

Introducción a la probabilidad

Carlos J. Gil Bellosta

2016-11-04

Índice

1. Eventos y probabilidades	3
1.1. Referencias	4
1.2. Ejercicios	4
2. Probabilidades conjuntas, marginales y condicionales, teorema de Bayes e independencia	4
2.1. Probabilidad conjunta, marginal y condicional	4
2.2. Teorema de Bayes	6
2.3. Independencia	7
2.4. Independencia y regla de la cadena	8
2.5. Independencia condicional	9
2.6. Apéndice: SamIam	9
2.7. Referencias	9
2.8. Ejercicios	9
3. Distribuciones de probabilidad	11
3.1. Referencias	12
4. Distribuciones de probabilidad discretas	12
4.1. Distribución de Dirac	12
4.2. Distribución de Bernoulli	12
4.3. La media de una variable aleatoria	13
4.4. Distribución binomial	14
4.5. Varianza y desviación estándar	14
4.6. La distribución multinomial	15
4.7. La distribución de Poisson	16
4.8. Urnas y muestras	17
4.9. Referencias	17
4.10. Ejercicios	18
5. Distribuciones de probabilidad continuas	19
5.1. De histogramas a funciones de densidad	19
5.2. Funciones de densidad, probabilidad y cuantiles	21
5.3. La distribución uniforme	25
5.4. La distribución beta	26
5.5. La distribución normal	28
5.6. La distribución t	29
5.7. Las distribuciones gamma y lognormal	30
5.8. Mezclas de distribuciones	31
5.9. Distribuciones jerárquicas	33
5.10. Consideraciones finales	33
5.11. Referencias	33
5.12. Ejercicios	34
6. Distribuciones de probabilidad multivariantes	35
6.1. Independencia	36

6.2. Teorema de Bayes	37
6.3. Covarianza y correlación	39
6.4. Referencias	40
6.5. Ejercicios	40
7. Introducción a la estadística	40
7.1. Universo, muestra y big data	40
7.2. El objeto de la estadística	41
8. Estadística descriptiva	42
8.1. Resúmenes numéricos	42
8.2. Visualización de datos	44
8.3. Análisis exploratorio de datos (EDA)	52
8.4. Referencias	53
8.5. Ejercicios	53
9. Estimación puntual	53
9.1. El método de los momentos	53
9.2. Funciones de pérdida	54
9.3. Estimación por máxima verosimilitud	56
9.4. Teoremas de convergencia	56
9.5. Variabilidad de las estimaciones e intervalos de confianza	57
9.6. Referencias	65
9.7. Ejercicios	65
10. Pruebas de hipótesis	65
10.1. Test de significancia	65
10.2. Pruebas de hipótesis	68
10.3. Zonas grises de las pruebas anteriores y NHST	69
10.4. El tamaño del efecto y las pruebas S y M	70
10.5. Pruebas de hipótesis e intervalos de confianza	70
10.6. Algunas pruebas de libro	71
10.7. Referencias	75
10.8. Ejercicios	75
11. Datos tabulares	76
11.1. Biplots y dependencias	79
11.2. Relaciones de dependencia y modelos loglineales	79
11.3. Referencias	80
12. Introducción a la modelización estadística	80
12.1. El modelo lineal	81
12.2. Regresión logística	91
12.3. Modelos lineales generalizados	94
12.4. Referencias	98
12.5. Ejercicios	98
13. Introducción a la estadística bayesiana	98
13.1. Teorema de Bayes	99
13.2. Distribuciones a priori	100
13.3. Construcción de las distribuciones a posteriori	101
13.4. Introducción a Stan	102
13.5. Caso práctico: prueba de la diferencia de medias	103
13.6. Caso práctico: suavizado estadístico	109
13.7. Caso práctico: regresión lineal simple	112

13.8. Referencias	115
13.9. Ejercicios	115
13.10TODO	115

1. Eventos y probabilidades

Según el DRAE, una contingencia es una *cosa que puede suceder o no suceder*. Nuestros **eventos** serán conjuntos de contingencias a los que luego asociaremos probabilidades de ocurrencia. Ocurrirá un evento cuando suceda alguno de las contingencias que lo integran.

Como conjuntos que son, podemos operar con eventos usando los operadores habituales de la teoría de conjuntos. Así, por ejemplo, $A \cup B$ es el evento consistente en que ocurra el evento A o el evento B ; y $A \cap B$, que ocurran ambos. Podremos usar los eventos especiales Ω y \emptyset , que corresponden a los eventos totales (que ocurren cualquier cosa) o que no ocurra ninguno. Que no ocurra el evento A (el **complementario** de A) se puede representar como $\Omega \setminus A$, aunque también se utiliza la notación \bar{A} .

Casi siempre consideraremos eventos generados por **variables aleatorias**. Las variables aleatorias son funciones (que por razones históricas suelen denotarse con letras mayúsculas, X , Y , etc.) que toman valores numéricos, ya sean discretos o continuos. X puede representar, por ejemplo, la suma de puntos en una tirada de cuatro dados; en tal caso, p.e., de los 6^4 posibles resultados, solo en 4 de ellos $X = 5$. O el número de caras al tirar una moneda 5 veces. En contextos más propiamente de ciencia de datos, indicar si un cliente se fuga o no. Como ejemplo de variable aleatoria continua, X podría ser la altura de una persona (elegida al azar) y nos podría interesar el evento en el que $X < 170$.

La **probabilidad** es una función que asigna a conjuntos (eventos) un número entre 0 y 1. Matemáticamente, para cada evento A ,

$$P(A) = x, \quad 0 \leq x \leq 1.$$

La función P está sujeta a ciertas reglas que fueron axiomatizadas por Kolmogorov en los años 30. Los axiomas de Kolmogorov son

- $P(A) \geq 0$
- $P(\Omega) = 1$
- $P(\cup A_i) = \sum_i P(A_i)$ si los A_i son eventos *mutuamente excluyentes*.

De los axiomas de probabilidad se deducen propiedades *razonables* de la probabilidad; p.e., si $A \subset B$, entonces $P(A) \leq P(B)$.

Reglas como las anteriores parecen obvias. Sin embargo, lo son mucho menos de lo que parecen: las personas tendemos a equivocarnos al operar con probabilidades en algunos contextos, como pusieron de manifiesto A. Tversky y D. Kahneman al plantear, junto con otros, el siguiente problema a un panel de voluntarios:

Linda is thirty-one years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations... The respondents are asked to rank in order of likelihood various scenarios: Linda is

1. an elementary school teacher,
2. active in the feminist movement,
3. a bank teller,
4. an insurance salesperson, or
5. a bank teller also active in the feminist movement.

Un porcentaje grande de ellos (y el experimento se ha repetido posteriormente con el mismo resultado) consideran más probable la opción 5 que la 3, a pesar de que el primer evento está contenido en el segundo.

Desde el punto de vista de la teoría de la probabilidad, las probabilidades asociadas a los eventos son, o bien conocidas, o se pueden calcular deductivamente aplicando determinadas reglas. En estadística, sin embargo, observaremos fenómenos aleatorios y nuestro problema consistirá en, inductivamente, tratar de esclarecer el mecanismo aleatorio subyacente.

1.1. Referencias

- Axiomas de probabilidad de Kolmogorov ([enlace](#))
- Gnedenko, B. *Teoría de las probabilidades*, Capítulo 1, que contiene una introducción accesible e ilustrada con ejemplos del material de esta sección.
- La falacia de la conjunción (el caso de Linda) ([enlace](#))
- H. Hoffrage et al., *Representation facilitates reasoning: what natural frequencies are and what they are not*, Cognition 84 (2002) 343–352 ([enlace](#))

1.2. Ejercicios

1.2.0.1. Ejercicio

Trata de comprender que $\bar{A} \cap \bar{B} = \overline{A \cup B}$

1.2.0.2. Ejercicio

Propón un ejemplo en el que se cumplan las siguientes condiciones (que no se cumplen en general):

- $P(A \cup B) = P(A)$
- $P(A \cup B) = P(A) + P(B)$
- $\max(P(A), P(B)) < P(A \cup B) < P(A) + P(B)$

1.2.0.3. Ejercicio

Prueba que $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Trata de obtener la expresión correspondiente para tres eventos y para n eventos.

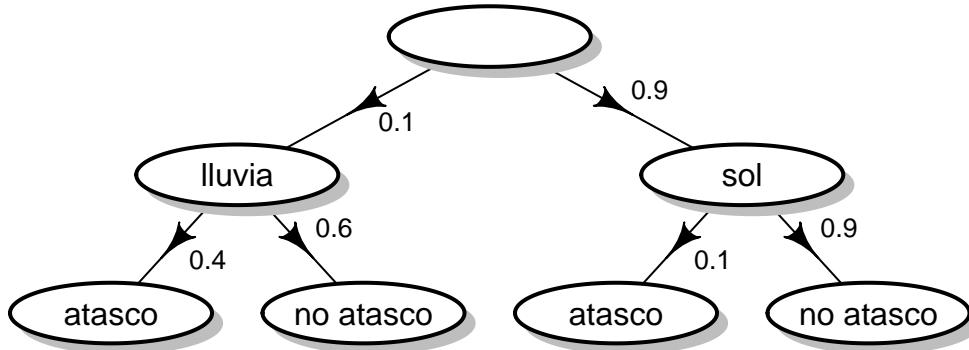
2. Probabilidades conjuntas, marginales y condicionales, teorema de Bayes e independencia

Como motivación para esta sección, vamos a describir un modelo probabilístico:

- En un determinado lugar, el 10 % de los días llueve (y el 90 % luce el sol).
- Cuando llueve, la probabilidad de que se produzcan atascos es del 40 %.
- Cuando no llueve, la probabilidad de que haya atascos es del 10 %.

2.1. Probabilidad conjunta, marginal y condicional

Podemos realizar una representación gráfica del modelo:



La probabilidad de que un día llueva y no haya atasco es $0,06 = 0,1 \times 0,6$: el 10 % de los días llueve y de ellos, el 60 % no hay atascos. Análogamente, se pueden calcular el resto de las combinaciones posibles y tabularlas así:

tiempo	atasco	prob
sol	sí	0.09
sol	no	0.81
lluvia	sí	0.04
lluvia	no	0.06

Esta tabla representa la llamada *probabilidad conjunta* (de las variables tiempo y atasco). Nótese cómo la suma de las probabilidades (la columna *prob*) es 1. La hemos construido a partir de las llamadas **probabilidades marginales** (del tiempo) y **condicionales** (de los atascos condicionados por el tiempo).

En efecto, expresiones del tipo *el 10 % de los días llueve* reciben el nombre de probabilidades marginales (por un motivo que luego quedará claro) y las del tipo *cuando llueve, la probabilidad de que haya atascos es del 40 %*, probabilidades condicionales, que se suelen representar así:

$$P(\text{atasco} \mid \text{lluvia}) = 0,4$$

La tabla de probabilidades conjuntas también puede representarse de la forma equivalente

P(T,A)	atasco	no atasco
lluvia	0.04	0.06
sol	0.09	0.81

En ella, si se suman filas y columnas, se obtienen las probabilidades marginales

P(T,A)	atasco	no atasco	P(T)
lluvia	0.04	0.06	0.1
sol	0.09	0.81	0.9
P(A)	0.13	0.86	1

En este caso, solo consideramos dos variables, T y A . En otras situaciones puede haber más variables y la probabilidad conjunta se representaría por medio de un (hiper)cubo de números (que seguirían sumando 1). En tales casos, las probabilidades marginales corresponderían a las sumas a lo largo de las distintas proyecciones. Por ejemplo, de tenerse las variables A , B y C , la probabilidad marginal $P(A, B)$ se calcularía a partir de la conjunta así:

$$P(A, B) = \sum_i P(A, B, C = c_i).$$

Los valores de la tabla de probabilidades conjuntas, $P(T, A)$, se han calculado así:

$$P(T, A) = P(A | T)P(T).$$

Despejando, se obtiene la relación

$$P(A | T) = \frac{P(T, A)}{P(T)}$$

que se usa a menudo como definición de la probabilidad condicional. La probabilidad condicional no es sino una colección de probabilidades (en nuestro caso, $P(A | T)$, una para cada condición climatológica). Lo que dice la fórmula anterior es que la probabilidad $P(A | \text{lluvia})$ se construye tomando la columna correspondiente de la tabla de probabilidades conjuntas y, como sus valores no suman la unidad, normalizando por su total, que es precisamente $P(\text{lluvia})$.

Las probabilidades condicionales son esenciales en ciencia de datos. Los modelos que se construyen en ciencia de datos están basados precisamente en $P(Y | X_1, \dots, X_n)$, es decir, la probabilidad condicional de la variable objetivo Y en función de las variables predictoras X_1, \dots, X_n .

2.2. Teorema de Bayes

El modelo plantado en esta sección partía del tiempo y, en función de él, definía la probabilidad de atasco. Pero es posible preguntar cuál es la probabilidad de que esté lloviendo a partir de la información de que hay un atasco. Es decir, calcular $P(\text{lluvia} | \text{ataasco})$ (y, en general, $P(T | A)$), i.e., *invertir* las probabilidades condicionales.

Es sencillo porque, por definición,

$$P(T | A) = \frac{P(T, A)}{P(A)}$$

y tanto $P(T, A)$ como $P(A)$ son conocidos. No obstante, podemos obviar la referencia a la probabilidad conjunta desarrollando el numerador $P(T, A)$ de la forma $P(T, A) = P(A | T)P(T)$, con lo que

$$P(T | A) = \frac{P(A | T)P(T)}{P(A)}$$

La expresión anterior se conoce como **teorema de Bayes**.

En nuestro problema,

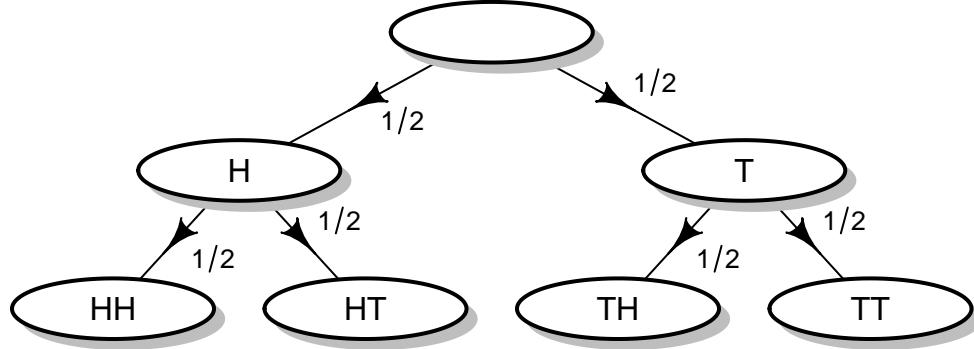
$$P(\text{lluvia} | \text{ataasco}) = \frac{P(\text{ataasco} | \text{lluvia})P(\text{lluvia})}{P(\text{ataasco})} = \frac{0,4 \times 0,1}{0,13} \approx 0,31$$

En general, llueve el 10 % de los días; no obstante, con la información adicional de que hay un atasco, la probabilidad estimada de lluvia crece hasta el 31 %.

2.3. Independencia

En el ejemplo discutido en la sección anterior, por definición, el tráfico dependía del tiempo. Había una correlación (y puede que relación causal) entre el tiempo y el estado del tráfico. En muchas circunstancias, sin embargo, encontramos sucesos independientes: en ellas, el valor de uno no influye en el del otro. Un ejemplo clásico es el de los lanzamientos sucesivos de una moneda.

Si tiramos una moneda al aire, podemos suponer que $P(H) = P(T) = 1/2$. Pero, ¿qué ocurre si la lanzamos al aire dos veces? Gráficamente,



En el ejemplo original, como el tiempo afecta al tráfico,

$$P(\text{atasco} \mid \text{lluvia}) \neq P(\text{atasco} \mid \text{sol})$$

Pero en los lanzamientos de monedas,

$$P(H_2 \mid T_1) = P(H_2 \mid H_1).$$

Es decir, la segunda tirada no está afectada por la primera y, de hecho,

$$P(H_2 \mid H_1) = P(H_2 \mid T_1) = P(H) = 1/2$$

En general, si A y B son **independientes**, relación que frecuentemente se denota mediante $A \perp B$, entonces $P(A|B) = P(A)$. Qualitativamente, conocer B no aporta información sobre A .

Puede haber muchas relaciones probabilísticas entre variables A y B . Una de ellas es la de que sean independientes. Es la relación más sencilla. Pero, desde el punto de vista de la ciencia de datos, la menos interesante de todas: si queremos predecir una variable Y a partir de otras variables X_1, \dots, X_n , nos interesa precisamente que las X_i no sean independientes de Y .

Si buscamos información acerca de un fenómeno Y , buscaremos otros X_1, \dots, X_n relacionados (i.e., no independientes) con Y .

La hipótesis de independencia, sin embargo, permite simplificar muchos modelos. Es gracias a la *sparsity* (*raleza*, de *ralo*, suena raro) que aportan las relaciones de independencia en las **redes bayesianas** o las **cadenas de Markov** que los modelos resultantes son tratables.

Otro ámbito en que la independencia juega un papel en la ciencia de datos tiene que ver con la relación entre las distintas observaciones y_i . Frecuentemente, son independientes entre sí; sin embargo, en muchos contextos, y_i no es independiente de otras observaciones y_j , por ejemplo:

- Si las y_i son observaciones ordenadas en el tiempo (p.e., temperaturas horarias)
- Si las y_i se refieren a ubicaciones (se correlacionan con otras y_j próximas)
- Si bloques de y_i se refieren a un mismo cliente, paciente o sujeto.

En esos casos, ignorar las relaciones de dependencia y asumir *de facto* la independencia de las observaciones conduce a modelos menos eficaces. El *quid* de la cuestión reside en que asumir independencia cuando no la hay equivale a descartar información potencialmente útil.

Si A y B son independientes, su probabilidad conjunta es

$P(A, B)$	b0	b1	$P(A)$
a0	$p_1 * p_3$	$p_1 * p_4$	p_1
a1	$p_2 * p_3$	$p_2 * p_4$	p_2
P(B)	p_3	p_4	1

Precisamente porque

$$P(A, B) = P(A | B)P(B) = P(A)P(B).$$

De hecho, a menudo, se definen como independientes las variables aleatorias cuando

$$P(A, B) = P(A)P(B).$$

Nosotros lo consideraremos una consecuencia de nuestra definición más que la definición en sí. En cualquier caso, la fórmula anterior proporciona un criterio rápido para determinar si dos variables aleatorias son o no independientes.

2.4. Independencia y regla de la cadena

En el ejemplo considerado en esta sección hemos estudiado una situación en la que se consideraban dos variables, T y A y hemos factorizado

$$P(T, A) = P(A | T)P(T)$$

En general, si tenemos A_1, A_2, A_3, A_4 , podemos factorizar

$$P(A_1, A_2, A_3, A_4) = P(A_4 | A_1, A_2, A_3)P(A_3 | A_1, A_2)P(A_2 | A_1)P(A_1)$$

Esto se puede demostrar dibujando el gráfico correspondiente: se comienza con A_1 , se hace depender A_2 de A_1 y así sucesivamente. Obviamente, la factorización tiene sentido cuando se conocen las probabilidades condicionales implicadas. Ocurre además con frecuencia que se puede suponer, por ejemplo, que A_4 no depende de A_1 y A_2 , por lo que la expresión potencialmente compleja $P(A_4 | A_1, A_2, A_3)$ se reduce a la más manejable $P(A_4 | A_3)$.

En un caso aún más extremo, si las variables aleatorias son independientes, entonces la expresión anterior se reduce a

$$P(A_1, A_2, A_3, A_4) = P(A_1)P(A_2)P(A_3)P(A_4)$$

Así, por ejemplo, en el lanzamiento de monedas,

$$P(HHTHT) = \frac{1}{2^5}.$$

2.5. Independencia condicional

Consideremos el siguiente **paseo aleatorio** X_t : una partícula, que en $t = 0$ está en la posición 0 se mueve aleatoriamente cada segundo dando saltos equiprobables de tamaño $-1, 0$ o 1 . Entonces, X_T depende de X_t si $t < T$, pero conocido X_{T-1} , X_T no depende de por dónde haya pasado previamente. Es decir, X_T no es independiente de X_t , pero ambas variables aleatorias son **condicionalmente independientes** dado X_{T-1} . La independencia condicional se representa así:

$$X_T \perp X_t \mid X_{T-1} \text{ si } t < T - 1.$$

En tales casos, la función de probabilidad conjunta puede factorizarse así:

$$P(X_0, X_1, \dots, X_T) = P(X_0) \prod_{i=1}^T P(X_i \mid X_{i-1}).$$

Es una expresión un poco más compleja que la correspondiente a eventos independientes pero mucho más simple que la que se deduce de la regla de la cadena general. Precisamente porque se explotan las relaciones de independencia condicional.

El anterior es un ejemplo simple de las **cadenas de Markov**, que son secuencias de variables aleatorias en las que cada una de ellas es condicionalmente independiente de todas las que la preceden no inmediatamente condicionalmente en, precisamente, su predecesora inmediata.

2.6. Apéndice: SamIam

SamIam es una herramienta para modelar y razonar sobre redes bayesianas desarrollado en la Universidad de California en Los Ángeles (UCLA). Permite representar relaciones de dependencia entre variables aleatorias a través de una interfaz gráfica y, luego, realizar cálculos sobre ellas.

SamIam se puede descargar libremente y los tutoriales en línea ilustran con detalle las dos funciones que, de momento, nos interesan:

- Cómo crear variables aleatorias y especificar las probabilidades marginales y condicionales que definen su estructura probabilística.
- Cómo realizar inferencias sobre la red (p.e., marcar un nodo que representa la variable aleatoria que indica el color de una bola extraída de una urna como blanco para ver cómo varían el resto de las probabilidades).

2.7. Referencias

- Gnedenko, B., *Teoría de las probabilidades*, Capítulo 1
- Gallier, J., *An Introduction to Discrete Probability*
- Darwiche, A., *SamIam: Sensitivity Analysis, Modelling, Inference and More*

2.8. Ejercicios

2.8.0.1. Ejercicio

En una empresa de seguros los clientes son hombres (60 %) y mujeres (40 %). Tienen coches de color rojo, gris u otros. Para los hombres, el porcentaje de coches grises y de otros colores es igual, pero el de coches rojos es el doble que los anteriores. Para las mujeres, sucede lo mismo, solo que el porcentaje de coches rojos es la mitad que los otros.

La tasa de siniestros (si un cliente tuvo un siniestros en un año dado) es del 10% para hombres y mujeres independientemente del color del coche con las siguientes excepciones:

- Es el doble para los hombres que conducen coches rojos.
- Es la mitad para las mujeres que conducen coches grises.

Dibuja la gráfica que describe las probabilidades anteriores.

2.8.0.2. Ejercicio

Construye la tabla de probabilidad conjunta asociada a las variables aleatorias descritas en el ejercicio anterior.

2.8.0.3. Ejercicio

- Calcula la probabilidad marginal de los colores de los coches.
- Calcula la probabilidad marginal de colores y siniestralidad (una tabla que contenga la probabilidad de que un coche de un determinado color tenga o no un accidente).

2.8.0.4. Ejercicio

Calcula la probabilidad de siniestro según el color del vehículo.

2.8.0.5. Ejercicio

De ocurrir un siniestro, calcula la probabilidad de que la afectada sea mujer.

2.8.0.6. Ejercicio

Hay tres urnas que contienen, respectivamente, 2, 3 y 5 bolas blancas y 2, 4 y 1 bolas negras. Alguien elige una urna al azar y extrae una bola, que resulta ser blanca. ¿Cuál es la probabilidad de que la urna de la que se ha extraído la bola sea la primera?

2.8.0.7. Ejercicio

Recalcula la probabilidad si se extrae una bola más y resulta ser negra.

2.8.0.8. Ejercicio

Prueba que si $A \perp B$, entonces $A \perp \bar{B}$ y $\bar{A} \perp \bar{B}$.

2.8.0.9. Ejercicio

Calcula la probabilidad de obtener 3 caras en 4 lanzamientos de monedas.

2.8.0.10. Ejercicio

- ¿Cuál es la probabilidad de sacar un 2 o un 6 tirando un dado? (Usa los axiomas de probabilidad)
- ¿Cuál es la probabilidad de sumar 7 puntos en dos tiradas de dados?
- ¿Cuál es la probabilidad de no sacar un 1 tirando un dado?
- ¿Cuál es la probabilidad de no sacar ninguno 1 después de tirar n dados? (Usa la independencia)

2.8.0.11. Ejercicio (problema del caballero de Méré)

¿Qué es más probable, sacar un as tirando cuatro dados (una vez), o sacar dos ases en alguna de 24 tiradas de dos dados?

2.8.0.12. Ejercicio

Modela el ejercicio anterior de las de las urnas usando SamIam y resuélvelo con él.

2.8.0.13. Ejercicio

Usa SamIam para modelar el problema de la empresa de seguros y úsalo para resolver los ejercicios relacionados con el teorema de Bayes planteados más arriba.

3. Distribuciones de probabilidad

Los axiomas de probabilidad mostrados más arriba son para las probabilidades lo mismo que la *gramática* para el lenguaje: establece las reglas con las que interactúan las probabilidades de la misma manera que la gramática fija las reglas con las que interactúan las palabras en el discurso. Las distribuciones, por otra parte, equivaldrían al léxico, lo que da significado a las construcciones. Porque no es suficiente con saber que si $A \subset B$, entonces $P(A) \leq P(B)$; frecuentemente queremos saber cuánto valen $P(A)$ y $P(B)$.

Por eso vamos a discutir la manera en que podemos asignar probabilidades a eventos.

Existe una escuela de pensamiento que defiende las llamadas **probabilidades subjetivas**. De acuerdo con L. J. Savage, la probabilidad de un evento sería:

La máxima cantidad de dinero que apostarías si alguien te ofreciese un euro si el evento en cuestión sucede.

Si estuvieses dispuesto a poner como máximo 20 céntimos por tener derecho a recibir un euro mañana si lloviese, estarías estimando la probabilidad de lluvia en un 20 %.

Se puede probar que si los agentes son racionales, las probabilidades asignadas de acuerdo con esta regla cumplen los axiomas de Kolmogorov.

La probabilidad subjetiva es la base de la construcción de probabilidades a partir de información obtenida a partir de opiniones (p.e., las expresadas por paneles de expertos) o del consenso de los jugadores en las casas de apuestas. Se usa fundamentalmente para estimar la probabilidad de fenómenos únicos o inhabituales, como el resultado de elecciones o partidos de fútbol, la posibilidad de accidentes nucleares graves, etc.

En ocasiones no hace falta recurrir a expertos: existen motivos *epistemológicos* para asignar determinadas probabilidades a determinados tipos de sucesos. Por ejemplo, el llamado **principio de indiferencia** dice que si n alternativas son indistinguibles entre sí excepto por su nombre, cada una de ellas tiene una probabilidad de $1/n$. Este tipo de razonamiento puede aplicarse a lanzamiento de monedas (¿hay algún motivo para que no se cumpla que $P(H) = P(T) = 1/2?$), de dados, etc.

Otros razonamientos pueden llevar a considerar como razonable que determinadas variables aleatorias sigan, por ejemplo, una distribución normal o exponencial.

Además de las anteriores, existen maneras de inferir las probabilidades asociadas a eventos basadas en datos. Particularmente cuando existe un registro histórico de ocurrencias de este tipo de eventos y muestran ciertas regularidades como, por ejemplo:

- La proporción de niñas entre los recién nacidos.
- La tasa de suicidios.
- La proporción de personas de 80 años que fallecen antes de cumplir los 81.
- La proporción de caras que se obtienen al lanzar al aire repetidamente una moneda.

