

# 6

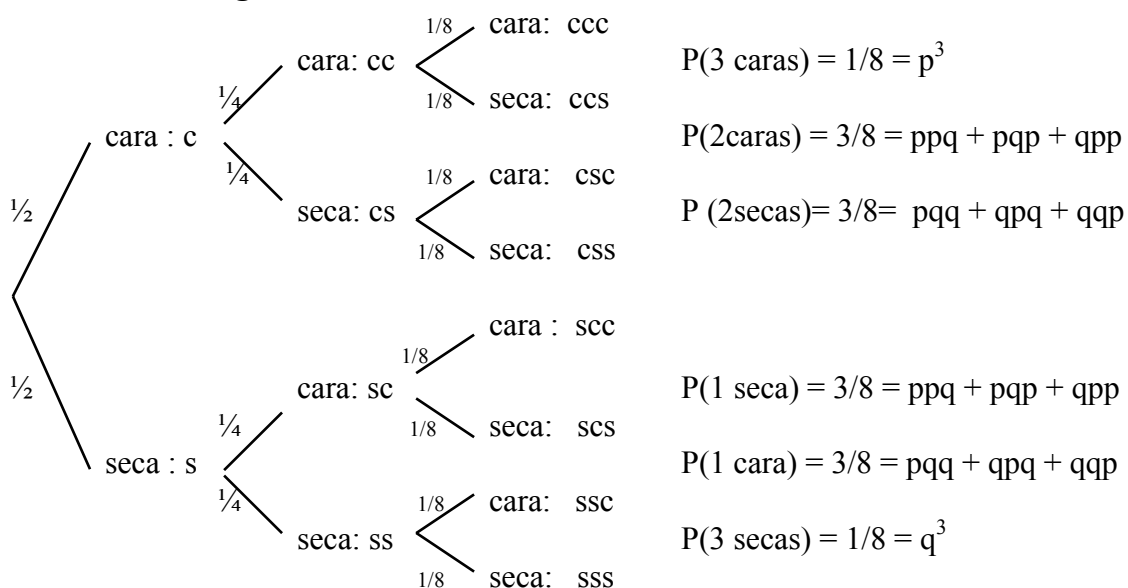
## Probabilidad Condicional

En este capítulo se introduce el concepto de la probabilidad condicional a la problemática probabilística. Se comienza con el enfoque clásico, para luego ir avanzando hacia las relaciones bayesianas, en particular entre índices clínicos. Se discuten los conceptos de independencia y condicionalidad para lograr un enfoque más general que el visto hasta ahora.

### 6.1 Introducción

La Teoría de Probabilidades comenzó con el estudio de los juegos de azar. El caso más simple es el lanzamiento de una moneda, donde pueden ocurrir dos casos posibles. (c: cara y s: seca). Al repetir la misma prueba, los casos posibles ahora serán cuatro, y al hacerlo una tercer vez serán ocho. Un diagrama de árbol ayudara a visualizar mejor la cuestión (Gráfico 6.1). Es de esperar que la probabilidad  $p = \frac{1}{2}$  de obtener una cara, o  $q = \frac{1}{2}$  de tener una seca, siempre sea la misma y no se altere por más veces que se lance la moneda al aire. Es decir, se trata de una probabilidad constante a lo largo de una serie de pruebas. Lo mismo ocurre con la mayoría de los juegos de azar, incluso con barajas cuando las extracciones se hacen con reposición. Resultó natural entonces, calcular la probabilidad de los lanzamientos sucesivos, como el producto de las probabilidades básicas  $p$  y  $q$ . Así, la probabilidad de sacar 3 caras seguidas es de  $\frac{1}{8}$  como se aprecia en el diagrama y se obtiene con  $p^3$ .

Gráfico 6.1: Diagrama del árbol.



Como puede verse más arriba, el producto de las probabilidades de cada rama, siguiendo el camino adecuado, permite obtener los valores de cada uno de los resultados posibles al hacer los tres lanzamientos. Ese método se conoció con el nombre de *regla del producto*. A su vez, cuando se busca obtener un resultado indirecto que se da en más de una rama, entonces se aplica la *regla de la suma*, esto es, se suman las probabilidades de cada una de las ramas involucradas. Los ejemplos de todo se han presentan en el Gráfico 6.1.

Sin embargo, esta manera de resolver problemas no sirve para todos los casos. Hay un ejemplo histórico de ello llamado “La paradoja de Marte”:

“... Muchos años atrás, se planteaba como problema la posible existencia de vida en el planeta Marte. Como no se sabía mucho del planeta, salvo que tenía atmósfera como la Tierra, se especulaba acerca de la existencia de alguna especie animal o vegetal con vida. Si se parte de la base de no saber nada, se puede asumir como  $\frac{1}{2}$  la probabilidad  $p$  de que haya vida, o  $q$  de que no la haya, para una especie cualquiera como la del Homo-Sapiens. Ahora, tomando otra especie como los peces, por un razonamiento análogo se llega al mismo valor,  $p = q$ . Y así sucesivamente, para  $n$  especies se tendrá que la probabilidad total de que haya especímenes vivos en Marte será:

$$P(\exists \text{ vida}) = P(\text{hombres y vacas y aves y .... peces vivos})$$

$$P(\exists \text{ vida}) = P(\text{hombres vivos}) \cap P(\text{vacas vivas}) \cap P(\text{aves vivas}) \cap \dots \cap P(\text{peces vivos})$$

$$P(\exists \text{ vida}) = p \cdot p \cdot p \dots p = p^n$$

Lo que conduce a una paradoja, pues 
$$P(\exists \text{ vida}) = p^n \xrightarrow[n \rightarrow \infty]{} 0$$

O sea, se parte de una ignorancia estimada en  $p = q = \frac{1}{2}$  y se llega a la certeza de que no hay vida. Y ocurre lo mismo tomado otros valores de  $p$  y  $q$ . Si se estima que la probabilidad de vida es muy baja  $p = 0,001$ , también tiende a cero con  $n$  muy grande. Si se estima lo contrario, una alta probabilidad de vida  $p = 0,9$ , entonces tiende más rápido”.

Otra paradoja: tomando una cepa de *Drosophila* que difiera de la cepa salvaje en dos caracteres: el color del cuerpo es negro ébano ( $E$ ), mientras el color normal es ( $E+$ ) y las alas son vestigiales ( $A$ ), mientras las normales ( $A+$ ). Si se cruza una hembra salvaje ( $e a$ ) con un macho normal ( $e+ a+$ ), en la primera generación F1 se obtienen únicamente moscas del tipo ( $e+ a+$ ) que son los caracteres dominantes. Luego, se realiza un segundo cruzamiento retrógrado, de los obtenidos en F1 con normales. En la segunda generación filial F2 se obtendrán cuatro tipos de casos posibles, todos igualmente probable  $p = \frac{1}{4}$  de acuerdo a las leyes de Mendel: “los gametos sólo transportan uno de los genes alelos presentes en los progenitores”, habiendo igualdad de frecuencias para ambos gametos. Los cuatro tipos de *Drosophilas* son:

		Mendel	Frec. obtenida
cuerpo normal y alas normales :	$E+ \cap A+$ o sea ( $e+ a+$ )	25%	240
cuerpo normal y alas vestigiales:	$E+ \cap A$ o sea ( $e+ a$ )	25%	266
cuerpo negro y alas normales :	$E \cap A+$ o sea ( $e a+$ )	25%	278
cuerpo negro y alas vestigiales:	$E \cap A$ o sea ( $e a$ )	25%	252

Habrá igual número de machos y hembras. Hecho el experimento se obtuvieron valores muy parecidos a los mendelianos, como era de esperar.  $p = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

Si se hubieran cruzado machos y hembras de la F1, entonces en la F2 se hubiesen obtenido también cuatro tipos de moscas, pero con diferentes proporciones, de acuerdo a la segunda ley de Mendel : 9/16, 3/16, 3/16 y 1/16. O sea ( 9:3:3:1), válida para la segregación de caracteres independientes cuando se cruzan entre sí dos individuos de la F1.

Si se analizan ahora otros caracteres, como por ejemplo una cepa salvaje con alas vestigiales y ojos pardos (**B**), en el primer cruzamiento F1, resultaran todos con los caracteres dominantes (**b+ a+**) Un cruzamiento retrógrado entre una hembra F1 y un macho salvaje dará lugar a cuatro tipos de moscas:

	Mendel	Frec. obtenida
ojos normales y alas normales : $B+ \cap A+$ o sea ( <b>b+ a+</b> )	25%	370
ojos normales y alas vestigiales: $B+ \cap A$ o sea ( <b>b+ a</b> )	25%	138
ojos pardos y alas normales : $B \cap A+$ o sea ( <b>b a+</b> )	25%	155
ojos pardos y alas vestigiales: $B \cap A$ o sea ( <b>b a</b> )	25%	350

Se da una paradoja, pues los valores obtenidos, distan de parecerse a los mendelianos esperados. Tomando esa evidencia, la conclusión sería que para esos caracteres dejan de cumplirse las leyes de Mendel, cosa que ningún genetista está dispuesto a aceptar.

En síntesis, se muestran dos paradojas que despertaron polémicas en su momento, consecuencia de aplicar la llamada *regla del producto*, lejos de su ámbito natural de los juegos de azar. Se puede pensar lo siguiente: si la probabilidad de encontrar una vaca viva en Marte es p, seguramente la probabilidad de encontrar la misma vaca conviviendo con tigre vivo no será la misma. Aumentará para el tigre y disminuirá para la vaca. Esto es, hay hechos cuya ocurrencia *condicionan* la ocurrencia de otros. En cambio, si al tirar una moneda sale cara, este hecho *no condicionará* lo que salga en la tirada siguiente. Entonces, había que replantearse el concepto de las probabilidades teniendo esto en cuenta. Y así, se llegó al concepto de *Probabilidad Condicional*. En la segunda paradoja, los genetistas lejos de sacrificar las leyes de Mendel, pensaron que los caracteres que no la cumplían no eran *independientes*, y así definieron un nuevo concepto el de los *caracteres ligados*.

