Bioestadística

1º Nutrición Humana y Dietética

Curso 2000-2001

Índice General

Capítulo 1

Introducción

La *Estadística* trata del estudio de un número finito de características de una determinada *población* (o conjunto de objetos). Por ejemplo: peso y altura de los individuos de un colectivo concreto. En la práctica, suele resultar imposible el estudio directo de dichas características en cada objeto de la población, debido normalmente a su tamaño. Por ejemplo: determinar el peso medio de los habitantes de una nación. Para evitar este contratiempo, se realiza un estudio previo en una parte de la población, llamada *muestra*, y esperamos que la información obtenida en dicha muestra nos permita determinar, de alguna manera, el comportamiento de dichas características en toda la población.

Con estas precisiones, al realizar un tratamiento estadístico de una población, los pasos a seguir son los siguientes:

1– Elección de la muestra, o subconjunto de la población. Aunque hay muchas maneras de elegir una muestra, habitualmente se supondrá que cada elemento de la población tiene la misma *probabilidad* de elegirse.

2– Estudio de los objetos de dicha muestra.

3– Organización de la información obtenida, es decir ordenar datos, dar tablas, representaciones gráficas, etc. Esta parte es lo que se conoce como *Estadística Descriptiva*.

4– Obtener conclusiones sobre la población general, partiendo de los datos obtenidos en la muestra. De esto se ocupa la *Inferencia Estadística*.

Finalmente, las conclusiones se formulan con un *margen de error*. En esta situación, dicho error se expresa en términos de *Probabilidad*.

Llegados a este punto lo primero que habría que empezar a estudiar es la

Estadística Descriptiva. De hecho, lo que se estudia son las llamadas *variables estadísticas*, o funciones que a cada objeto de una población le asocian un determinado valor. Por ejemplo peso, o peso y altura, o peso, altura y edad, etc. Así, la Estadística se ocupa del estudio de dichas variables en una población. Por lo tanto, el método a seguir consiste en estudiar dichas variables en una muestra y a partir de dicho estudio inferir el comportamiento de dichas variables en la población general. Pero, como las variables estadísticas están estrechamente relacionadas con las *variables aleatorias*, estudiaremos estas últimas y de paso introduciremos, al mismo tiempo, los conceptos más usuales de la Estadística Descriptiva. Previamente a su estudio, comenzaremos con el Cálculo de Probabilidades.

Capítulo 2

Cálculo de Probabilidades

2.1 Espacio Muestral. Sucesos

En el mundo real existen fenómenos que siguen unas leyes determinadas. Es decir, dadas unas condiciones, el resultado del fenómeno es previsible, salvo errores de medida. Por ejemplo, manteniendo la velocidad constante, podemos determinar “exactamente” la distancia recorrida por un coche en un determinado momento. Otro ejemplo sería determinar la velocidad, en un determinado instante, que tiene un objeto cuando cae desde una determinada altura. Estos fenómenos se llaman *deterministas*.

Sin embargo, hay otro tipo de fenómenos que no siguen leyes determinadas. Por ejemplo el lanzamiento de una moneda o de un dado, son fenómenos tales que no podemos predecir el resultado. Este tipo de fenómenos se llaman *aleatorios*, y su característica esencial es que en su realización pueden darse varios resultados sin que podamos predecir con certeza cual de ellos va a ocurrir.

En muchas disciplinas científicas se estudian fenómenos aleatorios. Por ejemplo, la frecuencia de aparición de un determinado rasgo étnico (en Genética), el volumen de ventas de una determinada mercancía (Mercadotecnia), el número de productos defectuosos en un proceso de fabricación (control de calidad), etc.

Al describir un fenómeno aleatorio, es esencial especificar que aspecto del resultado nos interesa observar. Es decir, hemos de fijar con precisión el criterio seguido para determinar cuando dos resultados son diferentes. Por ejemplo, al lanzar una moneda los dos resultados distintos que podemos obtener son que salga cara o cruz. Al lanzar una moneda dos veces, puede que nos interese solamente el número de caras y de cruces que hayan salido, en cuyo caso los resultados distintos son 2 caras, 2 cruces, cara y cruz. Sin embargo, si nos interesa también el orden de aparición, los resultados posibles son cara y cara, cara y cruz, cruz y cara, cruz y cruz. Esta especificación se logra mediante el espacio muestral. Vamos a precisar este concepto.

Dado un experimento aleatorio, el conjunto formado por todos los resultados diferentes de dicho experimento, o que consideramos diferentes, recibe el nombre de *espacio muestral* y se denota por Ω. Por ejemplo, al lanzar un dado, el espacio muestral es

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Cada uno de los elementos del espacio muestral se llama *suceso elemental*. Así por ejemplo, en el ejemplo anterior, un suceso elemental es "que salga 4". A todo subconjunto A del espacio muestral Ω se le llama *suceso* (aunque en general esto no es cierto para espacios muestrales con un número infinito de sucesos elementales). Concretamente, un suceso asociado a un experimento aleatorio es un posible resultado de dicho experimento. Vamos a dar algunos ejemplos de sucesos del experimento anterior, lanzar un dado.

A=que salga un número par=2,4,6,
B=que salga un número múltiplo de 3=3,6,
C=que salga un número menor que 4=1,2,3,
D=que salga un número impar=1,3,5,
E=que salga un número mayor que 4=5,6.

Dos sucesos "triviales" son el *suceso imposible*, denotado por $\emptyset$ (conjunto vacío), y el *suceso seguro*, o espacio muestral (que siempre ocurre).

Con los suceso se pueden realizar diversas operaciones, que pasamos a detallar:

Unión de sucesos: dados dos sucesos A y B de un mismo espacio muestral, se define la unión de A y B, y se representa por $A \cup B$, al suceso que ocurre siempre que ocurra el suceso A o el suceso B. En el ejemplo anterior $C \cup D$=1,2,3,5.

Intersección de sucesos: dados dos sucesos A y B de un mismo espacio muestral, se define la intersección de A y B, y se representa por $A \cap B$, al suceso que ocurre siempre que ocurran los sucesos A y B simultáneamente. En el ejemplo anterior $C \cap D$=1,3, $A \cap D$=$\emptyset$ (es decir los sucesos A y D no pueden ocurrir simultáneamente puesto que al lanzar un dado no puede salir un número que sea a la vez par e impar).

Suceso complementario: dado un suceso A, se define el suceso complementario de A, y se denota por A^c o A^*, al suceso que ocurre siempre que no ocurra A. En el ejemplo anterior $A^c = D$.

Es fácil comprobar que se cumplen las siguientes propiedades:

1) $A^c \cup A = \Omega$
2) $A^c \cap A = \emptyset$
3) $(A^c)^c = A$
4) $\Omega^c = \emptyset$.

Ejercicio: *Demostrar las leyes de Morgan para sucesos:*

1) $(A \cup B)^c = A^c \cap B^c$
2) $(A \cap B)^c = A^c \cup B^c$.

La unión e intersección de sucesos cumplen las siguientes propiedades:

$$\textit{son asociativas: } \begin{cases} A \cup (B \cup C) = (A \cup B) \cup C \\ A \cap (B \cap C) = (A \cap B) \cap C \end{cases}$$
$$\textit{son conmutativas: } \begin{cases} A \cup B = B \cup A \\ A \cap B = B \cap A \end{cases}$$
$$\textit{son distributivas: } \begin{cases} A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \\ A \cap (B \cup C) = (A \cap B) \cup (B \cap C) \end{cases}$$
$$\textit{tienen elemento neutro: } \begin{cases} \emptyset \text{ para la unión } A \cup \emptyset = A \\ \Omega \text{ para la intersección } A \cap \Omega = A \end{cases}$$

Dos sucesos A y B, de un mismo espacio muestral, se dicen que son *incompatibles* si su intersección es el suceso imposible, es decir $A \cap B = \emptyset$. Además, se dice que A está contenido en B, y se representa por $A \subset B$, si siempre que ocurra A ocurre B. Es claro que para cualquier suceso A siempre se cumple que $\emptyset \subset A \subset \Omega$.

Dado un suceso A de un determinado experimento aleatorio, se define la *frecuencia absoluta* del suceso A cuando el experimento se realiza n veces al número de veces que ocurre el suceso A en esa serie de repeticiones del experimento, y se denota por n_A. La *frecuencia relativa* de A se define como

$$f_A = \frac{n_A}{n}.$$

Observemos que para cualquier suceso A se cumple que $0 \leq f_A \leq 1$. Además $f_\Omega = 1$ y $f_\emptyset = 0$. Cuando se repite "muchas veces" un experimento (es decir cuando n es "grande"), la frecuencia relativa de un suceso tiende a estabilizarse al rededor de un número. Esta es la base de la probabilidad. Por ejemplo, al lanzar un dado varias veces, la frecuencia relativa del suceso salir 1 tiende a estabilizarse alrededor de $1/6$. Así diremos que la probabilidad de que al lanzar un dado salga el 1 es $1/6$. En la siguiente sección vamos a introducir el concepto de probabilidad con más rigor.

2.2 Probabilidad

En la sección anterior, al final, hemos indicado alguna idea sobre la posible definición de probabilidad (frecuencia relativa). No obstante hemos de observar que la posible definición de probabilidad ha de aplicarse no sólo a elementos de un conjunto (espacio muestral), sino también a subconjuntos de él mismo (sucesos). Por todo esto vamos a definir una estructura sobre "todos" los subconjuntos de un conjunto dado.

Dado un conjunto M, la clase de subconjuntos de M, $\mathcal{C}(M)$ que cumpla las propiedades siguientes:

$$\begin{array}{l} 1) M \in \mathcal{C}(M), \\ 2) \text{si } A, B \in \mathcal{C}(M) \Rightarrow A \cup B \in \mathcal{C}(M), \\ 3) \text{si } A \in \mathcal{C}(M) \Rightarrow A^c \in \mathcal{C}(M), \end{array}$$

recibe el nombre de *álgebra de Boole*. En el caso que nos ocupa, si Ω es un espacio muestral y $\mathcal{C}(\Omega)$ es la clase formada por todos los sucesos, o todos los subconjuntos, de Ω, entonces se cumple que

$$\begin{array}{l} 1) \Omega \in \mathcal{C}(\Omega), \text{ puesto que } \Omega \text{ es un suceso,} \\ 2) \text{si } A, B \in \mathcal{C}(\Omega) \Rightarrow A \cup B \in \mathcal{C}(\Omega), \text{ ya que la unión de sucesos es un suceso,} \\ 3) \text{si } A \in \mathcal{C}(\Omega) \Rightarrow A^c \in \mathcal{C}(\Omega). \end{array}$$

Por lo tanto, la clase formada por todos los sucesos de un experimento aleatorio tiene estructura de álgebra de Boole, llamada álgebra de Boole de sucesos o álgebra de sucesos. Con este concepto, pasamos a definir axiomáticamente la probabilidad.

Dado un experimento aleatorio, con espacio muestral Ω y álgebra de sucesos asociada $\mathcal{C}(\Omega)$, una aplicación

$$P : \mathcal{C}(\Omega) \to [0,1]$$

se dice que es una *probabilidad* sobre Ω si satisface los tres axiomas siguientes:

$$\begin{array}{l} 1) P(A) \geq 0, \forall A \in \mathcal{C}(\Omega), \\ 2) P(\Omega) = 1, \\ 3) \text{si } A \cap B = \emptyset, \text{ entonces } P(A \cup B) = P(A) + P(B). \end{array}$$

Observemos que el primer axioma establece que la probabilidad de cualquier suceso es mayor o igual que 0, al igual que ocurre con la frecuencia relativa. Notemos también que este axioma puede eliminarse puesto que hemos considerado que la función P (probabilidad) sólo puede tomar valores pertenecientes al intervalo [0,1]. El segundo axioma nos indica que la probabilidad del suceso seguro es 1, en concordancia con el hecho de que la frecuencia relativa del suceso seguro es 1. Finalmente, el tercer axioma establece que para sucesos incompatibles la probabilidad de la unión es la suma de las probabilidades. La probabilidad cumple diversas propiedades. Algunas de ellas son las siguientes:

$\boxed{1) P(\emptyset) = 0}$

En efecto, por el axioma 2 resulta que $1 = P(\Omega) = P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset)$ (por el axioma 3 puesto que $\Omega \cap \emptyset = \emptyset$). Luego $1 = P(\Omega) + P(\emptyset) = 1 + P(\emptyset) \Rightarrow P(\emptyset) = 0$.

$\boxed{2) P(A^c) = 1 - P(A)}$

$1 = P(A^c \cup A) = P(A^c + P(A))$.

