8

Funciones de probabilidad

En este capítulo se tratan las bases mínimas para comprender las distribuciones de probabilidad. Es un tratamiento matemático un poco más formal que el visto hasta ahora. Si bien no se requiere un nivel matemático avanzado para leer este capítulo, se recomienda a los lectores que no tengan aprobados cursos de análisis matemático, que eludan los desarrollos y se concentren solo en los conceptos tratados.

8.1 Fenómenos aleatorios en Bioquímica y Farmacia

Tanto en Bioquímica como en Farmacia y en otras ciencias se busca en forma sistemática la adquisición de conocimientos, cada una aplicada a su campo específico. Esta búsqueda de la verdad la efectúan a través del intelecto humano con la realización de experimentos que ayuden en la tarea. Los científicos reúnen sus hechos mediante una cuidadosa observación de los fenómenos que ocurren, los agrupan, los clasifican y los tratan de explicar a través de otros hechos que conozcan. Con las teorías ya adquiridas se intenta explicar un nuevo fenómeno analizado. Cuando no se puede, entonces se trata de elaborar una nueva teoría que se ajuste a las nuevas observaciones realizadas. Pero en ambos casos siempre deben verificar las teorías propuestas a través de experimentos, cuidadosamente planeados y controlados. Esto implica un trabajo permanente de comparación entre los valores observados y los valores esperados teóricamente. De eso se ocupan los tests de hipótesis estadísticos. Y para llegar allí, lo primero es desarrollar modelos matemáticos como se presentan en este apartado.

Con todo ese trabajo se puede elaborar una imagen que explique los fenómenos observados. Luego con la información proveniente de los experimentos se puede ir corrigiendo y ajustando la teoría, o bien, su campo de aplicación. Este ciclo permanente de teoría a experimento y de experimento a teoría -para estudiar los fenómenos- es buena parte del accionar científico. Se debe hacer así pues nunca se tiene una certeza total acerca del fenómeno estudiado.

Fenómeno aleatorio: *es todo fenómeno sobre el cual no se tiene la certeza absoluta de poder explicarlo, en por lo menos algún ámbito o sistema de referencia.*

Definido así, se deduce que todo fenómeno conocido es o fue aleatorio alguna vez. Por ejemplo, la determinación del sexo de un recién nacido fue aleatoria hasta antes del alumbramiento, momento en que se alcanza la certeza. No hay ciencia sin experimentos y tampoco hay experimentos sin ciencia.

Experimento aleatorio: *es todo experimento sobre cuyo resultado no se tiene "a priori" la certeza de su resultado.*

Por ejemplo, si el experimento es lanzar un dado, se tiene la certeza que hay seis resultados posibles pero nunca se sabe cuál cara saldrá si se trata de un dado normal. Existe un cierto grado de incertidumbre asociado a cada cara posible. En cambio, si se trata de contar los dedos de una mano, solo hay certeza si se trata de la mano del investigador; pero cuando se va a estudiar a un grupo de desconocidos aparece la incertidumbre.

A la indeterminación asociada al resultado de un experimento se la llama *error* y se la puede cuantificar con una probabilidad. Esa variabilidad de los resultados se produce por causas desconocidas, en la gran mayoría de los casos, a las que se les puso un nombre: *azar*. Como no se puede eliminar la variabilidad en los experimentos lo único que se puede hacer es acotarla a márgenes razonables que permitan mantenerlo controlado. Cuanto menor sea la indeterminación, menor será el error y mayor la precisión. La manera clásica es fijar todas las variables externas que se cree, pueden causar variaciones, menos una, la que se va a medir. Entonces, la repetición del experimento arrojará resultados lo más uniformes posibles ganando homogeneidad. Las mediciones diarias del laboratorio de análisis clínicos son experimentos aleatorios, tanto como la cantidad de clientes que llegan a una farmacia. El problema adicional en un laboratorio, a diferencia de la Física o la Química, reside en la poca estabilidad del material de trabajo. Una sangre entera se altera mucho más fácilmente que un trozo de metal. Por ende, sus resultados estarán asociados a una incertidumbre mayor. Lo mismo en una farmacia donde la llegada de clientes está influenciada por factores exógenos como la propaganda, su ubicación geográfica, etc. Esta falta de precisión no ayuda a la hora de efectuar diagnósticos o pronósticos y mucho menos para elaborar teorías. Sin embargo, se puede trabajar científicamente en la práctica diaria pero dentro de márgenes más amplios.

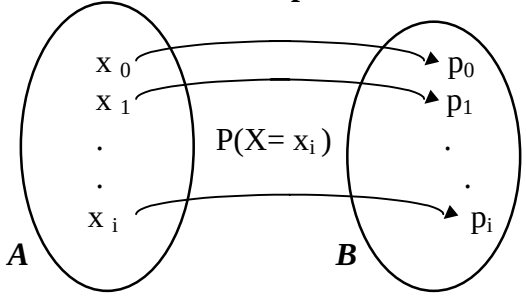
Si se imagina un resultado obtenido en una medición biológica (dato) como un valor de una variable matemática, entonces se puede asimilar la magnitud biológica medida a una:

Variable aleatoria: *son todas aquellas magnitudes donde cada uno de los valores que pueda tomar, en un sistema de referencia o población, tiene asociada una cierta probabilidad de ocurrencia.*

Así, toda variable biológica es una variable aleatoria. Cuando una magnitud cualquiera no varía en todo su ámbito de referencia, deja de ser variable aleatoria y se transforma en un *parámetro poblacional.* Como puede ser, la magnitud sexo en toda la población femenina actual de Posadas.

Sea el conjunto ***A*** de todos los resultados posibles de una medición de la magnitud biológica X, la cual es una variable aleatoria; entonces, cada valor medido x_i tiene asociada una cierta probabilidad de ocurrencia p_i, a través de una *función probabilidad* $P(X= x_i) = p_i$ con un dominio Dom_p definido como todos los valores obtenidos dentro de los números reales $\Re$ y con una imagen ***B*** perteneciente a los reales entre 0 y 1.

Gráfico 8.1: Función probabilidad.



$P : \boldsymbol{A} \rightarrow \Re$

Dominio: $\boldsymbol{A}$: $Dom_p \{ x_i \in \Re \ / \ x_i$: valor medido $\}$

Imagen: $\boldsymbol{B} \subset \Re$, $I_p = [0,1] \subset \Re$

Entonces, una *función de probabilidad* es una función tal que cumple dos condiciones:

1) $1 \geq P(X = x_i) \geq 0$

2) $\sum_i P(x_i) = 1$

La segunda condición refleja el hecho que el conjunto de sucesos correspondientes a los diferentes x_i es una partición. Como por ejemplo el caso de los cuatro casos posibles al efectuar un diagnóstico.

$\sum_i P(x_i) = vp / N + fp / N + vn / N + fn / N = (vp + vn + fp + fn) / N = 1$

8.2 Función de distribución

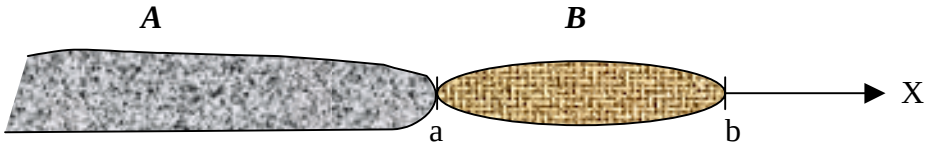
Un experimento aleatorio puede ser la determinación de una glucosa, un colesterol, la temperatura del cuerpo del paciente, su altura etc., donde se está midiendo una sola magnitud clínica aunque esta sea compuesta como una concentración o densidad. En estos casos se dice que se está trabajando con una variable simple o en un caso *unidimensional*. Al repetir sistemáticamente el experimento, bajo condiciones controladas por el observador, se dispone de una serie de valores para calcular la frecuencia relativa de los sucesos. O sea, una medida experimental de su probabilidad de ocurrencia. Otra forma es conociendo a través de la teoría todas las probabilidades asociadas a los valores posibles que toma la variable aleatoria, como la relación mendeliana 9:3:3:1 para la segunda generación filial.