## 6.2 Independencia

Se dice entonces que dos sucesos son *independientes* cuando la ocurrencia de uno de ellos, no modifica la ocurrencia del otro, ni esta influenciado por este. Si se realiza una serie de pruebas repetidas, las pruebas son *independientes*, cuando el resultado de una de ellas no está influenciada por el resultado de la prueba anterior, ni tampoco influenciará el resultado de la prueba siguiente.

Es importante no confundir independencia con exclusión. Pues en este último caso, la ocurrencia de un suceso hace imposible la ocurrencia del otro. O sea, se “obligan” entre sí. Esto lleva a pensar que dos sucesos son excluyentes cuando su intersección es el conjunto vacío, o sea, la probabilidad de que ocurra uno y el otro, será nula. En cambio, para que sean independientes deberán tener por lo menos un elemento en común, para que la probabilidad de su ocurrencia conjunta no sea nula. Por otro lado, para que un suceso condicione a otro también debe

haber algún elemento en común, pues si no hay ninguna relación entre ellos no podrá haber influencia de uno sobre el otro. Se pueden resumir estos casos con:

- . *Sucesos excluyentes* : No tienen ningún elemento en común. O sea,  $A \cap B = \emptyset$
- . *Sucesos independientes*: Deben tener algún elemento en común. O sea,  $A \cap B \neq \emptyset$
- . *Sucesos condicionales* : Deben tener algún elemento en común. O sea,  $A \cap B \neq \emptyset$

Se dice que dos sucesos  $A$  y  $B$  son *independientes* cuando la probabilidad de que ambos ocurran es igual al producto de sus probabilidades mutuas (independencia estadística). O sea:

Si se cumple,  $P(A \cap B) = P(A) \cdot P(B)$  entonces son independientes.

Esta es la vieja regla del producto. Se cumple en las reglas de la herencia humana, en los juegos de azar, en la teoría cinética de gases (Maxwell), y en un sinnúmero de casos de la vida real. Además, toda sucesión finita de experimentos en las cuales cada experimento tiene un número finito de resultados posibles, se llama *proceso estocástico* y es el caso típico de aplicación de la regla del producto donde las probabilidades en cada rama son independientes paso a paso; en cambio, si ocurre la secuencia de una rama, es imposible que ocurra la secuencia de otra. O sea, las ramas son excluyentes entre sí, y se les puede aplicar el axioma 3 o la regla de la suma para obtener la probabilidad de más de una rama.

Se puede generalizar con: Sean  $n$  sucesos  $C_i$  todos independientes entre sí, entonces se cumple que :

$$P(C_1 \cap C_2 \cap C_3 \cap \dots \cap C_n) = P(C_1) \cdot P(C_2) \cdot P(C_3) \dots P(C_n)$$

Si se realiza una prueba  $n$  veces, donde cada resultado es un suceso independiente de los demás, la probabilidad de ocurrencia de un resultado común a todos ellos es el producto de sus probabilidades. Por ejemplo, en el diagrama del árbol de Gráfico 6.1, la probabilidad que salgan 3 caras es igual al producto  $p \cdot p \cdot p = 1/8$ , y generalizando al caso de  $n$  lanzamientos, la probabilidad de obtener todas caras será  $(\frac{1}{2})^n$ .

Como corolario de la propiedad anterior, se cumple que si los  $n$  sucesos son independientes entre sí, entonces también lo serán para una cantidad menor de sucesos.

$$P(C_j \cap C_{j+1} \cap C_{j+2} \cap \dots \cap C_k) = P(C_j) \cdot P(C_{j+1}) \cdot P(C_{j+2}) \dots P(C_k)$$

Pero debe destacarse que lo *contrario no se cumple*, es decir si  $n$  sucesos son independientes de a pares, en todos los pares posibles, eso no permite afirmar que lo serán de a tres, de a cuatro, etc. Lo mismo para el caso inverso. Todo esto se muestra en los siguientes ejemplos:

*Ejemplo 1)* Una clínica tiene nueve pacientes en la sala de espera. De los cuales 3 son adultos, 3 son niños y 3 son ancianos. Cada paciente, de cada grupo, tiene un número de turno que va de 1 a 3. Además, el paciente adulto con el turno 1, el niño con el turno 2 y el anciano con el 3 son del sexo masculino y los demás del sexo femenino.

Sean los sucesos:  
 $A$  : un adulto



$$\begin{aligned}P(A) &= P(AC \cup AG) = P(AC) + P(AG) = 1/9 + 2/9 = 1/3 \text{ (por el axioma 3)} \\P(C) &= P(AC \cup GC \cup CG) = P(AC) + P(GC) + P(CG) = 2/3 \text{ (generalización de axioma 3)} \\P(G) &= P(GC \cup AG \cup CG \cup GG) = 1/9 + 2/9 + 1/9 + 2/9 = 2/3 \text{ (idem anterior)} \\P(U) &= P(UU \cup UU) = P(UU) = 2/9 \\P(A \cap C) &= P(AC) = 1/9 = P(A) \cdot P(C) \text{ entonces } A \text{ y } C \text{ son independientes} \\P(A \cap G) &= P(AG) = 2/9 = P(A) \cdot P(G) \text{ entonces } A \text{ y } G \text{ son independientes} \\P(C \cap G) &= P(CG) = 2/9 = P(C) \cdot P(G) \text{ entonces } C \text{ y } G \text{ son independientes}\end{aligned}$$

En cambio  $P(A \cap C \cap G) = P(\emptyset) = 0 \neq P(A) \cdot P(C) \cdot P(G)$  no son independientes. Se puede interpretar esto diciendo: Si un fragmento contiene adenina, la probabilidad de que contenga citosina o uracilo se mantiene constante; en cambio, si contiene adenina y guanina, es imposible que contenga citosina.

Los tres ejemplos anteriores ilustran la cuestión de independencia de a pares y de a triples. Puede verse que, si los pares lo son, no hay garantía que los tripletes lo sea, y viceversa, que lo sean los tripletes, no implica que lo sean de a pares. La única forma de asegurarse, es aplicando la definición a cada caso particular en estudio. *No se puede ver la independencia de un vistazo*

## 6.3 Condicionalidad

Recordando las anécdotas de la paradoja de Marte, o el concepto genético de carácter ligado, como así también la extracción de cartas de un mazo sin reposición, resulta clara la existencia de sucesos que condicionan la ocurrencia de otros. Las relaciones ya vistas de probabilidad resultan insuficientes para tratar estos casos. Se necesita entonces, definir un nuevo concepto: la *probabilidad condicional* que ocurra un suceso  $A$ , luego de que ha sucedido un suceso  $B$ , puede ser calculada con:

$$P(A / B) = P(A \cap B) / P(B)$$

O lo que es lo mismo,

$$P(A \cap B) = P(A / B) \cdot P(B)$$

Se dice entonces, la ocurrencia de  $B$  incide en la ocurrencia de  $A$ . Por ejemplo, la probabilidad de sacar dos cartas sucesivas de espadas en un mazo de 40 cartas españolas, se obtiene con:

$$P(\text{sacar } 1^{\text{a}} \text{ y } 2^{\text{a}} \text{ de espadas}) = P(2^{\text{a}} \text{ espada} / 1^{\text{a}} \text{ espada}) \cdot P(1^{\text{a}} \text{ espada}) = 9/39 \cdot 10/40 = 0,058$$

Esto permite encontrar otra manera de definir dos sucesos independientes. Si sucede que se cumple:  $P(A / B) = P(A)$ , entonces  $B$  no condiciona la ocurrencia de  $A$  y resultan independientes. Resumiendo :

$$\begin{aligned}\text{Condicionalidad} &: P(A \cap B) = P(A / B) \cdot P(B) \\ \text{Independencia} &: P(A \cap B) = P(A) \cdot P(B) \\ \text{Exclusión} &: P(A \cap B) = 0\end{aligned}$$

Como se ve, la condición más general es la condicionalidad, y como casos muy particulares de la misma, se da la condición de independencia o la de exclusión. La consistencia entre esta probabilidad de tipo teórica y la empírica se da con el lema de Bernoulli. Si se sabe que en una serie de

N experimentos hay resultados que verifican la condición  $A$ , acompañado por la ocurrencia de la condición  $B$ , entonces habrá un cierto número de pruebas  $N_{AB}$ , donde se verifican juntas ambas condiciones; entonces su frecuencia relativa se calcula con :

$$Fr_{AB} = N_{AB} / N_B$$

Pues ahora el número total de casos posibles no es más  $N$  sino  $N_B$ , porque se sabe que debe ocurrir primero  $B$ , para que ocurra  $A$ , de acuerdo a la condición impuesta "a priori". De todos los casos  $N_B$ , solo algunos  $N_{AB}$  se verificarán junto con  $A$ . Entonces,

$$\begin{array}{ccc} Fr_{AB} & \longrightarrow & P(A / B) \\ N & \longrightarrow & \infty \end{array}$$

Además,

$$Fr_{AB} = N_{AB} / N_B = (N_{AB} / N) / (N_B / N) = Fr_{(A \cap B)} / Fr_B$$

$$\begin{array}{ccc} Fr_{AB} = Fr_{(A \cap B)} / Fr_B & \longrightarrow & P(A \cap B) / P(B) = P(A / B) \\ N & \longrightarrow & \infty \end{array}$$

Se puede generalizar la definición de probabilidad condicional de la manera siguiente: Sean  $n$  conjuntos  $C_1, C_2, C_3, \dots, C_n$  pertenecientes al universo  $\phi$ , tales que no son mutuamente excluyentes, o sea si  $C_1 \cap C_2 \cap C_3 \cap \dots \cap C_n \neq \emptyset$  entonces,

$$P(C_1 \cap C_2 \cap C_3 \cap \dots \cap C_n) = P(C_1) \cdot P(C_2 / C_1) \cdot P(C_3 / C_1 \text{ y } C_2) \dots$$

Por ejemplo, para obtener la probabilidad de sacar flor de espadas en un juego de truco, dando las tres primeras cartas seguidas:  $P(\text{flor espadas}) = (10/40) \cdot (9/39) \cdot (8/38) = 0,0012$

### 6.3.1 Simulación para un test clínico

Se puede realizar una simulación con probabilidades condicionales siguiendo los lineamientos del experimento de Bernoulli. Imaginando que se tienen 20.000 fichas, la mitad rojas y la otra mitad blancas, cada una numerada consecutivamente de 1 a 20.000. Esto simula a 10.000 individuos enfermos (rojas) y el resto sanos (blancas), como si fueran la población general. El supuesto principal es que la *capacidad de la prueba clínica para detectar a los enfermos y a los sanos es constante en toda la población*. Y puede ser expresada con por ejemplo:

- 1) El método es capaz de detectar correctamente a 3 enfermos de cada 4 enfermos.
- 2) El método es capaz de detectar correctamente a 8 sanos entre 10 de ellos.

