Probabilidad y Simulación

Anabel Forte Deltell

Departamento de Estadística e Investigación Operativa. Universitat de Valencia.

En Construcción. Última actualización: septiembre 2020

Anabel Forte Detell

11/9/2020

Índice

1. Introducción a la probabilidad

1.1. Historia de la probabilidad

No me cabe la menor duda de que todas y todos habéis escuchado hablar de suerte, coincidencia, aleatoriedad, incertidumbre, riesgo, fortuna, azar... pero siempre utilizadas de una manera informal.

De hecho estos conceptos son intrínsecos al ser humano que siempre ha andado intrigado con cuestiones como si va a llover o no, si tiene alguna posibilidad de ganar a los dados (uno de los juegos de azar más antiguos que se conocen) o a las cartas, o a lo que sea... mientras sea ganar.

Pero para lo que estamos aquí es para formalizar estos conceptos porque, no olvidemos que la probabilidad es la rama de las matemáticas que trata de formalizar el concepto de incertidumbre (las matemáticas siempre empeñadas en formalizarlo todo).

Algunos de los primeros autores conocidos que trabajaron para formalizar probabilidad fueron Blaise Pascal (1623-1662) o Pierre Fermat (1601–1665) aunque en los trabajos de Cardano o Galileo Galilei ya aparecen algunos conceptos del cálculo de probabilidades.

Vale pero, ¿Por qué este empeño en formalizar la incertidumbre? Bien, la probabilidad es necesaria en muchas de las ciencias que conocemos (sino en todas). En particular, si pensamos en la Ciencia de Datos, vais a necesitar la probabilidad para aplicar y entender los conceptos estadísticos indispensables para el análisis de cualquier tipo de datos, desde los pequeños conjuntos de datos disponibles en cualquier ciencia (Medicina, Biología, Economía, etc.) al BigData de las redes sociales.

Pero empecemos por el principio: ¿qué es la probabilidad? No se trata de una definición sencilla por lo que vamos a empezar con algunos conceptos previos como son la aleatoriedad, los experimentos y los sucesos.

1.2. Aleatoriedad, experimentos y sucesos

Antes de empezar es importante que tengáis en mente que mucho de lo que vamos a ver en esta sección tiene que ver con conceptos que ya habéis estudiado en Matemática Discreta (conjuntos, subconjuntos, combinatoria, etc.)

Para empezar debemos entender que significa que algo sea aleatorio. Si yo os pregunto cuánto suman 2+2 seguro que tenéis clarísimo el resultado y, también, que este no va a cambiar por mucho que os lo pregunte saltando a la pata coja, dando volteretas, en un día de sol o si está lloviendo. Sin embargo, si lanzamos una moneda al aire no tendremos tan claro el resultado. Unas veces saldrá cara y otras, cruz.

A la situación del primer ejemplo la llamamos **determinista**, mientras que en la segunda hacemos referencia a un *proceso* **aleatorio**. En concreto, cuando realizamos una acción cuyo resultado es desconocido hasta que se ejecuta, decimos que estamos llevando a cabo un **experimento** y, más concretamente, un experimento aleatorio.

En el ejemplo anterior el experimento es lanzar la moneda al aire y observar si sale cara o cruz. Otros ejemplos clásicos de experimentos aleatorios son lanzar un dado de 6 caras al aire y observar el resultado numérico obtenido o sacar una carta de una baraja y ver de que palo es.

Estos son ejemplos sencillos pero podemos pensar en otros más elaborados como: extraer

un poco de sangre a una persona y ver cuál es su grupo sanguíneo; medir los minutos transcurridos entre el paso de dos tranvías...

Al conjunto de posibles resultados de un experimento se le llama **espacio muestral** y lo vamos a denotar con la letra griega Ω.

- En el ejemplo de la moneda el espacio muestral sería Ω ={Cara, Cruz}.
- En el del dado tenemos Ω ={1,2,3,4,5,6}.
- En el caso del grupo sanguíneo Ω ={A, B, AB, O}.
- En el ejemplo del tiempo transcurrido entre el paso de dos tranvías, el resultado puede ser cualquier número de minutos entre 0 e infinito, por tanto, $\Omega = [0, \infty)$

En ocasiones no nos interesa estudiar todo el espacio muestral, quizás nos interesa solo una parte de el, un subconjunto. Pues, bien, a cualquier subconjutno de un espacio muestral se le denomina **suceso**[1].

Volviendo a los ejemplos anteriores, posibles sucesos serían:

- Obtener una cara al lanzar la moneda.
- Obtener un número par al lanzar un dado. Fijaos que, en este caso, el suceso contiene más de un posible resultado.
- Tener una A en el grupo sanguíneo.
- Que el tiempo entre el paso de dos tranvías sea menor de 5 minutos.

Toda esta terminología probabilística puede formalizarse utilizando el lenguaje de la Teoría de Conjuntos que ya debéis conocer de matemática discreta. La Tabla 1 muestra la traducción entre ambos lenguajes.

Siguiendo con la Teoría de Conjuntos, una herramienta interesante para visualizar distintos sucesos con respecto al espacio muestral es lo que se conoce como Diagramas de Venn. Podemos ver un ejemplo en la Figura 7

La otra herramienta, no sólo interesante sino, fundamental es aprender a contar el número de resultados que caben dentro de un suceso. Esto ya lo habéis estudiado en matemática discreta y se llama combinatoria.

[1]Suceso y evento son sinónimos y, por tanto, es posible que encontréis esta misma definición bajo una u otra denominación. De hecho, el término en inglés es *event* (el termino success significa algo completamente diferente). En cualquier caso, la idea importante es que hablamos de un subconjunto del espacio muestral.

Lenguaje	**Notación**
Lo que podemos observar	Ω
Que no pase nada	$\emptyset$
s es un posible resultado	$s \in \Omega$
A es un suceso	$A \subseteq \Omega$
Ha pasado el suceso A	$s_{obs} \in A$
Observamos el suceso A o el B	$A \cup B$
Observamos el suceso A y el B	$A \cap B$
No observamos A	A^c
Observamos A o B pero no los dos	$(A \cap B^c) \cup (A^c \cap B)$
Observamos alguno de los sucesos $A_1, \ldots, A_n$	$A_1 \cup A_2 \cup \cdots \cup A_n$
Observamos todos los sucesos $A_1, \ldots, A_n$	$A_1 \cap A_2 \cap \cdots \cap A_n$
Que pase A implica que pase B	$A \subseteq B$
A y B no pueden suceder a la vez	$A \cap B = \emptyset$
$A_1, \ldots, A_n$ son una partición de Ω	$A_1 \cup \cdots \cup A_n = \Omega$ y $A_i \cap A_j = \emptyset \ \forall i \neq j$

Cuadro 1: Traducción de afirmaciones probabilísticas en lenguaje de teoría de conjuntos

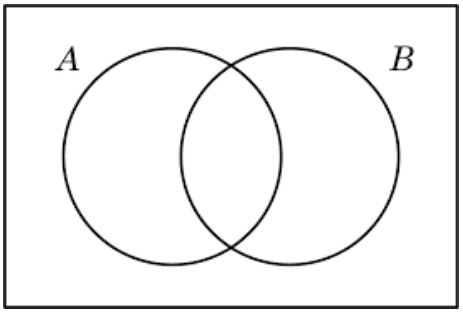


Figura 1: Diagrama de Venn

Un ejemplo sencillo. Imaginad que lanzamos una moneda dos veces, pensemos en el espacio muestral: los posibles resultados serán 4: CC, CX, XC, XX (siendo C cara y X cruz) y por tanto, el espacio muestral es Ω ={CC, CX, XC, XX}. Si pensamos en el suceso A "que salga al menos una cara" ¿Cuantos resultados posibles tiene dicho suceso? La solución en este caso es sencilla, mirando Ω vemos que hay tres resultados que contienen al menos una cara.

Bueno, ahora que conocemos los conceptos de experimento y suceso podemos plantearnos la formalización del termino **probabilidad**.

1.3. Definición y axiomas probabilísticos

Ahora ya estamos en disposición de definir que significa eso de probabilidad, ¿o no?. Definir la probabilidd no es una tarea fácil. Lo que es básico es tener claro que la probabilidad es un número entre 0 y 1 donde 0 significa que el suceso no es posible y 1 que es seguro que pasa.

Pero, ¿cómo se interpreta/calcula la probabilidad de un suceso concreto?

Siguiendo a DeGroot and Schervish [2012], existen tres formas distintas de entender la probabilidad. Tres que, además, se entremezclan y confunden fácilmente.

La probabilidad puede entenderse **como una frequencia**. En este sentido, la probabilidad de un determinado suceso sería la proporción de veces que observamos dicho suceso cuando realizamos el experimento un número grande de veces, bajo las mismas circunstancias.

En el ejemplo de la moneda esto sería, si lanzamos la moneda muchas veces cuántas caras y cuántos cruces obtendremos. Si lo pensamos intuitivamente, (teniendo en cuenta que la moneda no este trucada), esperaríamos obtener el mismo número de caras que de cruces y por tanto los dos sucesos serán igual de probables (esto es, como después veremos, una probabilidad de 0.5 para cada uno).

Esta definición de probabilidad tiene algunos inconvenientes: ¿qué entendemos por *un número grande de veces*? Y ¿ qué significa *bajo las mismas circunstancias*? Además, si lo repito bajo las mismas circunstancias ¿no obtendré siempre lo mismo? Por otra parte, ¿qué sucede cuando un experimento no puede ser repetido (ni un número grande ni un número pequeño de veces)? Por ejemplo, si el experimento es saber quien gana o pierde una competición de [ponga aquí su deporte favorito] ¿Es posible repetir esa competición bajo *las mismas circunstancias*?

Después tenemos la interpretación **clásica** de la probabilidad que se basa en el concepto de sucesos igual de probables. La idea es que, si todos los N resultados de un espacio muestral Ω son igual de probables, y un suceso A contiene m de esos resultados, la probabilidad de A puede calcularse mediante la fórmula:

$$P(A) = \frac{m}{N}$$

Conocida como **fórmula de Laplace**[2].

[2]Pierre-Simon Laplace (Beaumont-en-Auge, Normandía, Francia, 23 de marzo de 1749-París, 5 de marzo

Esta definición es, sin embargo, redundante puesto que estamos usando el concepto de probabilidad (sucesos igual de probables) dentro de su definición. Además, que pasa cuando dos sucesos no son igual de probables, por ejemplo, como asignamos la probabilidad a una cara o a una cruz si la moneda no está bien balanceada.

La tercera posible definición de probabilidad es el concepto de **probabilidad subjetiva** donde ésta se define como la medida de lo verosímil que es un suceso a partir del conocimiento que tenemos sobre el mismo. El problema, evidentemente, es que dos personas, con conocimientod distintos, pueden no estar de acuerdo en la asignación de dicha verosimilitud.

Esta interpretación de la probabilidad contiene, de una forma u otra, a las dos anteriores. Pensadlo, ¿Cómo establecemos la verosimilitud de un suceso? ¿Cómo creamos nuestra interpretación de esa verosimilitud? La mayor parte de nosotr@s lo haría pensando en las veces que dicho suceso se repite (definición frecuentista) o en que hay sucesos que son igual de verosímiles (interpretación clásica). Aunque también podemos hacerlo por comparación: Si me planteo apostar por cara en un lanzamiento de moneda antes que apostar por mi equipo en la competición es porque creo que la probabilidad de que mi equipo gane es inferior al 0.5. Del mismo modo, si apuesto antes por sacar un 1 en un lanzamiento de dados que por que gane mi equipo, es que confio bastante poco en que ganen (probabilidad inferior a 0.16).

Lo bueno es que, la formalización de la probabilidad que veremos a continuación es valida sin importar la interpretación utilizada[3].

Pues vamos a ver la definición de probabilidad y sus axiomas tal y como los enunció A.N.Kolmogorov[4] en 1933

de 1827) fue un astrónomo, físico y matemático francés. Continuador de la mecánica newtoniana, descubrió y desarrolló la transformada de Laplace y la ecuación de Laplace; como estadístico sentó las bases de la teoría analítica de la probabilidad

[3]Si bien es cierto que esta definición y sus axiomas derivados son ciertos sin importar la interpretación del termino probabilidad, debemos tener en cuenta que el paradigma estadístico resultante si será diferente. Las interpretaciones frequentista y clásica de la probabilidad dan lugar al paradigma conocido como **Frequentista** mientras que el enfoque **Bayesiano** se deriva de la visión *subjetiva* de la probabilidad.

[4]Andrey Nikolaevich Kolmogorov (Tambov, 25 de abril de 1903-Moscú, 20 de octubre de 1987) fue un matemático ruso que realizó aportes de primera línea en los contenidos de teoría de la probabilidad y de topología. Estructuró el sistema axiomático de la teoría de la probabilidad, utilizando el lenguaje teoría de conjuntos.

Definición: Un espacio de probabilidad consiste en un espacio muestral Ω y una función de probabilidad $P(\cdot)$.

La función P debe cumplir (axiomas):

- $0 \leq P \leq 1$
- $P(\Omega) = 1$
- Si $A_1, A_n \ldots$ son eventos que no pueden suceder a la vez (eventos **disjuntos** $\forall i \neq j$ $A_i \cap A_j = \emptyset$)

$$P(\bigcup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} P(A_j)$$

Para entender un poco mejor esta definición Blitzstein and Hwang [2015] nos animan a pensar en una caja llena de piedrecitas cuyo peso total sea 1 (1 kg, 1 g, lo que sea). La caja vacía pesa 0 mientras que la caja completa pesa 1 y cada una de las piedras tiene un peso diferente.

Evidentemente, el peso de cualquier subconjunto de piedrecitas será las suma del peso de cada piedra en ese conjunto.

A partir de esta definición podemos obtener las siguientes propiedades de la probabilidad.

Teorema: Con la anterior definición de probabilidad tenemos que:

- $P(\emptyset) = 0$,
- $P(A^c) = 1 - P(A)$, siendo A^c el suceso: que no pase A
- Si $A \subseteq B$, $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Podéis intentar demostrar el teorema vosotros mismos a partir de la definición de probabilidad o acudir a Blitzstein and Hwang [2015] para una prueba formal.

La tercera de estas reglas tiene una generalización conocida como la fórmula de inclusión-exclusión:

Teorema: (Inclusión-exclusión). Para cualquier grupo de sucesos $A_1, \ldots, A_n$,

$$P(\bigcup_{j=1}^{n} A_j) = \sum_{j=1}^{n} P(A_j) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) - \ldots (-1)^{n+1} P(A_1 \cap \cdots \cap A_n)$$

Fijaos que esta última regla parece estar en contradicción con la definición de probabilidad

inicial. pero OJO en la definición hacíamos referencia a sucesos disjuntos (que no tienen nada en común) y que juntos cubren todo el espacio muestral. Aquí, estamos hablando de sucesos que pueden compartir parte del espacio muestral y que, si sólo sumamos probabilidades, estaríamos considerando su probabilidad más de una vez.

Esto se ve muy bien en los siguientes diagramas de Venn

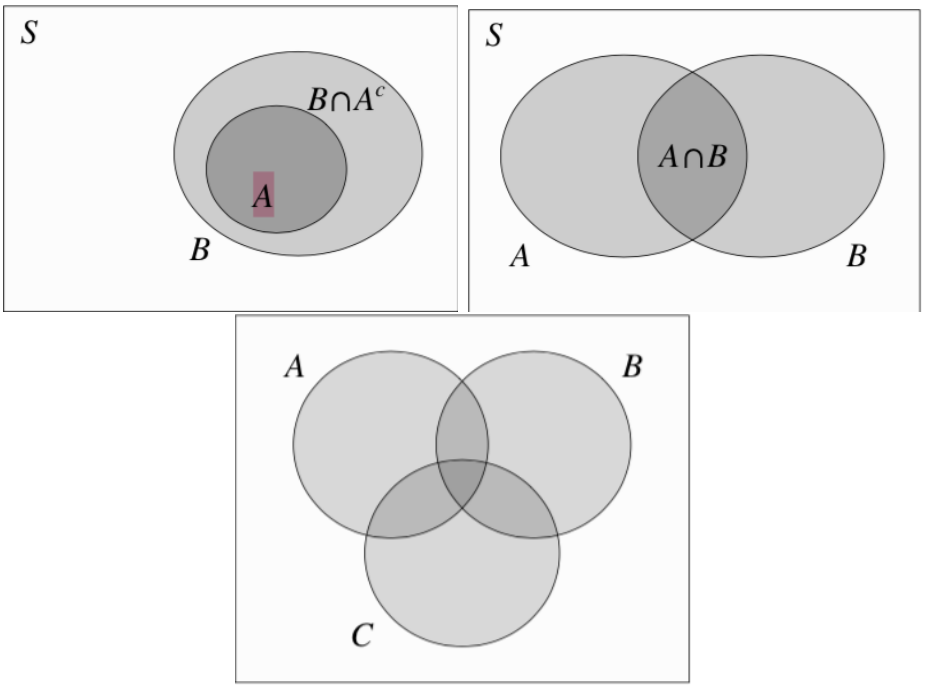


Figura 2: Diagrama de Venn

Una vez formalizado el concepto de probabilidad no debemos perder de vista que esta tiene una interpretación y una componente cotidiana que, como tal, la hace dependiente y cambiante según cambia el desarrollo de los acontecimientos.

Por ejemplo, si sabemos que hace 10 minutos que estamos en la parada del autobús, tendremos a pensar que la probabilidad de que el próximo autobús llegue en los próximos 5 minutos será mayor que si sólo llevamos esperando en la parada 1 minuto ¿no?

Bien, esta idea se formaliza en el concepto de probabilidad condicionada que vamos a estudiar en el Tema 2.

Ahora unos pocos ejercicios:

1.4. Ejercicios

1. Laura está planeado salir a cenar todas las noches de una semana, de lunes a viernes. Cada día quiere ir a uno de sus 10 restaurantes favoritos.

   (a) ¿Cuántas posibilidades tiene Laura de organizar sus cenas si no quiere ir al mismo restaurante más de una vez?

   (b) Y si no le importa repetir pero no quiere cenar en el mismo restaurante dos dias consecutivos?

2. Si tenemos 12 personas

   (a) ¿Cuántas formas hay de dividirlas en 3 equipos donde uno de los equipos tiene 2 personas y, los otros 2, 5 cada uno?

   (b) ¿Cuántas formas hay de dividirlas en tres equipos donde cada equipo tiene 4 personas?

3. Una familia tiene 3 hijos y 3 hijas. Asumiendo que todos tienen la misma probabilidad de haber nacido en cualquier orden. ¿Cuál es la probabilidad de que las tres mayores sean chicas?

4. Una ciudad con 6 barrios sufre 6 accidentes en una semana. Asumiendo que los accidentes pueden haber sucedido en cualquier parte de la ciudad con la misma probabilidad y que los tres barrios tienen el mismo tamaño. ¿Cual es la probabilidad de que un barrio haya tenido más de un accidente?

5. Sabemos que la probabilidad de que cierto estudiante A suspenda un examen de probabilidad es 0.5 y la probabilidad de que lo suspenda otro estudiante B es 0.2. Además sabemos que ambos suspenderán a la vez el examen con una probabilidad de 0.1,

   (a) ¿Cuál es la probabilidad de que, al menos uno, suspenda el examen?
   (b) ¿Cuál es la probabilidad de que ninguno de los estudiantes suspenda el examen?
   (c) ¿Cuál es la probabilidad de que exactametne 1 de ellos suspenda el examen?

6. Si el 50 % de las familias de una cierta ciudad se suscriben al periódico de la mañana, el 65 % se suscriben al de la tarde y el 85 % se suscribe al menos a uno de los dos. ¿Qué porcentaje de familias están suscritas a ambos periódicos?

7. Una persona llega al centro de salud con dolor de garganta y algo de fiebre. Después de examinarla le dicen que puede tener una infección bacteriana; una infección viral o ambas. En concreto tiene una probabilidad de 0.7 de que sea bacteriana y una probabilidad de 0.4 de que sea vírica ¿Cuál es la probabilidad de que tenga ambas?

8. Una caja contiene tres cartas. Una carta es roja por las dos caras, otra es verde por las dos caras y la tercera es roja por una cara y verde por la otra. Sacamos una carta de la caja y observamos que es verde por una cara. ¿Qué probabilidad hay de que la otra cara también lo sea?

9. Sea $A_i : i \in I$ una sucesión de conjuntos. Prueba las leyes de Morgan[5] que dicen:

$$(\bigcup_i A_i)^c = \bigcap_i A_i^c, \quad (\bigcap_i A_i)^c = \bigcup_i A_i^c$$

10. Tenemos un conjunto de tazas con sus correspondientes platos. Hay dos tazas (con sus platos) que son rojas, dos blancas y dos con estrellas. Si las tazas se asignan a los platos de forma aleatoria, encuentra la probabilidad de que ninguna taza coincida con el patron de su plato.

[5]Augustus De Morgan (Madurai, India; 27 de junio de 1806 - Londres, 18 de marzo de 1871) fue un matemático y lógico británico nacido en la India. Profesor de matemáticas en el University College de Londres entre 1828 y 1866; y primer presidente de la Sociedad Matemática de Londres. Conocido por formular las llamadas leyes de De Morgan, en su memoria, y establecer un concepto riguroso del procedimiento, inducción matemática

2. Probabilidad Condicionada

2.1. Probabilidad Condicionada

Cuando queremos calcular la probabilidad de un evento teniendo cierta información sobre lo que ya ha pasado, utilizamos lo que se conoce como Probabilidad condicionada

Definición: Dado un evento B del que sabemos que $P(B) > 0$, definimos la probabilidad de que suceda otro evento A condicionada a que ha sucedido B como:

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

Fijaros que en la definición de probabilidad condicionada se esconde cierta idea de aprendizaje. Estamos actualizando nuestro conocimiento sobre A dado que sabemos B.

Intuición matemática: Al considerar $P(A \cap B)$ estamos reduciendo nuestras posibilidades a que pasen A y B (dado que B ya ha sucedido) y después, al dividir por $P(B)$ estamos *enfocando* (consiguiendo que la probabilidad sume 1 en un nuevo espacio muestral, B). Si nos vamos al mundo de las piedrecitas del que hablamos en el tema anterior y miramos el siguiente gráfico, lo vemos más claro:

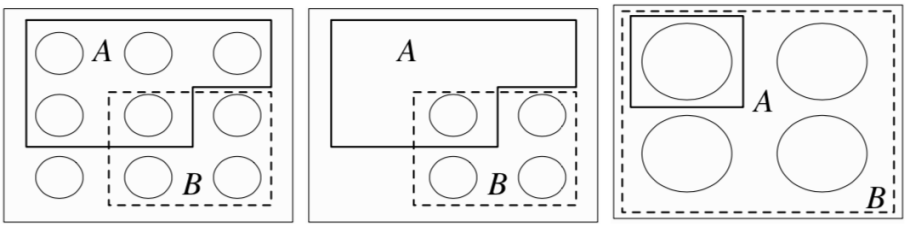


Figura 3: Probabilidad condicionada

Estamos reduciendo nuestras posibilidades a las piedras del grupo B (porque ya sabemos que es eso lo que ha pasado), B es nuestro nuevo espacio muestral. Ahora, como la suma de probabilidades debe ser 1 en ese nuevo espacio muestral, debemos dividir el peso de todas las piedras por $P(B)$ y el efecto es como hacer zoom (como vemos en la tercera imagen). Buscamos entonces las piedras que están en A y en B a la vez ($P(A \cap B)$) y sumamos sus nuevos pesos obteniendo así la $P(A \mid B)$.

Intuición en un ejemplo real: Una mujer es portadora de una enfermedad hereditaria ¿Cuál es la probabilidad de que su próximo hijo tenga la enfermedad?

Según las leyes de Mendel, todos los posibles genotipos del hijo de una madre portadora

(xX)[6] y un padre normal (XY) son xX, xY, XX, XY y tienen la misma probabilidad. El espacio muestral es $\Omega = \{\text{xX, xY, XX, XY}\}$

El suceso A=\{hijo enfermo\} corresponde al genotipo xY, por tanto, según la definición clásica de probabilidad $P(A) = 1/4 = 0{,}25$

La mujer tiene el hijo y es varón ¿qué probabilidad hay de que tenga la enfermedad?

Se define el suceso $B = \{\text{ser varón}\} = \{\text{xY, XY}\}$ con probabilidad $P(B) = 0{,}5$ la probabilidad que necesitamos es la del suceso $A \mid B$= {estar enfermo dado que es varón}; $P(A \mid B)$. Aplicando la definición de probabilidad condicional tenemos:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\text{xY})}{0{,}5} = \frac{0{,}25}{0{,}5} = 0{,}5$$

Es importante destacar que las probabilidades condicionales son probabilidades y, por tanto, se ciñen a las mismas reglas. Podemos definir, por ejemplo, la probabilidad de un suceso complementario $P(A^c \mid B)$ que será igual a $1 - P(A \mid B)$.

La definición de probabilidad condicional es sencilla pero da pie a dos teoremas muy importantes (el Teorema de la probabilidad total y el Teorema de Bayes) que permiten calcular probabilidades condicionadas en un gran número de situaciones.

2.1.1. Teorema de la probabilidad total

Partiendo de la definición de probabilidad condicionada, resulta sencillo observar que

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

y

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

Esta regla se puede extender a cualquier número de sucesos no disjuntos $\{A_1, \ldots, A_n\}$ de manera que

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \ldots P(A_n \mid A_1 \cap \cdots \cap A_{n-1})$$

[6] la x minúscula representa el cromosoma enfermo

Utilizando los dos resultados anteriores, la definición de probabilidad y teniendo en cuenta el resultado de teoría de conjuntos que nos dice que

$$A = (A \cap B) \cup (A \cap B^c)$$

obtenemos el siguiente resultado conocido como la **ley de la probabilidad total**

Teorema: (Ley de la probabilidad total) Dada una serie de sucesos disjuntos $B_1 \dots B_n$ cuya unión es el espacio muestral Ω, la probabilidad de cualquier suceso $A \subset \Omega$ puede calcularse como:

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) P(B_i)$$

En la Figura 4 podemos ver una versión gráfica de este teorema.

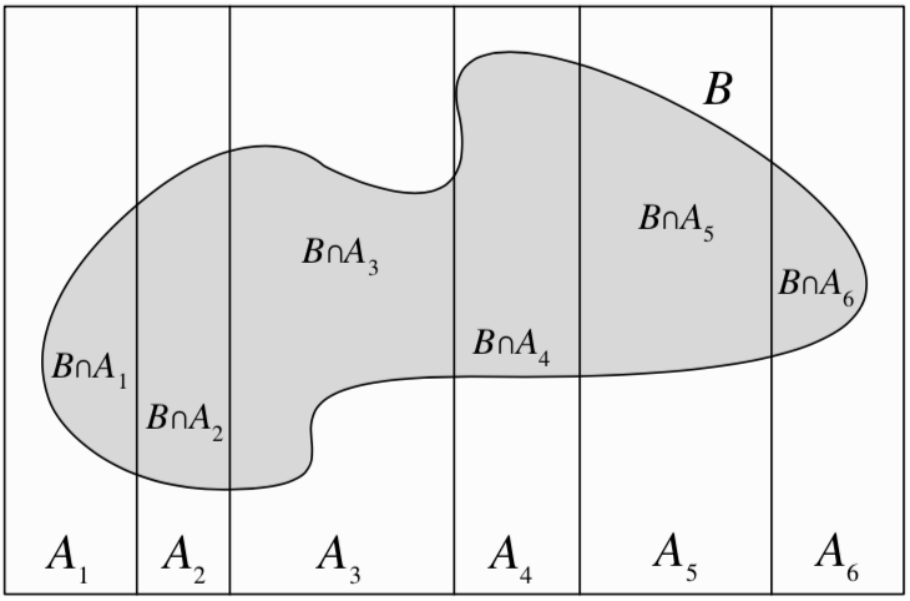


Figura 4: Teorema de la probabilidad total

Veamos un ejemplo: Imaginad que tenemos dos monedas. Una de ellas tiene cara y cruz mientras que la otra tiene dos caras. Escogéis una moneda al azar (de una bolsa) y la lanzáis. ¿Cuál es la probabilidad de cara?.

En este ejemplo resulta sencillo dividir el espacio muestral en dos, los resultados que vienen de la moneda trucada y los que vienen de la moneda no trucada. Además conocemos la probabilidad de cara bajo cada una de esas circunstancias 1 y 1/2 respectivamente. También

conocemos la probabilidad de que la moneda elegida esté trucada o no, 1/2 en ambos casos, puesto que las hemos elegido al azar.

Con todos estos elementos podemos definir el suceso $A =$ Cara, el suceso $B_1 =$ moneda trucada y el suceso B_2 moneda no trucada. Utilizando entonces el teorema anterior tenemos:

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) = 1 \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{3}{4}$$

Para trabajar este tipo de probabilidades, se puede recurrir a una herramienta muy visual llamada diagrama de árbol o árbol de probabilidad donde cada resultado viene representado por un circulo y el condicionamiento se representa mediante lineas de conexión. En cada arista suele aparecer un número que representa la probabilidad del resultado final dado el resultado anterior (es decir, la probabilidad condicionada). Para este ejemplo concreto, el diagrama de árbol sería:

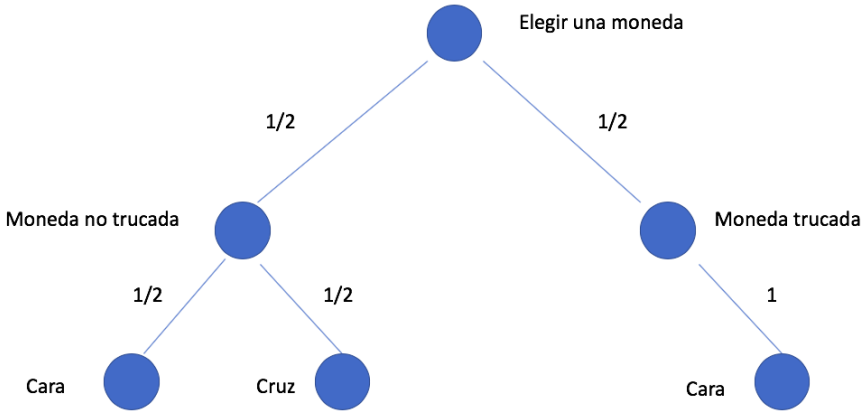


Figura 5: Árbol de decisión

La probabilidad de cualquier nodo puede calcularse como el producto de las probabilidades de las aristas que llevan hasta el y la probabilidad de un suceso concreto será la suma de las probabilidades de los nodos que conforman ese suceso.

2.1.2. Teorema de Bayes

Thomas Bayes fue un reverendo presbiteriano allá por el siglo XVIII. No publicó ningún trabajo sobre probabilidad en vida o al menos, no que se sepa. Sin embargo, tras su muerte, en 1763 su amigo y colega Richard Price publicaría un trabajo de Bayes que haría que su

nombre pasase a la historia aunque, cabe mencionar que la forma en la que hoy se conoce el Teorema de Bayes se la debemos a Laplace (el Newton Francés), así como muchos de los resultados que veremos en esta asignatura.

Teorema (Teorema de Bayes)

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Esta fórmula puede deducirse fácilmente a partir de la probabilidad condicionada y se trata de una manera muy útil de calcular probabilidades condicionadas ya que, habitualmente $P(B \mid A)$ resulta mucho más sencilla de calcular que $p(A \mid B)$ o viceversa.

Utilizando el teorema de la probabilidad total, la regla de Bayes puede expresarse también a partir de la una partición del espacio muestral $A_1 \dots A_n$ de manera que

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)}$$

Volviendo al ejemplo de las dos monedas, supongamos ahora que lo que nos interesa es saber cuál es la probabilidad de que la moneda escogida sea la trucada dado que nos ha salido una cara. Recordemos que una moneda tenía dos caras y la otra sólo una.

Veamos, al lanzar una moneda (cualquiera de las dos) los posibles resultados hubiesen sido $\Omega = \{\text{C,X,C,C}\}$. Este espacio muestral puede dividirse en dos sucesos A_1 que la moneda esté trucada y por tanto sólo pueda salir cara o que la moneda no esté trucada A_2 y que pueda salir cara o cruz. Cada uno de ellos tiene una probabilidad 1/2. El suceso B es, en este caso, haber obtenido cara.

Fijaros que en este caso es difícil determinar P(moneda esté trucada dado que he obtenido una cara) pero muy fácil dar la P(haber obtenido una cara dado que la moneda está trucada) (en concreto 1)

Tenemos entonces

$$P(A_1 \mid B) = \frac{P(B \mid A_1)P(A_1)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2)} = \frac{1 \times \frac{1}{2}}{1 \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2}} = \frac{2}{3}$$

Por tanto, la probabilidad de que la moneda escogida esté trucada es 2/3.

2.1.2.1. El teorema de Bayes en salud

Una aplicación muy común del teorema de Bayes en la salud es la detección de enfermedades.

Imaginad que existe una enfermedad muy rara que sólo afecta al 1 % de la población. Para poder detectar dicha enfermedad se elabora un test que tiene una efectividad de un 95 %, es decir, dará positivo para el 95 % de los enfermos y negativo para el 95 % de los no enfermos.

Si, tras hacerme la prueba obtengo un positivo, ¿Debería estar preocupada?

Fijaros que el espacio muestral en el que estamos trabajando contiene dos sucesos, estar enfermo E y no estarlo E^c. Sabemos que la probabilidad de estar enfermo es $P(E) = 0{,}01$, que obtener un positivo (+) cuando se está enfermo es $P(+ \mid E) = 0{,}95$ y que obtener un negativo (-) cuando no se está enfermo es $P(- \mid E^c) = 0{,}95$. Lo que nos interesa saber es la probabilidad de tener la enfermedad sabiendo que hemos obtenido un positivo, esto es: $P(E \mid +)$

Para poder calcular esta cantidad podemos usar el teorema de Bayes y tendremos:

$$P(E \mid +) = \frac{P(+ \mid E)P(E)}{P(+)}.$$

De esta expresión solo nos falta por conocer $P(+)$ que podemos obtener usando el teorema de la probabilidad total como:

$$P(+) = P(+ \mid E)P(E) + P(+ \mid E^c)P(E^c) = 0{,}95 \times 0{,}01 + (1 - 0{,}95) \times 0{,}99 = 0{,}059.$$

Con este resultado obtenemos que:

$$P(E \mid +) = \frac{P(+ \mid E)P(E)}{P(+)} = \frac{0{,}95 \times 0{,}01}{0{,}059} = 0{,}161.$$

Fijaros que, aunque el test era muy fiable, sólo hay un 16 % de posibilidades de que realmente esté enferma. Podemos verlo de manera intuitiva en el diagrama de árbol de la Figura 6.

Tanto en el caso de las aplicaciones en salud como en el mundo de las apuestas, un concepto que aparece muy a menudo es el conocido como odds (o momios en castellano).

Definición Los odds a favor de un evento A son el ratio entre la probabilidad del suceso y la probabilidad de su complementario. Esto es:

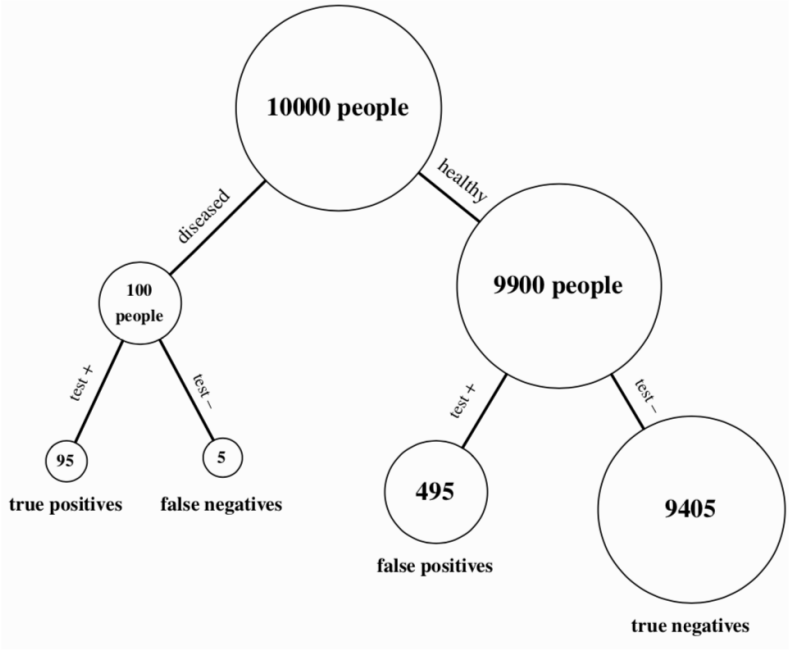


Figura 6: Detección de una enfermedad rara

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

En el fondo estamos estudiando cuando más probable es tener la enfermedad que no tenerla (por ejemplo). Un valor de 2 nos indicaría que es 2 veces más probable estar enfermo que no estarlo.

Los odds también se suelen utilizar mucho en términos de apuestas ya que nos indican cuanto más probable es ganar que perder y se pueden calcular "a priori", es decir, sin saber lo que ha pasado antes, o "a posteriori", una vez conocemos que ha tenido lugar el suceso B. Para ello podemos utilizar el Teorema de Bayes de la siguiente forma:

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

Este teorema nos dice que los odds a favor del suceso A dado el suceso B se pueden calcular como los odds sin condicionar multiplicados por

$$\frac{P(B \mid A)}{P(B \mid A^c)}.$$

A este ratio se le conoce como **Factor Bayes**.

2.1.3. Sucesos independientes

Durante toda esta sección hemos intentado dilucidar la probabilidad de un suceso condicionado a otro $P(A \mid B)$. Y bien ¿Qué sucede si el conocimiento de que ha sucedido B no cambia la probabilidad de que suceda A?

En tal caso podemos afirmar que el suceso A es independiente de B y por tanto $P(A \mid B) = P(A)$. Equivalentemente, si utilizamos la formula de la probabilidad condicionada, tenemos que, dados dos sucesos independientes $P(A \cap B) = P(A)P(B)$

De hecho, en términos matemáticos se dice que:

Definición: Dos sucesos son independientes si

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

Imaginad que lanzo dos monedas balanceadas y quiero saber la probabilidad de que una sea cara dado que la otra es cara. Es decir quiero calcular la probabilidad de $A = \{$cara en la moneda 1$\}$ dado que conozco $B = \{$cara en la moneda 2$\}$. El espacio muestral es Ω =$\{$CC, CX, XC, XX$\}$ y es fácil ver que $P(A \cap B) = 1/4$. Por otra parte, la probabilidad de A es idéntica a la probabilidad de B $P(A) = P(B) = 1/2$. Y, finalmente vemos que $P(A \cap B) = P(A)P(B) = 1/4$

En este ejemplo era fácil entender que ambos sucesos eran independientes puesto que, el resultado obtenido en una moneda no afecta al resultado obtenido en la otra. Pero pensemos en otro ejemplo.

Imaginad que lanzamos un dado y planteamos dos sucesos $A = \{$Obtener un número par$\}$ y $B = \{$Obtener 1, 2, 3 o 4$\}$

La probabilidad de A es claramente $1/2$ mientras que la probabilidad de B es igual a $2/3$. Por otra parte, la probabilidad de $A \cap B$ en este caso corresponde a los números pares de B, esto es $\{$2 y 4$\}$ y será $1/3$. Por tanto $P(A \cap B) = P(A)P(B) = 1/3$ y los sucesos son independientes.