Cuando se conoce la función probabilidad de una variable aleatoria, se puede obtener su acumulada: la *función distribución* de probabilidad. Sea r un valor cualquiera que puede adoptar una variable aleatoria X, y sea $P(X \leq r)$ la probabilidad que X adopte un valor menor o igual que r, entonces se llama:

$F(r) = P(X \leq r)$ *función distribución*

$P(r) = P(X = r)$ *función probabilidad*

Sean dos sucesos ***A***: $X \leq a$ (siendo a un valor cualquiera de la variable aleatoria X)
B: $a < X \leq b$ (donde $b > a$)



Entonces $P(A) = P(X \leq a) = F(a)$ (incluye al valor a)
$P(B) = P(a < X \leq b)$ (no incluye al valor a)

Como ambos sucesos ***A*** y ***B*** no pueden ocurrir a la vez, al no tener ningún punto en común son mutuamente excluyentes ($\boldsymbol{A} \cap \boldsymbol{B}$) = ***0***, luego se puede aplicar el Axioma 3 de probabilidad, para la unión de ambos sucesos:

$$P(\boldsymbol{A} \cup \boldsymbol{B}) = P(\boldsymbol{A}) + P(\boldsymbol{B}) = P(X \leq a) + P(a < X \leq b) = P\{(X \leq a) \cup (a < X \leq b)\}$$

$$P(\boldsymbol{A} \cup \boldsymbol{B}) = P(X \leq b) = F(b) = P(X \leq a) + P(a < X \leq b) = F(a) + P(a < X \leq b)$$

De donde $$P(a < X \leq b) = F(b) - F(a)$$

La igualdad anterior expresa lo siguiente: s*i se conoce la función distribución F (r) para todos los valores posibles de r, de la variable aleatoria X, entonces la función probabilidad queda completamente determinada.*

8.2.1 Distribuciones discretas unidimensionales

Si la variable aleatoria estudiada es de tipo discreto, es decir, solo puede tomar algunos valores $x_1, x_2, x_3 \ldots x_r$ dentro de un intervalo [a, b] cualquiera en los números reales, entonces se puede definir la probabilidad que ocurra un evento cualquiera $p_i = P(x_i)$ y su función distribución con la relación :

$$F(k) = \sum_i P(X \leq x_k) = P(x_1) + P(x_2) + \ldots + P(x_k)$$

Luego, si hay r eventos posibles será: $$\sum_{i=1}^{r} P(X = x_i) = p_1 + p_2 + \ldots + p_r = 1$$

Cuando r sea muy grande, los valores de probabilidad se hacen muy pequeños, y en el caso límite, cuando $r \to \infty$ los valores $p_i \to 0$. Pero aún así las expresiones anteriores siguen siendo válidas. Por ejemplo, sea el caso de lanzar tres veces una moneda al aire, representado en el Gráfico 6.1 mediante el diagrama del árbol. Se pueden graficar las funciones probabilidad y distribución para ese caso (Gráfico 8.2), tomando como variable aleatoria X al número de caras obtenidos al lanzar tres veces la moneda. Los casos posibles serán: X = 0 (no salió ninguna cara); X = 1 (salió una sola cara); X = 2 (salieron dos caras) y X = 3 (las tres fueron caras). O sea,

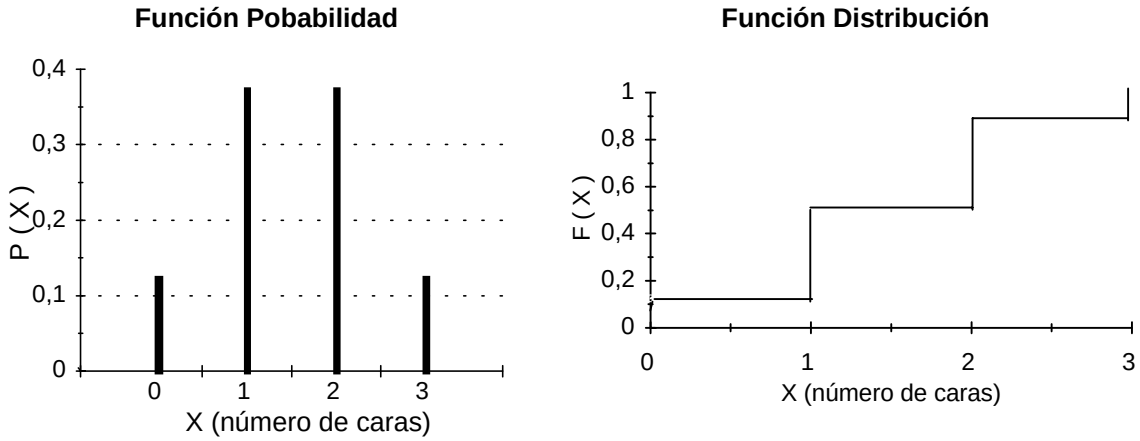
$$X = 0, 1, 2, 3 \Rightarrow P(X=0) = P(X=3) = 1/8 \quad ; \quad P(X=1) = P(X=2) = 3/8$$

En el caso de variable discreta, la función probabilidad es un diagrama de bastones parecido al diagrama de frecuencias relativas (histograma). Por su parte, la función distribución se parece al polígono de frecuencias acumuladas. En un intervalo cualquiera [1, 2] resulta

$F(1) = 0{,}125$ y $F(2) = 0{,}5$ por lo tanto $P(1 < X \leq 2) = 0{,}5 - 0{,}125 = 0{,}375 = P(2)$

Se pueden calcular así los valores de las respectivas funciones para cada valor de X. Porque si se conocen todos los valores posibles de la función probabilidad, quedan completamente determinados todos los valores de la función distribución.

Gráfico 8.2: Función probabilidad y distribución en el lanzamiento triple de una moneda.



En el caso de la Tabla de la Verdad para los resultados de un diagnóstico en N pruebas, aún sin conocer la función de probabilidad, se sabe que los cuatro resultados posibles tienen una probabilidad calculada con:

$$\sum_{i=1}^{r} P(X = x_i) = P(X = vp) + P(X = fp) + P(X = vn) + P(X = fn) = 1$$

8.2.2 Distribuciones continuas unidimensionales

Si ahora la variable aleatoria puede tomar cualquier valor en un intervalo cualquiera, entonces se trata de una variable continua. Pero, en un intervalo como [a, b] existen infinitos puntos de la variable aleatoria eso significa infinitos resultados posibles de experimento. Entonces los valores de la función probabilidad para un punto cualquiera, solo pueden ser infinitesimales para que la acumulada de todas ellas sea la unidad. Por ello, en variable continua la función de probabilidad *siempre se define para un intervalo y no para un punto*.

Sea un intervalo cualquiera [a, b] de una variable biológica X continua y aleatoria, definida en los números reales $\Re$, y sea x un punto dentro de ese intervalo, tal que $a < x < b$, la función probabilidad asociada $P(X = x)$ puede definirse para un pequeño intervalo Δx de manera tal de hacer: $a = x - \Delta x$ y $b = x + \Delta x$; o sea: $x - \Delta x < x < x + \Delta x$ entonces:

$$P(a < x < b) = F(b) - F(a) = P(x \pm \Delta x)$$

Tomando límites será

$$\lim_{\Delta x \to 0} P(x \pm \Delta x) = \int_a^b f(x)\, dx = F(b) - F(a)$$

Esa es la función distribución para el caso de variable continua. Por su parte, ahora se puede definir una *función densidad de probabilidad* o *función frecuencia* con:

$$f(x) = dF(x) / dx$$

donde: d F (x) : es un elemento de probabilidad
d x : es un pequeño incremento de la variable

Integrando la función frecuencia para todo el campo real resulta ser:

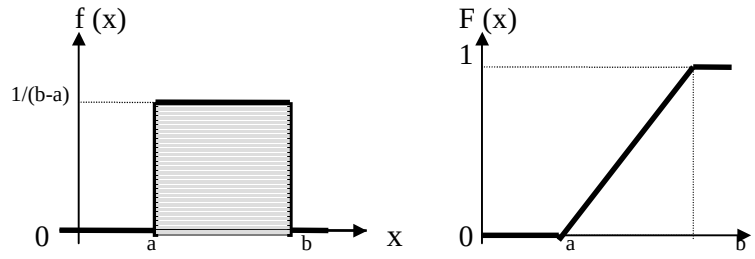
$$\int_{-\infty}^{+\infty} f(x)\, dx = 1$$

Análogamente, si $a \to \infty$ es $\int_{-\infty}^{b} f(x)\, dx = F(b)$

Gráfico 8.3: Función frecuencia y función distribución para la función "diente de sierra".

$$f(x) = \begin{cases} 1/(b-a) & \forall\ a < x < b \\ 0 & \text{para todo otro caso} \end{cases}$$

El área sombreada vale 1
F(a) = 0 y F(b) =1



8.2.3 Distribuciones conjuntas e independencia

Cuando se trabaja con dos variables aleatorias X e Y se trata de un espacio bidimensional, donde si las respectivas funciones de distribución P ($X \leq x$) y P ($Y \leq y$) existen para todo *x* y para todo *y*, entonces la función distribución *conjunta* de X e Y se define con:

$$F_{x,y}(x, y) = P[(X \leq x ; Y \leq y)] = \int\int f_{x,y}(x, y / \, dx\, dy \quad \text{(continua)}$$

$$F_{x,y}(x, y) = P[(X \leq x ; Y \leq y)] = \sum_i \sum_i P(X \leq x_k)\ P(Y \leq y_k) \quad \text{(discreta)}$$

Dos variables aleatorias X e Y son *independientes* (o mutuamente) si los eventos $X \leq x$ e $Y \leq y$ son mutuamente independientes para cada par de valores *x* e *y*. Esto significa que la distribución de valores de X no es afectada por los valores de Y, y viceversa. El concepto de independencia se puede expresar por la relación siguiente:

$$F_{x,y}(x, y) = P[(X \leq x ; Y \leq y)] = P[X \leq x] \,.\, P[Y \leq y]$$

Para variables discretas la función distribución es la sumatoria de los valores (la acumulada) y para las variables continuas se expresa con las integrales como se ilustró antes Generalizando si se tienen n variables aleatorias $X_1, X_2, ..., X_n$ la función distribución conjunta de todas ellas será:

$$F_{x1, x2,...xn}(x_1, x_2,...x_n) = P[(X_1 \leq x_1 ; X_2 \leq x_2....;X_n \leq x_n] = P[X_1 \leq x_1] \,.\, P[X_1 \leq x_1] .... P[X_n \leq x_n]$$

Siempre y cuando la función exista o sea convergente.

