```yaml
---
title: "Análisis de supervivencia"
author: "Grupo 01"
format: html
editor: visual
---
```

GRUPO 01 - Integrantes

- Bastidas Bendezu Ivan Luis
- Basurto Taype Marcelo Aaron
- Fonseca Moron José Kenneth
- Saravia Gutierrez Randy Esteban
- Talavera Ayllon Angel Ronaldo

Instalar (si es necesario)

```{r}
install.packages("broom")
install.packages("survival")
install.packages("survminer")
install.packages("ggsurvfit")
```

```{r}
install.packages("cardx")
```

Cargar paquetes

```{r}
library(tidyverse)
library(lubridate)
library(survival)
library(survminer)
library(gtsummary)
library(broom)
library(here)
library(rio)
library(ggsurvfit)
```

```{r}
library(cardx)
```

1 Analizando datos de tiempo a evento

El análisis de supervivencia, tambien conocido como análisis de tiempo a evento, es empleado para estudios donde el o los investigadores realizan un seguimiento (a los pacientes) hasta que ocurra un evento. Ejemplo de tales estudios caen en la categoría de estudios de cohorte prospectivo o retrospectivo.

El evento, en estudios de epidemiología, puede corresponder a muerte despues de diagnóstico, recurrencia de enfermedad, éxito de tratamiento, entre otros.

El análisis de supervivencia incluye datos de tiempo (dado que se hace seguimiento). Los datos de tiempo puede venir en cualquier forma: horas, días, meses, o años. Por ejemplo, meses hasta la recaída, años desde el diagnóstico hasta el fallecimiento, semanas antes de la recurrencia de la enfermedad, días desde el inicio del tratamiento al exito del tratamiento, años hasta el inicio de abuso de sustancias.

En esta sesión abordaremos 2 de las principales técnicas para realizar análisis de supervivencia:

- Análisis de supervivencia usando el método de Kaplan-Meier
- Regresión de riesgos proporcionales (PH) de Cox

1.1 Censura

Una característica clave en datos de supervivencia es la censura para un participante (una observación). La censura de un participante puede ocurrir por al menos 3 razones: 1) Perdida de seguimiento 2) Retiro del estudio 3) El evento no ocurre al término del periodo de estudio. Todos estos son ejemplos de "censura a la derecha", dado que ocurren luego del incio del estudio. Hay otros tipos de censura, menos frecuentes, pero estos no los consideraremos para esta sesión.

2 Estimaciones de supervivencia usando el método de Kaplan-Meier

2.1 El dataset para este ejercicio

El dataset almac_sangre contiene datos de 316 individuos. Para el primer ejecicio haremos uso de las siguientes 3 variables:

- Edad: tiempo de supervivencia observado en días.
- Evento: señala si el hubo el evento o no hubo el evento (sí o no).
- Raza_afroamericana: indica si el individuo es de raza afroamericana (sí o no).

Cargando el dataset almac_sangre

```{r}
almac_sangre <- import(here("data", "almac_sangre.csv"))
```

Vistazo al dataset

```{r}
almac_sangre |>
  select(Edad, evento, Raza_afroamericana) |>
  summary()
```

