Universidad Nacional Autónoma de México

Facultad de Estudios Superiores: Acatlán

ACTUARÍA

ESTADÍSTICA III

EXAMEN FINAL

Comportamiento Proteico en Personas con Cancer de MaMa

Un estudio estadístico en pacientes con cáncer de mama

21 de mayo del 2025

Índice

1. Introducción al dataset **3**

2. Análisis exploratorio **5**

3. Hipótesis y fundamentos teóricos **10**
3.1. Pruebas de bondad y Ajuste . . . 10
3.2. Pruebas de comparación entre dos distribuciones . . . 11
3.2.1. Descripción de la prueba de Mann–Whitney . . . 11
3.2.2. Prueba de Wilcoxon de los rangos con signo . . . 11
3.3. Pruebas de independencia para vectores aleatorios bivariados . . . 13
3.3.1. Prueba de Spearman . . . 13
3.3.2. Correlación de Kendall (τ) . . . 13
3.3.3. Prueba de Independencia de Hoeffding . . . 13
3.3.4. Prueba de Independencia de Genest–Rémillard . . . 14

4. Desarrollo estadístico (pruebas aplicadas) **15**
4.1. Pruebas de Bondad y ajuste . . . 15
4.1.1. Kolmogorov Smirnov . . . 15
4.1.2. Prueba de Cramér–von Mises . . . 15
4.1.3. Anderson-Darling . . . 16
4.2. Implementación de Pruebas para comparación de dos distribuciones . . . 18
4.2.1. Wilcoxon . . . 18
4.2.2. MannWhitney-Wilcoxon . . . 18
4.3. Pruebas de independencia para vectores aleatorios bivariados . . . 19
4.3.1. Prueba de Kendall . . . 19
4.3.2. Prueba de Spearman . . . 19
4.3.3. Prueba de Hoeffding . . . 20
4.3.4. Prueba de Genest–Rémillard . . . 21

1. Introducción al dataset

El cáncer de mama es una enfermedad compleja y multifactorial que representa una de las principales causas de morbilidad y mortalidad en mujeres a nivel mundial. Su detección oportuna, caracterización clínica y tratamiento adecuado son aspectos fundamentales para mejorar la esperanza y calidad de vida de las pacientes. En este contexto, el análisis estadístico de bases de datos clínicas resulta ser una herramienta clave para generar conocimiento, evaluar patrones y apoyar la toma de decisiones médicas basadas en evidencia.

En este trabajo se analiza el conjunto de datos denominado **Real Breast Cancer Data**, el cual fue extraído de la plataforma *Kaggle* y contiene información detallada de pacientes que fueron sometidas a procedimientos quirúrgicos para extirpar tumores malignos de mama. La base no solo registra datos sociodemográficos básicos, como edad y género, sino que también incluye variables clínicas cruciales como tipo de tumor, presencia de marcadores moleculares, tipo de cirugía realizada, y estado vital posterior a la operación.

Este tipo de datos permite el desarrollo de análisis exploratorios, descriptivos e inferenciales. Asimismo, facilita la identificación de asociaciones entre características clínicas y desenlaces, lo cual puede resultar de gran valor en el diseño de estrategias de detección temprana, en la planificación de tratamientos individualizados y en la investigación biomédica.

Entre los objetivos principales del análisis se encuentran:

- Describir la distribución de características demográficas y clínicas de las pacientes.
- Evaluar la prevalencia de distintos subtipos histológicos y etapas tumorales.
- Identificar patrones comunes en la asignación de tipo de cirugía según la etapa o histología.
- Explorar posibles factores asociados con la sobrevida o mortalidad.
- Proveer una base sólida para la generación de hipótesis futuras o incluso modelos predictivos en estudios más avanzados.

El data set es especialmente valioso debido a que incorpora múltiples dimensiones del cáncer de mama, incluyendo:

- Factores clínicos visibles al momento del diagnóstico (edad, etapa del tumor, tipo de histología).
- Indicadores moleculares fundamentales para el diagnóstico y tratamiento personalizado (receptores hormonales, HER2, proteínas).
- Información del proceso terapéutico (tipo de cirugía y seguimiento posterior).
- Estado final del paciente (vivo/muerto), permitiendo análisis de tipo longitudinal y de supervivencia.

Las variables incluidas en el conjunto de datos son las siguientes:

- **Patient_ID**: Identificador único del paciente.

- **Edad**: Edad al momento del diagnóstico, expresada en años.
- **Género**: Masculino o femenino.
- **Protein1, Protein2, Protein3, Protein4**: Niveles de expresión de cuatro proteínas clínicas (unidades no especificadas).
- **Etapa del tumor (Tumour_Stage)**: Clasificación clínica del tumor en etapas I, II o III.
- **Histología**: Tipo histológico del tumor; incluye carcinoma ductal infiltrante, lobulillar infiltrante y mucinoso.
- **ER status**: Estado del receptor de estrógeno (Positivo/Negativo).
- **PR status**: Estado del receptor de progesterona (Positivo/Negativo).
- **HER2 status**: Estado del receptor HER2 (Positivo/Negativo).
- **Tipo de cirugía (Surgery_type)**: Procedimiento realizado (Lumpectomía, Mastectomía simple, Mastectomía radical modificada, Otro).
- **Fecha de cirugía (Date_of_Surgery)**: Fecha exacta de la operación (formato DD-MM-AAAA).
- **Fecha de última visita (Date_of_Last_Visit)**: Última fecha de seguimiento clínico registrada.
- **Estado del paciente (Patient_Status)**: Vivo/Muerto. Puede estar vacía si no hay información posterior a la cirugía.

En conjunto, esta base de datos permite un análisis estadístico integral que va desde la descripción simple hasta el planteamiento de modelos explicativos o predictivos. Su riqueza radica en la combinación de información clínica objetiva con marcadores moleculares y datos de seguimiento, lo que ofrece una visión holística del comportamiento del cáncer de mama en una muestra realista y representativa.

La base fue tomada de Kaggle y se puede consultar en: `https://www.kaggle.com/datasets/amandam1/breastcancerdataset`

2. Análisis exploratorio

El análisis exploratorio de datos (EDA, por sus siglas en inglés) constituye una etapa fundamental en cualquier estudio estadístico o científico de datos. En esta fase, el objetivo principal es comprender la estructura de la información, detectar patrones generales, identificar valores atípicos, observar posibles relaciones entre variables y evaluar la calidad del conjunto de datos. A través de esta primera aproximación, se definen hipótesis preliminares que podrán ser contrastadas más adelante mediante análisis más rigurosos.