8.3 Valor esperado

Se llama también *esperanza matemática* o *momento de primer orden*. Se trata de un operador matemático que al ser aplicado a la función probabilidad permite el cálculo de ese valor en el caso discreto, mientras que en el caso continuo se lo aplica a la función frecuencia:

$\Xi (x) = \sum_i (x) \; P (X = x)$ en el caso discreto

$\Xi (x) = \int_{\Re} (x) \; f (x) \, dx$ en el caso continuo

Por ejemplo, si el experimento consiste en arrojar un dado normal, la variable aleatoria X podrá tomar seis valores equiprobables, y aplicando el operador anterior resultará:

$$\Xi (x) = \sum_{1}^{6} (x) \; P (X=x) = 1\,(1/6) + 2\,(1/6) + 3\,(1/6) + 4\,(1/6) + 5\,(1/6) + 6\,(1/6) = 3{,}5$$

Notar que el valor esperado coincide con el promedio de los dos datos centrales 3 y 4, o sea que este no tiene por qué coincidir con uno de los valores de la variable. Observando atentamente la definición anterior, puede verse que el *valor esperado es un valor promedio, ponderado* con las probabilidades de ocurrencia de cada caso y una medida de estas probabilidades es la frecuencia relativa de cada una.

Sea ahora el caso del ejemplo visto en el Gráfico 8.3 anterior

$$\Xi (x) = \int_{-\infty}^{+\infty} x \, . \, f (x) \, d x = \int_{a}^{b} x / (b - a) dx = 1/(b\text{-}a) \, | \, x^2/2 |_a^b = 1/(b\text{-}a)\{b^2/2 - a^2/2\} = (b+a)/2$$

El valor esperado resulta ser la semisuma de los límites del intervalo (una especie de centro de "gravedad" sobre el eje de abcisas, en términos probabilísticos).

Se demuestran fácilmente las propiedades siguientes:

1) El valor esperado de una constante es la constante:

$$\Xi (a) = a$$

2) El valor esperado del producto de una variable por una constante es igual al producto de la constante por el valor esperado de la variable.

$$\Xi (b \, . \, x) = b \, \Xi (x)$$

3) El valor esperado de una variable aleatoria, definida por una función indirecta: y = g (x), es distributivo respecto de la misma si el valor converge.

$$\Xi (y) = \Xi (g(x)) = g \{ \Xi (x) \}$$

Por ejemplo, si y = a + bx, entonces:

$$\Xi (y) = \Xi (g(x)) = \Xi (a + bx) = \Xi (a) + \Xi (b . x) = a + b \Xi (x)$$

4) El valor esperado de dos variables independientes X e Y es igual al producto de los valores esperados de cada una de ellas:

$$\Xi (X, Y) = \Xi (X) . \Xi (Y) \}$$

Generalizando para n variables aleatorias independientes:

$$\Xi (X_1, X_2, ..., X_n) = \prod_{1}^{n} \Xi (X_i) = \Xi (X_1) . \Xi (X_2) .... \Xi (X_n)$$

Esta propiedad es equivalente a la ya vista como *regla del producto* en el caso del lanzamiento de una moneda al aire tres veces y esquematizada con el diagrama del árbol en el capítulo 5.

Tabla 8.1: Valores esperados en Procesos Bernoulli, Poisson e Hipergeométricos.

Probabilidad **P (x)**	**Valor esperado** $\mu = \Xi_B (x)$
Binomial	n . p
Pascal	r . p
Binomial Negativa	r . q / p
Geométrica	q / p (Odds de fracaso)
Hipergeométrica	n . R / N
Poisson	μ

8.3.1 Aplicaciones de valor esperado

A continuación se presentan algunas aplicaciones muy comunes en este tema.

Ejemplo 1) Un bioquímico sabe por experiencia que el 20% de sus pacientes son enviadas a efectuarse un análisis de tipo Gravindex, los cuales le dejan una ganancia de $ 2,7 según sus cálculos. Si en su primer día del mes tuvo 20 pacientes:
a) ¿Cuánto espera ganar diariamente por los Gravindex?
b) ¿Cuál es la probabilidad de tratar 5 pacientes antes de tener primer Gravindex?
c) ¿Cuál es el número esperado de pacientes antes de tener el primer Gravindex?

a) Se estima que se trata de un proceso tipo Bernoulli donde hay una probabilidad de éxito p=0,2 y la cantidad diaria de pacientes es n = 20 por día, luego:

El valor esperado de prácticas hematológicas será μ = n . p = 20 . 0,2 = 4 Gravindex diarios;
y su ganancia promedio esperada será G = 2,7 . 4 = 10,8 $ diarios por esa práctica.

b) Se trata de un caso de probabilidad geométrica con r = 1 , g = 5 y q = 0,8, entonces
P_G (g=5 /p=0,2) = 0,2 $(0,8)^5$ = 0,066

c) El valor geométrico esperado es: $\mu = q / p = 0,8 / 0,2 = 4$ pacientes
O sea, espera 4 pacientes antes del quinto Gravindex del día.

Ejemplo 2) A un hospital llegan pacientes por la mañana a efectuarse extracciones de sangre. Se ha medido la frecuencia de llegada de los mismos, en intervalos elegidos al azar en intervalos de 10 minutos. La distribución de probabilidad empírica se muestra en la tabla siguiente. Para definir cuántos puntos de atención deben prepararse. Se desea saber:

a) El número esperado de pacientes en un tiempo de 10 minutos.
b) Si se puede suponer que los arribos se producen según un proceso del tipo Poisson.

Número de pacientes:	X	0	1	2	3	4	5	6 ó más
Probabilidad empírica:	P(X)	0,15	0,25	0,25	0,20	0,10	0,05	0,0

a) El número esperado de pacientes es la media aritmética ponderada

μ = 0 (0,15) + 1 (0,25) + 2 (0,25) + 3 (0,20) + 4 (0,10) + 5 (0,05) = 2 pacientes cada 10 minutos

Suponiendo que se trata de un modelo Poisson, el valor esperado de la frecuencia de arribos será de 2 pacientes cada 10 minutos. Entonces, midiendo cuánto se demora en atender a cada paciente en promedio, se puede saber el tiempo de espera de cada uno y colocar cierto número de puntos de atención para: 1) lograr que nunca haya más de x pacientes en la cola de espera; o 2) lograr que nunca se espere más de n minutos en la cola. Los modelos matemáticos para resolver estas dos cuestiones se ven en la Teoría de Colas (Investigación Operativa).

b) Una vez obtenido el valor esperado, se puede aplicar la fórmula para calcular la probabilidad Poisson, usando la manera iterativa, y comparar con la empírica.

Número de pacientes:	X	0	1	2	3	4	5	6 ó más
Probabilidad Poisson:	P_{PO}(X)	0,135	0,27	0,27	0,18	0,09	0,036	0,01

La concordancia entre la probabilidad empírica (Oi: observada) y la teórica (Ei: esperada) parece ser buena a simple vista. Sin embargo, se debe efectuar una *validación estadística* para comprobarlo como se mostrará más adelante en el tema Pruebas de Bondad de Ajuste.

Ejemplo 3) Se recibe un lote de mil ampollas con un cierto medicamento en una compra realizada. Para revisar la compra el farmacéutico decide controlar a 20 de ellos tomados al azar. Si el proveedor dice que el porcentaje de fallas no supera el 0,5% ¿cuántas ampollas falladas se espera encontrar en la muestra a revisar?

Es un caso de probabilidad Hipergeométrica, por lo tanto su valor esperado se obtiene con:

μ = n . R /N 20 . 5/1000 = 0,1 significa que no debería aparecer ninguna fallada.

