Parte II

PROBABILIDAD

Capítulo 3

TEORÍA ELEMENTAL DE LA PROBABILIDAD

La **Estadística** es la ciencia que se dedica al estudio de **fenómenos aleatorios** (o las variables asociadas a ellos) en los que se presenta un cierto grado de incertidumbre, en contraposición a los fenómenos deterministas estudiados en las ciencias clásicas (Biología, Física, Química).

Se define entonces un **experimento aleatorio** como áquel en el que no sabemos de antemano cuál es el resultado que vamos a obtener, es decir, existe una incertidumbre acerca del resultado que obtendremos. Sin embargo es habitual que, ante un experimento de este tipo, sí tengamos un cierto conocimiento sobre la mayor o menor posibilidad de aparición de los diferentes resultados posibles. Este conocimiento es el que vamos a tratar de recoger y formalizar mediante el concepto de probabilidad.

Para ello definimos en primer lugar el **espacio muestral** Ω asociado al experimento aleatorio como el conjunto de todos los posibles resultados que pudieran ocurrir. Si es posible hacer una lista o enumerar todos estos posibles resultados diremos que el espacio muestral Ω es **discreto** y por tanto necesariamente tiene que ser un conjunto finito o infinito pero numerable. En caso contrario, es decir, si hay una infinidad no numerable de posibles resultados, diremos que el espacio muestral Ω es **continuo**.

Cualquier subconjunto A contenido en Ω ($A \subset \Omega$) diremos que es un **suceso** asociado al experimento aleatorio y el objetivo será dar una medida de la mayor o menor posibilidad de aparición de cada uno de los posibles sucesos. La teoría elemental de la probabilidad empieza entonces con el manejo de la **teoría de conjuntos**, que necesita por tanto ser recordada. A continuación hacemos un rápido repaso de los conceptos necesarios.

Los subconjuntos de Ω formados por un único elemento $\{\omega\}$ reciben el nombre de **sucesos elementales** (subconjuntos unipuntuales) y el subconjunto formado por todo el espacio muestral Ω recibe el nombre de **suceso seguro** (recordar que todo conjunto puede ser considerado como un subconjunto de sí mismo). Al igual que en teoría de conjuntos también se considera el subconjunto vacío que no contiene ningún elemento ($\varnothing$) y recibe el nombre de **suceso imposible**. Todo suceso $A \subset \Omega$ tiene un **suceso**

contrario $\bar{A} \subset \Omega$ formado por todos los elementos de Ω que no están en A (conjunto complementario en la teoría de conjuntos). Dados dos subconjuntos $A, B \subset \Omega$ su **intersección** se define como el subconjunto formado por todos los elementos comunes a ambos y se denota por $A \cap B$ (o simplemente AB). En teoría de la probabilidad esta intersección AB se utilizará para denotar la ocurrencia simultánea de ambos sucesos, y si esta ocurrencia simultánea es imposible (es decir $AB = \varnothing$) diremos que los **sucesos** son **incompatibles** (conjuntos disjuntos). Del mismo modo, dados dos subconjuntos $A, B \subset \Omega$ su **unión** se define como el subconjunto formado por todos los elementos que pertenecen al menos a uno de los dos conjuntos y se denota por $A \cup B$. En teoría de la probabilidad esta unión $A \cup B$ se utilizará para denotar la ocurrencia de al menos uno de los dos sucesos. En el caso particular de que los sucesos implicados en esta unión sean incompatibles (es decir $AB = \varnothing$) utilizaremos la notación $A + B$ en lugar de $A \cup B$; de modo que siempre que escribamos $A + B$ estaremos suponiendo que los sucesos A y B no pueden ocurrir simultáneamente. Además, dados dos sucesos A y B diremos que A esta contenido en B ($A \subset B$) cuando todos los elementos de A pertenecen también a B y, por tanto, la ocurrencia del suceso A implica necesariamente la ocurrencia del suceso B. Finalmente, denotaremos por $B - A$ al conjunto formado por los elementos que están en B pero no están en A y este suceso representará la ocurrencia de B sin que ocurra el suceso A (observar que este conjunto pudiera representarse también como $\bar{A}B$).

Para terminar, recopilamos una serie de propiedades de la teoría de conjuntos que necesitaremos en teoría de la probabilidad:

1. $A \cup (B \cup C) = (A \cup B) \cup C$ (propiedad asociativa de la unión).
2. $A\,(BC) = (AB)\,C$ (propiedad asociativa de la intersección).
3. $A \cup (BC) = (A \cup B)\,(A \cup C)$ (propiedad distributiva de la unión respecto de la intersección).
4. $A\,(B \cup C) = (AB) \cup (AC)$ (propiedad distributiva de la intersección respecto de la unión).
5. $\overline{A \cup B} = \bar{A}\ \bar{B}$ (primera ley de De Morgan).
6. $\overline{AB} = \bar{A} \cup \bar{B}$ (segunda ley de De Morgan).

Otra materia importante en el desarrollo de la teoría de la probabilidad es la **Combinatoria**, que se dedica a estudiar las reglas de conteo en conjuntos finitos. Repasamos a continuación los conceptos necesarios.

Variaciones ordinarias (sin repetición). Dado un conjunto de n elementos distintos y dado un número natural $r \leq n$ se pretende calcular todas las posibles maneras de elegir r elementos de los n existentes sin repetir ninguno (suponemos que las extracciones se hacen sin devolución) y teniendo en cuenta el orden en que son elegidos (es decir las mismas extracciones en otro orden son consideradas distintas). Para ello se definen las variaciones sin repetición $V_{n,r}$ como:

$$V_{n,r} = n(n-1)...(n-r+1)$$

Por ejemplo, en una competición de 8 atletas las posibles maneras de elegir 3 para formar el podium serán $V_{8,3} = 8 \cdot 7 \cdot 6 = 336$, de modo que existen 336 posible resultados para el podium de la competición (el orden importa porque no es lo mismo quedar primero que quedar segundo y no se puede repetir porque un mismo atleta no puede ser a la vez primero y segundo).

Permutaciones ordinarias (sin repetición). Son un caso particular del anterior cuando $r = n$ y se representan por P_n. Es decir

$$P_n = V_{n,n} = n(n-1)...1 = n!$$

Por ejemplo, en la competición de 8 atletas del ejemplo anterior, el número total de posibles clasificaciones será $P_8 = 8! = 40320$.

Combinaciones ordinarias (sin repetición). Dado un conjunto de n elementos distintos y dado un número natural $r \leq n$ se pretende calcular todas las posibles maneras de elegir r elementos de los n existentes sin repetir ninguno y sin tener en cuenta el orden en que son elegidos. Por tanto, esto equivale al número posible de subconjuntos de r elementos que pueden extraerse de un conjunto formado por n elementos. Para ello se definen las combinaciones sin repetición $C_{n,r}$ como:

$$C_{n,r} = \frac{V_{n,r}}{P_r} = \frac{n(n-1)...(n-r+1)}{r(r-1)...1} = \frac{n!}{r!\ (n-r)!} = \binom{n}{r}$$

En particular, el número de subconjuntos que pueden formarse con un conjunto de n elementos (es decir, el número de sucesos que existen en un espacio muestral Ω con n sucesos elementales) será:

$$\binom{n}{0} + \binom{n}{1} + ... + \binom{n}{n-1} + \binom{n}{n} = (1+1)^n = 2^n$$

Es importante observar que, por la simetría de la definición, se verifica que

$$C_{n,r} = \binom{n}{r} = \binom{n}{n-r} = C_{n,n-r}$$

y además se tiene que

$$\begin{aligned}
\binom{n}{r} + \binom{n}{r+1} &= \frac{n(n-1)...(n-r+1)}{r(r-1)...1} + \frac{n(n-1)...(n-r+1)(n-r)}{(r+1)r(r-1)...1} \\
&= \frac{[n(n-1)...(n-r+1)(r+1)] + [n(n-1)...(n-r+1)(n-r)]}{(r+1)r(r-1)...1} \\
&= \frac{n(n-1)...(n-r+1)\,(r+1+n-r)}{(r+1)r(r-1)...1} = \frac{(n+1)n(n-1)...(n-r+1)}{(r+1)r(r-1)...1} \\
&= \binom{n+1}{r+1}
\end{aligned}$$

Por ejemplo, el número de posibles combinaciones ganadoras en el sorteo de la lotería primitiva será $C_{49,6} = \binom{49}{6} = \frac{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 13983816$ y el número total de sucesos que pueden definirse en el experimento aleatorio consistente en el lanzamiento de un dado y observar el resultado será $2^6 = 64$.

Variaciones con repetición. Dado un conjunto de n elementos distintos y dado un número natural r (no necesariamente con $r \leq n$) se pretende calcular todas las posibles maneras de elegir r elementos de los n existentes permitiendo su repetición (suponemos que las extracciones se hacen con devolución) y teniendo en cuenta el orden en que son elegidos (es decir las mismas extracciones en otro orden son consideradas distintas). Para ello se definen las variaciones con repetición $VR_{n,r}$ como:

$$VR_{n,r} = n^r$$

Por ejemplo, el número de posibles resultados para una quiniela de fútbol será $VR_{3,14} = 3^{14} = 4782969$.

Variaciones con repetición restringida. En las mismas circunstancias del caso anterior, podemos considerar todas las posibles maneras de elegir r elementos de los n existentes permitiendo su repetición, pero con la restricción de que el primer elemento tiene que aparecer r_1 veces, el segundo elemento r_2 y así sucesivamente hasta el último elemento que tiene que aparecer r_n veces (por tanto deberá ser $r = \sum_{i=1}^{n} r_i$). Para ello se definen las variaciones con repetición restringida de n elementos con r_i repeticiones para cada uno de ellos y $\sum_i r_i = r$ como

$$VR_{n,r}^{r_1,r_2,\ldots,r_n} = \frac{r!}{r_1!\ r_2!...r_n!} = \frac{(r_1 + r_2 + ... + r_n)!}{r_1!\ r_2!...r_n!}$$

Por ejemplo el número de posibles quinielas de fútbol con 7 unos, 4 equis y 3 doses son $VR_{3,14}^{7,4,3} = \frac{14!}{7!\ 4!\ 3!} = 120120$.

Combinaciones con repetición. Dado un conjunto de n elementos distintos y dado un número natural r (no necesariamente con $r \leq n$) se pretende calcular todos los posibles conjuntos de r elementos que pueden formarse con los n existentes permitiendo su repetición (suponemos que las extracciones se hacen con devolución y el orden no se tiene en cuenta). Para ello se definen las combinaciones con repetición $CR_{n,r}$ como

$$CR_{n,r} = C_{n+r-1,r} = \binom{n+r-1}{r}$$

Por ejemplo, el número posible de conjuntos de 3 elementos que puede formarse con las cifras $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ si permitimos la repetición de cifras es $CR_{10,3} = C_{12,3} = \binom{12}{3} = \frac{12 \cdot 11 \cdot 10}{3 \cdot 2 \cdot 1} = 220$, mientras que, sin repetir cifras, sería $C_{10,3} = \binom{10}{3} = 120$.

Concepto de probabilidad. Una vez conocidas estas herramientas de la Teoría de Conjuntos y la Combinatoria podemos comenzar con la Teoría de la Probabilidad. Históricamente esta teoría empieza con la definición de probabilidad en experimentos con espacio muestral finito y suponiendo que todos los elementos son igualmente probables. En estas circunstancias, si Ω es el espacio muestral finito con n elementos y A es un subconjunto de Ω se define la probabilidad de A como

$$p(A) = \frac{\text{n}^o \text{ de elementos de } A}{\text{n}^o \text{ de elementos de } \Omega} = \frac{\text{casos favorables}}{\text{casos posibles}} \qquad \textbf{(Regla de Laplace)}$$

Sin embargo esta definición no es siempre aplicable, porque puede ocurrir que Ω no sea finito o no todos sus elementos sean igualmente probables. Por ejemplo, si consideramos el experimento consistente en lanzar dos monedas indistinguibles a la vez es claro que el espacio muestral estaría formado por 3 elementos (dos caras, dos cruces o una cara y una cruz) pero no parece admisible que los 3 sean igualmente probables.

Si el experimento aleatorio puede repetirse indefinidamente en las mismas circunstancias, puede definirse la probabilidad mediante el límite de las frecuencias relativas de aparición de ese suceso cuando el número de repeticiones del experimento tiende a infinito (**interpretación frecuentista**). Por ejemplo, en el caso anterior de las dos monedas indistinguibles se observaría que la probabilidad de una cara y una cruz es 0.5, mientras que para el caso de dos caras o dos cruces la probabilidad es 0.25. Pero tampoco esta definición parece universal, ni práctica, porque puede ocurrir que el experimento no pueda repetirse en las mismas circunstancias, o sea muy difícil o costosa la repetición del experimento.

Por todo esto es necesario acudir a una **definición axiómática** y puramente matemática del concepto de probabilidad. En concreto, dado un experimento aleatorio con espacio muestral Ω se define una **probabilidad p sobre** Ω como cualquier función matemática que a cada subconjunto A le hace corresponder un número $p(A)$ de modo que obligatoriamente se verifiquen las siguientes propiedades (axiomas):

1. $p(A) \in [0,1]$ para todo $A \subset \Omega$.
2. $p(\Omega) = 1$.
3. Si A y B son dos sucesos incompatibles (es decir, $AB = \varnothing$) entonces $p(A+B) = p(A) + p(B)$.

Esta definición de modelo de probabilidad está sugerida por las propiedades de las frecuencias relativas, estudiadas en Estadística Descriptiva. Por supuesto, esta definición no resuelve el problema de asignar probabilidades a diferentes sucesos en una situación determinada; lo único que hace es imponer ciertas condiciones de coherencia para cualquier modelo de probabilidad. A partir de estas condiciones mínimas de coherencia se pueden obtener, sin embargo, muchas propiedades que serán muy útiles, en diferentes situaciones, para llevar a cabo el cálculo de probabilidades de sucesos complicados. A continuación recopilamos las fundamentales.

Propiedades de un modelo de probabilidad.

1. $p(\bar{A}) = 1 - p(A)$
2. $p(\varnothing) = 0$
3. Si $A \subset B$ entonces $p(A) \leq p(B)$ y $p(B-A) = p(B) - p(A)$
4. Si $A_1, A_2, ..., A_n ...$ son sucesos disjuntos dos a dos (incompatibles) entonces $p\left(\sum_i A_i\right) = \sum_i p(A_i)$
5. Si A y B son dos sucesos cualesquiera entonces $p(A \cup B) = p(A) + p(B) - p(AB)$
6. En general, si $A_1, A_2, ..., A_n$ son sucesos cualesquiera entonces
$$p(\cup_{i=1}^n A_i) = \sum_{i=1}^n p(A_i) - \sum_{i<j} p(A_i A_j) + \sum_{i<j<k} p(A_i A_j A_k) - ... + (-1)^{n+1} p(\cap_{i=1}^n A_i)$$
Como caso particular $p(A \cup B \cup C) = p(A) + p(B) + p(C) - p(AB) - p(AC) - p(BC) + p(ABC)$
7. En general, si $A_1, A_2, ..., A_n ...$ son sucesos cualesquiera entonces $p(\cup_i A_i) \leq \sum_i p(A_i)$

Teniendo en cuenta la definición axiomática y la propiedades anteriores, es claro que, cuando el espacio muestral es discreto (finito o infinito pero numerable) entonces el modelo de probabilidad queda perfectamente especificado dando la probabilidad de cada suceso elemental, es decir, sólo tendríamos que definir $p(a_1), p(a_2), ..., p(a_i), ...$ verificando $p(a_i) \geq 0$ para todo a_i y $\sum_i p(a_i) = 1$. En efecto, evidentemente, para cualquier suceso $B = \{b_1, ..., b_i, ...\} \subset \{a_1, ..., a_i, ...\} = \Omega$ tendríamos
$$p(B) = p\left(\sum_i \{b_i\}\right) = \sum_i p(b_i)$$

Probabilidad condicionada. En muchas situaciones prácticas podemos tener un conocimiento parcial de lo que ha ocurrido en el experimento aleatorio y esa información puede alterar la asignación de probabilidades para los sucesos del espacio muestral. Por ejemplo, en el experimento aleatorio consistente en el lanzamiento de un dado y anotar el resultado, si consideramos el suceso $A = \{1, 2, 4\}$ y admitimos que todos los resultados son igualmente probables, la asignación lógica de probabilidades vendría dada

por la Regla de Laplace y por tanto diríamos que $p(A) = \frac{3}{6} = 0.5$. Sin embargo, si tenemos el conocimiento parcial de que el resultado del lanzamiento ha sido un número par, los posibles resultados del lanzamiento se reducen al conjunto $B = \{2, 4, 6\}$, que es un suceso del espacio muestral inicial con $p(B) = 0.5 > 0$. ¿Seguiríamos diciendo ahora que la probabilidad de A es 0.5? Parece lógico pensar que no, porque los casos favorables ahora son 2 (el 2 y el 4) mientras que los casos posibles son 3 (el 2, el 4 y el 6); por tanto ahora diríamos que la probabilidad de A es $2/3 = 0.667$. Necesitamos por tanto definir un nuevo concepto de probabilidad para el caso de que tengamos un conocimiento parcial de lo que haya ocurrido en el experimento. Este es el motivo de la siguiente definición.

Dado un experimento aleatorio con espacio muestral Ω y probabilidad asociada p, y dado un suceso $B \subset \Omega$ con $p(B) > 0$, para cualquier suceso $A \subset \Omega$ se define la **probabilidad condicionada del suceso A sabiendo que ha ocurrido B** como

$$p(A/B) = \frac{p(AB)}{p(B)}$$

Por tanto, la definición de probabilidad condicionada sólo tiene sentido para sucesos B con $p(B) > 0$, y por ello en todas las probabilidades condicionadas que aparezcan se supondrá esto de manera implícita. Aunque hay veces que la asignación de probabilidades condicionadas es muy sencilla utilizando simplemente la intuición (ejemplo anterior), en otras circunstancias esto no es así y será en esos casos donde la definición anterior se hace totalmente necesaria

Independencia de sucesos. Otro concepto importante en teoría de probabilidad es la independencia de sucesos. Desde un punto de vista intuitivo, diremos que dos sucesos son independientes cuando la ocurrencia de uno de ellos no nos dice nada nuevo sobre la posible ocurrencia del otro. Esta es la idea que se intenta formalizar con la siguiente definición.

Diremos que dos **sucesos A y B son independientes** cuando $p(AB) = P(A)P(B)$. De esta definición, y utilizando el concepto de probabilidad condicionada, se deduce inmediatamente que, si $p(B) > 0$, entonces A y B son independientes si y sólo si $p(A/B) = p(A)$, lo cual refleja la idea intuitiva de que la ocurrencia de B no nos dice nada nuevo sobre la posible ocurrencia del suceso A, puesto que la probabilidad sigue siendo la misma. Del mismo modo, si $p(A) > 0$, entonces A y B son independientes si y sólo si $p(B/A) = p(B)$ y por tanto el concepto de independencia es una cosa mutua entre dos sucesos (si A es independiente de B también B es independiente de A).

Ejercicio. Consideramos el experimento aleatorio consistente en lanzar una moneda equilibrada dos veces. Sea el suceso $A = \{\text{cara en el primer lanzamiento}\}$ y el suceso $B = \{\text{cara en el segundo lanzamiento}\}$, ¿son ambos sucesos independientes?. Sea ahora $C = \{\text{diferente resultado en los dos lanzamientos}\}$, ¿son A y C independientes?.

El concepto de independencia puede generalizarse al caso de más de dos sucesos. En efecto, diremos que los sucesos $A_1, A_2, ..., A_n$ son independientes cuando, para cualquier colección de ellos, $A_i, ..., A_j$ se verifica que $p(A_i...A_j) = p(A_i) \cdot \cdot \cdot p(A_j)$. Es importante observar que puede ocurrir que los sucesos $A_1, A_2, ..., A_n$ sean independientes dos a dos, es decir $p(A_iA_j) = p(A_i)p(A_j)$ para cualesquiera i y j,

y sin embargo no sean globalmente independientes en el sentido de la definición dada. Por eso, cuando tengamos más de dos sucesos podemos hablar de independencia global o de independencia dos a dos.

Ejercicio. Comprobar que los sucesos A, B y C del ejercicio anterior son independientes dos a dos y sin embargo no son globalmente independientes.

Es importante no confundir sucesos independientes con sucesos incompatibles (que no pueden ocurrir simultáneamente). De hecho, si dos sucesos A y B, con $p(A) > 0$ y $p(B) > 0$, son incompatibles no pueden ser nunca independientes porque $p(AB) = 0 \neq p(A)p(B) > 0$.

Terminaremos la teoría de la probabilidad estudiando las tres reglas fundamentales que podemos utilizar para calcular probabilidades asociadas a experimentos aleatorios.

Regla de la multiplicación. Dados n sucesos cualesquiera $A_1, A_2, ..., A_n$ de un espacio muestral Ω con probabilidad asociada p, se verifica que

$$p\left(A_1A_2...A_n\right) = p\left(A_1\right)\ p\left(A_2/A_1\right)\ p\left(A_3/A_1A_2\right)\cdots p\left(A_n/A_1A_2...A_{n-1}\right)$$

Por tanto, esta regla nos permitirá calcular la probabilidad de ocurrencia simultánea de varios sucesos, recurriendo a las probabilidades condicionadas de cada suceso suponiendo que han ocurrido todos los anteriores. En muchas situaciones prácticas estas últimas son fácilmente calculables y sin embargo la probabilidad de ocurrencia simultánea es mucho más difícil de calcular. Observar que, en el caso particular de dos sucesos A y B, esta regla nos dice que $p(AB) = p(A)\ p(B/A)$, lo cual no es más que la definición de probabilidad condicionada.

Ejercicio. En una urna hay 20 bolas blancas y 10 bolas negras. Hacemos 3 extracciones sin devolución de la urna. ¿Cuál es la probabilidad de que las 3 sean blancas?

Se define una **partición del espacio muestral** Ω como un conjunto $\{A_1, A_2, ..., A_n\}$ de sucesos incompatibles dos a dos ($A_iA_j = \varnothing$ para todo $i \neq j$) y con $\sum_i A_i = \Omega$.

Regla de la probabilidad total. Dada una partición $\{A_1, A_2, ..., A_n\}$ del espacio muestral Ω y dado un suceso cualquiera $B \subset \Omega$ se verifica que

$$p(B) = \sum_{i=1}^{n} p(A_i)\ p\left(B/A_i\right)$$

En efecto, por ser $\{A_1, A_2, ..., A_n\}$ un partición de Ω se verifica que $B = \sum_{i=1}^{n} A_iB$ y por tanto se tiene que $p(B) = p\left(\sum_{i=1}^{n} A_iB\right) = \sum_{i=1}^{n} p(A_iB) = \sum_{i=1}^{n} p(A_i)\ p\left(B/A_i\right)$. Por tanto, esta regla nos permitirá calcular la probabilidad de cualquier suceso utilizando las probabilidades de ese suceso condicionadas por los elementos de la partición.

Ejercicio. Considerando el mismo experimento que en el ejercicio anterior, ¿cuál es la probabilidad de que la segunda bola sea blanca?. ¿Y la probabilidad de que la tercera bola sea blanca?.

Regla de Bayes. Dada una partición $\{A_1, A_2, ..., A_n\}$ del espacio muestral Ω y dado un suceso cualquiera $B \subset \Omega$ se verifica que, para cualquier j,

$$p\left(A_j/B\right) = \frac{p(A_j)\ p\left(B/A_j\right)}{\sum_{i=1}^{n} p(A_i)\ p\left(B/A_i\right)}$$

En efecto, aplicando la definición de probabilidad condicionada y la regla de la probabilidad total, se tiene que $p\left(A_j/B\right) = \frac{p(A_j B)}{p(B)} = \frac{p(A_j)\ p(B/A_j)}{\sum_{i=1}^{n} p(A_i)\ p(B/A_i)}$. Por tanto, esta regla nos permite actualizar de manera automática las probabilidades de los elementos de la partición (A_j) a la vista de la información adicional que vamos obteniendo con la experimentación (ocurrencia del suceso B).

Ejercicio. Supongamos que tenemos dos urnas. La urna 1 contiene 3 bolas blancas y 2 negras, mientras que la urna 2 contiene 2 bolas blancas y 3 negras. Se elige una urna al azar con probabilidades 1/3 para la primera y 2/3 para la segunda. A continuación se extrae una bola de la urna elegida. ¿Cuál es la probabilidad de que la bola extraída sea blanca?. Si al final nos comunican que la bola extraída ha sido blanca, ¿cuál es la probabilidad de que la bola haya sido extraída de la urna 1?.

3.1. PROBLEMAS

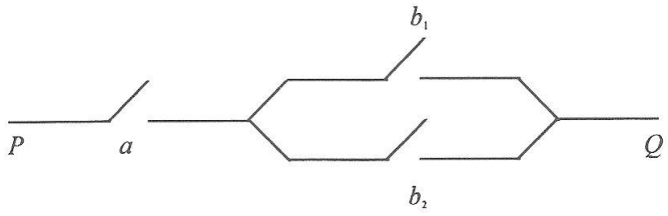
1. Sean A, B y C tres sucesos de un espacio muestral. Encontrar expresiones para los siguientes sucesos en función de A, B y C. Ilustrarlo mediante diagramas de Venn.
   (1) Ocurre A ó B.
   (2) Ocurren A y B.
   (3) Sólo ocurre A.
   (4) Ocurren tanto A como B pero no C.
   (5) Los tres ocurren.
   (6) Ocurre por lo menos uno de los tres.
   (7) Por lo menos dos ocurren.
   (8) Ocurre uno y no más de uno.
   (9) Ocurren dos y no más de dos.
   (10) No ocurre ninguno.
   (11) No ocurren más de dos.

2. Una moneda se lanza tres veces. Se pide:
   (a) Construir el espacio muestral asociado al experimento.
   (b) Expresar en función de los sucesos elementales los siguientes sucesos:
      (1) Los tres lanzamientos producen el mismo resultado.
      (2) El mismo resultado aparece exactamente dos veces.
      (3) Sale cara al menos dos veces
      (4) La cara aparece en el primero y en el segundo lanzamientos.
      (5) Sale cara exactamente dos veces.

3. Un número es seleccionado al azar entre los números 1 al 10. Se A el suceso "el número elegido es par", B el suceso "el número elegido es primo" y C el suceso "el número elegido es múltiplo de 3". Expresar verbalmente los siguientes sucesos y encontrar los elementos que los forman:

   $A \cap B$ $\quad A - B$ $\quad A \cup B$ $\quad \bar{A} \cap \bar{B}$ $\quad A \cap C$ $\quad A - C$ $\quad B \cap C$ $\quad \bar{A} \cap C$ $\quad (A \cup B) \cap \bar{C}$

4. En la figura se muestra el esquema de un circuito eléctrico entre dos puntos P y Q. Sea A el suceso "el interruptor a está cerrado" y B_i "el interruptor b_i está cerrado" $(i = 1, 2)$. Expresar el suceso de que el circuito esté cerrado entre P y Q.



5. Un sistema está construido con elementos de dos tipos. Sea A_k el suceso que la unidad k-ésima del primer tipo está útil y B_j que la unidad j-ésima del segundo tipo está útil. El sistema funciona si al menos están útiles una unidad del primer tipo y una del segundo. Expresar este hecho mediante los sucesos A_k y B_j.

6. De 25 micro computadoras disponibles en un almacén, 10 de ellas tienen tarjeta para impresora, 5 tienen tarjeta adaptadora para módem y 13 no tienen ninguna de éstas. Expresar simbólicamente los siguientes conjuntos y calcular el número de micro computadoras que hay en cada uno de ellos:
   (a) Las que tengan ambas tarjetas.
   (b) Las que no tengan ninguna tarjeta.
   (c) Las que sólo tengan tarjeta para impresora.
   (d) Las que tengan exactamente una de las tarjetas.

7. Simplificar la siguiente expresión: $(A \cup B) \cap (A \cup \bar{B}) \cap (\bar{A} \cup B)$

8. Una caja contiene seis bombillas de las cuales 2 son defectuosas. Se prueban las bombillas hasta encontrar la primera defectuosa. Encontrar el espacio muestral asociado al experimento y la probabilidad de cada uno de los sucesos elementales.

9. Una cadena comercial tiene dos establecimientos en una ciudad. Se sabe que el 30 % de los clientes potenciales compra productos sólo en la tienda 1, el 50 % compra solamente en la tienda 2, el 10 % compra en las tiendas 1 y 2, y el 10 % de los consumidores no compra en ninguna de las dos tiendas. Sea A el suceso en el que el cliente potencial compra en la tienda 1 y B el suceso en el que compra en la tienda 2. Calcular y explicar el significado de las siguientes probabilidades. Expresarlas como porcentajes.
   (a) $p(A)$ $\quad$ (b) $p(A \cup B)$ $\quad$ (c) $p(\bar{B})$ $\quad$ (d) $p(A \cap B)$ $\quad$ (e) $p(A \cup \bar{B})$ $\quad$ (f) $p(\bar{A} \cap \bar{B})$

10. Un experimentador desea averiguar el efecto de tres variables (temperatura, grado de abonado y tipo de semilla) sobre el rendimiento de un cultivo de invernadero. Si el investigador desea analizar cuatro niveles de temperatura, cinco dosis de diferentes de abonado y tres tipos de semilla, ¿cuántos ensayos experimentales ha de llevar a cabo si desea probar todas las combinaciones posibles de temperatura, abonado y tipo de semilla?

11. Un equipo de baloncesto cuenta con 3 bases, 6 aleros y 5 pivots. Si el quinteto inicial está formado por 1 base, 2 aleros y 2 pivots, ¿cuántos equipos titulares se pueden formar?

12. En un banco se pueden sentar 8 personas. Si entre un grupo de 8 amigos hay una pareja de novios, ¿cuál es la probabilidad de que se sienten juntos?

13. ¿Cuál es la probabilidad de que en una quiniela de fútbol haya exactamente cinco x's?

14. Un jugador italiano expresó su sorpresa a Galileo por observar que al jugar con tres dados la suma 10 aparece con más frecuencia que la suma 9. Según el jugador los casos favorables serían:

Casos favorables al 9	Casos favorables al 10
1 2 6	1 3 6
1 3 5	1 4 5
1 4 4	2 2 6
2 2 5	2 3 5
2 3 4	2 4 4
3 3 3	3 3 4

Galileo, en su libro *Considerazione sopra il giuco dei dadi*, vio que estas posibilidades no se pueden considerar igualmente probables. Explicar porqué y calcular las correspondientes probabilidades.

15. Obtener la probabilidad p de que al lanzar dos dados n veces se obtenga al menos un 6 doble. ¿Cuántas partidas habrá que jugar para que tengamos $p = 1/2$ de obtener un 6 doble? (*Problema propuesto a Pascal por el caballero de Mere*)

16. ¿Cuál es la probabilidad de torpedear un barco enemigo sabiendo que sólo pueden lanzarse 3 torpedos y que la probabilidad de hacer blanco con cada uno de ellos es 0.20?

17. Sea un dado tal que la probabilidad de las distintas caras es proporcional al número de puntos inscritos en ellas. Hallar la probabilidad de obtener con este dado un número par.

18. Un examen de reválida de licenciatura consta de 14 temas. Se eligen dos al azar, y el alumno deberá escoger uno para contestarlo. (a) Calcular la probabilidad de que a un alumno que ha preparado 5 temas le toque al menos uno que sabe. (b) ¿Cuál es el número mínimo de temas que debe de preparar para que tenga una probabilidad superior a 1/2 de superar el examen?

19. Se hacen tres disparos simultáneos con tres cañones distintos, siendo la probabilidad de alcanzar el objetivo 0.1, 0.2 y 0.3 , respectivamente. Calcular la probabilidad de cada uno de los números posibles de blancos. Calcular la probabilidad de obtener al menos un blanco.

20. En el lejano reino de Fabulandia a los condenados a muerte se les concedía la gracia de que su vida dependiera de que sacasen una bola blanca en el siguiente sorteo: se ponían 50 bolas blancas en una urna y 50 bolas negras en otra. Se vendaban los ojos al condenado, éste elegía en primer lugar una urna y luego extraía una bola. Mas en cierta ocasión un reo pidió la gracia de que se le dejara distribuir las bolas antes de hacer el sorteo. Tras alguna discusión con los magos de la corte se le concedió la gracia y colocó 1 bola blanca en una urna y en la otra 49 blancas y las 50 negras. ¿Mejoró el reo la probabilidad de salvar su vida?

21. Una empresa tiene establecido un programa que le permite servir todos sus pedidos con un retraso inferior a una semana. De datos anteriores se conoce que esta condición es cumplida en el 95 % de los casos, lo cual se considera como satisfactorio. Reclamaciones por parte de algunos clientes hacen sospechar a la dirección que actualmente hay un mayor número de demoras, por lo cual se plantea revisar el proceso. Para ello establece la siguiente norma: se seleccionan al azar tres pedidos y se procederá a observar cómo son servidos. A la vista de la información que se obtenga procederá de la siguiente manera: Si ninguna de la tres órdenes se retrasa, no se harán más comprobaciones; si una o más órdenes se retrasan se procederá a revisar el proceso. ¿Cuál es la probabilidad de que el proceso sea revisado sin necesidad?

22. En una parcela con 100 árboles sabemos que 30 están enfermos. Si se eligen, al azar, 20 árboles para ser talados, ¿cuál es la probabilidad de que 8 de esos 20 árboles estén enfermos?

23. Sean A y B dos sucesos de un espacio muestral. Supuestas conocidas $p(A)$, $p(B)$ y $P(A \cap B)$, calcular:

    (a) La probabilidad de que ocurran exactamente k de los sucesos A y B $(k = 0, 1, 2)$.

    (b) La probabilidad de que ocurran al menos k de los sucesos A y B $(k = 0, 1, 2)$.

    (c) La probabilidad de que ocurran a lo sumo k de los sucesos A y B $(k = 0, 1, 2)$.

    (d) La probabilidad de que ocurra A pero no B.

24. Se llama fiabilidad de un sistema a la probabilidad de que el sistema funcione con éxito durante un tiempo fijado. Normalmente un sistema está formado por varios subsistemas cuyo éxito o fallo afecta al éxito del sistema total. Un sistema en serie se caracteriza porque el sistema total opera con éxito si y sólo si todos los subsistemas operan con éxito. Un sistema en paralelo se caracteriza porque funciona si al menos uno de los subsistemas funciona.

    (a) Consideremos un sistema formado por n subsistemas con fiabilidad p_i $(i = 1, .., n)$. Obtener la fiabilidad del sistema en función de los p_i: (i) si el sistema está montado en serie. (ii) si el sistema está montado en paralelo.

    (b) Sea A un sistema formado por 10 subsistemas acoplados en serie. Calcular la fiabilidad del sistema si la fiabilidad de cada uno de los subsistemas es de 0.99.

    (c) Sea A un sistema formado por 3 subsistemas acoplados en paralelo. Calcular la fiabilidad del sistema si la fiabilidad de cada uno de los subsistemas es de 0.80.

25. Una flota de nueve taxis se destina al azar a tres aeropuertos A, B y C: 2 taxis van a A, 4 van a B y 3 taxis se destinan al aeropuerto C.

    (a) ¿De cuántos modos se puede hacer esta asignación?

    (b) ¿Cuál es la probabilidad de que el taxi que conduce Juan sea asignado al aeropuerto C?

26. Demostrar que si $1 \leq r \leq n$ entonces $\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r}$

27. Un inspector de calidad tiene a su cargo la inspección de 10 líneas de envasado. Cada mañana (de una semana de 5 días) selecciona al azar una de las líneas para inspeccionarla ese día. Calcular la probabilidad de que seleccione una línea más de una vez durante la semana.

28. El Servicio de Información de Incendios dio la siguiente tabla acerca de los incendios que ocurrieron en Estados Unidos durante los últimos diez años. Todas las cifras de la tabla son en porcentajes y, por ejemplo, el 22 de la esquina superior izquierda señala que el 22 % de los incendios que se registraron en hogares fueron provocados por la calefacción.

Causa del incendio	Hogares	Apartamentos	Casas móviles	Hoteles y moteles	Otros lugares
Calefacción	22	6	22	8	45
Al cocinar	15	24	13	7	0
Materias inflamables	10	15	7	16	8
Por fumar	7	18	6	36	19
Instalación eléctrica	8	5	15	7	28
Otras	38	32	37	26	0
Todas las causas	73	20	3	2	2

    Si en una estación de bomberos se recibe una llamada, determinar la probabilidad de que sea por:

    (a) un incendio en un hogar.

    (b) un incendio causado por la calefacción, dado que fue en un apartamento.

    (c) un incendio provocado por cada una de las causas (es decir, la probabilidad de un incendio originado por la calefacción, la probabilidad de un incendio provocado al cocinar, ...)

    (d) un incendio en un apartamento y originado por la calefacción.

    (e) un incendio en un hogar, dado que fue provocado por la calefacción.

    (f) en una casa móvil, dado que fue causado por fumar.

29. Una empresa produce resistencias que vende como resistencias de 10 omhnios. Sin embargo, los omhnios reales de los resistores pueden variar. Se ha observado que el 5 % de los valores son menores que 9.5 omhnios y el 10 % mayores que 10.5 omhnios. Si en un determinado sistema se usan dos de esas resistencias, seleccionadas al azar, calcular la probabilidad de que:

    (a) ambas tengan valores reales comprendidos entre 9.5 y 10.5 omhnios.

    (b) al menos una tenga un valor real mayor que 10.5 omhnios.

30. En una clase hay 23 estudiantes. ¿Cuál es la probabilidad de que el cumpleaños de por lo menos dos de ellos sea en la misma fecha (día y mes)? (suponer que el año tiene 365 días)

31. Supóngase que la probabilidad de exposición a la gripe durante una epidemia es 0.6. La experiencia ha demostrado que una vacuna tiene el 80 % de éxito en la prevención de la gripe en una persona que ha sido inoculada con ella, si la persona se expone al virus. Alguien que no esté vacunado se enfrenta a una probabilidad de 0.9 de contagiarse de la enfermedad, si se expone al virus.
    (a) ¿Cuál es la probabilidad de que una persona que ha sido vacunada coja la gripe?
    (b) ¿Cuál es la probabilidad de que se contagie una persona que no ha sido vacunada?

32. En un experimento epidemiológico con ratas se utilizan jaulas con una rata sana en cada jaula. Una segunda rata, infectada con uno de los microorganismos A, B y C se introduce en cada jaula. La probabilidad de que la rata lo esté con el microorganismo A, B o C es de $1/3$ en cada caso. Si se introduce en la jaula una rata con la infección A la probabilidad es de $1/2$ de que la rata sana se contagie. Estas probabilidades son de $1/3$ y $1/4$ si la rata introducida estaba infectada con los microorganismos B o C, respectivamente. La rata sana de la jaula enfermó. ¿Cuál es la probabilidad de que la rata introducida estuviera contagiada con A? ¿Y con B? ¿Y con C?

33. (a) Se considera una caja de 50 naranjas de las que el 10 % están heladas. Se extrae una naranja al azar. Sea A el suceso "la naranja está helada". Se extrae, a continuación, una segunda naranja (sin volver a reponer la primera naranja extraída). Sea B el suceso "la segunda naranja está helada". Calcular $p(A)$, $p(B)$, $p(B/A)$ y $p(B/\bar{A})$. ¿Son los sucesos A y B independientes?
    (b) Repetir los cálculos anteriores suponiendo que se considera un camión cargado con 100000 naranjas de las que el 10 % están heladas. ¿Puede considerarse que los sucesos A y B son independientes?