Dado que el presente data set contiene una amplia variedad de variables, tanto numéricas como categóricas, se optó por iniciar con una visualización general de las primeras filas, tal como se muestra a continuación:

	Patient_ID	Age	Gender	Protein1	Protein2	Protein3	Protein4	Tumour_Stage	Histology	ER status	PR status	HER2 status	Surgery_type	Date_of_Surgery	Date_of_Last_Visit	Patient_Status	Her Status Bin	Patient_Status Bin
0	TCGA-D8-A1XD	36	FEMALE	0.080353	0.42638	0.54715	0.273680	III	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Modified Radical Mastectomy	15-Jan-17	19-jun-17	Alive	0	1
1	TCGA-EW-A1OX	43	FEMALE	-0.420320	0.57807	0.61447	-0.031505	II	Mucinous Carcinoma	Positive	Positive	Negative	Lumpectomy	26-Apr-17	09-nov-18	Dead	0	0
2	TCGA-A8-A079	69	FEMALE	0.213980	1.31140	-0.32747	-0.234260	III	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Other	08-sep-17	09-jun-18	Alive	0	1
3	TCGA-D8-A1XR	56	FEMALE	0.345090	-0.21147	-0.19304	0.124270	II	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Modified Radical Mastectomy	25-Jan-17	12-jul-17	Alive	0	1
4	TCGA-BH-A0BF	56	FEMALE	0.221550	1.90680	0.52045	-0.311990	II	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Other	06-may-17	27-jun-19	Dead	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
329	TCGA-AN-A04A	36	FEMALE	0.231800	0.61804	-0.55779	-0.517350	III	Infiltrating Ductal Carcinoma	Positive	Positive	Positive	Simple Mastectomy	11-nov-19	09-feb-20	Dead	1	0
330	TCGA-A8-A085	44	MALE	0.732720	1.11170	-0.26952	-0.354920	II	Infiltrating Lobular Carcinoma	Positive	Positive	Negative	Other	01-nov-19	04-mar-20	Dead	0	0
331	TCGA-A1-A0SG	61	FEMALE	-0.719470	2.54850	-0.15024	0.339680	II	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Lumpectomy	11-nov-19	18-Jan-21	Dead	0	0
332	TCGA-A2-A0EU	79	FEMALE	0.479400	2.05590	-0.53136	-0.188480	I	Infiltrating Ductal Carcinoma	Positive	Positive	Positive	Lumpectomy	21-nov-19	19-feb-21	Dead	1	0
333	TCGA-B6-A40B	76	FEMALE	-0.244270	0.92556	-0.41823	-0.067848	I	Infiltrating Ductal Carcinoma	Positive	Positive	Negative	Lumpectomy	11-nov-19	05-Jan-21	Dead	0	0

Figura 1: Visualización de las primeras filas de la base

Este vistazo preliminar permite observar que los datos incluyen información clínica detallada como la edad, el tipo de tumor, la presencia o ausencia de ciertos receptores hormonales, así como variables temporales (fechas de cirugía y de última visita) y el estado vital del paciente.

Un primer grupo de variables que resulta de especial interés es el correspondiente a los niveles de expresión de proteínas (**Protein1, Protein2, Protein3, Protein4**). Estas variables son numéricas, lo que permite aplicar directamente herramientas estadísticas, y se presume que podrían estar relacionadas tanto entre sí como con el desenlace del paciente (es decir, si se encuentra vivo o no tras la cirugía).

Para comenzar con la exploración de estas variables, se generaron histogramas individuales que permiten visualizar sus distribuciones. A continuación, se presenta un ejemplo:

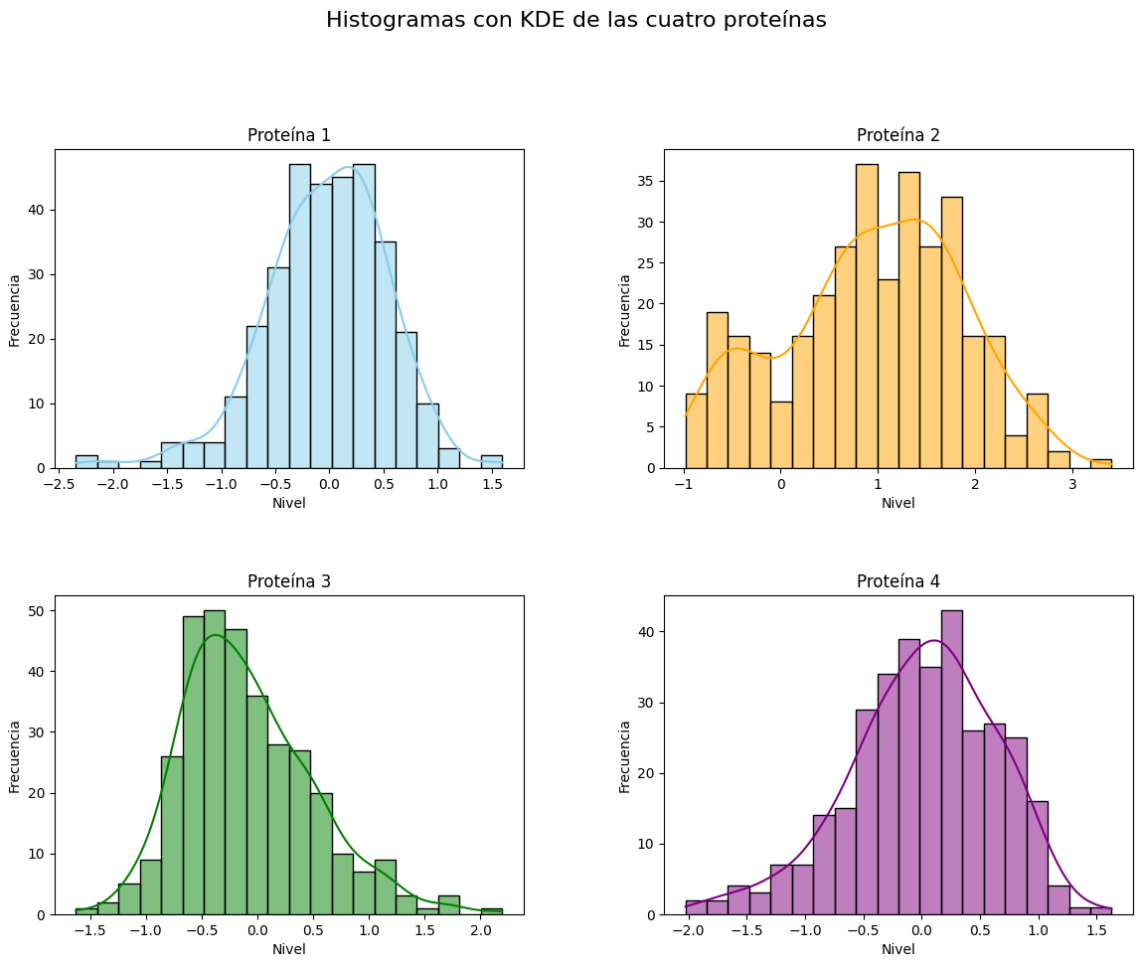


Figura 2: Histograma de la distribuciónn de las proteínas

La distribución de la edad muestra una concentración entre los 40 y 60 años, lo cual es consistente con la literatura médica que indica que el cáncer de mama es más frecuente en mujeres de mediana edad. Esta observación preliminar puede ser útil para segmentar a la población en subgrupos etarios y analizar si ciertas variables clínicas o desenlaces difieren entre ellos.

A lo largo de esta sección, se seguirán explorando otras variables categóricas relevantes (como el tipo de cirugía, la etapa del tumor o los receptores hormonales), así como posibles asociaciones entre ellas y variables numéricas como las proteínas y la edad. Esta etapa preparatoria permitirá enfocar de mejor manera los análisis inferenciales y predictivos en secciones posteriores del trabajo.

Es fácil ver que los histogramas emulan de manera cercana lo que es una distribución normal, por lo que más adelante proponer que cada una se distribuye normal es bastante razonable.

