

Fundamentos de Inferencia Estadística

Alex Sanchez Pla y Santiago Pérez Hoyos

2024-11-17

Contents

Presentación	3
Objetivo	3
Prerequisitos y organización del material	4
Referencias	4
Agradecimiento y fuentes utilizadas	4
El proyecto Statmedia	4
Otros materiales utilizados	5
1 Probabilidad y Experimentos aleatorios	5
1.1 Introducción	5
1.2 Función de probabilidad	6
1.3 ¿Cómo se calculan las probabilidades?	7
1.4 Sucesos elementales y sucesos observables	7
1.5 Propiedades inmediatas de la probabilidad	8
1.6 Espacios de probabilidad	9
1.7 Probabilidad condicionada	9
1.8 Dos Teoremas importantes	11
1.9 Introducción a los experimentos múltiples	11
1.10 Combinatoria	12
1.11 Frecuencia relativa y probabilidad	14
1.12 Caso de Estudio: Eficacia de una prueba diagnóstica	15
2 Variables aleatorias y Distribuciones de probabilidad	16
2.1 El espacio muestral y sus elementos	17
2.2 Representación numérica de los sucesos elementales. Variables aleatorias	17
2.3 Caracterización de una variable aleatoria a través de la probabilidad. Función de distribución	18
2.4 Propiedades de la función de distribución	18
2.5 Clasificación de las variables aleatorias	19
2.6 Variable aleatoria discretas	21
2.7 Variables aleatorias continuas	23
2.8 Caracterización de una variable aleatoria a través de parámetros	27
2.9 Esperanza de una variable aleatoria discreta	27
2.10 Esperanza de una variable aleatoria continua	27
2.11 Propiedades de la esperanza matemática	28
2.12 Varianza de una variable aleatoria	28
2.13 Momentos (de orden k) de una variable aleatoria	29
2.14 Definición formal de variable aleatoria	29
2.15 Caso práctico: Lanzamiento de dos dados	30
3 Distribuciones Notables	40

3.1	Distribuciones discretas	40
3.2	Distribuciones Continuas	50
3.3	Distribuciones con R (y Python)	56
3.4	La familia exponencial de distribuciones	57
4	Distribuciones de probabilidad multidimensionales	60
4.1	Distribuciones conjuntas de probabilidades	60
4.2	Variable aleatorias bivariantes discretas	62
4.3	La distribución multinomial	66
4.4	Distribuciones marginales	68
4.5	Distribuciones condicionales	68
4.6	Vectores aleatorios absolutamente continuos	69
4.7	Independencia de variables aleatorias	72
4.8	Momentos de vectores aleatorios	73
5	Grandes muestras	77
5.1	Introducción: Aproximaciones asintóticas	77
5.2	Ley de los Grandes Números (Ley débil)	77
5.3	El teorema central del límite	79
6	Introducción a la inferencia estadística	85
6.1	Inferencia estadística	85
6.2	Problemas de inferencia estadística	85
6.3	Distribución de la población	85
6.4	Muestra aleatoria simple	86
6.5	Estadísticos	88
6.6	Distribución en el muestreo de un estadístico	89
6.7	La distribución empírica	90
6.8	Los momentos muestrales	92
6.9	Distribución en el muestreo de los momentos muestrales	92
6.10	Propiedades asintóticas de los momentos muestrales	93
6.11	Muestreo en poblaciones normales	94
7	Estimación puntual	96
7.1	El problema de la estimación puntual	96
7.2	Estudio de las propiedades deseables de los estimadores	99
7.3	Propiedades de los estimadores consistentes	103
7.4	Información de Fisher y cota de CramerRao	104
7.5	Información y verosimilitud de un modelo estadístico	105
7.6	Información de Fisher	105
7.7	La desigualdad de Cramer-Rao	107
7.8	Caracterización del estimador eficiente	110
7.9	Estadísticos suficientes	111
7.10	Obtención de estimadores	115
7.11	El método de los momentos	115
7.12	El método del máximo de verosimilitud	117
8	Estimación por intervalos	121
8.1	Preliminares: estimación del error estándar e Introducción al bootstrap	122
8.2	Estimadores por intervalo: intervalos de confianza	122
8.3	Intervalos de confianza para características de una población normal (media, varianza), . . .	122
8.4	Intervalos de confianza bootstrap.	122
8.5	Intervalos de confianza para proporciones binomiales	122
8.6	Intervalos de confianza para parámetros en muestra grandes y para casos generales (tasas, OR, ...)	122

8.7	Aplicaciones: cálculo del tamaño muestral	122
9	Pruebas de hipótesis	122
9.1	Conceptos básicos: pruebas de hipótesis y de significación, pruebas unilaterales y bilaterales, tipos de error, valores críticos de test y p-valores	122
9.2	Potencia de un test. Cálculos de potencia y de tamaño de la muestra. Tamaño del efecto. . .	122
9.3	Métodos de construcción de tests.	122
9.4	Problemas asociados al uso de tests estadísticos. La crisis de la significación	122
10	Inferencia Aplicada	122
10.1	Pruebas de normalidad. Pruebas gráficas. El test de Shapiro-Wilks	123
10.2	Pruebas de hipótesis para contrastar variables cuantitativas: pruebas paramétricas t-test y Anova	123
10.3	Pruebas de hipótesis para contrastar variables cuantitativas: pruebas de hipótesis no paramétricas de Wilcoxon y Kruskal-Wallis	123
10.4	Contrastes para datos categóricos. Pruebas binomiales, ji cuadrado y test de Fisher.	123
10.5	Riesgo relativo y razón de «odds»	123
11	Computación Intensiva y <i>Multiple Testing</i>	123
11.1	Tests de permutaciones; ¿Qué?, ¿Cuándo?, ¿Cómo?	123
11.2	El bootstrap en contraste de hipótesis	123
11.3	El problema de las comparaciones múltiples	123
11.4	Métodos de control de error: FWER y FDR	123

Presentación

Objetivo

El objetivo de estas notas es presentar un material de soporte para la asignatura de “Inferencia Estadística” del Máster interuniversitario de Bioestadística y Bioinformática impartido conjuntamente por la Universitat Oberta de Catalunya (UOC) y la Universidad de Barcelona (UB).

Esta asignatura adolece de las características habituales de las asignaturas de posgrado, y especialmente de un posgrado de estadística (y bioinformática), que muestran algunas de las cosas que no debe de ser esta asignatura:

- No puede ser un primer curso de estadística, porque se supone que los estudiantes del máster ya lo han cursado en sus grados. Por no decir que, a quien viene a especializarse en estadística se le puede suponer una base mínima.
- Tampoco debe ser como los segundos cursos de estadística de algunos grados, que tratan temas como la regresión, el diseño de experimentos o el análisis multivariante, porque esto ya se trata en diversas asignaturas del máster.

¿Que debemos pues esperar que sea este curso?

- Puestos a pedir, este curso debería servir para repasar y consolidar los conceptos básicos que la mayoría de estudiantes traerán consigo.
- Además, y sobretodo, debe proporcionar una visión general, lo más completa posible dentro de las limitaciones de tiempo, del campo de la inferencia estadística
- Y, naturalmente, esto significa proporcionar aquellos conceptos sobre los que se apoyaran muchas de las restantes asignaturas como “Regresión modelos y métodos”, “Diseño de Experimentos”, “Análisis Multivariante”, “Análisis de la Supervivencia” o “Análisis de datos ómicos”.

Prerequisitos y organización del material

Uno de los problemas “eternos” en el estudio de la estadística ha sido siempre la falta de acuerdo, entre la comunidad de docentes, de cual debería ser el nivel matemático a que se impartan los cursos.

En los cursos de pre-grado ha habido un cierto consenso, y con los años el nivel de formalismo ha disminuido, incluso en estudios de tipo “STEM”, tendiendo a centrarse en la aplicación de los conceptos, por ejemplo usando R, más que en un tratamiento formal (“matemático”) de los mismos.

Aunque esto puede ser práctico para aquellos estudios en los que la estadística es una asignatura de un grado, es también obvio que dicha aproximación no permite profundizar en muchos de los puntos que se tratan.

Es por ello que en este curso seguiremos la indicación habitual en cursos similares de asumir que el estudiante:

- Se siente cómodo con el lenguaje algebraico, desarrollo de expresiones, sumatorios etc.
- Está familiarizado con el cálculo diferencial en una o varias variables, aunque esta familiaridad no será imprescindible para seguir la mayoría de los contenidos del curso.
- Conoce el lenguaje estadístico R, que en muchas ocasiones nos ofrecerá una solución directa a los problemas de cálculo.

Referencias

Los requisitos anteriores corresponden básicamente a las matemáticas del bachillerato. Algunas fuentes adicionales pueden ser:

- Iniciación a las matemáticas para la ingeniería. M. Besalú y Joana Villalonga
 - Colección de (100) videos de soporte a las matemáticas para la ingeniería

Agradecimiento y fuentes utilizadas

Salvo que uno desee escribir un libro sobre algo muy extraño, siempre habrán otros libros o manuales similares al que se está planteando.

La respuesta a la pregunta, “Y entonces, ¿porque hacer un nuevo material?” suele ser más una excusa que una explicación sólida.

Una posible razón puede ser *para ajustarlo al máximo al perfil del curso para al que se destinan dichos materiales*, condición que otros textos, pensados para cursos y audiencias distintas, pueden no satisfacer. En este caso adoptaremos esta explicación y el tiempo decidirá si el objetivo se alcanza.

Dicho esto, debemos agradecer a las distintas fuentes utilizadas, el que hayan puesto a disposición sus materiales para poder reutilizarlos. Entre estos destacamos:

El proyecto Statmedia

Statmedia es un grupo de innovación docente de la Universidad de Barcelona, cuyo objetivo es desarrollar nuevas herramientas que ayuden en la enseñanza de la estadística aplicada, mejorando así el rendimiento académico de los alumnos y su motivación hacia la estadística.

Partiendo de la idea que el aprendizaje debe basarse en casos prácticos para motivar y fomentar la participación de los estudiantes. Se desarrolló primer proyecto, Statmedia I, un texto multimedia de estadística que además de los contenidos, relativamente ampliados, para un curso de introducción a la estadística, incorporaba:

- Una serie de casos para motivar e ilustrar los conceptos introducidos.
- Un conjunto de gadgets interactivos con los que interactuar y experimentar y
- Ejercicios de respuesta múltiple para verificar los conceptos trabajados.

Aunque el proyecto Statmedia ha seguido evolucionando en múltiples direcciones, Statmedia I, como tantos otros, no sobrevivió al desarrollo tecnológico, y la evolución (o decadencia) del lenguaje Java lo llevó a dejar de ser funcional.

Para estos apuntes hemos recuperado, y en ocasiones adaptado o modificado, algunos de los contenidos de Statmedia I, que habían estado escritos con gran pulcritud. Esto se ha hecho siguiendo las indicaciones de la licencia (CC-Share-alike) que permite adaptar contenidos atribuyéndolo a sus autores y citando la fuente.

Los gadgets originales ya no son funcionales pero muchos de ellos han sido re-escritos en R como aplicaciones Shiny (disponibles en: https://grbio.upc.edu/en/software/teaching_apps) y se enlazaran desde los puntos necesarios del texto.

Dejando aparte (además) de la licencia, vaya nuestro agradecimiento explícito al equipo de profesores del Departamento de Estadística de la Universidad de Barcelona, redactor de la versión inicial del proyecto, que es la que hemos utilizado: Antonio Arcas Pons, Miquel calvo Llorca, Antonio Miñarro Alonso, Sergi Civit Vives y Angel Vilarroya del Campo.

Antoni Arcas, Antonio Miñarro and Miguel Calvo (2008) Statmedia projects in Statistical Education

Otros materiales utilizados

- Alex Sanchez y Francesc Carmona (2002). Apunts d'Estadística Matemàtica Licència CC0 1.0 Universal
- Molina Peralta, I. and García-Portugués, E. (2024). *A First Course on Statistical Inference*. Version 2.4.1. ISBN 978-84-09-29680-4. Licencia CC BY-NC-ND 4.0
- Peter K. Dunn (2024) *The theory of distributions*. Licencia CC BY-NC-ND 4.0

1 Probabilidad y Experimentos aleatorios

1.1 Introducción

1.1.1 Fenómenos deterministas y fenómenos aleatorios

Supongamos que disponemos de un dado regular con todas las caras pintadas de blanco y con un número, que irá de 1 a 6 sin repetir ninguno, en cada una de las seis caras.

Definamos los dos experimentos siguientes: Experimento 1: Tirar el dado y anotar el color de la cara resultante. Experimento 2: Tirar el dado y anotar el número de la cara resultante. ¿Qué diferencia fundamental observamos entre ambos experimentos? Muy simple! En el experimento 1, el resultado es obvio: saldrá una cara de color blanco. Es decir, es posible predecir el resultado. Se trata de un experimento o fenómeno determinista.

En cambio, en el experimento 2 no podemos predecir cuál será el valor resultante. El resultado puede ser : 1, 2, 3, 4, 5 o 6 . Se trata de un experimento o fenómeno aleatorio.

El conjunto de resultados se anotará con el símbolo: Ω . En este caso, $\Omega = \{1, 2, 3, 4, 5, 6\}$. En los fenómenos aleatorios, al hacer muchas veces la experiencia, la frecuencia relativa de cualquier elemento del conjunto de resultados debe aproximarse siempre hacia un mismo valor.

1.1.2 Sucesos

Supongamos que se ejecuta un experimento aleatorio. Se nos puede ocurrir emitir un enunciado que, una vez realizada la experiencia, pueda decirse si se ha verificado o no se ha verificado. A dichos enunciados los denominamos sucesos.

Por otro lado, los sucesos van asociados a subconjuntos del conjunto de resultados. Cada suceso se corresponderá exactamente con uno, y sólo con un, subconjunto del conjunto de resultados.

Veamos un ejemplo: Experimento: Tirar un dado regular. Conjunto de resultados : $\Omega = \{1, 2, 3, 4, 5, 6\}$
 Enunciado: Obtener múltiplo de 3. Subconjunto al que se asocia el enunciado: $A = \{3, 6\}$ Nos referiremos habitualmente al suceso A.

1.1.2.1 Sucesos y conjuntos Al conjunto de resultados Ω , se le denomina suceso seguro. Al conjunto \emptyset (conjunto sin elementos), se le denomina suceso imposible. Al complementario del conjunto A (A^c), se le denomina suceso contrario o complementario de A. A partir de dos sucesos A y B, podemos formar los sucesos siguientes:

- A intersección B, que anotaremos como:

$$A \cap B$$

- A unión B, que anotaremos como:

$$A \cup B$$

A intersección B, significa que se verifican a la vez A y B. A unión B, significa que se verifica A o B (se pueden verificar a la vez).

1.2 Función de probabilidad

Lógicamente, una vez tenemos un suceso, nos preocupa saber si hay muchas o pocas posibilidades de que al realizar la experiencia se haya verificado.

Por lo tanto, sería interesante el tener alguna función que midiera el grado de confianza a depositar en que se verifique el suceso.

A esta función la denominaremos función de probabilidad. La función de probabilidad será, pues, una aplicación entre el conjunto de resultados y el conjunto de números reales, que asignará a cada suceso la probabilidad de que se verifique.

La notación: $P(A)$ significará: probabilidad de que se verifique el suceso A. Pero claro, de funciones de probabilidad asociadas a priori a una experiencia aleatoria podrían haber muchas.

Lo que se hace para decir qué es y qué no es una función de probabilidad es construir una serie de propiedades (denominadas axiomas) que se exigirán a una función para poder ser catalogada como función de probabilidad.

Y, ¿cuáles son estos axiomas? Pues los siguientes: Sea S el conjunto de sucesos.

- Axioma 1: Para cualquier suceso A, la probabilidad debe ser mayor o igual que 0.
- Axioma 2: La probabilidad del *suceso seguro* debe ser 1: $P(\Omega) = 1$
- Axioma 3: Para sucesos A_i , de modo que cada par de sucesos no tengan ningún resultado común, se verifica que:

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

De este modo, pueden haber muchas funciones de probabilidad que se podrían asociar con la experiencia.

El problema pasa entonces al investigador para decidir cual o cuales son las funciones de probabilidad más razonables asociadas con la experiencia que está manejando.

1.2.1 ¿Diferentes funciones de probabilidad para una misma experiencia aleatoria?

Supongamos la experiencia de tirar un dado regular. A todo el mundo se le ocurriría pensar que la función de probabilidad se obtiene de contar el número de resultados que contiene el suceso dividido por 6, que es el número total de resultados posibles. Así pues, la probabilidad de obtener un múltiplo de 3 sería igual a $2/6$, la probabilidad de obtener el número 2 sería $1/6$ y la probabilidad de obtener un número par sería $3/6$. Es decir, parece inmediato construir la función de probabilidad que, además, parece única. A nadie se le ocurre decir, por ejemplo, que la probabilidad de obtener un número par es $5/6$!

En este caso, todo ha sido muy fácil. Hemos visto que existe una única función de probabilidad que encaje de forma lógica con la experiencia y, además, ha sido muy sencillo encontrarla.

Pero esto, por desgracia, no siempre es así. En muchísimas ocasiones resulta muy complejo el decidir cuál es la función de probabilidad.

En el tema de variables aleatorias y de función de distribución se explica el problema de la modelización de muchas situaciones reales.

1.3 ¿Cómo se calculan las probabilidades?

No siempre es fácil conocer los valores de la función de probabilidad de todos los sucesos. Sin embargo, muchas veces se pueden conocer las probabilidades de algunos de estos sucesos. Con la ayuda de ciertas propiedades que se deducen de manera inmediata a partir de la axiomática es posible calcular las probabilidades de más sucesos.

Por otro lado, en caso de que el número de resultados sea finito y de que todos los resultados tengan las mismas posibilidades de verificarse, la probabilidad de un suceso cualquiera se puede calcular a partir de la regla de Laplace:

Si A es un suceso :

$$\text{Probabilidad } (A) = \frac{\text{Número de casos favorables}}{\text{Número de casos posibles}}$$

donde: Número de casos favorables = Número de resultados contenidos en A (cardinal de A)
Número de casos posibles = Número total de resultados posibles (cardinal del conjunto total de resultados)

En este caso, el contar número de resultados, ya sean favorables o posibles, debe hacerse por medio de la combinatoria.

Veamos con unos ejemplos muy sencillos y visuales cómo se obtienen y qué representan los casos posibles y los casos favorables.

También es posible obtener de manera aproximada la probabilidad de un suceso si se puede repetir muchas veces la experiencia: la probabilidad del suceso sería el valor al que tendería la frecuencia relativa del suceso. Podéis consultar más detalles acerca de esta aproximación.

En este caso, la cuestión estriba en poder hacer muchas veces la experiencia en condiciones independientes.

1.4 Sucesos elementales y sucesos observables

En el contexto de la probabilidad, es fundamental diferenciar entre los **sucesos elementales** y los **sucesos observables**.

Los sucesos elementales son los resultados individuales que pueden ocurrir al realizar un experimento aleatorio, es decir, cada uno de los elementos que conforman el conjunto de resultados Ω . En nuestro ejemplo del dado, los sucesos elementales son los números 1, 2, 3, 4, 5 y 6.

Sin embargo, no todos los sucesos elementales son necesariamente observables. Un suceso observable es un subconjunto de estos sucesos elementales que permite formular afirmaciones verificables sobre el resultado del experimento.

Ejemplo

1. Podemos imaginar un dado en el que pintamos de blanco las caras pares y de negro las impares. En este caso los sucesos elementales serían los habituales 1, 2, 3, ... 6. Sin embargo tan solo “Par” (“blanco”) o impar (“negro”) se pueden observar.
2. Si repintamos el dado de forma que las caras 1 y 2 esten blancas, las 3 y 4, azules y las 5 y 6 rojas podremos observar el suceso “Sale 1 o 2 (=Sale blanco)” o “sale blanco o azul”, pero no el suceso “sale par” dado que cada color contiene un número par y uno impar

Para formalizar estos conceptos, definimos el **espacio de probabilizable** como el par de conjuntos formados por: (Ω, \mathcal{A})

- Ω es el conjunto de todos los resultados posibles (el conjunto de resultados o sucesos elementales).
- \mathcal{A} es el conjunto de todos los sucesos observables, que vienen definidos por el *nivel de observación* del experimento.

1.5 Propiedades inmediatas de la probabilidad

Veremos a continuación una serie de propiedades que se deducen de manera inmediata de la axiomática de la probabilidad.

1.5.1 Suceso imposible

El suceso imposible se identifica con el conjunto vacío, puesto que no hay ningún resultado asociado a él. La probabilidad del suceso imposible es:

$$P(\emptyset) = 0$$

1.5.2 Suceso implicado

Decimos que un suceso, B, esta implicado por otro suceso A, si siempre que se presenta A, también lo hace B. Por ejemplo, si al tirar un dado se obtiene un dos (suceso A), ello implica que ha salido un número par (suceso B). En terminos de conjuntos, A es un suceso que está contenido en B (todos los resultados de A también pertenecen a B), por lo que:

$$P(A) \leq P(B)$$

1.5.3 Complementario de un suceso

Sea A^c el suceso formado por todos los elementos de Ω que no pertenecen a A (Suceso complementario de A). La probabilidad de dicho suceso es igual a:

$$P(A^c) = 1 - P(A)$$

1.5.4 Ocurrencia de algun suceso

La probabilidad de la unión de dos sucesos A y B es igual a:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

1.5.5 Probabilidad de que ocurra algun suceso

Si tenemos una colección de k sucesos, la probabilidad de la unión de dichos sucesos será:

$$P\left(\bigcup_{i=1}^k A_i\right) = \sum_{i=1}^k P(A_i) - \sum_{i < j} P(A_i \cap A_j) + \sum P(A_i \cap A_j \cap A_k) + \dots + (-1)^{k+1} \cdot P(A_1 \cap \dots \cap A_k)$$

1.5.6 Probabilidad de que ocurran dos (o más) sucesos a la vez

No existe una expresión cerrada única para la probabilidad de que ocurran dos o más sucesos a la vez, pues esto depende de si los sucesos que consideramos son dependientes o independientes, conceptos éstos, que introduciremos en la próxima sección.

Lo que si que existe es una cota para dicha probabilidad, es decir, podemos decir que valor alcanza dicha probabilidad, *como mínimo*.

$$P\left(\bigcap_{i=1}^n A_i\right) \geq 1 - \sum_{i=1}^n P(\bar{A}_i)$$

1.6 Espacios de probabilidad

Para concluir esta introducción introduciremos los **espacio de probabilidad** que, extienden los **espacios probabilizables** definidos en la sección anterior

La terna (Ω, \mathcal{A}, P) donde:

- Ω es el conjunto de todos los resultados posibles (el conjunto de resultados o sucesos elementales),
- \mathcal{A} es el conjunto de todos los sucesos observables, que vienen definidos por el *nivel de observación* del experimento y
- P es una función de probabilidad, que asigna a cada suceso observable $A \in \mathcal{A}$ un número real $P(A)$ que representa la probabilidad de que ocurra dicho suceso

se conoce como **espacio de probabilidad**.

Es importante destacar que **la probabilidad se calcula exclusivamente para los sucesos observables**, lo que garantiza que la medida sea coherente y verificada a través de experimentos.

Los espacios de probabilidad proporcionan una estructura fundamental para analizar y medir las incertidumbres asociadas a los fenómenos aleatorios, facilitando el estudio de sus propiedades, la construcción, sobre ellos de diversos conceptos fundamentales como el de variables aleatorias, y, en general, la aplicación de teorías de la probabilidad a diversas áreas de conocimiento.

1.7 Probabilidad condicionada

Imaginemos que en la experiencia de tirar un dado regular supiéramos de antemano que se ha obtenido un número par. Es decir, que se ha verificado el suceso: $\{B = \text{número par}\}$.

Pregunta: ¿Cuál es ahora la probabilidad de que se verifique el suceso mayor o igual a cuatro? Lógicamente, el resultado sería: $2/3$. Por lo tanto, la probabilidad del suceso $A = \text{mayor o igual a cuatro}$ se ha modificado. Evidentemente, ha pasado de ser $1/2$ (cuando no tenemos ninguna información previa) a ser $2/3$ (cuando sabemos que se ha verificado el suceso B). ¿Cómo podemos anotar esta última probabilidad ($2/3$)? Muy sencillo. Anotaremos $P(A/B)$, que se lee como probabilidad de A condicionada a B . Así, en este ejemplo,

$$\begin{aligned} P(A/B) &= 2/3 \\ P(A) &= 1/2 \end{aligned}$$

En términos generales, estamos en condiciones de poder definir la probabilidad condicionada, y lo hacemos como:

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

Podemos ahora visualizar de una manera práctica y divertida el ejemplo anterior. Siguiendo con la notación utilizada, el suceso A será lo que denominamos suceso de obtención, mientras que el suceso B será lo que denominamos suceso condicionado. La pantalla nos proporcionará los casos posibles para el condicionante elegido y los casos favorables, calculando mediante la regla de Laplace la probabilidad del suceso.

- 1) Elegid suceso a estudiar. Desplazad, si procede, las barras de puntos.
- 2) Elegir suceso condicionante. Desplazad, si procede, las barras de puntos.
- 3) Comprobad los sucesos posibles y los favorables.

La probabilidad condicionada se comporta, entonces, como una función de probabilidad. Es decir, verifica los tres axiomas siguientes:

Axioma 1:

$$P(A/B) \geq 0$$

Axioma 2:

$$P(\Omega/B) = 1$$

Axioma 3:

$$P\left(\bigcup_{i=1}^{\infty} A_i/B\right) = \sum_{i=1}^{\infty} P(A_i/B)$$

para sucesos A_i con intersección vacía dos a dos.

1.7.1 Sucesos dependientes y sucesos independientes

Sean A y B dos sucesos con probabilidad mayor que 0 . Evidentemente, si

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

B no ha modificado la probabilidad de que suceda A. En este caso diremos que son sucesos independientes.

En caso contrario diremos que son sucesos dependientes. En el ejemplo del apartado anterior, se observa que los sucesos son dependientes puesto que las probabilidades anteriores no coinciden.

Se verifica que independencia de los sucesos A y B es equivalente a decir que la probabilidad de la intersección es igual a producto de probabilidades de los dos sucesos.

Se verifica también que si A y B son independientes: a) El complementario del suceso A y el suceso B son independientes. b) El complementario del suceso A y el complementario del suceso B son independientes. c) El complementario del suceso B y el suceso A son independientes.

1.7.2 Incompatibilidad e independencia

Dos sucesos con intersección vacía se denominan sucesos incompatibles. Esto, ¿qué implica? Pues, que si se verifica uno seguro que no se verifica el otro, ya que no tienen resultados en común. Por lo tanto es el caso extremo de dependencia. Obtenemos en este caso que:

$$P(A/B) = 0$$

y, en consecuencia, si $P(A)$ y $P(B)$ son diferentes de cero, la probabilidad condicionada anterior es diferente de $P(A)$, y así se deduce la dependencia.

La única posibilidad de que se dé incompatibilidad e independencia a la vez, es que alguno de los dos sucesos tenga probabilidad igual a cero.

1.8 Dos Teoremas importantes

1.8.1 Teorema de las probabilidades totales

Sea Ω el conjunto total formado por una partición (colección de sucesos con intersección vacía dos a dos):

$$\Omega = H_1 \cup \dots \cup H_n$$

La probabilidad de cualquier otro suceso A , se puede obtener a partir de las probabilidades de los sucesos de la partición y de las probabilidades de A condicionado a los sucesos de la partición, de la manera siguiente:

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

Esto es lo que se conoce como teorema de las probabilidades totales.

1.8.2 Teorema de Bayes

Es una consecuencia del teorema de las probabilidades totales. Sea Ω el conjunto total formado por una partición (colección de sucesos con intersección vacía dos a dos).

$$\Omega = H_1 \cup \dots \cup H_n$$

Ahora el interés se centrará en la obtención de la probabilidad de cualquier suceso de la partición condicionada a un suceso A cualquiera.

El resultado será:

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

Esto es conocido como teorema o regla de Bayes.

1.9 Introducción a los experimentos múltiples

Supongamos que tiramos a la vez un dado y una moneda. Tenemos una experiencia múltiple, puesto que la experiencia que se realiza es la composición de dos experiencias (experiencia 1 = tirar un dado regular; experiencia 2 = tirar una moneda regular). ¿Cuál es en este caso el conjunto de resultados? Si Ω_1 es el conjunto de resultados asociado con la experiencia tirar un dado y Ω_2 es el conjunto de resultados asociado con la experiencia tirar una moneda, el conjunto de resultados asociado a la experiencia múltiple será $\Omega_1 \times \Omega_2$.

Es decir, $\Omega_1 = \{1, 2, 3, 4, 5, 6\}$ $\Omega_2 = \{ \text{cara}, \text{cruz} \}$ $\Omega_1 \times \Omega_2 = \{(1, \text{cara}), (2, \text{cara}), (3, \text{cara}), (4, \text{cara}), (5, \text{cara}), (6, \text{cara}), (1, \text{cruz}), (2, \text{cruz}), (3, \text{cruz}), (4, \text{cruz}), (5, \text{cruz}), (6, \text{cruz})\}$

Si P_1 y P_2 son, respectivamente, las funciones de probabilidad asociadas a las experiencias 1 y 2, ¿es posible calcular probabilidades de la experiencia múltiple a partir de P_1 y P_2 ?

Efectivamente! Pero hemos de distinguir dos situaciones:

- Experiencias independientes: cuando el resultado de una no influya en la otra.
- Experiencias dependientes: cuando el resultado de una influya en la otra.

En nuestro caso se trata de experiencias independientes, puesto que el resultado que se obtenga al tirar el dado no influye sobre el resultado que se obtenga al lanzar la moneda y al revés. ¿Como se calculan, pues, las probabilidades de la experiencia múltiple? Sea un suceso de la experiencia múltiple: $A \times B$.

- Caso de experiencias independientes:

$$P(A \times B) = P_1(A) \times P_2(B)$$

- Caso de experiencias dependientes:

$$P(A \times B) = P_1(A) \times P_2(B/A)$$

Entendemos que existe una P_2 para cada suceso A .

Esto que hemos explicado se puede, lógicamente, generalizar a una experiencia múltiple formada por n experiencias.

1.10 Combinatoria

Veamos algunas fórmulas simples que se utilizan en combinatoria y que nos pueden ayudar a calcular el número de casos posibles o el número de casos favorables.

1.10.1 Permutaciones

Sea un conjunto de n elementos. A las ordenaciones que se pueden hacer con estos n elementos sin repetir ningún elemento y utilizándolos todos se las denomina permutaciones. El número de permutaciones que se pueden realizar coincide con el factorial de n , y su cálculo es:

$$n! = n \cdot (n - 1) \cdot (n - 2) \cdot \dots \cdot 2 \cdot 1$$

Ejemplo:

¿De cuántas maneras distintas podemos alinear a seis personas en una fila?

Respuesta

De $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$ maneras (permutaciones de 6 elementos).

1.10.2 Variaciones

Sea un conjunto de n elementos. Supongamos que deseamos ordenar r elementos de entre los n . A cada una de estas ordenaciones la denominamos variación. El número de variaciones que se pueden hacer de los n elementos tomados de r en r es:

$$V_n^r = n \cdot (n - 1) \cdot \dots \cdot (n - r + 1)$$

Ejemplo

En una carrera de velocidad compiten diez atletas. ¿De cuántas maneras distintas podría estar formado el podio? (el podio lo forman el primer, el segundo y el tercer clasificado)

Respuesta

Cada podio posible es una variación de diez elementos tomado de tres en tres. Por tanto, el número posible de podios es:

$$V_{10}^3 = 10 \cdot 9 \cdot 8 = 720$$

1.10.3 Variaciones con repetición

Sea un conjunto de n elementos. Supongamos que se trata de ordenar r elementos que pueden estar repetidos. Cada ordenación es una variación con repetición. El número de variaciones con repetición para un conjunto de n tomados de r en r es :

$$RV_n^r = n^r$$

Ejemplo

En una urna tenemos cinco bolas numeradas del 1 al 5 . Se extraen tres bolas sucesivamente con reposición (devolviendo cada vez la bola a la urna). ¿Cuántos resultados distintos es posible obtener?

Respuesta: Se trata de variaciones con repetición de un conjunto de cinco bolas tomadas de tres en tres. En total tendremos:

$$RV_5^3 = 5^3 = 125$$

1.10.4 Combinaciones

Quando se trata de contar el número de subconjuntos de x elementos en un conjunto de n elementos tenemos lo que se denomina combinaciones de x elementos en un conjunto de n . El cálculo del contaje se hace mediante el número combinatorio, de la manera siguiente:

$$C_n^x = \binom{n}{x} = \frac{n!}{x! \cdot (n-x)!}$$

Ejemplo

¿De cuántas maneras podemos elegir, en la urna anterior (recordemos que había cinco bolas), tres bolas en una única extracción?

Respuesta

Serán combinaciones de cinco elementos tomados de tres en tres, por tanto, tendremos:

$$C_5^3 = \binom{5}{3} = \frac{5!}{3!(5-3)!} = 10$$

1.10.5 Permutaciones con repetición

Sea un conjunto de n elementos, de entre los cuales tenemos a elementos indistinguibles entre sí, b elementos indistinguibles entre sí, c elementos indistinguibles entre sí, etc. Cada ordenación de estos elementos se denominará permutación con repetición. El número de permutaciones con repetición es:

$$RP_n^{a,b,c,\dots} = \frac{n!}{a!b!c!\dots}$$

Ejemplo

¿Cuántas palabras con sentido o sin él pueden formarse con las letras PATATA?

Respuesta: Tenemos tres veces la letra A, dos veces la T y una vez la P. Por tanto, serán:

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

1.11 Frecuencia relativa y probabilidad

La definición moderna de probabilidad basada en la axiomática de Kolmogorov (presentada anteriormente) es relativamente reciente. Históricamente hubo otros intentos previos de definir el escurridizo concepto de probabilidad, descartados por diferentes razones. Sin embargo conviene destacar aquí algunas ideas que aparecen en la antigua definición basada en la frecuencia relativa, ya que permiten intuir algunas profundas propiedades de la probabilidad.

Recordemos antes que si en un experimento que se ha repetido n veces un determinado suceso A se ha observado en k de estas repeticiones, la frecuencia relativa f_r del suceso A es:

$$f_r = k/n$$

El interés por la frecuencia relativa y su relación con el concepto de probabilidad aparece a lo largo de los siglos XVIII a XX al observar el comportamiento de numerosas repeticiones de experimentos reales.

A título de ejemplo de un experimento de este tipo, supongamos que se dispone de una moneda ideal perfectamente equilibrada. Aplicando directamente la regla de Laplace resulta claro que el suceso A = obtener cara tiene probabilidad:

$$p(A) = 1/2 = 0,5$$

Ilustración por simulación

En el enlace siguiente se accede a una simulación por ordenador de la *ley de los grandes números* en la que se basa precisamente la idea de asimilar “a la larga” (es decir a medida que crece el número de repeticiones) frecuencia relativa y probabilidad.

Enlace a la simulación

En la simulación podéis definir:

- La verdadera probabilidad” de que al tirar la moneda salga cara,
- EL número de tiradas.

Como podréis comprobar, sea cual sea la probabilidad (una moneda justa es un 0.5) a la larga la frecuencia relativa converge hacia el valor que habéis fijado.

Eso sí, observad lo que sucede si fijais probabilidades cercanas a 0.5 o muy alejadas de ell.

¿La idea de lo que sucede a la larga es la misma? ¿En que encontráis diferencias? Aunque no deje de llamar la atención el carácter errático del comportamiento de f_r entre los valores 0 y 1, estaréis seguramente de acuerdo que a mayor número de lanzamientos n , más improbable es que f_r se aleje mucho de $p(A)$.

La teoría moderna de la probabilidad enlaza formalmente estas ideas con el estudio de las leyes de los grandes números, que se discutan con más detalle en el capítulo dedicado a las “Grandes muestras”.

1.12 Caso de Estudio: Eficacia de una prueba diagnóstica

Para decidir la presencia (E) o ausencia (A) de sordera profunda a la edad de seis meses, se está ensayando una batería de tests.

Considerando el caso en que la prueba pueda dar positivo (+) o negativo (−), hay que tener en cuenta que en individuos con dicha sordera la prueba dará a veces positivo y a veces negativo, e igual ocurrirá con individuos que no presentan la sordera.

En este contexto todas las probabilidades pueden ser interpretadas en terminos de resultados positivos o negativos, correctamente o no y cada una ha recibido un nombre que la ha popularizado dentro de la literatura médica:

Así tenemos:

- $P(+/E)$
 - Probabilidad de test positivo en individuos que padecen la sordera.
 - Este valor se conoce como *sensibilidad del test*.
- $P(+/A) =$
 - Probabilidad de test positivo en individuos que no padecen la sordera.
 - Este valor se conoce como *probabilidad de falso-positivo*.
- $P(-/E) =$
 - Probabilidad de test negativo en individuos que padecen la sordera
 - Este valor se conoce como *probabilidad de falso-negativo*.
- $P(-/A) =$
 - Probabilidad de test negativo en individuos que no padecen sordera.
 - Este valor se conoce como *especificidad del test*.
- Finalmente a la probabilidad, $P(E)$, de presentar la enfermedad se le conoce como *prevalencia* de la enfermedad.

Lógicamente, en un “buen test” nos interesa que la sensibilidad y la especificidad sean elevadas, mientras que los falsos-positivos y falsos-negativos sean valores bajos.

Además no debemos olvidar que, el interés de aplicar el test, consiste en que sirva de elemento predictivo para diagnosticar la sordera.

Por lo tanto, interesa que las probabilidades:

- $P(E/+)$ = Probabilidad de padecer sordera si el test da positivo
- $P(A/-)$ = Probabilidad de no padecer sordera si el test da negativo

sean realmente altas.

A las probabilidades anteriores se las conoce como: *valores predictivos* del test, en concreto:

- $P(E/+)$ = es el *valor predictivo positivo* y
- $P(A/-)$ = es el *valor predictivo negativo*

1.12.1 Aplicación del Teorema de Bayes

Estamos en una situación en que, a partir de conocimiento de unas probabilidades, nos interesa calcular otras, para lo que utilizaremos el teorema de Bayes.

Habitualmente, a partir de estudios epidemiológicos y muestras experimentales, se estiman:

- La prevalencia

- La sensibilidad del test
- La especificidad del test
- La probabilidad de falso positivo
- La probabilidad de falso negativo

¿Cómo se obtiene entonces el valor predictivo del test?

Veamos como aplicar el teorema de Bayes a este problema:

Si dividimos a la población global (en este caso, el conjunto de todos los bebés de seis meses) entre los que padecen sordera y los que no la padecen, aplicando el teorema de Bayes resulta que:

$$P(E/+) = (P(+/E) \times P(E)) / (P(+/E) \times P(E) + P(+/A) \times P(A))$$

y

$$P(A/-) = (P(-/A) \times P(A)) / (P(-/A) \times P(A) + P(-/E) \times P(E))$$

1.12.2 Ejemplo numérico

Supongamos que en el ejemplo de la sordera, se sabe que:

- Prevalencia = 0,003, Es decir, que un tres por mil padece sordera profunda a esta edad.
- Sensibilidad = 0,98
- Especificidad = 0,95
- Probabilidad de falso positivo = 0,05
- Probabilidad de falso negativo = 0,02

¿Cuál es el valor predictivo del test?

$$P(E/+) = (0,98 \times 0,003) / (0,98 \times 0,003 + 0,05 \times 0,997) = 0,00294 / 0,05279 = 0,055692$$

$$P(A/-) = (0,95 \times 0,997) / (0,95 \times 0,997 + 0,02 \times 0,003) = 0,94715 / 0,94721 = 0,999936$$

En conclusión, Podemos afirmar que se trata de un test muy válido para decidir que no hay sordera en caso de que el resultado del test sea negativo.

Sin embargo, el valor tan bajo de $P(E/+)$ no permite poder considerar al test como un predictor válido para diagnosticar la sordera.

Obsérvese que:

- Probabilidad de falso positivo = $1 - \text{especificidad}$
- Probabilidad de falso negativo = $1 - \text{sensibilidad}$

2 Variables aleatorias y Distribuciones de probabilidad

En el capítulo anterior hemos introducido el concepto de probabilidad y como calcular probabilidades asociadas a sucesos observables, formados por uno o mas sucesos elementales, resultado de un experimento aleatorio.

En muchas ocasiones nos interesa representar los resultados de un experimento aleatorio mediante un valor numérico que lo caracterice. Por ejemplo si tiramos tres monedas y contamos el número de caras, nos será indiferente cuando salgan dos caras, en que monedas ha salido una cara y en cual ha salido una cruz.

En la práctica, esto significa que en dichas ocasiones, aunque haya un experimento aleatorio detras de los valores que observamos, tan sólo nos interesan los resultados que expresamos a traves de valores numéricos.

Las variables aleatorias son la forma que hemos desarrollado para *trasladar la estructura proporcionada por los espacios de probabilidad el espacio muestral, el conjunto de sucesos elementales, al conjunto de los números, en concreto a la recta real, haciéndolo de tal forma que podamos seguir calculando probabilidades de sucesos observables.*

En este capítulo veremos que las variables aleatorias permiten pues *transportar* la probabilidad del espacio de probabilidad original a la recta real. Para ello, introduciremos una función que es la que se ocupa de ello, la *función de distribución de probabilidad*.

2.1 El espacio muestral y sus elementos

Cuando llevamos a cabo un experimento aleatorio, el conjunto Ω de resultados posibles forman el denominado espacio muestral. Sus elementos ω (resultados o sucesos elementales) deben ser conocidos por el investigador que realiza la experiencia, aun cuando no podamos determinar a priori el resultado particular de una realización concreta.

Supondremos que también conocemos la manera de asignar una probabilidad sobre el conjunto de enunciados o *sucesos observables* que se pueden construir a partir de Ω . Es decir, supondremos la existencia de un espacio de probabilidad construido a partir de los resultados de Ω .

Generalmente, la estructura del espacio muestral no permite, o por lo menos no facilita, su tratamiento matemático. Pensemos en la inmensa variedad en la naturaleza de resultados posibles de diferentes experimentos. Además es bastante frecuente que no nos interesen los resultados en sí, sino una característica que, de alguna manera, resuma el resultado del experimento.

2.2 Representación numérica de los sucesos elementales. Variables aleatorias

La forma de resumen que adoptaremos es la asignación a cada suceso elemental de un valor numérico, en particular, de un número real.

En la práctica la asignación de un valor numérico a cada elemento del espacio muestral se hace siguiendo una regla o enunciado, según el interés concreto del experimentador. Evidentemente, podemos construir diversas maneras de asignar valores numéricos a los mismos resultados de un experimento.

Hablando en términos coloquiales, podemos decir que cada regla de asignación corresponde a una determinada variable que se puede medir sobre los sucesos elementales.

Nótese que es posible construir múltiples variables sobre un mismo espacio de probabilidad. En términos algo más formales, las reglas de asignación se pueden interpretar como una aplicación de Ω en el conjunto de números reales.

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

X representa la variable o regla de asignación concreta. El conjunto de valores numéricos que puede tomar una variable, y que depende de la naturaleza de la misma variable, recibe el nombre de recorrido de la variable.