Esto significa que la sensibilidad es  $S = 3/4 = 0,75$  y la especificidad es  $E = 8/10 = 0,8$ . Esta capacidad es la misma para todo individuo de la población, y esos valores son los teóricos para dicha población, que tiene una prevalencia  $\pi = 0,5$ . Luego se marcan al azar a las fichas para que cumplan con este requisito: 3 de cada 4 rojas se las identifica como verdadero positivo (vp), y 8 de cada 10 blancas se las identifica como verdadero negativo (vn). Luego cada ficha tendrá un número propio (como si fuese nombre y apellido del individuo) y su verdadero estado (uno de estos cuatro: vp, vn, fp y fn).

*Caso 1:* Se mezclan las fichas rojas y blancas, y de entre ellas se eligen al azar a  $N = 4.000$ . Luego se cuentan los casos encontrados y se tendrán las cuatro frecuencias experimentales:

$N^{\circ} vp$  : número de casos encontrados de fichas rojas con diagnóstico verdadero positivo.

$N^{\circ} vn$  : lo mismo que arriba pero de las fichas blancas (verdadero negativo).

$N^{\circ} fp$  : la frecuencia encontrada de fichas rojas con diagnóstico falso.

$N^{\circ} fn$  : lo mismo que arriba pero para las fichas blancas.

Con estas cuatro frecuencias se puede armar una Tabla de diagnóstico experimental. Pero, por el lema de Bernoulli, como  $N = 4.000$  es grande se sabe estas frecuencias experimentales tenderán a sus respectivos valores teóricos. Y entonces, con los valores experimentales se pueden obtener los índices clínicos experimentales, para ver que tanto se acercan a los valores teóricos. Por ejemplo, se puede deducir que:

$S = P(VP/TE)$  es la probabilidad condicional de encontrar un verdadero positivo, sabiendo que el individuo está enfermo. Lo que se puede expresar como:

$S = P(VP \cap TE) / P(TE)$  y en otros términos sería:

$S = \text{número de fichas rojas y positivas} / \text{número total de fichas rojas}$

Análogamente,  $E = \text{número total de fichas blancas y negativas} / \text{número total de fichas blancas}$ .

De esta forma se pueden obtener experimentalmente los valores de  $S$ ,  $E$ ,  $Y$ ,  $LR+$  y  $LR-$  muy parecidos a los esperados. Además la prevalencia experimental  $p$  es aproximadamente igual a la teórica  $p \approx \pi = 0,5$ . Mientras que los valores predictivos y la eficiencia obtenidos son similares a los esperados cuando la prevalencia  $p = 0,5$ .

Se puede repetir muchas veces esto hasta convencerse de que el lema de Bernoulli funciona bien.

*Caso 2:* Se separan las fichas rojas y blancas en dos conjuntos, y se eligen al azar 2.000 fichas rojas y otras tantas blancas, para tener un total de  $N = 4.000$  muestras en total. La diferencia con el caso anterior es que la muestra ha sido estratificada en mitad enfermos y mitad sanos, *antes de hacer la selección al azar*. Luego se cuentan los casos seleccionados en sus cuatro resultados posibles y se puede armar otra Tabla de la Verdad experimental. Los resultados son muy similares a los del caso anterior.

Para estudiar que pasa si se cambia el valor de la prevalencia en la muestra, se puede comenzar por ejemplo eligiendo 400 fichas rojas y 3.600 blancas. Esto simula el llamado diseño de estudio del tipo Caso-control, donde la relación es de 1 caso cada 9 controles. O sea, una prevalencia simulada en esta muestra  $p = 0,1$ . Se cuentan las frecuencias halladas y se arma la Tabla de la Verdad experimental, con lo que se calculan todos los índices clínicos. Los resultados son que:

- a)  $S$ ,  $E$ ,  $Y$ ,  $LR+$  y  $LR-$  son iguales a los anteriores.
- b)  $A$  y valores predictivos cambian.

Si se van simulando distintas prevalencias  $p = 0,2; 0,3; \dots; 0,9$  las consecuencias son iguales a las dos anteriores (a) y (b). Se puede repetir esto tantas veces como se quiera hasta convencerse que:

Hay un grupo de índices que no varían no importa la manera en que se tome la muestra (los llamados *parámetros* o características internas del método), y hay otros que varían con la prevalencia de una manera muy similar a la explicada en el capítulo 4. Si se grafican los valores variables, se pueden encontrar las mismas curvas detalladas en los Gráficos 4.2 y 4.3. Esto demuestra experimentalmente el razonamiento efectuado antes.

*Caso 3:* Se separan las fichas en dos grupos, uno con los resultados positivos (15.500 fichas) y otro con los negativos (4.500 fichas) sin que importe el color de las mismas. Luego se eligen al azar 2.000 casos de positivos y otros tantos de negativos, de manera tal que el tamaño de la muestra sea  $N = 4.000$ . Es decir, que acá se ha estratificado la muestra en positivos y negativos, para simular el llamado estudio por cohortes. En este caso, la relación entre positivos y negativos elegida es  $P = 50\%$ . Se cuentan los resultados obtenidos para obtener las cuatro frecuencias, con las cuales se arma la Tabla de la Verdad experimental para este caso. Los resultados obtenidos son  $p \approx \pi = 0,5$ , y los mismos resultados obtenidos en el *Caso 1*.

Luego para ver la influencia del porcentaje de positivos y negativos seleccionados, se eligen una serie de valores  $P = 10\%, 20\%, \dots, 90\%$  y se vuelven a obtener los mismos resultados anteriores. De donde se puede concluir que: *No importa el porcentaje de positivos y negativos, los índices clínicos no se alteran*. En otras palabras, no interesa la manera de estratificar la muestra basada en los resultados del diagnóstico, nada cambia (Ver Cuadro 5.2 del capítulo anterior).

En conclusión, *la manera de seleccionar las muestras no afecta los índices clínicos paramétricos como S, E, IY y Likelihood Ratios. Los únicos afectados son la eficiencia y los valores predictivos por el efecto de la prevalencia simulada en los estudios del tipo caso-control*. Todo esto desde el punto de vista experimental, la explicación desde un punto de vista teórico se alcanza a través del teorema de Bayes como se explica a continuación.

## 6.4 Teoremas de Probabilidad Total y de Bayes

En el apartado 5.2. se puede ver el Diagrama de Venn para la partición de un suceso cualquiera  $A$  por un grupo de  $n$  sucesos  $C_i$ . y en la propiedad (5) vista en 5.3.1, se puede ver la probabilidad del suceso  $P(A)$ , expresada como la sumatoria de las probabilidades de sus trozos componentes  $P(A \cap C_i)$ , formados por la partición del suceso. Ahora bien, reemplazando estas probabilidades compuestas con la definición de probabilidad, se tiene la expresión del *Teorema de la Probabilidad Total*.

Sean  $n$  sucesos pertenecientes al mismo espacio muestral  $S$ , tales que particionan al suceso  $A$  (ver Gráfico 5.2) entonces se cumple que :

$$P(A) = \sum_{i=1}^n P(A \cap C_i) \quad \Leftrightarrow \quad \left\{ \begin{array}{l} (C_i \cap C_j) \cap A = \emptyset \quad \forall i \neq j \\ \bigcup_{i=1}^n C_i \supseteq A \end{array} \right.$$

Reemplazando  $P(A \cap C_i)$  por la definición resulta:

$P(A) = \sum_{i=1}^n P(C_i) \cdot P(A / C_i)$	<i>Teorema de Probabilidad Total</i>
---	--------------------------------------

Como una consecuencia del teorema anterior, haciendo substituciones adecuadas, se obtiene el otro teorema. En efecto, de la definición resulta :

$$P(C_i) \cdot P(A / C_i) = P(A \cap C_i) = P(C_i \cap A) = P(A) \cdot P(C_i / A)$$

Despejando,

$$P(C_i / A) = \{ P(C_i) \cdot P(A / C_i) \} / P(A)$$

Y reemplazando  $P(A)$  con el Teorema de Probabilidad Total, se obtiene el *Teorema de Bayes*

$$P(C_i / A) = \frac{P(C_i) \cdot P(A / C_i)}{\sum_{i=1}^n P(C_i) \cdot P(A / C_i)}$$

Donde :

$P(C_i)$  : son las llamadas probabilidades “a priori” por ser las que tienen los sucesos  $C_i$  antes de saber que ha ocurrido el suceso  $A$ .

$P(A \cap C_i)$  : son las probabilidades conjuntas

$P(A / C_i)$  : son las probabilidades condicionales

$P(C_i / A)$  : son las llamadas probabilidades “a posteriori” porque son las que tienen los sucesos  $C_i$ , luego de saber que ha ocurrido el suceso  $A$ .

