26/04/2023

Prácticas: Probabilidad y Estadística

Luis Fernandez Fernandez - Carlos García Santa - Pablo Aguiño Guío

PRÁCTICA 1

Ejercicio 1

Parte 1: Importación de los datos a Rstudio. Limpieza de los datos.

1. Importe los datos teniendo en cuenta lo siguiente:

- Los datos están separados por ";"
- ˆLos decimales están separados por "."
- ˆLos valores "missings" o valores perdidos están denotados en la base de datos por "NaN".

Para importar los datos usamos la función read.csv2, en los argumentos usamos googleplay.csv que es el nombre del archivo, na.strings "NaN" para indicar que los campos vacíos están definidos con "NaN", y dec "." para indicar que los decimales vienen seguidos de una coma read.csv2("googleplay.csv", header=TRUE, na.strings="NaN", dec="."). La función por defecto establece que los datos están separados por ";".
Obtenemos:

2. Una vez importada la base de datos, defina mediante el código (o función de R) apropiado, cuáles son las variables numéricas. De lo contrario no se podría realizar la práctica.

Con as.character se crea un vector de caracteres con las distintas columnas a convertir a variable numérica, esta conversión se realiza en concreto con as.numeric que convierte dicho vector de caracteres a variable numérica.

```
googleplay$Rating <- as.numeric(as.character(googleplay$Rating))
googleplay$Reviews <- as.numeric(as.character(googleplay$Reviews))
googleplay$Price <- as.numeric(as.character(googleplay$Price))
```

3. Elimine de la base de datos TODAS las FILAS que tengan un "NaN" en alguna de sus columnas. La nueva base de datos ha de tener una dimensión de 8724 filas y 6 columnas (variables).

Con la función na.omit omitimos las las filas que contengan Na que hemos establecido con na.strings="NaN" anteriormente del dataset introducido.

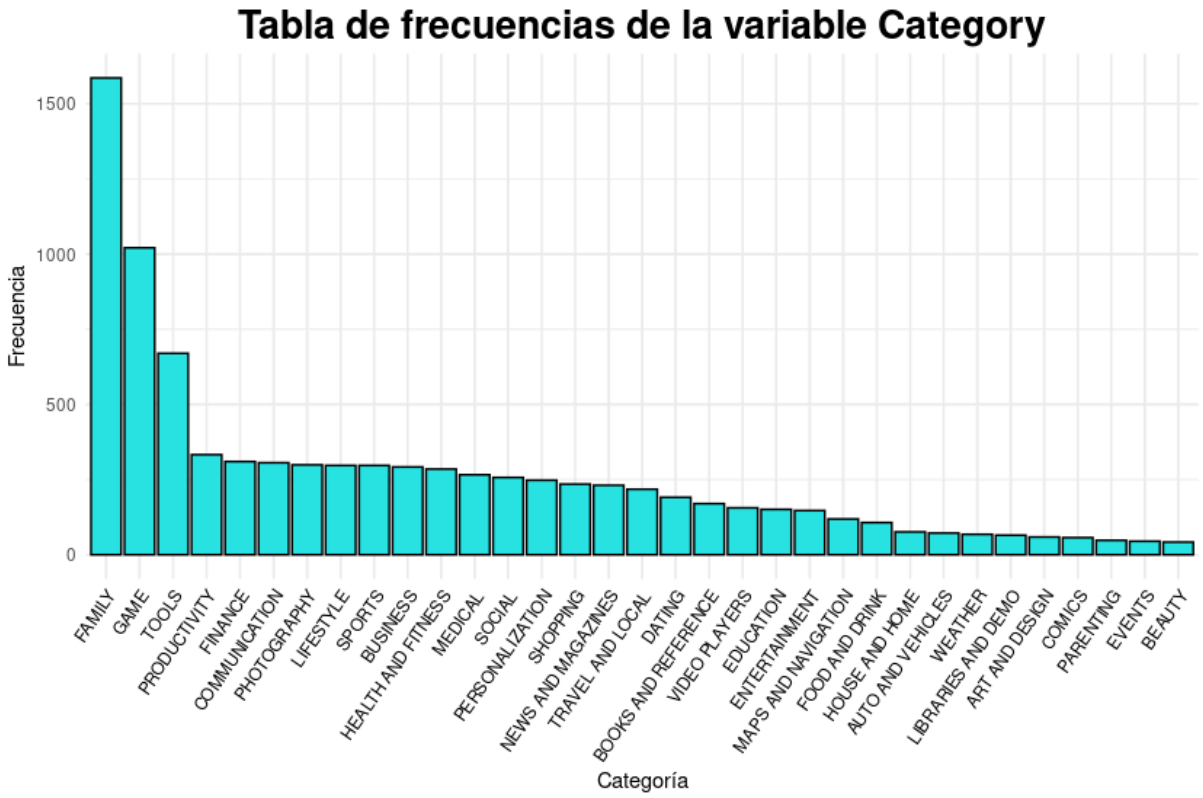
```
googleplay <- na.omit(googleplay)
```

El resultado es el siguiente conjunto de datos:

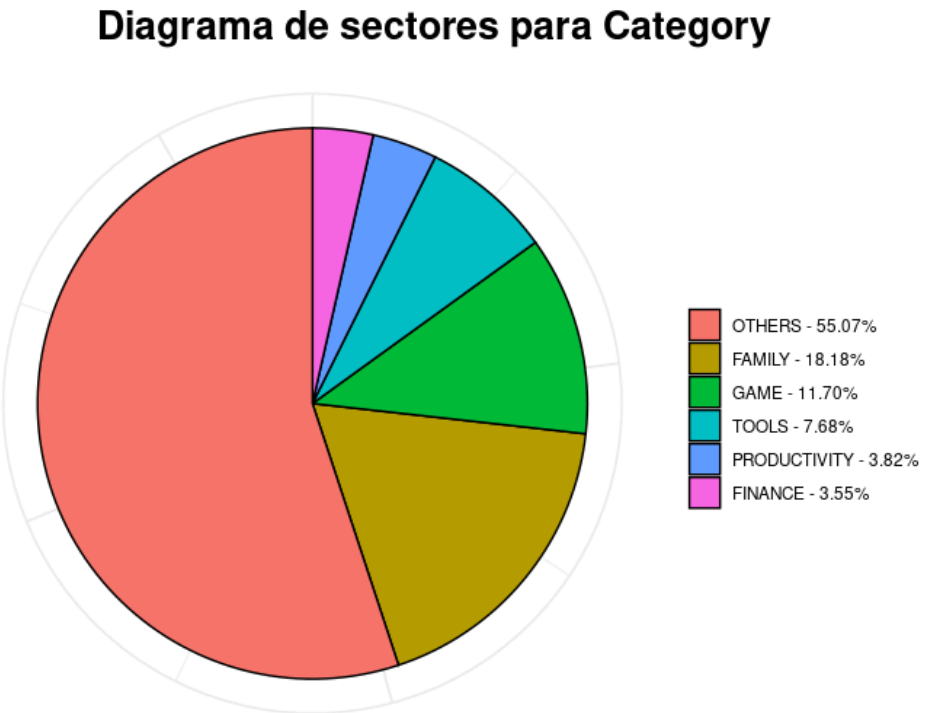
Parte 2: Análisis Descriptivo

1. Obtenga una tabla de frecuencias de la variable Category. Además, obtenga un diagrama de sectores para representar esta información. Para el diagrama de sectores, buscar alguna forma de mejorar el aspecto del gráfico con funciones R.

Para obtener la tabla de frecuencias se ordenan los datos por frecuencia de mayor a menor y se modifica el dataset para ajustarlo en función de la representación visual implementada.



Para la representación en diagrama de sectores modificamos el data frame empleado en la tabla de frecuencias de tal modo que obtenemos un data frame con las categorías principales (con mayor frecuencia) que elijamos y un conjunto de categorías denominado “Otros” que engloba las categorías con menor frecuencia que las principales. Esto se hace debido al elevado número de variables a representar que hay, siendo 33 categorías cuya representación en sectores generaría un diagrama complicado de leer.

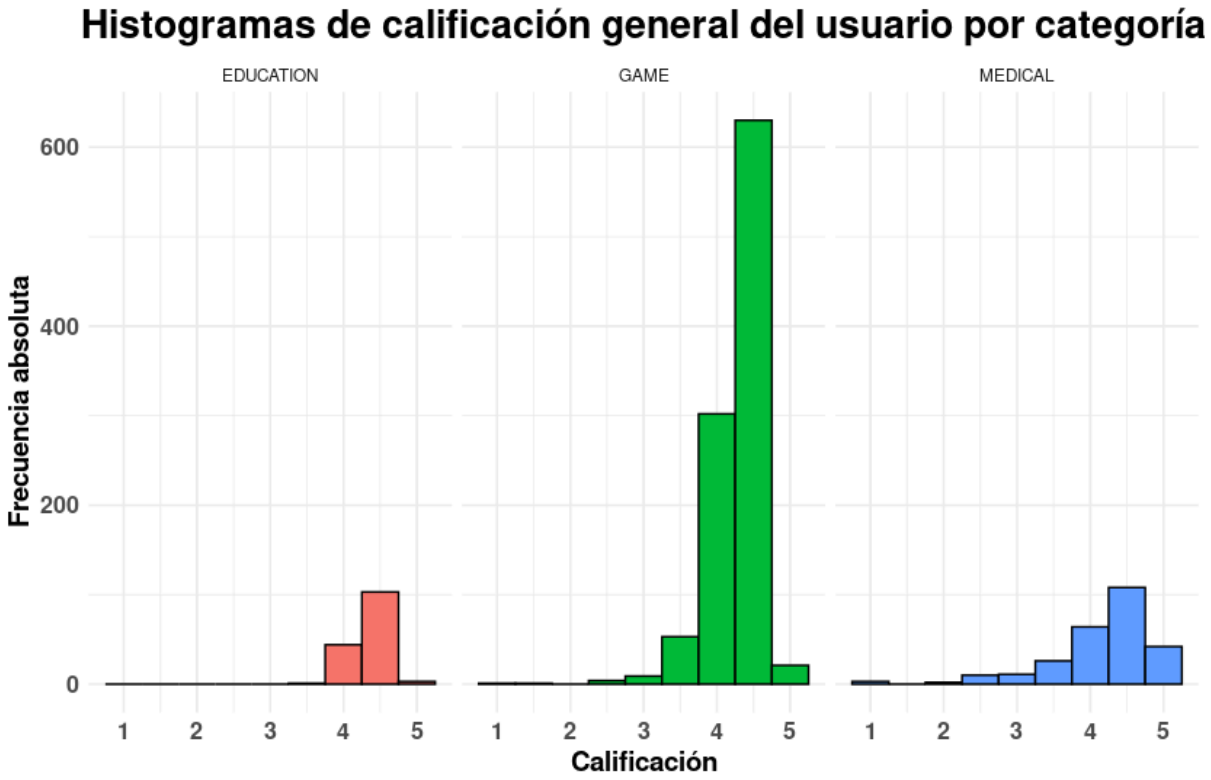


2. Usando la tabla de frecuencias anterior, ¿Cuál Categoría tiene el mayor número de apps? ¿Cuáles son las 5 categorías con menor número de apps? ¿Cuáles son las 5 categorías con más disponibilidad de apps?

La categoría con mayor número de apps es "FAMILY", mientras que las 5 primeras categorías con menor número de apps son de menor a mayor: "BEAUTY", "EVENTS", "PARENTING", "COMICS", "ART AND DESING". Por otra parte de mayor a menor las categorías con mayor disponibilidad de apps son: "FAMILY", "GAME", "TOOLS", "PRODUCTIVITY" y "FINANCE".

Para obtener estos resultados observamos el diagrama de barras según la frecuencia o el diagrama de sectores.

