Alexandra González Bermúdez
Estefanía Mora Contreras
Liz Salazar Amaya
José Miguel Rodríguez Gómez

Link al repositorio: https://github.com/lizsalazar17/TrabajoGrupal-CA0204.git

1 Fundamentación del proyecto

1.1 Pregunta de investigación

¿Cómo varía la expresión y la frecuencia de las mutaciones del gen TP53 en el tejido mamario de pacientes y personas en riesgo de desarrollar cáncer de pecho, considerando los distintos grupos etarios? ¿Cómo se relacionan los resultados obtenidos a los registrados en Costa Rica?

2 Objetivos

2.1 Objetivo general

Analizar la expresión y frecuencia de las mutaciones del gen TP53 en el tejido mamario de pacientes y personas en riesgo de desarrollar cáncer de pecho, diferenciando los resultados por grupos etarios.

2.2 Objetivos específicos

- Identificar las mutaciones más frecuentes del gen TP53 presentes en el tejido mamario de los individuos analizados.
- Describir la distribución etaria de los casos diagnosticados con mutaciones en el gen TP53.
- Determinar la relación existente entre los tipos de mutación y la edad al diagnóstico.
- Comparar la prevalencia de casos según las regiones geográficas y consorcios internacionales de origen.
- Formular hipótesis preliminares sobre la posible correlación entre las alteraciones del gen TP53 y la incidencia de cáncer de pecho en Costa Rica, a partir de los registros del Ministerio de Salud.

3 Hipótesis

3.1 Hipótesis general:

La expresión del gen TP53 en el tejido mamario presenta variaciones significativas según el grupo etario, siendo más frecuentes las mutaciones en adultos jóvenes y adultos medios.

Esta tendencia ha sido observada en múltiples estudios donde se reporta que las mutaciones somáticas y germinales de TP53 aparecen con mayor prevalencia en mujeres diagnosticadas a edades tempranas, particularmente menores de 40 años (Evans et al., 2020; Li et al., 2025; Kwong et al., 2020). Asimismo, se ha documentado que aproximadamente el 30 % de los cánceres de mama muestran mutaciones en este gen, lo que justifica su análisis detallado en cohortes estratificadas por edad (Silwal-Pandit et al., 2014; Hwang et al., 2024).

Alexandra González Bermúdez
Estefanía Mora Contreras
Liz Salazar Amaya
José Miguel Rodríguez Gómez

4 Hipótesis específicas:

- H_1: Las mutaciones más frecuentes del gen TP53 se concentran en individuos diagnosticados entre los 30 y 50 años.

Esta hipótesis se fundamenta en el hecho de que varias investigaciones han reportado picos de incidencia y prevalencia de mutaciones TP53 en grupos etarios jóvenes y de mediana edad. Por ejemplo, Li et al. (2025) encontraron prevalencias mayores en pacientes de ≤30 años que en grupos de 31–40 años, y Kwong et al. (2020) reportaron edades de diagnóstico cercanas a los 27–30 años en portadoras de mutaciones germinales. Aunque los rangos exactos varían entre poblaciones, la evidencia coincide en una mayor frecuencia de mutaciones TP53 en edades anteriores a los 50 años, lo cual respalda la formulación de esta hipótesis.

- H_2: Existe una relación significativa entre el tipo de mutación (MUT_ID) y la edad al diagnóstico (Age_at_diagnosis).

Esto coincide con hallazgos que señalan que el espectro mutacional de TP53 no es uniforme y puede asociarse con factores clínicos, incluyendo edad, subtipo tumoral y agresividad (Silwal-Pandit et al., 2014). Además, Hwang et al. (2024) destacan que la distribución de mutaciones TP53 en cáncer de mama presenta patrones clínicos definidos, lo cual sugiere que podrían encontrarse relaciones estadísticamente significativas entre los tipos de mutación (hotspots, missense, truncantes, etc.) y la edad en que se diagnostica el tumor.

- H_3: Los rangos etarios del escenario costarricense que poseen una mayor acumulación de casos, también presentan una mayor diversidad mutacional del gen TP53.

Aunque no existen estudios extensivos de TP53 en poblaciones costarricenses, las tendencias globales muestran que poblaciones con mayor carga de cáncer de mama en ciertos grupos etarios tienden también a presentar variabilidad mutacional significativa en genes clave como TP53 (Evans et al., 2020; Silwal-Pandit et al., 2014). Por lo tanto, extrapolando estas tendencias, es razonable plantear que si en Costa Rica existe un grupo etario con mayor prevalencia de casos, podría observarse también una mayor diversidad mutacional asociada al gen.

Estas hipótesis se generaron de acuerdo a lo estipulado por la evidencia científica existente sobre la distribución etaria de mutaciones del gen TP53 en cáncer de mama, así como por estudios que destacan su variabilidad mutacional y su relevancia clínica en poblaciones jóvenes y de mediana edad (Evans et al., 2020; Li et al., 2025; Silwal-Pandit et al., 2014).

5 Justificación

El estudio del gen TP53 es de vital importancia para la comprensión de los mecanismos moleculares del cáncer de mama. Las mutaciones en este gen pueden conducir a una proliferación celular descontrolada y a la pérdida de la función supresora tumoral, incrementando el riesgo de malignidad. Analizar su expresión por grupos etarios permitirá identificar patrones que podrían asociarse a factores de riesgo biológicos o ambientales específicos.

Además, la comparación con datos del Ministerio de Salud de Costa Rica ofrece la posibilidad de contextualizar los hallazgos en la realidad nacional, contribuyendo al diseño de estrategias de prevención y detección temprana más precisas.

6 Marco metodológico

El presente estudio se desarrollará bajo un enfoque cuantitativo de tipo descriptivo. Su propósito será analizar cómo varía la expresión y la frecuencia de las mutaciones del gen TP53 en el tejido mamario de pacientes y personas en riesgo de desarrollar cáncer de pecho, tomando en cuenta los diferentes grupos etarios. Este diseño permitirá observar las relaciones existentes entre variables genéticas, demográficas y clínicas sin manipularlas directamente, centrándose en la descripción, comparación y correlación de los datos obtenidos (Hernández-Sampieri & Mendoza, 2018).

La investigación se basará en el uso de bases de datos secundarias de carácter público y científico. La principal fuente será la base internacional The TP53 Dataset (versión R21), desarrollada por el Instituto Nacional de Cáncer de los Estados Unidos, la cual contiene información sobre más de 29,000 variantes tumorales y más de 2,000 individuos con mutaciones confirmadas en el gen TP53 (Bouaoun et al., 2016). De esta base se seleccionarán los registros cuya topografía corresponda al tejido mamario, a fin de centrar el análisis en los casos de cáncer de mama. Complementariamente, se considerarán los registros nacionales provenientes del Ministerio de Salud de Costa Rica (2022), con el objetivo de contrastar los hallazgos internacionales con la incidencia local del cáncer de mama.

En una primera etapa, se aplicarán estadísticas descriptivas con el fin de caracterizar la población en estudio. Se calcularán medidas de tendencia central (media, mediana y moda) y de dispersión (desviación estándar y rango intercuartílico) para variables como la edad al diagnóstico y la frecuencia de casos reportados. Asimismo, se elaborarán tablas de frecuencia, histogramas y diagramas de caja, que permitirán visualizar la distribución de las edades, la variación de las mutaciones y la presencia de posibles valores atípicos. Este análisis inicial ayudará a comprender la composición general de la muestra y a detectar patrones preliminares en la distribución de las mutaciones según la edad (Ott & Longnecker, 2016).

Posteriormente, se llevará a cabo un análisis comparativo entre grupos, con el propósito de determinar si existen diferencias significativas en la edad al diagnóstico según el tipo de mutación del gen TP53. Para ello se aplicará una prueba de análisis de varianza (ANOVA), en caso de que los datos cumplan con los supuestos de normalidad y homogeneidad de varianzas (Montgomery, 2017). Si dichos supuestos no se cumplen, se recurrirá a la prueba no paramétrica de Kruskal–Wallis, la cual permitirá contrastar las medianas de edad entre grupos sin asumir una distribución normal (Gibbons & Chakraborti, 2011). Este análisis facilitará identificar si ciertos tipos de mutación se asocian a edades de diagnóstico más tempranas o tardías.

Además, se aplicarán análisis de correlación con el fin de evaluar la relación existente entre la edad al diagnóstico (Age_at_diagnosis) y el número de casos reportados por los principales consorcios internacionales (TCGA, ICGC y GENIE). Dependiendo de la distribución de los datos, se utilizarán los coeficientes de correlación de Pearson (para datos normales) o Spearman (para datos no normales) (Benesty et al., 2009). Estos análisis permitirán establecer si la edad al diagnóstico se relaciona de manera significativa con la frecuencia o severidad de las mutaciones del gen TP53, aportando evidencia empírica a la hipótesis de que la edad influye en la expresión de dicho gen.

Asimismo, se llevará a cabo un análisis de frecuencias y proporciones para determinar la incidencia relativa de las mutaciones por grupo etario y por tipo de mutación. Esta técnica permitirá identificar cuáles grupos de edad presentan mayor prevalencia de mutaciones en el gen TP53, lo que resultará clave para establecer patrones de riesgo y posibles implicaciones clínicas (Agresti, 2018).

Antes de aplicar los procedimientos inferenciales, se realizará una limpieza y depuración de los datos (Esta se trabajó previamente en la Bitácora 2). Este proceso incluirá la identificación de valores faltantes, que serán tratados mediante eliminación o imputación según su relevancia; la detección de valores atípicos

a través del rango intercuartílico (IQR), para garantizar la coherencia interna del conjunto de datos (Field, 2018). Este paso será fundamental para asegurar la calidad y fiabilidad de los análisis posteriores.

