LECCIÓN 4: VARIABLES ALEATORIAS Y DISTRIBUCIONES DE PROBABILIDAD

SAAVEDRA, P.

1. Concepto de variable aleatoria y su distribución de probabilidad

En general, cualquier magnitud X cuyo valor no puede predecirse con certeza es por definición una ***variable aleatoria***. En ocasiones, su valor puede predecirse en términos de probabilidad a través de la llamada *función de distribución acumulativa de probabilidad,* o simplemente ***función de distribución***, la cual se define por:

$$F(t) = \Pr(X \leq t)$$

Para cada valor de t, $F(t)$ es la probabilidad de que la variable tome un valor *menor o igual* a t. Nótese que $F(t)$ toma sus valores en el intervalo $[0, 1]$ y que es no decreciente (según aumenta t, aumenta la probabilidad del suceso $\{X \leq t\}$).

Ejemplo 1. Loprinzi *et al* (1994) llevaron a efecto un estudio en pacientes con cáncer avanzado cuyo propósito era establecer un pronóstico de supervivencia a partir de la información suministrada por el propio paciente. Para el subgrupo de los que tenían cáncer de pulmón, consideramos la variable aleatoria: $X =$ *'Meses transcurridos entre el diagnóstico y la muerte del paciente'*. Tal periodo de tiempo recibe el nombre de ***supervivencia global***.

De los datos del estudio de Loprinzi se estimó la *función de distribución de probabilidad* de la variable aleatoria X, la cual tiene la forma:

$$F(t) = 1 - \exp\left(-(t/\lambda)^{\kappa}\right) \quad : \quad t \geq 0$$

siendo $\kappa \approx 1.3168$ y $\lambda \approx 13.742$. La gráfica de la función se muestra en la figura 1.

Nótese que la probabilidad de que un paciente tenga una supervivencia inferior o igual a 24 meses es: $F(24) \approx 0.876$. Ello supone que el porcentaje de pacientes que morirán antes de los dos años de evolución es del 87,6%.

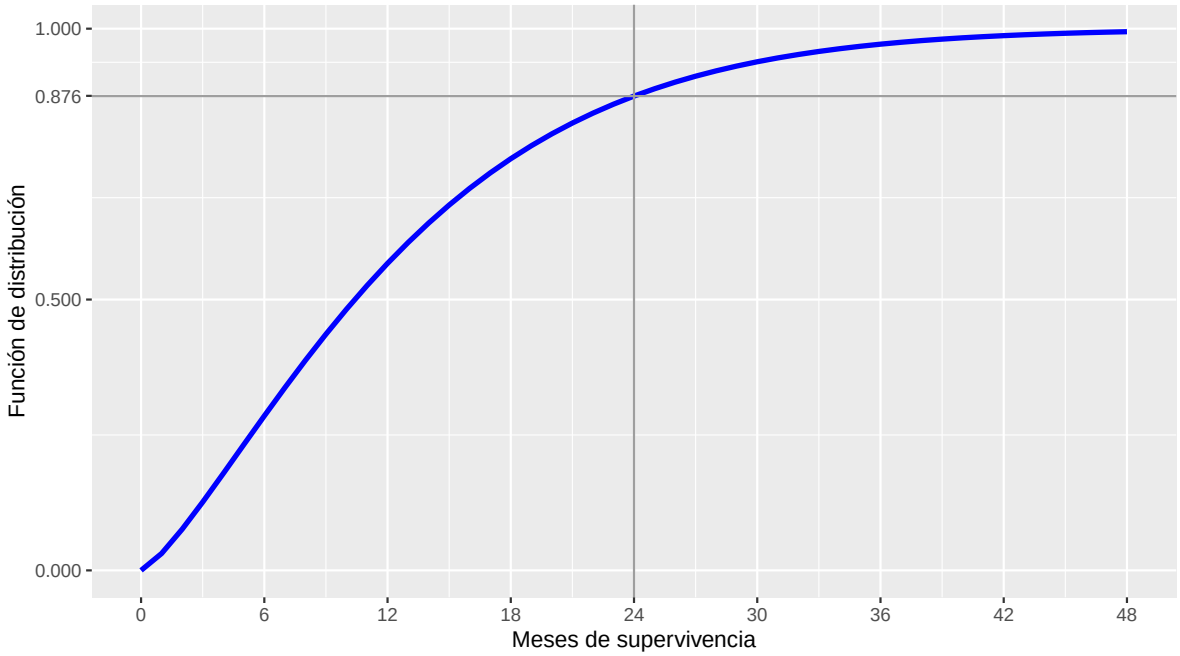

FIGURA 1. Supervivencia global de pacientes con cáncer de pulmón avanzado (Loprinzi, 1994)

De la definición de función de distribución, es fácil deducir la siguiente propiedad:

$$\Pr\left(a < X \leq b\right) = F\left(b\right) - F\left(a\right) \quad : \quad a < b$$

En efecto, si $a < b$, puede escribirse: $\{X \leq b\} = \{X \leq a\} \cup \{a < X \leq b\}$. Dado que los sucesos $\{X \leq a\}$ y $\{a < X \leq b\}$ son incompatibles, en virtud del segundo axioma de la probabilidad se satisface:

$$F\left(b\right) = \Pr\left(X \leq b\right) = \Pr\left(X \leq a\right) + \Pr\left(a < X \leq b\right) = F\left(a\right) + \Pr\left(a < X \leq b\right)$$

Para los pacientes con cáncer de pulmón avanzado descritos en el ejemplo 1, la probabilidad de que un paciente muera durante el segundo año de seguimiento es:

$$\Pr\left(12 < X \leq 24\right) = F\left(24\right) - F\left(12\right) \approx 0,309 \quad (30,9\%)$$

El propósito del estudio de Loprinzi *et al* fue establecer un pronóstico de supervivencia a partir de la información suministrada por el paciente. Para ello se le realizó a cada paciente una encuesta a partir de la que se determinó la llamada *puntuación Karnowsky* (PK), la cual oscila en un rango de 0 a 100, correspondiendo el 0 al peor estado y 100 al mejor. En la figura 2 se muestran las funciones de distribución de la *supervivencia global* en los grupos determinados por tener o no una PK inferior a 80.

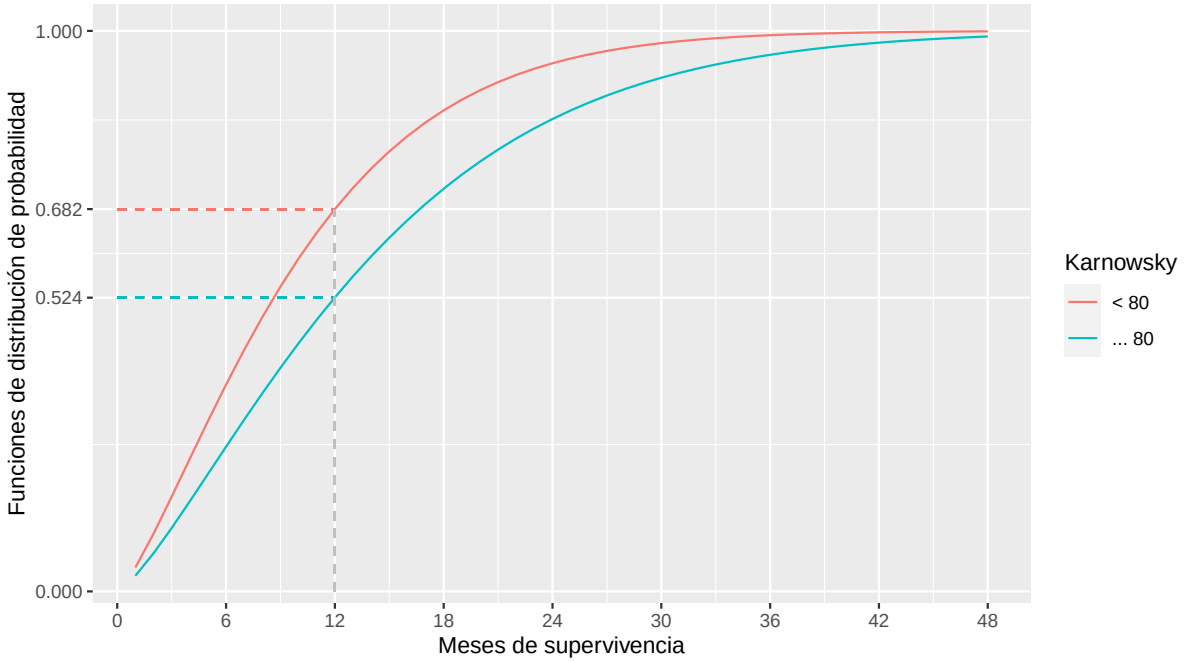

FIGURA 2. Supervivencia global de pacientes con cáncer de pulmón avanzado según puntuación Karnowski (Loprinzi, 1994)

Denotamos por $F_0(t)$ y $F_1(t)$ a las funciones de distribución de probabilidad correspondientes a las cohortes determinadas según la puntuación Karnowsky superase o no el valor de 80. Se obtuvo que $F_0(12) \approx 0,524$ y $F_1(12) \approx 0,682$. Ello significa que en aquellos pacientes con $PK \geq 80$, la probabilidad de morir en el primer año de seguimiento fue del 52,4%, mientras en el grupo $PK < 80$, esa probabilidad se elevó al 68,2%. Nótese que $F_1(12)/F_0(12) \approx 1,30$ es un riesgo relativo que representa cuánto más probable es morir en el primer año de seguimiento en los pacientes con un $PK < 80$ en relación con aquellos con $PK \geq 80$.

2. Clasificación de variables aleatorias

El cálculo con variables aleatorias difiere según sus distribuciones de probabilidad sean ***discretas*** o ***continuas***. Esta clasificación se analiza en la presente sección.

Variables numéricas discretas. Una variable aleatoria X se dice ***discreta*** cuando el conjunto de valores que puede tomar es finito o numerable. En tal caso, su distribución de probabilidad queda plenamente especificada por las probabilidades de la forma $\Pr(X = t)$, para cualquier valor t que pueda tomar la variable. Obviamente se tiene que: $\sum_t \Pr(X = t) = 1$. Esta sumatoria se entiende que se realiza a lo largo de todos los valores que puede tomar X.