También podemos visualizar la distribución de las edades:

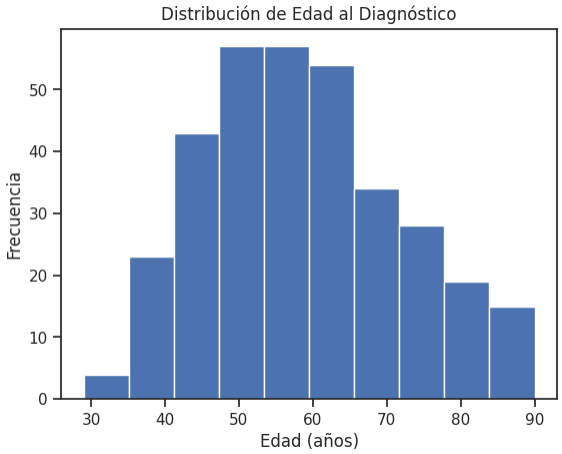


Figura 3: Histograma de las edades

Previamente se tiene el siguiente resumen:

Cuadro 1: Figura 4: Resumen de cada variable

	Age	Protein1	Protein2	Protein3	Protein4	Her_Status_Bin	Patient_Status_Bin
count	334.000000	334.000000	334.000000	334.000000	334.000000	334.000000	334.000000
mean	58.886228	-0.029991	0.946896	-0.090204	0.009819	0.086826	0.763473
std	12.961212	0.563588	0.911637	0.585175	0.629055	0.282003	0.425587
min	29.000000	-2.340900	-0.978730	-1.627400	-2.025500	0.000000	0.000000
25 %	49.000000	-0.358888	0.362173	-0.513748	-0.377090	0.000000	1.000000
50 %	58.000000	0.006129	0.992805	-0.173180	0.041768	0.000000	1.000000
75 %	68.000000	0.343598	1.627900	0.278353	0.425630	0.000000	1.000000
max	90.000000	1.593600	3.402200	2.193400	1.629900	1.000000	1.000000

En esta etapa del análisis se incorporaron las variables **Her Status Bin** y **Patient Status Bin**, ambas originalmente de naturaleza categórica. Con el fin de facilitar su manipulación en procedimientos estadísticos y modelos cuantitativos, se recodificaron como variables binarias.

En el caso de **Her Status Bin**, se asignó el valor de 1 para indicar positividad del marcador HER2 y 0 para casos negativos. Por su parte, la variable **Patient Status Bin** fue codificada con un 1 para representar a pacientes que se encontraban con vida en su última visita registrada, y con un 0 en los casos en que se reporta fallecimiento.

Este proceso binario permite aplicar técnicas estadísticas que requieren datos numéricos, como análisis de correlación, regresiones o clasificación supervisada.

Como mencionamos, algo que podría interesarnos es cómo se relacionan. De manera gráfica podemos verlo con scatters:

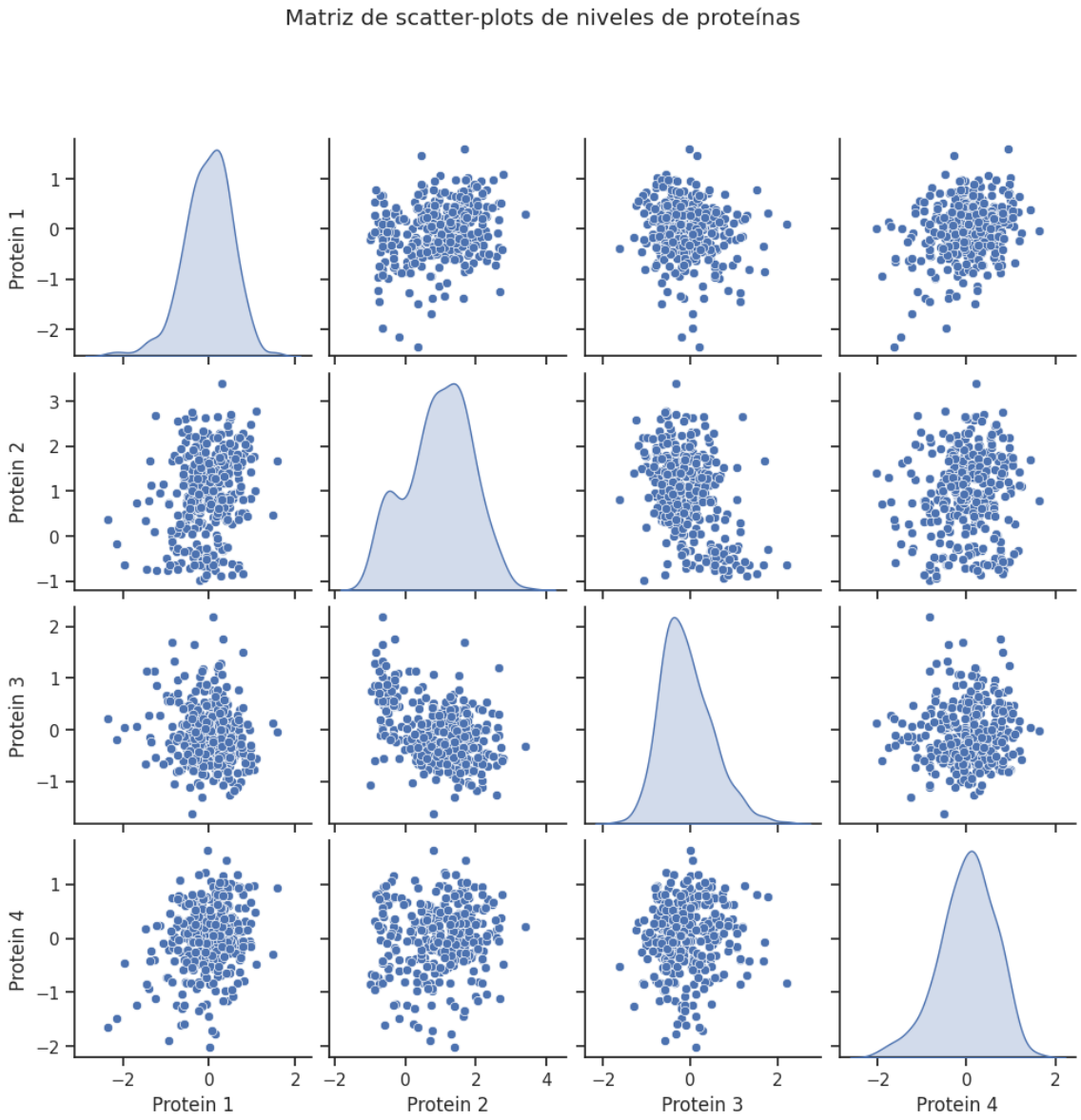


Figura 4: Scatters entre proteinas

Los scatters permiten explorar, visualizar y validar relaciones entre variables. A partir de la matriz de scatter-plots, se observa preliminarmente que no existe una correlación fuerte entre los niveles de proteínas. Si bien esta visualización no permite obtener valores con exactitud, es un indicio razonable de una dependencia débil. Además, la dispersión de los puntos sugiere la ausencia de relaciones lineales claras entre las variables consideradas

Ahora, considerando esto, y con el fin de obtener conclusiones mejor fundamentadas procedimos con el cálculo de los coeficientes de correlación :

Figura 5: Matriz de correlación Spearman

	Protein1	**Protein2**	**Protein3**	**Protein4**
Protein1	1.000000	0.236255	-0.125991	0.227072
Protein2	0.236255	1.000000	-0.352301	0.095852
Protein3	-0.125991	-0.352301	1.000000	0.069621
Protein4	0.227072	0.095852	0.069621	1.000000

Esto confirma que existe cierta dependencia entre las variables, aunque no es especialmente fuerte. Lo más destacado es la correlación entre la proteína 2 y la proteína 3, con un valor de $-0{,}352301$, lo cual sugiere que cuando una de estas proteínas se eleva, la otra tiende a disminuir.