34. Un hombre va de pesca. En un bote lleva tres carnadas de tipo A, 7 carnadas de tipo B y 10 carnadas de tipo C. La mejor carnada es la A: la probabilidad de pescar un pez con ella es $3/5$. La probabilidad de pescar un pez con las otras carnadas es sólo de $2/7$. Mete la mano en el bote y saca una carnada al azar. Calcular:
    (a) Probabilidad de que pesque un pez.
    (b) Si el hombre tiene éxito en la pesca, ¿cuál es la probabilidad de que utilizara una carnada de tipo B?
    (c) Si saca dos carnadas a la vez, ¿cuál es la probabilidad de que ninguna sea de tipo A?

35. La siguiente tabla da el número de muertes por accidente que se registraron en Estados Unidos durante 1984, para cuatro causas específicas más un apartado para otros tipos de accidente, desglosadas por edades:

	Tráfico	Atropello	Caídas	Ahogamientos	Otras
Menos de 5 años	1132	1024	114	638	744
De 5 a 14	2263	925	68	532	410
De 15 a 24	19738	2546	399	1353	765
De 25 a 44	15036	2396	936	1549	1581
De 45 a 64	6954	3521	1624	763	1411
De 65 a 74	4020	2954	1702	281	467
Más de 75	3114	4381	7067	272	1231

(a) Obtener la distribución de probabilidades marginales según la edad.

(b) Obtener la distribución de probabilidades marginales según el tipo de accidente.

(c) Obtener la tabla de probabilidades condicionales del tipo de accidente según las distintas edades de la población accidentada.

(d) ¿Cuál es la causa más probable de accidente para una víctima de más de 75 años?

(e) ¿Cuál es la causa más probable de accidente para una víctima de edad comprendida entre los 15 y los 24 años?

(f) ¿Cuál es el tipo de accidente menos probable?

(g) ¿Cuál es la probabilidad de que la víctima tenga una edad comprendida entre los 25 y 44 años, si se sabe que el accidente fue de tráfico?

36. El 30 % de los enfermos de hepatitis sufre hepatitis obstructiva que exige una intervención quirúrgica mientras que el otro 70 % tiene hepatitis infecciosa que puede curarse simplemente con el reposo y la medicación. Para discernir entre ambas situaciones cuando se ingresa en un hospital se realiza una determinada prueba clínica, que puede dar positiva o negativa. Se sabe que la probabilidad de que la prueba resulte positiva es de 0.95 cuando los enfermos tienen hepatitis obstructiva y de 0.10 cuando la tienen infecciosa. Supongamos que en un hospital se producen 150 ingresos por hepatitis durante cierto año y que se opera a todos aquellos que dan positivo en la prueba.

    (a) ¿Cuántas intervenciones, aproximadamente, se realizarán en dicho año? ¿Cuántas intervenciones que no deberían realizarse se practican ese año? ¿Cuántos diagnósticos erróneos se realizarán?

    (b) Sabiendo que un enfermo ha dado positivo en la prueba, ¿cuál es la probabilidad de que tenga hepatitis infecciosa?

    (c) Si la prueba da negativa en un paciente, ¿cuál es, pese a ello, la probabilidad de que realmente tenga una hepatitis obstructiva?

37. Un concursante debe elegir entre tres puertas detrás de una de las cuales se encuentra el premio. Hecha la elección y antes de abrir la puerta, el presentador le muestra que en una de las dos puertas no elegidas no está el premio y le da la posibilidad de reconsiderar su elección. ¿Qué debería hacer el concursante?

38. Un banco ha estimado, por experiencias anteriores, que la probabilidad de que una persona falle en los pagos de un préstamo personal es de 0.3. También ha estimado que el 30 % de los préstamos no pagados a tiempo se han hecho para financiar viajes de vacaciones y el 60 % de los préstamos pagados a tiempo también se han hecho para financiar viajes de vacaciones.

    (a) ¿Cuál es la probabilidad de que un préstamo que se haga para financiar un viaje de vacaciones no se pague a tiempo?

    (b) ¿Cuál es la probabilidad de que, si el préstamo se ha hecho para propósitos distintos a viajes de vacaciones, sea pagado a tiempo?

39. En el jardinero del Sr. Rodríguez no se puede confiar: la probabilidad de que se olvide de regar al rosal favorito durante la ausencia del Sr. Rodríguez es de 2/3. El rosal está en un estado inseguro: si se le riega tiene igual probabilidad de progresar que de secarse, pero solamente una probabilidad de 0.25 de progresar si no se le riega. Después de su regreso, el Sr. Rodríguez se encuentra con que su rosal favorito se ha secado. ¿Cuál es la probabilidad de que el jardinero no lo haya regado?

40. Un parque natural está separado en dos zonas A y B por un río. Hay 10 ciervos en la zona A y otros 10 en la zona B. Un biólogo está realizando investigaciones sobre la conducta de un cierto ciervo que vive en la zona A, al que cariñosamente se le conoce como Bambi. Por un descuido de los vigilantes del parque, 9 ciervos de la zona A se pasan a la zona B. Los vigilantes lo advierten, y devuelven 9 ciervos (escogidos al azar) al territorio A. Informado el biólogo de tal contingencia y deseando proseguir sus pesquisas sobre Bambi, ¿en cuál de las dos zonas es preferible que empiece a buscar a su ciervo?

41. En un sistema de emergencia, la probabilidad de que se produzca una situación de peligro es de 0.1. Si ésta se produce, la probabilidad de que el sistema de alarma funcione es de 0.95. La probabilidad de que la alarma funcione sin haber habido situación de peligro es de 0.03. Hallar:

    (a) La probabilidad de que habiendo funcionado la alarma, no haya habido situación de emergencia.

    (b) La probabilidad de que haya una situación de emergencia y la alarma no funcione.

    (c) La probabilidad de que, no habiendo funcionado la alarma, haya una situación de peligro.

42. Las causas por las que puede dejar de funcionar el motor de una automóvil se clasifican en tres categorías, que se suponen independientes, A, B y C. La probabilidad de fallo del motor por la causa A en su primer año de uso es 0.1 y las probabilidades análogas por las causas B y C son 0.2 y 0.3, respectivamente. Hallar la probabilidad de que un motor falle en su primer año de uso.

SOLUCIONES

1. (1) $A \cup B$ (2) $A \cap B$ (3) $A \cap \bar{B} \cap \bar{C}$ (4) $A \cap B \cap \bar{C}$ (5) $A \cap B \cap C$ (6) $A \cup B \cup C$
   (7) $(A \cap B) \cup (A \cap C) \cup (B \cap C)$ (8) $(A \cap \bar{B} \cap \bar{C}) \cup (\bar{A} \cap B \cap \bar{C}) \cup (\bar{A} \cap \bar{B} \cap C)$
   (9) $(A \cap B \cap \bar{C}) \cup (A \cap \bar{B} \cap C) \cup (\bar{A} \cap B \cap C)$ (10) $\overline{A \cup B \cup C}$ (11) $\overline{A \cap B \cap C}$

2. (a) Llamando C=sale cara y X=sale cruz, se tiene que $\Omega = \{e_i\}_{i=1,\ldots 8}$ en donde $e_1 = CCC$, $e_2 = CCX$, $e_3 = CXC$, $e_4 = CXX$, $e_5 = XCC$, $e_6 = XCX$, $e_7 = XXC$ y $e_8 = XXX$.
   (b) (1) $\{e_1, e_8\}$ (2) $\{e_2, e_3, e_4, e_5, e_6, e_7\}$ (3) $\{e_1, e_2, e_3, e_4\}$ (4) $\{e_1, e_2\}$ (5) $\{e_2, e_3, e_5\}$

3. $A \cap B$= {el número elegido es primo y par} = $\{2\}$
   $A - B$= { el número elegido es par pero no es primo} = $\{4, 6, 8, 10\}$
   $A \cup B$= {el número elegido es par o es primo} = $\{1, 2, 3, 4, 5, 6, 7, 8, 10\}$
   $\bar{A} \cap \bar{B}$ = { el número elegido ni es par ni es primo} = $\{9\}$
   $A \cap C$= { el número elegido es múltiplo de 6} = $\{6\}$
   $A - C$= { el número elegido es par no múltiplo de 3} = $\{2, 4, 8, 10\}$
   $B \cap C$= { el número elegido es primo múltiplo de 3} = $\{3\}$
   $\bar{A} \cap C$= { el número elegido es múltiplo impar de 3} = $\{3, 9\}$
   $(A \cup B) \cap \bar{C}$= { el número elegido es par o primo pero no múltiplo de 3} = $\{1, 2, 4, 5, 7, 8, 10\}$

4. $A \cap (B_1 \cup B_2)$

5. $(\cup A_k) \cap (\cup B_j)$

6. Llamando I= {la computadora tiene tarjeta para impresora} y M= {la computadora tiene tarjeta para módem}
   (a) $M \cap I; \#(M \cap I) = 3$ (b) $\overline{M \cup I}; \#\left(\overline{M \cup I}\right) = 13$ (c) $\bar{M} \cap I; \#(\bar{M} \cap I) = 7$
   (d) $(\bar{M} \cap I) + (M \cap \bar{I}); \#[(\bar{M} \cap I) + (M \cap \bar{I})] = 9$

7. $A \cap B$

8. Llamando D =bombilla defectuosa y N =bombilla no defectuosa se tiene que:
   $\Omega = \{D, ND, NND, NNND, NNNND\}$, $p(\{D\}) = 1/3$; $p((ND\}) = 4/15$; $p(\{NND\}) = 1/5$; $p(\{NNND\}) = 2/15$ y $p(\{NNNND\}) = 1/15$

9. (a) El 40 % de los clientes potenciales compra en la tienda 1.
   (b) El 90 % de los clientes potenciales compra en alguna de las dos tiendas.
   (e) El 40 % de los clientes potenciales no compra en la tienda 2.
   (d) El 10 % de los clientes potenciales compra en ambas tiendas.
   (e) El 50 % de los clientes potenciales o compra la tienda 1 o no compra.
   (f) El 10 %de los clientes potenciales no compra en ninguna de las dos tiendas.

10. 60 combinaciones posibles.

11. 450 equipos titulares.

12. $\dfrac{(7)\,(2)\,(6!)}{8!} = 0.25$

13. $\dfrac{\binom{14}{5} 2^9}{3^{14}} = 0.214$

14. $p(\text{suma}= 9) = 0.116$; $p(\text{suma}= 10) = 0.125$

15. $p = 1 - \left(\frac{35}{36}\right)^n$; Habrá que jugar 25 partidas.

16. 0.488

17. $\dfrac{12}{21} = 0.571$

18. (a) $\dfrac{110}{182} = 0.604$ (b) 4 temas.

19. p(exactamente 0) = 0.504; p(exactamente 1) = 0.398; p(exactamente 2) = 0.092
    p(exactamente 3) = 0.006; p(al menos un blanco) = 0.496

20. Siguiendo la costumbre de Fabulandia, la probabilidad de sacar bola blanca es 0.5. Con la iniciativa del reo, la probabilidad de sacar bola blanca es 0.747.

21. 0.143

22. $\dfrac{\binom{30}{8}\binom{70}{12}}{\binom{100}{20}} = 0.116$

23. (a) p(exactamente 0)= $1 + p(A \cap B) - p(A) - p(B)$
    p(exactamente 1)= $p(A) + p(B) - 2p(A \cap B)$
    p(exactamente 2)= $p(A \cap B)$
    (b) p(al menos 0)= 1
    p(al menos 1)= $p(A) + p(B) - p(A \cap B)$
    p(al menos 2)= $p(A \cap B)$
    (c) p(a lo sumo 0)= $1 + p(A \cap B) - p(A) - p(B)$
    p(a lo sumo 1)= $1 - p(A \cap B)$
    pea lo sumo 2)= 1
    (d) p(ocurra A pero no B)= $p(A) - p(A \cap B)$

24. (a) Fiabilidad del sistema en serie: $\prod_{i=1}^{n} p_i$; Fiabilidad del sistema en paralelo: $1 - \prod_{i=1}^{n} (1 - p_i)$
    (b) 0.904 (c) 0.992

25. (a) 1260 formas de asignar los taxis. (b) 1/3

26. Demostrado en la parte de teoría.

27. 0.6976

28. (a) 0.73 (b) 0.06 (d) 0.012 (e) 0.8462 (f) 0.018

Causa del incendio	Porcentaje de incendios originados por dicha causa (%)
Calefacción	19
Al cocinar	16
Materias inflamables	11
Por fumar	10
Instalación eléctrica	8
Otras	36
Todas las causas	100

29. (a) 0.7225 (b) 0.19

30. 0.5073

31. (a) 0.12 (b) 0.54

32. La probabilidad de que la rata introducida estuviera infectada con el organismo A, B o C, dado que la rata sana enfermó, es de 0.462, 0.308 y 0.231, respectivamente.

33. (a) $p(A) = 0.1$, $p(B) = 0.1$, $p(B/A) = 0.816$, $p(B/\bar{A}) = 0.1020$. No.
(b) $p(A) = 0.1$, $p(B) = 0.09990099$, $p(B/A) = 0.09990099$, $p(B/\bar{A}) = 0.100001$. Los sucesos A y B son prácticamente independientes.

34. (a) 0.333 (b) 0.300 (c) 0.716

35. (a), (b) y (c). Los resultados están expresados en porcentajes en la tabla. Por comodidad de lectura de la tabla se han eliminado los decimales. Las filas o las columnas pueden no sumar 100 por errores de redondeo.

	Tráfico	Atropello	Caídas	Ahogamientos	Otras	Todas las causas
Menos de 5 años	31	28	3	17	20	4
De 5 a 14	54	22	2	13	10	4
De 15 a 24	80	10	2	5	3	26
De 25 a 44	70	11	4	7	7	23
De 45 a 64	49	25	11	5	10	15
De 65 a 74	43	31	18	3	5	10
Más de 75	19	27	44	2	8	17
Todas las edades	56	19	13	6	7	100

La parte central de la tabla es la distribución condicionada del tipo de accidente dada la edad de la víctima. La última columna es la distribución marginal para las edades y la última fila la distribución marginal para cada tipo de accidente.

(d) La caída (e) El accidente de tráfico (f) El ahogamiento (g) 0.28

36. (a) Aproximadamente unas 53 intervenciones. Entre 10 y 11 pacientes serán operados sin necesitarlo. Unos 13 diagnósticos erróneos. (b) 0.197 (c) 0.023

37. Cambiar a la otra puerta. Una vez que el presentador ha abierto una puerta, la probabilidad de que el premio esté en la puerta que eligió el concursante es de 1/3 mientras que la probabilidad de que esté en la tercera puerta es de 2/3.

38. (a) 0.176 (b) 0.571

39. 3/4

40. Aunque hay poca diferencia parece mejor comenzar la búsqueda en la zona A: la probabilidad de que Bambi se encuentre en A es de 0.53 y la de encontrarle en la zona B es de 0.47

41. (a) 0.221 (b) 0.005 (c) 0.0057

42. 0.496

Capítulo 4

VARIABLES ALEATORIAS

Definición de variable aleatoria. Dado un experimento aleatorio con espacio muestral Ω y con probabilidad asociada p, una **variable aleatoria** X es una función

$$\begin{aligned} X: \quad & \Omega \longrightarrow \mathbb{R} \\ & \omega \longrightarrow X(\omega) \end{aligned}$$

Al conjunto imagen de esta función se le denomina **Soporte de X** y se representa por $Sop(X) = \mathrm{Im}(X) \subset \mathbb{R}$.

Podemos entonces construir una probabilidad sobre $\mathbb{R}$, a partir de la probabilidad p definida sobre Ω, del siguiente modo:

$$p(A) = p\left(X^{-1}(A)\right) = p\left(\{\omega \in \Omega / X(\omega) \in A\}\right) = p\left(X \in A\right)$$

A esta probabilidad se le denomina **probabilidad inducida por la variable aleatoria** y, como vemos, se representa con la misma notación p utilizada para la probabilidad sobre Ω, aunque ahora se trate de una probabilidad construida sobre $\mathbb{R}$.

Función de distribución. Consideremos entonces una variable aleatoria X definida sobre un espacio muestral Ω y con probabilidad inducida p. Se define la **función de distribución** de la variable aleatoria X como:

$$\begin{array}{rccl} F: & \mathbb{R} & \longrightarrow & [0,1] \\ & x & \longrightarrow & F(x) = p\left((-\infty, x]\right) = p\left(X \leq x\right) \end{array}$$

La función de distribución de una variable aleatoria X verifica las siguientes propiedades:

1. $\lim\limits_{x \to -\infty} F(x) = 0$
2. $\lim\limits_{x \to +\infty} F(x) = 1$
3. Si $x_1 < x_2$ entonces $F(x_1) \leq F(x_2)$ (es decir, es una función creciente).
4. $\lim\limits_{h \to 0^+} F(x+h) = F(x)$ para cualquier $x \in \mathbb{R}$ (es decir, es continua por la derecha).

A partir de esta función se pueden calcular las siguientes probabilidades sobre intervalos de $\mathbb{R}$:

- $p\left((-\infty, x)\right) = p(X < x) = \lim\limits_{h \to 0^-} F(x+h) = F(x^-)$
- $p\left((x, \infty)\right) = p\left(X > x\right) = 1 - p\left(X \leq x\right) = 1 - F(x)$
- $p\left([x, \infty)\right) = p\left(X \geq x\right) = 1 - p(X < x) = 1 - F(x^-)$

- $p\left((a,b]\right) = p\left(a < X \leq b\right) = p\left((-\infty,b]\right) - p\left((-\infty,a]\right) = p\left(X \leq b\right) - p\left(X \leq a\right) = F(b) - F(a)$
- $p\left((a,b)\right) = p\left(a < X < b\right) = p\left((-\infty,b)\right) - p\left((-\infty,a]\right) = p\left(X < b\right) - p\left(X \leq a\right) = F(b^-) - F(a)$
- $p\left([a,b]\right) = p\left(a \leq X \leq b\right) = p\left((-\infty,b]\right) - p\left((-\infty,a)\right) = p\left(X \leq b\right) - p\left(X < a\right) = F(b) - F(a^-)$
- $p\left([a,b)\right) = p\left(a \leq X < b\right) = p\left((-\infty,b)\right) - p\left((-\infty,a)\right) = p\left(X < b\right) - p\left(X < a\right) = F(b^-) - F(a^-)$
- $p\left(\{x\}\right) = p\left(X = x\right) = p\left((-\infty,x]\right) - p\left((-\infty,x)\right) = p\left(X \leq x\right) - p\left(X < x\right) = F(x) - F(x^-)$

Observar que si la función de distribución $F(x)$ es continua en el punto x entonces $F(x) = F(x-)$ y $p\left(X = x\right) = 0$, mientras que si $F(x)$ es discontinua en el punto x entonces $p\left(X = x\right)$ es la magnitud del salto de la función en dicho punto ($F(x) - F(x^-) = F(x^+) - F(x^-)$). Por tanto, cuando la función de distribución es continua entonces se verifica que $p\left(X = x\right) = 0$ para todos los puntos $x \in \mathbb{R}$, es decir, todos los puntos tienen probabilidad cero, tanto si están en el conjunto $Sop(X)$ (y por tanto pueden ocurrir) como si no lo están (y por tanto no pueden ocurrir).

Variables aleatorias discretas. Diremos que una **variable aleatoria** X es **discreta** cuando $Sop(X)$ es un conjunto finito o numerable de $\mathbb{R}$, es decir $Sop(X) = \{x_1, x_2, ..., x_n\}$ o $Sop(X) = \{x_1, x_2, ..., x_n, ...\}$, al cual podemos considerar ordenado, es decir, $x_i < x_{i+1}$ para cualquier i. Para este tipo de variables aleatorias podemos definir la **función de probabilidad** de X como:

$$\begin{array}{rccc} p: & Sop(X) & \longrightarrow & [0,1] \\ & x_i & \longrightarrow & p(x_i) = p(X = x_i) = p_i \end{array}$$

Observar que esta función necesariamente verifica que $\sum_i p_i = 1$ y además, a partir de ella, podemos construir la función de distribución del siguiente modo:

$$F(x) = p\left(X \leq x\right) = \begin{cases} 0 & \text{si } x < x_1 \\ p_1 & \text{si } x_1 \leq x < x_2 \\ p_1 + p_2 & \text{si } x_2 \leq x < x_3 \\ \cdot & \\ \cdot & \\ \cdot & \\ p_1 + p_2 + ... + p_{n-1} & \text{si } x_{n-1} \leq x < x_n \\ \sum_i p_i = 1 & \text{si } x \geq x_n \end{cases}$$

Por tanto la función de distribución de una variable aleatoria discreta es una función escalonada, continua por la derecha, con discontinuidades de salto finito en los puntos de $Sop(X)$ y las magnitudes de los saltos coinciden con las probabilidades puntuales de dichos puntos.

Si X es una variable aleatoria discreta se define la **esperanza de** X como:

$$EX = \sum_i x_i p\left(X = x_i\right) = \sum_i x_i p_i = \mu$$

Teniendo en cuenta que $p_i \in [0,1]$ y $\sum_i p_i = 1$ podemos asegurar que EX es un número real comprendido entre el valor mínimo y el valor máximo de los elementos de $Sop(X)$ (es decir, $x_1 \leq EX \leq x_n$), pero no necesariamente EX pertenece a $Sop(X)$. A veces a este valor también se le denomina media teórica de la variable aleatoria X, por su similitud en la definición a la media muestral en Estadística Descriptiva. Podemos decir entonces que EX es el concepto teórico que generaliza la idea de media muestral y que representa un valor de referencia alrededor del cual podemos decir que oscila la variable aleatoria.

Además, si X es una variable aleatoria discreta se define la **mediana de** X como:

$$Me = \text{mín}\left\{x_i \in Sop(X) / F(x_i) \geq 0.5\right\}$$

de modo que ahora sí podemos asegurar que $Me \in Sop(X)$ y también puede ser considerado como un valor de referencia alrededor del cual oscila la variable aleatoria.

Para cuantificar la dispersión de una variable aleatoria discreta se define la **varianza de** X como:

$$\begin{aligned} Var(X) &= \sum_i (x_i - EX)^2 p\left(X = x_i\right) = \sum_i (x_i - EX)^2 p_i \\ &= \sum_i x_i^2 p_i - (EX)^2 = E\left(X^2\right) - (EX)^2 = \sigma^2 \end{aligned}$$

Esta medida de dispersión no viene dada en las mismas unidades que la variable aleatoria X, sino en unidades cuadráticas. Es por ello que resulta más adecuado construir una nueva medida dispersión denominada **desviación típica de** X, definida como $\sigma = \sqrt{Var(X)}$ y que sí tiene las mismas unidades que la variable aleatoria X, o que EX y Me.

Para comparar la dispersión de variables aleatorias con diferente esperanza es interesante definir una medida de dispersión relativa que compare el valor de la desviación típica con el valor de la esperanza, siempre que ésta sea distinta de cero. En estos casos, se define el **coeficiente de variación de** X, en términos porcentuales, como:

$$CV = 100 \frac{\sigma}{|EX|}$$

(observar que no necesariamente $CV \in [0, 100]$, ya que puede ocurrir que $\sigma > |EX|$).

Ejercicio. Consideramos el experimento aleatorio consistente en lanzar una moneda equilibrada tres veces, donde el espacio muestral Ω tiene ocho elementos todos ellos igualmente probables (es decir, la probabilidad p sobre Ω queda definida por la Regla de Laplace). Sea entonces la variable aleatoria X =número de caras obtenidas en los tres lanzamientos. Se pide calcular $Sop(X)$, la función de probabilidad, la función de distribución y sus medidas de centralización y dispersión: esperanza, mediana, varianza, desviación típica y coeficiente de variación. ¿Cuál es la probabilidad de obtener más de una cara?

Variables aleatorias continuas. Diremos que una **variable aleatoria** X es **continua** cuando $Sop(X)$ es un conjunto infinito y no numerable de $\mathbb{R}$, habitualmente un intervalo cerrado finito o infinito de números reales de la forma $[a, b]$, $[a, \infty)$ o incluso $(-\infty, \infty) = \mathbb{R}$. Para este tipo de variables aleatorias no es posible construir la función de probabilidad que utilizamos en el caso discreto, puesto que no podemos enumerar todos los elementos de $Sop(X)$ con su respectiva probabilidad. Para poder calcular probabilidades con este tipo de variables aleatorias vamos a utilizar una función matemática que identifique la probabilidad de un intervalo de números reales con el área encerrada bajo la curva en dicho intervalo (integral de Riemann). Con esta idea, dada una variable aleatoria continua X, se define la **función de densidad de** X como una función matemática

$$\begin{array}{rccc} f: & \mathbb{R} & \longrightarrow & \mathbb{R} \\ & x & \longrightarrow & f(x) \end{array}$$

que verifica las siguientes propiedades:

1. $f(x) \geq 0$ para todo $x \in \mathbb{R}$.
2. $\int_{\mathbb{R}} f(x)\, dx = 1$.

3. Para cualquier intervalo finito o infinito I de $\mathbb{R}$ se verifica que $p(I) = \int_I f(x)\ dx$.

Observar que $f(x)$ no representa $p(X = x)$, como lo hacía la función de probabilidad de una variable discreta, ya que, como consecuencia de esta definición, podemos asegurar que si X es una variable aleatoria continua y x_0 es un punto cualquiera de $Sop(X)$ entonces necesariamente se verifica que

$$p(X = x_0) = p\left([x_0, x_0]\right) = \int_{x_0}^{x_0} f(x)\ dx = 0$$

Por tanto, en una variable aleatoria continua todos los puntos tienen probabilidad cero. Además, la función de distribución de X vendrá definida como

$$F(x) = p\left(X \leq x\right) = \int_{-\infty}^{x} f(t)\ dt$$

y, utilizando el Teorema Fundamental del Cálculo Integral, podemos asegurar que $F(x)$ es derivable (por tanto también continua) y además $F'(x) = f(x)$. Es decir, conocida la función de distribución podemos calcular la función de densidad sin más que derivar; y conocida la función de densidad podemos calcular la función de distribución buscando una primitiva de la función densidad y eligiendo la constante C de integración de modo que, si $Sop(X) = [a, b]$, se verifique $F(a) = 0$ o, alternativamente, $F(b) = 1$ (si a o b son infinitos cambiaríamos las imágenes por los respectivos límites). Además, para este tipo de variables aleatorias, por las propiedades de la integral de Riemann, es evidente que, para cualesquiera valores a y b de $\mathbb{R}$ con $a < b$, se verifica:

$$p\left(a \leq X \leq b\right) = p\left(a \leq X < b\right) = p\left(a < X \leq b\right) = p\left(a < X < b\right) = \int_a^b f(x)\ dx$$

En este caso las definiciones de esperanza y varianza de la variable aleatoria son las siguientes:

$$EX = \int_{-\infty}^{\infty} x\ f(x)\ dx = \mu$$

$$Var(X) = \int_{-\infty}^{\infty} (x - EX)^2\ \ f(x)\ dx = \int_{-\infty}^{\infty} x^2\ f(x)\ dx - (EX)^2 = E\left(X^2\right) - (EX)^2 = \sigma^2$$

Las definiciones de mediana, desviación típica y coeficiente de variación son las mismas que en el caso de las variables aleatorias discretas.

Regla de Tchebyscheff. Una propiedad importante que verifican todas las variables aleatorias (tanto discretas como continuas) y que nos ayuda a interpretar mejor los valores de la esperanza μ y la desviación típica σ es la conocida **Regla de Tchebyscheff**, la cual asegura que, para cualquier variable aleatoria X y para cualquier número real $k \in \mathbb{R}$, se verifica que

$$p\left(|X - \mu| \leq k\sigma\right) = p\left(\mu - k\sigma \leq X \leq \mu + k\sigma\right) \geq 1 - \frac{1}{k^2}$$

Por ejemplo, para $k = 2$, podemos asegurar que $p\left(\mu - 2\sigma \leq X \leq \mu + 2\sigma\right) \geq \frac{3}{4} = 0.75$ y, para $k = 3$, $p\left(\mu - 3\sigma \leq X \leq \mu + 3\sigma\right) \geq \frac{8}{9} = 0.89$.

Ejercicio. Supongamos que el pH, con el que se mide la acidez del agua de un lago, es una variable aleatoria continua X con función de densidad

$$f(x) = \begin{cases} \frac{3}{8}\left(7 - x\right)^2 & \text{si } 5 \leq x \leq 7 \\ 0 & \text{en el resto} \end{cases}$$

Se pide: demostrar que $f(x)$ es una función de densidad, calcular la función de distribución de X, calcular la probabilidad de que el pH sea menor 6, calcular EX, Me, $Var(X)$, σ y CV. Calcular también los

cuartiles, el rango intercuartílico y dar un intervalo, menor que $[5, 7]$, en el que estén, por lo menos, el 75 % de las posibles mediciones del pH del agua en el lago. ¿Es de esperar con mucha frecuencia valores del pH menores que 5.5?.

Definición de vector aleatorio. En muchas situaciones prácticas resulta necesario calcular probabilidades que involucran, simultáneamente, a varias variables aleatorias. Por eso es necesario también construir modelos de probabilidad para vectores aleatorios con valores en $\mathbb{R}^n$ en vez de $\mathbb{R}$. Nosotros vamos a ilustrar la teoría con el caso de $n = 2$, es decir, vectores aleatorios bidimensionales. Lo haremos primero en el caso de que las dos variables aleatorias que componen el vector sean discretas.

Vectores aleatorios discretos. Se define un **vector aleatorio bidimensional discreto** como una función

$$\begin{aligned}(X, Y): \quad & \Omega \longrightarrow \mathbb{R}^2 \\ & \omega \longrightarrow (X(\omega), Y(\omega))\end{aligned}$$

donde $X(\omega) \in Sop(X)$, $Y(\omega) \in Sop(Y)$ y suponemos que ambos conjuntos son finitos o infinitos pero numerables, de modo que $Sop(X) = \{x_1, x_2, ..., x_n\}$ y $Sop(Y) = \{y_1, y_2, ..., y_m\}$ (nosotros lo ilustraremos para el caso finito, pero igual se haría para el caso infinito pero numerable).

Al igual que en el caso de las variables aleatorias unidimensionales, la probabilidad p definida sobre Ω induce una nueva probabilidad (denotada también por p) sobre el conjunto $\mathbb{R}^2$. Para trabajar con esta probabilidad definimos la **función de probabilidad conjunta del vector** (X, Y) como

$$\begin{aligned}p: \quad & Sop(X) \times Sop(Y) \longrightarrow [0, 1] \\ & (x_i, y_j) \longrightarrow p(x_i, y_j) = p\left((X = x_i) \cap (Y = y_j)\right) = p_{ij}\end{aligned}$$

Observar que, por las propiedades de toda probabilidad, necesariamente se verifica que $p_{ij} \geq 0$ y $\sum_i \sum_j p_{ij} = 1$. Además, a partir de esta función de probabilidad conjunta podemos construir las funciones de probabilidad individuales de cada una de las variables aleatorias, a las que ahora se denomina **funciones de probabilidad marginales**, del siguiente modo:

$$p_X(x_i) = p(X = x_i) = \sum_j p\left((X = x_i) \cap (Y = y_j)\right) = \sum_j p_{ij} = p_{i\cdot}$$

$$p_Y(y_j) = p(Y = y_j) = \sum_i p\left((X = x_i) \cap (Y = y_j)\right) = \sum_i p_{ij} = p_{\cdot j}$$

Evidentemente, estas dos funciones verifican que $\sum_i p_X(x_i) = \sum_i p_{i\cdot} = 1$ y $\sum_j p_Y(y_j) = \sum_j p_{\cdot j} = 1$.

En el caso finito habitualmente se representa esta función mediante una tabla de doble entrada del modo siguiente:

$X \diagdown Y$	y_1	y_2	.	.	.	y_m	
x_1	p_{11}	p_{12}	.	.	.	p_{1m}	$p_{1\cdot}$
x_2	p_{21}	p_{22}	.	.	.	p_{2m}	$p_{2\cdot}$
.	.	.	.			.	.
.	.	.		.		.	.
.	.	.			.	.	.
x_n	p_{n1}	p_{n2}	.	.	.	p_{nm}	$p_{n\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	.	.	.	$p_{\cdot m}$	1

Además, si $p(Y = y_j) > 0$, se define la **función de probabilidad de X condicionada por $Y = y_j$** como

$$p_{X/Y=y_j}(x_i) = \frac{p\left((X = x_i) \cap (Y = y_j)\right)}{p(Y = y_j)} = \frac{p_{ij}}{p_{\cdot j}}$$

Del mismo modo, si $p(X = x_i) > 0$, se define la **función de probabilidad de Y condicionada por $X = x_i$** como

$$p_{Y/X=x_i}(y_j) = \frac{p\left((X = x_i) \cap (Y = y_j)\right)}{p(X = x_i)} = \frac{p_{ij}}{p_{i\cdot}}$$

Por tanto, si todos los puntos tiene probabilidad estrictamente positiva, tendremos un total de $n+m$ distribuciones de probabilidad condicionadas que podrán ser utilizadas para calcular probabilidades de una de las variables aleatorias sabiendo cuál es el valor obtenido para la otra.

Teniendo en cuenta la definición de independencia de sucesos en espacios de probabilidad, parece lógico considerar que **las variables X e Y son independientes** cuando se verifica que

$$p(x_i, y_j) = p\left((X = x_i) \cap (Y = y_j)\right) = p\left(X = x_i\right)\ p\left(Y = y_j\right) = p_X(x_i)\ p_Y(y_j)$$

Por tanto, cuando X e Y sean independientes necesariamente, en la tabla de doble entrada, se tiene que verificar que cada elemento p_{ij} debe coincidir con el producto $p_{i\cdot}\ p_{\cdot j}$.

Además, resulta interesante definir el **valor esperado del producto de las variables aleatorias X e Y** del siguiente modo

$$E(XY) = \sum_i \sum_j x_i\ y_j p(x_i, y_j)$$

y, a partir de él, construir una medida de relación entre las dos variables aleatorias, denominada **Covarianza de X e Y**, y definida como

$$Cov(X, Y) = E(XY) - (EX)\ (EY)$$

Teniendo en cuenta esta definición es fácil demostrar que, si X e Y son variables aleatorias independientes, necesariamente se verifica que $Cov(X, Y) = 0$, ya que se tiene:

$$\begin{aligned} E(XY) &= \sum_i \sum_j x_i\ y_j p(x_i, y_j) = E(XY) = \sum_i \sum_j x_i\ y_j p_X(x_i)\ p_Y(y_j) \\ &= \sum_i x_i\ p_X(x_i) \sum_j y_j\ p_Y(y_j) = (EY) \sum_i x_i\ p_X(x_i) = (EX)\ (EY) \end{aligned}$$

Sin embargo, es importante observar que puede ocurrir que $Cov(X, Y) = 0$ y sin embargo no siempre ocurra que $p_{ij} = p_{i\cdot}\ p_{\cdot j}$, y por tanto las variables X e Y no sean independientes. Es decir, el hecho de que la covarianza sea 0 no asegura la independencia de las variables, aunque sí podemos decir que, si $Cov(X, Y) \neq 0$ entonces es seguro que las variables no son independientes.

Ejercicio. Lanzamos 3 veces una moneda equilibrada y consideramos el vector aleatorio (X, Y) definido como X = número de caras e Y = diferencia, en valor absoluto, entre el número de caras y el número de cruces. Se pide: calcular la función de probabilidad conjunta, las funciones de probabilidad marginales, las funciones de probabilidad condicionadas, las esperanzas y varianzas de ambas variables aleatorias y también $Cov(X, Y)$. ¿Son X e Y independientes?.

Vestores aleatorios continuos. Si las variables aleatorias X e Y son continuas entonces diremos que (X, Y) es un **vector aleatorio bidimensional continuo**. En este caso, al no poder enumerar los

elementos del soporte de ambas variables aleatorias, no podemos construir la función de probabilidad conjunta. Para calcular probabilidades usaremos de nuevo una herramienta matemática que se conoce como **función de densidad conjunta del vector** (X, Y) y se define como una función matemática

$$\begin{array}{rccc} f: & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (x,y) & \longrightarrow & f(x,y) \end{array}$$

que verifica las siguientes propiedades:

1. $f(x,y) \geq 0$ para todo $(x,y) \in \mathbb{R}^2$.
2. $\int_{\mathbb{R}} \int_{\mathbb{R}} f(x,y)\ dx\ dy = 1$.
3. Para cualesquiera intervalos finitos o infinitos I_1 e I_2 de $\mathbb{R}$ se verifica que $p\left((X,Y) \in I_1 \times I_2\right) = \int_{I_1} \int_{I_2} f(x,y)\ dx\ dy$.

Del mismo modo que en el caso discreto, a partir de esta función podemos calcular las funciones de densidad individuales de cada una de las variables aleatorias, denominadas ahora **funciones de densidad marginales de** X **e** Y, del siguiente modo

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)\ dy \qquad\qquad f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)\ dx$$

Si $f_Y(y) > 0$ se define también la **función de densidad de la variable** X **condicionado por el valor** y de la variable Y como

$$f_{X/Y=y}(x) = \frac{f(x,y)}{f_Y(y)}$$

y se utiliza para calcular probabilidades sobre la variable X sabiendo qué valor ha tomado la variable Y.

Del mismo modo, si $f_X(x) > 0$,se define la **función de densidad de la variable** Y **condicionado por el valor** x de la variable X como

$$f_{Y/X=x}(y) = \frac{f(x,y)}{f_X(x)}$$

y se utiliza para calcular probabilidades sobre la variable Y sabiendo qué valor ha tomado la variable X.

De forma análoga al caso de las variables aleatorias discretas, diremos que las **variables** X **e** Y **son independientes** cuando se verique que

$$f(x,y) = f_X(x)\ f_Y(y)$$

y definiremos el valor esperado del producto XY como

$$E(XY) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x\ y\ f(x,y)\ dx\ dy$$

y la Covarianza de X e Y como

$$Cov(X,Y) = E(XY) - (EX)\ (EY)$$

Al igual que en el caso de las variables discretas, esta covarianza es una medida de la relación entre las dos variables que verifica además las mismas propiedades ya citadas en el caso discreto.

Propiedades generales. Enunciamos a continuación una lista de propiedades para esperanza, varianza y covarianza que resulta imprescindible conocer para trabajar en el resto de la asignatura.

1. $E(kX) = k\ EX$
2. $E\left(X \pm Y\right) = EX \pm EY$
3. $Var(kX) = k^2\ Var(X)$ y por tanto $\sigma_{kX} = |k|\ \sigma_X$
4. $Var(X \pm k) = Var(X)$
5. $Cov(X,k) = 0$

6. $Cov(X,Y) = E\left[(X - EX)\,(Y - EY)\right]$
7. $Cov\left(k_1X, k_2Y\right) = k_1k_2\ Cov(X,Y)$
8. $Cov\left(X \pm k_1, Y \pm k_2\right) = Cov(X,Y)$
9. $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X,Y)$
10. Si X e Y son variables aleatorias independientes se verifica que $Var(X \pm Y) = Var(X) + Var(Y)$
11. En general, si X_1, X_2, . . . , X_n son variables aleatorias independientes dos a dos, se verifica que $Var\left(\sum_i X_i\right) = \sum_i Var(X_i)$

Conocidas estas propiedades podemos interpretar aún mejor la covarianza entre dos variables aleatorias como medida de relación entre ambas. En efecto, como el produto $(X - EX)\ \ (Y - EY)$ es positivo o bien cuando $X \geq EX$ e $Y \geq EY$ o bien cuando $X \leq EX$ e $Y \leq EY$ (es decir, valores altos de X van acompañados de valores altos de Y y valores bajos de X van acompañados de valores bajos de Y) entonces, por la propiedad 6, podemos asegurar que cuando $Cov(X,Y) > 0$ hay una relación positiva o creciente porque el valor esperado de esas diferencias es positivo. En cambio, cuando $Cov(X,Y) < 0$ habrá una mayor tendencia a que valores bajos de X vayan acompañados de valores altos de Y y valores altos de X vayan acompañados de valores bajos de Y (es decir, productos $(X - EX)\ \ (Y - EY)$ negativos). Diremos entonces que hay una relación negativa o decreciente. El problema es que la mayor o menor magnitud positiva o negativa de $Cov(X,Y)$ no es indicativo de una mayor o menor relación positiva o negativa entre ambas variables, porque, teniendo en cuenta la propiedad 7, la covarianza cambia al variar las unidades de medida de las variables aleatorias (por ejemplo, si multiplicamos ambas variables por 10, teniendo en cuenta la propiedad 7, la covarianza se multiplica por 100, sin que eso signifique que la relación entre las variables sea mayor). Este es el motivo por el que se considera una nueva medida de relación entre dos variables que sea adimensional e independiente de las unidades en que vengan dadas.