De este modo, cada método estadístico aportará evidencia específica para responder la pregunta de investigación:

- Los análisis descriptivos permitirán conocer cómo se distribuyen la edad y las mutaciones del gen TP53.
- Las pruebas de comparación (ANOVA/Kruskal–Wallis) identificarán si existen diferencias significativas en la edad de diagnóstico según el tipo de mutación.
- Los análisis de correlación mostrarán si existe una relación entre la edad y la frecuencia de mutaciones reportadas.
- Finalmente, el análisis de frecuencias permitirá observar la incidencia relativa de las mutaciones en cada grupo etario.

Posteriormente, se llevará a cabo un análisis comparativo entre regiones geográficas, tomando como base la variable Country de la base internacional. Se describirá la distribución de mutaciones del gen TP53 en diferentes países y regiones, con especial énfasis en América Latina. Esta información se contrastará cualitativamente con la base de datos del Ministerio de Salud de Costa Rica, para explorar si los patrones internacionales guardan relación con las tasas de incidencia y mortalidad por cáncer de mama en la población costarricense. Aunque no se establecerán correlaciones estadísticas directas entre ambas bases debido a la diferencia en el tipo de información, este ejercicio permitirá contextualizar los resultados globales en el ámbito nacional, y formular hipótesis sobre la posible prevalencia de mutaciones similares en el país.

7 Construcción del código

8 Librerías

```
library(readr)

library(dplyr)

library(ggplot2)

library(DescTools)

library(tidyr)

library(readxl)

library(here)
```

9 Bases de datos

```
url <- "https://raw.githubusercontent.com/eetefy2311/TP53-data/refs/heads/main/GermlineDownloa

temp.file <- tempfile(fileext = ".csv")

download.file(url, destfile = temp.file, mode = "wb")

base.de.datos <- read_csv(temp.file)

base.datos.cr.masculino <- read_excel(here("data", "base_cr.xlsx"))

base.datos.cr.femenino <- read_excel(here("data", "base_cr.xlsx"), 2)

base.datos.cr.femenino.provincias <- read_excel(here("data", "base_cr.xlsx"), 3)

View(base.datos.cr.femenino)
View(base.datos.cr.femenino.provincias)
View(base.datos.cr.masculino)

base.de.datos.filtrada <- base.de.datos %>%
  select(Country, Population, Region, Germline_mutation,
         Individual_ID, MUT_ID, TCGA_ICGC_GENIE_count,
         Sex,Age_at_diagnosis, Topography, Short_topo, Morphology,
         Dead, Age) %>%
  filter(Topography == "BREAST")
```

10 Datos descriptivos de la base internacional

Frecuencias de MUT_ID

```
frecuencias.mutaciones <- base.de.datos %>%
  group_by(MUT_ID) %>%
  summarise(
    frecuencia = n(),
    porcentaje = round((n()/nrow(base.de.datos))*100, 3)
  ) %>%
  arrange(desc(frecuencia))
```

Las diez mutaciones con mayor frecuencia

```
top.mutaciones <- frecuencias.mutaciones %>%
  head(10)
```

10.1 Estadísticas Descriptivas de la Edad

Edad promedio: 34 Edad mediana: 33 Edad máxima: 94 Edad mínima: 1

Primer cuartil: 19 Segundo cuartil: 33 Tercer cuartil: 47

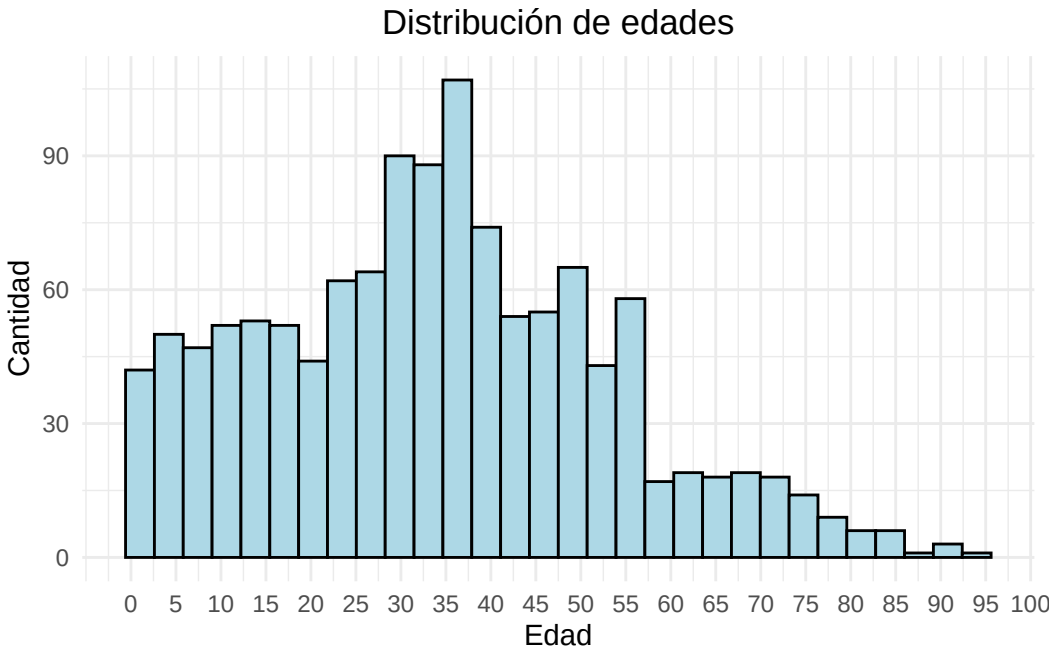
Desviación estándar: 19.2 Varianza: 369 IQR: 28

11 Gráficos

Histograma sobre la distribución de edades

```
base.de.datos %>%
  filter(!is.na(Age)) %>%
  ggplot(aes(x = Age)) +
  geom_histogram(color = "black", fill = "lightblue") +
  scale_x_continuous(breaks = seq(0,100,5)) +
  labs(title = "Distribución de edades",
       x = "Edad",
       y = "Cantidad") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

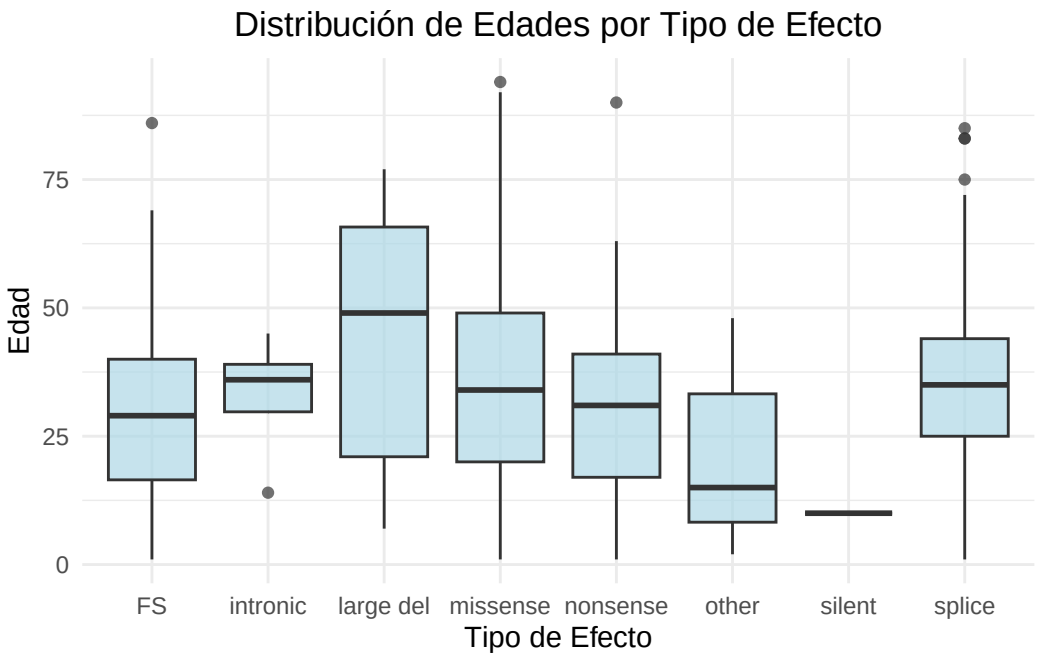
Alexandra González Bermúdez
Estefanía Mora Contreras
Liz Salazar Amaya
José Miguel Rodríguez Gómez



Boxplot de Edades con respecto al Efecto de Mutación

```
base.de.datos %>%
  filter(!is.na(Age), !is.na(Effect)) %>%
  ggplot(aes(x = Effect, y = Age)) +
  geom_boxplot(fill = "lightblue", alpha = 0.7) +
  labs(title = "Distribución de Edades por Tipo de Efecto",
       x = "Tipo de Efecto",
       y = "Edad") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

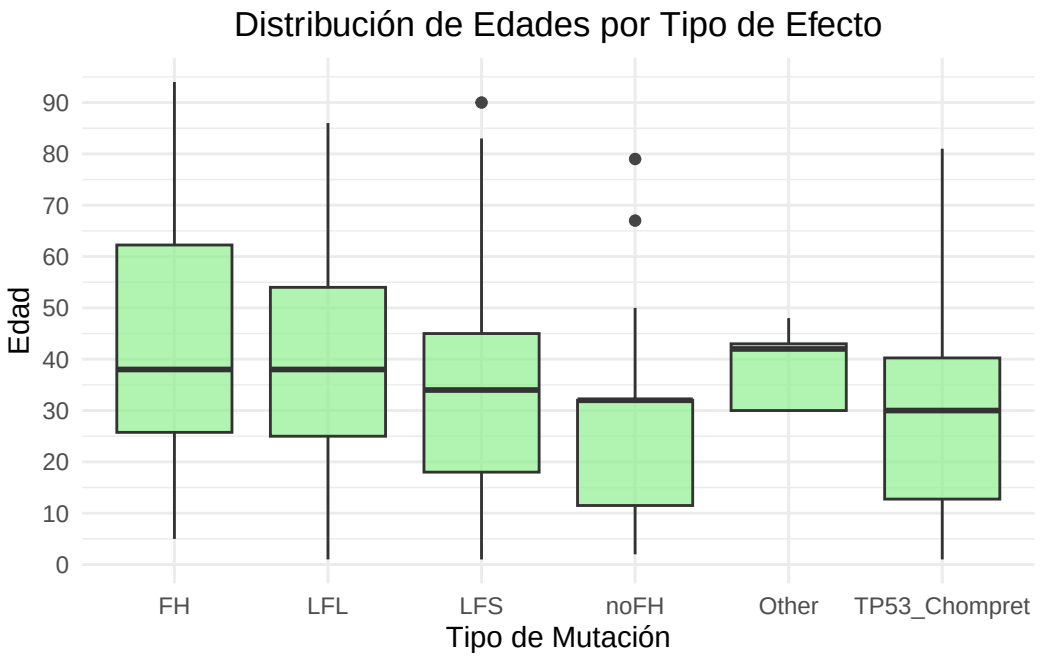
Alexandra González Bermúdez
Estefanía Mora Contreras
Liz Salazar Amaya
José Miguel Rodríguez Gómez