A partir de este momento, los sucesos elementales quedan substituidos por sus valores numéricos de acuerdo a una determinada variable y permiten un mayor tratamiento matemático en el marco de la teoría de la probabilidad.

El apelativo aleatoria que reciben las variables hace referencia al hecho de que los posibles valores que toman dependen de los resultados de un fenómeno aleatorio que se presentan con una determinada probabilidad.

Como un complemento al tema, al final del capítulo, presentamos la definición formal de variable aleatoria, donde se introducen las restricciones a las reglas de asignación numérica que posibilitan el tratamiento matemático de las variables.

2.3 Caracterización de una variable aleatoria a través de la probabilidad. Función de distribución

Una vez que tenemos definida una variable aleatoria, ésta queda totalmente caracterizada en el momento en que somos capaces de determinar la probabilidad de que la variable tome valores en cualquier intervalo de la recta real. Dado que los posibles valores que puede tomar la variable, es decir, su recorrido, pueden ser muy grandes (infinitos de hecho), el problema de caracterizar una variable aleatoria se resuelve introduciendo una función especial, *la función de distribución*.

Definición

La función de distribución de una variable aleatoria X es la aplicación que, a cada punto de la recta real, le asigna la probabilidad del suceso formado por los resultados del experimento que tienen asignado un valor de la variable aleatoria menor o igual a dicho punto.

$$\begin{aligned} F : \mathbb{R} &\rightarrow [0, 1] \\ x &\rightarrow F(x) = P(X \leq x) = P\{\omega \in \Omega \mid X(\omega) \leq x\} \end{aligned}$$

También podemos decir que es la probabilidad inducida en el intervalo de la recta $(-\infty, x]$

Hay que hacer notar que siempre será posible determinar dicha probabilidad gracias a los requerimientos exigidos en la definición formal de variable aleatoria. Por tanto, toda variable aleatoria tiene asociada una función de distribución. Nos referimos a esta función cuando decimos que conocemos la distribución de la variable aleatoria.

2.4 Propiedades de la función de distribución

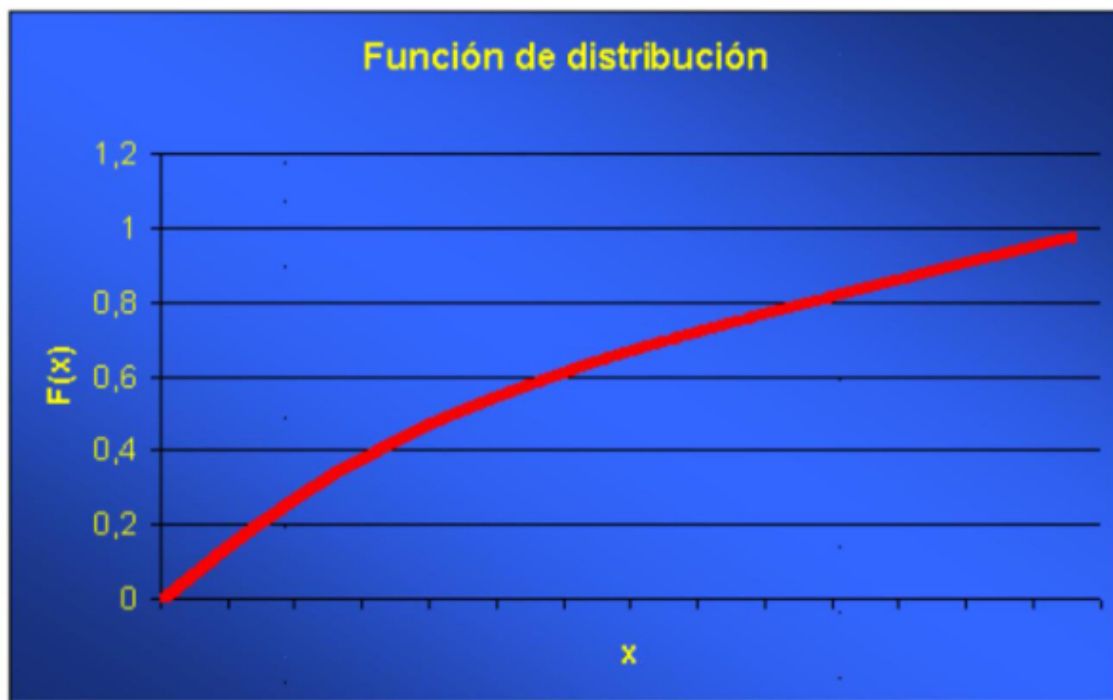
La forma en que hemos definido las funciones de distribución determina que dichas funciones deban de tener las siguientes propiedades:

1. $0 \leq F(x) \leq 1$. Efectivamente, se trata de una probabilidad, por lo que toma valores entre 0 y 1
2. $\lim_{x \rightarrow +\infty} F(x) = 1$. A medida que un valor se hace más y más grande, la probabilidad de encontrar valores anteriores a él crece y, en el límite, valdrá uno (el valor máximo para una probabilidad).
3. $\lim_{x \rightarrow -\infty} F(x) = 0$. A medida que un valor se hace más y más negativo, la probabilidad de encontrar valores anteriores a él disminuye, y en el límite es cero (el valor mínimo para una probabilidad).
4. $x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$. Por construcción, es una función monótona, es decir, si un valor es inferior a otro, la probabilidad de encontrar valores inferiores al menor de los dos será menor o igual que la de encontrarlos inferiores al mayor de los dos.
5. $\lim_{x \rightarrow a+} F(x) = F(a) \quad \forall a \in \mathbb{R}$. Por la forma en que se ha definido, la función de distribución es *continua por la derecha*.

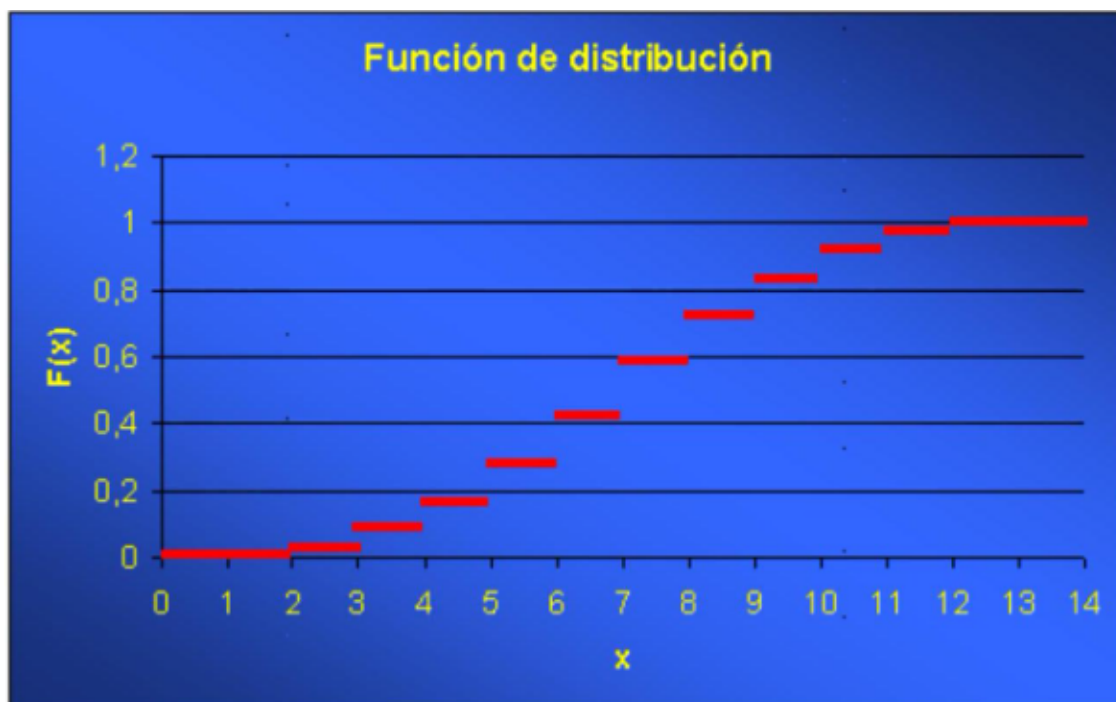
Toda función que verifique las propiedades anteriores es una función de distribución y toda función de distribución caracteriza una determinada variable aleatoria sobre algún espacio de probabilidad.

Las propiedades anteriores determinan la forma de la función de distribución. En concreto, según la variable sea continua o discreta, conceptos definidos a continuación en el capítulo, la forma de la función será :

Primer tipo (Variables continuas)



Segundo tipo (variables discretas)



2.5 Clasificación de las variables aleatorias

Para su estudio, las variables aleatorias se clasifican en variables discretas o variables continuas.

2.5.1 Variables aleatorias discretas

Definición: Variable aleatoria discreta

Diremos que una variable aleatoria es discreta si su recorrido, es decir, el conjunto de valores que puede tomar, es finito o infinito numerable.

Generalmente, este tipo de variables van asociadas a experimentos en los cuales se cuenta el número de veces que se ha presentado un suceso o donde el resultado es una puntuación concreta.

Los puntos del recorrido se corresponden con saltos en la gráfica de la función de distribución, que correspondería al segundo tipo de gráfica visto anteriormente.

2.5.2 Variables aleatorias continuas

Definición: Variable aleatoria continua

Diremos que una variable aleatoria es continua si su función de distribución es una función continua.

También puede definirse, de forma análoga a las variables discretas como aquellas cuyo recorrido, es decir, el conjunto de valores que puede tomar, es un intervalo o subconjunto no numerable de los números reales. En otras palabras, aquellas que pueden tomar cualquier valor dentro de un rango continuo, sin saltos entre los valores posibles.

Se corresponde con el primer tipo de gráfica visto.

Generalmente, se corresponden con variables asociadas a experimentos en los cuales la variable medida puede tomar cualquier valor en un intervalo; mediciones biométricas, por ejemplo.

Un caso particular dentro de las variables aleatorias continuas y al cual pertenecen todos los ejemplos usualmente utilizados, son las denominadas variables aleatorias absolutamente continuas.

Definición: Distribución absolutamente continua

Diremos que una variable aleatoria X continua tiene una distribución absolutamente continua si existe una función real f , positiva e integrable en el conjunto de números reales, tal que la función de distribución F de X se puede expresar como

$$F(x) = \int_{-\infty}^x f(t)dt$$

Una variable aleatoria con distribución absolutamente continua, por extensión, se la clasifica como variable aleatoria absolutamente continua.

Definición: función de densidad de probabilidad

A la función f se la denomina función de densidad de probabilidad de la variable X .

Hay que hacer notar que no toda variable continua es absolutamente continua, pero los ejemplos son complicados, algunos utilizan para su construcción el conjunto de Cantor, y quedan fuera del nivel y del objetivo de este curso.

Igualmente indicaremos que los tipos de variables comentados anteriormente forman únicamente una parte de todos los posibles tipos de variables, sin embargo contienen prácticamente todas las variables aleatorias que encontramos usualmente.

Tal como se estudiará más adelante, existen algunas familias de funciones de distribución, tanto dentro del grupo de las discretas como de las continuas, que por su importancia reciben un nombre propio y se estudiarán en los capítulos siguientes.

En ocasiones encontramos variables de tipo mixto, es decir que se comportan como discretas o continuas para distintos grupos de valores.

2.6 Variable aleatoria discretas

Tal como se ha definido, una variable aleatoria X discreta toma valores en un conjunto finito o numerables.

Indicaremos el recorrido de la variable X como: $\{x_1, x_2, \dots, x_k, \dots\}$.

El ejemplo más sencillo de variable aleatoria discreta lo constituyen las variables indicadoras. Sea A un suceso observable, se llama indicador de A a la variable aleatoria definida por

$$I_A : \Omega \rightarrow \mathbb{R}$$
$$\omega \rightarrow I_A(\omega) = \begin{cases} 1 & \text{si } \omega \in A \\ 0 & \text{si } \omega \notin A \end{cases}$$

2.6.0.1 Ejercicio propuesto Construir, a partir de las variables indicadoras de A y B , las siguientes variables indicadoras

$$I_{A \cap B}; I_{A \cup B}; I_{A^c}; I_{\Omega}$$

2.6.0.1.1 Solución

$$I_{A \cap B} = I_A \cdot I_B$$
$$I_{A \cup B} = I_A + I_B - I_{A \cap B}$$
$$I_{A^c} = 1 - I_A$$
$$I_{\Omega} = 1$$

2.6.1 Caracterización de las v.a. discretas

Una variable aleatoria discreta puede caracterizarse a través de la función que asocia cada elemento del recorrido su probabilidad. Dicha función recibe varios nombres según los autores: - función de probabilidad - ley de probabilidad, - función de densidad de la variable aleatoria discreta. - función de masa de probabilidad.

Aunque es habitual encontrar, en muchos libros el término *función de densidad* para variables (absolutamente) continuas y el término *función de masa de probabilidad* para variables discretas, también lo es referirse a ambas como “función de densidad”.

La función de probabilidad de una variable discreta se puede representar de la manera siguiente:

$$f : \mathbb{R} \rightarrow [0, 1]$$
$$x \rightarrow f(x) = P(X = x) = P\{\omega \in \Omega \mid X(\omega) = x\}$$

Obsérvese que, a diferencia de la función de distribución que toma valores para cualquier valor real, la función definida anteriormente es nula en todo punto que no pertenezca al recorrido.

En cambio, siguiendo con la analogía, y dado que se trata de una probabilidad, la función de densidad discreta está acotada $0 \leq f(x) \leq 1$.

Toda función de densidad discreta puede expresarse de manera explícita a través de una tabla que asocie directamente puntos del recorrido con sus probabilidades.

Ejemplo: Función de densidad de una variable indicadora

Consideremos la variable indicadora del suceso A :

$$I_A : \Omega \rightarrow \mathbb{R}$$
$$\omega \rightarrow I_A(\omega) = \begin{cases} 1 & \text{si } \omega \in A \\ 0 & \text{si } \omega \notin A \end{cases}$$

La función de densidad de esta variable sería la siguiente:

x	0	1
$f(x) = P(X = x)$	$1 - P(A) = P(A^c)$	$P(A)$

El recorrido está formado por dos valores: 1 y 0, con las mismas probabilidades que las del suceso A y su complementario, respectivamente.

En muchos casos será posible expresar la función de probabilidad mediante una fórmula matemática que define una regla de asignación de probabilidades para los valores del recorrido.

Ejemplo: Un modelo matemático para la función de probabilidad

$$P(X = x) = 0,2 \cdot 0,8^{x-1}, \quad x = 1, 2, \dots$$

es la función de densidad de una variable aleatoria discreta con recorrido numerable.

2.6.2 Propiedades de la función de densidad discreta

1.

$$0 \leq f(x) \leq 1$$

2. $\sum_{i=1}^n f(x_i) = 1$, si el recorrido es finito.

3. $\sum_{i=1}^{\infty} f(x_i) = 1$, si el recorrido es numerable.

2.6.3 Relaciones entre la función de distribución y la función de densidad discreta. Probabilidad de intervalos.

Existe una relación muy importante entre las funciones de distribución $F(x)$ y de densidad $f(x)$ de una variable aleatoria discreta. La función de distribución en un punto se obtiene acumulando el valor de la función de densidad para todos los valores del recorrido menores o iguales al punto en cuestión.

$$F(x) = \sum_{x_i \leq x} f(x_i) \quad \text{para todo } x_i \text{ perteneciente al recorrido de la variable.}$$

En efecto, supongamos que el recorrido de una variable discreta X es $\{x_1, x_2, \dots, x_k, \dots\}$ y que deseamos conocer el valor de la función de distribución en un punto x tal que $x_i \leq x < x_{i+1}$, entonces es inmediato que

$$F(x) = P(X \leq x) = P(X = x_1) + P(X = x_2) + \dots + P(X = x_i) = f(x_1) + f(x_2) + f(x_3) + \dots + f(x_i)$$

Por ejemplo, para una variable indicadora de un suceso A , tenemos la relación siguiente:

Valor de x	$f(x)$	$F(x)$
$(-\infty, 0)$		0
0	$P(A^c)$	$P(A^c)$
$(0, 1)$		$P(A^c)$
1	$P(A)$	$P(A^c) + P(A) = 1$
$(1, +\infty)$		1

A partir de las funciones de densidad y de distribución es posible expresar las probabilidades para cualquier posible intervalo de valores de la variable. Por ejemplo:

Intervalo
$P(X \leq a) = F(a)$
$P(X < a) = F(a) - f(a)$
$P(X > a) = 1 - F(a) = 1 - P(X \leq a)$
$P(X \geq a) = 1 - F(a) + f(a) = 1 - P(X > a)$
$P(a < X \leq b) = F(b) - F(a)$
$P(a < X < b) = F(b) - f(b) - F(a)$
$P(a \leq X \leq b) = F(b) - F(a) + f(a)$
$P(a \leq X < b) = F(b) - f(b) - F(a) + f(a)$

2.7 Variables aleatorias continuas

Una variable aleatoria X diremos que es continua si su función de distribución es una función continua. En la práctica, se corresponden con variables asociadas con experimentos en los cuales la variable medida puede tomar cualquier valor en un intervalo: mediciones biométricas, intervalos de tiempo, áreas, etc.

Ejemplo: Variables aleatorias continuas

- Resultado de un generador de números aleatorios entre 0 y 1. Es el ejemplo más sencillo que podemos considerar, es un caso particular de una familia de variables aleatorias que tienen una distribución uniforme en un intervalo $[a, b]$. Se corresponde con la elección al azar de cualquier valor entre a y b .
- Estatura de una persona elegida al azar en una población. El valor que se obtenga será una medición en cualquier unidad de longitud (m , cm , etc.) dentro de unos límites condicionados por la naturaleza de la variable. El resultado es impredecible con antelación, pero existen intervalos de valores más probables que otros debido a la distribución de alturas en la población. Más adelante veremos que, generalmente, variables biométricas como la altura se adaptan un modelo de distribución denominado distribución Normal y representado por una campana de Gauss.

Dentro de las variables aleatorias continuas tenemos las variables aleatorias absolutamente continuas.

Diremos que una variable aleatoria X continua tiene una distribución absolutamente continua si existe una función real f , positiva e integrable en el conjunto de números reales, tal que la función de distribución F de X se puede expresar como

$$F(x) = \int_{-\infty}^x f(t)dt$$

Una variable aleatoria con distribución absolutamente continua, por extensión, se clasifica como variable aleatoria absolutamente continua.

En cuanto a nuestro manual, todas las variables aleatorias continuas con las que trabajemos pertenecen al grupo de las variables absolutamente continuas, en particular, los ejemplos y casos expuestos.

2.7.1 Función de densidad continua

La función que caracteriza las variables continuas es aquella función f positiva e integrable en los reales, tal que acumulada desde $-\infty$ hasta un punto x , nos proporciona el valor de la función de distribución en x , $F(x)$. Recibe el nombre de función de densidad de la variable aleatoria continua.

$$F(x) = \int_{-\infty}^x f(t)dt$$

Las funciones de densidad discreta y continua tienen, por tanto, un significado análogo, ambas son las funciones que acumuladas (en forma de sumatorio en el caso discreto o en forma de integral en el caso continuo) dan como resultado la función de distribución.

La diferencia entre ambas, sin embargo, es notable.

- La función de densidad discreta toma valores positivos únicamente en los puntos del recorrido y se interpreta como la probabilidad de la que la variable tome ese valor $f(x) = P(X = x)$.
- La función de densidad continua toma valores en el conjunto de números reales y no se interpreta como una probabilidad. No está acotada por 1, puede tomar cualquier valor positivo. Es más, en una variable continua se cumple que probabilidades definidas sobre puntos concretos siempre son nulas.

$$P(X = x) = 0 \text{ para todo } x \text{ real.}$$

¿Cómo se interpreta, entonces, la función de densidad continua? Las probabilidades son las áreas bajo la función de densidad. El área bajo la función de densidad entre dos puntos a y b se interpreta como la probabilidad de que la variable aleatoria tome valores comprendidos entre a y b .

Por tanto, siempre se cumple lo siguiente:

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

La función de densidad se expresa a través de una función matemática. La forma específica de la función matemática generalmente pasa por considerar a la variable aleatoria como miembro de una determinada familia de distribuciones, un determinado modelo de probabilidad. Estas familias generalmente dependen de uno o más parámetros y serán objeto de un estudio específico en un capítulo posterior. La atribución a una determinada familia depende de la naturaleza de la variable en cuestión.

Podemos ver, únicamente con ánimo ilustrativo, la expresión analítica y la gráfica para los ejemplos comentados con anterioridad:

- Resultado de un generador de números aleatorios entre a y b . Modelo Uniforme. $f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & x \notin [a, b] \end{cases}$
- Estatura de una persona elegida al azar en una población. Modelo Normal.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-170)^2}{2}} \quad -\infty < x < \infty$$

2.7.2 Relaciones entre la función de distribución y la función de densidad.

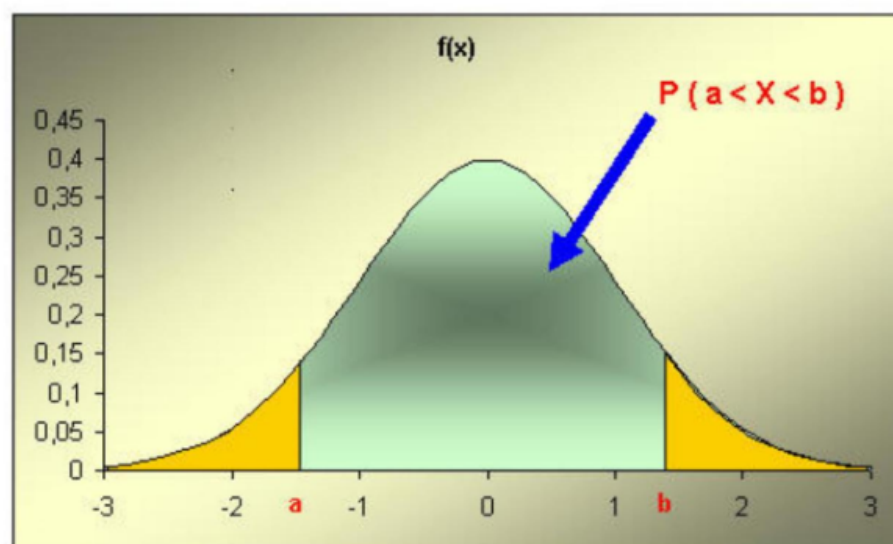
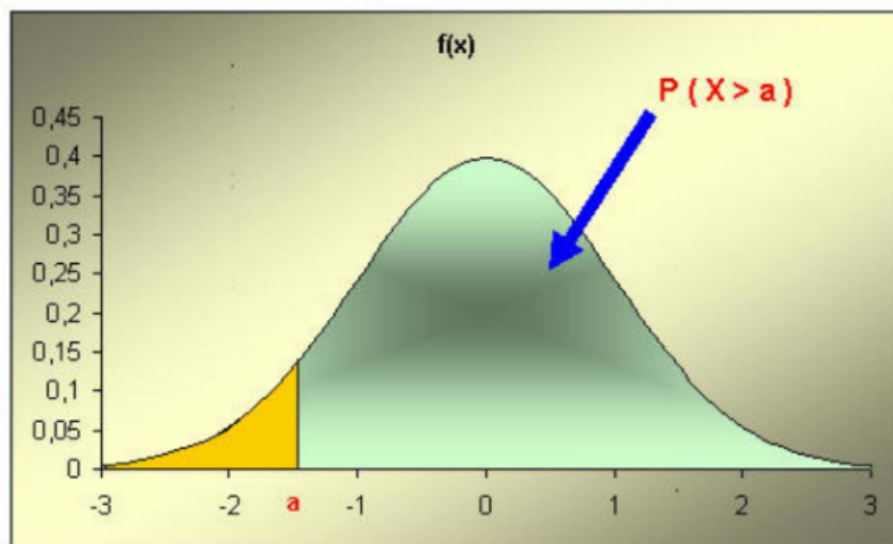
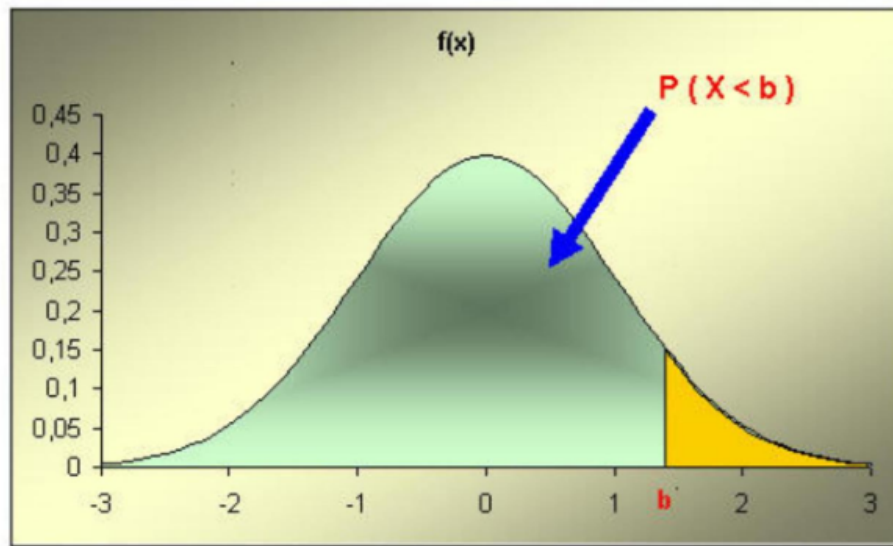
Para una variable continua, la relación entre las funciones de distribución y de densidad viene dada directamente a través de la definición. La función de distribución en un punto se obtiene integrando el valor de la función de densidad desde menos infinito hasta el punto en cuestión. Por ejemplo:

$$F(x) = \int_{-\infty}^x f(t)dt$$

2.7.2.1 Probabilidad de intervalos A partir de las funciones de densidad y de distribución, y teniendo en cuenta que $P(X = x) = 0$ para todo x real, es posible expresar las probabilidades para cualquier posible intervalo de valores de la variable. Por ejemplo:

Intervalo
$P(X \leq a) = P(X < a) = F(a) = \int_{-\infty}^a f(x)dx$ $P(X \geq a) = P(X > a) = 1 - F(a) = \int_a^{+\infty} f(x)dx$ $P(a < X \leq b) = P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b)$ $= F(b) - F(a) = \int_a^b f(x)dx$

Fijémonos que la probabilidad de los intervalos se corresponde con el área bajo la función de densidad dentro del intervalo considerado.



2.8 Caracterización de una variable aleatoria a través de parámetros

Hasta el momento hemos visto que toda variable aleatoria viene caracterizada a través de unas determinadas funciones matemáticas, las funciones de distribución y de densidad. Una vez caracterizada, y por tanto conocida, la distribución de una variable aleatoria, podemos obtener cualquier probabilidad asociada.

En ocasiones podemos acotar más el problema y reducir el estudio de una variable aleatoria a determinar una serie de características numéricas asociadas con la distribución de la variable. Dichas características tienen como propiedad fundamental el hecho de resumir gran parte de las propiedades de la variable aleatoria y juegan un papel muy destacado en las técnicas estadísticas que desarrollaremos a lo largo del curso.

Por ejemplo, supuesta la pertenencia de una variable aleatoria a una determinada familia de distribuciones de probabilidad, bien sea discreta o continua, los diferentes miembros de la familia diferirán en el valor de esas características numéricas. En este caso, denominaremos a tales características los parámetros de la distribución.

Existe un buen número de tales características, pero nos centraremos en las dos más importantes: la esperanza y la varianza. La primera nos informa sobre la localización de los valores de la variable y la segunda, sobre el grado de dispersión de estos valores.

2.9 Esperanza de una variable aleatoria discreta

La esperanza matemática de una variable aleatoria es una característica numérica que proporciona una idea de la localización de la variable aleatoria sobre la recta real. Decimos que es un parámetro de centralización o de localización.

Su interpretación intuitiva o significado se corresponde con el valor medio teórico de los posibles valores que pueda tomar la variable aleatoria, o también con el centro de gravedad de los valores de la variable supuesto que cada valor tuviera una masa proporcional a la función de densidad en ellos.

La definición matemática de la esperanza en el caso de las variables aleatorias discretas se corresponde directamente con las interpretaciones proporcionadas en el párrafo anterior. Efectivamente, supuesta una variable aleatoria discreta X con recorrido $\{x_1, x_2, \dots, x_k, \dots\}$ y con función de densidad $f(x)$, se define la esperanza matemática de X como el valor

$$E(X) = \sum_{x_i \in X(\Omega)} x_i f(x_i)$$

donde el sumatorio se efectúa para todo valor que pertenece al recorrido de X . En caso de que el recorrido sea infinito la esperanza existe si la serie resultante es absolutamente convergente, condición que no siempre se cumple.

La definición se corresponde con un promedio ponderado según su probabilidad de los valores del recorrido y, por tanto, se corresponde con la idea de un valor medio teórico.

2.10 Esperanza de una variable aleatoria continua

La idea intuitiva que más nos puede ayudar en la definición de la esperanza matemática de una variable aleatoria continua es la idea del centro de gravedad de los valores de la variable, donde cada valor tiene una masa proporcional a la función de densidad en ellos.

Dada una variable aleatoria absolutamente continua X con función de densidad $f(x)$, se define la esperanza matemática de X como el valor

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

suponiendo que la integral exista.

2.11 Propiedades de la esperanza matemática

1. Esperanza de una función de una variable aleatoria

- Variable discreta

$$E(h(X)) = \sum_{x_i \in X(\Omega)} h(x_i) f(x_i)$$

- Variable continua

$$E(h(X)) = \int_{-\infty}^{+\infty} h(x) f(x) dx$$

2.11.1 Linealidad de la esperanza matemática

- $E(X + Y) = E(X) + E(Y)$
- $E(k \cdot X) = k \cdot E(X)$ para todo número real k .
- $E(k) = k$ para todo número real k .
- $E(a \cdot X + b) = a \cdot E(X) + b$ para todo par de números reales a y b .

2.11.2 Esperanza del producto

- $E(X \cdot Y) = E(X) \cdot E(Y)$ únicamente en el caso de que X e Y sean variables aleatorias independientes.

2.12 Varianza de una variable aleatoria

La varianza de una variable aleatoria es una característica numérica que proporciona una idea de la dispersión de la variable aleatoria respecto de su esperanza. Decimos que es un parámetro de dispersión.

La definición es la siguiente:

$$\text{Var}(X) = E((X - E(X))^2)$$

Es, por tanto, el promedio teórico de las desviaciones cuadráticas de los diferentes valores que puede tomar la variable respecto de su valor medio teórico o esperanza.

En el caso de las variables discretas, la expresión se convierte en:

$$\text{Var}(X) = \sum_{x_i \in X(\Omega)} (x_i - E(X))^2 f(x_i)$$

mientras que para las variables continuas tenemos:

$$\text{Var}(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x) dx$$

En ambos casos existe una expresión equivalente alternativa y generalmente de cálculo más fácil:

$$\text{Var}(X) = E(X^2) - (E(X))^2$$

Una de las características de la varianza es que viene expresada en unidades cuadráticas respecto de las unidades originales de la variable. Un parámetro de dispersión derivado de la varianza y que tiene las mismas unidades de la variable aleatoria es la desviación típica, que se define como la raíz cuadrada de la varianza.

$$\sigma_X = \sqrt{\text{Var}(X)} = \sqrt{E((X - E(X))^2)}$$

2.12.1 Propiedades de la varianza

1. $\text{Var}(X) \geq 0$
2. $\text{Var}(k \cdot X) = k^2 \cdot \text{Var}(X)$ para todo numero real k .
3. $\text{Var}(k) = 0$ para todo numero real k .
4. $\text{Var}(a \cdot X + b) = a^2 \cdot \text{Var}(X)$ para todo par de números reales a i b .
5. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ únicamente en el caso que X y Y sean independientes.

2.13 Momentos (de orden k) de una variable aleatoria

- Dada una variable aleatoria X , definimos el momento de orden k como:

$$m_k = E(X^k)$$

suponiendo que tal esperanza exista. Podemos ver que la esperanza es el momento de orden 1, $E(X) = m_1$.

- Definimos el momento central de orden k como:

$$\mu_k = E((X - E(X))^k)$$

Con la denominación anterior, la varianza es el momento central de orden 2, $\text{Var}(X) = \mu_2$.

- Es posible también definir momentos mixtos de dos variables aleatorias. Dadas dos variables aleatorias X e Y definimos el momento mixto de orden (r, k) como

$$m_{rk} = E(X^r \cdot Y^k)$$

y el momento mixto central de orden (r, k) como

$$\mu_{rk} = E(X - E(X))^r \cdot (Y - E(Y))^k$$

- El momento mixto central más importante es el μ_{11} , denominado la covarianza de X e Y , y con una interpretación en el sentido de cuantificar el grado de dependencia entre dos variables aleatorias, puesto que si X e Y son independientes se verifica que $\mu_{11} = 0$, mientras que si $\mu_{11} \neq 0$ entonces las variables son dependientes.

2.14 Definición formal de variable aleatoria

Tal como hemos comentado, la definición formal de variable aleatoria impone una restricción matemática en la formulación vista hasta el momento.

Definiremos una variable aleatoria como una aplicación de Ω en el conjunto de números reales

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

que verifique la propiedad siguiente

$$\forall x \in \mathbb{R} \quad \text{el conjunto } A = \{\omega \mid X(\omega) \leq x\} \text{ es un suceso observable}$$

es decir, para todo número real x , el conjunto de resultados elementales tales que la variable aleatoria toma sobre ellos valores inferiores o iguales a x ha de ser un suceso sobre el cual podamos definir una probabilidad.

Dicha propiedad recibe el nombre de medibilidad y por tanto podríamos decir que una variable aleatoria es una función medible de Ω en los reales.

Esta condición nos asegura que podremos calcular sin problemas, probabilidades sobre intervalos de la recta real a partir de las probabilidades de los sucesos correspondientes.

$$P(X \leq x) = P\{\omega \mid X(\omega) \leq x\}$$

La expresión anterior se leería de la manera siguiente: La probabilidad de que la variable aleatoria tome valores inferiores o iguales a x es igual a la probabilidad del suceso formado por el conjunto de resultados elementales sobre los que el valor de la variable es menor o igual que x .

La probabilidad obtenida de esta manera se denomina probabilidad inducida. Se puede comprobar que, a partir de la condición requerida, se pueden obtener probabilidades sobre cualquier tipo de intervalo de la recta real. Por ejemplo:

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

La condición exigida para ser variable aleatoria discreta ahora puede ser expresada como:

$$\forall k = 1, 2, \dots \text{ el conjunto } A = \{\omega \mid X(\omega) = x_k\} = X^{-1}(\{x_k\}) \text{ es un suceso observable}$$

Toda variable aleatoria definida sobre un espacio de probabilidad finito es necesariamente discreta. La suma y el producto de variables aleatorias discretas, definido por:

$$(X + Y)(w) = X(w) + Y(w) \text{ y } (X \cdot Y)(w) = X(w) \cdot Y(w)$$

es también una variable aleatoria discreta.

2.15 Caso práctico: Lanzamiento de dos dados

2.15.1 Espacio muestral

Supongamos que estamos realizando un experimento consistente en el lanzamiento simultáneo de dos dados y en la observación del resultado obtenido.

El conjunto de resultados posibles forma el espacio muestral Ω asociado a dicho experimento. Sus elementos serán como los que se muestran a continuación:



En total, el espacio muestral estaría formado por 36 resultados posibles que, en principio y suponiendo los dados regulares, son todos ellos equiprobables con probabilidad $1/36$.

Nótese que consideramos diferentes resultados del tipo: un uno en el primer dado y un dos en el segundo o un dos en el primer dado y un uno en el segundo.

Una vez fijados los enunciados anteriores, es fácil asignar probabilidades a diferentes sucesos observables, por ejemplo:

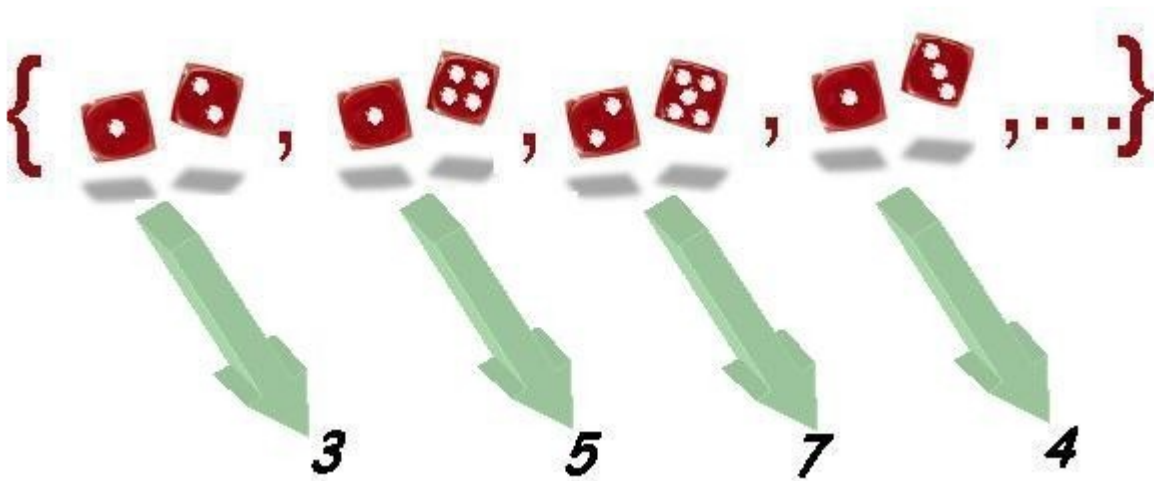
Suceso	Probabilidad
Que aparezcan dos cifras iguales	$6 \cdot 1/36 = 1/6$
Que la suma sea 10	$3 \cdot 1/36 = 1/12$

No entramos en detalles de la obtención de las probabilidades dado que se ha estudiado suficientemente en el tema anterior.

2.15.2 Representación numérica

Continuando con el experimento anterior, podemos representar los resultados obtenidos al lanzar dos dados por valores numéricos. ¿Cómo hacerlo? Definiendo una regla de asignación numérica para cada resultado.

Una posible regla sería, por ejemplo, asignar a cada resultado la suma de puntos de las caras. Este enunciado nos define una variable que representa cada suceso elemental por un valor numérico.

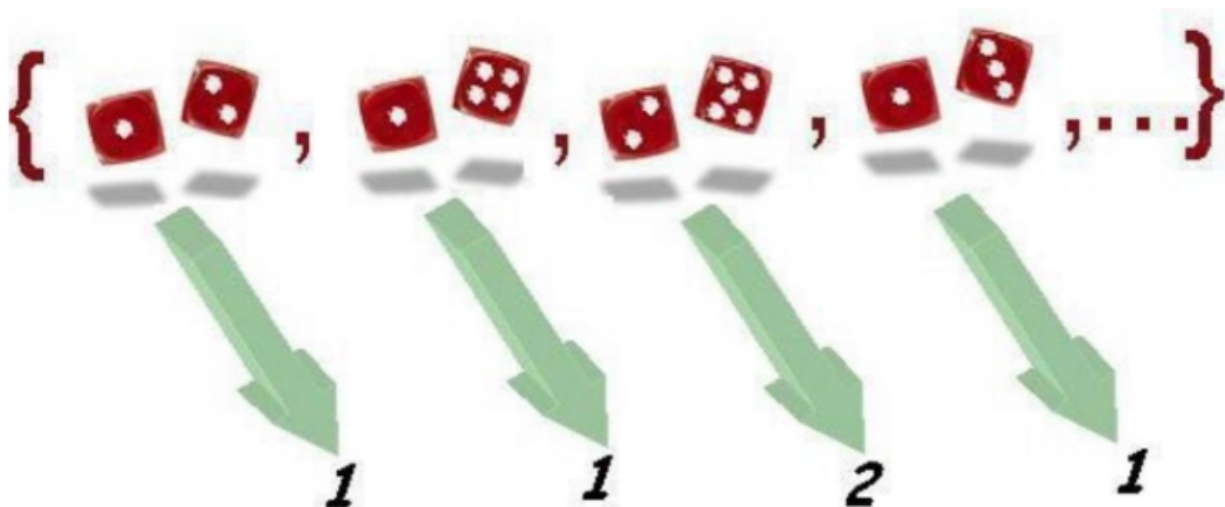


Los 36 posibles resultados del experimento se transforman en 11 posibles valores numéricos para la variable: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 y 12 .

Este conjunto de valores forman el *recorrido de la variable suma de puntos de las caras*. A partir de las probabilidades definidas sobre los sucesos observables es fácil extender las probabilidades a los diferentes resultados de la variable.

Por ejemplo, la probabilidad de que la variable tome el valor 10 es equivalente a la probabilidad del suceso observable que la suma sea 10 , calculada anteriormente e igual a $1/12$.

La variable considerada hasta el momento es sólo una de las múltiples variables que podríamos definir sobre el mismo experimento. Por ejemplo, podemos estar interesados no en la suma de puntos sino en el punto más bajo de cada tirada, de forma que podríamos construir una nueva variable a partir del enunciado o regla de asignación asignar a cada resultado el menor de los puntos de las dos caras. Tenemos una nueva variable sobre el mismo espacio anterior.

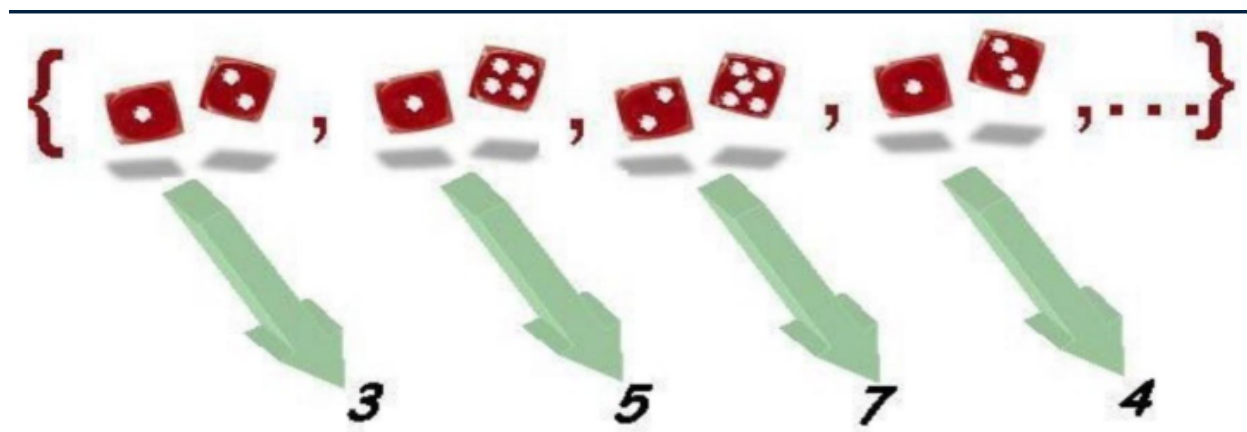


El recorrido, en este caso, está formado por los valores: 1, 2, 3, 4, 5 y 6 . Las dos variables estudiadas y otras muchas que se podrían definir sobre este experimento son ejemplos absolutamente equivalentes desde el punto de vista formal.