Coeficiente de correlación. Para cualesquiera variables aleatorias, se define el **coeficiente de correlación de X e Y**, como

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{Cov(X,Y)}{\sigma_X\ \sigma_Y}$$

Este coeficiente verifica las siguientes propiedades:

1. $\rho(X \pm k_1, Y \pm k_2) = \rho(X,Y)$
2. $\rho\left(k_1X, k_2Y\right) = \rho(X,Y)$
3. $signo\left[\rho(X,Y)\right] = signo\left[Cov(X,Y)\right]$
4. $\rho(X,Y) \in [-1,1]$
5. Si X e Y son independientes entonces $\rho(X,Y) = 0$ (pero el recíproco no es cierto).
6. Si $\rho(X,Y) = 1$ entonces $Y = a + bX$ con $b > 0$, es decir, hay una relación lineal perfecta y positiva entre ambas variables.
7. Si $\rho(X,Y) = -1$ entonces $Y = a + bX$ con $b < 0$, es decir, hay una relación lineal perfecta y negativa entre ambas variables.

Es importante señalar que el coeficiente de correlación es sólo una medida de la relación lineal entre dos variables aleatorias, pero pueden existir otro tipo de relación entre las dos variables que no queda reflejada en dicho coeficiente. Por ejemplo, puede ocurrir que el coeficiente de correlación sea cero (o esté próximo a cero) y sin embargo exista una fuerte relación no lineal entre ambas variables (por ejemplo, cuadrática o logarítmica, etc.).

Ejercicio. Supongamos que (X, Y) es un vector aleatorio continuo con función de densidad conjunta:

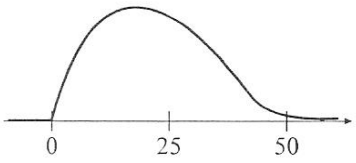
$$f(x,y) = \begin{cases} k & \text{si } 0 < y < x < 1 \\ 0 & \text{en el resto} \end{cases}$$

Se pide: calcular el valor de k, las funciones de densidad marginales y condicionadas, las esperanzas y varianzas y $Cov(X, Y)$. ¿Son X e Y independientes? Calcular $p\left((X > 0.5) \cap (Y < 0.5)\right)$.

4.1. PROBLEMAS

1. Una compañía de refrescos anuncia premio en las chapas asegurando que de cada 1000 chapas hay 400 con "inténtalo de nuevo", 300 con premio de 0.5 euros, 250 con premio de 1 euro, 40 con premio de 5 euros y 10 con premio de 10 euros. Un individuo, al que no le gusta el refresco. decide comprar una botella cuyo coste es de 1 euro. Caracterizar su ganancia mediante una variable aleatoria. ¿Cuál es la probabilidad de que el individuo no pierda dinero con la compra?

2. Para determinadas bacterias se ha estudiado la variable aleatoria $X =$*tiempo de vida de una bacteria* (horas) resultando ser la función de densidad de dicha variable la que se muestra en la figura.



   Sombrear la región bajo la curva de densidad que corresponde a cada una de las siguientes probabilidades:

   (a) Probabilidad de que una bacteria viva menos de 30 horas.

   (b) Probabilidad de que una bacteria viva entre 10 y 25 horas

3. Sea X una variable aleatoria con función de densidad:

$$f(x) = \begin{cases} k\left(1 + x^2\right) & \text{si} \quad x \in (0,3) \\ 0 & \text{si} \quad x \notin (0,3) \end{cases}$$

   Se pide:

   (a) Hallar el valor de la constante k para que $f(x)$ sea una función de densidad.

   (b) Obtener la función de distribución de la variable aleatoria X.

   (c) Calcular la probabilidad de que X esté comprendida entre 1 y 2.

(d) Obtener la mediana de la distribución.

(e) Sabiendo que X es mayor que 1, obtener la probabilidad de que sea menor que 2.

4. La eficacia de las calefacciones que funcionan mediante energía solar depende de la cantidad de radiación de sol. Para un mes de octubre típico, la radiación total diaria en Miami, Florida, es una variable aleatoria cuya función de densidad viene dada por (las unidades son cientos de kilocalorías)

$$f(x) = \begin{cases} \frac{3}{32}(x-2)(6-x) & \text{si } 2 \leq x \leq 6 \\ 0 & \text{en el resto} \end{cases}$$

(a) Calcular la probabilidad de que la radiación solar sea mayor que 300 kilocalorías en un día normal de octubre.

(b) Según este modelo, ¿qué cantidad de radiación solar queda rebasada exactamente el 50 % de los días de octubre?

5. El pH, con el que se mide la acidez del agua, es importante en los estudios sobre lluvia ácida. Para determinado lago se llevan a cabo mediciones testigo de la acidez para que se pueda notar cualquier cambio en la acidez del agua originado por la lluvia ácida. El pH del agua del lago es un variable aleatoria X que, según se ha podido comprobar a partir de las muestras tomadas, tiene una función de densidad:

$$f(x) = \begin{cases} \frac{3}{8}(7-x)^2 & \text{si } 5 \leq x \leq 7 \\ 0 & \text{en el resto} \end{cases}$$

(a) Demostrar que $f(x)$ es una función de densidad.

(b) Dibujar la curva $f(x)$.

(c) Calcular la función de distribución de la variable aleatoria X.

(d) Calcular la probabilidad de que el pH sea menor que 6 en una muestra del agua de este lago.

(e) Calcular $E(X)$ y $Var(X)$.

(f) Calcular la mediana. los cuartiles y el rango intercuartílico de la variable X.

(g) Calcular un intervalo menor que $[5, 7]$ en el que estén por lo menos el 75 % de las mediciones.

(h) ¿Es de esperar con mucha frecuencia valores del pH menores que 5.5?

6. La acidez X de cierto compuesto depende de la proporción Y de uno de sus componentes químicos y viene dada por la relación $X = (1+Y)^2$. La proporción Y es una variable aleatoria con función de densidad

$$f(y) = \begin{cases} 2y & \text{si } 0 \leq y \leq 1 \\ 0 & \text{en el resto} \end{cases}$$

Calcular la función de distribución y la función de densidad de la variable aleatoria X. Obtener la esperanza de X de dos formas: (i) utilizando la función de densidad de X, y (ii) utilizando la función de densidad de Y.

7. Al examinar los pozos de agua en un zona con respecto a tres impurezas encontradas frecuentemente en el agua potable, resultó que el 20 % de los pozos no revelaba impureza alguna, el 40 % tenía la impureza A, en el 60 % se encontró la impureza B yen el 30 % la impureza C. Además, de los pozos en los que se encontró la impureza A, se vio que el 60 % también contenía la impureza B y que un 30 % contenía la impureza C. Del mismo modo, en un 40 % de los pozos que tenían impureza B también se encontró la impureza C. Encontrar la función de probabilidad y la función de distribución de la variable aleatoria "número de tipos distintos de impurezas encontradas en un pozo de agua".

8. En una población, la cantidad de plomo que tiene una persona en la sangre, medida en ppm, es una variable aleatoria con función de densidad

$$f(x) = \begin{cases} \frac{x}{500} & \text{si } 0 \leq x < 20 \\ \frac{50-x}{750} & \text{si } 20 \leq x < 50 \\ 0 & \text{en el resto} \end{cases}$$

   (a) Representar $f(x)$ gráficamente.

   (b) Demostrar que $f(x)$ es una función de densidad.

   (c) Calcular la cantidad media de plomo en la sangre de los individuos de la población.

   (d) Elegimos una persona al azar. Obtener la probabilidad de que la cantidad de plomo en su sangre sea inferior a 15 ppm.

   (e) Obtener la probabilidad de que en 40 personas elegidas al azar, haya entre 5 y 12 personas con una cantidad de plomo en la sangre inferior a 15 ppm.

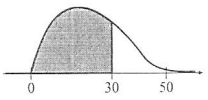
9. Para una determinada sección de un bosque de pinos, el número de árboles enfermos por hectárea se puede suponer que tiene una distribución cuyo promedio es de 10 árboles enfermos por hectárea. Los árboles con plaga se fumigan con un insecticida a un coste de 3$ por árbol, además de un coste fijo de administración por renta de equipo igual a 50$. Calcular el valor esperado y la desviación estándar del costo total C de la desinfección de una hectárea de bosque. ¿Dentro de qué intervalo se puede esperar que quede C con una probabilidad de por lo menos 0.75?

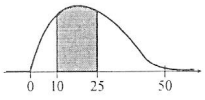
10. Sean X_1 y X_2 variables aleatorias que representen el alto y el ancho, respectivamente, de una pieza manufacturada. El alto es una variable aleatoria con media 2 cm y desviación típica de 0.1 cm, mientras que el ancho tiene media de 5 cm con varianza de 0.04 cm^2. Se supone que X_1 y X_2 son variables aleatorias independientes. Calcular la media y la desviación estándar del perímetro de la pieza.

SOLUCIONES

1. La probabilidad de que no pierda dinero es 0.3

x_i	-100	-50	0	400	900
$p\,(X = x_i)$	0.4	0.3	0.25	0.04	0.01

2. (a) 

(b) 

3. (a) $k = 1/12$ (b) $F(x) = \begin{cases} 0 & \text{si} \quad x \leq 0 \\ \frac{x}{36}\left(3 + x^2\right) & \text{si} \quad x \in (0,3) \\ 1 & \text{si} \quad x \geq 3 \end{cases}$

(c) 0.278 (d) $Me \cong 2.2422$ (e) 0.3125

4. (a) 0.8438 (b) 400 kilocalorías

5. (a) $f(x) \geq 0$ para todo $x \in \mathbb{R}$ y ademas $\int_{-\infty}^{\infty} f(x)dx = 1$

(c) $F(x) = \begin{cases} 0 & \text{si} \quad x \leq 5 \\ 1 - \frac{(7-x)^3}{8} & \text{si} \quad x \in (5,7) \\ 1 & \text{si} \quad x \geq 7 \end{cases}$

(d) 0.875 (e) $E(X) = 5.5$, $E\left(X^2\right) = 30.4$ y $Var\left(X\right) = 0.15$

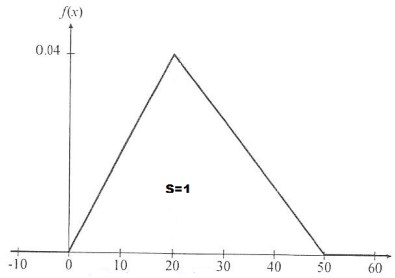
(f) $Me = 7 - \sqrt[3]{4}$, $Q_1 = 7 - \sqrt[3]{6}$, $Q_3 = 7 - \sqrt[3]{2}$ y $RI = \sqrt[3]{6} - \sqrt[3]{2}$

(g) La probabilidad de que el pH sea menor que 5.5 es 0.5781. Por lo tanto, es bastante probable que X tome valores menores que 5.5.

6. $F(x) = \begin{cases} 0 & \text{si} \quad x < 1 \\ \left(\sqrt{x} - 1\right)^2 & \text{si} \quad x \in [1,4] \\ 1 & \text{si} \quad x > 4 \end{cases}$ $f(x) = \begin{cases} 1 - \frac{1}{\sqrt{x}} & \text{si } x \in [1,4] \\ 1 & \text{en el resto} \end{cases}$ $E(X) = 17/6$

7.

x_i	0	1	2	3
$p\left(X = x_i\right)$	0.2	0.4	0.3	0.1

8. (a) y (b) 

(c) $E(X) = 70/3$ (d) 0.225 (e) 0.8636

9. $E(C) = 80\$$ y $\sqrt{Var(C)} = 9.5\$$. El coste queda en el intervalo $(61.03, 98.97)$ con una probabilidad de al menos 0.75

10. $E\left(P\right) = 14$ cm y $\sqrt{Var(P)} = 0.4472$

Capítulo 5

DISTRIBUCIONES DE PROBABILIDAD DISCRETAS

Hasta ahora hemos estado refiriéndonos a modelos de probabilidad en general, sin hacer referencia a ninguno en particular. Sin embargo, algunas distribuciones de probabilidad específicas juegan un papel muy importante tanto en la Probabilidad como en la Estadística. En este capítulo estudiaremos los modelos de probabilidad de tipo discreto más importantes y dejaremos los de tipo continuo para el siguiente capítulo.

5.1. Distribución de Bernouilli.

Consideramos un experimento aleatorio cuyos posibles resultados son agrupados en dos conjuntos excluyentes que llamaremos éxito $\{E\}$ y fracaso $\{F\}$. Esta división en éxito y fracaso puede ser algo que viene impuesto de manera natural o una división artificial que el experimentador está interesado en realizar. A un experimento de este tipo se le denomina **experimento de Bernouilli**.

Dado un experimento de este tipo, consideramos una variable aleatoria X definida como

$$X = \begin{cases} 1 & \text{si obtenemos éxito} \\ 0 & \text{si obtenemos fracaso} \end{cases}$$

Diremos entonces que la variable X tiene una distribución de Bernouilli de parámetro $p \in (0,1)$, siendo $p = p(X = 1)$ y, por tanto, $1 - p = p(X = 0)$. Lo denotaremos por $X \rightsquigarrow B(p)$. Observar entonces que la función de probabilidad puede escribirse como $p(X = x) = p^x \ (1-p)^{1-x}$ con $x \in \{0, 1\}$. Para esta variable aleatoria es fácil demostrar que $EX = p$, $Var(X) = p(1-p)$, $\sigma = \sqrt{p(1-p)}$, $CV = 100\sqrt{\frac{1-p}{p}}\,\%$ y $Me = \begin{cases} 1 & \text{si } p < 0.5 \\ 0 & \text{si } p \geq 0.5 \end{cases}$.

Aunque en situaciones prácticas pocas veces nos encontramos ante una situación tan simple, esta variable aleatoria es muy importante porque, a partir de ella, se generan diferentes modelos de probabilidad, algunos de ellos muy importantes. Concretamente las tres próximas distribuciones que vamos a estudiar se obtendrán a partir de repeticiones independientes de experimentos de Bernouilli.

5.2. Distribución Binomial.

Supongamos que se repite un experimento de Bernouilli n veces de forma independiente (es decir, el resultado de cada ensayo es independiente del resultado obtenido en los ensayos anteriores) y denotemos por p a la probabilidad de éxito en cada ensayo. Consideramos entonces la variable aleatoria definida como: X =número de éxitos obtenidos en los n ensayos. Diremos entonces que la variable aleatoria X tiene una **distribución binomial de parámetros** n **y** p, y lo denotaremos por $X \rightsquigarrow B(n,p)$. Es fácil demostrar que para esta variable aleatoria se tiene $Sop(X) = \{0, 1, ..., n\}$ y que su función de probabilidad es

$$p(X = x) = \binom{n}{x}\ p^x\ (1-p)^{n-x}$$

(Observar que $\sum_{x=0}^{n} p(X = x) = \sum_{x=0}^{n} \binom{n}{x}\ p^x\ (1-p)^{n-x} = [p + (1-p)]^n = 1$).

Si consideramos las variables de Bernouilli X_i definidas para cada ensayo (1 si obtenemos éxito y 0 si obtenemos fracaso) es evidente que $X = \sum_{i=1}^{n} X_i$ y entonces, por ser las variables X_i independientes, se tendrá que $E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = np$ y también $Var(X) = \sum_{i=1}^{n} Var(X_i) = np(1-p)$. Por tanto, $\sigma = \sqrt{np(1-p)}$ y $CV = 100\sqrt{\frac{1-p}{np}}\,\%$.

Para esta variable aleatoria es importante observar que, si $X \rightsquigarrow B(n,p)$, y definimos la variable aleatoria Y =número de fracasos obtenidos en los n ensayos entonces es claro $Y \rightsquigarrow B(n, 1-p)$ y además

$$p(Y = y) = p(X = n - y) = \binom{n}{n-y} p^{n-y}(1-p)^y = \binom{n}{y}(1-p)^y p^{n-y}$$

Teniendo en cuenta esta propiedad bastará con saber calcular probabilidades de una distribución Binomial con $p \leq 0.5$.

Otra propiedad importante para esta distribución es que, si $X_1 \rightsquigarrow B(n_1, p)$ y $X_2 \rightsquigarrow B(n_2, p)$ con X_1 y X_2 independientes, entonces $X_1 + X_2 \rightsquigarrow B(n_1 + n_2, p)$.

Ejercicio. La probabilidad de que un enfermo cualquiera se recupere de una determinada enfermedad es de 0.7. En un hospital hay ingresadas 15 personas con dicha enfermedad. ¿Cuál es la probabilidad de que se recuperen 5 personas o menos? ¿Y la probabilidad de que se recuperen exactamente 5 personas?. ¿Cuál es la probabilidad de que se recuperen al menos 10 personas?. ¿Y de que se recuperen entre 8 y 12 personas?. Calcular el número esperado de personas que se recuperará.

5.3. Distribución Geométrica.

Supongamos que se repite un experimento de Bernouilli de forma independiente hasta que se obtenga el primer éxito. Consideramos la variable aleatoria X=número de fracasos obtenidos antes del primer éxito. Diremos entonces que la variable aleatoria X tiene una **distribución geométrica de parámetro** p, y lo denotaremos por $X \rightsquigarrow G(p)$. Es fácil demostrar que para esta variable aleatoria se tiene $Sop(X) = \{0, 1, ..., n, ...\} = \mathbb{N}$ y que su función de probabilidad es

$$p(X = x) = (1-p)^x p$$

Además, para esta variable aleatoria, aunque no entraremos en la demostración, se puede demostrar que $EX = \frac{1-p}{p}$ y $Var(X) = \frac{1-p}{p^2}$. Por tanto $\sigma = \frac{\sqrt{1-p}}{p}$ y $CV = \frac{100}{\sqrt{1-p}}\,\%$.

Este modelo de probabilidad se puede construir también a través de la variable aleatoria Y=número de ensayos necesarios para obtener el primer éxito. En este caso es claro que $Sop(Y) = \{1, ..., n, ...\} = \mathbb{N} - \{0\} = \mathbb{N}^*$ y $p(Y = y) = (1-p)^{y-1}p$. Además, es obvio que $Y = X + 1$ y por tanto $EY = EX + 1 = \frac{1-p}{p} + 1 = \frac{1}{p}$ y $Var(Y) = Var(X + 1) = Var(X) = \frac{1-p}{p^2}$.

Dependiendo de cómo venga expresado el problema podremos utilizar una u otra formulación.

Ejercicio. Supongamos que un determinado momento del día la probabilidad de que una centralita telefónica esté libre (y por tanto tengamos línea) es de 0.6. Calcular la probabilidad de que se consiga línea a la quinta llamada. Calcular la probabilidad de que hagan falta por lo menos 5 llamadas para conseguir línea. ¿Cuál es el número medio de llamadas necesarias para obtener línea?.

5.4. Distribución Binomial Negativa.

Dado un número natural $r \geq 2$, supongamos que se repite un experimento de Bernouilli de forma independiente hasta que se obtenga el r-ésimo éxito. Consideramos la variable aleatoria X=número de fracasos obtenidos antes del r-ésimo éxito. Diremos entonces que la variable aleatoria X tiene una **distribución binomial negativa de parámetros** r y p, y lo denotaremos por $X \rightsquigarrow BN(r, p)$. Observar que excluimos el caso $r = 1$ porque estaríamos en el caso anterior de la distribución geométrica. Es fácil demostrar que para esta variable aleatoria se tiene $Sop(X) = \{0, 1, ..., n, ...\} = \mathbb{N}$ y que su función de probabilidad es

$$p(X = x) = \binom{x + r - 1}{r - 1} p^r (1-p)^x$$

Nota: Observar que $\binom{x + r - 1}{r - 1} = \binom{x + r - 1}{x}$

Además, si para cada $i \in \{1, 2, ..., r\}$ definimos la variable aleatoria X_i =número de fracasos obtenidos entre el éxito $i-1$ y el éxito i, es evidente que, por la independencia en las repeticiones de los ensayos de Bernouilli, se verifica que $X_i \rightsquigarrow G(p)$ y la variable X puede expresarse como suma de las variables X_i, siendo éstas independientes. Es decir $X = \sum_{i=1}^{r} X_i$ con $X_i \rightsquigarrow G(p)$ e independientes. Teniendo en cuenta entonces las propiedades de la esperanza y la varianza podemos asegurar que $E(X) = E\left(\sum_{i=1}^{r} X_i\right) = \sum_{i=1}^{r} E(X_i) = r\frac{1-p}{p}$ y $Var(X) = \sum_{i=1}^{r} Var(X_i) = r\frac{1-p}{p^2}$. Por tanto $\sigma = \frac{\sqrt{r(1-p)}}{p}$ y $CV = \frac{100}{\sqrt{r(1-p)}}\,\%$.

Este modelo de probabilidad se puede construir también a través de la variable aleatoria Y=número de ensayos necesarios para obtener el r-ésimo éxito. En este caso es claro que $Sop(Y) = \{r, r+1, ...\}$ y $p(Y = y) = \binom{y-1}{r-1}(1-p)^{y-r}p^r$. Además, es obvio que $Y = X + r$ y por tanto $EY = EX + r = r\frac{1-p}{p} + r = \frac{r}{p}$ y $Var(Y) = Var(X + r) = Var(X) = r\frac{1-p}{p^2}$. Entonces, en este caso, tenemos $\sigma = \frac{\sqrt{r(1-p)}}{p}$ y $CV = 100\sqrt{\frac{1-p}{r}}\,\%$.

Dependiendo de cómo venga expresado el problema podremos utilizar una u otra formulación.

Ejercicio. Un científico necesita encontrar 5 monos afectados de una cierta enfermedad para realizar un estudio. La incidencia de esta enfermedad en la población de monos de una reserva es del 30 %. El científico examinará uno a uno a los monos capturados hasta encontrar los 5 afectados por la enfermedad.

Calcular la probabilidad de que tenga que examinar al menos a 20 monos. ¿Cuál es el número medio de exámenes requeridos?

5.5. Distribución de Poisson.

Esta distribución de probabilidad constituye uno de los modelos más utilizados en Probabilidad y Estadística. Surge de forma natural al considerar el límite de la distribución binomial $B(n,p)$ cuando $n \to \infty$ y $p \to 0$ pero manteniendo el producto $np = \lambda$ constante (observar que esta constante resulta ser el valor esperado de la distribución binomial). Veamos entonces cuál es el límite de la distribución de probabilidad binomial en estas circunstancias. Para ello tendremos en cuenta que:

$$\begin{aligned}\binom{n}{x}p^x(1-p)^{n-x} &= \frac{n(n-1)\cdots(n-x+1)}{x!n^x}(np)^x\frac{(1-p)^n}{(1-p)^x} \\ &= \left[\frac{\lambda^x}{x!}\right]\left[\frac{n(n-1)\cdots(n-x+1)}{n^x}\right]\frac{\left[\left(1-\frac{1}{1/p}\right)^{1/p}\right]^\lambda}{(1-p)^x}\end{aligned}$$

Entonces, teniendo en cuenta que $\lim\limits_{n\to\infty}\frac{n(n-1)\cdots(n-x+1)}{n^x}=1$, $\lim\limits_{p\to 0}(1-p)^x=1$ y $\lim\limits_{p\to 0}\left(1-\frac{1}{1/p}\right)^{1/p}=e^{-1}$, podemos asegurar que

$$\lim_{n\to\infty,p\to 0,np=\lambda}\binom{n}{x}p^x(1-p)^{n-x}=\frac{e^{-\lambda}\lambda^x}{x!}$$

También es fácil demostrar (utilizando el desarrollo de Mc-Laurin de la función e^λ) que $\sum\limits_{x=0}^{\infty}\frac{e^{-\lambda}\lambda^x}{x!}=1$ y, por tanto, estos sumandos constituyen una distribución de probabilidad cuando x varía en el conjunto de los números naturales. Todo esto motiva la siguiente definición.

Diremos que una variable aleatoria discreta X tiene una **distribución de Poisson de parámetro** λ cuando $Sop(X)=\mathbb{N}$ y su función de probabilidad viene dada por

$$p(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}$$

y lo representaremos por $X \rightsquigarrow P(\lambda)$.

Por tanto, cuando nos encontremos con un modelo binomial con n grande y p próximo a cero podremos utilizar la distribución de Poisson con parámetro $\lambda = np$ para calcular las correspondientes probabilidades con una buena aproximación. A título orientativo, diremos que esta sustitución puede resultar aconsejable cuando $n \geq 30$ y $p \leq 0.1$. En resumen, se puede decir que la distribución de Poisson es un modelo bastante razonable cuando estamos interesados en estudiar el número de éxitos obtenidos en un número grande de pruebas independientes de Bernouilli y la probabilidad de éxito cada vez que se repite la prueba es pequeña. Además, por extensión, también se utiliza este modelo de probabilidad cuando estamos interesados en calcular el número de éxitos contabilizados en una unidad de medida de tipo continuo, como puede ser el tiempo, longitud, superficie, volumen, etc. Por ejemplo: número de llamadas recibidas en una centralita en una hora, número de clientes que llegan a la cola de un supermercado en un minuto, número de bacterias en un litro de agua, número de coches que pasan por un determinado punto de una carretera en un minuto, etc. En efecto, en todos estos casos podríamos considerar la unidad de medida

continua (por ejemplo una hora) dividida en una infinidad de pequeños instantes de tiempo ($n \to \infty$) de modo que en cada instante sólo pudiese ocurrir, a lo sumo, un éxito y la probabilidad de que ocurriese ese éxito fuese muy pequeña ($p \to 0$), pero manteniendo un promedio de éxitos por hora constante ($\lambda = np$).

Finalmente, teniendo en cuenta la definición dada para esta distribución de probabilidad, y sabiendo que para una distribución de probabilidad binomial se tiene que $EX = np = \lambda$ y $Var(X) = np(1-p) = \lambda(1-p)$, parece lógico pensar que si X tiene una distribución de Poisson de parámetro λ entonces $EX = \lambda$ y $Var(X) = \lambda$ (estos valores pueden obtenerse también fácilmente utilizando las definiciones de esperanza y varianza). Por tanto $\sigma = \sqrt{\lambda}$ y $CV = \frac{100}{\sqrt{\lambda}}$ %.

Una propiedad importante de esta distribución es que si $X_1, X_2, ..., X_n$ son variables aleatorias independientes cada una de ellas con distribución de Poisson de un determinado parámetro, $X_i \rightsquigarrow P(\lambda_i)$, entonces se verifica que $\sum_{i=1}^{n} X_i$ también tiene una distribución de Poisson con parámetro igual a la suma de los respectivos parámetros, $\sum_{i=1}^{n} X_i \rightsquigarrow P\left(\sum_{i=1}^{n} \lambda_i\right)$. En particular, si $X_1, X_2, ..., X_n$ son variables aleatorias independientes con $X_i \rightsquigarrow P(\lambda)$ entonces se tiene que $\sum_{i=1}^{n} X_i \rightsquigarrow P(n\lambda)$. Ahora bien, esta propiedad no debe entenderse nunca diciendo que si $X \rightsquigarrow P(\lambda)$ entonces $kX \rightsquigarrow P(k\lambda)$, ya que no es lo mismo una variable aleatoria multiplicada por k que la suma de k variables aleatorias con la misma distribución de probabilidad.

Ejercicio. El número esperado de llamadas que entran en una centralita telefónica es de cuatro por minuto. Calcular la probabilidad de que no lleguen llamadas en un periodo de un minuto y la probabilidad de que lleguen por lo menos cuatro llamadas en un periodo de un minuto. ¿Cuál sería la probabilidad de que lleguen cinco llamadas en un periodo de dos minutos?.

5.6. Distribución Hipergeométrica.

Consideramos una población finita con N elementos de los cuales D son éxitos (es decir, tienen una determinada característica) y el resto, $N - D$, son fracasos (no tienen esa característica). Suponemos que extraemos de la población, de forma aleatoria, una muestra de n elementos sin devolución (cada extracción se hace sobre los elementos aún no extraídos, por tanto $n \leq N$) y consideramos la variable aleatoria X =número de éxitos obtenidos en las n extracciones. Diremos entonces que la variable aleatoria X tiene una **distribución hipergeométrica de parámetros** n, N **y** D, y lo denotaremos por $X \rightsquigarrow H(n, N, D)$, o alternativamente $X \rightsquigarrow H(p, n, N)$ siendo $p = \frac{D}{N}$ la proporción de éxitos en la población (el programa Statgraphics utiliza esta última notación). Habitualmente el soporte de esta variable aleatoria será $Sop(X) = \{0, 1, ..., n\}$, porque lo más común será que el número de éxitos (D) y el número de fracasos ($N - D$) en la población sea superior al número de extracciones (n). Sin embargo, si $D < n$ entonces el mayor valor del soporte será D y no n, es decir, el mayor valor del soporte será mín $\{n, D\}$. Del mismo modo, si $N - D < n$, entonces el menor valor del soporte será $n - (N - D)$ y no 0, es decir, el menor valor del soporte será máx $\{0, n - (N - D)\}$. Por tanto, en general los posibles valores para esta variable

aleatoria serán todos los números naturales x con $\text{máx}\{0, n-(N-D)\} \leq x \leq \text{mín}\{n, D\}$.

Teniendo en cuenta que las extracciones se hacen de forma aleatoria, y utilizando la Combinatoria, es fácil demostrar que para esta variable aleatoria la función de probabilidad es

$$p(X=x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}$$

Si consideramos las variables de Bernouilli X_i definidas para cada ensayo (1 si obtenemos éxito y 0 si obtenemos fracaso) es evidente que $X = \sum_{i=1}^{n} X_i$. Además vamos a ver que todas ellas tienen parámetro $p = \frac{D}{N}$, es decir, $X_i \rightsquigarrow B(p)$. En efecto, es evidente que $X_1 \rightsquigarrow B(p)$ y además, por la Regla de la Probabilidad Total, se tiene

$$\begin{aligned} p(X_2 &= 1) = p(X_1=0)p(X_2=1/X_1=0) + p(X_1=1)p(X_2=1/X_1=0) \\ &= \frac{N-D}{N}\frac{D}{N-1} + \frac{D}{N}\frac{D-1}{N-1} = \frac{ND-D}{N(N-1)} = \frac{D(N-1)}{N(N-1)} = \frac{D}{N} = p \end{aligned}$$

y por tanto también $X_2 \rightsquigarrow B(p)$. Del mismo modo se prueba que

$$\begin{aligned} p\left(X_3=1\right) &= p\left[(X_1=0)(X_2=0)\left(X_3=1\right)\right] + p\left[(X_1=0)(X_2=1)\left(X_3=1\right)\right] \\ &\quad + p\left[(X_1=1)(X_2=0)\left(X_3=1\right)\right] + p\left[(X_1=1)(X_2=1)\left(X_3=1\right)\right] \\ &= \left\{p(X_1=0)p(X_2=0/X_1=0)p\left[X_3=1/(X_1=0)(X_2=0)\right]\right\} \\ &\quad + \left\{p(X_1=0)p(X_2=1/X_1=0)p\left[X_3=1/(X_1=0)(X_2=1)\right]\right\} \\ &\quad + \left\{p(X_1=1)p(X_2=0/X_1=1)p\left[X_3=1/(X_1=1)(X_2=0)\right]\right\} \\ &\quad + \left\{p(X_1=1)p(X_2=1/X_1=1)p\left[X_3=1/(X_1=1)(X_2=1)\right]\right\} \\ &= \frac{N-D}{N}\frac{N-D-1}{N-1}\frac{D}{N-2} + \frac{N-D}{N}\frac{D}{N-1}\frac{D-1}{N-2} + \frac{D}{N}\frac{N-D}{N-1}\frac{D-1}{N-2} + \frac{D}{N}\frac{D-1}{N-1}\frac{D-2}{N-2} \\ &= \frac{(N-D)D(N-2)}{N(N-1)(N-2)} + \frac{D(D-1)(N-2)}{N(N-1)(N-2)} = \frac{D(N-1)}{N(N-1)} = \frac{D}{N} = p \end{aligned}$$

y $X_3 \rightsquigarrow B(p)$. Por consiguiente, en general, $X_i \rightsquigarrow B(p)$.

Por tanto, haciendo $p = \frac{D}{N}$, al igual que para la distribución binomial, podemos asegurar que, si $X \rightsquigarrow H(n,N,D)$, entonces $X = \sum_{i=1}^{n} X_i$ con $X_i \rightsquigarrow B(p)$. La diferencia está en que ahora las variables X_i no son independientes ya que, por ejemplo, se tiene

$$\begin{aligned} p\left[(X_1=1)(X_2=1)\right] &= p(X_1=1)p(X_2=1/X_1=1) = \frac{D}{N}\frac{D-1}{N-1} \\ &\neq p(X_1=1)p(X_2=1) = \frac{D^2}{N^2} \end{aligned}$$

A pesar de ello, por las propiedades de la esperanza, podemos asegurar que

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = np = n\frac{D}{N}.$$

Sin embargo, ahora no es cierto que $Var(X) = \sum_{i=1}^{n} Var(X_i) = np(1-p)$. De hecho se puede demostrar (aunque nosotros no lo haremos) que

$$Var(X) = np(1-p)\frac{N-n}{N-1} = n\frac{D}{N}\left(1-\frac{D}{N}\right)\frac{N-n}{N-1}.$$

En situaciones en las que el tamaño de la población N es grande el cálculo de probabilidades para esta distribución de probabilidad se dificulta por los tres números combinatorios que aparecen en la función de probabilidad. Por ello es interesante conocer la siguiente propiedad matemática:

$$\lim_{N\to\infty,\,\frac{D}{N}=p} \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} = \binom{n}{x}\, p^x\, (1-p)^{1-x}.$$

De ello deducimos que, si N es grande y $p = \frac{D}{N}$ es la proporción de éxitos en la población, entonces se verifica que $H(n, N, D) \approx B(n, p)$ con $p = \frac{D}{N}$ y podremos utilizar la distribución binomial para aproximar las probabilidades de la distribución hipergeométrica. En general, esta aproximación puede considerarse satisfactoria cuando $N > 50$ y $\frac{n}{N} \leq 0.1$.

Ejercicio. Un lote de 200 piezas contiene un 40 % de defectuosas. Calcular la probabilidad de obtener 10 piezas defectuosas al sacar una muestra de 20 piezas del lote, tanto si el muestreo se hace sin reemplazamiento como si se hace con reemplazamiento. ¿Se puede considerar el tamaño del lote, $N = 200$, lo suficientemente grande para que la aproximación de la hipergeométrica por la binomial sea satisfactoria?

5.7. PROBLEMAS

1. Un club de automovilistas comienza una campaña telefónica para la captación de nuevos socios. De acuerdo con la experiencia previa se sabe que una de cada 10 personas que recibe la llamada se une al club. Si en un día 25 personas son contactadas por teléfono, se pide:
   (a) ¿Cuál es la probabilidad de que se unan exactamente 4 personas al club?
   (b) ¿Cuál es la probabilidad de que por lo menos 3 personas se unan al club?
   (c) ¿Cuál es el número esperado de personas que se van a unir al club?

2. En un taller hay 10 máquinas iguales. Se ha visto que cada máquina se suele averiar un día de cada cinco. Evaluar las probabilidades de que en un cierto día haya r o más máquinas averiadas. Si la pérdida diaria ocasionada por tener una máquina averiada es de 3000 euros, calcular la pérdida media diaria.

3. En una reunión de 300 individuos. calcular la probabilidad de que hayan nacido k de ellos el día de Navidad.

4. La probabilidad de que un enfermo se recupere de cierta enfermedad es de 0.70. En un hospital hay ingresadas 15 personas con dicha enfermedad. ¿Cuál es la probabilidad de que:
   (a) ¿se recuperen 5 personas o menos?
   (b) ¿sobrevivan exactamente 5 personas?

(c) ¿al menos 10 se recuperen?

(d) ¿se recuperen entre 8 y 12 personas?

(e) Calcular el número de personas que se espera se recuperen.

5. El número medio de resistencias que produce una fábrica es de 100 por día. De ellas, aproximadamente un 5 % no cumple con las normas de calidad mínimas que exige el mercado.

   (a) Obtener la distribución del número de resistencias diarias que no cumplen con las normas de calidad.

   (b) Calcular la probabilidad de que en un día todas las resistencias fabricadas cumplan con la normativa.

   (c) Calcular la probabilidad de que más de 5 resistencias no cumplan con las especificaciones en un día determinado.

6. Un fabricante de automóviles compra motores a una compañía. El fabricante recibe un lote de 40 motores. Su plan para aceptar el lote consiste en seleccionar 8 motores de entre los 40 de manera aleatoria y someterlos a prueba. Si encuentra que ninguno de los motores presenta serios defectos el fabricante acepta el lote; de otra forma, lo rechaza. Si el lote de 40 contiene cuatro motores con serios defectos, ¿cuál es la probabilidad de que sea aceptado?

7. Demostrar que la distribución hipergeométrica tiende a la binomial en el límite cuando $N \to \infty$ y $p = \frac{D}{N}$ permanece constante. Es decir, demostrar que, si $p = \frac{D}{N}$, entonces

$$\lim_{N\to\infty} \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} = \binom{n}{x} p^x (1-p)^{n-x}$$

8. Un lote de 200 piezas tiene un 40 % de defectuosas. Calcular la probabilidad de obtener 10 piezas defectuosas al sacar una muestra de 20 piezas del lote si

   (a) si el muestro se hace sin reemplazamiento.

   (b) si el muestreo se hace con reemplazamiento.

   (c) ¿Se puede considerar $N = 200$, el tamaño del lote, lo suficientemente grande para que la función de probabilidad binomial sea buena aproximación de la hipergeométrica?

9. El número de vehículos que pasa por determinado punto de una carretera es 10 por minuto.

   (a) Calcular la probabilidad de que pasen por lo menos 15 vehículos por este punto en cualquier minuto.

   (b) Calcular la probabilidad de que pasen por lo menos 15 vehículos por este punto en un intervalo de dos minutos.

10. Se está desarrollando una nueva variedad de maíz en una estación agrícola experimental. Se espera que germinen el 90 % de las semillas. Para verificar esta tasa se plantan 100 semillas en un suelo de idéntica composición y se les dedican los mismos cuidados. Si la tasa de germinación del 90 % fuera

correcta, ¿cuántas semillas habría que esperar que germinaran? Si germinan 84 semillas o menos, ¿habría razones para sospechar de que la tasa del 90 % fuera correcta?