Un estudio interesante sería analizar cómo se relacionan los niveles de expresión proteica con los distintos tipos de tumor, con el fin de identificar si existe una relación entre ellos, y en particular, determinar si algún grupo tiende a concentrar determinados niveles de expresión proteica:

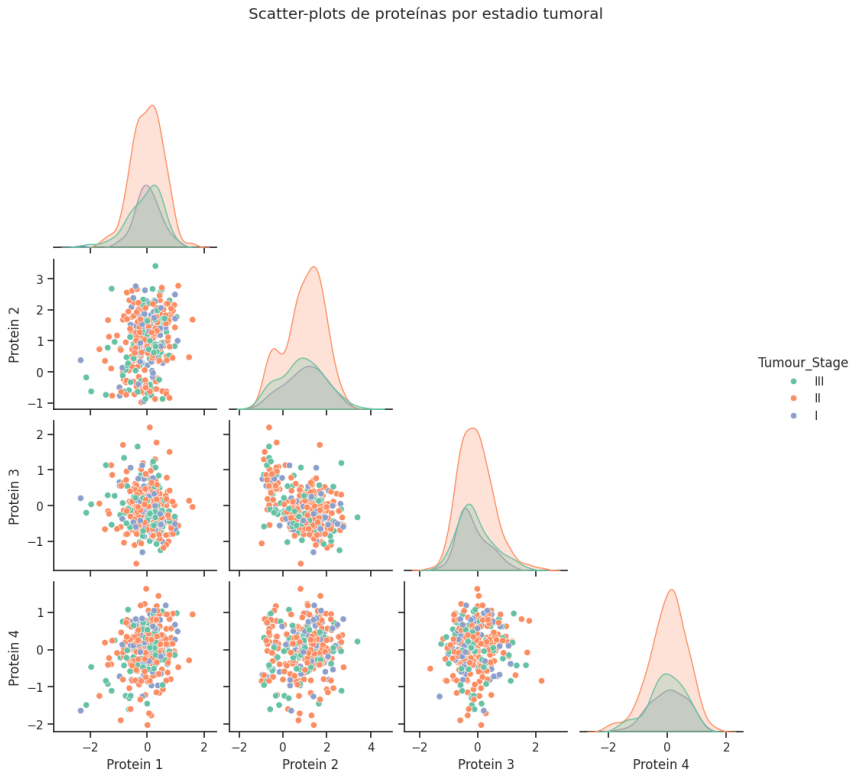


Figura 6: Agrupacion por tipos de tumor

En general, no se observa una relación evidente entre los niveles de expresión proteica y el tipo de tumor, ya que las distribuciones parecen similares entre los distintos estadios. No obstante, se aprecia ligeramente una mayor concentración de casos con tumores de tipo II asociados a niveles más altos de proteína 2.

3. Hipótesis y fundamentos teóricos

3.1. Pruebas de bondad y Ajuste

Cuando recibimos datos y decidimos hacer inferencias sobre una población, lo primero que debemos hacer es tratar de verificar si nuestros datos se ajustan a un modelo paramétrico conocido (por ejemplo, Normal, Poisson, Gamma, etc.).

Al procedimiento estadístico que utilizamos para ajustar un modelo paramétrico se le conoce como *Bondad de Ajuste*. En dicho procedimiento se pretende contrastar las siguientes hipótesis:

$$\begin{aligned} H_0: &\quad F_X(x) = F_X^*(x), \\ H_1: &\quad F_X(x) \neq F_X^*(x), \end{aligned}$$

donde F_X^* es una distribución que puede o no estar completamente especificada, y F_X es la distribución de donde provienen los datos.

Cuando hicimos el análisis exploratorio mencionamos que la gráfica nos indicaba que las proteínas y la edad se asemejaban a una distribución normal, por lo que podemos aplicar algunas pruebas de bondad y ajuste para determinar si es verdad o no.

Existen varias pruebas en la literatura para hacer bondad de ajuste; las que veremos son:

- Prueba de Kolmogorov–Smirnov.
- Pruebas basadas en la QEDF: Cramér–von Mises y Anderson–Darling.

Para cada una de estas pruebas se utiliza un estadístico:

- Kolmogorov–Smirnov: $D = \sup_x \big|F_X(x) - F_0(x)\big|$.
- Cramér–von Mises: $\omega^2 = \displaystyle\int_{-\infty}^{\infty} \big[F_X(x) - F_0(x)\big]^2 \, \mathrm{d}F_0(x)$.
- Anderson–Darling: $A^2 = n \displaystyle\int_{-\infty}^{\infty} \frac{\big[F_X(x) - F_0(x)\big]^2}{F_0(x)\big[1 - F_0(x)\big]} \, \mathrm{d}F_0(x)$.

Donde F_n es la función de distribución empirica, la cual se define dada $X_1, \ldots, X_n$ una muestra aleatoria de cierta distribución $F_X(x)$. Se define la función de distribución empírica como:

$$F_n(x) = \frac{\text{Número de observaciones } \leqslant x}{n} = \frac{\sum_{i=1}^{n} 1(X_i \leqslant x)}{n}$$

Dado que suponemos que la muestra aún no ha sido observada, entonces X_i es una variable aleatoria y, por tanto, $F_n(x)$ también es una variable aleatoria.

Rechazamos H_0 a nivel de significación α si el estadístico observado excede el valor crítico o si su p-valor es menor que α.

3.2. Pruebas de comparación entre dos distribuciones

3.2.1. Descripción de la prueba de Mann–Whitney

Supongamos que tenemos una muestra de n_x observaciones $\{x_1, x_2, \ldots, x_{n_x}\}$ pertenecientes a un grupo (es decir, provenientes de una población), y una muestra de n_y observaciones $\{y_1, y_2, \ldots, y_{n_y}\}$ pertenecientes a otro grupo (provenientes de otra población).

La prueba de Mann–Whitney se basa en la comparación de cada observación x_i de la primera muestra con cada observación y_j de la segunda muestra. El número total de comparaciones por pares que se pueden realizar es $n_x n_y$.

Si las muestras tienen la misma mediana, entonces cada x_i tiene la misma probabilidad (es decir, $\frac{1}{2}$) de ser mayor o menor que cada y_j.

Bajo la hipótesis nula:

$$H_0 : \mathbb{P}(x_i > y_j) = \frac{1}{2}$$

Y bajo la hipótesis alternativa:

$$H_1 : \mathbb{P}(x_i > y_j) \neq \frac{1}{2}$$

Contamos el número de veces que un x_i de la muestra 1 es mayor que un y_j de la muestra 2. Este número se denota como U_x. De manera similar, el número de veces que un x_i es menor que un y_j se denota como U_y. Bajo la hipótesis nula, se esperaría que U_x y U_y sean aproximadamente iguales.