$\boxed{3) \text{ si } A \subset B \Rightarrow P(A) \leq P(B)}$

$B = B \cap \Omega = B \cap (A^c \cup A) = (B \cap A^c) \cup (B \cap A) = (B \cap A^c) \cup A$, que son sucesos incompatibles. Por lo tanto (axioma 3) resulta que $P(B) = P(B \cap A^c) + P(A) \geq P(A)$.

$\boxed{4) P(A \cup B) = P(A) + P(B) - P(A \cap B)}$

Observemos en primer lugar que $A \cup B = (A \cup B) \cap (B \cup B^c) = [(A \cup B) \cap B] \cup [(A \cup B) \cap B^c] = B \cup (A \cap B^c)$, sucesos incompatibles. Luego, por el axioma 3 resulta que

$$P(A \cup B) = P(B) + P(A \cap B^c) \ (*)$$

De otro lado, como $A = (A \cap B) \cup (A \cap B^c)$, incompatibles, resulta que $P(A) = P(A \cap B) + P(A \cap B^c)$. Es decir

$$P(A \cap B^c) = P(A) - P(A \cap B).$$

Ahora, sustituyendo esta última expresión en $(*)$ queda demostrada la propiedad 4.

$\boxed{5)}$ Sean $A_1, A_2, \ldots, A_n$ sucesos incompatibles dos a dos, es decir $A_i \cap A_j = \emptyset$ para $i \neq j$ con $i, j = 1, 2, \ldots n$. Entonces se tiene que

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

La demostración se hace por inducción sobre n y aplicando el axioma 3.

En un espacio muestral finito, si cada elemento o suceso elemental tiene la misma probabilidad, caso de un espacio *equiprobable*, la probabilidad de un suceso es igual al número de elementos de dicho suceso dividido por el número de elementos del espacio muestral. Es decir, la regla clásica de casos favorables entre casos posibles. En el ejemplo del lanzamiento de un dado, cada suceso elemental (que salga un número concreto) tiene la misma probabilidad (igual a 1/6). Luego, en el ejemplo de la sección anterior obtenemos las siguientes probabilidades $P(A) = P(C) = P(D) = 1/2, P(B) = P(E) = 1/3$. Este tipo de espacios (equiprobables y finitos) son los que vamos a tratar con más detenimiento. Luego en ellos, el problema de la probabilidad de un suceso se reduce a determinar el número de casos favorables (o elementos del suceso) y el de casos posibles (o elementos del espacio muestral). Una herramienta útil para estos problemas es el cálculo combinatorio.

2.3 Cálculo Combinatorio

Supongamos que tenemos m elementos distintos y sea $n \leq m$. Llamamos $V_{m,n}$ *variaciones de* m *elementos tomados de* n en n, al número de subconjuntos ordenados formados cada uno por n de dichos m elementos, siendo dos variaciones distintas si difieren en algún elemento, o si constando de los mismos, se disponen en distinto orden. Si $n = m$ el número $V_{n,n}$ se denomina *permutaciones* de n elementos y se denota por P_n. El número de variaciones, y de permutaciones, viene dado por:

$$V_{m,n} = m(m-1)(m-2)\cdots(m-n+1)$$

$$P_n = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1 = n!$$

Así entonces

$$V_{m,n} = \frac{m!}{(m-n)!} = \frac{P_m}{P_{(m-n)}}.$$

Ejemplo: *¿Cuántos números de* 4 *cifras distintas se pueden formar con los números* $1,2,3,4,5,6,7,8$*?*
$8\cdot 7\cdot 6\cdot 5 = 8!/4!$

Ejemplo: *¿De cuántas formas distintas pueden sentarse* 5 *personas en fila?*
$P_5 = 5!$

$C_{m,n}$ *combinaciones de m elementos tomados de n* en n, al número de subconjuntos formados cada uno por n de dichos m elementos, siendo dos combinaciones distintas si difieren en algún elemento. El número de combinaciones viene dado por:

$$C_{m,n} = \binom{m}{n} = \frac{V_{m,n}}{P_n} = \frac{m!}{n!(m-n)!}.$$

Ejemplo: *¿De cuántas maneras se pueden extraer* 3 *cartas de un conjunto de* 40 *?*

$$C_{40,3} = \binom{40}{3} = \frac{V_{40,3}}{P_3} = \frac{40!}{3!(37)!} = 9880.$$

Sean ahora m elementos distintos y n arbitrario (pudiendo ser mayor que m). Llamamos
V_m^n *variaciones con repetición de m elementos tomados de n* en n, al número de subconjuntos ordenados formados cada uno por n de dichos m elementos, pudiéndose repetir uno o varios, siendo dos variaciones con repetición distintas si difieren en algún elemento, o si constando de los mismos, se disponen en distinto orden. $V_m^n = m^n$.

Ejemplo: *¿Cuántos números distintos de* 3 *cifras pueden formarse con los números* $1,2,3,4$*?*. SOLUCIÓN: $V_4^3 = 4^3 = 64$.

Supongamos ahora que $m_1 + m_2 + \cdots + m_r = m$. Llamamos $RP_m^{m_1,m_2,\ldots,m_r}$ *permutaciones con repetición* de m elementos de los cuales m_1 son iguales, m_2 son iguales, ..., m_r son iguales, al número de conjuntos ordenados formados por dichos elementos, siendo dos permutaciones distintas si los elementos se disponen en distinto orden. Su número es

$$RP_m^{m_1,m_2,\ldots,m_r} = \frac{m!}{m_1!m_2!\cdots m_r!}.$$

Ejemplo: *¿Cuántos números distintos constan de los números* $1, 1, 2, 3, 3, 3$*?*

$$RP_6^{2,1,3} = \frac{6!}{2!1!3!} = 60.$$

2.4 Probabilidad Condicionada. Independencia

Supongamos que en un pueblo viven 200 personas, distribuidas según su sexo y religión de la forma siguiente:

	Católicos	Musulmanes	
Hombres	40	60	100
Mujeres	30	70	100
	70	130	200

Elijamos ahora una persona al azar. ¿Cuál será la probabilidad de que sea hombre? Es claro que

$$P(\text{Hombre}) = \frac{100}{200} = 0'5.$$

Supongamos ahora que tenemos cierta información. Por ejemplo, sabemos que la persona es de religión musulmana. ¿Cuál será ahora la probabilidad de que sea hombre? En este caso resulta que

$$\text{``}P(\text{Hombre})'' = \frac{60}{130} = \frac{6}{13}.$$

Luego, al tener cierta información, la probabilidad varía. esta información adicional es lo que va a dar lugar al concepto de probabilidad condicionada.

Dados dos sucesos A y B, tal que $P(B) > 0$, definimos la probabilidad de que ocurra el suceso A sabiendo que se realiza B (o *probabilidad* del suceso A *condicionada* al suceso B) como

$$P(A/B) = \frac{P(A \cap B)}{P(B)}.$$

En el ejemplo anterior

$$P(\text{Hombre}/\text{Musulmán}) = \frac{P(\text{Hombre} \cap \text{Musulmán})}{P(\text{Musulmán})} = \frac{60/200}{130/200} = \frac{6}{13}.$$

En la noción de probabilidad condicionada surge una cuestión importante. En concreto, en el ejemplo anterior, la probabilidad de que ocurra el suceso A (Hombre) es distinta a la probabilidad de que ocurra A sabiendo que el suceso B (Musulmán) ocurre. Es decir $P(A/B) \neq P(A)$. O lo que es lo mismo, la información adicional que tenemos ("ocurre B") influye el la probabilidad de que ocurra A. Por lo tanto podemos pensar que ambos sucesos tienen alguna relación, es decir existe alguna dependencia entre ellos. En otros casos este hecho no tiene por que ocurrir. Por ejemplo, si consideramos el experimento aleatorio: lanzar una moneda dos veces y los sucesos A=cara en el primer lanzamiento, B=cruz en el segundo lanzamiento, como $A \cap B$=cara, cruz, resulta que

$$P(A) = P(B) = \frac{1}{2}, \quad P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{1/4}{1/2} = \frac{1}{2} = P(A).$$

Por lo tanto, en este ejemplo, $P(A/B) = P(A)$. Es decir, la información adicional salir cruz en el segundo lanzamiento no influye en la realización del suceso A. Así pues, estos dos sucesos, en principio, no parece que tengan ninguna relación. Observemos también que, de igual manera, obtenemos $P(B/A) = P(B)$. Además, en este caso se cumple la igualdad (interesante) siguiente: $P(A \cap B) = P(A)P(B)$. Este hecho es el que va a dar lugar al concepto de sucesos independientes.

Sean A y B dos sucesos, del mismo espacio muestral. Se dice que A y B son *independientes* si

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

o lo que es lo mismo, dos sucesos A y B son independientes si $P(A/B) = P(A)$, o bien $P(B/A) = P(A)$. Para tres sucesos, la definición de independencia es que sean independientes dos a dos y los tres conjuntamente. Es decir, los sucesos A, B y C son *independientes* si se cumplen las cuatro condiciones siguientes:

$$\begin{cases} P(A\cap B) = P(A)P(B) \\ P(A\cap C) = P(A)P(C) \\ P(B\cap C) = P(B)P(C) \\ P(A\cap B\cap C) = P(A)P(B)P(C). \end{cases}$$

Con cierta frecuencia aparecen en el cálculo de probabilidades problemas que consisten en la repetición de un determinado experimento. Dichos problemas se pueden reducir a un experimento consistente en ir sacando sucesivas bolas de una urna, con o sin reposición. Para resolverlos es útil el concepto de probabilidad condicionada. Más concretamente un resultado conocido como teorema de la probabilidad compuesta. Dicho resultado es:

Teorema de la Probabilidad Compuesta: *Sean $A_1, A_2, \ldots, A_n$, n sucesos de un mismo espacio muestral tales que $P(A_1\cap A_2\cap\ldots\cap A_n) > 0$. Entonces se tiene que*

$$\begin{aligned} P(A_1\cap A_2\cap\ldots\cap A_n) &= P(A_1)P(A_2/A_1)P(A_3/A_1\cap A_2)\cdots \\ &= \cdots P(A_n/P(A_1\cap A_2\cap\ldots\cap A_{n-1}). \end{aligned}$$

Para demostrar este resultado,por inducción, basta con aplicar la definición de probabilidad condicionada. Queda como ejercicio.

Veamos ahora un ejemplo en el que aplicar el teorema anterior.

Ejercicio: *consideramos una urna con 7 bolas blancas, 3 negras y 5 azules. Extraemos de dicha urna 3 sin reposición. Calcular la probabilidad de que las bolas extraídas sean, por este orden, blanca, azul y blanca. Calcular lo mismo si las extracciones se hacen con reposición.*

Para resolverlo, en primer lugar vamos a denotar por A al suceso blanca en la primera, B a azul en la segunda y C a blanca en la tercera. Así, hemos de calcular la probabilidad del suceso $A\cap B\cap C$. Luego, por el teorema anterior dicha probabilidad es

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

Sin Reposición: en este caso es fácil comprobar que $P(A) = 7/15$, $P(B/A) = 5/14$ y $P(C/A\cap B) = 6/13$. Luego $P(A\cap B\cap C) = 1/13 \simeq 0'0769230$.

Con Reposición: en esta situación resulta que $P(A) = 7/15$, $P(B/A) = 5/15$ y $P(C/A\cap B) = 7/15$. Luego $P(A\cap B\cap C) = 49/675 \simeq 0'0725925$.

Un problema bastante frecuente en el cálculo de probabilidades es el considerado esquema de urnas y bolas. Partimos de una serie de urnas, con una cierta probabilidad de ser elegidas, y cada una de ellas tiene bolas de diferentes colores. El experimento consiste en elegir una urna y extraer una bola. Una posible pregunta es la siguiente: ¿Cuál será la probabilidad de que la bola extraída sea de un determinado color? La otra pregunta es: si la bola es de un determinado color, ¿Cuál será la probabilidad de que hayamos elegido una determinada urna? Para resolver estas dos cuestiones vamos a demostrar dos resultados, íntimamente relacionados con la noción de probabilidad condicionada. Previamente vamos a plantear un ejemplo concreto, demostraremos los resultados y resolveremos el problema.