Ejemplo 2. Supóngase que la tasa de respuestas favorables de un tratamiento (*probabilidad de respuesta favorable*) es π. Supóngase además que el tratamiento se aplica a tres pacientes y sea la variable aleatoria: $X =$ *Número de pacientes que presentan una respuesta favorable*. El conjunto de valores que puede tomar X es: $\{0, 1, 2, 3\}$. Las correspondientes probabilidades se muestran en la tabla 1.

Tabla 1. Distribución de probabilidad correspondiente al ejemplo 2.

t	0	1	2	3
$\Pr(X = t)$	$(1-\pi)^3$	$3\pi(1-\pi)^2$	$3\pi^2(1-\pi)$	π^3

Nótese que las probabilidades de esta distribución corresponden al desarrollo del binomio $[(1-\pi)+\pi]^3$, y de ahí, $\sum_{t=0}^{3} \Pr(X = t) = 1$.

En la siguiente tabla se resume la función de distribución de probabilidad $F(t)$ de X la cuando $\pi = 0,6$.

t	0	1	2	3
$F(t)$	0,064	0,352	0,784	1

En la figura 3 se muestra la gráfica de la función de distribución de probabilidad $F(t) = \Pr(X \leq t)$.

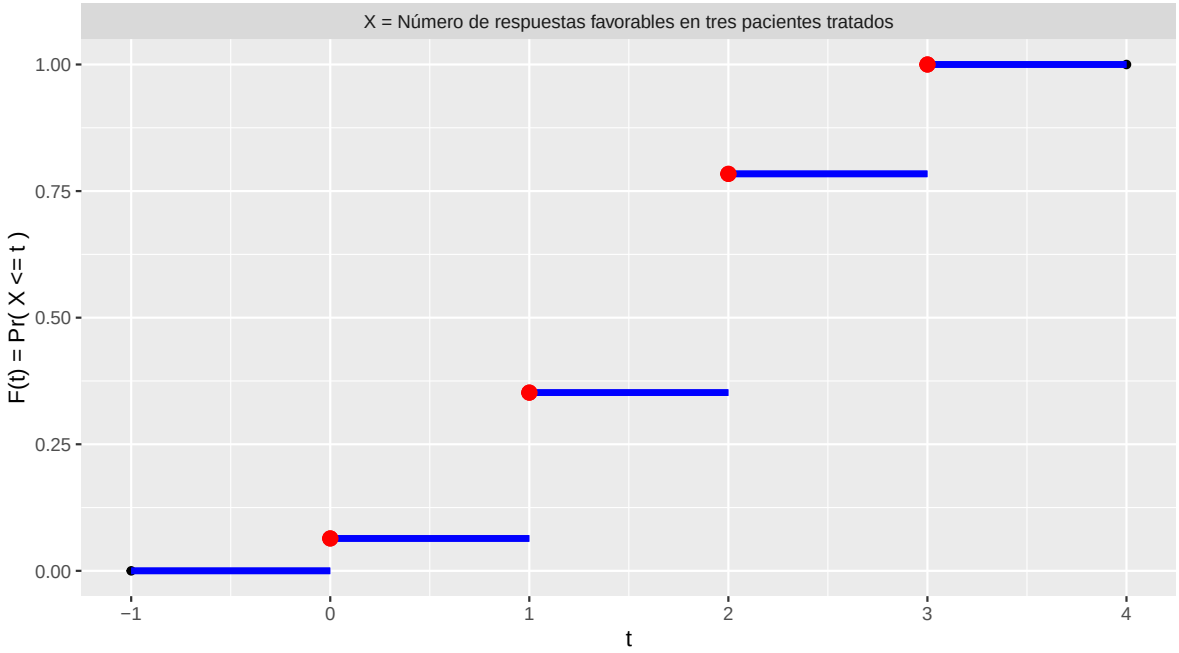

FIGURA 3. Función de distribución de probabilidad para la variable del ejemplo 2

Nótese que esta función de distribución de probabilidad es ***discontinua***. A partir de esta observación definimos el concepto de *variable continua*.

Variables aleatorias continuas. Una variable aleatoria X se dice ***continua*** cuando lo es su función de distribución. En tales casos ocurre que $\Pr(X = t) = 0$, para cualquier valor de t. Esta propiedad puede justificarse mediante la siguiente consideración:

$$\Pr(t - h < X \leq t) = F(t) - F(t - h)$$

Si ahora $h \to 0$, el primer miembro tiende a ser $\Pr(X = t)$ (*este es un hecho que requiere una discusión más profunda*) y el segundo satisface $F(t) - F(t - h) \to 0$ en virtud de la continuidad de la función de distribución $F(t)$.

3. Funciones de densidad de probabilidad

Para algunas variables aleatorias continuas, su distribución de probabilidad puede especificarse a través de la llamada *función de densidad de probabilidad*. Se dice que una función $f(t)$ es una *función de densidad de probabilidad* si satisface las siguientes propiedades:

- $f(t) \geq 0$ para todos los valores de t. Esto signfica que su gráfica está por encima del eje de abscisas.
- El área bajo toda la gráfica de $f(t)$ es igual a 1.

La variable aleatoria descrita en el ejemplo 1 posee función de densidad de probabilidad y su gráfica se muestra en la figura 4:

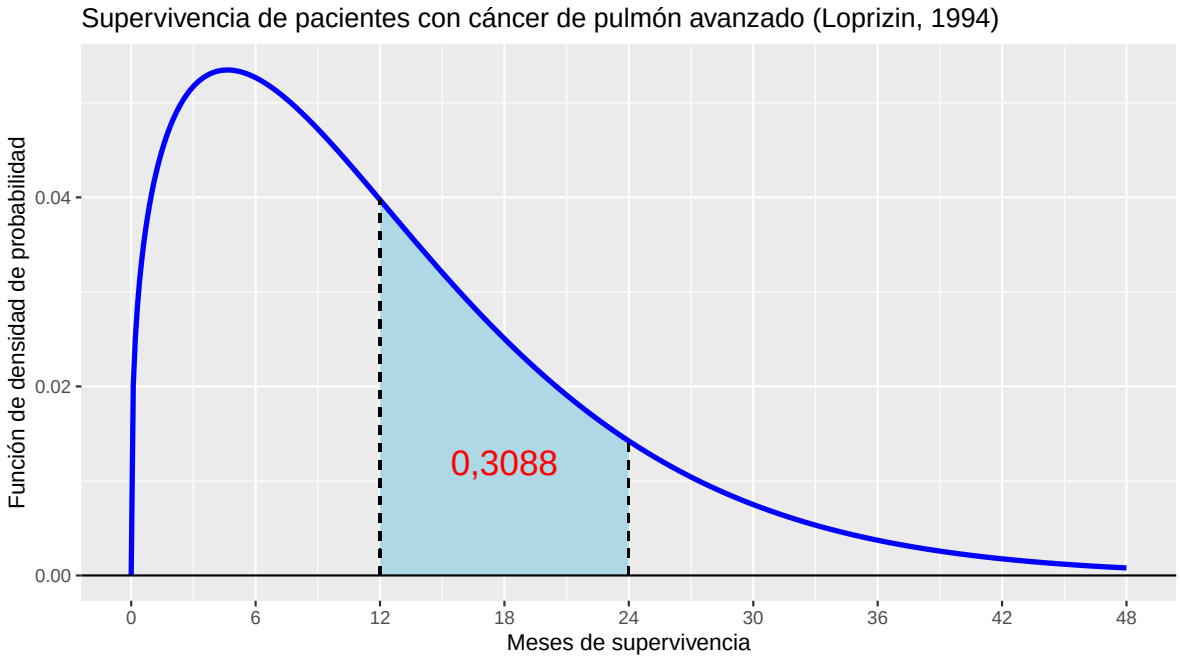

Figura 4. Función de densidad de probabilidad de la variable aleatoria '*Tiempo de supervivencia global de pacientes con cáncer de pulmón avanzado*' (ejemplo 1)

A partir de la función de densidad puede obtenerse la probabilidad de que la variable aleatoria tome un valor en un intervalo $[a, b]$ como el área comprendida bajo la gráfica de la función de densidad y las ordenadas trazadas por los extremos del intervalo.

Para los pacientes con cáncer de pulmón avanzado (ejemplo 1), puede verse que el área comprendida bajo la gráfica de la función de densidad y las ordenadas trazadas por los puntos 12 y 24 es 0,3088. Esto significa que la probabilidad de que el ***tiempo de supervivencia global*** esté comprendido entre los 12 y 24 meses es del 30,88%.

En la figura 5 se muestran las funciones de densidad de probabilidad correspondientes a las cohortes de pacientes con cáncer de pulmón avanzado según la PK reportada por el paciente fuese o no inferior a 80. Nótese que para los pacientes con $PK < 80$ la densidad tiene una mayor concentración en los valores menores, indicando de esa forma una supervivencia global menor.

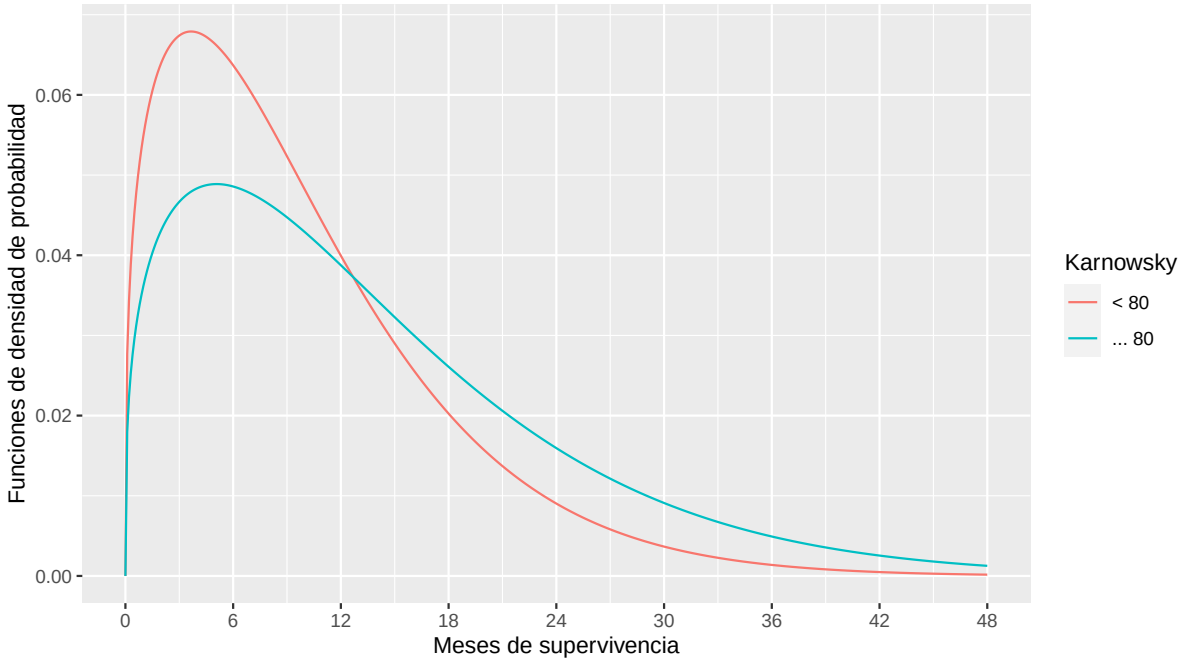

FIGURA 5. Densidades de probabilidad correspondientes a la supervivencia global de pacientes con cáncer de pulmón avanzado según la puntuación Karnowsky supere o no el umbral de 80.