Procedimiento para realizar la prueba:

1. Ordenar todas las observaciones en orden creciente.
2. Bajo cada observación, anotar X o Y (u otro símbolo relevante) para indicar a qué muestra pertenece.
3. Bajo cada x, escribir el número de y's que están a su izquierda (es decir, que son menores que él); esto indica que $x_i > y_j$. Bajo cada y, escribir el número de x's que están a su izquierda; esto indica que $y_j > x_i$.
4. Sumar el total de veces que $x_i > y_j$: denotado por U_x. Sumar el total de veces que $y_j > x_i$: denotado por U_y. Verificar que $U_x + U_y = n_x n_y$.
5. Calcular $U = \text{mín}(U_x, U_y)$.
6. Utilizar las tablas estadísticas de la prueba de Mann–Whitney U para encontrar la probabilidad de observar un valor de U o menor. Si la prueba es unilateral, este es el valor-p; si es una prueba bilateral, se debe duplicar esta probabilidad para obtener el valor-p.

3.2.2. Prueba de Wilcoxon de los rangos con signo

La prueba de Wilcoxon de los rangos con signo es un ejemplo de test no paramétrico o libre de distribución. Al igual que la prueba de los signos, se utiliza para contrastar la hipótesis nula de que la mediana de una distribución es igual a un valor dado. Puede emplearse:

1. En lugar de un t-test de una muestra.

2. En lugar de un t-test pareado.
3. Para datos categóricos ordenados donde no es adecuado un intervalo numérico, pero sí es posible asignar rangos.

Formulación de hipótesis.

H_0 : la mediana de las diferencias es 0 $\quad vs. \quad$ H_1 : la mediana de las diferencias no es 0.

Procedimiento para datos pareados.

1. Sea $\{(x_i, y_i)\}_{i=1}^{n}$ el conjunto de pares de observaciones. Planteamos
$$H_0 : \text{Mediana}(x_i - y_i) = 0.$$
2. Calculamos las diferencias $d_i = x_i - y_i$, $i = 1, \ldots, n$.
3. Ordenamos $|d_i|$ de menor a mayor y asignamos rangos R_i (ignorando signo).
4. Reasignamos a cada rango R_i el signo de d_i.
5. Definimos
$$W^+ = \sum_{d_i>0} R_i, \qquad W^- = \sum_{d_i<0} R_i.$$
Como verificación debe cumplirse $W^+ + W^- = \frac{n(n+1)}{2}$.
6. Tomamos el estadístico de prueba
$$W = \text{mín}(W^+, W^-).$$
7. Consultamos tablas de *critical values* para el test de Wilcoxon: obtenemos el valor-p exacto (unilateral o bilateral, éste último doblando el valor unilateral).

Procedimiento para una sola muestra.

1. Planteamos $H_0 : \text{Mediana}(X) = M$, para un valor hipotetizado M.
2. Calculamos $d_i = x_i - M$, $i = 1, \ldots, n$.
3. Seguimos los pasos 3–7 indicados para datos pareados.

Bajo H_0 se espera que la distribución de las diferencias sea aproximadamente simétrica en torno a cero y que los signos positivos y negativos se distribuyan al azar entre los rangos. De esta manera, el estadístico W tiene una distribución conocida, lo que permite *calcular exactamente* la probabilidad de obtener un valor tan extremo como el observado.

Un valor-p inferior al nivel de significación α conduce a rechazar H_0, concluyendo que existe evidencia estadística de que la mediana de las diferencias es distinta de cero.

3.3. Pruebas de independencia para vectores aleatorios bivariados

3.3.1. Prueba de Spearman

La correlación de Spearman mide la fuerza de una relación monótona (creciente o decreciente) entre dos variables continuas o ordinales.

$$\rho_s \;=\; 1 \;-\; \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

donde $d_i = R(x_i) - R(y_i)$ es la diferencia de rangos de las observaciones x_i e y_i.

- **Hipótesis nula** (H_0): $\rho_s = 0$ (no hay dependencia monótona).
- **Hipótesis alternativa** (H_1): $\rho_s \neq 0$.
- **Cálculo**: Se ordenan los datos, se asignan rangos, se computan las diferencias d_i y se evalúa la fórmula anterior.
- **Valor-p**: Se obtiene a partir de la distribución de ρ_s bajo H_0 (exacta para muestras pequeñas o aproximación normal para $n \gtrsim 10$).

$\rho_s \approx +1$ indica correlación monótona positiva; $\rho_s \approx -1$ correlación monótona negativa; $\rho_s \approx 0$ sugiere ausencia de dependencia monótona.

3.3.2. Correlación de Kendall (τ)

La correlación de Kendall cuantifica la concordancia entre pares de observaciones (x_i, y_i).

$$\tau \;=\; \frac{2}{n(n-1)} \sum_{i<j} \operatorname{sign}(x_i - x_j)\operatorname{sign}(y_i - y_j),$$

donde $\operatorname{sign}(x_i - x_j)$ es $+1$ si $x_i > x_j$, 1 si $x_i < x_j$ y 0 si $x_i = x_j$.

- **Hipótesis nula** (H_0): $\tau = 0$ (independencia ordinal).
- **Hipótesis alternativa** (H_1): $\tau \neq 0$.
- **Cálculo**: Se cuentan pares concordantes (C) y discordantes (D), luego $\tau = \frac{C-D}{\frac{1}{2}n(n-1)}$.
- **Valor-p**: Se basa en la distribución muestral de τ bajo H_0.

Interpretación $\tau \approx 1$ señala que casi todos los pares están concordantes (asociación positiva fuerte), $\tau \approx -1$ indica asociación negativa fuerte, $\tau \approx 0$ sugiere independencia en el ordenamiento.

3.3.3. Prueba de Independencia de Hoeffding

La prueba de independencia de Hoeffding es una prueba no paramétrica diseñada para detectar cualquier tipo de dependencia entre dos variables aleatorias continuas, incluyendo relaciones no lineales. Se basa en comparar la distribución conjunta empírica con el producto de las distribuciones marginales.

Hipótesis

$$H_0: \quad F_{XY}(x,y) = F_X(x)\,F_Y(y),$$
$$H_1: \quad F_{XY}(x,y) \neq F_X(x)\,F_Y(y).$$

Estadístico de prueba

$$D \;=\; \iint \big[F_{XY}(x,y) - F_X(x)\,F_Y(y)\big]^2 \, dF_{XY}(x,y).$$

En la práctica se emplea su estimación empírica a partir de los datos muestrales.

Si D es significativamente mayor que cero (valor-$p < \alpha$), se rechaza H_0 y se concluye que existe dependencia entre X e Y. Esta prueba es sensible a dependencias no lineales y no monótonas.

3.3.4. Prueba de Independencia de Genest–Rémillard

La prueba de Genest–Rémillard es un test no paramétrico basado en el proceso empírico de cópulas, adecuado para detectar dependencias complejas entre variables continuas.

Hipótesis

$$H_0: \quad C(u,v) = u\,v,$$
$$H_1: \quad C(u,v) \neq u\,v,$$

donde $C(u,v)$ es la cópula de las variables transformadas a uniformes $u = F_X(x)$, $v = F_Y(y)$.

Estadístico de prueba

$$S_n \;=\; n \iint\limits_{[0,1]^2} \big[C_n(u,v) - u\,v\big]^2 \, du\,dv,$$

donde $C_n(u,v)$ es la cópula empírica construida a partir de los datos.

Valores grandes de S_n (con valor-$p < \alpha$) indican rechazo de H_0, es decir, evidencia de dependencia entre las variables. Este test es consistente frente a alternativas generales y se implementa en R con las funciones `indepTestSim()` e `indepTest()` del paquete `copula`.