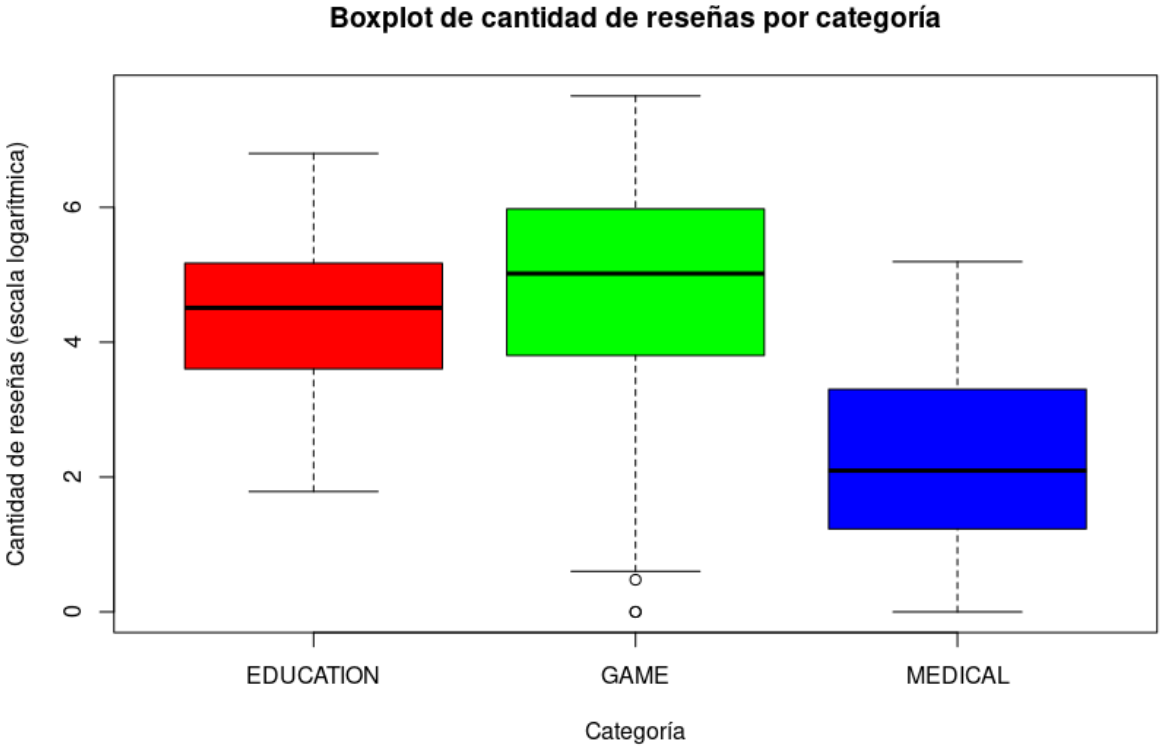
3. Elija 3 Categorías dentro de la variable Category. Obtenga sus histogramas de la calificación general del usuario de la aplicación. En el eje Y del histograma ha de ir la frecuencia absoluta.



4. Con la función de R "subset", construya un nuevo conjunto de datos (un subconjunto de datos) con únicamente los datos de todas las variables pero asociados únicamente a las 3 categorías seleccionadas en el apartado anterior.

Obtenemos usando las categorías GAME, MEDICAL y EDUCATION:

5. Con el subconjunto de datos obtenido anteriormente, construya un gráfico conjunto donde esten los 3 diagramas de caja asociados a las categorıas (eje X) y la cantidad de reseñas. En caso de ser necesario, aplique algún tipo de transformación al eje Y para cambiar la escala de los datos. ¿Hay datos atípicos? ¿Se observa algún tipo de asimetría?



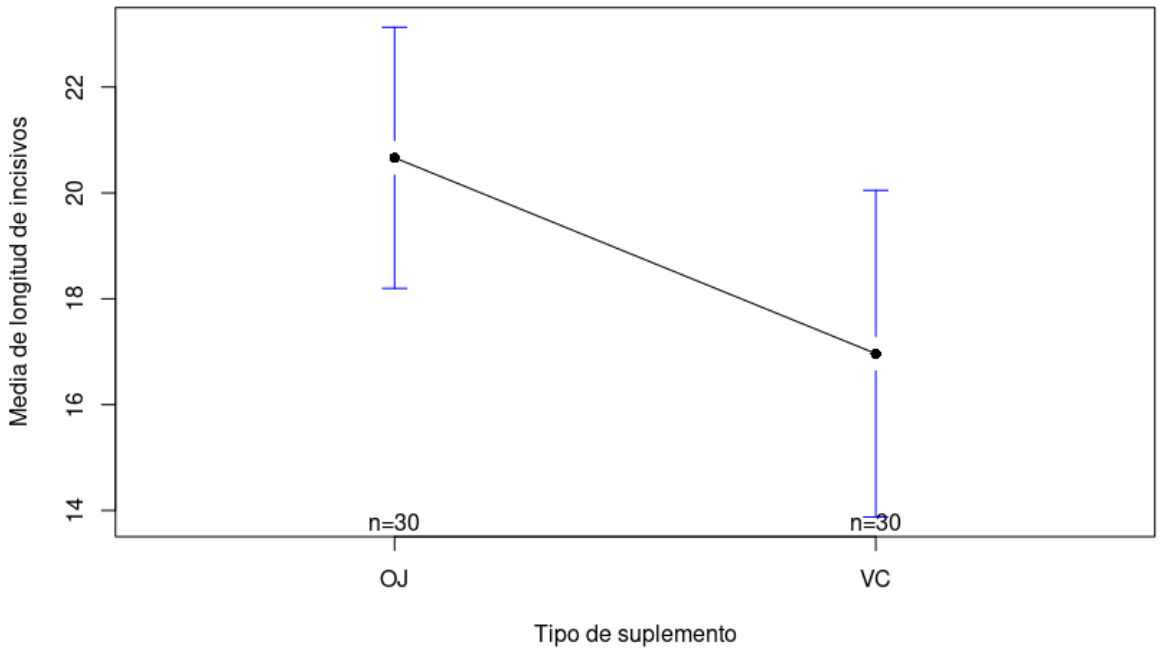
Datos atípicos:
En la categoría "GAME" hay dos puntos fuera de los bigotes y de la barra de error inferior, entonces esos dos puntos se pueden considerar datos atípicos, además pueden corresponder a aplicaciones con un número inusualmente bajo de reseñas en comparación con la mayoría de las aplicaciones en la categoría "GAME".