Este tipo de probabilidades son las que más a menudo se explotan en ciencia de datos. Están además relacionadas con un resultado famoso de la teoría de la probabilidad conocido como la **ley de los grandes números**.

Podría argüirse que un caso particular de este último procedimiento de asignación de probabilidades justifica el uso de las **distribuciones de probabilidad** propiamente dichas. En efecto, se ha observado que fenómenos aleatorios procedentes de ámbitos completamente diferentes siguen *patrones* similares. Muchos de estos patrones están estudiados y descritos y pueden servir de plantilla para describir determinados fenómenos aleatorios.

Hay que tener cuidado, en todo caso, en no atribuir a las distribuciones de probabilidad *con nombre* (las que aparecen en los libros) propiedades universales y dar por hecho que *todo* fenómeno aleatorio puede describirse con alguna de ellas. Lo más típico es que alguna de ellas (o alguna variante de ellas) pueda usarse como aproximación suficientemente buena para algún fin.

3.1. Referencias

- Principio de indiferencia
- La utilidad esperada subjetiva
- La ley de los grandes números
- Gnedenko, B. *Teoría de las probabilidades*, Capítulo 6, que contiene una discusión (y demostraciones) tanto de la ley débil como de la fuerte de los grandes números.
- Seneta, E., *A Tricentenary history of the Law of Large Numbers*

4. Distribuciones de probabilidad discretas

Se aplican a situaciones donde las variables aleatorias toman valores discretos, como por ejemplo los valores 0 y 1, las letras del abecedario, determinados colores, o los números 0, 1, Ejemplos de ellas son:

- el número de caras en 100 tiradas de una moneda,
- el número de siniestros mensual en una compañía de seguros o
- el número de apariciones de las palabras *viagra* u *oferta* en un correo electrónico.

En esta sección vamos a presentar una serie de distribuciones de probabilidad *de libro* que pueden resultar útiles para modelar fenómenos aleatorios discretos. No hay que olvidar, sin embargo, que un determinado conjunto de datos no tiene por qué seguir alguna de las que se discutirán aquí.

Además, junto con la presentación de una serie de distribuciones discretas y apoyándonos en ellas, iremos introduciendo conceptos estadísticos importantes, como la media, la desviación estándar, etc.

4.1. Distribución de Dirac

La distribución de Dirac puede considerarse *degenerada*: toma siempre (con probabilidad 1) un valor fijo a . Una variable aleatoria de Dirac, por lo tanto, no es aleatoria. Aunque parezca contraintuitivo, la distribución de Dirac tiene su importancia.

4.2. Distribución de Bernoulli

La distribución de Bernoulli es muy simple: es la de una moneda con probabilidad $P(H) = p$ de cara. Una variable aleatoria de Bernoulli toma valores 0 o 1 (que frecuentemente se usan para codificar otros tales como cara o cruz, éxito o fracaso, etc.). Si $X \sim \text{Bernoulli}(p)$, entonces

$$X = \begin{cases} 1 & \text{con probabilidad } p \\ 0 & \text{con probabilidad } 1-p \end{cases}$$

Puede considerarse una *mezcla* (un concepto que se introducirá más tarde) de dos distribuciones de Dirac. A la inversa, la distribución de Dirac es un caso degenerado de la de Bernoulli que ocurre cuando $p = 0$ o $p = 1$.

A pesar de su aparente trivialidad, la distribución de Bernoulli es la base de muchos modelos de clasificación. De hecho, una de las tareas más habituales de la ciencia de datos es encontrar el valor p_i asociado a una determinada acción (de resultado binario) de un sujeto i .

En R podemos obtener una muestra de la distribución de Bernoulli mediante

```
p <- 0.7
rbinom(1, 1, p)
```

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

o un conjunto de ellas haciendo

```
rbinom(10, 1, p)
```

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

4.3. La media de una variable aleatoria

La media de una variable aleatoria discreta X que toma valores x_i con probabilidad $P(X = x_i)$ es

$$E(X) = \mu(X) = \bar{X} = \sum_i x_i P(X = x_i).$$

Como ejemplo de la fórmula anterior, podemos calcular la media de una variable aleatoria de Bernoulli $X \sim \text{Bernoulli}(p)$, que es

$$E(X) = 0 \times (1 - p) + 1 \times p = p.$$

La media es una medida de **centralidad**: es un valor alrededor del cual podría decirse que pivota la distribución. No es, sin embargo, un valor típico: pudiera ser, incluso, un valor imposible para la distribución. Eso ocurre precisamente con la distribución de Bernoulli: su media es p , un valor típicamente estrictamente comprendido entre 0 y 1, mientras la variable solo puede tomar los valores 0 o 1.

La media de una variable aleatoria también puede *estimarse* mediante simulaciones. En situaciones triviales no merece la pena; en otros casos, sí. Por ejemplo, para una variable aleatoria de Bernoulli,

```
set.seed(1234)
p <- 0.7
mean(rbinom(1e6, 1, p))
```

```
## [1] 0.699833
```

En el ejemplo anterior hemos estimado p mediante un millón de simulaciones con R. Sirve también para ilustrar cómo a través de repeticiones de un experimento podemos llegar a estimar la probabilidad de un evento, la base de la aproximación frecuentista a la asignación de probabilidades discutida anteriormente: si desconocemos la probabilidad de cara de una moneda, podemos tirarla al aire muchas veces y dividir el número de éxitos entre el de ensayos. Probablemente, el conciente obtenido estará próximo a la probabilidad buscada.

Si tenemos dos variables aleatorias, entonces $E(X + Y) = E(X) + E(Y)$. Intuitivamente, si la fábrica A fabrica, en promedio, 100 unidades y la B 120, en promedio, conjuntamente, deberían fabricar 220. No obstante, es ilustrativo proporcionar una demostración matemática:

$$E(X+Y) = \sum_{ij} (x_i + y_j) P(X = x_i, Y = y_j) = \sum_{ij} x_i P(X = x_i, Y = y_j) + \sum_{ij} y_j P(X = x_i, Y = y_j) = \sum_i x_i P(X = x_i) + \sum_j y_j P(Y = y_j)$$

En la expresión anterior se ha usado la marginalización de la distribución conjunta $P(X = x_i, Y = y_j)$. En efecto,

$$\sum_{ij} x_i P(X = x_i, Y = y_j) = \sum_i x_i \sum_j P(X = x_i, Y = y_j) = \sum_i x_i P(X = x_i).$$

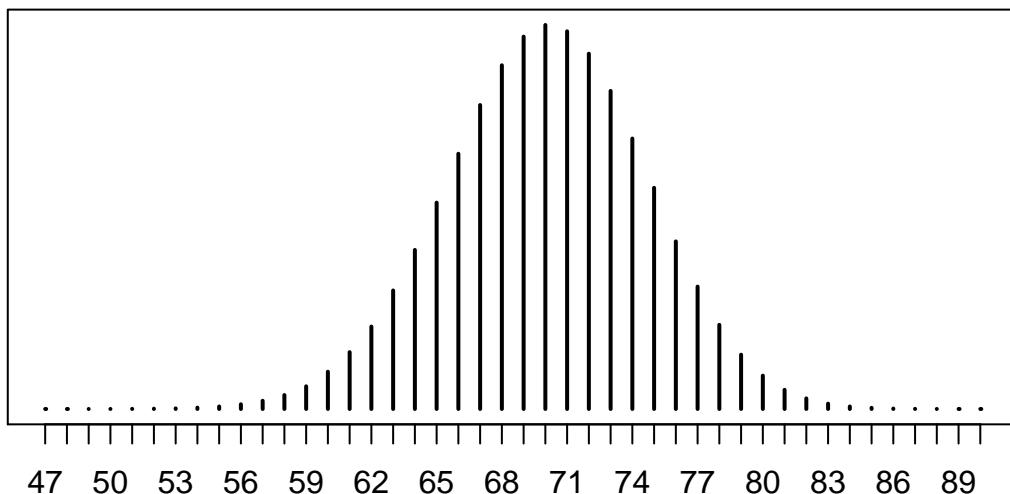
4.4. Distribución binomial

La distribución binomial es la de una suma de variables aleatorias de Bernoulli independientes. Permite modelar problemas como el número de caras que se obtienen después de tirar una moneda 15 veces. O el número de clientes que abandonarán la empresa si la tasa de fuga es del 12 %.

Por ser una suma de variables aleatorias de Bernoulli, podemos deducir su media: es np . Pero, además de su media, interesa saber cómo se distribuyen los valores alrededor de ese valor.

La función de probabilidad tiene una típica forma de campana (es *unimodal*) y es ligeramente asimétrica (es simétrica solo cuando $p = 0,5$):

Distribución binomial ($n = 100$, $p = 0.7$)



En realidad, la función toma valores (en nuestro ejemplo) para cada uno de los enteros entre 0 y 100. Sin embargo, casi toda la probabilidad está concentrada en un entorno de la media, 70.

4.5. Varianza y desviación estándar

La gráfica de la sección anterior muestra cómo se distribuye la probabilidad de la distribución binomial alrededor de su media. Nos da una idea de la **dispersión** de la distribución, es decir, si los valores esperados de la distribución están cerca o lejos de su media.

La varianza es un indicador tradicionalmente usado para cuantificar la dispersión. Tanto que en ocasiones se utiliza el término varianza cuando, en realidad, se quiere decir dispersión. La varianza se define así:

$$\sigma^2(X) = \text{Var}(X) = \sum_i (x_i - E(X))^2 P(X = x_i)$$

Se trata, pues, de $E[(X - E(X))^2]$, el promedio de las distancias al cuadrado de cada punto con la media. Lo cual puede ser problemático en algunas ocasiones: por ejemplo, cuando un punto está excesivamente alejado de la media, la varianza puede ser muy alta incluso cuando el resto de los valores no muestran gran dispersión. Por eso, a veces, en lugar de la varianza se usa la medida alternativa de la dispersión $E(|X - E(X)|)$.

Si X e Y son independientes, entonces

$$\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y).$$

Pero, en general, lo anterior no es cierto. Por ejemplo,

$$\sigma^2(X + X) = \sigma^2(2X) = 4\sigma^2(X)$$

No obstante, como aplicación de lo anterior, si $X \sim \text{Binom}(n, p)$, entonces $\sigma^2(X) = np(1 - p)$ porque $\sigma^2(Y) = p(1 - p)$ si $Y \sim \text{Bernoulli}(p)$.

La **desviación estándar** de X , $\sigma(X)$, es la raíz cuadrada de la varianza de X . Tiene como ventaja que está en las mismas unidades que X . Si $X > 0$, tiene sentido hablar del **coeficiente de variación**, $\frac{\sigma(X)}{E(X)}$, que compara la media con la desviación alrededor de la media.

4.6. La distribución multinomial

La distribución multinomial es una extensión de la distribución binomial que se aplica a situaciones en que la variable aleatoria X puede tomar más de dos valores. En particular, cuando existen

- n etiquetas ($n > 2$) y
- probabilidades p_1, \dots, p_n (por supuesto, tales que $\sum_i p_i = 1$) asociadas a ellas.

La distribución multinomial tiene muchas aplicaciones en ciencia de datos. Por ejemplo, para modelar los tipos de productos que comprará un cliente o las palabras que aparecerán en un texto en función de su asunto.

Podemos muestrear la distribución multinomial así:

```
rmultinom(10, 4, c(0.1, 0.4, 0.5))
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]     0    0    0    1    0    1    1    1    0    0
## [2,]     2    4    1    2    1    1    2    1    2    2
## [3,]     2    0    3    1    3    2    1    2    2    2
```

El resultado de la simulación son 10 vectores (columnas) de cuatro extracciones de tres elementos (indexados por las tres probabilidades, 0,1, 0,4 y 0,5). El que una columna sea, p.e., (0, 2, 2) significa que en el experimento se han obtenido dos elementos de la segunda etiqueta y otros dos de la tercera.

Por su parte,

```
rmultinom(10, 3, rep(1, 6))
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]     1    0    1    0    1    1    2    0    0    0
## [2,]     0    0    0    0    0    0    0    1    1    0
## [3,]     0    1    0    0    0    0    1    1    1    1
```

```
## [4,] 2 0 1 2 1 0 0 0 0 0
## [5,] 0 2 1 0 1 2 0 1 0 1
## [6,] 0 0 0 1 0 0 0 0 1 1
```

muestra 10 tiradas de 3 dados: cada columna cuenta el número de unos, doses, etc. obtenidos.

La función `dmultinom` permite comprobar lo atípica que es una determinada configuración. Por ejemplo,

```
dmultinom(c(0, 2, 2), prob = c(0.1, 0.4, 0.5))
```

```
## [1] 0.24
```

indica que la probabilidad de la configuración (0, 2, 2) es mucho mayor que la de (2, 1, 1):

```
dmultinom(c(2, 1, 1), prob = c(0.1, 0.4, 0.5))
```

```
## [1] 0.024
```

4.7. La distribución de Poisson

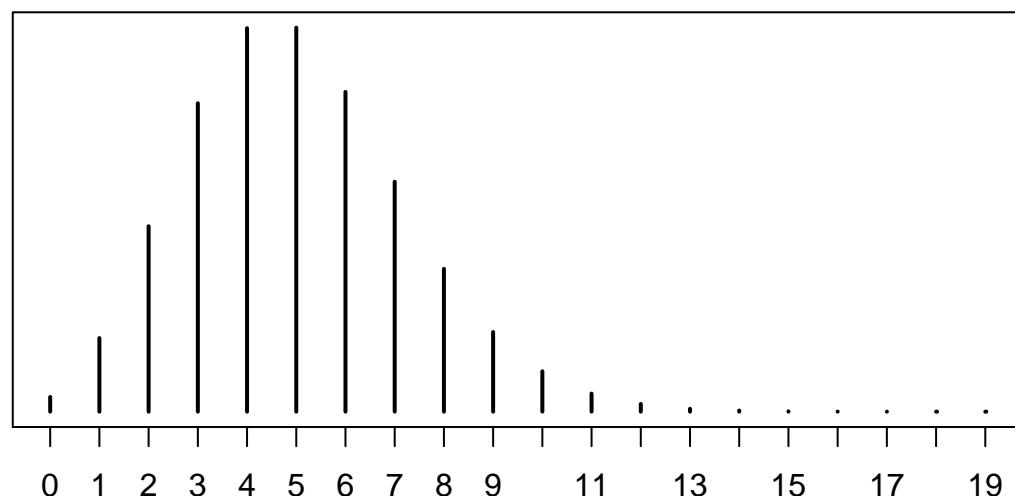
Hay situaciones en que interesa modelar conteos, como por ejemplo, el número de enfermos de hepatitis en una provincia y periodo, el número de llamadas entrantes a una centralita o el número de compras que realiza un cliente en un mes.

Dos distribuciones que pueden ser útiles para este tipo de problemas son la de Poisson (la más habitual) y la binomial negativa. La distribución de Poisson admite un parámetro, λ , que es a su vez la media y la varianza. Si $X \sim \text{Pois}(\lambda)$, entonces

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Gráficamente,

Distribución de Poisson (lambda = 5)



Ese es el perfil típico de la distribución de Poisson: acampanado y con una cola larga hacia la derecha. Porque, recuérdese, que teóricamente podría tener cualquier valor entero mayor que cero: a diferencia de la distribución binomial, no tiene límite superior.

La distribución binomial negativa tiene un perfil similar y los motivos por los que es preferible usar una u otra exceden el alcance de estas páginas.

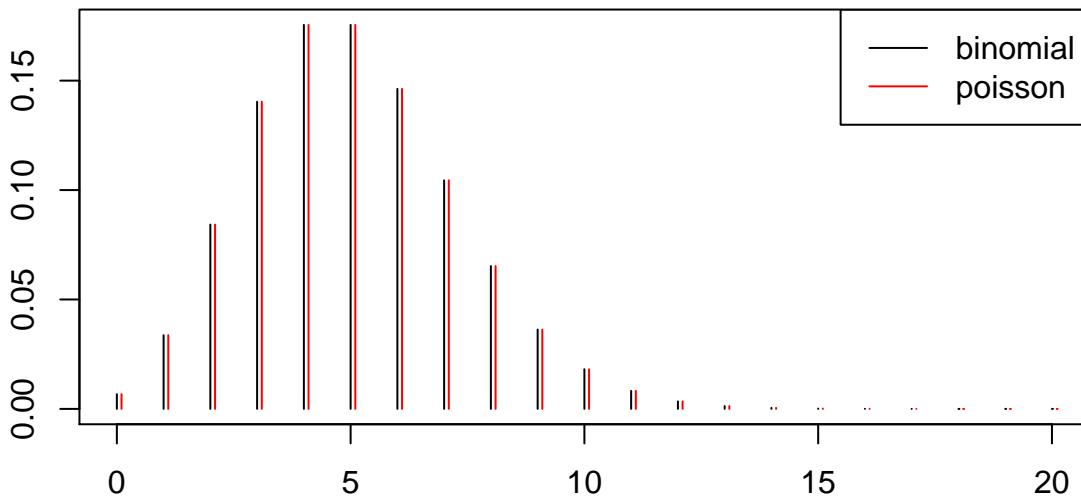
Una propiedad muy útil de la distribución de Poisson es que si $X_1 \sim \text{Pois}(\lambda_1)$ y $X_2 \sim \text{Pois}(\lambda_2)$, entonces

$$X_1 + X_2 \sim \text{Pois}(\lambda_1 + \lambda_2).$$

Dicho de otra manera, si el número de llamadas en una semana es Poisson de parámetro λ , el número de llamadas en dos semanas es también Poisson y con parámetro 2λ .

Existe una relación importante entre las distribuciones de binomial y de Poisson: una distribución binomial con n grande y p pequeño es aproximadamente Poisson de parámetro np , como ilustra el siguiente gráfico, construido con $n = 10000$ y $p = 5/n$:

Binomial vs Poisson



Por eso, si en un país de un millón de habitantes, si la tasa de suicidios es de 13 por cada 100000 habitantes, el número de suicidios puede modelarse con cualquiera de los dos modelos prácticamente equivalentes:

- Con una distribución binomial con parámetros $p = 13/100k$ y $n = 1M$.
- Con una distribución de Poisson con parámetro $\lambda = 130$.

De hecho, se tiende a preferir el segundo.

4.8. Urnas y muestras

Existen problemas de probabilidad discreta que pueden describirse (y, de hecho, tienden a describirse) en términos de extracciones al azar de bolas de colores de urnas. Algunos de ellos tienen aplicaciones *reales*, como el tipo de productos que comparará un cliente o la clasificación de las llamadas recibidas en un *call center*.

Muchos de ellos pueden modelarse utilizando alguna de las distribuciones contempladas más arriba, como la binomial o la multinomial. Sobre todo cuando existe *reemplazo*, i.e., las bolas son devueltas a la urna. En caso de reemplazo, las distintas extracciones son independientes entre sí. Sin embargo, de no haber reemplazo, tiradas sucesivas son dependientes: de extraerse una bola negra, en la siguiente tirada decrece la probabilidad de extraer otra bola negra (porque quedan menos).

En tales casos, se pueden simular los fenómenos aleatorios usando, por ejemplo, la función `sample` de R.

4.9. Referencias

- Las páginas de la Wikipedia (en inglés) correspondientes a las distribuciones mencionadas en la sección.

4.10. Ejercicios

4.10.0.1. Ejercicio

Demuestra que si $Y \sim \text{Bernoulli}(p)$, entonces $\sigma^2(Y) = p(1 - p)$.

4.10.0.2. Ejercicio

¿Cómo probarías que en R `sum(rbinom(15, 1, 0.5))` muestrea la misma variable aleatoria que `rbinom(1, 15, 0.5)`.

4.10.0.3. Ejercicio

Simula y representa gráficamente una **caminata aleatoria** simple. Comienza con el valor 0 en $t = 1$ y en $t + 1$ genera un valor que sea $X_t + 1$ con probabilidad 1/2 y $X_t - 1$ con probabilidad también 1/2. Prueba también con probabilidades desiguales. Haz caminatas aleatorias de distintas longitudes: 100, 1000, 10000 iteraciones.

4.10.0.4. Ejercicio

Crea muchas caminatas aleatorias (de la misma longitud) y representa gráficamente la distribución de la última posición. ¿Qué distribución sigue?

4.10.0.5. Ejercicio

En $t = 0$ arrancan 5 caminatas aleatorias. Una de ella, que llamaremos ladrón, lo hace desde la posición 2. Las otras cuatro, que llamaremos policías, lo hacen desde la posición 0. Estima la media y la varianza del tiempo que les lleva a alguno de los policías a *atrapar* al ladrón.

4.10.0.6. Ejercicio

Una empresa tiene un capital inicial de 5 euros. En cada ejercicio gana un euro con probabilidad 1/3, pierde un euro con probabilidad 1/3 y tiene un resultado nulo con probabilidad 1/3. Si algún año la empresa se queda con un capital negativo, la empresa quiebra y desaparece. Estima mediante simulaciones la probabilidad de que la empresa siga operando al cabo de 30 años.

4.10.0.7. Ejercicio

Si $X \sim \text{Binom}(n, p)$ y sabiendo que $\sigma^2(X) = np(1 - p)$, calcula el valor de p para el que la varianza es máxima, i.e., la dispersión de los datos alrededor de la media es máxima.

4.10.0.8. Ejercicio

Si X es una variable aleatoria binomial $\text{binom}(n, p)$, entonces la probabilidad $P(X = x)$ se calcula en R usando `dbinom(x, n, p)`. Si $n = 100$ y $p = 0,7$, calcula:

- El valor x para el que la probabilidad es máxima.
- El conjunto más pequeño de valores tales que la suma de sus probabilidades excede el 90 %; es decir, los más probables extendiendo la lista justo hasta que la suma de sus probabilidades rebase el 90 % (de modo que las probabilidades del resto de los valores sea inferior al 10 %).

4.10.0.9. Ejercicio

Repite el ejercicio anterior con la distribución de Poisson (de parámetro 10). La función correspondiente es `dpois`.

4.10.0.10. Ejercicio

Usa la distribución multinomial para estimar la media y la desviación estándar de la distribución de la variable aleatoria que cuenta el número de tiradas de 6 dados necesarias hasta lograr un puntaje ≥ 30 .

4.10.0.11. Ejercicio

En un lago hay 1000 peces. Capturas 100, los marcas y los tira de nuevo al lago. Luego capturas otros 100 y cuentas cuántos tienen marca. Calcula (mediante remuestreos) una aproximación a la probabilidad de que haya 10 peces marcados en la segunda captura.

Pista: selecciona dos veces 100 elementos de entre 1000 y cuenta las coincidencias.

Una mejora: si lo piensas bien, solo hace falta seleccionar una vez, no dos (y la simulación es más rápida).

Nota: este es un ejemplo de una distribución discreta, la **distribución hipergeométrica**.

5. Distribuciones de probabilidad continuas

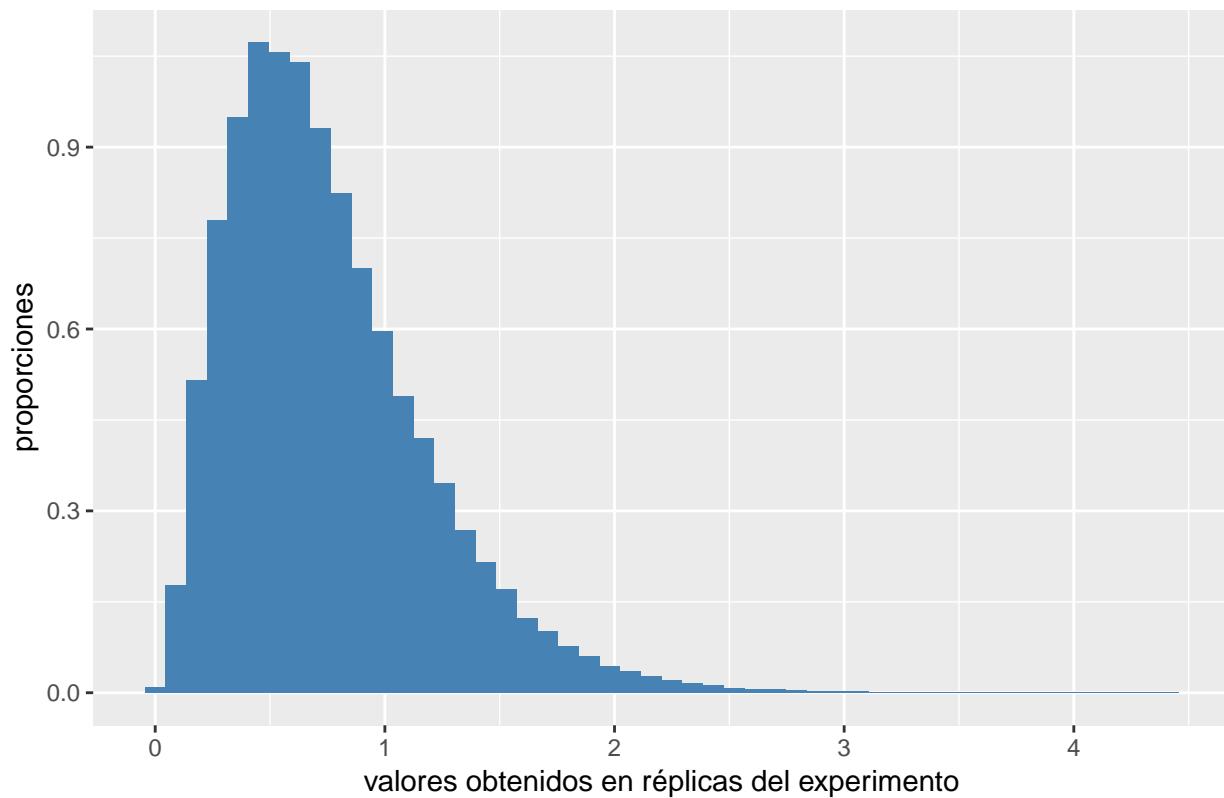
Las distribuciones de probabilidad continuas son las correspondientes a variables aleatorias relacionadas con eventos como los siguientes:

- Que mañana la bolsa baje más del 1%.
- Que alguien sano tenga una concentración de urea en la sangre superior a x .
- Que alguien sin estudios gane más de 3000 euros al mes; la probabilidad en correspondiente en este caso sería una probabilidad condicional (condicionada a que el sujeto no tenga estudios)
- Que alguien mida más de 1.90 y pese menos de 80 kilos; en este caso, a diferencia de los anteriores, la variable aleatoria es bidimensional: tiene en cuenta la altura y el peso.
- Que el tiempo que pasa hasta que alguien se entera de la noticia *Fidel Castro ha muerto* sea menor que una hora

5.1. De histogramas a funciones de densidad

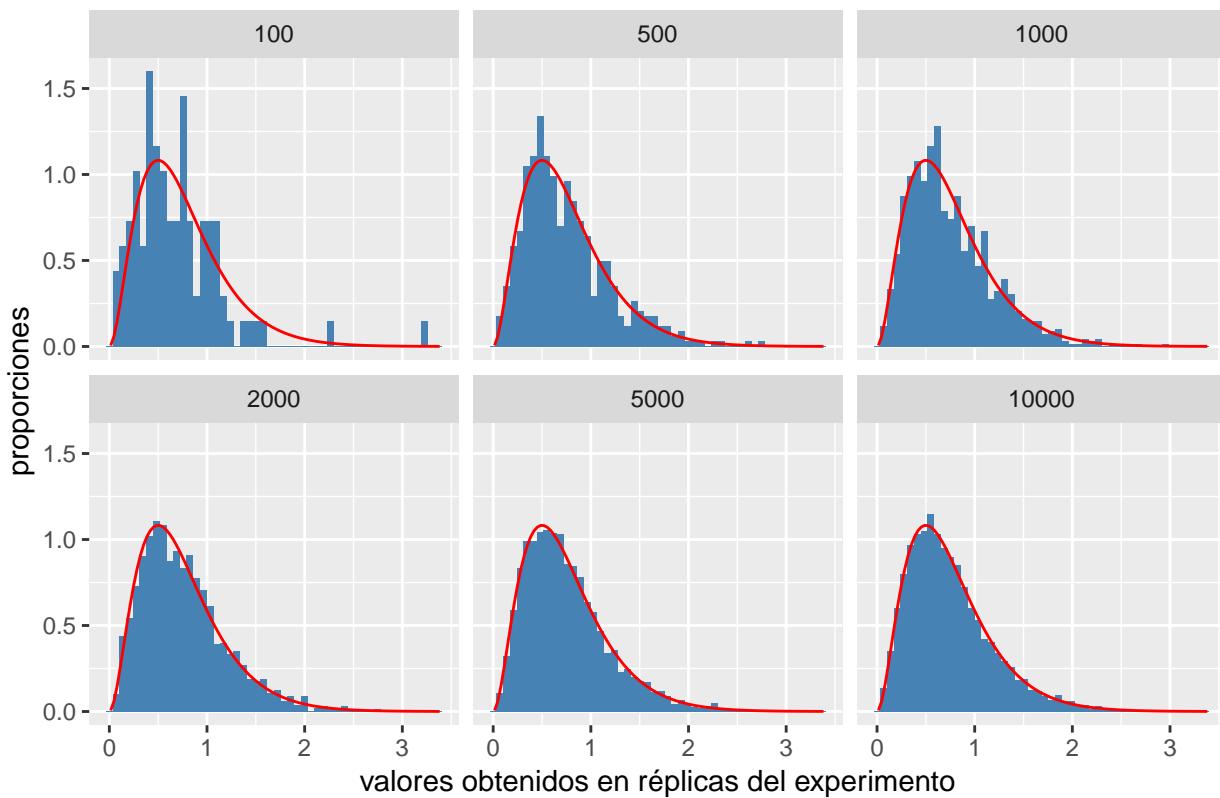
De una variable aleatoria continua, X , pueden obtenerse observaciones (sea simulando, sea realizando experimentos en laboratorio, etc.). Estas observaciones pueden representarse mediante un histograma, como el siguiente:

histograma de la muestra de valores experimentales



Los histogramas son representaciones de datos que parten el rango de variación de la variable aleatoria en segmentos (*bins*) y calcula la proporción de las observaciones obtenidas que caen en cada uno de ellos. Estas proporciones definen la altura de los rectángulos.

histograma de la muestra de valores experimentales



Conforme se obtienen más observaciones, estos histogramas suelen adoptar una forma suave: en el límite, *convergen* a una función suave. Esa función suave a la que convergen los histogramas conforme $n \leftrightarrow \infty$ se llama función de densidad. Si f es una función de densidad, como consecuencia de cómo se obtiene (es decir, a partir de histogramas conforme $n \leftrightarrow \infty$), se obtienen algunas propiedades suyas:

- $f \leq 0$
- $\int_{-\infty}^{\infty} n f(x) dx = 1$
- $\int_{-\infty}^a f(x) dx = P(X < a)$.