8.3.2 Aplicaciones en Epidemiología: OR y RR

En Epidemiología se emplean dos índices básicos llamados Odds Ratio (OR) y Riesgo Relativo (RR) que se usan de acuerdo al tipo de estudio que se está efectuando. En Medicina el *riesgo* puede ser definido como la probabilidad de que ocurra un determinado suceso que implique un peligro para la salud de un paciente. Tal como el riego de contraer una septicemia luego de ser operado, etc. El *factor de riesgo* es una magnitud cualitativa, de tipo dicotómica, donde sus dos resultados posibles indican si el factor de riesgo (o exposición) está presente o no. Usualmente esto se denota con: SÍ-NO. Son ejemplos de factor de riesgo características tales como: edad, sexo, tratamiento profiláctico antes de una cirugía, inmunización, medicación preventiva, etc. Cuando un grupo de pacientes recibe una droga que se está probando, será el caso SÍ del factor de riesgo, y el otro grupo usado como control o placebo será el caso NO. En cambio, cuando el primer grupo es cateterizado en su cirugía se piensa que está expuesto a una infección (SÍ) y otro grupo donde no fue necesaria (NO) será el de control.

La comparación del riesgo es uno de los objetivos principales de los estudios epidemiológicos, y su cuantificación implica la medición experimental de sus dos índices asociados: *Riesgo Relativo* (RR) y "*Odds-Ratio*" (OR).

El diseño experimental para medir ambos índices es el planteo de una Tabla de Contingencia del tipo 2x2. Cuando el investigador es quien decide el tamaño muestral y cuántos serán expuestos al factor de riesgo, es un caso de *muestreo basado en la exposición*. En cambio, cuando el investigador decide cuantos sujetos enfermos y sanos entran en su estudio se trata de un *muestreo basado en la enfermedad*. Por ejemplo, suponiendo que toda la población estudiada es homogénea en todas sus características excepto en una, luego una parte de esta se expone a un factor de riesgo que se cree es de importancia en causar una determinada enfermedad, y el diseño experimental usado generalmente se muestra esquemáticamente en la Tabla 15.2 siguiente:

Tabla 8.2: Frecuencias observadas entre la exposición a un factor de riesgo y la contracción de una enfermedad cualquiera.

Factor de Riesgo O Característica Grupal		Enfermedad SÍ	Enfermedad NO	Total
	SÍ	a	b	$n_1 = a+b$
	NO	c	d	$n_2 = c+d$
	Total	$n_3 = a+c$	$n_4 = b+d$	n

$$RR = (a \cdot n_2) / (c \cdot n_1)$$

$$OR = (a \cdot d) / (b \cdot c)$$

Riesgo relativo (RR) se define como el cociente entre la probabilidad de contraer la enfermedad de la población expuesta y la probabilidad de contraerla de los no expuestos.

$RR = P(E) / P(noE) = (a/n_1) / (c/n_2) = (a \cdot n_2) / (c \cdot n_1)$ que es la expresión de más arriba.

Cuando no hay pacientes de la población expuesta que contraigan la enfermedad (a = 0) el RR se anula. En cambio, si ningún paciente de los no expuestos contrajo la enfermedad (c = 0) entonces el RR se vuelve infinito. Si los factores son *independientes* el RR = 1. El RR solo puede ser estimado con estudios prospectivos. La interpretación se puede ver mejor con un ejemplo: se

sabe que entre los que sufrieron infarto de miocardio el nivel de colesterol medio es de 300 mg/dl. En la población general el nivel es de 200 mg/dl. Se conoce la distribución de colesterol en ambas poblaciones, la de los infartados y la de la población humana usadas como referencia. Desde el punto de vista médico se toma como valor referencial 240 mg/dl, se considera a un paciente con "colesterol alto" cuando supera dicho valor. Luego se cuentan los casos de encontrados con colesterol alto en los infartados y entre los no infartados. Sea por caso, una relación de 120/180 = 2/3 entre los infartados y de 1000/9000 = 1/9 entre los no infartados, entonces se calcula el RR = (2/3) / (1/9) = 6. Eso se interpreta así: "Si un paciente tiene un colesterol mayor que 240, su chance de infarto es 6 veces mayor que si tuviese un valor de 200 o menos". Así se construyen las simplificaciones en los diarios y revistas de divulgación.

Si se trata de un estudio del tipo caso-control, generalmente el RR no puede calcularse y se necesita del OR como una medida de asociación entre ambos factores analizados. Entonces, el *Odds Ratio* (OR) se define como el cociente entre dos "odds" posibles. Un "*odd*" es la relación entre la cantidad de "enfermos" y los "no enfermos" de una población dada. Como hay dos poblaciones, la expuesta y la no expuesta al factor de riesgo, hay dos tipos de "*odd*" posibles y la tasa entre ambos es el valor de OR. Sea:

$P_1 = a / n_1$: la probabilidad de contraer la enfermedad de la población expuesta.

$(1 - P_1) = b / n_1$: la probabilidad de no contraer la enfermedad de la población expuesta

El "*odd*" de la población expuesta es $O_1 = P_1 / (1 - P_1) = a / b$. A su vez,

$P_2 = c / n_2$: la probabilidad de contraer la enfermedad de la población no expuesta.

$(1 – P_2) = d / n_2$: la probabilidad de no contraer la enfermedad de la población no expuesta.

El "*odd*" de la población no expuesta es $O_2 = P_2 / (1 – P_2) = c / d$.

Por lo tanto, su cociente es $OR = O_1 / O_2 = (a . d) / (c . b)$ que es la expresión vista en la Tabla 8.2 de más arriba. Cuando el OR = 1 significa que ambos factores estudiados son independientes entre sí. Si a o d se anulan entonces el OR = 0; en cambio, si se anulan c o b, se hace infinito. Para el caso usado como ejemplo para el RR, se puede calcular:

OR = (120 / 1000) / (60 / 8000) = 16

Y se interpreta así: "cuanto más alto el OR, peor". Otra forma de ver este índice es como el cociente de dos pares de probabilidades. Una es el RR entre los expuestos y el otro es el RR entre los no expuestos:

$$RR1 = Pe/(1-Pe) = \frac{a/(a+b)}{b/(a+b)} = a/b \quad y \quad RR2 = Pno\ e/(1 – Pno\ e) = \frac{c/(c+d)}{d/(c+d)} = c/d$$

Luego es: $OR = RR1 / RR2$

La diferencia en el uso de uno u otro índice reside en el tipo de investigación que se está realizando. *RR* se emplea en un estudio de *incidencia,* donde la frecuencia de algún resultado se calcula entre dos grupos determinados por la presencia o ausencia de alguna característica. En cambio, *OR* se usa en un *estudio de caso-control* donde los resultados se obtienen en dos grupos, uno expuesto a un factor de riego y el otro usado como placebo o control.

El valor esperado de una variable aleatoria discreta X se denota con Ξ (X). Cuando la sumatoria abarca todos los valores posibles de X, la definición será

$$\Xi (X) = \sum_j Xj \; P(X = Xj)$$

Y cuando se aplica a n variables aleatorias mutuamente independientes, el valor esperado es

$\Xi \left(\prod_{j=1}^{n} Xj \right) = \prod_{j=1}^{n} E(Xj)$ Para este caso donde n = 4 será:

$$\Xi (X_1 X_2 X_3 X_{4)}) = \Xi (X_1) \Xi (X_2) \Xi (X_3) \Xi (X_4)$$

Por ejemplo si es : $X_1 = a$; $X_2 = d$; $X_3 = 1/b$; $X_4 = 1/c$ entonces el valor esperado de OR será

$$\Xi (OR = X_1 X_2 X_3 X_{4)}) = \Xi (a) \Xi (d) \Xi (1/b) \Xi (1/c) = [\Xi (a) \Xi (d)] / [\Xi (b) \Xi (c)]$$

Por otra parte se puede obtener el valor esperado de acuerdo a la definición de independencia:

$P(\boldsymbol{A} \cap \boldsymbol{B}) = P(\boldsymbol{A}) \; . \; P(\boldsymbol{B})$ donde el evento ***A*** son los individuos expuestos y ***B*** los sanos

Donde $P(\boldsymbol{A}) = n_1 / n$ y $P(\boldsymbol{B}) = n_3 / n$ de acuerdo a la notación de la Tabla 8.2

$P(\boldsymbol{A} \cap \boldsymbol{B}) = \Xi (a) / n$ de acuerdo a la definición clásica de probabilidad, entonces reemplazando

$\Xi (a) / n = (n_1 / n) \; . \; (n_3 / n)$ O sea $\Xi (a) = (n_1 \; . \; n_3) / n$

Si hay independencia el valor esperado de cada celda de la Tabla se puede calcular como el producto de sus totales marginales dividido el total muestral, como ya se vio en 6.6. Análogamente:

$\Xi (b) = (n_1 \; . \; n_4) / n$ $\Xi (c) = (n_2 \; . \; n_3) / n$ $\Xi (a) = (n_2 \; . \; n_4) / n$

Reemplazando en el valor esperado de OR es: $\Xi (OR) = [\Xi (a) \Xi (d)] / [\Xi (b) \Xi (c)] = 1$

Y la conclusión es: Si el factor es independiente (matemáticamente hablando) de la enfermedad el valor esperado de OR es igual a 1, como ya se vio en el punto 6.6. Análogamente:

$$\Xi (RR) = [\Xi (a / n_1)] / [\Xi (c/ n_2)] = [\Xi (a) / n_1] / [\Xi (c) / n_3] = 1$$

El factor de riesgo, o de protección en el caso de una vacuna, puede ser cualquiera. Sin embargo, cuando se considera a un método de diagnóstico (o a un test clínico) como si fuese el factor de riesgo, entonces la tabla epidemiológica usual, se transforma en una Tabla de la Verdad y pueden hallarse relaciones entre ambos índices.