Ejercicio: *se tienen tres urnas, A con 2 bolas blancas y 3 negras, B con 2 bolas blancas y 2 negras y C con 5 bolas blancas y 4 negras. La probabilidad de elegir cada urna es 1/2 para la A y 1/4 para la B y C. Elegimos una urna y sacamos una bola. Sea D el suceso sacar una bola blanca. Calcular la probabilidad de D. Supongamos ahora que hemos extraído una bola negra, suceso que vamos a denotar por E. Calcular la probabilidad de haber elegido la urna C.*

La primera pregunta se responde con el teorema de la probabilidad total, que dice lo siguiente:

Teorema de la Probabilidad Total: *Sea el conjunto de sucesos $A_1, A_2, \ldots,$ $, A_n$, que cumplen las siguientes condiciones:*

$$
\begin{array}{l}
i)\bigcup_{i=1}^{n} A_i = \Omega, \quad \text{el espacio muestral,} \\
ii) A_i \cap A_j = \emptyset, \quad i \neq j,\ i,j = 1,2,\ldots,n, \\
iii) P(A_i) > 0, \quad i = 1,2,\ldots,n.
\end{array}
$$

Consideremos ahora un suceso B. Entonces se cumple que

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B/A_i).$$

Para el caso de dos sucesos A y B con $P(A)P(A^c) > 0$, resulta

$$P(B) = P(A)P(B/A) + P(A^c)P(B/A^c)$$

DEMOSTRACIÓN

Observemos en primer lugar que como $B = B \cap \Omega$, resulta que

$$B = B \cap \Omega = B \cap \left(\bigcup_{i=1}^{n} A_i \right) = \bigcup_{i=1}^{n} (B \cap A_i)$$

siendo los sucesos $B \cap A_i$ incompatibles dos a dos, puesto que por hipótesis los A_i también lo son. Luego por la propiedad 5 de las probabilidades resulta que

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i).$$

Finalmente, como para todo i se tiene que $P(B \cap A_i) = P(A_i)P(B/A_i)$, de la igualdad anterior se deduce el teorema.

$\square$

Ahora, estamos en condiciones de responder a la primera pregunta del ejemplo de urnas y bolas, aplicando el teorema anterior. Para ello, vamos a denotar por A, B y C a los sucesos elegir las urnas A, B y C, respectivamente. Notemos que estos tres sucesos cumplen las condiciones exigidas a los A_i en el teorema anterior. Luego, aplicando dicho resultado, obtenemos para el suceso D lo siguiente:

$$\begin{aligned} P(D) &= P(A)P(D/A) + P(B)P(D/B) + P(C)P(D/C) = \\ &= \frac{1}{2}\frac{2}{5} + \frac{1}{4}\frac{2}{4} + \frac{1}{4}\frac{5}{9} = \frac{167}{360} \simeq 0'463888. \end{aligned}$$

Para la segunda cuestión, utilizaremos el teorema de Bayes que dice

Teorema de Bayes: *Sea el conjunto de sucesos $A_1, A_2, \ldots, A_n$, que cumplen las siguientes condiciones:*

$$\begin{aligned} &i) \bigcup_{i=1}^{n} A_i = \Omega, \quad \text{el espacio muestral,} \\ &ii) A_i \cap A_j = \emptyset, \quad i \neq j,\ i, j = 1, 2, \ldots, n, \\ &iii) P(A_i) > 0, \quad i = 1, 2, \ldots, n. \end{aligned}$$

Consideremos ahora un suceso B tal que $P(B) > 0$. Entonces, para todo $j = 1, 2, \ldots, n$ se cumple que

$$P(A_j/B) = \frac{P(A_j)P(B/A_j)}{\sum_{i=1}^{n} P(A_i)P(B/A_i)}.$$

Para el caso de dos sucesos A y B con $P(A)P(A^c) > 0$, resulta

$$P(B) = \frac{P(A)P(B/A)}{P(A)P(B/A) + P(A^c)P(B/A^c)}$$

DEMOSTRACIÓN

En primer lugar observemos que

$$P(A_j/B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(A_j)P(B/A_j)}{P(B)}.$$

Ahora, aplicando el teorema le la probabilidad total, se deduce este teorema.

□

Para responder a la segunda cuestión planteada en el problema de urnas y bolas, siguiendo la notación utilizada para la primera pregunta, y aplicando el teorema de Bayes, obtenemos:

$$P(C/E) = \frac{P(C)P(E/C)}{P(A)P(E/A) + P(B)P(E/B) + P(C)P(E/C)} = \frac{40}{193} \simeq 0'207253.$$

2.5 Ejercicios

1) Consideramos el experimento lanzar un dado. Sean los sucesos:

$$A = \{1, 2\}, \quad B = \{2, 4, 6\}, \quad C = \{1, 3\}, \quad D = \{2, 5, 6\}.$$

Determinar los sucesos siguientes:

1)$A \cup B$, 2)$A \cap B^c$, 3)$C \cap D$, 4)$A^c \cap B^c$, 5)$(A \cup B \cup D)^c$, 6)$C^c \cap A^c$.

2) Lanzamos una moneda 3 veces. Construir el espacio muestral y calcular la probabilidad de los siguientes sucesos:

a) salen 2 caras exactamente	b) salen al menos 2 cruces
c) sale como mucho una cara	d) la primera vez sale cara
e) la segunda y la tercera son cruces	f) la segunda vez no es cara

3) Un dado se lanza dos veces. Sean los sucesos:

$A = \{$en el primer lanzamiento, el número es menor o igual que 3$\}$
$B = \{$en el segundo lanzamiento, el número es mayor o igual que 5$\}$
$C = \{$en ambos lanzamientos, el número es par$\}$
$D = \{$al menos en uno de los dos lanzamientos, el número es impar$\}$.

Calcular

$$P(A),\ P(B),\ P(C),\ P(A \cup B),\ P(B \cup C),\ P(A \cap C),\ P(B \cap C),\ P(D).$$

4) ¿De cuántas maneras distintas se pueden poner en fila 5 fichas de colores distintos? ¿Y en círculo?

5) Calcular el número de permutaciones distintas que pueden formarse con las letras de las palabras

a) MEDICINA, b) ESTETOSCOPIO, c) FARMACIA.

6) Supongamos que un dietista dispone de los siguientes alimentos: 5 con la vitamina A, 7 con la vitamina B y 2 con la C. Calcular el número de comidas distintas que puede elaborar, suponiendo que contengan:

a) un alimento con la vitamina A, uno con la B y uno con la C,
b) 3 alimentos con la vitamina A,
c) 2 alimentos con la vitamina A y 3 con la B,
d) 2 con la vitamina A, 3 con la B y 1 con la C,
e) 4 con la vitamina A, 4 con la B y 1 con la C.

7) Se colocan 5 hombres y 4 mujeres en fila, de manera que las mujeres ocupen los lugares pares. ¿De cuántas maneras puede hacerse?

8) En una estantería vamos a colocar 4 libros diferentes de Matemáticas, 6 de Física y 2 de Química. Calcular el número de colocaciones distintas que pueden hacerse si

a) los libros de cada materia han de estar juntos,
b) los libros de Matemáticas han de estar juntos,
c) los libros de Química han de estar en los extremos de la estantería.

9) De entre 5 matemáticos y 7 físicos hay que constituir una comisión formada por 2 matemáticos y 3 físicos. Calcular las formas posibles de hacerse si:

a) todos son elegibles,
b) un físico en particular ha de estar en la comisión,
c) dos matemáticos tienen prohibido pertenecer a la comisión.

10) Calcular todas las formas posibles de colocar 7 libros en una estantería si

a) cualquier colocación es admitida,
b) 3 libros particulares han de estar juntos,
c) 2 libros particulares han de estar en los extremos de la estantería.

11) Tenemos los números 2,3,5,7,9, y queremos formar números de 3 cifras.

a) ¿Cuántos podemos formar sin que se repitan las cifras?,
b) De ellos ¿Cuántos son pares?,
c) ¿Cuántos son impares?,
d) ¿Cuántos son menores que 400?,
d) ¿Cuántos son múltiplos de 5?

Repetir las mismas cuestiones si se pueden repetir cifras.

12) Calcular la probabilidad de los siguientes sucesos:

a) que aparezca el as de copas, la sota de bastos o el 7 de espadas, al sacar una carta de una baraja de 40.

b) Que al lanzar dos dados, la suma sea 7, 11 o 12.

13) Un experimento consiste en lanzar un dado y una moneda. sean los sucesos:

$$A = \{\text{cara}\}, \quad B = \{3 \text{ o } 6\}, \quad C = \{\text{par en el dado}\}, \quad D = \{\text{impar en el dado}\}.$$

Describir los siguientes sucesos y calcular sus probabilidades:

a)A^c	b)B^c	c)$A \cap B$
d)$A \cup B$	a)$A \cap B \cap C$	f)$A^c \cap C$
g)$B \cap C$	h)$A \cap B \cap D$	i)$A \cup C \cup D$.

14) Una urna tiene 6 bolas rojas, 4 blancas y 5 azules. Si sacamos una bola sea

a) roja, b) blanca, c) azul, d) no roja, e) roja o blanca.

Si sacamos una bola, se repone y realizamos una nueva extracción, calcular la probabilidad de los siguientes sucesos:

a) ambas bolas rojas,	b) la primera no es azul,
c) al menos una blanca,	d) la primera es azul y la segunda no,
e) a lo sumo una roja,	f) ambas son rojas o azules o una de cada,
g) las dos son de colores distintos,	h) la primera es blanca y la segunda roja,
i) las dos son del mismo color,	j) la segunda es blanca.

Repetir las mismas cuestiones si no hay reposición.

15) Una caja contiene 8 bolas rojas, 3 blancas y 9 azules. Si sacamos 3 bolas al azar, calcular la probabilidad de que

a) las tres sean rojas,	b) las tres sean blancas,
c) dos sean rojas y una blanca,	d) al menos una blanca,
e) una de cada color,	f) las tres del mismo color.

16) Tres joyeros tienen dos relojes cada uno. El primero 2 de oro, 2 de plata el segundo y uno de oro y otro de plata el tercero. Elegimos un joyero y sacamos un reloj. Calcular la probabilidad de que el reloj sea de oro. Si el reloj es de oro, calcular la probabilidad de haber elegido el primer joyero. Si es de plata, calcular la probabilidad de haber elegido el tercero.

17) En un laboratorio tenemos drogas del tipo A, B y C. Si se inyecta la droga A, la probabilidad de desarrollar cierta antitoxina es 1/4, 1/8 para la B y 1/3 para la C. Hay 2 frascos con la droga A, 3 con la B y 1 con la C. Elegimos un frasco al azar e inyectamos a un paciente la droga. Calcular la probabilidad de que se produzca la antitoxina. Si no se produce la antitoxina, calcular la probabilidad de haber inyectado la droga B. Si se produce la antitoxina, calcular la probabilidad de haber inyectado la droga C.

18) Una urna contiene 5 bolas blancas y 7 negras. se extraen 2 bolas sin reposición. Calcular la probabilidad de que la segunda bola sea negra. Si la primera es blanca, calcular la probabilidad de que la segunda sea blanca. Resolver las mismas cuestiones si la extracción es sin reposición.

19) Se clasifican 400 personas de la forma siguiente:

		PESO(LIBRAS)				
		130-149	150-169	170-189	$\geq$190	
EDAD(AÑOS)	30-39	10	20	20	40	90
	40-49	10	15	50	70	145
	50-59	5	15	50	40	110
	60-69	5	10	15	25	55
		30	60	135	175	400

Elegimos un individuo. Calcular la probabilidad de que

a) su edad esté entre 40-49 años,
b) su edad esté entre 40-49 años y su peso entre 170-189 libras,
c) su edad esté entre 40-49 años o entre 60-69,
d) su edad esté entre 40-49 años si pesa entre 150-169 libras,
e) pese menos de 170 libras,
f) pese menos de 170 libras y tenga menos de 50 años,
g) pese menos de 170 libras, siempre que sea menor de 50 años.

Capítulo 3

Variables Aleatorias

3.1 Variable Aleatoria Discreta

Una *variable aleatoria unidimensional discreta* X es una función que puede tomar un conjunto *discreto* (finito o numerable) de valores $x_1, x_2, \ldots, x_n$, con probabilidades respectivas $p_1, p_2, \ldots, p_n$. Las variables aleatorias tienen asociadas dos funciones que pasamos a introducir.