4. Desarrollo estadístico (pruebas aplicadas)

4.1. Pruebas de Bondad y ajuste

4.1.1. Kolmogorov Smirnov

Para cada proteína y la edad, planteamos las hipótesis:

$$H_0 : F_X(x) = F_0(x), \quad \forall x,$$
$$H_1 : F_X(x) \neq F_0(x), \quad \text{para algún } x,$$

donde F_X es la función de distribución empírica de la muestra y F_0 es la función de distribución teórica de una normal con esperanza $\overline{X}$ media muestral y la varianza siendo la varianza muestral.

Cuadro 2: Resultados de Kolmogorov–Smirnov para las variables

Variable	Estadístico	p-valor	Interpretación
Age	0.0661	0.1035	No se rechaza normalidad ($p >$ 0,05).
Protein1	0.0407	0.6211	No se rechaza normalidad.
Protein2	0.0607	0.1641	No se rechaza normalidad.
Protein3	0.0719	0.0600	No se rechaza normalidad.
Protein4	0.0498	0.3671	No se rechaza normalidad.

Análisis KS La mayoría de las variables muestran p-valores superiores a 0.05, lo que sugiere que no hay evidencia suficiente para rechazar la hipótesis de normalidad según Kolmogorov–Smirnov. Solo *Protein3* está muy cerca del umbral ($p = 0{,}06$).

4.1.2. Prueba de Cramér–von Mises

Cuadro 3: Resultados de Cramér–von Mises para las variables

Variable	Estadístico	p-valor	Interpretación
Age	0.2669	0.1681	No se rechaza normalidad ($p >$ 0,05).
Protein1	0.1672	0.3412	No se rechaza normalidad.
Protein2	0.2639	0.1715	No se rechaza normalidad.
Protein3	0.4663	0.0485	Se rechaza normalidad ($p <$ 0,05).
Protein4	0.1407	0.4192	No se rechaza normalidad.

Análisis CvM Cramér–von Mises confirma la normalidad para la mayoría de las variables, excepto *Protein3* (estadístico=0.4663, $p = 0{,}0485$), donde $p < 0{,}05$ indica rechazo de la normalidad. Esto sugiere que *Protein3* podría requerir transformaciones o métodos no paramétricos adicionales.

4.1.3. Anderson-Darling

- **Variable: Edad**
  Estadístico A–D = 1.7506, p-valor aproximado $< 0{,}15$
  No se rechaza la hipótesis nula. La variable `Age` lo que indica puede seguir una distribución normal.

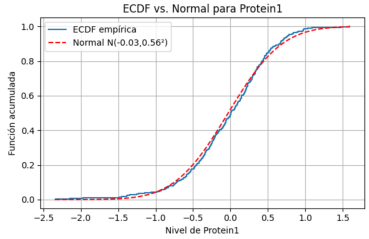
- **Variable: Protein1**
  Estadístico A–D = 1.2190, p-valor aproximado $< 0{,}15$
  Interpretación: No se rechaza la hipótesis nula. La variable `Protein1` **puede seguir una distribución normal**.

- **Variable: Protein2**
  Estadístico A–D = 1.9513, p-valor aproximado $< 0{,}15$
  Interpretación: No se rechaza la hipótesis nula. La variable `Protein2` **puede seguir una distribución normal**.

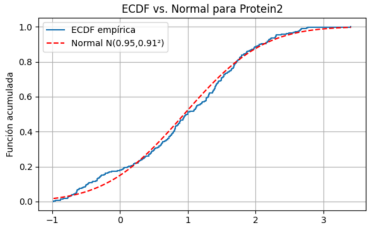
- **Variable: Protein3**
  Estadístico A–D = 2.7493, p-valor aproximado $< 0{,}15$
  Interpretación: No se rechaza la hipótesis nula. La variable `Protein3` **puede seguir una distribución normal**.

- **Variable: Protein4**
  Estadístico A–D = 1.0478, p-valor aproximado 0,01
  Interpretación: Se rechaza la hipótesis nula. La variable `Protein4` **no sigue una distribución normal**.

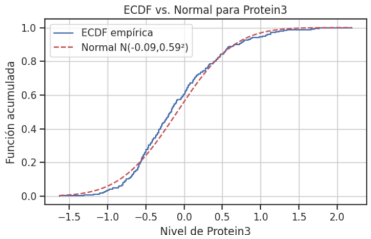
Después de realizar las pruebas vale la pena visualizar cada función de distribución empírica, y compararla con la distribución normal que se propone en cada caso.



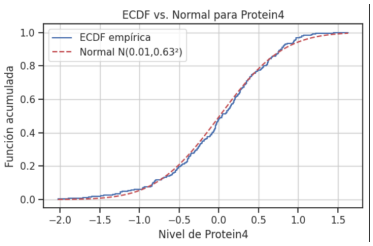
(a) Proteína 1



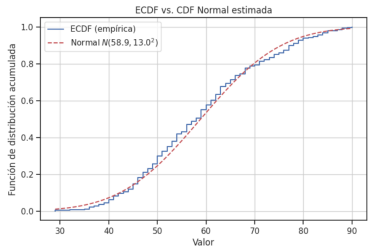
(b) Proteína 2



(c) Proteína 3



(d) Proteína 4



(e) Edad

Figura 7: Funciones de distribución empírica (EDF) frente a la distribución normal para cada variable.

4.2. Implementación de Pruebas para comparación de dos distribuciones

4.2.1. Wilcoxon

Cuadro 4: Resultados de la prueba de los rangos con signo de Wilcoxon entre pares de proteínas

Comparación	Estadístico	Valor p	Interpretación
Protein1 vs Protein2	4,250.0000	0.0000	**Rechazamos** H_0: diferencia significativa entre las muestras empareadas.
Protein1 vs Protein3	24,453.0000	0.0463	**Rechazamos** H_0: diferencia significativa entre las muestras empareadas.
Protein1 vs Protein4	25,677.0000	0.1937	**No rechazamos** H_0: no hay diferencia significativa entre las muestras empareadas.
Protein2 vs Protein3	7,581.0000	0.0000	**Rechazamos** H_0: diferencia significativa entre las muestras empareadas.
Protein2 vs Protein4	6,216.0000	0.0000	**Rechazamos** H_0: diferencia significativa entre las muestras empareadas.
Protein3 vs Protein4	22,923.0000	0.0042	**Rechazamos** H_0: diferencia significativa entre las muestras empareadas.

En la mayoria de los casos hay evidencia estadística suficiente para afirmar que las muestras tienen diferencias sistemáticas en sus valores.

Por ejemplo, Observamos que entre la proteína 2 y la proteína 3 existe una diferencia significativa en sus niveles de expresión, lo cual podría reflejar diferencias en su función biológica.

4.2.2. MannWhitney-Wilcoxon

Cuadro 5: Resultados de la prueba de Mann–Whitney Wilcoxon entre pares de proteínas

Comparación	Estadístico U	Valor p	Interpretación
Protein1 vs Protein2	21,469.0000	0.0000	**Rechazamos** H_0: diferencia significativa entre las distribuciones.
Protein1 vs Protein3	62,641.0000	0.0059	**Rechazamos** H_0: diferencia significativa entre las distribuciones.
Protein1 vs Protein4	53,211.0000	0.3034	**No rechazamos** H_0: no hay diferencia significativa entre las distribuciones.
Protein2 vs Protein3	91,072.5000	0.0000	**Rechazamos** H_0: diferencia significativa entre las distribuciones.
Protein2 vs Protein4	88,449.0000	0.0000	**Rechazamos** H_0: diferencia significativa entre las distribuciones.
Protein3 vs Protein4	47,424.5000	0.0008	**Rechazamos** H_0: diferencia significativa entre las distribuciones.