8.3.3 Caso especial: riesgo en los test clínicos

Cuando un test clínico es considerado como el factor de riesgo, entonces el factor Enfermedad (presente o ausente) divide a la muestra en dos conjuntos: sanos y enfermos. Por su parte el factor de riesgo ahora es el método diagnóstico y cuando está presente equivale al caso positivo; en cambio si está ausente equivale a un caso negativo. En el gráfico siguiente se muestra la equivalencia entre ambas tablas

Gráfico 8.4: Tabla epidemiológica y Tabla diagnóstica

Tabla epidemiológica

		Enfermedad		
		SÍ	NO	Total
Factor de Riesgo	SÍ	a	b	$n_1 = a+b$
	NO	c	d	$n_2 = c+d$
	Total	$n_3 = a+c$	$n_4 = b+d$	n

Tabla diagnóstica

Factor de Riesgo o Test Clínico		Enfermedad		
		SÍ	NO	Total
	(+)	vp	fp	TP
	(-)	fn	vn	TN
	Total	TE	TS	N

En el apartado 6.3.1 se mostró una simulación donde siempre se obtenían dos índices básicos (sensibilidad y especificidad) no importa de que manera se seleccionaban las muestras. Esto significa que es lo mismo una selección basada en la enfermedad (Caso control) que una selección basada en la exposición (estudio por Cohorte o Clinical trial), para el caso especial donde el factor de riesgo es el método del diagnóstico o test clínico. Por ejemplo, para determinar si un individuo estuvo expuesto alguna vez a la toxoplasmosis, se debe efectuar un análisis clínico por inmuno fluorescencia. Y si estuvo expuesto en su pasado, el resultado da positivo. Se define:

Diagnostic Ratio (DR): Es el OR cuando el factor de riesgo es un método de diagnóstico.

Se cambia la denominación para recordar que el DR es un caso especial del OR. Se puede calcular de tres maneras diferentes, según el significado que se busca, pero siempre se encuentra el mismo resultado:

$$DR = (vp \cdot vn) / (fp \cdot fn)$$

Significado estadístico 1) Es el odd de una enfermedad cuando se la predice (o diagnostica) dividido el odd de la enfermedad cuando no es predicha (o diagnosticada).

Para este caso se pueden suponer dos eventos opuestos: ***A*** (un enfermo con un diagnóstico positivo) y ***B*** (un enfermo con un diagnóstico negativo). Y sus probabilidades son:

$P(\boldsymbol{A}) = P(\boldsymbol{TE} \cap \boldsymbol{VP}) / P(\boldsymbol{VP}) = vp / (vp + fp)$

$P(\boldsymbol{B}) = P(\boldsymbol{TD} \cap \boldsymbol{VN}) / P(\boldsymbol{VN}) = fn / (vn + fn)$
$Odds(\boldsymbol{A}) = P(\boldsymbol{A}) / [1 - P(\boldsymbol{A})] = vp / fp$

Odds (***B***) = P(***B***) / [1 - P(***B***)] = fn / vn

Entonces el cociente de ambos términos es:

DR = Odds (***A***) / Odds (***B***) = (vp . vn) / (fp . fn)

Significado estadístico 2) Es el odd de no-enfermedad cuando se la diagnostica, dividido el odd de la no-enfermedad cuando no es diagnosticada.

Para este caso se pueden suponer dos eventos opuestos: ***C*** (un sano con un diagnóstico negativo) y ***D*** (un sano con un diagnóstico positivo). Y sus probabilidades son:

P (***C***) = P ***(TS ∩ TN***) / P (***TN***) = vn / (vn + fn)

P (***D***) = P (***TS ∩ TP***) / P (***TP***) = fp / (vp + fp)

Odds (***C***) = P(***C***) / [1 - P(***C***)] = vn / fn

Odds (***D***) = P(***D***) / [1 - P(***D***)] = fp / vp

Nuevamente el cociente de estos dos términos es:

DR = Odds (***C***) / Odds (***D***) = (vp . vn) / (fp . fn)

Significado clínico: Es el cociente entre los dos Likelihood Ratios (LR+ dividido LR-)

LR+ = S / (1- E) = (vp /TE) / [1- (vn /TS)] = (vp /TE) / (fp /TS)

LR- = (1- S) / E = [1 - (vp /TE)] / (vn /TS) = (fn /TE) / (vn /TS)

DR = LR+ / LR- = S . E / [(1-S) (1-E)] = (tp . tn) / (fp . fn)

Por ejemplo, si DR = 9 significa que los odd de enfermedad del método cuando diagnostica la enfermedad es nueve veces mayor que cuando no la diagnostica. O más simple, el LR+ es nueve veces mayor que el LR-. Por lo tanto, el DR puede ser considerado como un índice de calidad intrínseco al método que no varía con la prevalencia, ya que es función solo de S y E.

Por su parte, el riesgo relativo puede ser definido desde un punto de vista estadístico como: *La probabilidad de una enfermedad cuando es diagnosticada, dividida por la probabilidad de la enfermedad cuando no es diagnosticada:*

RR = [vp / (vp + fp)] / [fn / (vn + fn)] = (vp . TN) / (fn . TP)

Por ejemplo, si RR = 10 indica que la probabilidad de predecir la enfermedad es diez veces más grande que la de predecir la no-enfermedad. Por su parte, clínicamente se lo puede definir como:

RR = VPP / (1 – VPN)

Por lo tanto, este índice variará de acuerdo a la prevalencia de la población de referencia en la cual es usado, y puede ser considerado como una variable de la calidad del método clínico. Es conveniente informar su valor con una curva en lugar de un número. Puede verse, que para el caso de una enfermedad muy rara, como la meningitis, los valores vp y fp serán muy chicos y su producto será casi nulo. Por eso, los epidemiólogos lo aproximan con OR. Pero esto es solo una aproximación matemática porque conceptualmente son conceptos muy diferentes. Se puede encontrar la relación entre ellos, para este caso en particular con:

DR = RR [1 +(LR-). O)] / [1 + (LR+) . O] donde el odds de enfermedad es: O = p / (1-p), o bien:

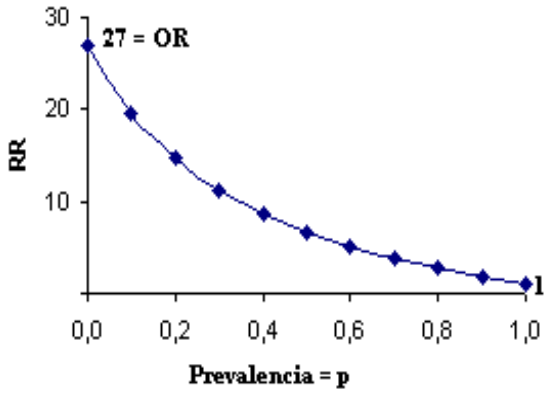
RR = DR [1 + (LR+) . O] / [1 + (LR-) . O]

Como DR es una constante, entonces se puede deducir de la última ecuación que:

- Cuando p tiende a cero, también O tiende a cero y así RR ≈ DR
- Cuando p tiende a uno, entonces O tiene a infinito y así RR tiene a uno.

Esto significa que, cuando la enfermedad es muy rara p tiene a cero, y cuando hay una epidemia muy grande entonces p tiene a uno. Para mostrar la variabilidad de RR con la prevalencia, se puede ver un caso donde S = 0,9 y E = 0,75. En el Gráfico 8.5 se puede ver que cuando la prevalencia es nula el valor de RR es uno, a medida que la prevalencia disminuye el valor de RR aumenta hasta un máximo de RR = DR = 27 cuando la prevalencia es nula.