2.2 El desenlace en el análisis de supervivencia

Kaplan-Meier estima la probabilidad de supervivencia para distintos puntos de tiempo. En R, usamos la función `Surv()` para crear la variable "respuesta" o desenlace. Este tipo de desenlace incluye a su vez evento del participante (con el evento o sin el evento) y edad.

```r
Surv(almac_sangre$Edad, almac_sangre$evento)
```

```
  [1] 72.1+ 73.6+ 67.5  65.8  63.2  65.4  65.5  67.1+ 63.9  63.0  59.0  58.5  56.2
 73.3  59.7  67.6  61.6  61.2
 [19] 59.7  65.2  67.1  64.4+ 61.5  57.9  69.7  40.5  69.8  68.3  67.9  56.4
 69.0+ 62.1  49.8  73.9  64.2  71.2
 [37] 60.3  58.6  71.7  69.7  60.0  71.5  43.6  61.9  61.5  63.4+ 70.4+ 67.4  62.4
 69.4  65.3+ 59.9  61.8  58.9
 [55] 67.2  74.6  65.0  56.2  57.0  60.4  68.7  60.8+ 56.8  63.6  60.0  64.4  54.9
 47.0  63.6  55.3  74.2  56.0
 [73] 72.1+ 71.8  64.7  68.0  66.1+ 59.0  61.8  52.8  57.5  74.4  49.6  68.7
 47.9+ 63.7  60.0  61.7  70.3  55.8
 [91] 70.2+ 68.1  69.8+ 57.1  65.2  60.8  49.9  55.2  57.6  57.4  59.2+ 59.9  56.3
 70.6  71.8  58.2+ 78.3+ 55.9
[109] 53.9  51.7  64.2  67.4  62.3  62.4  71.0  60.4  65.9  58.8+ 63.2  61.1
 51.1+ 57.9  63.4  67.7  58.2  62.1
[127] 55.6  64.5  52.8  55.8  68.9+ 63.5+ 68.0  64.9  62.7  57.5  70.7  61.7  64.2
 69.8  68.0  63.1  64.1  58.1
[145] 60.1  56.0  62.3  51.2  57.0  51.0  65.8  54.5  55.9+ 55.7  48.4  58.2  63.9
 66.2  62.9  69.1  67.7+ 55.4
[163] 55.6  54.6+ 64.4  69.7  50.3+ 57.1  64.2  66.9  71.8+ 63.9  53.1  62.4  64.9
 62.4  58.8+ 54.4  57.0  65.6
[181] 65.9  67.1+ 62.2  50.6  68.4  55.8  79.0+ 63.9+ 49.6  54.9  61.1+ 59.6+ 64.7
 60.1  64.2  67.6  51.0  58.1
[199] 49.0+ 50.8+ 55.9  62.3  49.0  59.2  65.7  57.1  62.7  68.6+ 65.0  66.1+ 64.7
 62.2  55.9  45.7  71.4  55.4
[217] 54.8  56.1+ 67.5  58.2  65.3+ 56.2  66.8  61.5  55.5  43.8  48.9  68.3  60.5
 68.5  46.7+ 70.8  72.7  47.0
[235] 64.4  69.3  51.8  60.4  67.5  63.6  64.3  42.5+ 56.3  65.7  58.9+ 43.4+ 50.9
 58.2  74.8  71.9  61.1+ 76.3+
[253] 53.5+ 43.8  49.9  60.5  45.2  58.3  58.5  57.6  62.3  55.7  62.7  64.4  76.9
 55.4  47.7  49.1  53.2  54.8
[271] 62.2  46.5  62.4  63.3  67.8  64.8  51.2  62.6  66.7  63.5  57.2+ 59.0  74.9
 59.0  56.5+ 62.7  53.0  58.5
[289] 54.5+ 67.1  38.4  68.2  59.6  60.8  54.2  60.8  72.8  64.8  51.9+ 55.1  52.9
 66.5  66.1+ 55.3  55.7  66.9
[307] 64.9+ 70.6  65.1  61.6+ 64.1+ 54.8  62.3  62.4+ 57.6  59.9
```

El resultado en este chunk indica el estado (desenlace) de los participantes. El participante 1 fue censurado a los 72.1 días y no tuvo el evento. El participante 2 fue censurado a los 73.6 días y no tuvo el evento. El parcipante 3 sobrevivió 67.5 días y tuvo el evento.

La variable "respuesta" la incluimos en la función survfit para estimar la probabilidad de supervivencia (una curva) usando el método de Kaplan-Meier. Aquí estimados la curva de superviencia global sin considerar ningun predictor. Es por ello que añadimos "~ 1" a la formula. Esa parte será reemplazada más adelante por algun predictor o covariable de interés.

```{r}
km = survfit(Surv(Edad, evento) ~ 1, data = almac_sangre)
```

¿Qué hay dentro del objeto km?

