

Análisis de Datos del Cáncer de Pulmón NCCTG

February 2, 2025

Introducción

El cáncer de pulmón es una de las principales causas de mortalidad en todo el mundo, siendo responsable de millones de muertes cada año. En este contexto, los datos clínicos resultan fundamentales para entender los factores que afectan la supervivencia de los pacientes y, en última instancia, mejorar las estrategias de tratamiento. El conjunto de datos del cáncer de pulmón NCCTG proporciona información valiosa sobre pacientes con cáncer de pulmón avanzado recopilada por el Grupo de Tratamiento del Cáncer del Norte Central. Este conjunto de datos incluye tanto variables clínicas como datos autoinformados por los pacientes, lo que permite una visión integral de las condiciones de los mismos.

El propósito de este análisis es explorar las características de este conjunto de datos, identificar patrones significativos y comprender cómo ciertas variables, como la edad, el sexo, las puntuaciones de rendimiento físico y la pérdida de peso, influyen en los tiempos de supervivencia. A través de este análisis, se espera contribuir a la identificación de posibles predictores de la supervivencia que podrían ser útiles en entornos clínicos y en investigaciones futuras.

Descripción del Conjunto de Datos

El conjunto de datos contiene información detallada sobre 228 pacientes y se estructura en las siguientes variables clave:

- **inst**: Código de la institución donde se atendió al paciente. Esta variable permite identificar la procedencia de los datos.
- **time**: Tiempo de supervivencia del paciente en días desde el inicio del estudio hasta el momento del fallecimiento o censura.
- **status**: Estado de censura al final del estudio. Se codifica como 1 para censurado (es decir, el paciente estaba vivo al final del seguimiento) y 2 para fallecido.
- **age**: Edad del paciente en años al inicio del estudio. Esta variable es crucial para evaluar la influencia de la edad en la supervivencia.
- **sex**: Sexo del paciente, codificado como 1 para masculino y 2 para femenino.
- **ph.ecog**: Puntuación de rendimiento según el Índice de Rendimiento ECOG (Eastern Cooperative Oncology Group):

- 0: Asintomático.
 - 1: Sintomático pero completamente ambulatorio.
 - 2: En cama menos del 50% del día.
 - 3: En cama más del 50% del día pero no postrado.
 - 4: Postrado en cama.
- **ph.karno:** Puntuación de rendimiento de Karnofsky asignada por el médico, que varía de 0 (muy mal estado) a 100 (excelente estado físico).
 - **pat.karno:** Puntuación de rendimiento de Karnofsky asignada por el propio paciente, lo que permite comparar la percepción del paciente con la evaluación del médico.
 - **meal.cal:** Cantidad de calorías consumidas durante las comidas diarias, una medida que puede reflejar el estado nutricional del paciente.
 - **wt.loss:** Pérdida de peso en los últimos seis meses, medida en libras. Esta variable es un indicador importante del deterioro físico en pacientes con cáncer avanzado.

Contexto y Antecedentes

El uso de los datos clínicos para el análisis de supervivencia se ha convertido en una herramienta esencial en oncología. En particular, el conjunto de datos NCCTG fue recopilado para evaluar factores pronósticos relacionados con la supervivencia de pacientes con cáncer de pulmón avanzado. Una característica notable de este conjunto de datos es el uso de codificaciones específicas (por ejemplo, 1 y 2 para vivo y muerto, respectivamente) en lugar de las convencionales 0 y 1. Este enfoque, aunque inusual, se adoptó por motivos técnicos durante la era de las tarjetas perforadas en sistemas como el IBM 360 Fortran, donde los valores en blanco se interpretaban como ceros. Aunque estos sistemas han quedado obsoletos, las prácticas asociadas persistieron durante años.

El estudio asociado con este conjunto de datos, liderado por Loprinzi et al. (1994), exploró cómo las puntuaciones de rendimiento y otras variables influían en la supervivencia. Estas puntuaciones, como las de Karnofsky y ECOG, han demostrado ser herramientas valiosas para evaluar el estado funcional de los pacientes y predecir resultados clínicos. Además, variables como la pérdida de peso y la ingesta calórica son indicadores del estado nutricional, que también juega un papel crítico en el manejo del cáncer.

Objetivos del Análisis

Este informe tiene como objetivos:

- Resumir y describir las características principales de las variables del conjunto de datos.
- Identificar patrones y tendencias en los tiempos de supervivencia y los estados de censura.
- Evaluar la relación entre las puntuaciones de rendimiento, el estado nutricional y los resultados de supervivencia.

- Proporcionar una base para análisis futuros más complejos, como modelos de regresión de supervivencia.

Análisis de las variables

Inst

Dado que este valor solo representa la institución donde se atendió al paciente, no se considera relevante para el análisis de supervivencia y, por lo tanto, no se incluirá en los análisis posteriores.

Time

La variable de tiempo de supervivencia es fundamental para el análisis de supervivencia y se presenta en días. **La Figura 1** muestra un histograma de los tiempos de supervivencia, que revela una distribución asimétrica con una cola larga hacia la derecha. La mayoría de los pacientes sobrevivieron menos de 500 días, con un pico alrededor de los 200 días. La censura también es evidente en los datos, ya que hay un número significativo de pacientes cuyo tiempo de supervivencia no se conoce debido a la censura.

Clasificación de la variable

En cuanto al tipo de variable es **cuantitativa discreta** debido a que se da en el número de días que sobrevivió el paciente y de **intervalos** ya que el 0 tiene un significado.

Medidas de tendencia central

A continuación se presentan las medidas de tendencia central para la variable de tiempo de supervivencia:

Medida	Valor
Media	305.2325
Moda	163
Mediana	255
Q1 (25%)	166
Q3 (75%)	396

Table 1: Medidas de tendencia central para el tiempo de supervivencia

Medidas de variabilidad

A continuación se presentan las medidas de variabilidad para la variable de tiempo de supervivencia:

Medida	Valor
Máximo	1022
Mínimo	5
Rango	1017
Varianza	44371.54
Desviación Estándar	210.6455
Coefficiente de Variación	0.6901152

Table 2: Medidas de variabilidad para el tiempo de supervivencia

Prueba de Distribución

El siguiente gráfico muestra la distribución de los tiempos de supervivencia comparada con diferentes tipos de distribución.