La función de densidad es fundamental en probabilidad y estadística y a partir de ella, como veremos, se pueden definir otras como la de probabilidad o la de cuantiles. Pero es importante tener en cuenta que la función de densidad se construye como el límite *ideal* de una sucesión de histogramas con un número creciente de datos. Este mecanismo constructivo no es solo interesante por el hecho de establecer su procedencia sino, además, como se ha indicado más arriba, porque las funciones de densidad heredan las propiedades conocidas de los histogramas.

Para todo fenómeno aleatorio discreto puede construirse una función de densidad que le es específica. Sin embargo, sucede que muchos fenómenos aleatorios comparten función de densidad. O, más bien, las funciones de densidad asociadas a muchos fenómenos aleatorios en principio distintos, son (aproximadamente) comunes: existen motivos, que se discutirán más adelante, por los que muchos histogramas convergen a la función de densidad normal u otras distribuciones *de libro*, algunas de las cuales se estudiarán luego.

5.2. Funciones de densidad, probabilidad y cuantiles

Para ilustrar los conceptos de esta sección, vamos a utilizar un caso hipotético en el que nuestra variable aleatoria será el tiempo transcurrido desde que ocurre cierto acontecimiento noticiable hasta que la gente, por el medio que sea, llega a conocer la noticia. Y supondremos que la variable aleatoria que mide ese tiempo

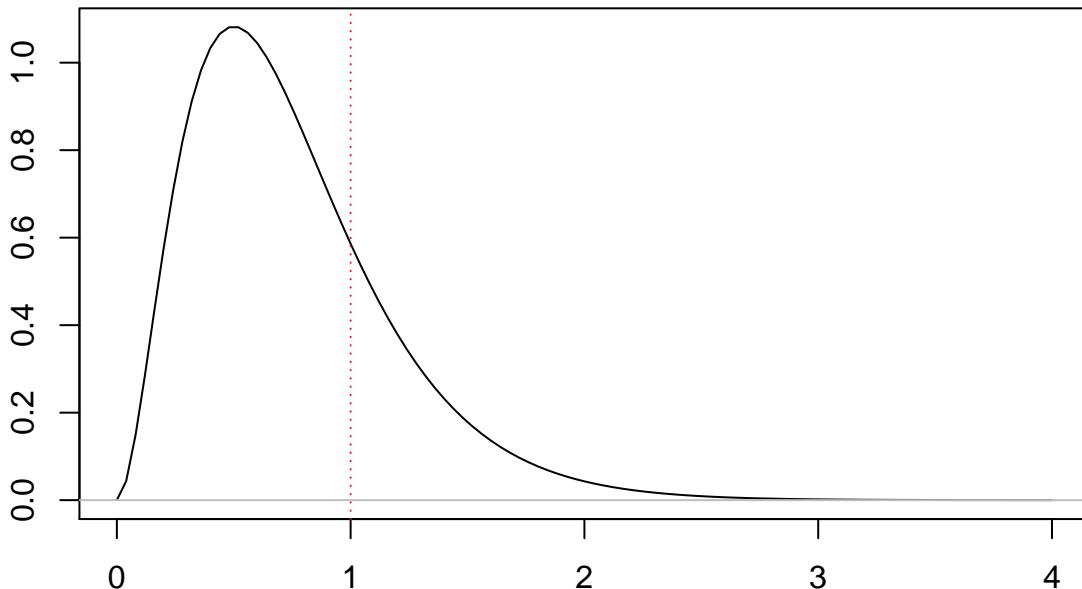
sigue una distribución gamma de parámetros 3 y 4, i.e., $\Gamma(3, 4)$.

Nota: existen dos maneras distintas de parametrizar la distribución gamma; en lo que sigue, sus parámetros serán los de forma e intensidad. En nuestro caso, el de escala sería el inverso del de intensidad, i.e., $1/4$.

5.2.1. Función de densidad

Gráficamente, la **función de densidad** es la siguiente:

Densidad de una gamma (3,4)



Como se ha indicado antes, la función de densidad es siempre positiva (≥ 0) y su integral es $P(\Omega) = 1$. También se puede deducir que su integral entre a y b , $\int_a^b p(x)d(x)$ es $P(a < X < b)$; en nuestro caso, la proporción de personas que se enteran de la noticia entre las horas a y b ; o, de otro modo, la probabilidad de que alguien se entere de ella en ese periodo de tiempo.

La forma de p indica cómo al principio se enteran de la noticia pocas personas, pero la velocidad de transmisión de la información crece hasta alcanzar un pico alrededor de los 40 minutos (que es cuando más probable es enterarse de ella) para luego decaer *lentamente* (al menos, con respecto a la velocidad inicial de transmisión). A partir de las 2 horas, son ya pocas las personas que desconocen la noticia.

Todas las distribuciones continuas tienen una función de densidad. Las funciones de R correspondientes son, por ejemplo, `dgamma`, `dnorm`, etc. Por convención, siguen esa nomenclatura: `d` seguido del nombre (tal vez abreviado) de la distribución.

En la gráfica se ha trazado una línea vertical punteada que pasa por el 1 (una hora). La integral de la curva de 0 a 1 es, precisamente, la proporción de personas que se enteran de la noticia en una hora o menos, i.e., $P(X < 1)$.

5.2.2. Función de probabilidad

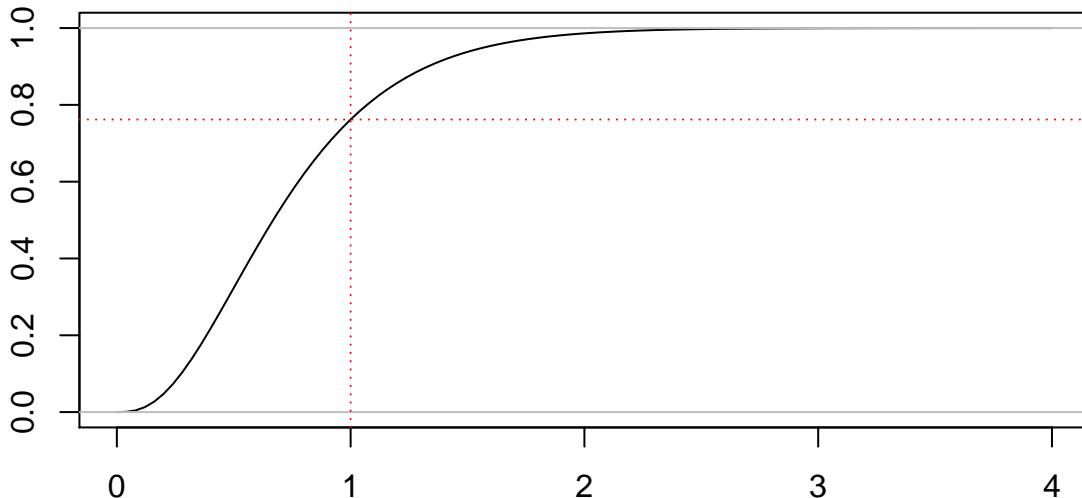
El tipo de eventos $X \leq a$ son muy importantes y por eso es útil contar con la llamada **función de probabilidad**, F , definida así:

$$F(a) = P(X \leq a).$$

Se deduce automáticamente que la función de probabilidad crece desde 0 hasta 1.

En R, la función de probabilidad sigue la misma nomenclatura que la de densidad, solo que usando **p** en lugar de **d** (p.e., **pgamma**). Usando, precisamente, **pgamma** podemos representar la función de probabilidad asociada a nuestro problema:

Función de probabilidad de una gamma (3,4)



Se ve cómo crece desde 0 y cómo se satura hacia el valor 1 por la derecha. La recta vertical anterior corta a la curva en el valor $F(1) = P(X \leq 1) \approx 0,762$: el 76.2% de las personas se enteran de la noticia en una hora o menos.

5.2.3. Cuantiles

Vamos a examinar detenidamente esa última expresión:

El 76.2% de la población se enteró de la noticia en menos de una hora.

Significa que, tal como ya sabemos, 0,762 es el valor de la función de probabilidad asociada a 1 (hora); pero, visto a la inversa, 1 (hora) es el **cuantil** al 76.2% de la distribución. Es decir, igual que podemos asociar probabilidades a momentos en el tiempo, invirtiendo la relación podemos asociar momentos en el tiempo a probabilidades. Los valores de X asociados de esa manera a probabilidades son los cuantiles. Y, por supuesto, existe una función, la de cuantiles (con prefijo **q** en R; p.e., **qnorm**), que permite responder a preguntas del tipo:

- ¿En cuánto tiempo se enteró el primer 10%?
- ¿A partir de cuándo conocía la noticia el 99%?

En R, la respuesta a las dos preguntas anteriores sería

```
qgamma(0.1, 3, 4)
```

```
## [1] 0.2755163
```

y

```
qgamma(0.99, 3, 4)
```

```
## [1] 2.101487
```

respectivamente.

5.2.4. Media y varianza

Si p es la función de densidad de una variable aleatoria continua X , entonces su media es

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

y su varianza,

$$\sigma^2(X) = \int_{-\infty}^{\infty} (x - E(X))^2 p(x)dx.$$

Las integrales anteriores también se pueden estimar por simulación, tal como hicimos con las distribuciones discretas. Por ejemplo, en nuestro problema y de acuerdo con la Wikipedia (que indica el valor de la media y la varianza en función de sus parámetros porque alguien calculó las integrales anteriores), $E(X) = 0,75$ y $\sigma^2(X) = 0,1875$.

Pero podemos aproximar mediante simulación esos valores usando las funciones correspondientes de R (con prefijo `r` como, p.e., `rgamma` o `rnorm`):

```
muestra <- rgamma(1e6, 3, 4)
mean(muestra)
```

```
## [1] 0.74983
var(muestra)

## [1] 0.1871121
```

5.2.5. Funciones d, p y q para distribuciones discretas

Las distribuciones discretas (casi todas: algunas no tienen sentido, p.e., para la distribución multinomial) también tienen asociadas estas funciones. Con la salvedad de que la función de densidad está concentrada en determinados puntos,

```
dbinom(2, 4, 0.5)
```

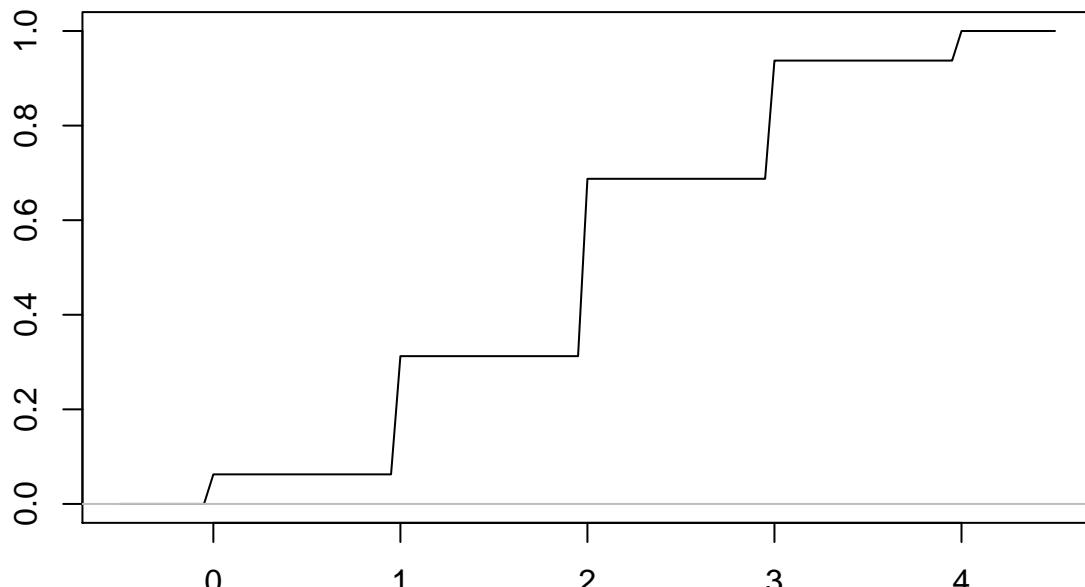
```
## [1] 0.375
dbinom(2.5, 4, 0.5)

## Warning in dbinom(2.5, 4, 0.5): non-integer x = 2.500000
## [1] 0
```

y la función de probabilidad es escalonada:

```
curve(pbinom(x, 4, 0.5), -0.5, 4.5, xlab = "", ylab = "",
      main = "Función de probabilidad Binom(4, 0.5)")
abline(h = 0, col = "gray")
```

Función de probabilidad Binom(4, 0.5)



Por supuesto, también está definida la función de cuantiles. Así, por ejemplo, si el número diario de fallecimientos en un hospital es Pois(20), típicamente, con un 90 % de probabilidad, el número de fallecidos estará en el rango

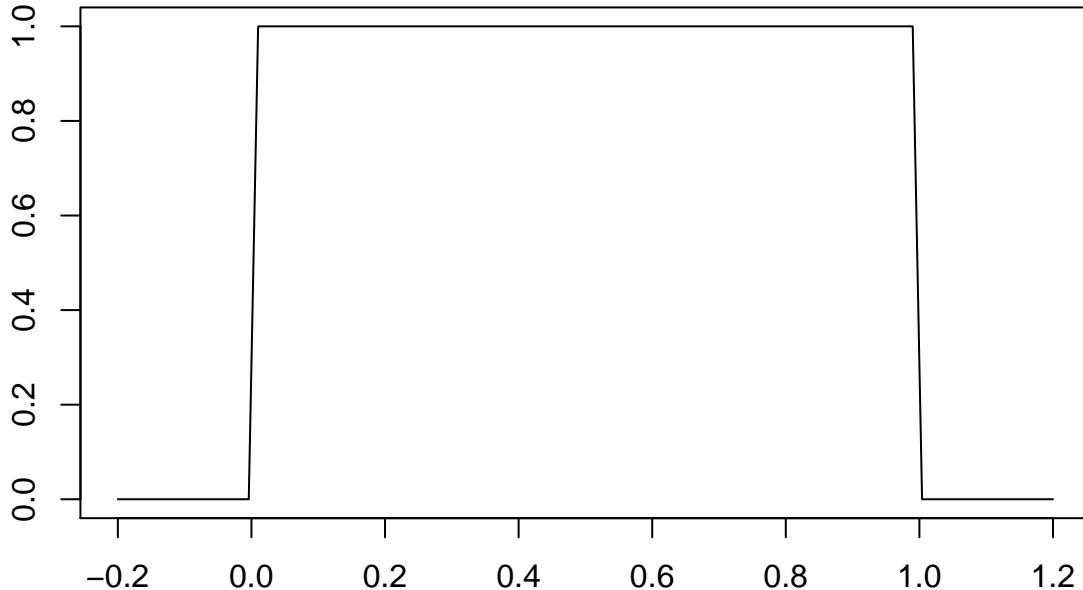
```
qpois(c(0.05, 0.95), 20)
```

```
## [1] 13 28
```

5.3. La distribución uniforme

La distribución uniforme es, posiblemente, la más sencilla entre las continuas: su densidad es 0 salvo en un determinado rango $[a, b]$ donde es constante (y, como consecuencia, toma el valor $1/(b - a)$). Es decir, solo puede tomar valores en ese rango y, dentro de él, todos son equiprobables.

Densidad uniforme en [0,1]



Es importante, entre otros motivos, por que los generadores de números seudoaleatorios tratan de muestrear una distribución uniforme en $[0, 1]$. Para generar valores de otras distribuciones es necesario realizar manipulaciones sobre esos valores. Un procedimiento práctico para simular determinadas distribuciones es el siguiente:

- Obtener una muestra de valores x_i de la distribución uniforme sobre $[0, 1]$.
- Aplicarles la función de cuantiles (i.e, F^{-1}) de la distribución objetivo para obtener la muestra $y_i = F^{-1}(x_i)$.

En efecto, $P(y_i \leq a) = P(F^{-1}(x_i) \leq a) = P(x_i \leq F(a)) = F(a)$. La tercera igualdad es consecuencia del hecho de que los x_i tienen una distribución uniforme en $[0, 1]$, por lo que $P(x_i \leq a) = a$.

Por ejemplo, la distribución exponencial tiene $F(x) = 1 - \exp(-\lambda x)$, por lo que $F^{-1}(x) = \frac{-1}{\lambda} \log(1 - x)$ y, como consecuencia, muestrear la distribución exponencial se reduce a tomar logaritmos de valores números seudoaleatorios en $(0, 1)$.

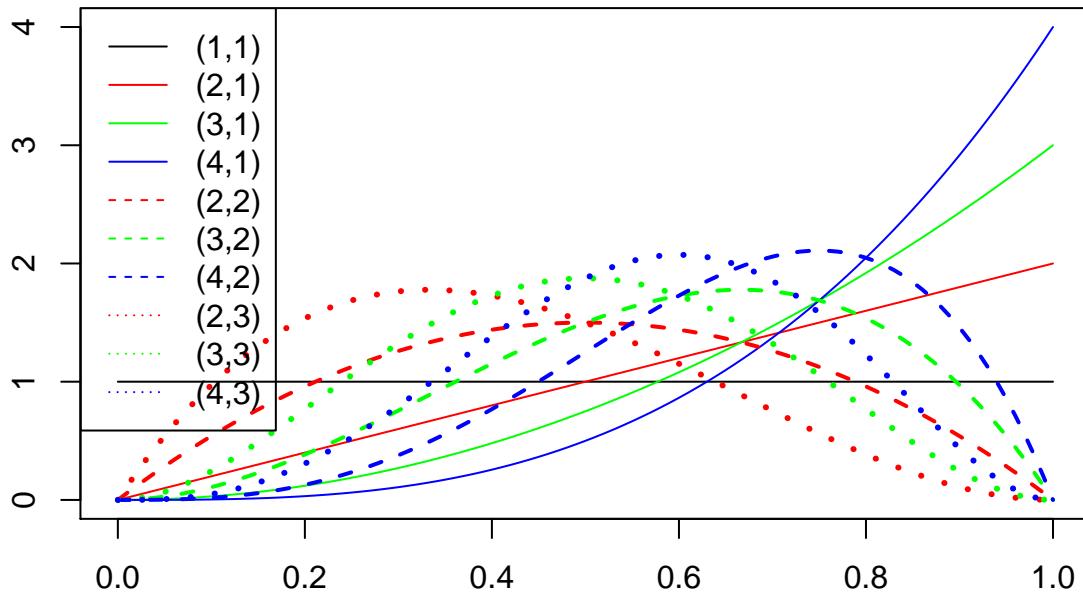
5.4. La distribución beta

La distribución beta es una generalización de la uniforme y también toma valores entre 0 y 1. Valores entre 0 y 1 pueden significar muchas cosas, pero habitualmente representan proporciones o probabilidades. De ahí que la distribución beta se utilice a menudo para modelar la incertidumbre sobre una probabilidad.

Esto se entiende mejor con un ejemplo. Los clientes de un banco pueden usar su tarjeta de débito para extraer dinero de cajeros y para pagar en comercios. Puede interesar conocer los hábitos de los clientes: ¿qué tipo de uso tienden a hacer? ¿Cuál es la proporción de veces que usan la tarjeta de uno u otro modo? Si un cliente la ha usado 100 veces y siempre para extraer dinero de cajeros, hay bastante certeza en que $P(\text{cajeros}) \approx 1$. ¿Pero qué pasa si la ha usado 3 veces, dos de ellas en cajeros y 1 en comercios? En tal caso, $P(\text{cajeros}) \approx 2/3$, pero el grado de certeza de esa probabilidad es menor que si el cliente la ha usado 300 veces, 200 de ellas en cajeros. A pesar de que la proporción estimada sea la misma.

La distribución beta tiene este aspecto:

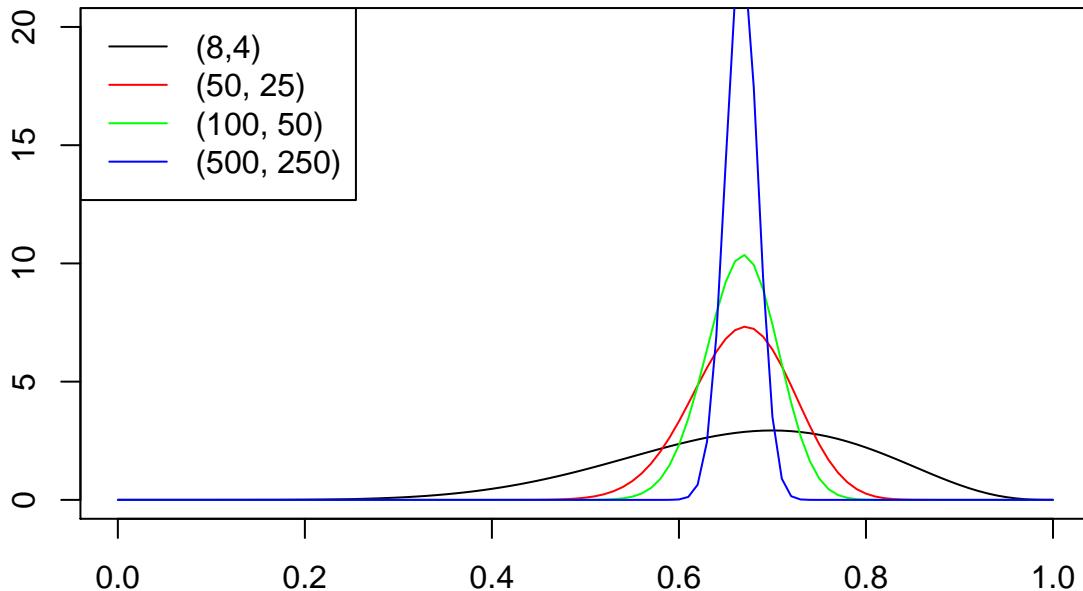
Distribución beta



Con parámetros (1,1) es uniforme. Conforme aumenta el primer parámetro, tiende a concentrar probabilidad en la parte derecha, alrededor del 1. Eso indicaría que crece nuestra certeza de que el valor de la proporción desconocida es 1.

Para valores iguales de los parámetros la distribución es simétrica y tiene media 1/2. Pero conforme aumenta su valor, la distribución se hace más picuda (i.e., decrece la dispersión). Porque, en efecto, la media y la varianza de la distribución beta es, en función de sus parámetros α y β , $E(X) = \frac{\alpha}{\alpha+\beta}$ y $\sigma^2(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Aunque la expresión de la varianza ya es de por sí lo suficientemente reveladora, puede mostrarse la *convergencia* de la distribución beta a una proporción dada en función de del tamaño de los parámetros gráficamente:

Distribución beta y aumento de certeza



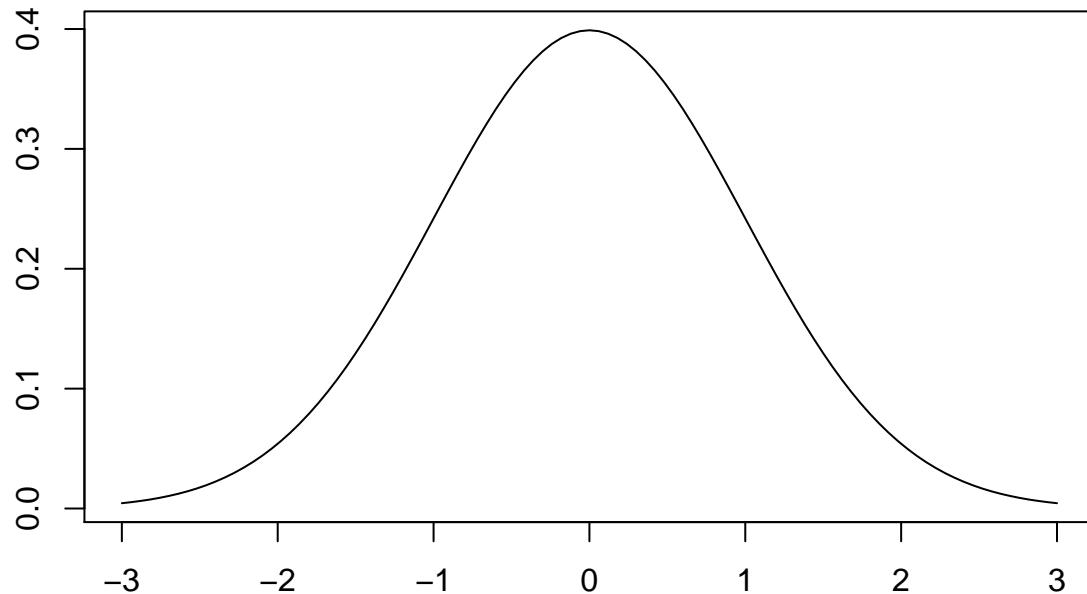
Se aprecia cómo al aumentar el tamaño relativo de los parámetros (guardando la misma proporción) aumenta

la certeza sobre el valor estimado subyacente ($2/3$ en nuestro caso). Se puede probar cómo, en la modelación de determinados problemas, los parámetros de la distribución crecen con la cantidad de información disponible (que está directamente relacionada con el número de usos de la tarjeta por parte de un cliente en nuestro ejemplo).