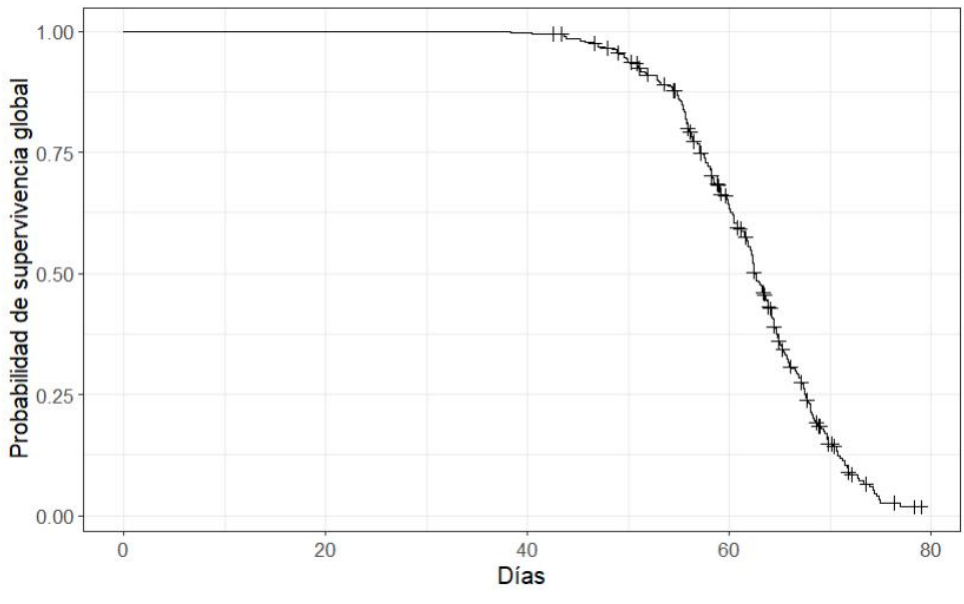
```{r}
str(km)
```

```
List of 17
 $ n         : int 316
 $ time      : num [1:182] 38.4 40.5 42.5 43.4 43.6 43.8 45.2 45.7 46.5 46.7 ...
 $ n.risk    : num [1:182] 316 315 314 313 312 311 309 308 307 306 ...
 $ n.event   : num [1:182] 1 1 0 0 1 2 1 1 1 0 ...
 $ n.censor  : num [1:182] 0 0 1 1 0 0 0 0 0 1 ...
 $ surv      : num [1:182] 0.997 0.994 0.994 0.994 0.99 ...
 $ std.err   : num [1:182] 0.00317 0.00449 0.00449 0.00449 0.00552 ...
 $ cumhaz    : num [1:182] 0.00316 0.00634 0.00634 0.00634 0.00954 ...
 $ std.chaz  : num [1:182] 0.00316 0.00448 0.00448 0.00448 0.00551 ...
 $ type      : chr "right"
 $ logse     : logi TRUE
 $ conf.int  : num 0.95
 $ conf.type: chr "log"
 $ lower     : num [1:182] 0.991 0.985 0.985 0.985 0.98 ...
 $ upper     : num [1:182] 1 1 1 1 1 ...
 $ t0        : num 0
 $ call      : language survfit(formula = Surv(Edad, evento) ~ 1, data =
almac_sangre)
 - attr(*, "class")= chr "survfit"
```

2.3 Gráficos de Kaplan-Meier

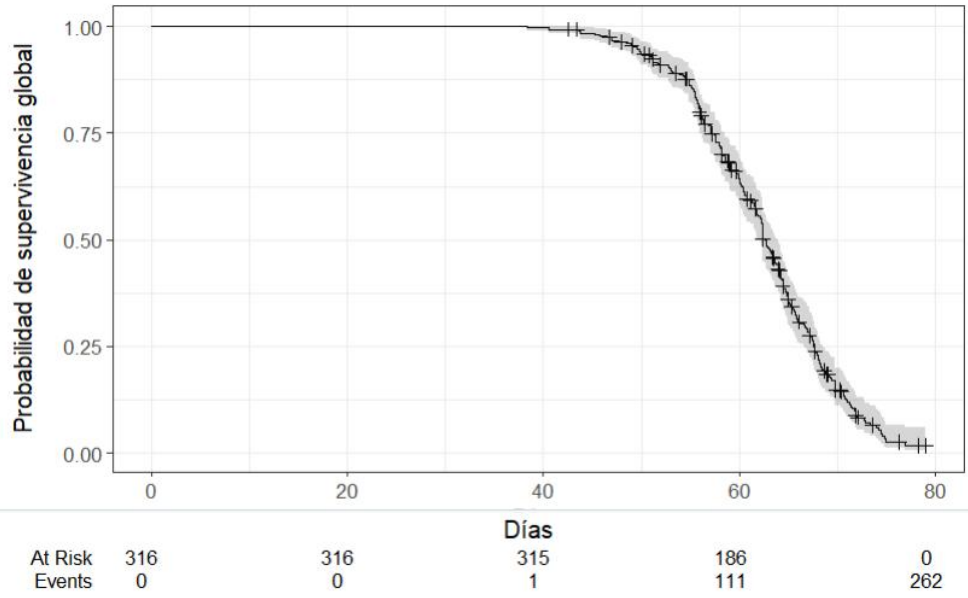
La información contenida en el objeto creado con las estimaciones puede ser mejor visualizada en los gráficos de Kaplan-Meier.

```{r}
survfit2(Surv(Edad, evento) ~ 1, data = almac_sangre) |>
  ggsurvfit() +
  labs(
    x = "Días",
    y = "Probabilidad de supervivencia global"
  ) +
  add_censor_mark()
```