En este caso resulta más difícil de entender porque estamos hablando del mismo dado y del mismo lanzamiento pero se puede ver de una forma muy sencilla. Imaginad que os hago apostar por si va a salir un número par o impar, como la probabilidad es la misma en los dos casos no me sabríais que decir. Ahora lanzo el dado y os digo que ha salido un número

del 1 al 4, ¿Sabríais entonces que decirme? ¿Os ha servido de algo la información que os he dado?. Evidentemente la respuesta a ambas preguntas es no y, de ahí, la independencia de ambos sucesos.

La independencia de dos sucesos puede extenderse también a sus complementarios. En concreto:

Proposición Si A y B son sucesos independientes, A y B^c también lo serán al igual que A^c y B y que A^c y B^c.

Podemos también hablar de independencia de más de dos sucesos. En el caso concreto de tres sucesos A, B y C, diremos que estos son independientes si se cumplen las cuatro condiciones siguientes:

1. $P(A \cap B) = P(A)P(B)$
2. $P(A \cap C) = P(A)P(C)$
3. $P(B \cap C) = P(B)P(C)$
4. $P(A \cap B \cap C) = P(A)P(B)P(C)$

Es importante tener en cuenta que las tres primeras condiciones (independencia dos a dos) no implican necesariamente la cuarta. Por ejemplo: Consideremos dos lanzamientos independientes de una moneda balanceada. Sea A el suceso *obtener primero cara*; B el suceso *el segundo lanzamiento es cara* y C el suceso *obtener el mismo resultado en ambos lanzamientos*. En este caso A, B Y C son sucesos independientes dos a dos pero no independientes ya que $P(A \cap B \cap C) = 1/4$ mientras que $P(A)P(B)P(C) = 1/8$. El punto es que, saber A o B no nos dice nada sobre C pero si conocer ambos $A \cap B$.

Siguiendo este argumento podemos extender el concepto de independencia a múltiples sucesos de la siguiente forma:

Definición (Independencia de múltiples sucesos) n sucesos $A_1, A_2, \ldots, A_n$ se consideran independientes si

- cualquier par de ellos, $P(A_i \cap A_j) = P(A_i)P(A_j)$ (para $i \neq j$).
- cualquier grupo de tres cumple $P(A_i \cap A_j \cap A_k) = P(A_i)P(A_j)P(A_k)$.
- y lo mismo para cualquier grupo de cuatro, cinco, seis etc.

En ocasiones varios sucesos no son independientes entre si pero sí lo son bajo la premisa de

otro suceso. A este tipo de independencia se le conoce como independencia condicional

Definición: Decimos que los sucesos $A_1 \dots A_k$ son condicionalmente independientes dado un suceso B si, para toda subcolección de sucesos $A_{i_1}, \dots, A_{i_j}$ con $j = 2, 3, \dots, k$

$$P(A_{i_1} \cap \cdots \cap A_{i_j} \mid B) = P(A_{i_1} \mid B) \dots P(A_{i_j} \mid B)$$

Es importante no confundir la independencia con la independencia condicional ya que una no tiene porque implicar a la otra. Veámoslo con algunos ejemplos.

La independencia condicional no implica independencia Supongamos que tenemos dos monedas, una balanceada y otra trucada con probabilidad de cara 3/4. Lanzamos la moneda un número de veces. Si hubiésemos elegido la moneda justa, los lanzamientos serian independientes con probabilidad de cara 1/2 mientras que, si la moneda elegida es la trucada, los lanzamientos son, de nuevo, independientes con probabilidad de cara 3/4. Sin embargo, sin saber cual ha sido la moneda elegida, no podemos afirmar que los lanzamientos sean independientes ya que, observar la secuencia de resultados, nos da información sobre la moneda elegida.

De manera formal. Sea B el suceso *haber elegido la moneda balanceada* y A_1 y A_2 los sucesos *obtener cara en el primer lanzamiento* y *obtener cara en el segundo lanzamiento* respectivamente. A_1 y A_2 son independientes condicionados a B pero no son lo son de forma incondicional puesto que A_1 nos da información sobre A_2

La independencia no implica independencia condicional Supongamos que mis amigas Asun y Belisa son dos las dos únicas personas que me llaman. Cada día, ellas deciden, de manera independiente, si llamarme o no. Sea A el suceso *me llama Asun* y B *me llama Belisa*. A y B son incondicionalmente independientes. Sin embargo, sea S el suceso *el teléfono está sonando ahora mismo* yo se que, o bien es Asun o bien es Belisa, y si no es una, será la otra. Por tanto, dado el suceso S, A y B ya no son independientes.

Independencia condicional y el complementario Supongamos que tenemos dos tipos de clases, las clases buenas y las malas. En las buenas clases, si trabajas duro es muy posible sacar un sobresaliente. En las clases malas, da igual lo que te esfuerces, el profesor asigna la nota aleatoriamente.

Si llamamos A al suceso *obtener un sobresaliente* B al suceso, *estar en una clase buena* y

C al suceso *haber trabajado duro*. En este caso, A y C no son independientes dado B pero si lo son dado B^c

2.2. Ejercicios

1. Cada vez que un cliente compra una determinada pasta de dientes elige entre la marca A o B. Supongamos que si ha comprado una determinada marca, la probabilidad de que repita en la siguiente compra es 1/3. Si es igual de probable que, en la primera compra elija A o B ¿Cual es la probabilidad de que en la primera y la segunda compra elija la marca A?

2. Una caja contiene tres monedas con una cara en ambos lados, cuatro monedas con una cruz en cada lado y dos con cara y cruz. Si elegimos al azar una de esas 9 monedas y la lanzamos, ¿Qué probabilidad hay de obtener cara?

3. El porcentaje de personas con gafas en los tres barrios de una ciudad son, 30 para el primero , 25 en el segundo y 45 en el tercero. Teniendo en cuenta que 1/4 de la población vive en el primer barrio, 2/4 en el segundo y 1/4 en el tercero ¿Qué probabilidad hay de que una persona elegida al azar tenga gafas?

4. De acuerdo con la cifras del INE (Instituto Nacional de Estadística), los hombres que fuman tienen 23 veces más probabilidad de desarrollar cáncer de pulmón que aquellos que no fuman. El mismo estudio informa de que un 21 % de los hombres españoles fuman. ¿Cual es la probabilidad de que un hombre fumase dado que ha desarrollado cáncer de pulmón?

5. Las pantallas que se usan en un tipo de móviles pueden ser fabricadas por tres compañías diferentes A, B o C. La proporción de pantallas elaboradas por cada una de ellas es 0.5, 0.3 y 0.2 respectivamente. Se sabe que el 0.01 de las fabricadas por A, el 0.02 de las que fabrica B y el 0.03 de las elaboradas por C son defectuosas. Dado que la pantalla de un teléfono es defectuosa, que probabilidad hay de que la haya fabricado A.

6. La compañía A ha desarrollado un test diagnostico para una determinada enfermedad que sólo afecta al 1 % de la población. La sensibilidad (probabilidad de dar positivo en alguien que tiene la enfermedad) y la especificidad (probabilidad de dar negativo en alguien que no tiene la enfermedad) del test son ambas del 95 %

Una nueva compañía B, para competir con A, ofrece un nuevo test que dice que detecta la enfermedad con mayor facilidad. En concreto, B afirma que la sensibilidad de su test es del 98 % aunque su especificidad se reduce al 90 %.

Ante un resultado positivo, ¿Con que test estarías más seguro/a de tener la enfermedad? ¿Se te ocurre cuando es mejor usar el test proporcionado por A y cuando el proporcionado por B?

7. Supongamos que hay 5 tipos de sangre en la probación cada uno con probabilidad $p_1, p_2, \ldots, p_5$. Sabemos que un crimen ha sido cometido por dos individuos. Tenemos un sospechoso que tiene un tipo de sangre 1 y una probabilidad p de ser culpable. En el escenario del crimen se ha descubierto que uno de los criminales tenía sangre de tipo 1 y, el otro, de tipo 2.

   Tras este descubrimiento, ¿La probabilidad de que el sospechoso sea culpable aumenta o disminuye? ¿Depende esta probabilidad *a posteriori* de las probabilidades $p, p_1, \ldots, p_5$?

8. Consideremos cuatro dados no estándar (el dado de Efron) cuyas caras están numeradas de la siguiente forma: (las seis caras de cada dado son igual de probables

   A: 4,4,4,4,0,0

   B: 3,3,3,3,3,3

   C: 6,6,2,2,2,2

   D: 5,5,5,1,1,1

   Estos cuatro dados son lanzados una vez cada uno. Sea A el resultado del dado A, B el resultado del dado B etc.

   a) Encuentra $P(A > B)$, $P(B > C)$, $P(C > D)$, y $P(D > A)$.

   b) ¿Es el suceso $A > B$ independiente del suceso $B > C$? ¿Es el suceso $B > C$ independiente del suceso $C > D$? Explica por qué

9. Supongamos que existen dos tipos de conductores, los buenos y los malos. Sea G el suceso de *cierta persona es un buen conductor*; A el suceso *el conductor se ve involucrado en un accidente este año* y B el suceso *el conductor se ve involucrado en un accidente el próximo año.*

Sea $P(G) = g$ y $P(A|G) = P(B|G) = p_1$, $P(A|G^c) = P(B|G^c) = p_2$, con $p_1 < p_2$. Supongamos que, dado que sabemos que es un buen conductor o no, A y B son independientes. (También que los accidentes son leves y que el conductor puede seguir conduciendo)

a) Explica intuitivamente si A y B son independientes o no

b) Calcula $P(G|A^c)$.

c) Calcula $P(B|A^c)$.

10. Una familia tiene tres hijos/as y cada uno de ellos es niño o niña con la misma probabilidad. Si definimos los eventos

A: todos tienen el mismo sexo

B: hay como máximo un chico

C: la familia tiene un chico y una chica al menos

a) Muestra que A es independiente de B y que B es independiente de C

b) ¿Es A independiente de C?

c) Se mantienen los resultados anteriores si la probabilidad de niño no es la misma que la de niña.

3. Variables aleatorias y distribución de probabilidad

3.1. Introducción.

Cuando los experimentos se complican la notación en términos de sucesos y, por tanto, de conjuntos puede ser complicada.

Imaginemos un experimento que consiste en observar cuantos coches pasan por las calles de la ciudad durante una hora de un día determinado.

Podríamos empezar definiendo A_{kj} como el suceso: en la calle A pasan k coches en el minuto j. Complicado ¿no? Y, si además, empezamos a interesarnos por cosas como el tráfico total entre las calles A y B o el tráfico máximo durante esa hora, resulta imposible expresarnos en términos de sucesos.

¿No sería mucho más fácil, en lugar de trabajar con sucesos y conjuntos, trabajar con números reales? Por ejemplo, que el evento A_{43} se representase simplemente por, digamos, el número 4. Así, podríamos sumarlo con el número de coches que pasan en ese mismo minuto por la calle B, digamos 3.

3.2. Variables aleatorias.

3.2.1. Definición de variable aleatoria.

Esta "simplificación" metodológica es la que se encuentra detrás del concepto de variable aleatoria.

Definición: (Variable Aleatoria). Dado un experimento con un espacio muestral Ω, una variable aleatoria es toda aquella función que transforma los sucesos del espacio muestral en números reales.

Por ejemplo, imaginad que lanzamos una moneda 10 veces. El espacio muestral tendrá 2^{10} elementos:

C C C C C C C C C C
X C C C C C C C C C
X X C C C C C C C C
. . . .

Ahora pensad que estamos interesad@s en el número de caras obtenidas. Podemos definir la variable aleatoria X evaluada sobre un suceso S, $X(S)$ como el número de caras de dicho

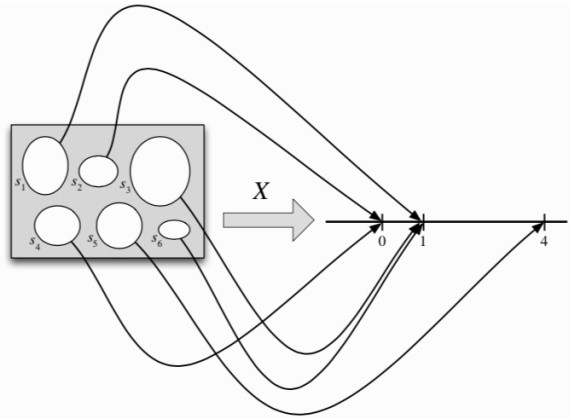
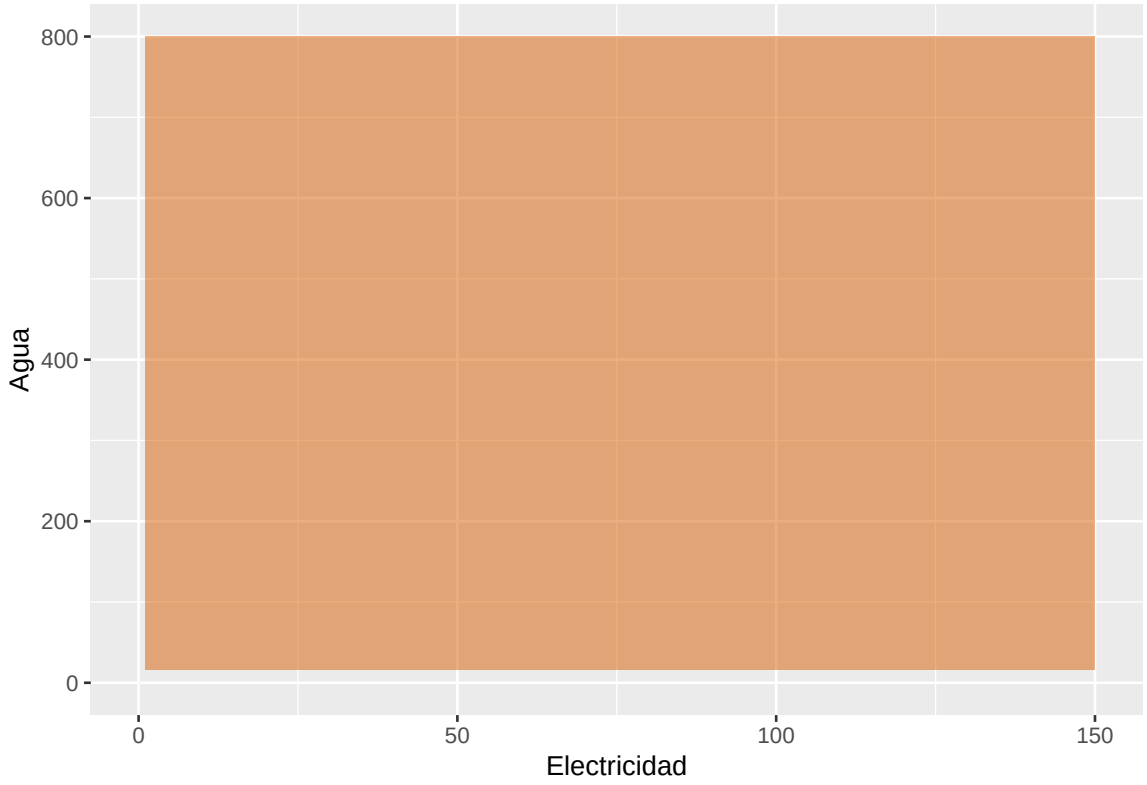


Figura 7: Una variable aleatoria convierte cada suceso en un valor de la recta real.

suceso. Por ejemplo, si S es el suceso CCXXCCXXXX, $X(S) = 4$. Los posibles valores de X serán $0, 1, \ldots, 10$

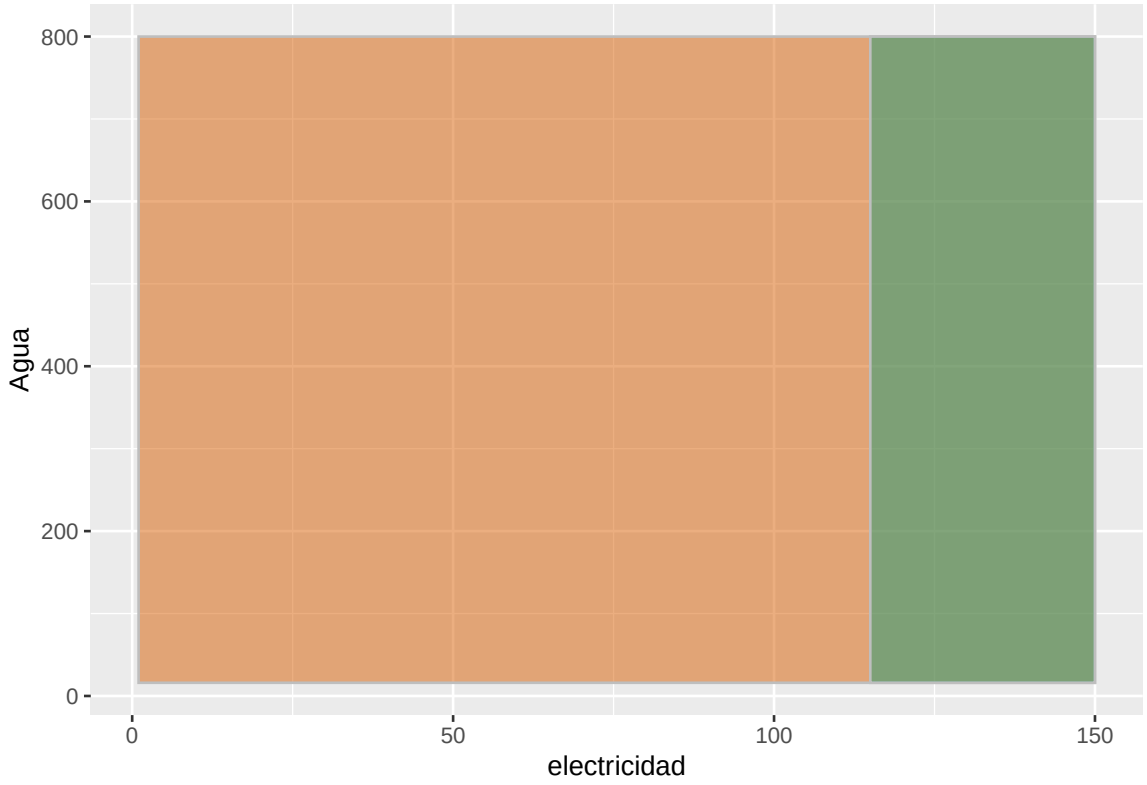
Pensemos ahora en una empresa de construcción que está preocupada por la posible demanda de agua y de electricidad en un nuevo edificio de viviendas. Se sabe que la demanda de agua se mueve entre los 16 y los 800 L por día mientras que la electricidad se mueve entre los 1 y los 150 kw/h al día:

```
d=data.frame(x1=c(1,0), x2=c(150,0), y1=c(16,0), y2=c(800,0),t=c("S","A"))


ggplot() + geom_rect(data=d, mapping=aes(xmin=x1, xmax=x2, ymin=y1, ymax=y2, fill=t), a
```



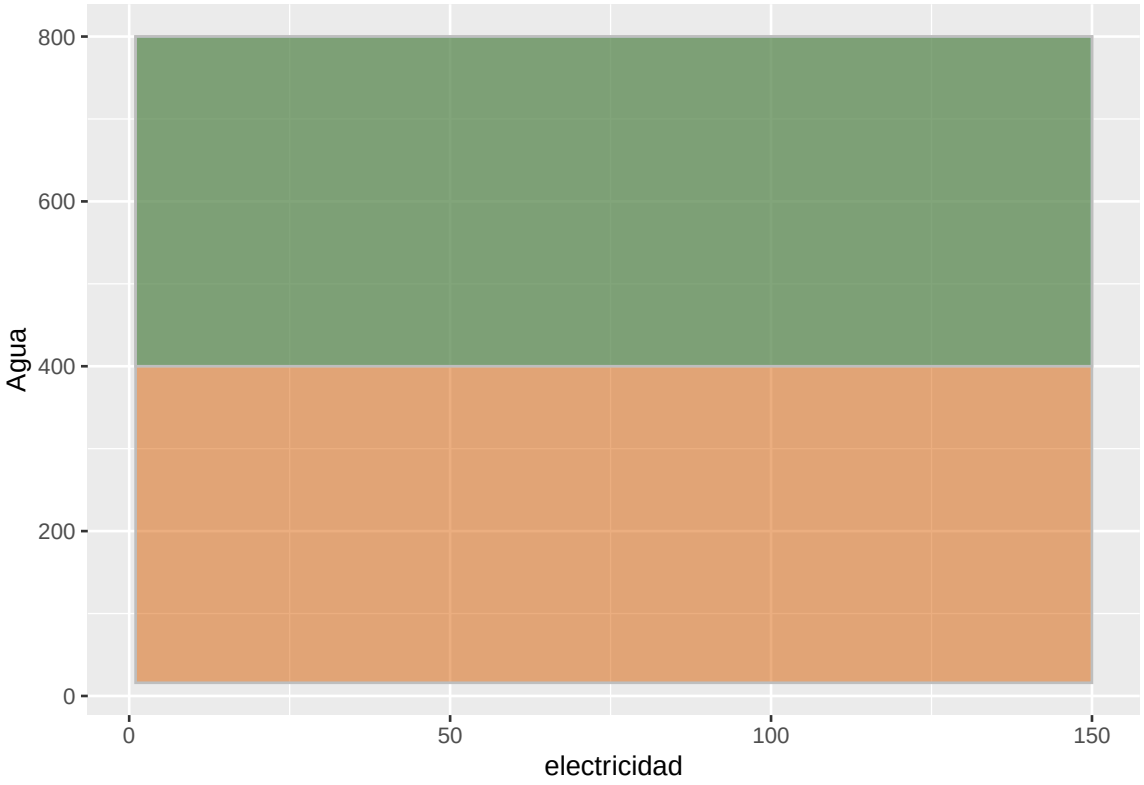
Evidentemente, pensar en sucesos en este escenario es complejo. Podríamos pensar en el suceso, que la demanda de electricidad sea alta (entendiendo como alta que sea mayor de 115 kw/h por día)

```
d=data.frame(x1=c(1,115), x2=c(150,150), y1=c(16,16), y2=c(800,800),t=c("S","A"))


ggplot() + geom_rect(data=d, mapping=aes(xmin=x1, xmax=x2, ymin=y1, ymax=y2,fill=t), co
```



O el suceso que la demanda de agua se alta (mayor de 400 litros al día)

```
d=data.frame(x1=c(1,1), x2=c(150,150), y1=c(16,400), y2=c(800,800),t=c("S","B"))


ggplot() + geom_rect(data=d, mapping=aes(xmin=x1, xmax=x2, ymin=y1, ymax=y2,fill=t), co
```



Fijaos que, todos los sucesos descritos se pueden expresar como un conjunto de puntos X, Y donde X es la demanda de agua e Y la demanda de electricidad, y cualquier punto del espacio muestral Ω se corresponde con un conjunto de números (x, y)

Incluso podríamos definir una tercera variable Z que, para cada punto del espacio muestral Ω, Indicará si la demanda es alta o no:

$$Z(s) = \begin{cases} 0 & \text{si } x > 115 \ \& \ y > 400 \\ 1 & \text{si no} \end{cases}$$

La aleatoriedad de la variable viene dada por la propia aleatoriedad del experimento ya que antes de realizarlo no sabíamos cual iba a ser el resultado para la variable. De la misma forma, podemos utilizar la definición de probabilidad para valorar como de verosímiles son los posibles resultados de la variable X con la ventaja de haber simplificado considerablemente la definición del espacio muestral.

Es importante, sin embargo, no perder de vista que existe una conexión entre el espacio muestral y la variable aleatoria. Esta conexión es la que nos permite utilizar la definición de probabilidad.

Por ejemplo, podemos resumir los resultados de 10 lanzamientos de una moneda mediante el número de caras. Pero, para pensar en como de probable es obtener 4 retornaremos, consciente o inconscientemente, a la idea inicial del espacio muestral intentando contar cuantos resultados son favorables de entre los 2^{10} posibles.

Nota: Fijaos que las variables aleatorias se denotan usando letras mayúsculas, por ejemplo, X mientras que los posibles valores de esta variable se representan con letras minúsculas, por ejemplo, x.

3.3. Distribución de una variable aleatoria.

Bien, hemos definido lo que es una variable aleatoria pero lo que realmente queremos es entender su comportamiento. ¿Cuál es el rango de valores que sucederá con mayor probabilidad? ¿Qué puedo esperar que suceda?

Pensemos en el ejemplo de la moneda, ¿cuál es la probabilidad de obtener al menos dos caras($X \geq 2$)? O, en el ejemplo de la constructora, ¿qué probabilidad hay de que la demanda de electricidad esté entre 100 y 120 kw/h?

Para poder describir estas y otras características probabilísticas de la variable que estamos estudiando utilizamos su *distribución.*

Definición: sea X una variable aleatoria, definimos su **distribución** como la colección de todas las probabilidades $P(X \in C)$ siendo C cualquier subconjunto de los números reales tal que $X \in C$ representa un suceso.

Como ya hemos comentado al final de la sección anterior, la probabilidad sobre los elementos del espacio muestral de la variable X viene inducida por la probabilidad definida sobre el espacio muestral Ω del experimento original. En este sentido $P(X \in C)$ se define como la probabilidad del suceso formado por los resultados del experimento s tal que $X(s) \in C$.

Volvamos al ejemplo en que lanzábamos la moneda 10 veces y donde la variable X representaba el número de caras obtenido. Esta variable puede obtener valores $\{0, \ldots, 10\}$ y cada uno de ellos tiene asociados un número r de los 2^{10} posibles resultados del experimento. Por ejemplo, elegimos $X = 2$ está asociado con todos los resultados en los que se obtienen exactamente dos caras, esto es $r = \binom{10}{2}$. Si suponemos que todos los resultados son igual de probables, la probabilidad de obtener dos caras será $r/2^{10}$ o, escrito de otra forma:

$$P(X=2) = \binom{10}{2}\left(\frac{1}{2}\right)^{10} \approx 0{,}044$$

En el ejemplo de la empresa constructora, podemos calcular la probabilidad haciéndola proporcional al área de los cuadrados asociados a los sucesos que nos interesan. Por tanto, La probabilidad de que la demanda de electricidad, Y, sea superior a 115 kw/h (Figura 2) será el área del cuadrado verde ($(150-115)\times(800-16) = 26440$) dividida por el área del cuadrado grande ($(150-1)\times(800-16) = 116816$). Esto es:

$$P(115 \leq Y \leq 150) = \frac{(150-115)\times 784}{116816} = 0{,}235.$$

Este cálculo podemos extenderlo a cualquier intervalo de valores $C = [c_1, c_2]$:

$$P(c_1 \leq Y \leq c_2) = \frac{(c_2-c_1)\times 784}{116816}$$

A lo largo de este capítulo y el siguiente veremos formas más elegantes de asignar probabilidades a los valores de las variables aleatorias. Para hacerlo, distinguiremos entre variables discretas y continuas.

3.3.1. Variables discretas. Función de probabilidad.

Empecemos por definir lo que entendemos por una variable discreta:

Definición: una variable **discreta** X es aquella que sólo puede tomar un conjunto de valores finitos $a_1, \ldots, a_n$ o infinitos pero contables $a_1, a_2, \ldots$.

Para entender la distribución de una variable aleatoria discreta debemos hablar de su función de probabilidad. Habitualmente podréis encontrarla escrita en inglés como *probability function* (p.f) o *probability mass function* (p.m.f), y se define como:

Definición: la función de probabilidad de una variable aleatoria discreta es la función p_X dada por: $p_X(x) = P(X = x)$. Al conjunto de valores de x donde $p_X(x) > 0$ se le denomina **soporte** de X.

Fijaos que, al escribir $X = x$ estamos denotando un suceso que consiste en todos los posibles resultados del experimento que asignen a X el valor x. De manera formal podríamos escribir: $s \in S : X(s) = x$, pero, escribir $X = x$ es más corto y más intuitivo.

Volviendo al ejemplo de las monedas, $X = 3$ denotaría todos aquellas tiradas en las que 3 de las 10 monedas hayan resultado ser cara. En concreto, para este ejemplo, podríamos

determinar las probabilidades de todos los valores que conforman el soporte de X, esto es: $0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$

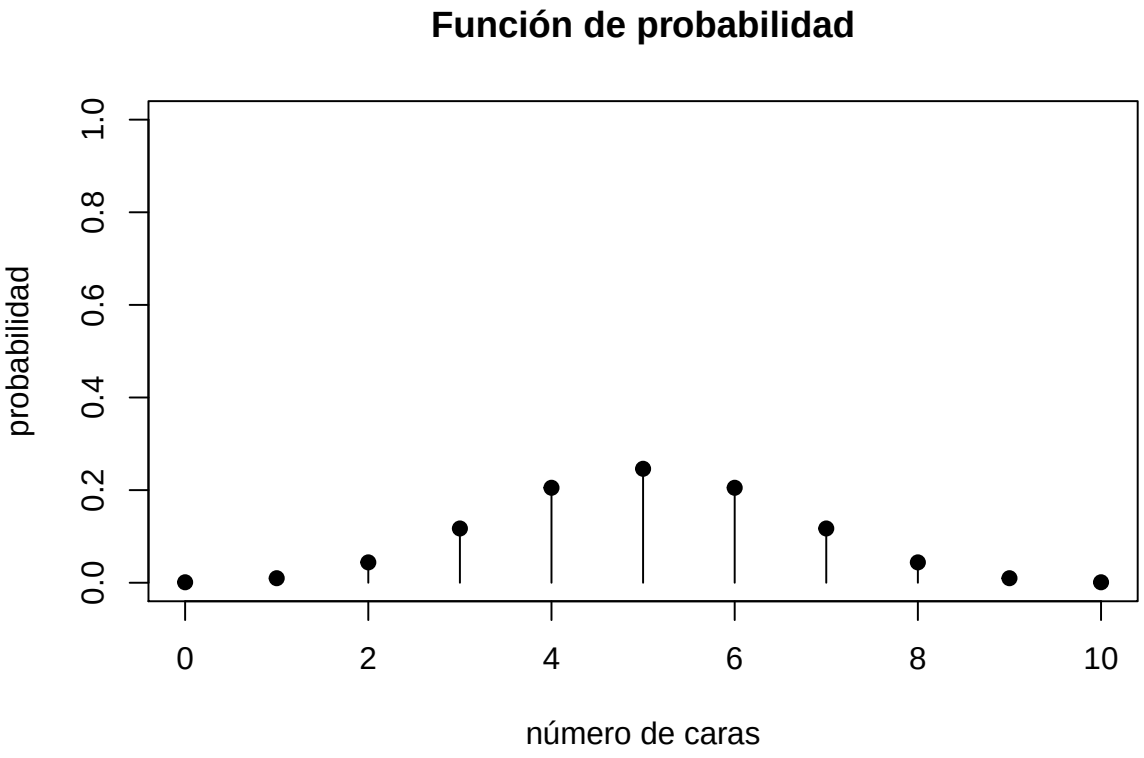
$$
\begin{aligned}
P(X = 0) &= \binom{10}{0} \frac{1}{2^{10}} \\
P(X = 1) &= \binom{10}{1} \times \frac{1}{2^1} \times \frac{1}{2^9} \\
P(X = 2) &= \binom{10}{2} \times \frac{1}{2^2} \times \frac{1}{2^8} \\
&\dots \\
P(X = k) &= \binom{10}{k} \times \frac{1}{2^k} \times \frac{1}{2^{10-k}}
\end{aligned}
$$

Algunas funciones de probabilidad son tan conocidas y usadas que tienen nombre propio. En concreto, la función que expresa la probabilidad de un número dado de éxitos en N intentos se llama Binomial. Esta y otras distribuciones conocidas las estudiaremos en el próximo tema.

Nota: dos variables aleatorias pueden tener la misma distribución (es decir, usar la misma función de probabilidad) sin ser la misma variable. Por ejemplo, cualquier variable aleatoria que sólo pueda tomar dos valores con la misma probabilidad, tendrá la misma distribución que el número de caras en el lanzamiento de una moneda (0 o 1).

Como toda función, una función de probabilidad puede dibujarse. Para la función de probabilidad del experimento de las monedas tenemos:

```
plot(0:10,dbinom(0:10, size = 10, prob = 1/2),type="h", ylab= "probabilidad", xlab = "n
points(0:10,dbinom(0:10, size = 10, prob = 1/2), pch=19)
```



Fijaos que se trata de una función que sólo toma valores en algunos puntos. La altura a la que se encuentran estos puntos es la probabilidad de ese valor. Suele añadirse una línea vertical para que se observe mejor la magnitud de dicha probabilidad.

En términos generales, cualquier función de probabilidad debe cumplir las siguientes condiciones:

Teorema: sea X una variable aleatoria discreta con función de probabilidad p_X cuyo soporte es $x_1, x_2, \ldots$:

1. $p_X(x) > 0$ si $x \in \{x_1, x_2, \ldots\}$ y $p_X(x) = 0$ si no.
2. $P(X \in C) = \sum_{x_i \in C} p_X(x_i)$
3. $\sum_{i=1}^{\infty} p_X(x_i) = 1$

3.3.2. Variables continuas. Función de densidad.

Una vez hemos visto lo que significa que una variable sea discreta la definición de variable continua parece obvia. De forma simple podríamos decir que una variable aleatoria es continua cuando puede tomar cualquier valor en un intervalo de la recta real. Pero, ¿qué significa que *puede tomar cualquier valor*? Bien, para formalizar esta definición debemos pensar, de nuevo, en términos de probabilidad:

Definición: decimos que una **variable aleatoria** es **continua** o que tiene una **distribución continua** si existe una función no negativa f_X, definida en la recta real, tal que, para todo intervalo de números reales (acotado o no acotado) la probabilidad de que X tome un valor en dicho intervalo es la integral de f sobre dicho intervalo:

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

De esta definición se deduce que la distribución de una variable aleatoria continua X (es decir, el comportamiento de ésta, que valores son más probables y cuales menos, etc.) queda totalmente caracterizado por la función f. Podemos decir entonces, que f_X juega el mismo papel para una variable aleatoria continua que la función de probabilidad en el caso de una variable aleatoria discreta y merece, por tanto, su propio nombre y definición.

Definición: sea X una variable aleatoria continua, la función f_X que caracteriza su distribución de probabilidad recibe el nombre de **función de densidad** (en inglés: *probability density function*, p.d.f) y el conjunto $\{x : f_X(x) > 0\}$ recibe el nombre de **soporte de** X.

Una función de densidad debe cumplir los siguientes requisitos:

1. Ser no negativa:

$$f_X(x) \geq 0 \ \ \forall x$$

2. Integrar 1 (o encerrar un área de 1) en la recta real:

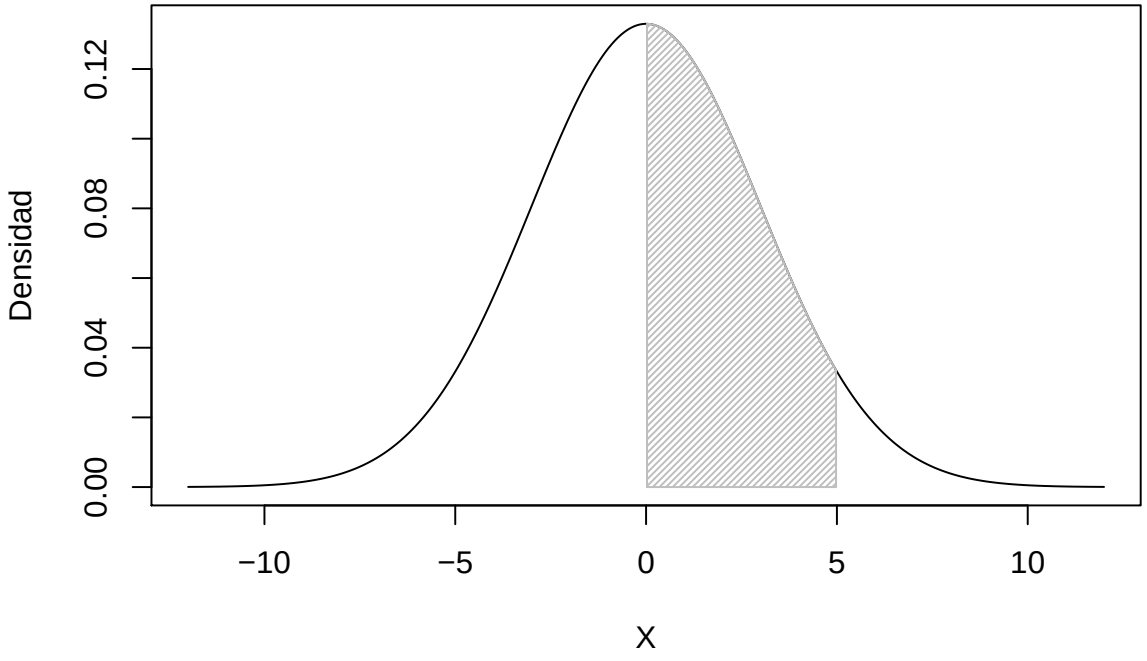
$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Un típico ejemplo de una función de densidad podemos verlo en la siguiente figura, donde el área sombreada representa la probabilidad de que X esté en el intervalo $[0, 5]$:

```
shadenorm <- function(below=NULL, above=NULL, pcts = c(0.025,0.975), mu=0, sig=1, numpt
                      justabove= FALSE, justbelow = FALSE, lines=FALSE,between=NULL,outsi

    if(is.null(between)){
         below = ifelse(is.null(below), qnorm(pcts[1],mu,sig), below)
         above = ifelse(is.null(above), qnorm(pcts[2],mu,sig), above)
    }
```

```
if(is.null(outside)==FALSE){
     below = min(outside)
     above = max(outside)
}
lowlim = mu - 4*sig
uplim  = mu + 4*sig

x.grid = seq(lowlim,uplim, length= numpts)
dens.all = dnorm(x.grid,mean=mu, sd = sig)
if(lines==FALSE){
      plot(x.grid, dens.all, type="l", xlab="X", ylab="Densidad")
}
if(lines==TRUE){
      lines(x.grid,dens.all)
}

if(justabove==FALSE){
    x.below    = x.grid[x.grid<below]
    dens.below = dens.all[x.grid<below]
    polygon(c(x.below,rev(x.below)),c(rep(0,length(x.below)),rev(dens.below)),col=c
}
if(justbelow==FALSE){
    x.above    = x.grid[x.grid>above]
    dens.above = dens.all[x.grid>above]
    polygon(c(x.above,rev(x.above)),c(rep(0,length(x.above)),rev(dens.above)),col=c
}

if(is.null(between)==FALSE){
     from = min(between)
     to   = max(between)

     x.between    = x.grid[x.grid>from&x.grid<to]
     dens.between = dens.all[x.grid>from&x.grid<to]
```

```
        polygon(c(x.between,rev(x.between)),c(rep(0,length(x.between)),rev(dens.betwee
    }

}
## ---------------------- ##
## Shading under a Normal ##
## in R                   ##
## ---------------------- ##

shadenorm(mu=0, sig=3, between=c(0, 5)) ## works with between and outside
```



En el ejemplo de la demanda de electricidad vimos que la probabilidad de que Y esté en un intervalo $C = [c_1, c_2]$ era:

$$P(c_1 \leq Y \leq c_2) = \frac{(c_2 - c_1) \times 784}{116816} = \frac{(c_2 - c_1)}{149} = \int_{c_1}^{c_2} \frac{1}{149} dy$$

Podemos definir, entonces, la función de densidad para Y como:

$$f(y) = \begin{cases} \frac{1}{149} & \text{si } 1 \leq y \leq 150 \\ 0 & \text{en otro caso} \end{cases}$$

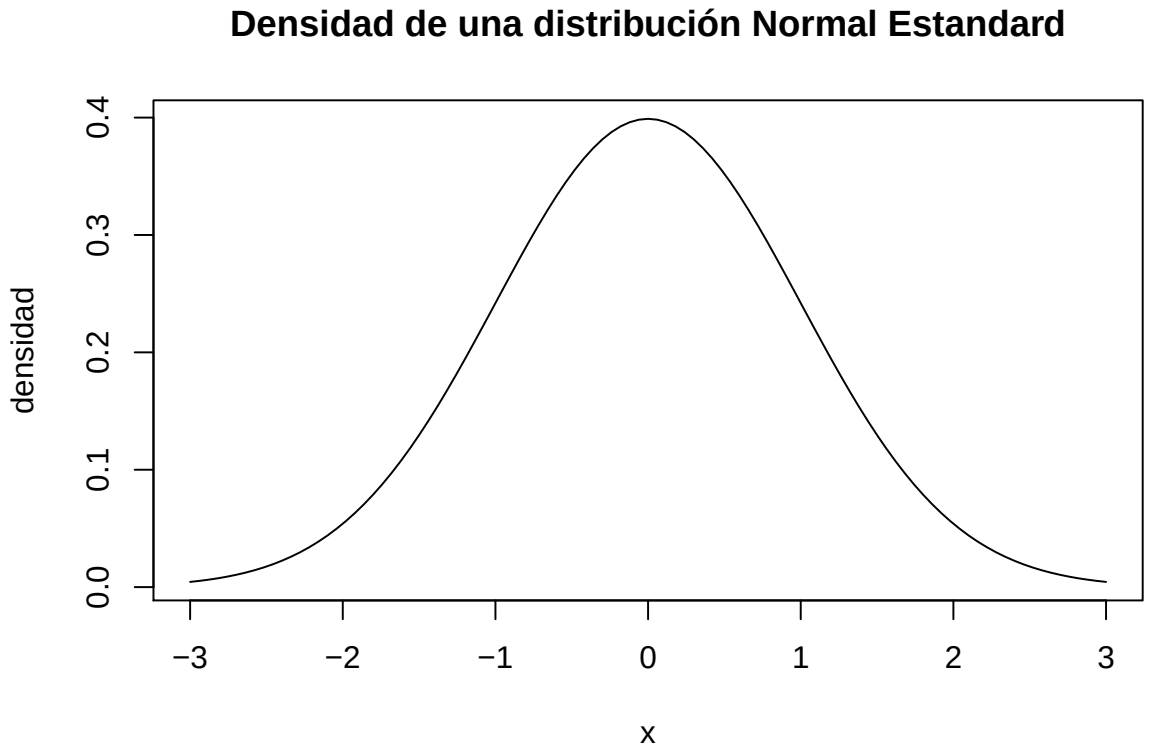
Al igual que comentamos en el caso de las funciones de probabilidad, en el Tema 3, veremos como algunas funciones de densidad se usan de forma habitual y reciben un nombre que

las identifica de forma única. En concreto, la distribución continua más conocida es la distribución *Normal* cuya función de densidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

donde μ y σ son parámetros cuyo significado estudiaremos más adelante. En particular, cuando $\mu = 0$ y $\sigma = 1$ se le conoce como distribución *Normal Estandard* y su densidad es:

```
curve(dnorm(x),from = -3, to = 3, ylab="densidad", main="Densidad de una distribución N
```



Nota 1: una distribución continua asigna probabilidad 0 a valores individuales. Es fácil ver que, dada la definición de la probabilidad a partir de la integral de una función continua, la probabilidad de un valor aislado será 0. Esto es:

$$P(X = a) = P(a \leq X \leq a) = \int_a^a f_X(x)dx = 0.$$

De esta forma, si X es una variable aleatoria continua $P(X \leq x) = P(X < x)$ al contrario de lo que sucede en el caso de una variable aleatoria discreta.