4. INDEPENDENCIA DE VARIABLES ALEATORIAS

La idea de **independencia** de dos variables aleatorias es que **la observación de una de ellas no permite hacer predicciones de la otra**. Formalmente se define de la siguiente forma:

Las variables aleatorias X e Y son ***independientes*** si y sólo si lo son los sucesos $\{X \leq a\}$ e $\{Y \leq b\}$ cualesquiera que sean los valores a y b.

En general, las variables aleatorias $X_1, \ldots, X_p$ son independientes si los sucesos asociados $\{X_1 \leq t_1\}$, ... ,$\{X_p \leq t_p\}$ son independientes, para cualesquiera $t_1, ..., t_p$.

En los estudios biomédicos, el concepto de independencia de variables aleatorias surge frecuentemente cuando se precisa repetir observaciones en las mismas condiciones. En el ejemplo 2 se repitió la intervención terapéutica considerada en tres pacientes. Asociada a la intervención en el i-ésimo paciente, podemos definir la variable aleatoria: $X_i =$ 1 ó 0 *según se produzca o no una respuesta favorable en ese paciente*. Dado que los resultados de las intervenciones no se condicionan entre ellos, las variables aleatorias X_1, X_2, X_3 son independientes y de esta forma puede escribirse:

$$\Pr\left(\{X_1 = 1\} \cap \{X_2 = 0\} \cap \{X_3 = 1\}\right) = \pi \times (1 - \pi) \times \pi$$

5. Características de las distribuciones de probabilidad

Esperanza matemática. La expresión *esperanza de vida* es de uso frecuente en los campos de la medicina, epidemiología o fiabilidad industrial. Si establecemos una analogía entre la *distribución de probabilidad* y la *distribución de masas*, la esperanza resume la distribución de probabilidad como su ***centro de gravedad***. De esta forma, el cálculo de la misma dependerá de que la variable aleatoria sea discreta o continua.

En la figura 6 se representa gráficamente la distribución de probabilidad dada en la tabla 1 para el caso en el que $\pi = 0.6$. Nótese que las longitudes de las barras corresponden a las probabilidades de los puntos. Podemos asumir que tales probabilidades son las ***masas*** de los puntos. En tal caso, puede probarse que el centro de gravedad del conjunto de los valores $\{0, 1, 2, 3\}$ es 1.8. Esta es la idea del concepto de esperanza que se define seguidamente.

Formalmente, la *esperanza matemática* de una variable aleatoria X se define por :

- Si X es discreta: $E[X] = \sum_t t \cdot \Pr(X = t)$
- Si X tiene función de densidad $f(t)$, entonces: $E[X] = \int_{-\infty}^{\infty} t \cdot f(t) \cdot dt$

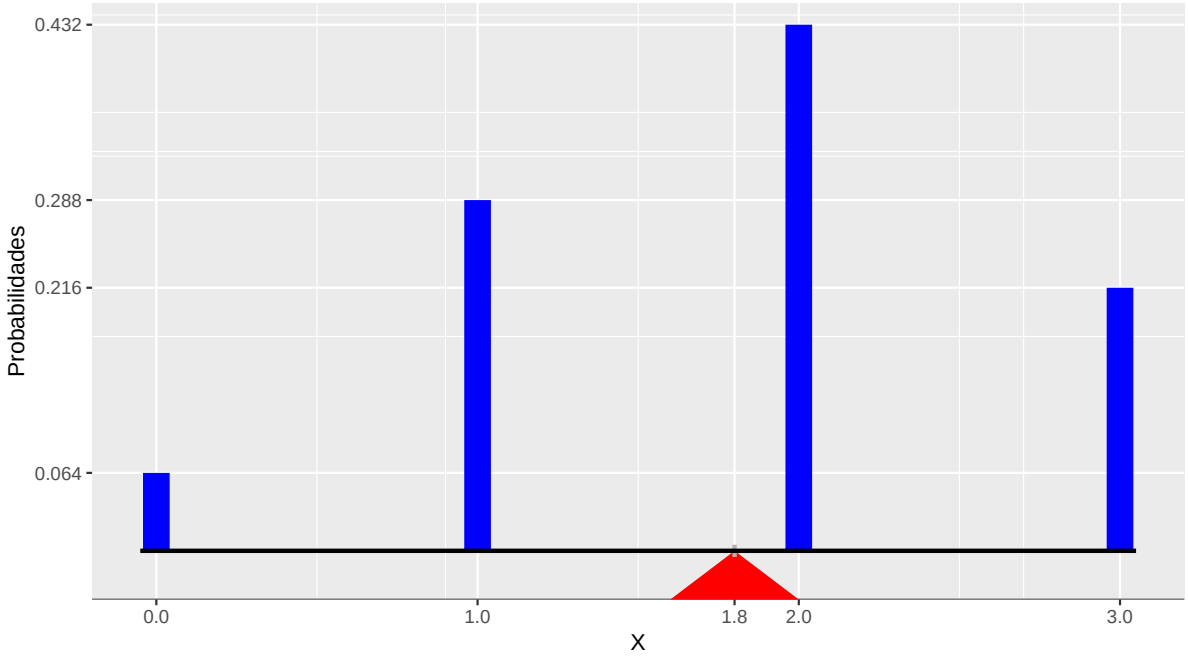

FIGURA 6. Distribución de probabilidad descrita en el ejemplo 2 para $\pi = 0,64$. El punto 1,8 es el *centro de gravedad* del conjunto de masas ó la *esperanza* de la variable aleatoria.

En la figura 7 se muestra la densidad de probabilidad correspondiente a la supervivencia global de los pacientes con cáncer de pulmón avanzado y su esperanza o centro de gravedad.

Para la ley de probabilidad discreta dada en el ejemplo 2, la esperanza se calcula en la forma que sigue:

$$E[X] = 0 \times (1-\pi)^3 + 1 \times 3\pi(1-\pi)^2 + 2 \times 3\pi^2(1-\pi) + 3 \times \pi^3 = 3\pi$$

Tal como de ha visto anteriormente, para aquellas distribuciones con función de densidad de probabilidad, la esperanza se obtiene mediante el cálculo integral. En cualquier caso, no se realizarán ejercicios de esta naturaleza dado que el cálculo integral no se considera como necesario para cursar esta materia.

La esperanza matemática satisface las siguientes propiedades (independientemente del tipo de distribución):

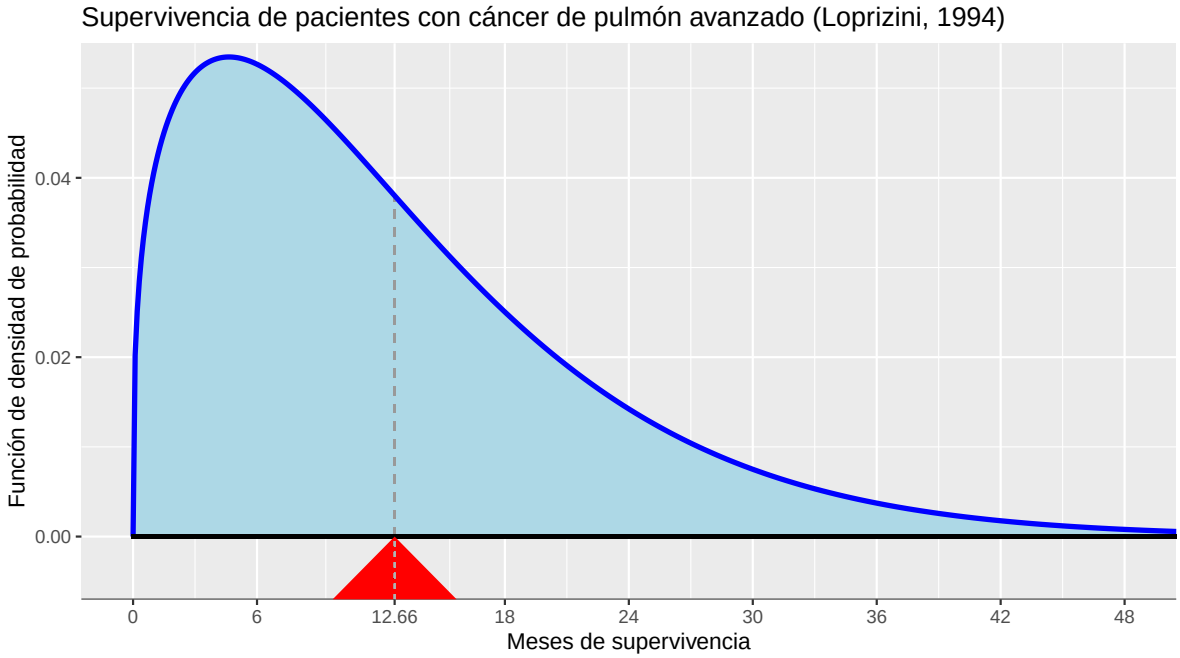

FIGURA 7. Meses esperados de *supervivencia global* para los pacientes con cáncer de pulmón avanzado. El valor 12,66 es el centro de gravedad de todos los posibles tiempos de supervivencia de los pacientes de esta cohorte

- $E[\kappa \cdot X] = \kappa \cdot E[X]$, siendo κ constante y X variable aleatoria.
- $E[X + Y] = E[X] + E[Y]$

Medidas de dispersión de variables aleatorias. En la figura 8 se representan simultáneamente tres funciones de densidad con el mismo centro de gravedad (esperanza), pero con diferente dispersión. En el análisis de datos, el concepto de dispersión juega un papel esencial. En tal sentido, definimos ahora los conceptos de *varianza* y *desviación estándar*.