Asimetría:
En “EDUCATION” la mediana está ligeramente desplazada hacia arriba pero cerca del centro, lo que indica que la distribución es simétrica y que la mayoría de las aplicaciones en esta categoría tienen una cantidad similar de reseñas. En “GAME” la mediana es cercana al centro, por otro lado la mayor distancia entre la barra de error inferior y el bigote inferior indica que hay más variabilidad en la cantidad de reseñas en el extremo inferior de la distribución. Es posible que haya algunas aplicaciones en la categoría "GAME" que tengan un número muy bajo de reseñas en comparación con el resto de las aplicaciones como se ha mencionado anteriormente, y que tenga una pequeña asimetría negativa. En “MEDICAL” la mediana es también muy próxima al centro, por otra parte la mayor distancia entre la barra de error superior y el bigote superior indica que hay más variabilidad en la cantidad de reseñas en el extremo superior de la distribución. Esto puede sugerir que hay algunas aplicaciones en la categoría "MEDICAL" que tienen un número muy alto de reseñas en comparación con el resto de las aplicaciones y que podría tener una ligera asimetría positiva.

Ejercicio 2

Parte 3: Análisis Bivariante

1. Obtenga un diagrama de medias tal que en el eje X estén las categorías de la variable supp y en eje Y las medias de longitud del incisivo. ¿Observa diferencias en la longitud de los incisivos y los dos formatos de vitamina C suministrado a los 60 niños?



Se observa que la media de longitud de incisivos en niños que recibieron zumo de naranja (OJ) como fuente de vitamina C es aproximadamente 21, mientras que para aquellos que recibieron ácido ascórbico (VC) es aproximadamente 17. Estas diferencias en las medias sugieren que podría haber una diferencia en la longitud de los incisivos entre los dos grupos de niños que recibieron diferentes formatos de vitamina C. En promedio, los niños que recibieron vitamina C en forma de zumo de naranja (OJ) parecen tener incisivos más largos que aquellos que recibieron ácido ascórbico (VC).

Sin embargo, para determinar si estas diferencias son estadísticamente significativas, sería necesario realizar pruebas estadísticas adicionales, como la prueba t de Student o la prueba no paramétrica de Wilcoxon-Mann-Whitney, que permiten comparar las medias de dos grupos y determinar si las diferencias observadas son simplemente resultado del azar o si realmente hay una diferencia significativa entre los grupos.

wilcox_test <- wilcox.test(vc_datalen, oj_datalen): El estadístico W obtenido en esta prueba es 324.5 y el valor p es 0.06449. En este caso, el valor p es 0.06449, que es ligeramente mayor que el nivel de significancia comúnmente utilizado de 0.05, esto significa que no podemos rechazar la hipótesis nula y no hay suficiente evidencia para concluir que hay

una diferencia estadísticamente significativa entre las longitudes de los incisivos en niños que recibieron vitamina C en forma de ácido ascórbico (VC) y aquellos que recibieron vitamina C en forma de zumo de naranja (OJ).

Sin embargo, es importante tener en cuenta que el valor p es bastante cercano al nivel de significancia de 0.05, lo que podría sugerir que hay alguna evidencia de una diferencia entre los dos grupos, pero no lo suficiente como para afirmar con seguridad que hay una diferencia significativa en este estudio en particular.

2. Calcular el coeficiente de correlación lineal entre:

- ˆLa longitud de los incisivos y la dosis de vitamina C administrada en formato ácido ascórbico.
- ˆLa longitud de los incisivos y la dosis de vitamina C administrada en formato de zumo de naranja.

Los resultados son para “OJ” y “VC” respectivamente: 0.79 y 0.89 aproximadamente.

```
cor_oj    0.788972536879534
cor_vc    0.892336683537425
```

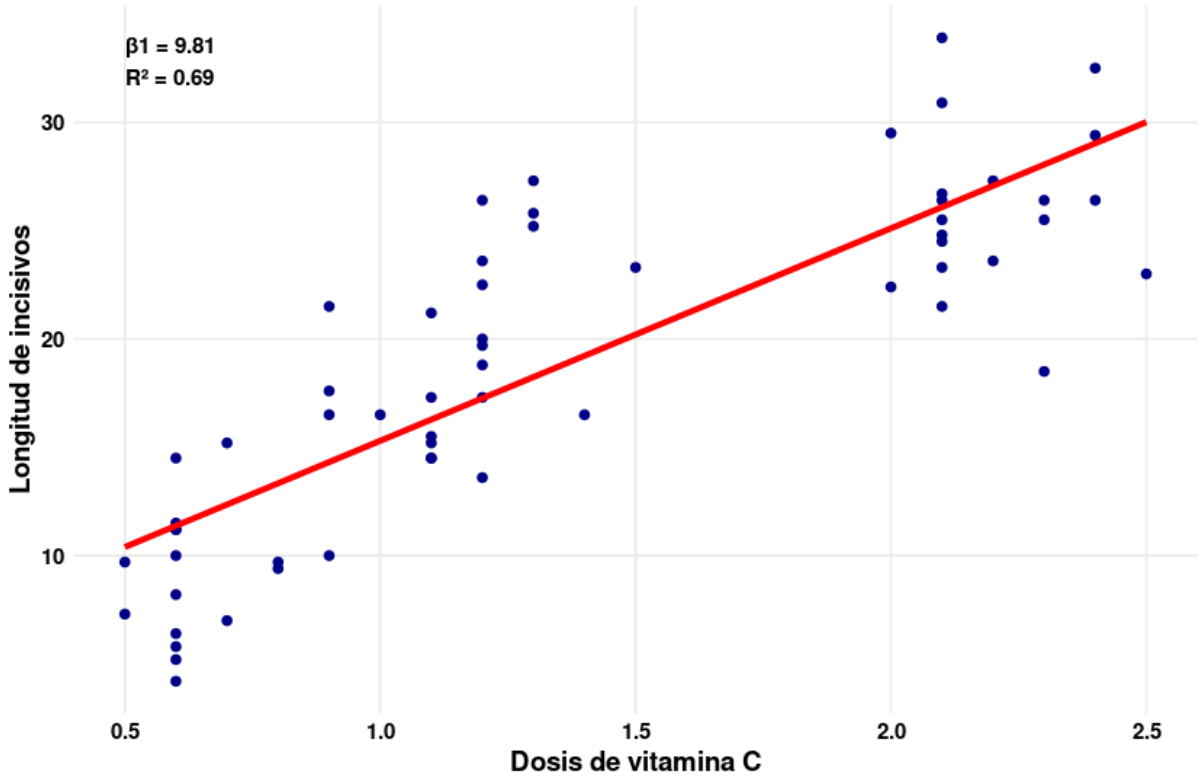
3. En base a los resultados del apartado anterior, ¿qué formato de vitamina C presenta más correlación lineal entre la longitud de los incisivos y la dosis de vitamina C administrada?