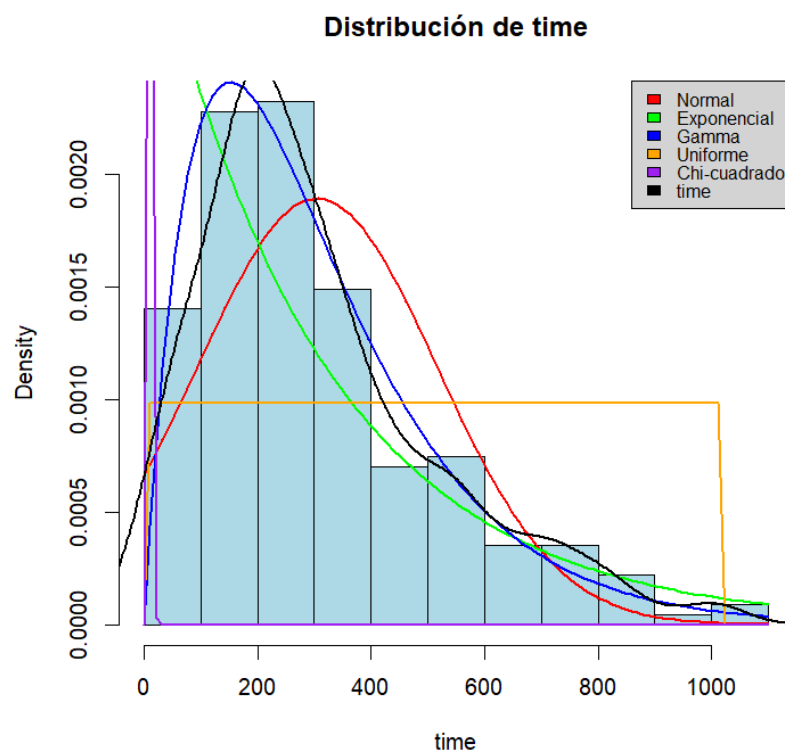


Figure 1: Distribución de los tiempos de supervivencia

Por tanto es conveniente realizar las pruebas de hipótesis para determinar si la variable sigue una distribución específica. Dicha prueba resultó en los siguientes valores:

Prueba	p-value
Anderson-Darling (Normal)	7.417945e-14
Kolmogorov-Smirnov (Normal)	0.001288452
Shapiro-Wilk (Normal)	5.114211e-10
Kolmogorov-Smirnov (Exponencial)	2.970448e-07
Kolmogorov-Smirnov (Gamma)	0.3606842
Kolmogorov-Smirnov (Chi-cuadrado)	3.916153e-192

Table 3: Resultados de las pruebas de distribución para la variable *time*

Como es posible apreciar solo la prueba de Kolmogorov-Smirnov para la distribución Gamma no rechaza la hipótesis nula, por lo que se puede concluir que no es posible negar la variable **time** posee una distribución Gamma y por tanto la asumiremos en el análisis siguiente.

Intervalo de confianza para la media

El intervalo de confianza para la media de la variable **time** es de 269.2987567105341 a 341.1661555701677 con un nivel de confianza del 99%. Esto lo podemos saber gracias al siguiente código :

```

1  import pandas as pd
2  import numpy as np
3  from scipy.stats import norm as z
4
5  # Leer el archivo CSV y seleccionar la columna 'time'
6  muestra = pd.read_csv('lung_dataset.csv')['time'].dropna().
    values
7
8
9  media_muestral = np.mean(muestra)
10 desviacion_muestral = np.std(muestra, ddof=1)
11 n = len(muestra)
12
13 # Nivel de confianza y grados de libertad
14 nivel_significancia = 0.01
15 confianza = 1 - nivel_significancia
16
17
18 z_critico = z.ppf(1 - nivel_significancia / 2)
19
20
21 margen_error = z_critico * (desviacion_muestral / np.sqrt(n))
22 limite_inferior = media_muestral - margen_error
23 limite_superior = media_muestral + margen_error
24
25 print(media_muestral, desviacion_muestral, (limite_inferior,
    limite_superior))

```

Listing 1: Código en Python para calcular el intervalo de confianza

Prueba de hipótesis para la media

Se plantea que el tiempo de supervivencia de los pacientes con cáncer es de un promedio de 250 días. Probemos la veracidad de esta proposición mediante el siguiente código:

```
1
2     import math
3     import pandas as pd
4     from scipy.stats import t as t_dist
5
6     muestra = pd.read_csv('lung_dataset.csv')['time'].dropna().
           values
7
8
9     # H0: La media de la edad es <= 250
10    # H1: La media de la edad es > 250
11    mu_0 = 250
12    alpha = 0.01 # Nivel de significancia
13
14
15    n = len(muestra)
16    sample_mean = sum(muestra) / n # Media muestral
17    sample_std = math.sqrt(sum((x - sample_mean) ** 2 for x in
           muestra) / (n - 1))
18    t_stat = (sample_mean - mu_0) / (sample_std / math.sqrt(n))
19
20    # Grados de libertad
21    df = n - 1
22
23
24    t_critical = t_dist.ppf(1 - alpha, df)
```

Listing 2: Código en Python para calcular el estadígrafo de la prueba de hipótesis

Este código nos da como resultado que se rechaza la hipótesis nula, por lo que el tiempo de supervivencia promedio de los pacientes con cáncer de pulmón es mayor a 250 días.

Status

La variable de estado de censura, codificada como 1 para censurado y 2 para fallecido, es crucial para el análisis de supervivencia. **La Figura 2** muestra la distribución de los estados de censura en el conjunto de datos. La mayoría de los pacientes están censurados, lo que refleja la naturaleza de los estudios de supervivencia donde no todos los pacientes experimentan el evento de interés (es decir, la muerte) durante el período de seguimiento.

Clasificación de la variable

En cuanto al tipo de variable es **cualitativa nominal** debido a que se da en dos categorías y de **intervalos** ya que el 0 tiene un significado.

Medidas de tendencia central

A continuación se presentan las medidas de tendencia central para la variable de estado de censura:

Medida	Valor
Media	1.723684
Moda	2
Mediana	2
Q1 (25%)	1
Q3 (75%)	2

Table 4: Medidas de tendencia central para la variable de estado de censura

Medidas de variabilidad

A continuación se presentan las medidas de variabilidad para la variable de estado de censura:

Medida	Valor
Máximo	2
Mínimo	1
Rango	1
Varianza	0.200846
Desviación Estándar	0.448159
Coefficiente de Variación	0.260001

Table 5: Medidas de variabilidad para la variable de estado de censura

Prueba de Distribución

El siguiente gráfico muestra la distribución de los estados de censura en el conjunto de datos, este tipo de variable aleatoria es una Bernoulli ya que solo puede tomar dos valores, en este caso 1 y 2.

