar verificar que, si no son iguales, cuál tiene un mayor nivel, por lo que a continuación se realiza un contraste de medias con $x < y$, donde la población de raza blanca tenga notas menores que la de raza negra.

```
datos <- contraste_medias(x, y, "<", 95)
print(t(t(datos)))
```

```
##                  [,1]
## mean1        2.678360
## mean2        2.255546
## t           10.079823
## t_critical  -1.650742
## p_value      1.000000
## alpha        0.050000
## df         259.700032
```

6.7 Interpretación del test

En el segundo test el *p_value* es mayor con respecto al *alpha* para no poder rechazar la nueva hipótesis, por lo que no contamos con información suficiente para asegurar que, la población de raza negra tiene mejores notas que aquella de raza blanca.

7 Proporción de atletas

Se define una función que calculará el contraste de proporciones con una única distribución.

```
contraste_proporciones_unidist <- function(p, p0, n, alt, CL) {
    z <- (p - p0)/(sqrt(p0 * (1 - p0)/n))
    alpha <- 1 - CL/100

    # Evaluación de la condición =
    if (alt == "bilateral") {
        p_value <- pnorm(abs(z), lower.tail = FALSE) * 2
        z_critical <- qnorm(alpha/2, lower.tail = FALSE)

        # Evaluación de la condición >
    } else if (alt == ">") {
        p_value <- pnorm(z, lower.tail = FALSE)
        z_critical <- qnorm(alpha, lower.tail = FALSE)

        # Evaluación de la condición <
    } else if (alt == "<") {
        p_value <- pnorm(z, lower.tail = TRUE)
        z_critical <- qnorm(alpha, lower.tail = TRUE)

    }

    datos <- c(p, p0, z, z_critical, alpha, p_value)
    names(datos) <- c("p", "p0", "z", "z_critical", "alpha", "p_value")
    return(datos)
}
```

7.1 Análisis visual

```
# total de atletas
athletes <- length(gpa[gpa$athlete == TRUE, ]$athlete)

# total de la muestra menos los atletas
non_athletes = length(gpa$colgpa) - athletes

# definición de los vectores para el Pie chart
proportions <- c(athletes, non_athletes)
labels <- c(TRUE, FALSE)

# Definición del gráfico
pie(proportions, labels = labels, main = "Proporción de Atletas")
```

7.2 Pregunta de investigación

¿La proporción de atletas es menor al 5% en la muestra?

7.3 Hipótesis nula y la alternativa

Hipótesis nula:

$$H_0 : p_{athletes} \leq p_{muestra}$$

Hipótesis alternativa:

$$H_1 : p_{athletes} > qp_{muestra}$$

7.4 Justificación del test a aplicar

Se realiza el contraste de proporciones a través de la función del comienzo del apartado, cuya firma establece los parámetros necesarios para poder realizar el procedimiento, devolviendo como resultado los valores que nos permitan rechazar la hipótesis nula o no. La función permite el contraste de proporciones con una única distribución, cuya proporción se contraste con un valor fijo.

7.5 Realizad los cálculos del test

```
n <- athletes + non_athletes
p_athletes <- athletes/n
datos <- contraste_proporciones_unidist(p_athletes, 0.05, n, "<", 95)
print(t(t(datos)))
```

```
##                  [,1]
## p           0.04689388
## p0          0.05000000
## z          -0.91667115
## z_critical -1.64485363
## alpha       0.05000000
## p_value     0.17965749
```

7.6 Interpretación del test

Como el *p_value* es mayor que *alpha*, no podemos rechazar la hipótesis nula puesto que el error que cometeríamos es suficientemente grande. Eso nos permite responder a la pregunta de investigación con un sí, la proporción de atletas es menor al 5%.

8 ¿Hay más atletas entre los hombres que entre las mujeres?