Podemos concluir que el formato de vitamina C que presenta una mayor correlación lineal entre la longitud de los incisivos y la dosis de vitamina C administrada es el ácido ascórbico (VC). Esto sugiere que, en este estudio, la relación entre la dosis de vitamina C y la longitud de los incisivos es más fuerte para el ácido ascórbico que para el zumo de naranja. Sin embargo, es importante tener en cuenta que la correlación no implica causalidad y que se deben considerar otros factores y estudios adicionales antes de llegar a conclusiones definitivas.

4. Ajuste una recta de regresión lineal a la longitud de los incisivos en función de la dosis administrada de vitamina C. De una interpretación β1 en el contexto de estos datos. ¿Es bueno el ajuste?

En este estudio, β1 (pendiente de la recta de regresión) es igual a 9.81. Esto significa que, en promedio y asumiendo una relación lineal, por cada unidad de aumento en la dosis de vitamina C administrada (ya sea en formato ácido ascórbico o zumo de naranja), se espera que la longitud de los incisivo en niños aumente en 9.81 unidades.

El coeficiente de determinación (R^2) es 0.69, lo que indica que aproximadamente el 69% de la variabilidad en la longitud de los incisivos se puede explicar por la dosis de vitamina C administrada en el modelo lineal ajustado. En general, un R^2 de 0.69 se considera un ajuste moderado a bueno. Sin embargo, es importante tener en cuenta que la calidad del ajuste también depende del contexto y del propósito del análisis. En algunas situaciones, un R^2 de 0.69 podría considerarse suficientemente bueno, mientras que en otras podría no serlo.

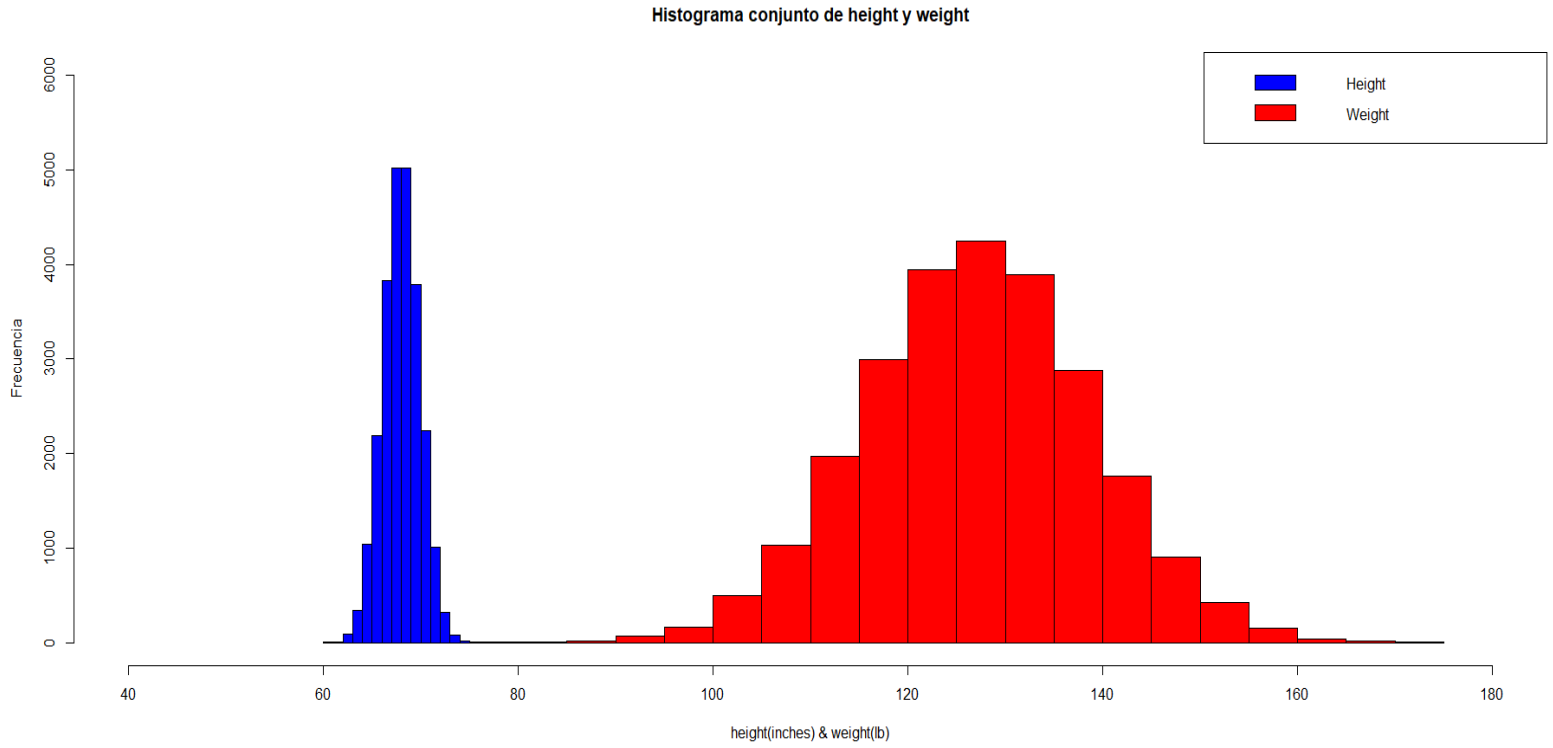

β1 = 9.81
R² = 0.69
Longitud de incisivos
30
20
10
0.5
1.0
1.5
2.0
2.5
Dosis de vitamina C