Función de densidad: $f_X(x) = P\{\omega | X(\omega) = x\} = P(X = x)$. Así, en la variable anterior, la función de densidad viene definida de la forma siguiente:

$$f_X(x) = \begin{cases} f_X(x_1) = P(X = x_1) = p_1 \\ f_X(x_2) = P(X = x_2) = p_2 \\ \vdots \\ f_X(x_n) = P(X = x_n) = p_n \\ 0 \text{ para el resto} \end{cases}$$

Consideremos el experimento aleatorio: lanzar dos monedas. Sea Y la variable aleatoria que a cada suceso de dicho experimento le asocia el número de caras que aparecen. La función de densidad f_Y viene definida de la forma: $f_Y(0) = 1/4$, $f_Y(1) = 1/2$, $f_Y(2) = 1/4$, 0 para el resto de valores. Su representación gráfica es:

Función de distribución: $F_X(x) = P\{\omega | X(\omega) \leq x\} = P(X \leq x)$. Por lo tanto, F_X viene definida de la forma siguiente:

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} f_X(x_i) = \sum_{x_i \leq x} p_i.$$

En el ejemplo anterior, la función de distribución de la variable Y viene definida de la forma siguiente:

$$F_Y(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \leq x < 1 \\ 2/4 & 1 \leq x < 2 \\ 1 & 2 \leq y \end{cases}$$

y su representación gráfica es

Hemos de hacer notar que $F_X(x)$ es una función creciente que tiende a 1 y 0, cuando x tiende a ∞ y $-\infty$ respectivamente.

En la introducción, ya habíamos avanzado la relación entre las muestras y las variables aleatorias. Veamos ahora dicha relación. Supongamos que en una muestra de tamaño N hemos obtenido los siguientes m valores diferentes:

valores	x_1	$<$	x_2	$<$	$\ldots$	$<$	x_m
número de veces que se repiten	n_1	,	n_2	,	$\ldots$	,	n_m

Es decir, si X es la variable estadística que a cada objeto de la población, de donde se ha extraído la muestra, le asigna un determinado carácter numérico x, por ejemplo peso, entonces, en esta muestra, la variable estadística toma los valores anteriores x_i repetidos tantas veces como las ya indicadas. Observemos ahora que X puede considerarse como una variable aleatoria que toma los valores $x_1 < x_2 < \ldots < x_m$ con probabilidades $n_1/N, n_2/N, \ldots, n_m/N$ respectivamente. Así, podemos definir las funciones de densidad y distribución asociadas a la variable estadística (o aleatoria) X. Veamos un ejemplo.

Supongamos que elegimos una muestra de 25 personas y a cada una le asignamos el número de calzado que usa, obteniendo los resultados siguientes:

resultados	31	32	33	34	35	36
número de veces que se repiten	5	4	1	3	7	5

Llamemos Y a la variable aleatoria asociada a esta muestra. Entonces, sus funciones de densidad y distribución son:

$$f_Y(x) = \begin{cases} f_Y(31) = P(Y = 31) = 5/25 \\ f_Y(32) = P(Y = 32) = 4/25 \\ f_Y(33) = P(Y = 33) = 1/25 \\ f_Y(34) = P(Y = 34) = 3/25 \\ f_Y(35) = P(Y = 35) = 7/25 \\ f_Y(36) = P(Y = 36) = 5/25 \end{cases} \qquad F_Y(x) = \begin{cases} 0 & x < 31 \\ 5/25 & 31 \leq x < 32 \\ 9/25 & 32 \leq x < 33 \\ 10/25 & 33 \leq x < 34 \\ 13/25 & 34 \leq x < 35 \\ 20/25 & 35 \leq x < 36 \\ 25/25 & 36 \leq x \end{cases}$$

Observaciones: para una muestra se definen los siguientes conceptos:

Frecuencia absoluta del valor x_i (n_i): es el número de veces que se repite dicho valor en la muestra. Observemos que $0 \leq n_i \leq N$. Además

$$\sum_{i=1}^{m} n_i = N.$$

Frecuencia relativa del valor x_i (f_i): es el cociente entre la frecuencia absoluta del valor x_i y el número de observaciones realizadas. Es decir $f_i = n_i/N$. Además, se cumple que

$$0 \leq f_i \leq 1, \qquad y \qquad \sum_{i=1}^{m} f_i = 1.$$

Frecuencia absoluta acumulada del valor x_i (N_i): es la suma de las frecuencias absolutas de los valores menores o iguales que x_i, es decir

$$N_i = \sum_{j=1}^{i} n_j$$

Observemos que $N_m = N$

Frecuencia relativa acumulada del valor x_i (F_i): es la suma de las frecuencias relativas de los valores menores o iguales que x_i, es decir

$$F_i = \sum_{j=1}^{i} f_j = \frac{N_i}{N}$$

Observemos que $F_m = 1$

Con estas definiciones se confeccionan las llamadas *tablas de frecuencias*. En nuestro ejemplo, dicha tabla es:

x_i	n_i	f_i	N_i	F_i
31	5	5/25	5	5/25
32	4	4/25	9	9/25
33	1	1/25	10	10/25
34	3	3/25	13	13/25
35	7	7/25	20	20/25
36	5	5/25	25	25/25=1
TOTAL	25	1		

Otra forma de agrupar datos, viene dada por los diagramas, que consisten en representaciones gráficas. Así, asociando a cada valor su frecuencia absoluta, tenemos el *diagrama de barras*, que coincide esencialmente con la función de distribución. De igual manera, asociando a cada valor su frecuencia absoluta, obtenemos el *diagrama de frecuencias acumuladas*, que coincide con la representación gráfica de la función de distribución.

3.1.1 Esperanza, Varianza.

Consideremos la variable aleatoria unidimensional discreta X, que toma los valores $x_1, x_2, \ldots, x_n$, con probabilidades respectivas $p_1, p_2, \ldots, p_n$. Definimos:

Esperanza de la variable X, EX, como:

$$EX = \sum_{i=1}^{n} x_i P(X = x_i) = \sum_{i=1}^{n} x_i p_i = \sum_{i=1}^{n} x_i f_X(x_i).$$

Para una muestra de tamaño N, suponiendo que hemos obtenido los valores $x_1, x_2, \ldots, x_m$ con frecuencias $n_1, n_2, \ldots, n_m$ respectivamente, llamando nuevamente X a la variable asociada, la esperanza de X se llama *media* y se denota por $\bar{x}$. Así entonces, tenemos para esta muestra general, que

$$\bar{x} = \sum_{i=1}^{m} x_i \frac{m_i}{N} = \sum_{i=1}^{m} \frac{m_i x_i}{N}.$$

En los ejemplos anteriores, tenemos:
<u>monedas:</u> $EX = 0\,\frac{1}{4} + 1\,\frac{1}{2} + 2\,\frac{1}{4} = 1,$
<u>medidas del pie</u> $EY = 31\,\frac{5}{25} + 32\,\frac{4}{25} + 33\,\frac{1}{25} + 34\,\frac{3}{25} + 35\,\frac{7}{25} + 36\,\frac{5}{25} = 33'72$

Varianza de la variable X, $(\mathrm{Var}X = \sigma^2)$, se define

$$\mathrm{Var}X = \sigma^2 = \sum_{i=1}^{n} (x_i - EX)^2 P(X = x_i) = \sum_{i=1}^{n} (x_i - EX)^2 p_i.$$

En el caso de muestras, usualmente el cálculo de la varianza viene dado por la expresión

$$\sigma^2 = \frac{\sum_{i=1}^{m} n_i (x_i - \bar{x})^2}{N}$$

Desviación típica de la variable X, es la raíz cuadrada positiva de la varianza, es decir σ.

Frecuentemente, en el caso de muestras, se utiliza la *varianza corregida* o *cuasi varianza* s, que viene definida como:

$$s^2 = \frac{\sum_{i=1}^{m} n_i (x_i - \bar{x})^2}{N-1} = \frac{N}{N-1}\,\sigma^2.$$

Para concluir esta subsección, vamos a introducir unos conceptos utilizados en muestras (es decir variables estadísticas), que nos dividen el intervalo de observaciones en "partes iguales".

Cuartiles $(P_{i/4}, i = 1, 2, 3)$: el cuartil $P_{i/4}$ es el valor de la variable que deja inferiores o iguales a él las $i/4$ partes de las observaciones.

En el ejemplo de las medidas del pie, observando la tabla de frecuencias acumuladas N_i, obtenemos:

$1/4\,N = 25/4 = 6'25$. Luego, como $5 < 6'25 < 9 \Rightarrow P_{1/4} = 32$,
$2/4\,N = 25/2 = 12'5$. Luego, como $10 < 12'5 < 13 \Rightarrow P_{2/4} = 34$,
$3/4\,N = 75/4 = 18'75$. Luego, como $13 < 18'75 < 20 \Rightarrow P_{31/4} = 35$.

Centiles o percentiles (P_k): se define el percentil $K-$ésimo como el valor de la variable que deja inferiores o iguales a él las $K/100$ partes de las observaciones.

En el ejemplo anterior, obtenemos
$P_{52} : 52/100\,N = 52/100\,25 = 13 \Rightarrow P_{52} = 34$.

3.1.2 Ejemplo de Variable Aleatoria Discreta.

Existen muchos ejemplos de variables aleatorias discretas: uniforme, polinomial, Poisson, etc.. En esta subsección solamente vamos a analizar la llamada distribución binomial, partiendo de un ejemplo.

Supongamos que tenemos 3 pacientes A, B y C. El tratamiento utilizado en su enfermedad es efectivo en un 70%, es decir, la probabilidad de que se cure es $0'7$. Vamos a calcular la probabilidad de que se curen exactamente 2 pacientes. Para ello, confeccionamos una tabla con los sucesos en los que se curan exactamente dos pacientes:

	CURADOS	NO CURADOS
E_1	AB	C
E_2	AC	B
E_3	BC	A

Luego

$$P(\text{dos curados}) = P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3).$$

Veamos la probabilidad de cada suceso E_i.

$$\begin{aligned} P(E_1) &= P(\text{A curado} \cap \text{B curado} \cap \text{C no curado}) \text{ independientes } = \\ &= P(\text{A curado})P(\text{B curado})P(\text{C no curado}) = 0'7 \cdot 0'7 \cdot 0'3 \\ &= 0'147. \end{aligned}$$

Luego, como la probabilidad de cada E_i es la misma, resulta finalmente:

$$P(\text{dos curados}) = 3 \cdot 0'7^2 \cdot 0'3^1 = 0'441.$$

ahora con la lista completa de todos los casos posibles:

$$\begin{array}{rcccl} P(0 \text{ curados}) & = & 1 \cdot 0'3^3 & = & \binom{3}{0} 0'7^0 \cdot 0'3^3 \\ P(1 \text{ curados}) & = & 1 \cdot 0'7^1 0'3^2 & = & \binom{3}{1} 0'7^1 \cdot 0'3^2 \\ P(2 \text{ curados}) & = & 1 \cdot 0'7^2 0'3^1 & = & \binom{3}{2} 0'7^2 \cdot 0'3^1 \\ P(3 \text{ curados}) & = & 1 \cdot 0'7^3 & = & \binom{3}{3} 0'7^3 \cdot 0'3^0, \end{array}$$

es decir

$$P(k \text{ curados}) = \binom{3}{k} 0'7^k \cdot 0'3^{3-k} \quad k = 0, 1, 2, 3.$$

Para extenderlo al caso general, supongamos que un determinado experimento aleatorio admite solamente dos resultados posibles:

$$\begin{array}{rcl} A & : & \text{éxito}, con probabilidad p \\ A^c & : & \text{fracaso, con probabilidad } q = 1 - p. \end{array}$$

Supongamos ahora que repetimos n veces dicho experimento, siendo cada prueba independiente de las anteriores. Sea X la variable aleatoria que a cada serie de n pruebas le asocia el número de éxitos obtenidos. Veamos las funciones de densidad y distribución de X.

Sea k un entero tal que $0 \leq k \leq n$. Vamos a calcular $P(X = k)$. Primero contamos todos los sucesos con k éxitos, es decir ¿de cuántas maneras se pueden elegir k sitios de un total de n?. Dicho número es

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

y la probabilidad de cada uno de ellos es $p^k(1-p)^{n-k}$. Luego, la función de densidad de X es:

$$f_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & k = 0, 1, 2, \ldots, n \\ 0 & \text{resto}, \end{cases}$$

y la de distribución

$$F_X(x) = \sum_{k \leq x} \binom{n}{k} p^k (1-p)^{n-k}.$$

A esta función de distribución se le llama *distribución binomial de parámetros* n, p, y se denota por $B(n,p)$. Se puede demostrar que su esperanza o media es $\mu_X = np$, y su varianza $\sigma^2 = np(1-p) = npq$.

Ejercicio: *lanzamos* 6 *veces una moneda. Calcular la probabilidad de obtener*

A: exactamente 2 caras, B: al menos 4 caras, C: como mucho 3 caras.