5.5. La distribución normal

La densidad de la distribución normal tiene el siguiente aspecto:

Distribución normal est\'andar

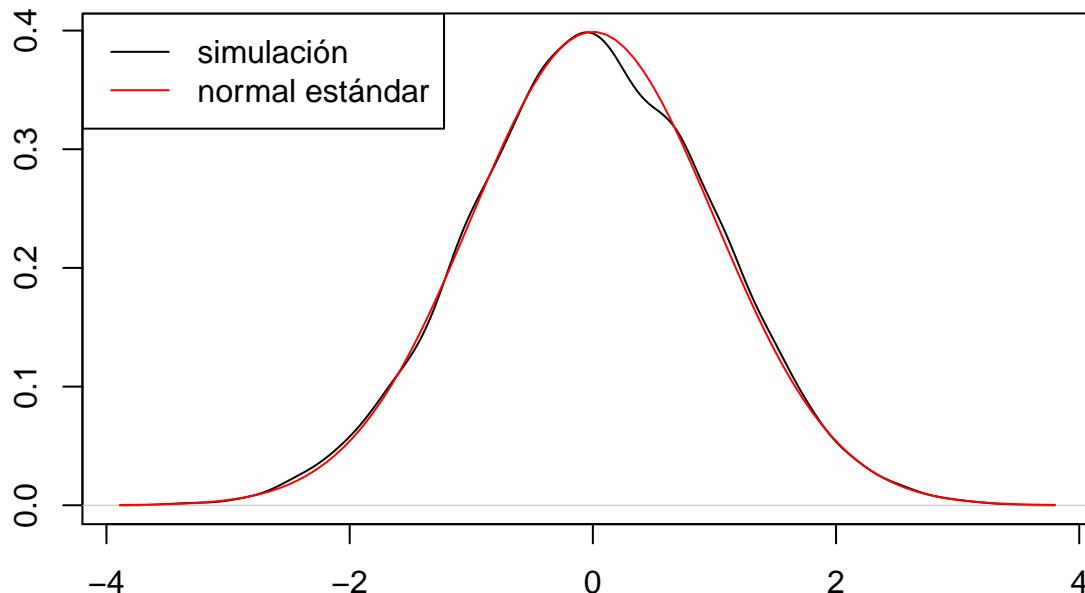


Se trata de la famosa *campana de Gauss*, que, por defecto, está centrada en 0 y tiene desviación est\'andar 1. Sin embargo, admite dos par\'ametros, μ , la media y σ , la desviaci\'on est\'andar) que o la desplazan o la contraen o expanden.

La distribución normal tiene una importancia fundamental en la teoría de la probabilidad porque es un *atractor* de distribuciones. No es solo que la suma de variables aleatorias independientes con una distribución normal tenga también distribución normal sino que, además, la suma de variables aleatorias indpendientes de otras distribuciones tiende a tener tambi\'en distribuci\'on normal. Por ejemplo, podemos sumar 12 variables aleatorias uniformes independientes y restar seis muchas veces para ver qué aspecto tiene la variable aleatoria resultante:

```
muestra <- replicate(10000, sum(runif(12))) - 6
plot(density(muestra), xlab = "", ylab = "", main = "Muestra vs normal")
curve(dnorm(x), col = "red", add = T)
legend("topleft",
       c('simulaci\'on', 'normal est\'andar'),
       lty = c(1, 1),
       col = c(par('fg'), 'red'))
```

Muestra vs normal



En efecto, la distribución uniforme tiene media 6 y desviación estándar $1/\sqrt{12}$. Sumando 12 variables aleatorias uniformes y restando seis, se obtiene una variable aleatoria con media 0 y desviación estándar 1. Pero no solo eso: su aspecto es muy similar a la de una normal.

De la misma manera, el aspecto de la distribución binomial que, recuérdese, es la suma de variables aleatorias independientes de Bernoulli, también es acampanado. Y se puede probar que la distribución normal (debidamente parametrizada para que concuerden la media y la varianza) es una aproximación legítima a la binomial cuando n es lo suficientemente grande.

Esto no es casualidad: el **teorema central del límite** garantiza en ciertos casos la convergencia de sumas (más bien, promedios) de variables aleatorias a una con la distribución normal. En el caso de la distribución uniforme, la convergencia es muy rápida: basta (para ciertos fines) sumar 12 de ellas.

El teorema central del límite tiene ciertas restricciones sobre las variables aleatorias que se promedian. Por ejemplo, que sean independientes (aunque esa exigencia se puede suavizar). Otra, que todas sean pequeñas, de manera que ninguna de ellas prevalezca sobre el resto. Esos criterios sirven tanto para intuir que la distribución correspondiente a cierto fenómeno es normal como para argüir en sentido contrario cuando se detecta alguna violación de las restricciones.

Por ejemplo, los errores de medida (varios técnicos midiendo una determinada magnitud independientemente) tienden a tener una distribución normal: errores de calibración del instrumental, cambios pequeños en las condiciones físicas, etc. influyen independientemente y en pequeña medida en las observaciones. Puede argumentarse que la altura de las personas sigue una distribución normal: está influida por una miríada de pequeños factores, sean genéticos, nutricionales, etológicos, etc. Pero también puede contraargumentarse: existe un factor genético muy importante, el sexo, que hace que los hombres tiendan a ser más altos que las mujeres. De hecho, la altura de las personas no es normal sino que puede modelarse más adecuadamente como una mezcla de dos normales: las correspondientes a los dos sexos.

5.6. La distribución t

La distribución normal tiene colas finas: es prácticamente imposible que ocurran eventos alejados de la media. Por ejemplo,

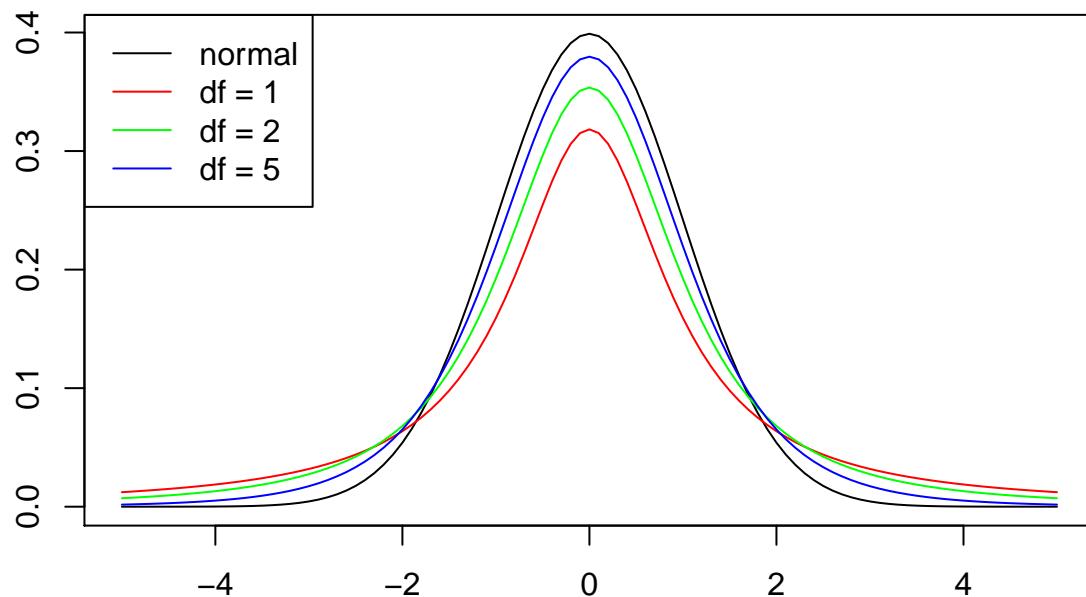
```
pnorm(-4)
```

```
## [1] 3.167124e-05
```

es un número muy pequeño: con la distribución normal solo ocurren eventos situados más allá de 4 desviaciones típicas de la media en 3.1 ocasiones de cada cien mil. Pero muchas variables aleatorias tienen un comportamiento más errático. Por ejemplo, se observan variaciones en el precio diario de las acciones de un tamaño mayor que 4 desviaciones típicas con una frecuencia muy superior a las indicadas arriba, que correspondería a una vez cada 125 años (suponiendo que los mercados abren 250 días al año).

La distribución t es similar a la normal (simétrica, unimodal, etc.) pero tiene colas más gruesas. De hecho, no es una distribución sino una familia de distribuciones parametrizadas por un parámetro, el número de grados de libertad (o df), según el cual las colas son más o menos gruesas:

Distribución t

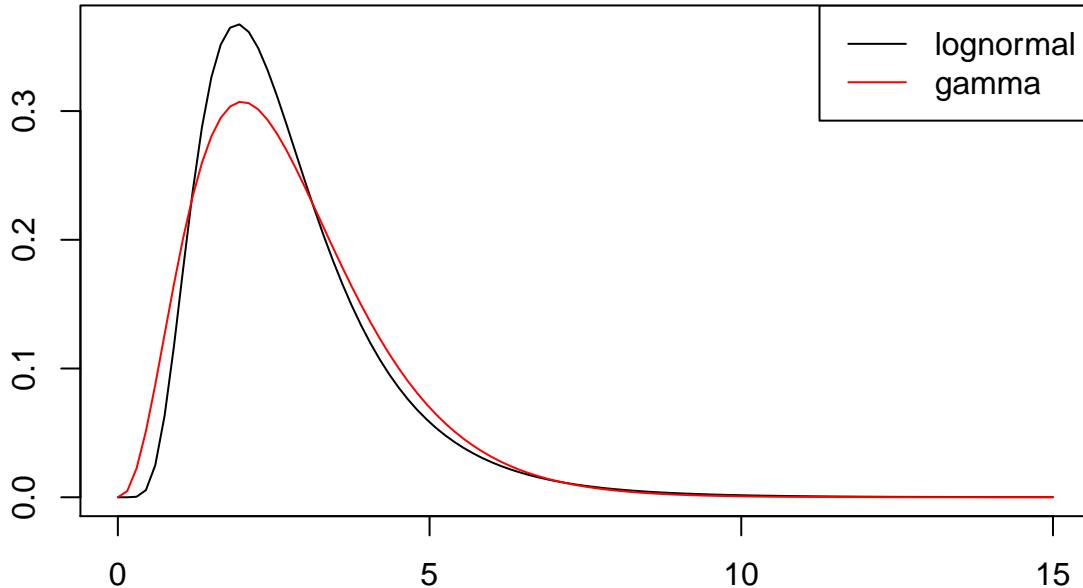


La distribución t con un grado de libertad, i.e., $df = 1$ se la conoce también como distribución de Cauchy. Tiene una peculiaridad: no tiene media. Eso se manifiesta, por ejemplo, en que un promedio de muestras de la distribución de Cauchy no converge como tienden a hacer los promedios en que aplica la ley de los grandes números sino que oscila. Eso se debe principalmente a los *outliers*: la cola de la distribución de Cauchy es tan gruesa que valores de un tamaño tan grande que hacen bascular todo el promedio ocurren frecuentemente.

5.7. Las distribuciones gamma y lognormal

Se trata de dos distribuciones con soporte en los valores $x > 0$ y que tienen una forma similar:

Distribuciones gamma y lognormal



Ambas son asimétricas y tienen una cola que desciende lentamente hacia la derecha. Se usan para modelar tiempos (hasta que ocurre algún evento) o magnitudes tales como ingresos, que se extienden a través de varios órdenes de magnitud.

La distribución lognormal, que es la exponencial de una distribución normal, ejerce el papel de *atractor de distribuciones* cuando, en lugar de sumarse, se multiplican. En los mercados financieros, por ejemplo, el precio de una acción que en t_0 vale A , sufre una variación de precio en t_1 que puede expresarse multiplicativamente: $A(1+x_1)$, donde x_1 es un valor positivo o negativo, próximo a 0, que indica el porcentaje de variación diario. Al cabo de n períodos, el precio se convierte en $A(1+x_1)\dots(1+x_n) \approx A \exp(x_1)\dots\exp(x_n) = A \exp(\sum_i x_i)$. Si las sumas de variaciones de precios son (vía el teorema central de límite) aproximadamente normales, la expresión $\exp(\sum_i x_i)$ será aproximadamente lognormal.

Por eso se usa en ocasiones la distribución lognormal para modelar los resultados bursátiles. Aunque hay que tener en cuenta la discusión anterior sobre la no normalidad de los movimientos diarios del precio de los activos financieros, que arrojan una sombra de sospecha sobre el uso de la distribución lognormal en estos contextos.

No obstante, e independientemente de la pertinencia del uso de la distribución lognormal en estos contextos, el ejemplo anterior ilustra cómo tal vez en otros en los que el efecto de las variables no es aditivo sino multiplicativo, la distribución lognormal puede resultar una herramienta de modelado útil.

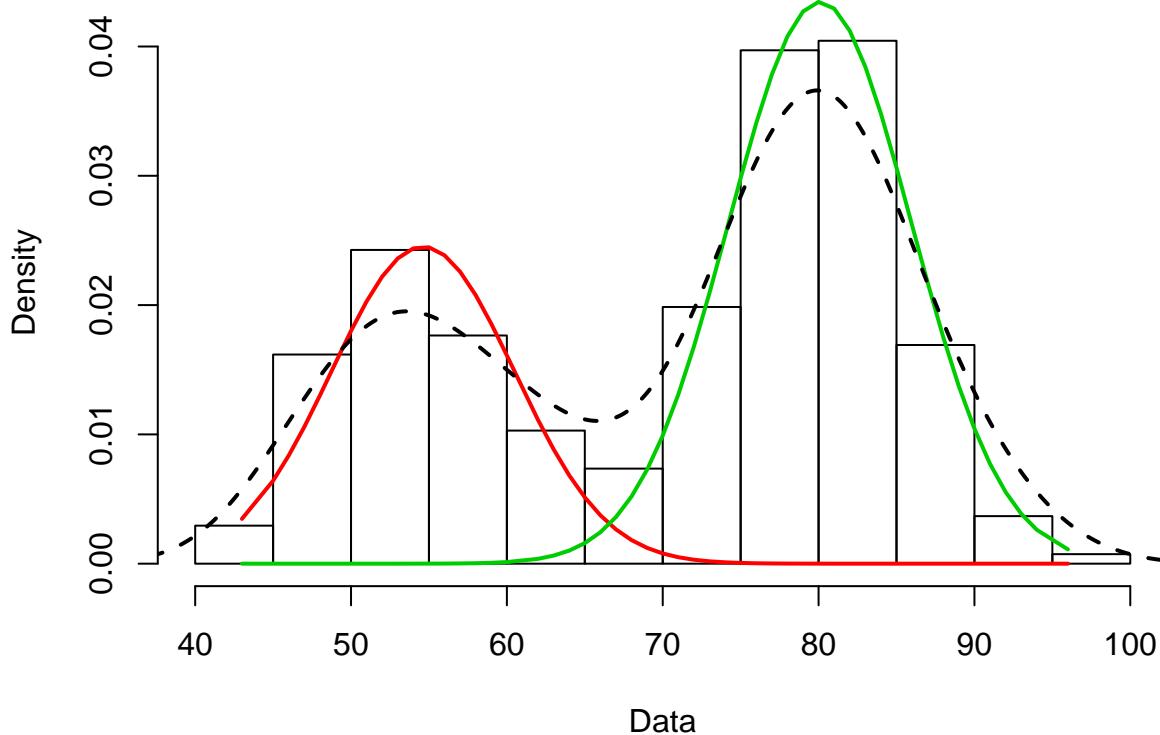
La distribución gamma se usa frecuentemente en el llamado análisis de la supervivencia: el estudio estadístico del tiempo que transcurre hasta que ocurre un fenómeno aleatorio: que falle una máquina, que fallezca un paciente, cierre su cuenta un cliente, etc.

5.8. Mezclas de distribuciones

Las distribuciones mencionadas más arriba, ni complementadas con las no consideradas por motivos de espacio, bastan para modelar *cualquier* fenómeno aleatorio. Deberían considerarse o bien como plantillas, o bien como aproximaciones, o bien, como en esta sección, como piezas para componer distribuciones más realistas.

Una manera de crear distribuciones más fieles a un fenómeno aleatorio es mediante mezclas (en ocasiones, *mixturas*) de distribuciones como en este ejemplo gráfico:

Mezcla de dos normales



En él se aproximan unos datos (el histograma) por una distribución (línea negra punteada) que es la mezcla de dos distribuciones normales (líneas continuas de color rojo y verde). Otro ejemplo habitual de mezclas de distribuciones es el de la altura de las personas que, aunque aparentemente normal, es (o es más fielmente) la mezcla de dos normales: las de las normales correspondientes a las alturas de hombres por un lado y mujeres por el otro. Otra situación en la que juegan un papel importante las mezclas de distribuciones es cuando en datos que siguen aparentemente una distribución de Poisson hay un exceso de ceros. Puede ser, por ejemplo, que se esté modelando el número de veces que los clientes de un banco usan su tarjeta de débito al mes; pero ocurre con frecuencia que un porcentaje importante de ellos no la usa nunca. Eso da lugar a los llamados modelos de Poisson con inflación de ceros, que no son otra cosa que una mezcla de la distribución de Poisson con otra de Dirac anclada en el cero.

Para describir una mezcla de distribuciones hace falta especificar dos cosas:

- Las distribuciones X_i que se mezclan.
- Sus correspondientes pesos, p_i (obviamente, $\sum_i p_i = 1$).

Para muestrear una mezcla de distribuciones, se itera el siguiente algoritmo tantas veces como muestras se quieran obtener:

1. Se obtiene un índice i al azar de acuerdo con las probabilidades p_i .
2. Se elige la variable aleatoria correspondiente X_i .
3. Se muestrea X_i .

Resulta evidente que la media de una mezcla de distribuciones es $E(X) = \sum_i p_i E(X_i)$. La expresión correspondiente a la varianza es un tanto más complicada.

5.9. Distribuciones jerárquicas

La mezcla de distribuciones es un caso particular de una técnica para construir las distribuciones con las que modelar fenómenos aleatorios complejos. Por ejemplo, el de las pérdidas por siniestros en una compañía de seguros en un periodo determinado (p.e., un mes), que se describe de la siguiente manera:

- El número de siniestros es Pois(λ)
- El impacto económico de cada uno de ellos es lognormal

Aunque tal vez la distribución así descrita no esté descrita en la literatura o tenga un nombre determinado, es posible simularla obteniendo muestras repitiendo cuantas veces sea necesario el siguiente algoritmo:

1. Se toma un valor n de una variable aleatoria Pois(λ)
2. Se toman n muestras de una variable aleatoria lognormal (con los parámetros adecuados), x_1, \dots, x_n
3. Se toma la suma $\sum_i x_i$

Otro ejemplo de utilidad práctica podría ser el siguiente: en un negocio de internet, el número de visitas diarias es una variable aleatoria de Poisson. Un porcentaje (pequeño) de esos visitantes realiza una compra y el precio de la venta es lognormal.

5.10. Consideraciones finales

En general, cada tipo de evento tiene su propia distribución de probabilidad. Si nos interesa el número de litros por metro cuadrado que lloverá mañana, probablemente sea adecuado modelarlo como una mezcla de una distribución discreta (de Dirac centrada en cero) con una continua.

En esta sección hemos presentado algunas distribuciones *con nombre*, que son útiles o inútiles según cómo se considere. Son útiles en tanto que algunos procesos (¡pocos!) siguen ese tipo de distribuciones. También porque aunque solo lo sea aproximadamente, la aproximación resulta lo suficientemente buena. Y, en cualquier caso, porque las propiedades conocidas de las distribuciones *con nombre* pueden extrapolarse a fenómenos aleatorios cuya distribución se parece a ellas. Finalmente, porque las distribuciones *con nombre* pueden combinarse de diversas maneras para modelar fenómenos complejos.

Por eso que el problema de determinar *qué distribución siguen mis datos*, especialmente cuando se formula en términos de *cuál de la lista de distribuciones conocidas* es mucho menos relevante de lo que muchos opinan.

Existen pruebas estadísticas y medidas de la bondad de ajuste para determinar en qué medida, por ejemplo, unos datos siguen o no la distribución normal. En ese caso se pueden usar pruebas estadísticas como la de Kolmogorov-Smirnov (`ks.test` en R) u otras técnicas similares.

Pero, en general, es recomendable replantear el problema en otros términos. En primer lugar, reflexionando acerca de si hay razones para suponer que unos determinados datos tienen una de esas distribuciones *de libro*. En ocasiones puede justificarse. En otras es posible describir la distribución como, por ejemplo, como hemos hecho más arriba, mediante una mezcla de distribuciones o mediante otro mecanismo que simule el mecanismo generativo de los datos usando las distribuciones conocidas como elemento constructivo.

En última instancia, siempre se puede trabajar sobre los datos mismos y estudiarlos sin construcciones matemáticas (las distribuciones de probabilidad) interpuestas. El recurso, tan habitual, a la distribución normal se debe a muchos motivos, de entre los que sobresale el teorema central del límite. Sin embargo, también obedece a motivos espurios: para describir una distribución normal basta con conocer su media y su desviación estándar, solo dos números. Esa concisión fue importante en la época en que tanto el proceso como la transmisión de información era muy onerosa. Actualmente, ese ya no es un problema y podemos operar sobre muestras grandes (o incluso completas) directamente.

5.11. Referencias

- Mezclas de distribuciones

5.12. Ejercicios

5.12.0.1. Ejercicio

Integra la función de densidad `dgamma(x, 3, 4)` entre 0 y 1 usando `integrate`. Compara el resultado con el obtenido usando la correspondiente función de probabilidad.

5.12.0.2. Ejercicio

Usa `optimize` para encontrar el valor máximo de `dgamma(x, 3, 4)`.

Nota: el valor máximo de una función de densidad se llama **moda**.

5.12.0.3. Ejercicio

Toma una muestra de tamaño n de la distribución $\Gamma(3,4)$ (usando `rgamma(n, 3, 4)`) y calcula la **función de probabilidad empírica** usando la función `ecdf`. Crea un gráfico que la compare con la función de probabilidad original. Utiliza distintos valores de n . ¿Ves algún patrón?

5.12.0.4. Ejercicio

El 80 % de la probabilidad de una $\Gamma(3,4)$ está entre los valores `qgamma(c(0, 0.8), 3, 4)`. Y también entre `qgamma(c(0.1, 0.9), 3, 4)`. Usa `optimize` para encontrar el intervalo más estrecho. ¿Qué utilidad piensas que puede tener este intervalo que lo hace preferible al resto?

5.12.0.5. Ejercicio

Representa gráficamente varios intervalos de los que comprenden el 80 % de la probabilidad de una $\Gamma(3,4)$. ¿Qué propiedad característica tiene el más corto comparado con el resto?

5.12.0.6. Ejercicio

En $t = 1$ quedan por enterarse de la noticia (en el ejemplo del texto) $1 - F(1)$ (en proporción) de la población. En $t = 1,1$ se han enterado (en proporción) $F(1,1) - F(1)$ de la población. Por lo tanto, en ese intervalo se han enterado de la noticia una proporción (o tasa)

$$\frac{F(1,1) - F(1)}{1 - F(1)}$$

de los que aún no se habían enterado. Considera la función

$$g(t) = \frac{F(t + 0,1) - F(t)}{1 - F(t)}$$

y represéntala gráficamente. ¿Qué aspecto tiene? ¿Cada vez es más fácil o más difícil enterarse?

Nota: esta función es una versión de la llamada función de riesgo (*hazard*) en el análisis de la supervivencia.

5.12.0.7. Ejercicio

Repite el ejercicio anterior pero considerando una distribución exponencial de parámetro 4 (puedes usar $\Gamma(1, 4)$ o, mejor, la distribución exponencial `Exp(4)`).

5.12.0.8. Ejercicio

El número de visitas a una página es, en promedio, de 240k al día. Podemos suponer que el número de visitas se distribuye según una distribución de Poisson. Los ingenieros quieren dimensionar la página para que solo se caiga en situaciones extremas, el 0.1 % de los días. ¿Para cuántas visitas diarias deberían dimensionar el servidor?

5.12.0.9. Ejercicio

Resuelve el ejercicio anterior mediante simulaciones: genera muchas muestras de la distribución de Poisson correspondiente y calcula el valor que deja a su derecha derecha solo el 0.1 % de ellas (i.e., si la muestra son x_i , el valor buscado es a tal que la proporción de los $x_i > a$ es el 0.1 %).

5.12.0.10. Ejercicio

Las visitas no tienen la misma *intensidad* a lo largo del día. La intensidad varía horariamente de acuerdo con una distribución exponencial $\text{Exp}(\lambda)$. Si en una hora el valor de λ es λ_0 , entonces el número de visitas en esa hora tiene distribución de Poisson con parámetro λ_0 .

Encuentra el valor de λ tal que la intensidad promedio diaria da 240k visitas. Después, calcula (mediante simulaciones) el valor extremo de interés, el que da una garantía de servicio del 99.9 % (de las horas). Compáralo con el obtenido antes.

5.12.0.11. Ejercicio

Calcula la media de una distribución normal estándar mediante simulaciones. Haz muchas simulaciones con 100, 500, 1000 y 10000 muestras y compara su dispersión.

5.12.0.12. Ejercicio

Haz lo mismo que antes, pero para la varianza.

5.12.0.13. Ejercicio

Repite el ejercicio de la estimación de la media con la distribución de Cauchy. Muestra su distribución obteniendo muchas estimaciones con 100, 500, 1000 y 10000 muestras y compara su dispersión.