PRÁCTICA 2

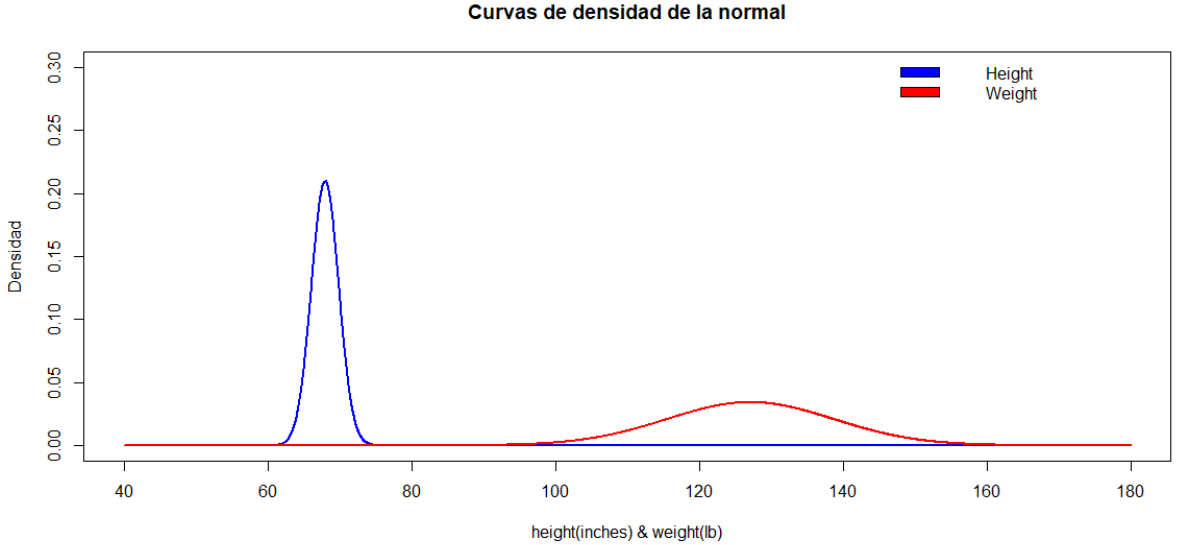
Ejercicio 1

1. Para el peso y la altura obtenga sus histogramas en un mismo gráfico. También sobre los histogramas obtenga la curva de densidad de la normal (use la función "lines" de R). En base a estos histogramas y las curvas de densidad asociadas a cada uno, ¿Que se puede concluir?

Los histogramas representados en un mismo gráfico tienen la siguiente forma:



Las curvas de densidad de la normal:



Conclusión:

Ambas son distribuciones normales simétricas. Sus medias, medianas y moda se encuentran prácticamente en el mismo punto.

Podemos concluir que los valores más comunes, es decir, donde encontramos los picos, sobre la altura, se encuentran entre 67 y 69 pulgadas. Los del peso, entre 120 y 135 libras aproximadamente. Los histogramas están realizados basándonos en una muestra de 25.000 habitantes, por lo que tendrá mayor semejanza con la distribución de la población, al ser de gran tamaño la muestra. No presentan asimetrías ni valores atípicos y el ajuste de la distribución es adecuado.

2. Calcule la media y la desviación estándar para cada una de las variables.

```
> # Imprimir los valores de media y desviación estándar en la pantalla
> cat("Media de height:", mean_height, "\n")
Media de height: 67.99311
> cat("Desviación estándar de height:", sd_height, "\n")
Desviación estándar de height: 1.901679
> cat("Media de weight:", mean_weight, "\n")
Media de weight: 127.0794
> cat("Desviación estándar de weight:", sd_weight, "\n")
Desviación estándar de weight: 11.6609
```

3. En base a los resultados poblacionales del apartado anterior, calcular:

(a) Calcule la probabilidad de que una persona tenga una altura entre 66 y 69 pulgadas.

```
> #SUBAPARTADO 3.A)
> # Calcular la probabilidad de que una persona tenga una altura entre 66 y 69 pulgadas
> prob_3A <- sum(data$height >= 66 & data$height <= 69) / length(data$height)
> # Imprimir la probabilidad en la pantalla
> cat("La probabilidad de que una persona tenga una altura entre 66 y 69 pulgadas es:", prob_3A, "\n")
La probabilidad de que una persona tenga una altura entre 66 y 69 pulgadas es: 0.55488
```

La probabilidad de que una persona mida entre 66 y 69 pulgadas será de 0.55488 o del **55.488%**.

(b) Calcule la probabilidad de que una persona tenga un peso mayor que 134 libras.

```
> #SUBAPARTADO 3.B)
> # Calcular el valor Z correspondiente a 134 libras
> zB <- (134 - mean_weight) / sd_weight
> # Calcular la probabilidad de que una persona tenga un peso mayor que 134 libras
> prob_3B <- pnorm(zB, lower.tail = FALSE)
> # Imprimir la probabilidad en la pantalla
> cat("La probabilidad de que una persona tenga un peso mayor que 134 libras es:", prob_3B)
La probabilidad de que una persona tenga un peso mayor que 134 libras es: 0.276428
```

La probabilidad de que una persona pese más de 134 libras será 0.276428 o del **27.6428%**.