Gráfico 8.5: Variabilidad del Riesgo Relativo con la prevalencia en la población



Método clínico: S = 0,9 y E = 0,75

RR	Prevalencia %
27,00	0
19,57	10
14,68	20
11,22	30
8,65	40
6,65	50
5,06	60
3,77	70
2,69	80
1,78	90
1,00	100

La simulación para obtener el gráfico anterior se comienza usando los valores de sensibilidad y especificidad del método, que son conocidos: S = 0,9 y E = 0,75. Con ellos quedan determinados los valores: DR = 27 y LR+ = 3,6. Luego se toma un valor p = 0,1 y se calcula el RR respectivo con la ecuación de más arriba y se obtiene RR = 19,57. Y así sucesivamente, para valores p = 0,2; 0,3; 0,4; ...; 0,9 con los cuales se calculan los respectivos valores de RR. Tal como se muestra en la tabla de datos del gráfico anterior. Todos los demás índices clínicos se pueden obtener a partir de S y E como se vio antes.

8.4 Momentos de orden k

Se llama *momento de orden k, respecto de un punto c* de una variable aleatoria, a un operador matemático que al ser aplicado a la función probabilidad, permite el cálculo de ese valor en el caso discreto, mientras que para el caso continuo se lo aplica a la función frecuencia:

$$\mu_k = \Xi \{(x - c)^k \}$$

Por ejemplo, si k = 0 es $\mu_0 = 1$
si k = 1 es $\mu_1 = \Xi (x) - c = \mu - c$
si k = 2 es $\mu_2 = \Xi \{(x - c)^2\} = \Xi \{(x^2)\} - 2 \mu c + c^2$

Si se toman momentos *respecto al origen*, esto es c = 0, entonces resulta

$$\mu'_k = \Xi \{(x)^k \}$$

Por ejemplo, si k = 0 es $\mu'_0 = 1$
si k = 1 es $\mu'_1 = \Xi (x) - 0 = \mu$
si k = 2 es $\mu'_2 = \Xi \{(x - 0)^2\} = \Xi \{(x^2)\}$

Si se toman momentos respecto al valor esperado, se llaman *momentos centrales* de orden k, como $\mu = c$ resultan ser

$$\mu''_k = \Xi \{(x - \mu)^k \}$$

Por ejemplo, si k = 0 es $\mu''_0 = 1$
si k = 1 es $\mu''_1 = \Xi (x) - \mu = 0$
si k = 2 es $\mu''_2 = \Xi \{(x - \mu)^2\} = \Xi \{(x^2)\} - 2 \mu \mu + \mu^2 = \Xi \{(x^2)\} - \mu^2$

En particular, el más importante es este último caso; se define como *varianza* de una función de probabilidad al momento central de segundo orden de la misma

$$\sigma^2 = \Xi \{(x^2)\} - \mu^2$$

$\sigma^2 = \sum_i (x - \mu)^2 P(X=x)$ en el caso discreto

$\sigma^2 = \int_{\Re} (x - \mu)^2 . f (x) d x$ en el caso continuo

Otra forma de escribirlo es:

$$\sigma^2 = \Xi \{(x^2)\} - \Xi^2 \{(x)\}$$

Resulta sencillo verificar las propiedades siguientes:

1) La varianza de una constante cualquiera es nula:

$\sigma^2 = \Xi \{(a^2)\} - \Xi^2 \{(a)\} = a^2 - a^2 = 0$

2) La varianza del producto de una variable por una constante es igual al producto del cuadrado de la constante por la varianza de la variable. Si $x = a \cdot y$

$\sigma_x^2 = \Xi \{(ay)^2\} - \Xi^2 \{(ay)\} = a^2 \Xi \{(y^2)\} - a^2 \Xi^2 \{(y)\} = a^2 \sigma_y^2$

3) La varianza de una variable aleatoria, definida por una función indirecta: $y = g(x)$. Es distributivo respecto de la misma si el valor converge. Por ejemplo, si $y = a + bx$, entonces:

$\sigma_y^2 = \sigma^2 (g(x)) = \sigma^2 (a + bx) = \sigma^2 (a) + \sigma^2 (b \cdot x) = b^2 \sigma_x^2$

Los momentos de órdenes superiores suelen usarse como índices de dispersión y asimetría. La varianza es el índice de dispersión por excelencia, mientras que los momentos de tercer orden indican la desviación en simetría de la función de probabilidad respecto de un eje cualquiera. Esto se verá más adelante un poco mejor.

Se puede calcular la varianza para las funciones de probabilidad ya vistas mediante un simple cálculo. En la Tabla 8.3 siguiente se presenta un cuadro con las varianzas de las probabilidades de los procesos Bernoulli, Poisson e Hipergeométrico.

Tabla 8.3 : Varianza en Procesos Bernoulli, Poisson e Hipergeométrico.

Probabilidad $P(x)$	**Varianza** $\sigma^2(x)$
Binomial	$n \cdot p \cdot q$
Pascal	$r \cdot q / p^2$
Binomial Negativa	$r \cdot q / p^2$
Geométrica	q / p^2
Hipergeométrica	$(N-n)\ n.p.q\ /(N-1)$
Poisson	μ

Se puede aplicar esta tabla a los tres ejemplos del punto anterior:

Ejemplo 1) Caso de los Gravindex era Binomial: $\sigma^2 (x) = n \cdot p \cdot q = 20 \cdot 0{,}2 \cdot 0{,}8 = 3{,}2$

Ejemplo 2) Caso de arribo de pacientes era Poisson: $\sigma^2 (x) = \mu = 2$

Ejemplo 3) Caso de muestreo de aceptación era Hipergeométrico: $\sigma^2 (x) = (N-n)\ n.p.q\ /(N-1)$
$\sigma^2 (x) = (1000-20)\ 20\ (0{,}995)(0{,}005) / (1000-1) = 0{,}098$

8.4.1 Variables aleatorias tipificadas

Las propiedades vistas más arriba pueden usarse para transformar una variable, de manera tal de facilitar la operatoria y los cálculos. Es un artificio matemático para lograr que una variable aleatoria cualquiera sea tipificada se usa la relación:

$$z = (x - \mu) / \sigma$$

Sea $a = 1/\sigma$ y sea $b = - \mu /\sigma$, entonces la variable $z = a x + b$, valdrá $z = (x - \mu) / \sigma$ y aplicando los operadores valor esperado y varianza se calcula:

$$\Xi (z) = \Xi \{ (x - \mu) / \sigma \} = 1/\sigma \ (\Xi (x) - \mu) = 1/\sigma (\mu - \mu) = 0$$

$$\sigma_z^2 = \Xi \{(z^2)\} - \Xi^2 \{(z)\} = \Xi \{((x - \mu)/\sigma)\} - \Xi^2 \{((x - \mu)/\sigma)\} = 1/\sigma_x^2 \Xi \{(z^2)\} - (\mu - \mu)/ \sigma$$

$$\sigma_z^2 = (1/\sigma_x^2) \sigma_x^2 = 1$$

Entonces, una variable tipificada se caracteriza por tener:

Variable tipificada
- . *Valor esperado nulo*
- . *Varianza unitaria*

8.5 Aplicaciones en Bioquímica

Existen muchas aplicaciones prácticas de los momentos estadísticos. La mayoría de ellas se refieren a los valores económicos esperados o a los productos fabricados en serie. De entre ellos, se seleccionaron los siguientes.

8.5.1 Índice de Agregación

Es un índice de muchas aplicaciones en campos diversos tales como biología, botánica, etología, ecología, etc. Mide el grado de agrupamiento de los individuos en una determinada región del espacio. Cuando este índice tiene un valor cercano a la unidad, significa que el número de individuos se distribuye al azar, en la unidad de espacio que ocupan (ver Gráfico 8.6); es decir, se ubican en el espacio siguiendo una distribución del tipo Poisson. En cambio, cuando es mucho mayor que uno significa que los individuos tienden a presentarse agrupados, como las colonias de bacterias, los cardúmenes de peces, grupos de hematíes, manadas, una especie de árbol en el bosque, etc. Puede ocurrir que, a su vez, el conjunto de individuos se distribuya al azar dentro del continuo en forma grupal. Esto es, los grupos se distribuyen en el continuo según una distribución de Poisson, pero los individuos dentro del grupo no lo hacen sino que muestran una especie de “distribución contagiosa”. Tal como el fenómeno de apilamiento celular en las cámaras de recuento, o una aglutinación, etc. Por regla general, se trata de una distribución del tipo binomial negativa. Finalmente, cuando el índice de agregación es muy pequeño, próximo a cero, se trata de una distribución uniforme dentro de la ubicación espacial considerada. El

se trata de una distribución uniforme dentro de la ubicación espacial considerada. El número de individuos por unidad de espacio es una constante, la varianza es casi nula. Esto muestra claramente una distribución artificial de los individuos y no una natural, como el caso de reforestación plantando árboles alineados, un campo sembrado, etc.