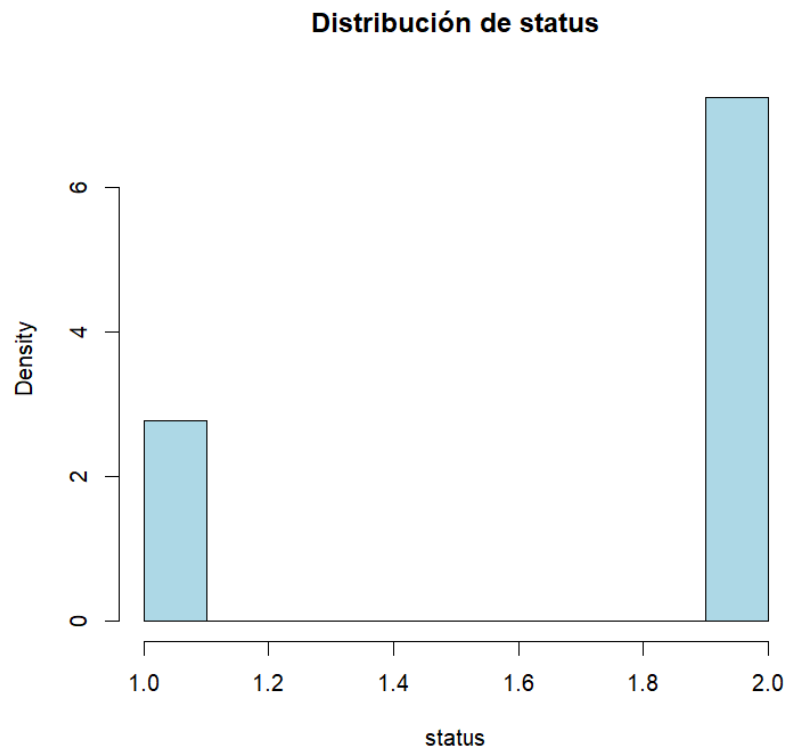


Figure 2: Distribución de los tiempos de supervivencia

Intervalo de confianza para la proporción

El intervalo de confianza para la proporción de la variable **status** es de 0.6474013349555908 a 0.7999670860970408 con un nivel de confianza del 99%. Esto lo podemos saber gracias al siguiente código:

```
1 import pandas as pd
2 import numpy as np
3 from scipy.stats import norm as z
4
5 # Leer el archivo CSV y seleccionar la columna 'status'
6 muestra = pd.read_csv('lung_dataset.csv')['status'].dropna().
    values
7
8 n = len(muestra)
9 proporcion_fallecidos = np.sum(muestra == 2) / n
10
11 # Nivel de confianza
12 nivel_significancia = 0.01
13 confianza = 1 - nivel_significancia
```



```

14
15     z_critico = z.ppf(1 - nivel_significancia / 2)
16
17     margen_error = z_critico * np.sqrt(proporcion_fallecidos * (1
18         - proporcion_fallecidos) / n)
19
20     limite_inferior = proporcion_fallecidos - margen_error
21     limite_superior = proporcion_fallecidos + margen_error
22
23     print(f"Intervalo de confianza al {confianza*100}%: ({
24         limite_inferior}, {limite_superior})")

```

Listing 3: Código en Python para calcular el intervalo de confianza

Age

La edad del paciente al inicio del estudio es una variable importante para evaluar la influencia de la edad en la supervivencia. **La Figura 3** muestra un histograma de las edades de los pacientes, que revela una distribución aproximadamente simétrica con un pico alrededor de los 65 años.

Clasificación de la variable

En cuanto al tipo de variable es **cuantitativa discreta** debido a que se da en el número de años que tiene el paciente y de **intervalos** ya que el 0 tiene un significado.

Medidas de tendencia central

A continuación se presentan las medidas de tendencia central para la variable de edad:

Medida	Valor
Media	62.44737
Moda	60
Mediana	63
Q1 (25%)	56
Q3 (75%)	69

Table 6: Medidas de tendencia central para la variable de edad

Medidas de variabilidad

A continuación se presentan las medidas de variabilidad para la variable de edad:

Medida	Valor
Máximo	82
Mínimo	39
Rango	43
Varianza	82.3276142
Desviación Estándar	9.0734566
Coefficiente de Variación	0.1452977

Table 7: Medidas de variabilidad para la variable de edad

Prueba de Distribución

El siguiente gráfico muestra la distribución de las edades de los pacientes comparada con diferentes tipos de distribución.

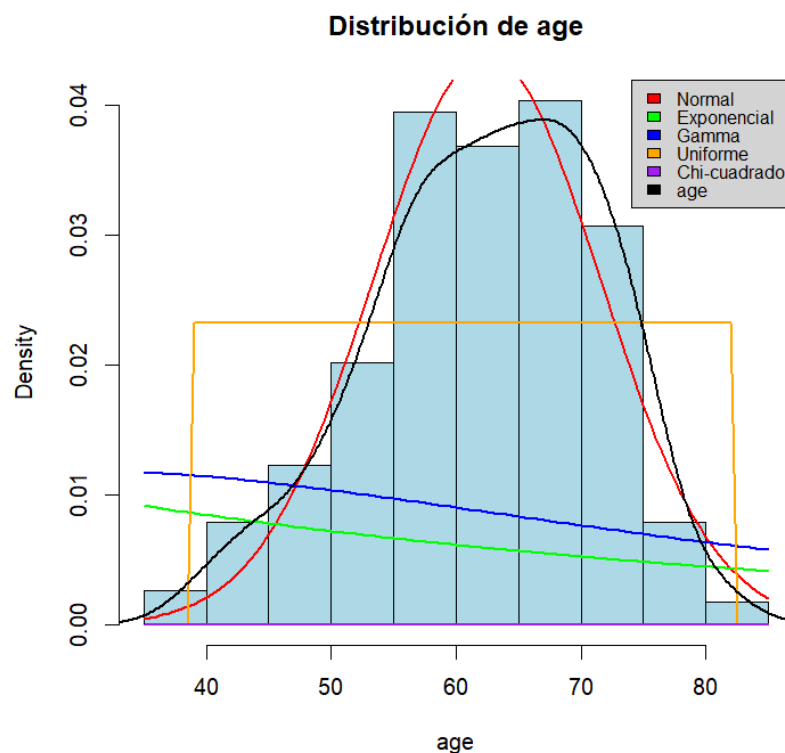


Figure 3: Distribución de la edad de los pacientes

Por tanto es conveniente realizar las pruebas de hipótesis para determinar si la variable sigue una distribución específica. Dicha prueba resultó en los siguientes valores:

Prueba	p-value
Anderson-Darling (Normal)	0.009280497
Kolmogorov-Smirnov (Normal)	0.2508968
Shapiro-Wilk (Normal)	0.00482907
Kolmogorov-Smirnov (Exponencial)	6.191832e-46
Kolmogorov-Smirnov (Gamma)	4.863935e-31
Kolmogorov-Smirnov (Chi-cuadrado)	1.831273e-198

Table 8: Resultados de las pruebas de distribución para la variable *age*

Como es posible apreciar bajo un nivel de significancia de 0.01 la prueba de Kolmogorov - Smirnov para la distribución normal no rechaza la hipótesis nula, por lo que se puede concluir que no es posible negar que la variable **age** posee una distribución normal y por tanto la asumiremos en el análisis siguiente.

Intervalo de confianza para la media

El intervalo de confianza para la media de la variable **age** es de 60.89954141062864 a 63.99519543147662 con un nivel de confianza del 99%. Esto lo podemos saber gracias al siguiente código, al ser análogo al código 1 utilizado para el intervalo de confianza de la variable **time** solamente sustituiríamos:

```
1 muestra = pd.read_csv('lung_dataset.csv')['age'].dropna().values
```

Listing 4: Código en Python para calcular el intervalo de confianza

Prueba de hipótesis para la media

Se plantea que la edad promedio de los pacientes con cáncer es de más de 60 años. Utilizando la implementación análoga para la variable **time** 2.

```
1
2
3 muestra = pd.read_csv('lung_dataset.csv')['age'].dropna().values
4
5 # H0: La media de la edad es <= 60
6 # H1: La media de la edad es > 60
7 mu_0 = 60
```

Listing 5: Código en Python para calcular el estimador de la prueba de hipótesis

Este código nos da como resultado que se rechaza la hipótesis nula, por lo que la edad promedio de los pacientes con cáncer de pulmón es más de 60 años.