La variable que a cada suceso le asocia el número de caras es una binomial $B(6,1/2)$. Así para cada uno de los sucesos anteriores, tenemos:

$$\begin{aligned}
&P(A) = P(X=2) = \sum_{k=2}^{2} \binom{6}{k} (1/2)^k (1/2)^{6-k} = \binom{6}{2} \frac{1}{2^6} \\
&P(B) = P(X \geq 4) = \sum_{k=4}^{6} \binom{6}{k} (1/2)^k (1/2)^{6-k} = \frac{1}{2^6} \left[\binom{6}{4} + \binom{6}{5} + \binom{6}{6} \right] \\
&P(C) = P(X \leq 3) = 1 - P(X > 3) = 1 - P(X \geq 4).
\end{aligned}$$

Uno de los problemas de la distribución binomial es el del cálculo de las sumas que aparecen. Por ejemplo, si lanzamos un dado 500 veces y nos piden la probabilidad de obtener entre 230 y 474 seises, dicha probabilidad viene dada por la expresión

$$\sum_{k=230}^{474} \binom{500}{k} \frac{5^{500-k}}{6^{500}},$$

cuyo cálculo es, aunque simple de concepto, trabajoso a efectos prácticos. Estos problemas trataremos de resolverlos con las llamadas distribuciones continuas, que aparecerán en la sección siguiente.

3.2 Variable Aleatoria Continua

Existen variables aleatorias que, a diferencia de las discretas, no toman valores aislados, sino que pueden tomar cualquier valor de un intervalo real.

Estas variables se llaman *continuas*. Por ejemplo, elegir un número entre 0 y 1. Para estas variables, la definición de función de distribución es análoga al caso de las discretas. En cambio, la definición de función de densidad es distinta. Así, decimos que la función f_X es una *función de densidad de la variable aleatoria* X, si cumple lo siguiente:

1) f_X está definida en toda la recta real y es *no negativa*,

2)

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1.$$

Para todo número real x, la función de distribución viene dada por

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}x.$$

Así, si consideramos la función de densidad, de la variable aleatoria Y, siguiente:

$$f_Y(x) = \begin{cases} 1 & \text{si } x \in [0,1] \\ 0 & \text{resto,} \end{cases}$$

su función de distribución viene dada por:

$$F_Y(x) = \begin{cases} 1 & \text{si } x \leq 0 \\ x & \text{si } 0 \leq x \leq 1 \\ 1 & \text{si } x \geq 1. \end{cases}$$

Una propiedad interesante, que cumplen las variables aleatorias, tanto discretas como continuas, es la siguiente:

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a),$$

siendo $a < b$ dos números reales.

Para probarlo, observemos que

$$\{\omega | X(\omega) \leq b\} = \{\omega | a < X(\omega) \leq b\} \cup \{\omega | X(\omega) \leq a\}$$

siendo incompatibles los dos sucesos del término de la derecha. Luego, basta aplicar probabilidades en la igualdad anterior, y aplicar el tercer axioma de la probabilidad.

3.2.1 Esperanza, Varianza.

Para una variable aleatoria continua X, con función de densidad f_X, la esperanza μ_X y varianza σ_X^2 se definen de la forma

$$\mu_X = \int_{-\infty}^{\infty} x f_X \, \mathrm{d}x, \qquad \sigma_X^2 = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X \, \mathrm{d}x,$$

suponiendo la convergencia absoluta de las integrales. Así mismo, la desviación típica (σ_X) es la raíz cuadrada positiva de la varianza.

En ejemplo anterior, la esperanza y varianza vienen dadas por

$$\mu_Y = \int_{-\infty}^{\infty} x f_X \, \mathrm{d}x = \int_0^1 x f_X \, \mathrm{d}x = \frac{1}{2}, \qquad \sigma_Y^2 = \int_0^1 (x-\frac{1}{2})^2 f_X \, \mathrm{d}x = \frac{1}{12}.$$

3.2.2 Ejemplos de Variables Aleatorias Continuas.

Normal o de **Laplace-Gauss.**

Una variable aleatoria continua X, se dice que tiene una distribución normal (o de Laplace-Gauss) de media μ y desviación típica σ y se denota por $N(\mu, \sigma)$, si su función de densidad es

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \, \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad \forall x \in \mathbb{R}.$$

Su gráfica es

Observemos que es simétrica respecto a la recta $x = \mu$. Su función de distribución viene dada por la expresión

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \, \mathrm{e}^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} \, \mathrm{d}t.$$

Se puede demostrar, utilizando el cálculo integral en dos variables, que

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \, \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \, \mathrm{d}x = 1.$$

En el caso $\mu = 0$ y $\sigma = 1$, la normal $N(0,1)$ se llama *forma normal canónica*. Su función de densidad viene dada por

$$f(x) = \frac{1}{\sqrt{2\pi}} \, \mathrm{e}^{-\frac{1}{2}\,x^2},$$

y su gráfica es

Como es simétrica respecto a la recta $x = 0$ y

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}x^2}\, \mathrm{d}x = 1$$

resulta que

$$\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}x^2}\, \mathrm{d}x = \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}x^2}\, \mathrm{d}x = \frac{1}{2}$$

Desgraciadamente, las integrales del tipo

$$\int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}\, \mathrm{d}t$$

no se pueden calcular por métodos algebraicos, incluso en los casos $\mu = 0$ y $\sigma = 1$. Por ello, dichas integrales se "calculan" de forma aproximada haciendo uso de tablas. Veamos cómo se hace esto. Primero vamos a calcular probabilidades del tipo $P(a \leq X \leq b)$, con X la normal canónica, y posteriormente lo extenderemos a cualquier tipo de normal.

Ejercicio. *Sea X una $N(0,1)$. Calcular aproximadamente las siguientes probabilidades:*

1) $P(X \leq 1'8)$ 2) $P(X \leq -0'6)$ 3) $P(X \geq 2'03)$
4) $P(X \geq -1'28)$ 5) $P(0'81 \leq X \leq 1'94)$ 6) $P(-1'21 \leq X \geq -0'1)$
7) $P(-0'46 \leq X \leq 2'21)$ 8) $P(0 \leq X \leq 1'2)$ 9) $P(-0'68 \leq X \geq 0)$

Dado que

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}x^2}\, \mathrm{d}x,$$

este problema se reduce al cálculo de áreas por integrales. Para cada caso, vamos a considerar la gráfica correspondiente.

Para el primer caso, observemos la figura-1

Es decir $P(X \leq 1'8) = P(X \leq 0) + P(0 \leq X \leq 1'8) = 0'5 + P(0 \leq X \leq 1'8)$. Ahora, observando en la tabla, vemos que $P(0 \leq X \leq 1'8) = 0'4641$, por lo tanto $P(X \leq 1'8) = 0'9641$.

El caso segundo se puede resolver observando la figura 2. Así, tenemos que $P(X \leq -0'6) = P(X \geq 0'6) = P(0 \leq X \leq \infty) - P(0 \leq X \leq 0'6) = 0'5 - 0'2258 = 0'2742$.

El resto de los casos se resuelven con consideraciones similares. Vamos a indicar analíticamente como se resuelven, quedando las consideraciones gráficas como ejercicio.

3) $P(X \geq 2'03) = P(X \geq 0) - P(X \leq 2'03) = 0'5 - 0'4788 = 0'0222.$

4) $P(X \geq -1'28) = 0'5 + P(0 \leq X \leq 1'28) = 0'5 + 0'3997 = 0'8997.$

5) $P(0'81 \leq X \leq 1'94) = P(0 \leq X \leq 1'94) - P(0 \leq x \leq 0'81) = 0'4738 - 0'2910 = 0'1828.$

6) $P(-1'21 \leq X \geq -0'1) = P(0'1 \leq X \leq 1'21) = 0'3471.$

7) $P(-0'46 \leq X \leq 2'21) = P(0 \leq X \leq 2'21) + P(0 \leq X \leq 0'46) = 0'4864 + 0'1772 = 0'6636.$

8) $P(0 \leq X \leq 1'2) = 0'3849.$

9) $P(-0'68 \leq X \geq 0) = P(0 \leq X \leq 0'68) = 0'2518.$

Veamos ahora el caso general. Sea X una normal $N(\mu, \sigma)$. Supongamos que queremos calcular $P(a \leq X \leq b)$, es decir, queremos calcular

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x.$$

Vamos a reducirlo al caso de una normal canónica. Para ello, hagamos el cambio de variable

$$\frac{x-\mu}{\sigma} = t \Rightarrow \mathrm{d}x = \sigma\, \mathrm{d}t.$$

Luego, la integral anterior se transforma en

$$\int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}t^2} \mathrm{d}t,$$

que podemos "calcular" usando las tablas de la normal canónica.

Ejercicio. *Sea $X = N(115, 15)$. Calcular las siguientes probabilidades:*

$$1)\ P(X \leq 140), \quad 2)\ P(X \geq 100), \quad P(110 \leq X \leq 120).$$

Haciendo el cambio de variable

$$\frac{x-115}{15}=t \Rightarrow \mathrm{d}x = 15\,\mathrm{d}t,$$

los límites de integración cambian de la forma siguiente: $140 \to \dfrac{140-115}{15} = \dfrac{5}{3} \simeq 1'67$. Luego,

$$P(X \leq 140) = \int_{-\infty}^{140} \frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\,\mathrm{d}x = \int_{-\infty}^{1'67} \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{1}{2}t^2}\,\mathrm{d}t = 0'9525.$$

Para el segundo caso, con el mismo cambio de variable, obtenemos lo siguiente: $P(X \geq 100) = P(N(0,1) \geq -1) = 0'8413$.

Finalmente $P(110 \leq X \leq 120) = P(-0'33 \leq N(0,1) \leq 0'33) = 0'2586$.

Para finalizar con la distribución normal, vamos a introducir una relación entre la distribución binomial y la normal. Supongamos que tenemos una distribución binomial $X = B(n,p)$. Entonces si se cumplen

$$\begin{cases} np \geq 5 \\ n(1-p) = nq \geq 5, \end{cases}$$

la binomial X se puede aproximar por una normal $Y = N(np, \sqrt{npq})$. En esta aproximación hay que tener en cuenta un hecho relevante. Dado que una binomial sólo puede tomar valores enteros, y una normal puede tomar cualquier valor, la aproximación debe hacerse de la forma siguiente: supongamos que a y b son dos enteros. Entonces $P(a \leq X \leq b) \simeq P(a-1/2 \leq Y \leq b+1/2)$. Veamos un ejemplo.

Ejercicio. *Lanzamos una moneda* 10 *veces. Calcular la probabilidad de obtener entre 3 y 6 caras, usando la binomial y la normal.*

<u>Binomial:</u> $X = B(10, 0'5)$.

$$P(3 \leq X \leq 6) = \sum_{k=3}^{6} \binom{10}{k} 0'5^{10} = 0'7734.$$

<u>Normal:</u> $np = nq = 5 \geq 5$. Luego X puede aproximarse por una normal $Y = N(5, \sqrt{2'5})$. Por lo tanto

$$P(3 \leq X \leq 6) \simeq P(2'5 \leq Y \leq 6'5) = P(-1'58 \leq N(0,1) \leq 0'95) = 0'7718.$$

$t-$**Student.**

Una variable aleatoria continua X, se dice que tiene una distribución *t de Student con n grados de libertad* si su función de densidad es

$$f_{t_n}(x) = \frac{(1+\frac{x^2}{n})^{-\frac{(n+1)}{2}}}{\sqrt{n}\beta(\frac{1}{2},\frac{n}{2})}, \quad n = 1, 2, ...$$

siendo

$$\beta(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}.$$

Su gráfica es

es decir, simétrica respecto a la recta $x = 0$. Para $n \geq 30$, la distribución $t-$Student se "aproxima mucho" a la normal canónica. Para esta función, los problemas que vamos a tratar van a consistir en determinar valores t de manera que $P(a \leq t_n \leq b)$ sea igual a un valor determinado de antemano. Nuevamente precisaremos del manejo de la tabla correspondiente a esta función. Veamos un ejemplo.

Ejercicio. *Sea X una variable aleatoria que sigue una distribución $t-$Student con 9 grados de libertad. Hallar t de manera que*

1) $P(X \geq t) = 0'05$, 2) $P(-t \leq X \leq t) = 0'99$, 3) $P(X \leq -t) = 0'01$,
4) $P(X \leq t) = 0'90$.

1) Notemos que $P(X \geq t) = 1 - P(X \leq t)$. Luego $P(X \geq t) = 0'05 \Leftrightarrow P(X \leq t) = 0'95$. Es decir $t = 1'83$

2) Como $P(-t \leq X \leq t) = 0'99 \Leftrightarrow P(X \leq t) = 0'995$, deducimos que $t = 3'25$.