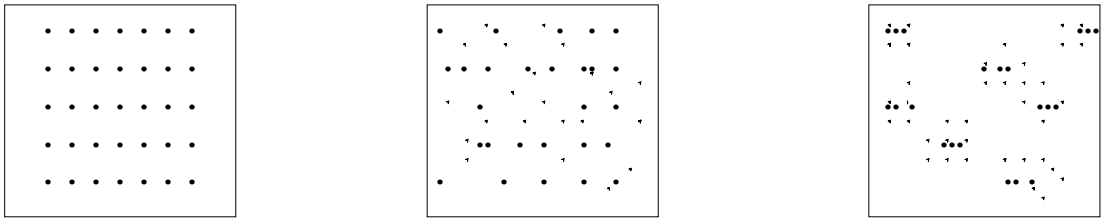
Se define como *Índice de Agregación* al cociente entre la varianza y el valor esperado de una variable aleatoria x cualquiera, o sea:

$$IA = \sigma_x^2 / \mu_x$$

Distribución de Poisson: $\sigma_x^2 = \mu_x \Rightarrow IA = 1$
Distribución Binomial Negativa: $\sigma_x^2 = r \cdot q / p^2$ y $\mu_x = r \cdot q / p \Rightarrow IA = 1 / p$
Distribución Uniforme: $\sigma_x^2 \approx 0 \Rightarrow IA \approx 0$

En el gráfico siguiente se esquematizan las tres situaciones planteadas, una plantación artificial muestra una distribución espacial uniforme y resulta nulo el Índice de agregación; en cambio, al mirar por el microscopio una dilución de glóbulos de sangre en el Hemocitómetro se puede ver una figura como la del cuadro del medio, donde IA = 1 pues se trata de una distribución poissoniana. Pero si no se le agrega EDTA en forma adecuada, las células se atraen por carga eléctrica, produciéndose el efecto *Rolleaux;* se agrupan como en un apilamiento y entonces el índice crece a valores mayores que la unidad. En Ecología, este agrupamiento de individuos se ve en los cardúmenes de peces, o en las manadas de animales, bandadas de pájaros, etc., según el continuo sea el agua, la tierra o el aire. Hay una especie de "contagio" entre los integrantes del grupo.

Gráfico 8.6 Índice de Agregación en diferentes casos.



IA = 0 Tipo uniforme | **IA = 1** Tipo Poisson | **IA > 1** Tipo Binomial Negativa

8.5.2 Muestreo de aceptación

En una relación de compraventa es usual tener un comprador que recibe un lote de piezas o elementos de un vendedor que se las provee. El problema básico es decidir si el comprador le acepta o no la mercadería enviada por el vendedor, de acuerdo con el resultado que obtenga al efectuarle una inspección para determinar la calidad de la misma. Cuando se trata de piezas muy valiosas como los diamantes, el comprador revisará con mucho cuidado cada pieza, una por una. Pero cuando se trate de una cantidad demasiado grande de elementos, como arroz a granel, o cuando el ensayo para probarla es destructivo como en el caso de un fósforo, o cuando el costo de la inspección sea muy elevado, entonces no revisará todo el lote sino una muestra del mismo.

El procedimiento usual se denomina *muestreo de aceptación* y consiste en tomar una muestra al azar de *n* unidades del lote comprado de tamaño *N*, revisar las piezas elegidas, y si se encuentran *r* o más piezas defectuosas se rechaza todo el lote. Caso contrario, se lo acepta.

Existen dos riesgos: el primero es *rechazar algo que se debería haber aceptado*; es el riesgo que corre el vendedor. Esto puede ocurrir porque en el muestreo de aceptación le encontraron las únicas piezas falladas del lote. A su probabilidad de ocurrencia se la suele llamar nivel de significación y se la simboliza con α. En Estadística, a ese tipo de equivocación se le designa con el nombre de Error de Tipo I. Por otra parte, la segunda manera de equivocarse es aceptar un lote con muchas piezas falladas porque en el muestreo de aceptación no apareció ninguna de ellas, esto es, *aceptar algo que se debería haber rechazado.* Ese es el riesgo del comprador y en estadística se lo llama Error de tipo II, y a su probabilidad de ocurrencia se la designa con el símbolo β (llamada la potencia del ensayo).

Cuando un profesional trabaje en un laboratorio de Control de Calidad en la industria farmacéutica, química, alimenticia, etc., se le presentará esta problemática en forma diaria. Por un lado, deberá hacer muestreos de aceptación para las compras de su industria, pero por otro lado deberá revisar la producción realizada para controlar la buena calidad del producto final que hace a la imagen de la empresa. Independientemente de los diferentes tipos de revisiones que le efectúe, el diagnóstico final será dicotómico: pieza aceptada o rechazada. Y el problema es que nunca se podrá conocer con certeza la cantidad total *R* de piezas falladas de todo el lote; solo una estimación de las mismas. Además, los riesgos de comprador y vendedor deben ser aceptables para ambas partes para que la operación de compraventa se concrete.

El procedimiento es simular el número de piezas falladas, para determinar los riesgos, como se muestra a continuación con un ejemplo numérico: un comprador recibe lotes de 250 unidades, de las cuales selecciona 50 al azar. Cuando encuentra 2 o más defectuosas rechaza el lote y se lo devuelve al vendedor. ¿Cómo se puede juzgar la bondad de esta regla de aceptación para ambas partes?

El método consiste en estimar la probabilidad de aceptar lotes con R defectuosos, tomando un valor esperado de p = R/N (donde p es el porcentaje de defectuosos). De esta forma, dando diferentes valores a R se van obteniendo los porcentajes, y se puede graficar esta situación en una gráfica denominada *Curva Característica de Operación* del plan de muestreo específico.
En este caso será una probabilidad hipergeométrica, en función del valor p, o sea:

$$P_H (r < 2 / N=250; R=250\ p, n=50) = P_H(r = 0) + P_H(r = 1)$$

$$P_H(p) = \frac{C(250-250\ p\ ;\ 50)}{C(250\ ;\ 50)} + \frac{C(250\ p\ ;\ 1)\ C\ (250-250\ p\ ;\ 49)}{C(250\ ;\ 50)}$$

Esta es la ecuación general de la probabilidad hipergeométrica en función del porcentaje de defectuosos. Ahora, dándole valores a p se obtienen los valores respectivos de $P_H(p)$ con los cuales se puede trazar la curva.

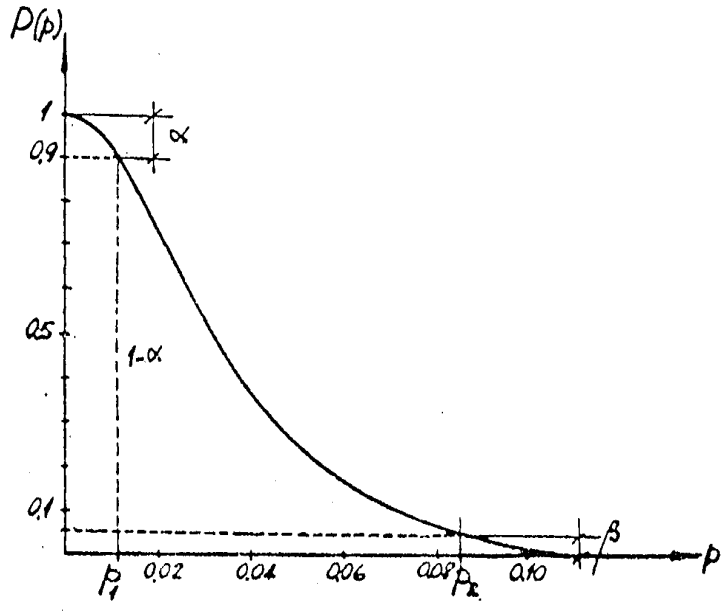
Por ejemplo, si p = 2% resulta:

$$P_H(p = 0{,}02) = \frac{C(245\,;\,50)}{C(250\,;\,50)} + \frac{C(5\,;\,1)\,C(245\,;\,49)}{C(250\,;\,50)} = \frac{245!\;200!}{250!\;195!}(1 + 250/196)$$

$$P_H(p=0{,}02) = \frac{245!\;200!\;446}{250!\;195!\;196} = 0{,}738 \qquad \text{O sea, un } 73{,}8\%$$

Análogamente, para valores de p variando de 0,4% cada vez, se logra la tabla del Gráfico 8.7. Para juzgar un plan de muestreo es necesario definir el porcentaje de defectuosos que el comprador está dispuesto a aceptar con mucha frecuencia. Por ejemplo, si el comprador está dispuesto a tolerar un total de R=3 piezas defectuosas por lote, entonces será $p_1 = 0{,}012$ el llamado: *Límite de Calidad Aceptable* (LCA). Para este problema, con un LCA del 1,2% se corre un riesgo $\alpha = 0{,}10$ de rechazar lotes, mientras que $P_H(0{,}012) = 0{,}9$. Lo cual es bastante razonable para el vendedor que en el 10% de los casos le rechazaran injustamente su venta.