Nota 2: la función de densidad no es una función de probabilidad. Es importante darse cuenta, también, que la función de densidad f_X no tiene porque estar acotada (es decir, puede tomar valores > 1) y, por tanto, no puede entenderse como una función de probabilidad.

Un ejemplo de esta situación es la siguiente función de densidad:

$$f_X(x) = \begin{cases} \frac{2}{3}x^{-1/3} & \text{para } 0 < x \leq 1 \\ 0 & \text{en otro caso} \end{cases}.$$

Es fácil ver que se trata de una función no acotada cerca de 0 (tiende a ∞) pero, sin embargo, si cumple los requisitos necesarios para ser una función de densidad

Nota 3: constante normalizadora. El hecho de que una función de densidad tenga que integrar 1 en toda la recta real, permite que esta pueda quedar definida a falta de una constante. Por ejemplo, si para la función de densidad de la nota 2, escribimos

$$f_X(x) = \begin{cases} cx^{-1/3} & \text{para } 0 < x \leq 1 \\ 0 & \text{en otro caso} \end{cases},$$

e integramos sobre la recta real, obtenemos

$$\int_{-\infty}^{\infty} f(x)dx = c\frac{3}{2}.$$

Es fácil ver, entonces que c debe ser $\frac{2}{3}$ si queremos que f sea una función de densidad.

Cuando definimos una función de probabilidad o densidad a falta de una constante hablamos de proporcionalidad y utilizamos el símbolo $\propto$. De esa forma si $f(x) = cx$ podemos escribir $f(x) \propto x$. Esta propiedad será muy útil en el último tema cuando aprendamos a simular de una distribución.

Nota 4: f_X para densidad p_X para probabilidad. Al definir la función de probabilidad de una variable aleatoria discreta usábamos la notación $p_X(x)$ mientras que la función de densidad de una variable aleatoria continua la denotábamos mediante $f_X(x)$. Es importante saber que, en algunos textos, es posible encontrar ambas denotadas como f_X. En ese caso el contexto nos dirá si estamos trabajando con la función de probabilidad o la de densidad.

Del mismo modo, es posible omitir el subíndice de f_X y p_X y denotarlas simplemente por f y p cuando queda claro a que variable nos estamos refiriendo.

3.3.3. Función de distribución acumulada.

Hemos visto que la distribución de una variable aleatoria discreta queda caracterizada por su función de probabilidad, mientras que la de una variable aleatoria continua lo hace por

su función de densidad. Sin embargo, ambos tipos de variables quedan caracterizados por la *función de distribución acumulada*.

Definición: dada una variable aleatoria X (continua o discreta) definimos su **función de distribución acumulada** (c.d.f. en inglés) F_X como $F_X(x) = P(X \leq x)$.

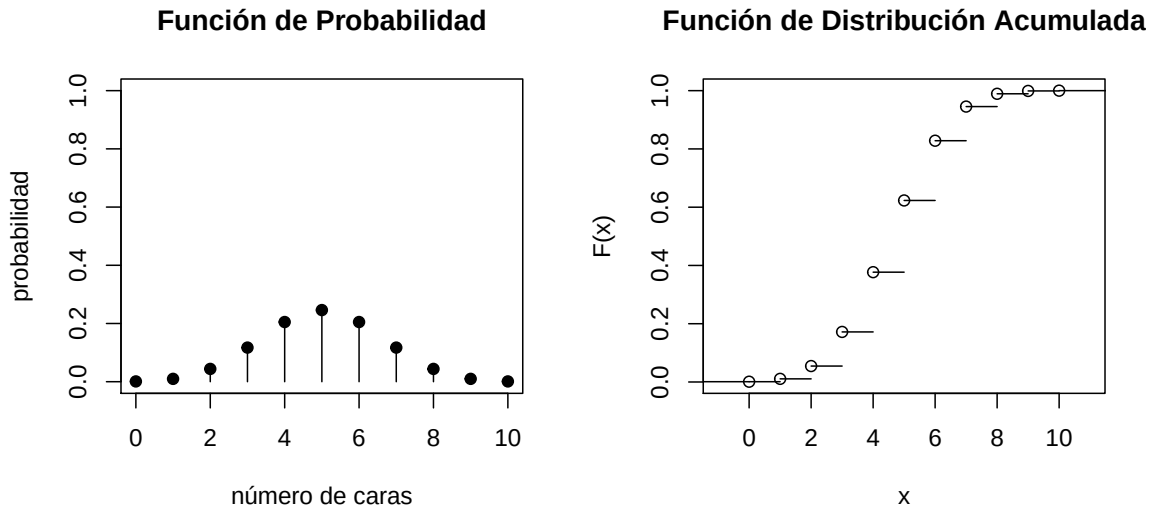
Fijaos que esta función esta definida tanto para variables discretas como para variables continuas. Lo que cambiará será su cálculo. Si X es una variable discreta, su función de distribución acumulada se calculará como:

$$F_X(x_j) = \sum_{x_i \leq x_j} p_X(x_i).$$

Si dibujamos esta función veremos que tendrá forma de escalera, es decir, será una función definida *a trozos*.

Volvamos por ejemplo al lanzamiento de las 10 monedas.

```
par(mfrow=c(1,2))

plot(0:10,dbinom(0:10, size = 10, prob = 1/2),type="h", ylab= "probabilidad", xlab = "n
points(0:10,dbinom(0:10, size = 10, prob = 1/2), pch=19)

plot(stepfun(0:10, pbinom(c(0,0:10), size=10, prob=.5)), verticals=FALSE, ylab="F(x)",m
```



En el caso de que X sea una variable continua, en lugar de utilizar la función de probabilidad deberemos usar la función de densidad. Del mismo modo, no usaremos una suma discreta

si no su equivalente continuo, la integral. Así, la función de distribución de una variable continua se calculará como:

$$F_X(x) = \int_{-\infty}^{x} f_X(x).$$

Podéis ver que esta definición tiene mucho que ver con la definición de la distribución de probabilidad en una variable aleatoria continua de la sección anterior.

Pensemos en una variable aleatoria X que representa el voltaje de un cierto sistema eléctrico. Se sabe que la función de densidad para esta variable es:

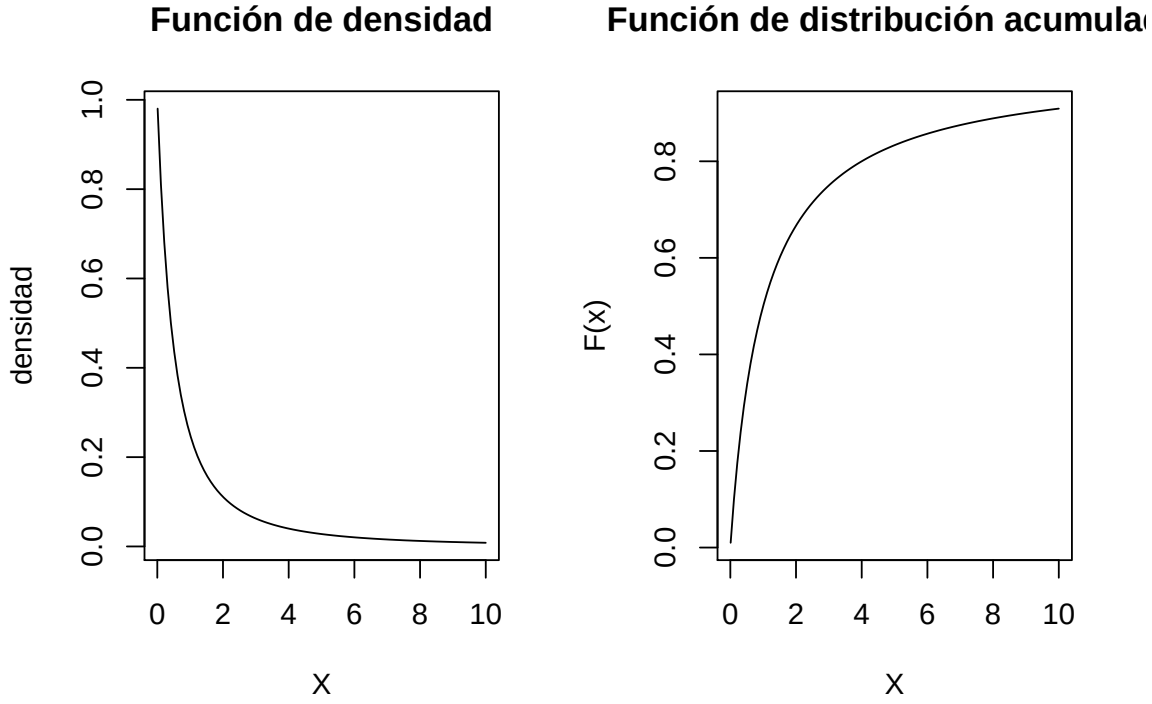
$$f_X(x) = \begin{cases} 0 & x \leq 0 \\ \frac{1}{(1+x)^2} & x > 0 \end{cases}.$$

Si lo que queremos conocer es la probabilidad de que X sea menor que un determinado valor x usaremos su función de distribución:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du = \begin{cases} 0 & x \leq 0 \\ \int_0^x \frac{1}{(1+u)^2}du & x > 0 \end{cases} = \begin{cases} 0 & x \leq 0 \\ 1 - \frac{1}{1+x} & x > 0 \end{cases}$$

así, la probabilidad de que el voltaje sea menor que 3 puede calcularse como $P(X \leq 3) = F_X(3) = 3/4$.

```
par(mfrow=c(1,2))
aux <- seq(0.01,10, length.out = 100)
plot(aux, 1/(1+aux)^2, type="l", main="Función de densidad", ylab = "densidad",xlab="X"


plot(aux, 1-1/(1+aux), type="l", main="Función de distribución acumulada", ylab = "F(x)
```



Teorema: cualquier función de distribución acumulada cumple las siguientes propiedades:

1. Es creciente: Si $x_1 \leq x_2$, se cumple que $F(x_1) \leq F(x_2)$.
2. Es continua por la derecha.
$$\lim_{x \to a^+} F(x) = F(a).$$
3. Converge a 0 y a 1 en los límites:
$$\lim_{x \to -\infty} F(x) = 0; \quad \lim_{x \to \infty} F(x) = 1.$$

A partir de la función de distribución, podemos calcular la probabilidad de cualquier suceso, en concreto:

Teorema: para cualquier valor x

$$P(X > x) = 1 - F_X(x)$$

Teorema: para dos valores x_1 y x_2 cualesquiera con $x_1 < x_2$,

$$P(x_1 \leq X \leq x_2) = F_X(x_2) - F_X(x_1).$$

Volviendo al ejemplo del voltaje, si queremos calcular la probabilidad de que X esté en el intervalo $[2, 4]$. Es decir, $P(2 \leq X \leq 4)$. Con la función de distribución acumulada podemos calcular $P(X \leq 4)$ y $P(X \leq 2)$ y, a partir de ellas la probabilidad buscada:

$$P(2 \leq X \leq 4) = F_X(4) - F_X(2) = \frac{4}{5} - \frac{3}{4} = \frac{1}{20}.$$

Otra particularidad interesante de la función de distribución acumulada es el cálculo de los **cuantiles**. Veámoslo con un ejemplo:

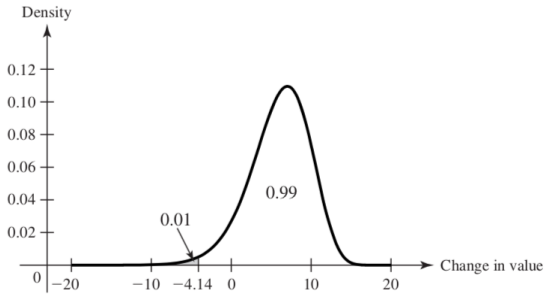
Supongamos que X es el número de lacasitos rojos que hay en un bote y que conocemos su función de distribución acumulada F. En mi grupo de amigas queremos apostar sobre cuantos hay. La dinámica del juego es, que si el bote tiene menos de x_0 lacasitos rojos $X \leq x_0$ ganamos un euro pero si es mayor $X \geq x_0$ perdemos un euro. Para que tengamos la misma probabilidad de ganar que de perder tenemos que buscar un x_0 tal que $P(X \leq x_0) = P(X > x_0) = 1/2$.

Podemos ponernos a buscar un número entero que cumpla que $F(x) = 1/2$ y elegir este como el valor por el que apostaremos (x_0) pero, si F es una función biyectiva tal que tiene una inversa F^{-1}, $x_0 = F^{-1}(1/2)$. A x_0 se le llama cuantil 0.5 o percentil 50 % de X

En general:

Definición: llamamos **Cuantil o Percentil** asociado a una probabilidad p, al valor $F_X^{-1}(p)$ definido como el valor más pequeño del soporte de X que cumple que $F(x) \geq p$. La función F_X^{-1} recibe el nombre de **función cuantil** de X.

Veamos un ejemplo en el caso continuo. El gestor de una cartera de inversiones esta interesado en cuanto dinero podría perder la cartera en un horizonte de tiempo dado. Para ello define X como el cambio en el valor de la cartera en un mes. Supongamos que X tiene la función de densidad que vemos en la siguiente figura:



El gestor quiere establecer un nivel de confianza sobre como de grande podría ser la perdida. En concreto quiere que saber cual es el valor por debajo del cual el cambio (X) sólo estará con probabilidad 0.01. Matemáticamente esto es: $P(X < x_0) = 0{,}01$. Vemos en la figura que este valor se establece en $x_0 = -4{,}14$.

De entre los cuantiles, hay 3 que son especialmente usados y conocidos, el cuantil 1/2 (o

percentil 50) conocido como la **mediana** el cuantil 1/4 (o percentil 25) y el cuantil 3/4 (o percentil 75). Estos se utilizan habitualmente para describir la distribución de una variable aleatoria y dan una buena idea de los valores que puede tomar esta.

3.4. Momentos de una variable aleatoria.

Como hemos visto, la distribución de una variable nos da una idea del comportamiento de la misma. Sin embargo, ésta es, a veces, difícil de entender y es necesario utilizar resúmenes más sencillos que nos permitan visualizar, sin mucho esfuerzo, la información que contiene.

Uno de estos resúmenes ya lo hemos estudiado al final de la sección anterior cuando hablábamos de los cuantiles y percentiles. Éstos nos dan una idea de como se reparte la probabilidad entre los posibles valores de la variable. Sin embargo, el valor al que estamos más habituados es a la *media* o *promedio* que suele entenderse como el *valor esperado* o *esperanza* de la variable. Además, en estadística, nos interesa interpretar correctamente la *variablilidad*, como de esparcidos están los valores de una variable. Esta característica de los datos queda reflejada en lo que se conoce como *desviación estandard* o *varianza*. En esta sección estudiaremos estos conceptos en términos de valores esperados por lo que, veremos que, la idea de esperanza va más allá del cálculo de un valor medio.

3.4.1. Esperanza.

Pensemos en una inversora que sabe que si compra un determinado stock por un valor de 18 euros, su ganancia tras un año será la variable aleatoria X tal que $18 + X$ sea el valor del stock en el mercado al cabo de un año. Parece lógico que esta inversora quiera saber el valor medio de X, pero que significa eso exactamente.

De forma intuitiva, el valor medio o esperado de X sería el promedio de todos los posibles valores que puede tomar X ponderados por la probabilidad de que ese sea el verdadero valor.

Pensadlo así, lanzamos una moneda, si sale cara nos dan 3 euros y si sale cruz nos quitan 1, cual será la ganancia esperada. Como la mitad de las veces ganaré 3 y la otra mitad ganaré -1, si juego muchas veces y saco el promedio del dinero ganado por tirada tendré:

$$3 \times \frac{1}{2} + (-1) \times \frac{1}{2} = 1$$

Por tanto, mi ganancia esperada será de 1 euro.

En general, se define la esperanza de una variable discreta como:

Definición: sea una variable discreta X con función de probabilidad p_X y soporte Ω. La **esperanza de la variable** X se denota por $E(X)$ y se calcula como:

$$E(X) = \sum_{x \in \Omega} x p_X(x)$$

Fijaos que la esperanza de una variable aleatoria discreta depende únicamente de la función de probabilidad p_X y que podría no existir si tenemos un soporte infinito y la suma de la serie definida por p_X no converge.

La idea de calcular una media ponderada de los valores de una variable puede extenderse al caso de variables continuas. En concreto:

Definición: sea X una variable aleatoria continua con función de densidad f_X. La **esperanza de la variable** X se denota por $E(X)$ y se calcula como:

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

Imaginemos que compramos una bombilla que tiene un tiempo máximo de funcionamiento de un año. El tiempo hasta que la bombilla falla, X, es una variable aleatoria continua con función de densidad:

$$f(x) = \begin{cases} 2x & \text{si } 0 < x < 1 \\ 0 & \text{en otro caso} \end{cases}.$$

La esperanza puede calcularse entonces como:

$$E(X) = \int_0^1 x 2x \, dx = 2 \int_0^1 x^2 \, dx = \frac{2}{3},$$

lo que nos indica que el tiempo medio de funcionamiento de la bombilla es de 2/3 de un año (unos 8 meses).

Al igual que en el caso de una variable aleatoria discreta, la esperanza de una v.a. continua depende únicamente de su función de densidad y podría no existir si, al integrar, no obtenemos un valor finito.

3.4.1.1. Esperanza de una función.

En algunas ocasiones no estamos interesados directamente en la esperanza una variable aleatoria sino en la esperanza una función de la misma. Por ejemplo, podemos conocemos

la tasa de fallos de una máquina en un año, X, pero estar interesados en el tiempo que la máquina tarda en fallar $Y = 1/X$.

En general, dada r una función en la recta real podemos definir $Y = r(X)$. La esperanza de esta nueva variable aleatoria podría calcularse usando la definición siempre y cuando la distribución de Y sea conocida. Sin embargo, en la mayoría de las situaciones esto no sucede. El siguiente teorema nos indica como calcular la esperanza de una función de una variable aleatoria a partir de la distribución de la variable original.

Teorema: sea X una variable aleatoria y r una función en la recta real. Si X tiene una distribución continua:

$$E[r(X)] = \int_{-\infty}^{\infty} r(x)f(x)dx,$$

si la integral es finita.

Si X tiene una distribución discreta:

$$E[r(X)] = \sum_{x\in\Omega} r(x)f(x),$$

si la suma es finita.

3.4.1.2. Propiedades de la esperanza.

Teorema: La esperanza de una variable aleatoria debe cumplir las siguientes propiedades: 1. sea $Y = aX + b$, utilizando el teorema que nos dice como calcular la esperanza de una función tenemos que

$$E(Y) = aE(X) + b.$$

2. Si existe a constante tal que $P(X \geq a) = 1$, entonces $E(X) \geq a$. Del mismo modo, si existe b constante tal que $P(X \leq b) = 1$, entonces $E(X) \leq b$. 3. Sean $X_1, \ldots, X_n$ n variables aleatorias tales que $E(X_i)$ es finita para todo $(i = 1, \ldots, n)$ entonces:

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n).$$

4. Sean $X_1, \ldots, X_n$ n variables aleatorias independientes tales que $E(X_i)$ es finita para todo $(i = 1, \ldots, n)$ entonces:

$$E(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n} E(X_i).$$

3.4.1.3. Comparación entre la mediana y la media.

Es usual cuando se define la media hablar de ella como el *centro de gravedad* de la distribución. Pero si pensamos en el centro, también tiene sentido pensar en un valor m_0 tal que $P(X \geq m_0) \geq 1/2$ y $P(X \leq m_0) \geq 1/2$. Es fácil ver existirá, al menos, un valor que cumpla estas características, se trata de la mediana o cuantil 0,5. Por tanto, tenemos dos valores que podríamos definir como centrales pero con distintas propiedades. Para entender sus diferencias veamos un ejemplo.

Supongamos que los ingresos medios anuales de una famila de una determinada comunidad son de 30 000 euros. Es posible que solo un número reducido de familias cobre más de esta cantidad pero que estas familias cobren mucho más de 30 000 euros. Como un caso extremo, pensemos en 100 familias de las cuales 99 cobra 1000 euros y la restante cobra 2 901 000 euros. En ese caso, efectivamente, la media es de 30 000 euros pero la mediana es de 1 000 euros. Sin embargo, si tuviésemos una mediana de 30 000 euros, podríamos estar seguros de que al menos la mitad de las familias cobra más de esa cantidad.

Y es que la mediana tiene una interesante propiedad que no tiene la media y es que:

Teorema: sea X una variable aleatoria y r una función biyectiva en la recta real, si m es la mediana de X, $r(m)$ será la mediana de $r(X)$.

3.4.2. Varianza

A pesar de que la esperanza es un resumen muy útil, no contiene demasiada información sobre la distribución de la variable. de la distribución de una variable. Por ejemplo, una variable constante que siempre vale 2 tendrá una media de 2, pero podemos encontrar otra variable aleatoria no constante con la misma media y queda claro que ambas variables tendrán distribuciones diferentes.

Definición: sea X una variable aleatoria con esperanza finita $\mu = E(X)$. La varianza de X se denota por $Var(X)$ se define como:

$$Var(X) = E[(X - \mu)^2].$$

Si la $E(X)$ no existe diremos que $Var(X)$ tampoco existe.

La desviación típica o estandard de X es la raíz cuadrada positiva de $Var(X)$ si ésta existe y se denota como σ_X.

Veamos un ejemplo, supongamos que tengamos una variable aleatoria discreta que sólo puede tomar valores -2, 0, 1, 3 y 4 con igual probabilidad. La media de esta variable X es claramente:

$$E(X) = \frac{1}{5}(-2+0+1+3+4) = 1{,}2.$$

Ahora definimos la variable $W = (X - 1{,}2)^2$ y calculo su esperanza como

$$E(W) = \frac{1}{5}\left[(-2-1{,}2)^2 + (0-1{,}2)^2 + (1-1{,}2)^2 + (3-1{,}2)^2 + (4-1{,}2)^2\right] = 4{,}56.$$

Teorema: $Var(X)$ también puede calcularse como:

$$Var(X) = E(X^2) - [E(X)]^2.$$

Volviendo al ejemplo anterior, podemos obtener el mismo resultado como $Var(X) = E(X^2) - [E(X)]^2$. Para ello calculamos $E(X)^2 = 1{,}2^2 = 1{,}44$ y

$$E(X^2) = \frac{1}{5}(4+0+1+9+16) = 6,$$

y obtenemos $Var(X) = 6 - 1{,}44 = 4{,}56$ igual que antes.

3.4.2.1. Propiedades de la varianza

Teorema: la varianza de una variable aleatoria X debe cumplir las siguientes propiedades: 1. La varianza de una variable aleatoria X, si existe, será siempre $Var(X) \geq 0$. 2. Si X es una variable aleatoria acotada, entonces $Var(X)$ existe y es finita.
3. $Var(X) = 0$ si y solo si existe una constante c tal que $Pr(X = c) = 1$. 4. Dadas dos constantes a y b, sea $Y = aX + b$ entonces:

$$Var(Y) = a^2 Var(X),$$

y la desviación estándar de Y será $\sigma_Y = \mid a \mid \sigma_X$.

3.4.3. Momentos

La esperanza y la varianza que ya hemos definido son casos particulares de un concepto más amplio que en probabilidad se llama momentos de una variable.

Definición: dada una variable aleatoria X, a la esperanza de X^k se le denomina **momento k-ésimo o de orden** k y diremos que este existe si $E(\mid X \mid^k) \leq \infty$.

Por otra parte, sea $\mu = E(X)$ a la magnitud $E[(x-\mu)^k]$ se le conoce como **momento central de orden** k

En esta asignatura no profundizaremos en el concepto de momento, sin embargo, caben destacar algunos momentos concretos que os serán muy útiles de cara al análisis de datos.

3.4.3.1. Asimetría (Skewness)

Una propiedad importante de la distribución de una variable aleatoria es su simetría. Si una variable aleatoria X es simétrica alrededor de su media μ quiere decir que la probabilidad de ser mayor de ese valor es igual a la probabilidad de ser menor que el mismo. En base a esta definición es fácil ver que el momento central de orden uno de la variable será 0:

$$E(X-\mu) = E(X) - \mu = \mu - \mu = 0$$

Se puede demostrar que sucede lo mismo para cualquier momento central de orden impar. De esta forma, los momentos centrales de orden impar pueden utilizarse para medir la simetría de la variable. En concreto

Definición: sea X una variable aleatoria con media μ y momento de orden 3 finito, se define su **Asimetría** o **Skewness** como

$$S(X) = \frac{E[(X-\mu)^3]}{\sigma^3}.$$

Cuanto más alejado de 0 esté este valor menos simétrica será la distribución de la variable indicando una mayor probabilidad para los los valores que están a la derecha de la media si es positivo y para los valores a la izquierda de la misma si es negativo.

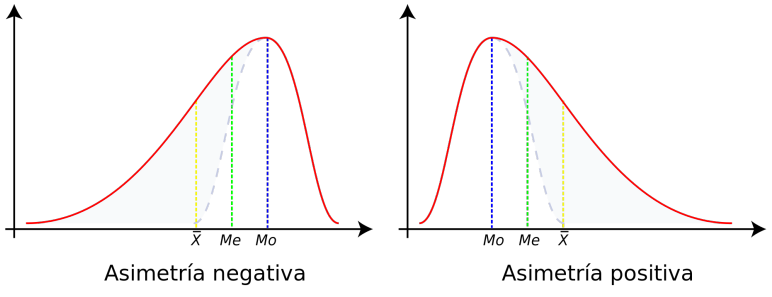


Figura 8: Simetría de una distribución continua.

3.4.3.2. Kurtosis

Otra característica importante de una distribución es como de concentrada está la probabilidad, es decir, si hay pocos valores que acumulan mucha probabilidad y el resto tienen una probabilidad menor o si todos los valores tienen más o menos la misma probabilidad. Para poder entender esta característica también nos apoyamos en una medida llamada *curtosis* y basada en el momento central de orden 4.

Definición: sea una variable aleatoria X, llamamos **curtosis** al valor:

$$K(X) = \frac{E[(X-\mu)^4]}{\sigma^4} - 3$$

El valor $K(X) = 0$ correspondería con la distribución de una variable normal (que es la que tomamos como referencia) y se les denomina distribuciones **mesocúrticas** o **normocúrticas**. Cuando $K(X) > 0$ hablamos de distribuciones **leptocúrticas**, en estas el pico central es más alto que en la normal y el soporte más corto Si $K(X) < 0$ hablamos de distribuciones **platicurticas** que suelen tener un pico más bajo y un soporte más largo que la normal.

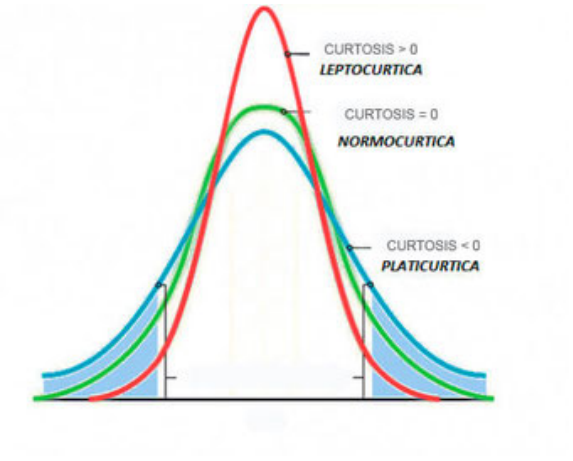


Figura 9: Curtosis de una distribución continua.

3.5. Ejercicios

1. Supongamos que se lanzan dos dados balanceados y llamamos X al valor absoluto de la diferencia entre los dos número obtenidos. Determina la función de probabilidad de la variable X.

2. Supongamos que una variable aleatoria discreta X tiene una función de probabilidad

$$f(x) = \begin{cases} cx & \text{para } x = 1, \ldots, 5 \\ 0 & \text{en otro caso} \end{cases}.$$

Calcula la constante c para que f sea realmente una función de probabilidad.

3. Un grupo de personas están llegando a una fiesta de uno en uno. Mientras esperan a que llegue más gente, se entretienen comparando sus fechas de cumpleaños. Sea X el número de personas que se necesitan para conseguir una coincidencia (es decir, antes de que llegase la persona X no había coincidencia y cuando llega si la hay). Calcula la función de probabilidad de la variable X.

4. En ocasiones, para detectar fraudes se hace uso de la ley de Benford. Esta ley dice que la primera cifra X de muchos de los números que manejamos a diario, sigue una distribución concreta que establece que el 30 % de las veces será $X = 1$, el 18 % $X = 2$ y, en general,

$$P(X = j) = \log_{10} \frac{j+1}{j}$$

para $j = 1, 2, 3, \ldots, 9$. Comprueba de que se trata de una función de probabilidad valida. (Inténtalo usando las propiedades de los logaritmos y no la calculadora).

5. Supongamos que la función de densidad de una variable aleatoria continua X es como sigue:

$$f(x) = \begin{cases} cx^2 & \text{para } 1 \leq x \leq 2 \\ 0 & \text{en otro caso} \end{cases}.$$

a. Encuentra el valor de la constante c y dibuja la función aproximadamente.

b. Encuentra el valor de $P(X > 3/2)$.

6. Supongamos que la función de densidad de una variable aleatoria continua X es :

$$f(x) = \begin{cases} \frac{1}{8}x & \text{para } 0 \leq x \leq 4 \\ 0 & \text{en otro caso} \end{cases}.$$

a. Encuentra el valor de t tal que: $P(X \leq t) = 1/4$.

b. Encuentra el valor de t tal que $P(X \geq t) = 1/2$.

7. Una vendedora de helado carga su carrito con 40 litros cada día. La cantidad de helado que ha conseguido vender al final del día es una variable aleatoria X con función de densidad

$$f(x) = \begin{cases} cx & \text{para } 0 < x < 40 \\ 0 & \text{en otro caso} \end{cases}.$$

Determina la constante c, dibuja como se comporta la venta de helado y calcula la probabilidad de que un día cualquiera venda menos de 10 litros.

8. Supongamos que una variable aleatoria X puede tomar únicamente los valores -2, 0, 1, y 4, con probabilidades: $P(X = -2) = 0{,}4$, $P(X = 0) = 0{,}1$, $P(X = 1) = 0{,}3$ y $P(X = 4) = 0{,}2$. Esboza la función de distribución acumulada de X.

9. Una moneda es lanzada repetidas veces hasta obtener una cara por primera vez. Sea X el número de lanzamientos necesarios. Esboza la función de distribución acumulada de X.

10. Calcula los principales cuantiles de la distribución de la variable del ejercicio 2.

11. Encuentra la función de distribución para el ejercicio 6. ¿Cuál es el valor de la mediana?

12. Imagina que debes elegir una palabra al azar de la frase *El yogurt griego es el mejor del mundo*. Si X denota el número de letras en la palabra seleccionada, ¿cuál es la $E(X)$? ¿y su varianza?

13. Encuentra la esperanza y la varianza de la variable del ejercicio 2.

14. Encuentra la esperanza y la varianza de la variable del ejercicio 6.

15. Supongamos que en una clase hay 10 chicos y 15 chicas. Si debemos elegir 8 estudiantes aleatoriamente y definimos X como el número de chicos seleccionados e Y como en número de chicas seleccionadas, calcula la esperanza de E(X-Y) y su varianza.

4. Principales distribuciones de probabilidad

4.1. Introducción

A lo largo del tema anterior vimos como se definía una variable aleatoria y como podía caracterizarse su distribución de probabilidad a través de su función de probabilidad (en el caso de v.a. discretas) o de su función de densidad (para v.a. continuas).

A lo largo de la historia, se han identificando algunas distribuciones que se repiten (de forma aproximada) en muchas situaciones. Estas han acabado recibiendo un nombre propio y siendo estudiadas en profundidad para identificar su comportamiento de forma muy especifica. De ellas se conocen su función de probabilidad o densidad, su función de distribución acumulada así como sus cuantiles y momentos de forma precisa.

Cuando hablamos de este tipo de distribuciones solemos referirnos a ellas como **familias** puesto que, en realidad, no se trata de una única distribución sino de un conjunto de ellas con características similares y que difieren, únicamente, en el valor de uno o varios **parámetros**.

A continuación estudiaremos algunas de las más importantes:

4.2. Distribuciones discretas.

4.2.1. Distribución Bernoulli y binomial.

El caso más simple de una v.a. discreta es aquella que sólo puede tomar dos valores que se suelen representar por 1 cuando sucede aquello que nos interesa y 0 cuando no.

Por ejemplo, si lanzamos una moneda y buscamos que nos salga una cara, 1 representará cara y 0 cruz. Si estamos estudiando la aparición de efectos secundarios tras tomar un medicamento, tendremos un 1 cuando estos aparezcan y 0 cuando no.

A este tipo de variables se les asocia una distribución de la familia Bernoulli:

Definición: Se dice que una variable aleatoria X sigue una distribución de Bernoulli con parámetro π $(0 \leq \pi \leq 1)$ si X sólo puede tomar los valores 0 y 1 con probabilidades:

$$P(X = 1) = p \text{ y } P(X = 0) = 1 - p.$$

La función de probabilidad para este tipo de variables puede escribirse en función del

parámetro π como

$$p(x|\pi) = \pi^x(1-\pi)^{(1-x)},$$

y es fácil deducir de ella que se cumplen las probabilidades de la definición. Del mismo modo, es fácil ver que:

- $E(X) = 1 \times \pi + 0 \times (1-\pi) = \pi$
- $E(X^2) = 1 \times \pi + 0 \times (1-\pi) = \pi$
- $Var(X) = E(X^2) - E(X)^2 = \pi - \pi^2 = \pi(1-\pi)$

A los experimentos en los que el resultado es una v.a. del tipo descrito se les conoce como **experimentos Bernoulli**.

El lanzamiento de 10 monedas que estudiábamos en el tema anterior puede verse como la repetición (de manera independiente) de 10 experimentos Bernoulli donde 1 representaba el suceso obtener cara. En ese caso la variable aleatoria X *número de caras*, puede verse como la suma de las 10 variables Bernoulli $X = X_1 + \cdots + X_{10}$

Del mismo modo, cuando tenemos un ensayo clínico con 30 pacientes para estudiar la aparición de efectos secundarios, estaremos ante 30 experimentos Bernoulli independientes $X_1, \ldots, X_30$. Si nuestro interés reside en saber cuantos pacientes desarrollaron efectos secundarios definiremos la variable X, de nuevo, como la suma de las 30 variables tipo Bernoulli.

De forma general, si tenemos un número N de experimentos Bernoulli independientes $X_1, \ldots, X_N$ con parámetro π y estamos interesados en conocer el número de veces que se repite una la característica de interés $X = X_1 + \cdots + X_N$, podremos decir que X sigue una distribución binomial de parámetros N y π.