Para una variable aleatoria X con esperanza μ (es decir, $E[X] = \mu$), la ***varianza*** se define por:

$$\text{var}(X) = E\left[(X - \mu)^2\right]$$

La varianza de una variable aleatoria satisface las siguientes propiedades:

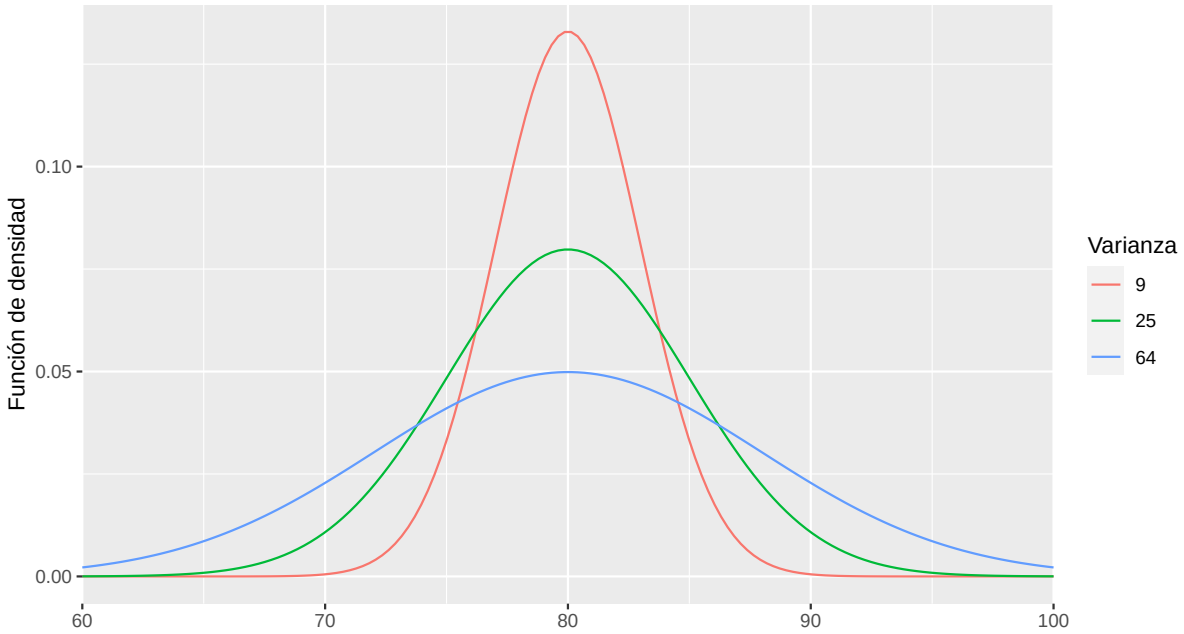

FIGURA 8. Funciones de densidad de probabilidad con esperanza común 80 y varianzas 9, 25 y 64.

- $\text{var}(\kappa X) = \kappa^2 \text{var}(X)$, siendo κ una constante y X una variable aleatoria
- Si X e Y son variables aleatorias independientes, entonces: $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$

La varianza puede obtenerse alternativamente como:

$$\text{var}(X) = E[X^2] - \{E[X]\}^2$$

En efecto:

$$\text{var}(X) = E\left[(X-\mu)^2\right] = E[X^2 - 2\mu X + \mu^2] =$$

$$E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - \mu^2$$

Adicionalmente a las propiedades para la esperanza y varianza, el lector puede comprobar fácilmente que la esperanza de una constante es la propia constante y la varianza, cero.

Para la distribución del ejemplo 2 se tiene:

$$E\left[X^2\right] = 0^2 \cdot (1-\pi)^3 + 1^2 \cdot 3\pi\,(1-\pi)^2 + 2^2 \cdot 3\pi^2\,(1-\pi) + 3^2 \cdot \pi^3 = 3\pi + 6\pi^2$$

De aquí se obtiene:

$$\text{var}\,(X) = E\left[X^2\right] - \{E\,[X]\}^2 = 3\pi + 6\pi^2 - (3\pi)^2 = 3\pi\,(1-\pi)$$

Finalmente, la ***desviación estandar*** (*sd*, por sus siglas en inglés *standard deviation*) se define por: $\text{sd}\,(X) = \sqrt{\text{var}\,(X)}$.

Cuantiles. Para una variable aleatoria X, con función de distribución acumulativa continua y creciente $F\,(t)$, el α-ésimo cuantil se define como el valor q_α, tal que $F\,(q_\alpha) = \Pr\,(X \leq q_\alpha) = \alpha$.

La mediana de la distribución de probabilidad se define como el cuantil 0.5. Los cuartiles 1 y 3 son los cuantiles 0,25 y 0,75 respectivamente.

La figura 9 corresponde a la función de distribución de probabilidad de la supervivencia global de los pacientes con cáncer de pulmón avanzado. Puede observarse que la mediana (cuantil 0.5) es 10,4 meses. Ello significa que la probabilidad de morir antes de ese tiempo es del 50%.

6. Distribuciones de probabilidad especiales

Bernoulli. Considérese un experimento aleatorio en el que un suceso A tiene probabilidad π de ocurrir. Asociado a éste se define una variable aleatoria X con valores 1 ó 0 según ocurra (1) o no (0) el suceso A. Nótese que la distribución de probabilidad se puede expresar en la forma:

$$\Pr\,(X = t) = \pi^t\,(1-\pi)^{1-t} \quad : \; t = 1, 0$$

Esta distribución recibe el nombre de *distribución de Bernoulli* de parámetro π. La esperanza y varianza pueden calcularse fácilmente en la forma que sigue:

$$E\,[X] = 1 \times \pi + 0 \times (1-\pi) = \pi$$

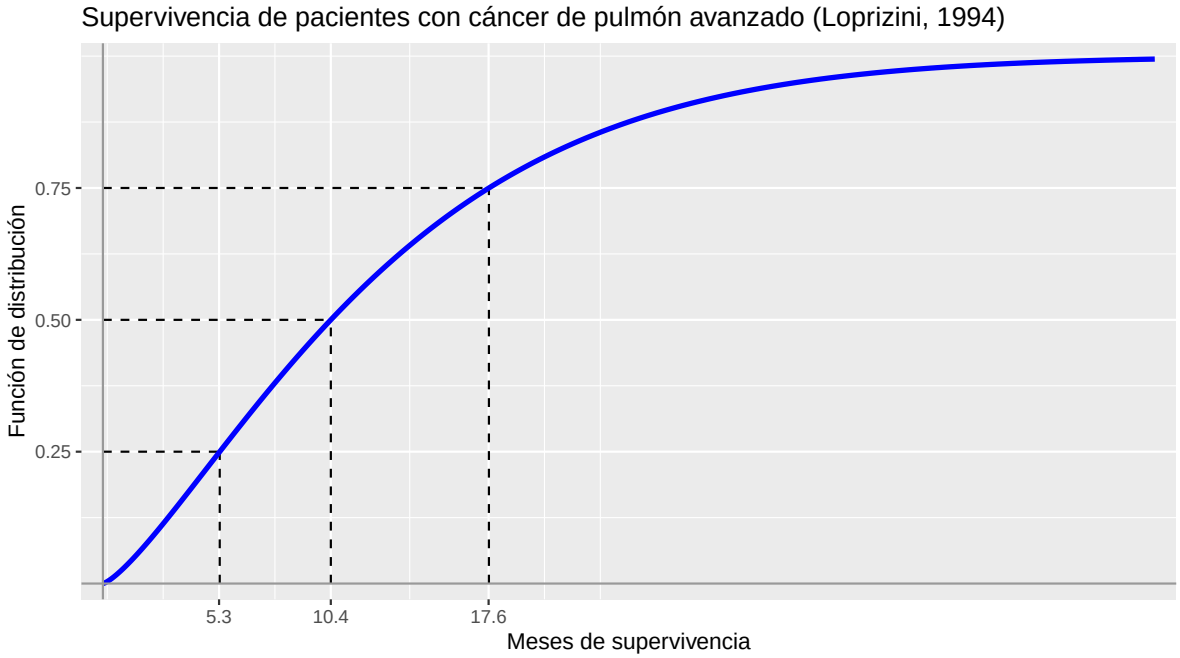

FIGURA 9. Para la variable aleatoria *supervivencia global* de los pacientes con cáncer de pulmón avanzado (ejemplo 1), los cuantiles 0.25, 0.5 y 0.75 son 5.3, 10.4 y 17.6 meses respectivamente.

$$\text{var}(X) = E[X^2] - E[X]^2 = 1^2 \times \pi + 0^2 \times (1-\pi) - \pi^2 = \pi - \pi^2 = \pi(1-\pi)$$

Binomial. Considérese ahora el mismo experimento aleatorio en el que el suceso A tiene probabilidad π de ocurrir. Supóngase que el experimento se repite n veces de forma independiente (los sucesivos resultados no se condicionan unos a otros) en las mismas condiciones. Se define la variable aleatoria X_i como 1 ó 0 según en el i-ésimo experimento ocurra o no A. De esta forma, $X_1, \ldots, X_n$ son variables aleatorias independientes y con ley de probabilidad de Bernoulli de parámetro π. La variable aleatoria $Y = \sum_{i=1}^{n} X_i$ representa el número de veces que ocurre A en las n repeticiones del experimentos.

Por definición, la variable aleatoria Y tiene distribución de probabilidad ***binomial*** de parámetros n y π la cual se expresa por $b(n, \pi)$. Nótese que la distribución de Bernoulli de parámetro π se corresponde con la distribución $b(1, \pi)$.

La esperanza y varianza de la distribución binomial pueden obtenerse fácilmente a partir de la distribución de Bernoulli. En efecto:

$$E[Y] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = n\pi$$

$$\text{var}(Y) = \text{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{var}(X_i) = n\pi(1-\pi)$$

Normal. Una familia de distribuciones de probabilidad clave para el análisis de datos es la *gausiana* o *normal*. Se dice que una variable aleatoria tiene distribución ***normal*** de esperanza μ y desviación estándar σ si su distribución de probabilidad puede expresarse por una función de densidad de la forma:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

La gráfica de esta distribución es la bien conocida *campana de Gauss*. La función representada en la figura 10 corresponde a una distribución normal de esperanza 80 y desviación estándar 5. Las distribuciones representadas en la figura 8 también son normales de esperanza 80 y distintas desviaciones estándar.