Boxplot de Edades con respecto a la Clase de Mutación

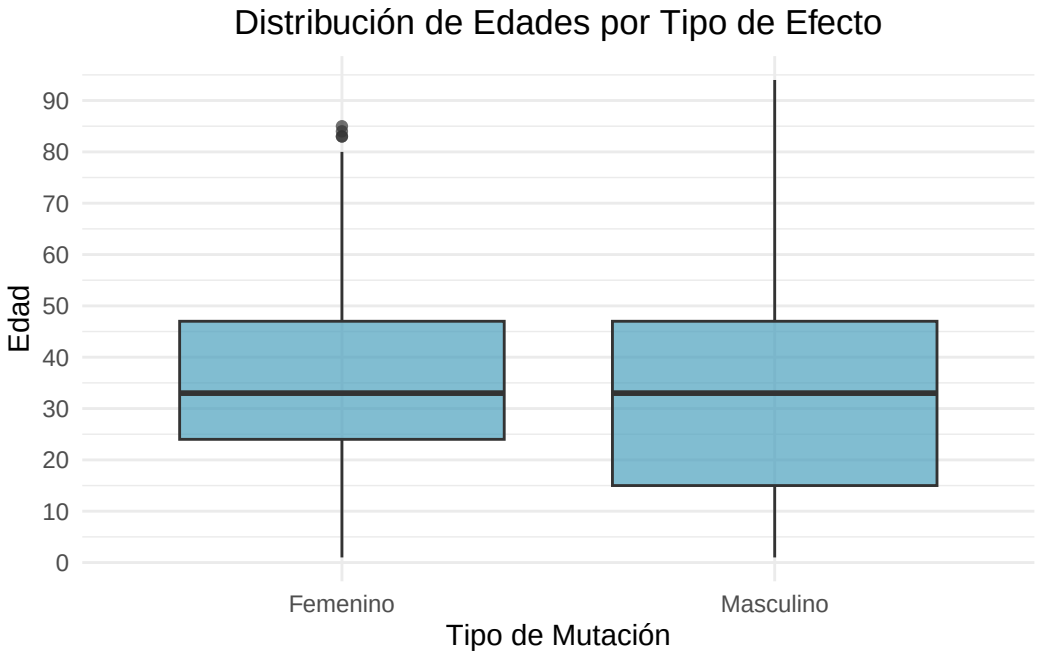
```
base.de.datos %>%
  filter(!is.na(Age), !is.na(Class)) %>%
  ggplot(aes(x = Class, y = Age)) +
  geom_boxplot(fill = "lightgreen", alpha = 0.7) +
  scale_y_continuous(breaks = seq(0,100,10)) +
  labs(title = "Distribución de Edades por Tipo de Efecto",
       x = "Tipo de Mutación",
       y = "Edad") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

Alexandra González Bermúdez
Estefanía Mora Contreras
Liz Salazar Amaya
José Miguel Rodríguez Gómez



Boxplot de Edades y Sexo

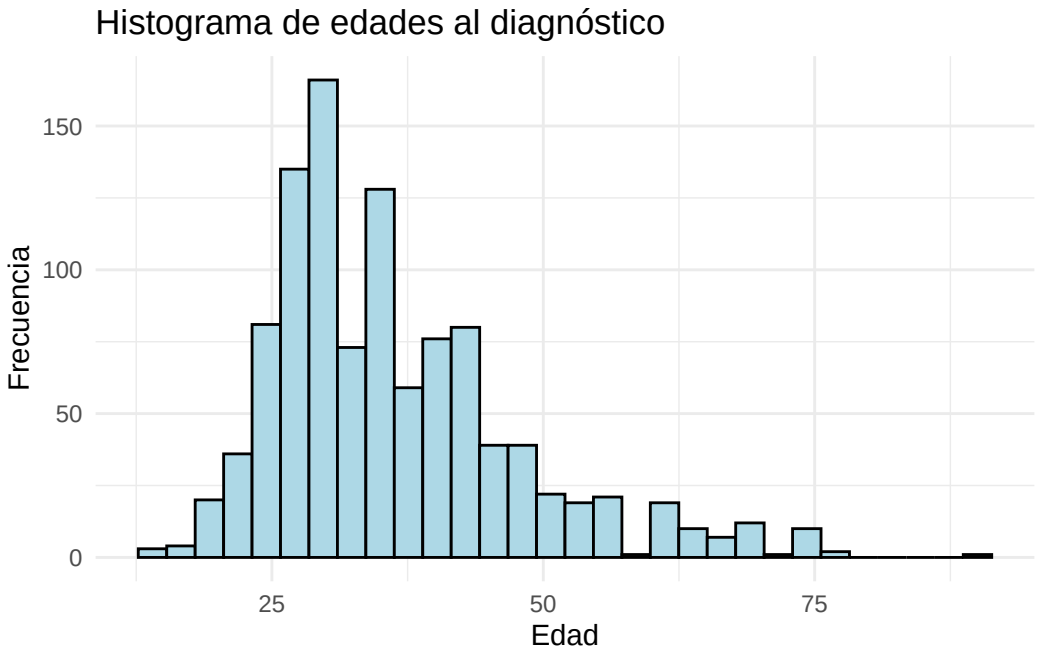
```
base.de.datos %>%
  filter(!is.na(Age), !is.na(Sex)) %>%
  mutate(Sexo = ifelse(Sex == "F", "Femenino", "Masculino")) %>%
  ggplot(aes(x = Sexo, y = Age)) +
  geom_boxplot(fill = "#4AA0BD", alpha = 0.7) +
  scale_y_continuous(breaks = seq(0,100,10)) +
  labs(title = "Distribución de Edades por Tipo de Efecto",
       x = "Tipo de Mutación",
       y = "Edad") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



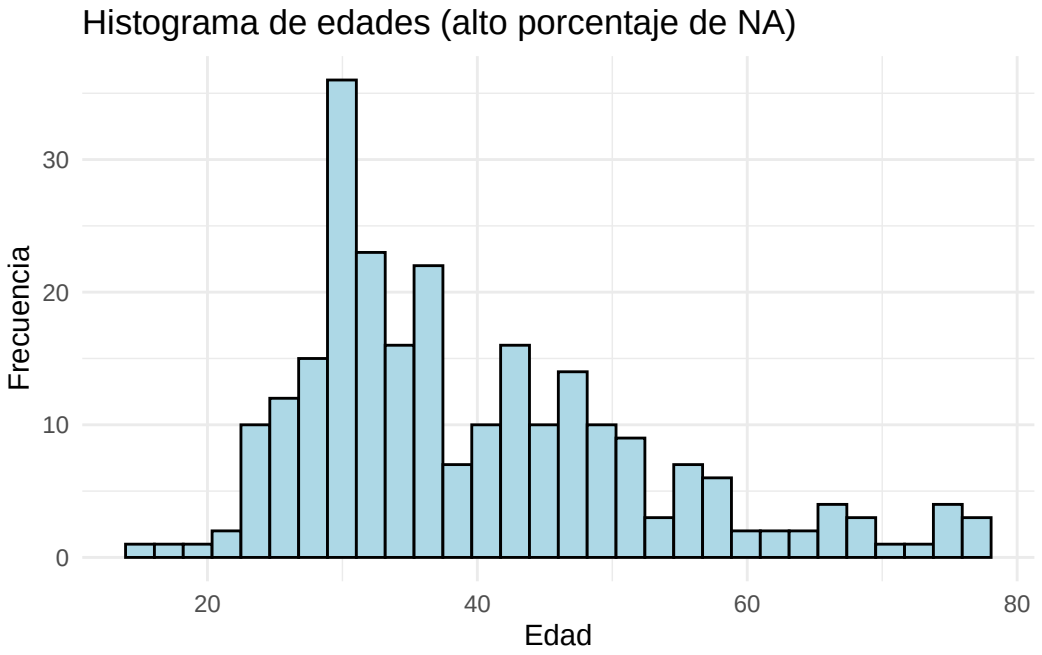
Histogramas y scatterplots de la correlación entre la edad y edad de diagnóstico y cantidad de casos reportados por consorcios internacionales

```
#Datos a utilizar para el proceso de correlación

cor.data.diagnostico <- base.de.datos.filtrada %>%
  filter(!is.na(Age_at_diagnosis), !is.na(TCGA_ICGC_GENIE_count)) %>%
  select(Age_at_diagnosis, TCGA_ICGC_GENIE_count)

cor.data <- base.de.datos.filtrada %>%
  filter(!is.na(Age), !is.na(TCGA_ICGC_GENIE_count)) %>%
  select(Age, TCGA_ICGC_GENIE_count)

#Histogramas de las edades
cor.data.diagnostico %>%
  ggplot(aes(x = Age_at_diagnosis)) +
  geom_histogram(color = "black", fill = "lightblue", bins = 30) +
  labs(title = "Histograma de edades al diagnóstico",
       x = "Edad",
       y = "Frecuencia") +
  theme_minimal()
```