Para ilustrar estos conceptos se desarrollan los ejemplos siguientes:

*Ejemplo 1)* En una sala de una clínica especializada solo se tratan tres tipos de enfermedades. Se sabe que en promedio ingresan un 50% de pacientes con la afección  $K$ , 30% con la enfermedad  $L$  y el resto con la afección  $M$  (datos obtenidos con las estadísticas de los últimos dos años). Realizando un relevamiento de historias clínicas se dedujo que un 70% de los ingresados con la enfermedad  $K$  se curan, mientras que para  $L$  y  $M$ , se obtuvieron 80% y 90% respectivamente. En la fecha, se dio de alta a un paciente: ¿Cuál es la probabilidad que se haya internado por la enfermedad  $K$  ?

- Sean  $C_1$  : pacientes con la enfermedad  $K$  . ∴  $P(C_1) = 0,5$  y la  $P(C / C_1) = 0,7$   
 $C_2$  : pacientes con la enfermedad  $L$  . ∴  $P(C_2) = 0,3$  y la  $P(C / C_2) = 0,8$   
 $C_3$  : pacientes con la enfermedad  $M$  . ∴  $P(C_3) = 0,2$  y la  $P(C / C_3) = 0,9$   
 $C$  : pacientes curados y dados de alta.  
 $C / C_i$  : pacientes curados y dados de alta que se internaron con la afección  $i$

Con los datos anteriores se puede armar la tabla del Tabla 6.1

**Tabla 6.1 : Cuadro bayesiano para resolución de problemas.**

Sucesos	A priori	Condicional	Conjunta	A posteriori
	(a)	(b)	(c)=(a) . (b)	(d)=(c)/ $\sum(c)$
$C_1$	0,5	0,7	0,35	0,4546
$C_2$	0,3	0,8	0,24	0,3117
$C_3$	0,2	0,9	0,18	0,2337
Totales	1		0,77	1

Las probabilidades *a priori*, se colocan en la primer columna de datos (a). Como verificación, el total de las mismas debe ser 1. Las probabilidades condicionales se colocan en la segunda (b). Luego se calculan las probabilidades conjuntas, multiplicando las dos anteriores entre sí. O sea, la columna (c) es igual a (a) .(b) La sumatoria de esta columna, es la probabilidad total de que el paciente se cure  $P(C) = 0,77$ . Finalmente, se pueden calcular las probabilidades *a posteriori*, dividiendo cada término de la columna (c) por la probabilidad total igual a 0,77. La respuesta a la pregunta del problema, se halla en el primer casillero de la columna (d). Esto es, la probabilidad de que el paciente dado de alta haya ingresado con la enfermedad **K** es de 45,46%.

*Ejemplo 2)* En una farmacia se sabe por experiencia que el 60% de su clientela es de sexo femenino. La estadística histórica muestra que un 4% de los hombres gasta más de \$100 en una compra, mientras que solo el 1% de las mujeres lo hace. Se elige al azar una compra mayor de \$100 y se desea saber cuál es la probabilidad de que el cliente sea mujer.

Definiendo los sucesos: **H** : hombre es  $P(H) = 0,4$       **C** : compra mayor a \$100  
**M** : mujer es  $P(M) = 0,6$

$$\text{La probabilidad pedida es: } P(M/C) = \frac{P(M) P(C/M)}{P(M) P(C/M) + P(H) P(C/H)}$$

$$P(M/C) = \frac{(0,6) (0,01)}{(0,6) (0,01) + (0,4) (0,04)} = 3/11 = 0,273$$

Hay un 27,3% de probabilidad que el cliente sea una mujer.

## 6.5. Diagnóstico y el Teorema de Bayes

Los cuatro casos posibles al efectuar un diagnóstico, basándose en la evidencia suministrada por el resultado de un análisis clínico, fueron presentados en el Tabla 5.1. (apartado 5.3.3). A la luz de los nuevos conceptos, conviene volver a revisar tal deducción, pero tratando de entender el significado de los índices clínicos, usando el concepto de la probabilidad condicional.

*VPP* : Se calcula con la fórmula del Teorema de Bayes. O sea, es una probabilidad bayesiana *a posteriori* y se puede imaginar como la probabilidad de diagnosticar al paciente en forma correcta un resultado positivo, sabiendo que el resultado de su análisis dio + Permite discriminar de entre todos los resultados positivos, la fracción de resultados correctos.

$$VPP = \frac{P(C_1 \cap TP)}{\sum_{i=1}^2 P(C_i \cap TP)} = P(C_1 / TP)$$

$$VPP = \frac{P(C_1) \cdot P(TP / C_1)}{P(C_1) \cdot P(TP / C_1) + P(C_2) \cdot P(TP / C_2)} = P(C_1 / TP)$$

*Sensibilidad* : es la probabilidad condicional  $P(TP / C_1)$  de que un paciente sea diagnosticado como enfermo, cuando se sabe que en realidad está enfermo.

*Prevalencia* : es la probabilidad de enfermarse  $P(C_1)$ . O sea, es la probabilidad de encontrar la enfermedad en la población.

*Sensibilidad . Prevalencia* : El producto de ambos índices es  $P(C_1 \cap TP)$ , una probabilidad conjunta que significa, la probabilidad de que un paciente este enfermo y al mismo tiempo su diagnóstico haya dado positivo. Difiere del concepto anterior, pues no se parte de un hecho condicionante conocido. Es la probabilidad de obtener un verdadero positivo  $P(VP)$ .

$$P(VP) = \text{Sensibilidad} \cdot \text{Prevalencia} = P(C_1 \cap TP) = P(TP / C_1) \cdot P(C_1)$$

*Especificidad* : Es la probabilidad condicional  $P(TS / C_2)$  de que un paciente sea diagnosticado como sano, cuando se sabe que en realidad está sano.

*Probabilidad de estar sano* : Es el complemento de la Prevalencia  $P(C_2) = 1 - P(C_1)$ . Además:

$$P(TP / C_2) = P(C_2 \cap TP) / P(C_2) = P(FP) / P(C_2) = (fp / N) / (TS / N) = fp / TS = (TS - vn) / TS$$

$$P(TP / C_2) = 1 - (vn / TS) = (1 - \text{especificidad}) : \text{Complemento de Especificidad}$$

$$\text{Entonces, } P(C_2) \cdot P(TP / C_2) = (1 - \text{especificidad}) (1 - \text{prevalencia})$$

$$\text{O bien, } P(C_1 \cap TP) = (1 - \text{Prevalencia}) (1 - \text{Especificidad}) = P(FP)$$

El Teorema de Bayes para los índices clínicos se puede escribir con:

$$VPP = \frac{\text{Prevalencia} \cdot \text{Sensibilidad}}{(1 - \text{Prevalencia})(1 - \text{Especificidad}) + \text{Prevalencia} \cdot \text{Sensibilidad}} = P(C_1 / TP)$$

$$VPP = S p / [(1-p) (1-E) + p S] = 1 / [1 + (O \cdot LR+)^{-1}] \text{ donde } O = p / (1 - p) \text{ y análogamente:}$$

$$VPN = 1 / [1 + (O \cdot LR-)] \text{ fórmulas más sencillas de recordar el Teorema de Bayes.}$$

Notar que los valores predictivos son función de la prevalencia, y reemplazando los valores del problema visto en el Gráfico 4.2 anterior, con  $S = 0,9$  y  $S = 0,75$ , se verifican teóricamente las dos curvas mostradas. Además, se verifican las relaciones experimentales halladas en 6.3.1, con este modelo teórico. O sea que el Teorema de Bayes puede ser comprobado en forma empírica.

## 6.5.1 Odds a posteriori

Aplicando el Teorema de Bayes a los nuevos índices clínicos, se pueden calcular los Odds a posteriori. Esto es, el Odds de la enfermedad luego de hacer el test clínico. Ya sea que los resultados encontrados modifiquen o no los Odds a priori disponibles. Esto se usa cuando hay un dilema en el diagnóstico a efectuarle al paciente y no se está seguro de cómo proceder. Es virtualmente imposible recordar la Sensibilidad y Especificidad de cada test clínico efectuado para hacer un diagnóstico. Sin embargo, cuando hay un dilema con el diagnóstico a efectuar, es muy útil usar la información de S; E; LR+; LR- y los Odds a priori como guía a seguir. En especial, cuando se hacen varios tests clínicos en secuencia y el Odds a posteriori del primero, se usa como a Odds a priori para la segunda prueba clínica y así sucesivamente. Un cuidadoso análisis de la información contenida en un Odds a priori, arroja luz sobre cual debería ser el paso siguiente hacia el diagnóstico. O sea, cuál debiera ser la prueba clínica a ordenar.

Como se vio en el capítulo anterior, el Odds de la enfermedad se calcula como el cociente entre la probabilidad de enfermedad y su complementaria. Como esta información se dispone al armar la tabla de decisión para el diagnóstico, se define como:

$$\text{Odds a priori} = TE / TS \text{ (para la enfermedad)}$$

El Odds a posteriori se calcula con el Odds del valor predictivo de positivos:

$$\text{Odds a posteriori} = VPP / (1 - VPP) = (p \cdot S / X) / [(1 - (p \cdot S / X))]$$

Donde  $X = p \cdot S + (1 - E)(1 - p)$ , simplificando queda

$$\text{Odds a posteriori} = p \cdot S / [X - (p \cdot S)] = p \cdot S / [p \cdot S + (1 - E)(1 - p) - p \cdot S]$$

Odds a posteriori =  $p \cdot S / [(1 - E)(1 - p)]$  y reagrupando resulta:

$$\text{Odds a posteriori} = \frac{\text{Sensibilidad} \times \text{Prevalencia}}{(1 - \text{Especificidad})(1 - \text{Prevalencia})} = \frac{\text{Sensibilidad}}{(1 - \text{Especificidad})} \times \frac{\text{Prevalencia}}{(1 - \text{Prevalencia})}$$

$$\text{Odds a posteriori} = LR+ \cdot \text{Odds a priori}$$

Concepto mucho más sencillo de recordar y usar que el Teorema de Bayes. El uso de los Odds a posteriori tanto en el caso de positivos como de negativos, se ejemplifica a continuación:

*Ejemplo 1)* Para diagnosticar una enfermedad en la arteria coronaria, un cardiólogo realiza un test de tolerancia a un ejercicio físico programado. De sus historias clínicas con resultados verificados escoge a 147 pacientes con problemas coronarios y los clasifica de acuerdo a su diagnóstico, junto con el estado real que se pudo verificar después. De esta forma, arma una Tabla de diagnóstico para ver que tan bien trabaja (datos tomados del trabajo de Simel). Los resultados fueron:

Test de tolerancia al ejercicio físico	Enfermedad Coronaria		
	Si	No	Total
(+)	73	9	82
(-)	28	37	65
Total	101	46	147

Se pueden calcular los principales índices clínicos con:

$$\begin{aligned} \text{Sensibilidad} &= tp / TE = 73 / 101 = 0,72 & \text{Prevalencia} &= TE / N = 101/147 = 0,69 \\ \text{Especificidad} &= tn / TS = 37 / 46 = 0,80 & \text{Eficiencia} &= (tp + tn) / N = (73+37)/147 = 0,75 \\ \text{LR+} &= \text{Sens.} / (1 - \text{Espec}) = 0,72 / 0,2 = 3,6 & \text{PPV} &= tp / T+ = 73 / 82 = 0,89 \\ \text{LR-} &= (1 - \text{Sens.}) / \text{Espec.} = 0,28/0,80 = 0,35 & \text{NPV} &= tn / T- = 37 / 65 = 0,57 \end{aligned}$$

*Ejemplo 2)* Luego de este trabajo, viene un paciente de 52 años de edad que tiene antecedentes familiares importantes de enfermedades coronarias y presenta los síntomas de una típica angina inducida por esfuerzo. Se desea saber si el test de tolerancia al ejercicio físico influirá en el diagnóstico. El cardiólogo estima en un 80% la sospecha de que tenga esta enfermedad y usando la información obtenida en el ejemplo anterior, calcula:

Su estimación de un 80% implica un Odds de 4 a 1 de que tenga la enfermedad. Y usando sus antecedentes vistos en el ejemplo anterior como referencia poblacional, deduce que:

. Si el paciente da (+) con el test, el Odds a posteriori será de  $4 \times 3,6 = 14,4$  su chance de estar enfermo. Esto implica que el Odds sube de 4 a 1 hasta 14,4 a 1 (o sea, de un 80% a un 93,5%).

. Si el paciente da (-) con el test, el Odds a posteriori será de  $4 \times 0,35 = 1,4$ . Esto implica que el Odds baja de 4 a 1 hasta 1,4 a 1 (o sea, de un 80% a un 58%).

Se concluye que un resultado (+) no influye mucho en el diagnóstico de enfermedad coronaria y además, un resultado (-) tampoco añade mucha más información. Por lo tanto, no conviene indicarle que se haga un test de tolerancia al ejercicio físico porque no ayuda mucho al diagnóstico. La probabilidad pre-test no varía demasiado y no agregará información importante.

*Ejemplo 3)* Otro hombre de 52 años de edad, robusto en apariencia tiene dolores en el pecho desde hace unos meses. No encuentra relación entre dolor y esfuerzo físico, además carece de los síntomas asociados. ¿Puede un test de tolerancia al ejercicio físico ser de utilidad para diagnosticar una enfermedad coronaria, al sospechar que el dolor es una señal de una angina no típica? Se sabe de los libros de texto que la prevalencia para una angina no típica es del 48%.

La estimación de un 48% implica a priori un Odds de 0,92 a 1 de que tenga una angina no típica. Los LR no varían porque Sensibilidad y Especificidad no dependen del paciente sino de la población tomada como referencia. Usando los antecedentes del ejemplo (1) anterior como referencia, se deduce que:

. Si el paciente da (+) con el test, el Odds a posteriori será de  $0,92 \times 3,6 = 3,3$ . Esto implica que el Odds sube de 0,92 a 1 hasta 3,3 a 1 (o sea, de un 48% a un 77%).

. Si el paciente da (-) con el test, el Odds a posteriori será de  $0,92 \times 0,35 = 0,32$ . Esto implica que el Odds baja de 0,92 a 1 hasta 0,32 a 1 ( o sea, de un 48% a un 24%).

Se concluye que un resultado (+) o (-) influye mucho en el diagnóstico de enfermedad coronaria. Por lo tanto, conviene indicarle que se haga un test de tolerancia al ejercicio físico porque sería de gran ayuda para el diagnóstico. En este caso la información pre-test resultará enriquecida y se ve la conveniencia de efectuar el ejercicio a la tolerancia física en ese paciente.

## 6.5.2 Simplificación del Teorema de Bayes

Una forma más sencilla de recordar el Teorema de Bayes es usando las relaciones:

$$NPV = 1 / \{1 + [ (LR- \cdot O) ]\} \quad \text{y} \quad PPV = 1 / \{1 + [1 / (LR+ \cdot O) ]\}$$

En la bibliografía clínica (JAMA) la recomendación actual es usar la metodología siguiente para calcular las probabilidades Post-test (a posteriori) a partir de las probabilidades Pre-test (a priori):

*Paso 1:* En forma subjetiva el clínico asigna una probabilidad a priori, luego de estudiar al paciente, de acuerdo a su experiencia en este tipo de casos. Esto es, la probabilidad sospechada que el mismo esté enfermo (Pre-test), que en el Ejemplo 2 anterior era 80%.

*Paso 2:* Luego transforma esta probabilidad en Odds ( $0,8/0,2 = 4$ ). Entonces tiene el llamado Pre-test Odds, que en el ejemplo vale 4 : 1

*Paso 3:* A continuación aplica el Teorema de Bayes para obtener los Odds a posteriori, denominados Post-test Odds, que para el ejemplo será: 14,4

*Paso 4:* Transforma este Odds en probabilidad con la relación  $p = O / (O + 1)$  resultando el valor de la llamada probabilidad Post-Test, que en el ejemplo es: 93%.

*Paso 5:* De acuerdo al resultado obtenido en el paso anterior, decide si le conviene, o no, hacerle el test al paciente y obtiene su primera conclusión. En el ejemplo, como la probabilidad sube de 80% a 93,5% decide que no ayudaría en mucho el hacerle la ergometría al paciente.

*Paso 6:* Para el caso en que el paciente de (-) en el test, aplica nuevamente el Teorema de Bayes para obtener el Odds a posteriori en el caso de los negativos. Es decir, multiplica el Pre-Test Odds de estar enfermo (4 : 1), por el Likelihood ratio de negativo (0,35) y obtiene  $1,4 = 4 \times 0,35$ .

*Paso 7:* Transforma este Odds en probabilidad y obtiene la probabilidad a posteriori de no estar enfermo (Post-test para el caso de negativo) de un 53%.

*Paso 8:* De acuerdo al resultado anterior decide si le conviene hacerle el test al paciente, en el caso de que este estuviera sano. Como la probabilidad antes de hacerle el test es del 80% y luego de hacerle la ergometría baja al 53%. Le puede ayudar mucho el hacerla, pero por otro lado sería riesgosa si el paciente estuviera enfermo como él sospecha (80%). Concluye finalmente, que no es conveniente efectuar la ergometría.

Como puede verse no es un procedimiento tan sencillo de seguir, a pesar de la simplificación hecha con el Teorema de Bayes. Un estudio realizado en Suiza en marzo del 2002 a más de 600 concurrentes a un curso de especialización, mostró que solo 30% de ellos había oído de este procedimiento y tan solo un 10% lo aplicaba en su práctica diaria. No parece lógico pedirle al clínico que tenga que hacer estos cálculos cada vez que estudie a un paciente. Por ese motivo, la cátedra propuso un método gráfico, que consta de un solo paso, y que no requiere del conocimiento de conceptos tales como Pre y Post-Odds de sanos y enfermos, ni del Teorema de Bayes.

Esta idea se basa en dos gráficos, donde se muestra la variabilidad de las probabilidades a posteriori, partiendo de la probabilidad a priori (un gráfico para el caso de positivos y otro para el de negativos). El concepto se basa en las siguientes relaciones:

*Caso positivo:* La probabilidad Pre-test y Post-test se relacionan con:

$$\text{Odds a posteriori (+)} = \text{LR+} \times \text{Odds a priori (+)} = [S / (1-E)] \times [p / (1-p)] = v_p / f_p \quad \text{O sea,}$$

$$\text{Odds a posteriori (+)} = \text{PPV} / (1 - \text{PPV})$$

De donde se puede deducir que la Post-test probabilidad de estar enfermo cuando el resultado del test fue positivo es simplemente el VPP.

*Caso negativo:* La probabilidad pre-test y Post test se relacionan con:

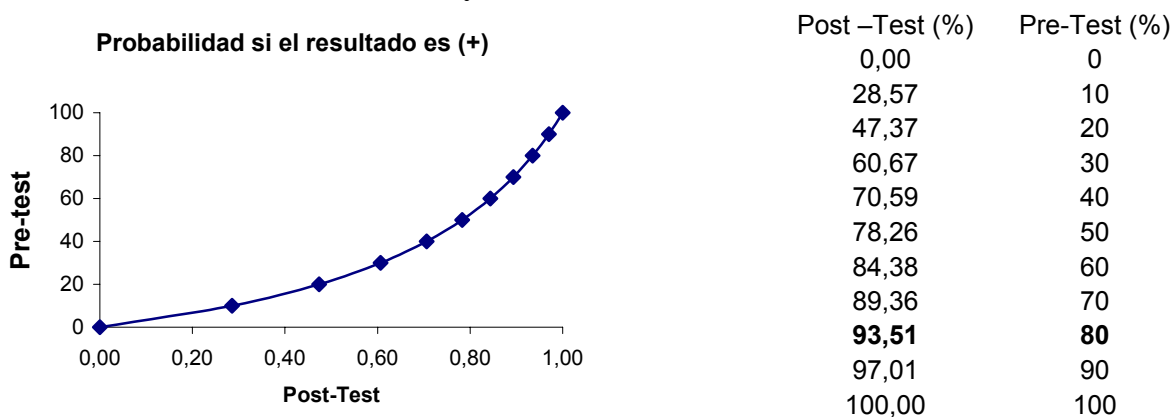
$$\text{Odds a posteriori (-)} = \text{LR-} \times \text{Odds a priori (+)} = [(1 - S) / E] \times [p / (1-p)] = f_n / v_n$$