11. A fin de aceptar o rechazar lotes de piezas se propone el siguiente plan de inspección por muestreo: los lotes tienen 1000 piezas y se toman 50 piezas al azar. Se aceptará el lote si la muestra da una o menos defectuosas y se rechazará en caso contrario. Calcular la probabilidad de rechazar el lote si p es la proporción de defectuosos en el lote, según que el muestreo se haga con o sin reemplazamiento. Particularizarlo para los casos en que la proporción de defectuosos del lote sea del 4 por mil y del 1 por ciento.

12. En una línea de montaje operada con robots industriales se pueden instalar cajas de engranajes en un minuto si los agujeros para los tornillos han sido barrenados de forma correcta. y en diez minutos si hay que volver a barrenar. Se dispone de 20 cajas de engranajes, de las cuales dos tienen los agujeros barrenados de forma incorrecta . Se deben seleccionar cinco cajas de las 20 disponibles para que las instalen los cinco robots de la línea.

    (a) Calcular la probabilidad de que las cinco cajas de engranajes se ajusten adecuadamente.

    (b) Calcular el valor esperado y la desviación estándar del tiempo que se necesita para instalar las cinco cajas de engranajes.

13. El número esperado de llamadas que entran en una centralita telefónica es de cuatro por minuto. Calcular la probabilidad de que:

    (a) no lleguen llamadas en un período de un minuto.

    (b) por lo menos lleguen cuatro llamadas en un período de un minuto.

    (c) por lo menos lleguen cinco llamadas en un período de dos minutos.

14. Se ha observado un telar durante un período de tiempo, anotándose el número de roturas por 10000 pasadas de lanzadera. obteniéndose la siguiente tabla:

Número de roturas	0	1	2	3	4	5	6
Frecuencia	39	48	39	16	6	1	1

    Ajustar a una ley de Poisson. Según la ley ajustada, ¿qué frecuencia cabría esperar para cada valor del número de roturas si realizáramos 150 observaciones?

15. Supongamos que la demanda de televisores de una cierta marca en un mes cualquiera sigue una distribución de Poisson de parámetro 25 . ¿Qué existencias ha de tener el comerciante al principio del mes para tener una probabilidad de 0.90 de satisfacer toda la demanda del mes? ¿ Y para satisfacerla con una probabilidad de 0.99?

16. El 60 % de los consumidores prefiere la pasta dental marca A. Si se entrevista a un grupo de consumidores, ¿cuál es la probabilidad de que se tenga que entrevistar exactamente a cinco personas para encontrar a una persona que prefiera la marca A? ¿y de entrevistar por lo menos a 5 personas?

17. Cierta centralita telefónica es de pequeña capacidad en cuanto al número de llamadas que puede recibir. De hecho, se sabe que en el período de mayor congestión la probabilidad de que la centralita esté libre es de 0.60. Calcular la probabilidad de que:

    (a) Se logre línea a la quinta llamada.

    (b) Hagan falta por lo menos 5 llamadas para conseguir línea.

    (c) ¿Cuál es el número medio de llamadas necesario para conseguir línea?

18. Un zoólogo desea capturar un ejemplar de cierta especie de arácnidos en un paraje en el que se sabe que dicha especie supone el 15 % de los arácnidos presentes. ¿Cuál es la probabilidad de que tenga que capturar 10 o más ejemplares para obtener el de la especie deseada?

19. Un científico necesita encontrar 5 monos afectados de cierta enfermedad para realizar un estudio. La incidencia de esta enfermedad en la población de monos de una reserva es del 30 %. El científico examinará uno a uno los monos hasta dar con los 5 afectados por la enfermedad. Calcular:

    (a) La probabilidad de que tenga que examinar almenas 20 monos.

    (b) El número medio de exámenes requeridos.

20. Se estima que sólo 1 de cada 50 loros capturados en la cuenca del Amazonas para su utilización como animales domésticos sobrevive al cambio.

    (a) Unos cazadores furtivos han capturado 700 pájaros. ¿Cuál es el número esperado de sobrevivientes? ¿Cuál es la probabilidad de que sobrevivan a lo sumo 10 pájaros?

    (b) El dueño de una tienda ilegal de mascotas paga 4$ a los cazadores furtivos por cada loro capturado. El precio de venta de un loro en su tienda es de 300$. ¿Cuántos loros deben capturar los cazadores si el dueño de la tienda espera tener al menos una ganancia de 1000$?

21. El número de colonias bacteriológicas por mm^2 en un determinado cultivo da la siguiente distribución de frecuencias:

Número de colonias	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Frecuencia	10	50	100	168	190	180	120	80	45	20	8	7	2	1	0

    ¿Qué distribución teórica podríamos emplear para ajustar los datos? Obtener las frecuencias teóricas utilizando dicha distribución.

22. Una máquina produce en serie cierto tipo de tornillos que son recogidos en cajas que contienen 1200. La experiencia ha demostrado que la proporción de estas cajas, según el número de tornillos defectuosos que contienen, es:

% de defectuosos	0	1	2	3	4	5	6
proporción con ese %	0.78	0.17	0.034	0.009	0.005	0.002	0.000

    Se considera aceptable una caja que contiene el 2 % o menos de defectuosos. El objeto de la inspección, por lo tanto, es tratar de rechazar aquellas que tengan un porcentaje de defectuosos superior al 2 %. La inspección normal consiste en el examen de 50 tornillos de cada caja.

(a) Calcular la probabilidad de encontrar 6 tornillos defectuosos entre los 50 inspeccionados según sea el porcentaje real de tornillos defectuosos que contenga la caja y según que el muestreo se haga con o sin reemplazamiento.

(b) Si decidimos rechazar toda caja que nos muestre exactamente 6 tornillos defectuosos entre los 50 inspeccionados, ¿qué porcentaje de cajas rechazadas mediante la inspección son realmente aceptables? (dejar indicados los resultados).

23. Un proceso de fabricación de cierto tipo de piezas es tal que el 2 % de las piezas que se producen son defectuosas. Las piezas se embalan en lotes de 100. A un cierto comprador no le importa aceptar el lote si lleva 4 o menos piezas defectuosas, pero no quiere aceptar lotes con más de 4 piezas defectuosas. El vendedor asegura que en ningún lote hay más de 4 piezas defectuosas, aunque realmente no se ha preocupado de que esto sea cierto y ha empaquetado todas las piezas producidas. El comprador, fiado de lo que le dicen, no lleva a cabo ninguna inspección . ¿Qué proporción de lotes con más de 4 piezas defectuosas aceptará el comprador?

24. Una muestra de sangre se examina al microscopio sobre una cuadrícula dividida en 400 cuadrados iguales. Suponiendo que la muestra ha sido diluida y homogeneizada de manera que los hematíes queden distribuidos al azar, y habiéndose observado 25 cuadrados vacíos, hallar el número medio de hematíes por cuadrado. Si la muestra ha sido diluida al 0.25 % y el volumen de disolución distribuido sobre la cuadricula es 0.1 mm^3 , ¿cuántos hematíes por mm^3 de sangre se puede esperar que tenga el paciente?

25. Un instituto de certificación de semillas desea garantizar que el porcentaje de semillas virosadas en las partidas que certifica, no supera el 5 por mil. Para ello se selecciona al azar de cada partida (que consta de un gran número de semillas) n semillas que tritura formando una pasta con la que inocula a una planta testigo. Se sabe que basta con que una de las n semillas esté virosada para que la planta testigo desarrolle síntomas de contagio, en cuyo caso, la partida correspondiente se rechazará.

(a) Determinar cuánto debe valer como mínimo n si se desea que la probabilidad de aceptar una partida que contenga el 5 por mil o más de semillas virosadas sea menor o igual que el 1 %.

(b) Calcular cuál es, con el plan de control establecido en el apartado anterior, la probabilidad de rechazar una partida que sólo tenga un 1 por mil de semillas infectadas.

SOLUCIONES

1. (a) 0.1384 (b) 0.4629 (c) 2.5 personas.

2.

r	0	1	2	3	4	5	6	7	8	9	10
$p(X \geq r)$	1	0.8926	0.6242	0.3222	0.1209	0.0328	0.0064	0.0009	0.0001	0.0000	0.0000

La pérdida media diaria por averías será 6000 euros.

3. $\binom{300}{k}\left(\frac{1}{365}\right)^k\left(\frac{364}{365}\right)^{300-k}$

4. (a) 0.0037 (b) 0.003 (c) 0.7216 (d) 0.8232 (e) 10.5 personas

5. (a) $X \rightsquigarrow P(5)$ (b) 0.0067 (c) 0.3840

6. 0.3935

7. Ver fichero de problema resueltos.

8. (a) 0.1184 (b) 0.1171 (c) Sí

9. (a) 0.0835 (b) 0.8951

10. (a) $E(X) = 90$ semillas (b) $p\,(X \leq 84) \cong 0.0336$, luego sí sería sospechoso.

11. Sin reemplazamiento: $p = 1 - \dfrac{\binom{100(1-p)}{50} + \binom{100p}{1}\binom{100(1-p)}{49}}{\binom{100}{50}}$

    Con reemplazamiento: $p = 1 - \left[(1-p)^{50} + 50p(1-p)^{49}\right]$

	$p = 0.004$	$p = 0.01$
Sin reemplazamiento	0.0138	0.085
Con reemplazamiento	0.0173	0.089

12. (a) 0.5526 (b) $E(T) = 9.5$ minutos con $\sqrt{Var(T)} = 5.4$ minutos.

13. (a) 0.0183 (b) 0.5665 (c) 0.9004

14. $X \rightsquigarrow P(1.4)$

Número de roturas	0	1	2	3	4	5	6 o más
Frecuencia teórica	37	52	36	17	6	2	0

15. Para satisfacer la demanda con una probabilidad de 0.90 ha de tener unas existencias de 30 televisores y para satisfacerla con una probabilidad de 0.99 debería tener unas 26.

16. (a) 0.0154 (b) 0.0256

17. (a) 0.0145 (b) 0.0256 (c) 1.67 llamadas.

18. 0.2316

19. (a) 0.282 (b) $E(X) = 16.67$

20. (a) $E(X) = 14$ loros, $p\,(X \leq 10) = 0.1757$ (b) Como mínimo, 500 loros.

21. Podemos suponer $X \rightsquigarrow P\,(4.5138)$

Número de colonias	0	1	2	3	4	5	6	7	8	9	10	11	12	13 o más
Frecuencia	11	49	109	165	186	168	126	81	46	23	10	4	2	1

22. (a) Sean $Y =$*porcentaje de piezas defectuosas que contiene la caja*

$X =$*número de piezas defectuosas que hay en la muestra de tamaño* 50

Con reemplazamiento: $X/Y = k \rightsquigarrow B(50, 0.01k)$, $p(X = 6/Y = k) = \binom{50}{6} \left(\frac{k}{100}\right)^6 \left(1 - \frac{k}{100}\right)^{44}$

Sin reemplazamiento: $X/Y = k \rightsquigarrow H(1200, 12k, 50)$, $p(X = 6/Y = k) = \dfrac{\binom{12k}{6}\binom{1200 - 12k}{44}}{\binom{1200}{50}}$

% defectuosas en la caja	0	1	2	3	4	5	6
Con reemplazamiento	0	0.00001	0.00042	0.003	0.011	0.026	0.049
Sin reemplazamiento	0	0.0000029	0.00026	0.0024	0.0096	0.024	0.047

(b) Con reemplazamiento: $p(Y \leq 2/X = 6) = \dfrac{\sum_{k=1}^{2} p(y = k)k^6(100 - k)^{44}}{\sum_{k=1}^{6} p(y = k)k^6(100 - k)^{44}} = 0.107$

Sin reemplazamiento: $p(Y \leq 2/X = 6) = \dfrac{\sum_{k=1}^{2} p(y = k)\binom{12k}{6}\binom{1200 - 12k}{44}}{\sum_{k=1}^{6} p(y = k)\binom{12k}{6}\binom{1200 - 12k}{44}} = 0.074$

23. Aproximadamente, el 5 % de los lotes tendrá más de 4 piezas defectuosas

24. Número medio de hematíes por cuadrado= $\ln(16)$

Número medio de hematíes por mm3 de sangre= $\frac{10}{0.0025} 400 \ln(16) = 4436.142$ hematíes/mm^3

25. (a) n ha de ser como mínimo 919 semillas.

(b) 0.6013

Capítulo 6

DISTRIBUCIONES DE PROBABILIDAD CONTINUAS

En este capítulo estudiaremos los modelos de probabilidad de tipo continuo más comunes.

6.1. Distribución uniforme continua.

Dados dos números reales a y b, con $a < b$, diremos que una variable aleatoria X tiene una **distribución uniforme en el intervalo** $[a, b]$, y lo denotaremos por $X \rightsquigarrow U[a, b]$, cuando su función de densidad sea

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{si } x < a \text{ o } x > b \end{cases}$$

Por tanto, esta distribución se caracteriza porque la densidad de probabilidad se reparte por igual en todo el intervalo $[a, b]$ y podemos decir que es la generalización al caso continuo de una probabilidad en un espacio muestral discreto en el que todos sus elementos son igualmente probables. En este caso, como todos los puntos del intervalo $[a, b]$ tiene una probabilidad 0 por tratarse de una variable continua, ahora lo que debemos decir es que cualquier subintervalo de una longitud fija d contenido en $[a, b]$ tiene igual probabilidad, y resulta ser $\frac{d}{b-a}$.

Es fácil demostrar que, para esta variable aleatoria, la función de distribución es

$$F(x) = \begin{cases} 0 & \text{si } x < a \\ \frac{x-a}{b-a} & \text{si } a \leq x \leq b \\ 1 & \text{si } x > b \end{cases}$$

Además, se tiene que

$$EX = \int_a^b \frac{x}{b-a} dx = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}$$

y

$$E\left(X^2\right) = \int_a^b \frac{x^2}{b-a}dx = \frac{b^3-a^3}{3\,(b-a)} = \frac{a^2+ab+b^2}{3}$$

Por tanto, $Var(X) = E\left(X^2\right) - (EX)^2 = \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} = \frac{(b-a)^2}{12}$ y $\sigma = \frac{b-a}{2\sqrt{3}}$. De ello se deduce que $CV = \frac{100(b-a)}{\sqrt{3}(a+b)}$. También es fácil comprobar que $Me = EX = \frac{a+b}{2}$. Finalmente, una propiedad de esta distribución es que para cualesquiera c, d con $a \leq c \leq d \leq b$ se verifica que

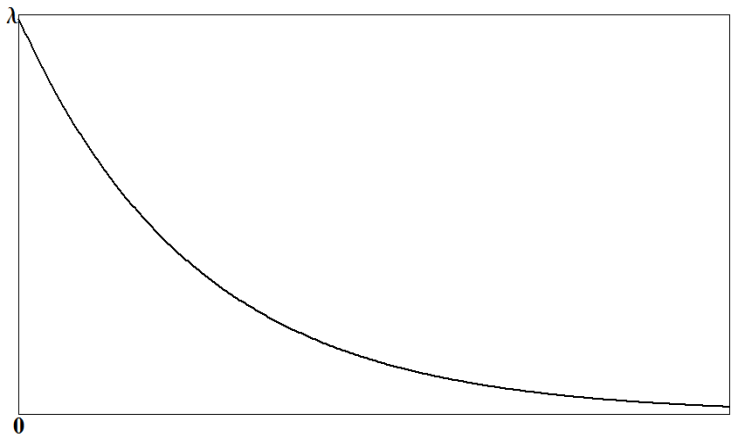
$$p\,(X > d \;/\; X > c) = \frac{b-d}{b-c}.$$

Ejercicio. De una estación parte un tren cada 20 minutos. Un pasajero llega de improviso. Hallar la función de distribución y la función de densidad de la variable aleatoria X=tiempo de espera del viajero (en minutos). Calcular la probabilidad de que el pasajero tenga que esperar menos de 7 minutos. ¿Cuál será la esperanza y la desviación típica del tiempo de espera?.

6.2. Distribución exponencial.

Diremos que una variable aleatoria X tiene una **distribución exponencial de parámetro** $\lambda > 0$, y lo representaremos por $X \rightsquigarrow \exp(\lambda)$, cuando su función de densidad está dada por

$$f(x) = \begin{cases} \lambda\, e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$



Esta distribución se utiliza a menudo para modelizar la duración o el tiempo de vida de sistemas con la particularidad de que la tasa de fallo permanece constante a lo largo del tiempo. En general, puede utilizarse siempre que estemos esperando a que ocurra algo (un éxito) y esa espera se mida en una unidad de medida de tipo continuo (tiempo, longitud, superficie, volumen, etc.), en lugar de una medida discreta como son las tiradas en la distribución geométrica. En cierto modo puede considerarse entonces como una extensión de la distribución geométrica al caso continuo. Pero resulta necesario suponer que el número esperado de fallos en dicha unidad de medida permanece constante. Por ejemplo, puede utilizarse para

modelizar el tiempo de vida o el tiempo de espera cuando no hay envejecimiento o desgaste, para que podamos decir que la tasa de fallo permanece constante (tiempo de vida de una bacteria, tiempo que pasa entre dos llamadas consecutivas en una centralita telefónica, etc.). También para modelizar tamaños en los cuáles los valores más bajos son los más probables, como el tamaño de las rentas familiares, o las precipitaciones en una región con escasas lluvias.

La función de distribución de esta variable aleatoria es

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

y sus medidas descriptivas son $EX = \frac{1}{\lambda}$, $Var(X) = \frac{1}{\lambda^2}$, $\sigma = \frac{1}{\lambda} = EX$, $CV = 100\,\%$ y $Me = \frac{-\ln\ 0.5}{\lambda}$.

Finalmente, una propiedad característica de esta distribución es que, para cualesquiera valores positivos a y b, se verifica que

$$p(X > a + b \ / \ X > a) = p(X > b)$$

o, equivalentemente,

$$p(X > a + b) = p(X > a)\ p(X > b)$$

Esta propiedad se puede interpretar diciendo que la probabilidad de que el tiempo de vida sea superior a b unidades adicionales es independiente del tiempo de vida ya vivido, a, y por ello se dice que la distribución exponencial no tiene memoria. Por tanto no sería adecuada, por ejemplo, para el tiempo de vida de personas humanas, en las que existe envejecimiento y la tasa de fallo no permanece constante.

(Nota. El programa Statgraphics utiliza $f(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}}$ y por tanto $E(X) = \lambda$)

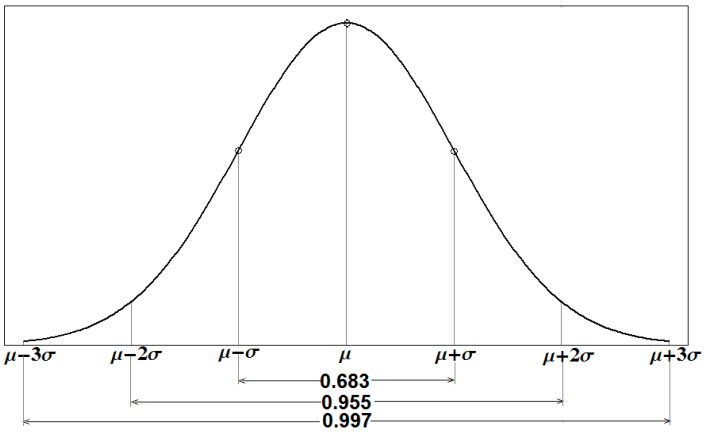
Ejercicio. Se ha comprobado que el tiempo de vida de unas determinadas bacterias sigue una distribución exponencial, siendo el tiempo de vida medio de una bacteria 225 días. Calcular la probabilidad de que una bacteria viva más de 8 meses (240 días) y obtener el percentil 95 de esta distribución. ¿Cuál será la probabilidad de que una bacteria que ha vivido ya 9 meses (270 días) viva como poco 3 meses más (90 días)?.

6.3. Distribución normal.

La distribución normal es el modelo más importante y más utilizado para variables aleatorias continuas. Su importancia proviene de que aparece (de forma exacta o aproximada) en muchas situaciones reales: medidas morfológicas en especies animales o vegetales (peso, altura, etc.), mediciones de experimentos físicos, etc. En general, la distribución normal surge siempre que los resultados de un experimento sean debidos a un conjunto muy grande de causas independientes que actúan sumando efectos, siendo cada efecto individual de poca importancia respecto al conjunto.

Diremos que una variable aleatoria X tiene una **distribución normal de parámetros** $\mu \in \mathbb{R}$ **y** $\sigma > 0$, y lo representaremos por $X \rightsquigarrow N(\mu, \sigma)$, cuando su función de densidad está dada por

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{para todo } x \in \mathbb{R}$$



La función de distribución de esta variable aleatoria no puede expresarse matemáticamente de forma explícita, porque en general la función $f(x) = e^{-x^2}$ no admite una primitiva que pueda expresarse mediante las funciones matemáticas elementales. Por eso simplemente podemos decir que

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

y cuando sea necesario evaluar esta función tendremos que utilizar integración numérica para obtener valores aproximados.

A continuación exponemos una serie de propiedades básicas que resulta necesario conocer para poder trabajar con esta distribución.

1. $EX = \mu$.
2. $Var(X) = \sigma^2$.
3. Es simétrica respecto de su esperanza, mediana y moda que coinciden con su parámetro μ. Por consiguiente podemos asegurar que, para cualquier $x \in \mathbb{R}$ se verifica que $p\,(X < \mu - x) = p\,(X > \mu + x)$.
4. El primer cuartil es $Q_1 = \mu - 0.675\sigma$ y, por simetría, el tercer cuartil es $Q_3 = \mu + 0.675\sigma$. Por tanto se verifica que $p\,(\mu - 0.675\sigma \leq X \leq \mu + 0.675\sigma) = 0.5$.
5. Las probabilidades alrededor de su valor esperado μ verifican: $p\,(\mu - \sigma \leq X \leq \mu + \sigma) = 0.683$, $p\,(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.955$ y $p\,(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.997$.
6. Si $X \rightsquigarrow N(\mu, \sigma)$ entonces la variable estandarizada $Z = \frac{X-\mu}{\sigma}$ tiene una distribución normal $(0, 1)$. Es decir, $Z \rightsquigarrow N(0, 1)$. Por tanto, para calcular probabilidades con cualquier distribución normal basta saber hacerlo con la distribución $Z \rightsquigarrow N(0, 1)$, ya que podremos asegurar que

$$p\,(a \leq X \leq b) = p\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = p\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right)$$

7. De forma general, si $X \rightsquigarrow N(\mu, \sigma)$ entonces $aX + b \rightsquigarrow N(a\mu + b, |a|\,\sigma)$ para cualesquiera números reales a, b.

8. Si $X_1 \rightsquigarrow N(\mu_1, \sigma_1)$, $X_2 \rightsquigarrow N(\mu_2, \sigma_2)$ entonces $X_1 + X_2 \rightsquigarrow N\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2 + 2Cov(X_1, X_2)}\right)$. En particular, si X_1 y X_2 son independientes se verifica que $X_1 + X_2 \rightsquigarrow N\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$.

9. Si $X_1 \rightsquigarrow N(\mu_1, \sigma_1)$, $X_2 \rightsquigarrow N(\mu_2, \sigma_2)$ entonces $X_1 - X_2 \rightsquigarrow N\left(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2 - 2Cov(X_1, X_2)}\right)$. En particular, si X_1 y X_2 son independientes se verifica que $X_1 - X_2 \rightsquigarrow N\left(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$.

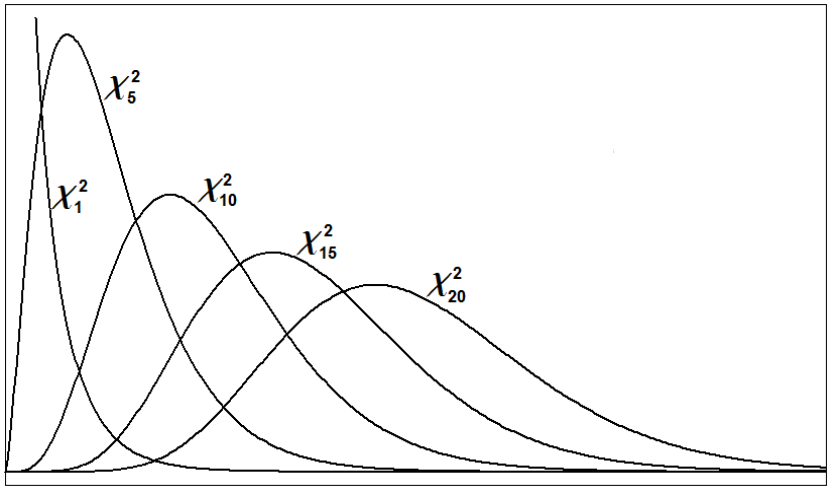
10. Si $X_1, X_2, ..., X_n$ son variables aleatorias independientes con $X_i \rightsquigarrow N(\mu_i, \sigma_i)$ para $i = 1, 2, ..., n$, entonces se verifica que $\sum_{i=1}^{n} X_i \rightsquigarrow N\left(\sum_{i=1}^{n} \mu_i, \sqrt{\sum_{i=1}^{n} \sigma_i^2}\right)$. Como caso particular, si $X_i \rightsquigarrow N(\mu, \sigma)$ para $i = 1, 2, ..., n$, entonces se verifica que $\sum_{i=1}^{n} X_i \rightsquigarrow N\left(n\mu, \sigma\sqrt{n}\right)$ y por tanto $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Ejercicio. Suponiendo que $X \rightsquigarrow N(5, 4)$ calcular la probabilidad de que X tome valores entre -1 y 7. ¿Cuáles serán el percentil 40 y el percentil 90 de esta distribución?.

Los modelos de probabilidad de tipo continuo que estudiaremos a continuación se derivan de efectuar determinadas operaciones matemáticas con variables que tienen distribución normal, y serán de gran utilidad en el último bloque de la asignatura, dedicado a la Inferencia Estadística.

6.4. Distribución ji-cuadrado de Pearson $\left(\chi^2\right)$.

Supongamos que $Z_1, Z_2, ..., Z_n$ son variables aleatorias independientes y todas con distribución $N(0, 1)$, es decir, $Z_i \rightsquigarrow N(0, 1)$. Entonces diremos que la variable aleatoria definida como $\sum_{i=1}^{n} Z_i^2$ tiene una **distribución ji-cuadrado (o chi-cuadrado) de Pearson con n grados de libertad**, y lo representaremos por $\sum_{i=1}^{n} Z_i^2 \rightsquigarrow \chi_n^2$. Aunque no estudiaremos la función de densidad de esta variable aleatoria, representamos a continuación su forma para diferentes valores del parámetro n.



Podemos observar que esta distribución presenta siempre una asimetría positiva que se va perdiendo a medida que el parámetro n crece. En general, podemos decir que si $n \geq 30$ la distribución ya es bastante simétrica y si $n \geq 300$ esta simetría ya es prácticamente perfecta. Además, para esta variable aleatoria se verifica que $E(\chi^2_n) = n$ y $Var(\chi^2_n) = 2n$. Por tanto $\sigma = \sqrt{2n}$ y $CV = 100\sqrt{\frac{2}{n}}\,\%$.

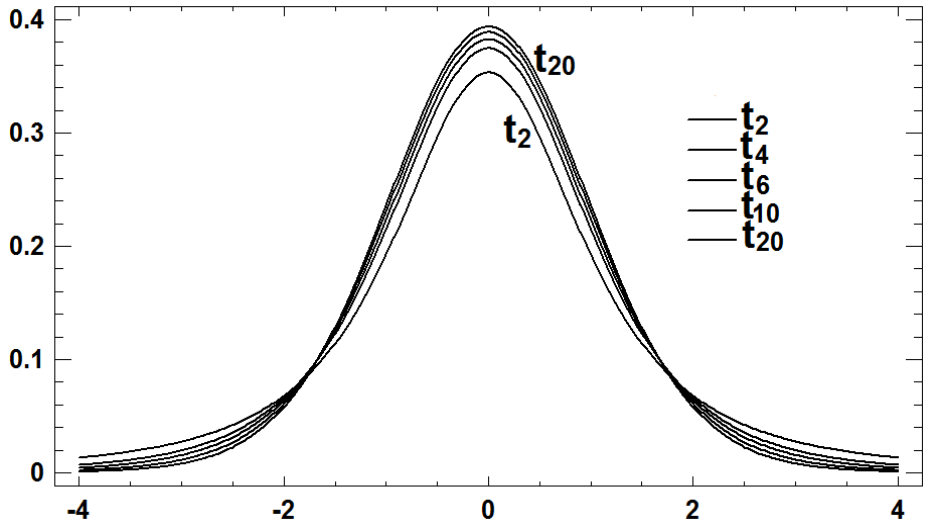
Las probabilidades de esta distribución se encuentran tabuladas, normalmente hasta $n = 30$, para diferentes valores $\alpha \in (0,1)$, de modo que $\chi^2_{n;\alpha}$ representa al punto que, para una distribución ji-cuadrado con n grados de libertad, deja una probabilidad a su derecha igual a α, es decir, $p\left(\chi^2 \geq \chi^2_{n;\alpha}\right) = \alpha$.

Una propiedad importante de esta distribución es que si X_1 y X_2 son variables aleatorias independientes cada una de ellas con distribución ji-cuadrado, es decir, $X_1 \rightsquigarrow \chi^2_{n_1}$ y $X_2 \rightsquigarrow \chi^2_{n_2}$, entonces se verifica que $X_1 + X_2$ también tiene una distribución ji-cuadrado, concretamente $X_1 + X_2 \rightsquigarrow \chi^2_{n_1+n_2}$.

Ejercicio. Encontrar los valores de $\chi^2_{16;0.025}$, $\chi^2_{16;0.975}$, $\chi^2_{16;0.05}$ y $\chi^2_{16;0.95}$. A partir de ellos, calcular $p\left(\chi^2_{16;0.025} \leq \chi^2_{16} \leq \chi^2_{16;0.975}\right)$ y $p\left(\chi^2_{16;0.05} \leq \chi^2_{16} \leq \chi^2_{16;0.95}\right)$.

6.5. Distribución t de Student.

Supongamos que $Z_1, Z_2, ..., Z_n$ y Z son variables aleatorias independientes y todas con distribución $N(0,1)$, es decir, $Z_i \rightsquigarrow N(0,1)$ y $Z \rightsquigarrow N(0,1)$. Entonces diremos que la variable aleatoria definida como $\frac{Z}{\sqrt{\frac{\sum_{i=1}^{n} Z_i^2}{n}}}$ tiene una **distribución t de Student con n grados de libertad**, y lo representaremos por $\frac{Z}{\sqrt{\frac{\sum_{i=1}^{n} Z_i^2}{n}}} \rightsquigarrow t_n$. Teniendo en cuenta que, como acabamos de ver, $\sum_{i=1}^{n} Z_i^2 \rightsquigarrow \chi^2_n$ podemos decir entonces que $t_n \equiv \frac{N(0,1)}{\sqrt{\frac{\chi^2_n}{n}}}$. Aunque no estudiaremos la función de densidad de esta variable aleatoria, representamos a continuación su forma para diferentes valores del parámetro n.



Podemos observar que esta distribución es simétrica y muy parecida a la distribución $N(0,1)$, de

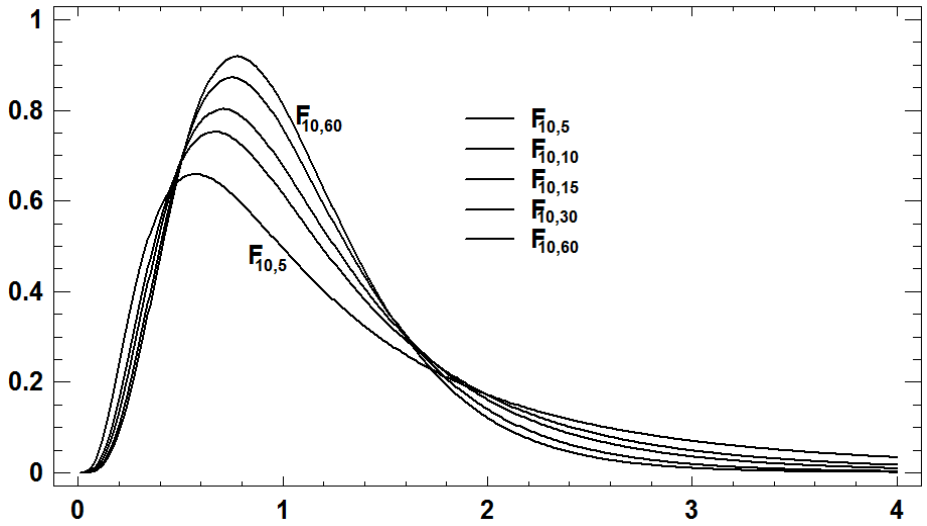
modo que para $n \geq 30$ son ya prácticamente iguales. Además, para esta variable aleatoria se verifica que $E(t_n) = 0$ y, si $n > 2$, $Var(t_n) = \frac{n}{n-2}$, presentando por tanto una mayor dispersión que la distribución $N(0,1)$.

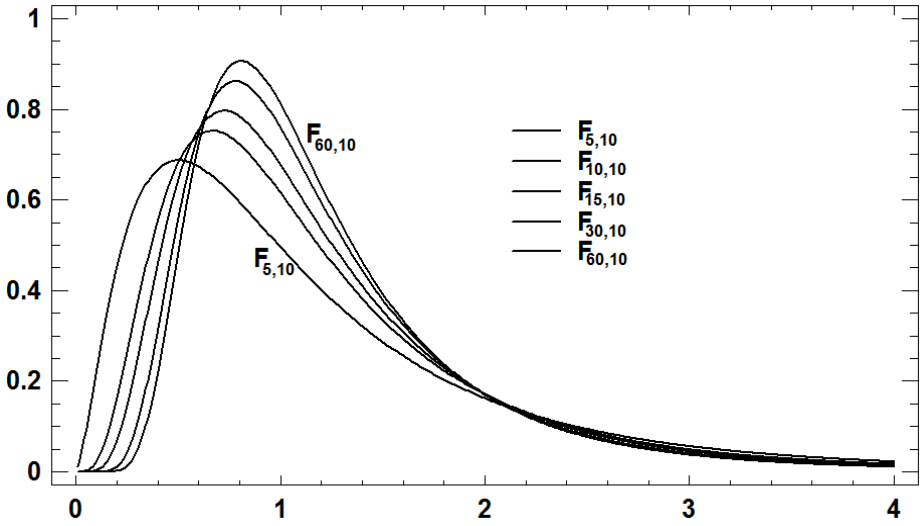
Las probabilidades de esta distribución se encuentran tabuladas, normalmente hasta $n = 30$, para diferentes valores $\alpha \in (0,1)$, de modo que $t_{n;\alpha}$ representa al punto que, para una distribución t de Student con n grados de libertad, deja una probabilidad a su derecha igual a α, es decir, $p\,(t \geq t_{n;\alpha}) = \alpha$. Además, en este caso, por la simetría también tendremos que $p\,(t \leq -t_{n;\alpha}) = \alpha$ y por este motivo sólo encontramos tabulados valores de α menores que 0.5, pudiendo obtenerse los correspondientes a α mayores que 0.5 sin más que cambiar el signo.

Ejercicio. Encontrar los valores de $t_{14;0.005}$, $t_{14;0.025}$ y $t_{14;0.05}$. A partir de ellos, calcular las siguientes probabilidades $p\,(-t_{14;0.025} \leq t_{14} \leq t_{14;0.025})$ y $p\,(-t_{14;0.05} \leq t_{14} \leq t_{14;0.05})$.

6.6. Distribución F de Fisher.

Supongamos que X_1 y X_2 son dos variables aleatorias independientes con $X_1 \rightsquigarrow \chi^2_{n_1}$ y $X_2 \rightsquigarrow \chi^2_{n_2}$. Entonces diremos que la variable aleatoria definida como $\frac{X_1/n_1}{X_2/n_2}$ tiene una **distribución F de Fisher con** n_1 y n_2 **grados de libertad** respectivamente, y lo representaremos por $\frac{X_1/n_1}{X_2/n_2} \rightsquigarrow F_{n_1,n_2}$. Por tanto podemos decir que $F_{n_1,n_2} \equiv \frac{\chi^2_{n_1}/n_1}{\chi^2_{n_2}/n_2}$ siempre que ambas distribuciones ji-cuadrado sean independientes. Aunque no estudiaremos la función de densidad de esta variable aleatoria, representamos a continuación su forma para diferentes valores de los parámetros n_1 y n_2. Lo haremos primero para $n_1 = 10$ y diferentes valores del parámetro n_2 y después para $n_2 = 10$ y diferentes valores del parámetro n_1.





Podemos observar entonces que en todos los casos la distribución F de Fisher es asimétrica positiva, pero la asimetría disminuye cuando ambos parámetros toman valores cada vez mayores. Además, para esta variable aleatoria se verifica que, si $n_2 > 2$, entonces $E(F_{n_1,n_2}) = \frac{n_1}{n_2-2}$ y, si $n_2 > 4$, entonces $Var(F_{n_1,n_2}) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$.

Las probabilidades de esta distribución se encuentran tabuladas para diferentes valores de los parámetros n_1 y n_2, fijado un valor $\alpha \in (0,1)$, de modo que $F_{n_1,n_2;\alpha}$ representa al punto que, para una distribución F de Fisher con n_1 y n_2 grados de libertad, deja una probabilidad a su derecha igual a α, es decir, $p\left(F_{n_1,n_2} \geq F_{n_1,n_2;\alpha}\right) = \alpha$. A la hora de utilizar estos valores tabulados es importante tener en cuenta la siguiente propiedad de esta distribución: si $X \rightsquigarrow F_{n_1,n_2}$ entonces $\frac{1}{X} \rightsquigarrow F_{n_2,n_1}$. Por tanto, podremos asegurar que $F_{n_1,n_2;1-\alpha} = \frac{1}{F_{n_2,n_1;\alpha}}$, como demostramos a continuación:

$$\begin{aligned} p\left(F_{n_1,n_2} \geq \frac{1}{F_{n_2,n_1;\alpha}}\right) &= 1 - p\left(F_{n_1,n_2} \leq \frac{1}{F_{n_2,n_1;\alpha}}\right) = 1 - p\left(\frac{1}{F_{n_1,n_2}} \geq F_{n_2,n_1;\alpha}\right) \\ &= 1 - p\left(F_{n_2,n_1} \geq F_{n_2,n_1;\alpha}\right) = 1 - \alpha \end{aligned}$$

Otra propiedad de esta distribución es la siguiente: si $X \rightsquigarrow t_n$ entonces $X^2 \rightsquigarrow F_{1,n}$, de modo que puede decirse $t_n^2 \equiv F_{1,n}$.

Ejercicio. Encontrar los valores de $F_{10,15;0.05}$, $F_{10,15;0.95}$, $F_{10,10;0.01}$ y $F_{10,10;0.99}$. A partir de ellos, calcular $p\left(F_{10,15;0.95} \leq F_{10,15} \leq F_{10,15;0.05}\right)$ y $p\left(F_{10,10;0.99} \leq F_{10,10} \leq F_{10,10;0.01}\right)$.

6.7. PROBLEMAS

1. Supóngase que la concentración de cierto contaminante se encuentra uniformemente distribuida en el intervalo 4 a 20 ppm. Si se considera como tóxica una concentración de 15 ppm. o más. ¿cuál es la probabilidad de que al tomarse una muestra la concentración de ésta sea tóxica?

2. De una estación parte un tren cada 20 minutos. Un pasajero llega de improviso. Hallar:

   (a) La función de distribución y de densidad de la variable tiempo de espera.

   (b) Probabilidad de que el pasajero tenga que esperar menos de 7 minutos.

   (e) La esperanza y la desviación típica de la variable aleatoria tiempo de espera.