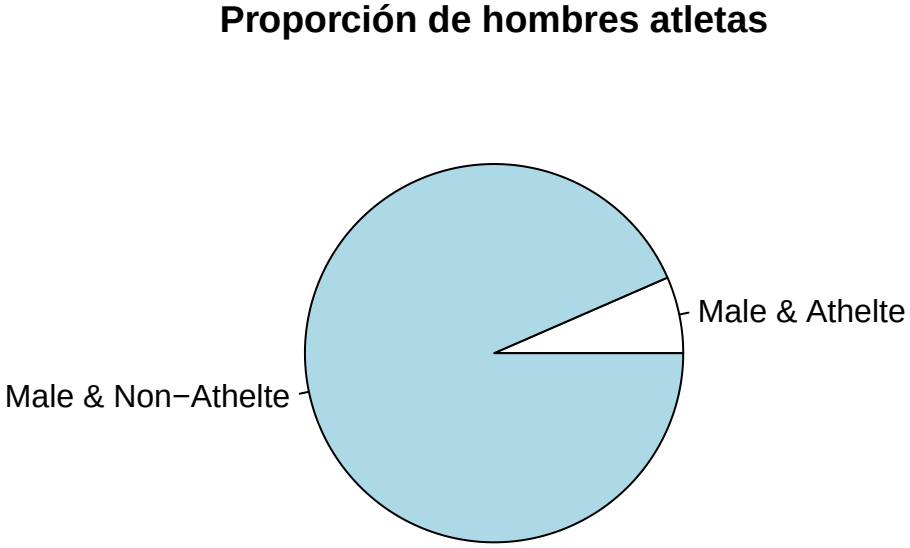
Función para la aplicación de controste de proporciones con dos distribuciones.

```
contraste_proporciones_bidist <- function(p1, p2, n1, n2, alt, CL) {
    p <- (n1 * p1 + n2 * p2)/(n1 + n2)
    z <- (p1 - p2)/(sqrt(p * (1 - p) * (1/n1 + 1/n2)))
    alpha <- 1 - CL/100

    # Evaluación de la condición =
    if (alt == "bilateral") {
        p_value <- pnorm(abs(z), lower.tail = FALSE) * 2
        z_critical <- qnorm(alpha/2, lower.tail = FALSE)

        # Evaluación de la condición >
    } else if (alt == ">") {
        p_value <- pnorm(z, lower.tail = FALSE)
        z_critical <- qnorm(alpha, lower.tail = FALSE)

        # Evaluación de la condición <
    } else if (alt == "<") {
        p_value <- pnorm(z, lower.tail = TRUE)
        z_critical <- qnorm(alpha, lower.tail = TRUE)

    }

    datos <- c(p1, p2, z, z_critical, alpha, p_value)
    names(datos) <- c("p1", "p2", "z", "z_critical", "alpha", "p_value")
    return(datos)
}
```

8.1 Análisis visual

```
# total de atletas
athletes_female <- length(gpa[gpa$athlete == TRUE & gpa$female == TRUE, ]$female)
non_athletes_female <- length(gpa[gpa$athlete == FALSE & gpa$female == TRUE, ]$female)

# total de la muestra menos los atletas
athletes_male = length(gpa[gpa$athlete == TRUE & gpa$female == FALSE, ]$female)
non_athletes_male = length(gpa[gpa$athlete == FALSE & gpa$female == FALSE, ]$female)

# definición de los vectores para el Pie chart
proportions <- c(athletes_female, non_athletes_female)
labels <- c("Female & Athelte", "Female & Non-Athelte")

# Definición del gráfico
pie(proportions, labels = labels, main = "Proporción de mujeres atletas")
```


```r
# definición de los vectores para el Pie chart
proportions <- c(athletes_male, non_athletes_male)
labels <- c("Male & Athelte", "Male & Non-Athelte")

# Definición del gráfico
pie(proportions, labels = labels, main = "Proporción de hombres atletas")
```



8.2 Pregunta de investigación

¿La proporción de hombres atletas es mayor que la proporción entre las mujeres?

8.3 Hipótesis nula y la alternativa

Hipótesis nula:

$$H_0 : p_{athletes\_female} \geq p_{athletes\_male}$$

Hipótesis alternativa:

$$H_1 : p_{athletes\_female} < p_{athletes\_male}$$

8.4 Justificación del test a aplicar

Se realiza el contraste de proporciones a través de la función *contraste_proporciones_bidist* del comienzo del apartado, cuya firma establece los parámetros necesarios para poder realizar el procedimiento, devolviendo como resultado los valores que nos permitan rechazar la hipótesis nula o no. Como se tiene que estudiar la proporción en dos distribuciones distintas, independientes y diferentes tamaños, se utiliza este test específico de proporciones de dos distribuciones.