Gráfico 8.7 Curva característica de operación



p	$P_H(p)$
0	1
0,004	1
0,008	0,96
0,012	0,897
0,016	0,821
0,02	0,738
0,024	0,655
0,028	0,575
0,032	0,501
0,036	0,432
0,04	0,371
0,044	0,316
0,048	0,268
0,052	0,226
0,056	0,19
0,06	0,159
0,064	0,132
0,068	0,11
0,072	0,091
0,076	0,075
0,08	0,062
0,084	0,05
0,088	0,041
0,092	0,034
0,096	0,027
0,1	0,022

Además, para que sea razonable para la otra parte, debe ocurrir que el plan de muestreo rechace con mucha frecuencia lotes con un porcentaje p_2 de defectuosos muy elevado. Este es un porcentaje que se está dispuesto a tolerar solo en raras ocasiones y se lo denomina: *Tolerancia porcentual de defectuosos en el lote* (TPD). Para este caso, si se toma $p_2 = 0{,}084$, la probabilidad de aceptar un lote que debería haber sido rechazado, o sea el riesgo del comprador, entonces será $\beta = 0{,}05$. En el 5% de las veces aceptará lotes con más de 3 defectuosos en total.

8.6 Teorema Central del Límite

Llamado el teorema fundamental de la estadística. Es el pilar que tomó casi un siglo hasta llegar a su demostración formal completa. Laplace lo publicó en su primera versión en 1812 y recién en 1901 Liapounoff los demostró bajo condiciones más generales, para todos los casos posibles. Su enunciado resumido dice:
"Sean n variables aleatorias independientes x_i con i =1,2,...n; cada una con distribuciones de probabilidad conocidas, en *condiciones muy generales*, la función distribución de la variable suma $(X = x_1 + x_2 + x_3 + ... + x_n)$, tipificada $\{Z = (X - \mu) / \sigma\}$ será aproximadamente igual a una función de Gauss tipificada $P_{Gauss}(z)$ si la secuencia de los x_i es ilimitada, esto es:

$$\lim_{n \to \infty} P\{((X - \mu) / \sigma) \leq a\} \approx F_{Gauss}(a)$$

En tal caso, se dice que la variable suma X es asintóticamente normal de parámetros N (μ ; σ).

¿Qué significa en condiciones muy generales? ... Por ejemplo, se cumple en los casos:
a) Es condición suficiente de validez, que todas las x_i tengan la misma función distribución.
b) Si ocurre que no tienen la misma función distribución, entonces es suficiente que cada una de ellas contribuya al total, con un aporte insignificante.
c) Cuando todas las variables tengan distribuciones de tipo continuo.
d) Si no son independientes, el teorema se sigue cumpliendo bajo ciertas condiciones.

Como se aprecia, esto es tan amplio que la mayoría de los casos reales se tomaban como teniendo distribuciones gaussianas con tal de tener un número muy grande de experimentos. Eso constituye la llamada Teoría de Grandes Muestras. Se deduce que si el número de casos n es lo suficientemente grande, las distribuciones vistas (Binomial, Poisson, Hipergométrica, Pascal, etc.) todas tienden a la función de Gauss. Lo que permite hacer aproximaciones para simplificar los cálculos de probabilidad.

Cuando una variable no parece distribuirse normalmente se pueden hacer transformaciones que ayuden a solucionar el problema, como por ejemplo tomar el logaritmo, el arco seno, etc. Cuando X es una variable aleatoria, tal que una función no lineal del tipo g (X) se distribuya normalmente, se dice que X tiene una *distribución normal transformada.*

Por ejemplo, si se supone que la indeterminación total (*error*) en una medición es producida por la suma de un gran número de causas aleatorias e independientes, cada una de ellas contribuyendo con un aporte despreciable frente al total (Hipótesis de Haeguen-Bessel), se está en el caso (b) descrito más arriba; por lo tanto, se puede asumir que la función distribución para los *errores casuales* es la función de Gauss o Normal, como se anunció al principio de este capítulo.

Desde su descubrimiento por Gauss en la primer década del siglo XIX, la curva se usó ampliamente en todas las mediciones físicas y químicas de aquel entonces. Prácticamente era la solución para cualquier caso práctico. Hasta 1907 donde Student muestra que el error de medición en el conteo con hemocitómetro (cámara de Neubauer), se parecía más a una distribución Poisson que a una de Gauss; fue el comienzo de la Estadística moderna.

8.7 Problemas propuestos

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

1) Todo fenómeno que no se puede explicar con certeza es de tipo aleatorio.	**V**	**F**
2) Un experimento es aleatorio cuando cada valor posible tiene una cierta probabilidad.	**V**	**F**
3) Una variable es aleatoria cuando algunos valores tienen asociadas ciertas probabilidades.	**V**	**F**
4) Si se conoce la función probabilidad, se conoce la función distribución.	**V**	**F**
5) Si se conoce la función distribución, se puede determinar la función de probabilidad.	**V**	**F**
6) La función distribución siempre es una sumatoria.	**V**	**F**
7) La función frecuencia es una especie de densidad de probabilidad.	**V**	**F**
8) La función densidad es una integral.	**V**	**F**
9) La esperanza matemática es un momento de segundo orden.	**V**	**F**
10) El valor esperado es un operador matemático.	**V**	**F**
11) La varianza es el momento central de segundo orden de una función probabilidad.	**V**	**F**
12) El riesgo relativo es lo mismo que el odd ratio.	**V**	**F**
13) Explicar la diferencia entre ambos conceptos desde un punto de vista epidemiológico.....	.....	
14) ¿ Que ocurre con los índices de riego OR y RR cuando el factor es un test clínico ?	.....	
15) Explicar la variabilidad de los índices de riesgo con la prevalencia de la población	.....	
16) Una variable tipificada tiene un valor esperado nulo.	**V**	**F**
17) La varianza de una variable aleatoria tipificada es 2.	**V**	**F**
18) Si la varianza es igual al valor esperado, se trata de un caso de Poisson.	**V**	**F**
19) Cuando el índice de agregación es la unidad se trata de una distribución uniforme.	**V**	**F**
20) Cuando se acepta algo que debía ser rechazado es un error de tipo I.	**V**	**F**
21) Cuando se rechaza algo que debía ser aceptado es un error del tipo II.	**V**	**F**
22) El error del tipo I es el riesgo del comprador.	**V**	**F**
23) El error del tipo II es el riesgo del vendedor.	**V**	**F**
24) La tolerancia porcentual de defectuosos en el lote es	.....	
25) El Límite de Calidad Aceptables es	.....	
26) Los valores esperados y las varianzas de un proceso Bernoulli y de Poisson son	.....	
27) Una variable se tipifica si se le resta el valor esperado y se la divide por su varianza.	**V**	**F**
28) El dominio de la función distribución son todos los números reales.	**V**	**F**

2) Calcular las varianzas y los valores esperados para todos los ejemplos desarrollados en el presente capítulo.

3) A una Farmacia llega cada hora una cantidad de pedidos que se muestran en la tabla siguiente. Se pide calcular para cada caso:
a) El número esperado de llegadas por hora.
b) La varianza de esta distribución de probabilidad.

Nº de pedidos	Probabilidad
0	0,05
1	0,10
2	0,15
3	0,25
4	0,30
5	0,10
6	0,05

NOTA: Usar la relación $\sigma^2 = \Xi \{(x^2)\} - \Xi^2 \{(x)\}$

4) Calcular el Riesgo Relativo y el Odds Ratio de las hipotéticas tablas siguientes:

Factor de Riesgo	Enfermedad SÍ	Enfermedad NO	Total
SÍ	10	90	100
NO	10	90	100
Total	20	180	200

Factor de Riesgo	Enfermedad SÍ	Enfermedad NO	Total
SÍ	10	10	20
NO	90	90	180
Total	100	100	200

Factor de Riesgo	Enfermedad SÍ	Enfermedad NO	Total
SÍ	90	10	20
NO	90	10	180
Total	180	20	200

Factor de Riesgo	Enfermedad SÍ	Enfermedad NO	Total
SÍ	90	90	180
NO	10	10	20
Total	100	100	200

A su vez calcular para cada caso la Prevalencia (probabilidad de estar enfermo), la probabilidad de exposición, el Odds de enfermos y el Odds de expuestos. Explicar clínicamente el significado de cada uno de los valores hallados para todos estos índices y extraer alguna conclusión.

5) Para los mismos datos de las cuatro tablas anteriores, suponer que el factor de riesgo es un método de diagnóstico (o un test clínico) y calcular:

a) Los índices básicos de calidad S y E

b) Los índices de calidad que no varían con la prevalencia (IY, LR+ y LR-)

c) Los índices de calidad que varían con la prevalencia (A, VPP y VPN)

d) Los índices de riesgo DR y RR.

e) Mostrar la variabilidad de RR con la prevalencia en forma gráfica.

6) Repetir los mismos cálculos que los del problema anterior para los datos siguientes:

Test Clínico	Enfermedad SÍ	Enfermedad NO	Total
(+)	180	50	230
(-)	20	150	170
Total	200	200	400