Teorema de la tipificación. Si una variable aleatoria X tiene distribución de probabilidad $N(\mu, \sigma)$, entonces $Z = (X-\mu)/\sigma$ tiene ley de probabilidad $N(0,1)$ (**normal estándar o tipificada**).

Para una variable aleatoria Z con distribución normal estándar ($Z \cong N(0,1)$), utilizaremos las siguientes representaciones:

- La función de distribución acumulativa por $\Phi(t)$, lo que significa que: $\Phi(t) = \Pr(Z \leq t)$.
- El cuantil α por z_α; esto es: $\Phi(z_\alpha) = \alpha$

Ejercicio. De acuerdo con los datos del estudio del Telde, asumiremos que la hemoglobina glicada en escala logarítmica ($X = \ln(HbA1c\%)$) sigue en el grupo de diabéticos

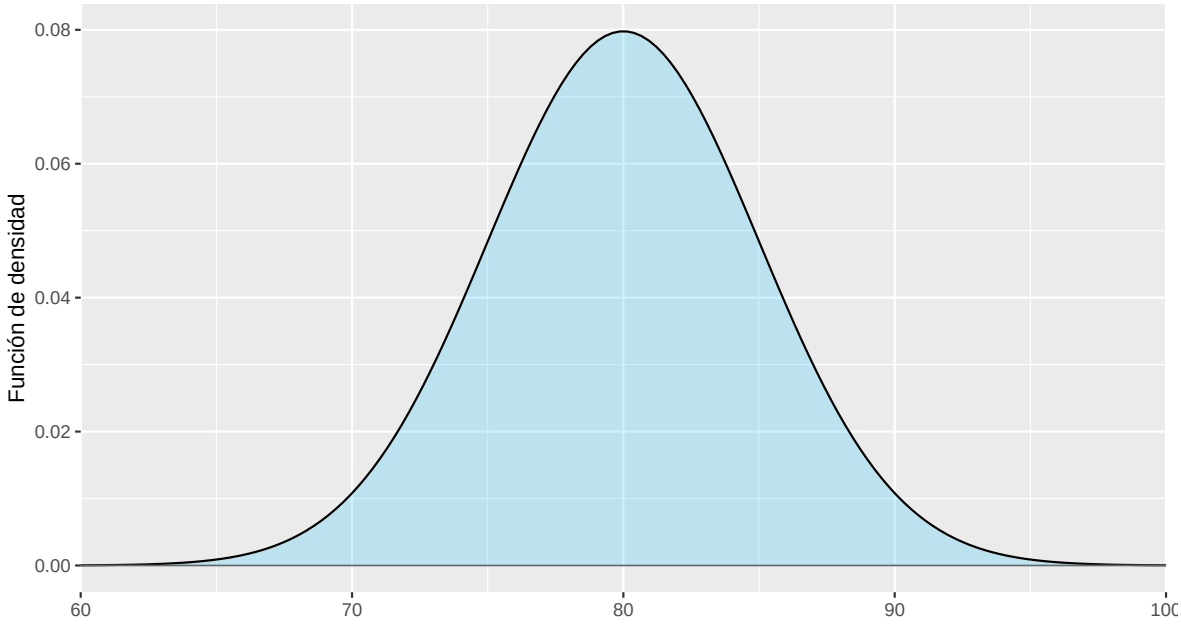

FIGURA 10. Función de densidad de probabilidad de una variable con distribución de probabilidad Normal de esperanza 80 y desviación estándar 5.

(E) una distribución $N(\mu_E = 1,96\,;\sigma_E = 0,22)$ mientras que en el de los no diabéticos (C) sigue una $N(\mu_C = 1.67\,;\sigma_C = 0.08)$. Consideramos entonces el siguiente test diagnóstico: *un sujeto es diabético si y sólo si* $X > K$, para un cierto umbral K.

1. Hallar la *sensibilidad* y *especificidad* del test para $K = \ln(6,5) \approx 1,872$ (el umbral del 6,5% fue el considerado inicialmente por la IDF).

   La *sensibilidad* del test es: $\Pr(X > K \mid E)$. Esta probabilidad puede obtenerse de la tabla de la distribución normal estándar $(N(0,1)$ previa tipificación de la variable X. Dado que condicionalmente a E, $X \cong N(\mu_E = 1,96\,;\sigma_E = 0,22)$, la variable tipificada es:

$$Z = \frac{X - 1.96}{0,22} \cong N(0,1)$$

   A partir de aquí podemos escribir:

$$\Pr(X > 1,872 \mid E) = \Pr\left(\frac{X - 1,96}{0,22} > \frac{1,872 - 1,96}{0,22} \mid E\right) \approx$$

$$\Pr(Z > -0,4) \approx 1 - \Phi(-0,4) \approx 0,655$$

La *especificidad* es: $\Pr(X \leq K \mid C)$. Dado que condicionalmente a C se satisface:

$$X \cong N(\mu_C = 1,67\,;\sigma_C = 0,08)$$

entonces, la especificidad puede obtenerse como:

$$\Pr(X \leq 1,872 \mid C) = \Pr\left(\frac{X-1,67}{0,08} \leq \frac{1,872-1,67}{0,08} \mid C\right) = \Phi(2,525) \approx 0,994$$

2. Hallar el valor de K para que la prueba tenga una *sensibilidad* del 80%.

   El valor de K es la solución de la ecuación: $\Pr(X > K \mid E) = 0.80$. Mediante la tipificación, la ecuación anterior se transforma en:

$$\Pr\left(\frac{X-1,96}{0,22} \leq \frac{K-1,96}{0,22} \mid E\right) = 0,20$$

   Por tanto, $(K-1,96)/0,22 = z_{0,2} \approx -0,8416$. De aquí se deduce: $K = 1,96 - 0,22 \cdot 8,8416 \approx 1,775$. Nótese que al deshacer la escala logarítmica, el umbral del test es: $\exp(1,775) \approx 5,9$.

3. Para el umbral K determinado en el punto anterior, hallar la *especificidad* del test resultante.

   Para el umbral $K = 1,775$, la especificidad de la prueba diagnóstica es:

$$\Pr(X \leq 1,775 \mid C) = \Phi\left(\frac{1,775-1,67}{0,08}\right) \approx 0,9053$$

Weibull. Una distribución de probabilidad frecuente en el análisis de supervivencia es la de Weibull. Su función de distribución acumulativa es:

$$F(t) = 1 - \exp\left(-(t/\lambda)^{\kappa}\right) \ : \ t \geq 0$$

Los parámetros λ, κ reciben el nombre de parámetros de escala y forma respectivamente. Esta distribución se ha utilizado más arriba para modelizar la supervivencia global de pacientes con cáncer de pulmón avanzado.

En aquellos casos en los que $\kappa = 1$, la función de distribución adopta la forma:

$$F(t) = 1 - \exp(-t/\lambda) \ : \ t \geq 0$$

Esta distribución de probabilidad recibe el nombre de *distribución de probabilidad exponencial* de parámetro λ.

Los tiempos de supervivencia de los pacientes afectados por ciertos tipos de cáncer pueden modelizarse a través de la distribución exponencial. Esta distribución presenta una curiosa paradoja que se analiza en la siguiente proposición.

Proposición: Sea X una variable aleatoria con distribución exponencial de parámetro λ. Entonces, para cualesquiera s, t ocurre:

$$\Pr(X > s + t \mid X > s) = \Pr(X > t)$$

Demostración: Dado que $\Pr(X > t) = \exp(-t/\lambda)$ se tiene:

$$\Pr(X > s + t \mid X > s) = \frac{\Pr(\{X > s\} \cap \{X > s + t\})}{\Pr(X > s)} = \frac{\Pr(X > s + t)}{\Pr(X > s)} =$$

$$\frac{\exp(-(s + t)/\lambda)}{\exp(-s/\lambda)} = \exp(-t/\lambda) = \Pr(X > t)$$

Supóngase ahora que X representa la supervivencia de un paciente afectado por un cierto cáncer del que se sabe que ya ha sobrevivido un tiempo s y por tanto, que ha ocurrido el suceso $\{X > s\}$. Se pregunta entonces si aún sobrevivirá al menos un tiempo adicional t; esto es: si ocurrirá el suceso $\{X > s + t\}$. Si X tiene distribución exponencial de parámetro λ, esta probabilidad condicionada por $\{X > s\}$ es, de acuerdo con la proposición anterior, igual a la probabilidad de que en total sobreviva al menos

un tiempo t. Ello significa que el tiempo que aún le queda por vivir (vida residual) es independiente del tiempo que ya ha sobrevivido. Las distribuciones de probabilidad que satisfacen una característica de esta naturaleza reciben el nombre de *distribuciones sin memoria*.

χ^2 (*ji*-cuadrado). Considérese un conjunto de variables aleatorias $Z_1, \ldots, Z_n$ independientes y con distribución de probabilidad común $.N(0,1)$. Entonces, la variable aleatoria definida por:

$$Y = Z_1^2 + \cdots + Z_n^2$$

sigue por definición una ley de

probabilidad $ji-cuadrado$ con n grados de libertad ($\chi^2(n)$). La forma de la gráfica de su función de densidad de probabilidad se muestra en la figura 11 para diversos valores de n.

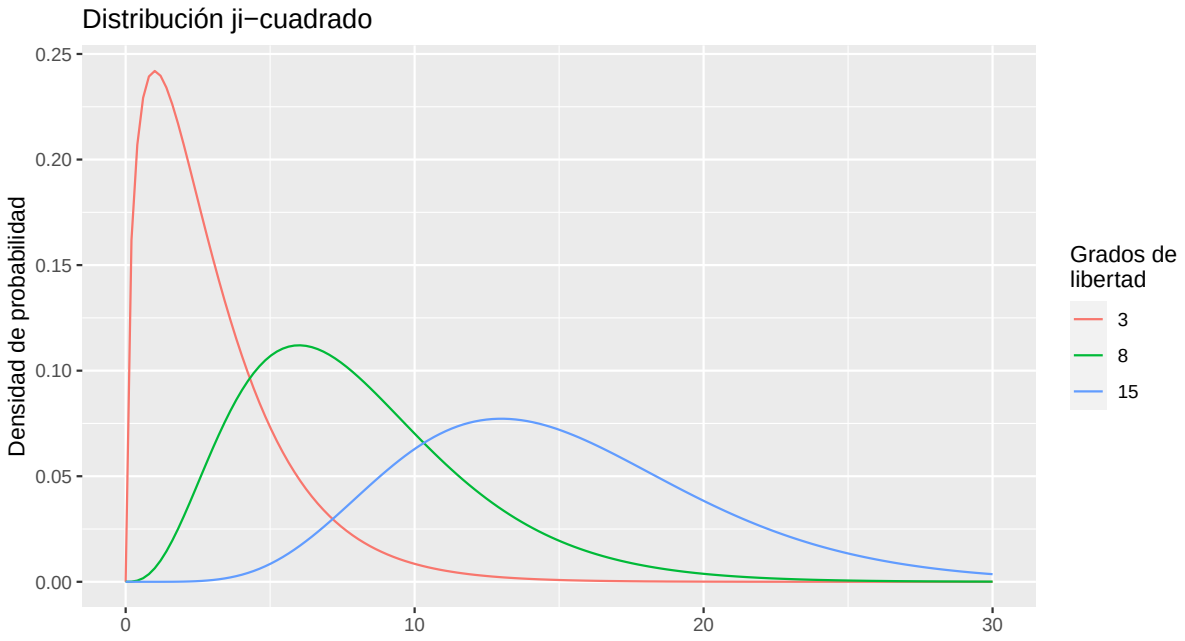

FIGURA 11. Densidades de probabilidad correspondientes a la distribución χ^2 con 3, 8 y15 grados de libertad.