3) $P(X \leq -t) = 0'01 \Leftrightarrow P(X \leq t) = 0'99$. Luego $t = 2'82$

4) $P(X \leq t) = 0'90$, en la tabla obtenemos directamente $t = 1'38$.

χ^2 (ji-cuadrado de Pearson).

Una variable aleatoria continua X, se dice que tiene una distribución χ^2 *ji-cuadrado de Pearson con n grados de libertad* (χ^2_n) si su función de densidad es

$$f_{\chi^2_n}(x) = \begin{cases} \dfrac{\mathrm{e}^{-x/2} x^{\frac{n}{2}-1}}{2^{n/2}\Gamma(n/2)} & 0 < x \\ 0 & x \leq 0 \end{cases} \qquad n = 1, 2, ...$$

siendo

$$\Gamma(p) = \int_0^\infty \mathrm{e}^{-x} x^{p-1}\,\mathrm{d}x, \quad p \geq 0.$$

Su gráfica es

Al igual que en la normal, nos interesa calcular áreas encerradas por dicha curva. Nuevamente las integrales no se pueden calcular algebraicamente, por lo que hemos de recurrir a las tablas. Veamos un ejemplo.

Ejercicio. *dada una distribución* χ^2 *con* $n = 12$ *grados de libertad, calcular* $\chi^2_{p_1}$ *y* $\chi^2_{p_2}$ *de manera que las áreas de las regiones A y B sean* $0'025$ *para A y* $0'01$ *para B.*

Observando la tabla para la función χ^2, vemos que para el caso A $p_1 = 0'025$. Luego, mirando el la fila correspondiente a 12 grados de libertad, obtenemos $\chi^2_{0'025,12} = 4'40$.
Para el caso B, notemos que $p_2 = 1 - 0'01 = 0'99$. Luego, $\chi^2_{0'99,12} = 26'2$.

Notación: con la expresión

$$\chi^2_{\alpha,n},$$

denotaremos al número tal que $P(\chi^2_n \geq \chi^2_{\alpha,n}) = p$. Gráficamente:

3.3 Ejercicios

1) Dadas las siguientes variables aleatorias, calcular su función de densidad y de distribución y representarlas gráficamente. Así mismo, calcular la esperanza, varianza, desviación típica y las probabilidades pedidas.

a) Lanzamos 3 monedas. X =número de caras-número de cruces.
$P(X < 0)$, $P(-1 \leq X)$, $P(X \leq 1)$.

b) De una urna con 4 bolas rojas y 6 blancas se extraen 3 bolas sin reposición. Sea Y =número de bolas rojas extraídas.

2) El número de hijos de 20 familias seleccionadas al azar, es el siguiente:

$$3, 1, 2, 2, 1, 5, 2, 2, 0, 6, 3, 2, 4, 3, 4, 2, 3, 1, 7, 6.$$

a) construir la tabla de frecuencias,
b) calcular la media, varianza, cuasi varianza y desviación típica,
c) calcular los cuartiles,
d) calcular los percentiles 18, 96, 31, 70, 84.

3) Dadas las siguientes funciones, comprobar que son funciones de densidad de laguna variable aleatoria. Calcular su función de distribución, la esperanza, varianza, desviación típica y las probabilidades pedidas.
a) Variable X definida por

$$f(x) = \begin{cases} \cos(x) & 0 < x < \pi/2 \\ 0 & \text{resto} \end{cases}$$

$P(X \leq 1)$, $P(X \geq -7)$, $P(-2 \leq X \leq 6)$, $P(X \leq 0)$.
b) Variable Y definida por

$$g(x) = \begin{cases} \mathrm{e}^x & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$P(Y \leq 1)$, $P(2 \leq Y \leq 7)$, $P(-4 \leq Y \leq -2)$, $P(-1 \leq Y)$.

4) Para cada una de las siguientes funciones, determinar el valor de la constante K, de manera que sean funciones de densidad de una variable aleatoria.

$$\text{a) } f(x) = \begin{cases} K/x^4 & x \geq 1 \\ 0 & x < 1 \end{cases}$$

$$\text{b) } g(x) = \begin{cases} \dfrac{K}{1+x^2} & 0 \leq x \leq 1 \\ 0 & \text{resto} \end{cases}$$

$$\text{c) } h(x) = \begin{cases} K^2 x & 0 \leq x \leq K \\ 0 & \text{resto} \end{cases}$$

5) Calcular la esperanza y varianza de las siguientes variables aleatorias, y las probabilidades pedidas.
a) X =número de hijos varones en una familia de 3 hijos. $P(X \leq 2)$, $P(X \geq 2)$.
b) Y =número de veces que aparece el número 3 al lanzar 5 veces un dado.

$P(Y \geq 2)$, $P(Y \leq 3)$, $P(2 \leq Y \leq 4)$.

6) Sea X la normal canónica $N(0,1)$. Calcular aproximadamente las siguientes probabilidades:

a) $P(X \leq 4'5)$, b) $P(-2 \leq X \leq 3)$, c) $P(X \leq 1'73)$, d) $P(X \geq -2'09)$,
e) $P(X \geq 1'27)$, f) $P(-2'34 \leq X \leq -0'98)$, g) $P(0'21 \leq X \leq 2'05)$.

7) Un examen de tipo test (verdadero,falso) consta de 10 preguntas. Respondiendo al azar, calcular la probabilidad de aprobar (responder correctamente a 5 o más preguntas).

8) Se lanza una moneda 500 veces. Hallar la probabilidad de que el número de caras no difiera de 250

a) en mas de 10,
b) en mas de 30.

9) Sea X una $t-$Student con n grados de libertad. Determinar t_c de manera que la probabilidad $P(-t_c \leq X \leq t_c)$ valga:

a) $0'9 \quad n = 10$, b) $0'99 \quad n = 8$, c) $0'95 \quad n = 6$, d) $0'6 \quad n = 4$.

10) Sea X una normal $N(0,1)$. Determinar z_c de manera que la probabilidad $P(-z_c \leq X \leq z_c)$ valga:

a) $0'9$, b) $0'99$, c) $0'95$.

11) Sea $\chi^2_{p,n}$ un número tal que $P(\chi^2_n \leq \chi^2_{p,n}) = p$, con χ^2_n una $ji-$cuadrado con n grados de libertad. Calcular $\chi^2_{p,n}$, siendo:

a) $p = 0'95, \quad p = 0'05 \qquad$ para $n = 15$,
b) $p = 0'90, \quad p = 0'10 \qquad$ para $n = 25$,
c) $p = 0'75, \quad p = 0'25 \qquad$ para $n = 28$.

12) Consideramos t_ν una función $t-$Student con ν grados de libertad. Hallar t_0 de manera que

a) $P(t_\nu \geq t_0) = 0'05 \quad$ para $\nu = 1, 9, 15, 30, 120$,
b) $P(-t_0 \leq t_\nu \leq t_0) = 0'99 \quad$ para $\nu = 1, 9, 30, 40, 60, 120$,
c) $P(t_\nu \geq -t_0) = 0'01 \quad$ para $\nu = 4, 9, 40, 60, 120$.

Comparar dichos resultados con los correspondientes a la normal $N(0,1)$.

13) Consideramos χ^2_n la distribución $ji-$cuadrado con n grados de libertad. Hallar χ_0 de manera que

a) $P(\chi_n^2 \geq \chi_0) = 0'05, \quad$ para $n = 2, 15, 21, 50, 60, 100,$
b) $P(\chi_n^2 \leq \chi_0) = 0'10, \quad$ para $n = 2, 5, 50, 100,$
c) $P(\chi_n^2 \leq \chi_0) = 0'95, \quad$ para $n = 2, 5, 50, 100,$
d) $P(\chi_n^2 \leq \chi_0) = 0'50, \quad$ para $n = 9, 28, 40, 100.$

Capítulo 4

Estimación

4.1 Introducción

La *estimación* es una parte de la inferencia estadística cuyo objetivo es "determinar de alguna manera", mas o menos aproximada, una determinada característica de una población, basándose en la información obtenida en una muestras (o parte de dicha población). Por ejemplo, si desconocemos la media de la altura de una población, podemos "asignar" el valor de la media de una muestra. Otro procedimiento, mas interesante, consiste en dar, en vez de un valor, un intervalo junto a un nivel de "confianza" de que la media de dicha población esté en dicho intervalo. Estos dos procedimientos van a dar lugar a la estimación puntual y estimación por intervalos.

Desde luego que parece mas interesante la estimación por intervalos, pero este problema precisa de la estimación puntual.Además, adelantándonos un poco, una pregunta que surge es la siguiente: ¿si deseamos estimar una característica de una población (media o varianza por ejemplo), qué tendremos que utilizar de las muestras? ¿la media, la varianza u otras medidas? Aunque en un principio parece que para estimar la media es preciso utilizar la media de una muestra, para la varianza la de la muestra, etc., en general no es así. Lo comentaremos más adelante.

Concretando más, el proceso que vamos a seguir es el siguiente: supongamos que deseamos determinar una característica de una población (ej. media). A cada muestra de la población, que denotamos por κ le asignamos un valor κ_0, que puede ser la media u otro valor. Así, la variable construida

$$X(\kappa) \mapsto \kappa_0$$

que a cada muestra le asocia un determinado valor (ej. media), es una variable aleatoria, que tendrá una función de distribución F_X, y por supuesto una función de densidad. Esta variable aleatoria es la que nos interesa. Concretamente, hemos de determinar lo que asociamos a cada muestra (media, varianza, etc.) para que la correspondiente variable aleatoria X, y por lo tanto F_X, tenga unas características "buenas" para inferir algo sobre la característica en estudio de la población. En esto consiste la estimación puntual, es decir: queremos estudiar una característica de una población (ej. media). ¿Qué tendremos que asignar a cada muestra (media, varianza, etc.) para que la variable aleatoria X sea "buena" para estimar la característica de la población? Comencemos con unas definiciones.

Una *muestra* de una población es la elección de unos cuantos elementos de dicha población. Las muestras pueden ser *con reposición* (se elige un elemento y se devuelve a la población, con lo que puede ser nuevamente elegido), o *sin reposición no se devuelve.* Cuando la población es muy grande, a efectos prácticos, se puede considerar infinita y las muestras con reposición. Aunque hay varias formas de considerar los elementos de una población para elegir una muestra, por simplificación supondremos que todos los elementos tienen la misma probabilidad de ser elegidos.

Se llama *parámetro* a una magnitud de la población, que queremos determinar. Por ejemplo, la desviación típica. Los valores en cada muestra se llaman *estadísticos.* Por ejemplo: queremos estimar la altura media de una población; ese valor sería el parámetro. Elegimos una muestra y le asignamos la media (o la varianza); eso es el estadístico.

Consideremos ahora en una población, todas las muestras de tamaño N (con o sin reposición). Para cada muestra, calculamos un estadístico. Entonces, llamando Ω al espacio muestral o población, y definiendo Ω_1 =todas las muestras de tamaño N de Ω, la variable

$$\begin{array}{rccc} X : & \Omega_1 & \longrightarrow & \mathbb{R} \\ & \kappa & \mapsto & \kappa_0 \end{array}$$

que a cada muestra le asocia un valor κ_0 (estadístico, igual para todas las muestras), es una variable aleatoria, llamada *distribución de muestreo* del estadístico en particular. Así, por ejemplo, tenemos:

ESTADÍSTICO	DISTRIBUCIÓN DE MUESTRO
Media	D. M. de medias
Varianza	D. M. de varianzas
Desviación Típica	D. M. desviación típica.

4.2 Estimación Puntual

Con las definiciones de la sección anterior, tenemos que una *estimación puntual* es el valor numérico que toma la variable (distribución de muestreo) en una muestra en particular. La variable utilizada se llama *estimador puntual* o más concretamente *estimador por punto*.

Por ejemplo, queremos "determinar" la media de una población. Entonces, elegimos una muestra y calculamos su media. Este valor es el que damos como una estimación (puntual) del valor de la media de la población. En este caso, el estimador por punto es la distribución de muestreo de medias. La estimación puntual es la media de una muestra.

En la introducción ya hemos avanzado la pregunta sobre qué estimador (por punto) usar. Los criterios que se siguen son:

a) que sea *centrado o insesgado*, es decir que su media coincida con el parámetro a estimar,

b) que sea de *varianza mínima*.

Por ejemplo, supongamos que queremos estimar la desviación típica (σ) de una población. Consideramos 3 estimadores por punto, que dan de resultas 3 variables aleatorias X, Y y Z. ¿Cuál elegiremos? Primero miramos cuales de las esperanzas (o medias) EX, EY y EZ son iguales a σ (es decir, cuales son centrados). Supongamos que $EX = EZ = \sigma$, $EY \neq \sigma$. Luego nos quedamos con X y Z. Después, miramos el de varianza mínima. Supongamos que $VarX > VarZ$. Luego, nos quedamos con Z. Veamos ahora algunos ejemplos de estimadores por punto.