2.15.3 Algunas probabilidades

En el ejemplo de los dados vamos a centrarnos en la variable aleatoria

$$X = \text{Suma de puntos de las caras}$$



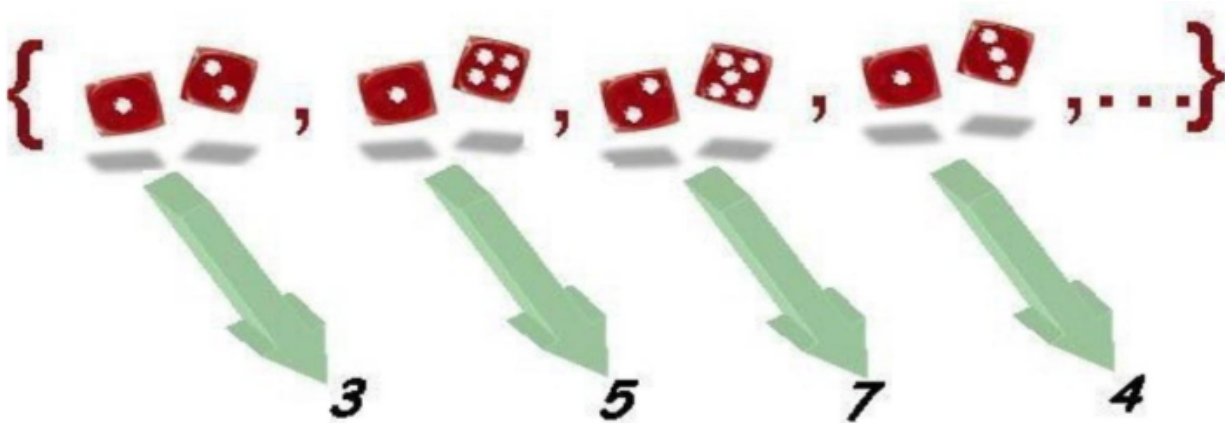
El recorrido de la variable está formado por los números $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11 \text{ i } 12\}$. Vamos a calcular algunas probabilidades:

- $P(X \leq 1) = P\{\emptyset\} = 0$ (Ningún resultado tiene asignado un valor menor o igual a 1)
- $P(X \leq 2) = P\{(1, 1)\} = 1/36$ (Sólo hay un caso al que se le asigne un valor inferior o igual a 2).
- $P(X \leq 3.5) = P\{(1, 1), (1, 2), (2, 1)\} = 3/36$ (Tres resultados elementales tienen asignado un valor menor o igual a 3.5)

Ahora podéis intentar calcular por vosotros mismos algunas probabilidades: (a) $P(X \leq 6)$ (b) $P(X \leq 8, 2)$; (c) $P(X \leq 12)$; (d) $P(X \leq 20)$ i (e) $P(2, 2 < X \leq 7)$

2.15.4 Función de distribución

Para calcular la función de distribución de la variable $X = \text{Suma de puntos de las caras}$:



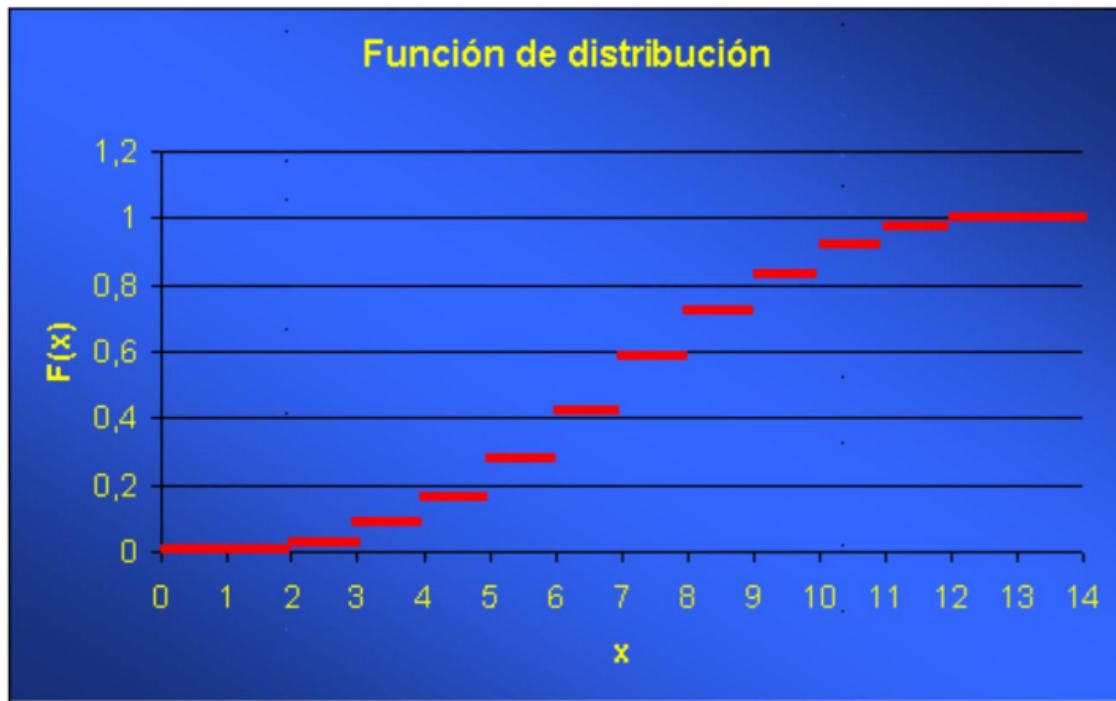
necesitamos conocer el recorrido de la variable, que es: $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ y, utilizando este recorrido como pauta, determinar *para todo punto x de la recta real* la probabilidad $P(X \leq x)$.

En nuestro ejemplo:

$$F(x) = P(X \leq x) = \begin{cases} 0 & x < 2 \\ 1/36 & 2 \leq x < 3 \\ 3/36 & 3 \leq x < 4 \\ 6/36 & 4 \leq x < 5 \\ 10/36 & 5 \leq x < 6 \\ 15/36 & 6 \leq x < 7 \\ 21/36 & 7 \leq x < 8 \\ 26/36 & 8 \leq x < 9 \\ 30/36 & 9 \leq x < 10 \\ 33/36 & 10 \leq x < 11 \\ 35/36 & 11 \leq x < 12 \\ 36/36 = 1 & x \geq 12 \end{cases}$$

Acabamos de construir la función de distribución de la variable suma de la puntuación al lanzar dos dados.

Vamos a ver su representación gráfica:



Ejercicio : Haced lo mismo para la variable aleatoria el menor de los puntos de las dos caras al lanzar dos dados.

2.15.5 Clasificación de las variables

En el experimento que estamos considerando, lanzar simultáneamente dos dados, cualquiera de las dos variables aleatorias que hemos considerado hasta el momento:

X = Suma los puntos de las dos caras

Y = El menor de los puntos de las dos caras

se clasifican dentro del tipo de variables aleatorias discretas, puesto que en ambos casos el recorrido es finito: $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ para la variable X y $\{1, 2, 3, 4, 5, 6\}$ para la variable Y .

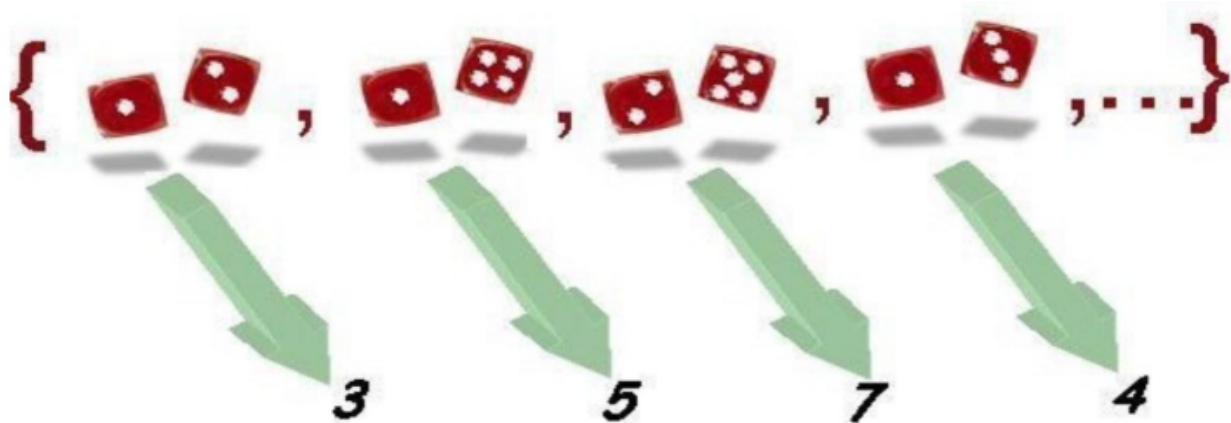
También son discretas aquellas variables aleatorias con recorrido infinito numerable.

Ejercicio: ¿Sabrías construir una variable aleatoria discreta con recorrido infinito numerable basada en el experimento que consiste en el lanzamiento de dos dados?

2.15.6 Función de densidad discreta

Para calcular la función de densidad de la variable

X = suma de puntos de las caras



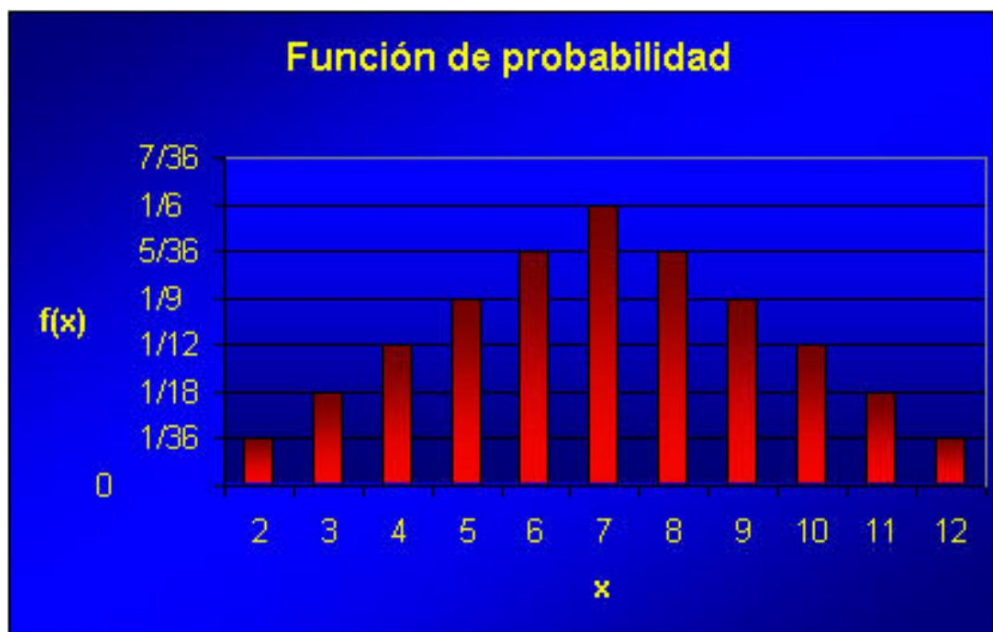
necesitamos conocer el recorrido de la variable, es decir: $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ y, a partir del recorrido, determinar para todo punto del recorrido la probabilidad $P(X = x)$.

En nuestro ejemplo

$$f(x) = P(X = x) = \begin{cases} 1/36 & x = 2 \\ 2/36 & x = 3 \\ 3/36 & x = 4 \\ 4/36 & x = 5 \\ 5/36 & x = 6 \\ 6/36 & x = 7 \\ 5/36 & x = 8 \\ 4/36 & x = 9 \\ 3/36 & x = 10 \\ 2/36 & x = 11 \\ 1/36 & x = 12 \end{cases}$$

Acabamos de construir la función de densidad de la variable suma de la puntuación al lanzar dos dados.

Vamos a ver su representación gráfica:



Hemos optado por la representación con barras en lugar de puntos para permitir una visualización de la función óptima.

Ejercicio: Haced lo mismo para la variable aleatoria el menor de los puntos de las dos caras al lanzar dos dados.

2.15.7 Probabilidad de intervalos

Vamos a centrarnos en la variable

$$X = \text{Suma de puntos de las caras}$$

Las funciones de distribución y de densidad son, respectivamente,

$$F(x) = P(X \leq x) = \begin{cases} 0 & x < 2 \\ 1/36 & 2 \leq x < 3 \\ 3/36 & 3 \leq x < 4 \\ 6/36 & 4 \leq x < 5 \\ 10/36 & 5 \leq x < 6 \\ 15/36 & 6 \leq x < 7 \\ 21/36 & 7 \leq x < 8 \\ 26/36 & 8 \leq x < 9 \\ 30/36 & 9 \leq x < 10 \\ 33/36 & 10 \leq x < 11 \\ 35/36 & 11 \leq x < 12 \\ 36/36 = 1 & x \geq 12 \end{cases} \quad f(x) = P(X = x) = \begin{cases} 1/36 & x = 2 \\ 2/36 & x = 3 \\ 3/36 & x = 4 \\ 4/36 & x = 5 \\ 5/36 & x = 6 \\ 6/36 & x = 7 \\ 5/36 & x = 8 \\ 4/36 & x = 9 \\ 3/36 & x = 10 \\ 2/36 & x = 11 \\ 1/36 & x = 12 \end{cases}$$

Puede observarse cómo los valores de la función de distribución se obtienen acumulando los valores de la función de densidad correspondientes.

Vamos a calcular algunas probabilidades utilizando las funciones anteriores. Compárese con los resultados obtenidos con anterioridad basados directamente en los resultados elementales.

- $P(X \leq 1) = F(1) = 0$

- $P(X \leq 3, 5) = F(3, 5) = 3/36 = f(2) + f(3)$
- $P(X < 6) = F(6) - f(6) = 15/36 - 5/36 = 10/36 = f(2) + f(3) + f(4) + f(5)$
- $P(2, 2 < X \leq 7) = F(7) - F(2, 2) = 21/36 - 1/36 = 20/36 = f(3) + f(4) + f(5) + f(6) + f(7)$
- $P(2 < X < 7) = F(7) - f(7) - F(2) = 21/36 - 6/36 - 1/36 = 14/36 = f(3) + f(4) + f(5) + f(6)$

2.15.8 Esperanza

Supongamos que estamos interesados en determinar cual sería el valor medio teórico de la variable

$$X = \text{Suma de puntos de las caras}$$

La función de densidad es:

$$f(x) = P(X = x) = \begin{cases} 1/36 & x = 2 \\ 2/36 & x = 3 \\ 3/36 & x = 4 \\ 4/36 & x = 5 \\ 5/36 & x = 6 \\ 6/36 & x = 7 \\ 5/36 & x = 8 \\ 4/36 & x = 9 \\ 3/36 & x = 10 \\ 2/36 & x = 11 \\ 1/36 & x = 12 \end{cases}$$

La misma función de densidad nos da información sobre el recorrido de la variable. Calcular el valor medio teórico de la variable quiere decir calcular la esperanza. A partir de la fórmula de la esperanza para variables discretas, tenemos

$$\begin{aligned} E(X) &= 2 \cdot 1/36 + 3 \cdot 2/36 + 4 \cdot 3/36 + 5 \cdot 4/36 + 6 \cdot 5/36 + \\ &\quad + 7 \cdot 6/36 + 8 \cdot 5/36 + 9 \cdot 4/36 + \\ &\quad + 10 \cdot 3/36 + 11 \cdot 2/36 + 12 \cdot 1/36 = \\ &= 7 \end{aligned}$$

Por tanto, 7 es la esperanza de la variable $X = \text{Suma de puntos de las caras}$. Fijaos que la esperanza para la variable Puntuación de un dado sería

$$1 \cdot 1/6 + 2 \cdot 1/6 + 3 \cdot 1/6 + 4 \cdot 1/6 + 5 \cdot 1/6 + 6 \cdot 1/6 = 3,5$$

y que se puede considerar la variable Suma de puntos de las dos caras como la suma de dos variables que representen la puntuación de cada dado. La esperanza de la suma es, efectivamente, la suma de las esperanzas de cada variable sumada.

En la aplicación siguiente, podéis calcular la esperanza de la variable Puntuación de un dado y modificar las probabilidades de las diferentes caras, de este modo se modifica la esperanza.

Ejercicio: ¿Podrías hacer lo mismo para la variable $X = \text{El menor de los puntos de las dos caras al lanzar dos dados}$?

2.15.9 Esperanza de un juego

Imaginemos que alguien os propone el juego siguiente: lanzad dos dados, si la suma obtenida es menor o igual a 6 ganáis 100 euros, sin embargo, si la suma obtenida es mayor que 6 tenéis que pagar 100 euros. ¿Nos conviene jugar a este juego?

Veamos, podemos considerar el resultado del juego como una variable aleatoria discreta que toma dos valores: +100 si ganamos y -100 si perdemos. Nos interesa conocer las probabilidades de los diferentes resultados. Consideremos la variable X = Suma de puntos de las caras, cuya función de densidad conocemos:

$$f(x) = P(X = x) = \begin{cases} 1/36 & x = 2 \\ 2/36 & x = 3 \\ 3/36 & x = 4 \\ 4/36 & x = 5 \\ 5/36 & x = 6 \\ 6/36 & x = 7 \\ 5/36 & x = 8 \\ 4/36 & x = 9 \\ 3/36 & x = 10 \\ 2/36 & x = 11 \\ 1/36 & x = 12 \end{cases}$$

A partir de aquí es fácil ver que la función de densidad de la variable Y = Resultado del juego será la siguiente:

$$f(100) = 15/36; f(-100) = 21/36$$

Por tanto, la esperanza del juego, que puede ser interpretada como la ganancia media por jugada, será

$$E(Y) = 100 \cdot 15/36 - 100 \cdot 21/36 = -100/6 \approx -16,667$$

Es decir, la ganancia media por jugada es negativa, por tanto no es favorable dicho juego para el jugador, es un juego no equitativo.

2.15.10 Esperanza con recorrido infinito

Vamos a tratar de calcular la esperanza de la siguiente variable aleatoria: X = Número de lanzamientos que hemos de hacer para conseguir que aparezca un doble seis. La variable que acabamos de definir es una variable discreta con recorrido infinito numerable. El recorrido sería el siguiente:

$$\{1, 2, 3, 4, \dots\}$$

Vamos a ver como calculamos la función de densidad: $P(X = 1)$ = Probabilidad de que aparezca un doble seis en el primer lanzamiento = $1/36$ $P(X = 2)$ = Probabilidad de que el doble seis no aparezca en el primer lanzamiento y sí en el segundo = $35/36 \cdot 1/36 = 35/36^2$ $P(X = 3)$ = Probabilidad de que el doble seis no aparezca ni en el primer ni en el segundo lanzamientos y sí en el tercero = $35/36 \cdot 35/36 \cdot 1/36 = 35^2/36^3$

En general, $P(X = k) = 35^{k-1}/36^k$. Para simplificar, vamos a llamar $p = 1/36$ y $q = 1 - p = 35/36$, con esta nomenclatura $P(X = k) = q^{k-1}p$. Por tanto, la esperanza será:

$$\begin{aligned}
E(X) &= \sum_{i=1}^{\infty} i q^{i-1} p = p \sum_{i=1}^{\infty} i q^{i-1} = p \frac{d}{dq} \sum_{i=1}^{\infty} q^i = \\
&= p \frac{d}{dq} \left(\frac{q}{1-q} \right) = p \frac{1}{(1-q)^2} = \\
&= \frac{1}{p}
\end{aligned}$$

En nuestro ejemplo el número medio de tiradas antes de salir un doble seis será 36 .

2.15.11 Esperanza infinita

Ahora calcularemos la esperanza del juego siguiente: lanzamos un dado hasta que aparece un número par, el jugador gana 2^n unidades monetarias si aparece un número par por primera vez en la tirada n -ésima.

El recorrido de la variable aleatoria X = Ganancia del juego, está formado por todos los números de la forma 2^n con $n = 1, 2, 3, \dots$. La probabilidad de cada valor del recorrido es la probabilidad de que aparezca un número par por primera vez en la tirada n -ésima, es decir $(1/2)^{n-1} \cdot (1/2) = (1/2)^n$. Por tanto, la esperanza del juego es la siguiente:

$$E(X) = \sum_{n=1}^{\infty} 2^n (1/2)^n = \sum_{n=1}^{\infty} 1 = \infty$$

Como vemos, la variable aleatoria X no tiene esperanza finita. El enunciado presentado es una versión del problema presentado alrededor de 1730 por el matemático Daniel Bernoulli a la Academia de San Petersburgo y conocido como la paradoja de San Petersburgo, dado que la esperanza del juego es aparentemente infinita.

2.15.12 Varianza

Si ahora queremos calcular la varianza de la variable

$$X = \text{Suma de puntos de las caras}$$

con función de densidad:

$$f(x) = P(X = x) = \begin{cases} 1/36 & x = 2 \\ 2/36 & x = 3 \\ 3/36 & x = 4 \\ 4/36 & x = 5 \\ 5/36 & x = 6 \\ 6/36 & x = 7 \\ 5/36 & x = 8 \\ 4/36 & x = 9 \\ 3/36 & x = 10 \\ 2/36 & x = 11 \\ 1/36 & x = 12 \end{cases}$$

Podemos aplicar la fórmula

$$\text{Var}(X) = E(X^2) - (E(X))^2$$

La esperanza ya la tenemos calculada con anterioridad

$$\begin{aligned} E(X) &= 2 \cdot 1/36 + 3 \cdot 2/36 + 4 \cdot 3/36 + 5 \cdot 4/36 + \\ &\quad + 6 \cdot 5/36 + 7 \cdot 6/36 + 8 \cdot 5/36 + 9 \cdot 4/36 + \\ &\quad + 10 \cdot 3/36 + 11 \cdot 2/36 + 12 \cdot 1/36 = \\ &= 7 \end{aligned}$$

Necesitamos calcular la esperanza de la variable al cuadrado, que en este caso resulta:

$$\begin{aligned} E(X^2) &= 2^2 \cdot 1/36 + 3^2 \cdot 2/36 + 4^2 \cdot 3/36 + 5^2 \cdot 4/36 + 6^2 \cdot 5/36 + \\ &\quad + 7^2 \cdot 6/36 + 8^2 \cdot 5/36 + 9^2 \cdot 4/36 + 10^2 \cdot 3/36 + \\ &\quad + 11^2 \cdot 2/36 + 12^2 \cdot 1/36 = 329/6 \\ &\approx 54,833 \end{aligned}$$

Con lo que la varianza resulta ser

$$\text{Var}(X) = 329/6 - 7^2 = 35/6 \approx 5,833$$

Nuevamente, para la variable Puntuación de un dado, la varianza se obtendría de la manera siguiente:

$$\begin{aligned} E(X) &= 1 \cdot 1/6 + 2 \cdot 1/6 + 3 \cdot 1/6 + 4 \cdot 1/6 + 5 \cdot 1/6 + 6 \cdot 1/6 = \\ &= 3,5 \\ E(X^2) &= 1^2 \cdot 1/6 + 2^2 \cdot 1/6 + 3^2 \cdot 1/6 + 4^2 \cdot 1/6 + \\ &\quad + 5^2 \cdot 1/6 + 6^2 \cdot 1/6 = 91/6 \\ &\approx 15,167 \\ \text{Var}(X) &= 91/6 - 3,5^2 = 35/12 \approx 2,9167 \end{aligned}$$

y se cumple que la varianza de la variable Suma de puntos de las dos caras es la suma de las varianzas de las puntuaciones de cada dado por separado. Recordemos que esto sólo sucede si las variables sumadas son independientes, como así ocurre con las puntuaciones de cada dado por separado.

3 Distribuciones Notables

3.1 Distribuciones discretas

3.1.1 La distribución de Bernoulli

Es el modelo discreto más sencillo en que podamos pensar. Hace referencia a situaciones en las que el resultado de un experimento sólo puede ser: se ha dado el suceso A ó no se ha dado el suceso A . Por ejemplo, en el lanzamiento de una moneda sólo puede darse el suceso sale cara o su complementario no sale cara (sale cruz).

Por lo tanto, definimos la variable aleatoria X de la siguiente manera:

- $X = 1$ si se ha dado A .
- $X = 0$ si no se ha dado A , es decir, se ha dado el complementario A^c .

Si además, conocemos la probabilidad de que suceda A :

$$P[A] = p$$

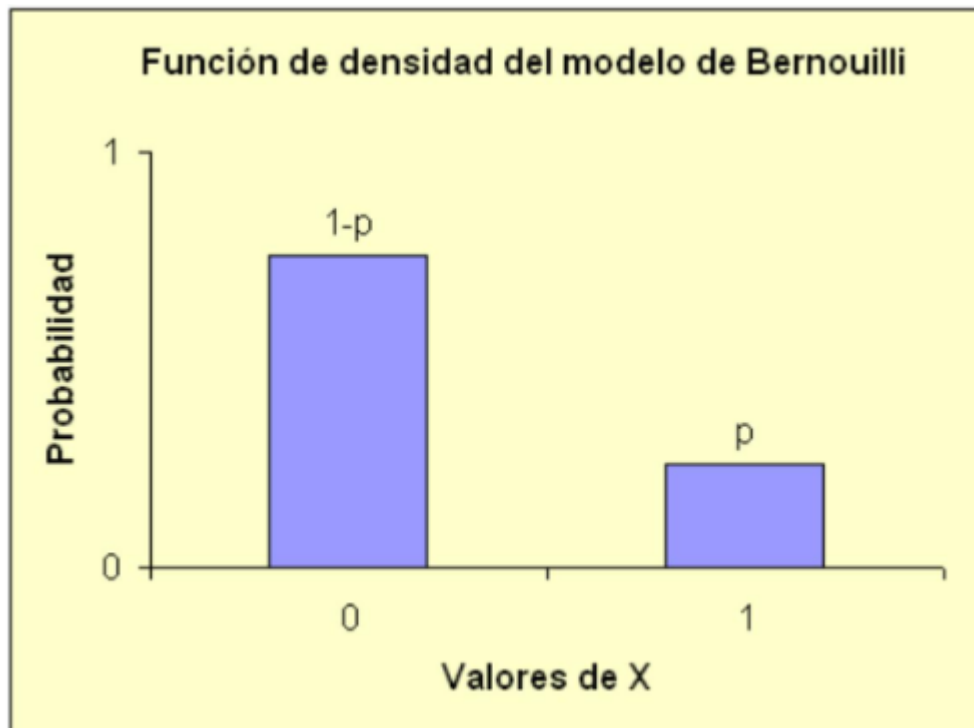
y, por tanto,

$$P[A^c] = 1 - p$$

ya podemos definir la distribución de la variable aleatoria X . En estas condiciones diremos que X sigue una distribución de Bernoulli de parámetro p , que abreviaremos así $X \sim \text{Bernoulli}(p)$, y su función de densidad se define así:

$$f(k) = P[X = k] = \begin{cases} p & \text{si } k = 1 \text{ (se ha dado } A \text{)} \\ 1 - p & \text{si } k = 0 \text{ (se ha dado } A^c \text{)} \end{cases}$$

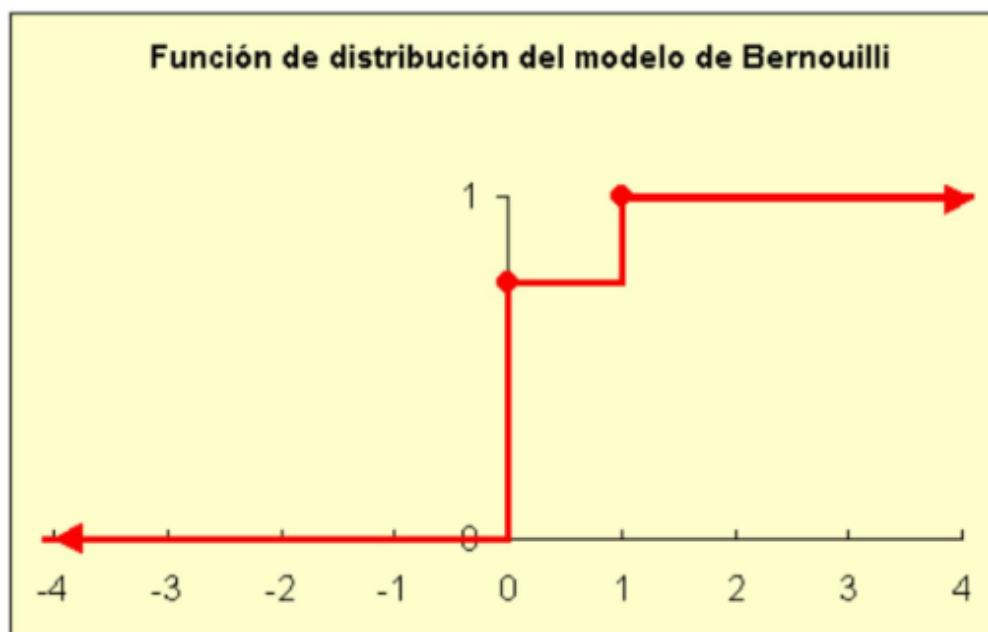
Gráficamente:



Mientras que la función de distribución será:

$$F(k) = P[X \leq k] = \begin{cases} 0 & \text{si } k < 0 \\ p & \text{si } 0 \leq k < 1 \\ 1 & \text{si } k \geq 1 \end{cases}$$

Gráficamente:



3.1.1.1 Propiedades del modelo de Bernoulli

- 1) La esperanza vale $E(X) = p$.
- 2) La varianza vale $V(X) = p(1 - p)$.

3.1.2 La distribución Binomial

Al igual que el modelo de Bernoulli, hace referencia a experiencias con resultados dicotómicos (el resultado sólo puede ser A o A^c). Sin embargo en este modelo estamos interesados en la repetición de n veces una experiencia de este tipo en condiciones independientes.

Tomemos el ejemplo del conteo del número de caras en el lanzamiento n veces de una moneda regular. Para concretar, vamos a suponer que disponemos de una moneda regular ($P[\text{cara}] = P[\text{cruz}] = 1/2$) que lanzamos cuatro veces. Es evidente que, en estas condiciones, la variable X : número de caras en cuatro lanzamientos independientes de una moneda regular es una variable aleatoria discreta que sólo puede tomar cinco posibles valores:

$$x = 0, 1, 2, 3, 4$$

Pasemos ahora a calcular la probabilidad de cada valor (en terminología estadística, vamos a calcular la función de densidad de la variable X).

Es evidente que la $P[X = 0]$ es igual a la probabilidad de salgan cuatro cruces seguidas:

$$P[X = 0] = P[\text{cruz}, \text{cruz}, \text{cruz}, \text{cruz}] = P[\text{cruz}]^4 = (1/2)^4 = 0,0625$$

ya que la moneda es regular y, por tanto, $P[\text{cara}] = P[\text{cruz}] = 1/2$. La $P[X = 3]$ corresponde al suceso de que salgan tres caras (c en adelante) y una cruz ($+$ en adelante). Sin embargo, en este caso tenemos hasta cuatro posibles maneras de obtener dicho resultado, según el orden en que aparezcan las tres caras y la cruz:

$+ccc$	$c + cc$	$cc + c$	$ccc +$
--------	----------	----------	---------

También debería resultar evidente que la probabilidad de cada uno de estos sucesos es la misma:

$$P[+ccc] = P[c + cc] = P[cc + c] = P[ccc+] = (1/2)^4 = (1/2)^4 = 0,0625$$

de manera que, finalmente, la probabilidad de que salgan tres caras y una cruz es la suma de las probabilidades de los 4 casos anteriores:

$$P[X = 3] = 4(1/2)^4 = 0,25$$

Y así podríamos ir calculando el resto de casos. Podemos ver que, en este ejemplo, todos los casos tienen la misma probabilidad (0,0625) y que el número total de casos posibles es 16. En términos de combinatoria dicho número se obtendría como variaciones con repetición de dos valores (cara o cruz) tomados de cuatro en cuatro (el número de lanzamientos de la moneda):

$$VR_2^4 = 2^4 = 16$$

En la siguiente tabla se muestran los dieciséis posibles resultados:

$k = \text{número de caras}$	Casos
0	+++++
1	+++c
	+ + c +
	+ c + +
	c + + +
	++cc
	+c + c
	c + +c +
	c + c +
	cc++
	ccc+
	c + cc

Si hacemos uso de nuestros conocimientos de combinatoria, comprobamos que el número de casos para cada posible valor k ($k = 0, 1, 2, 3, 4$) puede calcularse como permutaciones con repetición de cuatro elementos tomado de k y $4 - k$:

$$RP_4^{k,4-k} = \frac{4!}{k!(4-k)!} = \binom{4}{k}$$

y obtenemos finalmente el número combinatorio 4 sobre k . En efecto, para el caso $k = 3$, tendríamos:

$$\binom{4}{3} = \frac{4!}{3!1!} = 4$$

que son los cuatro posibles casos que nos dan tres caras y una cruz. Finalmente, recordando que todos los casos tienen la misma probabilidad, se construye la siguiente tabla:

$k = \text{número de caras}$	Número de casos	$P[X = k]$
0	1	0,0625
1	4	0,2500

2	6	0,3750
3	4	0,2500
4	1	0,0625
Total	16	1

3.1.2.1 Los parámetros de la distribución Binomial La última tabla de la página anterior es, justamente, la función de densidad de nuestra variable X .

Función de densidad de X	
k	$P[X = k]$
0	0,0625
1	0,2500
2	0,3750
3	0,2500
4	0,0625
En otro caso	0

Como hemos visto, para obtener los resultados anteriores, hemos tenido que definir dos valores:

1. n : el número de lanzamientos (repeticiones de la experiencia aleatoria en condiciones independientes), en nuestro caso $n = 4$.
2. p : la probabilidad de que salga cara ($P[c]$), en nuestro caso $p = 1/2$.

Se dice, por tanto, que la distribución Binomial depende de dos parámetros: n y p . En nuestro ejemplo, diremos que X sigue una distribución Binomial de parámetros $n = 4$ i $p = 1/2$. De forma abreviada:

$$X \sim B(n = 4; p = 1/2)$$

En el ejemplo que hemos visto, suponíamos que la moneda era regular y, por tanto,

$$P[c] = P[+] = 1/2$$

Si tenemos una moneda trucada con las siguientes probabilidades:

$$P[c] = 2/3 \quad \text{i} \quad P[+] = 1/3$$

diremos que en este caso la variable X : número de caras en cuatro lanzamientos independientes de nuestra moneda trucada sigue una distribución Binomial de parámetros:

$$X \sim B(n = 4; p = 2/3)$$

El problema se nos complica levemente ya que ahora no todos los posibles resultados tienen la misma probabilidad. Veamos dos ejemplos:

- La probabilidad de obtener cuatro caras es:

$$P[cccc] = (2/3)^4 = 0,1975$$

- La probabilidad de que el primer lanzamiento sea cara y el resto sean cruces valdrá:

$$P[c^{+++}] = (2/3)'(1/3)^3 = 0,0247$$

Sin embargo sí se cumplirá que la probabilidad de que todos los caso que resulten en el mismo número de caras y cruces tendrán la misma probabilidad. Por ejemplo, para los cuatro casos en los que el número total de caras es 1 y el de cruces 3 :

$$P[c + ++] = P[+c ++] = P[+ + c+] = P[+ + +c] = (2/3)'(1/3)^3 = 0,0247$$

Y, por tanto, la probabilidad de obtener una sola cara en el lanzamiento de nuestra moneda trucada será:

$$P[X = 1] = 4'0,0247 = 0,0988$$

O, generalizando, si $P[A] = p$ y $P[A^c] = 1 - p$ tenemos que

$$P[X = k] = c(n, k)p^k(1 - p)^{n-k} \quad \text{si } k = 0, 1, \dots, n$$

donde $c(n, k)$ representa el número de posibles resultados en los que obtenemos k caras y $n - k$ cruces en n lanzamientos. Tal como hemos visto, dicho número se puede calcular como permutaciones con repetición de n unidades tomadas de k y $n - k$.

Todo lo anterior nos lleva a formular el model binoial a traves de la siguiente función de densidad:

$$f(k) = P[X = k] = \begin{cases} \binom{n}{k} p^k (1 - p)^{n-k} & \text{si } k = 0, \dots, n \\ 0 & \text{en caso contrario} \end{cases}$$

con lo que la función de distribución se calcularía:

$$F(k) = P[X \leq k] = \begin{cases} 0 & \text{si } k < 0 \\ \sum_{i=0}^k \binom{n}{i} p^i (1 - p)^{n-i} & \\ 1 & \text{si } k \geq n \end{cases}$$

3.1.2.2 Propiedades del modelo Binomial

1. La esperanza vale $E(X) = np$.
2. La varianza es $V(X) = np(1 - p)$.
3. Es una generalización del modelo de Bernouilli. En efecto, la Binomial con $n = 1$ (una sola realización) coincide con la distribución de Bernouilli.
4. La suma de dos variables aleatorias binomiales independientes con igual parámetro p también sigue una distribución Binomial:

$$X_1 \sim B(n = n_1; p = p_0) \quad \text{i} \quad X_2 \sim B(n = n_2; p = p_0)$$

Si definimos $Z = X_1 + X_2$ entonces,

$$Z \sim B(n = n_1 + n_2; p = p_0)$$

3.1.3 La distribución de Poisson

Se trata de un modelo discreto, pero en el que el conjunto de valores con probabilidad no nula no es finito, sino numerable. Se dice que una variable aleatoria X sigue la distribución de Poisson si su función de densidad viene dada por:

$$f(k) = P[X = k] = \begin{cases} e^{-\lambda} \frac{\lambda^k}{k!} & \text{si } k = 0, 1, 2, \dots \\ 0 & \text{en caso contrario} \end{cases}$$

Como vemos, este modelo se caracteriza por un sólo parámetro λ , que debe ser positivo. Esta distribución suele utilizarse para contajes del tipo número de individuos por unidad de tiempo, de espacio, etc.

3.1.3.1 Propiedades del modelo de Poisson

1. Esperanza: $E(X) = \lambda$.
2. Varianza: $V(X) = \lambda$.

En esta distribución la esperanza y la varianza coinciden.

3. La suma de dos variables aleatorias independientes con distribución de Poisson resulta en una nueva variable aleatoria, también con distribución de Poisson, de parámetro igual a la suma de parámetros:

$$X_1 \sim P(\lambda = \lambda_1) \quad \text{y} \quad X_2 \sim P(\lambda = \lambda_2)$$

y definimos $Z = X_1 + X_2$, entonces,

$$Z \sim P(\lambda = \lambda_1 + \lambda_2)$$

Este resultado se extiende inmediatamente al caso de n variables aleatorias independientes con distribución de Poisson. En este caso, la variable suma de todas ellas sigue una distribución de Poisson de parámetro igual a la suma de los parámetros.

3.1.4 La distribución Uniforme discreta

Tenemos esta distribución cuando el resultado de una experiencia aleatoria puede ser un conjunto finito de n posibles resultados, todos ellos igualmente probables.

Un ejemplo puede ser la variable X , puntuación en el lanzamiento de un dado regular. Esta variable toma seis valores posibles, todos con la misma probabilidad $p = 1/6$. La función de densidad de esta variable será:

$$f(k) = P[X = k] = 1/6 \quad k = 1, 2, 3, 4, 5, 6$$



En general, si la variable X puede tomar n ($k = 1, 2, \dots, n$) valores, todos con igual probabilidad, su función de densidad será:

$$f(k) = P[X = k] = 1/n \quad k = 1, 2, \dots, n$$

3.1.4.1 Propiedades del modelo Uniforme discreto Sea n el número de valores equiprobables posibles:

3.1.4.2 Esperanza:

$$E(X) = \frac{n+1}{2}$$

3.1.4.3 Varianza:

$$V(X) = \frac{(n+1)[2(2n+1) - 3(n+1)]}{12}$$

3.1.5 La distribución Hipergeométrica

Este modelo presenta similitudes con el Binomial, pero sin la suposición de independencia de éste último. Veámoslo:

- Partimos de un conjunto formado por N individuos divididos en dos categorías mutuamente excluyentes: A y A^c ; de manera que N_1 individuos pertenecen a la categoría A y N_2 individuos, a la categoría A^c . Por tanto, se cumple que

$$N = N_1 + N_2$$

- Si del conjunto anterior extraemos n individuos sin reemplazamiento ($n \leq N$), la variable X que representa el número k de individuos que pertenecen a la categoría A (de los n extraídos) tiene por función de densidad:

$$f(k) = P[X = k] = \frac{\binom{N_1}{k} \binom{N_2}{n-k}}{\binom{N}{n}}$$

si $\max\{0, n - N_2\} \leq k \leq \min\{N_1, n\}$

La dependencia se debe al hecho de que N es finito y las extracciones se efectúan sin reemplazamiento. El caso de extracciones con reemplazamiento sería equivalente al de N infinito y se resolvería mediante el modelo Binomial.

3.1.5.1 Propiedades del modelo hipergeométrico

1. Esperanza: $E(X) = nN_1/N_2$.
2. Varianza: $V(X) = (nN_1N_2(N - n)) / (N_2(N - 1))$

3.1.6 La distribución Geométrica o de Pascal

Definamos una experiencia aleatoria cuyo resultado sólo puede ser el suceso A o su complementario A^c , y que se repite secuencialmente hasta que aparece el suceso A por primera vez.

Definamos la variable aleatoria X como el número de veces que repetimos la experiencia en condiciones independientes hasta que se dé A por primera vez. Bajo estas condiciones, decimos que la variable X sigue una distribución geométrica o de Pascal de parámetro $p = P(A)$.

La función de densidad puede deducirse fácilmente de la definición:

$$f(k) = P[X = k] = (1 - p)^k p \quad k = 0, 1, 2, \dots$$

En el programa siguiente podéis ver su forma y obtener los valores de la función de densidad y de la de distribución:

Algunas puntualizaciones de la definición de X :

- Notése que, en esta definición, condiciones independientes significa que p , la probabilidad de A , y $1 - p$, la de su complementario A^c , no varían a lo largo de las sucesivas repeticiones de la experiencia.
- Tal y como la hemos definido, X se refiere al número de lanzamientos hasta que se produce A , pero sin contabilizar el último caso en que se da A . Por dicha razón X puede tomar los valores $k = 0, 1, 2, \dots$ con probabilidad no nula.

Un ejemplo de este modelo podría ser la experiencia consistente en lanzar sucesivamente un dado regular hasta que aparezca el número 6 . Si definimos la variable aleatoria X como el número de lanzamientos de un dado regular hasta que aparezca un 6 , queda claro que X sigue una distribución geométrica de parámetro $p = 1/6$.

3.1.6.1 Propiedades del modelo Geométrico o de Pascal

- 1) Esperanza: $E(X) = (1 - p)/p$
- 2) Varianza: $V(X) = (1 - p)/p^2$

3.1.6.2 Preguntas:

- ¿A que suceso nos referimos cuando decimos $X = 0$? Respuesta.
 - Cuando decimos que $X = 0$ nos referimos al caso en que el 6 aparece en el primer lanzamiento. La probabilidad de que esto suceda, suponiendo un dado regular, es de $1/6$:

$$P[X = 0] = 1/6$$

- ¿Cuál es la probabilidad de que el primer 6 aparezca en el cuarto lanzamiento? Respuesta.
 - La probabilidad de que el primer 6 aparezca en el cuarto lanzamiento corresponde a:

$$P[X = 3] = (5/6)^3 \cdot 1/6 = 0,0965$$

Fijémonos en que, si definimos A como el suceso sale un 6, la probabilidad anterior corresponde a la del suceso: $\{A^c A^c A^c A\}$ (en este orden).