6. Distribuciones de probabilidad multivariantes

En esta sección vamos a volver a considerar la distribución conjunta de dos o más variables aleatorias. Si X_1, \dots, X_n son variables aleatorias continuas, podemos considerar su función de densidad conjunta,

$$f(x_1, \dots, x_n)$$

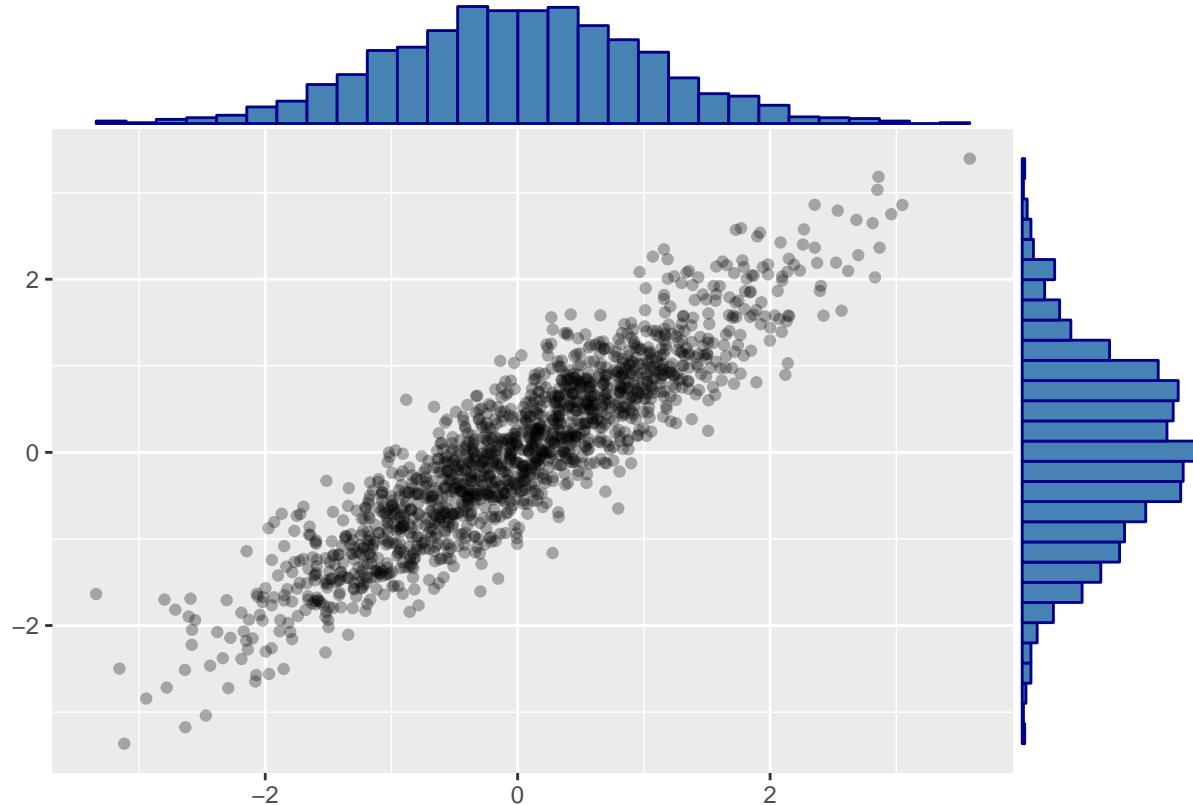
que tiene, obviamente, masa (o integral) 1 y la correspondiente función de probabilidad,

$$F(a_1, \dots, a_n) = P(X_1 \leq a_1, \dots, X_n \leq a_n) = \int_{-\infty}^{a_1} \cdots \int_{-\infty}^{a_n} f(x_1, \dots, x_n) dx_1 \dots dx_n$$

Las distribuciones marginales se obtienen por integración. P.e., para dos variables aleatorias,

$$f(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2$$

El siguiente gráfico muestra una distribución bidimensional (más bien, una muestra de ella) acompañada de sus correspondientes distribuciones marginales representadas como un histograma.

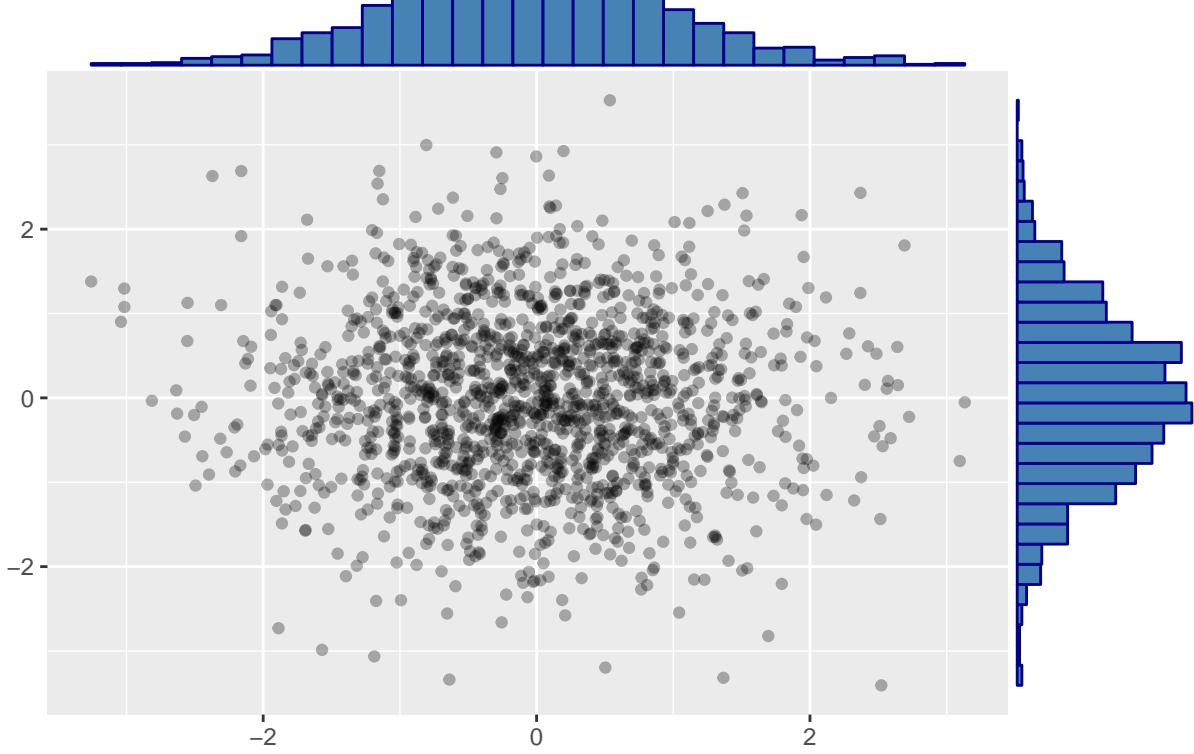


Merece la pena mencionar las cónulas: distribuciones construidas de forma que tengan unas distribuciones marginales dadas y además de una determinada correlación entre las variables. Las cónulas encuentran aplicaciones en finanzas y otros ámbitos muy concretos de la ciencia de datos y no nos ocuparemos de ellas.

6.1. Independencia

Sí que lo haremos de una relación muy importante, aunque ya conocida, entre variables aleatorias: la independencia. El gráfico anterior representaba dos variables aleatorias dependientes: valores bajos de la una correspondían a valores bajos de la otra, etc. Conocer el valor de una de ellas acotaba el rango de valores que podía tomar la otra.

Sin embargo, podemos generar una muestra similar, preservando las distribuciones marginales pero de manera que las variables aleatorias sean independientes:



En este nuevo caso, conocer X no aporta información en absoluto sobre Y . Como antes, esta descripción cualitativa de la probabilidad puede formalizarse en términos de la probabilidad condicional así:

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

Lo cual tiene como consecuencia directa que

$$f(x,y) = f(x)f(y)$$

y, en general, si las variables aleatorias X_1, \dots, X_n son independientes entre sí, entonces su función de densidad conjunta puede factorizarse de la siguiente forma:

$$f(x_1, \dots, x_n) = \prod_i f(x_i)$$

6.2. Teorema de Bayes

Reiteramos que la independencia, aunque importante, es una relación poco interesante entre variables aleatorias. El objetivo de la ciencia de datos es extraer información acerca de una variable aleatoria desconocida Y en función de otras conocidas X_i y eso es imposible si $Y \perp X_i$.

Vamos a ilustrar cómo explotar relaciones de dependencia simples a través de un ejemplo. Se trata de un juego en el que alguien elige al azar un valor $\theta \in [0, 1]$ y luego tira al aire cinco veces una moneda con $P(H) = \theta$. Si nos comunican el número de caras, ¿qué podemos decir acerca de θ ? De alguna manera, la información adicional, aunque indirecta, debería influir en nuestra idea acerca del valor (recuérdese: desconocido) de θ .

Este parece un ejercicio ocioso pero es la base de muchos problemas reales: un sujeto puede tener una propensión θ desconocida a realizar cierto tipo de acciones (de resultado binario) e interesa estimar θ .

La distribución del número de caras para un valor dado de θ (es decir, condicionado a θ) es binomial con parámetros 5 y θ . Matemáticamente, tenemos una variable aleatoria N con distribución condicional

$$N|\theta \sim \text{Binom}(5, \theta).$$

Conocemos pues $P(N = n|\theta)$, pero queremos conocer $P(\theta|N = n)$, es decir, la distribución de θ sabido que $N = n$. Podemos escribir

$$P(\theta|N = n) = \frac{P(N = n, \theta)}{P(N = n)} = \frac{P(N = n|\theta)P(\theta)}{P(N = n)}$$

para obtener de nuevo la expresión conocida como teorema de Bayes. En ella podemos sustituir $P(\theta)$ por 1 (y por 0 fuera del intervalo $[0, 1]$) dado que θ es uniforme en $[0, 1]$. También $P(N = n|\theta)$ por $\binom{5}{n}\theta^n(1 - \theta)^{5-n}$. Finalmente, $P(N = n)$ nos da un poco igual porque no depende de θ , así que podemos reemplazarlo por una constante de la que nos ocuparemos después.

Con esas sustituciones,

$$P(\theta|N = n) = \begin{cases} C\binom{5}{n}\theta^n(1 - \theta)^{5-n} & \text{si } \theta \in [0, 1] \\ 0 & \text{en otro caso} \end{cases}$$

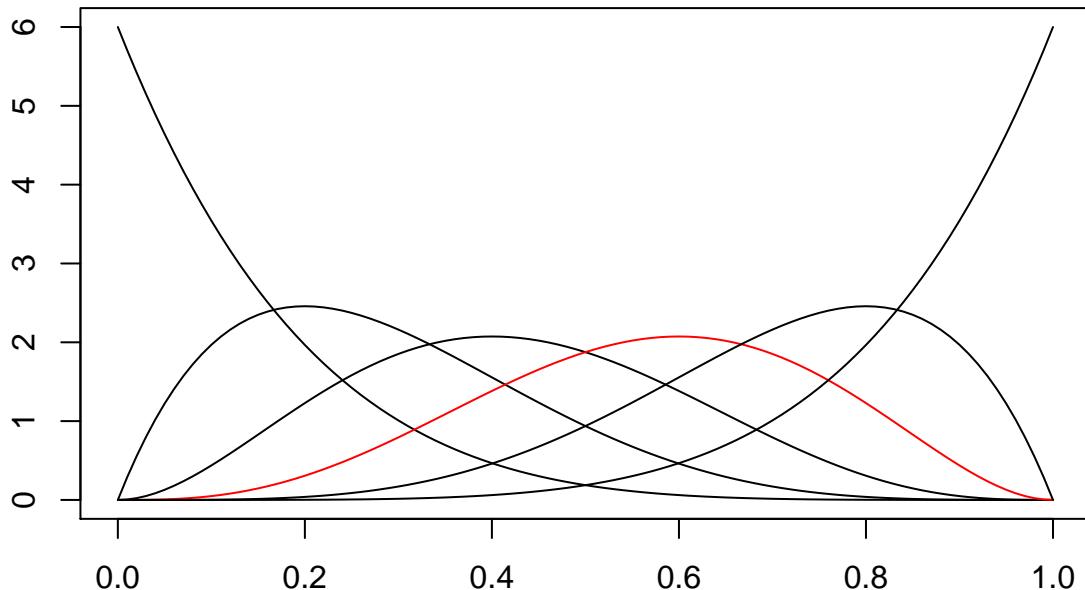
expresión donde todos los términos que no dependen de θ se han absorbido, tal vez con un abuso de la notación, en la constante C . Dado que la expresión obtenida es una función de densidad (de $\theta|N = n$), C tiene que ser tal que

$$C \int_0^1 \theta^n(1 - \theta)^{5-n} d\theta = 1$$

Se puede calcular C (aunque luego veremos que, realmente, no hace falta) y representar las 6 posibles distribuciones de densidad asociadas a θ según del valor de N :

```
curve(dbeta(x, 1, 6), main = "Distribuciones posteriores del parámetro según n", xlab = "", ylab = "")
curve(dbeta(x, 2, 5), add = T)
curve(dbeta(x, 3, 4), add = T)
curve(dbeta(x, 4, 3), add = T, col = "red")
curve(dbeta(x, 5, 2), add = T)
curve(dbeta(x, 6, 1), add = T)
```

Distribuciones posteriores del parámetro según n



En la gráfica se ha marcado en rojo la función de densidad correspondiente al caso $N = 3$, que se concentra alrededor del valor $3/5$. Por su forma, daremos más credibilidad a los valores próximos a $3/5$ de θ que a los más alejados de él.

No hace falta calcular el valor de C en la expresión anterior si se reconoce la distribución de $P(\theta|N = n)$. Por la forma de la función de densidad, se trata de una distribución beta (de parámetros $n + 1$ y $6 - n$). Esto no es casualidad: es un hecho conocido que la distribución beta es una **distribución conjugada** de la binomial. Eso significa que siempre que la información *a priori* sobre el parámetro tenga distribución beta (y la distribución uniforme es un caso especial de la beta) y nuestras observaciones sigan una distribución binomial, la distribución resultante (conocida como distribución **a posteriori**) será también beta.

Existen otras parejas de distribuciones conjugadas que puede ser útil conocer.

6.3. Covarianza y correlación

La relación entre dos variables aleatorias (continuas en este caso) puede medirse a través de su covarianza:

$$\sigma(X, Y) = \int (x - E(X))(y - E(Y))f(x, y) dx dy$$

Si $X \perp Y$, entonces es fácil probar que $\text{cov}(X, Y) = 0$. La relación inversa es falsa. Solo se cumple en algunos casos concretos, como cuando las variables aleatorias involucradas son normales. Pero en general, vale la pena repetirlo, no es cierta.

En general, si X e Y tienden a tomar valores altos o bajos a la vez, su covarianza será positiva. Si valores altos de X tienden a corresponder con valores bajos de Y (y a la inversa), la covarianza será negativa.

Por definición, además, $\sigma(X, X) = \sigma^2(X)$.

La covarianza está afectada por el valor relativo de las variables X e Y y puede tomar valores arbitrariamente altos (o bajos). Por eso, como medida de la relación entre dos variables es preferible el **coeficiente de correlación**, que se define así:

$$\rho(X, Y) = \frac{\sigma(X, Y)}{\sigma(X)\sigma(Y)}$$

El coeficiente de correlación toma valores entre -1 y 1 . Solo es 1 (-1) cuando $X = aY + b$ y $a > 0$ ($a < 0$), es decir, cuando las variables son transformaciones lineales la una de la otra. En esos casos *degenerados*, conocer una de ellas da información completa sobre la otra. Conforme más alto (en valor absoluto) es la correlación, más información proporciona conocer una de ellas sobre la otra. Por eso, en ocasiones, se utiliza el coeficiente de correlación para seleccionar variables predictoras para modelos (aunque hay técnicas mucho mejores para eso).

6.4. Referencias

- Distribuciones conjugadas

6.5. Ejercicios

7. Introducción a la estadística

Como hemos indicado más arriba, la teoría de la probabilidad es deductiva: o conocemos las probabilidades asociadas a los eventos son o podemos calcularlas aplicando ciertas reglas. La estadística trata de recorrer el camino inverso inductivamente: a partir de la observación de ciertos fenómenos aleatorios trata de revelar el mecanismo aleatorio subyacente.

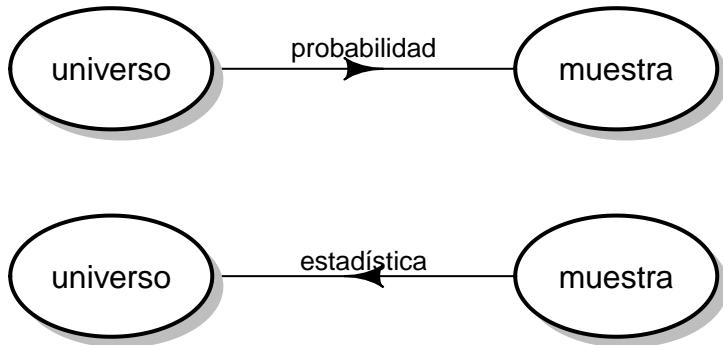
7.1. Universo, muestra y big data

Los conceptos de **universo** y **muestra** permiten ilustrar esa diferencia fundamental. En probabilidad se conoce (o se pretende conocer) el **universo**. Por ejemplo, puede *saberse* que en un país de 47 millones de habitantes hay 4 en paro. A partir de esos datos, puede deducirse la probabilidad de que en una **muestra** de cinco mil de ellos haya un determinado número de parados: se puede aproximar por una distribución binomial con una probabilidad de *éxito* conocida.

La estadística, sin embargo, se plantea un problema más real y de verdadera importancia práctica: a partir del número de parados en una muestra, estimar el de los existentes en la población total, en el universo.

Los universos no se extienden únicamente a través de poblaciones: también lo hacen, y esto es importante, en el tiempo y sobre lo contingente, lo que tal vez no ocurra nunca: cuando un probabilista asigna una probabilidad de 0.7 al evento de que un determinado futbolista marque un penalti, está refiriéndose implícitamente a un universo que incluye tanto los penaltis tirados como los aún no tirados por dicho futbolista. Lo mismo ocurre cuando se habla de la probabilidad de que un cliente *convierta* (donde se incluyen a los clientes del día de mañana) o a la de obtener cara en un lanzamiento de moneda (aun cuando no se realice ninguno).

Sin embargo, las muestras son concretas: los penaltis lanzados durante las últimas tres temporadas, el histórico de visitas y conversiones durante los últimos meses o el número de caras obtenidas en quinientos lanzamientos de una determinada moneda.



El *big data* ha replanteado la distinción entre universo y muestra. Si tienes *todos* los datos, si tu muestra es igual al universo, ¿cabe aún distinguir entre la una y el otro? ¿Tiene eso impacto sobre la inferencia, i.e., el proceso de razonar desde lo concreto a lo general?

En el caso del país y el paro descrito más arriba, tal vez. Podría pensarse en un sistema que automáticamente y fidedignamente pudiese proporcionar el número de parados (o el número de fumadores de marihuana) en un país en tiempo real. Sin embargo, en muchos otros casos de interés, *el universo no cabe en el big data*: no podría existir un registro de fenómenos jamás observados, como el efecto de una bajada de precios de un producto en sus ventas; o como las noticias que más pudieran interesar mañana a un lector determinado de un periódico; o, incluso, qué enfermedad puede tener un paciente que muestra un determinado cuadro clínico (antes de realizar pruebas adicionales). Porque el universo incluye eventos futuros, potenciales o nunca observados como el efecto de una intervención potencial o un comportamiento futuro. Además, frecuentemente, disponer de *toda* la información puede ser prohibitivamente caro: piénsese en pruebas médicas.

Por eso, siempre tendremos que tomar decisiones bajo incertidumbre, es decir, teniendo una visión parcial del universo. Que es, esencialmente, el objeto de la estadística.

7.2. El objeto de la estadística

Aunque existen precedentes previos de censos y otras *operaciones* estadísticas, la estadística (e incluso, su propio nombre: *Statistik*) nació en Alemania a mediados del s. XVIII y estaba relacionada con la recogida sistemática de información económica, demográfica, etc. por parte del estado. Existe una diferencia importante entre estas colecciones de datos y otras previas: por su propia estructura, formato y presentación estaban preparadas para la comparación. Típicamente, la información se resumía en tablas que permitían la comparación de variables climáticas, agropecuarias, económicas, demográficas, etc. entre las distintas unidades territoriales.

Esta *aritmética política* se convirtió en *aritmética social* en el siglo XIX tanto en Inglaterra como en Francia. En ambos países, aunque desde perspectivas y con intereses distintos, comenzaron a acumularse datos de interés social: censos, causas de muerte, impuestos, etc. Estos primeros estadísticos comenzaron a apreciar regularidades sorprendentes en los datos; por ejemplo, en la tasa de suicidios, que tenía a mantenerse tozudamente constante a través del tiempo. Estas regularidades y su estudio dieron lugar a disciplinas como la sociología. Pero también a un incipiente desarrollo matemático de la estadística.

La *estadística clásica*, la que se enseña en los cursos de iniciación, sin embargo, es un invento británico del periodo que aproximadamente va de 1880 a 1940 y está muy relacionada con los estudios agrícolas, aplicaciones industriales, etc. que proponían problemas de otra naturaleza: ¿es este tratamiento efectivo?, ¿son los rendimientos de estas semillas significativamente diferentes?, etc. Es en ese contexto que nacieron conceptos fundamentales como las pruebas de hipótesis, los p-valores, el análisis de la varianza, etc.

La *estadística moderna*, muy reciente, es la que construye sobre la estadística clásica pero, a diferencia de aquella, usa ordenadores. La estadística clásica y la moderna comparten problemas y objetivos, pero se diferencian en la forma. Muchos de los aspectos más abstrusos de la estadística clásica y que colean todavía en textos modernos obedecen a un motivo: responden a la falta de capacidad de cálculo en la época que los vio nacer. Es decir, resuelven un problema que hoy hemos aprendido a resolver de otra manera.

La emergencia de los ordenadores ha permitido, entre otras cosas, establecer una relación muy fructífera con otra disciplina que a veces se quiere ver confrontada a ella, la del aprendizaje automático (*machine learning*; antes *data mining*). Y también la del redescubrimiento de los métodos bayesianos, muy exigentes computacionalmente, pero que están adquiriendo una importancia fundamental hoy en día.

En cualquier caso, el hilo conductor de la estadística moderna, del s. XVIII hasta nuestros días, es el del desarrollo de técnicas para lograr dos objetivos fundamentales:

- Comparar
- Tomar decisiones

Estos objetivos están en conflicto con una visión reduccionista de la estadística (o, más propiamente, del análisis estadístico de datos) que es la que prevalece en libros y programas académicos. Para poder alcanzar los objetivos anteriores, esta visión debería trascenderse para incluir desde la adecuada recolección de datos hasta la discusión en términos económicos, sociales o políticos de sus resultados de cara a la toma de las decisiones últimas.

8. Estadística descriptiva

La estadística descriptiva comprende un conjunto de técnicas y procedimientos para un análisis primero de los datos que tiene como objetivo familiarizarse con ellos y descubrir y describir sus principales características. Es inseparable de la **visualización de datos**, casi una disciplina en sí misma pero que a menudo los mismos estadísticos marginan. Además, está muy relacionada con una disciplina emergente, el EDA (*exploratory data analysis*), que extiende la exploración gráfica de los datos hasta incluir prácticamente la fase de análisis.

8.1. Resúmenes numéricos

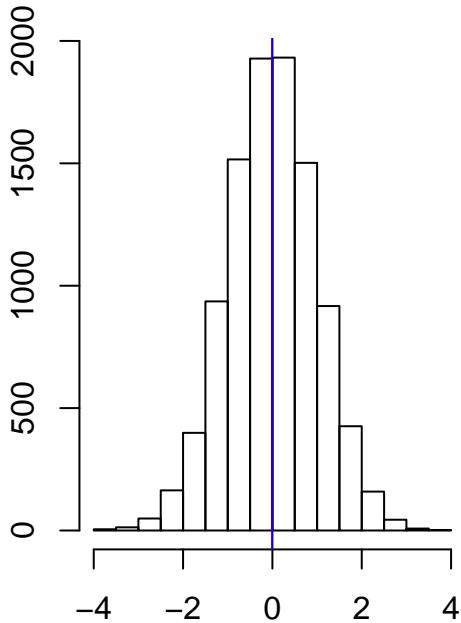
Aunque menos útiles de lo que suele considerarse y de la importancia que se les otorga, los resúmenes numéricos son rápidos y sirven para detectar fenómenos *gruesos* en los datos antes de un análisis más fino. Dados unos datos nuevos, en R, se pueden obtener resúmenes de interés usando, por ejemplo y entre otras funciones:

- **mean**, la media, que es un estadístico sobrevalorado y que puede verse muy severamente afectado por la presencia de *outliers*.
- **var**, la varianza o **sd**, la desviación estándar. Son más importantes cuando los datos son normales (o aproximadamente normales); en otros casos, su importancia como medida de la dispersión es relativa y más difícil de interpretar. Además, como la media (o incluso más que la media) se ven afectadas por los *outliers*.
- **quantile**, para los cuantiles, que son resúmenes más interesantes y que permiten capturar más información de los datos; de entre ellos, el más importante es la mediana, **median**, una medida de centralidad en general mucho más útil y robusta que la media.
- Relacionada con la anterior, **fivenum** proporciona los cinco números de Tukey, un resumen de un vector numérico que incluye el máximo, el mínimo, la mediana y los cuartiles. Es muy interesante porque el máximo y el mínimo, en ocasiones, de no ser razonables, pueden apuntar a problemas serios en los datos. Los cuartiles, por su parte, acotan la masa de la distribución, la región donde es más probable encontrar observaciones.
- **summary**, que para vectores numéricos proporciona los cinco números de Tukey más la media. Tiene la ventaja de que puede aplicarse directamente a tablas y que para columnas categóricas ofrece también la frecuencia de las clases más comunes.

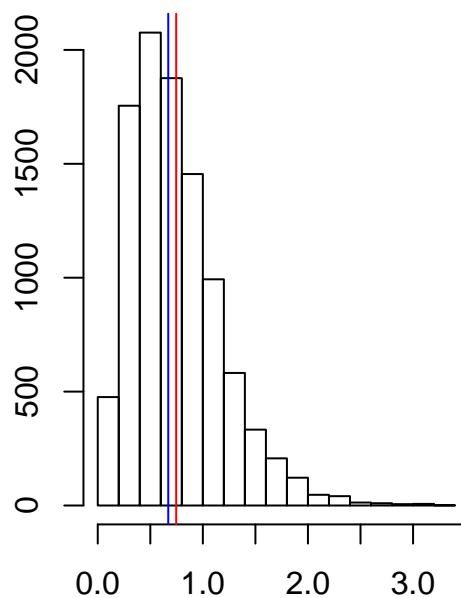
Existen muchas más funciones de ese tipo, pero se resumen en dos tipos según el tipo de información que proporcionen: información sobre la **centralidad** de la distribución o información sobre su **dispersión**. Las medidas de centralidad tratan de ofrecer un resumen de una distribución basado en un único valor, que frecuentemente se interpreta (casi siempre de manera errónea o engañosa) como el *sujeto representativo*. Las

medidas de centralidad más habituales son la media, la mediana y la moda (el valor más frecuente), aunque hay otras como las llamadas *medias winsorizadas*. Para distribuciones simétricas y unimodales (como la normal), todas estas medidas de centralidad coinciden y tienen una interpretación muy natural.

Distribución normal



Distribución gamma



En la gráfica anterior se han mostrado dos distribuciones de datos y se han indicado la media (en rojo) y la mediana (en azul) de los mismos. Para una distribución simétrica y unimodal como la normal, ambos valores coinciden y asumen un valor *notable* de la distribución (aunque, nótese, no es *representativo* de los valores). En el segundo caso, la situación es distinta y media y mediana difieren notablemente. En distribuciones como ella, que tienen una cola larga, la media suele ser superior a la mediana y ninguno de los dos valores puede calificarse de *notable* dentro de la distribución.

TODO: ¿Añadir la moda?

Una medida de centralidad es una descripción mínima de una distribución de probabilidad. Interesa, además, conocer la distribución de los datos alrededor del valor central. Los valores máximo y mínimo, los cuantiles (que pueden considerarse versiones menos *ruidosas* de los anteriores), la varianza y la desviación estándar y muchos otros tratan de dar una idea de como se distribuyen los valores de la muestra o población alrededor del valor central.

Es típico en estadística *caracterizar* una distribución por su media y su varianza (o desviación estándar). Este tipo de caracterización tiene sentido únicamente cuando los datos tienen una distribución normal (nótese cómo la distribución normal está perfectamente caracterizada cuando se conocen su media y su varianza) pero puede ser engañoso para otras. Además de antinatural: estamos acostumbrados en nuestra actividad diaria a describir la variabilidad mediante rangos de valores; si alguien nos pregunta, por ejemplo, cuánto vale una botella de vino en determinado restaurante, nunca contestaríamos con el valor medio y su desviación estándar; es mucho más natural e informativo ofrecer una *horquilla* de valores: *lo normal es que cueste entre tal y tal precio*.

De hecho, si hubiese una ley prohibiendo el uso de más de una cifra para describir distribuciones, la recomendación sería usar una medida de centralidad. Si se permitiesen dos, un rango de valores típicos. Antiguamente, con las limitaciones de ancho de banda de las tecnologías de intercambio de información, era imperativo usar ese tipo de resúmenes. Sin embargo, hoy en día, es posible mostrar la distribución (prácticamente) completa usando las técnicas de representación gráfica que se discutirán en las siguientes secciones.