Sex

La variable de sexo del paciente, codificada como 1 para masculino y 2 para femenino, es un factor importante a considerar en el análisis de supervivencia.

Clasificación de la variable

En cuanto al tipo de variable es **cualitativa nominal** debido a que se da en dos categorías y de **intervalos** ya que el 0 tiene un significado.

Medidas de tendencia central

A continuación se presentan las medidas de tendencia central para la variable *sex*:

Medida	Valor
Media	1.394737
Moda	1
Mediana	1
Q1 (25%)	1
Q3 (75%)	2

Table 9: Medidas de tendencia central para la variable *sex*

Medidas de variabilidad

A continuación se presentan las medidas de variabilidad para la variable *sex*:

Medida	Valor
Máximo	2
Mínimo	1
Rango	1
Varianza	0.2399722
Desviación Estándar	0.4898696
Coefficiente de Variación	0.3512272

Table 10: Medidas de variabilidad para la variable *sex*

Prueba de Distribución

La **Figura 4** muestra la distribución de los pacientes por sexo. Como se puede observar, hay una proporción ligeramente mayor de pacientes masculinos en comparación con los femeninos. Esta diferencia en la distribución de sexos puede influir en los resultados de supervivencia y, por lo tanto, es crucial tenerla en cuenta en cualquier análisis posterior.

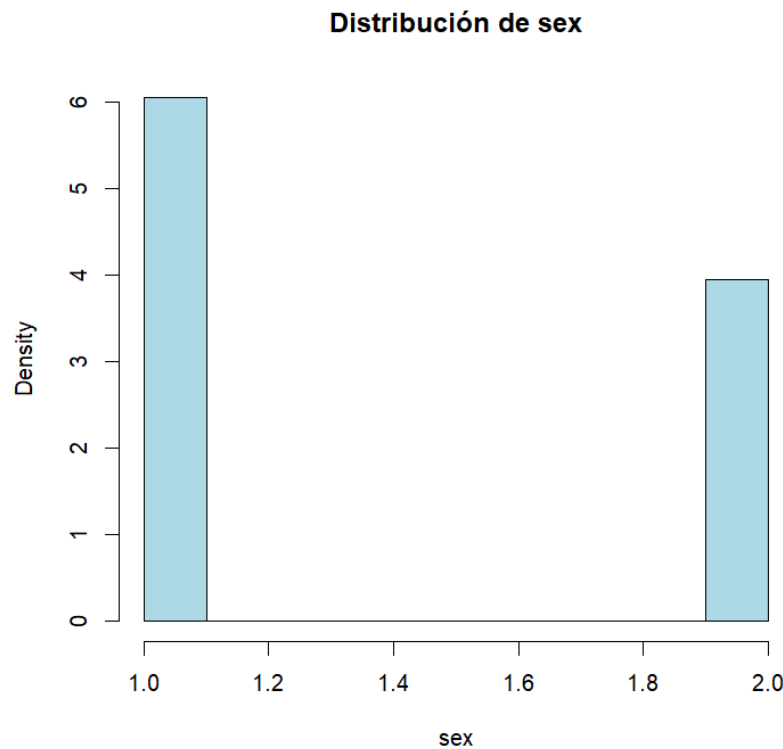


Figure 4: Distribución del género de los pacientes

Intervalo de confianza para la proporción

El intervalo de confianza para la proporción de la variable **sex** es de 0.31135420869499797 a 0.47811947551552836 con un nivel de confianza del 99%. Esto lo podemos saber gracias al siguiente código, al ser análogo al código 3 utilizado para el intervalo de confianza de la variable **status** solamente sustituiríamos:

1

```
muestra = pd.read_csv('lung_dataset.csv')['sex'].dropna()  
.values
```

Listing 6: Código en Python para calcular el intervalo de confianza

Prueba de hipótesis para dos muestras

Verificación de la existencia de una diferencia significativa en la proporción de personas fallecidas (**status** = 2) entre los géneros (**sex**: 1 para hombres, 2 para mujeres). Esto permitiría explorar si el género está asociado con la probabilidad de fallecimiento en el contexto del estudio. Nos apoyaremos en el siguiente código para darle solución a la proposición.

```

1      import math
2
3  from scipy.stats import norm
4  import pandas as pd
5
6  # Datos del dataset
7  # Status = 2 (Fallecidos)
8  # Sex = 1 (Hombres), 2 (Mujeres)
9
10 # Leer el archivo CSV y seleccionar las columnas necesarias
11 lung_data = pd.read_csv('lung_dataset.csv')[['sex', 'status']].
    dropna()
12
13 # Contar fallecidos y totales para hombres y mujeres
14 male_fallecidos = lung_data[(lung_data['sex'] == 1) & (lung_data[
    'status'] == 2)].shape[0]
15 female_fallecidos = lung_data[(lung_data['sex'] == 2) & (
    lung_data['status'] == 2)].shape[0]
16 male_total = lung_data[lung_data['sex'] == 1].shape[0]
17 female_total = lung_data[lung_data['sex'] == 2].shape[0]
18
19 # Proporciones de fallecidos para hombres y mujeres
20 p_male = male_fallecidos / male_total
21 p_female = female_fallecidos / female_total
22
23 p_pool = (male_fallecidos + female_fallecidos) / (male_total +
    female_total)
24
25 z_stat = (p_male - p_female) / math.sqrt(
26     p_pool * (1 - p_pool) * (1 / male_total + 1 / female_total)
27 )
28
29 alpha = 0.05
30 z_critical = norm.ppf(1 - alpha / 2)

```

Después de obtener los resultados del código vemos que se rechaza la hipótesis nula, por lo que hay diferencias significativas en la mortalidad entre sexos, lo cual podría tener implicaciones importantes para la investigación futura y las políticas de salud pública.

Pat.karno

La puntuación de rendimiento de Karnofsky asignada por el propio paciente es una medida importante del estado funcional percibido por el paciente. **La Figura 5** muestra un histograma de las puntuaciones de rendimiento de Karnofsky asignadas por los pacientes, que revela una distribución asimétrica con un pico alrededor de 80-90. Las puntuaciones de rendimiento de Karnofsky asignadas por los pacientes pueden diferir de las asignadas por los médicos y pueden proporcionar información adicional sobre la percepción del paciente sobre su estado funcional.

Clasificación de la variable

En cuanto al tipo de variable es **cuantitativa discreta** debido a que se da en el número de puntos asignados por el paciente de 0 a 100 y de **intervalos** ya que el 0 tiene un significado.