3.1.7 La distribución Binomial negativa

Puede definirse como una generalización del modelo Geométrico o de Pascal. Así, dado un suceso A y su complementario A^c , cuando X representa el número de veces que se da A^c (ausencias, fallos, etc.) hasta que se produce r veces el suceso A , en una serie de repeticiones de la experiencia aleatoria en condiciones independientes, decimos que X sigue la distribución Binomial negativa. Nótese que, cuando $r = 1$, tenemos exactamente el modelo geométrico.

Este modelo queda definido por dos parámetros p (la probabilidad de A : $p = P(A)$) y r (el número de veces que debe producirse A para que detengamos la experiencia).

La función de densidad viene dada por:

$$f(k) = P[X = k] = \binom{k+r-1}{r-1} p^r q^k \quad k = 0, 1, 2, \dots$$

donde q representa el complementario de p : $q = 1 - p$.

3.1.7.1 Propiedades del modelo Binomial negativo

1. Esperanza: $E(X) = r'q/p$
2. Varianza: $V(X) = r'q/p^2$
3. Se cumplen las siguientes propiedades respecto la función de densidad:

$$f(0) = p^r \quad \text{y} \quad f(k+1) = \frac{(1-p)(k+r)}{k+1} f(k)$$

4. Este modelo se ajusta bien a contajes (números de individuos por unidad de superficie) cuando se produce una distribución contagiosa (los individuos tienden a agruparse).
5. La distribución Binomial negativa puede definirse con mayor generalidad si tomamos r como un número real positivo cualquiera (no necesariamente entero). Pero, en dicho caso, se pierde el carácter intuitivo del modelo y se complican ligeramente los cálculos. Por dichas razones, se ha excluido dicha posibilidad en esta presentación.

3.1.8 Tabla resumen de las distribuciones discretas principales

Distribución	Parámetros	Función de densidad	Esperanza	Varianza
Bernouilli	$0 \leq p \leq 1$	$p^k(1-p)^{1-k}$ $k = 0, 1$	p	$p(1-p)$
Binomial	$0 \leq p \leq 1$ $n = 1, 2, \dots$	$\binom{n}{k} p^k (1-p)^{n-k}$ $k = 0, 1, \dots, n$	np	$np(1-p)$
Poisson	$\lambda > 0$	$e^{-\lambda} \frac{\lambda^k}{k!}$ $k = 0, 1, 2, \dots$	λ	λ
Multinomial	$0 \leq p_1, \dots, p_r \leq 1$ $(p_1 + \dots + p_r = 1)$ $n = 1, 2, \dots$	$\frac{n!}{k_1! k_2! \dots k_r!} p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$ $\sum_{i=1}^r k_i = n$	$\begin{pmatrix} np_1 \\ np_2 \\ \vdots \\ np_r \end{pmatrix}$	$\sigma_{ii} = np_i(1-p_i)$ $\sigma_{ij} = np_i p_j \quad i \neq j$
Uniforme discreta	$n = 1, 2, \dots$	$\frac{1}{n}$ $k = 1, 2, \dots, n$	$\frac{n+1}{2}$	$\frac{(n+1)[2(2n+1)-3(n+1)]}{12}$
Hipergeométrica	$\begin{cases} N = N_1 + N_2 \\ p = N_1/N \end{cases}$	$\frac{\binom{N_1}{k} \binom{N_2}{n-k}}{\binom{N}{n}}$ $\max\{0, n - N_2\} \leq k \leq \min\{N_1, n\}$	np	$np(1-p) \frac{N-n}{N-1}$
Pascal	$0 \leq p \leq 1$	$p(1-p)^k$ $k = 0, 1, 2, \dots$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Binomial negativa	$0 \leq p \leq 1$ $r > 0$		$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$

3.2 Distribuciones Continuas

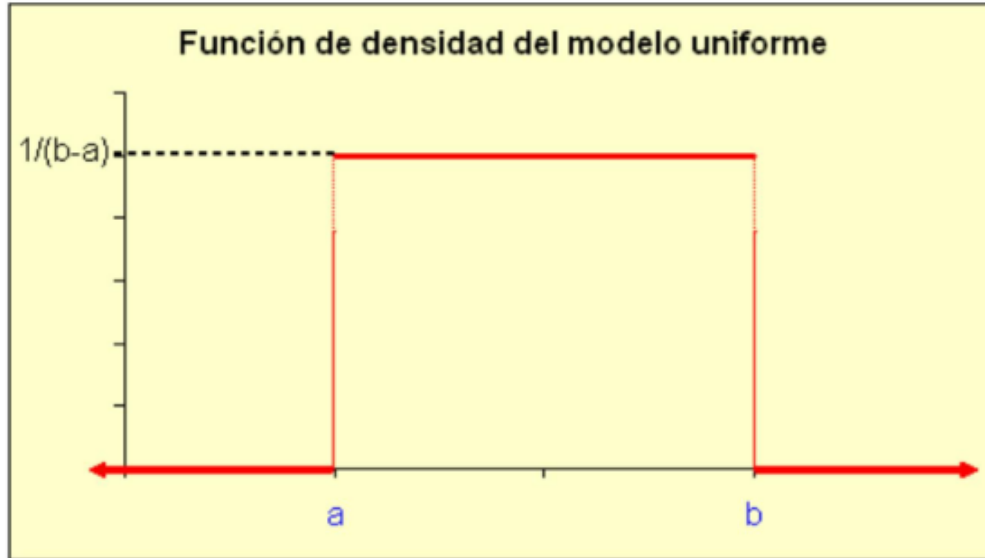
3.2.1 La distribución Uniforme

La distribución Uniforme es el modelo (absolutamente) continuo más simple. Corresponde al caso de una variable aleatoria que sólo puede tomar valores comprendidos entre dos extremos a y b , de manera que todos los intervalos de una misma longitud (dentro de (a, b)) tienen la misma probabilidad. También puede expresarse como el modelo probabilístico correspondiente a tomar un número al azar dentro de un intervalo (a, b) .

De la anterior definición se desprende que la función de densidad debe tomar el mismo valor para todos los puntos dentro del intervalo (a, b) (y cero fuera del intervalo). Es decir,

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in (a, b) \\ 0 & \text{si } x \notin (a, b) \end{cases}$$

Gráficamente:

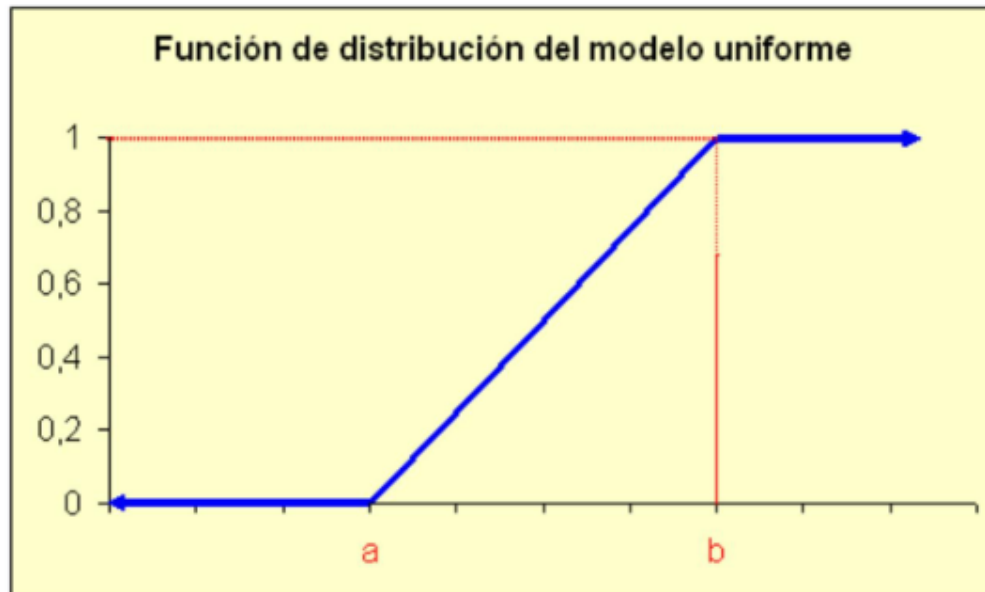


La función de distribución se obtiene integrando la función de densidad y viene dada por:

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{si } x \leq a \\ \frac{x-a}{b-a} & \text{si } x \in (a, b) \\ 1 & \text{si } x \geq b \end{cases}$$

Gráficamente:

Función de distribución del modelo uniforme



3.2.1.1 Propiedades del modelo Uniforme

1. Su esperanza vale $(b + a)/2$
2. Su varianza es $(b - a)^2/12$

3.2.1.2 Una aplicación del modelo Uniforme: el muestreo de Montecarlo En ciertos casos es útil simular el muestreo de una variable aleatoria con una distribución dada. El muestreo de Montecarlo es

un procedimiento general para obtener muestras aleatorias de cualquier tipo de variable (discreta o continua) si su función de distribución es conocida o se puede calcular.

Supongamos que queremos generar una muestra procedente de una variable aleatoria X con función de distribución $F(x)$. El proceso comprende los siguientes pasos:

1. Obtener un valor aleatorio y entre cero y uno. Es decir, obtener una muestra de una distribución Uniforme entre cero y uno. La mayoría de lenguajes de programación incorporan un generador de este tipo.
2. Considerar el valor obtenido como el valor de la función de distribución a generar: $y = F(x)$.
3. El valor $x = F^{-1}(y)$ (la inversa de la función de distribución en el punto y) es un valor procedente de la distribución de la que deseábamos generar la muestra.
4. Si queremos obtener una muestra con n individuos debemos repetir los pasos anteriores n veces.

3.2.1.3 Generación de una muestra procedente de una distribución Binomial Supongamos que queremos simular el experimento de contar el número de caras obtenidas en 5 lanzamientos de una moneda trucada con probabilidad de cara igual a 0,75 . Es decir, queremos obtener una muestra de una distribución Binomial con $n = 5$ y $p = 0,75$.

Siguiendo los pasos anteriores deberemos obtener un número al azar entre 0 y 1 (un valor procedente de una distribución Uniforme entre 0 y 1) y si este valor es menor o igual a 0,75 diremos que ha salido cara y, si es superior a 0,75 , cruz. Utiliza el siguiente programa para simular cinco lanzamientos con nuestra moneda trucada:

3.2.2 La distribución Exponencial

Este modelo suele utilizarse para variables que describen el tiempo hasta que se produce un determinado suceso.

Su función de densidad es de la forma:

$$f(x) = \begin{cases} \frac{1}{\alpha} \exp\left(-\frac{x}{\alpha}\right) & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$$

Como vemos este modelo depende de un único parámetro α que debe ser positivo: $\alpha > 0$. A continuación se muestra un programa que nos permite ver cómo cambia la forma de la función de densidad según el parámetro α .

La función de distribución se obtiene integrando la de densidad y es de la forma:

$$F(x) = \begin{cases} 1 - \exp\left(-\frac{x}{\alpha}\right) & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$$

Podemos utilizar el programa siguiente para calcular dicha función de distribución:

3.2.2.1 Propiedades del modelo Exponencial

1. Su esperanza es α .
2. Su varianza es α^2 .
3. Una propiedad importante es la denominada *carencia de memoria*, que podemos definir así: si la variable X mide el tiempo de vida y sigue una distribución Exponencial, significará que la probabilidad de que siga con vida dentro de 20 años es la misma para un individuo que a fecha de hoy tiene 25 años que para otro que tenga 60 años.

4. Cuando el número de sucesos por unidad de tiempo sigue una distribución de Poisson de parámetro λ (proceso de Poisson), el tiempo entre dos sucesos consecutivos sigue una distribución Exponencial de parámetro $\alpha = 1/\lambda$.

3.2.3 La distribución Normal

Se trata, sin duda, del modelo continuo más importante en estadística, tanto por su aplicación directa, veremos que muchas variables de interés general pueden describirse por dicho modelo, como por sus propiedades, que han permitido el desarrollo de numerosas técnicas de inferencia estadística. En realidad, el nombre de Normal proviene del hecho de que durante un tiempo se creyó, por parte de médicos y biólogos, que todas las variables naturales de interés seguían este modelo.

Su función de densidad viene dada por la fórmula:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \quad \text{donde } -\infty < x < +\infty$$

que, como vemos, depende de dos parámetros μ (que puede ser cualquier valor real) y σ (que ha de ser positiva). Por esta razón, a partir de ahora indicaremos de forma abreviada que una variable X sigue el modelo Normal así: $X \sim N(\mu, \sigma)$. Por ejemplo, si nos referimos a una distribución Normal con $\mu = 0$ y $\sigma = 1$ lo abreviaremos $N(0, 1)$.

A continuación vemos gráfica de esta función de densidad (podeis probar a cambiar los parámetros):

Como puedes ver, la función de densidad del modelo Normal tiene forma de campana, la que habitualmente se denomina campana de Gauss. De hecho, a este modelo, también se le conoce con el nombre de distribución gaussiana.

3.2.3.1 Propiedades del modelo Normal

1. Su esperanza es μ .
2. Su varianza es σ^2 y, por tanto, su desviación típica es σ .
3. Es simétrica respecto a su media μ , como puede apreciarse en la representación anterior.
4. Media, moda y mediana coinciden (μ).
5. Cualquier transformación lineal de una variable con distribución Normal seguirá también el modelo Normal. Si $X \sim N(\mu, \sigma)$ y definimos $Y = aX + b$ (con $a \neq 0$), entonces $Y \sim N(a\mu + b, |a|\sigma)$. Es decir, la esperanza de Y será $a\mu + b$ y su desviación típica, $|a|\sigma$.
6. Cualquier combinación lineal de variables normales independientes sigue también una distribución Normal. Es decir, dadas n variables aleatorias independientes con distribución $X_i \sim N(\mu_i, \sigma_i)$ para $i = 1, 2, \dots, n$ la combinación lineal: $Y = a_n X_n + a_{n-1} X_{n-1} + \dots + a_1 X_1 + a_0$ sigue también el modelo Normal:

$$Y \approx N \left(a_0 + \sum_{i=1}^n a_i \mu_i, \sqrt{\sum_{i=1}^n a_i^2 \sigma_i^2} \right)$$

La función de distribución del modelo Normal

La función de distribución del modelo Normal se debería calcular, como en el resto de distribuciones continuas, integrando la función de densidad:

$$F(x) = P[X \leq x] = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(t-\mu)^2}{2\sigma^2} \right\} dt$$

Pero nos encontramos con el problema de que no existe ninguna primitiva conocida para esta función, es decir, no sabemos resolver la anterior integral. Sin embargo, si somos incapaces de calcular la función distribución no podremos efectuar ningún cálculo con este modelo. ¿Cómo solucionamos el problema?

Una primera solución podría consistir en aproximar la integral a través de técnicas de cálculo numérico. Sin embargo, dado que el conjunto de valores que pueden tomar los parámetros μ y σ son infinitos, deberíamos repetir el proceso para cada valor diferente de algún parámetro. Afortunadamente, podemos ahorrarnos el esfuerzo aprovechando la propiedad de que cualquier transformación lineal de una variable Normal sigue también el modelo Normal. Por tanto, replantearemos cualquier problema en términos de una Normal concreta, que suele ser la $N(0, 1)$, de la siguiente manera:

Si $X \sim N(\mu, \sigma)$ y entonces definimos $Z = (X - \mu)/\sigma$ se cumplirá que $Z \sim N(0, 1)$

y, por tanto:

$$F_X(x) = P[X \leq x] = P\left[\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right] = P\left[Z \leq \frac{x - \mu}{\sigma}\right] = F_Z\left(\frac{x - \mu}{\sigma}\right)$$

A la distribución $N(0, 1)$, es decir, la que tiene por media cero y por desviación típica uno, se le denomina Normal reducida o tipificada. En cambio, al proceso de transformación del cálculo de la función de distribución de una Normal cualquiera a través de la Normal tipificada, se le denomina tipificación.

Debemos remarcar que el proceso de tipificación no resuelve el problema de la inexistencia de la función primitiva correspondiente. Sin embargo, sí es posible, mediante técnicas de cálculo numérico, obtener la integral numérica correspondiente y elaborar unas tablas que podemos consultar. Naturalmente, la tipificación permite que con una sola tabla, la de la $N(0, 1)$, tengamos suficiente.

Hoy en día, cada vez se utilizan menos tablas como la mencionada anteriormente, ya que los ordenadores, junto con los abundantes programas estadísticos existentes nos resuelven este problema. Sin embargo, la imposibilidad de integrar analíticamente la función de densidad persiste y, aunque nosotros no seamos conscientes, los programas informáticos realizan el proceso de tipificación para simplificar el problema.

3.2.4 La distribución Gamma

Este modelo es una generalización del modelo Exponencial ya que, en ocasiones, se utiliza para modelar variables que describen el tiempo hasta que se produce p veces un determinado suceso.

Su función de densidad es de la forma:

$$f(x) = \begin{cases} \frac{1}{\alpha^p \Gamma(p)} e^{-\frac{x}{\alpha}} x^{p-1} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$$

Como vemos, este modelo depende de dos parámetros positivos: α y p . La función $\Gamma(p)$ es la denominada función Gamma de Euler que representa la siguiente integral:

$$\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx$$

que verifica $\Gamma(p+1) = p\Gamma(p)$, con lo que, si p es un número entero positivo, $\Gamma(p+1) = p!$.

3.2.4.1 Propiedades de la distribución Gamma

1. Su esperanza es $p\alpha$.
2. Su varianza es $p\alpha^2$
3. La distribución Gamma ($\alpha, p = 1$) es una distribución Exponencial de parámetro α . Es decir, el modelo Exponencial es un caso particular de la Gamma con $p = 1$.

4. Dadas dos variables aleatorias con distribución Gamma y parámetro α común

$$X \sim G(\alpha, p_1) \text{ y } Y \sim G(\alpha, p_2)$$

se cumplirá que la suma también sigue una distribución Gamma

$$X + Y \sim G(\alpha, p_1 + p_2)$$

Una consecuencia inmediata de esta propiedad es que, si tenemos k variables aleatorias con distribución Exponencial de parámetro α (común) e independientes, la suma de todas ellas seguirá una distribución $G(\alpha, k)$.

3.2.5 La distribución de Cauchy

Se trata de un modelo continuo cuya función de densidad es:

$$f(x) = \frac{1}{\pi(1+x^2)} \quad \text{para} \quad -\infty < x < \infty$$

Cuya integral nos proporciona la función de distribución:

$$F(x) = \int_{-\infty}^x \frac{1}{\pi(1+t^2)} dt = \frac{1}{\pi} [\arctan(t)]_{t=-\infty}^{t=x} = \frac{1}{2} + \frac{\arctan(x)}{\pi}$$

El siguiente programa permite visualizar la forma de la función de densidad de este modelo y el valor de la función de distribución:

3.2.5.1 Propiedades de la distribución de Cauchy Se trata de un ejemplo de variable aleatoria que carece de esperanza (y, por tanto, también de varianza o cualquier otro momento), ya que la integral impropia correspondiente no es convergente:

$$E(X) = \int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)} dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{2x}{1+x^2} dx = \frac{1}{2\pi} \left[\lim_{x \rightarrow \infty} \ln(x^2) - \lim_{x \rightarrow -\infty} \ln(x^2) \right] = \frac{1}{2\pi} [\infty - \infty]$$

y nos queda una indeterminación. Por tanto, la esperanza de una distribución de Cauchy no existe. Cabe señalar que la función de densidad es simétrica respecto al valor cero (que sería la mediana y la moda), pero al no existir la integral anterior, la esperanza no existe.

3.2.6 La distribución de Weibull

Se trata de un modelo continuo asociado a variables del tipo tiempo de vida, tiempo hasta que un mecanismo falla, etc. La función de densidad de este modelo viene dada por:

$$f(x) = \begin{cases} \frac{\beta}{\alpha} \left(\frac{x}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x}{\alpha}\right)^{\beta}} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

que, como vemos, depende de dos parámetros: $\alpha > 0$ y $\beta > 0$, donde α es un parámetro de escala y β es un parámetro de forma (lo que proporciona una gran flexibilidad a este modelo).

La función de distribución se obtiene por la integración de la función de densidad y vale:

$$F(x) = 1 - e^{-\left(\frac{x}{\alpha}\right)^\beta}$$

El siguiente programa permite visualizar la forma de la función de densidad de este modelo y el valor de la función de distribución:

3.2.6.1 Propiedades de la distribución Weibull

1. Si tomamos $\beta = 1$ tenemos una distribución Exponencial.
2. Su esperanza vale:

$$E(X) = \alpha \Gamma\left(\frac{1}{\beta} + 1\right)$$

3. Su varianza vale:

$$V(X) = \alpha^2 \left\{ \Gamma\left(\frac{2}{\beta} + 1\right) - \left[\Gamma\left(\frac{1}{\beta} + 1\right) \right]^2 \right\}$$

donde $\Gamma(x)$ representa la función Gamma de Euler definida anteriormente.

3.2.7 Tabla resumen de las principales distribuciones continuas

Distribución	Parámetros	Función de densidad	Esperanza	Varianza
Uniforme	a, b	$\frac{1}{b-a} \quad a < x < b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponencial	$\alpha > 0$	$\frac{1}{\alpha} \exp\left(-\frac{x}{\alpha}\right) \quad x > 0$	α	α^2
Normal	$-\infty < \mu < \infty$ $\sigma > 0$	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$ $-\infty < x < +\infty$	μ	σ^2

Cauchy $| - | \frac{1}{\pi(1+x^2)} \quad -\infty < x < \infty \quad | - | - |$

Weibull $| \alpha > 0 \quad \beta > 0 \quad | \frac{\beta}{\alpha} \left(\frac{x}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x}{\alpha}\right)^\beta} \quad x \geq 0 \quad | \alpha \Gamma\left(\frac{1}{\beta} + 1\right) \quad | \alpha^2 \left\{ \Gamma\left(\frac{2}{\beta} + 1\right) - \left[\Gamma\left(\frac{1}{\beta} + 1\right) \right]^2 \right\} \quad |$

3.3 Distribuciones con R (y Python)

El lenguaje estadístico R es muy potente en cuanto al cálculo con distribuciones de probabilidad.

Dado que el trabajo con distribuciones de probabilidad usando R está muy estandarizado y explicado en múltiples fuentes no repetiremos aquí estas explicaciones.

Tan solo os referimos a dos buenas fuentes de información que podéis utilizar para aprender como hacer los cálculos con R y también una aplicación que os permite visualizar casi cualquier distribución conocida.

R Tutorials

Explicación detallada y de nivel básico del manejo de las principales distribuciones con R

<https://www.r-tutor.com/elementary-statistics/probability-distributions>

The distribution Zoo

Permite visualizar de forma interactiva distintas distribuciones y proporciona información diversa sobre sus propiedades e incluso su aplicación.

<https://ben18785.shinyapps.io/distribution-zoo/>

Distribution explorer

Más completo que los anteriores. No se basa en R sino en python.

<https://distribution-explorer.github.io/index.html>

3.4 La familia exponencial de distribuciones

En el estudio de las propiedades de los estimadores, vemos que algunas distribuciones se comportan mejor que otras. Muchas veces, este buen comportamiento refleja una estructura común que proviene de pertenecer a una misma familia de distribuciones llamada familia exponencial.

Definición: Sea f_θ una familia de probabilidades que depende de un parámetro unidimensional $\{f_\theta(x), \theta \in \Theta \subseteq \mathbb{R}\}$ tal que el soporte $S(\theta) = \{x \mid f_\theta(x) > 0\}$ no depende de θ . Si existen funciones de los parámetros $Q(\theta)$ y $C(\theta)$ y funciones de las muestras, $T(x)$ y $h(x)$, tales que la función de densidad puede escribirse como:

$$f_\theta(x) = C(\theta)h(x) \exp\{Q(\theta) \cdot T(x)\}$$

diremos que $f_\theta(x)$ pertenece a la familia exponencial de distribuciones.

La familia exponencial no representa un nuevo tipo de distribuciones, sino la constatación de que muchas distribuciones comunes, que pueden reformularse para ajustarse a la expresión anterior, pertenecen a esta familia.

Veamos algunos ejemplos de que esto es efectivamente así.

3.4.1 Ejemplos de distribuciones de esta familia

3.4.1.1 Distribución de Poisson La ley de Poisson pertenece a la familia exponencial uniparamétrica. Efectivamente,

$$f_\lambda(x) = e^{-\lambda} \frac{\lambda^x}{x!} = \exp\{-\lambda + x \log \lambda - \log(x!)\}$$

y si hacemos

$$Q(\lambda) = \log(\lambda) \quad T(x) = x \quad D(\lambda) = -\lambda \quad S(x) = -\log(x!)$$

se hace evidente que f_λ pertenece a la familia exponencial.

3.4.1.2 Distribución normal uniparamétrica La ley normal depende de dos parámetros μ y σ . Fijado uno de ellos, nos queda una distribución que depende de un solo parámetro, y de aquí la denominación “normal uniparamétrica”.

Si, con el subíndice “0”, indicamos el parámetro fijado, tenemos:

$$\begin{aligned} f_\sigma &= \{N(\mu_0, \sigma), \sigma > 0\} \text{ Normal uniparamétrica, de parámetro } \sigma^2, \\ f_\mu &= \{N(\mu, \sigma_0), \mu \in \mathbb{R}\} \text{ normal uniparamétrica, de parámetro } \mu. \end{aligned}$$

Si queremos considerar ambos parámetros a la vez, debemos extender la definición al caso de parámetros k -dimensionales. En estos materiales no trataremos esta extensión.

3.4.1.2.1 Caso 1: Fijando la media μ_0 Consideramos la distribución normal $N(\mu_0, \sigma^2)$, donde fijamos $\mu = \mu_0$ y σ^2 es el parámetro libre.

La función de densidad de probabilidad es

$$f_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu_0)^2}{2\sigma^2} \right\}$$

Vamos a reescribir esta función en forma de la familia exponencial. Primero, reorganizamos los términos de la densidad:

$$f_\sigma(x) = \frac{1}{\sqrt{2\pi}} \cdot \sigma^{-1} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu_0)^2 \right\}$$

Ahora identificamos las funciones que se corresponden con la forma de la familia exponencial $f_\theta(x) = C(\theta)h(x)\exp\{Q(\theta)T(x)\}$:

- $Q(\sigma) = -\frac{1}{2\sigma^2}$
- $T(x) = (x - \mu_0)^2$
- $C(\sigma) = \frac{1}{\sqrt{2\pi}\sigma}$
- $h(x) = 1$

Esto confirma que la distribución normal, con μ_0 fijo, pertenece a la familia exponencial.

3.4.1.2.2 Caso 2: Fijando la varianza σ_0^2 Ahora consideramos la distribución $N(\mu, \sigma_0^2)$, donde la varianza está fijada y el parámetro libre es μ .

La función de densidad es

$$f_\mu(x) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma_0^2} \right\}$$

Vamos a reescribir esta función de la misma manera:

$$f_\mu(x) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp \left\{ -\frac{1}{2\sigma_0^2} (x^2 - 2\mu x + \mu^2) \right\}$$

Identificamos las funciones correspondientes:

- $Q(\mu) = \frac{\mu}{\sigma_0^2}$
- $T(x) = x$
- $D(\mu) = -\frac{\mu^2}{2\sigma_0^2}$
- $S(x) = -\frac{x^2}{2\sigma_0^2}$

Esto prueba que la distribución normal con σ_0 fijo pertenece a la familia exponencial.

3.4.2 Distribución Binomial

La distribución binomial es un ejemplo interesante, puesto que, a priori, no parece tener la estructura propia de la distribución exponencial, cosa que si pasa con la distribución de Poisson o con la Normales uniparamétricas que acabamos de ver.

Sin embargo, tras aplicar algunas transformaciones se puede ver como, también esta distribución pertenece a la familia exponencial

La función de masa de probabilidad para la distribución binomial es

$$f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n$$

Reescribimos esta función en términos exponenciales:

$$f(x; n, p) = \binom{n}{x} \exp\{x \log(p) + (n-x) \log(1-p)\}$$

Agrupamos los términos dependientes de x :

$$f(x; n, p) = \binom{n}{x} \exp \left\{ x \log \left(\frac{p}{1-p} \right) + n \log(1-p) \right\}$$

Identificamos las funciones correspondientes a la familia exponencial:

- $Q(p) = \log \left(\frac{p}{1-p} \right)$
- $T(x) = x$
- $D(p) = n \log(1-p)$
- $S(x) = \log \binom{n}{x}$

Por lo tanto, la distribución binomial pertenece a la familia exponencial.

3.4.3 Importancia y utilidad de la familia exponencial

Muchas de las distribuciones usadas para modelar gran cantidad de situaciones prácticas pertenecen a esta familia.

Esto significa que es posible estudiar sus propiedades en conjunto. Es decir, si establecemos que una propiedad se verifica en una distribución que pertenece a la familia exponencial, automáticamente sabemos que todos los miembros de la familia verifican esa propiedad.

A continuación, se describen tres ventajas importantes de trabajar con esta familia:

3.4.4 Los modelos lineales generalizados (GLMs)

Una de las aplicaciones más importantes de la familia exponencial es su uso en los **Modelos Lineales Generalizados (GLMs)**.

Estos modelos nos permiten extender la regresión lineal clásica a diferentes tipos de datos, como los resultados binarios (por ejemplo, éxito o fracaso), mediante la *regresión logística*, recuentos de eventos (como el número de llamadas recibidas en una hora) mediante la *regresión de Poisson*, y muchos otros.

Gracias a la estructura de la familia exponencial, podemos conectar la media de la variable que estamos modelando con las variables explicativas de forma flexible, lo que hace posible aplicar GLMs en una amplia variedad de situaciones.

3.4.5 Estimación en la familia exponencial

Otra ventaja importante es que, al trabajar con distribuciones de la familia exponencial, los métodos que usamos para hacer inferencias estadísticas suelen tener **buenas propiedades**.

Esto, que se explicará con más detalle en capítulos siguientes, implica que los estimadores que obtenemos con estos modelos suelen ser precisos y reflejar correctamente la información que contienen los datos.

Naturalmente esto se puede ver al revés: Si podemos trabajar con distribuciones de la familia exponencial, solemos tener, de entrada, una serie de ventajas, como el buen comportamiento de los estimadores, por lo que siempre es una buena opción intentar utilizarlas en nuestros modelos.

4 Distribuciones de probabilidad multidimensionales

En este capítulo se extiende el concepto de variable aleatoria a un conjunto de variables que pueden interpretarse asociadas a un conjunto de medidas distintas y que pueden estar, o no relacionadas.

Tras introducir los conceptos de distribuciones multidimensionales, condicionales y marginales, se pasa a considerar el caso más habitual en inferencia estadística en el que las componentes de los vectores son independientes entre ellas.

Este es, de hecho, el punto de partida de muchos modelos y métodos en estadística.

4.1 Distribuciones conjuntas de probabilidades

- A menudo nos interesa estudiar múltiples características de un fenómeno aleatorio:
 - La altura, el peso y el sexo de un individuo.
 - La expresión coordinada de los genes que participan en una determinada vía metabólica.
 - El número de nucleótidos A, C, G, T en una región del genoma de tamaño n .
- Estas características numéricas que, de forma análoga al caso univariante, podemos suponer asociadas a los resultados de experimentos aleatorios se denominan *variables aleatorias multidimensionales* o, atendiendo a sus componentes, **vectores aleatorios**.

Las distribuciones de probabilidad que, siguiendo con la analogía, asociaremos a los vectores aleatorios se denominan **distribuciones de probabilidades conjuntas** o **multivariantes**.

Antes de desarrollar el tema es importante remarcar que consideraremos dos escenarios:

- El primero, el “natural” es considerar que si trabajamos con distintas variables asociadas a un mismo fenómeno, es razonable suponer que varíen de alguna forma coordinada. De ahí la expresión *distribución conjunta*.
- En ocasiones, sin embargo, dispondremos de vectores aleatorios que varían independientemente los unos de los otros. En este caso su distribución conjunta será de un tipo especial que se conoce *independencia*.

4.1.1 Variable aleatoria bivalente

Empezaremos por el caso más sencillo que, sin embargo permite estudiar la mayoría de los conceptos que nos interesan: Las distribuciones conjuntas de dos variables aleatorias.

Una **variable aleatoria bivalente** es una aplicación que, a cada resultado de un experimento, le asocia dos números:

$$(X, Y) : \Omega \rightarrow \mathbb{R}^2$$

$$w \mapsto (X(w), Y(w))$$

De modo que, para todo par de valores numéricos, $(x, y) \in \mathbb{R}^2$, se tiene

$$\{w \in \Omega \mid X(w) \leq x, \quad Y(w) \leq y\} \in \mathcal{A}$$

donde \mathcal{A} representa el conjunto de *sucesos observables* definido en el capítulo 1.

Lo que viene a significar esta definición es que una variable aleatoria bidimensional es un conjunto de medidas (números reales) a los que, por el hecho de poderse asociar con sucesos observables a través de los intervalos $X(w) \leq x, \quad Y(w) \leq y$ se les puede asociar (calcular) una probabilidad.

Fijémonos también que, como en el caso univariante, la función que *transporta* la probabilidad, del espacio de probabilidad al conjunto de los reales, será la función de distribución, que se define a continuación.

4.1.2 Función de distribución bivalente

La función de distribución conjunta de X y Y , F , es una generalización inmediata del caso univariado y se define como:

$$F(x, y) = P\{w \in \Omega \mid X(w) \leq x, Y(w) \leq y\} = P[X \leq x, Y \leq y]$$

Como en el caso univariante, esta es la función que define la forma en que podemos calcular probabilidades sobre los valores de las variables, en este caso de dimensión 2.

4.1.3 Ejemplo: Distribución conjunta del estado de infección y activación de células

Supongamos que estamos observando dos características de células en un experimento de inmunología. Las variables que describen las células son:

- X : La célula está infectada ($X = 1$) o no infectada ($X = 0$).
- Y : La célula está activada ($Y = 1$) o no activada ($Y = 0$).

La siguiente tabla muestra la probabilidad conjunta de observar cada combinación de infección y activación en una célula:

$X \backslash Y$	$Y = 0$ (No activada)	$Y = 1$ (Activada)
$X = 0$ (No infectada)	0.4	0.2
$X = 1$ (Infectada)	0.1	0.3

4.1.3.1 1. Función de distribución conjunta La función de distribución conjunta $F(x, y)$ para esta situación se calcula como:

$$F(x, y) = P(X \leq x, Y \leq y)$$

Los valores para los pares posibles de x y y son:

- $F(0, 0) = P(X = 0, Y = 0) = 0.4$
- $F(0, 1) = P(X = 0, Y \leq 1) = P(X = 0, Y = 0) + P(X = 0, Y = 1) = 0.4 + 0.2 = 0.6$
- $F(1, 0) = P(X \leq 1, Y = 0) = P(X = 0, Y = 0) + P(X = 1, Y = 0) = 0.4 + 0.1 = 0.5$
- $F(1, 1) = P(X \leq 1, Y \leq 1) = 1$

4.1.3.2 2. Cálculo de la probabilidad de eventos específicos Por ejemplo, la probabilidad de que una célula esté infectada pero no activada es:

$$P(X = 1, Y = 0) = 0.1$$

4.1.4 Implementación en R

Podemos visualizar esta distribución conjunta con un gráfico en R.

```
library(ggplot2)

# Crear los datos de la distribución conjunta
data <- expand.grid(X = c(0, 1), Y = c(0, 1))
data$Prob <- c(0.4, 0.2, 0.1, 0.3)

# Crear el gráfico
p <- ggplot(data, aes(x = factor(X, labels = c("No infectada", "Infectada")),
```

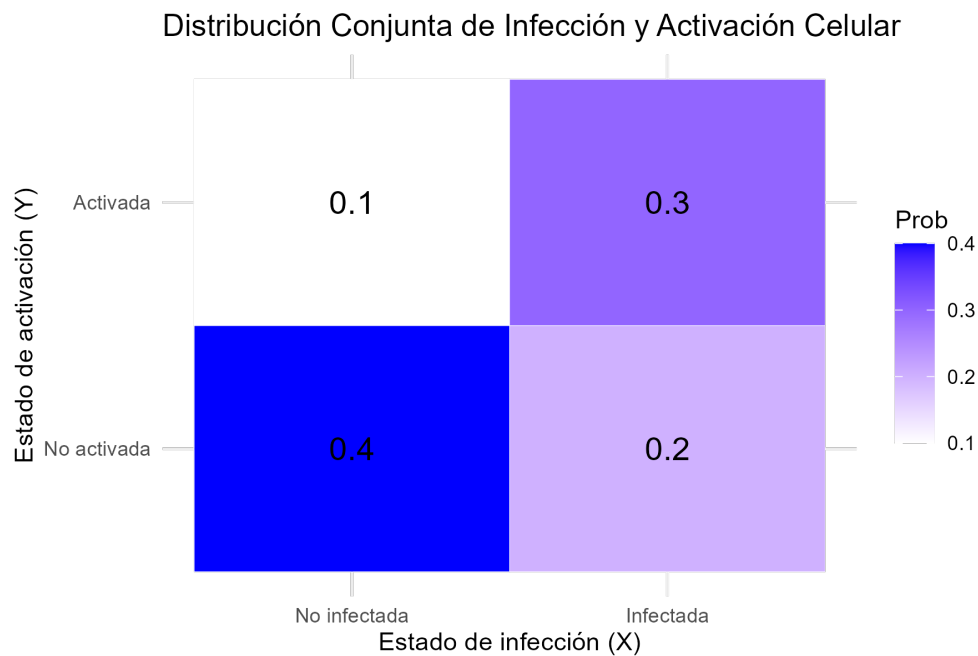
```

    y = factor(Y, labels = c("No activada", "Activada")))) +
  geom_tile(aes(fill = Prob), color = "white") +
  scale_fill_gradient(low = "white", high = "blue") +
  geom_text(aes(label = round(Prob, 2)), size = 5) +
  labs(x = "Estado de infección (X)", y = "Estado de activación (Y)",
       title = "Distribución Conjunta de Infección y Activación Celular") +
  theme_minimal()

# Guardar el gráfico en el subdirectorío imagenes
ggsave("imagenes/distribucion_conjunta.png", plot = p, width = 6, height = 4, dpi = 300)

knitr::include_graphics("imagenes/distribucion_conjunta.png")

```



4.2 Variable aleatorias bivariantes discretas

Una vez introducidos los conceptos de forma general pasamos a estudiar el problema en el caso discreto, que es muy intuitivo y, a la vez permite introducir todos los conceptos relevantes.

Un **vector aleatorio discreto**, (X, Y) es aquel cuyo recorrido o conjunto de valores posibles es finito o numerable.

En este caso, toda probabilidad

$$P\{(X, Y) \in B\}, \quad \text{donde } B \text{ es un conjunto de posibles valores de } X, Y,$$

se puede calcular a partir de la **función de masa de probabilidad discreta bivalente**.

4.2.1 Función de masa de probabilidad discreta (fmp)

La función de masa de probabilidad de los vectores aleatorios generaliza la función del mismo nombre en el caso univariante, es decir, es una función:

$$f : \mathbb{R}^2 \rightarrow [0, 1]$$

Que asigna la probabilidad a cada punto del plano: para todo $(x, y) \in \mathbb{R}^2$:

$$f(x, y) = P\{w \in \Omega \mid X(w) = x, Y(w) = y\} = P[X = x, Y = y]$$

4.2.2 Propiedades de la fmp bivalente

- La masa total de probabilidad sobre el plano es 1:

$$\sum_{(x_i, y_j) \in \mathbb{R}^2} f(x_i, y_j) = 1$$

- Para todo subconjunto $B \subseteq \mathbb{R}^2$, se verifica:

$$F(x, y) = P[X \leq x, Y \leq y] = \sum_{x_i \leq x, y_j \leq y} f(x_i, y_j)$$

Es decir, como en el caso univariante *la función de distribución se puede calcular a partir de la función de masa de probabilidad*.

4.2.2.1 Intuición frente a construcción La presentación de los conceptos anteriores suele generar cierto desasosiego entre los estudiantes que afrontan estos conceptos por primera (o siguientes) vez.

El motivo de este desasosiego es que el papel de la función de distribución no suele ser tan intuitivo como el de la función de masa de probabilidad.

Es decir, es más intuitivo pensar en como calcular la probabilidad que la variable tome un valor concreto ($P[X = x]$), que la probabilidad de que no alcance cierto valor ($P[X \leq x]$).

Sin embargo, la función que realmente permite transportar la probabilidad no es la función de masa de probabilidad (fmp) sino la función de distribución (fdd). De ahí el contraste entre intuición (fmp) y construcción (fdd)

4.2.3 Ejemplo de distribución bivalente discreta

Supongamos que un estudio mide el número de células infectadas y el número de linfocitos activados en un campo microscópico. Dado el tamaño del campo y el grado de infección los valores observados de cada variables son:

- X : Número de células infectadas ($X \in \{0, 1, 2, 3, 4, 5\}$).
- Y : Número de linfocitos activados ($Y \in \{0, 1, 2, 3\}$).

La distribución conjunta se refleja en la siguiente tabla de probabilidades conjuntas:

$P[X = x]$	$P[Y = 0]$	$P[Y = 1]$	$P[Y = 2]$	$P[Y = 3]$
0	0.12	0.06	0.02	0.00
1	0.10	0.10	0.04	0.01
2	0.06	0.12	0.08	0.02
3	0.03	0.12	0.10	0.05
4	0.01	0.08	0.12	0.06
5	0.00	0.03	0.10	0.07

Puede comprobarse como la suma de todos los valores de la tabla es 1, y calcular probabilidades de sucesos como

Probabilidad de que hayan dos células infectadas y un linfocito:

Para calcular la probabilidad de que haya exactamente 2 células infectadas y 1 linfocito activado, se puede usar el valor directamente de la tabla.

$$P(X = 2, Y = 1) = 0.12$$

Probabilidad de que hayan menos de tres células infectadas y menos de dos linfocitos:

Esta probabilidad es la suma de todas las combinaciones de X y Y que cumplen con la condición de $X < 3$ y $Y < 2$. Es decir, sumamos las probabilidades de los casos

$(X = 0, Y = 0)$, $(X = 0, Y = 1)$, $(X = 1, Y = 0)$, $(X = 1, Y = 1)$, $(X = 2, Y = 0)$, y $(X = 2, Y = 1)$.

$$P(X < 3, Y < 2) = P(X = 0, Y = 0) + P(X = 0, Y = 1) + P(X = 1, Y = 0) + P(X = 1, Y = 1) + P(X = 2, Y = 0) + P(X = 2, Y = 1)$$

$$P(X < 3, Y < 2) = 0.12 + 0.06 + 0.10 + 0.10 + 0.06 + 0.12 = 0.56$$

Recordemos que, al tratarse de variables discretas, no es lo mismo $P[X < x]$ que $P[X \leq x]$, por lo que si la pregunta fuera “Probabilidad de que hayan al menos tres células infectadas y al menos dos linfocitos” deberíamos calcular:

$$P(X \leq 3, Y \leq 2)$$

Esta última expresión se corresponde con la función de distribución evaluada en $(3, 2)$.