La función `add_confidence_interval()` añade los intervalos de confianza al 95% (sombreado en gris) para las estimaciones de probabilidad de supervivencia.

```{r}
survfit2(Surv(Edad, evento) ~ 1, data = almac_sangre) |>
  ggsurvfit() +
  labs(
    x = "Días",
    y = "Probabilidad de supervivencia global"
  ) +
  add_censor_mark() +
  add_confidence_interval() +
  add_risktable()
```



¿Cómo interpretar?

En la gráfica de Kaplan-Meier generada a partir de los datos del almacenamiento de sangre, se muestra la probabilidad de supervivencia global a lo largo del tiempo (días). La curva es escalonada, ya que representa una función de supervivencia estimada por intervalos de tiempo; cada escalón indica la ocurrencia de un evento (fallecimiento), reduciendo la probabilidad acumulada de supervivencia.

Las líneas horizontales reflejan la duración de los intervalos de tiempo entre eventos, mientras que las caídas verticales indican la magnitud de la reducción en la probabilidad de supervivencia al producirse un evento. Las marcas de censura, líneas verticales (añadidas con `add_censor_mark()`), señalan a aquellos pacientes que no experimentaron el evento al final del periodo de seguimiento o que fueron retirados del estudio antes de concluir el tiempo de observación. Estos pacientes censurados no afectan la probabilidad acumulada en el momento de su censura, pero reducen el número de sujetos en riesgo en los intervalos posteriores.

El intervalo de confianza del 95% (representado mediante bandas alrededor de la curva) proporciona una estimación de la incertidumbre asociada a la probabilidad de supervivencia en cada punto temporal. Nota que a medida que transcurre el tiempo, el intervalo de confianza al 95%, es más ancha, es decir, menos preciso.

Finalmente, la tabla de riesgo ubicada bajo la gráfica (generada con `add_risktable()`) muestra el número de pacientes que permanecen en riesgo en distintos momentos del seguimiento, lo que facilita la interpretación de la robustez de la estimación de la curva a lo largo del tiempo.

2.4 Estimación de la supervivencia a x años.

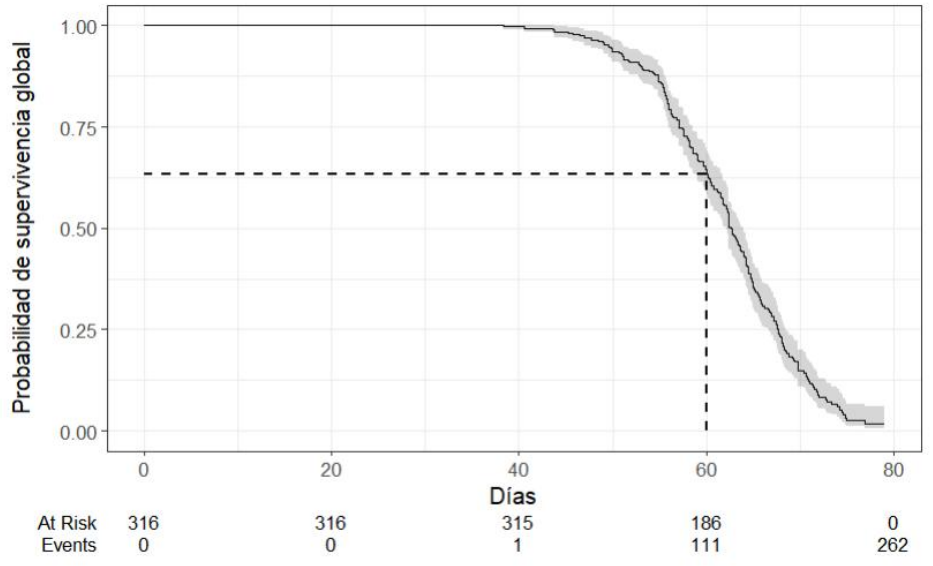
Al analizar datos de supervivencia es común que nos preguntemos, por ejemplo, ¿Cuál es la probabilidad de supervivencia después de 60 dias de seguimiento? Esto lo calculamos a partir de usar la función survfit(), añadiendo el argumento times.

```r
{r}
summary(survfit(Surv(Edad, evento) ~ 1, data = almac_sangre), times = 60)
```

```
Call: survfit(formula = Surv(Edad, evento) ~ 1, data = almac_sangre)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   60    186     111    0.633  0.0278        0.581        0.69
```