El cuantil α de la distribución $\chi^2(n)$ lo representaremos por $\chi^2_\alpha(n)$. Ello significa que si $Y \cong \chi^2(n)$, entonces $P(Y \leq \chi^2_\alpha(n)) = \alpha$.

Si una variable aleatoria $Y \cong \chi^2(n)$, entonces:

- $E[Y] = n$

- $\text{var}(Y) = 2n$

t de Student. Sean $Z \cong N(0,1)$ y $Y \cong \chi^2(n)$ dos variables aleatorias independientes. Entonces, la variable aleatoria:

$$T_n = \frac{Z}{\sqrt{Y/n}}$$

sigue por definición una ley de probabilidad t de *Student* con n grados de libertad ($t(n)$). Esta distribución de probabilidad tiene propiedades análogas a la distribución $N(0,1)$, siendo su función de densidad de probabilidad simétrica respecto del origen de coordenadas. Además:

$$\Pr(T_n \leq t) \to \Phi(t) \quad , \; n \to \infty$$

En la figura 12 se muestra (en rojo) la densidad de probabilidad de la distribución t de Student para distintos valores de los grados de libertad n. Se ha representado también (en azul) la densidad de la distribución normal estándar. Puede observarse como, a medida que crecen los grados de libertad, las densidades de la distribución t de Student convergen a la densidad de la distribución $N(0,1)$.

El cuantil α de la distribución $t(n)$ lo representaremos por $t_\alpha(n)$. Ello significa que si $T \cong t(n)$, entonces $P(T \leq t_\alpha(n)) = \alpha$.

7. Suma de variables aleatorias independientes: teorema central del límite

Un problema frecuente consiste en obtener la distribución de probabilidad de una suma de variables aleatorias independientes. Su solución requeriría conocer la ley de probabilidad de cada variable aleatoria. Sin embargo, este problema tiene una solución aproximada sorprendentemente sencilla que no requiere el conocimiento de las leyes

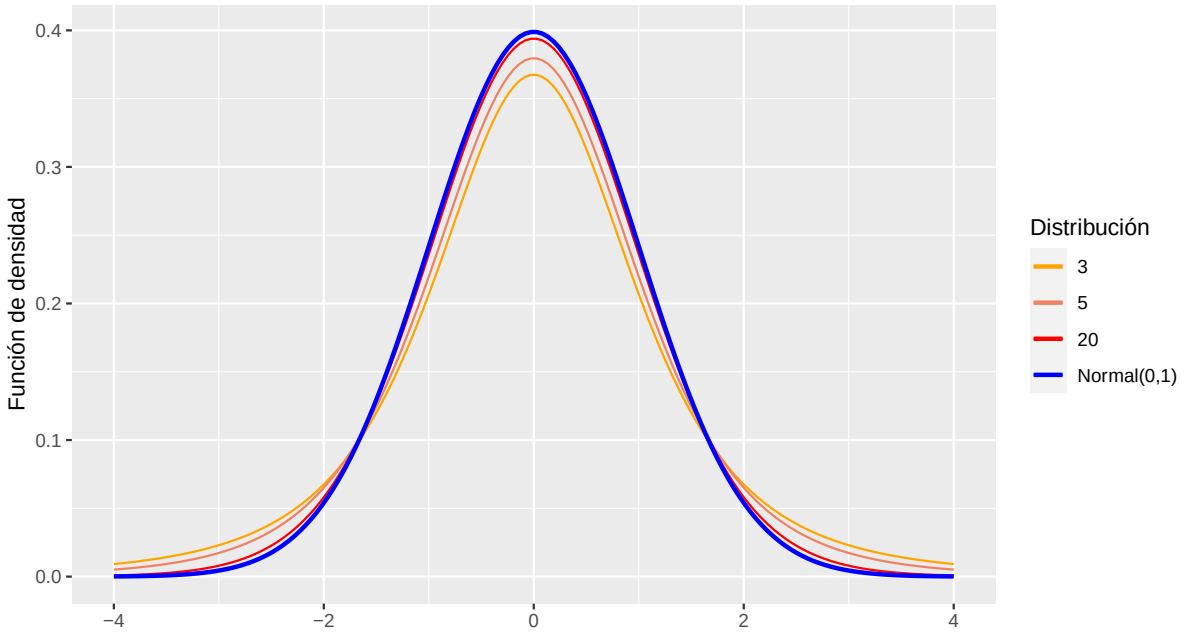

FIGURA 12. Distribución t de Student con distintos grados de libertad. Se ha representado también la distribución normal estándar.

de probabilidad de los sumandos. El teorema central del límite que se da a continuación establece que cualquier suma de variables aleatorias independientes e igualmente distribuidas tipificada tiene aproximadamente una distribución normal estándar.

Teorema: *(Central del límite) Considérese una secuencia de variables aleatorias* $X_1, \ldots, X_n$ *independientes y con la misma distribución de probabilidad, siendo* $E[X_i] = \mu$ *y* $\text{var}(X_i) = \sigma^2$ *para* $i = 1, \ldots, n$. *Entonces, para* $n \to \infty$

$$\Pr\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \leq t\right) \to \Phi(t)$$

Esta notación expresa que la distribución de probabilidad de la suma tipificada se aproxima a la normal estándar según aumenta el tamaño n. Nótese que:

- $E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = n\mu$
- $\text{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{var}(X_i) = n\sigma^2 \Rightarrow \text{sd}\left(\sum_{i=1}^{n} X_i\right) = \sigma\sqrt{n}$

Aproximación de la distribución binomial por la normal. Dado que una variable aleatoria $Y \cong b(n, \pi)$ puede expresarse como la suma de variables independientes $X_1, \ldots, X_n$ ($Y = \sum_{i=1}^{n} X_i$), donde $X_i \cong b(1, \pi)$, puede hacerse entonces, en virtud del teorema central del límite, la siguiente aproximación:

$$\frac{Y - n\pi}{\sqrt{n\pi(1-\pi)}} \approx N(0,1)$$

Supóngase que la prevalencia de una cierta enfermedad es del 12%. Ello significa que en una muestra aleatoria de 500 personas, el número enfermos es una variable aleatoria Y con distribución $b(n = 500; \pi = 0.12)$. Puede entonces hacerse la siguiente aproximación:

$$\frac{Y - 60}{7.266} \approx N(0,1)$$

o también: $Y \approx N(60; 7,233)$.

8. Distribuciones de probabilidad - distribución de datos

En el estudio de Telde se seleccionó una muestra aleatoria de 1.030 personas de esa población con edades superiores o iguales a 30 años. En cada una de las personas seleccionadas se observó, entre otros caracteres, el nivel de la *lipoproteína de baja densidad* (LDL, por sus siglas en inglés *Low Density Lipoprotein*). Cada vez que se seleccionaba una persona ***aleatoriamente*** y se determinaba su valor de LDL, se estaba observando una variable aleatoria cuya distribución de probabilidad asumimos que es aproximadamente $N(\mu = 134; \sigma = 32)$. Es obvio que una sóla observación no aporta información sobre el patrón que genera tales observaciones (en este caso, su distribución de probabilidad) o su relación con otros marcadores. Por tal motivo, la acción de seleccionar un sujeto de la población al azar se repitió n veces ($n = 1030$). Si representamos por X_i el *nivel de la LDL del sujeto que se obsevará en el i-ésimo lugar*, es obvio que $X_1, ..., X_n$ son variables aleatorias *independientes* (los valores observados no se condicionan entre sí), con distribución de probabilidad común $N(\mu = 134; \sigma = 32)$. Tal secuencia recibe el nombre de *muestra aleatoria simple* de la distribución de probabilidad $N(\mu = 134; \sigma = 32)$. El *histograma de frecuencias relativas* que aparece en la figura 13 corresponde a la representación gráfica de las 1.030 observaciones de la

LDL. Superpuesta al histograma, se representa la función de densidad de las variables aleatorias X_i.