Definición: se dice que una variable X tiene una **Distribución Binomial** de parámetros N y π cuando su función de probabilidad tiene la siguiente forma:

$$p(x \mid N, \pi) = \begin{cases} \binom{N}{x}\pi^x(1-\pi)^{(N-x)} & \text{Si } x = 0, 1, \ldots, N \\ 0 & \text{En otro caso} \end{cases}$$

La esperanza de una distribución binomial es

$$E(X) = \sum_{i=1}^{N} E(X_i) = N\pi,$$

mientras que su varianza será:

$$Var(X) = \sum_{i=1}^{N} Var(X_i) = N\pi(1-\pi)$$

Teorema La suma de p v.a. con distribución binomial de parámetros N_i y π siguen una distribución binomial con parámetros $N_1 + \cdots + N_p$ y π

Un ejemplo interesante (y real) Los juzgados americanos suelen utilizar la distribución binomial para determinar la composición de los jurados populares. En un caso concreto (Castaneda v. Partida, 430 U.S. 482, 1977), el acusado, de origen mejicano-americano, intentó alegar que la población a la que pertenecía estaba representada por debajo de la proporción real en los jurados populares. En concreto la población local era en 79.1 % Mejicana-Americana. Durante un periodo de 2,5 años había habido un total de 220 personas llamadas a participar como jurado pero sólo 100 fueron Mejicanas-Americanas.

Para certificar si la queja estaba justificada, el jurado tomo cada persona llamada a ser jurado como un experimento Bernoulli independiente con parámetro $\pi = 0{,}791$ y, como lo que alegaba era que 100 era un número muy bajo, se calculó la probabilidad de que una variable binomial X de parámetros $N = 220$ y $\pi = 0{,}791$ fuese igual o menor que 100. Realmente esta probabilidad es muy baja pero... es signo de discriminación hacia la población Mejicana-Americana?

Fijaros que estamos calculando la $P(X \leq 100 \mid N = 220, \pi = 0{,}791)$. Esto supone que estamos condicionando a que $\pi = 0{,}791$ o, equivalentemente, estamos condicionando a la situación en la que la población está representada en la proporción correspondiente. La probabilidad que nos gustaría tener es, sin embargo la probabilidad que nos interesaría realmente es la inversa, es decir, la probabilidad de que $\pi = 0{,}791$ dado que $X = 100$. Fijaros que esto podríamos hacerlo con el teorema de Bayes y veremos como hacerlo más adelante.

Nota: Distribución Binomial en R. La función de probabilidad de una distribución binomial puede obtenerse en R usando el comando *dbinom(x,size,prob)* donde *size*=N y *prob*=π. Por tanto, la probabilidad $P(X \leq 100 \mid N = 220, \pi = 0{,}791)$ puede calcularse como:

```
dbinom(100,size = 220,prob = 0.791)
```

```
## [1] 6.287453e-28
```

Del mismo modo, la función de distribución acumulada se calcula utilizando el comando *pbinom(x,size,prob)*

```
pbinom(100,size = 220,prob = 0.791)
```

```
## [1] 8.032817e-28
```

y los cuantiles pueden calcularse usando *qbinom(p,size,prob)* donde p será la probabilidad para la que queremos calcular el cuantil. Por ejemplo, el primer cuartil o cuantil 0.25 ($p = 0{,}25$) será:

```
qbinom(0.25,size = 220,prob = 0.791)
```

```
## [1] 170
```

Nota: toda variable con sólo dos posibles valores (0 y 1) tendrá una distribución Bernoulli pero, no toda suma de Bernoullis tendrá una distribución binomial. Por ejemplo, si estamos hablando de tener o no tener una enfermedad contagiosa, los experimentos Bernoulli no son independientes y la probabilidad de enfermar aumenta a medida que más gente se contagia.

Para estudiar este tipo de casos en los que los experimentos Bernoulli son dependientes, utilizamos la distribución hipergeométrica que estudiamos a continuación.

4.2.2. Distribución hipergeométrica.

Pensemos en el típico ejemplo de una urna que contiene A bolas rojas y B azules. Supongamos que seleccionamos $N \geq 0$ bolas de la urna sin reemplazamiento y estamos interesados en X *el número de bolas rojas.*

Claramente, debemos tener $N \leq A + B$ o nos quedaríamos sin bolas. Por otra parte, si $N = 0$ entonces $X = 0$ porque no hemos sacado ninguna bola. Centrándonos en $N \geq 1$ podemos pensar que, cada vez que sacamos una bola, tenemos una va.a. X_i que valdrá 0 si la bola es azul y 1 si es roja. Es fácil ver que cada X_i tiene una distribución Bernoulli pero, $X_1, \ldots, X_N$ no son independientes ya que la probabilidad de X_i cambia según lo que haya sucedido en los experimentos anteriores. Es por ello que no esperaremos que $X = X_1 + \cdots + X_N$ sea una distribución binomial.

Se puede demostrar que, cualquier variable que siga el esquema de este ejemplo tiene una función de probabilidad:

$$p(x \mid A, B, N) = \frac{\binom{A}{x}\binom{B}{N-x}}{\binom{A+B}{N}}$$

para todo $x = 0, 1, 2, \ldots, N$.

Definición: sean A, B y N números enteros no negativos tales que $A + B \geq N$, diremos que variable aleatoria X sigue una **Distribución hipergeometrica de parámetros A, B y N** si su función de probabilidad tiene la forma anterior.

Teorema Sea X una variable aleatoria con distribución hipergeométrica de parámetros A, B y N estrictamente positivos:

- $E(X) = \frac{NA}{A+B}$
- $Var(X) = \frac{NAB}{(A+B)^2} \cdot \frac{A+B-N}{A+B-1}$

Nota: fijaros que, si hubiésemos reemplazado las bolas en la urna, cada X_i podría ser considerado independiente y tendríamos una distribución binomial con $\pi = \frac{A}{A+B}$. En ese caso la media seguiría siendo $E(X) = \frac{NA}{A+B}$ aunque la varianza (es decir, la variabilidad en los resultados) sería diferente. Lo curioso es que ambas varianzas están relacionadas. De hecho, definiendo $T = A + B$, podemos escribir la varianza de una v.a. con distribución hipergeométrica como:

$$Var(X) = N\pi(1 - \pi)\frac{T - N}{T - 1}$$

Podemos entender T como el tamaño de la población de bolas que, en el caso de la hipergeométrica es finita (llega un momento que se nos acaban las bolas) y entonces que la varianza de una distribución hipergeométrica está corregida por un factor $\alpha = \frac{T-N}{T-1}$ que tiene el nombre de *corrección para una población finita.*

Fijaros, sin embargo que si T es muy grande en comparación con N, α se acercará a 1. Se demuestra, de hecho, que cuando esto sucede, las distribuciones hipergeométrica y binomial coinciden.

Es el caso del ejemplo de la población Mejicano-Americana. Realmente el número de personas en esa población es finito y la elección es sin reemplazamiento pero, por tratarse de una población muy grande con respecto al número de personas seleccionadas, lo tratamos

como si se tratasen de experimentos independientes y la distribución de la variable fuese binomial.

Nota: Distribución Hipergeométrica en R. La función de probabilidad de una distribución hipergeométrica de parámetros A, B y N puede calcularse en R usando el comando *dhyper(x,A, B, N)*. Por ejemplo, la probabilidad de sacar 3 bolas rojas de 5 extracciones cuando en la urna hay 10 rojas y 20 azules es:

```
dhyper(3,10,20,5)
```

```
## [1] 0.1599933
```

Y si lo que queremos es la función de distribución acumulada utilizaremos el comando *phyper(x, A, B, N)*. Es decir, la probabilidad de sacar 3 o menos bolas rojas de entre las 5 extracciones en el ejemplo anterior es:

```
phyper(3,10,20,5)
```

```
## [1] 0.9687592
```

y los cuantiles pueden calcularse usando *qhyper(p, A, B, N)* donde p será la probabilidad para la que queremos calcular el cuantil. Por ejemplo, el primer cuartil o cuantil 0.25 ($p = 0{,}25$) será:

```
qhyper(0.25,10,20,5)
```

```
## [1] 1
```

4.2.3. Distribución de Poisson.

En muchas ocasiones nos encontramos con experimentos que consisten en saber el número de veces que se repite un determinado evento. Por ejemplo, el número de llamadas que se reciben en una centralita en una hora, el número de visitas a una página web en un minuto o el número de inundaciones que sufre una población en un año.

Más concretamente, se trata de situaciones en las que tenemos un conjunto infinito de experimentos Bernoulli cuya probabilidad de éxito es muy baja. Por ejemplo:

¿Cuantos WhatsApp recibes en una hora? Hay muchas personas que podrían escribirte pero es poco probable que una persona especifica te escriba en esa hora. Visto de otra forma, si dividimos la hora en milisegundos es poco probable que en un milisegundo

concreto alguien te escriba un WhatsApp aunque es cierto que con $3{,}6 \times 10^6$ en una hora, alguno te llegará.

El número de terremotos en una región durante un año Evidentemente es poco probable que en una localización concreta y un tiempo determinado se produzca un terremoto pero lo cierto es que hay muchas localizaciones y ocasiones en un año para que suceda.

Este tipo de variables se estudian con lo que se conoce como *Paradigma de Poisson* o *ley de los eventos raros* y se deduce que siguen una *distribución de Poisson*:

Definición: una variable aleatoria X sigue una **Distribución de Poisson de parámetro** λ cuando la probabilidad de x se puede expresar como:

$$p(x \mid \lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!} & x = 0, 1, 2, \ldots \\ 0 & \text{en otro caso} \end{cases}$$

Notad que se trata de una función de probabilidad valida dado que la serie de Taylor $\sum_{x=1}^{\infty} \frac{\lambda^x}{x!} = e^{\lambda}$

Teorema Sea X una variable aleatoria con distribución de Poisson de parámetro λ:

- $E(X) = \lambda$
- $Var(X) = \lambda$

Teorema Sean $X_1, \ldots, X_k$ variables aleatorias independientes con media $\lambda_1, \ldots, \lambda_k$ respectivamente, $X_1 + \cdots + X_k$ sigue una distribución de Poisson con parámetro $\lambda_1 + \cdots + \lambda_k$

El parámetro λ representa la tasa a la que ocurre el evento, por ejemplo se reciben (de media) 20 WhatsApp por hora o se producen (de media) 2 terremotos en la región en un año. Es importante darse cuenta que, en estas situaciones, hablamos de eventos raros no porque λ sea pequeño sino porque la probabilidad de cada experimento Bernoulli (recibir un WhatsApp en un milisegundo) es muy baja.

Pero, volvamos un momento al ejemplo de los WhatsApp, ahora podríais decirme que si dividimos la hora en milisegundos y definimos el experimento Bernoulli como recibir un WhatsApp en un segundo, como mucho podría recibir $3{,}6 \times 10^6$ y se trataría, por tanto, de una distribución binomial. En realidad, la función de Poisson se puede expresar (como vemos en el siguiente teorema) como el límite de una distribución binomial cuando el

tamaño de la población N es muy grande y la probabilidad π es muy pequeña siendo $N\pi = \lambda$

Teorema Sea $X \sim Bin(N, \pi)$, si $N \to \infty$ y $\pi \to 0$ de forma que $N\pi = \lambda$, la distribución de X converge a una Poisson de parámetro λ.

Veámoslo con un ejemplo: la encargada de redes sociales de una gran empresa esta interesada en conocer la distribución del número de gente que retwittea su publicidad en un día. Cada día un millón de personas deciden, de manera independiente, si retwittear o no con probabilidad $\pi = 10^{-5}$

Imaginemos que quiere saber que probabilidad hay de que les retwitteen más de 15 personas ($P(X > 15) = 1 - P(X \leq 15)$. Utilizando la probabilidad binomial tendríamos:

```
1- pbinom(15,size = 10^6, prob = 10^(-5))
```

```
## [1] 0.04873954
```

En este caso π es muy pequeño, N muy grande y $N\pi = 10$ por lo que podríamos aplicar el teorema anterior y aproximar la probabilidad utilizando una distribución de Poisson de parámetro $\lambda = 10$

```
1-ppois(15,lambda = 10)
```

```
## [1] 0.0487404
```

y podemos comprobar que es una muy buena aproximación a la probabilidad anterior.

Pero, volviendo al ejemplo, es posible que, en lugar de estar interesado en el número de retwitts en un día de pronto se interese en la misma cantidad pero en una hora. Cabe preguntarse si la tasa de retwitts en ese tiempo es de 10/24. En este sentido se define lo que se conoce como *proceso Poisson*

Definición: Un **proceso de Poisson** con tasa λ es un proceso que satisface las siguientes dos propiedades:

- El numero de eventos en un intervalo de tiempo de tamaño t sigue una distribución de Poisson de parámetro λt.
- El número de eventos en intervalos de tiempos disjuntos son independientes.

Nota: A lo largo de esta sección hemos hablado de número de eventos a lo largo del tiempo,

sin embargo, la distribución de Poisson también puede asociarse al número de eventos o número de casos en una unidad espacial (como el número de fallos en un metro de tela) o, incluso, a una combinación de ambos (como el número de fallecimientos en una población de tamaño x en un tiempo t)

4.2.4. Distribución binomial negativa.

Si recapitulamos, todas las distribuciones vistas hasta ahora tienen que ver con el número total de éxitos X en N experimentos Bernoulli y distingamos entre experimentos independientes con la misma probabilidad de éxito π (distribución binomial); experimentos no independientes con probabilidad de éxito determinada por el resultado del experimento anterior (distribución hipergeométrica) y experimentos independientes con N muy grande π – común– muy pequeña (distribución de Poisson).

Pero existen muchos casos prácticos donde no me interesa observar N experimentos y contar el número de éxitos sino observar hasta que se produzca un determinado número de ellos.

Por ejemplo, podemos pensar en X_i como la v.a. resultante de un experimento Bernoulli que consiste en observar si una bombilla funciona ($X_i = 0$) o no ($X_i = 1$) en un determinado día. Nuestra variable de interés será el número de días transcurridos hasta que la bombilla se apaga.

Del mismo modo, si la encargada de una máquina esta pendiente de que ésta produzca 4 piezas defectuosas (para re-calibrarla) la variable aleatoria será el número de elementos producidos hasta que se producen 4 fallos. De nuevo, cada elemento i producido es un experimento Bernoulli con $X_i = 0$ si la pieza esta bien y $X_i = 1$ is la pieza es defectuosa.

Se puede demostrar que este tipo de variables siguen una distribución conocida como binomial negativa:

Definición: se dice que una v.a. X sigue una **Distribución Binomial Negativa** ($X \sim BN(r, \pi)$) con parámetros r (r=1,2,...) y $\pi \in (0, 1)$ si su función de probabilidad es de la forma:

$$p(x \mid r, \pi) = \binom{r + x - 1}{x} \pi^r (1 - \pi)^x$$

para cualquier $x = 0, 1, 2, \ldots$

En esta distribución π representa la probabilidad de *éxito* en cada experimento Bernoulli mientras que r representa el número de *éxitos* tras los cuales dejaremos de observar. En el caso de la encargada de la máquina $r = 4$ mientras que en el caso de la bombilla $r = 1$.

El caso $r = 1$ es un caso particular de la distribución binomial negativa que recibe el nombre de **Distribución Geométrica**

Definición: Diremos que una v.a. X sigue una **distribución Geométrica** cuando su función de probabilidad sea de la forma:

$$p(x \mid \pi) = \pi(1-\pi)^x.$$

para $x = 0, 1, 2, \ldots$.

Teorema: La suma de r v.a. con distribución geométrica de parámetro π siguen una distribución binomial negativa de parámetros r y π

Teorema: La media y la varianza de una v.a. X con parámetros r y π son:

$$E(X) = \frac{r(1-\pi)}{\pi} \text{ y } Var(X) = \frac{r(1-\pi)}{\pi^2}.$$

Veamos otro ejemplo. Imaginemos un juego de la lotería que implica elegir tres números del 0 al 9 de manera independiente y con reemplazamiento. Este juego se repite todos los días (también de forma independiente).

Un evento curioso es cuando los tres números obtenidos un día concreto son idénticos fenómeno se le denomina triplete y que se produce con una probabilidad $\pi = 0{,}01$ (fijaros que existen 10 posibles tripletes de los 10^3 posibles resultados).

Si queremos saber cuantos días transcurren antes de que se produzca un triplete estaremos ante una v.a. X con distribución geométrica de parámetro $\pi = 0{,}01$ cuya esperanza es $\frac{1-\pi}{\pi} = 0{,}99/0{,}01 = 99$ y por tanto tardaremos, de media, 100 días en ver un triplete.

Pero, imaginemos ahora que un jugador lleva 120 días sin ver un triplete y cree que debe estar a punto de suceder y para ello se dispone a calcular la probabilidad condicionada de X dado que $X \geq 120$. En ese momento se da cuenta de que no puede estar más lejos de la realidad como nos muestra el siguiente teorema:

Teorema: Sea X una v.a. con distribución geométrica con parámetro π y sea $k \geq 0$, entonces, para cualquier valor $t \geq 0$

$$P(X = k + t \mid X \geq k) = Pr(X = t).$$

A esta propiedad se le denomina *falta de memoria de la distribución geométrica*

Nota: Binomial Negativa en R La función de probabilidad de una distribución binomial negativa puede obtenerse en R usando el comando *dnbinom(x,size,prob)* donde *size*=r y *prob*=π. Por tanto, la probabilidad $P(X \leq 4 \mid r = 2, \pi = 0{,}5)$ puede calcularse como:

```
dnbinom(4, size = 2, prob =0.5 )
```

```
## [1] 0.078125
```

Del mismo modo, la función de distribución acumulada se calcula utilizando el comando *pnbinom(x,size,prob)*

```
pbinom(4, size = 2, prob =0.5 )
```

```
## [1] 1
```

y los cuantiles pueden calcularse usando *qnbinom(p,size,prob)* donde p será la probabilidad para la que queremos calcular el cuantil. Por ejemplo, el primer cuartil o cuantil 0.25 ($p = 0{,}25$) será:

```
qnbinom(0.25, size = 2, prob =0.5  )
```

```
## [1] 0
```

4.3. Distribuciones continuas.

4.3.1. Distribución Uniforme.

La distribución de probabilidad más sencilla para una variable continua que toma valores en un intervalo acotado (a, b) es aquella que da, a todos los valores, la misma densidad:

Definición: Una variable aleatoria X tiene una **Distribución Uniforme** en el intervalo (a, b) $X \sim Unif(a, b)$ si su función de densidad es

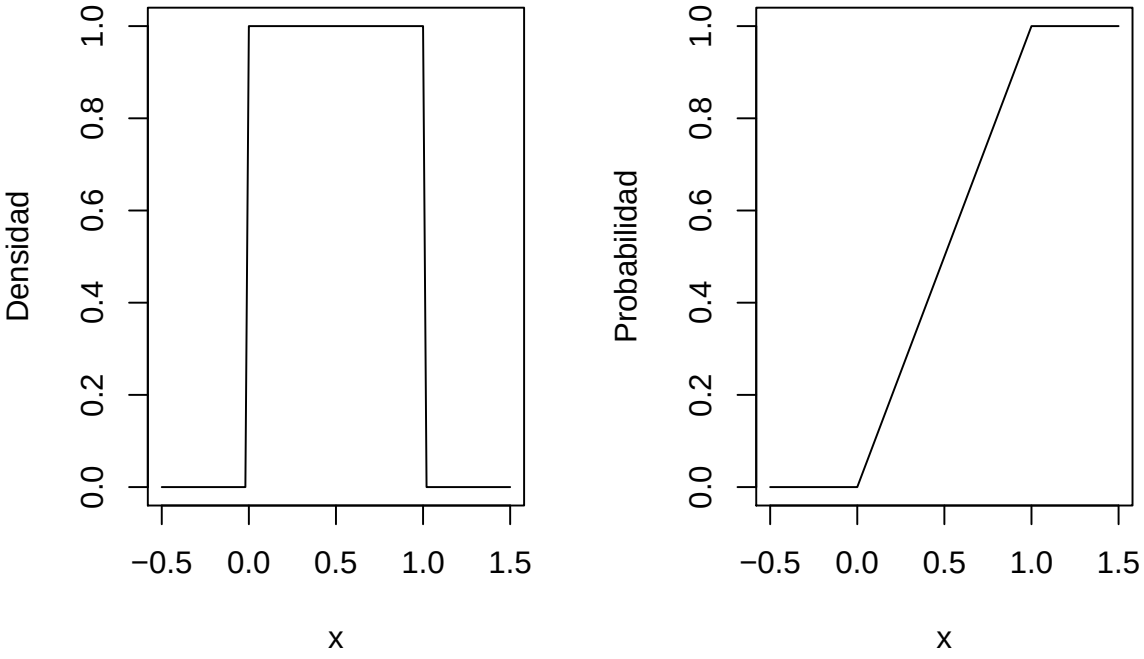
$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } a < x < b \\ 0 & \text{en otro caso} \end{cases}$$

La distribución uniforme más utilizada es la uniforme en el intervalo (0,1) ya que cualquier otra puede obtenerse usando una transformación lineal de esta.

Teorema: si $X \sim Unif(a, b)$ entonces $Y = cX + d$ sigue una distribución uniforme en el intervalo $(ca + d, cb + d)$

Nota a cualquier transformación del tipo $Y = cX + d$ se le denomina **transformación de localización y escala** donde d varia la *localización* (el centro) de la variable y c varia su *escala* o variabilidad.

El siguiente gráfico nos muestra la función de densidad y distribución de una $Unif(0, 1)$



Nota: Distribución uniforme en R La función de densidad de una distribución uniforme puede obtenerse en R usando el comando *dunif(x,min,max)*. Por tanto, sea $X \sim Unif(2, 4)$, la densidad en $X = 2{,}5$ puede calcularse como:

```
dunif(2.5,2,4)
```

```
## [1] 0.5
```

Del mismo modo, la función de distribución acumulada se calcula utilizando el comando *punif(x,min,max)* y, por tanto, la probabilidad de $X < 3$ será

```
punif(3, 2, 4)
```

```
## [1] 0.5
```

y los cuantiles pueden calcularse usando *qunif(p,min,max)* donde p será la probabilidad para la que queremos calcular el cuantil. Por ejemplo, el primer cuartil o cuantil 0.25 ($p = 0{,}25$) será:

```
qunif(0.25, 2)
```

```
## Warning in qunif(0.25, 2): NaNs produced
```

```
## [1] NaN
```

4.3.2. Distribución Normal.

Dentro de las distribuciones que nos sirven para describir variables continuas, existe una que tiene especial relevancia. Se trata de la que denominamos *Distribución Normal* también conocida como *campana de Gauss* (de lo que podemos deducir que tiene forma de campana y que fue "descubierta" por el matemático, físico y astrónomo alemán Karl Friedrich Gauss).

La importancia de la distribución de Gauss (o Gaussiana) reside en tres aspectos:

1. Sus propiedades matemáticas. Las funciones de densidad y distribución acumulada de la distribución normal tienen determinadas características que permiten simplificar muchos cálculos probabilísticos (y estadísticos).
2. Su *Naturalidad*. Muchas cantidades que medimos habitualmente muestran seguir una distribución normal. Sucede así, por ejemplo, con el peso o la altura de una población homogénea (aquella en la que todas las personas tienen unas características similares), con el número de granos de maíz en las mazorcas de una determinada especie o la resistencia de un determinado mineral.
3. Su relación con las *grandes muestras*. La distribución normal aparece automáticamente cuando tenemos una muestra muy grande ya que, si bien la muestra per-se no tiene un comportamiento normal, la suma de sus valores si lo tendrá. Está propiedad la estudiaremos con más cuidado en la siguiente sección y viene bajo el nombre de *Teorema central del límite*.

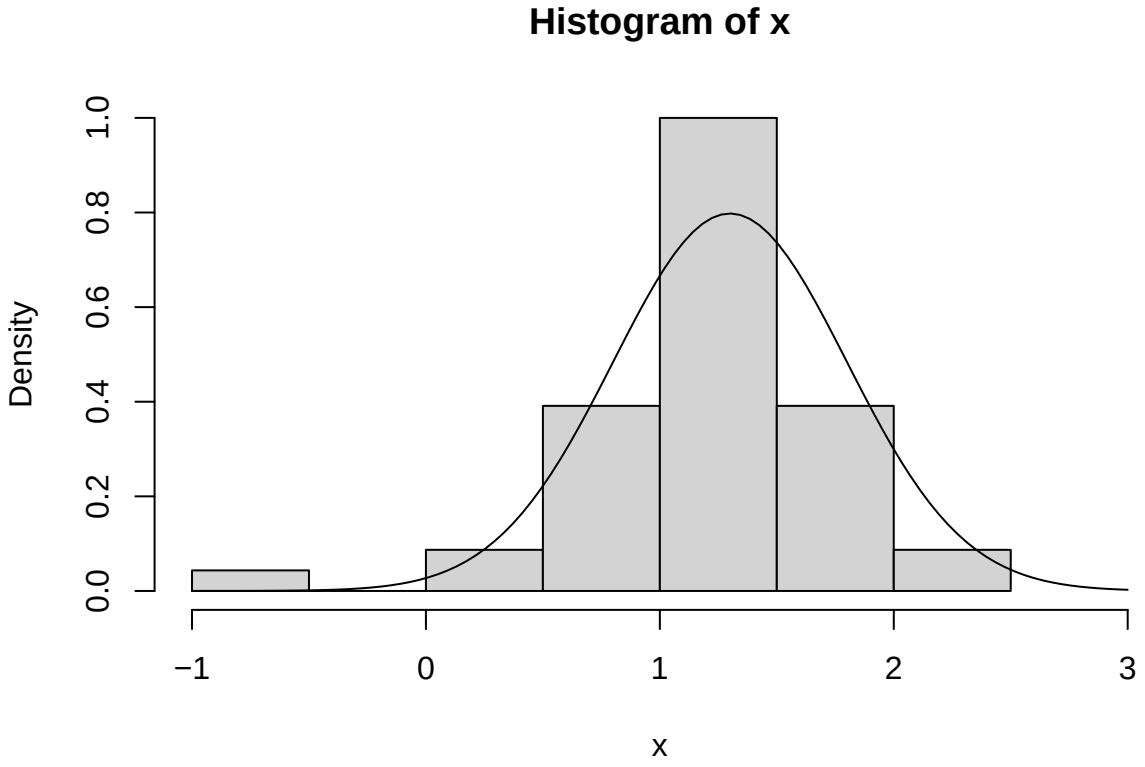
Pero veamos como se define:

Definición: decimos que una variable aleatoria X tiene una **distribución normal con media μ y varianza σ^2** ($X \sim N(\mu, \sigma^2)$) con $-\infty \leq \mu \leq \infty$, $\sigma > 0$ si su función de densidad puede expresarse como:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

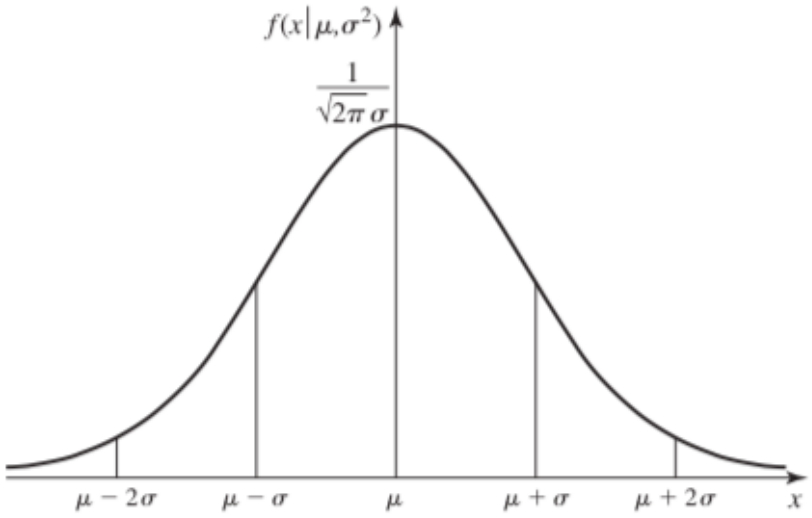
Haciendo integración por partes es relativamente sencillo probar que, efectivamente, la media de X será $E(X) = \mu$ mientras que su varianza será $Var(X) = \sigma^2$

Pensemos en una empresa de motores que necesita saber cual es la emisión de gases de un nuevo tipo de motor que están desarrollando y conocer cual es la probabilidad de que estos emitan más gases de lo permitido. El siguiente histograma muestra los datos recogidos para 46 motores así como una distribución normal que se aproxima bastante bien a los datos recogidos.



Otras propiedades de la distribución normal son: 1. Su función de densidad $f(x \mid \mu, \sigma^2)$ es una función simétrica alrededor del punto $x = \mu$ (que es también su máximo) y por tanto su media, su mediana y su moda son iguales. 2. La desviación estándar σ esta relacionada con determinados cuantiles de la distribución normal ya que, por ejemplo, el 95 % de la probabilidad queda entre (aproximadamente) $\mu + 2\sigma$ y $\mu - 2\sigma$ y es, prácticamente, imposible (probabilidad inferior a 0.01) encontrar valores a una distancia de más de 3 desviaciones estándar de la media.

Podemos ver todas estas características resumidas en la siguiente figura:



Otra propiedad importante de la distribución normal es que una combinación lineal de una variable normal siguen siendo normal:

Teorema: sea $X \sim N(\mu, \sigma^2)$, si definimos una nueva variable $Y = aX + b$, Y también tendrá una distribución normal con media $a\mu + b$ y varianza $a^2\sigma^2$

Dentro de la familia de distribuciones normales[7], existe una que es especialmente relevante. Se trata de la distribución conocida como **Normal Estandard** y que se corresponde con $\mu = 0$ y $\sigma^2 = 1$.

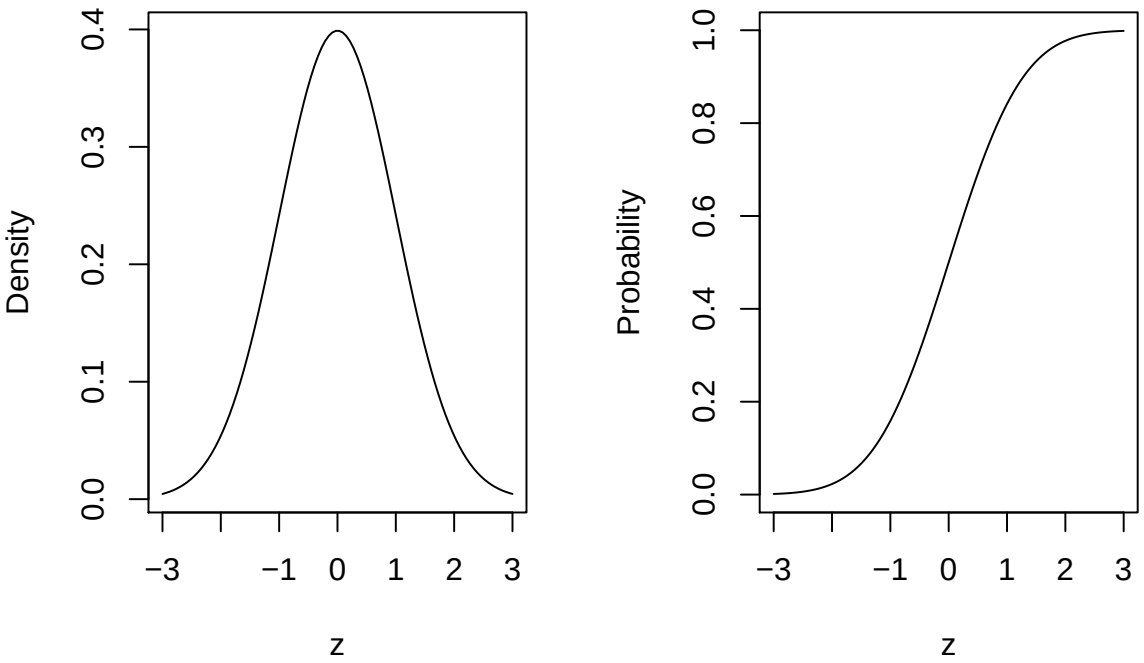
Por el teorema anterior es fácil ver que cualquier variable aleatoria X con distribución normal de media μ y varianza σ^2 puede transformarse en una variable Z con una distribución normal estándar simplemente restándole la media y dividiendo por la desviación estándar, esto es:

$$Z = \frac{X - \mu}{\sigma}.$$

A este proceso se le conoce como estandarización.

La distribución normal estándar se utiliza en muchos contextos y tiene, incluso, su propia nomenclatura. En particular, a una variable que tenga esta distribución la denotamos por Z, su función de densidad se denota por $\varphi(z)$ y su función de distribución acumulada por $\Phi(z)$. La siguiente figura muestra la función de densidad y distribución de una variable con distribución normal estándar.

[7] todas las posibles distribuciones normales que se obtiene al cambiar los parámetros μ y σ



Una aplicación particularmente importante de la estandarización de una variable aleatoria normal es la utilización de tablas de probabilidad. Y es que, cuando el acceso a un ordenador no era tan habitual como ahora, resultaba muy útil estandarizar para calcular probabilidades ya que estas están recogidas en tablas fáciles de utilizar.

Veamos un ejemplo: queremos calcular la $P(5 \leq X \leq 8)$ donde $X \sim N(4, 2)$.

$$P(5 \leq X \leq 8) = P(\frac{5-4}{2} \leq \frac{X-4}{2} \leq \frac{8-4}{2}) = P(0{,}5 \leq Z \leq 2)$$

donde Z es una variable aleatoria con distribución normal estándar. Podemos buscar entonces en las tablas los valores de $\Phi(2) = 0{,}9772$ y $\Phi(0{,}5) = 0{,}6914$ y calcular

$$(5 \leq X \leq 8) = 0{,}9772 - 0{,}6914 = 0{,}2858.$$

Otro aspecto importante de la distribución normal es la distribución de la combinación lineal de variables normales independientes entre si.

Teorema: sean $X_1, \ldots, X_k$ un conjunto de variables aleatorias independientes y normalmente distribuidas $X_i \sim N(\mu_i, \sigma_i^2$ (para $i = 1, \ldots, k$), su suma $Y = X_1 + \cdots + X_k$ sigue una distribución normal de media $\mu_1 + \cdots + \mu_k$ y varianza $\sigma_1^2 + \cdots + \sigma_k^2$.

Como consecuencia, sean $a_1, \ldots, a_k$ y b constantes tal que, al menos existe $a_j \neq 0$ la combinación lineal $Y = a_1X_1 + \cdots + a_kX_k + b$ sigue una distribución normal de media $a_1\mu_+ \cdots + a_k\mu_k + b$ y varianza $a_1^2\sigma_1^2 + \cdots + a_k^2\sigma_k^2$.

Una combinación lineal muy particular (y útil) de variables aleatorias es la **media muestral**

Definición sean un conjunto de n variables aleatorias $X_1, \ldots, X_n$ definimos su **media muestral** como la variable aleatoria $\frac{1}{n}\sum_{i=1}^{n} X_i$. Esta variable aleatoria suele denotarse por $\bar{X}_n$.

Dado el teorema anterior, si las variables $X_1, \ldots, X_n$ son independientes y vienen todas de la misma distribución $X_i \sim N(\mu, \sigma^2)$, se demuestra que $\bar{X}_n \sim N(\mu, \sigma^2/n)$

Pensemos, por ejemplo, que la altura de una determinada población X sigue una distribución normal de media 1.60 y varianza 0.05. Podemos suponer que la altura de cada persona X_i (antes de conocerla) será una variable aleatoria con esa misma distribución. Por tanto, si pensamos en la variable aleatoria que representa la media de la altura de 10 personas $\bar{X}_{10}$ de esa población será una variable aleatoria de media 1.60 y varianza $\sigma^2 = 0{,}05/10 = 0{,}005$.

Nota: Distribución normal en R La función de densidad de una distribución normal puede obtenerse en R usando el comando *dnorm(x,mean,sd)* donde *mean*=μ y *sd*=σ. Por tanto, la densidad en $X = 4$ puede calcularse como:

```
dnorm(4, mean = 2, sd =0.5 )
```

```
## [1] 0.0002676605
```

Del mismo modo, la función de distribución acumulada se calcula utilizando el comando *pnorm(x,mean,sd)* y podemos calcular la probabilidad de $X < 4$ como

```
pnorm(4, mean = 2, sd =0.5 )
```

```
## [1] 0.9999683
```

y los cuantiles pueden calcularse usando *qnorm(p,mean,sd)* donde p será la probabilidad para la que queremos calcular el cuantil. Por ejemplo, el primer cuartil o cuantil 0.25 ($p = 0{,}25$) será:

```
qnorm(0.25, mean = 2, sd =0.5 )
```

```
## [1] 1.662755
```

4.3.3. Distribución Lognormal

La primera distribución derivada de la distribución normal es la distribución conocida como Lognormal y que modeliza el comportamiento una variable cuyo logaritmo tiene una distribución normal, es decir:

Definición sea X una variable aleatoria tal que $\log(X) \sim N(\mu, \sigma^2)$ diremos que X sigue una distribución lognormal de parámetros μ y σ^2.

Se puede comprobar que la esperanza y la varianza de una variable lognormal son: - $E(X) = \exp(\mu + 0{,}5\sigma^2)$. - $Var(X) = exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1]$.

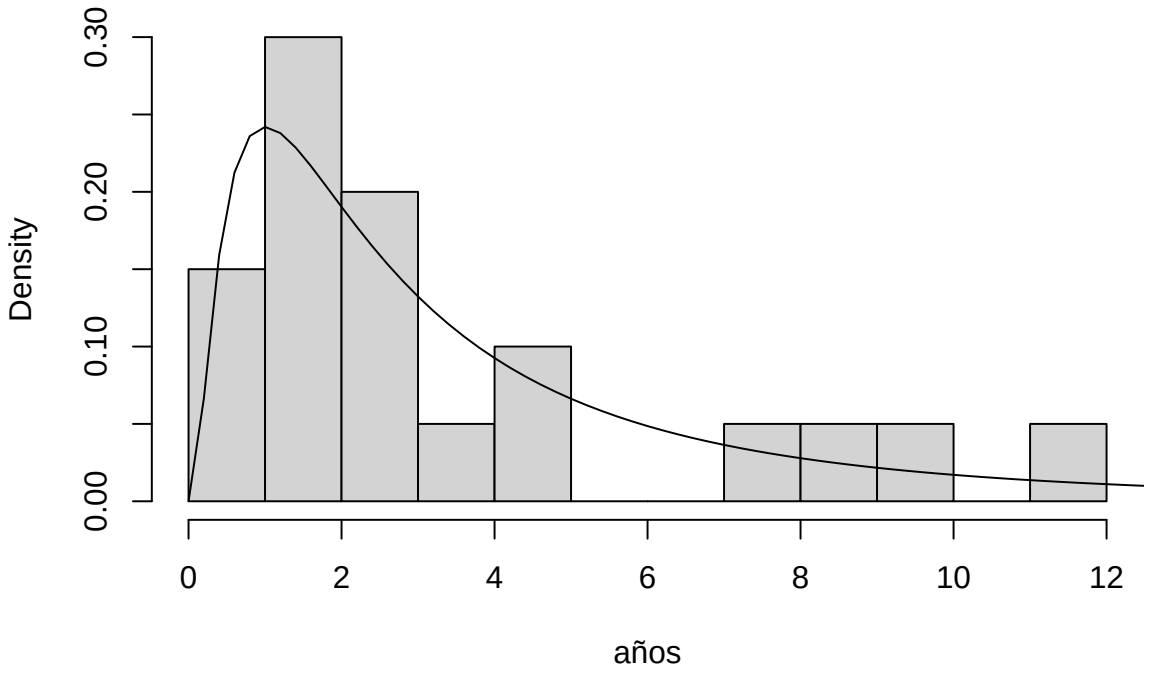
Pensemos, por ejemplo, en una fabrica de ropa que comprueba la resistencia de sus prendas y calcula el tiempo que van a durar (medido en años). Se sabe que el logaritmo del tiempo de vida de las prendas es normal de media $\mu = 1$ y desviación estándar $\sigma = 1$. Cual es la probabilidad de que una prenda dure más de 2 años.