(c) Considere un grupo de 20 personas seleccionadas al azar del municipio de Vizcaya. ¿Cuál es la probabilidad de que al menos 4 tengan una altura de más de 50 pulgadas?

```
> #SUBAPARTADO 3.C)
> # Calcular la probabilidad de éxito
> prob_exito <- sum(data$height > 50) / length(data$height)
> # Calcular la probabilidad de que al menos 4 de 20 personas tengan una altura de más de 50 pulgadas
> prob_3C <- pbinom(3, size=20, prob=prob_exito, lower.tail = FALSE)
> # Imprimir la probabilidad en la pantalla
> cat("La probabilidad de que al menos 4 de 20 personas tengan una altura de más de 50 pulgadas es:", prob_3C)
La probabilidad de que al menos 4 de 20 personas tengan una altura de más de 50 pulgadas es: 1
```

La probabilidad de que en un grupo de 20 personas, 4 o más de ellas midan más de 50 pulgadas es del **100%**

(d) Considere el mismo grupo de 20 personas. ¿Cuál es la probabilidad de que al menos 4 tengan una altura mayor que 70 pulgadas y que al menos 11 midan menos de 70 pulgadas?

```
> #SUBAPARTADO 3.D)
> # Calcular la proporción de personas con altura mayor a 70 pulgadas y menor a 70 pulgadas
> mayor_a_70 <- sum(data$height > 70) / length(data$height)
> menor_a_70 <- sum(data$height < 70) / length(data$height)
> # Calcular la probabilidad binomial acumulada
> prob_4_o_mas_mayores_a_70 <- 1 - pbinom(3, 20, mayor_a_70)
> prob_11_o_mas_menores_a_70 <- 1 - pbinom(10, 20, menor_a_70)
> # Calcular la probabilidad conjunta
> prob_3D <- 0
> for (i in 4:20) {
+   for (j in 11:20) {
+     if (i + j <= 20) {
+       prob_3D <- prob_3D + dbinom(i, 20, mayor_a_70) * dbinom(j, 20 - i, menor_a_70)
+     }
+   }
+ }
> # Imprimir la probabilidad en la pantalla
> cat("La probabilidad de que al menos 4 personas tengan una altura mayor que 70 pulgadas y al menos 11 midan menos de 70 pulgadas en u
n grupo de 20 personas seleccionadas al azar es:", prob_3D)
La probabilidad de que al menos 4 personas tengan una altura mayor que 70 pulgadas y al menos 11 midan menos de 70 pulgadas en un grupo
de 20 personas seleccionadas al azar es: 0.3135422
```

La probabilidad de ambos requisitos será de 0.3135422 o del **31.354%**

(e) Calcule el cuantil 0.10 de la variable peso y el cuantil 0.60 de la variable altura.

```
> #SUBAPARTADO 3.E)
> # Calcular el cuantil 0.10 de la variable peso
> cuantil_peso <- quantile(data$weight, 0.10)
> # Calcular el cuantil 0.60 de la variable altura
> cuantil_altura <- quantile(data$height, 0.60)
> # Imprimir los resultados en la pantalla
> cat("El cuantil 0.10 de la variable peso es:", cuantil_peso, "\n")
El cuantil 0.10 de la variable peso es: 112.1098
> cat("El cuantil 0.60 de la variable altura es:", cuantil_altura)
El cuantil 0.60 de la variable altura es: 68.48009
```

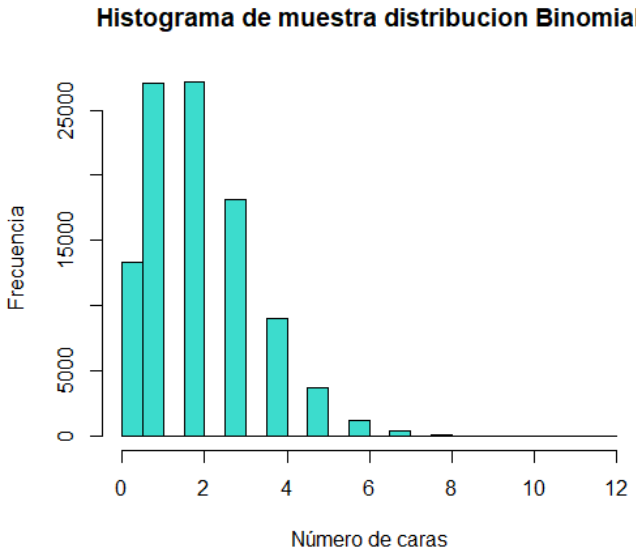
Los cuantiles pedidos serán 112.1098 libras y 68.48009 pulgadas.

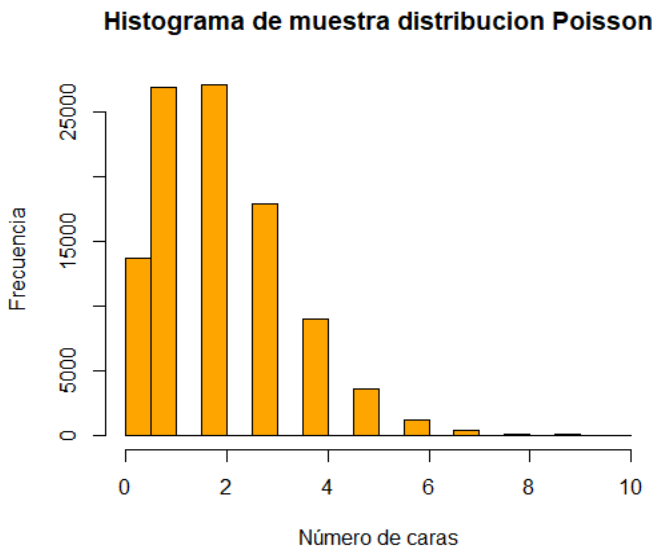
Ejercicio 2

Se lanza 100000 veces una moneda en la cual la probabilidad de salir cara es 0.002. Usando RStudio:

(a) Genere una muestra aleatoria de tamaño 100000 a partir de la distribución Binomial (1000, 0.002). Guarde esa muestra asignando un nombre, de lo contrario se perderá la muestra. Obtenga un histograma de dicha muestra.