3. En un estudio sobre la conducta animal, fueron liberados pájaros, de uno en uno, con determinadas condiciones que hacían muy difícil la orientación. Se vio que los pájaros elegían al azar la dirección en la que comenzaban a volar. La dirección puede definirse mediante el acimut α, es decir, mediante el ángulo formado por el Norte y la dirección elegida por el pájaro. Obtener la función de densidad y la función de distribución de la dirección seguida por el pájaro al ser liberado.

4. Se ha comprobado que el tiempo de vida de unas determinadas bacterias sigue una distribución exponencial, siendo el tiempo de vida medio de una bacteria de 225 días.

   (a) Calcular la probabilidad de que una bacteria viva más de 8 meses.

   (b) Obtener el 95-percentil de la distribución.

   (c) Calcular la probabilidad de que una bacteria que ha vivido al menos 9 meses viva como poco 3 meses más.

5. La concentración de monóxido de carbono durante una hora en una ciudad grande tiene una distribución exponencial cuya media es 3.6 ppm (partes por millón).

   (a) Calcular la probabilidad de que la concentración sea mayor que 9 ppm.

   (b) Una estrategia de control del tránsito de vehículos redujo el promedio a 2.5 ppm. Estimar en este caso la probabilidad de que la concentración sea mayor que 9 ppm.

6. Los totales semanales de precipitación para una determinada zona del medio oeste de los Estados Unidos sigue una distribución exponencial con promedio de 40 mm.

   (a) Hallar la probabilidad de que un total semanal de lluvia en esta zona sea mayor que 50 mm.

   (b) Calcular la probabilidad de que los totales semanales no sean mayores que 50 mm.

7. (a) Supongamos que el número de sucesos que tiene lugar a lo largo del tiempo constituye un proceso de Poisson con un número esperado de λ. ocurrencias por unidad de tiempo. Encontrar la distribución de la variable $T =$"tiempo transcurrido hasla la ocurrencia del primer suceso". (Sugerencia: nótese que $p[T > t]$ es igual a la probabilidad de que en el intervalo $[0, t]$ no haya ocurrido ningún suceso.)

   (b) El número esperado de llamadas telefónicas que entran en una centralita es de 4 por minuto. ¿Cuál es la probabilidad de que pase un minuto sin recibir una llamada telefónica?

8. Se ha medido el diámetro normal de los 9708 árboles de una hectárea de un bosque encontrándose los siguiente resultados:

Clase diamétrica (cm)	$5-15$	$15-25$	$25-35$	$35-45$	$45-55$	$55-65$
Frecuencia observada (n_i)	4966	2466	1208	616	305	147

   Tratar de ajustar los datos a una distribución exponencial.

9. Si Z es una variable aleatoria que tiene distribución normal estándar, encontrar la probabilidad de que Z asuma un valor:

   (a) mayor que 1.14 (b) mayor que -0.36 (c) menor que 2.13

   (d) menor que -1.23 (e) entre -0.46 y -0.09 (f) entre -0.58 y 1.12

10. Si $Z \rightsquigarrow N(0,1)$, encontrar el valor de z_i tal que:

    (a) $p(0 < Z < z_1) = 0.4306$ (b) $p(Z \geq z_2) = 0.7704$

    (c) $p(Z > z_3) = 0.2912$ (d) $p(-z_4 < Z < z_4) = 0.9700$

11. Si X es una variable aleatoria con distribución normal de media 10 y desviación típica de 2, calcular la probabilidad de que X tome un valor:

    (a) menor que 13 (b) mayor que 9 (c) entre 6 y 14 (d) entre 2 y 4

12. Sea X una variable aleatoria con distribución normal de media 5 y varianza 16. Calcular el valor de x_i que hace que:

    (a) $p(X > x_1) = 0.1075$ (b) $p(X > x_2) = 0.95$ (c) $p(5 - x_3 < X < 5 + x_3) = 0.7198$

13. En cierta comunidad de arenques comunes adultos (*Clupea harengus L.*) la longitud de los peces sigue una distribución normal con media 25 cm y desviación típica 4.5 cm.

    (a) ¿Qué porcentaje de los peces tiene una longitud comprendida entre 25 y 30 cm?

    (b) ¿Qué porcentaje de los peces tiene una longitud superior a 18 cm?

    (c) ¿Qué porcentaje de los peces tiene una longitud comprendida entre 32 y 36 cm?

    (d) Calcular los percentiles 20 y 70 de la población.

14. Un botánico ha observado que la anchura de las hojas de un álamo sigue una distribución normal con una media de 6 cm, y, además, que el 90 % de las hojas tiene una anchura inferior a 7.5 cm. Obtener una estimación de σ y calcular la probabilidad de que una hoja mida más de 8 cm.

15. La anchura, en milímetros, de los coleópteros de una población sigue una distribución normal. Se estima que el 77 % de la población mide menos de 12 mm y que el 84 % mide más de 7 mm. Dar una estimación de la anchura media y de la desviación típica de la población.

16. Sean X e Y variables aleatorias normales e independientes. La variable X tiene una media igual a 5 y la varianza igual a 4. La esperanza de la variable Y es de 10 y su varianza de 9. Calcular:

    (a) $p(3X + 2Y > 25)$ (b) $p(30 \leq 3X + 2Y \leq 50)$

17. Sean $X_1, X_2, ...X_7$, variables aleatorias independientes tales que $X_i \rightsquigarrow N(3,4)$.

    (a) Obtener la distribución de la variable $Y = 7X_1$.

    (b) Obtener la distribución de la variable $W = \sum_{i=1}^{7} X_i$.

18. El peso de las personas de una población sigue una distribución normal con media 72 kg y desviación típica 10 kg.

    (a) Cuatro personas elegidas al azar en esa población entran en un ascensor cuya carga máxima es de 350 kg. ¿Cuál es la probabilidad de que entre los cuatro se supere esta carga máxima?

(b) ¿Cuál es la probabilidad de que dos personas elegidas al azar en esa población puedan jugar en un balancín si sólo pueden hacerlo cuando sus pesos difieran en menos de 5 kg?

19. El peso de una gacela, en kilogramos, es una variable aleatoria que tiene una distribución normal con media 50 kg y desviación típica 6 kg. Si capturamos 10 gacelas:

    (a) Obtener la distribución de la variable aleatoria $Y =$"*número de gacelas, de las 10 capturadas, que pesan menos de* 40 *kg*". Calcular $p\,(Y = 2)$.

    (b) ¿Cuál es la probabilidad de que podamos transportar a las 10 gacelas en un vehículo que admite una carga máxima de 450 kg? ¿Y si admite una carga de 525 kg?

20. Una línea eléctrica se avería cuando la tensión sobrepasa la capacidad de la línea. Si la tensión es una variable aleatoria con distribución $N(100, 20)$ y la capacidad sigue una distribución $N(130, 10)$, calcular la probabilidad de que se produzca una avería.

21. Una máquina de empaquetado automático deposita en cada paquete 81.5 gr, por término medio, de cierto producto, con $\sigma = 8$ gr. El peso medio del paquete vacío es 14.5 gr con $\sigma = 6$ gr. Ambas distribuciones son normales e independientes.

    (a) Obtener la distribución del peso de los paquetes llenos.

    (b) Distribuimos los paquetes llenos, de 40 en 40, en cajas cuyo peso medio, vacías, es de 520 gr con $\sigma = 50$ gr (admitimos también normalidad). Hallar la distribución del peso de las cajas llenas medido en kg.

22. **Distribución de Weibull**: Se dice que una variable aleatoria X sigue una distribución de Weibull de parámetros a y b ($a > 0$ y $b > 0$) y lo denotaremos diciendo que $X \rightsquigarrow W(a, b)$, si su función de distribución viene dada por: $F(x) = \begin{cases} 0 & \text{si} \quad x < 0 \\ 1 - e^{-ax^b} & \text{si} \quad x \geq 0 \end{cases}$

    La distribución de Weibull aparece a menudo para modelizar la duración (se utiliza para caracterizar el tiempo de vida de sistemas con la particularidad de tener una tasa de fallo que crece, o decrece, de forma potencial a lo largo del tiempo). Debido al gran número de formas que adquieren las funciones de densidad de Weibull, según el valor de sus dos parámetros, la distribución de Weibull se utiliza en numerosas aplicaciones.(Nota, el programa Statgraphics utiliza $F(x) = 1 - e^{-\left(\frac{x}{a}\right)^b}$).

    (a) Demostrar que $F(x)$ es una función de distribución.

    (b) Obtener la función de densidad de una variable de Weibull.

23. Los niveles máximo de inundación, en millones de pies cúbicos por segundo, de determinado río siguen una distribución de probabilidad de Weibull con parámetros $a = 1/0.6$ y $b = 1.5$. Estimar la probabilidad de que el nivel máximo de inundación para el año próximo sea:

    (a) mayor que 500000 pies cúbicos por segundo.

    (b) menor que 800000 pies cúbicos por segundo.

24. La resistencia máxima a la tensión del alambre de acero que se usa para envolver a los tubos de concreto tuvo una distribución de Weibull con $a = 1/270$ y $b = 1.2$ (mediciones hechas en miles de

libras). En un determinado punto de un tubo, la presión existente necesita una resistencia máxima a la tensión de cuanto menos 300000 libras. ¿Cuál es la probabilidad de que el alambre tenga esta resistencia necesaria?

25. A partir de los datos de la velocidad media (medida en km/h) diaria del viento, medidos en el observatorio meteorológico de La Coruña durante varios años, se ha ajustado una ley de Weibull con parámetros $a = 1/117$ y $b = 1.4$.
    (a) ¿Qué velocidad media del viento se espera que se supere el 50 % de los días?
    (b) Calcular el porcentaje de días que se espera superen una velocidad media del viento de 80 km/h.

26. Se supone que las plantas de una determinada especie se distribuyen aleatoriamente en una región. con una densidad promedio de λ plantas por unidad de área. Para una planta seleccionada al azar en la zona, sea R la distancia a la planta vecina más próxima de la misma especie. Determinar la función de densidad de la variable R. (Sugerencia: nótese que $p\,(R > r)$ es igual a la probabilidad de que en el círculo de radio r no haya ninguna planta de la especie examinada).

27. A partir de una muestra de los diámetros de los árboles de un bosque (medidos en cm.), se ha ajustado satisfactoriamente la variable diámetro de un árbol a una Weibull de parámetros $a = 1.5581 \cdot 10^{-10}$ y $b = 5.45$. Si observamos una muestra de 1000 árboles, ¿qué frecuencias podríamos esperar bajo la distribución ajustada? Formar una tabla de frecuencias esperadas para clases diamétricas de 10 cm de amplitud.

SOLUCIONES

1. 0.3125

2. (a) $f(x) = \begin{cases} \frac{1}{20} & \text{si } x \in [0,20] \\ 0 & \text{en el resto} \end{cases}$ (b) $F(x) = \begin{cases} 0 & \text{si} \quad x < 0 \\ \frac{x}{20} & \text{si} \quad x \in [0,20] \\ 1 & \text{si} \quad x > 20 \end{cases}$ (c) $E(X) = 10$, $\sigma = 5.7735$

3. Si medimos los ángulos en grados y en el sentido de las agujas del reloj, tomando la dirección norte como origen de ángulos, entonces:

$$f(x) = \begin{cases} \frac{1}{360} & \text{si } x \in [0,360] \\ 0 & \text{en el resto} \end{cases} \qquad F(x) = \begin{cases} 0 & \text{si} \quad x < 0 \\ \frac{x}{360} & \text{si} \quad x \in [0,360] \\ 1 & \text{si} \quad x > 360 \end{cases}$$

4. (a) 0.3442 (b) 674 días (22.5 meses) (c) 0.6703

5. (a) 0.0821 (b) 0.0273

6. (a) 0.2865 (b) 0.5091

7. (a) $T \rightsquigarrow \exp(\lambda)$ (b) 0.0183

8. Definimos $X = D - 5$ y suponemos $X \rightsquigarrow \exp(\lambda)$. Teniendo en cuenta que $\bar{D} = 18.95$, deberíamos suponer $E(X) = \frac{1}{\lambda} = 13.95$ y por tanto $\lambda = \frac{1}{13.95}$. Usando esta distribución tendríamos:

Clase diamétrica (cm)	$5-15$	$15-25$	$25-35$	$35-45$	$45-55$	$55-65$
Frecuencia observada (n_i)	4966	2466	1208	616	305	147
$p_i = p\,(d_i - 10 \leq X \leq d_i)$	0.5118	0.2499	0.1220	0.0596	0.0291	0.0142
Frecuencia teórica (np_i)	4969	2426	1184	578	282	138

9. (a) 0.1271 (b) 0.6406 (c) 0.9834
   (d) 0.1093 (e) 0.1413 (f) 0.5876

10. (a) $z_1 = 1.48$ (b) $z_2 = -0.74$ (c) $z_3 = 0.55$ (d) $z_4 = 2.17$

11. (a) 0.9332 (b) 0.6915 (c) 0.9544 (d) 0.0013183

12. (a) $x_1 = 9.96$ (b) $x_2 = -1.58$ (c) $x_3 = 4.32$

13. (a) 0.3665 (b) 0.9406 (c) 0.0521 (d) $P_{70} = 27.34$, $P_{20} = 21.22$

14. $\sigma = 1.170$ cm., $p\,(X > 8) = 0.0436$

15. $\mu = 9.8691$ mm., $\sigma = 2.8847$ mm.

16. (a) 0.8810 (b) 0.6840

17. (a) $Y \rightsquigarrow N\,(21, 28)$ (b) $W \rightsquigarrow N\left(21, 4\sqrt{7}\right)$

18. (a) 0.000968 (b) 0.2736

19. (a) $Y \rightsquigarrow B\,(10, 0.0475)$, $p\,(Y = 2) = 0.0688$ (b) 0.00415 y 0.9049

20. 0.0901

21. (a) $N\,(96, 10)$ (b) $N\,(4.36, 0.0806)$

22. (b) $f(x) = \begin{cases} 0 & \text{si } x < 0 \\ abx^{b-1}e^{-ax^b} & \text{si } x \geq 0 \end{cases}$

23. (a) 0.554745 (b) 0.6966

24. 0.0309

25. (a)$Me = 23.1$ (b) 0.0193

26. $R \rightsquigarrow W\,(\pi\lambda, 2)$

27.

Clase diamétrica (cm)	$0-10$	$10-20$	$20-30$	$30-40$	$40-50$
Frecuencia esperada	0	2	15	63	166

Clase diamétrica (cm)	$50-60$	$60-70$	$70-80$	$80-90$	90 o más
Frecuencia esperada	288	295	145	25	1

Capítulo 7

TEOREMAS LÍMITE EN PROBABILIDAD. APROXIMACIONES

En este capítulo estudiaremos una serie de resultados que serán ampliamente utilizados en el Bloque III de esta asignatura, dedicado a la Inferencia Estadística, y en el que trabajaremos con observaciones independientes de una variable aleatoria para tratar de inferir cuál es su distribución o adivinar el valor de su parámetros característicos. Para ello empezamos introduciendo la siguiente definición:

Dada una variable aleatoria X se define una **muestra aleatoria simple de** X (abreviadamente **m.a.s**.) como una sucesión de variables aleatorias $X_1, X_2, ..., X_n$ independientes e igualmente distribuidas (**i.i.d.**) que X. Es decir, todas las variables X_i tienen la misma distribución de probabilidad que la variable original X y además son independientes entre sí.

La distribución de probabilidad normal que hemos estudiado en el capítulo anterior es especialmente importante cuando trabajamos con este tipo de muestras, como veremos a continuación.

Al igual que hicimos en Estadística Descriptiva trabajando con números, dada una m.a.s $X_1, X_2, ..., X_n$ de X podemos definir la media muestral de dicha muestra como una nueva variable aleatoria construida como $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Por tanto se hace necesario ahora distinguir entre la media muestral como variable aleatoria ($\overline{X}$) y el valor obtenido para esta variable aleatoria cuando se hayan observado unos valores concretos $x_1, x_2, ..., x_n$ para las variables aleatorias que componen la muestra, que será la media muestral observada de la que hablábamos en Estadística Descriptiva, y que ahora representaremos por $\bar{x}$.

Los dos resultados que enunciaremos a continuación nos permitirán deducir importantes propiedades para $\overline{X}$.

Teorema Central del Límite. Si $X_1, X_2, ..., X_n$ son variables aleatorias independientes con cua-

lesquiera distribuciones se verifica que $\sum_{i=1}^{n} X_i$ converge a una distribución de probabilidad normal cuando n tiende hacia infinito. Es decir, que para n grande podemos asegurar que $\sum_{i=1}^{n} X_i \overset{n\to\infty}{\rightsquigarrow} Normal$, sean cuales sean las distribuciones de probabilidad de las variables $X_1, X_2, ..., X_n$. Además, si $\mu_i = EX_i$ y $\sigma_i^2 = Var(X_i)$, por las propiedades de esperanza y varianza y teniendo en cuenta la independencia de las variables aleatorias, se verifica que $E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mu_i$ y $Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \sigma_i^2$. Por tanto podemos asegurar que $\sum_{i=1}^{n} X_i \overset{n\to\infty}{\rightsquigarrow} N\left(\sum_{i=1}^{n} \mu_i, \sqrt{\sum_{i=1}^{n} \sigma_i^2}\right)$.

En particular, si $X_1, X_2, ..., X_n$ es una m.a.s. de cualquier variable aleatoria X con $\mu = EX$ y $\sigma^2 = Var(X)$, se verifica que $\sum_{i=1}^{n} X_i \overset{n\to\infty}{\rightsquigarrow} N(n\mu, \sigma\sqrt{n})$, y de ello se deduce que la media muestral $\overline{X}$ verifica que $\overline{X} \overset{n\to\infty}{\rightsquigarrow} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Ley de los Grandes Números de Kolmogorov. Si $X_1, X_2, ..., X_n$ es una m.a.s. de cualquier variable aleatoria X con $EX = \mu$ y $\overline{X}_n$ es la media muestral se verifica que, para cualquier $\varepsilon > 0$, $\lim_{n\to\infty} p\left(\left|\overline{X}_n - \mu\right| < \varepsilon\right) = 1$.

Es decir, para cualquier variable aleatoria X con $EX = \mu$, la probabilidad de que la media muestral $(\overline{X})$ diste de la media teórica (μ) menos de una cierta cantidad (ε) infinitamente pequeña converge hacia 1 cuando el tamaño muestral (n) converge hacia infinito. Este resultado nos permite por tanto asegurar que, si el tamaño muestral n es suficientemente grande, la media muestral $(\overline{X})$ estará con gran probabilidad muy cerca de la media teórica (μ).

El Teorema Central del Límite nos va a permitir utilizar la distribución normal para aproximar dos distribuciones de probabilidad discretas: la distribución binomial y la distribución de Poisson. La primera de ellas porque, como ya sabemos, es suma de distribuciones de Bernouilli y la segunda porque se obtiene como límite de la distribución binomial cuando n tiende a infinito. A continuación estudiamos estas dos aproximaciones.

Aproximación de la distribución binomial por la distribución normal. Supongamos que $X \rightsquigarrow B(n,p)$ con $n > 30$ y $0.1 \leq p \leq 0.9$. Entonces se verifica que $X \widetilde{\rightsquigarrow} N\left(np, \sqrt{np(1-p)}\right)$. De este modo, para calcular probabilidades en una distribución binomial con n grande, podemos utilizar la distribución normal. También puede utilizarse esta aproximación cuando $p < 0.1$ o $p > 0.9$, siempre que $np > 5$ y $n(1-p) > 5$. No obstante, al estar aproximando una variable discreta por una variable continua, se hace necesario aplicar una **corrección por continuidad** cuando utilicemos esta aproximación, considerando que $p(X = x)$ para una distribución binomial debe ser aproximado por $p(x - 0.5 \leq X \leq x + 0.5)$ cuando utilicemos la correspondiente distribución normal. Del mismo modo, aplicando esta corrección por continuidad en la aproximación, tendremos $p(X \leq x) \simeq p(X \leq x + 0.5)$, $p(X < x) \simeq p(X \leq x - 0.5)$, $p(X \geq x) \simeq p(X \geq x - 0.5)$ y $p(X > x) \simeq p(X \geq x + 0.5)$.

Ejercicio. Tiramos 400 veces una moneda equilibrada. Hallar la probabilidad de que el número de caras esté comprendido entre 160 y 190. Obtener un intervalo centrado en 200, tal que la probabilidad de que el número de caras esté en dicho intervalo sea de 0.95.

Aproximación de la distribución de Poisson por la distribución normal. Supongamos que $X \rightsquigarrow P(\lambda)$ con $\lambda > 20$. Entonces se verifica que $X \widetilde{\rightsquigarrow} N\left(\lambda, \sqrt{\lambda}\right)$. De este modo, para calcular probabilidades en una distribución de Poisson con $\lambda > 20$, podemos utilizar la distribución normal. Al igual que en el caso anterior, también ahora es necesario aplicar una **corrección por continuidad** cuando utilicemos esta aproximación. Además, si recordamos la aproximación de la distribución binomial por la distribución de Poisson que ya estudiamos, podemos asegurar que, si $X \rightsquigarrow B(n,p)$ con n grande ($n > 30$) y p pequeño ($p < 0.1$, o $p > 0.9$ intercambiando éxitos con fracasos), pero $\lambda = np > 5$, también podemos aproximar esta distribución por la distribución normal con corrección por continuidad. En efecto, aplicando una doble aproximación tendremos: $X \rightsquigarrow B(n,p) \widetilde{\rightsquigarrow} P(np) \widetilde{\rightsquigarrow} N\left(np, \sqrt{np}\right)$.

Ejercicio. En un proceso de fabricación de papel aparece, por término medio, un defecto cada 20 metros de papel. Si se puede considerar que el número de defectos sigue una distribución de Poisson, calcular la probabilidad de que haya 30 o más defectos en un rollo de 500 metros de papel. Encontrar un intervalo, centrado en el valor medio, de modo que la probabilidad de que el número de defectos caiga en ese intervalo sea de 0.98.

Terminaremos este capítulo mostrando cómo las distribuciones χ^2 de Pearson, t de Student y F de Fisher que ya hemos estudiado aparecen cuando trabajamos con muestras aleatorias simples de variables aleatorias con distribución normal. El resultado fundamental en el que se basan todos los demás es el Teorema de Fisher que enunciamos a continuación.

Teorema de Fisher. Supongamos que $X_1, X_2, ..., X_n$ es una muestra aleatoria simple (**m.a.s.**) de tamaño n de una variable $X \rightsquigarrow N(\mu, \sigma)$. Sean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ y $S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ la media muestral y la varianza muestral corregida, respectivamente. Entonces se verifica que:

1. $\overline{X} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.
2. $\frac{(n-1)S_c^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$.
3. $\overline{X}$ y S_c^2 son independientes.

A partir de este resultado, y utilizando la propiedad de aditividad de la distribución χ^2 de Pearson con respecto a su parámetro, podemos obtener otro resultado importante para el caso de que estemos trabajando con dos m.a.s. independientes de dos distribuciones normales. En concreto, supongamos que $X_1, X_2, ..., X_{n_1}$ es una m.a.s. de $X \rightsquigarrow N(\mu_1, \sigma)$, $Y_1, Y_2, ..., Y_{n_2}$ es una m.a.s. de $Y \rightsquigarrow N(\mu_2, \sigma)$ y ambas muestras son independientes entre sí (observar que hemos supuesto que las dos variables aleatorias tienen la misma varianza). Sean $\overline{X}$, $\overline{Y}$ las dos medias muestrales respectivas, y S_1^2, S_2^2 las dos varianzas muestrales corregidas, respectivamente. Consideramos entonces la variable $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$, conocida como varianza muestral corregida promedio de ambas muestras. Entonces se verifica que

$$\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2} = \frac{(n_1+n_2-2)\,S_p^2}{\sigma^2} \rightsquigarrow \chi^2_{n_1+n_2-2}$$

y además S_p^2 y $\overline{X} - \overline{Y}$ son independientes.

Estos dos últimos resultados muestran la importancia de la distribución χ^2 de Pearson en el muestreo de variables aleatorias con distribución normal.

Ejercicio. Supongamos que $X \rightsquigarrow N(\mu, \sigma)$ con μ desconocido, pero con $\sigma = 2$. Utilizando una muestra de tamaño $n = 20$, ¿cuál es la probabilidad de que la media muestral diste de la media teórica, μ, más de la mitad de la desviación típica?. Si queremos que la media muestral diste de la media teórica menos de 0.1 con una probabilidad de 0.95, ¿qué tamaño muestral deberíamos elegir?. Si ahora suponemos que la desviación típica σ es desconocida y utilizamos también una muestra de tamaño 20, entre qué valores podríamos decir que caerá la varianza muestral corregida con una probabilidad de 0.95. ¿Cuál es la probabilidad de que la varianza muestral corregida supere el doble de la varianza teórica, es decir, $2\sigma^2$?.

Veamos ahora la importancia de la distribución t de Student. Supongamos que $X_1, X_2, ..., X_n$ es una muestra aleatoria simple (**m.a.s.**) de tamaño n de una variable $X \rightsquigarrow N(\mu, \sigma)$. Sean $\overline{X}$ y S_c^2 la media muestral y la varianza muestral corregida, respectivamente. Por el Teorema de Fisher sabemos que: (i) $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0,1)$; (ii) $\frac{(n-1)S_c^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$ y (iii) $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ y $\frac{(n-1)S_c^2}{\sigma^2}$ son independientes. Entonces, teniendo en cuenta que, como ya estudiamos, $t_n \equiv \frac{N(0,1)}{\sqrt{\frac{\chi^2_n}{n}}}$ podemos asegurar que

$$\frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{(n-1)S_c^2}{\sigma^2}}{(n-1)}}} = \frac{\overline{X}-\mu}{S_c/\sqrt{n}} \rightsquigarrow t_{n-1}$$

Del mismo modo, utilizando la propiedad de aditividad de la distribución t de Student con respecto a su parámetro, podemos obtener otro resultado importante para el caso de que estemos trabajando con dos m.a.s. independientes de dos distribuciones normales. En concreto, supongamos que $X_1, X_2, ..., X_{n_1}$ es una m.a.s. de $X \rightsquigarrow N(\mu_1, \sigma)$, $Y_1, Y_2, ..., Y_{n_2}$ es una m.a.s. de $Y \rightsquigarrow N(\mu_2, \sigma)$ y ambas muestras son independientes entre sí (observar que hemos supuesto que las dos variables aleatorias tienen la misma varianza). Sean $\overline{X}$, $\overline{Y}$ las dos medias muestrales respectivas, y S_1^2, S_2^2 las dos varianzas muestrales corregidas, respectivamente. Consideramos además la variable $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ que, como ya sabemos verifica que $\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \rightsquigarrow \chi^2_{n_1+n_2-2}$. Por otro lado, también se verifica que $\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \rightsquigarrow N(0,1)$. Por tanto, utilizando el mismo razonamiento que antes, podemos asegurar que

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \rightsquigarrow t_{n_1+n_2-2}$$

Ejercicio. Si disponemos de dos muestras aleatorias simples independientes entre sí, de tamaños $n_1 = 10$ y $n_2 = 15$ respectivamente, de una misma distribución normal $N(\mu, \sigma)$, ¿cuál es la probabilidad de que la diferencia entre las medias muestrales sea superior a la desviación típica corregida promedio S_p de ambas muestras? ¿Cuál es la probabilidad de que la varianza muestral corregida promedio supere el doble de la varianza teórica, es decir, $2\sigma^2$?

Finalmente, mostramos la importancia de la distribución F de Fisher en el muestreo de variables aleatorias con distribución normal. Consideramos de nuevo que $X_1, X_2, ..., X_{n_1}$ es una m.a.s. de $X \rightsquigarrow N(\mu_1, \sigma_1)$, $Y_1, Y_2, ..., Y_{n_2}$ es una m.a.s. de $Y \rightsquigarrow N(\mu_2, \sigma_2)$ y ambas muestras son independientes

entre sí (observar que ahora no hemos supuesto que las dos variables aleatorias tienen la misma varianza). Sean S_1^2 y S_2^2, respectivamente, las dos varianzas muestrales corregidas. Entonces, por el Teorema de Fisher, sabemos que $\frac{(n_1-1)S_1^2}{\sigma_1^2} \rightsquigarrow \chi^2_{n_1-1}$, $\frac{(n_2-1)S_2^2}{\sigma_2^2} \rightsquigarrow \chi^2_{n_2-1}$ y además ambas variables aleatorias son independientes. Entonces, teniendo en cuenta que, como ya estudiamos, $F_{n_1,n_2} \equiv \frac{\chi^2_{n_1}/n_1}{\chi^2_{n_2}/n_2}$ podemos asegurar que

$$\frac{\frac{(n_1-1)S_1^2}{(n_1-1)\sigma_1^2}}{\frac{(n_2-1)S_2^2}{(n_2-1)\sigma_2^2}} = \frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2} = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \rightsquigarrow F_{n_1-1,n_2-1}$$

Ejercicio. Si disponemos de dos muestras aleatorias simples independientes entre sí, de tamaños $n_1 = 10$ y $n_2 = 15$ respectivamente, de una misma distribución normal $N(\mu, \sigma)$, entre qué valores podríamos decir que caerá el cociente de las varianzas muestrales corregidas con una probabilidad de 0.90. ¿Cuál es la probabilidad de que el cociente entre las varianzas muestrales corregidas sea superior a 2?.

7.1. PROBLEMAS

1. La probabilidad de que un paciente se recupere de cierta enfermedad es 0.8 . Si 50 personas han contraído la enfermedad, ¿cuál es la probabilidad de que se recuperen 40? ¿ Y la probabilidad de que se recuperen más del 80 %?

2. En un proceso de fabricación de papel aparece, por término medio, un defecto cada 20 metros de papel. Si se puede considerar que el número de defectos sigue una distribución de Poisson, calcular la probabilidad de que haya 30 o más defectos en un rollo de 500 metros de papel.

3. Un distribuidor ha determinado a partir de numerosos ensayos que el 5 % de un grupo grande de semillas no germina. Para asegurarse, vende las semillas en paquetes de 200, garantizando al cliente la germinación del 90 %. ¿Cuál es la probabilidad de que un paquete no cumpla la garantía?

4. Tiramos 400 veces una moneda equilibrada.

   (a) Hallar la probabilidad de que el número de caras esté comprendido entre 160 y 190.

   (b) Obtener un intervalo centrado en 200, tal que la probabilidad de que el número de caras esté en dicho intervalo sea de 0.95.

5. Se supone que el número de bacterias por mm^3 de agua en un estanque, es una variable aleatoria X con distribución de Poisson de parámetro $\lambda = 0.13$.

   (a) ¿Cuál es la probabilidad de que en un mm^3 de agua del estanque no haya ninguna bacteria?

   (b) En 40 tubos de ensayo, se toman muestras del agua del estanque (10 mm3 de agua en cada tubo). ¿Qué distribución sigue la variable *número de tubos de ensayo entre los* 40 *que no contienen bacterias*? Calcular la probabilidad de que haya el 30 % o más de tubos en los que no hay bacterias.

   (c) De los tubos que contienen bacterias, ¿qué porcentaje contendrá al menos 4 bacterias?

6. Sean $X_1, X_2, ..., X_{100}$ variables aleatorias independientes tales que $E(X_i) = 2$ y $Var(X_i) = 9$. Calcular, aproximadamente, $p\left(225 \leq \sum_{i=1}^{100} X_i \leq 250\right)$.
7. En una asignatura del colegio la probabilidad de que te saquen a la pizarra en cada clase es del 10 %. A lo largo del año tienes 100 clases de esa asignatura. ¿Cuál es la probabilidad de tener que salir a la pizarra más de 15 veces?
8. Se considera que el número diario de ejemplares vendidos de una determinada publicación en un quiosco es una variable aleatoria que se distribuye según una Poisson de media 30. El precio de venta de cada ejemplar es de $1,3$€ euros. Obtener, razonadamente, la probabilidad de que en los próximos 40 días los ingresos por la venta de ejemplares superen los 1625€.
9. El dinero que se gastan los adolescentes de entre 16 y 18 años durante un fin de semana sigue una distribución desconocida de media 6.20€ y desviación típica $1,90$€. Se toma al azar una muestra de 60 de esos adolescentes.
   (a) ¿Qué distribución sigue la media del gasto en dicha muestra?
   (b) ¿Cuál es la probabilidad de que la media del gasto en esa muestra sea superior a 7€?
   (c) ¿Qué hubiese ocurrido si tomamos una muestra de solo 28 estudiantes?
10. Ante una medida adoptada por el Gobierno, se sabe que el 35 % de la población está a favor de dicha medida. Se toma una muestra al azar de 200 personas.
    (a) ¿Qué número de personas se espera que estén a favor?
    (b) Halla la distribución en el muestreo de la proporción de personas que están a favor de esta medida.
    (c) Halla la probabilidad de que en la muestra elegida, más del 40 % de los integrantes que la forman estén a favor.

SOLUCIONES

1. $p(X = 40) \simeq 0.1503$, $p(X > 40) \simeq 0.4298$
2. 0.1841
3. Aproximadamente, 3 paquetes de cada 10000 no cumplirán la garantía.
4. (a) 0.1707 (b) $[181, 219]$
5. (a) 0.8781 (b) $B \rightsquigarrow (40.0.2725)$, $p(X \geq 12) = 0.4168$
   (c) Aproximadamente, el 6 % de los tubos que contienen bacterias, contendrán por lo menos 4 bacterias.
6. 0.1558
7. 0.0334
8. 0.0724
9. (a) $N(6.2, 0.25)$ (b) 0.0004
10. (a) 70 personas. (b) $N(0.35, 0.034)$ (c) 0.0707

Parte III

INFERENCIA ESTADÍSTICA

Capítulo 8

ESTIMACIÓN ESTADÍSTICA

Hasta ahora, en la parte anterior de la asignatura, hemos creado modelos probabilísticos para determinados experimentos o variables aleatorias mediante razonamientos deductivos, es decir, se establecen hipótesis sobre el proceso que genera los datos y a partir de ellas se estudia el modelo probabilístico correspondiente, es decir, las probabilidades de los posibles valores que pueden observarse. Sin embargo ahora, por medio de la **inferencia estadística**, lo que pretendemos es, en cierto sentido, darle la vuelta al problema: si se conocen determinados datos experimentales sobre el fenómeno o sobre la variable aleatoria considerada, ¿cuál será el modelo probabilístico que ha generado los datos? Por tanto ahora tendremos que utilizar métodos inductivos para predecir cuál es el modelo probabilístico que subyace en un experimento o en una variable aleatoria.

Podemos definir entonces la **Inferencia** como la parte de la Estadística cuyo objetivo es predecir cuál es el modelo teórico que subyace a un fenómeno aleatorio a partir de cierta información obtenida de la experimentación de dicho fenómeno. Nosotros en esta asignatura sólo estudiaremos una parte de la Inferencia, puesto que estableceremos los siguientes requisitos en el proceso de estimación:

1. Suponemos que el tipo de modelo subyacente (normal, exponencial, Poisson, etc.) es conocido y sólo se trata de adivinar (predecir o **estimar**) los parámetros que caracterizan al mismo (μ y σ^2 en el modelo normal, p en el modelo de Bernouilli, λ en el modelo de Poisson o en el modelo exponencial, etc.). Esta parte de la Inferencia es la que se denomina **Inferencia Paramétrica**.

2. Los parámetros que queremos estimar son constantes fijas. Esta parte de la Inferencia se denomina **Inferencia Clásica**.

3. Suponemos que la información disponible para estimar los parámetros consiste siempre en una muestra aleatoria simple; tal y como ha sido definida en el capítulo anterior. Es decir, estudiaremos **Inferencia con muestreo aleatorio simple**.

Por tanto, este capítulo se centra sólamente en lo que se denomina **Inferencia paramétrica clásica con muestreo aleatorio simple**.

El problema general en este tipo de inferencia podría entonces formularse de la siguiente forma:

Problema 1 *Sea X una variable aleatoria con función de densidad $f(x,\theta)$ o función de probabilidad $p(x,\theta)$ (según se trate de una variable continua o discreta) cuya expresión es conocida salvo el parámetro (o vector de parámetros) θ. Extraemos una muestra aleatoria simple de la variable X: $X_1, X_2, ..., X_n$ independientes e igualmente distribuidas (i.i.d.) que X. A partir de ello pretendemos predecir el valor de θ (problema de **estimación** que abordaremos en este capítulo) o, establecida una hipótesis sobre su posible valor, decidir si ésta es verdadera o falsa (problema de **test de hipótesis** que abordaremos en el próximo capítulo).*

Comenzando con el problema de estimación, ésta podrá hacerse de dos formas: dando un valor único para el parámetro θ (**estimación puntual**) o estableciendo un intervalo entre dos valores que nos permite afirmar, con un cierto grado de confianza fijado previamente, que entre ellos se encuentra el verdadero valor del parámetro (**estimación por intervalos de confianza**). En la práctica, esta última opción resulta más interesante porque parece más útil decir que el parámetro está, por ejemplo, entre 10 y 12 con una confianza del 90 %, que decir que el valor del parámetro es 11 sin dar una medida de la confianza que nos merece esa afirmación. No obstante, la estimación por intervalos se basará siempre en la estimación puntual, puesto que dicho intervalo se construirá, como veremos, a partir de un estimador puntual del parámetro. Por ello comenzaremos estudiando la estimación puntual.

Definición de estadístico. Dada una muestra aleatoria simple $X_1, X_2, ..., X_n$ de una variable aleatoria X, se define un **estadístico** como una función matemática $T(X_1, X_2, ..., X_n)$ de las variables aleatorias que componen la muestra. Por tanto, todo estadístico es también una variable aleatoria que tendrá su función de distribución, su valor esperado, su varianza, etc. No debemos confundir lo que es la muestra aleatoria (como conjunto de variables aleatorias) con los datos obtenidos en la observación de la misma (a los que habitualmente representaremos con letras minúsculas $x_1, x_2, ..., x_n$, como venimos haciendo en estadística descriptiva). Del mismo modo, no debemos confundir un estadístico $T(X_1, X_2, ..., X_n)$, como variable aleatoria que es, con el valor proporcionado por la observación del mismo $T(x_1, x_2, ..., x_n)$, que será un valor numérico concreto como los que estamos obteniendo en estadística descriptiva.

Definición de estimador. Para el problema general de inferencia aquí planteado, dado un parámetro θ, llamaremos **estimador** de θ a un estadístico concreto $T(X_1, X_2, ..., X_n)$ que utilizaremos para estimar dicho parámetro y lo representaremos por $\widehat{\theta} = \widehat{\theta}\,(X_1, X_2, ..., X_n)$. Al valor obtenido para este estimador en una muestra concreta, $\widehat{\theta}\,(x_1, x_2, ..., x_n)$, lo llamaremos **estimación puntual** de θ.

Para la adecuada elección de estimadores, resulta necesario concretar qué propiedades sería coveniente pedir a cualquier estimador. Aunque no haremos un estudio detallado de las mismas, sí comentaremos brevemente las cuatro fundamentales: insesgadez, eficiencia, consistencia y normalidad asintótica.

Propiedades de los estimadores.

1. **Insesgadez.** Diremos que el estimador $\widehat{\theta}$ del parámetro θ es **insesgado** cuando verifica que $E\left(\widehat{\theta}\right) = \theta$, es decir, el valor esperado del estimador coincide con el verdadero valor del parámetro. En caso contrario, diremos que el estimador es **sesgad**o y definimos el **sesgo del estimador** $\widehat{\theta}$ como $sesgo\left(\widehat{\theta}\right) = E\left(\widehat{\theta}\right) - \theta$. Si un estimador es sesgado pero $sesgo\left(\widehat{\theta}\right)$ converge hacia cero cuando el tamaño muestral n tiende a infinito (es decir, el valor $E\left(\widehat{\theta}\right)$ converge hacia θ cuando n tiende a infinito), entonces diremos que el estimador es **asintóticamente insesgado**.