La probabilidad de supervivencia a los 60 días de seguimiento es del 63%. Dicho de otra manera, 63% de los pacientes estuvieron vivos 60 días despues del inicio del estudio.

```r
{r}
survfit2(Surv(Edad, evento) ~ 1, data = almac_sangre) |>
  ggsurvfit() +
  labs(
    x = "Días",
    y = "Probabilidad de supervivencia global"
  ) +
  add_confidence_interval() +
  add_risktable() +
  add_quantile(x_value = 60, linewidth = .7)
```



2.5 Estimación mediana del tiempo de supervivencia

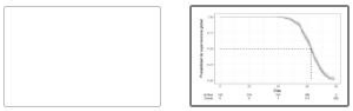
Otro dato importante a estimar es la mediana de supervivencia. Típicamente, los datos de supervivencia no tendrán una distribución normal. Asi que, la mediana es preferida sobre la media aritmética.

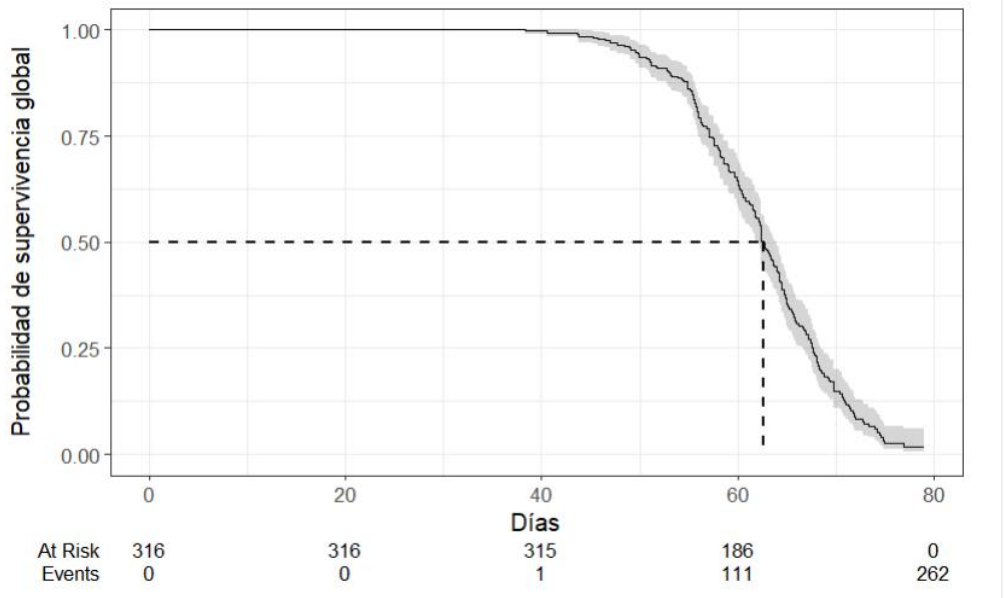
```{r}
survfit(Surv(Edad, evento) ~ 1, data = almac_sangre)
```

```
Call: survfit(formula = Surv(Edad, evento) ~ 1, data = almac_sangre)

       n events median 0.95LCL 0.95UCL
[1,] 316    262   62.6    62.1    63.9
```

La mediana de supervivencia corresponde a la probabilidad de supervivencia de 0.5. Para este dataset, la mediana es de 62.6 días. En el gráfico de Kaplan Meier

```{r}
survfit2(Surv(Edad, evento) ~ 1, data = almac_sangre) |>
  ggsurvfit() +
  labs(
    x = "Días",
    y = "Probabilidad de supervivencia global"
  ) +
  add_confidence_interval() +
  add_risktable() +
  add_quantile(y_value = 0.5, linewidth = .7)
```





¿Cómo reportar?