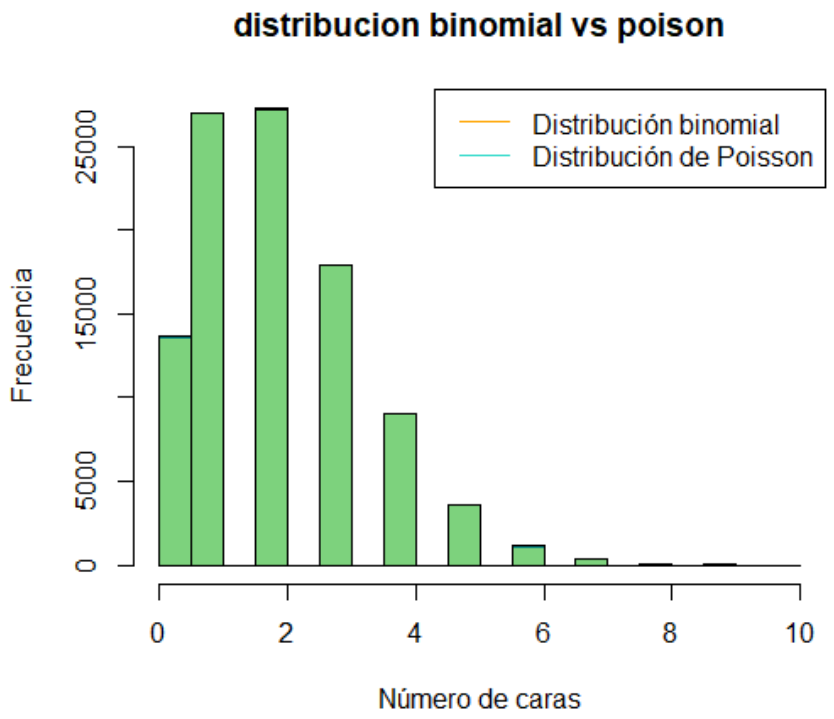
(b) Genere una muestra aleatoria de tamaño 100000 a partir de la distribución Poisson con $\lambda = 2$. Nuevamente, asignar un nombre. Obtenga un histograma.





(c) Compare los dos histogramas anteriores, ¿Qué conclusiones se obtienen? Explique sus conclusiones mediante razones o propiedades teóricas de estas distribuciones.

Al comparar los dos histogramas anteriores, observamos que la forma de la distribución binomial es prácticamente idéntica a la forma de la distribución Poisson. Como se observa en la siguiente gráfica.

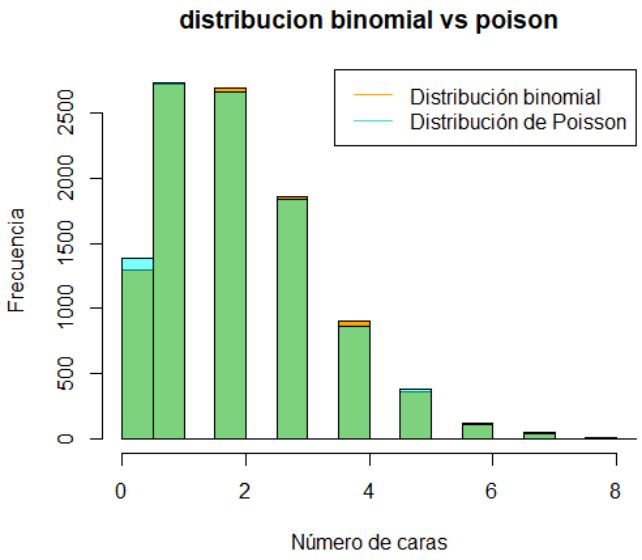


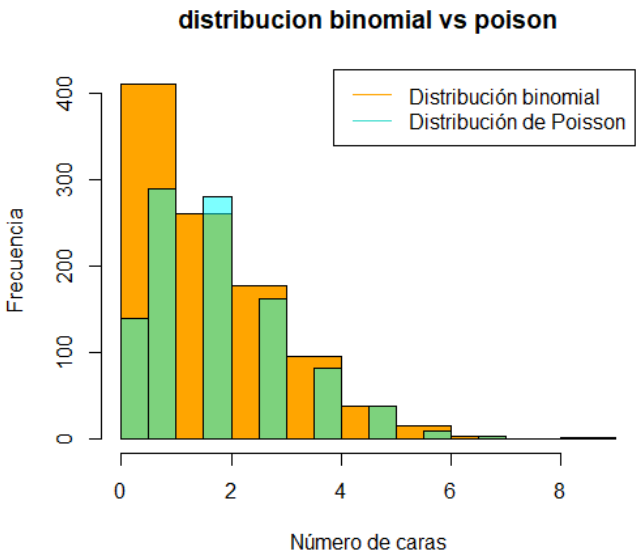
Pensamos que esta semejanza se debe a que la poisson es una aproximación de la distribución binomial. Mientras que normalmente las formas de las distribuciones binomial y de poisson simetrica y asimetrica, respectivamente, suelen ser fáciles de distinguir; en este caso se ha usado una muestra tan grande con número de intentos para la binomial tan grande que la aproximación de la poisson es prácticamente igual.

Esto también sucede porque la poisson trabaja con unos sucesos que se distribuyen homogéneamente a lo largo del tiempo, mientras que la binomial trabaja con sucesos simultáneos.

Así, la distribución poisson con lambda igual a 2 que, vendría a ser lo mismo que decir que el suceso tiene una probabilidad de suceder del 0.002; y la binomial con el valor de probabilidad de 0.002 con el mismo tipo de suceso y el mismo tamaño de muestra. quedan con la misma forma cuando se representan en forma de histograma.

Con valores grandes de las muestra la diferencia es mínima y no es apreciable pero modificando el tamaño de las distribuciones:





Se nota como en cuanto se reduce el número de muestras se desvela la diferencia entre un histograma, y otro. Forma normal de la distribución poisson es sesgada hacia la derecha y la binomial simétrica. Para el caso de usar tamaños de 1000 lanzamientos de moneda apreciamos con claridad las diferencias.