2. **Eficiencia.** Dado un estimador $\widehat{\theta}$ del parámetro θ se define la **eficiencia del estimador** $\widehat{\theta}$ como $eff\left(\widehat{\theta}\right) = \frac{1}{Var\left(\widehat{\theta}\right)}$. Por tanto, dados dos estimadores $\widehat{\theta_1}$ y $\widehat{\theta_2}$ del parámetro θ, diremos que $\widehat{\theta_1}$ es más eficiente que $\widehat{\theta_2}$ cuando $Var\left(\widehat{\theta_1}\right) < Var\left(\widehat{\theta_2}\right)$, y por tanto $eff\left(\widehat{\theta_1}\right) > eff\left(\widehat{\theta_2}\right)$. Es evidente que la eficiencia de un estimador depende del tamaño muestral n y suele ser habitual que un estimador sea más eficiente cuanto más grande sea dicho tamaño muestral. Si para un tamaño muestral dado, n, se verifica que $Var\left(\widehat{\theta_1}\right) < Var\left(\widehat{\theta_2}\right)$ para cualquier otro estimador $\widehat{\theta_2}$, entonces diremos que el estimador $\widehat{\theta_1}$ es el **estimador de mínima varianza** y, por tanto, $\widehat{\theta_1}$ es el estimador más eficiente con ese tamaño muestral.

3. **Consistencia.** Diremos que el estimador $\widehat{\theta}$ del parámetro θ es **consistente** cuando es insesgado (o asintóticamente insesgado) y además $Var\left(\widehat{\theta}\right)$ converge hacia cero cuando el tamaño muestral n tiende a infinito. Por tanto, si un estimador es consistente podemos asegurar que, con tamaños muestrales grandes, proporcionará siempre estimaciones cercanas al verdadero valor del parámetro, porque $E\left(\widehat{\theta}\right) = \theta$ y $Var\left(\widehat{\theta}\right) \simeq 0$.

4. **Normalidad asintótica.** Diremos que el estimador $\widehat{\theta}$ del parámetro θ es **asintóticamente normal** cuando su distribución de probabilidad converge hacia la distribución de probabilidad normal cuando el tamaño muestral n tiende a infinito. Por tanto, si un estimador insesgado $\widehat{\theta}$ (o asintóticamente insesgado), es asintóticamente normal entonces, para tamaños muestrales grandes, podemos asegurar que la distribución de probabilidad del estadístico $\frac{\widehat{\theta}-\theta}{\sqrt{Var\left(\widehat{\theta}\right)}}$ es, aproximadamente, una normal estándar, es decir, $\frac{\widehat{\theta}-\theta}{\sqrt{Var\left(\widehat{\theta}\right)}} \rightsquigarrow N(0,1)$.

Además, podemos definir el **error cuadrático medio de un estimador** $\widehat{\theta}$ como $E\left[\left(\widehat{\theta}-\theta\right)^2\right]$ y tenemos que

$$\begin{aligned} E\left[\left(\widehat{\theta}-\theta\right)^2\right] &= E\left[\widehat{\theta}^2\right] - \left[E\left(\widehat{\theta}\right)\right]^2 + \left[E\left(\widehat{\theta}\right)\right]^2 - 2\theta E\left(\widehat{\theta}\right) + \theta^2 \\ &= Var\left(\widehat{\theta}\right) + \left[sesgo\left(\widehat{\theta}\right)\right]^2 \end{aligned}$$

Entonces, diremos que el estimador $\widehat{\theta}$ es el **estimador mínimo cuadrático** del parámetro θ cuando su error cuadrático medio sea lo más pequeño posible. Teniendo en cuenta la fórmula anterior, siempre que utilicemos estimadores insesgados, podemos asegurar que el estimador de mínima varianza de un parámetro θ resulta ser el estimador mínimo cuadrático. Es decir, el estimador mínimo cuadrático es el más eficiente.

Estimación puntual de $\mu = EX$. Dada una variable aleatoria X cualquiera con $EX = \mu$ y $Var(X) = \sigma^2$, y una muestra aleatoria simple $X_1, X_2, ..., X_n$ de dicha variable, la media muestral $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ es un estimador del parámetro μ, es decir $\hat{\mu} = \overline{X}$. Teniendo en cuenta las propiedades estudiadas hasta ahora sabemos que $E\overline{X} = \mu$, $Var\left(\overline{X}\right) = \frac{\sigma^2}{n}$ y además $\overline{X} \rightsquigarrow N(\mu, \frac{\sigma}{\sqrt{n}})$ cuando el tamaño muestral n tiende a infinito. Por tanto podemos asegurar que la media muestral $\overline{X}$ verifica las propiedades 1, 2 y 4 anteriores como estimador de μ. Además, se puede demostrar que también es el estimador mínimo cuadrático de μ y, por ser insesgado, podemos asegurar que es el de mínima varianza, es decir, el más eficiente. Por tanto, como norma general, utilizaremos la media muestral $\overline{X}$ siempre que queramos estimar el valor esperado μ de una variable aleatoria cualquiera X, es decir, $\hat{\mu} = \overline{X}$.

Estimación puntual de $\sigma^2 = Var(X)$. Dada una variable aleatoria X cualquiera con $EX = \mu$ y $Var(X) = \sigma^2$, y una muestra aleatoria simple $X_1, X_2, ..., X_n$ de dicha variable, parece intuitivo utilizar la varianza muestral $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$ como estimador puntual del parámetro σ^2. Para este estimador se tiene que

$$\begin{aligned} E\left(S^2\right) &= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i^2\right) - E\left(\overline{X}^2\right) = \frac{1}{n}\sum_{i=1}^{n}\left[Var\left(X_i\right) + \left(E\left(X_i\right)\right)^2\right] - \left[Var\left(\overline{X}\right) + \left(E\left(\overline{X}\right)\right)^2\right] \\ &= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2 \end{aligned}$$

y por tanto S^2 no es un estimador insesgado de σ^2, aunque sí es asintóticamente insesgado. Ahora bien, teniendo en cuenta que $S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{n}{n-1}S^2$, podemos asegurar que $E\left(S_c^2\right) = \frac{n}{n-1}E\left(S^2\right) = \sigma^2$ y por tanto la varianza muestral corregida S_c^2 es un estimador insesgado de la varianza teórica σ^2. Sin embargo, se tiene que $Var\left(S_c^2\right) = \left(\frac{n}{n-1}\right)^2 Var\left(S^2\right) > Var\left(S^2\right)$ y por tanto S_c^2, aunque es insesgado, es menos eficiente que S^2. Cuando el tamaño muestral n tiende a infinito se tiene que $S_c^2 = S^2$ y además el Teorema Central del Límite nos permite asegurar que ambos estimadores verifican la propiedad de normalidad asintótica. Aunque no lo demostraremos, ambos estimadores son también consistentes, es decir, su varianza converge hacia cero cuando n tiende a infinito, y podemos asegurar que los dos convergen hacia el verdadero valor σ^2 cuando n tiende hacia infinito. Como norma general, utilizaremos la varianza muestral corregida S_c^2 siempre que queramos estimar la varianza teórica σ^2 de una variable aleatoria cualquiera X, es decir, $\widehat{\sigma^2} = S_c^2$.

Estimación puntual de $\sigma = \sqrt{Var(X)}$. Dada una variable aleatoria X cualquiera con $EX = \mu$ y $Var(X) = \sigma^2$, y una muestra aleatoria simple $X_1, X_2, ..., X_n$ de dicha variable, teniendo en cuenta el razonamiento anterior, parece lógico utilizar la desviación típica muestral corregida $S_c = \sqrt{S_c^2}$ como estimador puntual del parámetro σ. No obstante hay que observar que este estimador no es insesgado, es decir, $E\left(S_c\right) \neq \sigma$, ya que si fuese $E\left(S_c\right) = \sigma$ se tendría que $Var\left(S_c\right) = E\left(S_c^2\right) - \left[E\left(S_c\right)\right]^2 = \sigma^2 - \sigma^2 = 0$ y esto es imposible por ser S_c una variable aleatoria. Por tanto estamos utilizando un estimador que no es insesgado y tampoco es el más eficiente, porque $Var\left(S\right) = Var\left(\sqrt{\frac{n-1}{n}}S_c\right) < Var\left(S_c\right)$. Además, por ser $Var\left(S_c\right) > 0$, podemos asegurar que $\left[E\left(S_c\right)\right]^2 < \sigma^2$ y por tanto $E\left(S_c\right) < \sigma$, es decir, el sesgo es negativo. Del mismo modo, $\left[E\left(S\right)\right]^2 < E\left(S^2\right) = \frac{n-1}{n}\sigma^2 < \sigma^2$ y por tanto $E\left(S\right) < \sigma$, es decir, S también es un estimador sesgado de σ con sesgo negativo. Pero si cuantificamos la magnitud del sesgo de ambos

estimadores en valor absoluto se tiene que

$$|sesgo\,(S_c)| = \sigma - E\,(S_c) = \sigma - \sqrt{\sigma^2 - Var\,(S_c)} > \sigma - \sqrt{\sigma^2 - Var\,(S)} = \sigma - E\,(S) = |sesgo\,(S)|$$

por ser $Var\,(S_c) > Var\,(S)$. Por tanto, la desviación típica muestral $S = \sqrt{S^2}$ es un estimador de la desviación típica teórica σ más eficiente y con menos sesgo que S_c, lo cual nos permite asegurar que también el error cuadrático medio de S es menor que el de S_c. Entonces, como norma general, utilizaremos la desviación típica muestral S como estimador puntual de la desviación típica teórica σ de una variable aleatoria cualquiera X, es decir, $\widehat{\sigma} = S$. A pesar de ello algunos paquetes estadísticos (como Statgraphics que nosotros usamos) utilizan la desviación típica muestral corregida S_c como estimador de σ.

Concepto de intervalo de confianza. Dada una variable aleatoria X cualquiera con parámetro a estimar θ y una muestra aleatoria simple $X_1, X_2, ..., X_n$ de dicha variable, para cada valor fijo $\alpha \in (0,1)$ (generalmente próximo a cero) se buscan dos estadísticos $\widehat{\theta_1}\,(X_1, X_2, ..., X_n)$ y $\widehat{\theta_2}\,(X_1, X_2, ..., X_n)$ de modo que se verifique la siguiente igualdad

$$p\left(\widehat{\theta_1}\,(X_1, X_2, ..., X_n) \leq \theta \leq \widehat{\theta_2}\,(X_1, X_2, ..., X_n)\right) = 1 - \alpha$$

Diremos entondes que $\left(\widehat{\theta_1}\,(X_1, X_2, ..., X_n), \widehat{\theta_2}\,(X_1, X_2, ..., X_n)\right)$ es un intervalo de confianza para el parámetro θ con un grado de confianza $1 - \alpha$. Observar que los extremos de este intervalo son variables aleatorias, y por tanto no debemos interpretar la igualdad probabilística anterior diciendo que la probabilidad de que θ esté entre $\widehat{\theta_1}\,(X_1, X_2, ..., X_n)$ y $\widehat{\theta_2}\,(X_1, X_2, ..., X_n)$ es $1 - \alpha$, sino más bien que la probabilidad de que entre $\widehat{\theta_1}\,(X_1, X_2, ..., X_n)$ y $\widehat{\theta_2}\,(X_1, X_2, ..., X_n)$ esté θ es $1 - \alpha$. Una vez obtenida una muestra y evaluados los extremos del intervalo, el verdadero valor de θ está o no está entre esos valores, y no hay una probabilidad de que esté o no esté porque el verdadero valor del parámetro no es una variable aleatoria, sino que es una constante desconocida pero fija. Por ello es más correcto decir que tenemos una confianza de $1 - \alpha$ en que el verdadero valor del parámetro esté entre los extremos del intervalo.

Construcción de un intervalo de confianza. Los pasos que seguiremos para la construcción de un intervalo de confianza para un parámetro θ con un grado de confianza $1 - \alpha$ serán los siguientes:

1. Elegir un estimador puntual $\widehat{\theta}\,(X_1, X_2, ..., X_n)$ del parámetro θ.

2. Obtener una función del estimador puntual $T\left(\widehat{\theta}\,(X_1, X_2, ..., X_n)\right)$ que dependa sólo de la muestra y del parámetro θ, y cuya distribución de probabilidad sea conocida. Este nuevo estadístico $T\left(\widehat{\theta}\,(X_1, X_2, ..., X_n)\right)$ recibe el nombre de **estadístico pivote**.

3. Obtener el $\alpha/2$-cuantil y el $(1 - \alpha/2)$-cuantil de la distribución de probabilidad del estadístico pivote. Es decir, encontrar dos números que representaremos por $c_{\alpha/2}$ y $c_{1-\alpha/2}$ de modo que se verifique

$$p\left(T\left(\widehat{\theta}\,(X_1, X_2, ..., X_n)\right) \leq c_{\alpha/2}\right) = p\left(T\left(\widehat{\theta}\,(X_1, X_2, ..., X_n)\right) \geq c_{1-\alpha/2}\right) = \frac{\alpha}{2}$$

y por tanto tendremos que $p\left(c_{\alpha/2} \leq T\left(\widehat{\theta}\,(X_1, X_2, ..., X_n)\right) \leq c_{1-\alpha/2}\right) = 1 - \alpha$.

4. Despejar en la doble desigualdad anterior el valor del parámetro θ, de modo que tendremos

$$p\left(\widehat{\theta_1}\left(X_1, X_2, ..., X_n\right) \leq \theta \leq \widehat{\theta_2}\left(X_1, X_2, ..., X_n\right)\right) = 1 - \alpha$$

siendo $\widehat{\theta_1}\left(X_1, X_2, ..., X_n\right)$ y $\widehat{\theta_2}\left(X_1, X_2, ..., X_n\right)$ dos estadísticos que ya no dependerán de θ, sino sólo de la muestra y de los cuantiles $c_{\alpha/2}$ y $c_{1-\alpha/2}$ obtenidos en función del valor fijo del parámetro α.

Es importante observar que, una vez obtenida la muestra y obtenidos unos valores concretos $x_1, x_2, ..., x_n$, no es correcto escribir $p\left(\widehat{\theta_1}\left(x_1, x_2, ..., x_n\right) \leq \theta \leq \widehat{\theta_2}\left(x_1, x_2, ..., x_n\right)\right) = 1 - \alpha$ porque en esta igualdad probabilística no tendríamos ninguna variable aleatoria.

Intervalo de confianza para el parámetro μ de una distribución normal con σ conocida. Como primer ejemplo de este tipo de estimación vamos a considerar el problema de calcular un intervalo de confianza para el parámetro $\mu = EX$ de una distribución normal con desviación típica conocida σ, fijado un grado de confianza $1 - \alpha$. Si $X_1, X_2, ..., X_n$ representa a la muestra aleatoria simple de la variable aleatoria $X \rightsquigarrow N(\mu, \sigma)$, seleccionaremos el estimador puntual $\hat{\mu} = \overline{X}$, del cual sabemos que $\overline{X} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ y por tanto $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \rightsquigarrow N(0,1)$. Teniendo en cuenta que σ es conocida, el estadístico $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ tiene una distribución de probabilidad conocida y sólo depende de la muestra y del parámetro μ. Por tanto tenemos un estadístico pivote para este problema y utilizando su distribución de probabilidad podemos encontrar los cuantiles $c_{\alpha/2}$ y $c_{1-\alpha/2}$. En este caso concreto, y con la notación que venimos utilizando, tendremos que $c_{1-\alpha/2} = z_{\alpha/2}$ y $c_{\alpha/2} = -z_{\alpha/2}$, de modo que $p\left(-z_{\alpha/2} \leq N(0,1) \leq z_{\alpha/2}\right) = 1 - \alpha$. Por tanto, se tiene que

$$p\left(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

y despejando el parámetro μ en la doble desigualdad se tiene

$$p\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

En consecuencia, el intervalo aleatorio $\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ constituye un intervalo de confianza para el parámetro μ con un grado de confianza $1-\alpha$. Observar que en este caso, debido a la simetría de la distribución $N(0,1)$ y a que el parámetro σ es conocido, los dos estadísticos que definen el intervalo de confianza han sido obtenidos como el estimador puntual $\overline{X}$ mas-menos una constante que depende del tamaño muestral n y el valor α del grado de confianza. Es decir, el intervalo de confianza es $\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Es importante observar que si el grado de confianza aumenta entonces α disminuye y por tanto $z_{\alpha/2}$ será más grande, es decir, el intervalo de confianza será más amplio. Del mismo modo, si el grado de confianza disminuye, $z_{\alpha/2}$ se acercará a cero y el intervalo de confianza se estrecha. En el límite, el radio del intervalo de confianza converge hacia cero cuando el grado de confianza converge hacia cero y converge hacia infinito cuando el grado de confianza converge hacia uno. Además, si el tamaño muestral aumenta es evidente que el intervalo de confianza se estrecha, de modo que cuando n tiende hacia infinito la longitud del intervalo tiende hacia cero. Por tanto, con un tamaño muestral suficientemente grande

podemos conseguir que el intervalo sea tan estrecho como queramos. Concretamente, si queremos asegurar con una confianza de $1-\alpha$ que la media muestral distará de la media teórica, en valor absoluto, menos de una cierta cantidad $e_{\text{máx}}$ bastará elegir el tamaño muestral n de modo que $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq e_{\text{máx}}$, es decir, $n \geq \left(\frac{z_{\alpha/2}\sigma}{e_{\text{máx}}}\right)^2$.

Ejercicio. En un estudio sobre el problema de la lluvia ácida, causada por la reacción de ciertos contaminantes en el aire con el agua de la lluvia, en un determinada región se analizaron 40 muestras de agua de diferentes lluvias con respecto a su pH y se obtuvo una media muestral igual a 3.7 (la lluvia pura que se precipita a través del aire limpio tiene un pH de 5.7). Suponiendo que el pH de lluvia tiene una distribución normal con desviación típica $\sigma = 0.5$, obtener razonadamente un intervalo de confianza del 99 % para el pH medio de la lluvia en esa región e interpretar el resultado. Si queremos obtener una estimación del pH medio con un error menor que 0.1 y una confianza del 99 %, ¿qué tamaño muestral deberíamos utilizar?

Intervalo de confianza para el parámetro μ de una distribución normal con σ desconocida. Supongamos ahora que queremos resolver el mismo problema anterior pero sin conocer el valor de la desviación típica σ. En este caso el estadístico pivote anterior no es adecuado puesto que no sólo depende de la muestra y del parámetro a estimar μ, sino también del valor desconocido σ. Pero en este caso, como ya estudiamos en el capítulo anterior, se sabe que $\dfrac{\overline{X}-\mu}{S_c/\sqrt{n}} \rightsquigarrow t_{n-1}$ y tenemos también un estadístico pivote para este problema. Además, como la distribución t_{n-1} es también simétrica, y con la notación que venimos utilizando, tenemos que $c_{1-\alpha/2} = t_{n-1;\alpha/2}$ y $c_{\alpha/2} = -t_{n-1;\alpha/2}$, de modo que $p\left(-t_{n-1;\alpha/2} \leq t_{n-1} \leq t_{n-1;\alpha/2}\right) = 1-\alpha$. Por tanto, se tiene que

$$p\left(-t_{n-1;\alpha/2} \leq \frac{\overline{X}-\mu}{S_c/\sqrt{n}} \leq t_{n-1;\alpha/2}\right) = 1-\alpha$$

y despejando el parámetro μ en la doble desigualdad se tiene

$$p\left(\overline{X} - t_{n-1;\alpha/2}\frac{S_c}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1;\alpha/2}\frac{S_c}{\sqrt{n}}\right) = 1-\alpha$$

En consecuencia, el intervalo aleatorio $\left(\overline{X} - t_{n-1;\alpha/2}\frac{S_c}{\sqrt{n}}, \overline{X} + t_{n-1;\alpha/2}\frac{S_c}{\sqrt{n}}\right)$ constituye un intervalo de confianza para el parámetro μ con un grado de confianza $1-\alpha$. En este caso el intervalo también se construye sumando y restando una cierta cantidad al estimador puntual $\overline{X}$, pero esta vez esta cantidad no es fija sino aleatoria. Es decir, el intervalo de confianza es $\overline{X} \pm t_{n-1;\alpha/2}\frac{S_c}{\sqrt{n}}$. Teniendo en cuenta que, cuando n tiende hacia infinito, $t_{n-1;\alpha/2}$ converge hacia $z_{\alpha/2}$ y S_c converge hacia σ, podemos asegurar que también en este caso el radio del intervalo converge hacia cero y podemos estimar el parámetro μ con la precisión que queramos y el grado de confianza que queramos, sin más que elegir el tamaño muestral suficientemente grande.

Ejercicio. En el mismo caso del ejercicio anterior, si suponemos ahora que la desviación típica σ es desconocida y hemos obtenido unos valores muestrales $\overline{x} = 3.7$ y $s_c = 0.5$, ¿cuál sería ahora el intervalo de confianza al 99 % para el pH medio de la lluvia en esa región? ¿Podemos ahora asegurar con un 99 % de confianza que existe un problema de lluvia ácida?

Intervalo de confianza para el parámetro σ^2 (o para σ) de una distribución normal. Para este mismo caso de la distribución normal, podemos abordar también el problema de encontrar un intervalo de confianza para la varianza σ^2 de la variable aleatoria con un grado de confianza fijado $1-\alpha$. En efecto, utilizando el apartado (ii) del Teorema de Fisher estudiado en el capítulo anterior, sabemos que $\dfrac{(n-1)S_c^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$ y por tanto disponemos de un estadístico pivote, construido a partir del estimador puntual S_c^2, que sólo depende de la muestra y del parámetro que queremos estimar y cuya distribución de probabilidad es conocida. En este caso los cuantiles que necesitamos son $c_{1-\alpha/2} = \chi^2_{n-1;\alpha/2}$ y $c_{\alpha/2} = \chi^2_{n-1;1-\alpha/2}$, de modo que, $p\left(\chi^2_{n-1} \leq \chi^2_{n-1;1-\alpha/2}\right) = p\left(\chi^2_{n-1} \geq \chi^2_{n-1;\alpha/2}\right) = \frac{\alpha}{2}$. Podemos asegurar entonces que

$$p\left(\chi^2_{n-1;1-\alpha/2} \leq \frac{(n-1)S_c^2}{\sigma^2} \leq \chi^2_{n-1;\alpha/2}\right) = 1-\alpha$$

y despejando el parámetro σ^2 en la doble desigualdad se tiene

$$p\left(\frac{(n-1)S_c^2}{\chi^2_{n-1;\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S_c^2}{\chi^2_{n-1;1-\alpha/2}}\right) = 1-\alpha$$

En consecuencia, el intervalo aleatorio $\left(\dfrac{(n-1)S_c^2}{\chi^2_{n-1;\alpha/2}}, \dfrac{(n-1)S_c^2}{\chi^2_{n-1;1-\alpha/2}}\right)$ constituye un intervalo de confianza para el parámetro σ^2 con un grado de confianza $1-\alpha$. Observar que en este caso la distribución de probabilidad del estadístico pivote χ^2_{n-1} no es simétrica, y el intervalo de confianza no es simétrico respecto al estimador puntual S_c^2, aunque siempre se verifica que el estimador puntual cae dentro del intervalo de confianza.

Veamos ahora que a partir del intervalo anterior podemos obtener también un intervalo de confianza para la desviación típica σ de la distribución normal. En efecto, de modo general, si

$$\left(\widehat{\theta_1}\left(X_1, X_2, ..., X_n\right), \widehat{\theta_2}\left(X_1, X_2, ..., X_n\right)\right)$$

es un intervalo de confianza con grado $1-\alpha$ para el parámetro θ y $h(x)$ es una función estrictamente creciente, podemos asegurar que

$$\widehat{\theta_1}\left(X_1, X_2, ..., X_n\right) \leq \theta \leq \widehat{\theta_2}\left(X_1, X_2, ..., X_n\right) \Longleftrightarrow h\left(\widehat{\theta_1}\left(X_1, X_2, ..., X_n\right)\right) \leq h(\theta) \leq h\left(\widehat{\theta_2}\left(X_1, X_2, ..., X_n\right)\right)$$

y por tanto

$$p\left(h\left(\widehat{\theta_1}\left(X_1, X_2, ..., X_n\right)\right) \leq h(\theta) \leq h\left(\widehat{\theta_2}\left(X_1, X_2, ..., X_n\right)\right)\right) = 1-\alpha$$

Es decir, $\left(h\left(\widehat{\theta_1}\left(X_1, X_2, ..., X_n\right)\right), h\left(\widehat{\theta_2}\left(X_1, X_2, ..., X_n\right)\right)\right)$, es un intervalo de confianza con grado $1-\alpha$ para el parámetro $h(\theta)$. Teniendo en cuenta entonces que $\sigma = \sqrt{\sigma^2}$ y que la función $h(x) = \sqrt{x}$ es estrictamente creciente, podemos asegurar que $\left(\sqrt{\frac{n-1}{\chi^2_{n-1;\alpha/2}}}S_c, \sqrt{\frac{n-1}{\chi^2_{n-1;1-\alpha/2}}}S_c\right)$ es un intervalo de confianza con grado $1-\alpha$ para la desviación típica σ de una variable aleatoria con distribución normal.

Ejercicio. En el mismo caso del ejercicio anterior, ¿cuál sería la estimación puntual de la desviación típica σ del agua de lluvia en la citada región?. Obtener razonadamente un intervalo de confianza al 95 % para la citada desviación típica. ¿Podríamos decir con una confianza del 95 % que la desviación típica del agua de lluvia es inferior a 0.6?

Intervalo de confianza asintótico para el parámetro $\mu = EX$ de una distribución de probabilidad cualquiera. Supongamos ahora que la variable aleatoria X no tiene una distribución normal, sino otra distribución cualquiera con $EX = \mu$ y $Var(X) = \sigma^2$. En este caso el Teorema Central del Límite estudiado en el capítulo anterior nos permite asegurar que $\overline{X} \overset{n\to\infty}{\rightsquigarrow} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ y por tanto $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \widetilde{\rightsquigarrow} N(0,1)$ si el tamaño muestral n es grande (si la distribución de probabilidad de la variable aleatoria X no es muy asimétrica bastaría $n \geq 30$). Si σ^2 fuese conocida tendríamos ya un estadístico pivote con distribución de probabilidad conocida aproximadamente para poder obtener un intervalo de confianza aproximado. Si, como es habitual, σ^2 es desconocida necesitaríamos sustituir el parámetro σ por su estimación puntual S, pero como estamos suponiendo un tamaño muestral grande y el estimador S es, como ya hemos dicho, consistente podríamos asegurar que $\sigma \simeq S \simeq S_c$. Por tanto, también sería cierto que $\dfrac{\overline{X}-\mu}{S_c/\sqrt{n}} \widetilde{\rightsquigarrow} N(0,1)$ y podemos utilizar este estadístico pivote para obtener un intervalo de confianza para el parámetro μ. Concretamente, se tiene que

$$p\left(-z_{\alpha/2} \leq \frac{\overline{X}-\mu}{S_c/\sqrt{n}} \leq z_{\alpha/2}\right) \simeq 1-\alpha$$

y despejando el parámetro μ en la doble desigualdad se tiene

$$p\left(\overline{X} - z_{\alpha/2}\frac{S_c}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{S_c}{\sqrt{n}}\right) \simeq 1-\alpha$$

En consecuencia, el intervalo aleatorio $\left(\overline{X} - z_{\alpha/2}\frac{S_c}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{S_c}{\sqrt{n}}\right)$ constituye un intervalo de confianza aproximado para el parámetro μ con un grado de confianza $1-\alpha$. Es decir, el intervalo de confianza aproximado es $\overline{X} \pm z_{\alpha/2}\frac{S_c}{\sqrt{n}}$. A este tipo de intervalos de confianza aproximados se les conoce con el nombre de intervalos de confianza asintóticos.

Ejercicio. En el mismo caso del ejercicio anterior, si no podemos suponer que el agua de lluvia tiene una distribución normal, encontrar un intervalo de confianza asintótico al 99 % para el pH medio de la lluvia en esa región.

Intervalo de confianza asintótico para una proporción poblacional p. Un caso especialmente interesante en la aplicación de intervalos de confianza asintóticos es el problema de estimación de una proporción dentro de una población donde cada individuo puede clasificarse como éxito o fracaso. Elegido un individuo aleatoriamente en la población, podemos considerar entonces que tenemos una variable aleatoria de Bernouilli con valor 1 si el individuo es un éxito y el valor 0 si es un fracaso. Es decir, $X \rightsquigarrow B(p)$ donde $p = EX$ sería la proporción de éxitos dentro de la población. Además, sabemos que para esta distribución se tiene que $Var(X) = p(1-p)$. Entonces, dada una muestra aleatoria simple $X_1, X_2, ..., X_n$ de la variable X con n grande, se tiene que $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\textit{número de éxitos en la muestra}}{n} = \hat{p}$ (es decir, la proporción muestral de éxitos) y $S_c \simeq S = \sqrt{\hat{p}(1-\hat{p})}$. Por tanto, el intervalo asintótico para el parámetro p con un grado de confianza $1-\alpha$ sería $\hat{p} \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$. Entonces, si queremos asegurar que la proporción muestral diste de la verdadera proporción como mucho una cierta cantidad $e_{\text{máx}}$ con una confianza de $1-\alpha$, debemos elegir n de modo que $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq e_{\text{máx}}$, es decir, $n \geq \hat{p}(1-\hat{p})\left(\dfrac{z_{\alpha/2}}{e_{\text{máx}}}\right)^2$.

Si no se dispone de un estimador $\hat{p}$ a priori, teniendo en cuenta que $\hat{p} \in [0,1]$ y $\hat{p}(1-\hat{p}) \leq 0.25$, bastará elegir $n \geq 0.25 \left(\frac{z_{\alpha/2}}{e_{\text{máx}}} \right)^2$.

Ejercicio. Un periódico señala en su número del 23 de mayo de 1979 que "los estudiantes de derecho se oponen a la pena de muerte". Esta declaración se hizo en base a una encuesta para la cual se escogieron al azar y se entrevistaron 86 estudiantes de derecho. El 52 % de los entrevistados declararon que se oponían a la pena de muerte. A partir de esta información, obtener razonadamente un intervalo de confianza del 95 % para la proporción real de estudiantes de derecho que se oponen a la pena de muerte. ¿Se justifica la afirmación de este periódico? ¿Cuántos estudiantes se tendrán que entrevistar por parte del periódico para estimar la proporción de estudiantes en contra de la pena de muerte con un error máximo del 2 % y una confianza del 95 %?

Intervalo de confianza para la diferencia de medias en dos poblaciones normales con varianzas conocidas. En algunas situaciones prácticas lo que se pretende estimar no es el valor esperado de una variable aleatoria sino la diferencia entre los valores esperados de dos variables aleatorias, generalmente una misma variable medida en dos poblaciones distintas. Considerando en primer lugar el caso más sencillo en el que las dos variables aleatorias tienen distribución normal y las dos desviaciones típicas son conocidas, el problema puede plantearse de la siguiente forma: supongamos que tenemos dos variables aleatorias $X \rightsquigarrow N(\mu_1, \sigma_1)$ e $Y \rightsquigarrow N(\mu_2, \sigma_2)$ con σ_1 y σ_2 conocidas y disponemos de sendas muestras aleatorias simples independientes entre sí, $X_1, X_2, ..., X_{n_1}$ para X e $Y_1, Y_2, ..., Y_{n_2}$ para Y (por tanto las variables X_i son independientes entre sí, las variables Y_j son independientes entre sí y además las variables X_i e Y_j también son independientes). El propósito ahora consiste en estimar la diferencia $\mu_1 - \mu_2$ tanto puntualmente como utilizando un intervalo de confianza con un grado de confianza fijado $1-\alpha$. Teniendo en cuenta que los estimadores puntuales de μ_1 y μ_2 son, respectivamente, $\widehat{\mu_1} = \overline{X}$ y $\widehat{\mu_2} = \overline{Y}$, parece lógico utilizar el estimador puntual $\widehat{\mu_1 - \mu_2} = \overline{X} - \overline{Y}$ puesto que $E\left(\overline{X} - \overline{Y}\right) = E\left(\overline{X}\right) - E\left(\overline{Y}\right) = \mu_1 - \mu_2$. Además, con las hipótesis establecidas, es evidente que $\overline{X} \rightsquigarrow N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right)$, $\overline{Y} \rightsquigarrow N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$ y también $\overline{X}$ e $\overline{Y}$ son independientes. Por tanto, podemos asegurar que $\overline{X} - \overline{Y} \rightsquigarrow N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$ y estandarizando podemos obtener un estadístico pivote con distribución de probabilidad conocida para este problema:

$$\frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightsquigarrow N(0,1)$$

Calculando los cuantiles $c_{1-\alpha/2} = z_{\alpha/2}$ y $c_{\alpha/2} = -z_{\alpha/2}$ se tiene que

$$p\left(-z_{\alpha/2} \leq \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2} \right) = 1 - \alpha$$

y despejando el parámetro que queremos estimar $\mu_1 - \mu_2$ en esa doble desigualdad se tiene que

$$p\left(\left(\overline{X} - \overline{Y}\right) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \left(\overline{X} - \overline{Y}\right) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) = 1 - \alpha$$

y disponemos ya del intervalo de confianza buscado. Es decir un intervalo de confianza para $\mu_1 - \mu_2$ con grado de confianza $1-\alpha$ es $\left(\overline{X}-\overline{Y}\right) \pm z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}$. Una vez extraídas las muestras y evaluado el intervalo, si el valor 0 no cae dentro del intervalo podremos asegurar, con una grado de confianza $1-\alpha$, que los valores μ_1 y μ_2 son distintos. Más concretamente, si $\left(\overline{X}-\overline{Y}\right) - z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}} > 0$ podremos asegurar que $\mu_1 > \mu_2$, y si $\left(\overline{X}-\overline{Y}\right) + z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}} < 0$ podremos asegurar que $\mu_1 < \mu_2$. Tenemos por tanto un procedimiento que nos permite comparar los valores esperados de las dos variables aleatorias con distribución normal suponiendo que las dos varianzas son conocidas y que las dos muestras utilizadas son independientes entre sí.

Si utilizando la misma confianza quisiésemos estimar la diferencia entre μ_1 y μ_2 con un error máximo $e_{\text{máx}}$ el tamaño muestral necesario, si suponemos que ambas muestras son del mismo tamaño, debería verificar $z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2+\sigma_2^2}{n}} \leq e_{\text{máx}}$ y por tanto obtenemos que $n \geq \left(\dfrac{z_{\alpha/2}}{e_{\text{máx}}}\right)^2 \left(\sigma_1^2+\sigma_2^2\right)$.

Ejercicio. La tasa de consumo de oxígeno es una medida importante a la hora de estudiar la actividad fisiológica de los atletas. Se realizó un estudio para intentar observar las diferencias entre el consumo de oxígeno en dos grupos de varones universitarios entrenados con dos métodos diferentes: en el tipo I el entrenamiento se realizaba de forma continua y en el tipo II de forma intermitente. El consumo de oxígeno se mide en mililitros por kilogramo-minuto y se supone que esta variable aleatoria tiene una distribución normal con desviación típica $\sigma_1 = 6$ para el entrenamiento I y $\sigma_2 = 8$ para el entrenamiento II. Se obtuvieron muestras aleatorias independientes de los dos tipos de entrenamiento con tamaños $n_1 = 9$ y $n_2 = 7$ y se observaron las siguientes medias muestrales: $\bar{x} = 43.71$ para la primera e $\bar{y} = 39.63$ para la segunda. Estimar la diferencia entre las medias poblacionales de ambos tipos de entrenamiento con una confianza del 95 %. ¿Podemos asegurar que ambas medias poblacionales son distintas? Si queremos estimar esta diferencia con una confianza del 95 % y un error máximo $e_{\text{máx}} = 4$, ¿qué tamaño muestral deberíamos usar? (el mismo para las dos muestras).

Intervalo de confianza para la diferencia de medias en dos poblaciones normales con varianzas desconocidas pero iguales. Podemos plantearnos el mismo problema anterior suponiendo que las desviaciones típicas son ahora desconocidas pero admitiendo que son iguales, es decir, $\sigma_1 = \sigma_2 = \sigma$ con σ desconocida. En este caso, tal y como estudiamos en el capítulo anterior, se puede estimar σ^2 como $\widehat{\sigma^2} = S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ siendo S_1^2 y S_2^2 las dos varianzas muestrales corregidas de las dos muestras. Además, también sabemos que $\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \rightsquigarrow \chi^2_{n_1+n_2-2}$ y por tanto

$$\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \rightsquigarrow t_{n_1+n_2-2}$$

Por tanto tenemos un estadístico pivote con distribución de probabilidad conocida para este problema. Calculando los cuantiles $c_{1-\alpha/2} = t_{n_1+n_2-2;\alpha/2}$ y $c_{\alpha/2} = -t_{n_1+n_2-2;\alpha/2}$ se tiene que

$$p\left(-t_{n_1+n_2-2;\alpha/2} \leq \frac{\left(\overline{X}-\overline{Y}\right)-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leq t_{n_1+n_2-2;\alpha/2}\right) = 1-\alpha$$

y, despejando el parámetro que queremos estimar $\mu_1 - \mu_2$ en esa doble desigualdad, obtenemos

$$p\left(\left(\overline{X}-\overline{Y}\right) - t_{n_1+n_2-2;\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq \mu_1-\mu_2 \leq \left(\overline{X}-\overline{Y}\right) + t_{n_1+n_2-2;\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right) = 1-\alpha$$

y disponemos ya del intervalo de confianza buscado. Es decir un intervalo de confianza para $\mu_1 - \mu_2$ con grado de confianza $1-\alpha$ suponiendo las varianzas desconocidas pero iguales es

$$\left(\overline{X}-\overline{Y}\right) \pm t_{n_1+n_2-2;\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Ejercicio. Resolver el ejercicio anterior suponiendo ahora que las desviaciones típicas son desconocidas pero iguales y que las desviaciones típicas muestrales de ambas muestras son, respectivamente, $s_1 = 5.88$ y $s_2 = 7.68$ (recordar que teníamos $n_1 = 9$, $\bar{x} = 43.71$, $n_2 = 7$ e $\bar{y} = 39.63$). ¿Podemos decir ahora que ambas medias poblacionales son distintas?

