Análisis de Datos del Cáncer de Pulmón NCCTG

December 18, 2024

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
Coeficiente de Variación	0.6901152

Table 2: Medidas de variabilidad para el tiempo de supervivencia

Prueba de Distribución

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

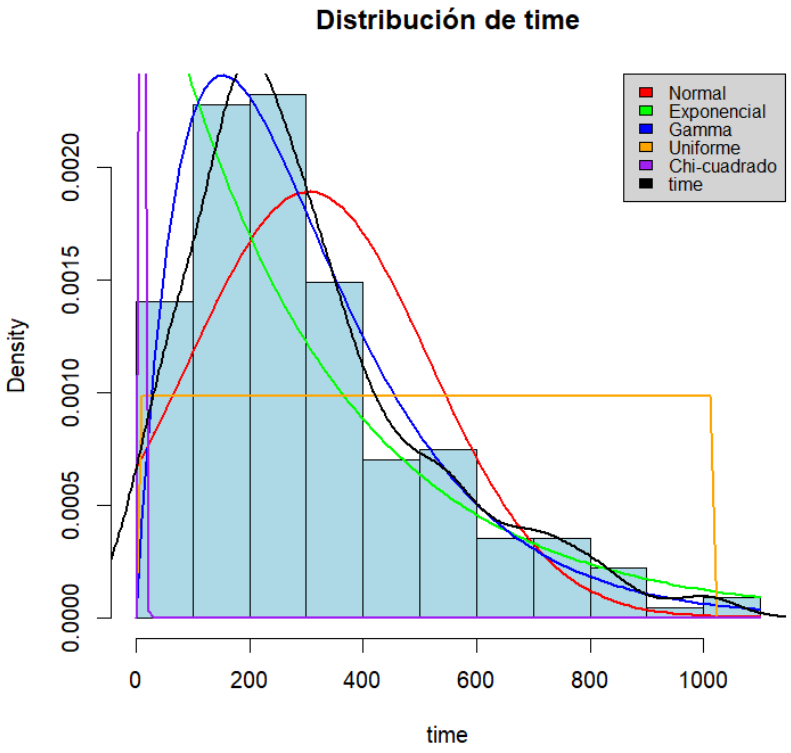

Figure 1: Distribución de los tiempos de supervivencia

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:

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

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*

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 :

```
import pandas as pd
import numpy as np
from scipy.stats import norm as z

# Leer el archivo CSV y seleccionar la columna 'time'
muestra = pd.read_csv('lung_dataset.csv')['time'].dropna().
    values


media_muestral = np.mean(muestra)
desviacion_muestral = np.std(muestra, ddof=1)
n = len(muestra)

# Nivel de confianza y grados de libertad
nivel_significancia = 0.01
confianza = 1 - nivel_significancia


z_critico = z.ppf(1 - nivel_significancia / 2)


margen_error = z_critico * (desviacion_muestral / np.sqrt(n))
limite_inferior = media_muestral - margen_error
limite_superior = media_muestral + margen_error

print(media_muestral, desviacion_muestral, (limite_inferior,
    limite_superior))
```

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

Hipótesis atractiva (luego creas una)

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 hipotesis

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
Coeficiente 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.