4.3. Pruebas de independencia para vectores aleatorios bivariados

Uno de los objetivos fundamentales en el análisis multivariado es determinar si existe alguna forma de dependencia entre pares de variables. En nuestro caso, nos enfocamos en los niveles de expresión de las proteínas (`Protein1`, `Protein2`, `Protein3` y `Protein4`), que constituyen variables continuas y que podrían estar relacionadas entre sí debido a su origen biológico común.

Para estudiar la independencia entre pares de proteínas se aplicaron tres pruebas no paramétricas ampliamente utilizadas en estadística: **Kendall**, **Spearman** y **Hoeffding**. Estas pruebas no requieren suposiciones de normalidad y se basan en rangos u ordenamientos, lo cual las hace ideales en contextos exploratorios como este.

4.3.1. Prueba de Kendall

La prueba de Kendall (τ) evalúa la concordancia entre pares de observaciones. Si dos variables están asociadas de forma monotónica, el valor de τ será significativamente diferente de cero. Al aplicarla a los pares de proteínas, obtuvimos los siguientes resultados:

Cuadro 6: Correlación de Kendall entre proteínas

Par	τ	p-valor	Significancia
Protein1 vs Protein2	0.1626	0.0000	**
Protein1 vs Protein3	−0,0826	0.0242	*
Protein1 vs Protein4	0.1510	0.0000	**
Protein2 vs Protein3	−0,2428	0.0000	**
Protein2 vs Protein4	0.0659	0.0725	n.s.
Protein3 vs Protein4	0.0449	0.2206	n.s.

** $p < 0{,}01$, * $p < 0{,}05$, n.s. no significativo.

Estos resultados indican una dependencia positiva moderada entre estos pares, sugiriendo que a mayor expresión de una proteína, tiende a haber mayor expresión en la otra.

4.3.2. Prueba de Spearman

La prueba de Spearman (ρ) mide la correlación de rangos. Sus resultados fueron consistentes con los obtenidos por Kendall. El par con mayor correlación inversa fue:

Cuadro 7: Correlación de Spearman entre proteínas

Par	ρ	p-valor	Significancia
Protein1 vs Protein2	0.2363	0.0000	**
Protein1 vs Protein3	−0,1260	0.0213	*
Protein1 vs Protein4	0.2271	0.0000	**
Protein2 vs Protein3	−0,3523	0.0000	**
Protein2 vs Protein4	0.0959	0.0803	n.s.
Protein3 vs Protein4	0.0696	0.2044	n.s.

** $p < 0{,}01$, * $p < 0{,}05$, n.s. no significativo.

Este resultado confirma una dependencia inversa moderada entre estas variables.

4.3.3. Prueba de Hoeffding

La prueba de independencia de Hoeffding es más general y puede detectar relaciones no lineales no necesariamente monotónicas. Al aplicarla, los valores D fueron en general pequeños (menores a 0.1), lo cual sugiere que las relaciones entre las proteínas pueden explicarse mediante dependencias de tipo ordenado y que no existen relaciones complejas adicionales entre ellas.

Algo a mencionar que para los coeficientes de Hoeffding y Genest–Rémillard , utilizamos R, ya que en Python no hay las mismas opciones de librerías para realizar estas pruebas, por lo que se optó mejor por este otro lenguaje de programación para esta parte de las pruebas aplicadas

Para evaluar la dependencia no lineal entre cada par de proteínas (*Protein1–Protein4*), aplicamos la prueba de Hoeffding's D con la función `hoeffd()` del paquete `Hmisc`. Dado que algunos pares tenían pocos valores únicos, en esos casos se utilizó como respaldo la prueba de Kendall (τ) mediante `cor.test(method="kendall")`. Los resultados se muestran a continuación:

Cuadro 8: Resultados pairwise de Hoeffding's D y prueba de Kendall de respaldo

Par	D	p_D	Método	Estadístico	p-valor	p–BH
Protein1 vs Protein2	0.01670	0.00001	Kendall	4.4340	0.0000	0.00002
Protein1 vs Protein3	0.00296	0.04356	Kendall	−2,2538	0.0242	0.05227
Protein1 vs Protein4	0.01208	0.00022	Kendall	4.1187	0.0000	0.00044
Protein2 vs Protein3	0.03979	0.00000	Kendall	−6,6226	0.0000	0.00000
Protein2 vs Protein4	0.00354	0.03010	Kendall	1.7958	0.0725	0.04516
Protein3 vs Protein4	0.00061	0.22488	Kendall	1.2250	0.2206	0.22488

D = estadístico de Hoeffding; p_D = valor-p original; τ (Kendall) = estadístico de respaldo; p–BH = valor-p de Hoeffding ajustado por Benjamini–Hochberg.

En todos los pares con suficiente variabilidad ($n_{\text{úni}} \geqslant 15$), el estadístico de Hoeffding mostró $p < ,05$, lo que sugiere dependencia no lineal significativa; para los pares con p_D = NA o pocos valores únicos, la prueba de Kendall también indicó dependencia significativa excepto en *Protein3 vs Protein4* ($p = ,2206$).

Por último, la prueba de Friedman para las cuatro proteínas en datos pareados (mismo paciente) arrojó $\chi^2(3) = 249{,}59$, $p < ,0001$, y el post-hoc de Nemenyi reveló diferencias significativas entre casi todos los pares de proteínas, lo que corrobora la heterogeneidad de sus niveles de expresión.

4.3.4. Prueba de Genest–Rémillard

Cuadro 9: Prueba de independencia multivariada (Genest–Rémillard) para Protein1–Protein4

Subconjunto	**Statistic**	**p-value**	**CRVMS$_{0,95}$**	**Rechazo**
Global (todas)	0.06983	.00050	—	Sí
{1,2}	0.19819	.00050	0.08997	Sí
{1,3}	0.06474	.03347	0.08997	Sí
{1,4}	0.16801	.00050	0.08997	Sí
{2,3}	0.43429	.00050	0.08997	Sí
{2,4}	0.06239	.04745	0.08997	Sí
{3,4}	0.03302	.26424	0.08997	No

Nota. Statistic = estadístico de Cramér–von Mises; CRVMS$_{0,95}$ = valor crítico al 95 % bajo independencia; "Rechazo" indica $p < ,05$.

Para evaluar la independencia conjunta de los niveles de *Protein1*, *Protein2*, *Protein3* y *Protein4*, utilizamos el enfoque de Genest–Rémillard, que se basa en comparar la cópula empírica de las variables con la cópula independiente (producto de marginales). Primero, simulamos la distribución nula de la estadística de Cramér–von Mises mediante la función `indepTestSim()` del paquete `copula`, especificando un tamaño de muestra n, dimensión $p = 4$, máximo tamaño de subconjuntos $m = 2$ y $N = 1000$ réplicas bajo independencia. A continuación, la función `indepTest()` calcula el valor observado de la estadística global y el correspondiente valor-p.

El estadístico global resultó ser:

$$D_{\mathrm{CvM}} = 0{,}06983, \quad p = 0{,}00050$$

Lo cual al ser menor que el umbral $\alpha = 0{,}05$ indica un rechazo claro de la hipótesis nula de independencia entre las cuatro proteínas: existen dependencias conjuntas que no se explican por el azar. Para profundizar en qué pares o grupos de dos variables contribuían más a esa dependencia global, consultamos el dependograma, donde cada subconjunto de tamaño dos (por ejemplo $\{1, 2\}$, $\{2, 3\}$, etc.) tiene su propia estadística y valor-p: casi todos mostraron valores-$p < 0{,}05$, salvo el subconjunto $\{3, 4\}$ (Protein3 vs Protein4, $p = 0{,}2642$), donde no se detectó dependencia significativa.