Intervalo de confianza para el cociente de varianzas (o de desviaciones típicas) en dos poblaciones normales. Teniendo en cuenta el intervalo de confianza que acabamos de estudiar surge enseguida la necesidad de decidir si podemos admitir o no que las desviaciones típicas teóricas son iguales. Este problema puede abordarse mediante la estimación con un intervalo de confianza del cociente entre las varianzas teóricas $\dfrac{\sigma_1^2}{\sigma_2^2}$. Parece lógico utilizar como estimador puntual el cociente de las varianzas muestrales corregidas de ambas muestras, es decir, $\widehat{\left(\dfrac{\sigma_1^2}{\sigma_2^2}\right)} = \dfrac{S_1^2}{S_2^2}$. Además, por ser ambas muestras independientes entre sí, sabemos por el capítulo anterior que $\dfrac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \rightsquigarrow F_{n_1-1,n_2-1}$ y por tanto disponemos de un estadístico pivote para este problema. Calculando los cuantiles $c_{1-\alpha/2} = F_{n_1-1,n_2-1;\alpha/2}$ y $c_{\alpha/2} = F_{n_1-1,n_2-1;1-\alpha/2} = \dfrac{1}{F_{n_2-1,n_1-1;\alpha/2}}$ se tiene que

$$p\left(\frac{1}{F_{n_2-1,n_1-1;\alpha/2}} \leq \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \leq F_{n_1-1,n_2-1;\alpha/2}\right) = 1-\alpha$$

y, despejando el parámetro que queremos estimar $\dfrac{\sigma_1^2}{\sigma_2^2}$ en esa doble desigualdad, obtenemos

$$p\left(\frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1;\alpha/2}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \left(S_1^2/S_2^2\right)F_{n_2-1,n_1-1;\alpha/2}\right) = 1-\alpha$$

y disponemos ya del intervalo de confianza buscado. Es decir un intervalo de confianza para $\mu_1 - \mu_2$ con grado de confianza $1-\alpha$ suponiendo las varianzas desconocidas pero iguales es

$$\left(\frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1;\alpha/2}}, \frac{S_1^2}{S_2^2}F_{n_2-1,n_1-1;\alpha/2}\right)$$

Una vez extraídas las muestras y evaluado el intervalo, si el valor 1 no cae dentro del intervalo podremos asegurar, con una grado de confianza $1-\alpha$, que los valores σ_1^2 y σ_2^2 son distintos. Más concretamente, si $\dfrac{S_1^2/S_2^2}{F_{n_1-1,n_2-1;\alpha/2}} > 1$ podremos asegurar que $\sigma_1^2 > \sigma_2^2$, y si $\dfrac{S_1^2}{S_2^2}F_{n_2-1,n_1-1;\alpha/2} < 1$ podremos asegurar que $\sigma_1^2 < \sigma_2^2$. En caso contrario, es decir, si se verifica

$$\frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1;\alpha/2}} \leq 1 \leq \frac{S_1^2}{S_2^2}F_{n_2-1,n_1-1;\alpha/2}$$

no podemos asegurar que las varianzas sean distintas y en este caso es común admitir que son iguales para poder aplicar el intervalo de confianza que obtuvimos anteriormente para la diferencia $\mu_1 - \mu_2$. Tenemos por tanto también un procedimiento que nos permite comparar las varianzas de dos variables aleatorias con distribución normal suponiendo que las dos muestras utilizadas son independientes entre sí.

Además, por el mismo razonamiento que utilizamos para el caso del intervalo de confianza para una varianza, también podemos asegurar que un intervalo de confianza con grado $1-\alpha$ para el cociente de las desviaciones típicas teóricas $\dfrac{\sigma_1}{\sigma_2}$ es

$$\left(\frac{S_1/S_2}{\sqrt{F_{n_1-1,n_2-1;\alpha/2}}}, \frac{S_1}{S_2}\sqrt{F_{n_2-1,n_1-1;\alpha/2}} \right)$$

Ejercicio. Utilizando los mismos datos de los ejercicios anteriores, encontrar un intervalo de confianza al 90 % para el cociente de las desviaciones típicas teóricas de ambos tipos de entrenamiento. ¿Podemos admitir que ambas desviaciones típicas son iguales?

Intervalo de confianza para la diferencia de medias en dos poblaciones normales con varianzas desconocidas y distintas. Si las dos variables aleatorias con distribución normal tienen varianzas poblacionales desconocidas y distintas podemos encontrar también un intervalo de confianza para la diferencia entre las medias poblacionales $\mu_1 - \mu_2$ utilizando un estadístico pivote cuya distribución aproximada es una t de Student. Concretamente, si S_1^2 y S_2^2 son las varianzas muestrales corregidas de ambas muestras independientes y definimos las constantes $\Delta_1 = \dfrac{S_1^2}{n_1}$, $\Delta_2 = \dfrac{S_2^2}{n_2}$ y $d = \left[\dfrac{(\Delta_1 + \Delta_2)^2}{\frac{\Delta_1^2}{n_1+1} + \frac{\Delta_2^2}{n_2+1}} \right]$ (donde $[x]$ representa a la parte entera de un número x) se verifica, aproximadamente, que

$$\frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \rightsquigarrow t_d$$

y este estadístico resulta adecuado para el intervalo que buscamos. Concretamente, si calculamos los cuantiles $c_{1-\alpha/2} = t_{d;\alpha/2}$ y $c_{\alpha/2} = -t_{d;\alpha/2}$ se tiene que

$$p\left(-t_{d;\alpha/2} \leq \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \leq t_{d;\alpha/2} \right) = 1 - \alpha$$

y, despejando el parámetro que queremos estimar $\mu_1 - \mu_2$ en esa doble desigualdad, obtenemos

$$p\left(\left(\overline{X} - \overline{Y}\right) - t_{d;\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \left(\overline{X} - \overline{Y}\right) + t_{d;\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right) = 1 - \alpha$$

y disponemos ya del intervalo de confianza buscado. Es decir un intervalo de confianza para $\mu_1 - \mu_2$ con grado de confianza $1-\alpha$ suponiendo las varianzas desconocidas y distintas es

$$\left(\overline{X} - \overline{Y}\right) \pm t_{d;\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Ejercicio. Con el mismo ejemplo que venimos utilizando, encontrar un intervalo de confianza con $1 - \alpha = 0.95$ suponiendo ahora que las varianzas son desconocidas y distintas. ¿Es muy distinto al calculado suponiendo varianzas desconocidas pero iguales?

Intervalo de confianza asintótico para la diferencia de medias de dos variables aleatorias con distribuciones de probabilidad no normales. Si consideramos ahora que las variables aleatorias X e Y tienen distribuciones de probabilidad cualesquiera (no necesariamente normales) pero los tamaños muestrales n_1 y n_2 son suficientemente grandes (si las distribuciones no son muy asimétricas bastaría mayores que 30) también podemos encontrar un intervalo de confianza asintótico para la diferencia $\mu_1 - \mu_2$. En efecto, teniendo en cuenta el Teorema Central del Límite y la propiedad de consistencia de la varianza muestral corregida como estimador de la varianza teórica, podemos decir que, cuando los tamaños muestrales tienden hacia ∞, se verifica

$$\frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \widetilde{\rightsquigarrow} N(0,1)$$

y razonando de la forma habitual tendremos que $\left(\overline{X} - \overline{Y}\right) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ es un intervalo de confianza asintótico para la diferencia $\mu_1 - \mu_2$ cuando las variables aleatorias X e Y no tienen distribución normal pero los tamaños muestrales son suficientemente grandes.

Ejercicio. Se desea comparar la altura media de dos poblaciones de *Pinus sylvestris* ubicadas en dos localizaciones de características a priori diferentes. Se midieron, en cada localización, las alturas totales de 51 ejemplares obteniéndose $\bar{x} = 10.576$ y $s_1^2 = 6.952$ para la primera localidad e $\bar{y} = 8.259$ y $s_2^2 = 8.479$ para la segunda. Construir un intervalo de confianza con $1 - \alpha = 0.99$ para la diferencia de medias entre las dos poblaciones. ¿A qué conclusiones podemos llegar a la vista de dicho intervalo?

Intervalo de confianza asintótico para la diferencia de dos proporciones poblacionales p_1 y p_2. Un caso especialmente interesante en la aplicación del anterior intervalo de confianza asintótico es el problema de estimación de una diferencia de proporciones en dos grupos de una determinada población. Elegido un individuo aleatoriamente en el primer grupo, podemos considerar entonces que tenemos una variable aleatoria de Bernouilli con valor 1 si el individuo es un éxito y el valor 0 si es un fracaso. Es decir, $X \rightsquigarrow B(p_1)$ donde $p_1 = EX$ y $Var(X) = p_1(1-p_1)$. Del mismo modo, para el segundo grupo, tendríamos $Y \rightsquigarrow B(p_2)$ donde $p_2 = EY$ y $Var(Y) = p_2(1-p_2)$. Entonces, dadas sendas muestras aleatorias simples independientes de ambos grupos $X_1, X_2, ..., X_{n_1}$ de la variable X e $Y_1, Y_2, ..., Y_{n_2}$ de la variable Y con n_1 y n_2 suficientemente grandes, se tiene que $\overline{X} = \hat{p}_1$ (es decir, la proporción muestral de éxitos en la primera muestra), $S_1 \simeq \sqrt{\hat{p}_1(1-\hat{p}_1)}$, $\overline{Y} = \hat{p}_2$ (es decir, la proporción muestral de éxitos en la segunda muestra) y $S_2 \simeq \sqrt{\hat{p}_2(1-\hat{p}_2)}$. Entonces tendríamos que

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \widetilde{\rightsquigarrow} N(0,1)$$

y razonando de la forma habitual obtenemos que $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ es un intervalo de confianza asintótico para la diferencia entre las proporciones poblacionales con una confianza $1 - \alpha$.

Ejercicio. Supongamos que, en relación con el ejercicio anterior, se observó que de los 51 pinos de la primera localidad 16 estaban atacados por la procesionaria, mientras que en los 51 de la segunda localidad sólo 8 lo estaban. ¿Podemos concluir con una confianza del 99 % que las proporciones de pinos atacados

en ambas localidades son diferentes? ¿Y con una confianza del 90 %? ¿Cuántos pinos deberíamos utilizar en cada localidad para poder afirmar que son diferentes con una confianza de 0.95?

Intervalo de confianza para la diferencia de medias en dos variables aleatorias con distribución normal pero utilizando muestras pareadas. En algunas situaciones prácticas las dos muestras aleatorias simples $X_1, X_2, ..., X_n$ de la variable $X \rightsquigarrow N(\mu_1, \sigma_1)$ e $Y_1, Y_2, ..., Y_n$ de la variable $Y \rightsquigarrow N(\mu_2, \sigma_2)$ no son independientes porque cada pareja de observaciones (X_i, Y_i) ha sido tomada sobre una misma unidad experimental (por eso necesariamente $n_1 = n_2 = n$). Se dice entonces que disponemos de dos muestras aleatorias pareadas de las variables X e Y. En este caso la estimación mediante intervalos de confianza de la diferencia $\mu_1 - \mu_2$ no puede hacerse como hemos estudiado hasta ahora por la falta de independencia entre las dos muestras. La falta de independencia no afecta a la estimación puntual $\mu_1 - \mu_2$ puesto que sigue siendo cierto que $E\left(\overline{X} - \overline{Y}\right) = \mu_1 - \mu_2$, pero sí a la elección de un estadístico pivote adecuado. Para este problema, si definimos la nueva variable aleatoria $D = X - Y$, podemos asegurar que $D \rightsquigarrow N(\mu_1 - \mu_2, \sigma_D)$ con un cierto parámetro σ_D desconocido y que, a priori no podemos evaluar a partir de las desviaciones típicas σ_1 y σ_2. No obstante, disponemos de una muestra aleatoria simple $D_1, D_2, ..., D_n$ de la variable aleatoria D, siendo $D_i = X_i - Y_i$ y podemos utilizar la teoría estudiada para estimar el valor esperado $ED = \mu_1 - \mu_2$ utilizando una muestra aleatoria simple con varianza desconocida σ_D. Concretamente, tendremos que $\overline{D} = \overline{X} - \overline{Y}$ y $\dfrac{\overline{D} - ED}{S_D/\sqrt{n}} = \dfrac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{S_D/\sqrt{n}} \rightsquigarrow t_{n-1}$ siendo S_D la desviación típica muestral corregida de la muestra de diferencias $X_i - Y_i$. Eligiendo entonces los cuantiles $c_{1-\alpha/2} = t_{n-1;\alpha/2}$ y $c_{\alpha/2} = -t_{n-1;\alpha/2}$ se tiene que

$$p\left(-t_{n-1;\alpha/2} \leq \frac{\left(\overline{X} - \overline{Y}\right) - (\mu_1 - \mu_2)}{\frac{S_D}{\sqrt{n}}} \leq t_{n-1;\alpha/2}\right) = 1 - \alpha$$

y, despejando el parámetro que queremos estimar $\mu_1 - \mu_2$ en esa doble desigualdad, obtenemos

$$p\left(\left(\overline{X} - \overline{Y}\right) - t_{n-1;\alpha/2}\frac{S_D}{\sqrt{n}} \leq \mu_1 - \mu_2 \leq \left(\overline{X} - \overline{Y}\right) + t_{n-1;\alpha/2}\frac{S_D}{\sqrt{n}}\right) = 1 - \alpha$$

Es decir un intervalo de confianza para $\mu_1 - \mu_2$ con grado de confianza $1 - \alpha$ es

$$\left(\overline{X} - \overline{Y}\right) \pm t_{n-1;\alpha/2}\frac{S_D}{\sqrt{n}}$$

Ejercicio. Un grupo de investigadores afirma haber descubierto un tipo de alimentación para las gallinas, bajo la cual éstas producen huevos que no aumentan el colesterol en las personas que los consumen. Para comprobar dicha teoría se seleccionaron al azar 10 personas a las que se les midió el colesterol antes (X) y después (Y) de ser sometidos a una dieta a base de dichos huevos. Suponiendo normalidad, encontrar un intervalo de confianza al 95 % para la diferencia entre los valores medios del colesterol antes y después de la dieta, si los datos obtenidos son los siguientes:

X (antes)	120	312	243	161	314	234	143	287	423	155
Y (después)	130	306	255	168	310	250	158	290	440	140

¿Podemos considerar jusficada la afirmación de los investigadores? En base a los datos obtenidos, ¿cuántas personas necesitaríamos añadir a la muestra para rebatir la afirmación de los investigadores?

8.1. PROBLEMAS

1. Según el Environmental News (Septiembre de 1975), la "lluvia ácida" causada por la reacción de ciertos contaminantes en el aire con el agua de la lluvia es un problema creciente en ciertas regiones. La lluvia pura que se precipita a través del aire limpio tiene un pH de 5.7 (el pH es una medida de la acidez: 0 ácido y 14 alcalino). Se analizan 40 muestras de agua de diferentes lluvias con respecto a su pH dando una media igual a 3.7 y desviación típica de 0.5. Determinar un intervalo de confianza del 99 % para el pH medio de las lluvias e interpretar el resultado. ¿Que supuestos deben hacerse para que el intervalo sea válido?.

   Si se desea estimar el promedio de PH de las lluvias en un área que experimenta una gran contaminación debido a la presencia de una fábrica que emite gran cantidad de humos, y se sabe que $\sigma = 0.5$, ¿cuántas lluvias deben tomarse para que la estimación difiera de la media teórica como mucho en 0.1 con una probabilidad de al menos 0.95?

2. Un periódico señala en su número del 23 de mayo de 1979 que "los estudiantes de derecho se oponen a la pena de muerte". Esta declaración se hizo en base a una encuesta para la cual se escogieron al azar y se entrevistaron 86 estudiantes de derecho. El 52 % de los entrevistados declararon que se oponían a la pena de muerte. A partir de esta información obtenga un intervalo de confianza del 95 % para la proporción real de estudiantes de derecho que se oponen a la pena de muerte. ¿Se justifica la afirmación de este periódico? ¿Cuántos estudiantes se tendrán que entrevistar por parte del periódico para estimar la proporción de estudiantes en contra de la pena de muerte, con un error máximo de 0.01 con una probabilidad de almenos 0.95?

3. La dirección médica de una clínica desea estimar el número promedio de días necesarios para el tratamiento de pacientes con edades entre 25 y 34 años. Una muestra aleatoria de 500 pacientes con estas características proporcionó una media y una desviación estándar (utilizando la cuasi-varianza) de 5.4 y 3.1 días, respectivamente. Obtener un intervalo de confianza del 95 % para el promedio de estancia en el hospital de la población de pacientes de la que se obtuvo la muestra.

4. Una encuesta intenta estudiar la relación entre la participación en los deportes y la destreza manual. De una muestra aleatoria de 37 alumnos de Enseñanza Secundaria que participaban regularmente en actividades deportivas se obtuvo una calificación media en asignaturas que medían la destreza manual de 32.19 con una desviación estándar de 4.34. De otra muestra aleatoria independiente de 37 alumnos de Enseñanza Secundaria que no participaban en deportes se calculó una calificación media de 31.68 y una desviación estándar de 4.56. Estimar la diferencia de los verdaderos promedios de ambos grupos con un intervalo de confianza del 90 %. ¿Se desprende de los resultados que la calificación promedio de ambos grupos es diferente?

   Si debe efectuarse otro estudio similar, ¿cuántas observaciones deben incluirse en cada grupo (igual número en ambos) para producir un intervalo de confianza del 90 % con una amplitud de 2 unidades?

5. Una encuesta realizada en Otoño de 1979 con respecto a la política de jubilaciones, reveló el pesimismo de la población respecto a sus perspectivas cuando lleguen a jubilarse. El 62.9 % de los 6100 trabajadores entrevistados, indicaron que creían que sus ingresos al jubilarse no serían suficientes. Calcular un intervalo de confianza del 95 % para esta proporción e interpretarlo.

6. Se realizó una encuesta preguntando si el sexo de un candidato político era decisivo a la hora de recibir el voto. El 62 % por ciento de 241 hombres y el 49 % de 256 mujeres opinaron que el sexo del candidato no importaba. Construir un intervalo de confianza para la diferencia de proporciones de hombres y mujeres que opinan que el sexo del candidato no tiene importancia en la votación.

7. Los resultados de un estudio sobre la concentración de plomo en el agua indicaron que el 20 % de los 248 hogares que se analizaron tenían una concentración de plomo en el agua que superaba el nivel permitido. En otro punto geográfico se observó que solo el 5 % de los 110 hogares analizados superaban este nivel. ¿Existe diferencia entre las dos zonas? Utilizar para responder a esta pregunta un intervalo de confianza del 95 %.

8. La tasa de consumo de oxigeno es una medida importante a la hora de estudiar la actividad fisiológica de los atletas. Se realizó un estudio para intentar observar las diferencias entre el consumo de oxigeno en dos grupos de varones universitarios entrenados con dos métodos diferentes, en uno el entrenamiento se realizaba de forma continua y en otro de forma intermitente. El consumo de oxigeno se mide en mililitros por kilogramo-minuto, y los resultados obtenidos vienen dados en la siguiente tabla:

	Entrenamiento I	Entrenamiento II
Tamaño muestral	9	7
Media muestral	43.71	39.63
Desviación típica corregida	5.88	7.68

   Si se supone que las mediciones provienen de poblaciones normales con varianzas iguales, estimar la diferencia entre las medias poblacionales de ambos tipos de entrenamiento con una confianza del 95 %. ¿Como podría comprobar que las varianzas se pueden considerar iguales?

   Responder a las mismas preguntas sin suponer que las mediciones provienen de dos poblaciones normales, esto es, provienen de dos distribuciones arbitrarias.

9. La Agencia para la Protección Ambiental reunió información respecto a las mediciones de $CL50$ (concentración letal de una sustancia que mata el 50 % de los animales utilizados en la experimentación) para ciertos productos que se encuentran en ocasiones en rios y lagos de agua dulce. Para cierta especie de peces las mediciones de $CL50$ para el DDT en 12 experimentos fueron las siguientes (en partes por millón): 16, 5, 21, 19, 10, 5, 8, 2, 7, 2, 4, 9 Estimar el verdadero promedio de $CL50$ para el DDT utilizando un intervalo de confianza del 90 % suponiendo que las mediciones de $CL50$ siguen aproximadamente una distribución normal.

Otro insecticida común, Diazinón, dio las siguientes mediciones de $CL50$ en tres experimentos: 7.8, 1.6 y 1.3. Calcular en este caso un intervalo de confianza del 90 % para la media poblacional. Estimar la diferencia entre los promedios de $CL50$ para el DDT y el Diazinón utilizando un intervalo de confianza del 90 %. ¿Qué supuestos hay que realizar para que el intervalo construido sea válido?

10. Se registraron las áreas (en hectáreas) que ocupan los caimanes, según las estaciones, en un lago de las cercanías de Gainesville (Florida) por la Comisión de Caza y Pesca de Florida. Cinco caimanes registrados en la primavera mostraron áreas de 8.0, 12.1, 8.1, 18.2 y 31.7. Cuatro caimanes diferentes registrados en el verano mostraron áreas de 102.0, 81.7, 54.7 y 50.7. Estimar la diferencia en el promedio de las áreas ocupadas por los caimanes en la primavera y en el verano utilizando un intervalo de confianza del 95 %. ¿Qué supuestos deben establecerse?

11. En el trabajo de un laboratorio es deseable verificar cuidadosamente la variabilidad de las lecturas obtenidas de muestras estándar. En un estudio de la concentración de calcio en agua potable como parte de la valoración de la calidad del agua se pasó el mismo patrón de medida seis veces por el laboratorio en intervalos aleatorios. Las lecturas en partes por millón fueron 9.54, 9.61, 9.32, 9.48, 9.70 y 9.26. Estimar la varianza de las lecturas y dar para ella un intervalo de confianza del 90 %.

12. Se seleccionó una muestra aleatoria de 21 ingenieros de un grupo mayor que trabaja para una determinada empresa. La desviación estándar de la muestra de las horas de trabajo por semana fue de 7 horas. Determinar un intervalo de confianza del 95 % para la varianza de la población suponiendo que las medidas siguen una distribución normal.

13. El crecimiento del tronco principal para una muestra de 17 pinos rojos de 4 años tiene una media de 11.3 pulgadas y una desviación estándar de 3.4 pulgadas. Obtener un intervalo de confianza del 90 % para la media del crecimiento del tronco principal para una población de pinos rojos de 4 años sujeta a condiciones ambientales similares. Supóngase que el crecimiento tiene un distribución normal.

14. Se aplicaron dos métodos para enseñar la lectura a dos grupos de niños de la escuela primaria, comparándose los resultados mediante una prueba de lectura y comprensión al final del periodo de aprendizaje. Obtener un intervalo de confianza del 95 % para la diferencia de promedios. ¿Qué supuestos hay que hacer?

	Método I	Método II
Tamaño muestral	11	14
Media muestral	64	69
Varianza corregida	52	71

15. Una comparación de los tiempos de reacción a dos estímulos diferentes en un experimento con 16 animales produjo los resultados (en segundos) dados por la siguiente tabla:

Estímulo 1:	1	3	2	1	2	1	3	3
Estímulo 2:	4	3	3	3	1	2	3	3

Obtener un intervalo de confianza del 90 % para la diferencia de medias de los tiempos de reacción. ¿Qué supuestos hay que hacer?

16. El tiempo transcurrido entre la facturación y el pago recibido se registró para una muestra aleatoria de 100 clientes de una empresa de contadores públicos. La media y la desviación estándar de las 100 cuentas fueron respectivamente 39.1 días y 17.3 días. Obtener un intervalo de confianza del 90 % para el tiempo medio que transcurre entre la facturación y el pago recibido para todas las cuentas de la firma de contadores públicos. Interpretar el resultado.

17. Se supone que los diámetros normales de los pies de una determinada masa forestal siguen una distribución normal. Se eligen 10 pies al azar obteniéndose las siguientes medidas de sus diámetros:

    25.5 26.8 24.2 25.0 27.3 26.1 23.2 28.4 27.8 25.7 $\left(\sum x_i = 260 \text{ y } \sum x_i^2 = 6783.76\right)$

    a) Obtener las estimaciones puntuales para la media y la varianza del diámetro de la masa.

    b) Determinar un intervalo de confianza al 90 % para el diámetro medio de la masa.

    c) Determinar un intervalo de confianza al 90 % para la varianza y la desviación típica de la población.

18. En un estudio sobre el crecimiento de una determinada especie vegetal se plantaron 10 semillas, 5 en una solución de nutrientes estándar y otras 5 en una solución de nutrientes que contenía una cantidad extra de nitrógeno. Después de 22 días, las plantas fueron recogidas y pesadas en seco, obteniéndose los siguientes resultados:

	Estándar	Nitrógeno extra
Tamaño muestral	5	5
Media muestral	3.62	4.17
Desviación típica corregida	0.54	0.67

    a) Construir un intervalo de confianza del 90 % para la razón de varianzas entre ambas poblaciones. ¿A qué conclusión podemos llegar sobre las varianzas de estas poblaciones? ¿Por qué?

    b) A partir de la conclusión obtenida en el apartado (a), construir un intervalo de confianza para la diferencia de medias entre las dos poblaciones a un nivel de confianza del 95 %. ¿Podemos asegurar que los tratamientos producen diferentes rendimientos en la producción? ¿Por qué?

    (En ambos apartados suponer que las dos poblaciones siguen una distribución normal)

19. Se desea comparar la altura de dos poblaciones de *Pinus sylvestris* ubicadas en dos localizaciones de características a priori diferentes. Se midieron, en cada localización, las alturas totales de 51 ejemplares. El resumen de los resultados aparece en la siguiente tabla:

	Localidad 1	Localidad 2
Tamaño muestral	51	51
Media muestral	10.576	8.259
Varianza corregida	6.952	8.479

Considerando que ambas poblaciones siguen una distribución normal:

a) Construir un intervalo de confianza del 95 % para la razón de varianzas entre ambas poblaciones. ¿A qué conclusión podemos llegar sobre las varianzas de estas poblaciones? ¿Por qué?

b) A partir de la conclusión obtenida en el apartado anterior, construir un intervalo de confianza del 95 % para la diferencia de medias entre las dos poblaciones. ¿A qué conclusiones podemos llegar sobre la altura de las dos poblaciones? ¿Por qué?

c) Además se observó que de los 51 pinos observados en la primera localización 16 estaban atacados por la procesionaria, mientras que en los 51 de la segunda población tan sólo 8 lo estaban. ¿Podemos concluir con una confianza del 99 % que las proporciones de pinos atacados en ambas poblaciones es diferente? ¿Y con una confianza del 90 %?

20. En un estudio sobre la evolución de una plantación de *Pinus pinaster* fueron medidas las alturas del fuste de 20 ejemplares, obteniéndose los siguientes resultados:

    5.3, 7.6, 7.7, 8.2, 8.3, 9.2, 9.3, 9.4, 9.7, 10.1, 10.5, 10.7, 10.8, 11.0, 11.3, 12.1, 12.2, 12.5, 13.3, 16.3

    ($\sum x_i = 205.5$ y $\sum x_i^2 = 2221.01$)

    a) Suponiendo que los datos siguen una distribución normal calcular intervalos de confianza, con una confianza del 95 %, para la altura media y variabilidad de la población.

    b) Calcular una intervalo de confianza al 90 %, para la proporción de árboles de la plantación con altura superior a 10 metros.

21. Se disponen de los siguientes resultados sobre el punto de ebullición (^oK) de un aldehído: 376, 351, 346, 363, 380, 372, 365, 356, 393, 367, 358, 342, 370, 385, 360, 365 (suma de valores= 5849 y suma de valores al cuadrado= 2141023).

    Se asegura que, añadiendo al aldehído un determinado aditivo, aumenta su temperatura de ebullición. Para verificar esta afirmación se realizan 10 pruebas con el aldehído más el aditivo y se obtienen los siguientes puntos de ebullición (también en $^\circ K$): 355, 378, 390, 400, 379, 381, 411, 374, 396, 366 (suma de valores= 3830 y suma de valores al cuadrado= 1469380). Suponiendo que en ambos casos el punto de ebullición sigue una distribución normal, se pide:

    a) Encontrar un intervalo de confianza al 95 % para la temperatura media de ebullición en ambos casos. En base a estos intervalos, ¿podríamos decir que el aditivo aumenta la temperatura de ebullición? Razona la respuesta.

    b) Encontrar un intervalo de confianza al 90 % para el cociente de las varianzas teóricas del punto de ebullición del aldehído según se use o no aditivo. En vista del resultado, ¿podríamos decir, con un 90 % de confianza, que dichas varianzas son distintas?

    c) Teniendo en cuenta el apartado anterior, encontrar un intervalo de confianza al 95 % para la diferencia entre los puntos medios de ebullición en ambos casos. ¿Podríamos afirmar, con un 95 % de confianza, que el aditivo aumenta la temperatura de ebullición?

22. Para comparar el pH del suelo de dos explotaciones forestales (A y B) se tomaron 8 muestras de tierra de cada una de ellas en puntos elegidos al azar sobre su superficie. Se analizaron las muestras para determinar el pH, obteniéndose los siguientes datos:

Muestra	1	2	3	4	5	6	7	8		
Explotación A	6.55	5.98	5.59	6.17	5.92	6.18	6.43	5.68	$\sum x_i = 48.5$	$\sum x_i^2 = 294.826$
Explotación B	6.78	6.14	6.80	6.91	6.10	6.01	6.18	6.88	$\sum y_i = 51.8$	$\sum y_i^2 = 336.513$

Suponiendo que en ambas explotaciones el pH tiene una distribución normal, se pide:

a) Encontrar un intervalo de confianza al 90 % para el cociente de las varianzas teóricas del pH de ambas explotaciones. En base a este intervalo, ¿podríamos decir, con un 90 % de confianza, que dichas varianzas son distintas?

b) Teniendo en cuenta el resultado obtenido en el apartado anterior, encontrar un intervalo de confianza al 99 % para la diferencia de los valores medios teóricos del pH del suelo de ambas explotaciones. ¿Podríamos afirmar, con un 95 % de confianza, que dichos valores son distintos? ¿Y con una confianza del 99 %?

c) Si en la explotación A de 40 pies se observaron 23 atacados por un determinado hongo, mientras que en la explotación B se detectaron 18 pies infectados de otros 40 analizados ¿Cuál es el mayor nivel de confianza con el que podemos afirmar que las proporciones de pies infectados es diferente en ambas poblaciones?

23. Un grupo de investigadores afirma haber descubierto un tipo de alimentación para las gallinas, bajo la cual éstas producen huevos que no aumentan el colesterol en las personas que los consumen. Para comprobar dicha teoría se seleccionaron al azar 10 personas a las que se les midió el colesterol antes (X) y después (Y) de ser sometidos a una dieta a base de dichos huevos. Suponiendo normalidad, encontrar un intervalo de confianza al 95 % para la diferencia entre los valores medios del colesterol antes y después de la dieta, si los datos obtenidos son los siguientes:

X (antes)	120	312	243	161	314	234	143	287	423	155
Y (después)	130	306	255	168	310	250	158	290	440	140

¿Podemos considerar jusficada la afirmación de los investigadores? En base a los datos obtenidos, ¿cuántas personas necesitaríamos añadir a la muestra para rebatir la afirmación de los investigadores?

Capítulo 9

TEST DE HIPÓTESIS

En el capítulo anterior hemos aprendido a efectuar inferencias acerca de los valores paramétricos poblacionales a partir de información muestral, y lo hemos hecho tanto a través de estimadores puntuales como de intervalos de confianza. Sin embargo, en muchas ocasiones el investigador tiene una creencia a priori, posiblemente basada en su experiencia previa con el fenómeno que está estudiando, acerca de los valores numéricos de dichos parámetros y, sobre la base de la información proporcionada por una o más muestras, desea someter a prueba esa suposición para tomar la decisión de rechazarla o no rechazarla. Esta forma de razonar, denominada contrastación de hipótesis, es uno de los procedimientos estadísticos más usados en las ciencias naturales por ser un procedimiento que le proporciona al investigador un criterio objetivo para tomar decisiones en base a un número siempre limitado de observaciones.

Concepto de test de hipótesis. Es un procedimiento estadístico de decisión sobre una determinada hipótesis formulada previamente por el investigador en relación a un fenómeno aleatorio o una variable aleatoria. Esta hipótesis previa se denomina **hipótesis nula**, se representa por H_o y es la que el investigador está dispuesto a creer a priori. Para diseñar este procedimiento es necesario especificar también una **hipótesis alternativa**, que representaremos por H_1, y que es aquella que pasaremos a aceptar si se rechaza la hipótesis nula. Ahora bien, la idea previa para la formulación de un test de hipótesis es que existen razones para creer que la hipótesis nula es cierta y sólo si en los datos muestrales hay mucha evidencia de lo contrario será rechazada para concluir que la hipótesis alternativa es cierta.

Por tanto, en la investigación científica interesa que la proposición que el investigador quiere probar esté recogida en la hipótesis alternativa H_1, mientras que la hipótesis nula H_o asume la proposición que se quiere negar. La utilidad de plantear las hipótesis de esta manera se explica porque el rechazo de H_o es un veredicto mucho más robusto que su no rechazo, puesto que es necesario acumular evidencia científica muy fuerte para poder rechazar una hipótesis nula. Por tanto, la consecuencia de rechazar una hipótesis nula es un gran apoyo a la hipótesis alternativa.

Ilustraremos esta situación con la analogía de un proceso judicial donde hay alguien acusado de un delito y dos hipótesis: inocente (hipótesis nula H_o) o culpable (hipótesis alternativa H_1). Si el fiscal público

tiene interés en probar que el acusado es culpable necesitará presentar suficientes evidencias que garanticen que la decisión es correcta. Por ello empezará suponiendo la hipótesis nula de que el acusado es inocente, de modo que, sólo si se tiene una fuerte evidencia de lo contrario, será rechazada para concluir que el acusado es culpable. Si no se tienen evidencias fuertes la hipótesis nula no puede ser rechazada, pero esto no significa que se probó la inocencia del acusado, sino sólo que no se logró acumular suficientes elementos para rechazar la hipótesis nula H_o de la inocencia. De hecho es posible que con nuevas investigaciones se determine la culpabilidad del acusado. Por el contrario, si se han obtenido suficientes evidencias de culpabilidad, se acepta la hipótesis alternativa H_1, y esta decisión es mucho más difícil de rebatir. En otras palabras, la probabilidad de cometer un error es mucho menor al rechazar H_o que al no rechazarla. En la práctica jurídica, si la evidencia es débil es preferible equivocarse declarando inocente a alguien culpable que condenando a un inocente.

Un razonamiento similar a éste es el que usan los investigadores cuando plantean como hipótesis alternativa la proposición que se quiere probar. Si los datos usados para probar las hipótesis proporcionan suficiente evidencia para rechazar la hipótesis nula, entonces la hipótesis alternativa recibe un respaldo muy fuerte. Pero si el investigador hubiese planteado la proposición que quiere probar como la hipotesis nula, su no rechazo no demuestra que la proposición sea cierta, sino que los datos no proporcionan evidencia suficiente para rechazarla, dejando abierta la posibilidad de ser refutada con otro conjunto datos o de que otra hipótesis sea la verdadera. Por esta razón, la sustitución del término *no rechazar* H_o por el término *aceptar* H_o no es muy conveniente y, si se hace, se debe estar consciente de que la aceptación es sólo temporal.

Estadístico de prueba. Es una variable aleatoria construida a partir de la o las muestras involucradas en el test de hipótesis y que utilizaremos para tomar la decisión más adecuada. Es el equivalente al estadístico pivote en el problema de intervalos de confianza y en general lo representaremos con T. Siempre será necesario conocer la distribución de probabilidad del estadístico de prueba suponiendo que la hipótesis nula H_o sea cierta. Esto condicionará la elección del estadístico de prueba y también por ello es muy habitual que, cuando la hipótesis nula se refiera a un parámetro, H_o especificará un valor concreto del mismo (**hipótesis nula simple**). A veces también se establecen hipótesis nulas compuestas, que incluyen un rango de valores del parámetro, pero nosotros nos centraremos en las hipótesis simples.

Región crítica. Es el conjunto de valores del estadístico de prueba que conducirán al rechazo de la hipótesis nula y, en general, la representaremos por C. De este modo, si T_o es el valor observado del estadístico de prueba y $T_o \in C$ rechazaremos la hipótesis nula, mientras que si $T_o \notin C$ no rechazaremos. El conjunto complementario de la región crítica es lo que denominaremos región de aceptación, o mejor, región de no rechazo.

Nivel de significación. Se define como la probabilidad de rechazar la hipótesis nula suponiendo que sea cierta y se representa por α, es decir, $\alpha = p\,(T \in C/H_o \text{ cierta})$. Siempre es fijado por el investigador y suele ser un número próximo a cero puesto que, como ya hemos dicho, la hipótesis nula sólo será rechazada si hay una fuerte evidencia de que es falsa. Habitualmente, los niveles de significación más usados son

$\alpha = 0.05$, $\alpha = 0.01$ y $\alpha = 0.001$, aunque en algunas investigaciones, cuando las consecuencias de rechazar la hipótesis nula suponiendo que es cierta no son muy graves, se puede usar también $\alpha = 0.10$.

Construcción de un test de hipótesis. Los pasos que seguiremos para diseñar un test de hipótesis serán los siguientes:

1. Establecer la **hipótesis nula** H_o y la **hipótesis alternativa** H_1.
2. Seleccionar el **estadístico de prueba** T con la condición de que su distribución de probabilidad sea conocida suponiendo que la hipótesis nula H_o es cierta.
3. Fijar el **nivel de significación** α que se desea utilizar.
4. Elegir la **región crítica** C de modo que $p\,(T \in C) = \alpha$ suponiendo que H_o es cierta.
5. Seleccionar la muestra o las muestras necesarias y calcular el **valor del estadístico de prueba** T_o con los datos observados.
6. **Tomar la decisión**: rechazar la hipótesis nula si $T_o \in C$ y no hacerlo en caso contrario.

Una vez que hemos diseñado el procedimiento de decisión, y teniendo en cuenta que éste se basa en unos datos que son aleatorios, tenemos que ser conscientes de que podemos estar equivocados en la decisión tomada. Se hace entonces necesario valorar los riesgos del test de hipótesis en la toma decisión. Teniendo en cuenta que hay dos situaciones reales posibles y dos posibles resultados del test de hipótesis, podrán presentarse cuatro posibilidades de modo que en dos de ellas la decisión será correcta y en otras dos no. Concretamente, en la siguiente tabla resumimos las cuatro posibilidades:

		Resultado del test	
		No rechazar H_o	Rechazar H_o
Situación real	H_o cierta	Correcto	Error de tipo I
	H_o falsa	Error de tipo II	Correcto

Por tanto pueden presentarse dos decisiones incorrectas: rechazar H_o siendo cierta (**error de tipo I**) o no rechazar H_o siendo falsa (**error de tipo II**). Con las definiciones que hemos dado, es claro que la probabilidad de cometer un error de tipo I es el nivel de significación α del test, y podemos asegurar que será pequeña porque habrá sido fijada previamente por el investigador. No ocurre lo mismo con la probabilidad de error de tipo II, que suele representarse por β. Es decir, en términos probabilísticos los dos tipos de error se expresan de la siguiente forma:

$$p\,(\text{error tipo I}) = p\,(T \in C / H_o \text{ cierta}) = \alpha = \text{nivel de significación}$$

$$p\,(\text{error tipo II}) = p\,(T \notin C / H_1 \text{ cierta}) = \beta$$

Como puede observarse, tanto α como β son probabilidades condicionadas y las probabilidades de ambos errores no pueden valorarse en un sentido absoluto. Para calcular α es necesario suponer que

H_o es cierta y para calcular β se asume que H_1 es cierta. Sería conveniente que ambas probabilidades fuesen pequeñas pero esto no es fácil de conseguir porque al intentar disminuir el valor de α generalmente aumentaremos el valor de β. Es lógico que así sea, porque si queremos tener poco riesgo de rechazar la hipótesis nula siendo cierta, tendremos que asumir un mayor riesgo de no rechazarla siendo falsa. En el caso extremo, si queremos que α sea igual a 0, siempre tendremos que aceptar H_o y por tanto nunca la rechazaremos siendo falsa, es decir β será igual a 1. Sin embargo, al incrementar el tamaño muestral utilizado para resolver el test de hipótesis, suele ser posible conseguir que el valor β sea pequeño manteniendo el valor α próximo a 0.