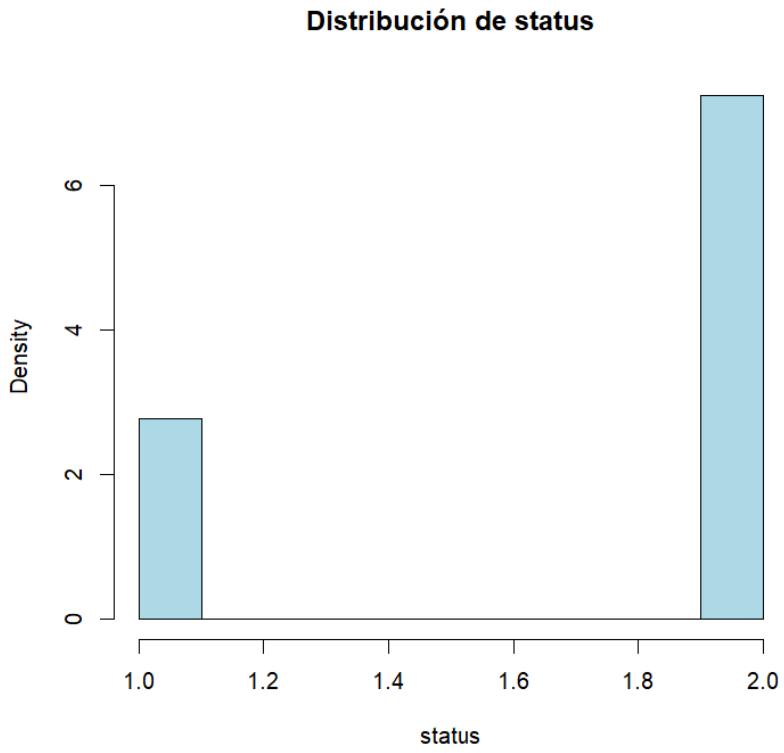

Figure 2: Distribución de los tiempos de supervivencia

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:

```
    import pandas as pd
    import numpy as np
    from scipy.stats import norm as z

    # Leer el archivo CSV y seleccionar la columna 'status'
    muestra = pd.read_csv('lung_dataset.csv')['status'].dropna().
       values

    n = len(muestra)
    proporcion_fallecidos = np.sum(muestra == 2) / n

    # Nivel de confianza
    nivel_significancia = 0.01
    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
            - proporcion_fallecidos) / n)
18
19      limite_inferior = proporcion_fallecidos - margen_error
20      limite_superior = proporcion_fallecidos + margen_error
21
22      print(f"Intervalo de confianza al {confianza*100}%: ({
           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
Coeficiente 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.

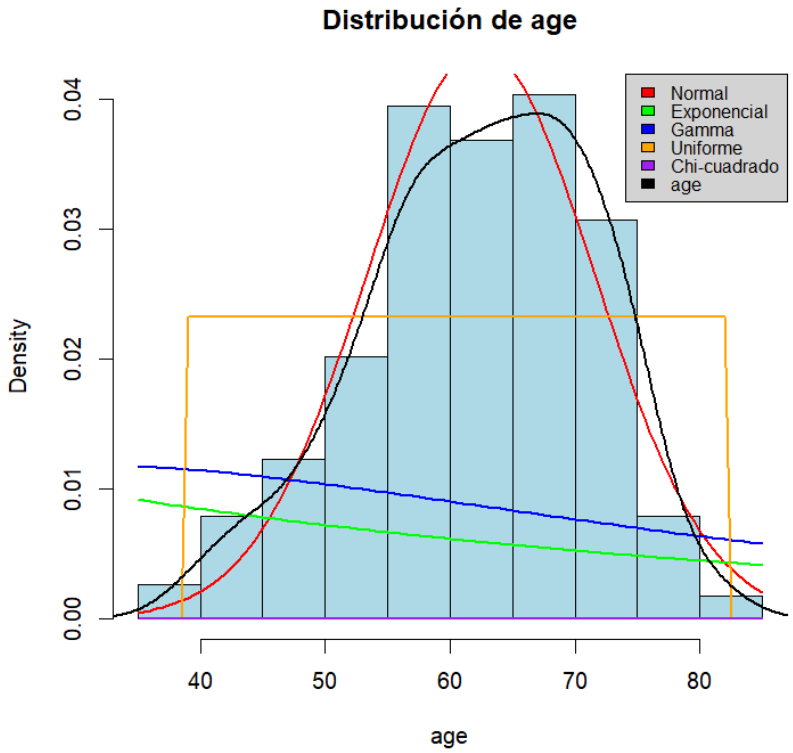

Figure 3: Distribución de la edad de los pacientes

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:

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 siguienete.

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*

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:

```
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.

```


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

# H0: La media de la edad es <= 60
# H1: La media de la edad es > 60
mu_0 = 60
```

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

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
Coeficiente 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.