Parámetro *p de una distribución binomial.*
Consideremos un experimento con dos resultados posibles:

$$\begin{array}{cccc} A & \text{éxito} & \text{con probabilidad} & p \\ A^c & \text{fracaso} & \text{con probabilidad} & 1-p \end{array}$$

(por ejemplo la efectividad de un fármaco para el tratamiento de una enfermedad.) Elegimos una muestra de tamaño N. Entonces, el estimador puntual de p es

$$\hat{p}\ :\ \frac{\text{número de veces que aparece } A}{N} = \frac{\text{número de éxitos}}{N}$$

es decir, tomamos N pacientes y dividimos el número de curados por N.

Propiedades de este estimador:

a) es insesgado (es decir $E\hat{p} = p$),

b) es de varianza mínima

$$\mathrm{Var}\hat{p} = \frac{p(1-p)}{N}$$

c) si $N \geq 30$ entonces $\hat{p}$ se distribuye aproximadamente según una normal

$$N\Big(p, \sqrt{\frac{p(1-p)}{N}}\Big).$$

Parámetros μ, σ *de una población normal.*

Consideremos una población cuya característica en estudio siga una distribución normal $N(\mu, \sigma)$. Sea κ una muestra de tamaño N y $(\xi_1, \xi_2, \ldots, \xi_n)$ los valores asociados a los elementos de la muestra. Entonces, los estimadores puntuales (insesgados y de varianza mínima) para μ y σ^2 (media y varianza), son

$$\begin{aligned}
\hat{\mu} &= \frac{\displaystyle\sum_{i=1}^{N}\xi_i}{N} &= \bar{x} \quad \text{(estadístico: media muestral)} \\
\hat{\sigma}^2 &= \frac{\displaystyle\sum_{i=1}^{N}(\xi_i - \bar{x})}{N} &= s^2 \quad \text{(estadístico: cuasi varianza muestral).}
\end{aligned}$$

Obviamente, para la desviación típica σ obtenemos como estimador s.

Veamos la distribución de muestreo de dichos estadísticos.

Para $\hat{\mu}$.

1) si σ es conocida, $\hat{\mu}$ se distribuye según una normal

$$N\Big(\mu, \frac{\sigma}{\sqrt{N}}\Big),$$

o lo que es lo mismo, la variable

$$\frac{\hat{\mu} - \mu}{\sigma/\sqrt{N}}$$

sigue una normal canónica $N(0,1)$.

2) si σ es desconocida y $N \geq 30$ (grandes muestras), $\hat{\mu}$ se distribuye aproximadamente según una normal

$$N\Big(\mu, \frac{s}{\sqrt{N}}\Big),$$

o lo que es lo mismo, la variable

$$\frac{\hat{\mu} - \mu}{s/\sqrt{N}}$$

sigue una normal canónica $N(0,1)$.

3) si σ es desconocida y $N < 30$ (pequeñas muestras), la variable

$$\frac{\hat{\mu} - \mu}{s/\sqrt{N}}$$

sige una distribución t de Student con $N-1$ grados de libertad, es decir, sigue una t_{N-1}.

Para σ. La variable

$$\frac{(N-1)s^2}{\sigma^2}$$

sigue una distribución χ^2 con $N-1$ grados de libertad, es decir, sigue una χ^2_{N-1}.

4.3 Estimación por Intervalo

En las estimación puntual, lo que hacemos es dar de alguna manera una aproximación del parámetro de una población, siendo el resultado un valor. La estimación por intervalos lo que hace es dar un intervalo $[a,b]$, donde estará el parámetro de la población, con un cierto nivel de confianza, o digamos probabilidad. Mas concretamente, un intervalo de confianza para la media μ (por ejemplo), es un intervalo $[a, b]$ que incluye en su interior a la media μ de la población, con un grado de incertidumbre. Por ejemplo, un intervalo de confianza del 95% es un intervalo $[a, b]$ tal que

$$P(a \leq \mu \leq b) = 0'95.$$

Es decir, si consideramos todas las muestras, de un determinado tamaño, cada una de ellas nos dará un intervalo de confianza al 95%. Lo que queremos decir es que la media de la población estará en el 95% de todos los intervalos que podemos considerar. Así, calculando un intervalo en particular (para una muestra concreta), la media de la población puede que esté (o no) en dicho intervalo. Pero tenemos la seguridad que hay un 95% de posibilidades (o una confianza del 95%) de que la media esté en dicho intervalo. Veamos algunos ejemplos de intervalos de confianza. Previamente, introducimos una notación.

$z_{\alpha/2}$ es el número tal que $P(N(0,1) \geq z_{\alpha/2}) = \alpha/2$.
$t_{N-1,\alpha/2}$ es el número tal que $P(t_{N-1} \geq t_{N-1,\alpha/2}) = \alpha/2$, siendo t_{N-1} una t de Student con $N-1$ grados de libertad.
$\chi^2_{N-1,\alpha/2}$ es el número tal que $P(\chi^2_{N-1} \geq \chi^2_{N-1,\alpha/2}) = \alpha/2$, siendo χ^2_{N-1} una χ cuadrado con $N-1$ grados de libertad.

Intervalo de confianza para el parámetro p de una distribución binomial (grandes muestras, $N \geq 30$).

El intervalo aleatorio

$$I = \left[\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}\right]$$

se llama *estimador por intervalo, de nivel de confianza* $(1-\alpha)$. Para una muestra en particular, si $\hat{p}_0$ es el valor de $\hat{p}$ en dicha muestra, el intervalo

$$I_0 = \left[\hat{p}_0 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{N}}\right]$$

recibe el nombre de *estimación por intervalo de confianza de nivel* $(1-\alpha)$. En resumen, el $(1-\alpha)$% de los I_0 contiene al parámetro p en su interior.

Ejercicio. *Un sondeo de* 100 *votantes, elegidos al azar en un distrito electoral, indica que el* 55% *de ellos está a favor de un candidato determinado.*

Hallar el intervalo de confianza al
a) 95%,
b) 99%,
c) $99'73$%,
para la proporción de todos los votantes favorables a dicho candidato.

Como $\hat{p}_0 = 0'55$, el intervalo de confianza, general a estor tres casos, es

$$I_0 = \left[0'55 \pm z_{\alpha/2}\sqrt{\frac{0'55 \cdot 0'45)}{100}}\right]$$

a) 95%, $1-\alpha = 0'95$, $\alpha/2 = 0'025$, $z_{\alpha/2} = 1'96$. Luego, con este valor, el intervalo de confianza al 95% es

$$I_0 = \left[0'55 \pm 1'96\sqrt{\frac{0'55 \cdot 0'45)}{100}}\right] = [0'55 \pm 0'0975087] \simeq [0'55 \pm 0'1].$$

b) 99%, $1-\alpha = 0'99$, $\alpha/2 = 0'005$, $z_{\alpha/2} = 2'575$. Luego, con este valor, el intervalo de confianza al 95% es

$$I_0 = \left[0'55 \pm 2'575\sqrt{\frac{0'55 \cdot 0'45)}{100}}\right] = [0'55 \pm 0'128...] \simeq [0'55 \pm 0'13].$$

a) $99'73$%, $1-\alpha = 0'9973$, $\alpha/2 = 0'00135$, $z_{\alpha/2} = 3$. Luego, con este valor, el intervalo de confianza al 95% es

$$I_0 = \left[0'55 \pm 3\sqrt{\frac{0'55 \cdot 0'45)}{100}}\right] = [0'55 \pm 0'1492...] \simeq [0'55 \pm 0'15]$$

Intervalos de confianza para los parámetros μ y σ de una población normal.

media, con varianza conocida.

$$I = \left[\hat{\mu} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right], \quad I_0 = \left[\hat{\mu}_0 \pm z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right].$$

media, con varianza desconocida. Grandes muestras ($N \geq 30$).

$$I = \Big[\hat{\mu} \pm z_{\alpha/2}\frac{s}{\sqrt{N}}\Big], \quad I_0 = \Big[\hat{\mu}_0 \pm z_{\alpha/2}\frac{s_0}{\sqrt{N}}\Big].$$

media, con varianza desconocida. Pequeñas muestras ($N < 30$).

$$I = \Big[\hat{\mu} \pm t_{N-1,\alpha/2}\frac{s}{\sqrt{N}}\Big], \quad I_0 = \Big[\hat{\mu}_0 \pm t_{N-1,\alpha/2}\frac{s_0}{\sqrt{N}}\Big].$$

Varianza.

$$I = \Big[\frac{(N-1)s^2}{\chi^2_{N-1,\alpha/2}}, \frac{(N-1)s^2}{\chi^2_{N-1,1-\alpha/2}}\Big], \quad I_0 = \Big[\frac{(N-1)s_0^2}{\chi^2_{N-1,\alpha/2}}, \frac{(N-1)s_0^2}{\chi^2_{N-1,1-\alpha/2}}\Big].$$

Obviamente, para la desviación típica σ, basta con considerar la raíz cuadrada de los extremos del intervalo de confianza de la varianza.

Ejercicio. *En un instituto se está investigando una determinada enfermedad. Se eligen* 20 *ratas aleatoriamente y se les provoca dicha enfermedad. Posteriormente, se les trata con un medicamento. Se elige como variable respuesta*

número de horas que tarda en curarse,

obteniéndose para la media $\hat{\mu}_0 = 10$ *horas. Si sabemos que la desviación típica* $\sigma = 99$ *horas, calcular el intervalo de confianza al* 99% *para la media. ¿Qué tamaño de muestra se necesita para que la longitud del intervalo sea de* 5 *horas?*

90%, $1 - \alpha = 0'90$, $\alpha/2 = 0'05$, $z_{\alpha/2} = 1'645$. Luego, con este valor, el intervalo de confianza al 95% es

$$I_0 = [10 \pm 1'645\frac{99}{\sqrt{20}}] = [10 \pm 36'4154].$$

De otro lado, como la longitud del intervalo de confianza (para una muestra de tamaño N) es $2 \cdot 1'645 \cdot 99/\sqrt{N}$, si queremos que dicha longitud sea 5, resulta que $N = 4243'48$. Luego $N = 4244$.

Ejercicio. *Queremos conocer el intervalo de confianza al* 95% *de la estancia media de los enfermos de un hospital. Para ello, tomamos una muestra de tamaño* 300, *obteniendo para la media* $\hat{\mu}_0 = 8$ *y para la desviación típica* 12. *Calcular el intervalo de confianza.*

95%, $1-\alpha = 0'95$, $\alpha/2 = 0'025$, $z_{\alpha/2} = 1'96$. De otro lado, para la cuasi varianza muestral, resulta

$$s_0^2 = \frac{N}{N-1}\sigma^2, \Rightarrow s_0^2 = \frac{300}{299}12^2.$$

Luego, con este cálculo, el intervalo de confianza al 90% es

$$I_0 = \left[8 \pm 1'96\frac{\sqrt{\frac{300}{299}}12}{\sqrt{300}}\right] = [8 \pm 1'36].$$

Ejercicio. *Se quiere probar la efectividad de un antitérmico en la reducción de la temperatura. Para ello, se tomó la temperatura a* 10 *niños con gripe, y después de haberles suministrado el antitérmico se obtuvieron los siguientes resultados:*

$$1'2,\ 1'7,\ 1'6,\ 1'7,\ 1,\ 1,\ 1,\ 2'6,\ 3,\ 1.$$

Hallar el intervalo de confianza al 95% *para la media.*

Notemos que de la muestra, obtenemos los siguientes datos: media $\hat{\mu}_0 = 1'58$, varianza muestral $0'4576$, luego la cuasi varianza es $s_0^2 = \frac{10}{9}0'4576 = 0'50844$, por lo tanto $s_0 = 0'710$. De otro lado, como el nivel de confianza es 95%, $1-\alpha = 0'95$, $\alpha/2 = 0'025$. Por lo tanto $t_{9,0'025} = 2'26$. Finalmente, el intervalo de confianza es

$$I_0 = \left[1'58 \pm 2'26\frac{0'71}{\sqrt{10}}\right] = [1'58 \pm 0'5090] \simeq [1'58 \pm 0'51].$$

Ejercicio. *De una muestra de tamaño* 10*, se ha obtenido 0'000204999 de varianza. Encontrar el intervalo de confianza al* 95% *para la varianza y la desviación típica.*

En primer lugar tenemos para la cuasi varianza muestral $s_0^2 = \frac{10}{9}0'000204999 = 2'7776 \cdot 10^{-4}$. Ahora, como el nivel de confianza es 95%, $1-\alpha = 0'95$, $\alpha/2 = 0'025$. Por lo tanto $\chi^2_{9,0'025} = 2'70$, y $\chi^2_{9,0'975} = 19$. Finalmente, el intervalo de confianza para la varianza es

$$I_0 = \left[\frac{9s_0^2}{19}, \frac{9s_0^2}{2'7}\right] = [0'0001078, 0'000759].$$

Para la desviación típica $I_0 = [0'010387, 0'0275]$.