Finalmente, y por su relación el tipo de resúmenes numéricos discutidos más arriba, está el problema de su **tabulación** para ser presentados a terceros o, más propiamente, su *adecuada* tabulación. Las tablas de datos son frecuentemente alternativas válidas (y recomendadas) a ciertos gráficos. Pero crear tablas efectivas es un arte que pocas veces se ejerce con el dominio que requiere. Aunque abundar sobre las características deseables de una buena tabla queda al margen del alcance de estas páginas, en las referencias se incluyen enlaces a algunos manuales valiosos.

8.2. Visualización de datos

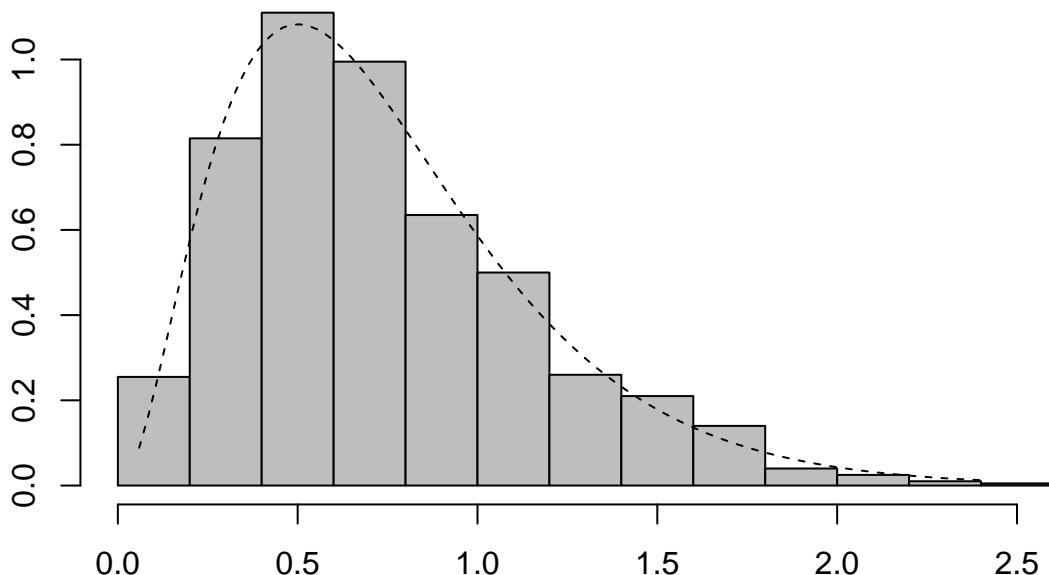
En esta sección vamos a repasar algunas visualizaciones clásicas y menos clásicas de información cuantitativa. En primer lugar y como evolución de los resúmenes estadísticos básicos de la sección anterior, presentaremos los **histogramas**. Los histogramas describen la forma de una distribución y dan cuenta no solo de sus valores centrales (i.e., los que indicarían la media o la mediana) o su dispersión (a través de la varianza o la desviación estándar) sino, además, su forma, el aspecto de sus colas y sus valores máximo y mínimo.

Es, casi indudablemente, el primer resumen estadístico (previo incluso a los cuantitativos de la sección anterior) que realizar sobre un vector de datos de interés. Además, en R, prácticamente con el mismo esfuerzo (medido, si se quiere, en términos del número de teclas pulsadas), se obtiene un resumen mucho más informativo.

En el gráfico que aparece a continuación se representa el histograma de una muestra de la distribución gamma y se le superpone la gráfica de la función de densidad correspondiente.

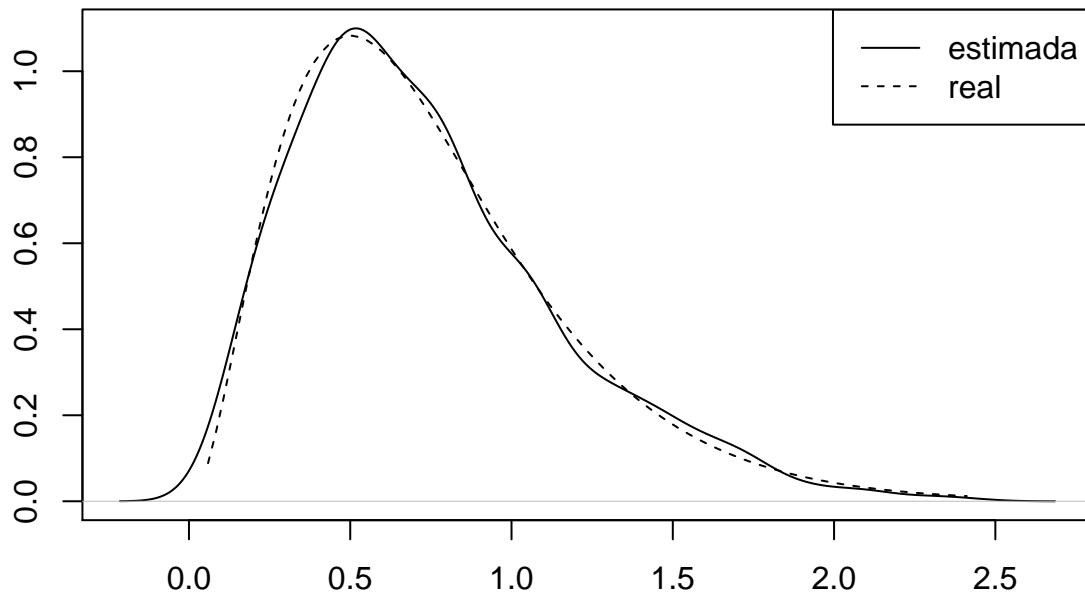
```
x <- rgamma(1000, 3, 4)
hist(x, main = "Histograma de una muestra de una Gamma(3,4)",
      probability = TRUE,
      xlab = "", ylab = "", col = "gray")
curve(dgamma(x, 3, 4), min(x), max(x), add = T, lty = 2)
```

Histograma de una muestra de una Gamma(3,4)



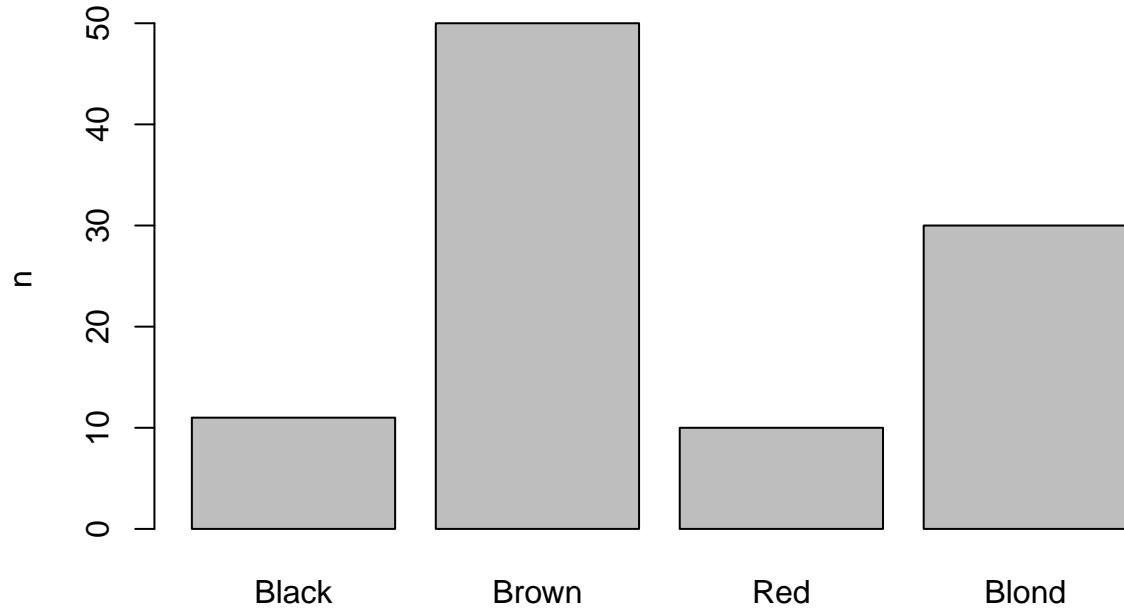
Los histogramas, en el fondo, son herencia de una época en que la capacidad computacional no era la misma que hoy en día: son representaciones sencillas de hacer incluso a mano. Una alternativa a ellos son las representaciones directas de la **curva de densidad empírica estimada**.

Densidad de una muestra de una Gamma(3,4)



Las dos técnicas anteriores permiten visualizar vectores numéricos. Para información categórica, la representación más habitual es la de los **diagramas de barras**.

Número de sujetos por color del pelo



En estas páginas omitiremos la discusión sobre la conveniencia o no de ordenar las barras por tamaño o la conveniencia o no de apilar o adjuntar barras para realizar comparaciones más complejas en que intervienen más variables. Sí que recogeremos, sin embargo, el parecer de quienes opinan que las barras son representaciones excesivamente aparatosas para la escasa información que proporcionan: en el gráfico anterior, apenas 4 números y sus correspondientes etiquetas.

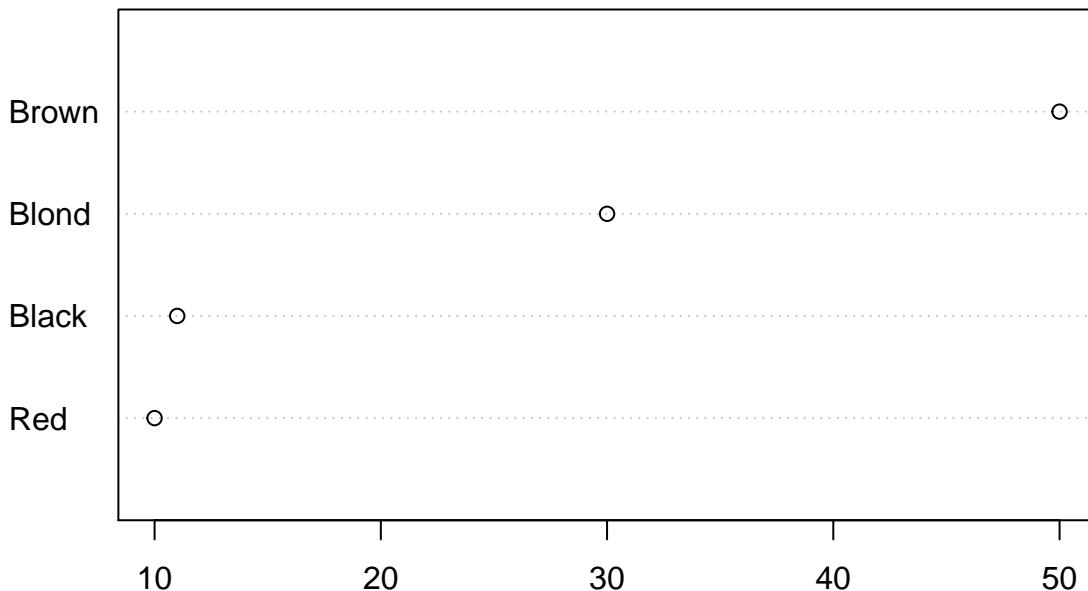
Como alternativa a ellas, sugieren los gráficos de puntos en que se reemplaza la barra completa por un único punto.

```

tmp <- HairEyeColor[, 2,1 ]
dotchart(sort(tmp), main = "Número de sujetos por color del pelo")

```

Número de sujetos por color del pelo



Una diferencia notable entre las barras y los puntos es que las primeras exigen, para una correcta interpretación de los datos, que estén basadas en el nivel 0. Subrayan, por tanto, las medidas absolutas. En cambio, para los diagramas de puntos esa restricción se suaviza y pueden representar mejor las variaciones relativas (p.e., entre medidas muy parecidas, con variaciones del orden del 1% entre ellas, que serían inapreciables con barras basadas en cero).

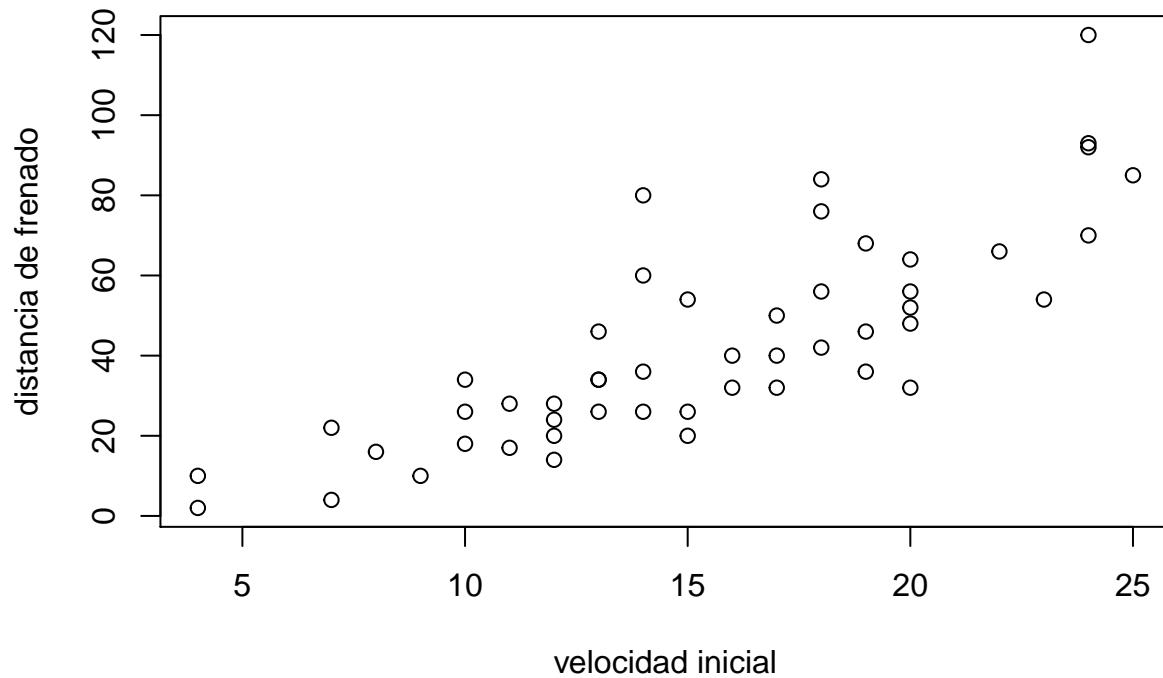
Los gráficos anteriores sirven para representar una única variable aleatoria, sea continua o categórica. En ocasiones hay que comparar dos de ellas. Si son numéricas, el gráfico por excelencia es el de **dispersión**:

```

plot(cars$speed, cars$dist,
      xlab = "velocidad inicial",
      ylab = "distancia de frenado",
      main = "Distancia de frenado en función de la velocidad inicial")

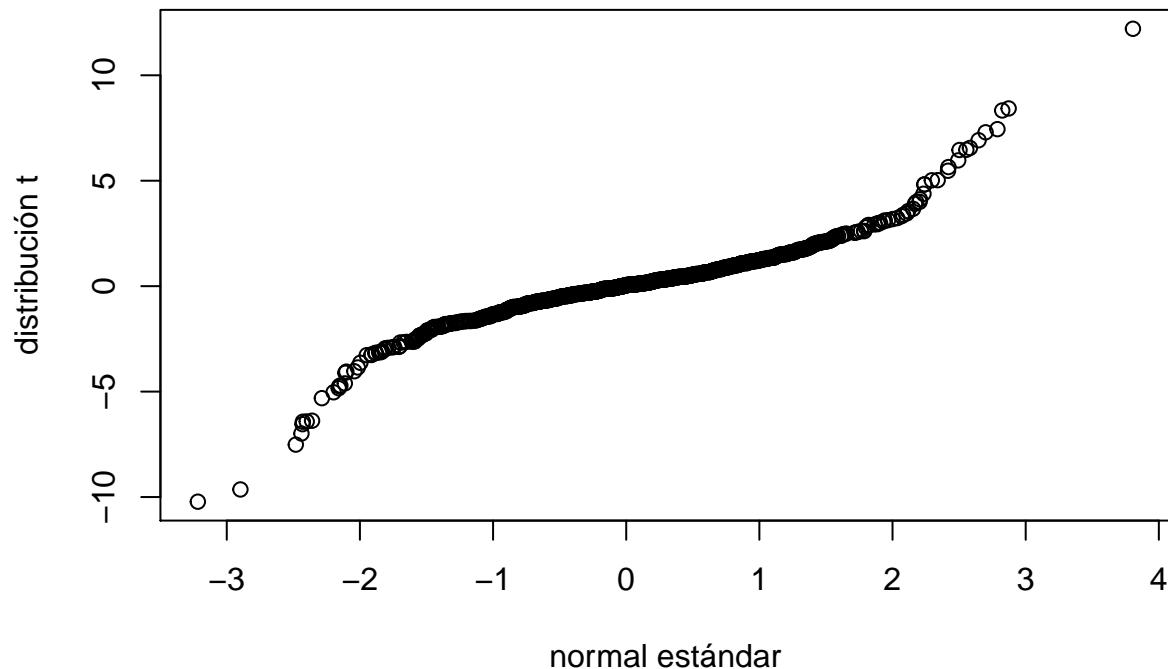
```

Distancia de frenado en función de la velocidad inicial



Sin embargo, a veces solo interesa comparar las distribuciones de dos variables. Para eso son útiles los **gráficos de cuantiles** (o *qq*), como el siguiente:

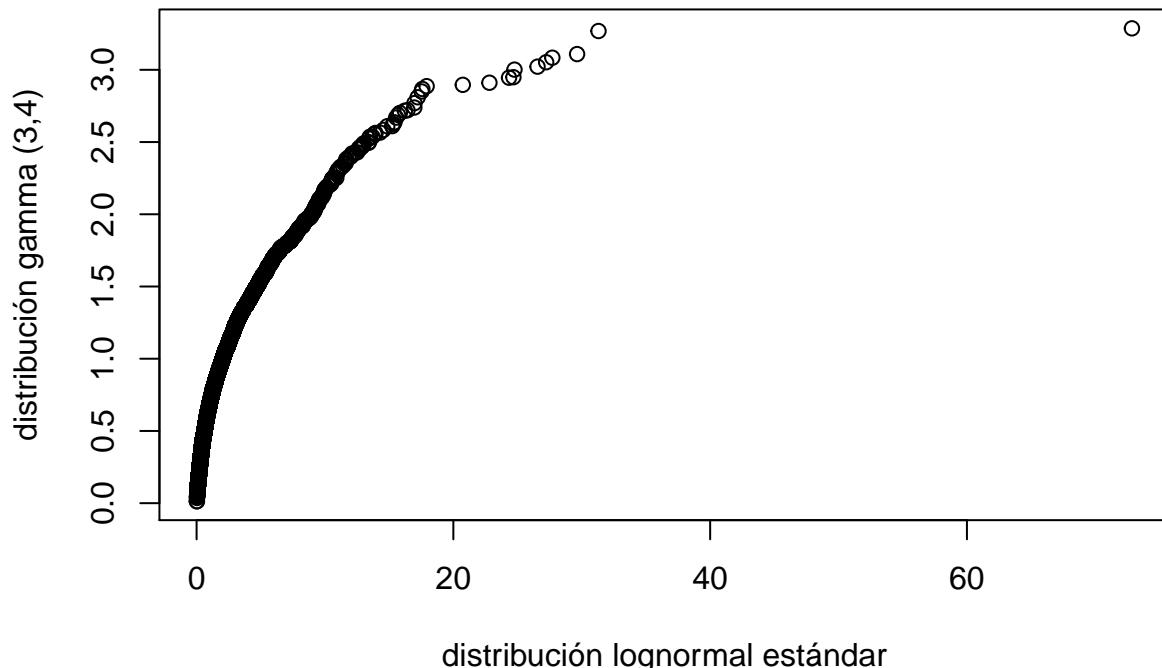
```
qqplot(rnorm(1000), rt(1000, 3),
       xlab = "normal est\'andar", ylab = "distribuci\'on t")
```



En él se comparan los cuantiles de dos muestras: una normal est\'andar y una distribuci\'on *t* con 3 grados de libertad. En la zona central la gr\'afica es casi una recta, lo que significa que las distribuciones son muy

parecidas (salvo en la escala: los cuantiles que para la normal recorren los valores entre -2 y 2, cubren el intervalo entre -5 y 5 para la distribución t). La gran diferencia se observa en las colas: las de la t son más gruesas y eso se refleja en la característica curvatura en los extremos de la gráfica: los cuantiles de la distribución t crecen más aprisa que los de la normal y eso curva la gráfica.

El gráfico siguiente muestra otro ejemplo en el que se comparan las distribuciones gamma y lognormal. Se aprecia cómo la cola de la segunda es más pesada que la de la primera, tiene una mayor dispersión.

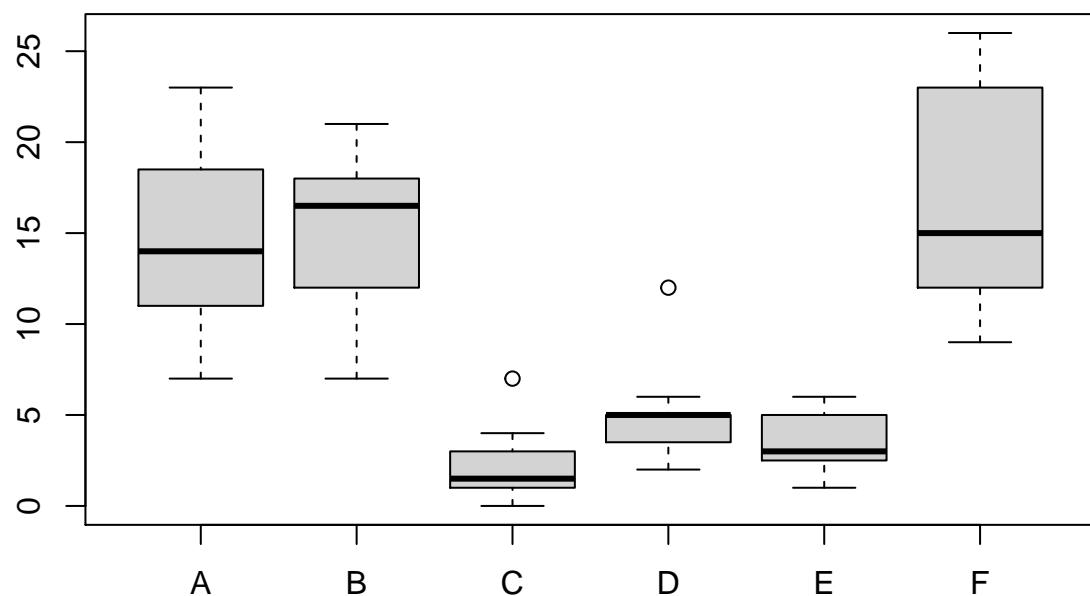


En cualquier caso, hay que tener cuidado con extraer conclusiones acerca de las colas de distribuciones con este tipo de representaciones gráficas porque, por definición, en las colas hay pocas observaciones, y, como consecuencia, la inestabilidad de la representación en los extremos puede ser grande.

La representación gráfica estándar para representar conjuntamente un vector continuo y otro categórico son los **diagramas de caja**:

```
boxplot(count ~ spray, data = InsectSprays, col = "lightgray",
       main = "Número de insectos según el tipo de tratamiento")
```

Número de insectos según el tipo de tratamiento



Los diagramas de caja están íntimamente relacionados con los cinco números de Tukey (además, este tipo de diagramas son invención de ese estadístico). Muestran la distribución de la variable continua a través de los distintos niveles de la categórica resumiendo su distribución y disponiéndola en paralelo al resto. Así se pueden comparar sus valores relativos.

Las cajas son representaciones sucintas de la distribución de la variable continua. Reflejan su rango intercuartílico (límites de la caja), donde se acumula la masa de la distribución, su mediana (trazo horizontal grueso) y unos indicadores, los bigotes, que indican hasta donde se extienden los valores no considerados anómalos. Estos, de ocurrir, se representan como puntos aislados.

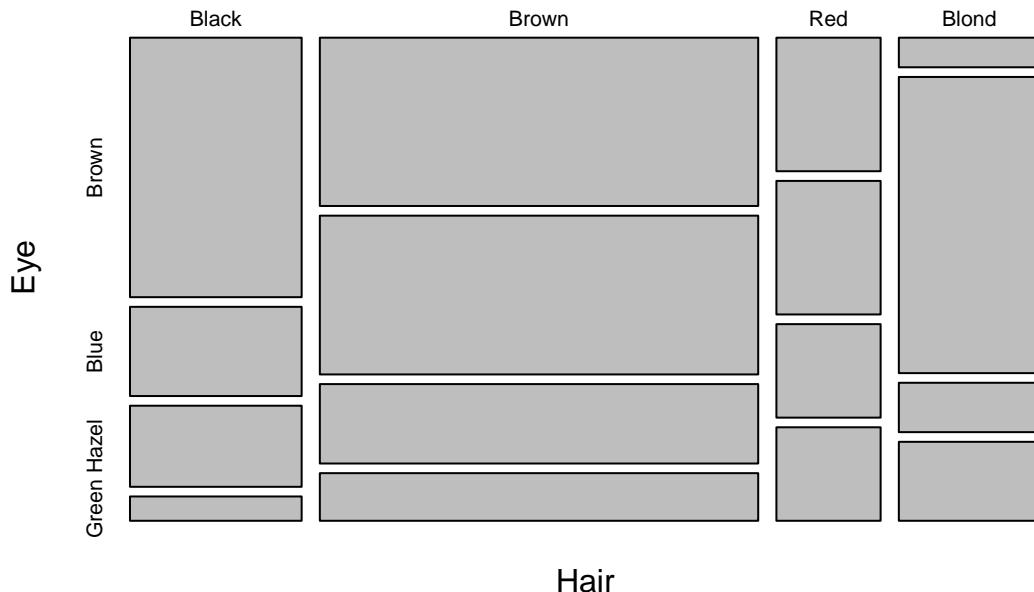
De hecho, en la gráfica anterior aparece un valor anómalo (>10) para la categoría D. Ese valor no parece atípico dentro de la distribución agregada de los datos (es parecido a muchas observaciones de las categorías A, B y F), pero destaca cuando se lo compara únicamente con los de su propio grupo. Lo mismo puede decirse de la observación anómala (>5) de la categoría C.

Características adicionales de la distribución (como la bimodalidad, de existir) quedan desdibujados por esta representación tan sencilla como efectiva. Por eso, en algunas circunstancias, son preferibles los gráficos de violín (ver referencias) o la superposición de las densidades.