Queremos calcular $P(X \geq 2)$ que, por la monotonía del logaritmo es equivalente a calcular $P(\log(X) \geq \log(2)) = P(\log(X) \geq 0{,}6931472)$

```
1- pnorm(log(2),1,1)
```

```
## [1] 0.6205223
```

Por tanto, la probabilidad de que la prenda dure más de dos años es 0.6205223.



Nota: Distribución lognormal en R La función de densidad de una distribución lognormal también puede obtenerse en R usando el comando *dlnorm(x,meanlog,sdlog)*. Por tanto, sea $X \sim logN(1,1)$, la densidad en $X = 2$ puede calcularse como:

```
dlnorm(2,1,1)
```

```
## [1] 0.1902978
```

Del mismo modo, la función de distribución acumulada se calcula utilizando el comando *plnorm(x,meanlog,sdlog)* y, por tanto, la probabilidad de $X < 2$ será

```
plnorm(2,1,1)
```

```
## [1] 0.3794777
```

y los cuantiles pueden calcularse usando *qlnorm(p,meanlog,sdlog)* donde p será la probabilidad para la que queremos calcular el cuantil. Por ejemplo, el primer cuartil o cuantil 0.25 ($p = 0{,}25$) será:

```
qlnorm(0.25, 1, 1)
```

```
## [1] 1.384737
```

4.3.4. Distribución Gamma

La distribución gamma es un modelo común para variables que sólo pueden tomar valores positivos.

Un aplicación muy común de la distribución gamma se da en el estudio de los tiempos entre sucesos Poisson. Por ejemplo, si pensamos en el número de llamadas a una centralita en una hora (variable Poisson de parámetro λ) el tiempo que transcurre entre dos llamadas consecutivas diremos que sigue una distribución gamma.

Pero, para poder definir la función de densidad de una distribución gamma debemos definir primero una función muy conocida en matemáticas y que tiene que ver con la generalización continua del concepto de factorial. Se trata de la *Función Gamma*

Definición: la función gamma para cualquier valor positivo α, $\Gamma(\alpha)$ viene definida por la siguiente integral:

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

y, en particular

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

Teorema: sea $\alpha > 1$

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

De este modo se comprueba que si $\alpha = n$ con n un numero entero

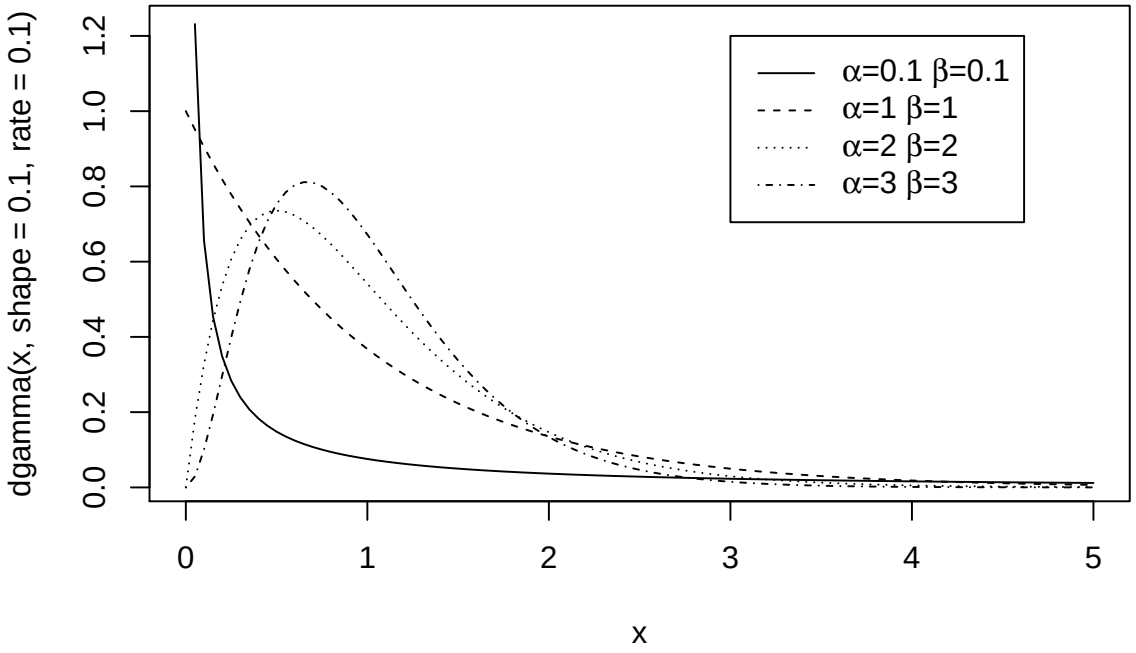
$$\Gamma(n) = (n-1)!$$

Una vez definida la función gamma podemos pasar a definir la distribución gamma.

Definición: decimos que una variable X sigue una distribución gamma de parámetros α y β ($X \sim Ga(\alpha, \beta)$) si su función de densidad es

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

La media y la varianza de una distribución gamma son: - $E(X) = \frac{\alpha}{\beta}$ - $Var(X) = \frac{\alpha}{\beta^2}$



Teorema: Si tenemos un conjunto de variables aleatorias $X_1, \ldots, X_k$ independientes tales que $X_i \sim Ga(\alpha_i, \beta)$, la suma $X_1 + \cdots + X_k$ tiene una distribución gamma de parámetros $\alpha_1 + \cdots + \alpha_k$ y β.

4.3.4.1. Distribución exponencial

Un caso particular de la distribución gamma se da cuando el parámetro $\alpha = 1$. Se trata de un modelo que se aplica muy habitualmente a tiempos de espera.

Definición: una variable aleatoria X sigue una **Distribución Exponencial** de parámetro β si su función de densidad es:

$$f(x \mid \beta) = \begin{cases} \beta e^{-\beta x} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}.$$

Es fácil deducir que $E(X) = 1/\beta$ y $Var(X) = 1/\beta^2$.

Una propiedad muy importante de esta distribución es la falta de memoria

Teorema: sea X una variable aleatoria con distribución exponencial de parámetro β y sea $t > 0$. Entonces, para todo $h > 0$,

$$P(X \geq t + h \mid X \geq t) = P(X \geq h).$$

De forma intuitiva, el tiempo de espera restante es, en cierto modo, independiente del tiempo ya transcurrido.

Otra propiedad interesante de la distribución exponencial es que podemos determinar la distribución del mínimo entre un grupo de variables exponenciales. Por ejemplo, si tenemos un conjunto de n bombillas cuyo tiempo de duración X_i es exponencial de parámetro β para cada $i = 1, \ldots, n$ ¿Cual será el mínimo tiempo que tendremos que esperar para ver un fallo?

Teorema: sean $X1, \ldots, X_n$ un conjunto de variables aleatorias independientes e idénticamente distribuidas (iid) exponencial de parámetro β. La distribución de $Y_1 = \text{mín}\{X_1, \ldots, X_n\}$ será exponencial de parámetro $n\beta$.

Ahora, ¿Cual será el tiempo hasta que falle la siguiente bombilla? Dada la propiedad de falta de memoria de la distribución exponencial, el tiempo hasta que la siguiente bombilla falle Y_2 tendrá también una distribución exponencial pero, esta vez, de parámetro $(n-1)\beta$ y de manera recursiva:

Teorema el tiempo entre dos sucesos consecutivos ($k-1$ y k) de un total de n donde cada uno de ellos era exponencial de parámetro β sigue una distribución exponencial de parámetro $(n+1-k)\beta$

4.3.4.2. Relación con el proceso de poisson

Es importante tener en cuenta que en el teorema anterior tenemos un numero fijo de elementos y sabemos que todos tuvieron un mismo tiempo inicial. Sin embargo, estas no son siempre las circunstancias.

Imaginemos, por ejemplo, que el encargado de una tienda de ropa quiere saber cuanto tiempo transcurrirá hasta que entre la siguiente persona. Se trata de una situación parecida a la del teorema pero con la particularidad de que no sabemos cuantas personas van a entrar en total ni cuando ha salido cada una de su casa.

Teorema: Supongamos que las llegadas suceden según un proceso de Poisson de parámetro λ, sea Z_k el tiempo hasta que se produce la k-ésima llegada, definimos el tiempo entre llegadas: $Y_1 = Z_1$ e $Y_k = Z_k - Z_{k-1}$. Se puede demostrar que $Y_1, Y_2, \ldots$ son variables independientes e idénticamente distribuidas con distribución exponencial de parámetro $\beta = \lambda$.

Como consecuencia, la distribución del tiempo hasta la k-ésima llegada, Z_k es una Gamma de parámetros k y β.

Nota: Distribución Gamma en R La función de densidad de una distribución gamma también puede obtenerse en R usando el comando *dgamma(x,shape, rate)* donde *shape*=α y *rate*=β. Por tanto, sea $X \sim Ga(2,2)$, la densidad en $X = 2$ puede calcularse como:

```
dgamma(2,2,2)
```

```
## [1] 0.1465251
```

Del mismo modo, la función de distribución acumulada se calcula utilizando el comando *pgamma(x,shape, rate)* y, por tanto, la probabilidad de $X < 2$ será

```
pgamma(2,2,2)
```

```
## [1] 0.9084218
```

y los cuantiles pueden calcularse usando *qchisq(p,shape,rate)* donde p será la probabilidad para la que queremos calcular el cuantil. Por ejemplo, el primer cuartil o cuantil 0.25 ($p = 0{,}25$) será:

```
qgamma(0.25, 2,2)
```

```
## [1] 0.4806394
```

4.3.5. Distribución Beta

La distribución beta es un modelo habitual para variables que se encuentran en el intervalo $[0, 1]$. Su uso es muy común, por ejemplo, para estudiar la proporción de veces que sucede determinado evento, es decir, como distribución para la probabilidad π de una variable Bernoulli (o binomial) cuando ésta es desconocida.

Al igual que con la distribución gamma, antes de pasar a definir la distribución beta debemos conocer la función matemática con el mismo nombre.

Definición: para todo α y β positivos, se define la función beta:

$$B(\alpha,\, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$$

Una de las propiedades de la función beta es que puede expresarse en términos de la función gamma como:

Teorema dados $\alpha > 0$ y $\beta > 0$

$$B(\alpha,\, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Y una vez definida esta función podemos pasar a definir la distribución beta: **Definición:** una variable aleatoria X tiene una **Distribución Beta** con parámetros $\alpha > 0$ y $\beta > 0$ si su función de densidad es:

$$f(x \mid \alpha,\, \beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)\Gamma(\alpha)} x^{\alpha-1}(1-x)^{\beta-1} & \text{si } 0 < x < 1 \\ 0 & \text{en otro caso} \end{cases}$$

La esperanza y la varianza de una variable aleatoria con distribución beta son: - $E(X) = \frac{\alpha}{\alpha+\beta}$ - $Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

4.3.5.1. Relación con la distribución gamma.

Teorema: Sean U y V variables aleatorias independientes y sea $U \sim Ga(\alpha,\, 1)$ y $V \sim Ga(\beta, 1)$:

1. $X = U/(U+V)$ e $Y = U + V$ son v.a. independientes ,
2. $X \sim Be(alpha, \beta)$ y
3. $Y \sim Ga(alpha + beta, 1)$

4.3.5.2. Relación con la distribución Uniforme

una distribución beta de parámetros $\alpha = \beta = 1$ es una distribución uniforme en el intervalo $[0, 1]$.

4.3.5.3. Teorema de Bayes para variables aleatorias. El proceso Beta-Binomial

Recordemos el ejemplo del juicio de Castaneda vs Partida en el que en una población con un 79.1 % mejicano-americanos, de de las 220 personas qua habían sido elegidas para ser jurado popular sólo 100 tenían dicha procedencia.

Sabemos que la variable X que mide el número de personas mejicano-americanas en el jurado popular sigue una Binomail de parametros $Bi(N = 220, \pi)$ donde π es la proporción de personas con esta procedencia en el jurado. Si suponíamos que esta proporción es

la misma que en la población ($\pi = 0{,}791$) podíamos calcular la probabilidad de que $X = 100$ pero, en realidad, nos interesaba conocer cual era realmente la proporción π dado que hay $X = 100$. Bien, ahora podemos definir P como la proporción de personas mejicano-americanas en el jurado y suponer que sigue, a priori, una distribución beta $P \sim Be(\alpha, \beta)$.

Nos interesa saber, una vez observado $X = 100$ (a posteriori), cuál es la probabilidad de que P sea menor que $0{,}8 \times 0{,}791 = 0{,}6328$ lo que, para nosotros, supondría un claro caso de discriminación

Si modificamos ligeramente el teorema de Bayes para adaptarlo a variables aleatorias tenemos que dado un valor p para P:

$$f(p \mid N = 220,\, X = 100,\, \alpha,\, \beta) = \frac{f(X = 100 \mid N = 220,\, \pi = p) f(p \mid \alpha,\, \beta)}{f(X = 100 \mid N = 220)}$$

El numerador de de esta ecuación es:

$$\binom{220}{100} p^{100}(1-p)^{120} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

Y el denominador $f(X = 100)$ se puede calcular utilizando la versión continua del teorema de la probabilidad total como

$$f(X = 100 \mid N = 220) = \int_0^1 f(X = 100 \mid N = 220, \pi = p) f(p \mid \alpha, \beta) dp$$

Podemos ver que $f(\pi \mid N = 220,\, X = 100,\, \alpha,\, \beta)$, como función de P es una constante (que no depende de P) multiplicada por $P^{100+\alpha-1}(1-P)^{220+\beta-1}$ que es, claramente, el núcleo de una distribución beta de parámetros $100 + \alpha$ y $220 + \beta$ por lo que podemos decir que, una vez observado $X = 100$ la distribución de P se ha transformado en una beta con esos parámetros.

Si hubiésemos elegido, a priori, una distribución uniforme para P, es decir, una beta con $\alpha = \beta = 1$ la probabilidad *a posteriori* de que $P < 0{,}6328$ es:

```
pbeta(0.6328,101,221)
```

```
## [1] 1
```

por lo que, claramente, existe discriminación hacia las personas de procedencia mejicano-americana.

Este resultado puede generalizarse con el siguiente teorema

Teorema: Proceso Beta Binomial. Supongamos que π es una variable aleatoria con distribución beta de parámetros α y β y que X es una variable aleatoria que, condicionada a que $\pi = p$ sigue una distribución binomial de parámetros N y p. Entonces, la distribución de π condicionada a $X = x$ es $Be(\alpha + x,\ \beta + N - x)$.

Nota: La versión del teorema de Bayes para variables aleatorias se utiliza, sobre todo, en el paradigma Bayesiano de la estadística aunque también en el su versión clásica o frecuentista.

Ya hemos comentado que a la función de distribución sobre π antes de observar los datos se le conoce como **distribución a priori** mientras que el resultado de aplicar el teorema de Bayes se conoce como **distribución a posteriori**.

Cabe destacar también que la función de densidad aplicada sobre los datos observados $f(X \mid \pi)$ es una función de π que recibe el nombre de **Verosimilitud** y que es muy importante tanto en probabilidad como en estadística ya que nos ayudará a determinar el valor del parámetro más *verosímil* según los datos observados.

Nota: Distribución Beta en R La función de densidad de una distribución beta también puede obtenerse en R usando el comando *dbeta(x,shape1,shape2)* donde *shape1*=α y *shape2*=β. Por tanto, sea $X \sim Be(3, 2)$, la densidad en $X = 2$ puede calcularse como:

```
dbeta(2,3,2)
```

```
## [1] 0
```

Del mismo modo, la función de distribución acumulada se calcula utilizando el comando *pbeta(x,shape1,shape2)* y, por tanto, la probabilidad de $X < 2$ será

```
pbeta(2,3,2)
```

```
## [1] 1
```

y los cuantiles pueden calcularse usando *qbeta(p,shape1,shape2)* donde p será la probabilidad para la que queremos calcular el cuantil. Por ejemplo, el primer cuartil o cuantil 0.25 ($p = 0{,}25$) será:

```r
qbeta(0.25,3,2)
```

```
## [1] 0.4563217
```

4.4. Ejercicios

1. Supongamos que la probabilidad de que un cierto experimento sea un éxito es $\pi = 0{,}4$ y sea X el número de éxitos obtenidos en 15 repeticiones independientes del experimento. Usando R calcula la $P(6 \leq X \leq 9)$.

2. Tres personas A, B y C lanzan a canasta. Supongamos que A lanza 3 veces y la probabilidad de que enceste es 1/8, B lanza 5 veces y su probabilidad de encestar es 1/4, y C dispara dos veces y su probabilidad de encestar de 1/2. ¿Cuál es el número esperado de canastas?

3. En un ensayo clínico, la probabilidad de éxito para un tratamiento A es 0,5 y la probabilidad de éxito para el tratamiento B es 0,6. Suponiendo que hay cinco pacientes en cada grupo, calcula la probabilidad de que el grupo A tenga, al menos, tantos éxitos como el grupo B.

4. Supongamos que una caja contiene 5 bolas rojas y 10 azules. Si sacamos 7 bolas al azar sin reemplazamiento, ¿Qué probabilidad hay de que, al menos, 3 sean rojas?

5. Considera un grupo de T personas cuyas alturas son $a_1, \ldots, a_T$. Supongamos que se eligen N personas de este grupo al azar y sin reemplazamiento. Sea X la suma de las alturas de esas N personas. Determina la media de X.

6. Supongamos que el número de defectos en un rollo de tela producidos durante el proceso de fabricación sigue una distribución de Poisson con media 0,4. Si inspeccionamos una muestra aleatoria de 5 rollos, ¿cuál es la probabilidad de encontrar al menos 6 defectos?

7. Supongamos que X_1 y X_2 son v.a. independientes con distribución de Poisson de medias λ_1 y λ_2 respectivamente. Para un valor fijo $k = 1, 2, \ldots$, determina la probabilidad condicional de X_1 dado que $X_1 + X_2 = k$.

8. Supongamos que en una secuencia de lanzamientos independientes de una moneda con probabilidad de obtener cara 1/30:

a. ¿Cuál es el número esperado de cruces antes de obtener 5 caras?

b. ¿Cuál es la varianza del número de cruces obtenido antes de obtener 5 caras?

9. Si la temperatura en grados Fahrenheit de una determinada zona sabemos que sigue una distribución normal con media 68 grados Fahrenheit y desviación estándar de 4 grados, ¿cual será la distribución de la temperatura de esa misma zona en grados Celsius?

10. Supongamos que los diámetros de una serie de tornillos almacenadas en una caja sigue una distribución normal con media 2 cm y desviación estándar 0.03 cm. Del mismo modo, los diámetros de una serie de tuercas en otra caja, siguen una distribución normal de media 2,02 cm y desviación estándar 0,04 cm. Un tornillo y una tuerca encajarán juntos si el diámetro de la tuerca es mayor que el diámetro del tornillo pero la diferencia entre ambos no es mayor de 0.05 cm. Si una tuerca y un tornillos son seleccionados al azar, cual es la probabilidad de que encajen juntos?

11. Supongamos que el voltaje en un determinado circuito eléctrico sigue una distribución normal con media 120 kw y desviación estándar 2 kw. Si se toman 3 medidas de manera independiente, cual es la probabilidad de que las tres estén entre 116 y 118 kw?

12. Supongamos que se están testando n elementos independientes y que el tiempo de vida de cada uno de ellos (X_i) sigue una distribución exponencial de parámetro β. Determina la esperanza del tiempo hasta que fallen tres de ellos. Pista: El valor requerido es $E(Y_1 + Y_2 + Y_3)$ donde Y_i es el tiempo hasta el i-ésimo fallo.

13. Cinco estudiantes deben hacer un examen, cada uno de manera independiente. Si tiempo que cada uno/a tarda en realizarlo es exponencial de media 80. Sabiendo que el examen ha comenzado a las 9:00 a.m. ¿Cuál es la probabilidad de que, al menos uno/a lo acabe antes de las 9:40 a.m.?

5. Teoremas de Convergencia y distribuciones derivadas

5.1. Introducción

Hasta ahora hemos visto algunas distribuciones conocidas para las que sabemos prácticamente todo (su función de densidad o probabilidad, su función de distribución acumulada, su esperanza, su varianza...). Sin embargo, es habitual encontrarse en situaciones en las que nuestra variable de interés no tiene, en principio, una distribución conocida.

La simulación es una de las técnicas más útiles a la hora de aproximar esas distribuciones *desconocidas* y veremos como utilizarla dentro de un par de temas.

Otra de las opciones para aproximar estas distribuciones son lo que se conocen como *teoremas límite* y que son dos de las herramientas más utilizadas en probabilidad y estadística. Se trata de la ley de los grandes números y del teorema central del límite.

A continuación introduciremos estos teoremas y, como consecuencia de los mismos, estudiaremos dos distribuciones continuas, la χ^2 y la t de Student.

5.2. Ley de los grandes números

Para esta sección y la siguiente vamos a suponer que tenemos una serie de variables aleatorias $X_1, X_2, \ldots$ independientes e idénticamente distribuidas, con media μ y desviación estándar σ. Como ya definimos cuando hablábamos de la distribución normal, la media de un conjunto de n de estas variables es:

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

Utilizando las propiedades de la esperanza y la varianza, es fácil comprobar que $E(\bar{X}_n) = \mu$ y $Var(\bar{X}_n) = \sigma^2/n$.

Lo que dice la ley de los grandes números es que, a medida que tengo más datos, la media muestral converge a la verdadera media de la variable. Formalmente, esta *convergencia* puede darse de dos maneras, fuerte y débil y, de ahí, las dos versiones de la ley de los grandes números:

Teorema: Ley fuerte de los grandes números. La media muestral $\bar{X}_n$ converge a la verdadera media μ con probabilidad 1 o, lo que es lo mismo, el evento $\bar{X}_n \to \mu$ tiene probabilidad 1.

Teorema: Ley débil de los grandes números. Para todo $\epsilon > 0$, $P(|\bar{X}_n - \mu| > \epsilon) \to 0$ a medida que $n \to \infty$. (A este tipo de convergencia se le denomina **convergencia en probabilidad**)

La ley de los grandes números es esencial para la ciencia y es algo que usamos sin apenas darnos cuenta.

Cada vez que aproximamos la probabilidad de que algo pase a través de la proporción de veces que lo hemos observado o cada vez que estimamos la media de una cantidad a partir de la media de nuestras observaciones, estamos, implícitamente, usando la ley de los grandes números.

Volveremos a esta ley cuando entremos en el tema de simulación.

5.3. Teorema central del limite.

Bien, en la versión fuerte de ley de los grandes números decíamos que $\bar{X}_n$ convergía a μ con probabilidad 1 pero, cual es la distribución de $\bar{X}_n$ en su camino a convertirse en una constante (μ)?

Teorema central del límite. Sean $X_1, \ldots, X_n$ variables iid con media μ y varianza σ^2. Para todo z

$$\lim_{n\to\infty} \left[P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z \right) \right] = \Phi(z).$$

Lo que es equivalente a decir que la variable aleatoria $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ converge a una distribución normal estándar.

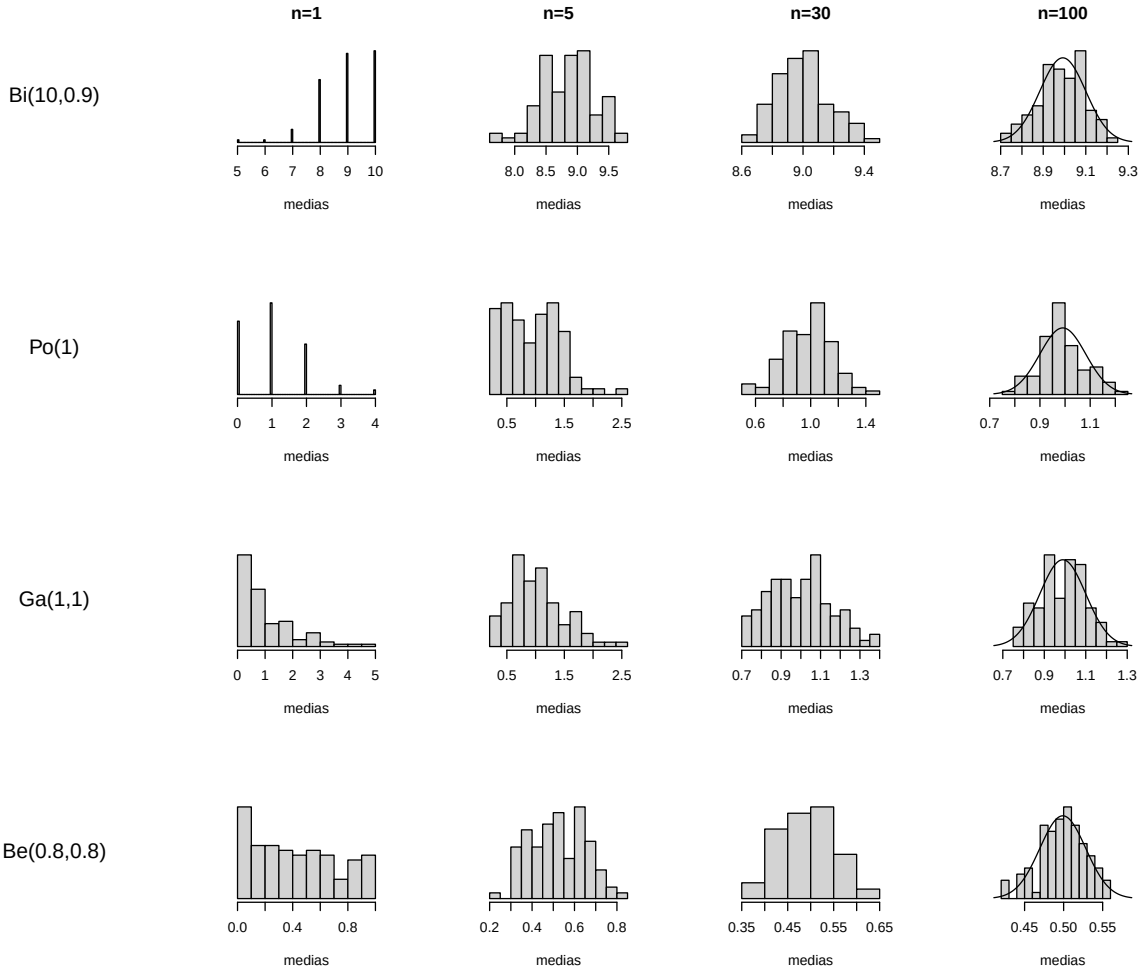
El teorema central del límite expresa la convergencia a la normal estándar en términos de $n \to \infty$ pero no es necesario llegar a infinito para que la distribución de la media muestral sea aproximadamente normal. Es por esto que se suele utilizar una versión aproximada del teorema que dice:

Teorema central del límite. Versión aproximada. Para una muestra de n variables independientes e idénticamente distribuidas con n grande tenemos que su media $\bar{X}_n$ se comporta aproximadamente normal $N(\mu, \sigma^2/n)$.

Es importante recalcar la importancia de este teorema. No importa cual sea la distribución de X_i, podría ser incluso discreta, sólo necesitamos que su media y su varianza sean finitas

para que, teniendo una cantidad suficientemente grande de ellas, podamos suponer que la media muestral se comporta de manera normal.

Lo podemos ver en el siguiente gráfico e el cual partimos de variables aleatorias con distribuciones diversas. Los histogramas muestran la media de muestras aleatorias de tamaño n (para $n = 1, 5, 30$ y 100) procedentes de dicha distribución.



Convergencia de una Poisson a una normal. Sea $Y \sim Po(n)$, dadas las propiedades de la suma de distribuciones Poisson podemos considerarla como la suma de n variables $X_i \sim Po(1)$. De esta forma, por el teorema central del limite, para un n grande la distribución de Y puede considerarse

$$Y \sim N(n, n)$$

Convergencia de una gamma a una normal. Sea $Y \sim Ga(n, \lambda)$. De nuevo, por las propiedades de la suma de distribuciones Gamma Y puede considerarse como la suma de

$X_i \sim Ga(1, \lambda)$ y, por el teorema central del límite, para n grande

$$Y \sim N\left(\frac{n}{\lambda}, \frac{n}{\lambda^2}\right)$$

Convergencia de una binomial a una normal Sea $Y \sim Bi(n, \pi)$, sabemos que Y puede considerarse la suma de n variables Bernoulli de parámetro π. Por tanto, para n grande, podemos considerar que

$$Y \sim N(n\pi,\, n\pi(1-\pi))$$

La aproximación normal a la distribución binomial es una de las más utilizadas en estadística. Sin embargo, cabe tener en cuenta que Y es una variable discreta pero al calcular $P(Y = y)$ utilizando la aproximación obtendríamos un valor de 0. Para corregir esta situación, calcularemos $P(y - 1/2 < Y < y + 1/2)$ de forma que tendremos la probabilidad de un intervalo de longitud distinta de 0. Esta solución es conocida como **corrección por continuidad** y conlleva la siguiente aproximación a la función de probabilidad de una distribución binomial:

$$P(Y = y) = P(y - 1/2 < Y < y + 1/2) = \Phi\left(\frac{y + 1/2 - n\pi}{\sqrt{n\pi(1-\pi)}}\right) - \Phi\left(\frac{y - 1/2 - n\pi}{\sqrt{n\pi(1-\pi)}}\right)$$

Recordemos que ya habíamos visto una aproximación a la distribución binomial cuando n era grande usando la distribución de Poisson. La distribución normal funcionará mejor como aproximación a la distribución binomial cuando π este alrededor de un $1/2$, situación en la cual la distribución binomial será prácticamente simétrica, mientras que la aproximación de Poisson funciona mejor cuando π es pequeño.

5.3.1. Distribuciones derivadas de la distribución normal

Continuando con la distribución de la suma de variables aleatorias podemos definir dos distribuciones más que se pueden derivar de la distribución normal. Estas son la distribución χ^2 (Ji-cuadrada o chi-squared) y la distribución t de Student

5.3.1.1. Distribución χ^2

Definición sean $V = Z_1^2 + \cdots + Z_n^2$ con $Z_i \sim N(0, 1)$ decimos que V sigue una distribución χ^2 con n grados de libertad y lo denotamos por $V \sim \chi_n^2$.

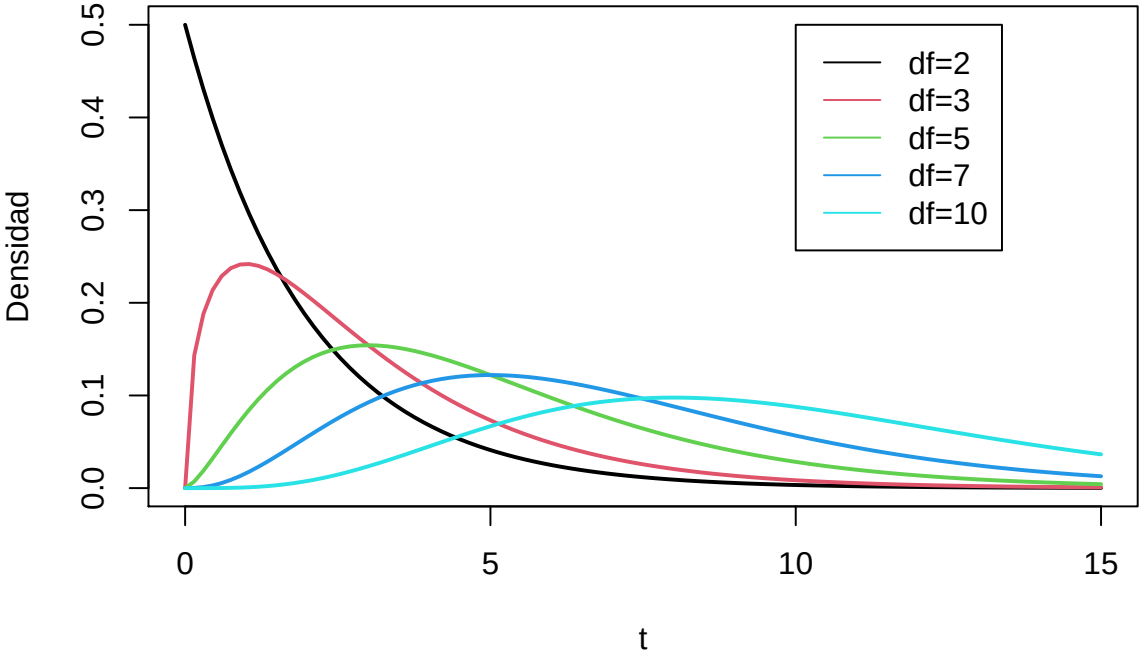
la función de densidad de una distribución χ^2_n puede deducirse del siguiente teorema

Teorema La distribución χ^2_n es un caso particular de la distribución Gamma $Ga(n/2, 1/2)$

A partir de este teorema es fácil ver que:

- $E(V) = n$
- $Var(V) = 2n$

La distribución χ^2 tiene gran importancia en estadística y aparece en el proceso de estimación de la varianza de una distribución normal cuando esta es desconocida.



Nota: Distribución χ^2-cuadrada en R La función de densidad de una distribución χ^2_n también puede obtenerse en R usando el comando *dchisq(x,df)* donde *df* son los grados de libertad. Por tanto, sea $X \sim \chi^2(10)$, la densidad en $X = 2$ puede calcularse como:

```
dchisq(2,10)
```

```
## [1] 0.007664155
```

Del mismo modo, la función de distribución acumulada se calcula utilizando el comando *pchisq(x,df)* y, por tanto, la probabilidad de $X < 2$ será

```
pchisq(2,10)
```

```
## [1] 0.003659847
```

y los cuantiles pueden calcularse usando *qchisq(p,df)* donde p será la probabilidad para la

que queremos calcular el cuantil. Por ejemplo, el primer cuartil o cuantil 0.25 ($p = 0{,}25$) será:

```
qchisq(0.25, 10)
```

```
## [1] 6.737201
```

5.3.1.2. Distribución t-Student

También relacionada con la distribución normal y con la χ^2 encontramos la distribución t de Student, t-Student o simplemente t.

Definición sea una variable

$$T = \frac{Z}{\sqrt{V/n}},$$

donde $Z \sim N(0,1)$ y $V \sim \chi^2_n$, decimos que T sigue una distribución t de Student con n grados de libertad y lo denotamos por $T \sim t_n$.

La densidad de esta distribución viene dada por:

$$f(t \mid n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

La distribución t de Student fue introducida por William Gosset en 1908. William Gosset era maestro cervecero en la compañía Guiness y trabaja en control de calidad. La compañía le pidió que publicará sus resultados bajo seudónimo y él adopto el nombre Student. Esta distribución es también muy importante en estadística y en particular en la metodología de contraste de hipótesis.

Cuando $n = 1$ la distribución t iguala a la distribución de Cauchy

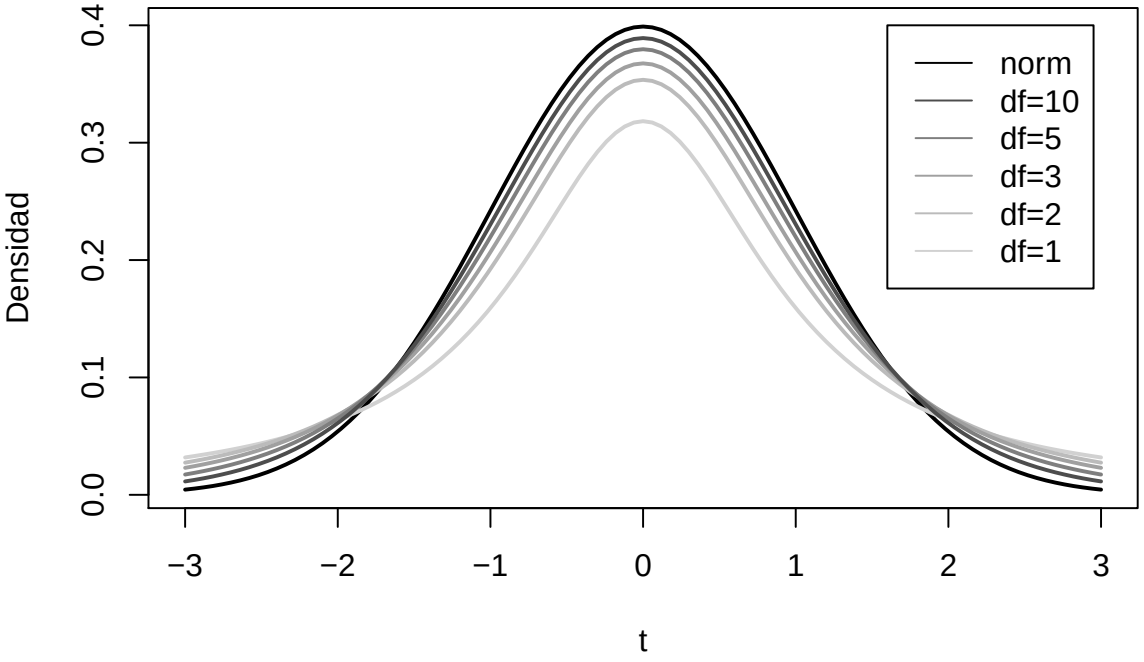
Definición diremos que X/Y tiene una distribución de Cauchy si X e Y son variables aleatorias independientes con distribución normal estándar.

Otras propiedades de la distribución t son:

Teorema Una variable aleatoria $T \sim t_n$ es simétrica y se aproxima a una $N(0,1)$ cuando $n \to \infty$.

La media de una distribución t solo existe para $n > 1$ y es $E(T) = 0$ al igual que su moda y su mediana. En cuanto a la varianza, sólo existe para $n > 2$ y es $Var(T) = n/(n-2)$

En el siguiente gráfico podemos ver la evolución de la densidad de una distribución t_n para $n = 1, 2, 3, 5$ y 10 (de más claro a más oscuro) comparada con una normal estándar (en negro)



Observamos que la distribución t tiene la misma forma que una normal pero con colas más pesadas, es decir, con mayor probabilidad para valores más alejados del 0.

Nota: Distribución t-Student en R La función de densidad de una distribución t-Student también puede obtenerse en R usando el comando *dt(x,df)* donde *df* son los grados de libertad. Por tanto, sea $X \sim St(10)$, la densidad en $X = 2$ puede calcularse como:

```
dt(2,10)
```

```
## [1] 0.06114577
```

Del mismo modo, la función de distribución acumulada se calcula utilizando el comando *pt(x,df)* y, por tanto, la probabilidad de $X < 2$ será

```
pt(2,10)
```

```
## [1] 0.963306
```

y los cuantiles pueden calcularse usando *qt(p,df)* donde p será la probabilidad para la que queremos calcular el cuantil. Por ejemplo, el primer cuartil o cuantil 0.25 ($p = 0{,}25$) será:

```
qt(0.25, 10)
```

```
## [1] -0.6998121
```

5.4. Ejercicios

1. Supongamos que tenemos una muestra aleatoria de tamaño n de variables con distribución normal de media μ y desviación estándar 3. Usa el teorema central del límite para determinar, aproximadamente, el valor más pequeño de n para el cual se cumple la siguiente relación:

$$P(|\bar{X}_n - \mu| < 0{,}3) \geq 0{,}95$$

2. Una máquina produce, cada hora, una cuerda cuya longitud tiene una media de 4 metros y una desviación estándar de 5 cm. Asumiendo que la cantidad de cuerda producida en distintos minutos es independiente e idénticamente distribuida aproxima la probabilidad de que la maquina fabrique, al menos, 250 metros de cuerda en una hora.