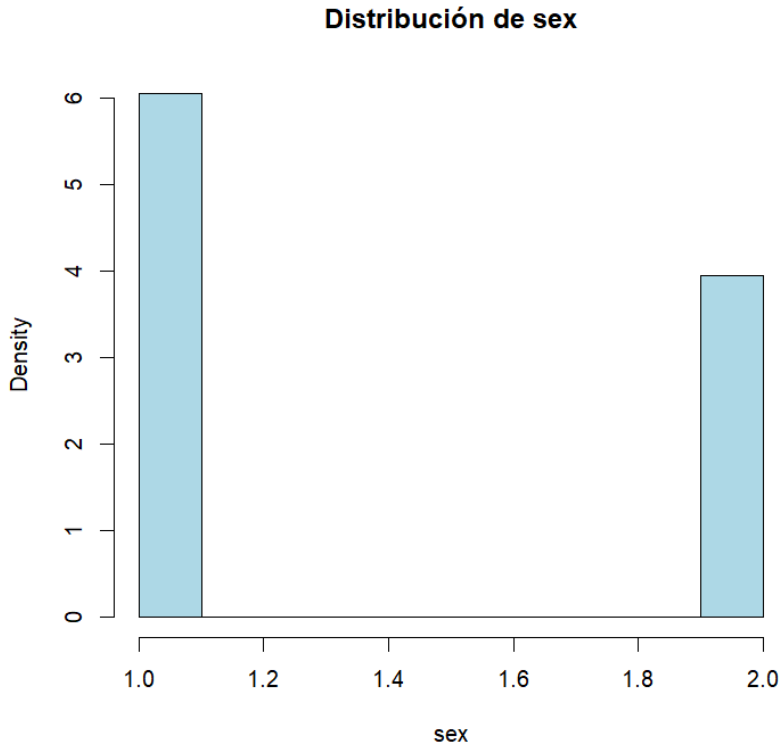

Figure 4: Distribución del género de los pacientes

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:

```
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.

```
        import math
from scipy.stats import norm
import pandas as pd

# Datos del dataset
# Status = 2 (Fallecidos)
# Sex = 1 (Hombres), 2 (Mujeres)

# Leer el archivo CSV y seleccionar las columnas necesarias
lung_data = pd.read_csv('lung_dataset.csv')[['sex', 'status']].
    dropna()

# Contar fallecidos y totales para hombres y mujeres
male_fallecidos = lung_data[(lung_data['sex'] == 1) & (lung_data[
    'status'] == 2)].shape[0]
female_fallecidos = lung_data[(lung_data['sex'] == 2) & (
    lung_data['status'] == 2)].shape[0]
male_total = lung_data[lung_data['sex'] == 1].shape[0]
female_total = lung_data[lung_data['sex'] == 2].shape[0]

# Proporciones de fallecidos para hombres y mujeres
p_male = male_fallecidos / male_total
p_female = female_fallecidos / female_total

p_pool = (male_fallecidos + female_fallecidos) / (male_total +
    female_total)

z_stat = (p_male - p_female) / math.sqrt(
    p_pool * (1 - p_pool) * (1 / male_total + 1 / female_total)
)

alpha = 0.05
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
Coeficiente 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.

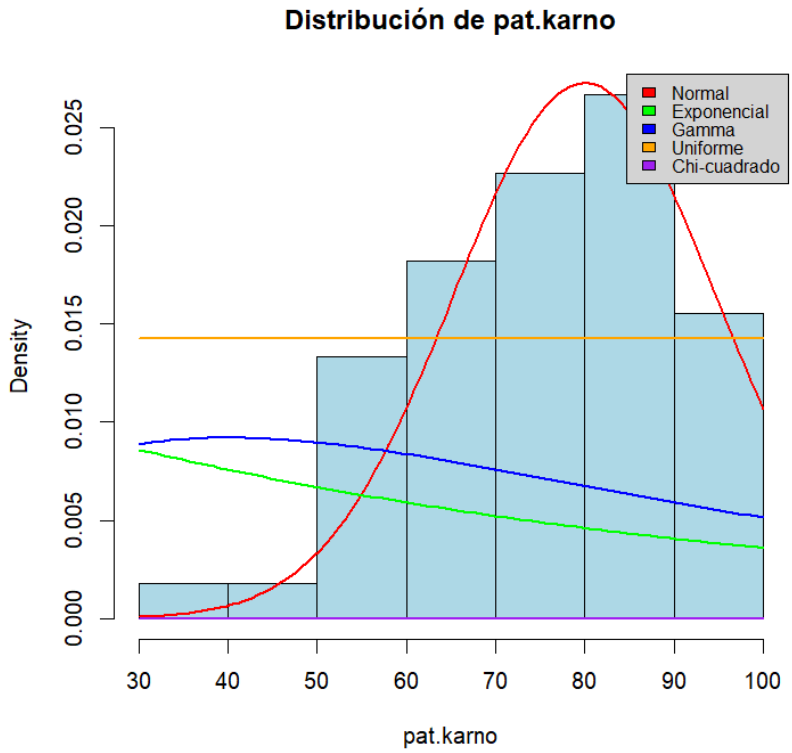

Figure 5: Distribución de las puntuaciones de rendimiento de Karnofsky asignadas por los pacientes

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.

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*.