Medidas de tendencia central

A continuación se presentan las medidas de tendencia central para la variable *pat.karno*:

Medida	Valor
Media	79.95556
Moda	90
Mediana	80
Q1 (25%)	70
Q3 (75%)	90

Table 11: Medidas de tendencia central para la variable *pat.karno*

Medidas de variabilidad

A continuación se presentan las medidas de variabilidad para la variable *pat.karno*:

Medida	Valor
Máximo	100
Mínimo	30
Rango	70
Varianza	213.8373016
Desviación Estándar	14.6231769
Coficiente de Variación	0.1828913

Table 12: Medidas de variabilidad para la variable *pat.karno*

Prueba de Distribución

El siguiente gráfico muestra la distribución de las puntuaciones de rendimiento de Karnofsky asignadas por los pacientes comparada con diferentes tipos de distribución.

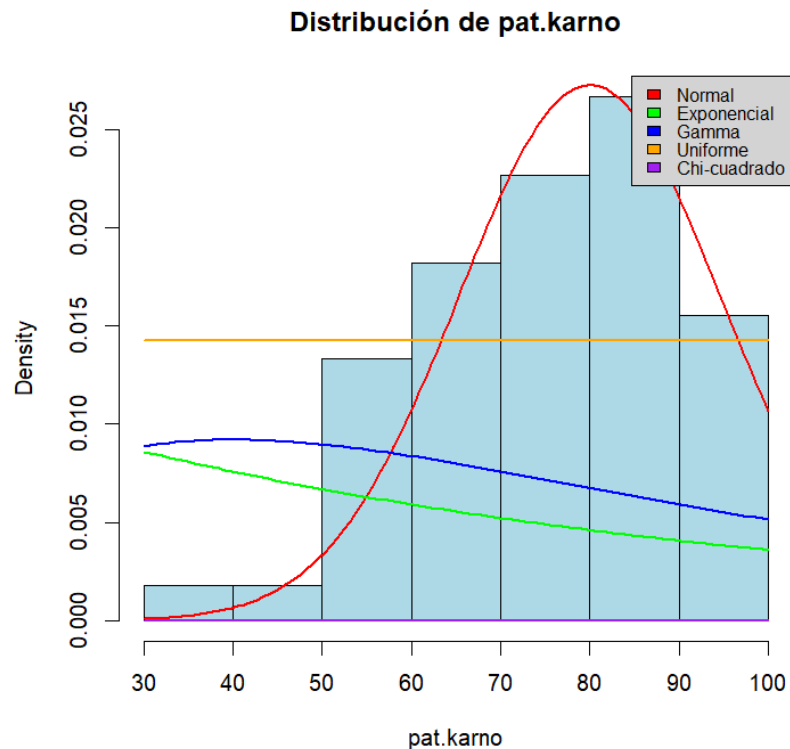


Figure 5: Distribución de las puntuaciones de rendimiento de Karnofsky asignadas por los pacientes

Intervalo de confianza para la media

Prueba de hipótesis para la media

Se plantea que la puntuación promedio asignada por los pacientes con cáncer de pulmón, en una escala de 1 a 100, en función de su bienestar corporal es de 75. Utilizando la implementación análoga para la variable time 2.

```
1
2 muestra = pd.read_csv('lung_dataset.csv')['pat.karno'].dropna().
  values
3
4 # H0: La media de la edad es <= 75
5 # H1: La media de la edad es > 75
6 mu_0 = 75
```

Listing 7: Código en Python para calcular el estimador de la prueba de hipotesis

El resultado de este código nos dice que se rechaza la hipótesis nula, por lo que la calificación que dada por cada paciente es superior a 75, lo que nos dice que en promedio los pacientes sienten mejoría ante su estado de enfermedad.

Wt.loss

La variable `wt.loss` representa la pérdida de peso en los pacientes con cáncer de pulmón. Específicamente, esta variable indica la cantidad de peso que ha perdido un paciente en los últimos seis meses antes de ser diagnosticado o evaluado.

La pérdida de peso es un indicador importante en oncología, ya que puede reflejar el estado nutricional del paciente y la progresión de la enfermedad. En muchos casos, una pérdida significativa de peso puede asociarse con un pronóstico más desfavorable y puede influir en las decisiones sobre el tratamiento y el manejo del paciente.

Es importante considerar esta variable junto con otras características clínicas y demográficas para obtener una comprensión más completa del estado del paciente y su supervivencia. Esta variable tiene valores negativos, lo que indica que algunos pacientes ganaron peso en lugar de perderlo. Esto puede deberse a una variedad de factores, como la retención de líquidos, la ganancia de peso relacionada con el tratamiento o la variabilidad en la medición del peso. En cualquier caso, es importante tener en cuenta esta variabilidad al interpretar los resultados del análisis.

Clasificación de la variable

En cuanto al tipo de variable es **cuantitativa discreta**, aunque el peso generalmente se toma como una variable continua, en este caso se hace un redondeo a los **enteros** que representa el número de libras que ha perdido el paciente. Es de **intervalos** ya que el 0 tiene un significado.

Medidas de tendencia central

A continuación se presentan las medidas de tendencia central para la variable `wt.loss`:

Medida	Valor
Media	9.831776
Moda	0.000000
Mediana	7.000000
Q1 (25%)	0.000000
Q3 (75%)	15.750000

Table 13: Medidas de tendencia central para la variable `wt.loss`

Medidas de variabilidad

A continuación se presentan las medidas de variabilidad para la variable `wt.loss`:

Prueba de Distribución

El siguiente gráfico muestra la distribución de la pérdida de peso en los pacientes comparada con diferentes tipos de distribución.

Medida	Valor
Máximo	68.000000
Mínimo	-24.000000
Rango	92.000000
Varianza	172.657014
Desviación Estándar	13.139902
Coeficiente de Variación	1.336473

Table 14: Medidas de variabilidad para la variable *wt.loss*

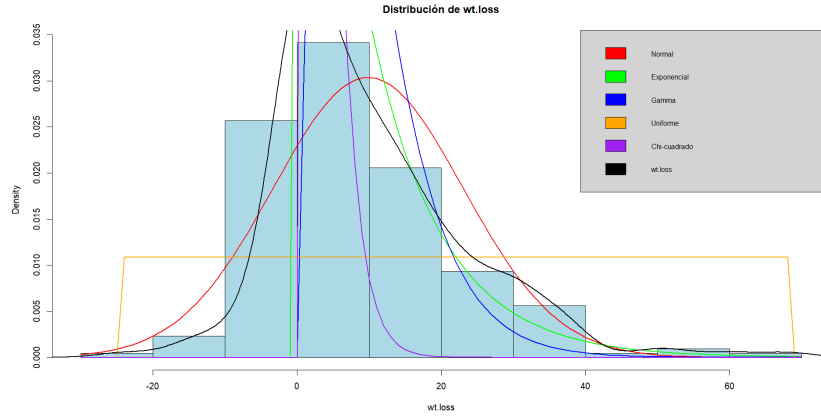


Figure 6: Distribución de la pérdida(o ganancia) de peso

Como es posible notar la distribución empírica de la pérdida de peso no sigue ninguna de las distribuciones conocidas por lo que realizaremos un análisis diferenciado entre las personas que perdieron peso y las que ganaron peso. Haciendo dicho análisis podemos diferenciar los siguientes gráficos:

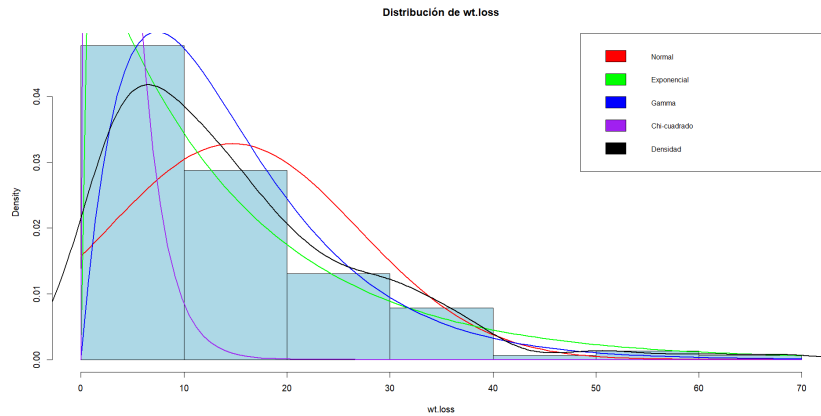


Figure 7: Distribución de la pérdida de peso

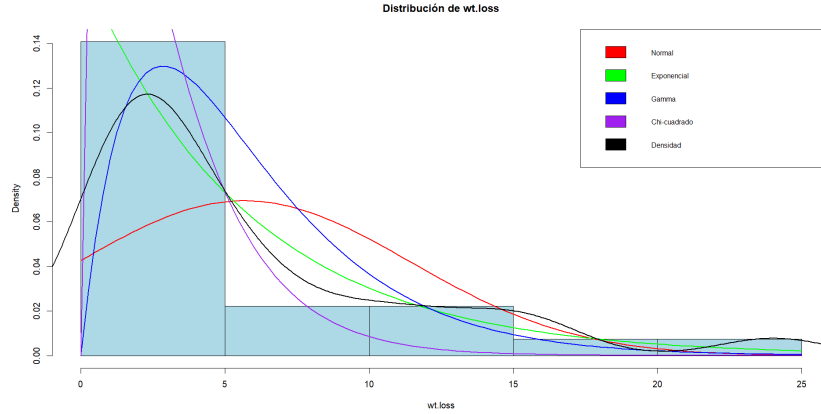


Figure 8: Distribución de la ganancia de peso

Por lo tanto, es conveniente realizar las pruebas de hipótesis para determinar si las variables siguen una distribución específica. Dicha prueba resultó en los siguientes valores:

Prueba	p-value
Anderson-Darling (Normal)	1.197951e-11
Kolmogorov-Smirnov (Normal)	0.005456181
Shapiro-Wilk (Normal)	4.066897e-10
Kolmogorov-Smirnov (Exponencial)	0.09287115
Kolmogorov-Smirnov (Gamma)	0.03941509
Kolmogorov-Smirnov (Chi-cuadrado)	7.692063e-54

Table 15: Resultados de las pruebas de distribución para la variable *wt.loss* de los que perdieron peso

Prueba	p-value
Anderson-Darling (Normal)	1.613223e-05
Kolmogorov-Smirnov (Normal)	0.06862226
Shapiro-Wilk (Normal)	6.581438e-05
Kolmogorov-Smirnov (Exponencial)	0.4796857
Kolmogorov-Smirnov (Gamma)	0.03938671
Kolmogorov-Smirnov (Chi-cuadrado)	0.03611553

Table 16: Resultados de las pruebas de distribución para la variable *wt.loss* de los que ganaron peso

Por lo que podemos concluir que tanto la distribución de la pérdida de peso como la de la ganancia de peso siguen una distribución Exponencial.

Intervalo de confianza para las media de las personas que perdieron peso

El intervalo de confianza para la media de la variable **wt.loss** de las personas que perdieron peso es de 12.086570460717061 a 18.346939910801844 con un nivel de confianza del 99%. Esto lo podemos saber gracias al siguiente código:

```

1  import pandas as pd
2  import numpy as np
3  from scipy.stats import chi2
4
5  # Leer el archivo CSV y seleccionar la columna 'wt.loss'
6  df = pd.read_csv('lung_dataset.csv')
7  wt_loss = df['wt.loss'].dropna()
8
9  # Filtrar los valores mayores que 0
10 wt_loss_filtered = wt_loss[wt_loss > 0].values
11
12 # Parametros
13 n = len(wt_loss_filtered)
14 alpha = 0.01 # Nivel de significancia (para un intervalo de
15               confianza del 99%)
16
17 # Suma de los datos
18 sum_data = np.sum(wt_loss_filtered)
19
20 # Valores criticos de chi-cuadrado
21 chi2_lower = chi2.ppf(alpha / 2, 2 * n)
22 chi2_upper = chi2.ppf(1 - alpha / 2, 2 * n)
23
24 # Intervalo de confianza
25 lower_bound = (2 * sum_data)/chi2_upper
26 upper_bound = (2 * sum_data)/chi2_lower
27
28 print(f"Intervalo de confianza para la media de una
29       distribucion exponencial: ({lower_bound}, {upper_bound})")

```

Listing 8: Código en Python para calcular el intervalo de confianza para la media de una distribución exponencial

Intervalo de confianza para la media de la personas que ganaron peso

Análogamente el intervalo de confianza para las media de las personas que ganaron peso está entre 3.621220157012075 y 9.87694081167703 con un 99% de confianza.

Prueba de Hipótesis

El cáncer de pulmón avanzado es uno de los que estadísticamente provoca más pérdida de peso en aquellos que lo sufren. Según estudios, perder más del 5% del peso corporal en 6 meses o menos es un indicador de un deterioro físico avanzado, dado que el promedio de masa corporal humana a nivel mundial es de 137 lbs, probaremos que aquellos pacientes que fallecen tienden a tener una pérdida de peso corporal de 6.85 lbs (5% de 137 lbs).

Tomemos como hipótesis nula: El valor medio de la pérdida de peso en 6 meses de los pacientes que fallecieron durante el estudio es menor de 6.85 lbs.

Por tanto la hipótesis alternativa es que dicho valor es mayor que 6.85 lbs.