6. Vectores aleatorios y distribuciones multivariantes

6.1. Introducción

En muchas ocasiones los procesos en los que estamos interesados no se reducen al estudio de una única variable si no de varias que se distribuyen de manera, digamos, *coordinada*.

Uno de los aspectos fundamentales será extender el concepto de independencia –que ya definimos para experimentos/eventos– al caso de las variables. Se trata, en el fondo, de conceptos que ya hemos trabajado intuitivamente cuando estudiamos las distintas distribuciones a las que pueden dar lugar una serie de sucesos Bernoulli (Binomial si son independientes, hipergeométrica si no), o cuando hablábamos de las n variables aleatorias que aparecen a la hora de tomar una muestra y de las que decíamos que eran **independientes e idénticamente distribuidas**.

Para terminar de comprender el por qué de este tema, veamos algunos ejemplos en los que el estudio conjunto de diversas variables es especialmente necesario:

- **Medicina**: Para evaluar la efectividad de un tratamiento es mucho más informativo trabajar de manera conjunta con distintas medidas biológicas (presión sanguínea, glucosa en sangre...).
- **Genética**: En el estudio de biomarcadores que nos ayuden a estudiar diversas patologías es muy importante entender que los genes no actúan de manera independiente y trabajar con ellos consecuentemente.
- **Series temporales/ datos longitudinales**: A veces el interés no reside en la variable en un momento concreto (estudio transversal) sino entender su evolución temporal (en uno o más sujetos). En ese caso, la observación en cada instante es una variable aleatoria y es natural pensar que todas estarán relacionadas (lo que pasa en un instante no puede ser muy diferente de lo que sucede el anterior o el siguiente).

Extenderemos, por tanto, todo lo que hemos visto sobre la distribución de una variable al caso en que denominaremos **conjunto** o **multivariante**.

6.2. Distribución conjunta

6.2.1. Vector aleatorio

En el estudio de distribuciones continuas es conveniente el uso de notación vectorial $\boldsymbol{X} = (X_1, \dots X_n)$ donde X_i son variables aleatorias y $\boldsymbol{X}$ recibe el nombre de **vector aleatorio**. Es importante que, cuando usemos esta notación, no olvidemos que se trata de un vector n dimensional y que, por tanto, cualquier función que definamos sobre este (función de densidad o de distribución acumulada) será una función cuyo soporte será un subconjunto de $\mathbb{R}^n$.

Un vector aleatorio puede estar compuesto tanto íntegramente por variables discretas o continuas o de manera híbrida por ambos tipos. Esta composición será fundamental para la definición de las distintas funciones que caracterizan su distribución.

Extendiendo el concepto de soporte que hemos estudiado para una variable aleatoria, hablaremos de **soporte** de un vector aleatorio como el conjunto de todos los posibles valores de $\boldsymbol{X}$ y lo denotamos por $S_{\boldsymbol{X}}$.

6.2.2. Distribución conjunta

De forma breve, diremos que la *distribución conjunta* de un vector aleatorio $\boldsymbol{X}$ se define como la colección de todas las probabilidades de la forma $P(\boldsymbol{X} \in C), \quad \forall\, C \in S_{\boldsymbol{X}}$.

6.2.2.1. Vectores aleatorios discretos

Empecemos por el caso discreto. Imaginemos que tenemos un vector aleatorio con todas sus componentes v.a. discretas. La distribución de este vector quedará caracterizada, al igual que en el caso *univariante* por su función de probabilidad a la que en este caso añadiremos el *apellido* "conjunta".

Definición La **función de probabilidad conjunta** de un vector aleatorio discreto $\boldsymbol{X} = (X_1, \dots, X_n)$ se define como

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = p_{\boldsymbol{X}}(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n) = P(\boldsymbol{X} = \boldsymbol{x})$$

En virtud de los axiomas de probabilidad, toda función de probabilidad conjunta verificará:

$$0 \leq p_{\boldsymbol{X}}(\boldsymbol{x}) \leq 1$$

y

$$\sum_{S_{\boldsymbol{X}}} p_{\boldsymbol{X}}(\boldsymbol{x}) = 1.$$

Cuando decimos que la función de probabilidad conjunta caracteriza la distribución de un vector aleatorio es porque para cualquier subconjunto C del soporte de $\boldsymbol{X}$ podemos calcular su probabilidad utilizando el siguiente teorema:

Teorema Sea $\boldsymbol{X}$ un vector aleatorio discreto con función de probabilidad conjunta $p_{\boldsymbol{X}}(\boldsymbol{x})$, la probabilidad de cualquier conjunto $C \subset \mathbb{R}^n$ se puede calcular como:

$$P(\boldsymbol{X} \in C) = \sum_{\boldsymbol{x} \in C} p_{\boldsymbol{X}}(\boldsymbol{x})$$

De esta forma, la **función de distribución acumulada conjunta**

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = p_{\boldsymbol{X}}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

puede calcularse a partir de su función de probabilidad conjunta teniendo en cuenta, simplemente que el evento $(X_1 \leq x_1, \ldots, X_n \leq x_n)$ define un subconjunto $C \subset \mathbb{R}^n$.

A modo de ilustración, pensemos en un ensayo clínico donde tenemos m pacientes que pueden (o no) tener una recaída de una determinada enfermedad. Definimos X_i como la variable que vale 1 si el paciente i tiene una recaída y 0 en caso contrario.

Suponiendo que conocemos que la probabilidad de recaer es la misma para todos los pacientes (π) y que estos recaen o no de manera independiente, la probabilidad de un vector $(x_1, \ldots, x_m)$ (de ceros y unos) será:

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \pi^{x_1 + \cdots + x_m}(1-\pi)^{m - x_1 - \cdots - x_m}$$

para todo $x_i \in \{0, 1\}$ o 0 en otro caso.

Fijaros que, para llegar a esta función de probabilidad hemos usado el concepto de independencia que estudiábamos en el tema 1 así como la distribución de probabilidad de una distribución Bernoulli de parámetro π. Cabe señalar, además, que no se trata de una distribución binomial (no aparece el número combinatorio) porque aquí si que importa el orden ya que estamos calculando la probabilidad de que cada paciente tenga recaída o no y no el número de pacientes que sufren una recaída.

6.2.2.2. Vectores aleatorios continuos

Un vector aleatorio $\boldsymbol{X}$ decimos que es continuo cuando todas sus componentes lo son. En el caso de vectores aleatorios continuos, su distribución queda caracterizada por su función de densidad y, más concretamente, por su función de densidad conjunta.

Definición decimos que un vector aleatorio $\boldsymbol{X}$ tiene una distribución continua si existe una función f no negativa tal que, para todo $C \subset \mathbb{R}^n$

$$P(\boldsymbol{X} \in C) = \int \cdots \int_C f(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

cuando esta integral existe. A f se le denomina **función de densidad conjunta** y, en virtud de los axiomas de probabilidad estudiados en el tema 1, debe cumplir que $f(\boldsymbol{x}) \geq 0$ para todo $\boldsymbol{x} \in S_{\boldsymbol{X}}$ y

$$P(\boldsymbol{X} \in \mathbb{R}^n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) dx_1 \ldots dx_n = 1$$

La **función de distribución acumulada conjunta** se define como

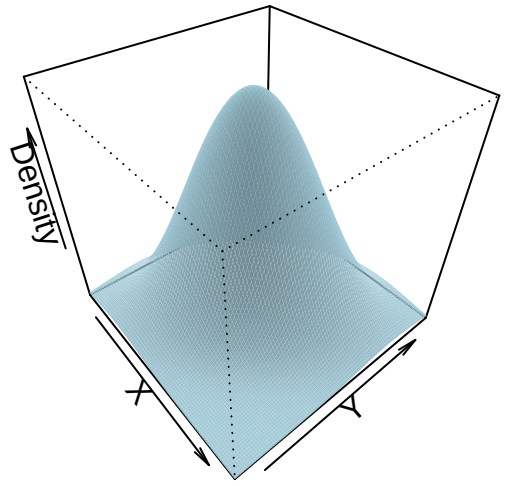
$$F(\boldsymbol{x}) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

y queda caracterizada, de nuevo, por la función de densidad conjunta.

El siguiente gráfico nos muestra un función de densidad bivariante ($n = 2$ en nuestro vector aleatorio) y, en concreto, a lo que se conoce como distribución **normal multivariante** que estudiaremos con detalle un poco más adelante. Este tipo de gráficos sólo pueden realizarse para visualizar la distribución de dos variables aleatorias.

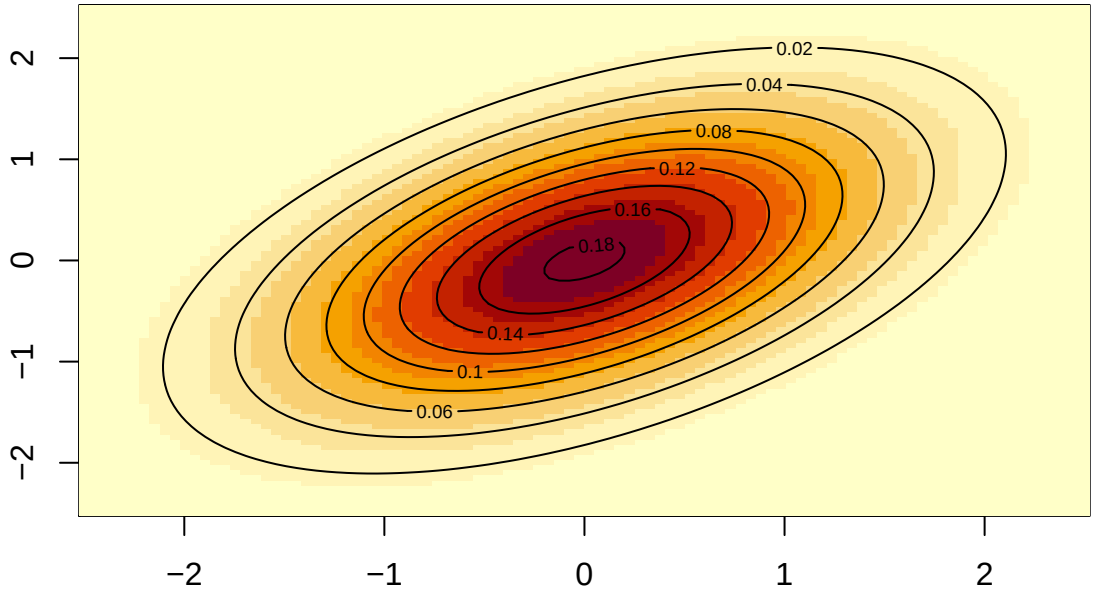
```
library("mvtnorm")
range = seq(-2.5,2.5,length.out = 100)
mean = c(0,0)
Sigma = matrix(c(1, .5, .5, 1), 2)

out = list(x=range,y=range, z=matrix(rep(0,100*100),100))

for (i in 1:length(range)){
    for (j in 1:length(range)){
        out$z[i,j] = dmvnorm(c(range[i],range[j]),mean=mean,sigma=Sigma)
```

```
    }
}
persp(out,theta = 50,phi = 40,col="lightblue", shade = .1, border = NA, xlab="X", zlab=
```



En este caso también se suele recurrir a gráficos de contorno

```
image(out); contour(out, add = T)
```



Veamos un ejemplo. Pensemos en la cola que se produce a la hora de pagar en un supermercado y, más concretamente, en las colas únicas donde, aunque tengamos diferentes cajas, los clientes se ordenan en una única fila. Supongamos que llegan n clientes y sea X_i el tiempo que se tarda en cobrar al cliente i.

En estas circunstancias estamos en disposición de utilizar una función conjunta para

$\boldsymbol{X} = (X_1, \ldots, X_n)$:

$$f(\boldsymbol{x}) = \begin{cases} \frac{c}{(2+\sum_{i=1}^{n} x_i)^{n+1}} & \text{para todo } x_i > 0 \\ 0 & \text{en otro caso} \end{cases}$$

Nos faltaría ahora encontrar el valor de c para que la función de distribución acumulada integre 1. Para ello, vamos a integrar sucesivamente $x_1, \ldots, x_n$ empezando por x_n. La primera integral nos da

$$\int_0^\infty \frac{c}{(2 + x_1 + \cdots + x_n)^{n+1}} dx_n = \frac{c/n}{(2 + x_1 + \cdots + x_{n-1})^n}.$$

Podemos observar que el resultado es igual a la función de densidad original pero con n reducido a $n - 1$. Si hacemos esta integración de forma iterativa llegamos a x_1 teniendo

$$\frac{c/n!}{(2 + x_1)^2}$$

e integrando para x_1 tenemos $c/(2n!)$ por lo que $c = 2n!$ si queremos que la función integre 1.

6.2.2.3. Vectores aleatorios mixtos

Es posible que un investigador/a se encuentre en la circunstancia en que su vector aleatorio de interés contenga tanto variables aleatorias continuas como discretas. En ese caso la función de distribución acumulada deberá obtenerse mediante una suma para aquellas variables de naturaleza discreta y mediante integración para la continuas.

Continuando con el ejemplo de la cola, el tiempo de espera de un cliente dado dependerá de la rapidez del proceso de cobro Y, la tasa a la que llegan los clientes Z y cuantos clientes van al supermercado W. En este ejemplo, Y y Z son variables continuas mientras que W es una variable discreta.

Una posible función de densidad conjunta en este caso puede ser:

$$f(\boldsymbol{x}) = \begin{cases} 6e^{-3z-10y}(8y)^w/w! & \text{para todo } z, y > 0 \text{ y } w = 0, 1, \ldots \\ 0 & \text{en otro caso} \end{cases}$$

pero, lo primero será comprobar que verdaderamente es una función de densidad. Es fácil ver que podemos separarla en dos funciones $h_1(z) = 6e^{-3z}$ y $h_2(y, w) = e^{-10y}(8y)^w/w!$

Podemos empezar integrando z y tendremos

$$h_2(y,w)\int_0^\infty 6e^{-3z}dz = 2h_2(y,w).$$

A continuación podemos sumar para w y teniendo en cuenta la expansión en serie de Taylor de e^x tenemos

$$\sum_{i=0}^{\infty} 2h_2(y,w) = 2e^{-10y}\sum_{i=0}^{\infty}\frac{(8y)^w}{w!} = 2e^{-10y}e^{8y} = 2e^{-2y}.$$

Ahora solo nos falta integrar sobre y que, claramente nos dará 1, tal y como esperábamos.

6.3. Distribución marginal y distribución condicional

Y que pasa cuando conocemos la distribución conjunta de un vector aleatorio pero nos interesa saber el comportamiento de una (o algunas) de las variables de forma individual.

En ese caso la probabilidad cuenta con dos herramientas, la probabilidad condicional y la probabilidad marginal. Como resumen podríamos decir que la probabilidad condicional consiste en obtener la distribución de un subconjunto de las variables para un valor concreto del resto mientras que la probabilidad marginal *integra* aquellas variables que no nos interesan.

De forma intuitiva veremos que nos referimos a los conceptos de probabilidad condicional y probabilidad total que vimos en el tema 1.

Para definir mejor estas herramientas vamos a empezar por la distribución marginal pasando después a la distribución condicional y retomando, por último, el Teorema de Bayes y el concepto de independencia.

6.3.1. Distribución marginal

Para entender el concepto de distribución marginal vamos a empezar con un ejemplo.

El siguiente cuadro nos muestra las probabilidades de estar muerto, presuntamente muerto, resucitado o vivo, en la serie Juego de Tronos, según el género.

	muerto	p.muerto	resucitado	vivo	Total
mujer	0.21	0	0	0.10	0.31
hombre	0.54	0.01	0.02	0.12	0.69
Total	0.75	0.01	0.02	0.22	1

Si quiero saber la proporción de mujeres en la serie bastará con sumar todas las probabilidades de esa fila: un 31 % de los personajes son mujeres. Si lo que quiero saber el la probabilidad de estar muerto, esto es un 75 %. se trata de lo que llamamos probabilidades marginales y, como hemos visto, se obtienen sumando las probabilidades para todos los valores de la variable que no nos interesa.

De forma general, La distribución marginal de una variable aleatoria puede obtenerse mediante la integración (suma) de su función de densidad (probabilidad) conjunta.

Definición: Dado un vector aleatorio $\boldsymbol{X}$ discreto de dimensión n con función de probabilidad $p_{\boldsymbol{X}}$ podemos calcular la función de probabilidad para X_i como:

$$p_{X_i}(x_i) = \underbrace{\sum\sum\cdots\sum}_{n-1} p_{\boldsymbol{X}}(x_1, \ldots, x_i, \ldots, x_n)$$

Equivalentemente:

Definición: Dado un vector aleatorio $\boldsymbol{X}$ continuo de dimensión n con función de densidad $f_{\boldsymbol{X}}$ podemos calcular la función de densidad para X_i como:

$$f_{X_i}(x_i) = \underbrace{\int\int\cdots\int}_{n-1} f_{\boldsymbol{X}}(x_1, \ldots, x_i, \ldots, x_n) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n$$

A partir de estas definiciones se puede demostrar que la función de distribución acumulada marginal se puede calcular como:

$$F_{X_i}(x_i) = Pr(X_1 < \infty, \ldots, X_i \leq x_i, \ldots, X_n < \infty) = \lim_{x_j \to \infty \ j \neq i} F_{\boldsymbol{X}}(x_1, \ldots, x_n)$$

Intuitivamente, situamos en su máximo todas las variables que no son de nuestro interés, integrando por tanto en todo el espacio muestral para esas variables y dejamos sólo la variable de interés como desconocida.

6.3.1.1. Independencia de variables aleatorias

Si pensamos en la definición de independencia que estudiamos en el tema 1, tiene sentido pensar que dos o más variables aleatorias serán independientes si

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P(X_1 \in A_1) P(X_2 \in A_2) \cdots P(X_n \in A_n)$$

Esta condición se satisface si y solo sí la función de densidad (probabilidad) conjunta puede expresarse como el producto de de las distribuciones marginales para cada una de las variables. Es decir:

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n)$$

6.3.1.2. Muestras aleatorias

El concepto de independencia de variables aleatorias nos permite definir lo que entendemos por una muestra aleatoria.

Supongamos que tenemos una cantidad medida en la recta real con función de densidad(probabilidad) f. Un conjunto de n variables aleatorias $X_1, \ldots, X_n$ forman una **muestra aleatoria** de f de tamaño n si cada una de estas variables tienen función de densidad (probabilidad) f.

En ese caso, su función de densidad (probabilidad) conjunta puede expresarse como

$$f_{\boldsymbol{X}}(x_1, \ldots, x_n) = f(x_1)\cdots f(x_n)$$

y decimos que las variables son independiente e idénticamente distribuidas, termino que ya hemos utilizado y que denotamos por i.i.d.

6.3.2. Distribución condicional

Como ya explicamos en el tema 1, a veces conocemos lo que ha pasado en un determinado experimento, es decir, sabemos el valor de una o varias variables y nos interesa conocer el valor de otras.

A este tipo de probabilidad se le conoce como probabilidad condicional y se aplican las mismas reglas, definiciones y propiedades que ya vimos cuando hablábamos de experimentos y sucesos.

Volvamos al ejemplo de Juego de Tronos, imaginemos que sabemos que un personaje está muerto y queremos adivinar si era hombre o mujer. Las probabilidad de ser hombre y estar muerto es 0.54 y la de ser mujer y estar muerta es 0.21. Estas suman 0.75 y parece lógico que, si nuestra variable de interés sólo puede tomar los valores hombre/mujer, la probabilidad de ambas debería sumar 1. Esto lo conseguimos dividiendo por la suma:

$$P(X = \text{mujer} \mid Y = \text{muerto}) = \frac{0{,}21}{0{,}75} = 0{,}28 \quad P(X = \text{hombre} \mid Y = \text{muerto}) = \frac{0{,}54}{0{,}75} = 0{,}72$$

En general:

Definición Dado un vector aleatorio $\boldsymbol{X}$ de dimensión n, la **función de probabilidad (densidad) condicionada** de la variable X_i se obtiene como: $f_{X_i}(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = \frac{f_{\boldsymbol{X}}(x_1,\ldots,x_n)}{f(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)}$

Notad que la función que aparece en el denominador es la marginal para todas las variables eliminado/integrando X_i. Esto es, la suma para todos los posibles valores de X_i, si es discreta, o la integral sobre el espacio de todos los posibles valores de X_i si es continua.

6.3.2.1. Teorema de Bayes para variables aleatorias

Fijaos que, utilizando la distribución condicional de una variable, podemos llegar a la versión del teorema de Bayes que ya vimos en el tema anterior. Simplificando al caso de dos variables aleatorias Esto es:

$$f(x_1 \mid x_2) = \frac{g(x_2 \mid x_1)h(x_1)}{m(x_2)}$$

Notad que cada distribución tiene una nomenclatura diferente f, g, h, m dado que cada una de ellas representa una densidad distinta. f y g son densidades condicionales sobre $X_1 \mid X_2$ y $X_2 \mid X_1$ respectivamente; h es lo que se suele conocer como distribución *a priori* y se trata de una función de densidad marginal sobre X_1. Por último, m es la marginal de X_2 que se obtiene integrando X_1 en la función de densidad conjunta para X_1 y X_2 que es $g(x_2 \mid x_1)h(x_1)$

$$m(x_2) = \int_{\Theta_{X_1}} g(x_2 \mid x_1)h(x_1)d(x_1)$$

6.4. Relación entre variables

Al igual que cuando hablamos de una variable aleatoria, conocer los momentos de un vector aleatorio es útil para resumir su distribución. Sin embargo, la media, la mediana o la varianza resumen el comportamiento individual de cada variable sin decirnos nada de la relación existente entre ellas.

De hecho, la esperanza de un vector aleatorio se define como el vector con las esperanzas marginales de cada una de las variables que lo componen. Esto es:

$$E(\boldsymbol{X}) = (E(X_1), \ldots, E(X_n)) = (\mu_1, \ldots, \mu_n)$$

Y pasa lo mismo con la mediana o con la varianza.

Necesitamos, entonces, una medida de la capacidad de dos variables de *variar* juntas. Con este fin pasamos a definir los términos covarianza, correlación y esperanza condicional.

6.4.1. Covarianza

Definición Sean X_1 y X_2 dos variables aleatorias con una distribución conjunta determinada dónde $E(X_i) = \mu_i$ y $Var(X_i) = \sigma_i$ para $i = 1, 2$. La **covarianza** de X_1 y X_2 se denota por $Cov(X_1, X_2)$ y se define como:

$$Cov(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)]$$

Se puede demostrar que si σ_1 y σ_2 son finitas, entonces la covarianza existe y es un valor real que puede ser negativo positivo o cero.

Un valor positivo de la covarianza indicará que cuanto mayor sea una de las variables mayor será la otra. Por el contrario, un valor negativo indica una relación inversa, cuando mayor sea una menor será la otra.

Teorema Para cualesquiera dos variables aleatorias X_1 y X_2 tales que sus varianzas existen y son finitas:

$$Cov(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2)$$

La covarianza permite también calcular la varianza de la suma de dos variables tal y como muestra el siguiente teorema:

Teorema Si X_1 y X_2 son variable aleatorias con varianza finita:

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2) + 2Cov(X_1, X_2)$$

Este teorema puede extenderse a cualquier número de variables como:

Teorema Sean $X_1, \ldots, X_n$ un conjunto de variables aleatorias con varianza finita:

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum\sum_{i<j} Cov(X_i, X_j)$$

6.4.2. Correlación

A partir de la covarianza se puede definir el coeficiente de **correlación** como

$$\rho(X_1, X_2) = \frac{Cov(X_1, X_2)}{\sigma_1\sigma_2}$$

La correlación es un valor que se encuentra en el intervalo $[-1, 1]$ y tiene el mismo signo que la covarianza. Decimos que dos variables están positívamente correladas cuando

$\rho(X_1, X_2) > 0$, que están negativamente correladas cuando $\rho(X_1, X_2) < 0$ y no correladas o incorreladas cuando $\rho(X_1, X_2) = 0$

Se puede demostrar que si dos variables, con varianzas finitas, son independientes la correlación será 0. Sin embargo, el caso contrario no es cierto en general, como podemos ver en el siguiente ejemplo:

Supongamos que una variable aleatoria X_1 puede tomar sólo 3 valores -1, 0 y 1, con igual probabilidad. Sea X_2 la variable aleatoria que se define como $X_2 = X_1^2$. En este caso, ambas variables están completamente correladas (una está definida a partir de la otra) pero, dado que la media de ambas variables era 0:

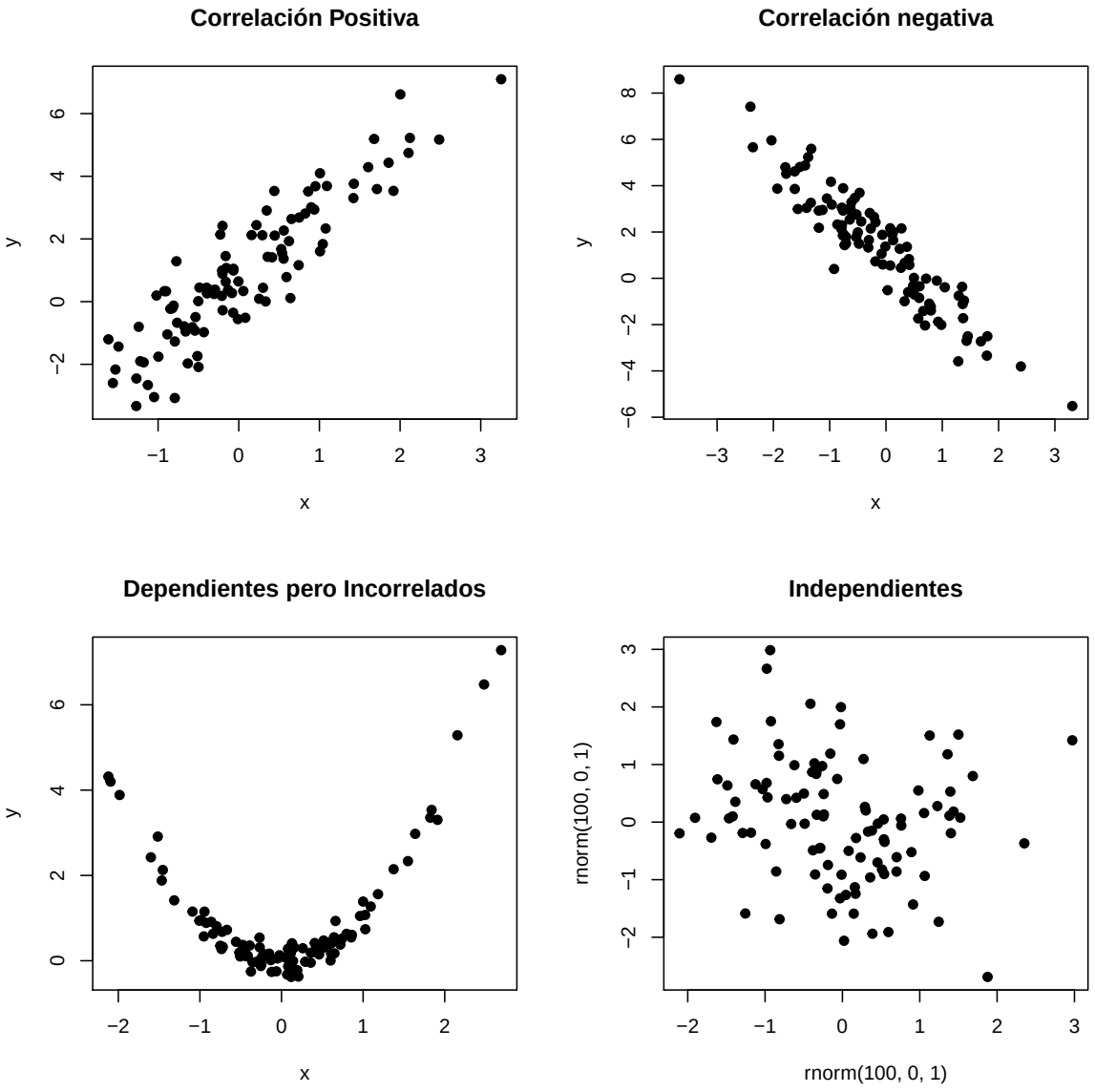
$$Cov(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] = E[X_1^3] = E[X_1] = 0$$

Se demuestra, por tanto, que podemos tener variables dependientes pero incorreladas. De hecho, el coeficiente de correlación está especialmente indicado para medir la correlación lineal de las variables, como muestra el siguiente teorema, pero no tanto para otro tipo de relaciones de dependencia.

Teorema Sea X_1 es una variable aleatoria con varianza finita y sea X_2 otra variable aleatoria definida como una función lineal de X_1: $X_2 = aX_1 + b$ para determinadas constantes $a \neq 0$ y b. Si $a > 0$, $\rho(X_1, X_2) = 1$ y $\rho(X_1, X_2) = -1$ si $a < 0$.

Para entender el concepto de dependencia y correlación podemos ver las siguientes gráficas

```
par(mfrow=c(2,2))
set.seed(22)
x<- rnorm(100,0,1)
y<- rnorm(100,2*x+1)
plot(x,y,pch=19,xlab="x",ylab="y", main="Correlación Positiva")
x<- rnorm(100,0,1)
y<- rnorm(100,-2*x+1,1)
plot(x,y,pch=19,xlab="x",ylab="y", main="Correlación negativa")
x<- rnorm(100,0,1)
y<- rnorm(100,x^2,0.2)
plot(x,y,pch=19,xlab="x",ylab="y", main="Dependientes pero Incorrelados")
plot(rnorm(100,0,1),rnorm(100,0,1),pch=19, main="Independientes")
```



6.4.3. Esperanza condicional

Otra de las medidas que ayuda a entender la relación entre dos variables aleatorias es la esperanza condicional.

Definición dadas dos variables aleatorias X_1 y X_2 con función de densidad conjunta $f(x_1, x_2)$, la esperanza condicional de X_2 dada X_1 se denota $E(X_2 \mid X_1)$ y se define como una función de la variable aleatoria X_1 cuyo valor en $X_1 = x_1$ es igual a

$$E(X_2 \mid x_1) = \int_{-\infty}^{\infty} x_2 g(x_2 \mid x_1) dx_2$$

En otras palabras, $E(X_2 \mid x_1)$ es la media de la distribución condicional de X_2 cuando $X_1 = x_1$.

Evidentemente, si X_2 es una variable discreta, la integral será sustituida por una suma.

Notese que $E(X_2 \mid X_1)$ es una variable aleatoria y, por tanto, se puede calcular tanto su esperanza como su varianza. En concreto:

Teorema Para cualesquiera dos variables aleatorias X_1 y X_2

$$E[E(X_2 \mid X_1)] = E(X_2)$$

Por otra parte, la varianza de esta variable aleatoria $Var[E(X_2 \mid X_1)]$ tiene una interpretación muy interesante como la medida de cuanto más conocemos de X_2 tras conocer X_1.

Más concretamente, imaginemos que queremos predecir X_2. Si no tenemos ninguna información al respecto, la mejor predicción resulta ser $E(X_2)$. Pero, ¿Qué pasa si ya conocemos el valor de $X_1 = x_1$? ¿Nos da éste información a cerca de X_2?

En tal caso, la mejor predicción para X_2 es $E(X_2 \mid x_1)$. Y cuando decimos mejor, a lo que nos referimos es a que es el valor que minimiza el error cometido que, en este caso se denota por $Var(X_2 \mid x_1)$ y que es la varianza de la función condicional de X_2 dado $X_1 = x_1$.

Nuestro interés entonces es saber si el error cometido al predecir X_2 sin saber nada de X_1, esto es $Var(X_2)$ es mucho mayor que el error cometido en promedio al predecir X_2 conociendo X_1, esto es: $E[Var(X_2 \mid X_1)]$

En concreto, se demuestra que esta mejora viene representada por la varianza de la variable aleatoria $E(X_2 \mid X_1)$

$$Var[E(X_2 \mid X_1)] = Var(X_2) - E[Var(X_2 \mid X_1)]$$

6.5. Ejercicios

1. Supongamos que un dispositivo eléctrico que tiene tres bombillas en la primera fila y 4 en la segunda. Sea X el número de bombillas de la primera fila que se habrán apagado en un tiempo t dado y sea Y el número de bombillas de la segunda fila que se apagan en el mismo tiempo. La probabilidad conjunta de X e Y se da en la siguiente tabla:

	0	1	2	3	4
0	0.08	0.07	0.06	0.01	0.01
1	0.06	0.10	0.12	0.05	0.02
2	0.05	0.06	0.09	0.04	0.03
3	0.02	0.03	0.03	0.03	0.04

determina las siguientes probabilidades:

(a) $P(X = 2)$ (b) $P(Y \geq 2)$ (c) $P(X \leq 2 \text{ y } Y \leq 2)$

(d) $P(X = Y)$ (e) $P(X > Y)$

2. Supongamos que X e Y son variables continuas con función de distribución conjunta:

$$f(x, y) = \begin{cases} cy^2 & \text{para } 0 \leq x \leq 2 \text{ y } 0 \leq y \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$

Determina el valor de la constante para que f sea una función de densidad conjunta valida.

3. Supongamos que un punto (X, Y) es elegido al azar del la región S de un plano que contiene todos los puntos que cumplen $x \geq 0$, $y \geq 0$ y $4y + x \leq 4$.

(a) Determina la función de densidad conjunta de X e Y

(b) Supongamos que S_0 es un subconjunto de la región S con un área α. Determina $P((x, y) \in S_0)$

4. Supongamos que X e Y son variables continuas con función de distribución conjunta:

$$f(x, y) = \begin{cases} c(x^2 + y) & \text{para } 0 \leq y \leq 1 - x^2 \text{ y } 0 \leq x \leq 1 \\ 0 & \text{en otro caso} \end{cases}.$$

Determina el valor de la constante para que f sea una función de densidad conjunta valida.

5. Supongamos que tenemos tres variables aleatorias X_1, X_2 y X_3. Su función de densidad conjunta es

$$f(x_1, x_2, x_3) = \begin{cases} c(x_1 + 2x_2 + 3x_3) & \text{para } 0 \leq x_i \leq 1 \;\; i = 1, 2, 3 \\ 0 & \text{en otro caso} \end{cases}$$

(a) Calcula el valor de la constante c

(b) Determina la distribución marginal para (X_1, X_2)

(c) Calcula

$$P\left(X_3 \leq \frac{1}{2} \mid X_1 = \frac{1}{4}, X_2 = \frac{3}{4}\right)$$

6. Supongamos que tenemos tres variables aleatorias X_1, X_2 y X_3. Su función de densidad conjunta es

$$f(x_1, x_2, x_3) = \begin{cases} ce^{-(x_1+2x_2+3x_3)} & \text{para } x_i \geq 0 \;\; i = 1, 2, 3 \\ 0 & \text{en otro caso} \end{cases}$$

(a) Calcula el valor de la constante c

(b) Determina la distribución marginal para (X_1, X_3)

(c) Calcula $P(X_2 \leq 1 \mid X_1 = 2, X_3 = 1)$

7. Supongamos que un sistema electrónico contiene n componentes que funcionan de manera independiente unas de otras. La probabilidad de que cada componente funcione correctamente es π_i, $i = 1, \ldots, n$. Se dice que las componentes están conectadas en *serie* si una condición necesaria y suficiente para que el sistema funcione es que todas las componentes funcionen. Del mismo modo, decimos que están conectadas en *paralelo* si una condición necesaria y suficiente para que el sistema funcione es que, al menos una de las componentes funcione. A la probabilidad de que el sistema funcione se le conoce como *fiabilidad*. Determina la fiabilidad del sistema:

(a) Asumiendo que las componentes están conectadas en serie.

(b) Asumiendo que están conectadas en paralelo.

8. Sean X_1 y X_2 dos variables aleatorias cuya función de densidad conjunta es:

$$f(x_1, x_2) = \begin{cases} \frac{1}{3}(x_1 + x_2) & \text{para } 0 \leq x_1 \leq 1 \quad \text{y} \quad 0 \leq x_2 \leq 2 \\ 0 & \text{en otro caso} \end{cases}$$

Determina el valor de $Var(2X_1 - 3X_2 + 8)$

9. Supongamos que la nota de la asignatura de probabilidad la medimos como un valor entre 0 y 1. La nota de una persona elegida al azar es una variable aleatoria X. Del mismo modo, la nota en optimización se mide también en el intervalo (0,1) y la nota para esta misma persona es la variable aleatoria Y. Sabiendo que la función de

densidad conjunta de ambas notas es:

$$f(x,y) = \begin{cases} \frac{2}{5}(2x+3y) & \text{para } 0 \leq x \leq 1 \quad \text{y} \quad 0 \leq y \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$

Si seleccionamos a una persona al azar, ¿Cual será el valor predicho para su nota en optimización? y si conocemos que su nota en probabilidad es $x = 0{,}7$.

7. Distribuciones multivariantes conocidas

7.1. Algunas distribuciones multivariantes conocidas

Veamos a continuación las generalizaciones multivariantes de algunas distribuciones que ya estudiamos en el tema anterior. La distribución multinomial como generalización de una distribución binomial y la distribución normal multivariante.

7.1.1. Distribución Multinomial

La distribución binomial media el número de éxitos en un conjunto de N pruebas cuando el resultado del experimento sólo podía tomar dos valores. Pero qué pasa si mi variable puede tomar más de dos valores (recordemos el caso de Juego de Tronos, donde una persona puede estar muerta, presuntamente muerta, resucitada o viva). En ese caso nos interesa saber cuantas observaciones de cada categoría tenemos en un total de N pruebas. De esta manera definimos un vector aleatorio multinomial como: **Definición** un vector aleatorio $\boldsymbol{X} = (X_1, \ldots, X_k)$ con $\sum_{i=1}^{k} X_i = N$ sigue una distribución multinomial $\boldsymbol{X} \sim Mult_k(N, \boldsymbol{\pi})$ donde $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ con $\sum_{i=1}^{k} \pi_i = 1$ y su función de probabilidad conjunta es:

$$p(X_1 = n_1, \ldots, X_k = n_k) = \frac{N!}{n_1! n_2! \cdots n_k!} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_k^{n_k}$$

El siguiente teorema nos muestra que cada una de las variables que conforman el vector multinomial se comporta, marginalmente como una binomial.

Teorema Dado un vector aleatorio con distribución multinomial de parámetros N y $\boldsymbol{\pi}$, la función marginal para cada variable aleatoria X_i es $X_i \sim Bi(N, \pi_i)$.

Del mismo modo, si decidimos juntar varias categorías en una única variable, por ejemplo $X_i + X_j$, el vector resultante sigue siendo multinomial con el parámetro de probabilidad correspondiente transformado a $\pi_i + \pi_j$ y la distribución marginal en ese caso, es $X_i + X_j \sim Bi(N, \pi_i + \pi_j)$.