Alexandra González Bermúdez
Estefanía Mora Contreras
Liz Salazar Amaya
José Miguel Rodríguez Gómez




```
#ggsave("../res/Histograma_edad_diag.png", width = 8, height = 6, dpi = 900)

cor.data %>%
  ggplot(aes(x = Age)) +
  geom_histogram(color = "black", fill = "lightblue", bins = 30) +
  labs(title = "Histograma de edades (alto porcentaje de NA)",
       x = "Edad",
       y = "Frecuencia") +
  theme_minimal()
```




```
#ggsave("../res/Histograma_edad.png", width = 8, height = 6, dpi = 900)
```

Prueba de normalidad de la correlación de la edad y edad de diágnostico y el conteo de casos de cáncer por consorcios internacionales utilizando el Shapiro-Wilk test.

```
shapiro.test(cor.data$Age)
```

```
    Shapiro-Wilk normality test

data:  cor.data$Age
W = 0.92968, p-value = 1.334e-09
```

```
shapiro.test(cor.data.diagnostico$Age_at_diagnosis)
```

```
    Shapiro-Wilk normality test

data:  cor.data.diagnostico$Age_at_diagnosis
W = 0.91686, p-value < 2.2e-16
```

```
shapiro.test(cor.data$TCGA_ICGC_GENIE_count)
```

```
    Shapiro-Wilk normality test
```

```
data:  cor.data$TCGA_ICGC_GENIE_count
W = 0.65364, p-value < 2.2e-16
```

```
cor.test(cor.data$Age, cor.data$TCGA_ICGC_GENIE_count,
         method = "spearman")
```

```
    Spearman's rank correlation rho

data:  cor.data$Age and cor.data$TCGA_ICGC_GENIE_count
S = 2501553, p-value = 0.2463
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
0.07315695
```

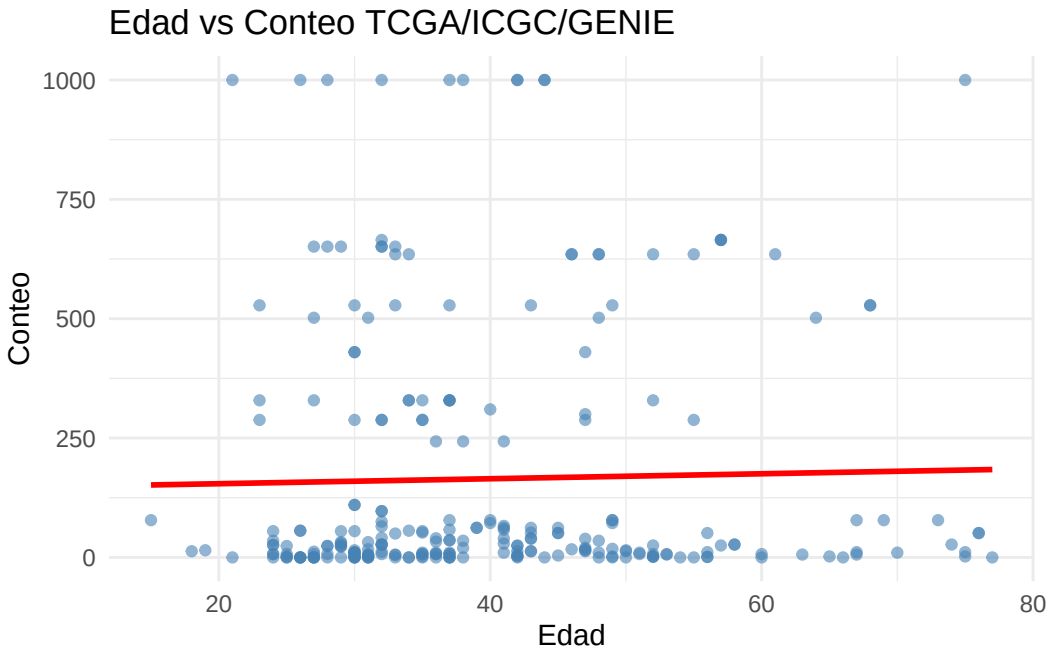
```
cor.test(cor.data.diagnostico$Age_at_diagnosis, cor.data.diagnostico$TCGA_ICGC_GENIE_count,
         method = "spearman")
```

```
    Spearman's rank correlation rho

data:  cor.data.diagnostico$Age_at_diagnosis and cor.data.diagnostico$TCGA_ICGC_GENIE_count
S = 216318513, p-value = 0.01144
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.07750784
```

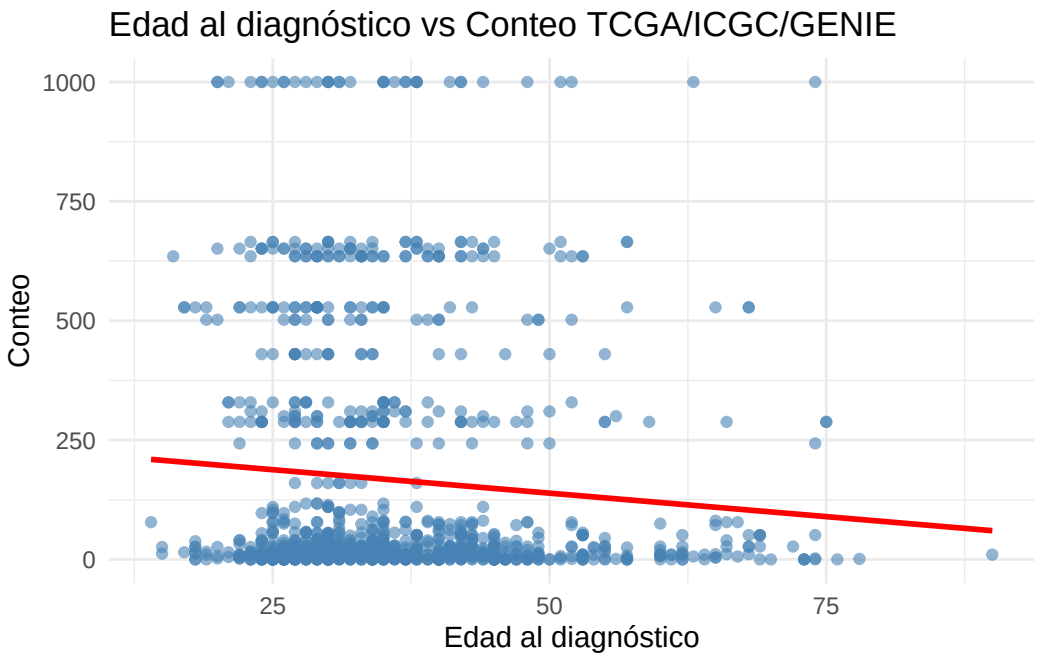
Correlación utilizando método spearman

```
#Gráficos de correlación
cor.data %>%
  ggplot(aes(x = Age, y = TCGA_ICGC_GENIE_count)) +
  geom_point(alpha = 0.6, color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Edad vs Conteo TCGA/ICGC/GENIE",
       x = "Edad",
       y = "Conteo") +
  theme_minimal()
```

Alexandra González Bermúdez
Estefanía Mora Contreras
Liz Salazar Amaya
José Miguel Rodríguez Gómez




```
#ggsave("../res/correlacion_edad.pdf", width = 8, height = 6, dpi = 900)

cor.data.diagnostico %>%
  ggplot(aes(x = Age_at_diagnosis, y = TCGA_ICGC_GENIE_count)) +
  geom_point(alpha = 0.6, color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Edad al diagnóstico vs Conteo TCGA/ICGC/GENIE",
       x = "Edad al diagnóstico",
       y = "Conteo") +
  theme_minimal()
```




```
#ggsave("../res/Correlación_edad_diag.pdf", width = 8, height = 6, dpi = 900)
```

Relación entre la edad del diágnostico y las mutaciones más frecuentes

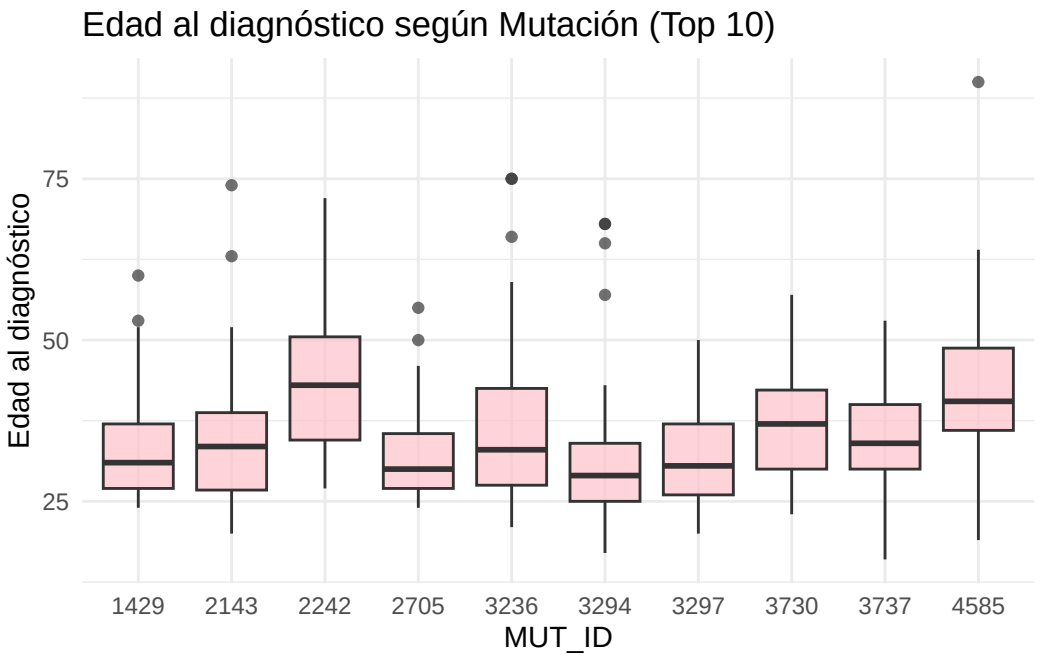
```
#Datos a utilizar para el análisis de la relación entre la edad del diagnóstico y las mutacione

base.infer <- base.de.datos %>%
  filter(Topography == "BREAST")

base.infer <- base.infer %>%
  select(MUT_ID, Age_at_diagnosis) %>%
  filter(!is.na(MUT_ID), !is.na(Age_at_diagnosis))

frecuencias.mutaciones.infer <- base.infer %>%
  group_by(MUT_ID) %>%
  summarise(frecuencia = n()) %>%
  arrange(desc(frecuencia))

top10.infer <- frecuencias.mutaciones.infer %>%
  slice_head(n = 10) %>%
  pull(MUT_ID)


base.infer.2 <- base.infer %>%
  filter(MUT_ID %in% top10.infer)

#Veamos si los datos se distribuyen normalmente
```

```
base.infer.2 %>%
  group_by(MUT_ID) %>%
  summarise(
    n = n(),
    shapiro_p = if (n() >= 3 && n() <= 5000) {
      shapiro.test(Age_at_diagnosis)$p.value
    } else {
      NA_real_
    }
  )
```

```
# A tibble: 10 x 3
   MUT_ID     n  shapiro_p
    <dbl> <int>      <dbl>
 1   1429    21 0.00405
 2   2143    36 0.00271
 3   2242    31 0.0483
 4   2705    20 0.00582
 5   3236    43 0.000180
 6   3294    38 0.00000521
 7   3297    26 0.211
 8   3730    20 0.238
 9   3737    38 0.203
10   4585    42 0.00355
```