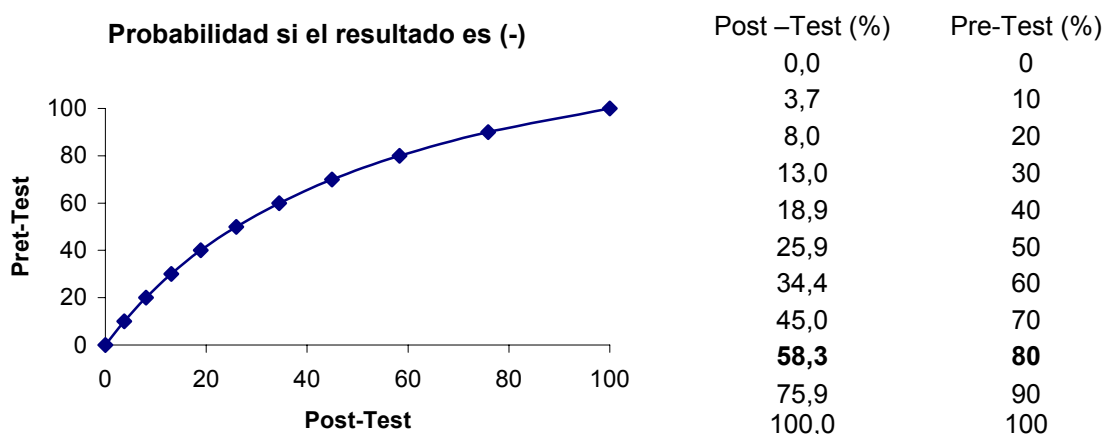
$$\text{Odds a posteriori (-)} = (1 - \text{NPV}) / \text{NPV}$$

De donde se puede deducir que la Post-test probabilidad de estar sano cuando el resultado del test fue negativo es simplemente (1 - NPV).

Entonces, recordando la simulación para verificar experimentalmente el Teorema de Bayes, se pueden graficar las variabilidades de los valores predictivos con una probabilidad cualquiera (tomada como la prevalencia del estudio antes, y que ahora es sencillamente la probabilidad asignada por el clínico al paciente en forma subjetiva)

**Gráfico 6.2. Probabilidades Pre-test y Post-test**





Se puede observar de estos gráficos, que entrando con la sospecha del clínico del ejemplo anterior (80%) se obtienen directo ambas probabilidades Post-test, sin hacer ninguna cuenta. Para obtener este par de gráficos en forma rápida, la cátedra ha desarrollado un programa en Excel para computadoras, donde entrando con los valores de sensibilidad y especificidad se pueden obtener directamente ambos gráficos, y luego imprimirlos para tenerlos disponibles.

Sin embargo, a pesar de esta nueva simplificación del problema, no es razonable esperar que cada clínico tenga a mano el par de gráficos para decidir si le efectúa la paciente el test. Por lo tanto, se presenta a continuación una manera simplificada de hacer todo esto de manera sencilla, rápida y fácil de recordar:

La idea es jugar con el par de gráficos para derivar criterios clínicos generales, que a modo de guía faciliten la tarea de ver si conviene hacer un test clínico para un caso individual. A modo de ejemplo se puede usar el caso de la ergometría anterior:

*Paso 1:* Se toma una probabilidad subjetiva alta de que el paciente esté enfermo cuando viene a la consulta, como ser del 80%. Mirando en los gráficos se obtienen las dos probabilidades posteriores al test asociadas de un 93,5% si en realidad está enfermo, y de un 58,3% si estaba sano.

*Paso 2:* Se toma una probabilidad subjetiva baja respecto a que el paciente esté enfermo, como ser del 20%. Ahora las probabilidades posteriores al test son del 47,4% si estaba enfermo y del 8% si estaba sano.

*Paso 3:* Se toma una probabilidad subjetiva intermedia, como ser del 50% (esto significa en el fondo que el clínico no tiene una sospecha clara en ningún sentido, sería como el tirar una moneda al aire, o bien usar estos casos para cuando no se tiene mucha idea acerca del estado real del paciente). Mirando en los gráficos se puede obtener una probabilidad Post-test del 78,3% si en realidad estaba enfermo, y del 26% si estaba sano.

De estos 3 pasos se pueden deducir unas reglas generales, a modo de protocolo, para saber como proceder en este tipo de casos. Por ejemplo, las reglas generales pueden ser:

- 1) Si la sospecha de que el paciente está enfermo es alta, no conviene hacerle la ergometría.
- 2) Si la sospecha de que esté enfermo es baja, conviene efectuar la ergometría.
- 3) Si no se tiene una idea clara del estado del paciente conviene hacer el test.

## 6.6 Estudio de la independencia en tablas de riesgo

En el primer capítulo se vieron los estudios de riesgo, donde el objeto era ver el efecto de cierto factor (de exposición, de protección, etc.) sobre la condición buscada. Es el típico estudio epidemiológico que se clasificaba en: Caso-control, Cohortes o RCT. En todos ellos se usaba una tabla de 2 x 2 para volcar los datos del estudio como la siguiente:

**Tabla 6.2.** Tabla de riesgo: datos observados

Factor	Condición observada		Total
	Enfermos	Sanos	
Expuesto	$a$	$b$	$a + b$
No expuesto	$c$	$d$	$c + d$
Total	$a + c$	$b + d$	$N$

Como puede verse, se trata de una partición de la población usada en el estudio en cuatro conjuntos básicos, mutuamente independientes y colectivamente exhaustivos. Esta tabla está construida con los datos observados. Haciendo el siguiente supuesto:

- . El factor de riesgo es independiente de la condición observada. O bien,
- . El factor y la condición no están asociados.

Aquí se habla de *asociación estadística*, que no es una asociación del tipo *causa-efecto*, sino que podría llegar a serlo. Es muy importante tener muy en claro esta diferencia, porque es común ver que se confunde una cosa con la otra.

Entonces, bajo ese supuesto se pueden calcular las frecuencias esperadas, usando la teoría de que hay independencia estadística, como se muestra en la tabla siguiente:

**Tabla 6.3.** Tabla de riesgo: datos esperados teóricamente si hay independencia

Factor	Condición observada		Total
	Enfermos	Sanos	
Expuesto	$\frac{(a+c)(a+b)}{N}$	$\frac{(b+d)(a+b)}{N}$	$a + b$
No expuesto	$\frac{(a+c)(c+d)}{N}$	$\frac{(b+d)(c+d)}{N}$	$c + d$
Total	$a + c$	$b + d$	$N$

Notar que cada la frecuencia esperada de cada celda, se saca con el producto de sus respectivos totales marginales, dividido el tamaño muestral total  $N$ .

La manera de demostrar esto es como sigue, sean los sucesos siguientes:

$E$  : la condición de enfermo.  
 $S$  : la condición de sano.

$X$  : la condición de expuesto.  
 $Y$  : la condición de no-expuesto

Los conjuntos siguientes son los cuatro casos posibles:

$(E \cap X)$ : el conjunto de los enfermos que estuvieron expuestos, cuya frecuencia observada es  $a$   
 $(S \cap X)$ : el conjunto de los sanos que estuvieron expuestos, cuya frecuencia observada es  $b$   
 $(E \cap Y)$ : el conjunto de los enfermos que no estuvieron expuestos, con frecuencia observada  $c$   
 $(S \cap Y)$ : el conjunto de los sanos que no estuvieron expuestos, cuya frecuencia observada es  $d$

Sus probabilidades se pueden estimar con:

$$P(E \cap X) = a^* / N \qquad P(S \cap X) = b^* / N$$

$$P(E \cap Y) = c^* / N \qquad P(S \cap Y) = d^* / N$$

Donde el asterisco denota que son valores teóricos y no observados. Luego, para que haya independencia se tiene que cumplir la condición:

$$P(E \cap X) = P(E) \cdot P(X) = [(a + c) / N] \cdot [(a + b) / N] = a^* / N$$

$$P(E \cap Y) = P(E) \cdot P(Y) = [(a + c) / N] \cdot [(c + d) / N] = c^* / N$$

$$P(S \cap X) = P(S) \cdot P(X) = [(b + d) / N] \cdot [(a + b) / N] = b^* / N$$

$$P(S \cap Y) = P(S) \cdot P(Y) = [(b + d) / N] \cdot [(c + d) / N] = d^* / N$$

Despejando y simplificando se obtienen las frecuencias esperadas de la Tabla 6.3 anterior. Por ejemplo, si las frecuencias observadas fueron:

Factor	Condición observada		Total
	Enfermos	Sanos	
Expuesto	20	80	100
No exp.	50	150	200
Total	70	230	300

Las frecuencias esperadas si hay independencia serán:

Factor	Condición observada		Total
	Enfermos	Sanos	
Expuesto	23,3	76,7	100
No exp.	46,7	153,3	200
Total	70	230	300

## 6.7 Probabilidad hipergeométrica

Se trata de un caso muy frecuente en los muestreos de aceptación donde se debe decidir si el comprador acepta comprar un lote de mercadería, a partir del análisis de una pequeña muestra del lote. Los profesionales que trabajan en el sector Control de Calidad de la industria farmacéutica, química, cosméticos, alimentos, etc. suelen aplicarla desde otro punto de vista. Allí interesa saber si se da por bueno un lote de producción, a partir del control de una muestra del mismo. O sea, el control no lo hace el consumidor sino el fabricante. En ambos casos, la idea central es la misma, no se pueden revisar todas las piezas fabricadas, ya sea porque llevaría mucho tiempo, o porque el ensayo es destructivo, o porque es muy caro, etc.