Por otra parte, el siguiente teorema nos muestra cual será la función de probabilidad condicionada:

Teorema sea $\boldsymbol{X} \sim Mult_k(N, \boldsymbol{\pi})$ tenemos que

$$X_2, \ldots, X_k \mid X_1 = n_1 \sim Multk - 1(N - n_1, (\pi_2', \ldots, \pi_k'))$$

donde $\pi_j' = \frac{\pi_j}{\pi_2 + \cdots + \pi_k}$

Por supuesto, $X_1, \ldots, X_k$ son variables dependientes (pensemos que deben sumar N) por lo que tiene sentido estudiar su covarianza

Teorema sea $(X_1, \ldots, X_k) \sim Mult_k(N, \boldsymbol{\pi})$, para todo $i \neq j$,

$$Cov(X_i, X_j) = -N\pi_i\pi_j$$

Vemos que la covarianza es negativa algo que, de hecho, tiene mucho sentido ya que, cuando más valores *caigan* en una categoría, menos caerán en otra.

7.1.2. Multinomial en R

En R, la densidad conjunta de una función multinomial con parámetros puede estudiarse mediante las ordenes:

```
x <- c(2,0,3)
N <- 5
p <- c(1/3,1/3,1/3)
dmultinom(x,N,p)
```

```
## [1] 0.04115226
```

7.1.3. Distribución Normal multivariante

La definición formal de una distribución normal multivariante dice que un vector $(X_1, \ldots, X_n)$ tiene una distribución normal multivariante si cualquier combinación lineal $(t_1X_1 + \cdots + t_kX_k)$ se distribuye siguiendo una normal univariante.

Partiendo de esta definición, podemos decir que una distribución normal multivariante queda completamente determinada si conocemos:

- El vector de medias para cada $(\mu_1, \ldots, \mu_k)$ donde $\mu_i = E(X_i)$. (Nótese que por la propiedad anterior $X_i \sim N(\mu_i, \sigma_i^2)$)
- La matriz de varianzas-covarianzas Σ que se define como la matriz cuya entrada i, j es el valor $Cov(X_i, Y_j)$. Los valores de la diagonal serán simplemente la varianza marginal de cada una de las componentes, σ_i^2.

En concreto, diremos que un vector aleatorio sigue una distribución normal multivariante

$\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \Sigma)$ cuando su función de densidad conjunta es:

$$f(X_1, \ldots, X_k) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{n/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)$$

donde $|\Sigma|$ denota el determinante de la matriz de varianzas covarianzas.

Las funciones de densidad marginales para cada una de las variables aleatorias X_i es, como ya hemos visto más arriba $X_i \sim N(\mu_i, \sigma_i^2)$ mientras que la distribución de un subgrupo de variables $X_1, \ldots, X_q$ condicionadas a $X_{q+1}, \ldots, X_k$ es también normal con media

$$E(X_1, \ldots, X_q \mid x_{q+1}, \ldots, x_k) = (\mu_1, \ldots, \mu_q) + \Sigma_{12}\Sigma_{22}^{-1}\left((x_{q+1}, \ldots, x_k) - (\mu_{q+1}, \ldots, \mu_k)\right)$$

Y matrix de varianzas-covarianzas

$$\overline{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T.$$

donde Σ_{12} es el bloque de la matriz Σ correspondiente a las covarianzas de las variables del bloque 1 con las del bloque 2 mientras que y Σ_{11} y Σ_{22} contienen las varianzas y covarianzas del bloque 1 y 2 respectivamente, esto es:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

7.1.3.1. Normal Multivariante en R

Para poder usar la normal multivariante en R debemos tener instalada la librería *mvtnorm* y usar las ordenes:

```
library(mvtnorm)
medias <- c(0,0)
rho <- 0.7
covmatrix <- matrix(c(1,rho,rho,1), nrow = 2, ncol = 2)
x <- c(2,1)
dmvnorm(x, mean = medias, sigma = covmatrix)
```

```
## [1] 0.02578229
```

```
pmvnorm(x, mean = medias, sigma = covmatrix)[[1]]
```

```
## [1] 0.0184354
```

8. Simulación y Métodos Monte Carlo

8.1. Introducción

Hasta ahora, todo lo que hemos visto es la formalización matemática de la incertidumbre existente al hablar de una o varias variables aleatorias.

Sin embargo, la abstracción detrás de todos estos conceptos es, en ocasiones, difícil de seguir y es necesario "bajar a la tierra" todos esos conceptos. En esta tarea, resultan muy útiles las técnicas de simulación.

Por ejemplo, puede ser difícil hacer entender a una persona que, en el programa de Monty-Hall es más probable ganar si se cambia de puerta. Sin embargo, podemos *simular* el proceso y conseguir convencerla viendo la proporción de resultados favorables. Al simular, lo que estaremos haciendo es convertir en datos los resultados teóricos ya obtenidos.

Pero las técnicas de simulación no son sólo útiles para la concreción (lo contrario de abstracción). Estás técnicas también pueden usarse para la aproximación de valores que no se pueden obtener de forma exacta como la media, la varianza de una variable aleatoria. En estos casos bastará con obtener muestras aleatorias independientes de la variable $X_1, \ldots, X_n$ y calcular su media y su varianza muestral:

$$E(X) \approx \frac{1}{N} \sum_{i=1}^{n} X_i = \bar{X}$$

$$Var(X) \approx \frac{1}{N-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

La ley de los grandes números nos asegura que estas aproximaciones serán mejores a medida que aumente N. De esta manera, Si los valores de X_i se obtienen por ordenador, obtener una buena aproximación será cuestión de dejarlo trabajar el tiempo suficiente.

Aplicando ideas similares podremos también aproximar integrales difícilmente tratables de otra forma.

Pero, ¿Qué significa simular? Y, lo más importante, ¿Cómo simulamos?

Literalmente, simular significa imitar un proceso real, en nuestro contexto, obtener realizaciones de una variable aleatoria de la que sólo se conoce (total o parcialmente) su función de densidad. Para simular partiremos siempre de algún generador de números (pseudo)aleatorios pero necesitaremos técnicas que nos permitan transformar esos números aleatorios en valores de la variable que queremos estudiar. Sobretodo, cuando las variables de las que queramos simular nos son independientes.

Este tema se centra en el estudio de dichas técnicas y particularmente de dos que se engloban en lo que se conoce como métodos MCMC (Markov Chain Monte Carlo): Metrópolis Hastings y Gibbs Sampling. Para llegar a los métodos MCMC empezaremos por estudiar con un poco de detalle a que nos referimos cuando hablamos de métodos Monte Carlo (en honor al casino) y de Cadenas de Markov (Markov Chain). Aunque, antes, vamos a describir una propiedad de la distribución uniforme que nos será muy útil a la hora de simular.

8.2. Transformada integral de probabilidad.

La distribución uniforme tiene una gran ventaja que la hace muy importante. A partir de una variable aleatoria uniforme en el intervalo (0, 1) podemos simular cualquier variable aleatoria continua y viceversa. De manera formal:

Teorema: 1. Dada una variable aleatoria $U \sim Unif(0,1)$, $X = F^{-1}(U)$ es una variable aleatoria continua con función de distribución acumulada F 2. Dada X una variable aleatoria continua con función de distribución F_X, $U = F_X(X)$ es una variable aleatoria continua con distribución $Unif(0,1)$.

La primera parte de este teorema nos dice que si tenemos una variable aleatoria uniforme y la transformamos usando la inversa de una función de distribución acumulada, el resultado es una variable aleatoria con dicha función de distribución. Gracias a esta propiedad, si queremos simular de una variable aleatoria con función de distribución acumulada F, nos bastará con conocer su inversa.

La segunda parte afirma que, si tenemos una variable aleatoria cualquiera, X, y consideramos su función de distribución (que toma valores en [0,1])) se comporta (cuando X es desconocida) como una variable aleatoria continua con distribución $Unif(0,1)$.

NOTA 1: Aunque suene un poco redundante, tomar $F_X(X)$ como una variable aleatoria es bastante natural. Se trata simplemente de pensar en F_X como una función (e.g $1 - e^{-x}$) y aplicarla sobre la variable aleatoria X $(1 - e^{-X})$. Recordemos que ya hemos visto en otras ocasiones como trabajar con funciones de una variable aleatoria.

Veamos un ejemplo. Decimos que una variable aleatoria X tiene una distribución logística si su función de distribución acumulada es:

$$F_X(x) = \frac{e^x}{1 + e^x}$$

Supongamos que tenemos valores de una distribución uniforme:

```
U <- runif(1000)
```

Para poder simular de la distribución logística bastará con obtener su inversa:

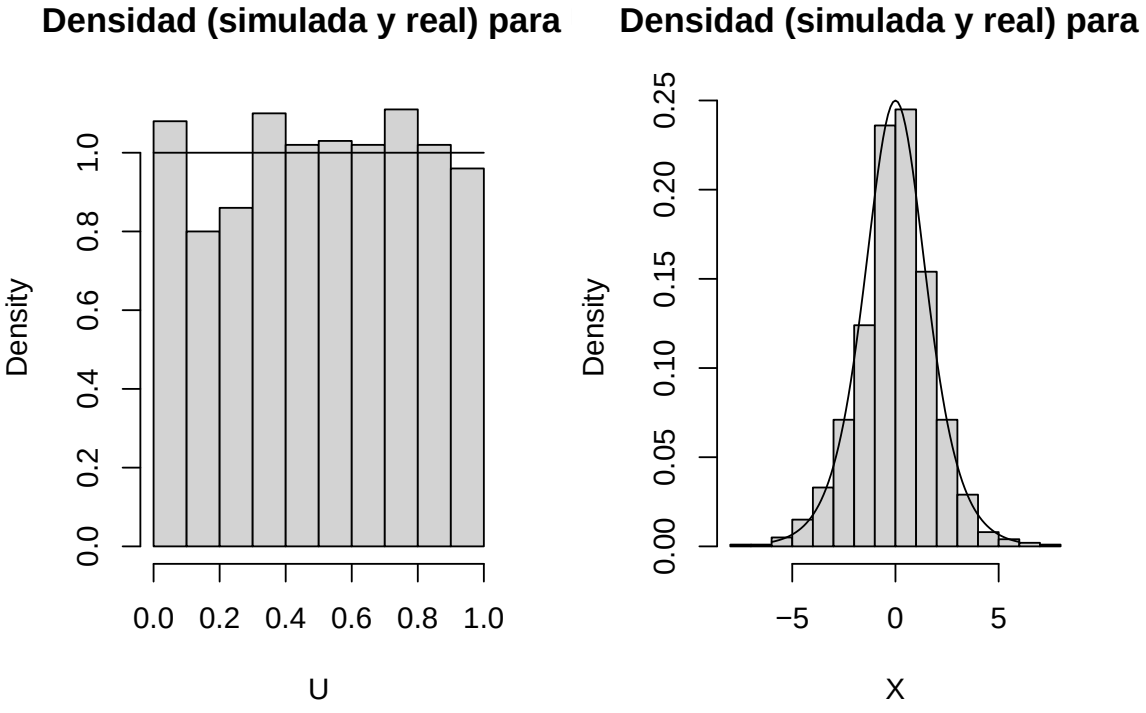
$$F^{-1}(u) = \log\left(\frac{u}{1-u}\right)$$

Si aplicamos esta función en los valores de la variable aleatoria U, $F^{-1}(U)$ que ya hemos simulado, tenemos:

```
X <- log(U/(1-U))
```

Las siguientes gráficas visualizan este proceso:

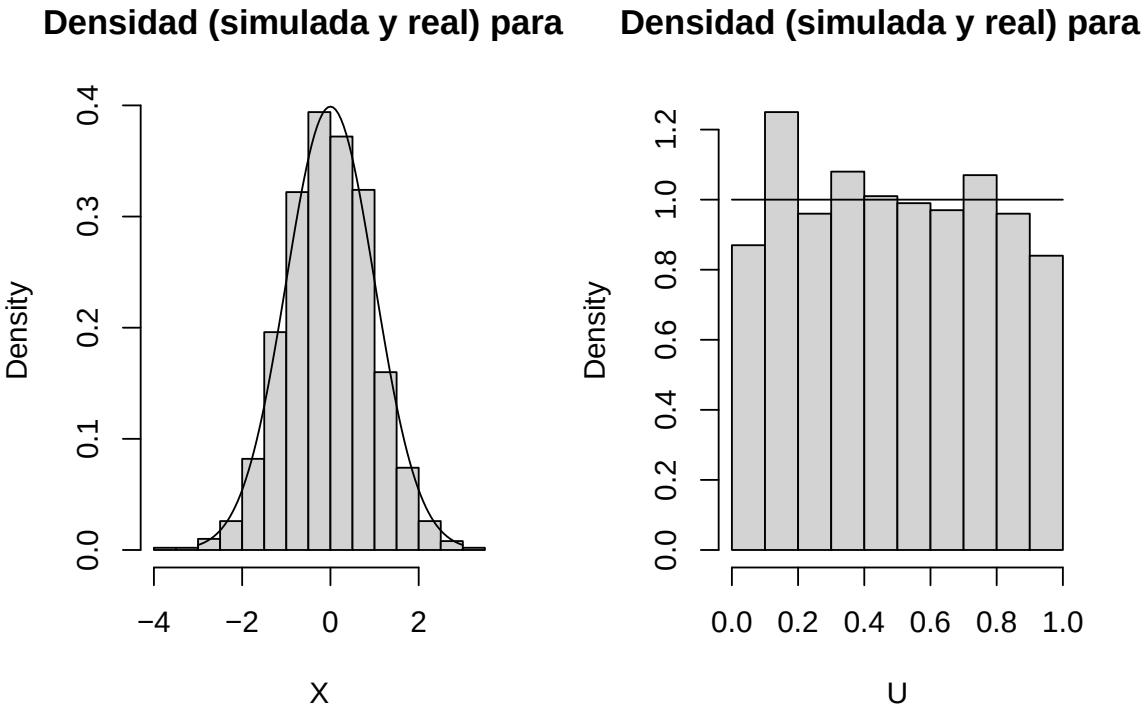
```
par(mfrow=c(1,2))
hist(U, main= "Densidad (simulada y real) para U",freq = FALSE)
curve(dunif(x),from=0, to=1, add=TRUE)
hist(X, main= "Densidad (simulada y real) para X",freq = FALSE)
curve(dlogis(x),from=-6, to=6, add=TRUE)
```



Un ejemplo de la segunda parte del teorema se puede obtener de forma sencilla simulando de una distribución normal y obteniendo valores de U como la función de distribución acumulada en cada uno de los valores de X:

```
X <- rnorm(1000)
U <- pnorm(X)
```

```
par(mfrow= c(1,2))
hist(X, main= "Densidad (simulada y real) para X",freq = FALSE)
curve(dnorm(x),from=-3, to =3, add=TRUE)
hist(U, main= "Densidad (simulada y real) para U",freq = FALSE)
curve(dunif(x),from=0, to=1, add=TRUE)
```



NOTA 2: El teorema de la trasformada integral de probabilidad funciona, en particular, para variables continuas. Cuando se trata de variables discretas, la segunda parte del teorema nunca será cierta. Sin embargo, aunque F es una función a trozos y F^{-1} no existe, la primera parte se podrá aplicar utilizando la función de probabilidad en lugar de la función de distribución acumulada. En concreto, para obtener una variable aleatoria discreta con función de probabilidad $p(X = j) = p_j$ para $j = 0, \ldots, k$ a partir de una variable $U \sim Unif(0, 1)$ bastará con dividir el intervalo (0,1) en k subintervalos donde, el j-ésimo intervalo tendrá una longitud p_j. Así, si el valor de U cae en el intervalo j-ésimo asignaremos a X el valor j.

NOTA 3: Lo más importante de esta propiedad es que, cualquier función f que cumpla las características para ser una función de densidad o probabilidad, lo será. Es decir, existirá

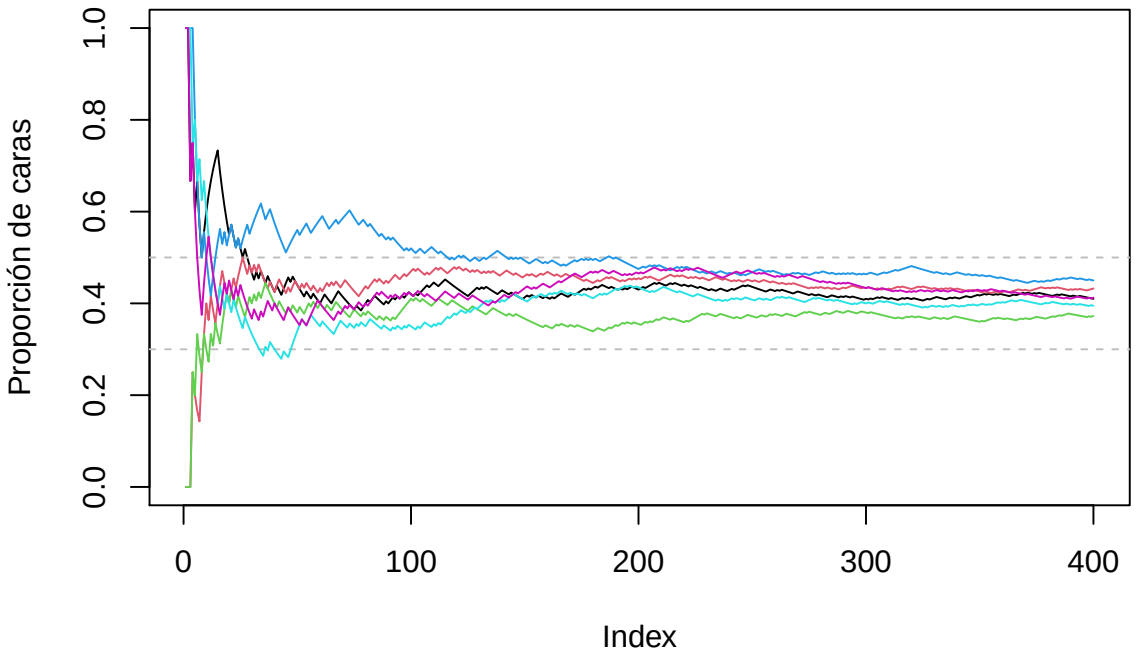
una variable aleatoria X cuya función de densidad o probabilidad sea f.

8.3. Métodos Monte Carlo y la ley de los grandes números

Cuando hablamos de Métodos Monte Carlo simplemente nos estamos refiriendo al hecho de usar números aleatorios para aproximar alguna cantidad desconocida. Estos números aleatorios pueden haberse obtenido mediante observación (de una muestra obtenida al azar) o mediante la utilización de números (pseudo)aleatorios.

Ya hemos visto que R proporciona este tipo de números utilizando las funciones *runif*, *rnorm*, *rbinom* o *rgamma*, entre otras, que simulan aleatoriamente de una variable con la distribución correspondiente.

Ejemplo Imaginemos que queremos estudiar la probabilidad de que una moneda (que sospechamos, está trucada) salga cara. Se trata de una probabilidad desconocida (no sabemos si realmente esta trucada ni en que sentido). Empezamos tirando la moneda 1, 2, 3... hasta N veces y con cada nueva tirada calculamos la proporción (acumulada) de veces que nos ha salido cara. Esto es, si la primera nos sale cara tendremos una proporción de 1, si la segunda vez nos sale cruz, la proporción acumulada de caras será 1/2, si añadimos un tercer valor y este vuelve a ser cara el nuevo valor para la proporción de caras será de 2/3 y así sucesivamente. Estos valores calculados a medida que realizamos un mayor número de tiradas convergerán al verdadero valor de la probabilidad de cara. El siguiente gráfico muestra la evolución de esta proporción en 6 experimentos realizados en las mismas condiciones ¿Cuál es la probabilidad de cara en este caso?



8.3.1. Integración Monte Carlo

Una de las aplicaciones más útiles de la metodología Monte Carlo es el cálculo de integrales complejas. Por ejemplo, supongamos que os interesa que calcular el área por debajo de una función $f(x)$ que no sabéis integrar. Podría parecer que se trata de un problema completamente determinista y que los números aleatorios no tienen nada que ver aquí, sin embargo, las técnicas Monte Carlo generan una aleatoriedad *fictícia* y se sirve de esta para resolver la integral en cuestión.

Supongamos que f es una función positiva y acotada $0 \leq f(x) \leq c$ tal que la integral entre a y b existe y es finita. Sea A el rectángulo $[a, b] \times [0, c]$ con área $(b - a)c$ y sea B la región de dicho cuadrado que se encuentra entre el eje de x y la curva $y = f(x)$. La integral deseada: $\int_a^b f(x)dx$ es el área de la región B. Para calcular este área bastará con simular dentro del cuadrado utilizando una distribución uniforme y, después, determinar la proporción de valores por debajo de la curva. Al multiplicar esta proporción por el área original del rectángulo, obtendremos el valor deseado.

De forma esquemática, el procedimiento es el siguiente:

1. Simulamos un punto (x, y) dentro del cuadrado utilizando una distribución uniforme.
2. Para cada punto obtenemos $I(y_i \leq f(x_i))$ que valdrá 1 si se cumple la condición y 0 si no.

3. Calcularemos el área de B como:

$$\int_a^b f(x)dx = (b-a)c\frac{1}{N}\sum_{i=1}^{N} I(y_i < f(x_i))$$

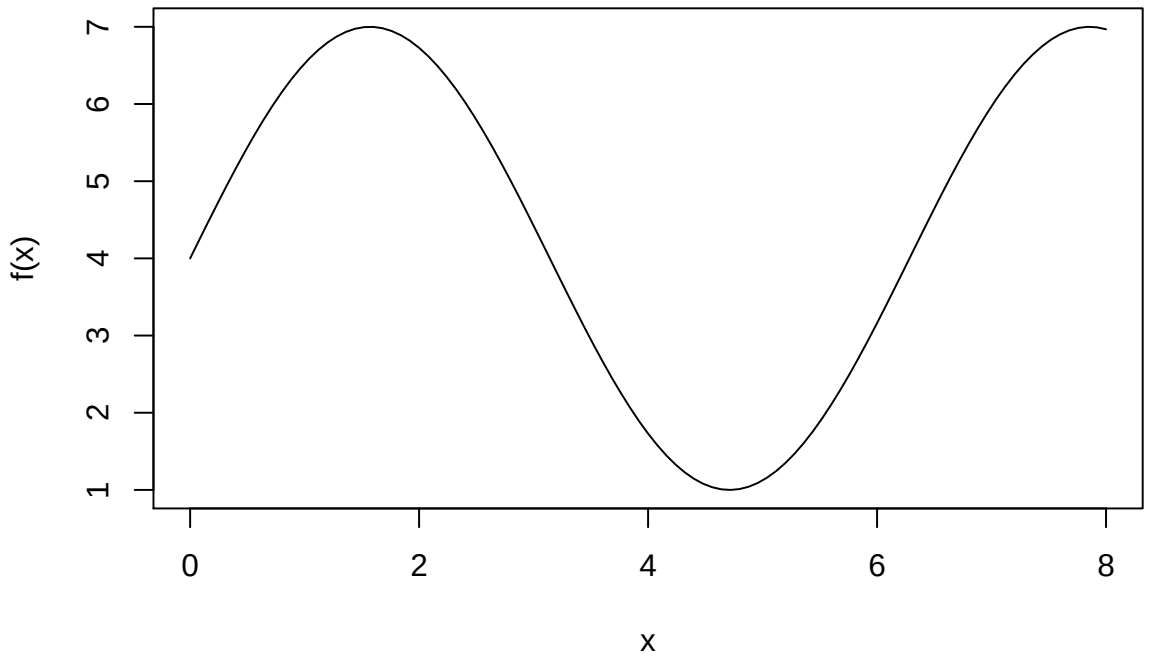
Si lo pensamos desde el punto de vista de la transformada integral de probabilidad, estamos generando una variable aleatoria discreta binaria ($I(y_i \leq f(x_i))$) que toma valor 1 con probabilidad p proporcional al área bajo la curva. Después estamos calculando su media que, coincide con la probabilidad indicada, utilizando Monte Carlo. Esto es:

$$\hat{E}[I(y_i < f(x_i))] = \hat{p} = \frac{1}{N}\sum_{i=1}^{N} I(y_i < f(x_i)).$$

La ley de los grandes números nos asegura que, si N es lo suficientemente grande, esta aproximación convergerá al verdadero valor.

Veamos un ejemplo: Sea $f(x)$ la función que genera la siguiente curva:

```
f <- function(x){
 4+ 3*sin(x)
}


curve(f(x),from = 0, to = 8)
```



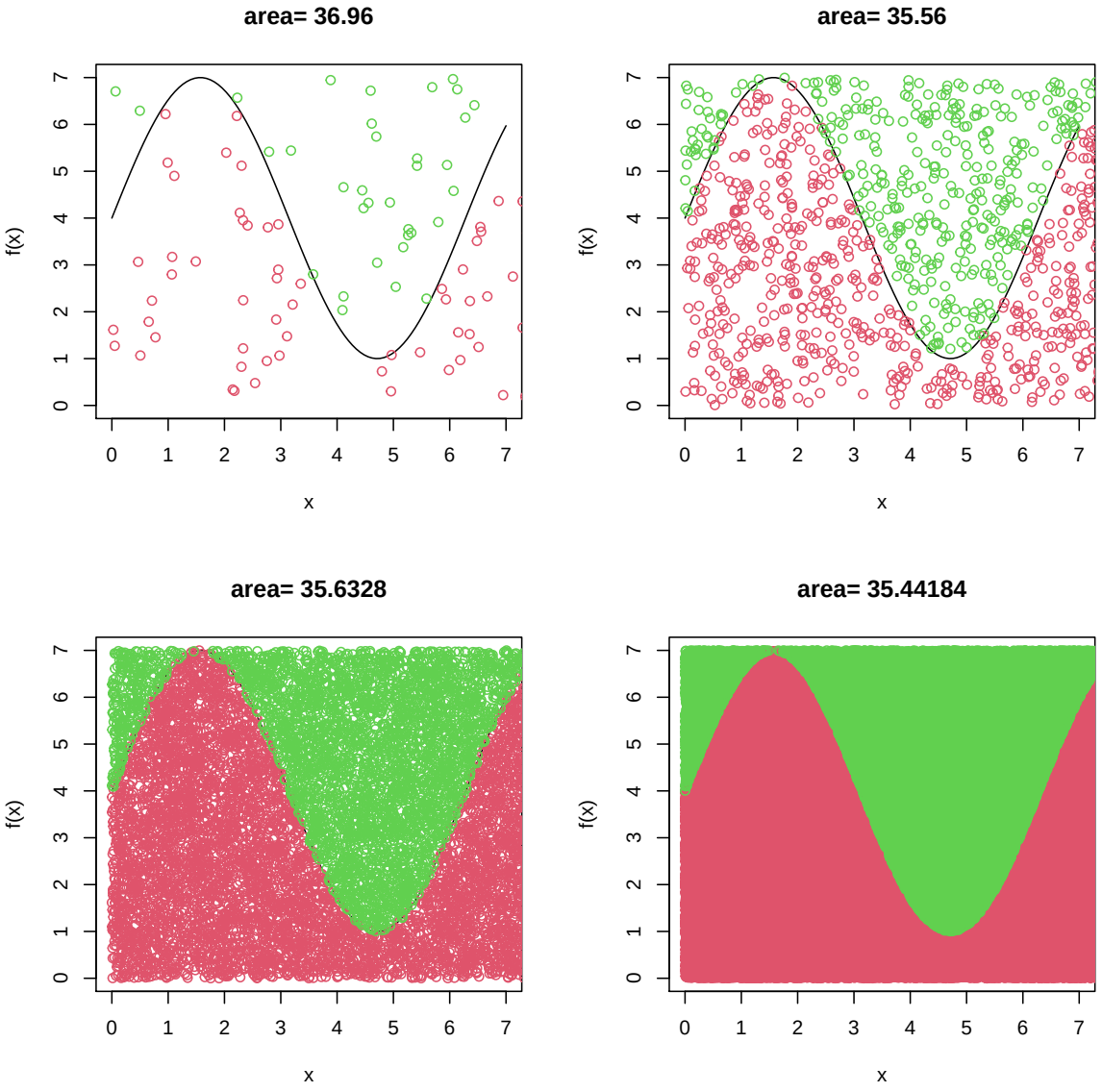
Podemos calcular el valor de la integral en R usando la función *integrate*

```
integrate(f,0,8)
```

```
## 35.4365 with absolute error < 3.9e-13
```

Pero también podemos calcular el área bajo la curva utilizando métodos Monte Carlo como vemos en las siguientes figuras:

```
par(mfrow=c(2,2))
for(N in c(10^2,10^3,10^4, 10^5)){
 x <- runif(N,0,8)
y <- runif(N,0,7)
area <- 7*8*sum(y<=f(x))/N



curve(f(x), from=0, to =7, ylim=c(0,7), main=paste("area=", area))
points(x[y<=f(x)],y[y<=f(x)],col=2)
points(x[y>f(x)],y[y>f(x)],col=3)
}
```



8.3.2. Estimación Monte Carlo de π

De una forma similar podemos encontrar otras magnitudes como, por ejemplo, una aproximación al número π. Para ello, de nuevo, sólo tenemos que simular en el cuadrado $[0,1] \times [0,1]$ y quedarnos con los puntos que cumplan la ecuación $x^2 + y^2 < 1$. Estos supondrán la proporción del área del cuadrado (2*2=4) que pertenece al circulo.

Las siguientes gráficas nos muestran el valor aproximado de π al aumentar π

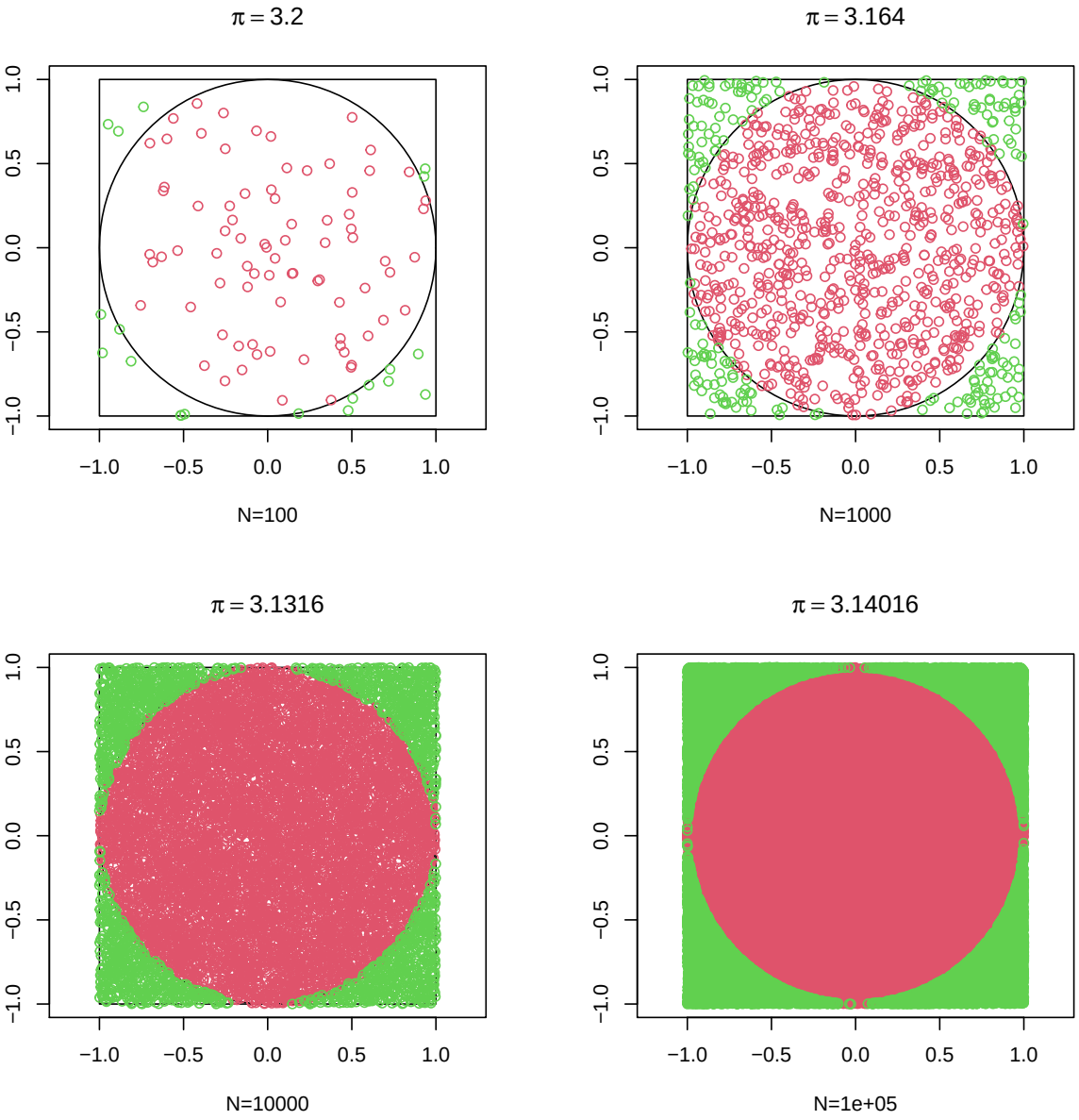
```
require(plotrix)
```

```
## Loading required package: plotrix
```

```
require(grid)
```

```
## Loading required package: grid
```

```
par(mfrow=c(2,2))
for(N in c(10^2,10^3,10^4,10^5)){
  x <- runif(N,-1,1)
  y <- runif(N,-1,1)
   pi_aprox <- 4*sum((x^2+y^2)<=1)/N
  plot(c(-1, 1), c(-1,1), type = "n", asp=1, main=bquote(pi == .(pi_aprox)), xlab = pas
   rect( -1, -1, 1, 1)
  draw.circle( 0, 0, 1 )
  points(x[(x^2+y^2)<=1],y[(x^2+y^2)<=1], col=2)
  points(x[(x^2+y^2)>1],y[(x^2+y^2)>1], col=3)
}
```



8.4. Introducción a las cadenas de Markov

Las cadenas de Markov fueron introducidas por Andrey Markov en 1906. Su objetivo principal era poder aplicar la ley de los grandes números cuando las variables aleatorias que componen la muestra no son independientes.

Pero para poder definir lo que son las cadenas de Markov es necesario entender primero que es un proceso estocástico.

8.4.1. Procesos estocásticos.

La mejor forma de entender que es un proceso estocástico es mediante un ejemplo.

Imaginemos que cada cinco minutos la gerente de un supermercado se acerca a la cola de las cajas y observa cuantas personas hay en ella con el objetivo de controlar que no llegue a 7 ya que, en el momento que lo haga se abrirá una nueva caja.

La primera vez que sale estará observando la variable aleatoria X_0: número de personas en la cola en el instante inicial $t = 0$, en el tiempo $t = 1$ observará X_1 y así sucesivamente. A X_0 se le denomina *estado inicial*, mientras que, a cada una de las observaciones en un tiempo $t = n$, X_n, se les conoce como *estado del proceso en el tiempo* $t = n$

En este escenario, llamamos *espacio de estados* al conjunto de los posibles valores que puede tomar cada una de las variables X_n que en nuestro caso sería $\{0, 1, 2, \ldots, 7\}$.

Si bien es cierto que, tanto el tiempo (de observación) como el espacio de estados podrían ser de naturaleza continua, en este tema nos centraremos en procesos estocásticos como los del ejemplo donde tanto el parámetro de tiempo como el espacio de estados son discretos y, en concreto, en los procesos estocásticos denominados Cadenas de Markov.

8.4.2. Cadenas de Markov.

Una de las características principales de los procesos estocásticos que los diferencian de la sucesión de variables aleatorias de la que hablamos cuando hacemos referencia a la ley de los grandes números (y que ya han aparecido varias veces en este tema), es que lo que sucede en un instante concreto de tiempo X_n estará relacionado con lo que haya pasado en instantes anteriores de tiempo. Se trata, por tanto, de una sucesión de variables NO independientes.

Los tipos de relación temporal pueden ser muy variados y las cadenas de Markov representan un caso particular. En concreto:

Definición: Decimos que un proceso estocástico $X_0, X_1, X_2, \ldots$ con espacio de estados $\{1, \ldots M\}$ es una **Cadena de Markov** cuando la función de probabilidad de X_i para todo $n > 0$ y para cualquier secuencia de valores $\{x_0, \ldots, x_n\}$ en el espacio de estados, sólo depende de lo que sucedió en el instante anterior. Esto es:

$$P(X_{n+1} = x_{n+1} \mid X_0 = x_0,\, X_1 = x_1, \ldots, X_n = x_n) = P(X_{n+1} = x_{n+1} \mid X_n = x_n)$$

Además, cuando el espacio de estados es finito $\{s_1, \ldots, s_k\}$ hablaremos de una *Cadena de Markov finita.*

Fijaos que en el caso de una cadena de Markov finita, el interés reside en saber cual es la probabilidad de pasar del estado s_i al estado s_j en un instante de tiempo n. A este tipo de probabilidades se les denomina **probabilidades de transición** y diremos que las probabilidades de transición son **estacionarias** cuando no cambian con el instante de tiempo sino, únicamente, con el hecho de pasar del estado s_i al estado s_j. Esto es:

$$p_{i,j} = P(X_{n+1} = s_j \mid X_n = s_i)$$

Volvamos a nuestro ejemplo. Para que pudiésemos considerarlo una cadena de Markov necesitaríamos que el número de personas en la cola sólo dependiese del número de personas en la cola en el instante anterior (es decir, 5 minutos antes). Las probabilidades de transición nos indicarían, por ejemplo, que probabilidad hay de pasar a tener 5 personas en la cola cuando actualmente hay 3 ($p_{3,5}$). Además, podríamos considerar que las probabilidades de transición son estacionarias si estas se mantienen constantes en el tiempo (cosa que no es muy probable en un supermercado).

Por su notación, $p_{i,j}$, parece bastante razonable ordenar las probabilidades de transición en forma de matriz:

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,k} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k,1} & p_{k,2} & \cdots & p_{k,k} \end{bmatrix}$$

a la que denominaremos **matriz de transición** y nos permite especificar, de forma sencilla, las probabilidades de pasar de un estado s_i a otro s_j en t pasos, simplemente multiplicando $P \dots P = P^t$. Esto es:

$$P(X_{n+t} = s_j \mid X_n = s_i) = p_{i,j}^{(t)}$$

siendo $p_{i,j}^{(t)}$ el elemento (i, j) de la matriz P^t.

Veamos otro ejemplo. Supongamos que ir al cine en un día determinado depende de si hemos ido al cine el día anterior o no. En concreto, iremos al cine con probabilidad 1/3 si ya hemos ido hoy y con probabilidad 1/2 si hoy no hemos ido. Con esta premisa la matriz de transición sería:

$$P = \begin{array}{c} \\ \text{sí} \\ \text{no} \end{array} \begin{array}{c} \begin{array}{cc} \text{sí} & \text{no} \end{array} \\ \begin{pmatrix} 1/3 & 2/3 \\ 1/2 & 1/2 \end{pmatrix} \end{array}$$

Si queremos saber que probabilidad tenemos de ir al cine dentro de dos días sabiendo que hoy sí que hemos ido podemos hacerlo simplemente multiplicando P por si misma y obtenemos:

$$P^2 = \begin{array}{c} \\ \text{sí} \\ \text{no} \end{array} \begin{array}{c} \begin{array}{cc} \text{sí} & \text{no} \end{array} \\ \begin{pmatrix} 4/9 & 5/9 \\ 5/12 & 7/12 \end{pmatrix} \end{array}$$

Y la probabilidad que estamos buscando es 4/9.