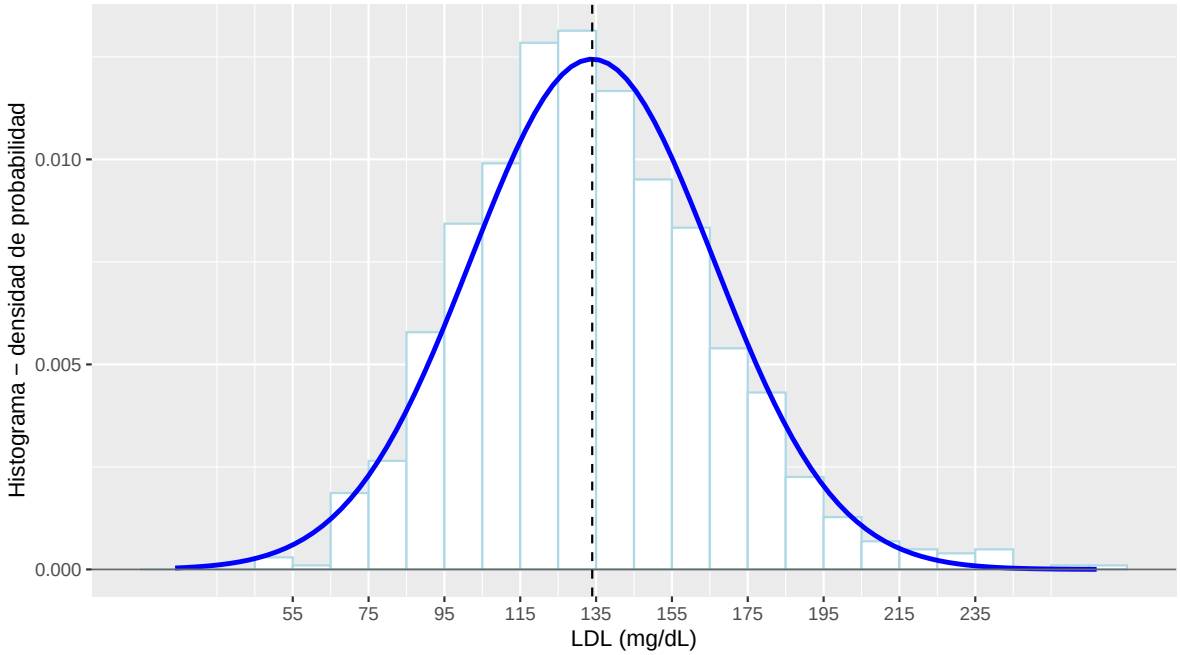

FIGURA 13. Histograma de frecuencias relativas de la LDL en el estudio de Telde. Se ha superpuesto la función de densidad de una $N(\mu = 134; \sigma = 32)$

Obsérvese que la función de densidad de probabilidad es un instrumento que permite una ***mirada hacia el futuro:*** ¿cuál es la probabilidad de que la LDL de un paciente que se seleccione aleatoriamente (en el futuro) sea superior a 160 mg/dL? Los datos que han sido generados por la distribución de probabilidad $N(\mu = 134; \sigma = 32)$ se han obtenido en el pasado.

Si queremos predecir el valor de LDL que tendrá un sujeto que se seleccione aleatoriamente de esta población (en el futuro), tendríamos que hacer uso de la *función de densidad.* Por ejemplo, la probabilidad de que su LDL esté entre 105 y 115 se obtiene como el área bajo la gráfica de la función de densidad sobre el intervalo 105 - 115 cuyo valor es 0.0939. Por otro lado, el *histograma* permite ver que la proporción de personas cuya LDL tomó un valor entre 105 y 115 fue $0,010 \times 10 = 0,10$ (aproximadamente 0.0939).

Así pues, los datos generados por la distribución de probabilidad contienen información para estimar tal distribución de probabilidad. Este es el propósito de la **inferencia estadística** cuyo estudio se iniciará en la siguiente lección.

Ejercicios

1. En una cierta población de mujeres con cáncer de mama, la *supervivencia libre de enfermedad* medida (periodo entre remisión y recidiva) evaluada en meses es una variable aleatoria X con distribución de probabilidad especificada por la función de distribución de probabilidad $F(t) = 1 - \exp(-t/\lambda) \,:\, t > 0$, siendo $\lambda = 117$. Para una mujer de esta población, hallar las probabilidades de que recidive:

   a) después de los 40 meses.

   b) entre los 40 y 60 meses.

   c) después de los 60 meses sabiendo que a los 40 meses seguía libre de enfermedad.

2. Del banco de tumores de Rotterdam se extrajeron datos correspondientes a 2.982 mujeres con cáncer de mama primario. Para cada una de las pacientes se disponía de los niveles del receptor de progesterona (RPG) medido en fmol/L. Las funciones de distribución de probabilidad correspondientes a la variable aleatoria $X =$ *meses de supervivencia libre de enfermedad* según los niveles del RPG fuesen o no superiores a 10 fmol/L estimadas de los datos fueron (*Weibull*):

$$F_0(t) = \Pr(X \leq t \mid RPG < 10) = 1 - \exp\left(-(t/120)^{\kappa}\right)$$

$$F_1(t) = \Pr(X \leq t \mid RPG \geq 10) = 1 - \exp\left(-(t/152)^{\kappa}\right)$$

   siendo en ambos casos $\kappa = 0.918$. En la figura 14 se muestran sus representaciones gráficas.

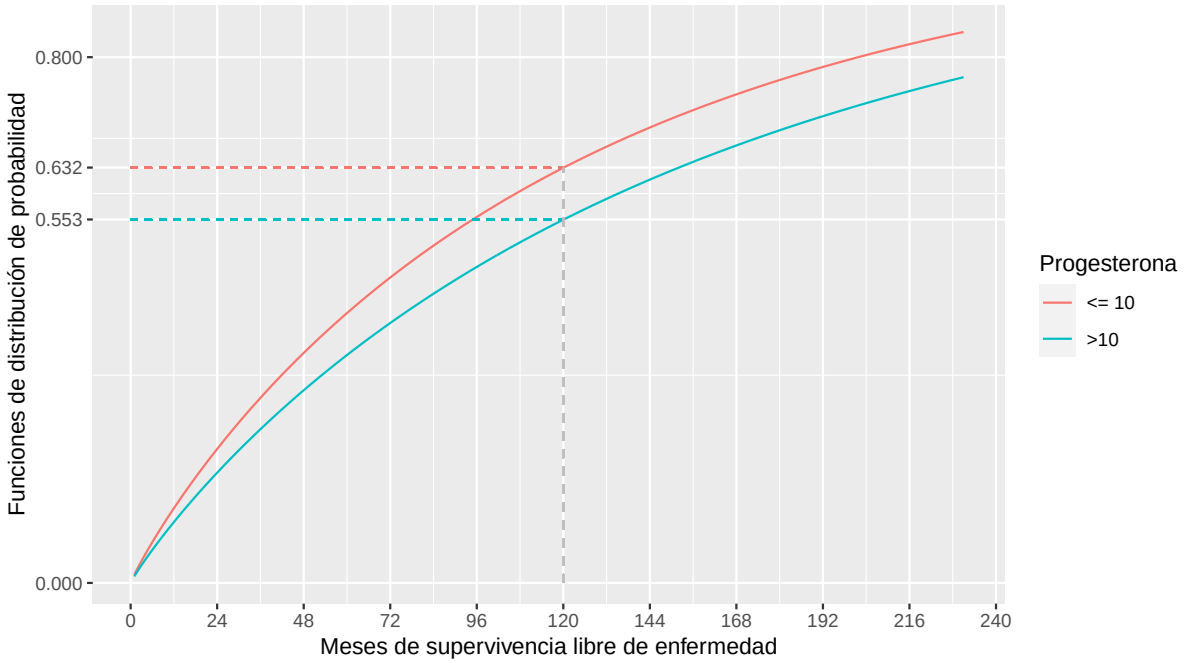

FIGURA 14. Curvas de supervivencia del estudio de Rotterdam

a) Hallar las probabilidades de recidiva antes de los 10 años según $RPG < 10$ fmol/L ó $RPG \geq 10$ fmol/L.

b) Cuánto más probable es que a los 10 años, una paciente con $RPG < 10$ recidive en relación a una paciente con $RPG \geq 10$ fmol/L.

3. Laurie *et al* llevaron a efecto un ensayo clínico en pacientes con carcinoma colorrectal resecado en estadios B y C. Los pacientes fueron asignados aleatoriamente a no recibir más terapia (***Obs***) o a recibir tratamiento adyuvante con levamisol solo (***Lev***), 150 mg/durante 3 días cada 2 semanas durante 1 año, o levamisol más fluorouracilo (5-FU), 450 mg/m2/d por vía intravenosa (IV) durante 5 días y, a partir de los 28 días, 450 mg/m2 semanalmente durante 1 año. Las funciones de distribución de probabilidad correspondientes a la ***supervivencia global*** (X) estimadas según tratamiento fueron (ver fig-colon):

$$F_0(t) = \Pr(X \leq t \mid Obs) = 1 - \exp(-t/98.5)$$

$$F_1(t) = \Pr(X \leq t \mid Lev) = 1 - \exp(-t/102)$$

$$F_2(t) = \Pr(X \leq t \mid Lev + FU) = 1 - \exp(-t/145.7)$$

a) Hallar $F_0(60)$, $F_1(60)$ y $F_2(60)$ e interpretar estas cantidades.

b) Obtener $F_1(60)/F_0(60)$ y $F_2(60)/F_0(60)$ e interpretar los valores.

c) Establecer una comparación de resultados a los 5 años de seguimiento.

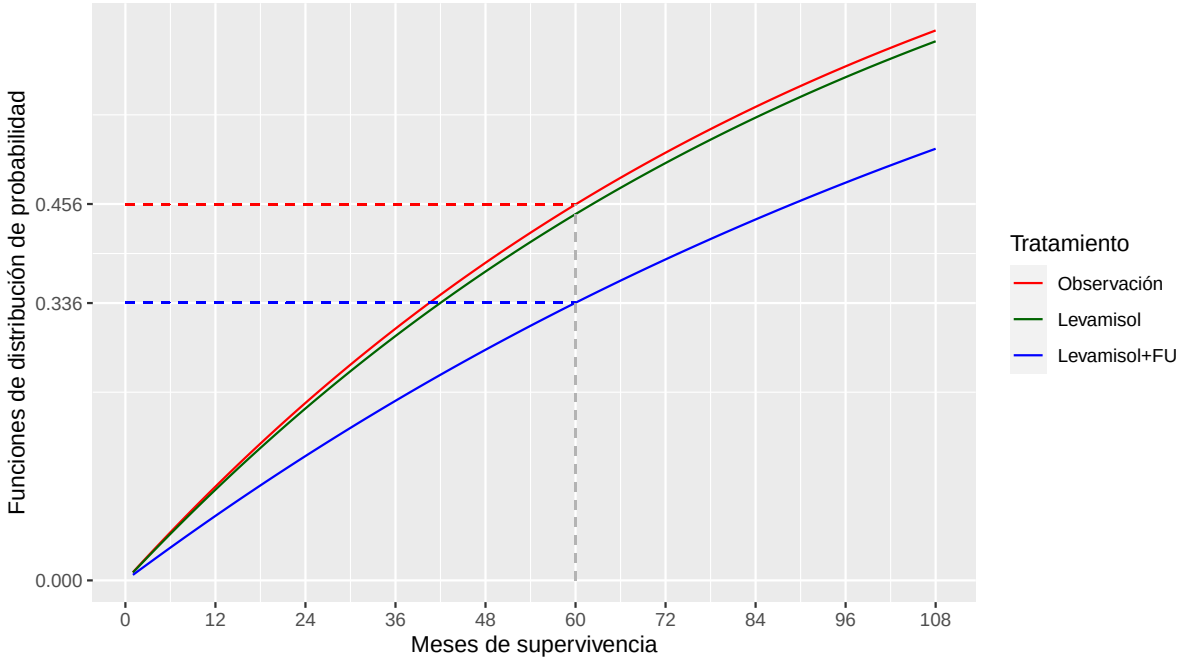

FIGURA 15. Funciones de distribución de probabilidad para la supervivencia global de pacientes con cáncer de colon según tratamiento

4. Consideremos un juego de azar en el que participan 60 sujetos, aportando cada uno 100 ¤. A cada sujeto participante se le asigna un número. Se seleccionan entonces tres números consecutivamente al azar sin reemplazamiento. El individuo que tenga el primer número seleccionado recibe entonces un premio de 3.000 ¤, el que tenga el segundo de 2.000 ¤ y el que tenga el tercero, 1.000 ¤. Describir la variable aleatoria que representa la ganancia neta de cada jugador y calcular su esperanza.