Existen Normas IRAM para el control de calidad, que se usan como una solución de compromiso entre los intereses encontrados de comprador y vendedor. Por un lado, el comprador quiere minimizar su riesgo de comprar un lote malo aumentando el tamaño de la muestra a revisar. Y por el otro lado, el vendedor desea minimizar el costo del control revisando la menor cantidad posible de piezas. Entonces, la norma ayuda a definir el tamaño de la muestra a controlar, que es el cuello de botella de toda negociación de esa naturaleza, a partir de riesgos de compra y venta aceptables para ambas partes. Tal riesgo se calcula a partir de la probabilidad hipergeométrica.

Sea un lote de producción de tamaño  $N$  que se lleva al laboratorio de Control de Calidad para su revisión, antes de lanzarlo a la venta. El profesional a cargo, sabe por su larga experiencia, el porcentaje de piezas falladas que encontrará al efectuarle una serie de análisis que permitan dos conclusiones: pieza aceptada o pieza rechazada. Multiplicando tal porcentaje de fallas  $p$  por el tamaño del lote  $N$ , obtiene el número esperado  $R$  de piezas rechazadas. De acuerdo a las normas vigentes, se elige una muestra control de tamaño  $n$ . ¿ Cuanto vale la probabilidad de encontrar exactamente  $r$  piezas falladas en la muestra control ? Tal pregunta se responde con la *probabilidad hipergeométrica* denotada con  $P_H(r/n, N, R)$ . Como toda probabilidad, esta también resulta el cociente entre el número de casos favorables y el número total de casos posibles.

El numerador son todos los casos donde al revisar una pieza se la rechaza. Hay que hacer un conteo ordenado, pues habrá  $C(R, r)$  combinaciones diferentes de encontrar las  $r$  piezas falladas de la muestra, dentro de las  $R$  piezas totales del lote de producción. Pero a su vez, esto ocurre con cada una de las formas posibles de encontrar la  $(n-r)$  piezas buenas, restantes de la muestra, de entre todas las piezas buenas del lote  $(N-R)$ . O sea,  $C(N-R, n-r)$ . En síntesis, el número total de casos donde hay rechazo se obtendrá con el producto de ambos valores, esto es:  $C(R, r) \cdot C(N-R, n-r)$ . El denominador esta dado por todas las combinaciones posibles que existen al tomar  $n$  elementos, de un total de  $N$  elementos del lote:  $C(N, n)$

$$P_H(r/n, N, R) = \frac{C(R, r) \cdot C(N-R, n-r)}{C(N, n)}$$

*Ejemplo 1)* Se fabricó un lote de 1000 ampollas conteniendo un medicamento liofilizado. Se sabe que un 0,2 % tiene problemas con el cierre hermético del vidrio. Se elige al azar una muestra control de 10 ampollas. ¿ Cuánto vale la probabilidad de encontrar una fallada ?

$$P_H ( r=1 / n=10, N=1000, R=2) = C(2, 1) C ( 998 , 9 ) / C (1000 , 10 )$$

$$P_H ( r=1 / n=10, N=1000, R=2) = \frac{2 ( 998 )! ( 990)! 10!}{(989)! 9! 1000!} = \frac{2 \cdot 998! \cdot 990 \cdot 989! \cdot 10 \cdot 9!}{989! \cdot 9! \cdot 1000 \cdot 999 \cdot 998!}$$

$$P_H ( r=1 / n=10, N=1000, R=2) = 0,0198$$

La probabilidad de no encontrar ninguna fallada será

$$P_H ( r=0 / n=10, N=1000, R=2) = C(2, 0) \cdot C ( 998 , 10 ) / C (1000 , 10) = 0,98$$

En síntesis, hay un 98% de probabilidades de no encontrar ampollas falladas y 1,98% de encontrar una sola fallada. Entonces habrá:  $(100-98-1,98)\% = 0,02\%$  de encontrar dos o más falladas.

*Ejemplo 2)* El mismo caso de sacar tres barajas de espadas del mazo, visto antes y resuelto con el concepto de probabilidad condicional, ahora se puede resolver con:

$$P_H ( r=3 / n=3, N=40, R=10) = C(10, 3) \cdot C ( 30 , 0 ) / C ( 40 , 3) = \frac{10! 37! 3!}{40! 7! 3!}$$

$$P_H ( r=3 / n=3, N=40, R=10) = ( 10 \cdot 9 \cdot 8 ) / ( 40 \cdot 39 \cdot 38) = 0,012$$

*Ejemplos 3)* Un médico suministra un fármaco a siete enfermos de similares características. Seis de ellos experimentan una gran mejoría, en cambio el restante permanece igual. Por otra parte, durante todo el tiempo transcurrido se ha controlado a un grupo de cinco pacientes, de características similares a los otros sometidos al tratamiento. La diferencia es que, a este grupo control se le ha suministrado un placebo. Resultó que sólo uno mejoró, mientras los otros cuatro permanecieron iguales. El problema es calcular si la mejoría observada en el grupo tratado, se debe al azar o a la acción del medicamento.

La probabilidad hipergeométrica puede emplearse en estos casos para contestar tal pregunta. En efecto, tomando en conjunto total de pacientes ( $N=12$ ), de los cuales ( $n=7$ ) fueron tratados; los resultados finales fueron ( $R=7$ ) han mejorado y entre éstos ( $r=6$ ) fueron los medicados, mientras que al restante ( $R-r =1$ ) se le había suministrado un placebo inocuo para tratar la enfermedad. Se puede presentar lo mismo en forma esquemática con el diagrama del Gráfico 6.3. Asumiendo que los pacientes fueron distribuidos al azar entre ambos grupos, para evitar cualquier tipo de tendencia, y tratando de mantener todas sus características lo más parecidas posibles para no introducir fluctuaciones extrañas. Se puede asumir, que se trata de un problema de “ n extracciones al azar sin reemplazamiento, de un grupo de tamaño N”.

**Gráfico 6.2 Diagrama esquemático para el problema 3.**

Con mejoría	Sin mejoría	Total
-------------	-------------	-------

Con fármaco (medicados)	6 = r	1	7 = n
Sin fármaco (placebo)	1	4	5
Total	7 = R	5	12 = N

Entonces, la probabilidad de ocurrencia por azar en una tabla del tipo 2 x 2 como las vistas en índices clínicos, de los sucesos esquematizados en el diagrama, se calcula con:

$$P_H ( r=6 / n=7, N=12, R=7) = C(7, 6) \cdot C(5, 1) / C(12, 7) = 0,0442$$

La probabilidad de que no ocurra por azar, sino por la acción del medicamento será el complemento de la anterior, o sea 0,9558. Se tiene entonces un 95,58% de probabilidad de que el medicamento (o tratamiento) sea eficaz y un 4,42% que no lo sea. Se puede afirmar que es muy improbable, que la mejoría de los pacientes sea casual. Sin embargo, todavía esto no es una *prueba científica* con evidencia estadística. Para que así ocurra, se deben plantear tests de hipótesis como se verá en capítulos posteriores.

## 6.8 Problemas propuestos

1) Marcar la respuesta correcta a cada una de las afirmaciones siguientes, o completar la frase:

- |  |   |   |
|--|---|---|
| 1) La regla del producto se puede aplicar en todos los casos.                                  | V | F |
| 2) La probabilidad de sacar tres caras al lanzar al aire tres monedas es 1/8.                  | V | F |
| 3) Dos sucesos son independientes si la ocurrencia de uno no afecta al otro.                   | V | F |
| 4) Dos sucesos independientes no tienen ningún elemento en común.                              | V | F |
| 5) Si tres sucesos son independientes, también lo son todos los pares posibles.                | V | F |
| 6) Si todos los pares posibles son independientes, también lo son si se toman de a tres.       | V | F |
| 7) Dos sucesos son excluyentes si no tienen ningún elemento en común.                          | V | F |
| 8) Es lo mismo el concepto de independencia que el de exclusión.                               | V | F |
| 9) En sucesos condicionales, la probabilidad de uno se afecta si se sabe que ocurrió el otro.  | V | F |
| 10) Exclusión e independencia, son casos particulares de la Condicionalidad.                   | V | F |
| 11) El Teorema de Bayes se puede aplicar si un suceso, es particionado por otros.              | V | F |
| 12) Ídem para el Teorema de Probabilidad Total.  | V | F |
| 13) Los Índices Clínicos se relacionan entre sí, a través del Teorema de Bayes.                | V | F |
| 14) Algunos Índices Clínicos se pueden interpretar como probabilidades condicionales.          | V | F |
| 15) En los muestreos de aceptación se emplea la probabilidad Hipergeométrica.                  | V | F |
| 16) La Normas IRAM el Control de Calidad aplican la probabilidad Hipergeométrica.              | V | F |
| 17) Si el producto de las probabilidades de dos sucesos es igual a la conjunta, son .....      |   |   |
| 18) Si la probabilidad condicional de dos sucesos es igual a la del producto de ambas, hay.... |   |   |
| 19) El diagrama del árbol se usa para visualizar .....   |   |   |
| 20) La sensibilidad es la capacidad de un test de detectar como positivos a los enfermos       | V | F |
| 21) La especificidad es la capacidad de un test de detectar como negativos a los sanos         | V | F |
| 22) Sensibilidad, especificidad, Youden y los Likelihood ratio no varían con la prevalencia.   | V | F |
| 23) Se puede mostrar experimentalmente lo anterior.  | V | F |