Usando el paquete gtsummary podemos generar una tabla con datos de la supervivencia a los 60 dias.

```{r}
theme_gtsummary_language(language = "es")
```

```
Setting theme "language: es"
```

```{r}
survfit(Surv(Edad, evento) ~ 1, data = almac_sangre) %>%
  tbl_survfit(
    times = 60,
    label_header = "**Supervivencia a los 60 días (IC 95%)**"
  )
```

```{r}
survfit(Surv(Edad, evento) ~ 1, data = almac_sangre) |>
  tbl_survfit(
    probs = 0.5,
    label_header = "**Supervivencia a los 60 días (IC 95%)**"
  )
```

Comparando tiempos de supervivencia entre dos grupos

En el conjunto de datos `almac_sangre` se incluyen tanto afroamericanos como no afroamericanos. Un análisis de interés consiste en evaluar si los tiempos de supervivencia difieren significativamente entre ambos grupos. Para ello, se utiliza la función `survdiff()`, que permite aplicar la prueba de log-rank y estimar si existen diferencias en las curvas de supervivencia.

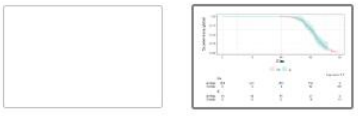
```{r}
survdiff(Surv(Edad, evento) ~ Raza_afroamericana, data = almac_sangre)
```

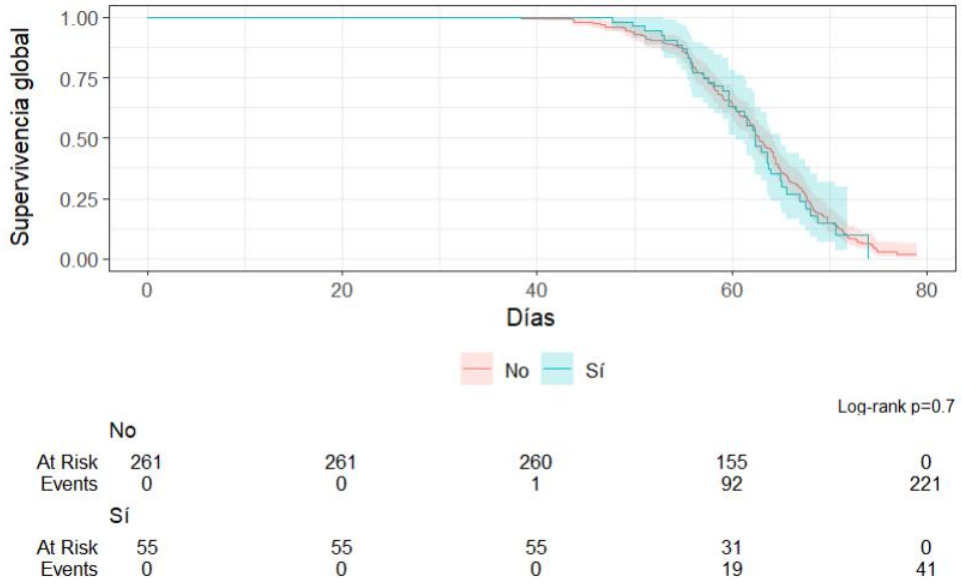
```
Call:
survdiff(formula = Surv(Edad, evento) ~ Raza_afroamericana, data = almac_sangre)

                         N Observed Expected (O-E)^2/E (O-E)^2/V
Raza_afroamericana=No 261      221    223.5    0.0279     0.194
Raza_afroamericana=Sí  55       41     38.5    0.1621     0.194

 Chisq= 0.2  on 1 degrees of freedom, p= 0.7
```

El valor de p = 0.7 de la prueba de log-rank indica que no se encontraron diferencias estadísticamente significativas entre los grupos comparados, en la supervivencia global de afroamericanos o no afroamericanos.