4.2.3.1 Código R para el cálculo de la pmf Podemos hacer los cálculos usando R:

```
prob_table <- matrix(c(0.12, 0.06, 0.02, 0.00,
                      0.10, 0.10, 0.04, 0.01,
                      0.06, 0.12, 0.08, 0.02,
                      0.03, 0.12, 0.10, 0.05,
                      0.01, 0.08, 0.12, 0.06,
                      0.00, 0.03, 0.10, 0.07),
                    nrow = 6, byrow = TRUE)
```

Asignar nombres a las filas y columnas

```
rownames(prob_table) <- 0:5
```

```
colnames(prob_table) <- 0:3
```

Mostrar la tabla

```
prob_table
```

```
##      0      1      2      3
## 0 0.12 0.06 0.02 0.00
## 1 0.10 0.10 0.04 0.01
## 2 0.06 0.12 0.08 0.02
## 3 0.03 0.12 0.10 0.05
## 4 0.01 0.08 0.12 0.06
## 5 0.00 0.03 0.10 0.07
```

Calcular la probabilidad de $(X = 2, Y = 1)$

```
prob_X2_Y1 <- prob_table["2", "1"]
```

```
cat("P(X = 2, Y = 1) =", prob_X2_Y1, "\n")
```



```
## P(X = 2, Y = 1) = 0.12
# Calcular la probabilidad de (X < 3, Y < 2)
prob_X_lt_3_Y_lt_2 <- sum(prob_table[1:3, 1:2])
cat("P(X < 3, Y < 2) =", prob_X_lt_3_Y_lt_2, "\n")
```

```
## P(X < 3, Y < 2) = 0.56
```

4.2.3.2 Código R para visualizar la distribución conjunta Para visualizar la distribución conjunta, podemos usar el código siguiente;

```
# Es preciso instalar y cargar el paquete scatterplot3d si no lo tienes instalado
# install.packages("scatterplot3d")
library(scatterplot3d)

# Crear una matriz con los datos de la tabla de probabilidades
X_vals <- as.numeric(rownames(prob_table))
Y_vals <- as.numeric(colnames(prob_table))

# Crear un grid de valores X e Y
X_grid <- rep(X_vals, each = length(Y_vals))
Y_grid <- rep(Y_vals, times = length(X_vals))

# Extraer las probabilidades como un vector
Z_vals <- as.vector(prob_table)

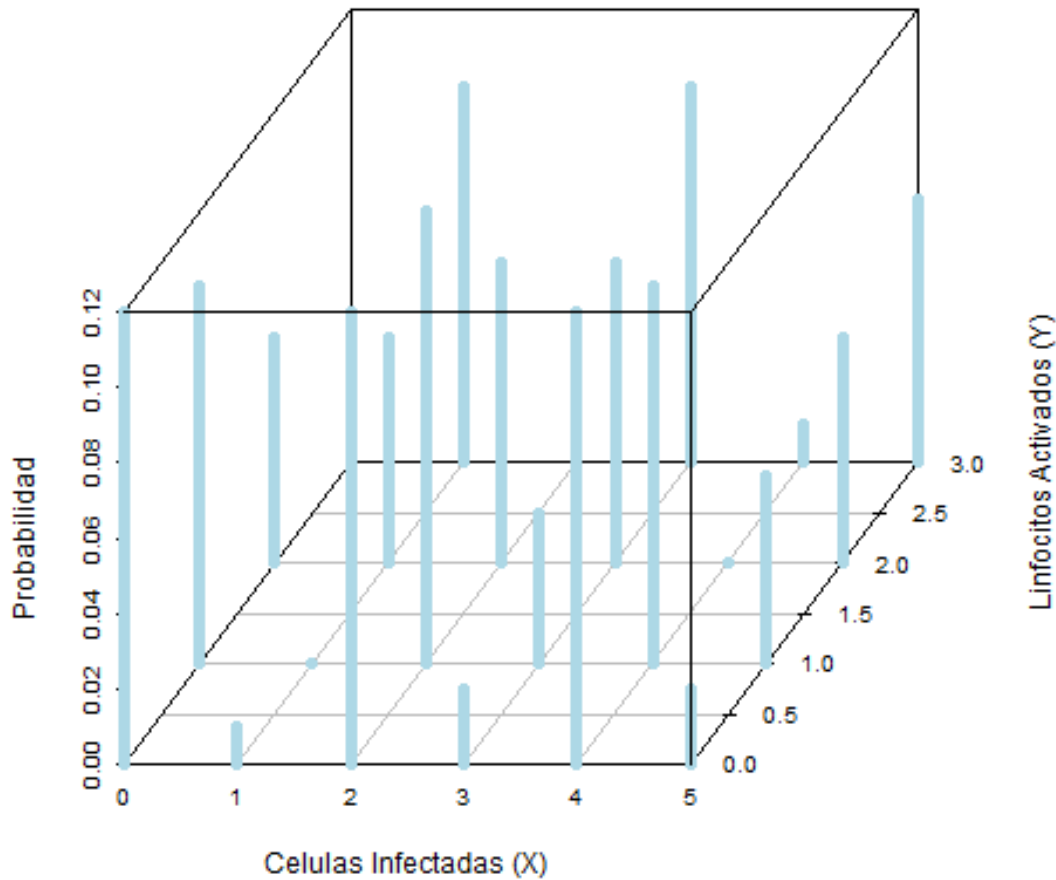
# Enviar el gráfico 3D de barras simuladas a pdf
png("images/pmfTrinomial.png")
scatterplot3d(X_grid, Y_grid, Z_vals,
              type = "h", color = "lightblue",
              pch = 16, lwd = 5,
              cex.symbols = 1,
              angle=60,
              xlab = "Celulas Infectadas (X)",
              ylab = "Linfocitos Activados (Y)",
              zlab = "Probabilidad",
              main = "Distribución Conjunta de \n Celulas Infectadas y Linfocitos Activados")
dev.off()

## pdf
## 2

# Añadir texto con los valores de las probabilidades en la parte superior de las barras
# s3d$text(X_grid, Y_grid, Z_vals, labels = round(Z_vals, 2), pos = 3, col = "black")

knitr::include_graphics("images/pmfTrinomial.png", rel_path = TRUE )
```

Distribución Conjunta de Celulas Infectadas y Linfocitos Activados



4.3 La distribución multinomial

Antes de seguir con el estudio de las distribuciones discretas presentamos un caso importante de distribución multivariante discreta, la **distribución multinomial**.

4.3.1 Generación de las observaciones

Supongamos un experimento aleatorio que puede producir k resultados posibles A_1, A_2, \dots, A_k con probabilidades p_1, p_2, \dots, p_k , tales que $p_1 + p_2 + \dots + p_k = 1$.

Repetimos el experimento n veces y llamamos X_1, X_2, \dots, X_k al número de veces que se presenta A_1, A_2, \dots, A_k .

La distribución conjunta de X_1, X_2, \dots, X_k recibe el nombre de **multinomial**.

4.3.2 Función de masa de probabilidad de la distribución multinomial

El vector $\mathbf{X} = (X_1, \dots, X_k)$ tiene distribución multinomial de parámetros n y $\mathbf{p} = (p_1, \dots, p_k)$, denotado por $\mathbf{X} \sim M(n, \mathbf{p})$, con n entero positivo, $p_i \geq 0$ y $\sum_{i=1}^k p_i = 1$.

Su función de densidad conjunta es:

$$f(\mathbf{x}) = P[\mathbf{X} = \mathbf{x}] = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

donde x_i son enteros no negativos tales que $\sum_{i=1}^k x_i = n$.

4.3.3 Relación con la distribución binomial

Esta distribución puede verse como una generalización de la distribución binomial en el que, en lugar de tener dos posibles resultados, tenemos r resultados posibles.

4.3.4 Un caso particular: La distribución trinomial

Veamos un ejemplo propio del análisis de secuencias en el que se aplica esta distribución:

Si consideramos el alineamiento de dos secuencias x, y de tamaño n , podemos observar:

- A_1 : x_i alineado con y_i , con $P(A_1) = p_1$
- A_2 : x_i alineado con “-”, con $P(A_2) = p_2$
- A_3 : “-” alineado con y_i , con $P(A_3) = 1 - p_1 - p_2$

La variable (X_1, X_2) , que cuenta el número de veces que se observa A_1, A_2 (con $X_3 = n - X_1 - X_2$), sigue una distribución trinomial de parámetros n, p_1, p_2 .

Obsérvese que, dado que el total de observaciones n está prefijado, aunque haya tres categorías, A_1, A_2, A_3 el número de observaciones de A_3 es el total menos la suma de las observaciones de $A_1 + A_2$. O dicho de otra forma el número de probabilidades que són parámetros de la distribución es $n - 1 = 2$, lo que junto con n que es otro parámetro determina que “trinomial” se refiera tanto al total de categorías como al número de parámetros, aunque, en realidad tan sólo hay dos componentes X_1 y X_2 independientes (concepto este que se definirá con precisión más adelante).

Estudiamos los posibles alineamientos de dos secuencias de 5 nucleótidos, en un contexto en el que las probabilidades de A_1 y A_2 son, respectivamente 0.6 y 0.2, es decir una Trinomial $M(5; 0.6, 0.2)$ que dan lugar a la tabla siguiente.

$X_1 \backslash X_2$	0	1	2	3	4	5
0	(0,0,5)	(0,1,4)	(0,2,3)	(0,3,2)	(0,4,1)	(0,5,0)
1	(1,0,4)	(1,1,3)	(1,2,2)	(1,3,1)	(1,4,0)	
2	(2,0,3)	(2,1,2)	(2,2,1)	(2,3,0)		
3	(3,0,2)	(3,1,1)	(3,2,0)			
4	(4,0,1)	(4,1,0)				
5	(5,0,0)					

A partir de la tabla anterior podemos determinar las probabilidades conjuntas:

$X_1 \backslash X_2$	0	1	2	3	4	5
0	0.0003	0.0016	0.0032	0.0032	0.0016	0.0003
1	0.0048	0.0192	0.0288	0.0192	0.0048	
2	0.0288	0.0864	0.0864	0.0288		
3	0.0864	0.1728	0.0864			
4	0.1296	0.1296				
5	0.0778					

4.4 Distribuciones marginales

- Dado un vector aleatorio, puede interesar el comportamiento individual de una o cada una de sus componentes X_i .
- La distribución de la componente i -ésima se denomina **distribución marginal** de X_i .
- Representa el comportamiento de X_i sin tener en cuenta las otras componentes, es decir, como si fuera una variable aleatoria unidimensional.

4.4.1 Las marginales están en los márgenes

- El nombre de **distribución marginal** proviene del hecho de que en una distribución bivariada discreta como la trinomial, los valores de una fila coinciden con los valores de X_2 , y todos los de una columna con los de X_1 . Los valores en la fila 0 o columna 0 (los márgenes) representan precisamente las distribuciones marginales.

4.4.2 Densidades marginales discretas

- La densidad marginal de X es:

$$f_X(x) = f_1(x) = \sum_j f(x, y_j)$$

y la de Y es:

$$f_Y(y) = f_2(y) = \sum_i f(x_i, y)$$

4.4.3 Trinomial M(5; 0.6, 0.2): Distribuciones marginales

$X_1 \backslash X_2$	0	1	2	3	4	5	X_2	$P[X_2 = x]$
0	(0,0,5)	(0,1,4)	(0,2,3)	(0,3,2)	(0,4,1)	(0,5,0)	0	0.0102
1	(1,0,4)	(1,1,3)	(1,2,2)	(1,3,1)	(1,4,0)		1	0.0768
2	(2,0,3)	(2,1,2)	(2,2,1)	(2,3,0)			2	0.2304
3	(3,0,2)	(3,1,1)	(3,2,0)				3	0.3456
4	(4,0,1)	(4,1,0)					4	0.2592
5	(5,0,0)						5	0.0778
X_2	0	1	2	3	4	5		1.0000
$P[X_2 = x]$	0.3277	0.4096	0.2048	0.0512	0.0064	0.0003	1.0000	

4.5 Distribuciones condicionales

- A veces nos interesa la distribución de una componente si conocemos que la otra ha tomado un valor determinado.
- En el ejemplo de los alineamientos, podríamos querer conocer los posibles valores y probabilidades de un alineamiento, si sabemos que hay exactamente un “gap” en la secuencia de prueba.

4.5.1 Densidad condicional

¿Qué podemos decir de la distribución de Y si conocemos el valor de X ?

$$f(y | X = x) = P[Y = y | X = x] = \frac{P[X = x, Y = y]}{P[X = x]} = \frac{f(x, y)}{f_X(x)}$$

siempre que $f_X(x) > 0$.

4.5.2 Trinomial M(5; 0.6, 0.2): Distribución condicional

Distribución de X_1 condicionada a que $X_2 = 1$.

$(X_1, 1)$	$P(X_1, 1)$	$P_{X_2}(1)$	$P(X_1 X_2 = 1)$
(0,1,4)	0.002	0.41	0.004
(1,1,3)	0.019	0.41	0.047
(2,1,2)	0.086	0.41	0.211
(3,1,1)	0.173	0.41	0.422
(4,1,0)	0.13	0.41	0.316
Total			1

4.6 Vectores aleatorios absolutamente continuos

- Diremos que (X, Y) es absolutamente continua si existe una función $f(x, y)$, llamada **función de densidad conjunta absolutamente continua** o **bivariada**, tal que, para todo $(x, y) \in \mathbb{R}^2$,

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv$$

- Si existe, la función de densidad absolutamente continua es única.

4.6.1 Propiedades de la función de densidad conjunta

- $f(x, y) \geq 0$
- La masa total de probabilidad es 1:

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

- Para cualquier conjunto S :

$$P\{(X, Y) \in S\} = \int_S f(x, y) dx dy$$

En particular, la probabilidad de que (X, Y) esté en un rectángulo:

$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f(x, y) dx dy$$

4.6.2 Densidades marginales en el caso continuo

- Las densidades marginales son:

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

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

4.6.3 Densidad condicional en el caso continuo

- La densidad de Y condicionada a un valor de X es:

$$f(y | X = x) = \frac{f(x, y)}{f_X(x)}$$

siempre que $f_X(x) > 0$.

4.6.4 La Distribución Normal Bivariante

El ejemplo más importante de una distribución de probabilidad absolutamente continua para vectores aleatorios es la **distribución normal bivalente**. Esta distribución describe dos variables aleatorias continuas, X y Y , cuya relación está modelada por una correlación lineal y tiene forma de campana (gaussiana) en dos dimensiones.

4.6.4.1 Función de Densidad Conjunta La función de densidad conjunta de la distribución normal bivalente con medias μ_X , μ_Y , desviaciones estándar σ_X , σ_Y y coeficiente de correlación ρ es:

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} \right]\right)$$

Esta expresión se generaliza fácilmente de la distribución normal univariante, pero en este caso incluye términos adicionales que representan la interacción entre X y Y .

4.6.4.2 Ejemplo En vez de proporcionar un código para visualizar la distribución normal bivalente podéis seguir este enlace: <https://datasciencegenie.com/3d-contour-plots-of-bivariate-normal-distribution/> en donde se extiende lo que acabamos de discutir y se proporciona algunos ejemplos con R.

4.6.4.3 Distribuciones Marginales Para obtener las **distribuciones marginales** a partir de una **normal bivalente**, debemos integrar la densidad conjunta sobre una de las variables. Dado que estamos trabajando con una distribución normal bivalente, su densidad conjunta está dada por:

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} \right]\right)$$

Para obtener la **marginal de X** , debemos integrar sobre Y :

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$$

Al realizar esta integral, se obtiene que la distribución marginal de X es:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{(x - \mu_X)^2}{2\sigma_X^2}\right)$$

Esto muestra que X sigue una distribución normal con media μ_X y varianza σ_X^2 , es decir, $X \sim N(\mu_X, \sigma_X^2)$.

Del mismo modo, para la **marginal de** Y , integramos sobre X :

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx$$

La solución de esta integral da:

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left(-\frac{(y - \mu_Y)^2}{2\sigma_Y^2}\right)$$

Lo que significa que Y sigue una distribución normal con media μ_Y y varianza σ_Y^2 , es decir, $Y \sim N(\mu_Y, \sigma_Y^2)$.

4.6.4.4 Ejemplo Supongamos que tenemos una distribución normal bivalente con los siguientes parámetros:

- $\mu_X = 100, \sigma_X = 15$
- $\mu_Y = 50, \sigma_Y = 10$
- $\rho = 0.5$

La densidad conjunta es:

$$f_{X,Y}(x, y) = \frac{1}{2\pi(15)(10)\sqrt{1-0.5^2}} \exp\left(-\frac{1}{2(1-0.5^2)} \left[\frac{(x-100)^2}{15^2} + \frac{(y-50)^2}{10^2} - \frac{2(0.5)(x-100)(y-50)}{(15)(10)} \right]\right)$$

Integrando sobre Y , obtenemos la distribución marginal de X :

$$f_X(x) = \frac{1}{\sqrt{2\pi(15^2)}} \exp\left(-\frac{(x-100)^2}{2 \cdot 15^2}\right)$$

De manera análoga, la marginal de Y es:

$$f_Y(y) = \frac{1}{\sqrt{2\pi(10^2)}} \exp\left(-\frac{(y-50)^2}{2 \cdot 10^2}\right)$$

4.6.5 Distribuciones Condicionales

La distribución condicional de una variable dado un valor específico de la otra también es normal univariante. Por ejemplo, la distribución condicional de X dado $Y = y$ es:

$$X | Y = y \sim N\left(\mu_X + \rho \frac{\sigma_X}{\sigma_Y}(y - \mu_Y), (1 - \rho^2)\sigma_X^2\right)$$

De forma análoga, la distribución condicional de Y dado $X = x$ es:

$$Y | X = x \sim N\left(\mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X), (1 - \rho^2)\sigma_Y^2\right)$$

4.6.5.1 Ejemplo Podemos calcular la distribución condicional de X dado que $Y = 180$ cm, y mostrar cómo cambia la distribución de X bajo esta condición:

```
# Valores originales
mu <- c(100, 50)
sigma <- c(15, 10)
rho <- 0.5

# Condicionar X dado Y = 180
y_cond <- 180
mu_cond <- mu[1] + 0.6 * (10/7) * (y_cond - mu[2])
sigma_cond <- sqrt(1 - 0.6^2) * 10

# Mostrar la media y desviación estándar condicionales
mu_cond

## [1] 211.4286

sigma_cond

## [1] 8
```

Esto nos dice que el peso medio de una persona con altura de 180 cm es mayor que el peso medio de la población total, y su desviación estándar es menor debido a la correlación positiva entre peso y altura.

4.7 Independencia de variables aleatorias

Una vez introducido el concepto de distribución conjunta pasamos a estudiar un caso particularmente importante de distribución conjunta, la independencia. De forma aparentemente contradictoria, en este caso, las variables se caracterizan por el hecho de que *no varían conjuntamente* sino que lo hacen *independientemente* las unas de las otras.

De manera intuitiva podemos decir que dos variables aleatorias son independientes si los valores que toma una de ellas no afectan a los de la otra ni a sus probabilidades.

En muchas ocasiones la independencia será evidente a partir del experimento, por ejemplo, es independiente el resultado del lanzamiento de un dado y el de una moneda tres veces. Por tanto las variables:

- X_1 : “Puntuación obtenida con el dado” y
- X_2 : “Número de caras obtenidas al lanzar tres veces una moneda” serán variables independientes.

En otras ocasiones tenemos una dependencia clara, por ejemplo, al lanzar un dado consideremos las variables

- Y_1 =: puntuación del dado,
- Y_2 =: variable indicadora de puntuación par.

Es evidente que existe una clara dependencia, si sabemos que $Y = 1$, la variable X sólo puede tomar los valores 2, 4 o 6; si sabemos que $X = 3$, entonces, $Y = 0$ forzosamente.

Algunas veces podemos suponer la existencia de una cierta relación entre variables, aunque sea en forma algo abstracta y sin concretar. Por ejemplo si realizamos unas mediciones sobre unos individuos, las variables altura en cm y peso en Kg probablemente estarán relacionadas, los valores de una influirán en los valores de la otra. Intentar determinar la naturaleza exacta de la relación entre ambas es lo que en estadística conocemos como un problema de correlación (si nos interesa únicamente la asociación) o de regresión (si queremos modelizar una variable en función de la otra).

Si queremos una definición algo más formal, basta con que recordemos que dos sucesos son independientes si la probabilidad de la intersección es igual al producto de probabilidades, aplicando esta definición a sucesos del tipo $X \leq a$ tenemos la definición siguiente:

4.7.1 Primera caracterización de la independencia

Diremos que dos variables aleatorias X e Y son independientes si y sólo si su función de distribución conjunta puede expresarse como el producto de las funciones de distribución marginales, es decir si

$$F_{X,Y}(x, y) = P((X \leq x) \cap (Y \leq y)) = P(X \leq x) \times P(Y \leq y) = F_X(x) \times F_Y(y)$$

Fijémonos que, como en otros casos, la función que nos permite caracterizar una condición de forma general es la función de distribución.

4.7.1.1 Variables discretas independientes En el caso de las variables discretas la caracterización de la independencia puede hacerse, además, por las funciones de masa de probabilidad:

Diremos que dos variables aleatorias **discretas** X e Y son independientes si y sólo si su función de masa de probabilidad conjunta puede expresarse como el producto de las funciones de masa de probabilidad marginales, es decir si

$$f_{X,Y}(x, y) = P((X = x) \cap (Y = y)) = P(X = x) \times P(Y = y) = f_X(x) \times f_Y(y)$$

4.7.2 Propiedades de las variables independientes

Como consecuencia inmediata de la independencia de X e Y , se cumple lo siguiente:

$$P(a < X \leq c \cap b < Y \leq d) = P(a < X \leq c) \cdot P(b < Y \leq d)$$

Que podría re-enunciarse diciendo que la probabilidad conjunta en un rectángulo definido por los valores “a, c, b, d” es el producto de las probabilidades marginales en los segmentos “ac”, para X y “bd” para Y .

4.8 Momentos de vectores aleatorios

Una vez hemos introducido los vectores aleatorios, que como hemos señalado, son variables aleatorias bi, tri o n -dimensionales tiene sentido preguntarse como se extienden a dichos vectores los conceptos y propiedades que introdujimos para variables aleatorias unidimensionales.

Ya hemos visto como, para las funciones de probabilidad, la función de densidad o la función de distribución, existen extensiones inmediatas, la función de densidad conjunta o la función de distribución conjunta.

Hemos visto también que, además de dichas extensiones, aparecen nuevos conceptos, que sólo tienen sentido en dos o más dimensiones, como las funciones de densidad condicionales o funciones de densidad marginales.

Al considerar conceptos como la media o la varianza veremos que sucede algo similar:

- Por un lado conceptos como el de esperanza se extiende inmediatamente al vector de medias.
- Por otro, conceptos como la varianza, han de tener en cuenta ahora, la posibilidad de variación conjunta entre dos o más variables lo que lleva a introducir magnitudes como la covarianza y la correlación.
 - La extensión del concepto de varianza pasa ahora a combinar extensiones y conceptos nuevos en lo que se conoce como matriz de varianzas-covarianzas.

4.8.1 Esperanza de un vector aleatorio o vector de medias

La **esperanza matemática** de un vector aleatorio es un vector que contiene las esperanzas matemáticas de cada una de las componentes de dicho vector.

Si tenemos un vector aleatorio bivalente $\mathbf{X} = (X_1, X_2)$, su esperanza $\mathbb{E}(\mathbf{X})$ está dada por:

$$\mathbb{E}(\mathbf{X}) = \begin{pmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \end{pmatrix}$$

Consideremos un experimento en el que estamos midiendo el nivel de expresión génica de dos genes X_1 y X_2 en una muestra de células. Si los niveles promedio de expresión son $\mu_1 = 5$ y $\mu_2 = 8$, entonces la esperanza del vector aleatorio sería:

$$\mathbb{E}(\mathbf{X}) = \begin{pmatrix} 5 \\ 8 \end{pmatrix}$$

4.8.2 Covarianza entre dos variables aleatorias

La **covarianza** entre dos variables aleatorias X_1 y X_2 es una medida del grado de dependencia *lineal* entre ellas.

La covarianza se define como

$$\text{Cov}(X_1, X_2) = \mathbb{E}[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))]$$

Supongamos que estamos midiendo la cantidad de dos metabolitos X_1 y X_2 en una muestra, y queremos saber si sus concentraciones tienden a aumentar o disminuir juntas. Si obtenemos una covarianza de 0.5, y conocemos la escala en que varían los datos, podemos concluir que existe ligera tendencia a que los aumentos en X_1 estén asociados con aumentos en X_2 .

4.8.3 Covarianza y correlación

El ejemplo anterior es claramente insatisfactorio, puesto que valores de 0.5 pueden sugerir una gran dependencia o casi ninguna, según cual sea la escala o el rango de variación de los valores que se consideran.

Para evitar esta arbitrariedad se introduce la correlación lineal.

La **correlación** entre dos variables aleatorias es una medida estandarizada del grado de dependencia lineal entre dos variables (es decir de la covarianza), que toma valores entre -1 y 1 y que se define como:

$$\text{Corr}(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)\text{Var}(X_2)}}$$

En el caso de los metabolitos mencionados anteriormente, si $\text{Cov}(X_1, X_2) = 0.5$, $\text{Var}(X_1) = 2$ y $\text{Var}(X_2) = 3$, podemos calcular la correlación, que valdría:

$$\text{Corr}(X_1, X_2) = \frac{0.5}{\sqrt{2 \times 3}} = \frac{0.5}{\sqrt{6}} \approx 0.204$$

Esto indica una correlación positiva débil entre las concentraciones de los dos metabolitos.

Obsérvese, sin embargo que si en vez de los valores anteriores para las varianzas de X e Y hubiéramos tenido $\text{Var}(X_1) = 1$ y $\text{Var}(X_2) = .5$ el valor de la correlación habría sido:

$$\text{Corr}(X_1, X_2) = \frac{0.5}{\sqrt{1 \times 0.5}} = \frac{0.5}{\sqrt{0.5}} \approx 0.7071$$

Este ejemplo muestra como la correlación aporta más información sobre la dependencia lineal, puesto que, además de tener en cuenta la variación conjunta, tiene en cuenta la variabilidad individual de cada componente.

4.8.4 Matriz de varianzas-covarianzas

La **matriz de varianzas-covarianzas** de un vector aleatorio $\mathbf{X} = (X_1, X_2)$ es una matriz que contiene las varianzas de las componentes en la diagonal y las covarianzas fuera de la diagonal. Está definida como:

$$\text{Cov}(\mathbf{X}) = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) \end{pmatrix}$$

Siguiendo con el ejemplo de los metabolitos, si $\text{Var}(X_1) = 2$, $\text{Var}(X_2) = 3$, y la covarianza es 0.5, la matriz de covarianzas sería:

$$\text{Cov}(\mathbf{X}) = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 3 \end{pmatrix}$$

Esto nos indica la dispersión de cada variable y la relación entre ambas.

La distribución normal bivalente

Una de las distribuciones más importantes que describe el comportamiento conjunto de dos variables aleatorias es la **distribución normal bivalente**.

Un vector aleatorio $\mathbf{X} = (X_1, X_2)$ tiene una distribución normal bivalente si su función de densidad conjunta está dada por:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right]\right)$$

Aquí, μ_1 y μ_2 son las medias de X_1 y X_2 , σ_1^2 y σ_2^2 son las varianzas, y ρ es el coeficiente de correlación.

4.8.5 Matriz de correlaciones

La **matriz de correlaciones** de un vector aleatorio bivalente $\mathbf{X} = (X_1, X_2)$ es una matriz simétrica 2×2 que contiene los coeficientes de correlación entre las componentes X_1 y X_2 . La correlación mide la relación lineal entre las variables y se define como:

$$\text{Corr}(X_1, X_2) = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\text{Var}(X_1)\text{Var}(X_2)}}$$

La matriz de correlaciones $\text{Corr}(\mathbf{X})$ está dada por:

$$\text{Corr}(\mathbf{X}) = \begin{pmatrix} 1 & \text{Corr}(X_1, X_2) \\ \text{Corr}(X_2, X_1) & 1 \end{pmatrix}$$

Dado que $\text{Corr}(X_1, X_2) = \text{Corr}(X_2, X_1)$, la matriz es simétrica, y los elementos diagonales son siempre 1 porque la correlación de una variable consigo misma es 1.

4.8.5.1 Relación con la matriz de covarianzas La matriz de correlaciones está relacionada con la **matriz de covarianzas** de la forma siguiente:

Si Σ es la matriz de covarianzas de $\mathbf{X} = (X_1, X_2)$, con $\Sigma = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) \end{pmatrix}$, la matriz de correlaciones se obtiene “normalizando” cada covarianza dividiendo por el producto de las desviaciones estándar de las respectivas variables:

$$\text{Corr}(\mathbf{X}) = \begin{pmatrix} 1 & \frac{\text{Cov}(X_1, X_2)}{\sigma_1 \sigma_2} \\ \frac{\text{Cov}(X_2, X_1)}{\sigma_1 \sigma_2} & 1 \end{pmatrix}$$

donde $\sigma_1 = \sqrt{\text{Var}(X_1)}$ y $\sigma_2 = \sqrt{\text{Var}(X_2)}$.

Supongamos que medimos dos variables, como la altura X_1 y el peso X_2 de un grupo de personas. Sabemos que:

- $\text{Var}(X_1) = 25$ (varianza de la altura),
- $\text{Var}(X_2) = 100$ (varianza del peso),
- $\text{Cov}(X_1, X_2) = 40$ (covarianza entre altura y peso).

La **matriz de covarianzas** sería:

$$\Sigma = \begin{pmatrix} 25 & 40 \\ 40 & 100 \end{pmatrix}$$

La correlación entre X_1 y X_2 se calcula como:

$$\text{Corr}(X_1, X_2) = \frac{40}{\sqrt{25 \times 100}} = \frac{40}{50} = 0.8$$

Por lo tanto, la **matriz de correlaciones** será:

$$\text{Corr}(\mathbf{X}) = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

Esto indica una fuerte correlación positiva entre la altura y el peso de las personas en este grupo. La matriz de correlaciones nos proporciona una forma normalizada de comparar la dependencia entre las variables, sin depender de las unidades de medida.

4.8.6 Segunda caracterización de la independencia

La **independencia** entre dos variables aleatorias X_1 y X_2 puede caracterizarse también a través de sus **esperanzas** de la siguiente manera:

Dos variables son **independientes** si la esperanza del producto de ambas es igual al producto de las esperanzas de cada una por separado. Es decir si se verifica que:

$$\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1] \mathbb{E}[X_2]$$

Esta propiedad refleja que, cuando las variables son independientes, el valor esperado del producto no se ve afectado por la interacción entre ellas, lo que implica que no hay dependencia entre las dos.

Una consecuencia importante de esta propiedad es cómo afecta a la **covarianza** entre X_1 y X_2 .

Si X_1 y X_2 son **independientes**, entonces, por la propiedad anterior, $\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1] \mathbb{E}[X_2]$ lo que, a su vez, significa que la covarianza es cero:

$$\text{Cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1] \mathbb{E}[X_2] = 0$$

Por lo tanto, **si dos variables son independientes, necesariamente su covarianza es cero.**

Sin embargo, la inversa no es cierta: el hecho de que la covarianza sea cero no implica que las variables sean independientes.

4.8.7 Relación entre incorrelación e independencia

Cuando la covarianza entre dos variables es cero, se dice que las variables son **incorreladas**. Aunque la **independencia** implica que las variables son incorreladas, lo contrario no siempre es verdad: dos variables pueden ser incorreladas (tener covarianza cero) pero **no independientes**.

Un ejemplo clásico es el siguiente: si consideramos una variable aleatoria X y definimos $Y = X^2$, entonces, aunque la covarianza entre X y Y puede ser cero (especialmente si X tiene una distribución simétrica alrededor de 0, como la normal estándar), X y Y no son independientes, porque el valor de Y está completamente determinado por X .

Consideremos dos variables aleatorias X_1 y X_2 que siguen una distribución normal conjunta bivalente con media cero:

$$(X_1, X_2) \sim \mathcal{N}(\mathbf{0}, \Sigma)$$

Si la **matriz de covarianzas** Σ es diagonal, es decir, $\text{Cov}(X_1, X_2) = 0$, entonces X_1 y X_2 son incorreladas.

En este caso particular, cuando las variables son normales, la incorrelación **sí** implica independencia, porque en distribuciones normales la ausencia de correlación (covarianza cero) también implica que no hay ninguna dependencia entre las variables.

Sin embargo, en otras distribuciones que no son normales, la incorrelación no garantiza la independencia, lo que subraya la importancia de distinguir entre los dos conceptos.

5 Grandes muestras

5.1 Introducción: Aproximaciones asintóticas

En estadística y teoría de la probabilidad, el estudio de las grandes muestras juega un papel crucial debido a su relevancia tanto en la definición frecuentista de probabilidad como en la construcción de estimadores en la práctica estadística.

- Desde la perspectiva de la probabilidad frecuentista, la probabilidad se define como el límite de la frecuencia relativa de un evento cuando el número de ensayos tiende a infinito.
- En el contexto de la estadística, las grandes muestras sirven como base para muchas aproximaciones importantes, como las distribuciones de muestreo, las estimaciones de parámetros y la validación de inferencias.

La ley de los grandes números y el teorema central del límite son ejemplos clave de teoremas que se fundamentan en el comportamiento de las muestras grandes, proporcionando las bases para muchos de los métodos estadísticos utilizados en la inferencia moderna.

5.2 Ley de los Grandes Números (Ley débil)

La **ley de los grandes números** establece que, a medida que el tamaño de la muestra aumenta, la media muestral se aproxima a la media de la población.

Formalmente, la ley de los grandes números en su versión débil se enuncia de la siguiente manera:

Sea X_1, X_2, \dots, X_n una secuencia de variables aleatorias independientes e idénticamente distribuidas (i.i.d.) con esperanza $\mu = \mathbb{E}[X_i]$ y varianza $\sigma^2 = \text{Var}(X_i)$, entonces para cualquier $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq \epsilon \right) = 0.$$

Esto significa que, con alta probabilidad, la media muestral $\frac{1}{n} \sum_{i=1}^n X_i$ se aproxima a μ a medida que n crece.

5.2.1 Ejemplo

Imaginemos un dado equilibrado. Sabemos que la esperanza de cada lanzamiento es el valor promedio de los números en el dado, que es

$$\mu = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5.$$

Ahora, supongamos que lanzamos el dado repetidamente y calculamos la media de los resultados. Al principio, con pocos lanzamientos, la media puede estar alejada de 3.5, pero a medida que aumentan los lanzamientos, la media se acercará más y más a 3.5, como lo predice la ley de los grandes números. Es decir, a medida que lanzamos más veces el dado, la probabilidad de que la media de los resultados se aleje de 3.5 por más de una cantidad arbitraria disminuye.

Podemos ilustrarlo con el siguiente código de R

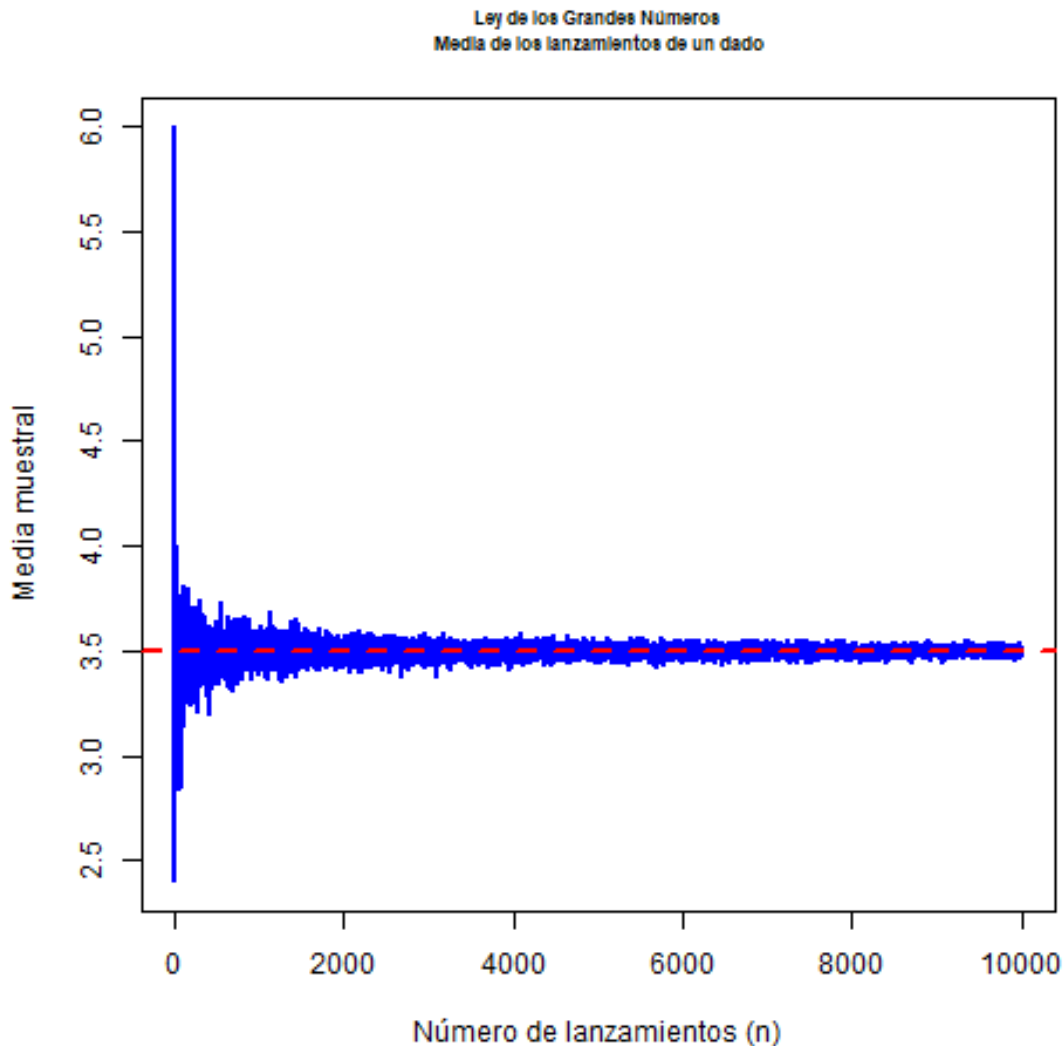
```
# Definir la función para simular lanzamientos de un dado
simular_dado <- function(max_n) {
  medias <- numeric(max_n) # Vector para almacenar las medias muestrales
  for (n in 1:max_n) {
    lanzamientos <- sample(1:6, n, replace = TRUE) # Lanzar el dado n veces
    medias[n] <- mean(lanzamientos) # Calcular la media de los lanzamientos
  }
  return(medias)
}

# Simular para un tamaño máximo de muestra de 10000 lanzamientos
max_n <- 10000
medias <- simular_dado(max_n)

# Graficar las medias muestrales a medida que n aumenta
png("images/LLN1.png")
plot(1:max_n, medias, type = "l", col = "blue", lwd = 2,
     xlab = "Número de lanzamientos (n)", ylab = "Media muestral",
     main = "Ley de los Grandes Números\nMedia de los lanzamientos de un dado", cex.main=0.7)
abline(h = 3.5, col = "red", lwd = 2, lty = 2) # Línea horizontal en 3.5
dev.off()

## pdf
## 2

knitr::include_graphics("images/LLN1.png",rel_path = TRUE)
```



Este comportamiento es una manifestación intuitiva de la ley débil de los grandes números, ya que nos garantiza que la media muestral se acercará a la media poblacional a medida que el número de observaciones aumente.

5.3 El teorema central del límite

El teorema central del límite (a partir de ahora, TCL) presenta un doble interés. Por un lado, proporciona a la estadística un resultado crucial para abordar el estudio de la distribución asintótica de muchos tipos de variables aleatorias. Como se verá en próximos capítulos, va a resultar básico en la construcción de contrastes de hipótesis y de intervalos de confianza, dos herramientas esenciales en estadística aplicada.

Además, el TCL proporciona una explicación teórica fundamentada a un fenómeno habitual en experimentos reales: las variables estudiadas presentan muchas veces una distribución empírica aproximadamente normal.

El TCL forma parte de un conjunto de propiedades relativas a las convergencias de variables aleatorias. En este tema se estudia sólo un tipo de convergencia, la convergencia en ley, ya que es necesaria para entender el enunciado del TCL. Se descarta, pues, en este documento el estudio de los otros tipos de convergencias (en probabilidad, casi segura, etc.) y el estudio de las leyes de los grandes números.

Posiblemente el lector con poca formación en análisis matemático hallará alguna dificultad en la primera

lectura de la definición de convergencia en ley y en el enunciado del TCL. Si es este el caso, los ejemplos incluidos han de ayudar en su comprensión. Consideramos al TCL un resultado básico con el que hay que familiarizarse, ya que se aplicará repetidamente en los próximos temas.

5.3.1 Sumas de variables aleatorias

El TCL estudia el comportamiento de las sumas de variables aleatorias. En temas anteriores se han visto ya ejemplos de sumas de variables aleatorias.

Formalmente, la suma de dos variables aleatorias corresponde a la siguiente aplicación: si X_1 y X_2 son dos variables aleatorias definidas sobre Ω , la suma es:

$$\begin{aligned} X_1 + X_2 : \Omega &\rightarrow \mathbb{R} \\ \omega &\mapsto X_1(\omega) + X_2(\omega) \end{aligned}$$

La suma de dos variables puede extenderse sin dificultad a sumas de tres, cuatro, ... y, en general, n variables aleatorias.

El TCL se ocupa de las sucesiones de variables aleatorias. En el contexto del TCL una sucesión corresponde a un conjunto donde el primer elemento es una variable aleatoria, el segundo elemento es la suma de dos variables aleatorias, el tercero es la suma de tres variables aleatorias, y así sucesivamente.

Una sucesión es un conjunto de elementos infinitos, que se designan simbólicamente mediante $\{X_n\}$. Cada uno de los elementos de la sucesión (que es una variable aleatoria) lleva asociada una determinada función de distribución:

$$X_n \rightarrow F_n$$

Así pues, la sucesión de variables aleatorias lleva asociada una secuencia paralela de funciones de distribución.

5.3.2 Definición de convergencia en ley

La siguiente definición se ocupa del comportamiento de las sucesiones. Sea $\{X_n\}$ una sucesión de variables aleatorias, y sea $\{F_n\}$ la correspondiente sucesión de funciones de distribución. Se dice que $\{X_n\}$ converge en ley a una variable aleatoria X de función de distribución F si:

$$\lim_{n \rightarrow \infty} F_n(x) = F(x) \quad \text{para todo } x \text{ donde } F \text{ es continua.}$$

Se indica que la sucesión converge en ley mediante el símbolo:

$$X_n \xrightarrow{L} X$$

El significado de la definición es que, al aumentar arbitrariamente n , las sucesivas funciones de distribución de la secuencia se aproximan a la distribución F de la variable X .

En los ejemplos se presentan gráficamente algunas situaciones donde diferentes sucesiones de variables aleatorias convergen en ley a una variable aleatoria normal.