Esto lo podemos calcular con el siguiente código, para un nivel de significación del 0.01

```

1 import numpy as np
2 import pandas as pd
3 from scipy.stats import chi2
4
5 valor_hipotetico = 6.85
6 # Leer el archivo CSV y seleccionar la columna 'wt.loss' y '
   status'
7 df = pd.read_csv('lung_dataset.csv')
8 wt_loss = df[df['status'] == 2]['wt.loss'].dropna()
9
10 # Filtrar los valores mayores que 0
11 wt_loss_filtered = wt_loss[wt_loss > 0].values
12
13 def prueba_hipotesis_exponencial(datos, media_poblacional, alfa
   =0.05):
14     """
15     Realiza una prueba de hipotesis para verificar si la media
       poblacional
16     es menor que un valor dado en una distribucion exponencial.
17
18     Parametros:
19     - datos: lista o array de datos muestrales.
20     - media_poblacional: valor de media poblacional bajo la
       hipotesis nula.
21     - alfa: nivel de significancia (por defecto 0.05).
22
23     Retorna:
24     - Resultado de la prueba y el valor p.
25     """
26     # Calculo de la media muestral y el tamano de la muestra
27     media_muestral = np.mean(datos)
28     n = len(datos)
29
30     # Hipotesis:
31     # H0: Media poblacional <= media_poblacional
32     # H1: Media poblacional > media_poblacional (prueba
       unilateral)
33
34     # Varianza muestral para una exponencial (estimador de maxima
       verosimilitud)
35     varianza_muestral = np.var(datos, ddof=1)
36
37     # Estadistico de prueba (estimador 1/x)
38     estadistico_prueba = n * (media_muestral / media_poblacional)
39     #print(estadistico_prueba)
40     # Valor critico y comparacion
41     valor_critico = chi2.ppf(1 - alfa, df=n)
42     #print(valor_critico)
43     p_valor = 1 - chi2.cdf(estadistico_prueba, df=n)
44

```

```

45     # Resultado
46     if estadistico_prueba > valor_critico:
47         resultado = "Rechazamos la hipotesis nula"
48     else:
49         resultado = "No podemos rechazar la hipotesis nula"
50
51     return resultado, p_valor, estadistico_prueba
52
53 resultado, p_valor, estadistico = prueba_hipotesis_exponencial(
54     wt_loss_filtered, 6.85, 0.01)
55 print(f"Resultado: {resultado}")
56 print(f"Valor p: {p_valor}")
57 print(f"Estadistico de prueba: {estadistico}")

```

Bajo esta prueba se rechaza la hipótesis nula bajo un p-value de 4.779732165616224e-12 por lo que podemos decir que podemos asumir la hipótesis alternativa con un 1% de significancia, y por tanto estadísticamente las personas que fallecen presentan una disminución de peso de más de 6.85 lbs

Análisis de Correlación

Para entender mejor las relaciones entre las variables numéricas en nuestro conjunto de datos, realizamos un análisis de correlación utilizando dos métodos diferentes: el coeficiente de correlación de Pearson y el coeficiente de correlación de Spearman.

Metodología

Se utilizaron dos matrices de correlación:

- **Correlación de Pearson:** Mide la fuerza y dirección de la relación lineal entre variables.
- **Correlación de Spearman:** Evalúa la relación monótona entre variables, siendo más robusta a valores atípicos.

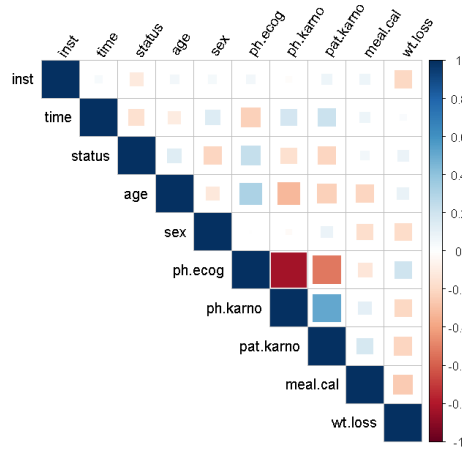


Figure 9: Matriz de correlación de Pearson

Correlaciones Detalladas

Par de Variables	Correlación de Pearson
inst - time	0.02
inst - status	-0.13
inst - age	0.05
inst - sex	0.08
inst - ph.ecog	0.06
inst - ph.karno	-0.02
inst - pat.karno	0.04
inst - meal.cal	0.10
inst - wt.loss	-0.17
time - status	-0.16
time - age	-0.08
time - sex	0.11
time - ph.ecog	-0.19
time - ph.karno	0.09
time - pat.karno	0.18
time - meal.cal	0.07
time - wt.loss	0.03
status - age	0.16
status - sex	-0.22
status - ph.ecog	0.24
status - ph.karno	-0.16
status - pat.karno	-0.19
status - meal.cal	0.02
status - wt.loss	0.05

Table 17: Correlaciones de Pearson (Parte 1)

Par de Variables	Correlación de Spearman
ph.ecog - ph.karno	-0.84
ph.ecog - pat.karno	-0.53
ph.karno - pat.karno	0.51
age - ph.ecog	0.32
age - ph.karno	-0.33
age - pat.karno	-0.24
status - ph.ecog	0.24
time - ph.ecog	-0.24
status - sex	-0.22
pat.karno - wt.loss	-0.22
meal.cal - wt.loss	-0.25
time - pat.karno	0.21
time - ph.karno	0.19
sex - wt.loss	-0.19
ph.karno - wt.loss	-0.20

Table 18: Correlaciones de Spearman más relevantes ($-r-$ ¿ 0.15)

Hallazgos Principales

Las correlaciones más significativas (con umbral de 0.7) encontradas fueron:

- **ph.karno y pat.karno:** Correlación de 0.72. Esto indica una fuerte relación positiva entre las evaluaciones de rendimiento realizadas por médicos y pacientes. Sugiere consistencia entre las percepciones de los profesionales y los pacientes sobre el estado funcional.
- La correlación negativa más fuerte se observa entre **ph.ecog** y **ph.karno** ($r = -0.84$), lo cual es esperado ya que miden aspectos similares pero en escalas opuestas.
- El **status** muestra correlaciones moderadas negativas con **sex** y positivas con **ph.ecog**, sugiriendo diferencias en la supervivencia entre géneros.
- La **edad** muestra correlaciones moderadas con las medidas de rendimiento (**ph.ecog**, **ph.karno**, **pat.karno**), indicando que el estado funcional tiende a deteriorarse con la edad.

Esta correlación tiene sentido desde una perspectiva clínica, ya que ambas medidas evalúan el estado funcional del paciente, aunque desde diferentes perspectivas. La moderada pero no perfecta correlación sugiere que, si bien hay acuerdo general entre las evaluaciones de médicos y pacientes, también hay diferencias importantes en las percepciones que podrían ser clínicamente relevantes.

Análisis de Regresión

El análisis de regresión con un modelo OLS en este caso no es factible por ciertas razones expuestas posteriormente, por tanto proponemos dicho análisis utilizando una técnica conocida, el modelo Weibull AFT.

La regresión con modelo Weibull AFT (Accelerated Failure Time) es una técnica utilizada en análisis de supervivencia para modelar el tiempo hasta que ocurre un evento de

interés, como la muerte o la falla de un componente. En este contexto, el modelo Weibull AFT asume que el tiempo de supervivencia se puede modelar como una función de los efectos de las covariables y que sigue una distribución Weibull.

¿En qué consiste un modelo Weibull AFT?

1. **Estructura del modelo:** En un modelo AFT, se supone que el tiempo hasta el evento (T) está relacionado con un vector de covariables (X) a través de una relación multiplicativa. Específicamente, se puede expresar como:

$$T = \exp(X\beta) \cdot Z$$

donde Z es una variable aleatoria que sigue una distribución Weibull.