El siguiente gráfico de Kaplan-meier muestra el resultado de la prueba de log-rank.

```{r}
survfit2(Surv(Edad, evento) ~ Raza_afroamericana, data = almac_sangre) |>
  ggsurvfit() +
  labs(
    x = "Días",
    y = "Supervivencia global"
  ) +
  add_confidence_interval() +
  add_risktable() +
  add_pvalue(caption = "Log-rank {p.value}")
```





No					
At Risk	261	261	260	155	0
Events	0	0	1	92	221
Sí					
At Risk	55	55	55	31	0
Events	0	0	0	19	41

3 El modelo de regresión de Cox

La prueba de log-rank no ofrece una medida de efecto y solo permite evaluar una única variable independiente a la vez. Sin embargo, en investigación suele ser necesario cuantificar el tamaño del efecto de una o varias variables, e incluso realizar un análisis multivariable, aspecto que abordaremos en la siguiente sesión. Para este propósito, el modelo de regresión de Cox es una herramienta adecuada para analizar datos con desenlaces de supervivencia.

En R, este modelo se puede ajustar utilizando la función `coxph()` del paquete **survival**.

3.1 El dataset para este ejercicio

El dataset para esta parte de la sesión incluye información de 316 pacientes. Entre las variables estan:

- Edad_mediana_GR: es la mediana de días de los GR
- evento: indica si el evento de interés esta presente.
- Historia familiar - y otras covariables

Seguiremos empleando el dataset `almac_sangre`

Usaremos a la variable Historia familiar como la variable independiente de interés

```{r}
coxph(Surv(Edad_mediana_GR, evento == "1") ~ Historia_familiar, data = almac_sangre
)
```

```
Call:
coxph(formula = Surv(Edad_mediana_GR, evento == "1") ~ Historia_familiar,
    data = almac_sangre)

                      coef exp(coef) se(coef)     z    p
Historia_familiarSí 0.1030    1.1084   0.1458 0.706 0.48

Likelihood ratio test=0.49  on 1 df, p=0.484
n= 316, number of events= 262
```

En el análisis de regresión de Cox, la presencia de historia familiar no se asoció significativamente con un mayor riesgo del evento (HR: 1.11; IC 95% no reportado; p = 0.48). El modelo no mostró una mejora significativa en la verosimilitud global (test de razón de verosimilitudes p = 0.484).

3.2 Interpretación y reporte

Estas tablas de resultados pueden obtenerse con la función `tbl_regression()` del paquete gtsummary, utilizando la opción `exponentiate = TRUE` para mostrar la razón de riesgos (hazard ratio, HR) en lugar del logaritmo del riesgo.

```{r}
coxph(Surv(Edad_mediana_GR, evento == "1") ~ Historia_familiar, data = almac_sangre
) %>%
  tbl_regression(exp = TRUE)
```

La presencia de historia familiar no se asoció significativamente con un mayor riesgo del evento. Aunque el hazard ratio fue de 1.11, el intervalo de confianza del 95% (0.83–1.48) incluye el valor nulo y el valor p fue de 0.5, indicando ausencia de significancia estadística.

3.3 Reporte para multiple variables

Es frecuente que en un estudio que incluya datos de supervivencia sea de interés evaluar multiples covariables. En R, usando la función `tbl_uvregression()` podemos generar modelos univariables simples para todas las covariables. A cambio, obtenemos la HR cruda para todas las covariables de interés.

```{r}
tabla_cox <- almac_sangre |>
  tbl_uvregression(
    include = c(Edad, Historia_familiar, Raza_afroamericana,Confinamiento_organo,
Terapia_previa),
    y = Surv(Edad_mediana_GR, evento == "1"),
    method = coxph,
    exponentiate = TRUE,
    conf.int = TRUE,
    hide_n = TRUE,
    add_estimate_to_reference_rows = FALSE,
    pvalue_fun = ~ style_pvalue(.x, digits = 3),
    estimate_fun = ~ style_number(.x, digits = 2),
    label = list(
      Edad ~ "Edad días",
      Historia_familiar ~ "Antecedente familiar",
      Raza_afroamericana ~ "Raza",
      Confinamiento_organo ~ "Confinamiento de organo",
      Terapia_previa ~ "Terapia previa"
    )
  ) |>
  bold_p(t = 0.05) |>
  modify_header(estimate = "**HR no ajustado**", p.value = "**Valor P**")
```

Imprimimos la tabla

```{r}
tabla_cox
```