Si utilizamos el lenguaje de las funciones de probabilidad del tema anterior, podríamos decir que la matriz de transición P *codifica* la función de probabilidad condicionada al valor del estado inicial X_0. En concreto, la fila i-ésima de P contiene los valores de la función de probabilidad para $X_1 \mid X_0 = s_i$ mientras que la misma fila de P^n contiene los de $X_n \mid X_0 = s_i$.

Pero, si queremos recuperar la distribución marginal de cada uno de los X_n necesitaremos conocer, además de P, las condiciones iniciales con las que se inició la cadena, es decir, X_0 o la función de probabilidad que las genera, es decir $P(X_0 = s_i) = t_i$ para $i = 1, \ldots, k$

Resultado Conocido el vector $\boldsymbol{t} = (t_1, \ldots, t_k)$ La probabilidad marginal de $X_n = s_j$ puede obtenerse como el j-ésimo valor del vector $\boldsymbol{t}P^n$.

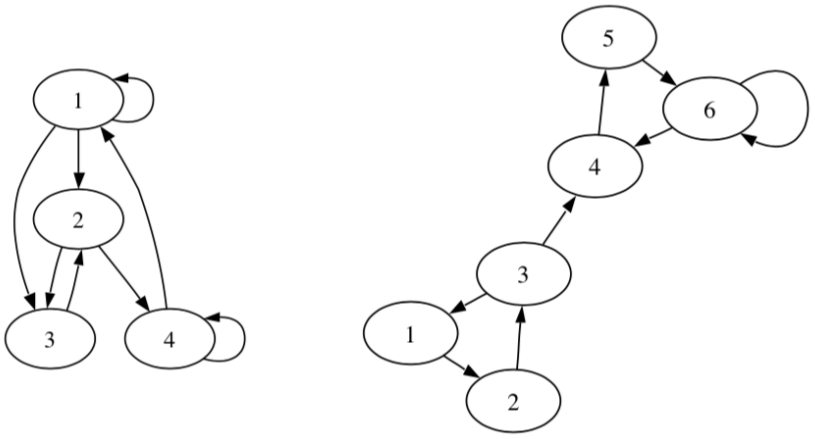
Volviendo al ejemplo, si sabemos que el primer día que empecé mi *política* de ir al cine, elegí ir con probabilidad 1/2, la probabilidad de ir al cine el cuarto día será 0.4286265:

$$\boldsymbol{t}P^4 = \begin{bmatrix} 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 0{,}4290123 & 0{,}5709877 \\ 0{,}4282407 & 0{,}5717593 \end{bmatrix} = \begin{bmatrix} 0{,}4286265 & 0{,}5713735 \end{bmatrix}$$

8.4.3. Tipos de estados

Las cadenas de Markov pueden clasificarse según sus estados. A su vez, los estados de una cadena de Markov finita pueden clasificarse según el tiempo que se permanezca en el y las veces que se visite.

En general, definiremos un estado **recurrente** como aquel que volveremos a visitar, eventualmente, con probabilidad 1. Sin embargo, un estado **transitorio** será aquel al que podríamos no volver nunca más. Podemos ilustrarlo con la siguiente representación gráfica de dos cadenas de Markov.



Podemos ver que, para la cadena de Markov de la izquierda, siempre existe la probabilidad pasar de unos estados a otros, y una "partícula" que entre en el sistema no dejaría de moverse por todos los estados. Sin embargo, la cadena de la derecha presenta un conjunto de estados transitorios 1, 2 y 3 por los que se puede mover hasta que caiga en el estado 4, momento a partir del cual, se quedará entre los estados 4, 5 y 6 que serán nuestros estados recurrentes

Cuando todos los estados de la cadena son recurrentes o lo que es lo mismo, siempre puedo llegar de un estado a otro en un número finito de pasos (como en la figura de la izquierda), diremos que la cadena de Markov es **irreducible**.

Por otra parte, también podemos hablar del **periodo de un estado** i como el tiempo que se tarda en volver a ese estado después de visitarlo. Para calcularlo bastará con coger el máximo común divisor de todos los valores n tales que el elemento (i, i) de la matriz P^n es distinto de 0.

Por supuesto, el periodo no podrá definirse si es imposible volver a un determinado estado, y llamaremos aperiodicos a los estados con periodo 1. Del mismo modo, definiremos una cadena de Markov aperiodica como la cadena en la que todos sus estados lo son.

8.4.4. Distribución Estacionaria

Cuando tenemos una cadena de Markov resulta inevitable plantearse que pasará en el largo plazo, ¿Acabaré yendo al cine todos los días? ¿Dejaré de ir?

Definición Decimos que un vector $\boldsymbol{l}^t = (l_1, \ldots, l_k)$ de probabilidades tales que $\sum_{i=1}^{k} l_k = 1$

es una distribución estacionaria de una cadena de Markov con matriz de transición P si

$$\boldsymbol{l}P = \boldsymbol{l}.$$

Fijaos que esta condición implica que si la función de probabilidad inicial de X_0 es su distribución estacionaria, $\boldsymbol{t} = \boldsymbol{l}$, las probabilidades marginales se mantendrán constantes para $X_1, X_2, \ldots$.

NOTA La distribución estacionaria es una distribución marginal y no una distribución condicional. En el ejemplo del cine, se puede comprobar que la distribución estacionaria es $\boldsymbol{l}^t = (3/7, 4/7)$. Esto lo podemos hacer fácilmente estableciendo la ecuación

$$\boldsymbol{t}P = \begin{bmatrix} l & 1-l \end{bmatrix} \begin{bmatrix} 1/3 & 2/3 \\ 1/2 & 1/2 \end{bmatrix} = \begin{bmatrix} l & 1-l \end{bmatrix}$$

de la que se comprueba que la (única) solución es $3/7$.

Pero, ¿Se da siempre esta circunstancia? ¿Existe la distribución estacionaria? ¿Es única?

Teorema Toda cadena de Markov irreducible tiene distribución estacionaria única.

Hemos visto que, en el ejemplo del cine, existía la distribución estacionaria, además, con el teorema anterior, podemos decir que la solución es única. Sin embargo, de cara a calcular dicha distribución, no siempre será tan fácil.

El siguiente teorema nos muestra que podemos obtener el mismo resultado de forma empírica calculando P^n con $n \to \infty$.

Teorema Para cualquier cadena de Markov irreducible y aperiodica, la probabilidad marginal

$$P(X_n = s_i)$$

converge a la distribución estacionaria con $n \to \infty$. Equivalentemente, P^n converge a una matriz de transiciones donde cada fila es la distribución estacionaria.

Por último, resulta interesante conocer cual es la probabilidad de volver a un estado dado y su relación la distribución estacionaria.

Teorema Sea $X_0, X_1, \ldots$ una cadena de Markov irreducible con distribución estacionaria $\boldsymbol{l}$. Si r_i es el tiempo esperado que tardará la cadena en volver al estado s_i dado que empezó en s_i, la distribución estacionaria se puede calcular como $l_i = 1/r_i$.

8.5. Simulación de una variable aleatoria

Tal y como vimos en la práctica 4, R dispone de una función específica que permite simular valores aleatorios para cada una de las distribuciones de probabilidad más habituales. En la práctica 4 también se presentó el listado de distribuciones de probabilidad disponibles en R.

Así, por ejemplo, la función *rexp* es la función que genera un vector del tamaño solicitado de números aleatorios de una distribución exponencial, mientras que *rbinom* es la función que genera vectores de números aleatorios de una binomial.

```
> sims.binom <- rbinom(500,10,0.5) # Simulación de 500 valores de una Bin(10,0.5)
```

Pero existen muchas otras funciones de probabilidad o densidad para las que R no lleva incorporada una función específica que permita simular de ellas. Si la función de distribución acumulada F y su inversa F^{-1} son conocidas, podemos utilizar el teorema de la **Transformada Integral de Probabilidad** para simular de ella. Esto es.

Teorema Sea U una v.a. distribuida según una $Unif(0,1)$ y F una función que cumple las condiciones para ser una función de distribución acumulada. La variable aleatoria $X = F_X^{-1}(U)$ es una v.a. con función de distribución acumulada F.

En general, dada una muestra aleatoria $\{U_1, U_2, \ldots, U_n\}$ donde cada U_i es i.i.d. $Unif(0,1)$, entonces $X_1 = F_X^{-1}(u_1), X_2 = F_X^{-1}(u_2), \ldots, X_n = F_X^{-1}(u_n)$ son v.a.'s i.i.d. con función de distribución acumulada F.

Es decir, basta saber simular de una Un(0,1), lo cual podemos hacer con la función *runif* y luego transformar los resultados según la inversa de la función de distribución acumulada.

Ejercicio 1.- Simular 10000 valores de una exponencial de parámetro (tasa) 10 utilizando este teorema. Comparar el resultado con el que se obtiene con R, utilizando para ello un histograma de los valores simulados superponiendo la densidad de la Exp(10).

Para ello basta con simular $n = 10000$ de una Un(0,1) con R y utilizar el resultado anterior para obtener valores simulados de la Exponencial,

$$F_X(x) = 1 - e^{-\lambda x} \longrightarrow F_X^{-1}(u) = \frac{-\log(1-u)}{\lambda} \sim \text{Ex}(\lambda)\,,$$

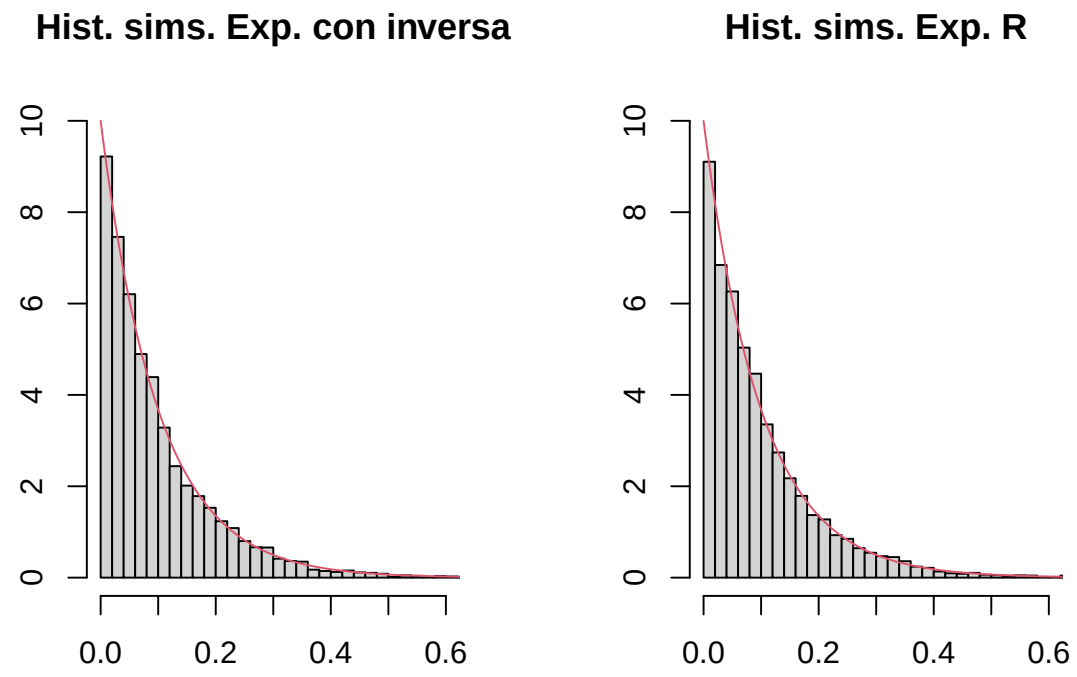
```r
# Función casera para simulación por inversión de una exponencial
rexpcasera <- function(n=1,lambda = 1){
  u <- runif(n)
  -log(1-u)/lambda
}
sim1 <- rexpcasera(10000,10)
```

Ahora basta con simular de una exponencial con la función de R

```r
sim2 <- rexp(10000,10)
```

y representar gráficamente ambas simulaciones incorporando la función de densidad

```r
par(mfrow=c(1,2))
hist(sim1, prob=T,xlim=c(0,0.6),ylim=c(0,10), nclass = 40,
     xlab =" ", ylab =" ", main="Hist. sims. Exp. con inversa")
valores <- seq(0,1,by=0.001)
lines(valores, dexp(valores,10),col=2)
hist(sim2, prob=T,xlim=c(0,0.6),ylim=c(0,10), nclass = 40,
     xlab =" ",ylab =" ", main="Hist. sims. Exp. R")
lines(valores, dexp(valores,10),col=2)
```



8.5.1. Integración Monte Carlo

La metodología Monte Carlo nos permite aproximar cantidades desconocidas utilizando valores simulados y la ley de los grandes números. En concreto, nos permite obtener áreas por debajo de una función $f(x)$ de la que no sabemos integrar. La idea es simular valores de una superficie con área conocida y ver la proporción de estos que se encuentran por debajo de la función en la zona en la que queremos calcular la integral.

Ejercicio 2.- Calcular aproximadamente el valor del número e.

Para ello vamos a calcular el área bajo la curva $1 + \mathrm{e}^x$ entre $[0, 1]$ ya que su valor es precisamente ese valor:

$$\int_0^1 (1 + \mathrm{e}^x)\mathrm{d}x = [x + \mathrm{e}^x]_0^1 = \mathrm{e}$$

Podemos calcular dicho valor con R:

```
f <- function(x){ 1 + exp(x) } ; integrate(f,0,1)
```

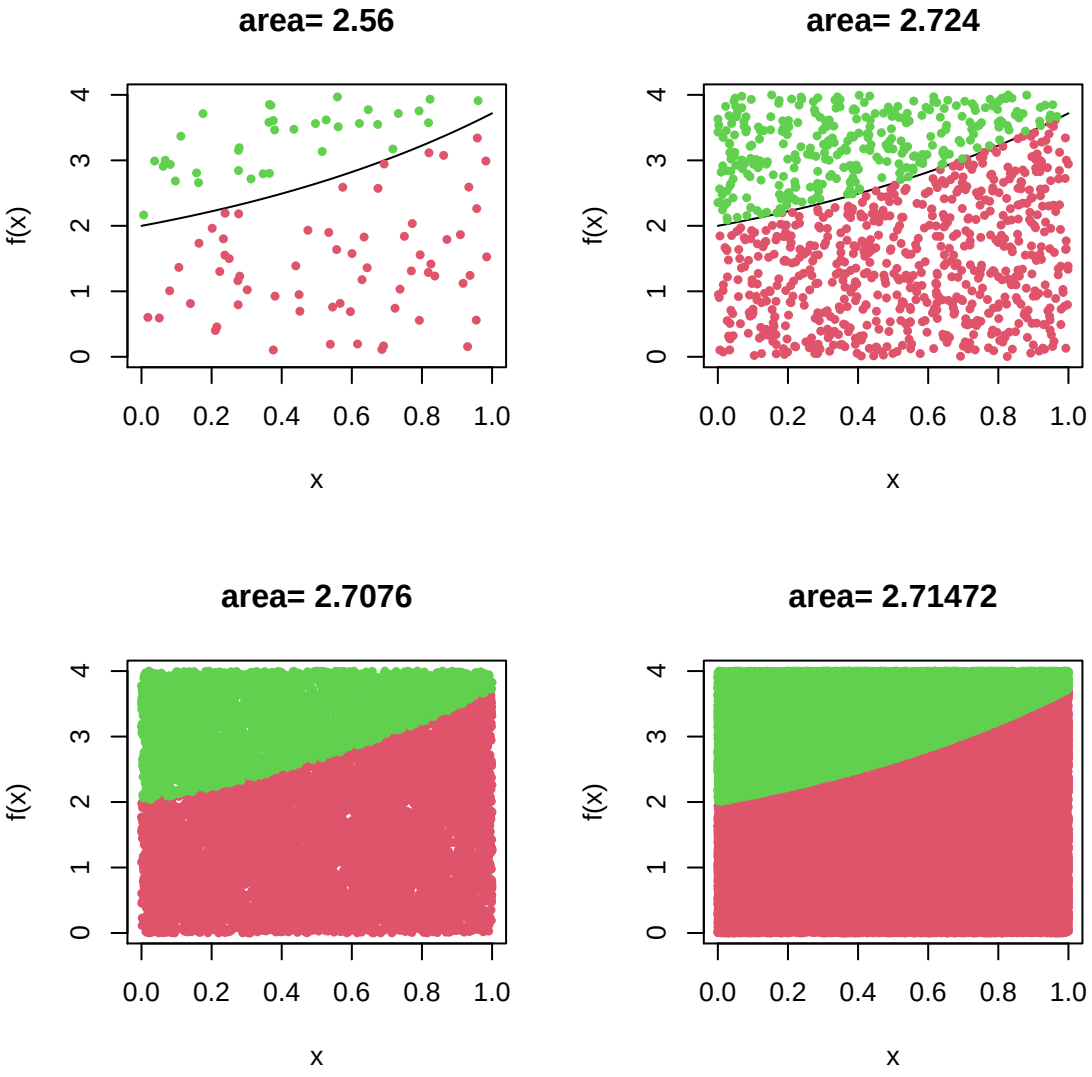
```
## 2.718282 with absolute error < 3e-14
```

```
exp(1)
```

```
## [1] 2.718282
```

Ahora vamos a ver como podemos hacerlo simulando valores y calculando cuantos puntos quedan por debajo de la curva. La idea es simular de valores del rectángulo $[0, 1] \times [0, 4]$ con área $(1 - 0) \times 4 = 4$, y contar los puntos bajo la curva. Lo haremos para 10^2, 10^3, 10^4 y 10^5 puntos.

```
par(mfrow=c(2,2))
for(N in c(10^2,10^3,10^4,10^5)){
x <- runif(N,0,1)
y <- runif(N,0,4)
area <- 1*4*sum(y<=f(x))/N
curve(f(x), from=0, to =1, ylim=c(0,4), main=paste("area=", area))
points(x[y<=f(x)],y[y<=f(x)],col=2,pch=16,cex=0.75)
points(x[y>f(x)],y[y>f(x)],col=3,pch=16,cex=0.75)
}
```



Observa que aún con 10^5 puntos no se consigue una aproximación buena, y además se ha tardado mucho en obtener el resultado. Bueno, eso depende de la máquina con la que hayas trabajado. Si tienes tiempo, prueba con 5×10^5 o con 10^6. Con 10^8 igual colapsas el ordenador.

8.6. Simulación por métodos MCMC

Acabamos de ver como simular de distribuciones de probabilidad univariantes de las que conocemos su función de distribución acumulada. Luego hemos visto como, utilizando dichas simulaciones podemos calcular integrales que no serían fáciles de calcular.

Ahora vamos a unir ambas ideas al concepto de Cadenas de Markov para simular de variables aleatorias de las que la distribución F no es fácilmente calculable. Es decir, para aquellas situaciones en las que no es posible simular directamente de la distribución de

probabilidad.

Antes de proceder, recordemos que una cadena de Markov es un proceso estocástico donde la variable aleatoria en un instante n, X_n depende de la variable aleatoria en el instante inmediatamente anterior. En las clases de teoría habéis estudiado Cadenas de Markov con espacio de estados discreto y finito. Una de las caracteristicas más importantes de las Cadenas de Markov es la existencia de distribución estacionaria.

La **distribución estacionaria** de una Cadena de Markov es la función de probabilidad (en el caso discreto) que nos indica que la probabilidad marginal de estar en un estado concreto se estabiliza con el tiempo y se vuelve independiente del valor del estado en el instante anterior.

Si el espacio de estados es continuo, en lugar de una función de probabilidad, la distribución estacionaria quedará determinada por una función de densidad.

En el contexto de la simulación, el objetivo será obtener una muestra de una distribución con función de densidad f, partiendo de una cadena de Markov $X^{(0)}, X^{(1)}, \ldots$ cuya distribución estacionaria esté determinada por esa misma f. El procedimiento utilizado, en terminos generales, es el siguiente:

- Partiendo de un valor inicial $X^{(0)}$,
- encontrar una función de densidad condicionada $p(\cdot|\cdot)$ tal que en cada paso i podamos obtener el siguiente valor $X^{(i+1)}$ como una simulación de la distribución $\mathrm{p}(X^{(i+1)}|X^{(i)})$,
- repetir el procedimiento obteniendo una Cadena de Markov $X^{(0)}, \ldots, X^{(n)}$ hasta que se alcanza la convergencia a la distribución estacionaria f (período de **calentamiento** de la cadena),
- seleccionar las siguientes $X^{(n+1)}, \ldots, X^{(n+p)}$ como la muestra buscada.

La gran dificultad de este procedimiento reside en encontrar la función intermedia $p(\cdot|\cdot)$ de la que simular para conseguir llegar a la distribución estacionaria f (conocida o parcialmente conocida a falta de la constante de integración). Sobre todo teniendo en cuenta que la complejidad aumenta al tener datos multivariantes y/o parámetros desconocidos.

A continuación, estudiaremos dos aproximaciones: Gibbs-Sampling y Metropolis-Hastings.

Nota: fijaos que utilizamos el super índice para indicar el instante de la cadena de Markov en el que estamos sin confundirlo con el posible subíndice a la hora de trabajar con vectores aleatorios.

8.6.1. Gibbs-Sampling

Supongamos que tenemos un vector de v.a's $\boldsymbol{X}$ con función de densidad conjunta $f(\boldsymbol{x})$ de la que queremos simular. Puede que no sea fácil simular de la función conjunta pero si lo sea hacerlo de sus condicionales $f(x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_d)$. Estas serán, por tanto, nuestras funciones p intermedias y el procedimiento de simulación será:

1. Inicializar con $\boldsymbol{x}^{(0)}=(x_1^{(0)},\ldots,x_d^{(0)})$.
2. Simular $x_1^{(1)}$ de la **distribución condicional** $x_1|(x_2^{(0)},\ldots,x_d^{(0)})$.
3. Simular $x_2^{(1)}$ de la **distribución condicional** $x_2|(x_1^{(1)},x_3^{(0)},\ldots,x_d^{(0)})$.
4. ...
5. Simular $x_d^{(1)}$ de la **distribución condicional** $x_d|(x_1^{(1)},\ldots,x_{d-1}^{(1)})$.
6. volver al paso 1 sustituyendo $\boldsymbol{x}^{(0)}$ por $\boldsymbol{x}^{(1)}$

Tras eliminar las n primeras observaciones, el resultado es una **muestra aleatoria** $\{(x_1^{(i)},\ldots,x_d^{(i)})\}_{i=n+1}^{n+p}$ de la **distribución conjunta** $f(\boldsymbol{x})$.

Ejercicio 3.- Obtener una muestra de la distribución conjunta:

$$f(x,y)\propto\binom{16}{x}y^{x+1}(1-y)^{19-x}\qquad x=0,1,\ldots,16\qquad 0\leq y\leq 1$$

y de la marginal $f(x)$, que en realidad es una distribución beta-binomial de parámetros $\alpha=2$, $\beta=4$ y $n=16$ (de la que es posible simular y así poder comparar resultados).

Las condicionales tienen una expresión de la que es sencillo simular:

$$\begin{aligned}(x|y) &\sim \text{Binomial}(16,y)\\ (y|x) &\sim \text{Beta}(x+2,20-x)\end{aligned}$$

Procedimiento:

1. Inicializar con (x_0,y_0).
2. Simular x_1 de la distribución condicional $(x|y_0)\sim\text{Binomial}(16,y_0)$.

3. Simular y_1 de la distribución condicional $(y|x_1) \sim \text{Beta}(x_1 + 2, 20 - x_1)$.

4. Repetir la simulación.

Construimos una función para simular con este algoritmo y simulamos 1000 valores después de dejar 1000 de calentamiento:

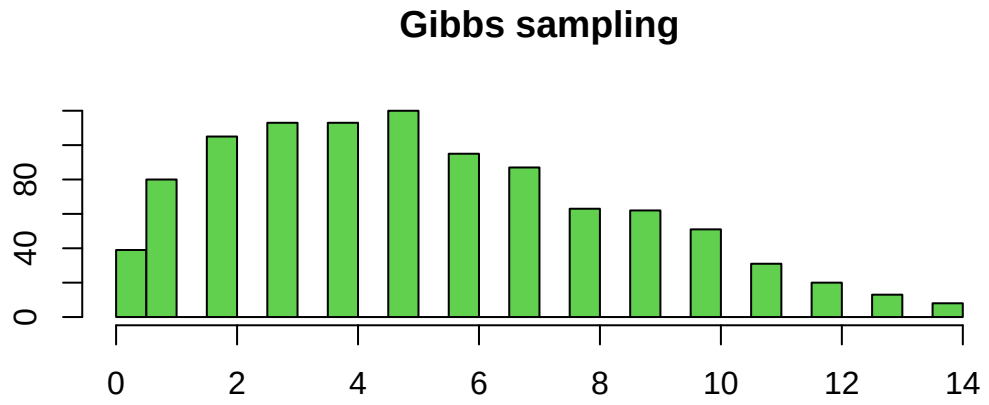
```
gibbs <- function(nsim=1000,ncal=100)
{
xc <- matrix(0,nrow=nsim,ncol=2)
xc[1,2] <- 0
for(t in 1:(nsim-1)){
  xc[t+1,1] <- rbinom(1,16,xc[t,2])
  xc[t+1,2] <- rbeta(1,xc[t+1,1]+2,20-xc[t+1,1])
}
xc[(ncal+1):nsim,]
}
sim3 <- gibbs(2000,1000)
```

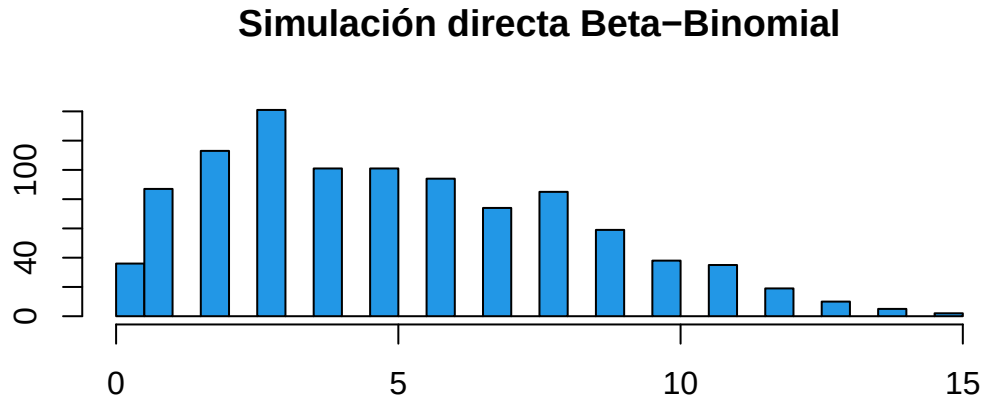
Teniendo en cuenta que la primera componente de los vectores simulados es una simulación de la marginal $f(x)$, podemos comparar el resultado obtenido con una simulación de la beta-binomial correcta:

```
# función para simular de una beta-binomial
rbetabin <- function(length,alpha,beta,n) {
  x <- rbinom(length,n,rbeta(length,alpha,beta))
  return(x)
}
bbsim <- rbetabin(1000,2,4,16) # simulación de la beta-binomial
```

Para ello utilizamos los histogramas de ambas muestras:

```
par(mfrow=c(2,1))
hist(sim3[,1],nclass=40,col=3,xlab=" ", ylab=" ",
     main="Gibbs sampling")
hist(bbsim,nclass=40,col=4,xlab=" ", ylab=" ",
     main="Simulación directa Beta-Binomial")
```





Como además la esperanza de una densidad condicional se puede expresar como

$$E[f(x|Y)] = \int f(x|y)f(y)\mathrm{d}y = f(x)\,,$$

podemos utilizar Monte Carlo para aproximar esta integral y por tanto aproximar los valores de $f(x) = P(X = x)$ mediante

$$\tilde{f}(x) = \frac{1}{m}\sum_{i=1}^{m} f(x|y_i)\,,$$

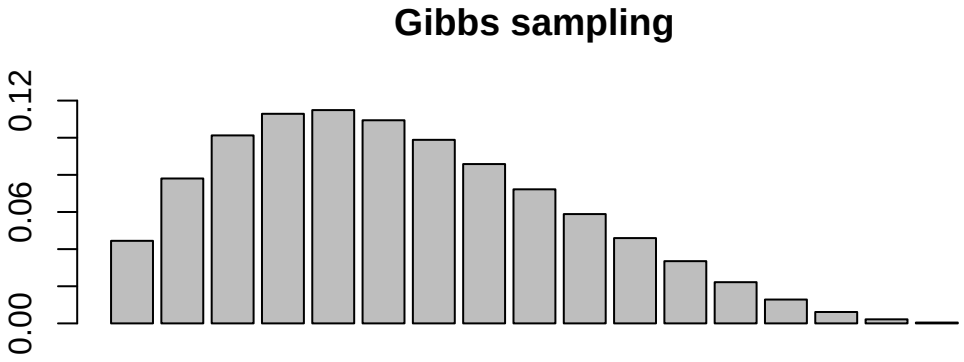
donde $y_1, \ldots, y_m$ son las simulaciones de $f(y)$ obtenidas por Gibss.

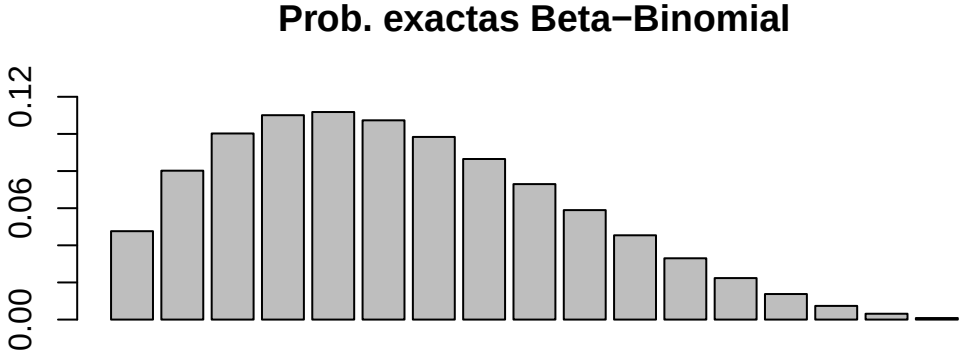
Así las simulaciones obtenidas por Gibbs nos permiten aproximar las probabilidades de la marginal $f(x)$:

```
fxtilde <- function(x,sims) {1/length(sims)*sum(dbinom(x,16,sims))}
fxtildex <- numeric(17)
for(i in 1:17){fxtildex[i] <- fxtilde(i-1,sim3[,2])}
```

y compararlas con las obtenidas con la distribución beta-binomial real:

```
dbetabin <- function(x,a,b,n) { choose(n,x) * beta(x+a,n-x+b)/beta(a,b) }
fx <- numeric(17)
for(i in 1:17){fx[i] <- dbetabin(i-1,2,4,16)}

par(mfrow=c(2,1))
barplot(fxtildex, ylim=c(0,0.12), main="Gibbs sampling")
barplot(fx,ylim=c(0,0.12), main="Prob. exactas Beta-Binomial")
```





8.6.2. Metropolis-Hastings

El algoritmo de Metropolis-Hastings es otro método MCMC que permite simular de distribuciones de las que no conocemos (por la dificultad de la integral necesaria para obtenerla) la constante que hace que la función de densidad integre 1. Este método se utilizará cuando tampoco sea posible simular de las distribuciones condicionales.

La idea general es simular de una distribución *conveniente* y decidir si el punto simulado cumple los requisitos para ser parte de la muestra de la función objetivo o no.

El procedimiento en el paso i sería el siguiente.

1. Simular un candidato $\boldsymbol{X}^{(*)}$ de una distribución $q(\cdot|\boldsymbol{X}^{(i)})$ que tenga una forma conveniente
2. Calcular la probabilidad de aceptación:
$$\alpha(\boldsymbol{x}^{(*)}, \boldsymbol{x}^{(i)}) = \min\left(1, \frac{f(\boldsymbol{x}^{(*)})\, q(\boldsymbol{x}^{(i)}|\boldsymbol{x}^{(*)})}{f(\boldsymbol{x}^{(i)})\, q(\boldsymbol{x}^{(*)}|\boldsymbol{x}^{(i)})}\right)$$
3. Simular un valor u de una Uniforme$(0, 1)$
4. Considerar $\boldsymbol{x}^{(i+1)} = \begin{cases} \boldsymbol{x}^{(*)} \text{ si } u \leq \alpha \\ \boldsymbol{x}^{(i)} \text{ si } u > \alpha \end{cases}$
5. Iterar este procedimiento.

Ejercicio 4.- Obtener una simulación de la siguiente distribución de la que solo conocemos su densidad menos la constante de integración:

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

y utilizar la simulación obtenida para calcular la esperanza de dicha distribución. Aprovecharse del hecho de que esta distribución es conocida para comprobar que la distribución estacionaria a la que ha convergido la cadena es la correcta N(1,1).

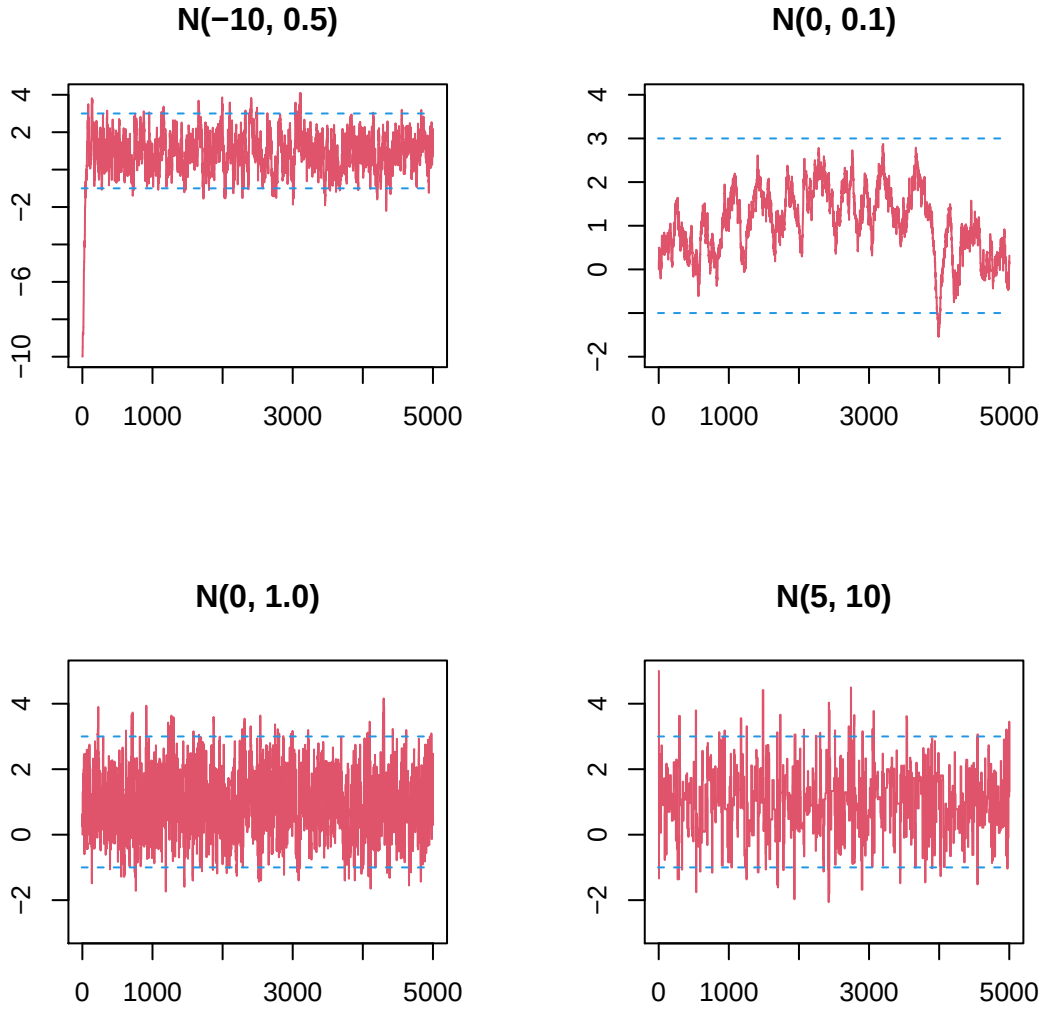
Para ello utilizaremos como distribuciones propuesta $q(.|x)$ distribuciones normales con desviaciones típica 0.5, 0.1, 1 y 10. Utilizaremos también diferentes valores iniciales de las cadenas y diferentes periodos de calentamiento para ver que ocurre en cada caso.

```
# la función de densidad de la que queremos simular
fno <- function(x){ exp(-1/2 * (x-1)^2)}

# El algoritmo M-H
mh <- function(nsim,inicial,std)
{
  sims <- numeric(nsim); sims[1]<-inicial
  for(t in 1:nsim)
    {
    rn <- rnorm(1,sims[t],std)
    alpha <- min(1,fno(rn)/fno(sims[t]))
    ru <- runif(1,0,1)
    if (ru<=alpha) sims[t+1] <- rn else sims[t+1] <- sims[t]
    t <- t+1
    }
  sims
}
nsims <- 5000 ; ncal <- 1000 # probar a cambiar todos estos valores
mhsims1 <- mh(nsims,-10,0.5)
mhsims2 <- mh(nsims,  0,0.1)
mhsims3 <- mh(nsims,  0,1.0)
mhsims4 <- mh(nsims,  5, 10)
```

Vamos a ver los resultado en una gráfica común incorporando en color azul líneas que marquen los valores habituales entre los que encontraríamos simulaciones de una N(1,1):

```
par(mfrow=c(2,2))
plot.ts(mhsims1,ylim=c(-10,4),xlab=" ",ylab=" ", col=2, main = "N(-10, 0.5)")
lines(c(-10,nsims+10),c(-1,-1),lty=2,col=4)
lines(c(-10,nsims+10),c(3,3),lty=2,col=4)

plot.ts(mhsims2,ylim=c(-2,4),xlab=" ",ylab=" ", col=2, main = "N(0, 0.1)")
lines(c(-10,nsims+10),c(-1,-1),lty=2,col=4)
lines(c(-10,nsims+10),c(3,3),lty=2,col=4)
```

```
plot.ts(mhsims3,ylim=c(-3,5),xlab=" ",ylab=" ", col=2, main = "N(0, 1.0)")
lines(c(-10,nsims+10),c(-1,-1),lty=2,col=4)
lines(c(-10,nsims+10),c(3,3),lty=2,col=4)

plot.ts(mhsims4,ylim=c(-3,5),xlab=" ",ylab=" ", col=2, main = "N(5, 10)")
lines(c(-10,nsims+10),c(-1,-1),lty=2,col=4)
lines(c(-10,nsims+10),c(3,3),lty=2,col=4)
```



Para calcular la esperanza basta con calcular la media de los valores simulados (tal y como hemos visto en integración Monte Carlo):

```
mean(mhsims1[(ncal+1):nsims])
```

```
## [1] 0.9175292
```

```r
mean(mhsims2[(ncal+1):nsims])
```

```
## [1] 1.177105
```

```r
mean(mhsims3[(ncal+1):nsims])
```

```
## [1] 0.9916431
```

```r
mean(mhsims4[(ncal+1):nsims])
```

```
## [1] 1.000446
```

Referencias

Joseph K. Blitzstein and Jessica Hwang. *Introduction to Probability*. CRC Press, 2015.

M. H. DeGroot and M.J. Schervish. *Probability and Statistics*. Addison-Wesley, 4 edition, 2012.