- 24) La eficiencia y los valores predictivos varían con la prevalencia. **V F**  
 25) La forma de extraer las muestras de la población altera a los índices clínicos. **V F**  
 26) Para un estudio por cohortes se hace un muestreo basándose en la exposición. **V F**  
 27) Para un estudio caso-control se hace un muestreo basándose en la condición buscada. **V F**  
 28) Explicar como hacer un experimento, el estilo Bernoulli, para verificar lo anterior .....  
 29) Expresar el Teorema de Bayes en forma más simplificada. ....  
 30) No hay relación entre el teorema de probabilidad total y el de Bayes. **V F**  
 31) Las probabilidades Pre-test se relacionan con las Post-test. **V F**  
 32) Los Odds a priori multiplicados por el LR+ dan los Odds a posteriori en sanos. **V F**  
 33) Se pueden transformar las probabilidades en Odds y viceversa. **V F**  
 34) La probabilidad a posteriori en el caso de enfermos es igual al VPP. **V F**  
 35) La probabilidad a posteriori en el caso de sanos es igual al VPN. **V F**  
 36) Con dos gráficos se pueden resumir los 8 pasos recomendados por JAMA para Post-Test. **V F**  
 37) Es mejor diseñar un protocolo general que usar los gráficos en cada paciente. **V F**  
 38) La independencia estadística significa: .....  
 39) En una tabla de riesgo la asociación estadística significa: .....  
 40) Si hay asociación estadística, hay una relación del tipo causa-efecto. **V F**  
 41) Si hay una relación causa-efecto, tiene que haber una asociación estadística. **V F**  
 42) Los valores de frecuencias esperadas, bajo el supuesto de independencia, se calculan con: .....  
 43) El producto de los totales marginales dividido N es la frecuencia esperada de una celda. **V F**  
 44) Las tablas de independencia 2 x 2 son un caso general de la tabla de riesgo. **V F**  
 45) Las tablas de riesgo pueden llegar a considerarse como tablas diagnósticas **V F**

2) La clasificación por edades de los profesionales de una zona de salud es:

Servicios	Menor de 25	Entre 26 y 30	Entre 31 y 35	Mayor de 35	Total
Médicos	0	5	25	75	105
Bioquímicos	20	30	35	35	120
Alimentación	3	6	6	10	25
Registración	7	15	8	12	42
Enfermeros	200	375	442	203	1220
Farmacéuticos	1	12	8	3	24
Radiología	4	10	19	12	45
Terapeutas	5	25	15	10	55
Otros servicios	20	35	50	25	130
Total	260	513	608	385	1766

a) Calcular las probabilidades para cada uno de los cuatro grupos de edades.

b) Lo mismo para los nueve tipos de servicios.

c) Si se escoge una persona al azar. Averiguar la probabilidad de que:

- Sea un médico de más de 35 años
- Sea un enfermero menor de 25 años.
- Sea un farmacéutico de 28 años.

c1) Sabiendo que tiene 33 años, cuanto vale la probabilidad de que sea:

- de Radiología
- de otros servicios

c2) Sabiendo que es un bioquímico, cuanto vale la probabilidad de que tenga:

- más de 35 años.
- 30 años.
- 21 años

d) Si se escogen dos personas al azar, cuanto vale la probabilidad de que:

- Uno sea médico y el otro bioquímico.

- Los dos sean farmacéuticos

e) Verificar el Teorema de Probabilidad Total en los puntos (a) y (b)

3) Una compra en una farmacia fue de 7 tabletas de aspirinas y 3 paquetes de algodón. Si se saca de la bolsa de a uno, averiguar la probabilidad de que se saquen dos tabletas y la tercera sea un algodón.

4) Calcular los índices clínicos para una prueba de laboratorio, que en un estudio de 400 personas arrojó los resultados siguientes:

Test Clínico	Enfermedad		Total
	Si	No	
(+)	30	190	220
(-)	20	160	180
Total	50	350	400

Con los valores obtenidos de S y E, simular una prevalencia variable entre  $p = 0,1; 0,2; \dots; 0,9$  para obtener las curvas que muestren la variabilidad de la eficiencia y los valores predictivos con la prevalencia. Verificar si se cumplen las relaciones teóricas expuestas en el cuadro resumen 5.2.

5) Resolver la siguiente sopa de letras:

B	R	N	O	R	A	L	I	N	A	C	O	N	H	D	O	X	A	G	U	E	S	W	O	U	M	T	R	A	X
X	B	S	C	H	Y	N	A	S	V	A	C	X	Q	I	R	F	O	I	T	A	L	Q	U	P	Y	O	L	X	S
E	F	Y	O	J	L	C	H	T	M	V	B	C	W	A	E	X	C	L	U	Y	E	N	T	E	S	X	Ñ	L	A
R	C	B	N	I	V	N	U	I	Y	R	A	I	C	N	E	U	C	E	R	F	R	E	S	A	X	P	L	S	R
O	V	R	D	D	M	H	R	J	N	M	T	R	W	A	D	E	S	L	A	X	V	N	R	O	D	E	D	Y	T
J	W	O	I	E	L	P	S	U	R	E	S	T	O	C	A	S	T	I	C	O	Ñ	K	D	P	I	S	M	S	A
I	T	E	C	Q	C	X	O	T	W	D	F	Q	B	F	J	L	S	H	L	J	P	F	I	I	Q	E	D	I	V
O	S	S	I	S	U	C	C	R	F	N	T	O	T	N	V	A	C	E	C	Z	O	D	E	C	A	T	A	L	S
U	I	T	O	O	I	S	I	E	P	S	I	P	Y	Q	C	C	N	L	I	C	A	S	O	T	T	N	D	S	Ñ
G	S	A	N	B	N	V	T	O	I	O	M	R	K	R	X	C	Z	A	F	T	I	T	T	O	P	E	I	K	N
N	D	L	A	D	R	E	S	T	R	U	T	O	R	E	S	N	I	O	I	I	N	S	I	O	A	I	S	L	E
A	O	A	L	C	U	S	O	I	T	R	I	D	Z	F	R	E	D	M	U	R	R	M	D	A	D	D	E	V	S
M	S	A	S	M	Z	X	N	R	D	I	N	U	L	E	S	S	C	V	I	G	O	I	A	L	O	N	C	O	T
B	K	U	S	S	R	O	G	T	W	S	W	C	L	R	A	L	A	S	O	N	O	L	L	A	R	E	S	C	G
E	A	J	Ñ	M	G	P	A	O	B	R	I	T	E	O	R	E	M	A	R	I	B	L	O	S	I	P	E	S	U
I	G	Y	I	N	A	N	I	A	S	A	D	O	O	U	M	I	L	A	A	D	M	I	S	T	R	E	U	I	Z
D	U	G	E	K	A	Ñ	D	C	I	N	A	A	U	L	A	D	I	M	U	A	U	B	Ñ	T	A	D	M	V	I
A	S	R	I	S	B	D	G	O	J	C	E	Ñ	F	U	L	S	I	R	A	R	S	O	T	A	R	N	S	E	R
N	Q	U	A	R	T	H	I	P	E	R	G	E	O	M	E	T	R	I	C	A	Z	T	O	R	T	I	P	L	A
K	I	B	W	D	A	D	I	L	I	B	I	S	N	E	S	N	G	H	D	A	Ñ	O	S	C	A	R	I	O	D

a) El tipo de probabilidad más general de todas.

b) Teorema de ..... es el que permite relacionar los índices clínicos entre sí.

c) La probabilidad usada en los muestreos de aceptación.

- d) Dos sucesos cuya probabilidad conjunta es igual al producto de sus probabilidades individuales.
- e) Dos sucesos cuya probabilidad conjunta es nula.
- f) Proceso conformado por una sucesión finita de experimentos, cada uno con probabilidad finita.
- g) La regla del .... permite calcular la probabilidad conjunta de dos sucesos independientes.
- h) ..... de Probabilidad Total se aplica cuando un grupo de sucesos particiona a otro.
- i) Conclusiones de los médicos cuando examinan a un paciente.
- j) La probabilidad de sacar cara es la ..... de la probabilidad del suceso seguro.

6) Usando los datos del ejercicio (4) anterior, e imaginándolos como las frecuencias observadas en el experimento, se pide:

- a) Calcular las 4 frecuencias esperadas teóricamente si hay independencia total entre el test clínico y el hecho de que exista o no la enfermedad.
- b) Recalcular los índices clínicos obtenidos en el ejercicio 5 con los datos teóricos bajo el supuesto de independencia.
- c) Discutir los resultados obtenidos y explicar como se puede relacionar la independencia estadística con la “independencia clínica”.

7) En un estudio epidemiológico se determinaron los datos siguientes:

		Enfermedad		
		Si	No	Total
Exposición a una fuente de contagio	Si	50	150	200
	No	50	150	200
Total		100	300	400

- a) Haciendo el supuesto de que la enfermedad no tiene nada que ver con la fuente de contagio (también llamado factor de riesgo), esto es que son estadísticamente independientes. Calcular las cuatro frecuencias esperadas.
- b) Calcular la Prevalencia en el caso de frecuencias esperadas y también en el caso de las frecuencias observadas. ¿ Qué conclusión puede sacarse ?
- c) Calcular el odd de los enfermos y de los que no enfermaron entre los pacientes expuestos.
- d) Calcular el odd de los enfermos y de los no enfermos entre los pacientes no expuestos.

8) Se pide repetir el ejemplo anterior pero con estos nuevos datos:

Enfermedad		
Si	No	Total

---

Exposición a una fuente de contagio	Si	150	50	200
	No	150	50	200
	Total	300	100	400

9) Comparar los resultados de los ejercicios 7 y 8 para sacar algunas conclusiones entre la independencia estadística y la realidad.

10) Explicar la diferencia entre variabilidad de los índices clínicos debidas a la prevalencia y la debida a la elección de un punto de corte en un método clínico.

11) Se efectuaron dos estudios epidemiológicos con un total de 400 individuos. De entre todos los casos disponibles, se seleccionaron 200 enfermos al azar y otros tantos sanos. Efectuando un estudio retrospectivo se separaron los casos que estuvieron expuestos a la fuente de contagio de los otros, y las frecuencias encontradas fueron las siguientes:

		Enfermedad		
		Si	No	Total
Exposición a una fuente de contagio	Si	50	50	100
	No	150	150	300
	Total	200	200	400

Se pide calcular:

- a) El Odds Ratio
- b) El riesgo relativo

12) En un estudio similar al anterior se encontraron los datos siguientes:

		Enfermedad		
		Si	No	Total
Exposición a una fuente de contagio	Si	150	150	300
	No	50	50	100
	Total	200	200	400

Se pide calcular los índices OR y RR y compararlos con el caso anterior.