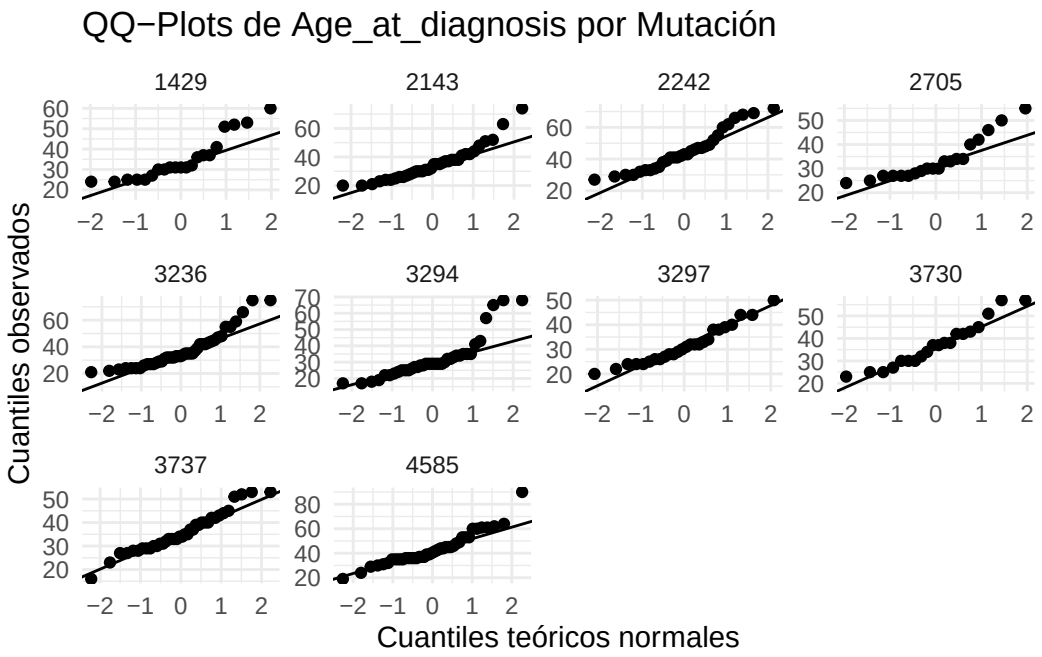
Boxplot de Edad al diagnóstico respecto a las mutaciones más frecuentes

```
ggplot(base.infer.2, aes(x = as.factor(MUT_ID), y = Age_at_diagnosis)) +
  geom_boxplot(fill = "pink", alpha = 0.7) +
  theme_minimal() +
  labs(title = "Edad al diagnóstico según Mutación (Top 10)",
       x = "MUT_ID",
       y = "Edad al diagnóstico")
```



QQ-plots de Edad al diagnóstico con respecto a las mutaciones más frecuentes

```
base.infer.2 %>%
  mutate(MUT_ID = as.factor(MUT_ID)) %>%
  ggplot(aes(sample = Age_at_diagnosis)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap(~ MUT_ID, scales = "free") +
  theme_minimal() +
  labs(title = "QQ-Plots de Age_at_diagnosis por Mutación",
       y = "Cuantiles observados",
       x = "Cuantiles teóricos normales")
```



Aplicación del método Kruskal-Wallis

```
kruskal.test(Age_at_diagnosis ~ MUT_ID, data = base.infer.2)
```

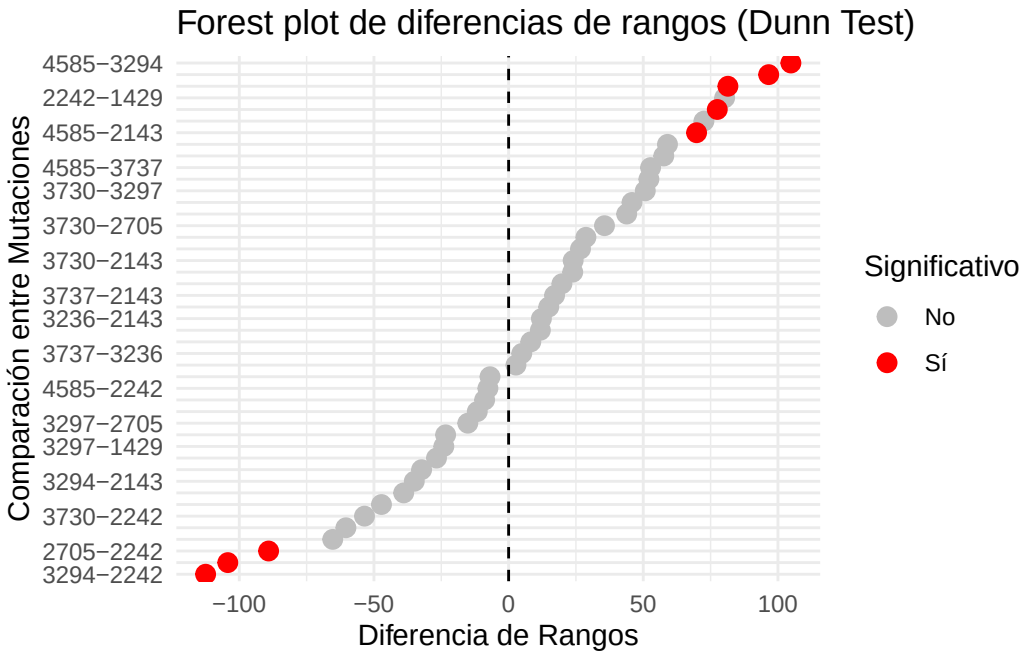
```
    Kruskal-Wallis rank sum test

data:  Age_at_diagnosis by MUT_ID
Kruskal-Wallis chi-squared = 50.525, df = 9, p-value = 8.579e-08
```

```
dunn <- DunnTest(Age_at_diagnosis ~ MUT_ID, data = base.infer.2, method="bonferroni")

df.kruskal <- as.data.frame(dunn[[1]])
df.kruskal$Comparison <- rownames(df.kruskal)
rownames(df.kruskal) <- NULL

df.kruskal <- df.kruskal %>%
  mutate(Comparison = as.character(Comparison),
         Signif = ifelse(pval < 0.05, "Sí", "No"))
```

Forest plot: diferencia de rangos y comparación entre mutaciones

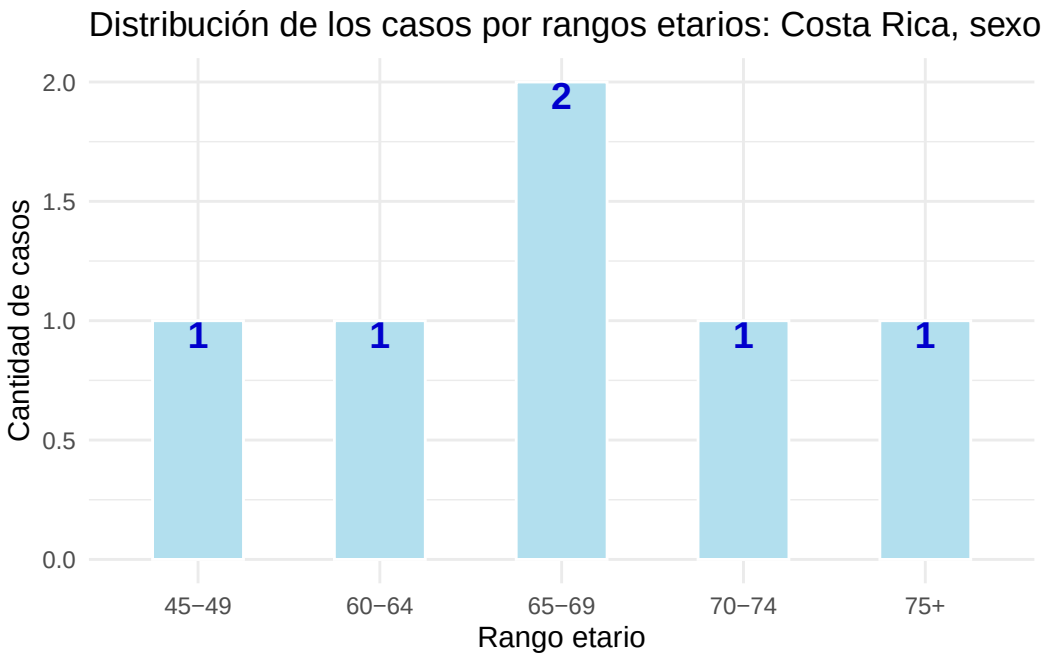
```
ggplot(df.kruskal, aes(x = `mean rank diff`,
                       y = reorder(Comparison, `mean rank diff`),
                       color = Signif)) +
  geom_point(size = 3) +
  geom_vline(xintercept = 0, linetype = "dashed") +
```

```
  scale_y_discrete(guide = guide_axis(check.overlap = TRUE)) +  #   AQUÍ
  theme_minimal() +
  labs(
    title = "Forest plot de diferencias de rangos (Dunn Test)",
    x = "Diferencia de Rangos",
    y = "Comparación entre Mutaciones",
    color = "Significativo"
  ) +
  scale_color_manual(values = c("Sí" = "red","No" = "gray"))
```