5.3.3 Enunciado del teorema central del límite

A continuación se presenta el enunciado del TCL en la versión de Lindeberg y Lévy. Teorema: Sea X_1, X_2, \dots, X_n , un conjunto de variables aleatorias independientes idénticamente distribuidas, cada una de ellas con función de distribución F , y supongamos que $E(X_k) = \mu$ y $\text{var}(X_k) = \sigma^2$ para cualquier elemento del conjunto. Si designamos a la suma normalizada de n términos con el símbolo:

$$S_n^* = \frac{X_1 + X_2 + \dots + X_n - n\mu}{\sigma\sqrt{n}}$$

entonces la sucesión de sumas normalizadas converge en ley a la variable aleatoria normal tipificada $Z \sim N(0, 1)$, es decir:

$$S_n^* \xrightarrow{L}$$

El teorema anterior tiene dos importantes corolarios:

1. Si consideramos la suma ordinaria de las n variables aleatorias, es decir, $S_n = X_1 + X_2 + \dots + X_n$, entonces la sucesión de sumas ordinarias converge en ley a una normal de media $n\mu$ y varianza $n\sigma^2$.
2. Si consideramos el promedio de las n variables aleatorias, es decir, $n^{-1}S_n$, entonces la sucesión de promedios converge en ley a una normal de media μ y varianza $n^{-1}\sigma^2$.

5.3.3.1 Comentarios al teorema:

1. La convergencia a la normal tipificada se produce con cualquier tipo de variable que cumpla las condiciones del teorema, sea discreta o absolutamente continua.
2. Un sinónimo para indicar que una sucesión converge en ley a una normal es señalar que es asintóticamente normal.
3. El TCL presenta el comportamiento de sumas infinitas de variables aleatorias. Veremos posteriormente como interpretar el resultado para valores finitos.
4. Existen otras versiones del TCL donde se relajan las condiciones de la versión de Lindeberg y Lévy, que, como se ha visto, obliga a las variables aleatorias a tener idénticas medias y varianzas. Dichas versiones del TCL necesitan el conocimiento de conceptos matemáticos que exceden el nivel al que se orienta Statmedia, y por esta razón se omite su enunciado.

5.3.4 Algunos ejemplos de aplicación del TCL

```
# Parámetros de la distribución binomial
n <- 1000 # Número de ensayos
p <- 0.5 # Probabilidad de éxito
size <- 10000 # Número de simulaciones

# Generar una variable aleatoria binomial
binomial_sample <- rbinom(size, n, p)

# Estimación de la media y la desviación estándar de la distribución binomial
mean_binom <- n * p
sd_binom <- sqrt(n * p * (1 - p))

# Generar la distribución normal aproximada
normal_sample <- rnorm(size, mean = mean_binom, sd = sd_binom)

# Graficar los histogramas de la binomial y la normal
par(mfrow = c(1, 2)) # Organizar gráficos en dos paneles

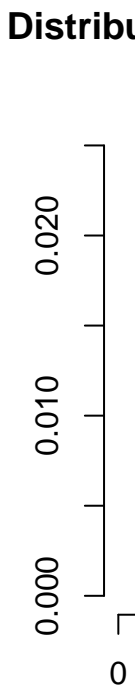
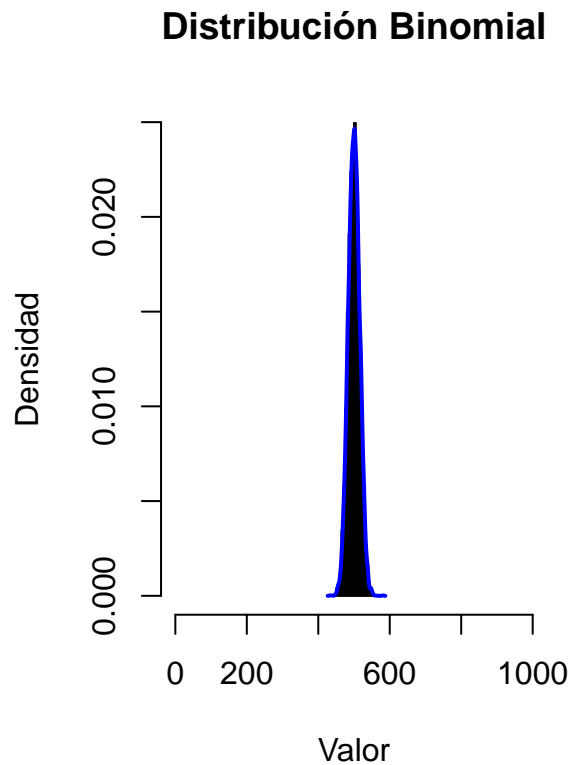
# Histograma de la muestra binomial
hist(binomial_sample, breaks = 50, probability = TRUE,
     col = rgb(0, 0, 1, 0.5), xlim = c(0, n),
     main = "Distribución Binomial", xlab = "Valor",
```

```

    ylab = "Densidad")
lines(density(binomial_sample), col = "blue", lwd = 2)

# Histograma de la distribución normal aproximada
hist(normal_sample, breaks = 50, probability = TRUE,
     col = rgb(1, 0, 0, 0.5), xlim = c(0, n),
     main = "Distribución Normal Aproximada", xlab = "Valor",
     ylab = "Densidad")
lines(density(normal_sample), col = "red", lwd = 2)

```



5.3.4.1 Normalidad asintótica de la Binomial.

```

# Parámetros de la simulación
num_simulaciones <- 10000 # Número de simulaciones
num_lanzamientos <- c(10, 100, 1000, 10000) # Diferentes tamaños de muestra

# Función para simular la suma de las puntuaciones de un dado
simular_suma_dado <- function(n) {
  suma <- rowSums(matrix(sample(1:6, n * num_simulaciones, replace = TRUE),
                        ncol = n)) # Simulación de las sumas
  return(suma)
}

# Graficar las distribuciones de las sumas para diferentes tamaños de muestra
par(mfrow = c(2, 2)) # Organizar gráficos en 2x2

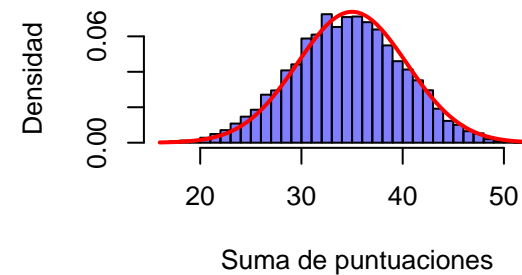
```

```

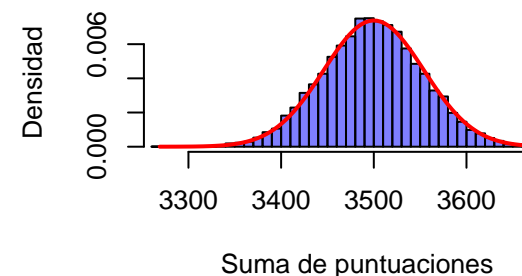
for (n in num_lanzamientos) {
  suma_dado <- simular_suma_dado(n)
  # Histograma de la suma de las puntuaciones del dado
  hist(suma_dado, breaks = 50, probability = TRUE,
       col = rgb(0, 0, 1, 0.5), xlim = c(min(suma_dado), max(suma_dado)),
       main = paste("Suma de", n, "lanzamientos de un dado"),
       xlab = "Suma de puntuaciones", ylab = "Densidad")
  # Superponer la curva de densidad normal (aproximación asintótica)
  mean_dado <- 3.5 * n # Media esperada de la suma (media de un dado es 3.5)
  sd_dado <- sqrt(n * (35 / 12)) # Desviación estándar de la suma (varianza de un dado es 35/12)
  curve(dnorm(x, mean = mean_dado, sd = sd_dado),
        col = "red", lwd = 2, add = TRUE)
}

```

Suma de 10 lanzamientos de un dado



Suma de 1000 lanzamientos de un dado



5.3.4.2 Normalidad asintótica de la suma de puntuaciones de un dado

5.3.5 Casos particulares más notables

Aunque el TCL tiene multitud de casos particulares interesantes, son especialmente relevantes para el desarrollo de los próximos temas los siguientes casos:

5.3.5.1 Promedio de n variables aleatorias Al considerar n variables independientes, todas con la misma distribución, cada una de ellas con esperanza igual a μ y varianza igual a σ^2 , el promedio es asintóticamente normal con media μ y varianza $n^{-1}\sigma^2$. Este resultado proporciona una distribución asintótica a la media de n observaciones en el muestreo aleatorio simple que se estudiará en el próximo tema.

5.3.5.2 Binomial de parámetros n y p Es asintóticamente normal con media np y varianza $np(1-p)$. Históricamente (de Moivre, 1733), es el primer resultado demostrado de convergencia a una normal.

5.3.5.3 Poisson de parámetro $n\lambda$ Es asintóticamente normal con media $n\lambda$ y varianza $n\lambda$.

5.3.6 Interpretación del teorema central del límite

El TCL hace referencia a sucesiones infinitas, por tanto, la igualdad de las distribuciones se alcanza sólo en el límite, y hace mención a una distribución final teórica o de referencia.

Sin embargo, puede utilizarse esta distribución final de referencia para aproximar distribuciones correspondientes a sumas finitas. Algunos casos particulares importantes (binomial, Poisson, etc.) alcanzan grados de aproximación suficientes para sumas con no demasiados términos.

Los resultados que se indican a continuación son, por tanto, aproximaciones que se consideran usualmente suficientes, pero conllevan errores numéricos de aproximación.

1. Binomial: aproximar si $n \geq 30$ y $0.1 \leq p \leq 0.9$ a una normal de media np , varianza $np(1-p)$. Ver aquí más detalles.
2. Poisson: aproximar si $\lambda \geq 10$ a una normal de media λ y varianza λ . Ver aquí más detalles.

Para evaluar aproximadamente el error cometido en las aproximaciones, puede consultarse los cuadros gráficos de los ejemplos de este tema.

El TCL permite aproximar funciones de distribución, independientemente del carácter (continuo o discreto) de las variables sumadas. No sirve, por tanto, para aproximar la funciones de densidad discretas por una normal. En el caso continuo sí puede establecerse también una convergencia de las densidades asociadas.

Finalmente, es conveniente mencionar que existen resultados teóricos que permiten estudiar la velocidad de convergencia de una suma de variables aleatorias a la normal, sin embargo la dificultad técnica que conllevan trasciende el nivel marcado para el conjunto de documentos marcado para Statmedia.

5.3.7 Acerca de las variables aproximadamente normales

En general, cuando se estudia en experimentos reales una determinada variable no se conoce su distribución teórica. Sin embargo, puede establecerse su distribución empírica a partir de una muestra más o menos amplia.

Una forma habitual de presentar la distribución empírica es construir el histograma de clases de dicha variable. Es un hecho conocido desde el siglo XIX que esta distribución empírica presenta muchas veces una forma que es aproximadamente normal. Por ejemplo, al realizar un estudio sobre el peso de adultos varones de dieciocho años en Catalunya, se observó la distribución siguiente en la muestra:

El TCL permite dar una explicación a este fenómeno. La variable peso de un adulto viene determinada en cada individuo por la conjunción de multitud de diferentes factores. Algunos de estos factores son ambientales (dietas, ejercicio, enfermedades, etc.) y otros son congénitos. Con el nivel actual de conocimiento no se pueden desglosar completamente todos los factores que intervienen, pero puede aceptarse en cambio que la variable peso es el resultante de la suma de diferentes variables primarias, congénitas o ambientales, y que posiblemente no todas tienen el mismo grado de influencia. Seguramente, estas variables primarias tampoco tienen la misma media, varianza o, incluso, la misma distribución.

La versión del TCL que se ha presentado aquí exige estas condiciones para la convergencia a la normal, pero, como ya se ha comentado antes otras versiones más elaboradas del TCL permiten modelar la suma de variables de forma menos restringida. En este contexto, al considerar la variable peso como una suma más o menos extensa (pero finita) de diferentes variables primarias, es esperable que ocurra que la variable resultante, el peso, siga una distribución aproximadamente normal.

De forma similar es explicable la normalidad aproximada que se observa en muchas variables biométricas (pesos, alturas, longitudes, concentraciones de metabolitos, distribuciones de edad, etc.) así como en muchos otros contextos (distribución de rentas, errores de medición, etc.). A pesar de esta ubicuidad de la distribución normal, el lector no debe inferir que es forzosamente, ni mucho menos, la distribución de referencia en todo estudio aplicado.

6 Introducción a la inferencia estadística

6.1 Inferencia estadística

Para comenzar, vamos a definir cuál es el ámbito de estudio de la inferencia estadística desde su relación con el cálculo de probabilidades. El cálculo de probabilidades proporciona una teoría matemática que permite analizar (o modelizar) las propiedades de los fenómenos donde interviene el azar. El cálculo de probabilidades utiliza como modelo básico para cualquier situación aleatoria el concepto de espacio de probabilidades (Ω, \mathcal{A}, P) y una variable aleatoria $X : \Omega \rightarrow \mathbb{R}$ definida sobre él. El conocimiento de la distribución de la variable aleatoria permite:

1. Análisis deductivo de situaciones. Por ejemplo: si asumimos que el peso de los recién nacidos se distribuye según una distribución $N(\mu = 3 \text{ kg}, \sigma = 0.25 \text{ kg})$, nos puede interesar calcular la probabilidad de que un recién nacido pese entre 2.9 y 3.1 kg, o encontrar unos valores centrados en la media entre los cuales esperemos que se encuentren el 10%(25%, 50%, 95%, ...) de los recién nacidos.
2. Modelización de situaciones aleatorias. Por ejemplo: si asumimos que el tiempo, en años, hasta que se estropea un componente de un ordenador se distribuye según una distribución exponencial $T \sim \xi(\lambda = 0.3)$, nos puede interesar calcular la probabilidad de que un componente dado dure más de 4 años.

En los casos anteriores nos encontramos en una situación muy común, donde ya disponemos de un modelo sobre el cual efectuamos los cálculos, pero del cual desconocemos la procedencia. Parece razonable, y de hecho es precisamente así, que si queremos adaptar un modelo a una situación debemos basarnos únicamente en las observaciones del fenómeno. Si queremos saber cómo se distribuyen los pesos de los recién nacidos tomaremos unos cuantos, los pesaremos y después observaremos la distribución de estos. Puede que no sea necesario pesar a todos los recién nacidos (jde hecho, no es posible!), pero tampoco es posible deducir la ley por consideraciones puramente teóricas. Ahora, en lugar de partir de un espacio de probabilidades, partiremos de unas observaciones (x_1, \dots, x_n) y el objetivo que perseguiremos será obtener información sobre la distribución de probabilidades de un fenómeno a partir de una observación no exhaustiva del mismo.

6.2 Problemas de inferencia estadística

Hemos presentado como objetivo de la inferencia estadística inducir propiedades del modelo probabilístico que representa la población a partir de un conjunto de observaciones. Según el tipo de conclusión que queramos extraer, diferenciaremos diferentes tipos de problemas:

1. Si queremos utilizar la información proporcionada por la muestra para obtener un pronóstico numérico único (es decir, una única aproximación numérica) de una o más características de la población, tenemos un problema de estimación puntual.
2. Si queremos obtener información sobre un rango de valores dentro del cual podamos afirmar, con un cierto grado de confianza, que podemos capturar un parámetro desconocido de la distribución, hablamos de estimación por intervalo.
3. Si lo que queremos hacer es decidir si podemos aceptar o debemos rechazar una afirmación sobre la distribución de probabilidad del fenómeno estudiado, hablamos de contraste de hipótesis. Este contraste puede ser:
 - Paramétrico: si la afirmación (la hipótesis) se refiere a los parámetros de la distribución.
 - No paramétrico: si la afirmación es sobre la forma de la distribución.

6.3 Distribución de la población

Todo problema de inferencia está motivado por un cierto grado de desconocimiento de la ley de probabilidades que rige un determinado fenómeno aleatorio. El caso más sencillo que encontramos es cuando nos interesa una cierta variable X con una función de distribución F desconocida en mayor o menor grado. La distribución que teóricamente sigue la variable de interés X en la población recibe el nombre de distribución teórica o distribución de la población. La distribución de la población es importante ya que, a menudo, se utiliza para determinar la distribución de alguna característica de los individuos de una población. En los modelos de la inferencia estadística indicamos el relativo grado de desconocimiento sobre la distribución F en función de su

pertenencia a una familia \mathcal{F} de distribuciones. Por ello, en lugar de explicar que $X \sim F = F_0$ indicaremos que $X \sim F \in \mathcal{F}$, donde \mathcal{F} puede ser un conjunto más o menos amplio de distribuciones de probabilidad, como todas las distribuciones normales o las distribuciones simétricas o las distribuciones discretas sobre \mathbb{N} . Muchas veces, la distribución poblacional F está completamente especificada excepto por el valor de algún parámetro o parámetros. En este caso, podemos concretar más la forma de la familia de distribuciones:

$$X \sim F \in \mathcal{F} = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$$

donde Θ es el espacio de los k parámetros. La familia de posibles distribuciones de probabilidad para X se denomina, genéricamente, modelo estadístico y se indica como: $\{X \sim F_\theta : \theta \in \Theta\}$. Veamos algunos ejemplos.

Ejemplo 1.3.1 Supongamos que X representa la duración de un componente electrónico que no envejece, solo se estropea. Es decir, si en un instante t está funcionando, su estado es el mismo que en cualquier momento del pasado y la distribución del tiempo hasta que se estropee es la misma que al principio. Esta propiedad se denomina falta de memoria. Un modelo razonable para esta situación lo da la distribución de Weibull que, en este caso, podemos definir a través de la siguiente función de densidad:

$$f_\theta(x) = \begin{cases} \alpha \beta x^{\beta-1} e^{-\alpha x^\beta} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

La familia de distribuciones asociada es

$$\mathcal{F} = \{F_\theta : \theta = (\alpha, \beta) \in (0, \infty) \times (0, \infty)\}$$

Ejemplo 1.3.2 Supongamos que queremos determinar la masa de un cierto tipo de partículas elementales a partir de las observaciones en una cámara de burbujas. En cada observación obtenemos un dato de la masa de la partícula x_i y asociado con ella un cierto error de medida ε_i . Si la masa común de cada una de ellas es μ , entonces podemos escribir:

$$x_i = \mu + \varepsilon_i \quad i = 1, \dots, n$$

donde la distribución $\varepsilon_i \sim F$ es desconocida. Nuestro objetivo es obtener información sobre F . Si admitimos que $P(\varepsilon_i < 0) = P(\varepsilon_i > 0)$, según el grado de exigencia que queramos tener, podemos suponer:

- Con un enfoque de inferencia paramétrica:

$$X \sim F \in \mathcal{F} = \{N(0, \sigma) : \sigma \in \mathbb{R}^+\}$$

- Con un enfoque de inferencia no paramétrica:

$$X \sim F \in \mathcal{F} = \{ \text{Distribuciones simétricas} \}$$

6.4 Muestra aleatoria simple

6.4.1 Definición

Para estudiar un problema de inferencia estadística analizamos una muestra de tamaño n . Se trata de escoger n individuos o elementos de la población Ω

$$\omega_1, \omega_2, \dots, \omega_n$$

que sean representativos. El valor de n y la forma de elección de los individuos de la muestra es una materia de Estadística llamada Muestreo estadístico. Por ahora y para simplificar, solo hace falta decir que la elección

se hace de forma que todos los individuos tienen la misma probabilidad de estar presentes en la muestra, si es necesario con reemplazo, y que el valor de n está dado. En realidad, lo que nos interesa verdaderamente no son los individuos de la muestra sino las mediciones de una característica X sobre ellos. Es decir, los valores de una variable aleatoria X sobre estos individuos

$$X(\omega_1) = x_1, X(\omega_2) = x_2, \dots, X(\omega_n) = x_n$$

También podemos pensar que los valores muestrales x_1, x_2, \dots, x_n son generados directamente desde la variable aleatoria. En todo caso, los valores muestrales no son únicos y podemos generar varias muestras

$$\begin{array}{ccccc} x_1^1 & x_2^1 & x_3^1 & \dots & x_n^1 \\ x_1^2 & x_2^2 & x_3^2 & \dots & x_n^2 \\ \vdots & \vdots & \vdots & & \vdots \\ x_1^s & x_2^s & x_3^s & \dots & x_n^s \end{array}$$

Si todos los valores son independientes, de la misma forma que $x_1, x_2, x_3, \dots, x_n$ es una muestra generada por X , podemos considerar todos los x_1^i $i = 1, \dots, s$ provenientes de una variable aleatoria X_1 con la misma distribución que X $X_1 \stackrel{d}{=} X$ y que genera los primeros valores, los x_i^2 provenientes de una variable aleatoria $X_2 \stackrel{d}{=} X$ que genera los segundos y así sucesivamente. Todo esto nos lleva a definir el concepto de muestra aleatoria de una forma muy conveniente para trabajar con ella:

Definición 1.1 Una muestra aleatoria simple de tamaño n de una variable aleatoria X con distribución F es una colección de n variables aleatorias independientes X_1, X_2, \dots, X_n con la misma distribución F que X . Esto se suele indicar como:

$$\mathbf{X} = X_1, X_2, \dots, X_n \stackrel{i.i.d}{\sim} X$$

Definición 1.2 El conjunto $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ de observaciones concretas de X_1, X_2, \dots, X_n se denomina realización de la muestra.

6.4.2 Distribución de la muestra

Una muestra aleatoria simple, como vector aleatorio n -dimensional que es, tiene una distribución conjunta o distribución de la muestra que depende de F , pero que obviamente es diferente, ya que en particular X y \mathbf{X} tienen dimensiones diferentes. Sin embargo, gracias a la independencia de las variables X_1, X_2, \dots, X_n , la función de distribución conjunta de \mathbf{X} , que podría ser muy complicada, toma una forma muy sencilla. En resumen:

Definición 1.3 Se llama distribución de la muestra de una variable aleatoria $X \sim F$ a la distribución del vector aleatorio n -dimensional (X_1, X_2, \dots, X_n)

$$G(x_1, x_2, \dots, x_n) = F(x_1) F(x_2) \cdots F(x_n)$$

En los casos particulares en que X sea discreta o absolutamente continua, la distribución conjunta de la muestra suele expresarse mediante la función de masa de probabilidad o la función de densidad:

- Para variables discretas:

$$\begin{aligned} p_G(x_1, x_2, \dots, x_n) &= P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\ &= \prod_{i=1}^n P(X = x_i) = \prod_{i=1}^n p_F(x_i), \end{aligned}$$

- Para variables absolutamente continuas:

$$g(x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i)$$

Ejemplo 1.4.1 Una moneda tiene una probabilidad θ de salir cara. Queremos estudiar la variable aleatoria:

$$X = \begin{cases} 1 & \text{si sale cara} \\ 0 & \text{si sale cruz} \end{cases}$$

con densidad $P\{X = 1\} = \theta, P\{X = 0\} = 1 - \theta$. Es decir

$$X \sim F_\theta \in \mathcal{F} = \{F_\theta = B(1, \theta) : \theta \in (0, 1)\}$$

Supongamos que hacemos tres lanzamientos. Las posibles muestras son:

X_1	X_2	X_3	Probabilidad
1	1	1	θ^3
1	0	0	$\theta(1 - \theta)^2$
0	1	0	$\theta(1 - \theta)^2$
0	0	1	$\theta(1 - \theta)^2$
1	0	1	$\theta^2(1 - \theta)$
1	1	0	$\theta^2(1 - \theta)$
0	1	1	$\theta^2(1 - \theta)$
0	0	0	$(1 - \theta)^3$

El muestreo ha especificado la distribución conjunta de la muestra a través de la distribución desconocida F_θ . Si escribimos la función de probabilidades de la variable aleatoria como $f_\theta(x) = \theta^x(1 - \theta)^{1-x}$, entonces la función de probabilidades de la muestra la podemos expresar como:

$$g_\theta(x_1, x_2, x_3) = \theta^{x_1+x_2+x_3}(1 - \theta)^{3-(x_1+x_2+x_3)}$$

6.5 Estadísticos

6.5.1 Definición

Para lograr el objetivo de realizar inferencias sobre la población a partir de la muestra, solemos basarnos en la realización de cálculos sobre la muestra para tratar de obtener la información que deseamos. En este proceso aparecen los conceptos de estadístico y el caso particular, que más nos interesa a nosotros, de estimador. Un estadístico es una función de la muestra que no depende del valor del parámetro.

Definición 1.4 Dada una muestra aleatoria simple X_1, X_2, \dots, X_n y una función medible $T : \mathbb{R}^n \rightarrow \mathbb{R}^k$, entonces $T(X_1, X_2, \dots, X_n)$ es un vector aleatorio (variable aleatoria cuando $k = 1$). Si T no depende de θ (donde θ es un parámetro a especificar en F_θ), entonces T recibe el nombre de estadístico.

Solo por su nombre, parece evidente que un estimador de un parámetro θ será alguna función de la muestra que sirva para aproximar, en algún sentido, el valor desconocido de θ . Si añadimos la condición razonable de que un estimador no pueda tomar valores que no puede tomar el parámetro, podemos dar la siguiente definición.

Definición 1.5 Un estimador de un parámetro θ es un estadístico T cuyo recorrido es el espacio de los parámetros, es decir:

$$T : \begin{matrix} \mathbb{R}^n & \longrightarrow \\ (x_1, x_2, \dots, x_n) & \longrightarrow \\ (t_1, \dots, t_k) & \in \Theta \subset \mathbb{R}^k \end{matrix}$$

Aquí tienes el texto traducido al castellano manteniendo toda la notación en LaTeX:

6.6 Distribución en el muestreo de un estadístico

Dado un estadístico $T(X_1, X_2, \dots, X_n)$ nos interesa conocer su distribución de probabilidad, ya que para hacer inferencia necesitaremos hacer cálculos del tipo

$$P[T(X_1, X_2, \dots, X_n) > t_0]$$

La distribución de probabilidad del estadístico se denomina distribución muestral o distribución en el muestreo del estadístico. Encontrarla es un problema que puede ser desde bastante sencillo hasta extremadamente complicado. Algunas de las técnicas utilizadas para intentar resolverlo son las siguientes:

- Uso de la técnica de cambio de variable.
- Uso de la función generadora de momentos.
- Aplicación del Teorema Central del Límite.

Ejemplo 1.5.1 Sea $X \sim F_\theta$ una variable aleatoria absolutamente continua con densidad

$$f_\theta(x) = e^{-(x-\theta)} e^{-e^{-(x-\theta)}} \quad \theta \in \mathbb{R}$$

y consideremos el estadístico

$$T(X_1, X_2, \dots, X_n) = \sum_{i=1}^n e^{-X_i}$$

Si aplicamos el teorema de cambio de variable unidimensional, se obtiene fácilmente que la variable aleatoria $Y = e^{-X}$ sigue una distribución exponencial de parámetro $e^{-\theta}$, de donde la suma seguirá una distribución gamma $T \sim \Gamma(e^{-\theta}, n)$.

Ejemplo 1.5.2 Supongamos que X representa el número de averías en una máquina al cabo de un mes. Este valor varía mes a mes. Sea \bar{X} la media de averías en n meses. Si X sigue una distribución de Poisson $P(\lambda)$, ¿cuál es la distribución de \bar{X} ? Como la suma de Poisson i.i.d. es $\sum_{i=1}^n X_i \sim P(n\lambda)$

$$P[\bar{X} = r] = P\left[\sum_{i=1}^n X_i = nr\right] = \frac{e^{-n\lambda} (n\lambda)^{nr}}{(nr)!}$$

Como ocurre en este ejemplo, uno de los estadísticos para el cual a menudo deseamos calcular la distribución en el muestreo es la media aritmética. Una manera útil de hacerlo es con la función generadora de momentos y la aplicación del siguiente lema.

Lema 1 Si X es una v.a. con $M_X(t)$ como función generadora de momentos, entonces la f.g.m. de $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ es

$$M_{\bar{X}_n}(t) = [M_X(t/n)]^n$$

6.6.1 Demostración:

La demostración es inmediata a partir de la definición o por las propiedades de la función generadora de momentos.

Si aplicamos directamente la definición de la f.g.m tenemos:

$$\begin{aligned} E\left(e^{t\bar{X}_n}\right) &= E\left(e^{t\frac{1}{n}\sum_{i=1}^n X_i}\right) = E\left(\prod_{i=1}^n e^{\frac{t}{n}X_i}\right) = \prod_{i=1}^n E\left(e^{\frac{t}{n}X_i}\right) \\ &= \prod_{i=1}^n M_{X_i}(t/n) = [M_X(t/n)]^n \end{aligned}$$

Si usamos las propiedades de la f.g.m tenemos:

1. Dado que $M_{aX}(t) = M_X(at)$ y si $a = \frac{1}{n}$, entonces $M_{\bar{X}}(t) = M_{\sum_{i=1}^n X_i}(t/n)$.
2. $M_{\sum_{i=1}^n X_i}(t/n) \stackrel{\text{ind}}{=} \prod_{i=1}^n M_{X_i}(t/n) \stackrel{\text{id}}{=} [M_X(t/n)]^n$.

Ejemplo 1.5.3 Para una variable aleatoria $X \sim N(\mu, \sigma)$ y por tanto $M_X(t) = \exp\left(t\mu + \frac{t^2\sigma^2}{2}\right)$, entonces

$$\begin{aligned} M_{\bar{X}_n}(t) &= \left[\exp\left(\frac{t\mu}{n} + \frac{t^2\sigma^2}{n^2 2}\right)\right]^n \\ &= \exp\left[n\left(\frac{t\mu}{n} + \frac{t^2\sigma^2}{n^2 2}\right)\right] \\ &= \exp\left[t\mu + \frac{1}{2}t^2\left(\frac{\sigma}{\sqrt{n}}\right)^2\right] \end{aligned}$$

que es la función generadora de momentos de una variable $N(\mu, \sigma/\sqrt{n})$.

6.7 La distribución empírica

6.7.1 Definición

En el apartado anterior hemos visto que a partir de una muestra X_1, X_2, \dots, X_n es interesante considerar la distribución muestral como la distribución conjunta del vector aleatorio (X_1, X_2, \dots, X_n) , sin que intervenga una realización concreta de la muestra x_1, x_2, \dots, x_n . Un enfoque diferente consiste en asociar una distribución particular directamente a las observaciones x_1, x_2, \dots, x_n con la intención de que, en tanto que la muestra “representa” la v.a. X , esta distribución asociada a la muestra $F_n(x)$ emule la distribución de la población. Esta distribución se denomina distribución empírica o distribución muestral y se define así:

$$F_n(x) = \frac{k(x)}{n}$$

donde $k(x)$ es el número de datos muestrales menores o iguales que x . En la práctica se construye por ordenación de la muestra

$$x_1, x_2, \dots, x_n \longrightarrow x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

y con la siguiente definición:

$$F_n(x) = \begin{cases} 0 & \text{si } x < x_{(1)} \\ \frac{k}{n} & \text{si } x_{(k)} \leq x < x_{(k+1)} \\ 1 & \text{si } x_{(n)} \leq x \end{cases}$$

Ejemplo 1.6.1 Extraemos una muestra y obtenemos:

x_1	x_2	x_3	x_4	x_5	x_6	x_7
5.1	3.4	1.2	17.6	2.1	16.4	4.3

Una vez ordenada queda:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$
x_3	x_5	x_2	x_7	x_1	x_6	x_4
1.2	2.1	3.4	4.3	5.1	16.4	17.6

y si hacemos la representación gráfica:

```
x <- c(5.1 , 3.4 , 1.2 , 17.6 , 2.1 , 16.4 , 4.3 , 1.2 , 2.1 , 3.4 , 4.3 , 5.1 , 16.4 , 17.6 )
plot(ecdf(x))
```

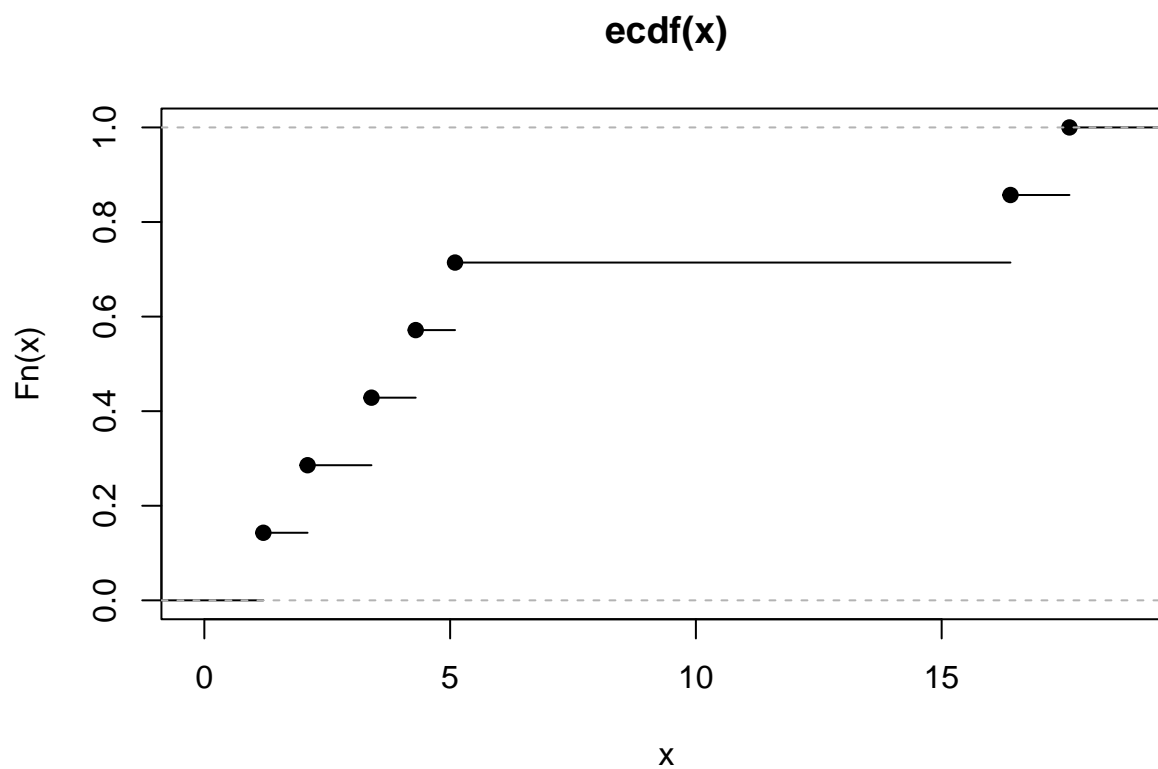


Figura 1.1: Función de distribución empírica con los datos del ejemplo

La distribución empírica refleja exclusivamente los valores observados en la muestra y, por lo tanto, no se relaciona directamente ni con la distribución conjunta de la muestra $G(x_1, x_2, \dots, x_n)$ ni con la distribución de la población F .

6.8 Los momentos muestrales

6.8.1 Definición

Sea F_n la v.a. que tiene $F_n(x)$ por distribución. La función de densidad de probabilidad de F_n es una densidad discreta que asigna probabilidades $1/n$ a cada una de las observaciones muestrales x_1, x_2, \dots, x_n . Así pues, tiene sentido calcular sus momentos, que se conocen como momentos muestrales a_k , y también sus momentos muestrales centrados respecto a la media b_k .

$$a_k = E(F_n^k) = \sum_{i=1}^n x_i^k \cdot P(F_n = x_i) = \sum_{i=1}^n x_i^k \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^n x_i^k$$
$$b_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$$

Observamos que dos medidas conocidas de la estadística descriptiva adquieren un significado diferente:

- Media muestral = Media de la distribución muestral

$$a_1 = \frac{1}{n} \sum_{i=1}^n x_i$$

- Varianza muestral = Varianza de la distribución muestral

$$b_2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

6.9 Distribución en el muestreo de los momentos muestrales

Dada una m.a.s. X_1, X_2, \dots, X_n , los momentos muestrales son estadísticos y, como tales, tienen su distribución en el muestreo. Por ejemplo, $a_k = \frac{1}{n} \sum_{i=1}^n X_i^k$.

La distribución en cada caso puede ser compleja y depender de la distribución poblacional subyacente. Lo que sí es posible calcular son los momentos de los momentos muestrales o, mejor dicho, los momentos de las distribuciones en el muestreo de los momentos muestrales.

1. Si consideramos $a_k = \frac{1}{n} \sum_{i=1}^n X_i^k$ y escribimos $\alpha_k = E(X^k)$ como el momento poblacional de orden k , tenemos:

$$E(a_k) = E\left(\frac{1}{n} \sum_{i=1}^n X_i^k\right) = \frac{1}{n} \cdot n \cdot \alpha_k = \alpha_k$$
$$\text{var}(a_k) = \text{var}\left(\frac{1}{n} \sum_{i=1}^n X_i^k\right) = \frac{1}{n^2} \sum_{i=1}^n \text{var}(X_i^k) = \frac{1}{n} \text{var}(X^k)$$
$$= \frac{1}{n} \left[E(X^{2k}) - (E(X^k))^2 \right] = \frac{\alpha_{2k} - \alpha_k^2}{n}$$

2. Si consideramos $s^2 = b_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2$, podemos calcular:

$$E(s^2) = \frac{1}{n} \sum_{i=1}^n E(X_i^2) - E(\bar{X})^2 = \alpha_2 - \left(\frac{\sigma^2}{n} + \mu^2\right)$$
$$= (\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right) = \frac{n-1}{n} \sigma^2$$

El cálculo de la varianza de s^2 es laborioso ¹ y no lo haremos aquí. Su valor es

$$\text{var}(s^2) = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}$$

donde μ_k es el momento poblacional centrado de orden k .

6.10 Propiedades asintóticas de los momentos muestrales

6.10.1 Convergencia de los momentos muestrales

Los momentos muestrales, tanto respecto al origen como respecto a la media, convergen hacia los momentos poblacionales. Es posible establecer la convergencia basándose en la ley fuerte de los grandes números (convergencia casi [^0]segura) o en la ley débil (convergencia en probabilidad). Si nos limitamos a esta última podemos afirmar que

$$a_k \xrightarrow{P} \alpha_k \quad \text{es decir} \quad \lim_{n \rightarrow \infty} P[|a_k - \alpha_k| \geq \epsilon] = 0$$

La prueba se basa en la desigualdad de Tchebychev. Si suponemos que $\alpha_{2k} < \infty$, tenemos

$$P[|a_k - \alpha_k| \geq \epsilon] \leq \frac{E|a_k - \alpha_k|^2}{\epsilon^2} = \frac{\text{var}(a_k)}{\epsilon^2} = \frac{\alpha_{2k} - \alpha_k^2}{n\epsilon^2} \rightarrow 0$$

Esta propiedad es importante porque hará posible el concepto de estimador consistente y en ella se basa un método de estimación llamado método de los momentos.

6.10.2 Distribución asintótica

Si consideramos el momento muestral $a_k = \frac{1}{n} \sum_{i=1}^n X_i^k$, entonces $n \cdot a_k$ es una suma de variables aleatorias i.i.d. a la que podemos aplicar el Teorema Central del Límite. Como hemos visto:

$$E(na_k) = n\alpha_k \quad \text{var}(na_k) = n^2 \text{var}(a_k) = n^2 \frac{\alpha_{2k} - \alpha_k^2}{n}$$

y por el Teorema Central del Límite de Lindeberg-Levy la variable

$$\frac{na_k - E(na_k)}{\sqrt{\text{var}(na_k)}} = \frac{na_k - n\alpha_k}{n\sqrt{\text{var}(a_k)}} = \frac{a_k - \alpha_k}{\sqrt{\text{var}(a_k)}}$$

verifica

$$\frac{a_k - \alpha_k}{\sqrt{\text{var}(a_k)}} \xrightarrow{\mathcal{L}} N(0, 1)$$

es decir

$$a_k \sim AN\left(\alpha_k, \sqrt{\frac{\alpha_{2k} - \alpha_k^2}{n}}\right)$$

6.11 Muestreo en poblaciones normales

Como hemos visto, a partir de una m.a.s. X_1, X_2, \dots, X_n y si consideramos un estadístico $T(X_1, X_2, \dots, X_n)$, puede resultar complicado obtener su distribución en el muestreo. Esta distribución depende de:

- La forma funcional de $T(X_1, X_2, \dots, X_n)$.
- La distribución subyacente de X , es decir, la distribución de la población.

Hay un caso especial en el que el problema se ha estudiado en profundidad para algunos estadísticos de gran importancia práctica. Si $X \sim N(\mu, \sigma)$ es posible encontrar la distribución de los estadísticos más utilizados como \bar{X} y $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2$. De hecho, obtendremos la distribución de funciones de estos estadísticos como

$$\frac{\bar{X} - \mu}{s/\sqrt{n-1}}; \quad \frac{ns^2}{\sigma^2}; \quad \bar{X}_1 - \bar{X}_2; \quad \frac{S_1^2/(n_1-1)}{S_2^2/(n_2-1)}$$

donde $s^2 = (1/n)S^2$. En el estudio de las distribuciones de estos estadísticos aparecen algunas distribuciones de probabilidad que han resultado ser de gran utilidad. Son las llamadas “distribuciones derivadas de la normal” y se conocen por el nombre del investigador que las formuló:

- la χ^2 chi-cuadrado de Pearson
- la t de Student (Gosset)
- la F de Fisher-Snedecor

6.11.1 La distribución chi-cuadrado

Sean X_1, X_2, \dots, X_k un conjunto de v.a. independientes sobre un mismo espacio de probabilidad (Ω, \mathcal{A}, P) y con distribución común $N(0, 1)$. Consideremos la variable

$$Y = X_1^2 + X_2^2 + \dots + X_k^2$$

La distribución de la variable Y se llama chi-cuadrado con k grados de libertad. La función de densidad de la variable aleatoria Y es

$$f(x) = \frac{1}{\Gamma(k/2)2^{k/2}} e^{-x/2} x^{k/2-1} \quad \text{si } x > 0$$

De modo que resulta que $Y = \sum_{i=1}^k X_i^2$ tiene una distribución gamma $G(\frac{1}{2}, \frac{k}{2})$ y su f.g.m. es

$$M(t) = (1 - 2t)^{-k/2} \quad \text{si } t < 1/2$$

6.11.1.1 Propiedades

1. Si recordamos que para $X \sim G(p, \alpha)$ entonces $E(X) = \frac{p}{\alpha}$ y $\text{var}(X) = \frac{p}{\alpha^2}$, resulta

$$E(Y) = \frac{k/2}{1/2} = k \quad \text{var}(Y) = \frac{k/2}{1/4} = 2k$$

2. De la aditividad (reproductividad) de las leyes gamma se deduce también la reproductividad de la chi-cuadrado χ^2 , es decir

$$Y_1^2 \sim \chi_{n_1}^2, Y_2^2 \sim \chi_{n_2}^2 \quad \text{indep.} \longrightarrow Y_1^2 + Y_2^2 \sim \chi_{n_1+n_2}^2$$

3. Como Y es la suma de v.a. independientes $X_i^2 \sim \chi_1^2$ se verifica

$$\frac{Y - k}{\sqrt{2k}} \xrightarrow{\mathcal{L}} N(0, 1)$$

Pero es mejor la aproximación de Fisher

$$\sqrt{2\chi_k^2} - \sqrt{2k - 1} \xrightarrow{\mathcal{L}} N(0, 1)$$

de donde se obtiene para valores de $k \geq 30$

$$\chi_k^2 \stackrel{\text{aprox}}{=} \frac{1}{2}(Z + \sqrt{2k - 1})^2$$

donde $Z \sim N(0, 1)$.