8.5 Realizad los cálculos del test

```
n1 <- athletes_male + non_athletes_male
n2 <- athletes_female + non_athletes_female

p1 <- athletes_male/n1
p2 <- athletes_female/n2

datos <- contraste_proporciones_bidist(p1, p2, n1, n2, "<", 95)
print(t(t(datos)))
```

```
##                   [,1]
## p1          0.06543698
## p2          0.02419355
## z           6.24196417
## z_critical -1.64485363
## alpha       0.05000000
## p_value     1.00000000
```

8.6 Interpretación del test

Tal como podía preverse a partir de las gráficas del apartado 1, la proporción de hombres atletas es mayor que la de mujeres atletas, lo que se ve confirmado a través del test de contraste de proporciones anterior, cuyo *p_value* es mayor que el *alpha*, por lo que no podemos rechazar la hipótesis nula puesto que el error que cometeríamos sería significativamente mayor.

9 Resumen y conclusiones

N	Pregunta	Resultado	Conclusión
P1	¿Cuál es el intervalo de confianza de la nota entre los estudiantes?	*sat*: [1030.028 1030.634], [1030.188 1030.474], *colgpa*: [2.652702 2.65556], [2.653456 2.654806]	El intervalo de confianza de *sat* al 95% es [1030.188 1030.474]. El intervalo de confianza de *sat* al 90% es [1030.028 1030.634]. El intervalo de confianza de *colgpa* al 95% es [2.653456 2.654806]. El intervalo de confianza de *colgpa* al 90% es [2.652702 2.65556].
P2	¿Ser atleta influye en la nota?	p_value <- 6.828951e-10, alpha <- 5.000000e-02	Las notas no son iguales. Se rechaza la hipótesis nula.
P3	¿Las mujeres obtienen mejor nota que los hombres?	p_value <- 1.154434e-04, alpha <- 5.000000e-02	La media de la nota de las mujeres es menor igual, a la de los hombres. Se rechaza la hipótesis nula.
P4	¿Hay diferencias significativas en la nota según la raza?	p_value <- 2.214163e-20, alpha <- 5.000000e-02	La población de raza negra tiene mejores notas que aquella de raza blanca. Se rechaza la hipótesis nula.
P5	¿La proporción de atletas en la población es inferior al 5%?	p_value <- 0.17965749, alpha <- 0.05000000	La proporción de atletas es menor al 5%. Se acepta la hipótesis nula.
P6	¿Hay más atletas entre los hombres que entre las mujeres?	p_value <- 1.00000000, alpha <- 0.05000000	La proporción de atletas hombres es mayor al de mujeres atletas. Se acepta la hipótesis nula.

10 Resumen ejecutivo

La masividad en los datos que comúnmente se obtienen en todas las esferas de la sociedad facilita la toma de decisiones y el reconocimiento de situaciones a corregir o a promover. Sin embargo, la información implícita en ellos es la que más valor proporciona, puesto que el conjunto de datos es quien debe responder a las tendencias y no cada caso aislado.

La utilización de métodos estadísticos permite escalar esta obtención de infomación somera a datos intrínsecos, capaces de dibujar una realidad que pudiera ser desoconocida o, al menos, pensada. Es por esta razón que, al utilizar test de verificaciones como los que se utilizan en esta práctica, se es capaz de dar ciertas afirmaciones por buenas bajo un nivel de confianza suficientemente alto.

La pregunta 4, por ejemplo, es capaz de comparar dos medias de series distintas, completamente desbalanceadas, ya que los casos de atletas representan menos del 5% del total de la muestra. Los datos pudieran verse afectados por otros factores que quizá no están siendo tomados en cuenta, sin embargo, bajo las condiciones en las que se establecen, se considera una representación de la población total.

La pregunta 6 que busca confirmar la igualdad de las notas entre la raza blanca y la raza negra, tiene un resultado de lo más peculiar, puesto que en el enunciado se podría dar a entender que, de no ser iguales, pudiera verse sesgado a situaciones sociales no consideradas en este juego datos. Sin embargo la sorpresa llegó cuando en realidad, los test arrojarían resultados que la media de estudiantes de raza negra tiene mejores notas que aquellos de raza blanca. Podríamos, quizás, echar así por tierra la existencia de factores ajenos al contexto o, al menos, asumir que no son lo suficientemente influyentes.

Finalmente, es importante destacar que se considera que esta muestra, sobre la cual se establecen los test que facilitan la información extraída de los datos, es lo suficientemente representativa como para poder generalizar estos resultados frente al resto de la población.