Se define entonces la **potencia del test** como la probabilidad de rechazar la hipótesis nula siendo falsa, es decir, $1 - \beta$, de modo que elegiremos test de hipótesis que tengan una potencia alta para un nivel de significación α pequeño y fijado previamente. Si queremos que la potencia sea muy alta y el nivel de significación muy pequeño tendremos que utilizar tamaños muestrales muy grandes. Es interesante observar que generalmente, cuando la hipótesis nula se refiere a un parámetro, la hipótesis alternativa H_1 contiene un conjunto infinito de posibles valores de ese parámetro y por tanto la potencia del test es en realidad una función del parámetro que a cada posible valor le asigna la correspondiente probabilidad. Es decir, en general habrá que hablar de la **función de potencia** del test.

Definición de p-valor. De acuerdo con lo estudiado hasta ahora es evidente que en todo test de hipótesis el rechazo o no rechazo de la hipótesis nula depende del nivel de significación fijado previamente, y si cambiamos su valor necesitamos volver a calcular la nueva región crítica para poder tomar una desición. Esto hace que el procedimiento sea demasiado dependiente del nivel de significación. Para evitar este inconveniente se introduce el concepto de p-valor, que se define como la probabilidad de error de tipo I que cometeríamos si rechazamos la hipótesis nula con los datos observados, es decir, eligiendo una región crítica que contenga al valor observado del estadístico de prueba y todos los demás valores de dicho estadístico más desfavorables para la hipótesis nula. La importancia del p-valor viene dada porque nos proporciona un resultado mucho más informativo que el resultado del contraste, que únicamente termina diciendo si rechazamos o no la hipótesis nula para ese nivel de significación. Sin embargo, el p-valor cuantifica la probabilidad de error de tipo I que tendremos que asumir si queremos rechazar la hipótesis nula con los datos observados. Por tanto, podemos interpretar el p-valor como una medida de la evidencia que aportan los datos a favor de la hipótesis nula y valores bajos del p-valor se corresponden con datos que no apoyan dicha hipótesis, ya que la probabilidad de cometer un error de tipo I en caso de que la rechazáramos sería baja. Además, el p-valor nos proporciona un criterio de decisión para cualquier valor que tome el nivel de significación. Concretamente, si el p-valor es menor que el nivel de significación prefijado, rechazaremos la hipótesis nula porque la probabilidad de error de tipo I que cometeríamos rechazándola sería menor que el nivel de significación. En caso contrario, es decir, si el p-valor está por encima del nivel de significación, la hipótesis nula no debería rechazarse porque, si lo hiciésemos, la probabilidad de error de tipo I que cometeríamos estaría por encima del nivel de significación. Hoy en día es habitual utilizar el p-valor para resolver un test de hipótesis, en lugar de utilizar la región crítica

calculada para un nivel de significación prefijado.

Test de hipótesis para el parámetro μ de una distribución normal con σ conocida. Como primer ejemplo de este tipo de estimación vamos a considerar el problema de diseñar un test de hipótesis para el parámetro $\mu = EX$ de una distribución normal con desviación típica conocida σ, fijado un nivel de significación α para el test. Supongamos que la hipótesis nula es $H_o : \mu = \mu_o$ y que la hipótesis alternativa es $H_1 : \mu \neq \mu_o$. Si $X_1, X_2, ..., X_n$ representa a la muestra aleatoria simple de la variable $X \rightsquigarrow N(\mu, \sigma)$, y suponiendo que la hipótesis nula H_o es cierta, podemos asegurar que $\dfrac{\overline{X} - \mu_o}{\sigma/\sqrt{n}} = T \rightsquigarrow N(0,1)$. Por tanto tendremos un estadístico de prueba T cuya distribución de probabilidad es conocida si H_o es cierta. Ahora tendremos que elegir la región crítica C de modo que $p\left(T \in C/\mu = \mu_o\right) = \alpha$ y teniendo en cuenta cuál es la hipótesis alternativa H_1. En nuestro caso es claro que la región crítica debería contener valores grandes del estadístico de prueba tanto en la zona positiva como en la zona negativa, ya que la hipótesis alternativa es $\mu \neq \mu_o$, es decir, es bilateral. Por tanto debería ser $C = \left(-\infty, -z_{\alpha/2}\right) \cup \left(z_{\alpha/2}, \infty\right)$ de modo que $p\left(T \in C/\mu = \mu_o\right) = 2p\left(T > z_{\alpha/2}\right) = \alpha$. Es decir, rechazaremos la hipótesis nula H_o si el valor del estadístico de prueba es, en módulo, mayor que $z_{\alpha/2}$. Para tomar la decisión en función del p-valor, si T_o es el valor del estadístico de prueba observado en la muestra, deberemos calcular $p-valor = 2p\left(T > |T_o|\right)$ y rechazar la hipótesis nula siempre que este p-valor esté por debajo del nivel de significación que queramos utilizar.

En los problemas prácticos, si queremos demostrar que $\mu > \mu_o$, interesa utilizar la hipótesis alternativa $H_1 : \mu > \mu_o$ y en este caso la región crítica debería ser $C = (z_\alpha, \infty)$ con $p - valor = p\left(T > T_o\right)$. Del mismo modo, si la hipótesis alternativa fuese $H_1 : \mu < \mu_o$, la región crítica debería ser $C = (-\infty, -z_\alpha)$ y $p - valor = p\left(T < T_o\right)$. Estos dos contrastes de hipótesis se denominan tests unilaterales, frente al test bilateral resuelto inicialmente.

Ejercicio. Consideramos de nuevo el ejemplo de la lluvia ácida utilizado en el capítulo anterior. Para demostrar que el problema realmente existe, construir un test de hipótesis para contrastar las hipótesis nula $H_o : \mu = 5.7$ frente a la hipótesis alternativa $H_1 : \mu < 5.7$ con un nivel de significación $\alpha = 0.01$ (recordar que con una muestra de tamaño $n = 40$ habíamos obtenido una media muestral $\bar{x} = 3.7$). Calcular el p-valor del test y construir también el test bilateral con su p-valor.

Test de hipótesis para el parámetro μ de una distribución normal con σ desconocida. Para el mismo test de hipótesis que en el caso anterior pero con σ desconocida, ahora el estadístico de prueba sería $T = \dfrac{\overline{X} - \mu_o}{S_c/\sqrt{n}} \rightsquigarrow t_{n-1}$ (suponiendo que la hipótesis nula es cierta, es decir, $\mu = \mu_o$) y por tanto la región crítica para el test bilateral sería $C = \left(-\infty, -t_{n-1;\alpha/2}\right) \cup \left(t_{n-1;\alpha/2}, \infty\right)$ de modo que $p\left(T \in C/\mu = \mu_o\right) = 2p\left(t_{n-1} > t_{n-1;\alpha/2}\right) = \alpha$. Si T_o es el valor del estadístico de prueba observado en la muestra, ahora tendríamos $p - valor = 2p\left(t_{n-1} > |T_o|\right)$. Del mismo modo, para los tests unilaterales, si $H_1 : \mu > \mu_o$, la región crítica debería ser $C = (t_{n-1;\alpha}, \infty)$ con $p - valor = p\left(t_{n-1} > T_o\right)$, y si $H_1 : \mu < \mu_o$ la región crítica debería ser $C = (-\infty, -t_{n-1;\alpha})$ con $p - valor = p\left(t_{n-1} < T_o\right)$.

Ejercicio. Resolver el mismo ejercicio anterior suponiendo ahora que la desviación típica σ es desconocida y en la muestra elegida hemos observado $s_c = 0.5$.

Test de hipótesis para la desviación típica σ de una distribución normal. En la misma situación que en los casos anteriores, consideramos ahora el test $H_o : \sigma = \sigma_o$ frente a $H_1 : \sigma \neq \sigma_o$ con un nivel de significación α. Como ya conocemos, si H_o es cierta, se verifica que $\dfrac{(n-1)S_c^2}{\sigma_o^2} = T \rightsquigarrow \chi^2_{n-1}$ y por tanto tenemos un estadístico de prueba apropiado para este problema y que proporciona la siguiente región crítica: $C = \left(0, \chi^2_{n-1;1-\alpha/2}\right) \cup \left(\chi^2_{n-1;\alpha/2}, \infty\right)$. Además, si T_o es el valor observado para este estadístico, el p-valor se calcula como:

$$p - valor = 2\,\text{mín}\left\{p\left(\chi^2_{n-1} < T_o\right), p\left(\chi^2_{n-1} > T_o\right)\right\}$$

Del mismo modo, para los tests unilaterales, si $H_1 : \sigma > \sigma_o$, la región crítica debería ser $C = \left(\chi^2_{n-1;\alpha}, \infty\right)$ con $p - valor = p\left(\chi^2_{n-1} > T_o\right)$, y si $H_1 : \sigma < \sigma_o$ la región crítica vendría dada por $C = \left(0, \chi^2_{n-1;1-\alpha}\right)$ con $p - valor = p\left(\chi^2_{n-1} < T_o\right)$.

Este test puede utilizarse en cualquier caso, tanto si, como es habitual, el valor esperado μ es desconocido como si lo suponemos conocido. No obstante, para este último caso se puede construir un test que resulta más eficaz. Concretamente, si consideramos el estadístico definido como $T = \dfrac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma_o^2}$ se verifica que $T \rightsquigarrow \chi^2_n$ y podemos utilizar la región crítica definida como $C = \left(0, \chi^2_{n;1-\alpha/2}\right) \cup \left(\chi^2_{n;\alpha/2}, \infty\right)$. El p-valor se calcularía del mismo modo que antes pero utilizando una distribución χ^2_n en lugar de χ^2_{n-1}.

Ejercicio. Considerando el mismo problema de la lluvia ácida utilizado en los casos anteriores, construir un test de hipótesis para contrastar $H_o : \sigma = 0.6$ frente a $H_1 : \sigma < 0.6$ con un nivel de significación $\alpha = 0.05$. Calcular el p-valor para este test y también para el test bilateral con $H_1 : \sigma \neq 0.6$.

Test de hipótesis para una proporción poblacional p. Si consideramos $X \rightsquigarrow B(p)$ y $X_1, X_2, ..., X_n$ una muestra aleatoria simple de X el test de hipótesis para contrastar la hipótesis nula $H_o : p = p_o$ frente a $H_1 : p \neq p_o$ con un nivel de significación α puede obtenerse con el estadístico test

$$T = \sum_{i=1}^{n} X_i \rightsquigarrow B(n, p_o)$$

La región crítica sería $C = \{0, ..., k_1\} \cup \{k_2, ..., n\}$ siendo k_1 el cuantil $\alpha/2$ y k_2 el cuantil $1 - \alpha/2$ de una distribución binomial $B(n, p_o)$. Estos valores deberían calcularse de modo que

$$\sum_{i=0}^{k_1} \binom{n}{i} p_o^i (1-p_0)^{n-i} \leq \frac{\alpha}{2} < \sum_{i=0}^{k_1+1} \binom{n}{i} p_o^i (1-p_0)^{n-i}$$

y

$$\sum_{i=k_2}^{n} \binom{n}{i} p_o^i (1-p_0)^{n-i} \leq \frac{\alpha}{2} < \sum_{i=k_2+1}^{n} \binom{n}{i} p_o^i (1-p_0)^{n-i}$$

Para los test unilaterales con $H_1 : p < p_o$ y $H_1 : p > p_o$ las regiones críticas serían, respectivamente, $C = \{0, ..., k_1\}$ y $C = \{k_2, ..., n\}$ siendo k_1 el cuantil α y k_2 el cuantil $1 - \alpha$ de una distribución binomial $B(n, p_o)$. Estos valores se calcularían ahora de modo que

$$\sum_{i=0}^{k_1} \binom{n}{i} p_o^i (1-p_0)^{n-i} \leq \alpha < \sum_{i=0}^{k_1+1} \binom{n}{i} p_o^i (1-p_0)^{n-i}$$

y

$$\sum_{i=k_2}^{n} \binom{n}{i} p_o^i (1-p_0)^{n-i} \leq \alpha < \sum_{i=k_2+1}^{n} \binom{n}{i} p_o^i (1-p_0)^{n-i}$$

Ejercicio. En el ejercicio de los estudiantes de derecho del capítulo anterior, construir el test para contrastar $H_o : p = 0.5$ frente a $H_1 : p > 0.5$, utilizando un nivel de significación $\alpha = 0.05$. ¿Cuál sería el p-valor de este test para la muestra obenida?

Test de hipótesis para el parámetro $\mu = EX$ de una variable aleatoria cualquiera X. En este caso, si el tamaño muestral es grande, podemos utilizar el estadístico $T = \dfrac{\overline{X} - \mu_o}{S_c/\sqrt{n}}$ cuya distribución de probabilidad es, asintóticamente, $N(0,1)$ como ya comentamos en el capítulo anterior. Utilizando entonces la región crítica $C = \left(-\infty, -z_{\alpha/2}\right) \cup \left(z_{\alpha/2}, \infty\right)$ tendremos un test de hipótesis aproximado. Según cuál sea la hipótesis alternativa, los p-valores aproximados podrían calcularse entonces de la forma habitual.

Test de hipótesis asintótico para una proporción poblacional p. Es un caso particular del test anterior teniendo en cuenta que $p = EX$ siendo X una variable aleatoria de Bernouilli de parámetro p. En este caso, si la hipótesis nula es $H_o : p = p_o$ y $\hat{p}$ representa a la proporción muestral, el estadístico de prueba asintótico sería $T = \dfrac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}} \rightsquigarrow N(0,1)$ y podríamos calcular la región crítica y el p-valor de la forma habitual.

Ejercicio. En el ejercicio de los estudiantes de derecho del capítulo anterior, construir el test asintótico para contrastar $H_o : p = 0.5$ frente a $H_1 : p > 0.5$, utilizando un nivel de significación $\alpha = 0.05$. ¿Cuál sería el p-valor de este test para la muestra obenida?.

Test de hipótesis para la diferencia de medias en dos poblaciones normales independientes. Al igual que en el capítulo anterior, si tenemos dos variables aleatorias $X \rightsquigarrow N(\mu_1, \sigma_1)$ e $Y \rightsquigarrow N(\mu_2, \sigma_2)$ con σ_1 y σ_2 conocidas y disponemos de sendas muestras aleatorias simples independientes entre sí, $X_1, X_2, ..., X_{n_1}$ para X e $Y_1, Y_2, ..., Y_{n_2}$ para Y, podemos construir un test de hipótesis para contrastar la hipótesis nula $H_o : \mu_1 - \mu_2 = \Delta$ utilizando el estadístico de prueba $T = \dfrac{(\overline{X} - \overline{Y}) - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightsquigarrow N(0,1)$. si las varianzas son conocidas. Si las varianzas son desconocidas pero iguales utilizaríamos el estadístico $\dfrac{(\overline{X} - \overline{Y}) - \Delta}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightsquigarrow t_{n_1+n_2-2}$ y finalmente, si son desconocidas y distintas, utilizaríammos $\dfrac{(\overline{X} - \overline{Y}) - \Delta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \rightsquigarrow t_d$, donde el valor d se calcularía como en el capítulo anterior. Utilizando esos estadísticos, la región crítica y el p-valor se calcularían de la forma habitual según cuál sea la hipótesis alternativa (test de dos lados o test de un lado).

Ejercicio. Considerar de nuevo el ejercicio de los atletas universitarios del capítulo anterior ($n_1 = 9$, $\bar{x} = 43.71$, $s_1 = 5.88$, $n_2 = 7$, $\bar{y} = 39.63$ y $s_2 = 7.68$). Construir un test de hipótesis para contrastar $H_o : \mu_1 - \mu_2 = 0$ frente a $H_1 : \mu_1 - \mu_2 \neq 0$ con un nivel de significación $\alpha = 0.1$ suponiendo que las varianzas son desconocidas pero iguales. ¿Cual seria el p-valor para este test?

Test de hipótesis para el cociente de varianzas en dos poblaciones normales independientes. Considerando la misma situación que en el caso anterior, podemos construir un test de hipótesis para $H_o : \frac{\sigma_1^2}{\sigma_2^2} = \Delta$ (equivalentemente, $H_o : \frac{\sigma_1}{\sigma_2} = \sqrt{\Delta}$) frente a la hipótesis alternativa $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq \Delta$. El estadístico de prueba sería ahora $T = \dfrac{S_1^2/S_2^2}{\Delta} \rightsquigarrow F_{n_1-1, n_2-1}$ y la región crítica para el nivel de significación

α sería

$$C = \left(0, F_{n_1-1,n_2-1;1-\alpha/2}\right) \cup \left(F_{n_1-1,n_2-1;\alpha/2}, \infty\right) = \left(0, \frac{1}{F_{n_2-1,n_1-1;\alpha/2}}\right) \cup \left(F_{n_1-1,n_2-1;\alpha/2}, \infty\right)$$

Si T_o es el valor observado del estadístico de prueba, el p-valor se calcularía entonces como

$$\begin{aligned} p-valor &= 2\,\text{mín}\left\{p\left(F_{n_1-1,n_2-1} < T_o\right), p\left(F_{n_1-1,n_2-1} > T_o\right)\right\} \\ &= 2\,\text{mín}\left\{p\left(F_{n_2-1,n_1-1} > \frac{1}{T_o}\right), p\left(F_{n_1-1,n_2-1} > T_o\right)\right\} \end{aligned}$$

Para los tests de un lado, si $H_1 : \frac{\sigma_1^2}{\sigma_2^2} > \Delta$, la región crítica sería $C = (F_{n_1-1,n_2-1;\alpha}, \infty)$ y $p-valor = p\left(F_{n_1-1,n_2-1} > T_o\right)$, mientras que se si $H_1 : \frac{\sigma_1^2}{\sigma_2^2} < \Delta$ serían $C = \left(0, \frac{1}{F_{n_2-1,n_1-1;\alpha}}\right)$ y $p-valor = p\left(F_{n_2-1,n_1-1} > \frac{1}{T_o}\right)$.

Ejercicio. En el mismo caso del ejercicio anterior de los atletas universitarios, construir el test $H_o : \frac{\sigma_1^2}{\sigma_2^2} = 1$ frente a $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$ con un nivel de significación $\alpha = 0.1$. ¿Cuál sería el p-valor del test? ¿Podemos aceptar que las varianzas son iguales?

Test de hipótesis para la diferencia de medias de dos variables aleatorias independientes con distribución cualquiera. Si consideramos ahora que las variables aleatorias independientes X e Y tienen distribuciones de probabilidad cualesquiera (no necesariamente normales) pero los tamaños muestrales n_1 y n_2 son suficientemente grandes también podemos construir un test de hipotesis para la diferencia $\mu_1 - \mu_2$ utilizando una distribución de probabilidad asintótica. Concretamente, para el test de hipótesis $H_o : \mu_1 - \mu_2 = \Delta$, teniendo en cuenta el Teorema Central del Límite y la propiedad de consistencia de la varianza muestral corregida como estimador de la varianza teórica, podemos decir que, cuando los tamaños muestrales tienden hacia ∞ y la hipótesis nula es cierta, se verifica que

$$\frac{\left(\overline{X} - \overline{Y}\right) - \Delta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \widetilde{\rightsquigarrow} N(0,1)$$

y razonando de la forma habitual podemos construir la región crítica y calcular los p-valores según que queramos resolver el test de dos lados o el test de un lado.

Test de hipótesis para la diferencia de dos proporciones poblacionales p_1 y p_2. Es un caso particular del anterior, teniendo en cuenta que $p_1 = EX$ y $p_2 = EY$ siendo $X \rightsquigarrow B(p_1)$ e $Y \rightsquigarrow B(p_2)$ con X, Y independientes. En este caso, para el test de hipótesis $H_o : p_1 - p_2 = \Delta$, el estadístico de prueba sería

$$\frac{(\hat{p}_1 - \hat{p}_2) - \Delta}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \widetilde{\rightsquigarrow} N(0,1)$$

siendo $\hat{p}_1$ y $\hat{p}_2$ las proporciones muestrales de ambas muestras. A partir de este estadístico podemos resolver el test deseado de la forma habitual.

En el caso especial de que $\Delta = 0$ (es decir, $H_o : p_1 = p_2$) el test anterior puede mejorarse teniendo en cuenta que, si la hipótesis H_o es cierta, podemos obtener una mejor estimación de la proporción

poblacional común utilizando $\hat{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$ y entonces un estadístico de prueba más apropiado sería

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \widetilde{\rightsquigarrow} N(0,1)$$

La región crítica y el p-valor se calcularían de la forma habitual.

Ejercicio. Utilizando el mismo ejercicio del capítulo anterior de los pinos atacados por la procesionaria en dos localidades distintas ($\hat{p}_1 = \frac{16}{51}$ y $\hat{p}_1 = \frac{8}{51}$), resolver el test de hipótesis $H_o : p_1 = p_2$ frente a $H_1 : p_1 > p_2$ con un nivel de significación $\alpha = 0.01$. ¿Cuál sería el p-valor del test? ¿Podemos afirmar con un 99 % de confianza que el porcentaje de pinos atacados es mayor en la primera localidad?

Test de hipótesis para la diferencia de medias en dos variables aleatorias con distribución normal pero utilizando muestras pareadas. Supongamos ahora que las dos muestras aleatorias simples $X_1, X_2, ..., X_n$ de la variable $X \rightsquigarrow N(\mu_1, \sigma_1)$ e $Y_1, Y_2, ..., Y_n$ de la variable $Y \rightsquigarrow N(\mu_2, \sigma_2)$ no son independientes, sino que son pareadas, y queremos contrastar la hipótesis nula $H_o : \mu_1 - \mu_2 = \Delta$. Al igual que en el capítulo anterior para este mismo caso, si definimos la nueva variable aleatoria $D = X - Y$, podemos asegurar que $D \rightsquigarrow N(\mu_1 - \mu_2, \sigma_D)$ con un cierto parámetro σ_D desconocido y que, a priori no podemos evaluar a partir de las desviaciones típicas σ_1 y σ_2. No obstante, disponemos de una muestra aleatoria simple $D_1, D_2, ..., D_n$ de la variable aleatoria D, siendo $D_i = X_i - Y_i$ y podemos utilizar la teoría estudiada para construir un test de hipotesis para el valor esperado $\mu = ED = \mu_1 - \mu_2$ utilizando una muestra aleatoria simple con varianza desconocida σ_D. Concretamente, si H_o es cierta, se verifica que $\overline{D} = \overline{X} - \overline{Y}$ y $\dfrac{\overline{D} - \Delta}{S_D/\sqrt{n}} = \dfrac{\left(\overline{X} - \overline{Y}\right) - \Delta}{S_D/\sqrt{n}} \rightsquigarrow t_{n-1}$, siendo S_D la desviación típica muestral corregida de la muestra de las diferencias. A partir de este estadístico de prueba podemos resolver el test de la forma habitual, según que utilizemos la prueba de dos lados o la prueba de un lado.

Ejercicio. Considerar el mismo ejercicio del capítulo anterior referente a la alimentación para las gallinas. Construir un test de hipótesis para constrastar la hipótesis nula $H_o : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 < \mu_2$ utilizando un nivel de significación $\alpha = 0.05$. ¿Cuál sería el p-valor de este test?

9.1. PROBLEMAS

1. Un investigador ha preparado el nivel de dosificación de un fármaco que afirma provocará el sueño en por lo menos el 80 % de las personas que padecen insomnio. Después de examinar la dosificación, se considera que su afirmación acerca de la efectividad del fármaco es exagerada. En un intento de refutar su afirmación se administra la dosificación prescrita a 20 personas que padecen insomnio, y se observa $Y =$"Número de personas que se adormecen debido al fármaco". Se desea realizar el contraste de hipótesis $H_o : p = 0.8$ frente a $H_1 : p < 0.8$, utilizando la región crítica $C = \{Y \leq 12\}$.

   a) Calcular el nivel de significación del test utilizado.

   b) Encontrar la potencia para $p = 0.6$ y $p = 0.4$.

c) ¿Cuál es la probabilidad de rechazar la hipótesis nula cuando es cierta? ¿Cuál es la probabilidad de no rechazar la hipótesis cuando p toma los valores del apartado anterior?

d) Se desea contrastar la hipótesis a un nivel de significación $\alpha = 0.01$. ¿Cuál es la región crítica que se debería utilizar? Responder a los dos apartados anteriores utilizando este nuevo test.

2. El voltaje de salida de un cierto circuito eléctrico debe ser igual a 130, segun las especificaciones. Una muestra de 40 lecturas independientes para este circuito dio una media muestral $\bar{x} = 128.6$. Suponiendo que los datos provienen de una población normal con desviación típica $\sigma = 2.1$, contrastar la hipótesis de que el voltaje de salida promedio es de 130 frente a la alternativa de que es menor, con un nivel de significación $\alpha = 0.05$. ¿Cuál es el p-valor que nos da la muestra obtenida para este test?

3. Para determinar el índice de dureza Rockwell del acero se oprime una punta de diamante en acero y se mide la profundidad de penetración. Para 50 muestras de un determinado acero se obtuvo un índice de dureza promedio $\bar{x} = 62$. El fabricante afirma que el acero tiene un índice promedio de dureza de por lo menos 64. Suponiendo que los datos provienen de una distribución normal con desviación típica $\sigma = 8$, contrastar la afirmación con un nivel de significación $\alpha = 0.01$. Calcular el p-valor para la muestra obtenida.

   El acero tiene la dureza suficiente cuando la media del índice de dureza Rockwell no baja de 60. En el test construido anteriormente, calcular la probabilidad de cometer un error de tipo II en la alternativa $\mu = 60$. Para contrastar $H_o : \mu = 64$ frente a $H_1 : \mu = 60$, determinar el tamaño de muestra necesario para que $\alpha = 0.01$ y $\beta = 0.05$.

4. El pH del agua que sale de una planta depuradora debe ser 7.0 por especificación. Se tomaron de forma independiente 30 muestras de agua de esa planta y se obtuvo un promedio de pH de $\bar{x} = 6.8$ y un desviación típica muestral corregida $s_c = 0.9$. Suponiendo que el pH del agua tiene una distribución normal, ¿hay razón para dudar del cumplimiento de la especificación de la planta? Ver la significatividad de los datos a un nivel $\alpha = 0.05$. Calcular el p-valor del test.

5. Se afirma que la dureza, en grados Shore, de un determinado caucho debe ser 65. Se tomaron 14 pruebas independientes y se obtuvo $\bar{x} = 63.1$ con $s_c = 1.4$. ¿Hay evidencia suficiente como para rechazar la afirmación anterior a un nivel de significación $\alpha = 0.05$? ¿Qué hipótesis se necesita asumir para que la respuesta sea válida?

6. La dispersión de los tiempos de descarga de material de un proyecto de construcción es de gran importacia, ya que si estos tiempos son muy variables es muy difícil organizar las tareas de forma eficiente. El encargado del transporte de una obra afirma que la diferencia entre los tiempos máximo y mínimo de descarga no debe ser mayor de 40 minutos. Si se supone que estos tiempos están distribuidos de forma aproximadamente normal, el encargado de la obra cree que esta afirmación quiere decir que la desviación típica σ debe ser aproximadamente 10 minutos. Se midieron, en

minutos, 15 tiempos en los que se obtuvo $\bar{x} = 142$ con $s_c = 12$. ¿Podría rechazarse la hipótesis de que la desviación típica es igual a 10 con un nivel de significación $\alpha = 0.05$? Calcular el p-valor del test.

7. Las pruebas de aptitud deben producir calificaciones con una gran cantidad de variación para que se pueda distinguir a las personas con pocas aptitudes de aquellas que tienen muchas. La prueba normal que se emplea en una determinada empresa ha dado calificaciones con una desviación típica $\sigma = 5$ puntos. Se ensaya una prueba en 20 aspirantes, obteniéndose una muestra con $s_c = 8$ puntos. ¿Son las calificaciones de la prueba apreciablemente más variables que la prueba normal? Usar un nivel de significación $\alpha = 0.05$ y calcular el p-valor de la prueba.

8. En una encuesta realizada entre 2207 agricultores, el 54 % respondieron que el sistema de ayudas a la producción tiene unos trámites demasiado complicados. ¿Se puede concluir que la mayoría de los agricultores piensan que el sistema de ayudas es muy complicado? Utilizar un nivel de significación $\alpha = 0.05$ y calcular el p-valor. Calcular la potencia del test en la alternativa dada por que una tercera parte de lo agricultores piensan que los trámites a seguir son demasiado complicados.

9. Mediciones respecto del esfuerzo cortante obtenidas a partir de pruebas de compresión independientes para dos tipos de suelos dieron los siguientes resultados (en toneladas por pie cuadrado):

	Suelo I	Suelo II
Tamaño muestral	30	35
Media muestral	1.65	1.43
Desviación típica corregida	0.26	0.22

¿Son los datos significativamente diferentes a un nivel $\alpha = 0.01$. Calcular el p-valor del test.

10. Unos cohetes se fabrican con un supuesto alcance de 2500 metros. Teóricamente se supone que el alcance se reduce cuando se almacenan durante algún tiempo. Seis de esos cohetes se almacenaron durante un determinado periodo y a continuación se probaron. Los alcances obtenidos fueron los siguientes: 2490, 2510, 2360, 2410, 2300 y 2440. ¿Es más corto el alcance después del almacenamiento? Suponiendo distribución normal, contrastar la hipótesis con un nivel de significación del 1 % y calcular el p-valor de la prueba para los datos obtenidos.

    En este problema, la variabilidad de los alcances es también importante. Los cohetes nuevos tienen una desviación típica $\sigma = 50$ metros. ¿Podemos decir que el almacenamiento aumenta la variabilidad de los alcances? Usar un nivel de significación $\alpha = 0.05$ y calcular el p-valor del test.

11. Se efectuó un estudio por parte de una Comisión de Caza y Pesca para estimar las cantidades de residuos químicos encontrados en los tejidos cerebrales de pelícanos. En una prueba sobre mediciones de DDT, muestras aleatorias de 10 pelícanos jóvenes y 13 polluelos dieron los siguientes resultados:

	Jóvenes	Polluelos
Tamaño muestral	10	13
Media muestral	0.041	0.026
Desviación típica corregida	0.017	0.006

Suponiendo que las mediciones de DDT provienen de distribuciones normales con la misma desviación típica, contrastar la hipótesis de que no existe diferencia en las cantidades promedio de DDT encontradas en los pelícanos jóvenes y en los polluelos, frente a la alternativa de que los pelícanos jóvenes presentan un mayor promedio. Utilizar un nivel de significación $\alpha = 0.05$ y calcular el p-valor. ¿Son los datos estadísticamente significativos a un nivel del 1 %?

¿Hay suficiente evidencia, a un nivel de significación del 5 %, para concluir que la varianza en las mediciones de los niveles de DDT es mayor en los pelícanos jóvenes que en los polluelos? ¿Qué implicaciones tiene en el contraste realizado con anterioridad? De acuerdo con esto, volver a realizar el test anterior sin suponer que las varianzas son iguales.

12. Se compararon dos proyectos para la construcción de un laboratorio con respecto a la cantidad media de luz que se tiene en la superficies de las mesas. Se tomaron 40 mediciones independientes en cada laboratorio obteniéndose los siguientes resultados:

	Diseño I	Diseño II
Tamaño muestral	40	40
Media muestral	28.8	32.6
Desviación típica corregida	15.1	15.8

Suponiendo que los datos provienen de dos poblaciones con distribución normal de varianzas desconocidas (pero iguales), contrastar la igualdad de los promedios con un nivel de significación $\alpha = 0.05$ y calcular el p-valor del test.

13. Los datos obtenidos en una prospección geológica del Departamento de Interior de los Estados Unidos presentan los caudales de un río pequeño en el norte de Florida. El interés está en comparar los caudales de marzo y abril, puesto que son dos meses relativamente secos. Tomadas 31 observaciones para el mes de marzo dieron como resultado $\bar{x} = 6.85$ con $s_1 = 1.2$ pies por segundo. Las 30 mediciones de abril dieron $\bar{y} = 7.47$ con $s_1 = 2.3$ pies por segundo. ¿Hay evidencia suficiente para decir que esos dos meses tuvieron diferentes caudales a un nivel de significación del 5 %? Calcular el p-valor que determinan los datos. ¿Te parece adecuada la hipótesis de igualdad de varianzas con estos datos? (responder utilizando un test de hipótesis con $\alpha = 0.05$ y calcular el p-valor de la prueba).

14. La retención del nitrógeno por el suelo es un aspecto importante en los métodos de cultivo, incluyendo el cultivo de bosques. Se compararon dos métodos de preparación de cepas para plantar pinos después de limpiar el terreno, analizando el porcentaje de nitrógeno marcado que se recupera. El método A deja intanto gran parte del piso del bosque, mientras que el método B remueve la mayor parte de la materia orgánica. Está claro que con el método B se recuperará una cantidad menor

de nitrógeno del piso del bosque. El asunto de interés es si el método B hará que la retención de nitrógeno sea mayor en la biomasa microbiana en compensación por tener menos material orgánico disponible. El porcentaje de nitrógeno recuperado en la biomasa microbiana se midio en 6 cepas de prueba para cada método. Las cepas del método mostraron un promedio de 12 y una desviación típica corregida de 1. Las cepas del método B mostraron un promedio 15 y una desviación típica corregida de 2. A un nivel de significación del 10 %, ¿se podría decir que el porcentaje promedio de nitrógeno recuperado es mayor para el método B? Calcular el p-valor del contraste utilizado para los datos obtenidos.

15. En una revista se afirma que "las técnicas y los dispositivos de biorretroalimentación para controlar las funciones fisiológicas ayudan a los astronautas a controlar la tensión". Seis personas se sometieron a una situación de tensión (mediante videojuegos) seguidos por un periodo de biorretroalimentación y adaptación. Otro grupo de seis personas fueron sometidas a la misma tensión, y después se les dijo que simplemente se relajaran. El primer grupo obtuvo 70.4 latidos por minuto de promedio, con una desviación típica corregida de 15.3, mientras el segundo grupo tuvo un promedio de 74.9 latidos por minuto, con una desviación típica corregida de 16.0. ¿Se puede decir que los latidos cardíacos promedio con biorretroalimentación son menos rápidos que sin la biorretroalimentación? ¿Qué hipótesis se necesita asumir para que la respuesta sea válida? Calcular el p-valor del contraste.

16. Se probaron dos máquinas A y B para estudiar la resistencia a la torsión de un alambre de acero en 12 pares distintos de alambre. En cada máquina se probó un alambre de cada uno de los pares. Los resultados de la resistencia a la torsión (punto de ruptura) en cada par fueron los siguientes:

Alambre	1	2	3	4	5	6	7	8	9	10	11	12
Máquina A	32	35	38	28	40	42	36	29	33	37	22	42
Máquina B	30	34	39	26	37	42	35	30	30	32	20	41

    a) ¿Hay significatividad a un nivel del 5 % de que las máquinas A y B dan lecturas diferentes?

    b) ¿Hay significatividad a un nivel del 5 % de que la máquina B de una lectura menor que la máquina A?

    En ambos casos calcular el p-valor para los datos obtenidos.

17. Supongamos que se está planeando un experimento en un invernadero para estudiar el crecimiento de las plantas de pimiento. Se plantan n de estas plantas en tierra normal, y otras n en tierra tratada. Después de 21 días se mide la longitud del tallo de cada una de las plantas. Si el efecto del tratamiento de la tierra es que se incrementa la longitud media de los tallos en 2 cm., en el correspondiente contraste se debería tener una probabilidad del 90 % de rechazar la hipótesis nula con el test de la t de Student de un lado. Los datos de un estudio previo en 15 plantas crecidas en tierra normal muestran una media muestral de 12.5 cm. y una desviación típica muestral corregida de 0.8 cm.

a) Si se quiere utilizar un test a un nivel de significación $\alpha = 0.05$, ¿cuál es el valor de n que se debe tomar?

b) ¿Qué condiciones deben cumplir los datos para que los cálculos del apartado anterior sean válidos?

c) Si se decide adoptar una postura más conservadora y utilizar un nivel de significación $\alpha = 0.01$, ¿qué valor de n se debe utilizar?

18. El siguiente problema trata sobre la eficiencia en la combustión que cabe esperar de un calentador de petróleo. el *Environmental News* (enero de 1977) afirma que el 80 % o más es excelente, del 75 % al 79 % bueno, del 70 % al 74 % mediano y menos del 70 % malo. Un contratista de calefacción que vende dos tipos de calentadores (A y B) decidió comparar las eficiencias de estos dos tipos de calentadores. Los datos obtenidos en una muestra de cada tipo son los siguientes:

    Tipo A: 72 78 73 69 75 74 69 75

    Tipo B: 78 76 81 74 82 75

    Contrastar la diferencia en la eficiencia promedio de ambos tipos de calentador, suponiendo que en ambos casos se tiene una distribución normal y utilizando un nivel de significación $\alpha = 0.05$.

19. Para comprobar la potencia en cuanto a vitamina D de un aceite de hígado de pescado, se alimentó a 20 polluelos de un día con un alimento al que le fue añadido un 1 % de este aceite como única fuente de vitamina D. Otros 10 polluelos fueron alimentados con el mismo alimento base, pero añadiendo un 1 % de un aceite de hígado de bacalao (estándar) del que se conoce su potencia en cuanto a la vitamina D. Después de 8 días los pesos de los polluelos (en gramos) fueron los siguientes:

    Aceite estándar: 69, 66, 71, 69, 68, 71, 74, 73, 70, 69

    Nuevo aceite: 68, 69, 66, 64, 67, 69, 65, 68, 70, 65, 63, 66, 68, 66, 69, 67, 66, 61, 71, 72

    Contrastar la hipótesis de que el nuevo aceite es mejor utilizando un nivel de significación $\alpha = 0.05$ y obtener el p-valor que proporcionan los datos (suponer distribuciones normales).

20. La cantidad de oxígeno que puede inspirar una persona representa la mayor cantidad de oxígeno que puede consumir dicha persona. Se llevó a cabo un test de fatiga para determinar la capacidad pulmonar de 9 niñas en edad escolar antes y después de un programa de 10 semanas de ejercicio físico. Los resultados expresados en ml. de oxígeno por minuto y por kg. de peso fueron:

Individuo	1	2	3	4	5	6	7	8	9
Antes	48.6	38.0	31.2	45.5	41.7	41.8	37.9	39.2	47.2
Después	38.8	40.7	32.0	45.4	43.2	45.3	38.9	43.5	45.0

    Contrastar la hipótesis de que existe diferencia entre las cantidades de oxígeno consumidad "Antes" y "Después" del test de fatiga con un nivel de significación $\alpha = 0.05$ y obtener el p-valor que proporcionan los datos (suponer distribuciones normales).

21. Los siguientes datos muestran 30 medidas del voltaje eléctrico tomadas durante un cierto proceso industrial. El proceso se considera satisfactorio cuando el voltaje es mayor que 9.2 voltios (las lecturas más grandes son mejores que las más pequeñas).

Vieja localización			Nueva localización		
9.98	10.12	9.84	9.19	10.01	8.82
10.26	10.05	10.15	9.63	8.82	8.65
10.05	9.80	10.02	10.10	9.43	8.51
10.29	10.15	9.80	9.70	10.03	9.14
10.03	10.00	9.73	10.09	9.85	9.75
8.05	9.87	10.01	9.60	9.27	8.78
10.55	9.55	9.98	10.05	8.83	9.35
10.26	9.95	8.72	10.12	9.39	9.54
9.97	9.70	8.80	9.49	9.48	9.36
9.87	8.72	9.84	9.37	9.64	8.68

Contrastar la hipotesis de si hay diferencia entre la nueva y la vieja localización con un nivel de significación $\alpha = 0.01$ y calcular el p-valor que proporcionan los datos. ¿Es mejor la nueva localización que la vieja?