4.4 Ejercicios

1) Las medidas de los diámetros de una muestra aleatoria de 200 bolas de rodamientos dieron una media de $0'824$ cm, y una desviación típica de $0'042$ cm. Hallar los intervalos de confianza al

$$90\%,\ 95\%,\ 98\%,\ 99\%,\ 99'73\%,$$

para el diámetro medio de las bolas.

2) Al medir los tiempos de reacción, un psicólogo estima que la desviación típica $\sigma = 0'05$ segundos. ¿De qué tamaño ha de tomarse una muestra de medias para tener una confianza

a) al 95%, b) al ,

de que el error de la estimación de medias no supera los $0'01$ segundos.

3) Una muestra al azar de 50 notas de examen, revela una media de 75 y una desviación típica de 10. Hallar el intervalo de confianza al 95% para la media.

4) La desviación típica de una muestra de 20 bombillas es de 100 horas. Hallar los intervalos de confianza al

$$90\%,\ 98\%,\ 99\%,$$

para la desviación típica de ese tipo de bombillas.

5 Una muestra de 10 medidas del diámetro de una esfera da de media $4'38$ cm, y de desviación típica $0'06$ cm. Hallar el intervalo de confianza al

$$95\%,\ 99\%,$$

para el diámetro de la esfera.

6) Repetir el problema anterior para una muestra de tamaño 12, con media $7'38$ cm, y desviación típica $1'24$ cm.

7) La desviación típica de las alturas de una muestra de 25 estudiantes es de $2'40$. Hallar el intervalo de confianza al

$$95\%,\ 99\%,$$

para la desviación típica.

8) La desviación típica de la vida media de una muestra de 10 bombillas es de 120 horas. Hallar el intervalo de confianza al

$$95\%, 99\%,$$

para la desviación típica.

9) Un nuevo medicamento ha curado a 80 de 200 enfermos. Calcular el intervalo de confianza al 99% para la proporción de personas curadas, si dicho medicamento se hubiera aplicado a todos los individuos con la misma enfermedad.

Capítulo 5

Regresión. Correlación

5.1 Regresión

Consideramos un experimento tal que cada realización produce dos valores reales, por ejemplo peso y altura. Llamando X a la variable que asigna el primer valor, e Y a la que asigna el segundo, nos preguntamos sobre la posible relación entre ambas variables. En concreto, si representamos los datos obtenidos en el plano

el problema de la *regresión* trata de ajustar esa nube de puntos obtenida, a una ecuación conocida, con el objeto de predecir una variable en función de las otras.

Una de las primeras cuestiones es elegir ese tipo de ecuación, y la segunda cuestión es establecer el criterio de la elección de la ecuación concreta. Por ejemplo, si buscamos una recta, de entre todas ¿cuál elegiremos?

Una vez decidido el tipo de ecuación a buscar, el criterio que se sigue es el *método de los mínimos cuadrados.*

Dicho método consiste en minimizar las distancias d_i, que no son las distancias de los puntos a la recta.

Comentaremos que se utilizan, normalmente, varios tipos de funciones o ecuaciones, llamadas *líneas de regresión*, dando lugar a las regresiones: *lineal* cuando se busca una recta, *parabólica* (parábola), *exponencial* (función exponencial), *polinomial* (polinomio de un determinado grado), *hiperbólica* (una hipérbola), etc.

En este curso vamos a exponer con detalle solamente la regresión lineal llamada también *regresión lineal mínimo cuadrática.* Concretando, supongamos que tenemos una nube de puntos

$$(x_i, y_i), \quad i = 1, 2, ..., N.$$

Buscamos una recta $y = \beta x + \alpha$ de manera que la suma de las d_i sea mínima. Pero como $d_i = |y_i - (\beta x_i + \alpha)|$, y el valor absoluto es una función que presenta ciertas complicaciones en su tratamiento, lo que se hace en realidad es minimizar la función

$$f(\alpha, \beta) = \sum_{i=1}^{N} \big(y_i - (\beta x_i + \alpha)\big)^2,$$

que es mas manejable. Por ejemplo, es derivable. Para calcular el mínimo de dicha función (de dos variables), en primer lugar la derivamos respecto a α y β (derivadas parciales), obteniendo:

$$\begin{aligned}
\frac{\partial f}{\partial \beta} &= \sum_{i=1}^{N} \frac{\partial}{\partial \beta} (y_i - \beta x_i - \alpha)^2 &= \sum_{i=1}^{N} -2x_i(y_i - \beta x_i - \alpha) \\
\frac{\partial f}{\partial \alpha} &= \sum_{i=1}^{N} \frac{\partial}{\partial \alpha} (y_i - \beta x_i - \alpha)^2 &= \sum_{i=1}^{N} -2(y_i - \beta x_i - \alpha).
\end{aligned}$$

Ahora, igualándolas a 0, resulta

$$\left.\begin{array}{c}
\displaystyle\sum_{i=1}^{N} x_i(y_i - \beta x_i - \alpha) = 0 \\
\displaystyle\sum_{i=1}^{N} (y_i - \beta x_i - \alpha) = 0
\end{array}\right\} \Rightarrow \left\{\begin{array}{c}
\displaystyle\sum_{i=1}^{N} x_i y_i - \beta \sum_{i=1}^{N} x_i^2 - \alpha \sum_{i=1}^{N} x_i = 0 \\
\displaystyle\sum_{i=1}^{N} y_i - \beta \sum_{i=1}^{N} x_i - \alpha \sum_{i=1}^{N} 1 = 0.
\end{array}\right.$$

Luego, tenemos un sistema de dos ecuaciones con dos incógnitas: α y β. Consideremos ahora dos variables aleatorias X, Y. Entonces, tenemos que, denotando por

$$\begin{cases} \bar{x} = EX = \sum_{i=1}^{N} \frac{x_i}{N} & \text{media de X,} \\ \bar{y} = EY = \sum_{i=1}^{N} \frac{y_i}{N} & \text{media de Y,} \\ \sigma_x^2 & \text{varianza de X,} \end{cases}$$

y llamando *covarianza* de X, Y a la expresión

$$\text{Covar}(X,Y) = \frac{1}{N} \sum_{i=1}^{N} (x_i y_i - \bar{x}\bar{y}),$$

se puede demostrar que la solución del sistema es

$$\beta = \frac{\text{Covar}(X,Y)}{\sigma_x^2}, \quad \alpha = \bar{y} - \frac{\text{Covar}(X,Y)}{\sigma_x^2}\bar{x}.$$

Finalmente, la expresión de la recta de regresión es

$$(y - \bar{y}) = \frac{\text{Covar}(X,Y)}{\sigma_x^2}(x - \bar{x}).$$

Dicha recta se llama *recta de regresión de Y sobre X*. La pendiente de dicha recta

$$\beta = \frac{\text{Covar}(X,Y)}{\sigma_x^2} = \beta_{y_x}$$

es el *coeficiente de regresión de Y sobre X*.

De igual manera, se puede construir la recta de regresión de X sobre Y, sin más que intercambiar los papeles. Así, la ecuación de dicha recta es

$$(x - \bar{x}) = \frac{\text{Covar}(X,Y)}{\sigma_y^2}(y - \bar{y}).$$

Observación: en la ecuación de la recta de regresión de Y sobre X, aparece σ_x^2 dividiendo. ¿Qué ocurrirá si $\sigma_x^2 = 0$? En este caso, si la varianza de X es igual a 0, esto quiere decir que $x_1 = x_2 = \cdots = x_N$. O lo que es lo mismo, los puntos están alineados, en una recta vertical, concretamente, la recta $x = \bar{x}$.

Ejercicio. *La siguiente tabla, muestra la natalidad por* 1000 *individuos en España, durante el periodo* $1969-1973$*:*

X	1969	1970	1971	1972	1973
Y	56	59	64	65	73

Calcular la recta de regresión de Y *sobre* X*. ¿Cuántos nacimientos se esperan en* 1978*?*

La recta de regresión es $(y-63'4)=4(x-1971)$. Luego, sustituyendo $x=1978$ en dicha ecuación, resulta que el número de nacimientos que se esperan en 1978 por 1000 habitantes es $91'4$.

5.2 Correlación

Hemos visto que la regresión estudia la predicción de una variable, en función de otra. La *correlación* estudia el grado de dependencia que existe entre las variables, intentando cuantificarla mediante en cálculo de los coeficientes de correlación. De entre todos ellos, en este curso trataremos solamente el *coeficiente de correlación lineal.*

<u>Coeficiente de correlación lineal:</u> (ρ) es un número abstracto, que nos determina el grado de ajuste a una recta de regresión. Viene definido por:

$$\rho=\sqrt{\beta_{x_y}\beta_{y_x}}=\frac{\mathrm{Covar}(X,Y)}{\sigma_x\sigma_y}.$$

Dos propiedades del coeficiente de correlación lineal, son las siguientes:
1ª) $-1\leq\rho\leq 1$,
2ª) si $\rho=\pm 1$, entonces los puntos están en una recta.

Observaciones:

$\rho=-1$, los puntos se encuentran en una recta. Existe una relación funcional $y=\beta x+\alpha,\quad \beta<0$. La recta de regresión es decreciente, puesto que $\mathrm{Covar}(X,Y)<0$.

$\rho=1$, igual que antes, salvo que la recta de regresión es creciente.

$\rho=0$, entonces $\mathrm{Covar}(X,Y)=0$. Esto quiere decir que las rectas de regresión son perpendiculares, y las variables *incorreladas.*

En consecuencia, el grado de correlación de las variables aumenta cuando ρ tiende a ± 1.

En el ejemplo de la sección anterior, se comprueba que $\rho = 0'972433$. Por lo tanto, podemos concluir que las variables están bastante correladas, en sentido creciente.

Ejercicio. *Calcular las rectas de regresión y el coeficiente de correlación de las variables dadas por la siguiente tabla:*

X	$1, 1, 2, 2, 4, 4, 5, 5$
Y	$1, 5, 2, 4, 2, 4, 1, 5$

$\bar{x} = \bar{y} = 3$, $\sigma_x^2 = 2$, $\sigma_y^2 = 2'5$.

$$\text{Covar}(X, Y) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y} = 0.$$

Recta de regresión de Y sobre X: $y = 3$,
Recta de regresión de X sobre Y: $x = 3$,
$\rho = 0$. Luego las variables son incorreladas.

Ejercicio. *Calcular las rectas de regresión y el coeficiente de correlación de las variables dadas por la siguiente tabla:*

X	$65, 50, 55, 65, 55, 70, 65, 70, 55, 70, 50, 50$
Y	$85, 79, 76, 90, 85, 87, 94, 98, 81, 95, 76, 74$

Cuando X toma el valor 80, ¿qué valor podemos predecir para Y? Cuando Y toma el valor 100, ¿qué valor podemos predecir para X?

$\bar{x} = 60$, $\bar{y} = 85$, $\sigma_x^2 = 62'5$, $\sigma_y^2 = 59'5$, $\text{Covar}(X, Y) = 53'75$.
Recta de regresión de Y sobre X:

$$y = \frac{43}{50}(x - 60) + 85, \qquad x = 60 \Rightarrow y = 102'2.$$

Recta de regresión de X sobre Y

$$x = \frac{215}{238}(y - 85) + 60, \qquad y = 100 \Rightarrow x = 73'5504.$$

$\rho = 0'881413$, lo que indica que las variables tienen una dependencia funcional creciente y “grande”.

Capítulo 6

Bibliografía

Curso y Ejercicios de Estadística, V. Quesada, A. Isidoro, L.A. López, Ed. Alhambra Universidad, Madrid (1992).

Es un libro que se adapta a los contenidos del curso, con teoría y problemas.

Curso y Ejercicios de Bioestadística, J. R. Vizmanos, R. Asensio, CPR, Madrid (1976).

Sobre todo problemas.

Estadística (segunda edición), M.R. Spiegel, Ed. McGraw-Hill, Madrid (1991).

Una colección de problemas de ecuaciones diferenciales. Libro de la serie Schaum.