Distribución de los casos por rangos etarios Costa Rica, sexo masculino

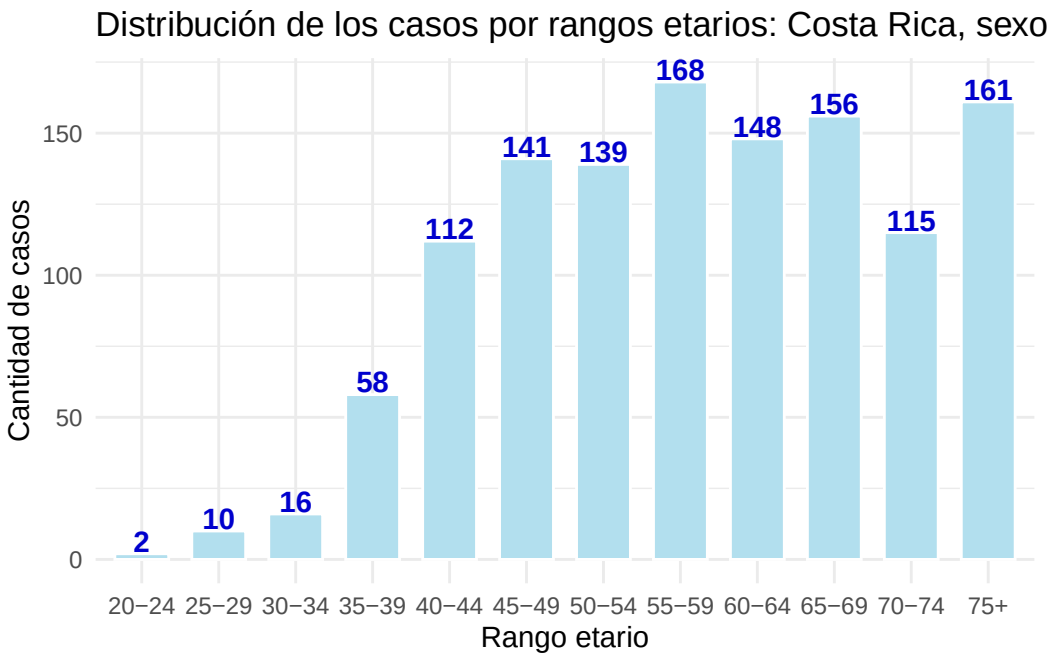
```
ggplot(subset(base.datos.cr.masculino, Cantidad_casos != 0),
       aes(x=Rango_etario, y = Cantidad_casos))+
  geom_bar(stat = "identity", fill = "lightblue2", color = "white", width = 0.5)+
  labs(x = "Rango etario", y = "Cantidad de casos",
       title = "Distribución de los casos por rangos etarios: Costa Rica, sexo masculino")+
  theme_minimal()+
  geom_text(aes(label = Cantidad_casos),
            vjust = 1, color = "blue3", size = 5, fontface = "bold")
```




```
#ggsave("C:/Users/alego/OneDrive/Documentos/Proyecto grupal/TrabajoGrupal-CA0204/res/Distribuci
```

Distribución de los casos por rangos etarios Costa Rica, sexo femenino

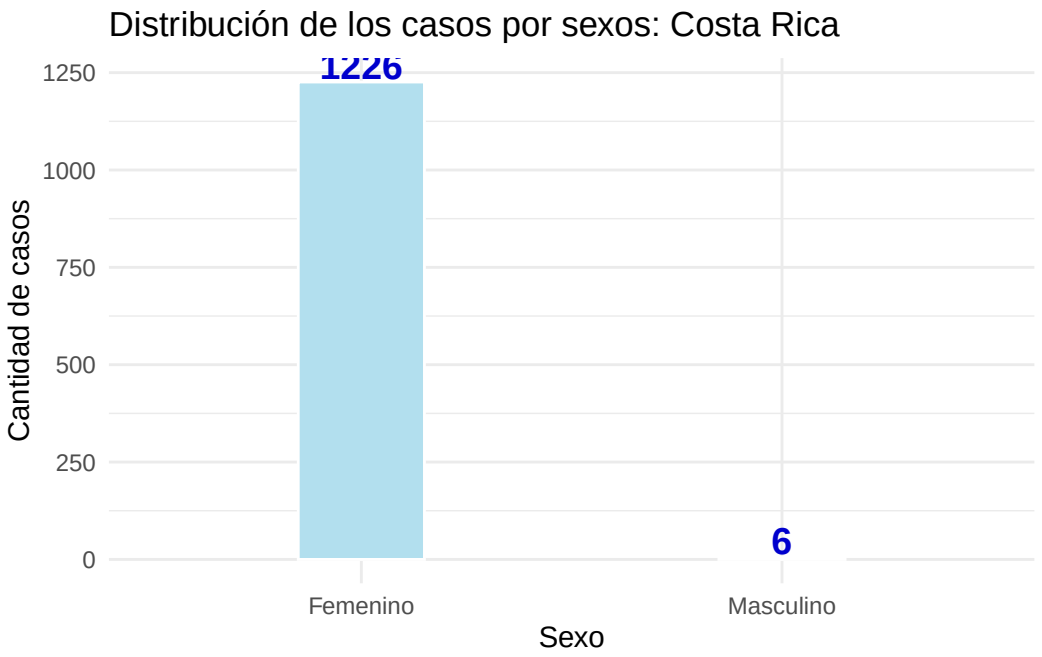
```
ggplot(subset(base.datos.cr.femenino, Cantidad_casos != 0),
       aes(x=Rango_etario, y = Cantidad_casos))+
  geom_bar(stat = "identity", fill = "lightblue2", color = "white", width = 0.7)+
  labs(x = "Rango etario", y = "Cantidad de casos",
       title = "Distribución de los casos por rangos etarios: Costa Rica, sexo femenino")+
  theme_minimal()+
  geom_text(aes(label = Cantidad_casos),
            vjust = -0.1, color = "blue3", size = 4, fontface = "bold")
```




```
#ggsave("C:/Users/alego/OneDrive/Documentos/Proyecto grupal/TrabajoGrupal-CA0204/res/Distribuci
```

Cantidad de casos por sexo: Costa Rica

```
total.femenino <- sum(base.datos.cr.femenino$Cantidad_casos)
total.masculino <- sum(base.datos.cr.masculino$Cantidad_casos)

df.sexos <- data.frame(
            Sexo = c("Femenino", "Masculino"),
            Cantidad_casos = c(total.femenino, total.masculino))

ggplot(df.sexos, aes(x=Sexo, y=Cantidad_casos))+
  geom_bar(stat = "identity", fill = "lightblue2", color = "white", width = 0.3)+
  labs(x = "Sexo", y = "Cantidad de casos",
       title = "Distribución de los casos por sexos: Costa Rica")+
  theme_minimal()+
  geom_text(aes(label = Cantidad_casos),
            vjust = -0.1, color = "blue3", size = 5, fontface = "bold")
```




```
#ggsave("C:/Users/alego/OneDrive/Documentos/Proyecto grupal/TrabajoGrupal-CA0204/res/Distribuci
```

12 Análisis de los hallazgos

13 Discusión y resultados

13.1 Relación de edad al diagnóstico y mutaciones más frecuentes

Al analizar la relación entre la variable Age_at_diagnosis y MUT_ID usando las 10 mutaciones más frecuentes, se utilizó shapiro.test para comprobar si las variables se dstribuyen normalmente. Esto se puede ver en la gráfica 10. Debido a que se presentaron pocos datos que si cumplen con esta condición, se escogió Kruskal-Wallis para determinar si hay una diferencia significativa entre los distintos grupos de mutaciones.

13.2 Evaluación del método Kruskal-Wallis para encontrar diferencias significativas entre edades al diagnóstico de distintos grupos.

Mediante la aplicación de Kruskal-Wallis se detectó un valor menor que 0,05 lo que significa que entre algunos de los grupos analizados hay una diferencia significativa respecto a las edades al diagnóstico. A partir de ellos, se aplicó dunnTest que ayuda a separar pares de grupos para que diferencias hay entre ellos y así detectar los tipos de mutaciones con mayor o menor edad de diagnóstico.

13.3 Diferencia de rangos y comparación entre mutaciones

En este gráfico se presentan diferencias de rangos promedio entre pares de mutacione. Cada punto es una comparación y los colores son determindados de acuerdos a si el valor p es menor a 0,05 (rojos) y o sino (grises), es decir, si se presenta una diferencia estadística significativa.

El eje vertical permite observar que mutación presenta un diagnóstico mayor respecto a la mutación con la que se compara, si la diferencia es positiva, el primera mutación del par es la que tiende a asociarse con una edad de diagnóstico tardío. Si la diferencia es negativa, el primer par es, en cambio, el que se relaciona con una edad temprana de diagnóstico.

Por lo tanto, las mutaciones 4585 y 2242 exponen una tendencia a asociarse con edad más altas de diagnóstico, al compararse con las mutaciones 3294, 3297, 2705 y 2143. Además, se encontró que las mutaciones 2705, 3297 y 3294, se asocian a edades más tempranas al compararse con 2242.