6.11.2 Distribución t de Student

Sean Y, Z dos variables aleatorias independientes con distribuciones $Z \sim N(0, 1)$ y $Y \sim \chi_m^2$, entonces se dice que la variable aleatoria

$$t = \frac{Z}{\sqrt{Y/m}}$$

tiene una distribución t de Student con m grados de libertad. Su función de densidad es

$$f(t) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\sqrt{m\pi}} \left(1 + \frac{t^2}{m}\right)^{-(m+1)/2} \quad t \in \mathbb{R}$$

Esta expresión se obtiene de la resolución del correspondiente problema de cambio de variable para encontrar la distribución de un cociente.

Se trata de una distribución unimodal y simétrica respecto al cero. La distribución depende de m , que llamamos los grados de libertad (g.l.). A medida que m crece, la forma acampanada se va “cerrando”, acercándose a la ley normal:

$$\left(1 + \frac{t^2}{m}\right)^{-(m+1)/2} \xrightarrow{m \rightarrow \infty} e^{-t^2/2}$$

Este hecho es muy relevante en inferencia estadística.

6.11.2.1 Propiedades

1. Si $m = 1$, entonces la t es una Cauchy y, en particular, no tiene esperanza.
2. Para $m > 1$, $E(t) = 0$ y para $m > 2$, $\text{var}(t) = m/(m - 2)$.
3. Cuando $m \rightarrow \infty$, entonces $t \xrightarrow{P} N(0, 1)$.

6.11.3 La distribución F de Fisher

Esta distribución aparece cuando se considera un cociente entre dos distribuciones chi-cuadrado $U \sim \chi_m^2, V \sim \chi_n^2$ con m y n g.l. respectivamente. En concreto decimos que la variable aleatoria

$$F = \frac{U/m}{V/n}$$

sigue una distribución F de Fisher con m y n grados de libertad. La función de densidad tiene la forma:

$$f(x) = \frac{m^{m/2} n^{n/2} \Gamma[(m+n)/2]}{\Gamma(m/2) \Gamma(n/2)} \cdot \frac{x^{m/2-1}}{(mx+n)^{(m+n)/2}} \quad \text{para } x > 0$$

6.11.3.1 Propiedades

1. La esperanza y la varianza son

$$E(F) = \frac{n}{n-2} \quad \text{var}(F) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$$

2. Esta distribución tiene una moda en $x = \frac{m-2}{m} \cdot \frac{n}{n+2}$, siempre que $m > 2$.
3. Si $F \sim F_{m,n}$, entonces resulta que $1/F \sim F_{n,m}$ y por lo tanto:

$$P(F \leq x) = P\left(\frac{1}{F} \geq \frac{1}{x}\right) = 1 - P\left(\frac{1}{F} \leq \frac{1}{x}\right)$$

Esta propiedad es de gran utilidad en el uso de las tablas. 4. Cuando $n \rightarrow \infty, F_{m,n} \xrightarrow{\mathcal{L}} \chi_m^2$. 5. Cuando $m \rightarrow \infty$ y $n \rightarrow \infty$, entonces $F_{m,n} \xrightarrow{\mathcal{L}} 1$.

7 Estimación puntual

7.1 El problema de la estimación puntual

Informalmente, la estimación de parámetros consiste en buscar aproximaciones a los valores de estos, calculables a partir de una muestra, que sean lo más precisas posible. El problema, claro, es que para medir cuán precisas son estas aproximaciones sería necesario conocer los valores de los parámetros y, como estos son siempre desconocidos, debemos basarnos en el uso de estimadores con buenas propiedades que, en algún sentido, nos garanticen esa proximidad. Más formalmente podemos plantear el problema de la siguiente manera: Sea X una v.a. con distribución F_θ donde $\theta = (\theta_1, \dots, \theta_k) \in \Theta \subset \mathbb{R}^k$ y sea X_1, X_2, \dots, X_n una muestra de n v.a. de X . El problema de la estimación puntual consiste en obtener alguna aproximación de θ en base a la información disponible en la muestra mediante un estimador de θ que definimos a continuación. Definición 2.1 Sea X_1, X_2, \dots, X_n una muestra aleatoria simple de X con distribución F_θ donde $\theta \in \Theta \subset \mathbb{R}^k$. Un estadístico $T(X_1, X_2, \dots, X_n)$ se denomina un estimador puntual de θ si T es una aplicación de \mathbb{R}^n en Θ , es decir, si toma valores sobre el mismo conjunto que los parámetros.

Ejemplo 2.1.1 Sea X_1, X_2, \dots, X_n una muestra aleatoria simple de una v.a. de Poisson $X \sim P(\lambda)$. Para estimar λ podemos utilizar:

$$T_1 = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$T_2 = s^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

ya que $E(X) = \text{var}(X) = \lambda$, pero también

$$T_3 = \frac{2}{n(n+1)} \sum_{i=1}^n X_i \cdot i$$

$$T_4 = X_i$$

Ejemplo 2.1.2 Sea X_1, X_2, \dots, X_n una m.a.s. de $X \sim B(1, p)$, con p desconocido. Podemos estimar p de las siguientes maneras:

$$T_1 = \bar{X} = (1/n) \sum_{i=1}^n X_i$$

$$T_2 = 1/2$$

$$T_3 = (X_1 + X_2) / 2$$

En cada caso resulta claro que algunos estimadores no son muy razonables mientras que la decisión entre los otros no está necesariamente clara. Básicamente debemos ocuparnos de dos problemas:

- Dado un modelo estadístico $\{X \sim F_\theta : \theta \in \Theta\}$, ¿cómo podemos obtener estimadores de θ que tengan “buenas” propiedades?
- Dado varios estimadores para un mismo parámetro ¿cómo podemos escoger el mejor en base a algún criterio?

Para poder alcanzar estos dos objetivos empezaremos por estudiar las propiedades de los estimadores, así como las medidas de optimalidad que podremos utilizar para decidir entre varios estimadores. De entrada nos restringiremos al caso en que $\Theta \subseteq \mathbb{R}$ o en que queremos aproximar alguna función $g(\theta)$ de los parámetros donde g es del tipo $g : \Theta \rightarrow \mathbb{R}$.

7.1.1 Criterios de optimalidad de estimadores. El Riesgo

Una forma de poder comparar entre diversos estimadores consiste en definir una función de pérdida que nos permita cuantificar de alguna manera la pérdida, o coste asociado, al estimar el valor real del parámetro, es decir, θ , mediante la aproximación que proporciona un estimador, es decir, t .

Definición 2.2 Una función de pérdida es una aplicación

$$L : \Theta \times \Theta \rightarrow \mathbb{R}$$

$$(\theta, t) \rightarrow L(\theta, t)$$

que verifica: a) $L(\theta, t) \geq 0$, $\forall \theta, t \in \Theta$ b) $L(\theta, t) = 0$, si $\theta = t$ c) $L(\theta, t) \leq L(\theta, t')$, si $d(\theta, t) \leq d(\theta, t')$ donde d es una distancia en Θ .

Por ejemplo, son funciones de pérdida:

$$L_1(\theta, t) = |\theta - t| \quad L_2(\theta, t) = (\theta - t)^2$$

$$L_3(\theta, t) = \left| \frac{\theta - t}{\theta} \right| \quad L_4(\theta, t) = \left(\frac{\theta - t}{\theta} \right)^2$$

$$L_5(\theta, t) = \begin{cases} c > 0 & \text{si } |\theta - t| > \epsilon \\ 0 & \text{si } |\theta - t| \leq \epsilon \end{cases}$$

7.1.2 El error cuadrático medio

Una de las funciones de pérdida más usuales es la función de pérdida cuadrática $L_2(\theta, t) = (\theta - t)^2$. Uno de los motivos de su uso es que el riesgo asociado a esta función de pérdida $E_\theta [(\theta - T)^2]$, que llamamos error cuadrático medio EQM_T , representa una medida de la variabilidad del estimador T en torno a θ semejante a la medida de dispersión en torno a la media que representa la varianza. Además, del desarrollo de esta expresión se obtiene un interesante resultado que muestra cuáles pueden ser las propiedades más interesantes para un estimador. Sea $\{X \sim F_\theta : \theta \in \Theta\}$ y sea T un estimador de θ . El error cuadrático medio de T para estimar θ vale

$$EQM_T(\theta) = E_\theta [(\theta - T)^2] = E [\theta^2 - 2\theta T + T^2] = \theta^2 - 2\theta E_\theta(T) + E_\theta(T^2)$$

Ahora, sumando y restando $(E_\theta(T))^2$, obtenemos

$$\begin{aligned} EQM_T(\theta) &= E_\theta(T^2) - (E_\theta(T))^2 + (E_\theta(T))^2 + \theta^2 - 2\theta E_\theta(T) = \\ &= \text{var}(T) + (E_\theta(T) - \theta)^2 \end{aligned}$$

El término $(E_\theta(T) - \theta)^2$ es el cuadrado del sesgo de T , que se define como

$$b_\theta(T) = E_\theta(T) - \theta$$

Definición 2.5 El error cuadrático medio $EQM_T(\theta)$, o simplemente EQM , de un estimador T para estimar el parámetro θ es la suma de su varianza más el cuadrado de la diferencia entre su valor medio y el verdadero valor del parámetro, que llamamos sesgo.

Si en la búsqueda de estimadores de mínimo riesgo nos basamos en la función de pérdida cuadrática, parece que los estimadores más deseables deberían ser aquellos en los que la varianza y el sesgo sean lo más pequeños posibles. Idealmente, quisiéramos reducir ambas cantidades a la vez. En la práctica, sin embargo, observamos que, en general, no suele ser posible reducir simultáneamente la varianza y el sesgo. Además, incluso si fuera práctico calcular el EQM para cada estimador, encontraríamos que, para la mayoría de las familias de probabilidad P_θ , no existiría ningún estimador que minimizase el EQM para todos los valores de θ . Es decir, que un estimador puede tener un EQM mínimo para algunos valores de θ , mientras que otro lo tendrá en otros valores de θ .

Ejemplo 2.1.4 Sea X_1, X_2, \dots, X_n una muestra aleatoria simple de $X \sim N(\mu, \sigma)$, donde suponemos σ conocida, y sean

$$T_1 = \bar{X} \quad T_2 = \frac{\sum_{i=1}^n X_i}{n+1}$$

Calculando la media y la varianza de los estimadores, tenemos

$$\begin{aligned} E_\mu(T_1) &= \mu & \Rightarrow b_{T_1}(\mu) &= 0 & \text{var}_\mu(T_1) &= \frac{\sigma^2}{n} \\ E_\mu(T_2) &= \frac{n}{n+1}\mu & \Rightarrow b_{T_2}(\mu) &= \frac{-1}{n+1}\mu & \text{var}_\mu(T_2) &= \frac{n}{(n+1)^2}\sigma^2 \end{aligned}$$

de donde

$$\begin{aligned} EQM_\mu(T_1) &= \text{var}(T_1) = \frac{\sigma^2}{n} \\ EQM_\mu(T_2) &= \frac{1}{(n+1)^2}\mu^2 + \frac{n}{(n+1)^2}\sigma^2 \end{aligned}$$

que son respectivamente una recta y una parábola. De manera que para algunos valores de μ tenemos que $EQM_\mu(T_1) < EQM_\mu(T_2)$ y para otros, al revés. La figura 2.1 muestra esta diferencia.

Ejemplo 2.1.5 Un ejemplo trivial bastante interesante es el siguiente. Para estimar un parámetro θ , el estimador que consiste en un valor fijo θ_0 , tiene riesgo 0 en $\theta = \theta_0$. Sin embargo, el riesgo aumenta considerablemente al alejarnos del valor real de θ . Por lo tanto, no resulta un estimador razonable, aunque su riesgo pueda ser mínimo para algún (único) valor de θ .

Figura 2.1: Comparación del riesgo de dos estimadores

Los ejemplos anteriores nos muestran que los criterios de preferencia entre estimadores basados en el riesgo o en el EQM no son de gran utilidad general ya que muchos estimadores pueden ser incomparables. Ante este hecho nos planteamos si es posible completar el criterio de minimizar el riesgo mediante alguna propiedad o criterio adicional. Las posibles soluciones obtenidas a esta cuestión siguen dos vías:

1. Restringir la clase de estimadores considerados a aquellos que cumplan alguna propiedad adicional de interés, eliminando estimadores indeseables para que el criterio de minimizar el riesgo permita seleccionar uno preferible a los demás. Este criterio lleva a considerar las propiedades deseables de los estimadores como falta de sesgo, consistencia, eficiencia y analizar cómo combinarlas con el criterio de mínimo riesgo. Este proceso culmina con el estudio de los Estimadores Sin Sesgo Uniformemente de Mínima Varianza (ESUMV).
2. Reforzar el criterio de preferencia de estimadores mediante la reducción de toda la función de riesgo $R_T(\theta)$ a un único valor representativo que permita ordenar linealmente todos los estimadores. Este criterio nos lleva a los Estimadores Bayes y a los Estimadores Minimax.

7.2 Estudio de las propiedades deseables de los estimadores

7.2.1 El sesgo

Supongamos que tenemos un modelo estadístico $\{X \sim F_\theta : \theta \in \Theta\}$ y un estimador $T(X_1, X_2, \dots, X_n)$ de una función medible $g(\theta)$ del parámetro. Una forma razonable de valorar qué tan próximos son los valores de T a los de $g(\theta)$ es ver si, en promedio, los valores de T coinciden con el valor medio de $g(\theta)$.

Definición 2.6 Bajo las condiciones mencionadas, si $E_\theta(T)$ es la esperanza de $T(X_1, X_2, \dots, X_n)$ y $g(\theta)$ es una función del parámetro (en particular la identidad), la diferencia

$$b_T(\theta) = b_T(\theta) = E_\theta(T) - g(\theta)$$

se denomina sesgo del estimador T para estimar $g(\theta)$. Si el sesgo es nulo, es decir, si:

$$E_\theta(T) = g(\theta), \quad \forall \theta \in \Theta$$

diremos que T es un estimador insesgado de $g(\theta)$. **Ejemplo 2.2.1** Los dos ejemplos más conocidos son el de la media y la varianza muestrales.

- La media muestral es un estimador insesgado de μ .
- La varianza muestral es un estimador con sesgo de la varianza poblacional. En concreto, su sesgo vale:

$$b_{s^2}(\sigma^2) = E_{\sigma^2}(s^2) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = \frac{-1}{n}\sigma^2$$

El uso de estimadores insesgados es conveniente en muestras de tamaño grande. En estas, $\text{var}_\theta(T)$ es a menudo pequeña y entonces, como $E_\theta(T) = g(\theta) + b_T(\theta)$, es muy probable obtener estimaciones centradas en este valor en lugar de en el entorno de $g(\theta)$.

Ejemplo 2.2.2 Sea X_1, X_2, \dots, X_n una muestra aleatoria simple de $X \sim U(0, \theta)$. Tomemos $T = \max\{X_1, X_2, \dots, X_n\}$ como el estimador del máximo de la distribución. Obviamente podemos decir que

$T < \theta$ y, por lo tanto, la estimación siempre está sesgada. Como hemos visto en el ejemplo ??, la distribución en el muestreo de T es

$$H_{\theta}(t) = P_{\theta}[T \leq t] = \left(\frac{t}{\theta}\right)^n$$

y su función de densidad es

$$f_{\theta}(\theta) = H'_{\theta}(\theta) = \frac{n}{\theta} \left(\frac{t}{\theta}\right)^{n-1}$$

Su esperanza (ver ejemplo ??) vale

$$E_{\theta}(T) = \int_0^{\theta} t \cdot \left[\frac{n}{\theta} \left(\frac{t}{\theta}\right)^{n-1} \right] dt = \frac{n}{n+1} \theta$$

de donde el sesgo de T para estimar θ es

$$b_T(\theta) = \frac{n}{n+1} \theta - \theta = -\frac{1}{n+1} \theta$$

Podemos preguntarnos si podríamos mejorar este estimador corrigiendo el sesgo de forma análoga a lo que hacíamos con \hat{s}^2 , es decir, tomando un estimador corregido para el sesgo

$$T' = \frac{n+1}{n} T \text{ que, por construcción, verifica: } E(T') = \theta.$$

Consideremos el estimador de mínimo riesgo en el sentido del error cuadrático medio, es decir, el estimador que minimiza $E[(\theta - T)^2]$. De hecho, como hemos visto en el ejemplo ??, conviene elegir el que minimice $E[(\theta - T)^2/\theta^2]$, porque también minimiza el EQM, pero alcanza un mínimo absoluto. Este estimador es

$$T'' = \frac{n+2}{n+1} T$$

y, por tanto, es más adecuado que T' , ya que tiene un menor riesgo respecto al error cuadrático medio. Cuando, como aquí, nos encontramos con que dado un estimador podemos encontrar otro de menor riesgo, decimos que el primero no es admisible respecto de la función de pérdida. En este caso decimos que T' no es admisible respecto al EQM. ¡Cuidado! Esto no significa que no podamos usarlo, sino que existe otro con menor riesgo, ya que existe otro T'' preferible a él que, por cierto, no es centrado. Efectivamente

$$E_{\theta}(T'') = \frac{n+2}{n+1} E_{\theta}(T) = \frac{(n+2)n}{(n+1)^2} \theta$$

El ejemplo anterior muestra que, debido a la descomposición $EQM_T(\theta) = \text{var}_{\theta}(T) + b_T^2(\theta)$, puede ser preferible un estimador con sesgo a otro que no lo tenga. En general, sin embargo, eliminar el sesgo no es una mala estrategia, sobre todo porque al restringirnos a la clase de los estimadores insesgados obtenemos una solución constructiva que permitirá obtener estimadores insesgados de mínima varianza en condiciones bastante generales. Los siguientes ejemplos ilustran dos propiedades interesantes del sesgo. Por un lado, muestran que no siempre existe un estimador insesgado. Por otro lado, vemos cómo a veces, incluso teniendo un estimador insesgado para un parámetro $E_{\theta}(T) = \theta$, una función $g(T)$ no es necesariamente un estimador insesgado de $g(\theta)$.

Ejemplo 2.2.3 Consideremos una variable X con distribución de Bernoulli $B(1, p)$. Supongamos que deseamos estimar $g(p) = p^2$ con una única observación. Para que un estimador T no tenga sesgo para estimar p^2 sería necesario que

$$p^2 = E_p(T) = p \cdot T(1) + (1 - p) \cdot T(0), \quad 0 \leq p \leq 1$$

es decir, para cualquier valor de $p \in [0, 1]$ se debería verificar

$$p^2 = p \cdot (T(1) - T(0)) + T(0)$$

Esto claramente no es posible, ya que la única forma en que una función lineal y una función parabólica coincidan en todo el intervalo $[0, 1]$ es cuando los coeficientes $T(0)$ y $T(1)$ valen cero.

Ejemplo 2.2.4 El parámetro α de una ley exponencial con función de densidad

$$f(x) = \alpha e^{-\alpha x} \mathbf{1}_{(0, \infty)}(x)$$

es el inverso de la media de la distribución, es decir, $\alpha = 1/E(X)$. Un estimador razonable de $\alpha = g(\mu)$ puede ser $\hat{\alpha} = g(\hat{\mu})$, es decir, $\hat{\alpha} = 1/\bar{X}$. Si aplicamos la propiedad de que la suma de variables aleatorias i.i.d. exponenciales sigue una ley gamma de parámetros n y α , se obtiene que este estimador tiene sesgo. Su esperanza es

$$E(\hat{\alpha}) = \frac{n}{n-1} \alpha$$

El sesgo se corrige simplemente con

$$\hat{\alpha}' = \frac{n-1}{n} \hat{\alpha}$$

7.2.2 Consistencia

La consistencia de un estimador es una propiedad bastante intuitiva que indica, de manera informal, que cuando aumenta el tamaño muestral, el valor del estimador se aproxima cada vez más al verdadero valor del parámetro.

Definición 2.7 Sea $X_1, X_2, \dots, X_n, \dots$ una sucesión de variables aleatorias i.i.d. $X \sim F_\theta, \theta \in \Theta$. Una sucesión de estimadores puntuales $T_n = T(X_1, X_2, \dots, X_n)$ se denomina consistente para $g(\theta)$ si

$$T_n \xrightarrow[n \rightarrow \infty]{P} g(\theta)$$

para cada $\theta \in \Theta$, es decir, si

$$\forall \varepsilon > 0 \quad \lim_{n \rightarrow \infty} P\{|T_n - g(\theta)| > \varepsilon\} = 0$$

Observemos que:

1. Se trata de un concepto asintótico: Hablamos de sucesiones de estimadores consistentes? más que de estimadores propiamente dichos.
2. La definición puede reforzarse si, en lugar de considerar convergencia en probabilidad (consistencia débil), consideramos convergencia casi segura o en media cuadrática:
 - T_n es fuertemente consistente si $T_n \xrightarrow{\text{c.s.}} g(\theta)$
 - T_n es consistente en media- r si $E_\theta[|T_n - g(\theta)|^r] \rightarrow 0$

Ejemplo 2.2.5 Muchos estimadores consistentes lo son como consecuencia de las leyes de los grandes números. Recordemos que la Ley débil de los Grandes Números (Tchebychev) afirma que, dada una sucesión de v.a. independientes e idénticamente distribuidas con medias $\mu < \infty$ y varianzas $\sigma^2 < \infty$, entonces

$$\bar{X}_n \xrightarrow{P} \mu$$

Como consecuencia de esta ley y dado que una muestra aleatoria simple es i.i.d., por definición, podemos afirmar que \bar{X}_n es consistente para estimar μ .

Ejemplo 2.2.6 La sucesión $T_n = \max_{1 \leq i \leq n} \{X_i\}$ es consistente para estimar el máximo de una distribución uniforme en $[0, \theta]$:

$$P \left[\left| \max_{1 \leq i \leq n} \{X_i\} - \theta \right| > \varepsilon \right] = P \left[\theta - \max_{1 \leq i \leq n} \{X_i\} > \varepsilon \right]$$

ya que $X_i \in [0, \theta]$, por lo tanto, podemos escribir:

$$\begin{aligned} P \left[\theta - \varepsilon > \max_{1 \leq i \leq n} \{X_i\} \right] &= P \left[\max_{1 \leq i \leq n} \{X_i\} < \theta - \varepsilon \right] \\ &= \left(\frac{\theta - \varepsilon}{\theta} \right)^n = \left(1 - \frac{\varepsilon}{\theta} \right)^n \xrightarrow{n \rightarrow \infty} 0 \end{aligned}$$

Es inmediato comprobar que

$$E \left[(\theta - T_n)^2 \right] = \left(1 - \frac{2n}{n+1} + \frac{n}{n+2} \right) \theta^2$$

que también tiende a cero cuando $n \rightarrow \infty$, y por lo tanto $T_n = \max_{1 \leq i \leq n} \{X_i\}$ también es consistente en media cuadrática.

Normalmente, cuando se habla de consistencia, se hace referencia a la convergencia en probabilidad, es decir, T_n es consistente si $\lim_{n \rightarrow \infty} P(|T_n - g(\theta)| > \varepsilon) = 0$. Si el estimador no tiene sesgo, estamos en la situación de aplicar la desigualdad de Tchebychev ¹ : Si $E(T_n) = g(\theta)$, entonces

$$P(|T_n - g(\theta)| > \varepsilon) = P(|T_n - E(T_n)| > \varepsilon) \underset{\text{Tchebychev}}{\leq} \frac{\text{var}(T_n)}{\varepsilon^2}$$

Así, para intentar establecer la consistencia de T , debemos probar que

$$\frac{\text{var}(T_n)}{\varepsilon^2} \xrightarrow{n \rightarrow \infty} 0$$

Ejemplo 2.2.7 Sea $M_n = \sum_{i=1}^n a_i X_i$ una combinación lineal de los valores de la muestra con coeficientes tales que $\sum_{i=1}^n a_i = 1$ y algún $a_i > 0$. ¿Es consistente M_n para estimar $E(X)$? Comencemos por ver que M_n no tiene sesgo

$$\begin{aligned} E(M_n) &= E \left(\sum_{i=1}^n a_i X_i \right) = \sum_{i=1}^n E(a_i X_i) \\ &= \sum_{i=1}^n a_i E(X_i) \stackrel{\text{i.i.d.}}{=} \sum_{i=1}^n a_i E(X) = E(X) \end{aligned}$$

[^1] Calculemos la varianza

$$\begin{aligned}\text{var}(M_n) &= \text{var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n \text{var}(a_i X_i) \\ &= \sum_{i=1}^n a_i^2 \text{var}(X_i) = \text{var}(X) \sum_{i=1}^n a_i^2\end{aligned}$$

Si aplicamos ahora la desigualdad de Tchebychev tenemos:

$$P(|M_n - \mu| > \varepsilon) \leq \frac{\sigma^2 \sum a_i^2}{\varepsilon^2}$$

lo cual no tiene por qué tender a 0 cuando $n \rightarrow \infty$, y por lo tanto no podemos afirmar que el estimador es consistente. Por ejemplo, si $a_1 = \frac{1}{2}, a_2 = a_3 = \dots = a_n = \frac{1}{2(n-1)}$ tendremos que $\lim_{n \rightarrow \infty} \sum a_i^2 = \frac{1}{4}$. Observamos que el resultado obtenido no puede asegurar la consistencia de M_n para cualquier familia de coeficientes a_1, \dots, a_n , aunque, obviamente, el estimador es consistente para alguno (caso $a_i = 1/n$).

7.3 Propiedades de los estimadores consistentes

Muchas de las propiedades de los estimadores son consecuencia directa de las propiedades de la convergencia en probabilidad, que se pueden revisar, por ejemplo, en Martin Pliego (1998a) capítulo 11.

1. Si T_n es consistente para estimar θ y $g: \mathbb{R} \rightarrow \mathbb{R}$ es una función continua, entonces $g(T_n)$ es consistente para estimar $g(\theta)$.
2. Si T_{1n} y T_{2n} son consistentes para estimar θ_1 y θ_2 respectivamente, entonces $aT_{1n} \pm bT_{2n}$ es consistente para estimar $a\theta_1 \pm b\theta_2$. $T_{1n} \cdot T_{2n}$ es consistente para estimar $\theta_1 \cdot \theta_2$. T_{1n}/T_{2n} es consistente para estimar θ_1/θ_2 , si $\theta_2 \neq 0$.
3. Sea $a_r = (1/n) \sum X_i^r$ el momento muestral de orden r . Como se ha visto en el capítulo 1, la esperanza de a_r es

$$E(a_r) = E\left[\frac{1}{n} \sum X_i^r\right] = \frac{1}{n} \sum E(X^r) = \frac{1}{n} n \alpha_r = \alpha_r$$

donde α_r es el momento poblacional de orden r . Así pues, a_r no tiene sesgo para estimar α_r . Su varianza es

$$\begin{aligned}\text{var}(a_r) &= \text{var}\left(\frac{1}{n} \sum X_i^r\right) = \frac{1}{n^2} \sum \text{var}(X^r) = \frac{1}{n} E[X^r - E(X^r)]^2 \\ &= \frac{1}{n} E[X^r - \alpha_r]^2 = \frac{1}{n} E(X^{2r} + \alpha_r^2 - 2\alpha_r X^r) \\ &= \frac{1}{n} (\alpha_{2r} - \alpha_r^2).\end{aligned}$$

Y si aplicamos la desigualdad de Tchebychev, se obtiene

$$P(|a_r - \alpha_r| \geq \varepsilon) \leq \frac{E(a_r - \alpha_r)^2}{\varepsilon^2} = \frac{\text{var}(a_r)}{\varepsilon^2} = \frac{\alpha_{2r} - \alpha_r^2}{n\varepsilon^2} \xrightarrow{n \rightarrow \infty} 0$$

Así pues, hemos visto que los momentos muestrales son estimadores consistentes de los momentos poblacionales.

7.3.1 Eficiencia

Como ya hemos visto, un objetivo deseable en la búsqueda de estimadores óptimos es considerar estimadores de “mínimo riesgo” o, si nos basamos en la función de pérdida cuadrática, estimadores que minimicen el error cuadrático medio $E(\theta - T)^2$. En general, es difícil encontrar estimadores que hagan mínimo el EQM para todos los valores de θ ; sin embargo, si nos restringimos a los estimadores sin sesgo, el problema tiene solución en una gama más amplia de situaciones. Supongamos que T_1, T_2 son dos estimadores sin sesgo de un parámetro θ . Para estos estimadores tenemos que

$$\begin{aligned}EQM_{T_1}(\theta) &= \text{var}_\theta(T_1) + b_{T_1}^2(\theta) \\EQM_{T_2}(\theta) &= \text{var}_\theta(T_2) + b_{T_2}^2(\theta)\end{aligned}$$

Si los estimadores no tienen sesgo $b_{T_1}(\theta) = b_{T_2}(\theta) = 0$, el que tenga menor varianza tendrá el menor riesgo para estimar θ . Si, por ejemplo, $\text{var}(T_1) \leq \text{var}(T_2)$, diremos que T_1 es más eficiente que T_2 para estimar θ . Para dos estimadores con sesgo cero $b_{T_i}(\theta) = 0$, el cociente

$$ER = \frac{EQM_{T_1}(\theta)}{EQM_{T_2}(\theta)} = \frac{\text{var}_\theta(T_1)}{\text{var}_\theta(T_2)}$$

se denomina eficiencia relativa de T_1 respecto a T_2 . Si solo hay dos estimadores de θ puede ser fácil ver cuál es el más eficiente. Si hay más, la cosa se complica. El “más eficiente”, en caso de que exista, se llamará el estimador sin sesgo de mínima varianza.

Figura 2.2: Comparación de la eficiencia de dos estimadores para un θ dado

Definición 2.8 Sea $\mathcal{S}(\theta)$ la clase de los estimadores sin sesgo de θ y con varianza. Si para todos los estimadores de esta clase $T \in \mathcal{S}(\theta)$ se verifica que

$$\text{var}_\theta(T) \leq \text{var}_\theta(T^*) \quad \forall T \in \mathcal{S}(\theta)$$

diremos que T^* es un estimador sin sesgo de mínima varianza de θ . Si la desigualdad es cierta $\forall \theta \in \Theta$, diremos que T^* es un estimador sin sesgo uniforme de mínima varianza (ESUMV) ².

7.4 Información de Fisher y cota de CramerRao

Obviamente, en un problema de estimación lo ideal es disponer de un ESUMV, pero esto no siempre es posible. Nos enfrentamos a varios problemas:

1. ¿Existen ESUMV para un parámetro θ en un modelo dado?
2. En caso de que exista el ESUMV, ¿sabremos cómo encontrarlo?

Este problema tiene solución, bajo ciertas condiciones, utilizando los teoremas de Lehmann-Scheffé y Rao-Blackwell y el concepto de suficiencia, que se discute más adelante.

[^2]Una solución parcial aparece gracias al Teorema de Cramer-Rao, que permite establecer una cota mínima para la varianza de un estimador. Cuando un estimador alcanza esta cota, sabemos que es un estimador de varianza mínima. Informalmente, este resultado sugiere que, bajo ciertas condiciones de regularidad, si T es un estimador insesgado de un parámetro θ , su varianza está acotada por una expresión que llamamos cota de Cramer-Rao $CCR(\theta)$

$$\text{var}(T) \geq CCR(\theta)$$

Antes de establecer con precisión este teorema, consideremos el concepto de información de un modelo estadístico introducido por Fisher.

7.5 Información y verosimilitud de un modelo estadístico

Una idea bastante razonable es esperar que un estimador funcione mejor en su intento de aproximarse al valor de un parámetro cuanto más información tenga para hacerlo. Por este motivo, la varianza del estimador y la información se presentan como cantidades opuestas: a mayor información, menor error (varianza) en la estimación:

$$\text{var}(T_n) \propto \frac{1}{I_n(\theta)}$$

Ahora nos encontramos con el problema de cómo definir la cantidad de información (contenida en una muestra/de un modelo), para que se ajuste a la idea intuitiva de información. Fisher lo hizo a través de la función de verosimilitud. Sea un modelo estadístico $\{X \sim F_\theta : \theta \in \Theta\}$ y una m.a.s. (X_1, X_2, \dots, X_n) , que toma valores $\mathbf{x} = (x_1, x_2, \dots, x_n)$. Si X es discreta, la función de masa de probabilidad indica, en términos generales, la probabilidad de observar la muestra, dado un valor del parámetro. Si X es absolutamente continua, esta interpretación ya no es tan directa.

$$f(x_1, x_2, \dots, x_n; \theta) = \begin{cases} P_\theta[X = x_1] \cdots P_\theta[X = x_n], & \text{si } X \text{ es discreta} \\ f_\theta(x_1) \cdots f_\theta(x_n), & \text{si } X \text{ es abs. continua} \end{cases}$$

La función de verosimilitud se obtiene si consideramos, en la expresión anterior, que lo que queda fijado es la muestra y no el parámetro. Es decir, fijada una muestra \mathbf{x} , la función de verosimilitud indica qué tan verosímil resulta, para cada valor del parámetro, que el modelo la haya generado.

Ejemplo 2.3.1 Supongamos que tenemos una m.a.s. x_1, x_2, \dots, x_n de tamaño n de una variable aleatoria X , que sigue una ley de Poisson de parámetro λ desconocido.

$$X \sim F_\lambda = P(\lambda), \quad \lambda > 0$$

La función de probabilidad de la muestra, fijado λ , es:

$$g_\lambda(x_1, x_2, \dots, x_n) = \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum x_i}}{\prod_{i=1}^n x_i!}$$

y la función de verosimilitud del modelo, fijada \mathbf{x} , es:

$$L(x_1, x_2, \dots, x_n; \lambda) = \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum x_i}}{\prod_{i=1}^n x_i!}$$

Aunque la forma funcional de $g_\lambda(\mathbf{x})$ y $L(\mathbf{x}; \lambda)$ es la misma, su aspecto es diferente, como se puede comprobar en la figura 2.3, donde damos valores a $g_\lambda(\mathbf{x})$, variando \mathbf{x} o a $L(\lambda; \mathbf{x})$ variando λ .

7.6 Información de Fisher

Para calcular la cantidad de información de Fisher contenida en una muestra sobre un parámetro, es necesario considerar modelos estadísticos regulares, es decir, donde se cumplen las siguientes condiciones de regularidad.

Definición 2.9 Diremos que $\{X \sim F_\theta : \theta \in \Theta\}$ es un modelo estadístico regular si se verifican las siguientes condiciones:

1. La población de donde proviene la muestra presenta un ?campo de variación? o soporte $S_\theta = \{x \mid f(x; \theta) > 0\} = S$ que no depende de θ .
2. La función $L(\mathbf{x}; \theta)$ admite, al menos, las dos primeras derivadas.

3. Las operaciones de derivación e integración son intercambiables.

Definición 2.10 Sea $\{X \sim F_\theta : \theta \in \Theta\}$ un modelo estadístico regular, es decir, donde se verifican las condiciones de regularidad 1-3 anteriores. Si $Z = \frac{\partial}{\partial \theta} \log L(\mathbf{X}; \theta)$, la cantidad de información de Fisher es

$$I_n(\theta) = \text{var}_\theta(Z) = \text{var}_\theta \left(\frac{\partial}{\partial \theta} \log L(\mathbf{X}; \theta) \right)$$

Figura 2.3: Probabilidad de la suma de $n = 5$ valores muestrales para 10 muestras de la ley de Poisson con $\lambda = 3$ versus la función de verosimilitud para una muestra observada.

Las condiciones de regularidad son necesarias para calcular $E_\theta(Z^2)$. A continuación, presentamos algunas propiedades de la información de Fisher. Puedes ver la demostración en Ruiz-Maya y Pliego (1995).

1. La información de Fisher se puede expresar como:

$$I_n(\theta) = E_\theta \left[\left(\frac{\partial \log L(\mathbf{X}; \theta)}{\partial \theta} \right)^2 \right]$$

Esto se puede comprobar, ya que si aplicamos las condiciones de regularidad

$$\begin{aligned} E(Z) &= E \left(\frac{\partial \log L(\mathbf{X}; \theta)}{\partial \theta} \right) = \int_{S^n} \frac{\partial \log L(\mathbf{x}; \theta)}{\partial \theta} L(\mathbf{x}; \theta) d\mathbf{x} \\ &= \int_{S^n} \frac{\frac{\partial L(\mathbf{x}; \theta)}{\partial \theta}}{L(\mathbf{x}; \theta)} L(\mathbf{x}; \theta) d\mathbf{x} = \int_{S^n} \frac{\partial L(\mathbf{x}; \theta)}{\partial \theta} d\mathbf{x} \\ &= \frac{\partial}{\partial \theta} \left(\int_{S^n} L(\mathbf{x}; \theta) d\mathbf{x} \right) = \frac{\partial}{\partial \theta} 1 = 0 \end{aligned}$$

De forma que $E(Z) = 0$, y por lo tanto, tendremos que $\text{var}_\theta(Z) = E_\theta(Z^2)$. 2. $I_n(\theta) = 0$ si y solo si $L(\mathbf{x}; \theta)$ no depende de θ . 3. Dadas dos m.a.s. $\mathbf{x}_1, \mathbf{x}_2$ de tamaños n_1, n_2 de la misma población, se verifica:

$$I_{n_1, n_2}(\theta) = I_{n_1}(\theta) + I_{n_2}(\theta)$$

De manera que podemos considerar una muestra de tamaño n como n muestras de tamaño 1 :

$$I_n(\theta) = \sum_{i=1}^n I_1(\theta) = n \cdot i(\theta), \text{ siendo } i(\theta) = I_1(\theta)$$

Es decir

$$E \left(\frac{\partial \log(L(\mathbf{X}; \theta))}{\partial \theta} \right) = n E \left(\frac{\partial \log f(X; \theta)}{\partial \theta} \right)$$

4. Se verifica la siguiente relación:

$$I_n(\theta) = E \left[\left(\frac{\partial \log L(\mathbf{X}; \theta)}{\partial \theta} \right)^2 \right] = -E \left[\frac{\partial^2 \log L(\mathbf{X}; \theta)}{\partial^2 \theta} \right]$$

Ejemplo 2.3.2 Vamos a calcular la cantidad de información de Fisher contenida en una m.a.s. extraída de una población $N(\mu, \sigma)$ con $\sigma = \sigma_0$ conocida. La función de verosimilitud es

$$L(\mathbf{x}; \mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(x_i - \mu)^2}{2\sigma_0^2}} = (2\pi\sigma_0^2)^{-n/2} \exp\left(-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma_0^2}\right)$$

y su logaritmo

$$\log L(\mathbf{x}; \mu) = -\frac{n}{2} \log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu)^2$$

Si derivamos respecto a μ

$$\frac{\partial \log L(\mathbf{x}; \mu)}{\mu} = \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma_0^2}$$

de donde

$$\begin{aligned} I_n(\mu) &= E\left(\frac{\partial \log L(\mathbf{X}; \mu)}{\partial \mu}\right)^2 = E\left(\frac{\sum_{i=1}^n (X_i - \mu)}{\sigma_0^2}\right)^2 \\ &= \frac{1}{\sigma_0^4} E\left[\sum_{i=1}^n (X_i - \mu)^2 + \sum_{i \neq j} (X_i - \mu)(X_j - \mu)\right] \\ &= \frac{1}{\sigma_0^4} n\sigma_0^2 = \frac{n}{\sigma_0^2} \end{aligned}$$

Este cálculo también puede hacerse a partir de la tercera propiedad de la información de Fisher:

$$I_n(\mu) = nE\left[\frac{\partial \log f(X; \mu)}{\partial \mu}\right] = n\frac{1}{\sigma_0^2} = \frac{n}{\sigma_0^2}$$

7.7 La desigualdad de Cramer-Rao

Una vez establecidas las condiciones de regularidad y características anteriores podemos enunciar el teorema de Cramer-Rao (1945).

Teorema 2.1 Dado un modelo estadístico regular $\{X \sim F_\theta : \theta \in \Theta\}$, es decir, un modelo donde se verifican las condiciones de regularidad enunciadas, cualquier estimador $T \in \mathcal{S}(\theta)$ de la clase de los estimadores no sesgados y con varianza verifica

$$\text{var}_\theta(T) \geq \frac{1}{I_n(\theta)}$$

Demostración: El estimador $T \in \mathcal{S}(\theta)$ no tiene sesgo, es decir que

$$E(T) = \int_{S^n} T(\mathbf{x}) \cdot L(\mathbf{x}; \theta) d\mathbf{x} = \theta$$

Si derivamos e introducimos la derivada bajo el signo de la integral, obtenemos

$$\begin{aligned} \frac{\partial}{\partial \theta} E(T) &= \int_{S^n} \frac{\partial}{\partial \theta} (T(\mathbf{x}) \cdot L(\mathbf{x}; \theta)) d\mathbf{x} = \int_{S^n} T(\mathbf{x}) \frac{\partial}{\partial \theta} L(\mathbf{x}; \theta) d\mathbf{x} \\ &= \int_{S^n} T(\mathbf{x}) \left(\frac{\frac{\partial}{\partial \theta} L(\mathbf{x}; \theta)}{L(\mathbf{x}; \theta)} \right) L(\mathbf{x}; \theta) d\mathbf{x} \end{aligned}$$

Así pues

$$1 = \frac{\partial}{\partial \theta} \theta = \frac{\partial}{\partial \theta} E(T) = E(TZ) = \int_{S^n} T(\mathbf{x}) \cdot ZL(\mathbf{x}; \theta) d\mathbf{x}$$

En resumen

$$E(T) = \theta, E(TZ) = 1, E(Z) = 0, \text{var}(Z) = I_n(\theta)$$

Si ahora consideramos el coeficiente de correlación al cuadrado entre T y Z , tenemos

$$\rho^2(T, Z) = \frac{[\text{cov}(T, Z)]^2}{\text{var}(T) \cdot \text{var}(Z)} = \frac{[E(TZ) - E(T)E(Z)]^2}{\text{var}(T) \cdot \text{var}(Z)} \leq 1$$

Si sustituimos los resultados hallados antes, obtenemos

$$\frac{1}{\text{var}(T) \cdot I_n(\theta)} \leq 1$$

de donde se deduce la desigualdad enunciada.

Definición 2.11 Si un estimador alcanza la CCR (Cota de Cramer-Rao), diremos que es un estimador eficiente.

Todo estimador eficiente es de mínima varianza en la clase $\mathcal{S}(\theta)$. Sin embargo, también puede suceder que exista un estimador de mínima varianza sin alcanzar necesariamente la CCR.