5. Supongamos que el 30% de los individuos de una cierta población pertenecen al grupo sanguíneo A. Para una muestra aleatoria de tamaño 20 extraída de esta población, encontrar las siguientes probabilidades:

   a. Exactamente tres personas pertenezcan al grupo A.

b. Al menos tres personas pertenecen al mencionado grupo.

c. Menos de tres personas son del grupo A.

6. Si X es una variable aleatoria constante tal que $\Pr(X = \kappa) = 1$, probar que:

   a. $E[X] = \kappa$.

   b. $\text{var}(X) = 0$.

7. Para dos variables aleatorias X e Y, probar las siguientes propiedades:

   a. $E[X - Y] = E[X] - E[Y]$.

   b. Si X e Y son independientes, entonces $\text{var}(X - Y) = \text{var}(X) + \text{var}(Y)$.

8. De acuerdo con los datos del estudio de Telde, puede admitirse que para la población estudiada, los niveles de la lipoproteína de baja densidad (LDL) siguen una ley de probabilidad $N(\mu = 134\,;\, \sigma = 32)$. En condiciones normales, los valores de este marcador se consideran elevados cuando superan el umbral de 160 mg/dL. Resolver las siguientes cuestiones.

   a. Si de la población se selecciona aleatoriamente un sujeto al azar, determinar la probabilidad de que su nivel de LDL esté elevado.

   b. Si se seleccionan 20 sujetos de la población, ¿cuál es la probabilidad de que a lo sumo tres tengan elevados los niveles de LDL? ¿Y de qué al menos tres tengan elevados los valores del referido marcador.

   c. Si de la misma población se seleccionan 300 sujetos al azar, ¿cuál es la probabilidad de que a lo sumo 70 tengan elevados los niveles del marcador? ¿Cuál es el número esperado de sujetos con el marcador elevado?

9. La osteoporosis es una patología que afecta a una de cada tres mujeres a partir de la menopausia (33%). Su determinación se realiza habitualmente mediante la técnica DXA. Recientemente se ha propuesto el uso de marcadores basados en ultrasonografía en el calcáneo. El *qui-stiffness* (X) resume los valores dados por esta técnica. De acuerdo con estudios desarrollados recientemente en mujeres postmenopáusicas, puede admitirse que este marcador en la población de mujeres sanas (no osteoporóticas) se distribuye $N(77; 14)$ mientras que en las mujeres con osteoporosis (D) su distribución es $N(65; 14)$ lo que supone que la osteoporosis se asocia con disminución en el *qui-stiffness*. En orden a utilizar esta variable como marcador, debe obtenerse un valor de corte C (cut-off), tal

que, una mujer será diagnosticada como osteoporótica si $X \leq C$ y como normal en caso contrario.

a. Determinar el valor de C para que la prueba diagnóstica alcance una sensibilidad del 95%.

b. Para el valor C determinado en el apartado anterior, hallar la tasa de falsos positivos.

c. Admitiendo que en la población de mujeres postmenopáusicas la prevalencia de osteoporosis es $1/3$, hallar el valor predictivo positivo de la prueba.

10. El CYFRA-21 es un marcador que se utiliza para el diagnóstico del cáncer de pulmón. De acuerdo con los datos de Pujol *et al* (1993), puede admitirse que en un grupo de pacientes con cáncer de pulmón, $X = \log(CYFRA - 21)$ es tal $X \cong N(1.45; 1.1)$ que mientras que en un grupo de controles, $X \cong N(0.087; 0.704)$. Se considera entonces como verosímil que un sujeto tenga cáncer de pulmón si , para un cierto valor de corte K.

a. Determinar el valor de K para que la prueba tenga una sensibilidad del 80%.

b. Para el valor de K obtenido en el apartado anterior, hallar la especificidad de la prueba.

c. Hallar el valor predictivo negativo de la prueba en una población en la que el 30% tiene cáncer de pulmón.

11. La proteína C reactiva (PCR) es una proteína plasmática que aumenta sus niveles en respuesta a la inflamación. En orden a examinar su capacidad diagnóstica para la neumonía adquirida en la comunidad (NAC), Almirall *et al* llevaron a efecto un estudio en el que encontraron que en el grupo de NAC confirmada (E), $PCR \cong N(111; 45)$ mientras que en el grupo no confirmada (C), $PCR \cong N(50; 35)$. Considérese la siguiente regla diagnóstica: un sujeto es susceptible de NAC si y sólo si $PCR > K$.

a. Obtener el valor de K para que la prueba tenga una sensibilidad del 80%.

b. Hallar la especificidad para el valor de K obtenido en el apartado anterior.

12. Mentese *et al* (2008) realizaron un estudio para investigar el efecto de la trombosis venosa profunda (TVP) sobre los niveles de la albúmina modificada por la

isquemia (IMA). La media (SD) los niveles plasmáticos de la IMA fueron 0.259 (0.066) unidades de absorbancia (ABSU) en el grupo de TVP y 0.171 (0.045) ABSU en el grupo control. Consideramos entonces que un sujeto puede haber sufrido una TVP si su nivel de IMA es superior a un cierto umbral K. Admitiendo que el marcador está normalmente distribuido en cada uno de los grupos, determinar:

a. Valor de K para que la prueba diagnóstica tenga una sensibilidad del 80%.

b. Especificidad de la prueba para el valor de K determinado en el apartado anterior.

c. Valor predictivo positivo en una población en la cual, la prevalencia de la TVP es del 40%.

13. El $NGAL$ es un biomarcador que se utiliza para el diagnóstico de la lesión renal aguda. Su ventaja principal es que responde antes que otros marcadores del estado renal y muestra una respuesta proporcional a la lesión. En un estudio llevado a cabo en la UMI del Hospital General de Gran Canaria se observó que en los pacientes con lesión renal aguda (E), el marcador en escala logarítmica ($X = \ln(NGAL)$) sigue una distribución de probabilidad $N(\mu_E = 5.05; \sigma_E = 1.69)$ mientras que en el grupo de control (C), $X \sim N(\mu_C = 3.93\,; \sigma_C = 1.37)$. El marcador predice que existe lesión renal aguda si $X > K$.

    a. Determinar K para que la prueba tenga una sensibilidad del 70%.

    b. Hallar la especificidad para el valor de K determinado en el apartado anterior.

    c. Hallar el valor predictivo positivo de la prueba diagnóstica para una población en la que la tasa de lesión renal aguda es el 37%.

14. Sean Y_1 e Y_2 variables aleatorias independientes con distribuciones de probabilidad $\chi^2(n_1)$ y $\chi^2(n_2)$ respectivamente. Demostrar que $Y_1 + Y_2$ sigue una distribución $\chi^2(n_1 + n_2)$.

15. Se lanza un dado al azar dos veces, ¿cuál es la probabilidad de que la suma de los números obtenidos sea menor o igual a ocho? Y si el dado se lanza 500 veces, ¿cuál es la probabilidad de que la suma de los números obtenidos sea menor o igual a 1.900 ?

Referencias

1. Loprinzi CL, Laurie JA, Wieand HS, Krook JE, Novotny PJ, Kugler JW, Bartel J, Law M, Bateman M, Klatt NE, et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. J Clin Oncol. 1994 Mar;12(3):601-7. doi: 10.1200/JCO.1994.12.3.601. PMID: 8120560.

2. Hofman A, Grobbee DE, de Jong PT, van den Ouweland FA. Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. Eur J Epidemiol 1991;7:403–22. pdf

3. Laurie JA, Moertel CG, Fleming TR, Wieand HS, Leigh JE, Rubin J, McCormack GW, Gerstner JB, Krook JE, Malliard J, et al. Surgical adjuvant therapy of large-bowel carcinoma: an evaluation of levamisole and the combination of levamisole and fluorouracil. The North Central Cancer Treatment Group and the Mayo Clinic. J Clin Oncol. 1989 Oct;7(10):1447-56. doi: 10.1200/JCO.1989.7.10.1447. PMID: 2778478.

4. Boronat, M., Varillas VF, Saavedra P, Suárez V, Bosch E, Carrillo A, Nóvoa FJ. Diabetes mellitus and impaired glucose regulation in the Canary Islands (Spain): prevalence and associated factors in the adult population of Telde, Gran Canaria. Diabet Med. 2006 Feb;23(2):148-55. doi: 10.1111/j.1464-5491.2005.01739.x. PMID: 16433712.

5. Pujol JL, Grenier J, Daurès JP, Daver A, Pujol H, Michel FB. Serum fragment of cytokeratin subunit 19 measured by CYFRA 21-1 immunoradiometric assay as a marker of lung cancer. Cancer Res. 1993 Jan 1;53(1):61-6. PMID: 7677981. pdf

6. Almirall J, Bolíbar I, Toran P, Pera G, Boquet X, Balanzó X, Sauca G; Community-Acquired Pneumonia Maresme Study Group. Contribution of C-reactive protein to the diagnosis and assessment of severity of community-acquired pneumonia. Chest. 2004 Apr;125(4):1335-42. doi: 10.1378/chest.125.4.1335. PMID: 15078743.

7. Mentese A, Mentese U, Turedi S, Gunduz A, Karahan SC, Topbas M, Turan A, Patan T, Turkmen S, Okur G, Eminagaoglu MS. Effect of deep vein thrombosis on ischaemia-modified albumin levels. Emerg Med J. 2008 Dec;25(12):811-4. doi: 10.1136/emj.2007.056614. PMID: 19033496.