13.4 Distribución de edades y edades al diagnóstico

El primer histograma (Figura 5) muestra que la distribución de la edad al diagnóstico es asimétrica hacia la derecha (sesgo positivo). La mayoría de los casos se concentran entre los 25 y 40 años, con un pico marcado alrededor de los 30 años, lo que indica que este grupo etario es el más común en los registros disponibles.

A partir de los 45–50 años la frecuencia comienza a disminuir de manera notable, y existe una cola extendida hacia edades más avanzadas (hasta alrededor de 75–80 años), pero con muy pocos casos. Esta forma confirma que la distribución no es normal y que la base contiene una gran proporción de diagnósticos en edades relativamente jóvenes.

El segundo histograma (Figura 6), aunque basado en un subconjunto más pequeño debido a la presencia de muchos valores NA (esto debido a la alta mortalidad que presentan los pacientes con cáncer), presenta

una forma muy similar: un claro sesgo hacia la derecha y mayor concentración de casos entre los 25 y 40 años, con un máximo alrededor de los 30 años.

La menor cantidad total de observaciones hace que las barras sean más variables y que la forma no se vea tan suave, pero la tendencia general se mantiene: hay pocos casos en edades muy jóvenes (<20) y muy pocos casos por encima de los 60 años.

Ambos histogramas reflejan una tendencia consistente:

- La mayoría de los diagnósticos se dan en personas entre 25 y 40 años.
- La distribución presenta una cola hacia la derecha, indicando sesgo positivo.
- No se observa una distribución normal.
- La presencia de NA afecta la claridad del segundo histograma, pero no modifica la tendencia central.

Estos patrones apoyan parcialmente la hipótesis de que las alteraciones del gen TP53 en tejido mamario tienden a concentrarse en grupos etarios relativamente jóvenes.

13.5 Evaluación de la normalidad de la correlación entre la edad y edad al diagnóstico con el conteo de casos de cáncer de acuerdo a consorcios internacionales.

Se evaluó la normalidad de las variables Edad (Age), Edad al diagnóstico (Age_at_diagnosis) y el conteo de casos (TCGA_ICGC_GENIE_count) mediante la prueba de Shapiro–Wilk. En los tres casos se obtuvo un valor de p extremadamente pequeño ($p < 0.05$), indicando que ninguna de las variables sigue una distribución normal:

$W_($Age$) = 0.929, p = 1.3 \times 10^{-9}$

$W_($Age_at_diagnnosis$) = 0.916, p < 2.2 \times 10^{-16}$

$W_($TCGA_ICGC_GENIE_count$) = 0.653, p < 2.2 \times 10^{-16}$

Dado que las variables no cumplen el supuesto de normalidad (y considerando además que el conteo es una variable discreta) se utilizó el coeficiente de correlación de Spearman (ρ) para analizar la relación entre la edad y el número de mutaciones reportadas por los consorcios TCGA, ICGC y GENIE.

En primer lugar, la correlación entre la edad (Age) y el conteo combinado de mutaciones (TCGA_ICGC_GENIE_count) no fue significativa:

$\rho = 0.073, p = 0.246$

El valor de ρ, cercano a cero, indica que no existe una relación relevante entre la edad del paciente y la cantidad de mutaciones reportadas por los consorcios.

Por otro lado, se evaluó la correlación entre la edad al diagnóstico (Age_at_diagnosis) y el mismo conteo. En este caso se obtuvo un coeficiente negativo muy pequeño:

$\rho = -0.077, p = 0.011$

Aunque la asociación es estadísticamente significativa, la magnitud del coeficiente es extremadamente baja, lo que indica que la relación (aunque detectable debido al tamaño de la muestra) carece de relevancia práctica. En términos reales, la edad al diagnóstico apenas se relaciona con el número de mutaciones reportadas.

En conjunto, ambos análisis sugieren que ni la edad ni la edad al diagnóstico se asocian de manera consistente o relevante con el conteo de mutaciones TP53 provenientes de los consorcios TCGA, ICGC y GENIE.

Gráficamente, se puede observar mediante Figura 7 y Figura 8.

13.6 Relación entre la edad y edad al diagnóstico con el conteo de mutaciones reportadas por los consorcios internacionales

Los gráficos de dispersión (Figura 7 y 8) muestran visualmente la relación entre la edad y el conteo de mutaciones reportadas por los consorcios TCGA, ICGC y GENIE. En el caso de la variable Age, los puntos se encuentran ampliamente dispersos y la línea de tendencia lineal es casi horizontal, lo que confirma la ausencia de una relación clara entre ambas variables.

Para la variable Age_at_diagnosis, aunque la línea de tendencia presenta una leve pendiente negativa, la dispersión sigue siendo elevada, indicando que la relación es muy débil. Esto coincide con el valor de ρ obtenido en la correlación, que muestra una asociación estadísticamente significativa pero prácticamente irrelevante.

13.7 Incidencia del cáncer de mama en la población costarricense durante el año 2022

13.7.1 Nivel general

Durante el año 2022 la población costarricense presentó una totalidad de 1232 casos de cáncer de mama, lo anterior con una particularidad: a partir del gráfico de la distribución de casos según el sexo para Costa Rica, se determina la existencia de una brecha considerable en la incidencia del cáncer de mama en las personas de sexo masculino respecto a las personas de sexo femenino; los masculinos presentaron únicamente seis casos, mientras que las féminas acumularon una cantidad de 1226 casos. El fenómeno descrito no corresponde a un hecho aislado, tal como menciona Soto Flores (2015) ” el cáncer de mama es la causa líder de muerte en mujeres de países en vías de desarrollo y la segunda causa de muerte en países desarrollados, siendo segundo al cáncer de pulmón” (p. 799). Asimismo, este autor se refiere al escenario nacional: “dentro de la población costarricense el cáncer de mama es la causa más común de mortalidad en mujeres por neoplasia maligna” (p. 800).

Del párrafo precedente se resalta un hecho importante, el cual consiste en que la mayoría de diagnósticos de cáncer de mama en el 2022 se presentaron en el sexo femenino, con un 99.51% del total.

13.7.2 Sexo masculino

Como se mencionó antes, de los 1232 casos solo seis fueron diagnosticados en sujetos de sexo masculino, equivalente a un 0.49% del total. En lo que respecta a la distribución de los casos, considerando el gráfico que describe dicho aspecto de acuerdo a los rangos etarios en el sexo masculino, se determina que la mayoría de casos se situaron en la categoría que comprende de los 65 a los 69 años, con una acumulación de dos elementos. Además, es importante destacar la inexistencia de diágnosticos de cáncer de mama en masculinos menores a 45 años, a la vez que se visualiza una concentración de los casos a partir de los 60 años.

13.7.3 Sexo femenino

En el caso del sexo femenino, se destaca la inexistencia de diagnósticos en personas menores a 20 años, lo cual representa un rango mucho menor en comparación al sexo masculino. Asimismo, la agrupación de edades que acumula la mayor cantidad de casos corresponde a aquella que comprende de los 55 a los 59 años con un total de 168. La variación de esta cifra respecto a otras agrupaciones es muy pequeña: el grupo a partir de los 75 años posee una cantidad de 161 elementos, de igual manera, el rango de edad que va de los 65 a los 69 años contiene 156 elementos.

Por otro lado, se visualiza una concentración de los casos a partir de los 40 años, donde las cifras acumuladas en los rangos etarios son mayores a 100. Previo a los 40 años, a excepción del rango que va de los 35 a los 39 años, todos acumulan cantidades menores a 20 elementos. En particular, se destaca que la menor cantidad de diagnósticos se sitúa en el rango etario que comprende de los 20 a los 24 años, con apenas dos elementos.

13.8 Relación del escenario costarricense en el 2022 y la expresión mutada del gen TP53

A partir de lo analizado en los subapartados anteriores, en conjunto con lo visualizado en los gráficos correspondientes a la incidencia del cáncer de mama en Costa Rica durante el 2022, se determina que en la población costarricense los tres rangos etarios con una mayor incidencia del cáncer estudiado fueron: de los 55 a los 59 años, a partir de los 75 años y de los 65 a los 69 años.

Se analizará la relación del boxplot que detalla la relación entre las diez principales mutaciones del gen TP53 y la edad al diagnóstico, con los tres rangos etarios de mayor incidencia en el escenario costarricense durante el año 2022.

13.8.1 Rango etario con mayor incidencia: de los 55 a los 59 años

En primera instancia, es preciso destacar que para todas las mutaciones se cumple que el 50% de los datos centrales se sitúan por debajo de los 50 años, además, se presentan casos particulares en los cuales los bigotes de las cajas están por debajo de esa misma edad o apenas la superan, tal es el escenario para las mutaciones de identificador 3297, 2143, 3294, 3737 y 3879. En particular, para la mutación con identificador 1429, se identifica la presencia de valores atípicos en edades que comprenden o son cercanas al rango etario de mayor incidencia en el país. Es importante aclarar que cuando se menciona que las mutaciones poseen valores atípicos en ciertas edades, se refiere a que la expresión de dicha mutación no es común para esas edades.

Por otra parte, la extensión de los bigotes de las mutaciones 2242 y 4585 comprenden el rango etario que va de los 55 a los 59 años: si bien, la mayoría de expresiones de ambas mutaciones no se sitúan en las edades versadas, su presencia en ellas no se considera un valor atípico e inclusive para la mutación 4585 representan las edades más altas consideradas como valores normales o típicos.

13.8.2 Rango etario con la segunda mayor incidencia: a partir de los 75 años

Para el rango etario bajo estudio, se presenta un fenómeno particular respecto a la expresión de las diez mutaciones del gen TP53 más comunes: en la totalidad de estas diez mutaciones, no se considera un valor típico su expresión en edades iguales o superiores a los 75 años. De hecho, en el gráfico no se muestra la presencia de valores atípicos en estas edades, a excepción de la mutación con identificador 4584, que posee

un valor atípico en dicha área. Lo anterior no es sinónimo de que la expresión de las mutaciones sea común para las edades comprendidas en el rango etario, en su lugar sugiere que estas no se presentan del todo, ni siquiera como valor atípico.