Ejemplo 2.3.3 Sea $X \sim F_\theta = P(\lambda)$, $\lambda > 0$ (Poisson). Buscamos la CCR de los estimadores de λ .

$$\begin{aligned} L(\mathbf{x}; \lambda) &= \prod_{i=1}^n e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum x_i}}{\prod_{i=1}^n x_i!} \\ \log L(\mathbf{x}; \lambda) &= -n\lambda + \left(\sum x_i \right) \log \lambda - \log \left(\prod_{i=1}^n x_i! \right) \\ \frac{\partial \log(L(\mathbf{x}; \lambda))}{\partial \lambda} &= -n + \frac{\sum x_i}{\lambda} \\ E \left[\frac{\partial \log L(\mathbf{x}; \lambda)}{\partial \lambda} \right]^2 &= E \left[n^2 + \left(\frac{\sum X_i}{\lambda} \right)^2 - \frac{2n \sum X_i}{\lambda} \right] \\ &= n^2 + \frac{1}{\lambda^2} E \left(\sum X_i \right)^2 - \frac{2n}{\lambda} nE(X) \end{aligned}$$

Aquí recordamos que la suma de variables de Poisson también es una Poisson, es decir:

$$\sum X_i \sim P(n\lambda)$$

por lo que

$$E \left(\sum X_i \right)^2 = \text{var} \left(\sum X_i \right) + \left[E \left(\sum X_i \right) \right]^2 = n\lambda + (n\lambda)^2$$

Finalmente, se obtiene:

$$E(Z^2) = n^2 + \frac{n\lambda}{\lambda^2} + \frac{n^2\lambda^2}{\lambda^2} - 2n^2 = \frac{n}{\lambda}$$

De esta forma,

$$I_n(\lambda) = \frac{n}{\lambda} \implies \text{var}(T) \geq \frac{\lambda}{n}$$

Sabemos que la media aritmética verifica

$$\text{var}(\bar{X}_n) = \frac{\lambda}{n}$$

lo cual coincide con la cota de Cramer-Rao, indicando que \bar{X}_n es el estimador eficiente de λ .

Ejemplo 2.3.4 Para calcular la CCR o, dicho de otro modo, para que el inverso de

$$E \left[\frac{\partial \log L(\mathbf{x}; \theta)}{\partial \theta} \right]^2$$

sea realmente la cota mínima de $\text{var}(\hat{\theta})$ en la clase $\mathcal{S}(\theta)$, es necesario que se verifiquen las condiciones de regularidad. De lo contrario, se pueden obtener resultados absurdos. Consideremos, por ejemplo, una variable aleatoria X con función de densidad

$$f(x; \theta) = \frac{3}{\theta^3} x^2 \mathbf{1}_{[0, \theta]}(x)$$

y esperanza

$$E(X) = \int_0^\theta x \cdot \frac{3}{\theta^3} x^2 dx = \frac{3}{4} \theta$$

Ya que $\theta = \frac{4}{3} E(X)$, esto sugiere estimar θ mediante $\hat{\theta} = \frac{4}{3} \bar{X}$, que no tiene sesgo. Por otro lado, si calculamos la varianza de X , tenemos

$$\text{var}(X) = E(X^2) - E(X)^2 = \frac{3}{80} \theta^2$$

Sabemos que $E(\hat{\theta}) = \theta$, y además

$$\text{var}(\hat{\theta}) = \text{var}\left(\frac{4}{3} \bar{X}\right) = \frac{\theta^2}{15n}$$

Si evaluamos $I_n(\theta)$ en su forma más sencilla, obtenemos

$$I_n(\theta) = nI(\theta) = n \frac{9}{\theta^2}$$

Así, la CCR resulta ser mayor que la varianza de este estimador:

$$\text{var}(\hat{\theta}) = \frac{\theta^2}{15n} < \frac{\theta^2}{9n}$$

lo cual es un resultado absurdo. Este error se debe a no considerar que el soporte de X depende de θ , por lo que no se cumplen las condiciones de regularidad, y la cota de Cramer-Rao no existe.

También ocurre que la varianza de un estimador es inferior a la CCR aunque esta exista. Esto puede pasar, por ejemplo, con algún estimador sesgado.

7.8 Caracterización del estimador eficiente

Calcular la cota de Cramer-Rao es una cosa; encontrar el estimador que alcanza esta cota y, en consecuencia, tiene varianza mínima es otra. La siguiente caracterización permite, en algunos casos, obtener directamente la forma del estimador eficiente.

Teorema 2.2 Sea T el estimador eficiente de θ , entonces se verifica

$$\sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(X_i; \theta) = K(\theta, n)(T - \theta)$$

donde $K(\theta, n)$ es una función que depende de θ y de n y que suele coincidir con la información de Fisher. Demostración: Si T es el estimador eficiente, entonces

$$\text{var}(T) = \frac{1}{I_n(\theta)}$$

y, por lo tanto, $\rho^2(T, Z) = 1$. En general, dadas dos variables aleatorias X e Y , se sabe que si $\rho(X, Y) = 1$, entonces

$$Y - E(Y) = \beta(X - E(X))$$

Si aplicamos este resultado a T y Z , tenemos

$$\begin{aligned} Z - E(Z) &= \beta(T - E(T)) \\ \frac{\partial \log L(\mathbf{x}; \theta)}{\partial \theta} &= K(\theta, n)(T - \theta) \end{aligned}$$

Ejemplo 2.3.5 En el caso de la distribución de Poisson, tenemos

$$\begin{aligned} f(x; \lambda) &= e^{-\lambda} \frac{\lambda^x}{x!} \\ \log f(x; \lambda) &= -\lambda + x \log(\lambda) - \log(x!) \\ \frac{\partial \log f(x; \lambda)}{\partial \lambda} &= -1 + x \frac{1}{\lambda} \\ Z = \sum_{i=1}^n \frac{\partial \log f(X_i; \lambda)}{\partial \lambda} &= \sum_{i=1}^n \left(-1 + \frac{X_i}{\lambda} \right) \end{aligned}$$

Queremos ver que

$$\sum_{i=1}^n \left(\frac{X_i}{\lambda} - 1 \right) = K(\theta, n)(T - \theta)$$

Si reescribimos esta expresión, obtenemos

$$\frac{1}{\lambda} \sum_{i=1}^n X_i - n = \frac{1}{\lambda} \left(\sum_{i=1}^n X_i - n\lambda \right) = \frac{n}{\lambda} \left(\frac{1}{n} \sum_{i=1}^n X_i - \lambda \right)$$

Así, $K(\lambda, n) = \frac{n}{\lambda}$, que coincide con la información de Fisher $I_n(\lambda)$. Por el teorema anterior, se deduce que $T = \bar{X}$ es el estimador eficiente y, por lo tanto, de mínima varianza.

7.9 Estadísticos suficientes

En un problema de inferencia puede suceder que los datos contengan información superflua o irrelevante a la hora de estimar el parámetro. También puede ocurrir lo contrario, que intentemos hacer la estimación sin utilizar toda la información disponible en la muestra. Ambas situaciones son indeseables. Parece razonable que, para estimar un parámetro, dada la dificultad derivada de disponer de varios estimadores entre los que queremos elegir el óptimo, nos basemos únicamente en aquellos que utilizan (solo) toda la información relevante.

Ejemplo 2.4.1 Supongamos que queremos estimar la proporción de piezas defectuosas θ en un proceso de fabricación. Para ello, examinamos n piezas extraídas al azar a lo largo de una jornada y asignamos un 1 a las piezas defectuosas y un 0 a las que no lo son. Así, obtenemos una muestra aleatoria simple X_1, X_2, \dots, X_n donde

$$X_i = \begin{cases} 1 & \text{con probabilidad } \theta \\ 0 & \text{con probabilidad } (1 - \theta) \end{cases}$$

Intuitivamente, está claro que para estimar θ solo nos interesa el número de ceros y unos, es decir, el valor del estadístico

$$T(\mathbf{X}) = \sum_{i=1}^n X_i$$

En este caso, un estadístico que considere la posición de los unos y los ceros en la muestra no aportaría nada relevante. En cambio, un estadístico que no considere todos los valores, como por ejemplo $T(\mathbf{X}) = X_1$, sería claramente menos adecuado.

Las observaciones del ejemplo anterior se justifican al observar que todas las muestras de tamaño n con el mismo número t de unos (1) tienen la misma probabilidad. En concreto, la función de probabilidad de una muestra x_1, x_2, \dots, x_n es

$$f_{\theta}(x_1, x_2, \dots, x_n) = \theta^t (1 - \theta)^{n-t}$$

donde $t = \sum_{i=1}^n x_i, x_i \in \{0, 1\}, i = 1, 2, \dots, n$. Como se puede ver, la probabilidad de la muestra solo depende del número de unos (o ceros) y no del orden en que aparecen en la muestra. El hecho de que la posición de los unos y los ceros en la muestra no aporte información relevante equivale a decir que el estadístico

$$T(\mathbf{X}) = \sum_{i=1}^n X_i$$

contiene la misma información que X_1, X_2, \dots, X_n para estimar θ . Observamos, sin embargo, varias diferencias entre basarse en $T(\mathbf{X})$ o en X_1, X_2, \dots, X_n :

- Al pasar de X_1, X_2, \dots, X_n a $\sum_{i=1}^n X_i$ hay una reducción de los datos que no implica pérdida de información.
- Muchas muestras diferentes dan lugar al mismo valor de T .

Fisher formalizó esta idea con el cálculo de la probabilidad condicionada de la observación muestral con $T(\mathbf{X}) = \sum_{i=1}^n X_i$ y para todo $t = 0, 1, \dots, n$:

$$\begin{aligned} P_{\theta}[\mathbf{X} = \mathbf{x} \mid T = t] &= \frac{P_{\theta}[\mathbf{X} = \mathbf{x}, T = t]}{P_{\theta}(T = t)} \\ &= \frac{\theta^t (1 - \theta)^{n-t}}{\binom{n}{t} \theta^t (1 - \theta)^{n-t}} = \frac{1}{\binom{n}{t}} \end{aligned}$$

Es decir, dados $(x_1, x_2, \dots, x_n) \in \{0, 1\}^n$ y $t \in \{0, 1, \dots, n\}$, tenemos

$$P_\theta[\mathbf{X} = \mathbf{x} \mid T = t] = \begin{cases} 0 & \text{si } t \neq \sum_{i=1}^n x_i \\ \frac{1}{\binom{n}{t}} & \text{si } t = \sum_{i=1}^n x_i \end{cases}$$

Obviamente, $P_\theta[\mathbf{X} = \mathbf{x}]$ depende de θ , que es el parámetro que queremos estimar. Sin embargo, la probabilidad condicionada $P_\theta[\mathbf{X} = \mathbf{x} \mid T = t]$ no depende de θ . Tenemos entonces la siguiente expresión de la función de probabilidad de la muestra:

$$P_\theta(\mathbf{X} = \mathbf{x}) = P_\theta(T = t) \cdot P_\theta[\mathbf{X} = \mathbf{x} \mid T = t]$$

Esta expresión muestra que $P_\theta(\mathbf{X})$ se puede descomponer en dos factores, uno que depende de θ , $P_\theta(T = t)$, y otro que no depende de θ ,

$$P_\theta[\mathbf{X} = \mathbf{x} \mid T = t].$$

Una forma de ver esta descomposición es pensar que el estadístico $T = \sum_{i=1}^n X_i$?acumula? o ?absorbe? toda la información relativa a θ , lo que se refleja en que la probabilidad de la muestra, dado $T = t$, ya no depende de θ . Es decir, podemos imaginar la construcción de la muestra en dos etapas:

- En una primera etapa se elige el valor t para T con distribución $B(n, \theta)$.
- A continuación, se sitúan aleatoriamente t unos y $n - t$ ceros en las n posiciones.

Cuando la estructura del estadístico $T(\mathbf{X})$ hace que el segundo factor en la expresión anterior no dependa de θ , significa que la observación adicional de la muestra es irrelevante. En este caso diremos que $T(\mathbf{X})$ es suficiente para la estimación de θ . Dado que esta propiedad de T queda caracterizada por la independencia de $P_\theta[\mathbf{X} = \mathbf{x} \mid T = t]$ respecto a θ , se utiliza esta independencia para definir la suficiencia.

7.9.1 Definición de estadístico suficiente

Dado un modelo estadístico $\{X \sim F_\theta : \theta \in \Theta\}$ y un estadístico T , diremos que T es suficiente para θ si, dada una muestra $\mathbf{X} = (X_1, X_2, \dots, X_n)$, se verifica que la distribución de \mathbf{X} condicionada por el valor de T no depende de θ .

- No es necesario que F_θ sea discreta, como en el ejemplo introductorio, o que la muestra sea una muestra aleatoria simple.
- El estadístico suficiente para un parámetro puede ser k -dimensional.

Ejemplo 2.4.2 Dada una muestra X_1, X_2, \dots, X_n de una distribución de Poisson, la función de probabilidad de la muestra es

$$P_\theta(X_1 = x_1, \dots, X_n = x_n) = \frac{e^{-n\lambda} \lambda^{\sum x_i}}{x_1! \cdots x_n!}$$

Calculemos la probabilidad de la muestra condicionada por el valor del estadístico $T = \sum_{i=1}^n X_i$:

$$\begin{aligned} P_\theta[X_1 = x_1, \dots, X_n = x_n \mid T = t] &= \frac{P_\theta(X_1 = x_1, \dots, X_n = x_n, T = t)}{P_\theta(T = t)} \\ &= \frac{t!}{x_1! \cdots x_n!} \left(\frac{1}{n}\right)^t \mathbf{1}_{\{\sum x_i = t\}}(x_1, \dots, x_n) \end{aligned}$$

La probabilidad condicional no depende de λ , por lo tanto, T es suficiente para λ . Conviene observar que, en este ejemplo, no todas las muestras tienen la misma probabilidad.

7.9.2 Teorema de factorización

La justificación de la suficiencia de un estadístico mediante la definición no siempre es sencilla, ya que la distribución condicional puede ser intratable con las herramientas disponibles. El teorema que se presenta a continuación proporciona un método sencillo para comprobar la suficiencia de un estadístico y, a menudo, sugiere cuál es el estadístico suficiente de menor dimensión posible.

Teorema 2.3 Neyman-Fisher. Sea $\{X \sim F_\theta : \theta \in \Theta\}$ un modelo estadístico y X_1, X_2, \dots, X_n una muestra aleatoria simple de X . Sea $f_\theta(\mathbf{x})$ la función de probabilidad o la función de densidad de la muestra, según si X es discreta o absolutamente continua. Un estadístico T es suficiente para θ si y solo si existen dos funciones medibles g_θ y h tales que

$$f_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x})) \cdot h(\mathbf{x})$$

donde h no depende de θ y g depende de θ y, además, solo depende de la muestra a través de T .

Veamos ahora la demostración del teorema de factorización, restringida al caso de variables discretas.

Demostración: Comenzaremos suponiendo que T es suficiente y concluiremos que es posible la factorización. Si $T(\mathbf{X})$ es suficiente para la familia de distribuciones $\{F_\theta; \theta \in \Theta\}$, la función de probabilidad de la muestra condicionada por T no depende de θ . Dado que

$$f_\theta(\mathbf{x}) = P_\theta[T = T(\mathbf{x})] \cdot f_\theta[\mathbf{x} \mid T = T(\mathbf{x})]$$

solo es necesario tomar $g_\theta(t) = P_\theta[T = T(\mathbf{x}) = t]$ y $h(\mathbf{x}) = f_\theta[\mathbf{x} \mid T = T(\mathbf{x})]$ para obtener el resultado. Ahora supongamos que es posible la factorización y deduzcamos la suficiencia. Si $f_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x})) \cdot h(\mathbf{x})$ y llamamos $A_t = \{\mathbf{x} \in X(\Omega)^n \mid T(\mathbf{x}) = t\}$, entonces

$$P_\theta[T(\mathbf{x}) = t] = \sum_{A_t} g_\theta(T(\mathbf{x})) \cdot h(\mathbf{x}) = g_\theta(t) \cdot \sum_{A_t} h(\mathbf{x})$$

Consideremos ahora la distribución de la muestra condicionada a $T = t$. El Teorema de Bayes para densidad permite escribir:

$$\begin{aligned} f_\theta(\mathbf{x} \mid T = t) &= \frac{f_\theta(\mathbf{x}, T = t)}{P_\theta(T = t)} \\ &= \begin{cases} \frac{g_\theta(t) \cdot h(\mathbf{x})}{g_\theta(t) \cdot \sum_{A_t} h(\mathbf{x})} = \frac{h(\mathbf{x})}{\sum_{A_t} h(\mathbf{x})} & \text{si } T(\mathbf{x}) = t \\ 0 & \text{si } T(\mathbf{x}) \neq t \end{cases} \end{aligned}$$

De modo que la distribución de \mathbf{X} condicionada por el valor de T no depende de θ , y, en consecuencia, T es suficiente.

Ejemplo 2.4.3 Si X sigue una distribución de Bernoulli, tenemos:

$$f_\theta(\mathbf{x}) = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i} = g_\theta \left(\sum_{i=1}^n x_i \right).$$

Si tomamos $h(\mathbf{x}) = 1$, queda probado que $T = \sum_{i=1}^n X_i$ es suficiente. **Ejemplo 2.4.4** Si consideramos una muestra de una distribución de Poisson

$$f_\lambda(\mathbf{x}) = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{x_1! x_2! \cdots x_n!}$$

y tomamos $T(\mathbf{x}) = \sum_{i=1}^n x_i$, podemos escribir

$$f_{\lambda}(\mathbf{x}) = e^{-n\lambda} \lambda^{T(\mathbf{x})} \cdot (x_1!x_2! \cdots x_n!)^{-1} = g_{\lambda}(T(\mathbf{x})) \cdot h(\mathbf{x})$$

donde

$$g_{\lambda}(T(\mathbf{x})) = e^{-n\lambda} \lambda^{T(\mathbf{x})}, \quad h(\mathbf{x}) = (x_1!x_2! \cdots x_n!)^{-1}$$

De modo que $g_{\lambda}(t) = e^{-n\lambda} \lambda^t$ depende de la muestra solo a través de $T = \sum_{i=1}^n x_i$ y $h(\mathbf{x}) = (x_1!x_2! \cdots x_n!)^{-1}$ no depende de λ .

Ejemplo 2.4.5 Supongamos que \mathbf{X} es una muestra aleatoria simple de una población $X \sim N(\mu, \sigma)$, cuya función de densidad es

$$f_{\mu, \sigma^2}(x_1, x_2, \dots, x_n) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\}$$

Para evidenciar la factorización, utilizamos que

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2.$$

Entonces,

$$\begin{aligned} f_{\mu, \sigma^2}(x_1, x_2, \dots, x_n) &= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp \left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \right) \right\} \\ &= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp \left\{ -\frac{1}{2\sigma^2} (ns^2 + n(\bar{x} - \mu)^2) \right\} \\ &= g_{\mu, \sigma^2}(\bar{x}, s^2) \cdot 1 \end{aligned}$$

Así, vemos que el estadístico (\bar{X}, s^2) es suficiente para la estimación de (μ, σ^2) . Si suponemos conocido uno de los dos parámetros σ^2 o μ , podemos obtener una factorización en la que se ve que $\sum_{i=1}^n (x_i - \mu)^2$ es suficiente para σ^2 (conocido μ) o \bar{x} es suficiente para μ (conocido σ^2).

En el ejemplo anterior se observa que el estadístico suficiente para un problema puede tener una dimensión superior a 1. En general, buscaremos el estadístico suficiente de menor dimensión posible, ya que a menor dimensión se elimina más información superflua. Si no es posible encontrarlo así, siempre podemos basarnos en el estadístico $T = (X_1, X_2, \dots, X_n)$, que es suficiente pero de dimensión máxima y, por lo tanto, no aporta ninguna reducción al problema de información. Estas reflexiones llevan a enunciar el principio de suficiencia, que aconseja condensar al máximo la información relevante en un estadístico suficiente T de la menor dimensión posible (“mínima”) y seleccionar un estimador T' entre los estadísticos que sean función de la muestra a través de $T : T'(\mathbf{X}) = \varphi(T(\mathbf{X}))$.

7.9.3 Propiedades de los estadísticos suficientes

Las siguientes propiedades se prueban de manera sencilla utilizando el teorema de factorización:

1. Si T es un estadístico suficiente para θ y φ es una función inyectiva (o monótona diferenciable), entonces $T_1 = \varphi(T)$ también es suficiente para θ .

Ejemplo 2.4.6 En la familia de la Poisson hemos visto que $\sum_{i=1}^n X_i$ es suficiente para λ . Entonces $\bar{X} = \varphi(\sum_{i=1}^n X_i)$, donde $\varphi(z) = (1/n)z$ es inyectiva, es suficiente para λ . 2. Si T es un estadístico suficiente para θ y φ es una función paramétrica monótona diferenciable, entonces $\varphi(T)$ también es suficiente para $\varphi(\theta)$. 3. Si T_1, T_2 son dos estadísticos suficientes para θ , entonces T_1 es función de T_2 .

7.10 Obtención de estimadores

En el capítulo anterior hemos analizado el problema de la estimación puntual desde el punto de vista de, dado un estimador, ver ¿qué tan bueno es? para estimar un parámetro. Otra cuestión que nos podemos plantear, de hecho la primera cuestión que hay que plantearse en la práctica, es cómo obtener un estimador ¿razonablemente bueno? de un parámetro. De hecho, desde el punto de vista práctico parece razonable empezar por ver cómo se obtiene un estimador y, una vez obtenido, analizar ¿cuán bueno resulta?. Existen muchos métodos para obtener estimadores, cada uno de los cuales puede llevarnos a unos resultados de diferente calidad. Los principales métodos de estimación son:

1. Método de los momentos
2. Método de la máxima verosimilitud
3. Método de Bayes
4. Otros métodos

7.11 El método de los momentos

Este método fue introducido por K. Pearson a finales del siglo XIX y es el principio en que nos basamos cuando hacemos una estimación de la media o de la varianza poblacional a partir de la media o la varianza muestrales. La idea del método de los momentos es bastante intuitiva. Si lo que queremos estimar (uno o varios parámetros) es una función de los momentos poblacionales, entonces una estimación razonable puede consistir en tomar como estimador la misma función en la que los momentos poblacionales han sido sustituidos por los momentos muestrales. Dado que estos últimos son estimadores consistentes de los momentos poblacionales, en condiciones bastante generales se puede garantizar que los estimadores obtenidos serán estimadores consistentes para las funciones de los momentos poblacionales estimadas. Algunos ejemplos típicos de estimadores basados en el método de los momentos son:

$$\hat{\mu} = \bar{X}_n \quad \hat{\sigma} = \sqrt{S^2} \quad \hat{\sigma}^2 = S^2$$

Sea un modelo estadístico, $\{X \sim F_\theta : \theta \in \Theta\}$, y X_1, X_2, \dots, X_n una muestra aleatoria simple de X . Sean m_1, m_2, \dots, m_k los momentos poblacionales de orden 1, 2, ..., k de X , que suponemos que existen,

$$m_k = E(X^k)$$

y a_1, a_2, \dots, a_k los momentos muestrales respectivos

$$a_k(X_1, X_2, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i^k$$

Suponemos que estamos interesados en estimar:

$$\theta = h(m_1, m_2, \dots, m_p),$$

donde h es una función conocida. Definición 3.1 El método de los momentos consiste en estimar θ por el estadístico

$$T(\mathbf{X}) = h(a_1, a_2, \dots, a_p)$$

7.11.1 Observaciones

- El método se extiende de forma sencilla a la estimación de momentos conjuntos. Podemos usar $\frac{1}{n} \sum_{i=1}^n X_i Y_i$ para estimar $E(XY)$, etc.
- Por la ley débil de los grandes números,

$$a_k(X_1, X_2, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow{P} E(X^k),$$

de modo que si lo que queremos es estimar los momentos muestrales, el método garantiza que los estimadores son consistentes y sin sesgo.

En este caso, además, los estimadores son asintóticamente normales. Si lo que se desea estimar es una función h continua de los momentos, entonces el método garantiza que el estimador $T(\mathbf{X})$ es consistente y, bajo ciertas condiciones de regularidad, también es asintóticamente normal.

Ejemplo 3.1.1 Sea $X \sim \Gamma(p, \alpha)$. Queremos estimar p y α . En lugar de conocer la función $h(\theta_1, \theta_2)$ sabemos que:

$$\begin{aligned} m_1 &= \frac{p}{\alpha} = E(X) \\ m_2 &= \frac{p(p+1)}{\alpha^2} = E(X^2) \\ &= V(X) + [E(X)]^2 = \frac{p}{\alpha^2} + \left(\frac{p}{\alpha}\right)^2 = \frac{p^2 + p}{\alpha^2} = \end{aligned}$$

De modo que podemos obtener las funciones deseadas ?aislando? p y α como funciones de m_1 y m_2 :

$$\begin{aligned} \alpha^2 &= \frac{p^2}{m_1^2} \\ \alpha^2 &= \frac{p(p+1)}{m_2} \end{aligned}$$

Procediendo por igualación:

$$\begin{aligned} \frac{p^2}{m_1^2} &= \frac{p(p+1)}{m_2} \\ \frac{p}{m_1} &= \frac{p+1}{m_2} \\ pm_2 &= pm_1^2 + m_1^2 \\ p(m_2 - m_1^2) &= m_1^2 \frac{m_1}{m_2 - m_1^2} \\ p &= \frac{m_1^2}{m_2 - m_1^2} \\ \alpha &= \frac{m_1^2}{m_2 - m_1^2} \\ m_1 & \end{aligned}$$

Los estimadores por el método de los momentos se obtendrán ahora sustituyendo p y α por \hat{p} y $\hat{\alpha}$ en la expresión anterior, es decir:

$$\hat{p} = \frac{a_1^2}{a_2 - a_1^2}$$

Hacemos lo mismo para el parámetro α :

$$\hat{\alpha} = \frac{a_1}{a_2 - a_1^2}$$

7.12 El método del máximo de verosimilitud

7.12.0.1 Introducción El método de la máxima verosimilitud, introducido por Fisher, es un método de estimación que se basa en la función de verosimilitud, presentada en el capítulo anterior. Básicamente consiste en tomar como estimadores de los parámetros aquellos valores que hagan más probable observar precisamente lo que se ha observado, es decir, que hagan que la muestra observada resulte más verosímil.

Ejemplo 3.2.1 Tomemos 5 papeles. En cada uno de ellos ponemos o bien un $++$ o bien un $-$, sin que se sepa qué hay en cada papel, y los guardamos en una bolsa. Nuestro objetivo es estimar el número de papeles con el signo $-$ escrito. Extraemos tres papeles, devolviéndolos a la bolsa después de cada extracción, y observamos que ha salido lo siguiente: $++-$. Los valores posibles para la probabilidad de $-$, llamémosla p , son:

En la bolsa hay	p
4 $+$?, 1 $-$?	0,2
3 $+$?, 2 $-$?	0,4
2 $+$?, 3 $-$?	0,6
1 $+$?, 4 $-$?	0,8

Supongamos que la variable X mide el número de $-$ en tres extracciones consecutivas y que, por tanto, sigue una distribución binomial:

$$X \sim B(3, p(?-))$$

La probabilidad de sacar un $-$ es:

$$P_p[X = 1] = \binom{3}{1} \cdot p^1(1-p)^2$$

Para cada uno de los valores de p , las probabilidades quedan así:

p	$P_p[X = 1]$
0.2	$3 \cdot 0.2 \cdot 0.8^2 = 0.384$
0.4	$3 \cdot 0.4 \cdot 0.6^2 = 0.432$
0.6	$3 \cdot 0.6 \cdot 0.4^2 = 0.288$
0.8	$3 \cdot 0.8 \cdot 0.2^2 = 0.096$

El valor de p que da una probabilidad mayor a la muestra, es decir, que la hace más verosímil, es $p = 0.4$. El método del máximo de verosimilitud consiste precisamente en tomar este valor como estimación de p .

7.12.0.2 La función de verosimilitud Una vez introducido el método con un ejemplo, podemos pasar a definirlo con mayor precisión. Para ello, comenzaremos con el concepto de función de verosimilitud. En el capítulo anterior presentamos la función de verosimilitud como la función que resulta de considerar que, en la función de probabilidad de la muestra, el parámetro es variable y la muestra queda fija. Es decir:

$$\underbrace{f(x_1, x_2, \dots, x_n; \theta)}_{\text{x variable, } \theta \text{ fijo}} \longrightarrow \underbrace{L(\theta; x_1, x_2, \dots, x_n)}_{\text{x fija, } \theta \text{ variable}}$$

Esta definición es básicamente correcta. En el caso de las variables discretas, donde $f(x_1, x_2, \dots, x_n; \theta)$ representa la probabilidad de la muestra, fijado θ , resulta intuitivamente claro decir que la verosimilitud representa la ?probabilidad de la muestra para cada valor del parámetro?. Refiriéndonos al ejemplo introductorio, resulta sencillo ver que se trata de ?dos puntos de vista? sobre la misma función. Fijado un valor del parámetro, por ejemplo, 0.4 , podemos considerar la probabilidad de diversas muestras posibles, como $x = 0, x = 1, \dots$, hasta $x = 3$:

$$\begin{aligned} f(x_1, x_2, \dots, x_n; \theta) &= P_{0.4}[X = x], x = 0, 1, \dots, 3 \\ &= \binom{3}{x} \cdot 0.4^x (0.6)^{3-x}. \end{aligned}$$

Análogamente, fijada una muestra, por ejemplo, $x = 1$, podemos considerar la probabilidad de esta para diversos valores del parámetro, $p = 0, 0.2, \dots, 1$.

$$\begin{aligned} L(x_1, x_2, \dots, x_n; \theta) &= P_p[X = 1], x = 0, 0.2, 0.4, \dots, 1 \\ &= 3 \cdot p(1 - p)^2. \end{aligned}$$

En el caso de las distribuciones absolutamente continuas, el significado de la función de verosimilitud ya no es intuitivamente tan claro como en el caso de las discretas. En este caso, la función de densidad de la muestra ya no representa la probabilidad de esta como en el caso de las discretas. Algunos autores intentan solucionar esto explicando que existe una conocida aproximación en que la función de densidad es la probabilidad de un suceso ?infinitesimal?. Lo que es importante en la función de verosimilitud, a la hora de hacer inferencias, es la parte que es función del parámetro. Esto hace que a menudo se considere que la expresión de la función de verosimilitud mantenga solo aquella parte de $f(x_1, x_2, \dots, x_n; \theta)$ que depende de θ , ignorando la parte que dependa solo de la muestra. Es decir, si podemos factorizar $f(x_1, x_2, \dots, x_n; \theta)$ como

$$f(\mathbf{x}; \theta) = c(\mathbf{x}) \cdot g(\mathbf{x}; \theta)$$

podremos prescindir de la ?constante? $c(\mathbf{x})$ (constante porque no depende de θ) al considerar la verosimilitud.

$$L(\theta; \mathbf{x}) = g(\mathbf{x}; \theta) \propto f(\mathbf{x}; \theta)$$

Esto implica que $L(\theta; \mathbf{x})$ no tiene por qué integrar a 1 , como en el caso de las probabilidades, y que depende de las unidades de medida.

Ejemplo 3.2.2 Si X es discreta, $X \sim \mathcal{P}(\lambda)$, y suponemos $n = 1$ (muestras de tamaño 1), tenemos que la f.d.p. de la muestra es:

$$P[x; \lambda] = e^{-\lambda} \frac{\lambda^x}{x!}$$

con $x = 0, 1, \dots$. Ahora, si hemos observado $x = 5$, la función de verosimilitud vale:

$$L(\lambda; 5) = e^{-\lambda} \lambda^5 \left[\frac{1}{5!} \right]$$

Como solo nos interesa la parte que es función de λ , podemos ignorar $\frac{1}{5!}$, es decir:

$$L(\lambda; 5) = e^{-\lambda} \lambda^5 \propto P[\mathbf{x}; \lambda].$$

Ejemplo 3.2.3 Si dada una muestra de tamaño 1, por ejemplo, $x = 2$, de una ley de Poisson $\mathcal{P}(\lambda)$ queremos comparar sus verosimilitudes respecto de los valores del parámetro $\lambda = 1.5$ o $\lambda = 3$, lo que haremos será basarnos en la razón de verosimilitudes:

$$\begin{aligned}\Lambda(\mathbf{x}) &= \frac{L(\lambda_1; x)}{L(\lambda_2; x)} = \frac{L(1.5; 2)}{L(3; 2)} \\ &= \frac{e^{-1.5} 1.5^2 \left[\frac{1}{2!}\right]}{e^{-3} 3^2 \left[\frac{1}{2!}\right]} = \frac{e^{-1.5} 1.5^2}{e^{-3} 3^2} = \frac{0.5020}{0.4481} = 1.12.\end{aligned}$$

Como se observa, al basarnos en la razón de verosimilitudes, la parte correspondiente solo a la muestra no se toma en cuenta. La razón de verosimilitudes sugiere que el valor $\lambda = 1.5$ hace la muestra más verosímil.

7.12.0.3 El método del máximo de verosimilitud Si partimos de las dos ideas que hemos visto en la introducción:

- Escoger como estimación el valor que maximice la probabilidad de la muestra observada.
- La verosimilitud de la muestra es una aproximación a la probabilidad de esta como función del valor del parámetro.

Una forma razonable de definir el EMV es entonces como aquel que maximice la verosimilitud.

Definición 3.2 Un estimador $T : \Omega \longrightarrow \Theta$ es un estimador del máximo de verosimilitud para el parámetro θ si cumple:

$$L(T(\mathbf{x}); \mathbf{x}) = \sup_{\theta \in \Theta} L(\theta; \mathbf{x})$$

Como suele ocurrir en problemas de maximización, este valor ni existe necesariamente ni tiene por qué ser único. Ahora bien, bajo ciertas condiciones (las habituales para los problemas de máximos y mínimos) el problema se podrá reducir a buscar un máximo para la función de verosimilitud.

Ejemplo 3.2.4 Supongamos que x_1, \dots, x_n es una muestra de una población de Bernoulli, $X \sim Be(p)$, donde queremos estimar p . La función de masa de la probabilidad de X es:

$$P[X = x_i] = P(x_i; p) = p^{x_i} (1 - p)^{1-x_i} \text{ donde } x_i \in \{0, 1\}; i = 1, \dots, n$$

La función de verosimilitud es:

$$L(p; \mathbf{x}) = \prod_{i=1}^n p^{x_i} (1 - p)^{1-x_i} = p^{\sum_{i=1}^n x_i} (1 - p)^{\sum_{i=1}^n (1-x_i)}$$

Debemos buscar el máximo de $L(p; \mathbf{x})$. En este caso, como en otros, es más sencillo buscar el máximo de su logaritmo, que, dado que es una función monótona, es el mismo que el máximo de L

$$\ln L(p; \mathbf{x}) = \left(\sum_{i=1}^n x_i \right) \cdot \ln p + \left(n - \sum_{i=1}^n x_i \right) \cdot \ln(1 - p)$$

Derivamos respecto a p :

$$\frac{\partial \ln L(p; \mathbf{x})}{\partial p} = \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1 - p}$$

e igualamos a cero la derivada, planteando lo que se denomina la ecuación de verosimilitud, cuyas soluciones nos conducirán eventualmente al estimador del máximo de verosimilitud.

$$\frac{\sum_{i=1}^n x_i - n\hat{p}}{\hat{p}(1 - \hat{p})} = 0 \Rightarrow \hat{p} = \frac{\sum_{i=1}^n x_i}{n}$$

Si la segunda derivada es negativa en \hat{p} entonces será un máximo:

$$\begin{aligned} \frac{\partial^2 \ln L(p; x)}{\partial p^2} &= \frac{\partial}{\partial p} \left(\frac{\sum_{i=1}^n x_i - np}{p(1 - p)} \right) = \frac{-n[p(1 - p)] - (\sum_{i=1}^n x_i - np) \cdot (1 - 2p)}{p^2(1 - p^2)} = \\ &= \frac{-np + np^2 - \sum_{i=1}^n x_i - np - 2p \sum_{i=1}^n x_i - 2np^2}{p^2(1 - p)^2} = \\ &= \frac{[\sum_{i=1}^n x_i(1 + 2p) - np^2]}{p^2 \cdot (1 - p)^2} \end{aligned}$$

que es negativa cuando $p = \hat{p}$, de forma que \hat{p} es efectivamente un máximo. El método analítico expuesto en el ejemplo anterior, consistente en el cálculo de un extremo de una función, no se puede aplicar en todas las situaciones. En estos casos, una alternativa puede ser estudiar directamente la función de verosimilitud. Veamos un ejemplo:

Ejemplo 3.2.5 Sea $X_1, \dots, X_n \stackrel{iid}{\sim} X \sim U(0, \theta)$ $\theta > 0$ desconocido. Sabemos que:

$$f(x; \theta) = \begin{cases} \frac{1}{\theta} & \text{si } 0 < \min \{x_i\} \leq \max \{x_i\} \leq \theta \\ 0 & \text{en caso contrario} \end{cases}$$

La derivada respecto a θ es $-\frac{n}{\theta^{n+1}}$, que se anula cuando $\theta \xrightarrow{n \rightarrow \infty} \infty$ que lleva a una solución sin sentido de la ecuación de verosimilitud. Una inspección de la gráfica de la función de verosimilitud revela que el EMV, en este caso,

Figura 3.1: Función de verosimilitud para una distribución uniforme es $\max \{X_i, \dots, X_n\}$. Efectivamente, consideremos cualquier otro valor θ^* diferente del máximo:

$$\begin{aligned} \text{Si } \theta^* > X_{(n)} &\Rightarrow \frac{1}{(\theta^*)^n} < \frac{1}{(X_{(n)})^n}, \\ \text{Si } \theta^* < X_{(n)} &\Rightarrow L(\theta^*; \mathbf{x}) = 0 \end{aligned}$$

ya que si un estimador toma un valor inferior al máximo de la muestra habrá algún valor muestral, x_i para el cual se verificará que $\theta^* < x_i$, lo que hace la muestra inverosímil, y por tanto el estimador no es admisible. A la vista de lo anterior, deducimos que el valor que maximiza $L(\theta; \mathbf{x})$ es el máximo de la muestra.

Ejemplo 3.2.6 El método del máximo de verosimilitud se extiende de forma inmediata a los parámetros K -dimensionales. Consideremos el caso de la ley normal $X \sim N(\mu, \sigma^2)$. Aquí el parámetro θ es bidimensional, es decir: $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}^+$

1. La función de verosimilitud de una muestra de tamaño n es:

$$L((\mu, \sigma^2); \mathbf{x}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} = \frac{1}{(2\pi)^{n/2} (\sigma^2)^{n/2}} e^{-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}}$$

2. Sacando logaritmos

$$\log L((\mu, \sigma^2); \mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}$$

3. La derivada de $L()$ es la matriz de derivadas:

$$D \log L((\mu, \sigma^2); \mathbf{x}) = \left(\frac{\partial \log L((\mu, \sigma^2); \mathbf{x})}{\partial \mu}, \frac{\partial \log L((\mu, \sigma^2); \mathbf{x})}{\partial \sigma^2} \right) = \left\{ \begin{array}{l} \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2} \\ \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2} \end{array} \right.$$

4. Planteando y resolviendo la ecuación de verosimilitud tenemos:

$$D \log L((\hat{\mu}, \hat{\sigma}^2); \mathbf{x}) = \left\{ \begin{array}{l} \frac{\sum_{i=1}^n (x_i - \hat{\mu})}{\hat{\sigma}^2} = 0 \\ \frac{\sum_{i=1}^n (x_i - \hat{\mu})^2}{2\hat{\sigma}^4} = \frac{n}{2\hat{\sigma}^2} \end{array} \right.$$

de donde las raíces de la ecuación de verosimilitud son:

$$\hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = s^2.$$

5. Para decidir si las raíces de la ecuación de verosimilitud corresponden a un máximo, analizamos la matriz de derivadas segundas, denominada Hessiana.

$$H = \begin{pmatrix} \frac{\partial^2 z}{\partial x^2} & \frac{\partial^2 z}{\partial x \partial y} \\ \frac{\partial^2 z}{\partial y \partial x} & \frac{\partial^2 z}{\partial y^2} \end{pmatrix}$$

Una condición suficiente para que un punto (x_0, y_0) sea un máximo es que el determinante de H sea positivo y el menor en la posición 11? negativo, es decir: $S_i |H| > 0 y \frac{\partial^2 z}{\partial x^2} \Big|_{(x_0, y_0)} < 0 \implies$ Hay un máximo relativo en (x_0, y_0) . Si evaluamos el Hessiano en el punto (\bar{x}, s^2) tenemos:

$$H = \begin{pmatrix} -\frac{n}{s^2} & 0 \\ 0 & -\frac{n}{2s^4} \end{pmatrix}.$$

Las condiciones de extremo que hemos dado más arriba se verifican: $H_{11} < 0 y |H| > 0$, de manera que podemos concluir que el estimador del máximo de verosimilitud de (μ, σ^2) es, efectivamente, (\bar{x}, s^2) .

8 Estimación por intervalos

Este capítulo está pendiente de ser introducido en los apuntes.

La versión actualizada estará disponible en el momento de inicio de la actividad, durante el semestre actual (2024-25-S1).

Se plantea el problema de la estimación como una forma de aproximación a las características de las distribuciones a partir de muestras aleatorias simples.

Se abordan las distintas filosofías para la construcción de estimadores.

- 8.1 Preliminares: estimación del error estándar e Introducción al bootstrap
- 8.2 Estimadores por intervalo: intervalos de confianza
- 8.3 Intervalos de confianza para características de una población normal (media, varianza),
- 8.4 Intervalos de confianza bootstrap.
- 8.5 Intervalos de confianza para proporciones binomiales
- 8.6 Intervalos de confianza para parámetros en muestra grandes y para casos generales (tasas, OR, ...)
- 8.7 Aplicaciones: cálculo del tamaño muestral

9 Pruebas de hipótesis

Este capítulo está pendiente de ser introducida en los apuntes.

La versión actualizada estará disponible en el momento de inicio de la actividad, durante el semestre actual (2024-25-S1).

Se plantea el problema de las pruebas de hipótesis. Se discuten las aproximaciones y los conceptos asociados. Se trata el problema de la crisis de la significación.

- 9.1 Conceptos básicos: pruebas de hipótesis y de significación, pruebas unilaterales y bilaterales, tipos de error, valores críticos de test y p-valores
- 9.2 Potencia de un test. Cálculos de potencia y de tamaño de la muestra. Tamaño del efecto.
- 9.3 Métodos de construcción de tests.
- 9.4 Problemas asociados al uso de tests estadísticos. La crisis de la significación

10 Inferencia Aplicada

Este capítulo está pendiente de ser introducida en los apuntes.

La versión actualizada estará disponible en el momento de inicio de la actividad, durante el semestre actual (2024-25-S1).

Se muestra como deducir y aplicar algunos de los tests mas populares.

- 10.1 Pruebas de normalidad. Pruebas gráficas. El test de Shapiro-Wilks
- 10.2 Pruebas de hipótesis para contrastar variables cuantitativas: pruebas paramétricas t-test y Anova
- 10.3 Pruebas de hipótesis para contrastar variables cuantitativas: pruebas de hipótesis no paramétricas de Wilcoxon y Kruskal-Wallis
- 10.4 Contrastes para datos categóricos. Pruebas binomiales, ji cuadrado y test de Fisher.
- 10.5 Riesgo relativo y razón de «odds»

11 Computación Intensiva y *Multiple Testing*

Este capítulo está pendiente de ser introducida en los apuntes.

La versión actualizada estará disponible en el momento de inicio de la actividad, durante el semestre actual (2024-25-S1).

Se introducen distintos métodos cuyo nexo común es la computación intensiva.

- 11.1 Tests de permutaciones; ¿Qué?, ¿Cuándo?, ¿Cómo?
- 11.2 El bootstrap en contraste de hipótesis
- 11.3 El problema de las comparaciones múltiples
- 11.4 Métodos de control de error: FWER y FDR