En conjunto, estos resultados nos enseñan que, aunque algunas parejas de proteínas (como Protein2 vs Protein3 o Protein1 vs Protein2) presentan correlaciones moderadas positivas o negativas en análisis univariados, la prueba multivariada pone de manifiesto que la estructura de dependencia abarca casi todas las combinaciones, confirmando que el perfil de expresión conjunto de estas cuatro proteínas no corresponde al de un conjunto de variables independientes.

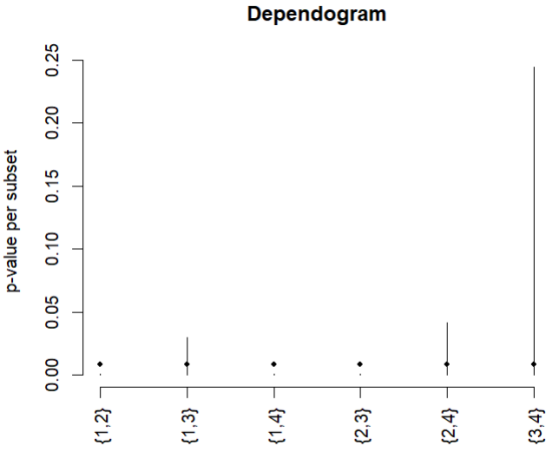


Figura 8: Dependograma

Aunque este estudio se centró en el análisis estadístico de la dependencia entre proteínas, una línea interesante de investigación es explorar si estas proteínas pueden actuar como predictores del tipo de tumor. En nuestro análisis preliminar no se observaron diferencias significativas en los niveles de expresión de proteínas según el tipo histológico del tumor. Esto puede deberse a la naturaleza limitada del dataset o a la necesidad de modelos más complejos para capturar dicha relación.

Conclusiones

El análisis estadístico realizado sobre el conjunto de datos *Real Breast Cancer Data* proporciona una evaluación de las características clínicas y moleculares de pacientes con cáncer de mama, con un enfoque particular en los niveles de expresión de cuatro proteínas (Protein1, Protein2, Protein3, Protein4) y su relación con variables clínicas. A continuación, se presentan las conclusiones más relevantes, organizadas desde los resultados más básicos hasta los más complejos, con interpretaciones derivadas de las pruebas estadísticas y visualizaciones.

1. **Análisis Exploratorio de Datos (EDA)**
   El análisis exploratorio reveló que la edad de las pacientes se concentra entre los 40 y 60 años, alineándose con la literatura médica que identifica esta franja etaria como de mayor riesgo para el cáncer de mama. Los histogramas de las proteínas (Figura 2) muestran distribuciones que visualmente se asemejan a una normal, especialmente para Protein1, Protein2 y Protein4, lo que justificó la aplicación de pruebas de bondad de ajuste. La matriz de dispersión (Figura 4) y la matriz de correlación de Spearman (Figura 5) indicaron correlaciones débiles a moderadas entre las proteínas, destacando una correlación inversa moderada entre Protein2 y Protein3 ($\rho = -0{,}3523$), sugiriendo que el aumento en una tiende a coincidir con una disminución en la otra. Sin embargo, no se observaron patrones claros de asociación entre los niveles de proteínas y el tipo de tumor (Figura 6), lo que sugiere que estas variables podrían no ser determinantes directas del estadio tumoral.

2. **Pruebas de Bondad de Ajuste**
   Las pruebas de Kolmogorov-Smirnov, Cramér-von Mises y Anderson-Darling (Cuadros 2, 3 y resultados en página 16) evaluaron si las distribuciones de la edad y las proteínas se ajustan a una normal. Los resultados de Kolmogorov-Smirnov (p-valores ¿0.05) no rechazaron la normalidad para ninguna variable, mientras que Cramér-von Mises rechazó la normalidad para Protein3 ($p = 0{,}0485$) y Anderson-Darling para Protein4 ($p \approx 0{,}01$). Esto indica que, aunque la mayoría de las variables son compatibles con una distribución normal, Protein3 y Protein4 podrían requerir métodos no paramétricos o transformaciones para análisis posteriores. Las funciones de distribución empírica (Figura 7) refuerzan visualmente estos hallazgos, mostrando un ajuste cercano a la normal para la mayoría de las variables, salvo desviaciones leves en Protein3 y Protein4.

3. **Pruebas de Comparación de Distribuciones**
   Las pruebas de Wilcoxon de rangos con signo y Mann-Whitney-Wilcoxon (Cuadros 4 y 5) compararon las distribuciones de los niveles de proteínas por pares. La mayoría de las comparaciones mostraron diferencias significativas ($p < 0{,}05$), excepto para Protein1 vs. Protein4 ($p > 0{,}05$ en ambas pruebas), lo que sugiere que estas dos proteínas podrían compartir características distribucionales similares. Estos resultados destacan la heterogeneidad en los niveles de expresión de las proteínas, especialmente entre Protein2 y las demás, lo que podría reflejar diferencias biológicas en su rol dentro del cáncer de mama.

4. **Pruebas de Independencia**
   Las pruebas de Kendall, Spearman y Hoeffding (Cuadros 6, 7 y 8) confirmaron dependencias significativas entre varios pares de proteínas, particularmente entre Protein1 vs. Protein2 ($\tau = 0{,}1626$, $\rho = 0{,}2363$, $p < 0{,}05$) y Protein2 vs. Protein3 ($\tau = -0{,}2428$, $\rho = -0{,}3523$, $p < 0{,}05$). La prueba de Hoeffding, más sensible a relaciones no lineales, mostró valores D pequeños, indicando que las dependencias son principalmente monótonas. La prueba de Genest-Rémillard (Cuadro 9) rechazó la independencia conjunta de las cuatro proteínas ($p = 0{,}00050$), salvo para el par Protein3 vs. Protein4 ($p = 0{,}2642$), lo que sugiere una estructura de dependencia compleja entre los niveles de expresión proteica, probablemente influenciada por factores biológicos subyacentes.

5. **Implicaciones y Consideraciones**
   Los resultados sugieren que las proteínas analizadas no son independientes y presentan diferencias significativas en sus distribuciones, lo que podría estar relacionado con procesos biológicos específicos del cáncer de mama. Sin embargo, la falta de asociación clara entre los niveles de proteínas y el tipo de tumor indica que estas variables podrían no ser marcadores directos del estadio tumoral, o que se requieren modelos más complejos para capturar dichas relaciones. El uso combinado de pruebas paramétricas y no paramétricas proporcionó una evaluación robusta, destacando la utilidad de métodos no paramétricos como Hoeffding y Genest-Rémillard para detectar dependencias no lineales.

En conclusión, este estudio demuestra que los niveles de expresión de las proteínas en pacientes con cáncer de mama presentan distribuciones cercanas a la normal, con algunas excepciones, y exhiben dependencias moderadas entre sí, aunque no se asocian claramente con el tipo de tumor. Estos hallazgos sientan una base sólida para investigaciones futuras que podrían explorar modelos predictivos o incorporar variables adicionales para profundizar en las relaciones biológicas y clínicas del cáncer de mama.