Cuando ambas variables son categóricas, pueden usarse barras (sean apiladas o yuxtapuestas) o los llamados **gráficos de mosaico**, una de cuyas variantes más básicas es:

```
mosaicplot(HairEyeColor[,1], main = "Color del pelo según el de los ojos")
```

Color del pelo según el de los ojos

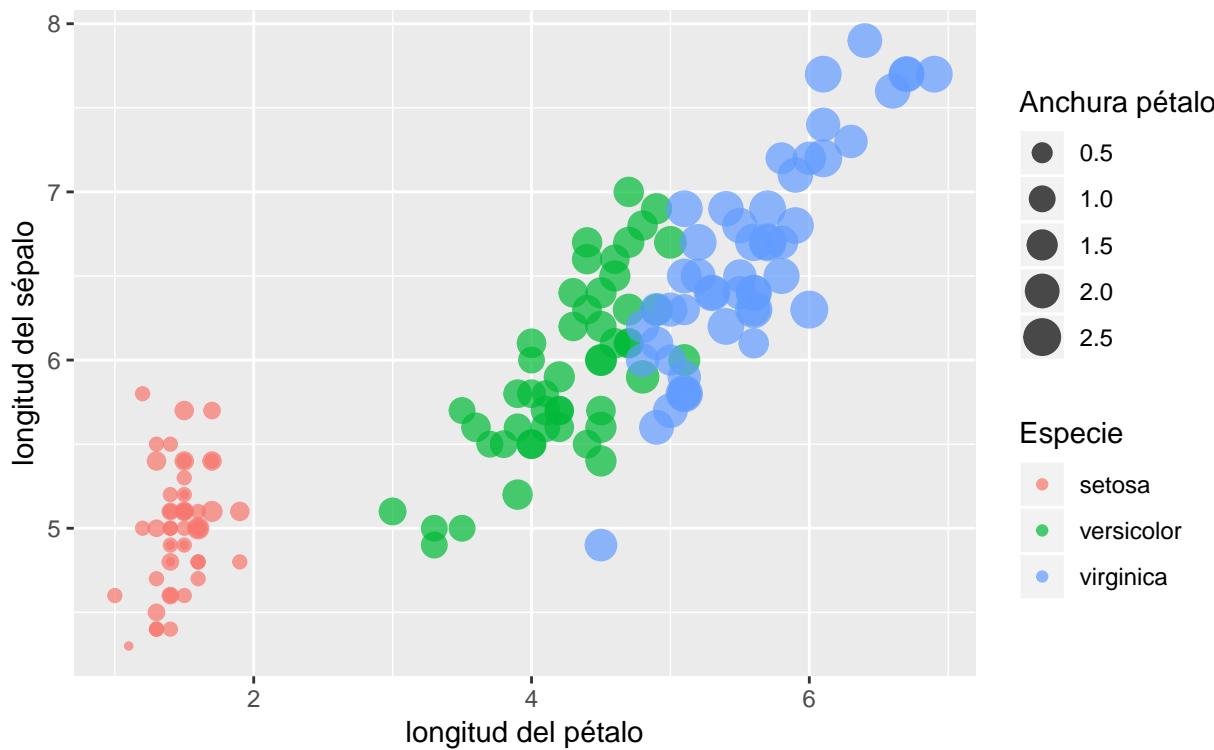


Existen alternativas que usan gradientes de color para indicar excesos o defectos relativos de conteos con respecto a los esperados, algo sobre lo que se abundará más adelante, al tratar los datos tabulares.

De todos modos, los anteriores son gráficos excesivamente simples que muestran a las variables una a una o por parejas. Sin embargo, en el estudio de fenómenos reales, es conveniente en ocasiones examinar el efecto simultáneo de varias variables y existen técnicas para mostrar relaciones más complejas, como el siguiente:

```
library(ggplot2)

ggplot(iris, aes(x = Petal.Length,
                  y = Sepal.Length,
                  size = Petal.Width,
                  color = Species)) +
  geom_point(alpha = 0.7) +
  xlab("longitud del pétalo") +
  ylab("longitud del sépalo") +
  guides(color = guide_legend(title = "Especie")) +
  guides(size = guide_legend(title = "Anchura pétalo"))
```

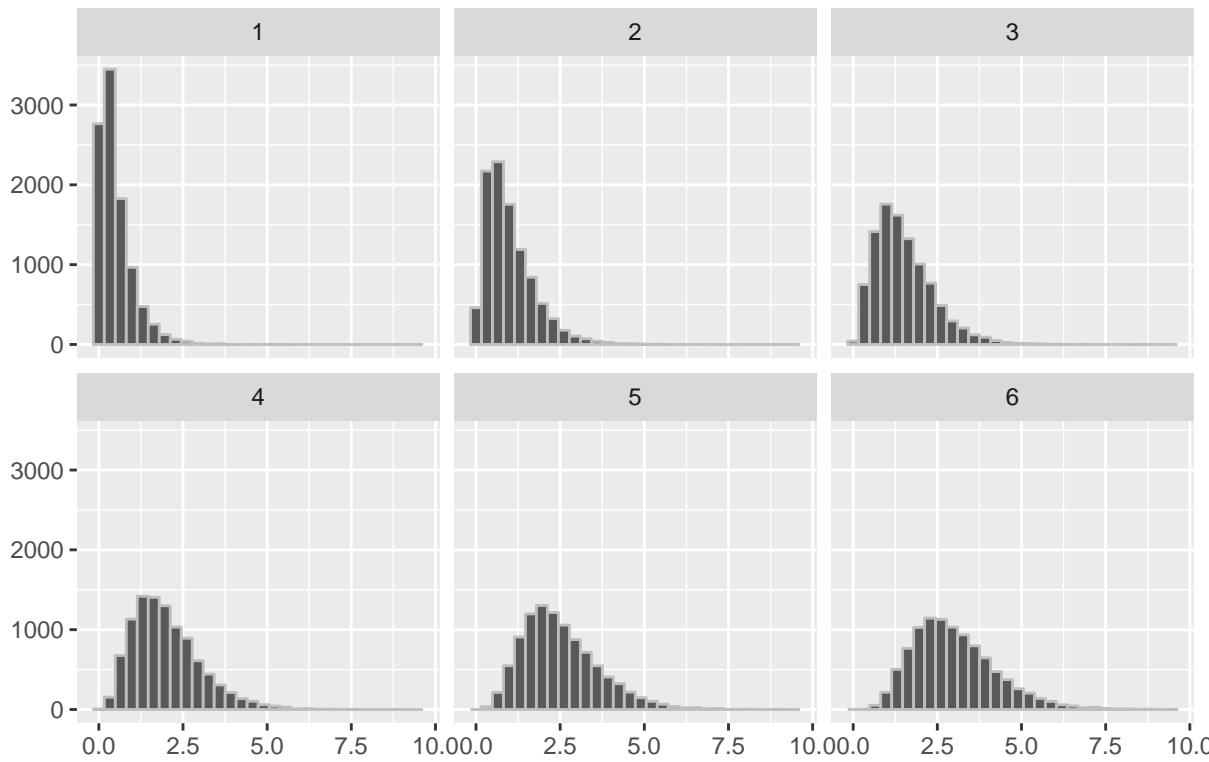


En el se muestran cuatro variables simultáneamente y se aprecia cómo la especie *setosa* está claramente diferenciadas del resto y cómo pétalos y sépalos son claramente más pequeños que los de las otras.

Una de las maneras más eficaces de incluir variables adicionales en un gráfico es usando la técnica conocida como de los **pequeños múltiplos**. La técnica de los pequeños múltiplos sugiere dividir el lienzo en varios paneles dispuestos en una retícula y ordenados de una determinada manera para poder apreciar los cambios en una representación gráfica básica a través de los niveles de otra u otras variables.

En el ejemplo que aparece a continuación se muestran los histogramas de muestras de una distribución gamma a lo largo de diversos valores de uno de sus parámetros.

Histograma de una Gamma(i , 2)



Los pequeños múltiplos son alternativas tanto a la superposición de figuras como, en algunos casos, a las animaciones. El principal problema de estas últimas, a pesar de su creciente popularidad en un mundo cada vez más audiovisual, es que ordenan las vistas a los datos en el tiempo y no en el espacio, lo que dificulta las comparaciones.

Para terminar, un área que está cobrando gran importancia hoy en día es el de los gráficos interactivos. Permiten tanto la incorporación de muchas variables simultáneamente como la exploración de la información por parte del usuario. Tienen ciertas ventajas sobre los gráficos estáticos, frecuentemente infravalorados, pero resultan muy eficaces en determinados contextos.

A pesar de su interés, en estas páginas no discutiremos más los últimos tipos de gráficos, que exigirían una monografía específica. Eso sí, se invita al lector a explorar y familiarizarse con esas técnicas y las herramientas necesarias.

8.3. Análisis exploratorio de datos (EDA)

El análisis exploratorio de datos (o EDA, de *exploratory data analysis*) es un conjunto de técnicas para explorar e incluso modelar datos usando fundamentalmente técnicas gráficas de manera interactiva.

Muchas de estas técnicas se basan en la representación simultánea de los mismos datos desde distintas perspectivas con herramientas específicas. Estas herramientas permiten realizar dos operaciones muy potentes y relacionadas.

- **Enlazado de datos** (*linked data*): al seleccionar interactivamente un conjunto de datos en una de las perspectivas, las observaciones correspondientes se marcan también en el resto. Esto permite por ejemplo seleccionar una categoría determinada en un diagrama de barras y ver qué región del espacio ocupan esos puntos en una serie de diagramas de dispersión enlazados. O, a la inversa, ver a qué categoría pertenecen unas observaciones destacadas en un diagrama de dispersión.

- **Brushing:** típicamente, sobre un diagrama de barras, se puede aplicar una paleta de colores que se transmiten al resto de las perspectivas. Es similar al enlazado de datos, solo que aplica simultáneamente para todas las categorías.

Existen libros enteros dedicados al EDA y herramientas libres como GGobi o Mondrian con las que familiarizarse con estas técnicas.

8.4. Referencias

- *What is a good resource on table design?*, con comentarios y enlaces sobre el diseño efectivo de tablas.
- *Grahping resources: designing tables*, con consejos para la creación de tablas.
- *Too many digits: the presentation of numerical data*, sobre cuántos dígitos mostar en tablas y otros tipos de resúmenes estadísticos.
- *Is it meaningful to talk about a probability of “65.7%” that Obama will win the election?*, sobre cómo el redondeo permite transmitir información sobre la incertidumbre asociada a mediciones y predicciones.
- *40 years of boxplots*, donde se discute la historia de los diagramas de caja así como alternativas recientes a los mismos, incluyendo los gráficos de violín.
- Dos técnicas fundamentales para gráficos dinámicos de datos: *linking* y *brushing*
- Superposición de densidades
- Mondrian, una herramienta desarrollada en Java para el análisis exploratorio de datos (EDA)
- GGobi, una herramienta para el análisis exploratorio de datos (EDA)

8.5. Ejercicios

8.5.0.1. Ejercicio

Probar que la media de una muestra de observaciones independientes de una distribución $N(\mu, \sigma)$ tiene media μ y desviación estándar σ/\sqrt{n} . (Nota: además de eso, ya sabemos que la distribución de esa media es normal).

9. Estimación puntual

El objetivo de la estimación puntual es el de obtener estimaciones razonables de parámetros de alguna distribución (p.e., la media) desconocida (o parcialmente desconocida) de la que se tiene una muestra, así como las correspondientes estimaciones del error cometido. Un ejemplo paradigmático de la estimación puntual es el de la estimación de la tasa de paro, que es una magnitud desconocida θ . Para ello, sobre la población española se realiza una encuesta, la EPA, con la que se construye un estimador $\hat{\theta}$ de θ .

En este capítulo presentaremos varios métodos de para la estimación puntual de parámetros. Comenzaremos con uno de los más antiguos, el de los momentos, para seguir con los basados en funciones de pérdida, uno de los favoritos de la ciencia de datos, que nos conducirá a los M-estimadores. Terminaremos con los basados en la función de verosimilitud y la estimación por máxima verosimilitud.

Finalmente, nos ocuparemos del error cometido y de la convergencia de estos estimadores a sus valores *reales*.

Queda fuera de este capítulo, y no porque sea irrelevante sino porque se cubrirá en uno posterior específico, la perspectiva bayesiana del problema.

9.1. El método de los momentos

Históricamente, es el primero (aunque con salvedades) que se propuso para resolver este problema. Fue propuesto por Karl Pearson en 1894.

Podemos ilustrarlo con un ejemplo. Podemos suponer que se tiene una muestra de una variable aleatoria con distribución $X \sim \Gamma(a, b)$ y el problema consiste en estimar los parámetros desconocidos a y b . Como sabemos que $E(X) = a/b$ y $\sigma^2(X) = a/b^2$, podemos calcular la media y la varianza muestral, plantear las ecuaciones correspondientes y despejar nuestras estimaciones \hat{a} y \hat{b} .

En lugar de resolver este problema analíticamente, podemos plantearlo en R directamente. En primer lugar, extraeremos una muestra de una distribución gamma con parámetros 3 y 4 y calcularemos la media y la varianza de esta muestra:

```
muestra <- rgamma(1000, 3, 4)
media    <- mean(muestra)
varianza <- var(muestra)
```

Después usaremos la función `nleqslv` del paquete homónimo para resolver el sistema de ecuaciones correspondiente. En realidad, la función no resuelve ecuaciones sino que encuentra ceros de funciones, por lo que definiremos la función auxiliar `foo` que tenga sus ceros en la solución del sistema de ecuaciones:

```
library(nleqslv)

foo <- function(x){
  c(x[1] / x[2] - media,
    x[1] / x[2]^2 - varianza)
}

nleqslv(c(1,2), foo)$x

## [1] 2.991873 4.114773
```

Para la distribución normal $N(\mu, \sigma)$, $E(X) = \mu$ y $\sigma^2(X) = \sigma^2$, por lo que en tal caso la estimación de los parámetros por el método de los momentos es trivial.

9.2. Funciones de pérdida

De nuevo, ilustraremos cómo estimar parámetros de una distribución mediante un ejemplo. Si P_μ es una distribución con media μ , entonces podemos definir la función del parámetro θ como

$$l(\theta) = \int (x - \theta)^2 dP_\mu(x)$$

Se puede probar que esta función tiene su mínimo en $E(X) = \mu$. Así que dada una muestra x_i de esa distribución, como

$$l(\theta) \approx \frac{1}{N} \sum_i (x_i - \theta)^2$$

y el mínimo de la expresión de la derecha es la media muestral, $\frac{1}{n} \sum_i x_i$, podemos usarla como aproximación a μ .

En general, si existe una función $f(x, \theta)$ tal que

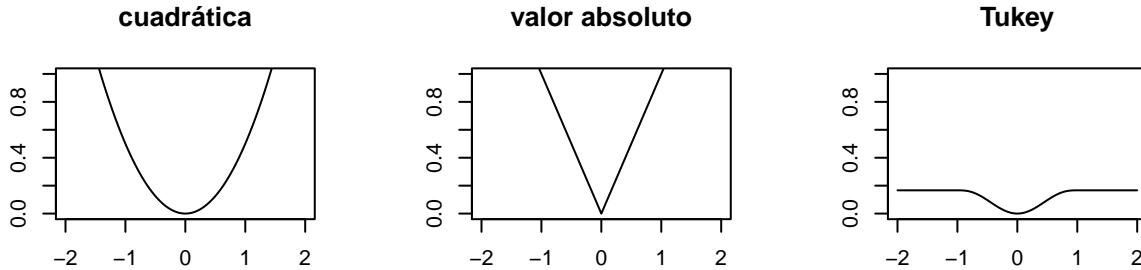
$$l(\theta) = \int f(x, \theta) dP_\mu(x)$$

tiene un mínimo en el parámetro de interés, podemos obtener una aproximación a él minimizando la expresión

$$\sum_i f(x_i, \theta)$$

En el caso anterior, $f(x, \theta) = (x - \theta)^2$ es la llamada **pérdida cuadrática**. Que se usa frecuentemente pero que es solo una de las posibles. Como alternativa se puede usar $f(x, \theta) = |x - \theta|$, que tiene su mínimo en la mediana de la distribución y que es menos sensible a *outliers*.

Estas dos son las más populares y cubren la práctica totalidad de las usadas en el mundo real. Sin embargo, es posible usar otras distintas con propiedades deseables. Por ejemplo, que sean aún más robustas a *outliers*. Existe, de hecho, una subdisciplina entera de la estadística, la de los M-estimadores, que estudia estas funciones alternativas y sus propiedades y que conecta con la teoría de los estimadores robustos, que son alternativas a los tradicionales cuando las observaciones presentan patrones de ruido excesivo.



En el gráfico anterior se muestra el perfil de tres funciones de pérdida: la cuadrática, la del valor absoluto y la de Tukey. Se observa cómo el impacto de una observación muy alejada en el último caso tiene el mismo peso que una *razonablemente alejada*; no ocurre lo mismo con las otras dos. El cuadrado de la cuadrática, además, da un peso superior a estas observaciones que el que les asigna la pérdida basada en el valor absoluto.

En concreto, con pérdidas cuadráticas, un *outlier* puede tener una influencia enorme sobre la estimación. El uso de otro tipo de funciones de pérdida reduce el impacto de los *outliers* y por lo tanto, proporcionar estimaciones más robustas.

El uso de funciones de pérdida es muy habitual cuando se ajustan modelos, algo a lo que se volverá más adelante. Supongamos que dos variables aleatorias X e Y dependen la una de la otra de manera que, p.e., $Y|X \sim N(a + bX, \sigma)$. Eso significa que Y depende de X y de parámetros a y b : para cada valor de X , Y tiene una distribución normal y su media (supuesto $a > 0$) es tanto mayor cuando mayor sea X .

Dada una muestra (x_i, y_i) de la distribución conjunta de X e Y , podemos estimar a y b minimizando la función de pérdida

$$l(a, b) = \sum_i (y_i - a - bx_i)^2.$$

De hecho, ese procedimiento es muy habitual, prácticamente el procedimiento por defecto, en ciencia de datos.

El uso de funciones de pérdida tiene una relación muy estrecha con la toma de decisiones. Las funciones de pérdida consideradas más arriba (p.e., la cuadrática, $(x - \theta)^2$) son tan comunes como arbitrarias. En situaciones reales pueden existir funciones de pérdida propiamente dichas que cuantifiquen el error cometido en, por ejemplo, euros. En tales circunstancias, es perfectamente razonable utilizarlas directamente en lugar de cualquiera de las *de libro*. Por ejemplo, en un hospital pueden querer estimar el número de camas que reservar para el servicio de urgencias. En tal caso, una sobreestimación (que dilapida recursos), aunque indeseable, es mucho menos grave que una infraestimación (que dejaría a pacientes con una atención inadecuada). En particular, la pérdida no es simétrica y la función de pérdida debería penalizar más severamente el error por defecto que el por exceso. Ese tipo de funciones de pérdida puede llevar a elegir como parámetro no ya una medida de centralidad de la distribución sino un cuantil de la misma.

Este ejemplo ilustra la estrecha vinculación entre el análisis de los datos y el proceso de toma de decisiones. Desafortunadamente, en demasiadas situaciones existe una importante brecha de comunicación entre quienes analizan los datos y los responsables de tomar las últimas decisiones. Este es el motivo, de hecho, de muchas de las decisiones incorrectas que toman instituciones públicas y privadas.

9.3. Estimación por máxima verosimilitud

La **estimación por máxima verosimilitud** está relacionada con el concepto, sumamente intuitivo, de *valor más verosímil*, que podemos ilustrar con el siguiente ejemplo. Supongamos que hay dos urnas; la primera contiene cinco bolas blancas y un negra; la segunda, cinco bolas negras y una blanca. Alguien extrae de una de ellas (y no nos dice cuál) una bola que resulta ser negra. Parece que la opción *más verosímil* es la segunda, la que contiene cinco bolas negras.

Esta idea intuitiva se puede formalizar para luego poder ser extendida. En efecto, la probabilidad de que la bola sea negra de haber sido extraída de la primera urna es $P(N | U = 1) = 1/6$. La probabilidad correspondiente a la segunda urna es $P(N | U = 2) = 5/6$. Por ese motivo nos decantamos por $U = 2$.

En general, si tenemos unos datos extraídos de una distribución P_θ , donde θ es un parámetro desconocido, podemos construir la función $P(\text{datos} | \text{parametros})$ y, de optar por algún valor de θ es razonable hacerlo por aquel que maximiza esa expresión. Esa expresión se conoce como **función de verosimilitud** y se representa así:

$$L(\theta) = L(\theta, X) = P(X | \theta).$$

Vamos a usar la estimación por máxima verosimilitud para estimar el parámetro desconocido en un problema de lanzamiento de monedas: se tira una 100 veces y se obtienen 60 caras. El parámetro desconocido es el parámetro θ de una distribución binomial. La función de verosimilitud es

$$L(\theta) = \binom{100}{60} \theta^{60} (1 - \theta)^{40}$$

que es la que tenemos que maximizar. Es evidente que maximizar esa expresión equivale a maximizar $\theta^{60}(1 - \theta)^{40}$ porque el factor faltante no depende de θ . También lo es que maximizar esa expresión equivale a maximizar su logaritmo, $60 \log \theta + 40 \log(1 - \theta)$. Ahora es sencillo derivar con respecto a θ , igualar a cero, despejar y obtener la estimación $\hat{\theta} = 0,6$.

De hecho, es habitual que la verosimilitud tenga forma de producto de términos (por ejemplo, cuando las observaciones, como en este caso, son independientes) y es muy habitual representar y operar con el logaritmo de la verosimilitud en lugar de con esa función directamente. Incluso por motivos de precisión cuando las operaciones se realizan con un ordenador.

También es posible resolver analíticamente otro caso general: X sigue una distribución normal $N(\mu, \sigma)$ y se obtiene una muestra x_1, \dots, x_n de X , entonces se puede probar que la estimación por máxima verosimilitud de la media, $\hat{\mu}$ no es otra cosa que $\frac{1}{n} \sum_i x_i$, la media muestral de las observaciones. De hecho, al operar sobre la verosimilitud, se observa cómo maximizarla se reduce a minimizar la suma $\sum_i (x_i - \theta)^2$. Es decir, estimar por máxima verosimilitud en este caso se reduce a minimizar una pérdida cuadrática. De hecho, en este caso, las estimaciones por máxima verosimilitud, con una pérdida cuadrática y por el método de los momentos, coinciden.

9.4. Teoremas de convergencia

En las secciones anteriores hemos planteado y justificado una serie de procedimientos para obtener estimaciones de parámetros. Pero surge una pregunta: ¿funcionan verdaderamente? Más concretamente, si hemos estimado $\hat{\theta}$, ¿tenemos garantías de que esa estimación esté próxima al valor *verdadero* θ ?

Existe una serie de teoremas, llamados de convergencia, que estudian qué sucede si construimos el estimador $\hat{\theta}_n$ a partir de muestras crecientes en tamaño (n) de la distribución subyacente. Una de las preguntas que intentan responder es si como se espera (o en qué casos), $\hat{\theta}_n$ tiende a θ , i.e., si al crecer el tamaño de la muestra, más se aproxima el valor estimado al *real*.

Estos teoremas se preocupan también del *sentido* que tiene esa convergencia. En concreto, si lo es en probabilidad o *casi seguro*. Esta es una discusión muy técnica que excede el alcance de este manual.

Una última cuestión que se plantean, refinando las anteriores, es acerca de la distancia esperada entre $\hat{\theta}_n$ y θ , que desarrollaremos en la siguiente sección.

En realidad, en algunos de los casos vistos anteriormente, el estimador $\hat{\theta}$ no es otra cosa que una proporción o, más en general, una media. En esos casos (y bajo ciertas condiciones no particularmente estrictas), la ley de los grandes números garantiza la convergencia. Los teoremas de convergencia arriba mencionados permiten extender estos resultados a (prácticamente) cualquier estimador $\hat{\theta}$ de los considerados antes.

9.5. Variabilidad de las estimaciones e intervalos de confianza

No solo interesa tener garantías de que $\hat{\theta}_n$ tiende a θ ; también queremos tener alguna indicación de la distancia entre ambos valores. Teóricamente, tendremos una distribución (usualmente no conocida o solo conocida aproximadamente) de las diferencias entre el estimador y el valor real. En la práctica, estas distribuciones se traducen en *rangos* alrededor del estimador derivados de aquellas distribuciones y cuya longitud decrecerá con el número de observaciones, dentro de los cuales cabe encontrar el parámetro verdadero.

Comenzaremos con una discusión sobre los intervalos de confianza que extenderemos con una mención a una serie de resultados teóricos y concluiremos con una introducción al principio del *plug-in* y el *bootstrap*, una técnica moderna para construir esos intervalos.

9.5.1. Intervalos de confianza para la distribución normal

Si X es una variable aleatoria normal de media μ y desviación estándar σ y se extrae de ella una muestra x_i de tamaño n , entonces, la media \bar{X} de la muestra, típicamente, se encuentra cerca de μ . Esto es así porque ya sabemos que

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

tiene distribución normal estándar. Es decir, el 95 % de las veces estará comprendida entre $q_{0,025}$ y $q_{0,975}$ de dicha distribución, es decir los valores

```
qnorm(c(0.025, 0.975))
```

```
## [1] -1.959964 1.959964
```

Que es lo mismo que decir que, muy frecuentemente, \bar{X} está dentro del rango $[\mu - 1,96\sigma/\sqrt{n}, \mu + 1,96\sigma/\sqrt{n}]$. O a la inversa, que muy frecuentemente, el 95 % de las veces aproximadamente (por el redondeo), μ estará comprendido en el intervalo

$$\left[\bar{X} - \frac{1,96\sigma}{\sqrt{n}}, \bar{X} + \frac{1,96\sigma}{\sqrt{n}} \right].$$

Eso es lo que en este caso se conoce como **intervalo de confianza** (al 95 %) para la media. Si se construyen muchos de ellos, aproximadamente el 95 % de ellos contendrán el valor μ . En efecto,

```
n <- 100; mu <- 0.5; sigma <- 0.7
mean(replicate(10000, {
  x <- rnorm(n, mu, sigma)
  mean(x) - 1.96 * sigma / sqrt(n) < mu & mu < mean(x) + 1.96 * sigma / sqrt(n)}))
```

```
## [1] 0.9496
```

Hay que tener en cuenta la siguiente sutileza. Así como podemos decir que la probabilidad de que \bar{X} esté contenido en el intervalo

$$\left[\mu - \frac{1,96\sigma}{\sqrt{n}}, \mu + \frac{1,96\sigma}{\sqrt{n}} \right]$$

es del 95 %, no se puede decir lo mismo de la probabilidad de que μ lo esté en

$$\left[\bar{X} - \frac{1,96\sigma}{\sqrt{n}}, \bar{X} + \frac{1,96\sigma}{\sqrt{n}} \right].$$

En este segundo caso, μ estará dentro o no. Además, el intervalo depende de la muestra obtenida. Lo único que se puede decir, como ilustra la simulación anterior, es que de repetirse el experimento muchas veces, en el 95 % de los casos el intervalo contendría el valor verdadero. Pero el intervalo cambia (porque cambia \bar{X}) en cada simulación.

Si no se conoce σ , se puede reemplazar en la fórmula anterior por su estimación, la desviación estándar de la muestra, i.e.,

$$s = \frac{1}{n-1} \sum_i (x_i - \hat{X})$$

y entonces se sabe que la expresión

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$

tiene una distribución t de Student (de $n - 1$ grados de libertad). Por lo que ya sabemos que para valores de n relativamente grandes, eso significa que es, esencialmente, normal de nuevo. Por eso los intervalos de confianza usados más arriba pueden aplicarse en este caso con dos salvedades:

- El valor σ desconocido se reemplaza por su aproximación s .
- Se pueden usar los cuantiles de la distribución normal solo si n es grande; en otro caso habría que usar `qt(c(0.025, 0.975), df = n-1)`.

Los dos ejemplos planteados más arriba son sumamente particulares. Pero haciendo abstracción, hemos considerado en ambos un estimador $\hat{\theta}_n$ de un parámetro de interés θ_0 y hemos podido deducir la distribución de $\hat{\theta}_n - \theta_0$, que resulta ser $N(0, \sigma/\sqrt{n})$. Además, a partir de ella hemos podido construir intervalos de confianza.

En el resto de la sección vamos a extender estos resultados a otros casos más generales. En los dos casos planteados nos interesa un parámetro de la distribución subyacente, la media. Además, también en ambos, la distribución subyacente es normal. Pero en ambos casos hemos obtenido un resultado similar: la distancia entre el estimador y el valor verdadero es normal de media cero y con una desviación estándar inversamente proporcional a \sqrt{n} .