13.8.3 Rango etario con la tercer mayor incidencia: de los 65 a los 69 años

Se identifican dos aspectos principales. En primer lugar, la presencia de valores atípicos situados en el rango etario bajo estudio para tres mutaciones distintas: 2143, 3294 y 3236. En segundo lugar, la extensión del bigote de la mutación de identificador 2242 comprende las edades estudiadas; si bien, la mayoría de expresiones de la mutación 2242 no se sitúa en ellas, su presencia en tal rango etario no se considera un valor atípico.

14 Conclusiones

14.1 Primera hipótesis

- H_1: Las mutaciones más frecuentes del gen TP53 se concentran en individuos diagnosticados entre los 30 y 50 años.

Nuestra primera hipósis hipótesis fue que las edades estaban concentradas en un rango entre los 30 y 50 años, basados en los resultados obtenidos en el gráfico boxplot, se puede llegar a la conclusión de que el rango de edad concentrado entre los 55 y 20 años, con mayores frecuencias entre los 30 y 40 años. Por otro lado se obtuvo que la edad promedio fue 34, mientras que la mediana de 33, lo que indica que no hay un sesgo a edades muy jóvenes o muy avanzadas; también se puede decir que en general las mutaciones del gen TP53 no se está en gran medida en pacientes de más de 50 años, sino que se está concentrando en edades de menos de 50.

Además, se obtuvo que la mutación más frecuente fue la del identificador 4585 con 314 observaciones, representando un 6.754% del total de observaciones, casi el doble que la mutación en el segundo lugar, la cual tiene el identificador 2143 con 180 observaciones, que representa un 3.872%, y la tercer mutación con mayor frecuencia fue el 3236 con 160 observaciones, representando un 3.442% del total de mutaciones.

14.2 Segunda hipótesis

- H_2: Existe una relación significativa entre el tipo de mutación (MUT_ID) y la edad al diagnóstico (Age_at_diagnosis).

El análisis comparativo hecho por el método Kruskal-Wallis revela que que existen cierto tipos de mutaciones que se asocian a edades de diagnóstico tempranas y tardías, es decir, exisste una diferencia significativa entre algunos tipos de mutación del gen TP53. Esto indica que la distribucón por edad no es homogénea entre las mutaciones más frecuentes del gen.

Tanto el análisis post-hoc que se usó luego de determinar el valor p de comparación general entre grupos y el gráfico forest plot, determinaron cuáles son las mutacioenes que tienden a diagnosticarse a una edad tardía respecto a los demás grupo, como lo son las mutaciones 4585 y 2242. También, revelaron las que teneden a asociarse a edades más tempranas que son las mutaciones 2705, 3297 y 3294.

Estos resultados respaldan en parte la hipótesis 2, que plantea una relación significativa entre la edad al diagnóstico y el tipo de mutación del TP53, sin embargo, esta relación no está presente en todos los tipos de mutación del gen, por lo tanto, se necesita otra evidencia para respaldar totalmente esta hipótesis, considerando estudiar las mutaciones específicas que mostraron tendencias importantes.

Aunque la hipótesis H_2 plantea una relación entre el tipo de mutación (MUT_ID) y la edad al diagnóstico, en esta sección se analiza una variable sustituta relacionada con la frecuencia mutacional (TCGA_ICGC_GENIE_count) para evaluar si existe una asociación entre la edad y la presencia o cantidad de mutaciones reportadas por los consorcios internacionales. Este análisis permite explorar parcialmente la hipótesis H2, identificando si la edad podría estar relacionada con la carga mutacional observada en los pacientes.

El análisis de correlación realizado revela que la edad de los individuos no se asocia de manera relevante con el número de mutaciones del gen TP53 reportadas por los consorcios TCGA, ICGC y GENIE. Tanto los resultados numéricos como los gráficos de dispersión muestran que las variables presentan una relación prácticamente nula.

Por otra parte, la edad al diagnóstico mostró una correlación estadísticamente significativa con el conteo de mutaciones (p = 0.011); sin embargo, el coeficiente obtenido (ρ = –0.077) indica que esta asociación es extremadamente débil y carece de importancia práctica. Esto sugiere que la edad al diagnóstico ejerce un efecto mínimo o inexistente en la cantidad de mutaciones registradas.

En conjunto, estos resultados permiten concluir que la variabilidad en el conteo de mutaciones del gen TP53 no se explica por la edad ni por la edad al diagnóstico, lo que implica que otros factores clínicos, biológicos o ambientales podrían tener un papel más determinante en la frecuencia mutacional.

A partir del análisis realizado, no se encontró evidencia sólida que respalde la hipótesis de que la edad al diagnóstico tiene una relación importante con la carga mutacional asociada a TP53. La correlación entre Age_at_diagnosis y el conteo combinado de mutaciones provenientes de los consorcios TCGA, ICGC y GENIE fue estadísticamente significativa (ρ = –0.077, p = 0.011); sin embargo, la magnitud del coeficiente es extremadamente baja, lo que indica que la relación, aunque detectable debido al tamaño de la muestra, no posee relevancia práctica.

Además, la correlación entre la edad general (Age) y el conteo mutacional tampoco mostró significancia estadística (ρ = 0.073, p = 0.246), reforzando la idea de que la edad no explica la variación en la cantidad de mutaciones registradas.

Por lo tanto, con base en estos resultados, la hipótesis H2 no puede considerarse apoyada, al menos desde la perspectiva del análisis correlacional asignado. Los datos sugieren que la edad al diagnóstico no está asociada de manera relevante con la frecuencia de mutaciones observadas, por lo que otros factores biológicos, genéticos o clínicos podrían ser más determinantes en la variabilidad mutacional.

14.3 Tercera hipótesis

- H_3: Los rangos etarios del escenario costarricense que poseen una mayor acumulación de casos, también presentan una mayor diversidad mutacional del gen TP53.

El análisis del escenario costarricense durante el año 2022 muestra que el sector etario con mayor incidencia de cáncer de mama, corresponde al rango que va de los 55 a los 59 años. En segundo lugar se encuentra las edades comprendidas a partir de los 75 años y en tercer lugar se sitúa el grupo de las edades de los 65 a los 69 años. Además, es preciso destacar la existencia de una considerable brecha entre la cantidad de diagnósticos reportados en el sexo masculino respecto al sexo femenino: alrededor del 99.51% del total están asociados al sexo femenino.

Al contrastar los resultados obtenidos en el escenario costarricense durante el año 2022, con el comportamiento de expresión observado en las diez mutaciones más comunes del gen TP53, se identifican ciertas variaciones de acuerdo al rango etario bajo estudio, sin embargo, se determina de forma general que las expresiones mutacionales del gen TP53 no suelen presentarse principalmente o de manera típica dentro de los rangos etarios en los cuales Costa Rica presenta la mayor incidencia. El análisis del boxplot que detalla la relación entre las diez principales mutaciones del gen TP53 y la edad al diagnóstico, evidencia que para la totalidad de las diez mutaciones principales del gen, se cumple que el 50% de los datos centrales se sitúan por debajo de los 50 años, asimismo, son escasas las ocasiones en las que los bigotes alcanzan los rangos etarios de mayor incidencia para el escenario nacional durante el 2022. Así, la presencia de las mutaciones en los rangos especificados no es significativa y en su mayoría representa valores normales extremos o valores atípicos. En particular, en los grupos de 65 a 69 años y de 75 años en adelante, las mutaciones prácticamente tienden a no expresarse, a excepción de ciertos valores atípicos.

En conjunto, estos hallazgos sugieren que no existe una fuerte concordancia entre las agrupaciones etarias que acumularon la mayor incidencia del cáncer de mama en Costa Rica y la distribución de las edades asociadas a las mutaciones más comunes del gen TP53: las mutaciones tienden a expresarse con mayor frecuencia en adultos jóvenes y adultos de mediana edad, mientras que en el contexto costarricense los casos se concentran en edades más avanzadas.

15 Referencias

Agresti, A. (2018). *Statistical Methods for the Social Sciences* (5th ed.). Pearson. https://www.pearson.com/en-us/subject-catalog/p/statistical-methods-for-the-social-sciences/P200000006673/9780134507101

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson Correlation Coefficient. En *Noise Reduction in Speech Processing* (pp. 1–4). Springer. https://doi.org/10.1007/978-3-642-00296-0_5

Bouaoun, L., Sonkin, D., Ardin, M., Hollstein, M., Byrnes, G., Zavadil, J., & Olivier, M. (2016). TP53 variations in human cancers: New lessons from the IARC TP53 database. *Human Mutation*, 37(9), 865–876. https://doi.org/10.1002/humu.23035

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Sage. https://uk.sagepub.com/en-gb/eur/discovering-statistics-using-ibm-spss-statistics/book257518

Gibbons, J. D., & Chakraborti, S. (2011). *Nonparametric Statistical Inference* (5th ed.). Chapman & Hall/CRC. https://doi.org/10.1201/b10905

Hernández-Sampieri, R., & Mendoza, C. (2018). *Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta.* McGraw-Hill. https://www.mheducation.com.mx/metodologia-de-la-investigacion-las-rutas-cuantitativa-cualitativa-y-mixta-9781456266669.html

Ministerio de Salud de Costa Rica. (2022). *Anuario de estadísticas vitales y de salud.* Ministerio de Salud de Costa Rica. https://www.ministeriodesalud.go.cr

Montgomery, D. C. (2017). *Design and Analysis of Experiments (9th ed.)*. Wiley. https://www.wiley.com/en-us/Design+and+Analysis+of+Experiments%2C+9th+Edition-p-9781119113478

Ott, R. L., & Longnecker, M. (2016). *An Introduction to Statistical Methods and Data Analysis (7th ed.)*. Cengage Learning. https://www.cengage.com/c/an-introduction-to-statistical-methods-and-data-analysis-7e-ott/

Soto Flores, W. (2015). Cáncer de mama. *Revista Médica de Costa Rica y Centroamérica*, (617), 799-802. https://www.binasss.sa.cr/revistas/rmcc/617/art20.pdf