2. **Distribución Weibull:** La función de supervivencia para la distribución Weibull se define como:

$$S(t) = e^{-(t/\lambda)^\kappa}$$

donde λ es el parámetro de escala y κ es el parámetro de forma. Dependiendo del valor de κ , la distribución puede ser monótonamente creciente o decreciente.

3. **Interpretación:** En el contexto de la regresión AFT, los coeficientes estimados (β) indican cómo las covariables afectan el tiempo hasta el evento. Un coeficiente positivo implica que un aumento en la covariable está asociado con un aumento en el tiempo hasta el evento (es decir, una vida útil más larga).

¿Por qué usar Weibull AFT en vez de OLS?

1. **Naturaleza del dato:** OLS (Mínimos Cuadrados Ordinarios) asume que los errores son normalmente distribuidos y que la relación entre las variables es lineal. Sin embargo, en análisis de supervivencia, los datos suelen estar censurados (es decir, no todos los individuos experimentan el evento antes de que termine el estudio), lo que hace que OLS no sea apropiado.

2. **Censura:** El modelo AFT maneja adecuadamente la censura, permitiendo que se utilicen todos los datos disponibles, incluso aquellos para los cuales no se ha observado el evento. OLS no puede manejar la censura correctamente, lo que puede llevar a estimaciones sesgadas.

3. **Distribución del tiempo:** El modelo Weibull AFT permite modelar explícitamente la distribución del tiempo hasta el evento, lo cual es fundamental en análisis de supervivencia. OLS no tiene en cuenta la naturaleza del tiempo hasta el evento y puede dar resultados engañosos.

4. **Interpretación más adecuada:** Los modelos AFT proporcionan interpretaciones más intuitivas en términos del tiempo hasta un evento, lo que es especialmente útil en contextos clínicos o industriales donde el tiempo es una variable crítica.

En resumen, el uso del modelo Weibull AFT es preferido en situaciones donde se está analizando el tiempo hasta un evento y se presentan datos censurados, mientras que OLS no es adecuado debido a sus supuestos y limitaciones en este contexto.

Fórmula del Modelo

El modelo Weibull AFT se plantea como:

$$\ln(T) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{sex} + \beta_3 \cdot \text{pat.karno} + \beta_4 \cdot \text{wt.loss} + \sigma\epsilon \quad (1)$$

Componentes

- **Variables del dataset:**

- *T*: **time** (tiempo de supervivencia en días)
- **status**: Indicador de censura (1 = evento observado, 2 = censurado)
- **age**: Edad en años
- **sex**: Género (1 = hombre, 0 = mujer) (Se realizó un ajuste para el modelo)
- **pat.karno**: Puntuación Karnofsky (0-100)
- **wt.loss**: Pérdida de peso (lbs)

- **Parámetros:**

- β_0 : Intercepto
- β_1, \dots, β_4 : Coeficientes de regresión
- σ : Parámetro de escala Weibull
- ϵ : Error con distribución de valor extremo (extreme value para Weibull)

Fórmula del Modelo Ajustado

Luego de analizar los p valores (descartar aquellos mayores a 0.05, no son significativos para el modelo) el modelo queda:

$$\ln(T) = \beta_0 + \beta_1 \cdot \text{sex} + \beta_2 \cdot \text{pat.karno} + \sigma \epsilon \quad (2)$$

Resultados de la Regresión

Predictor	Coeficiente (β)	Error Estándar	z-value	p-value
Intercepto	5.16967	0.37717	13.71	$< 2 \times 10^{-16}$
sex (Hombre=1)	-0.33580	0.14435	-2.33	0.0200
pat.karno	0.01373	0.00461	2.98	0.0029
Log(scale)	-0.31650	0.07224	-4.38	1.2×10^{-5}

Escala (scale): 0.729

Distribución: Weibull

Interpretación Clave

Efectos de los Predictores

- **Sexo** ($\beta_{\text{sex}} = -0.336$):

$\exp(-0.336) \approx 0.714 \implies$ Hombres tienen 28.6% menos tiempo de supervivencia que mujeres

Ejemplo: Si una mujer tiene un tiempo de supervivencia estimado de 200 días, un hombre comparable tendría un tiempo de supervivencia estimado de $200 \times 0.714 \approx 143$ días. Esto sugiere que el género masculino está asociado con una menor supervivencia en comparación con el género femenino en este conjunto de datos.

- **Puntuación Karnofsky** ($\beta_{\text{pat.karno}} = 0.0137$):

$\exp(0.0137) \approx 1.0138 \implies$ Cada punto adicional en la puntuación Karnofsky aumenta la supervivencia en un 1.38%.

Ejemplo: Si un paciente mejora su puntuación Karnofsky en 10 puntos, su tiempo de supervivencia se incrementaría en aproximadamente un 14.6%. Esto indica que una mejor percepción del estado funcional por parte del paciente está asociada con una mayor supervivencia.

Parámetro de Forma Weibull

El parámetro de forma de la distribución Weibull es crucial para entender la naturaleza del riesgo a lo largo del tiempo. En este caso:

$$\alpha = \frac{1}{\text{scale}} = \frac{1}{0.729} \approx 1.372 \quad (\text{Riesgo creciente con el tiempo})$$

Un parámetro de forma $\alpha > 1$ indica que el riesgo de fallecimiento aumenta con el tiempo. Esto es consistente con la progresión de enfermedades como el cáncer de pulmón, donde el riesgo de mortalidad tiende a incrementarse a medida que la enfermedad avanza.

Conclusiones del Modelo

El análisis de regresión con el modelo Weibull AFT ha revelado varios hallazgos importantes:

- El género masculino está asociado con una menor supervivencia en comparación con el género femenino.
- Una mejor puntuación Karnofsky, que refleja una mejor percepción del estado funcional por parte del paciente, está asociada con una mayor supervivencia.
- El riesgo de fallecimiento aumenta con el tiempo, lo cual es consistente con la naturaleza progresiva del cáncer de pulmón avanzado.

Estos resultados subrayan la importancia de considerar tanto las características demográficas como las percepciones del estado funcional en el análisis de supervivencia de pacientes con cáncer de pulmón. Además, el uso del modelo Weibull AFT ha permitido manejar adecuadamente la censura en los datos y proporcionar una interpretación más precisa de los factores que influyen en la supervivencia.

Conclusiones

El conjunto de datos del cáncer de pulmón NCCTG proporciona información valiosa sobre los resultados de supervivencia en pacientes con cáncer de pulmón avanzado. Los datos destacan la importancia de las puntuaciones de rendimiento y la pérdida de peso como posibles predictores de la supervivencia. Los estudios futuros podrían utilizar este conjunto de datos para modelado pronóstico más profundo y validación de herramientas clínicas.

Referencias

- Loprinzi CL, Laurie JA, Wieand HS, Krook JE, Novotny PJ, Kugler JW, et al. "Evaluación prospectiva de variables pronósticas a partir de cuestionarios completados por los pacientes." *Journal of Clinical Oncology*. 12(3):601-7, 1994.
- Therneau T. Documentación del paquete *survival*.