9.5.2. Resultados teóricos

Igual que existen resultados teóricos que extienden la ley de los grandes números y que garantizan la convergencia de $\hat{\theta}_n$ a θ_0 , las extensiones correspondientes del teorema central del límite concluyen que los ejemplos de la sección anterior son casos particulares de una propiedad bastante habitual: es frecuente que $\hat{\theta}_n - \theta_0$, al menos para valores de n bastante altos, tenga una distribución normal con desviación estándar del orden de $1/\sqrt{n}$. O, escrito de otra forma, que

$$\sqrt{n}(\hat{\theta}_n - \theta) \longrightarrow N(0, \sigma),$$

es decir, que las diferencias entre los estimadores y el valor real tienen una distribución aproximadamente normal (para valores altos de n) con una varianza que decrece como $1/\sqrt{n}$.

Estos resultados no son universales y dependen del problema, de las distribuciones implicadas, etc. Pero es relativamente seguro darlo por bueno en general.

Por supuesto, la enumeración y demostración de los teoremas relevantes queda fuera del alcance de estas páginas.

9.5.3. El principio plug-in e intervalos de confianza

Antes, al estudiar los intervalos de confianza para estimaciones sobre la distribución normal, hemos pasado de una expresión del tipo

$$\left[\mu - \frac{1,96\sigma}{\sqrt{n}}, \mu + \frac{1,96\sigma}{\sqrt{n}} \right]$$

que es teóricamente sólida pero que depende, precisamente, de los parámetros desconocidos μ y σ , a otra del tipo

$$\left[\bar{X} - \frac{1,96s}{\sqrt{n}}, \bar{X} + \frac{1,96s}{\sqrt{n}} \right]$$

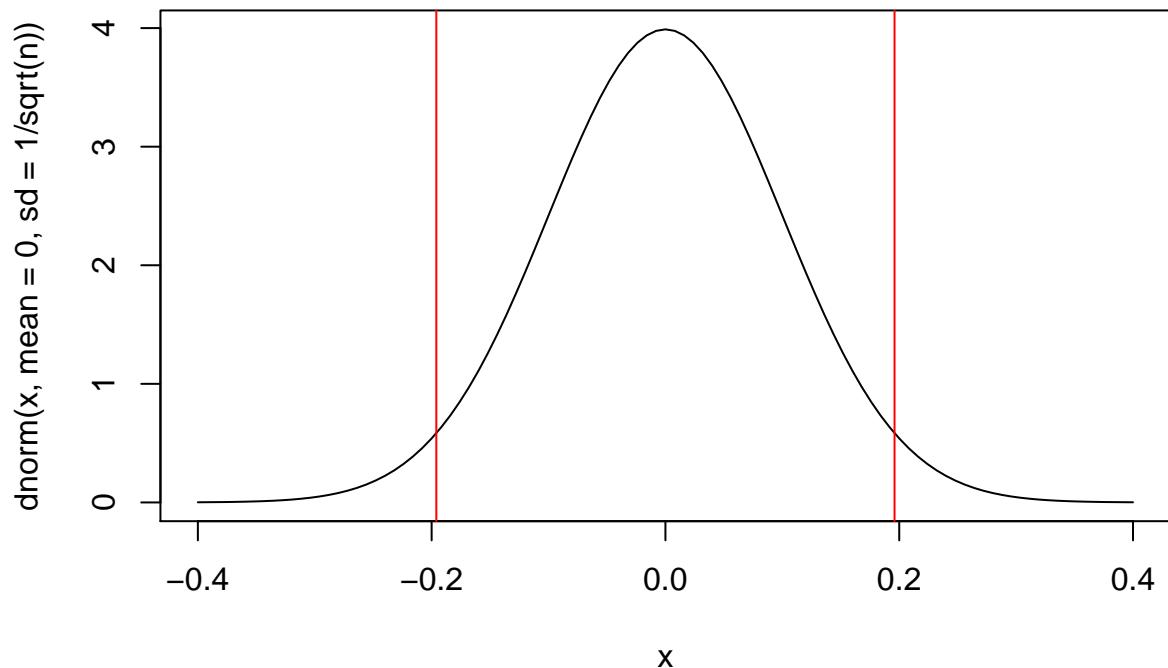
que depende únicamente de parámetros estimados (a partir de los datos de la muestra) \bar{X} y s . Esta técnica es un ejemplo del llamado *principio del plug-in*, que consiste precisamente en reemplazar parámetros desconocidos por sus estimaciones muestrales.

En el caso de la distribución normal con media $\mu = 0$ y desviación estándar $\sigma = 1$ (supuestamente desconocidas), de observarse una muestra de valores x_1, \dots, x_n , la distribución de \bar{X} es, precisamente,

```
set.seed(155)
```

```
n <- 100
curve(dnorm(x, mean = 0, sd = 1 / sqrt(n)), from = -.4, to = .4, main = "Densidad de la media de la muestra")
abline(v = qnorm(.025, mean = 0, sd = 1/sqrt(n)), col = "red")
abline(v = qnorm(.975, mean = 0, sd = 1/sqrt(n)), col = "red")
```

Densidad de la media de la muestra (e intervalo de confianza al 95%)



Si obtenemos una muestra de tamaño $n = 100$ de dicha distribución, podemos comparar las densidades originales y las construidas a partir de la estimaciones de μ y σ :

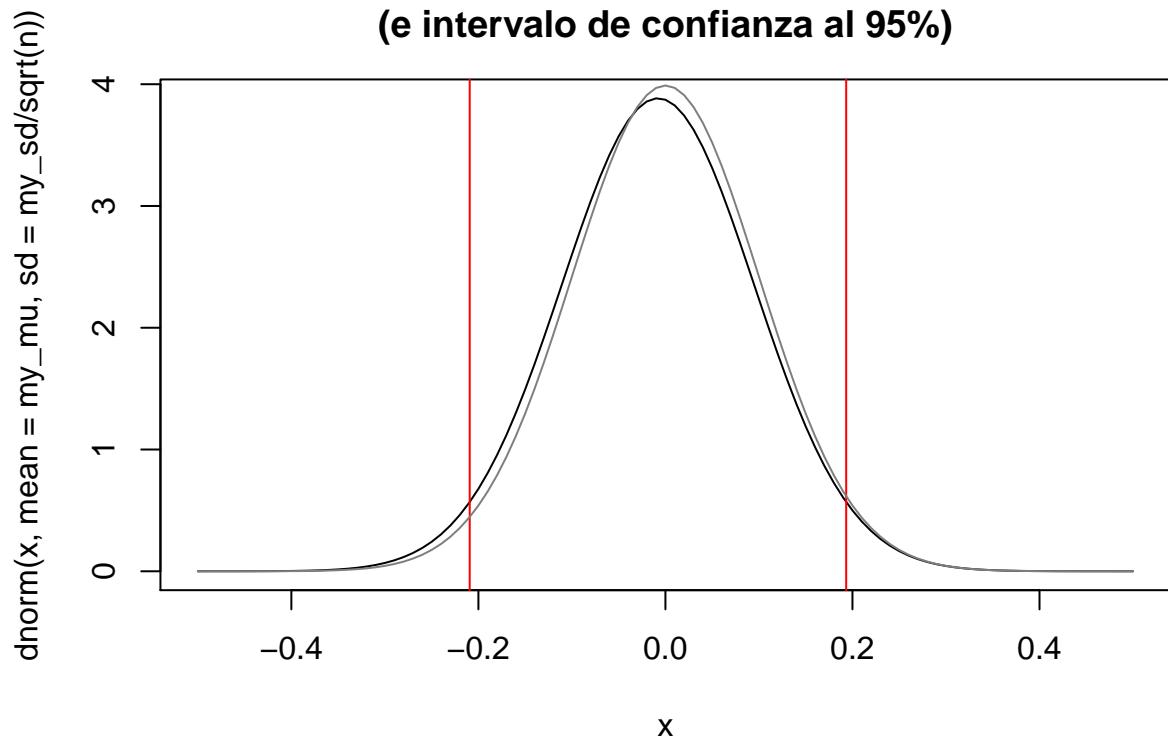
```
muestra <- rnorm(n, mean = 0, sd = 1)

my_mu <- mean(muestra)
my_sd <- sd(muestra)

curve(dnorm(x, mean = my_mu, sd = my_sd / sqrt(n)), from = -.5, to = .5, main = "Densidad de la media de la muestra")
abline(v = qnorm(.025, mean = my_mu, sd = my_sd/sqrt(n)), col = "red")
abline(v = qnorm(.975, mean = my_mu, sd = my_sd/sqrt(n)), col = "red")

curve(dnorm(x, mean = 0, sd = 1 / sqrt(n)), col = "gray50", add = TRUE)
```

Densidad de la media de la muestra (e intervalo de confianza al 95%)



Las dos densidades son muy parecidas, luego también lo son los correspondientes intervalos de confianza (indicados más arriba). De hecho, si obtenemos réplicas de 100 observaciones de la densidad original, podemos estimar la distribución de los extremos de los intervalos de confianza obtenidos,

```

mu0 <- 0
sd0 <- 1

lower_limit <- mu0 + sd0 * qnorm(0.025, mean = mu0, sd = sd0)
upper_limit <- mu0 + sd0 * qnorm(0.975, mean = mu0, sd = sd0)

res <- replicate(10000, {
  x <- rnorm(n, mean = mu0, sd = sd0)
  c(mean(x), sd(x))
})

res <- t(res)

lower_limits <- sapply(1:nrow(res), function(i) res[i, 1] + res[i, 2] * qnorm(0.025, mean = res[i, 1], sd = res[i, 2]))
upper_limits <- sapply(1:nrow(res), function(i) res[i, 1] + res[i, 2] * qnorm(0.975, mean = res[i, 1], sd = res[i, 2]))

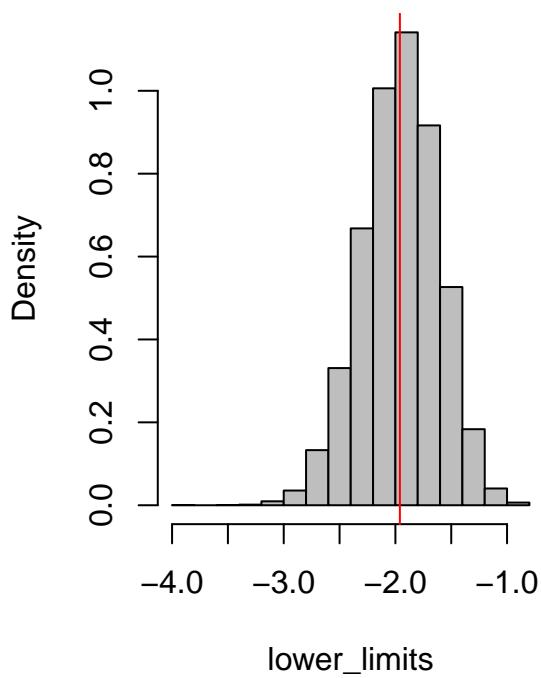
par(mfrow = c(1, 2))

hist(lower_limits, freq = FALSE, col = "gray")
abline(v = lower_limit, col = "red")

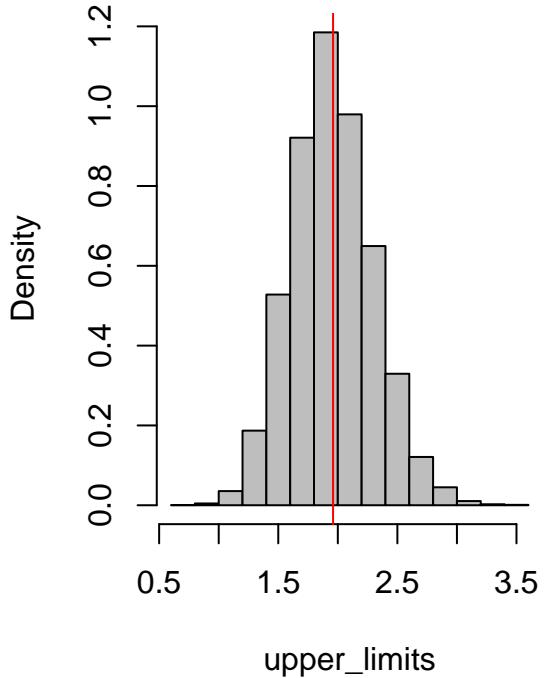
hist(upper_limits, freq = FALSE, col = "gray")
abline(v = upper_limit, col = "red")

```

Histogram of lower_limits



Histogram of upper_limits



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

En esto consiste fundamentalmente el principio del *plug-in*: reemplazar en las fórmulas teóricas los valores desconocidos por los estimados a partir de una muestra. Desde un punto de vista teórico, consiste en utilizar la distribución $D(\hat{\theta}_n)$ en lugar de la $D(\theta)$, donde $\hat{\theta}_n$ es una estimación de θ obtenida a partir de una muestra.

Por ejemplo, si se tira una moneda al aire 100 veces y salen 60 caras, podemos construir simplemente un intervalo de confianza para el parámetro p con los límites

```
qbinom(c(0.025, 0.975), 100, 0.6) / 100
```

```
## [1] 0.50 0.69
```

cuando que $\hat{\theta}_n = 0.6$ (que es un valor obtenido, por ejemplo, por máxima verosimilitud).

9.5.4. Intervalos de confianza mediante remuestreos

La técnica de los remuestreos (*bootstrap*) apareció a finales de los años 70 y es más intensiva computacionalmente (aunque en el s. XXI, ya no mucho) que las anteriores. Va un paso más allá que las anteriores: en lugar de considerar la distribución $D(\hat{\theta}_n)$, la aproxima por remuestreos a partir de la muestra original (con que se construye $\hat{\theta}_n$).

En concreto, la técnica simula otras muestras *posibles* de la distribución subyacente realizando remuestreos (con reemplazamiento) de la muestra original. Por ejemplo, si la muestra original está contenida en la variable `x` de R y el estimador buscado $\hat{\theta}$ está dado por la media muestral `mean(x)`, entonces se puede obtener una distribución aproximada de este estimador repitiendo la operación

```
mean(sample(x, length(x), replace = TRUE))
```

tantas veces como sea preciso. Finalmente, a partir de esas muestras se pueden calcular los cuantiles, etc.

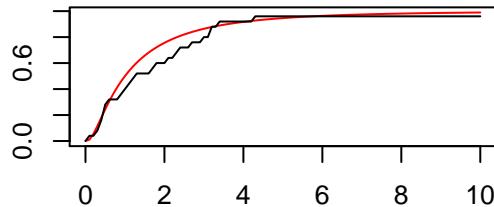
Este procedimiento está justificado en la siguiente observación (que, de hecho, es un teorema, el de Glivenko-Cantelli): que la función de distribución empírica F_n construida a partir de muestras x_1, \dots, x_n de una distribución F converge a F . Gráficamente, si F es la distribución lognormal,

```
foo <- function(i){
  curve(plnorm(x), from = 0, to = 10, col = "red", main = paste0("n = ", i), xlab = "", ylab = "")
  my_f <- ecdf(rlnorm(i))
  curve(my_f, add = TRUE)
}

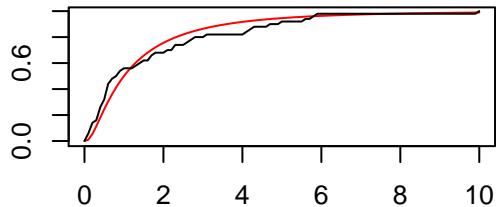
par(mfrow = c(2, 2))

foo(25)
foo(50)
foo(100)
foo(1000)
```

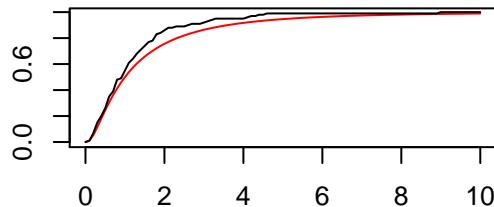
n = 25



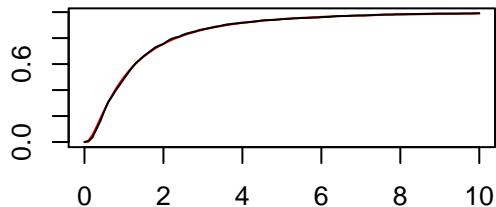
n = 50



n = 100



n = 1000



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

Por ejemplo, supongamos que tenemos una muestra de tamaño 100 de una distribución $N(1, 0.5)$:

```
muestra.orig <- rnorm(100, mean = 1, sd = 0.5)
```

Pueden obtenerse cuantiles aproximados mediante el principio del *plug-in* haciendo

```
qnorm(c(0.025, 0.975), mean = mean(muestra.orig), sd = sd(muestra.orig) / sqrt(100))
```

```
## [1] 0.8740544 1.0741009
```

Incluso podrían obtenerse muestras simuladas haciendo

```
muestras.plugin <- replicate(1000, mean(
  rnorm(100, mean = mean(muestra.orig),
  sd = sd(muestra.orig))))
```

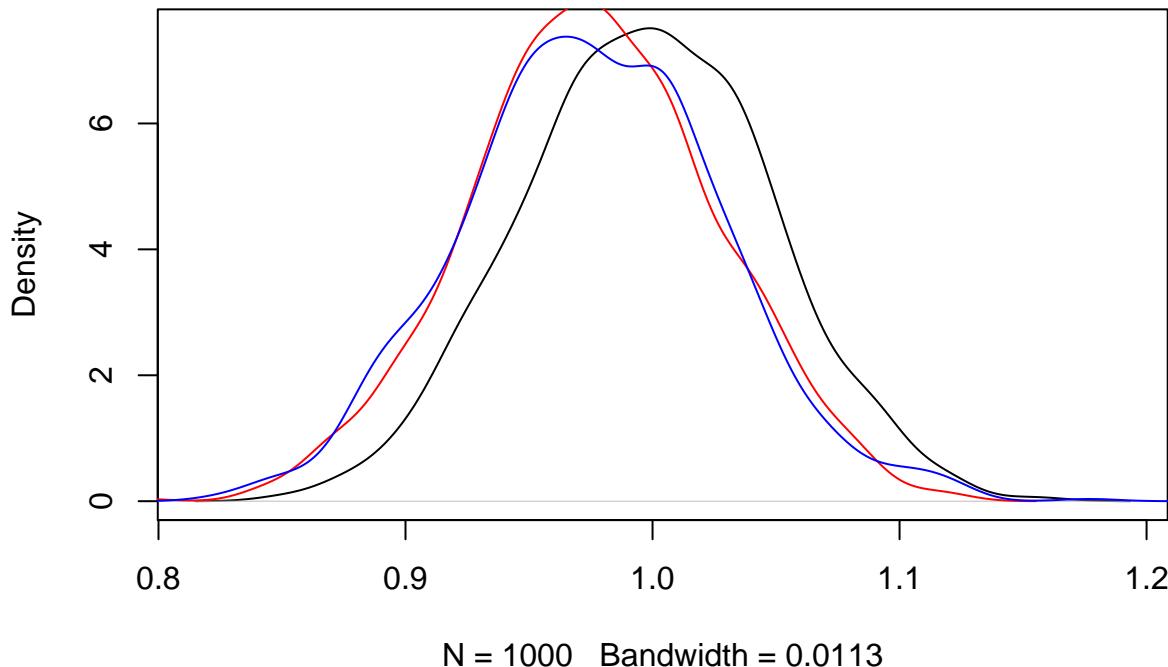
Pero el *bootstrap* es más radical y propugna remuestrear la muestra original así:

```
muestras.bootstrap <- replicate(1000, mean(sample(muestra.orig, 100, replace = TRUE)))
```

De hecho, todas las densidades son similares:

```
muestras <- replicate(1000, mean(rnorm(100, mean = 1, sd = 0.5)))
plot(density(muestras))
lines(density(muestras.bootstrap), col = "red")
lines(density(muestras.plugin), col = "blue")
```

density.default(x = muestras)



N = 1000 Bandwidth = 0.0113

Además, esta aproximación es tanto mejor cuanto mayor sea la muestra original.

Al igual que ocurre con el principio del *plug-in*, existen muchas aplicaciones del *bootstrap*. La más relevante para este capítulo es la de la construcción de intervalos de confianza. Estos intervalos de confianza se construyen a partir de la aproximación de la distribución de $\hat{\theta}_n - \theta_0$ a partir de remuestreos.

En el caso analizado anteriormente, por ejemplo, un intervalo de confianza estimado para la media de la distribución original es el que tiene por extremos

```
quantile(muestras.bootstrap, c(0.025, 0.95))
```

```
##      2.5%      95%
## 0.873683 1.058555
```

que puede compararse con el obtenido con el principio del *plug-in* más arriba, i.e.,

```
qnorm(c(0.025, 0.975), mean = mean(muestra.orig), sd = sd(muestra.orig) / sqrt(100))
```

```
## [1] 0.8740544 1.0741009
```

9.6. Referencias

- Sköld, M., *Computer Intensive Statistical Methods*, Lund University

9.7. Ejercicios

9.7.0.1. Ejercicio

Estimar la media de una normal cuando hay un outlier.

9.7.0.2. Ejercicio

Algo sobre la función de pérdida de Tukey con distribuciones bimodales.

9.7.0.3. Ejercicio

Crear intervalos con el principio *plug-in* para algunos estimadores.

10. Pruebas de hipótesis

Las pruebas de hipótesis tienen un papel fundamental en estadística y sus aplicaciones. Una de las más en boga recientemente es la de los llamados tests A/B: una empresa tiene un portal en internet en el que vende sus productos. Un buen día se plantea la posibilidad de realizar un rediseño de la página y se pregunta si será beneficioso o no para sus ventas. Para esclarecerlo, plantea un test A/B: a la mitad de sus visitantes muestra la versión original y a la otra mitad, la rediseñada. Después de un tiempo, compara las ventas realizadas a uno y otro grupo y en función de dicha comparación, tiene que decidir si mantener la versión existente o implantar definitivamente la rediseñada.

Existen varias alternativas a la hora de realizar pruebas de hipótesis:

- La de Fisher, o prueba de significancia
- La de Neyman-Pearson, que es una prueba de (o más bien, entre) dos hipótesis
- La llamada NHST (*Null Hypothesis Significance Testing*), muy habitual a la vez que muy cuestionada
- La bayesiana, que discutiremos en una sección posterior

10.1. Test de significancia

El test de significancia es el primero cronológicamente y se debe a Fisher. Lo ilustra el ejemplo siguiente, en el que alguien tira una moneda al aire 100 veces y obtiene 60 caras. Entonces, se pregunta si eso es *raro*, es decir, si es suficiente como para sospechar que la moneda tiene un sesgo o, dicho de otra manera, si $p \neq 0,5$.

En el contexto de la prueba de significancia, se plantea lo que se conoce como **hipótesis nula**. En este caso, que $p = 0,5$. El objetivo de la prueba de significancia es *obtener evidencia contra la hipótesis nula*, es decir, cuantificar cómo de inverosímil es. El procedimiento que propone el test de significancia para hacerlo es calcular la probabilidad de haber obtenido esas 60 caras en cien tiradas, es decir, algo que puede resumirse en la siguiente expresión:

$$P(D | H_0)$$

O dicho de otra manera, y en general, la probabilidad de los datos obtenidos, D , dada la hipótesis nula. Esa cantidad (veremos inmediatamente los detalles de su cálculo) es la que se denomina **p-valor**.

El razonamiento aplicado implícitamente es análogo al de la prueba por contradicción y es aproximadamente este:

- Si la moneda fuese tal que $p = 0,5$ podríamos realizar muchas tiradas de ella y analizar los resultados típicos y sus frecuencias.
- Entonces, podemos comparar esos valores con los obtenidos en el experimento. Si estos últimos son *normales*, no podemos descartar en absoluto que $p = ,5$; pero, ahora bien, si son extraños, probablemente esté ocurriendo algo (cuya naturaleza desconocemos) y, estaremos legitimados a pensar que $p \neq ,5$.

Más adelante veremos cómo pueden simularse muchas tiradas de moneda para calcular los p-valores. Sin embargo, en este caso, dado $p = 0,5$ podemos calcular exactamente la probabilidad de cada resultado (i.e., $0,1,\dots,99,100$ caras); por ejemplo, la probabilidad de obtener 50 caras es

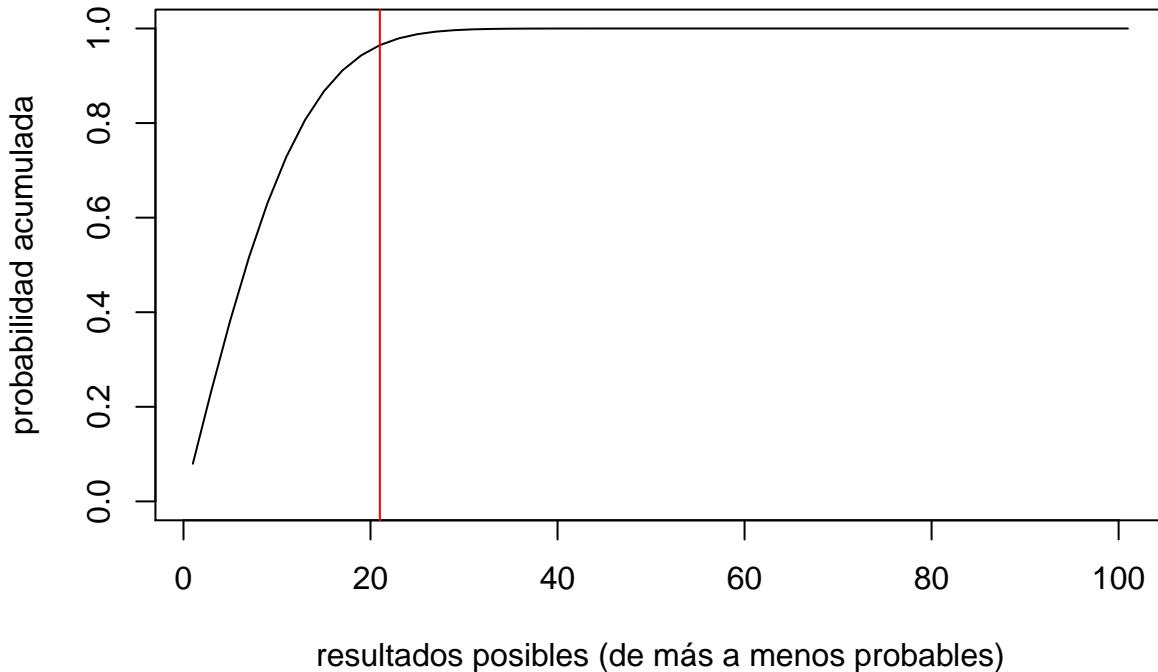
```
dbinom(50, 100, 0.5)
```

```
## [1] 0.07958924
```

o algo así como el 8 %, mientras que la de sacar 60 caras es apenas del 1 %. Así las cosas, ¿qué método puede plantearse para determinar si 60 caras es un resultado *raro*? O dicho de otra manera, ¿qué forma concreta adopta la expresión $P(D | H_0)$ en el contexto de las pruebas de significancia?

La estrategia, sumamente cuestionada, que propone la prueba de significancia es calcular la probabilidad de obtener un resultado *tanto o más raro* que el observado. Si ordenamos los posibles resultados obtenidos de mayor a menor probabilidad de ocurrencia y calculamos sus probabilidades acumuladas, obtenemos

Probabilidad acumulada de los eventos posibles (ordenados de mayor a menor probabilidad de ocurrencia)



donde se ha indicado con una línea roja la posición de nuestro caso de interés, $n = 60$. De acuerdo con la estrategia antes señalada, serían tanto o más extraños que $n = 60$ los valores $60, 61, \dots, 100$, y también, por simetría, los valores $0, 1, \dots, 40$. La suma de la probabilidad de estos valores, que puede representarse brevemente como $P(|X - 50| \geq 10)$ sería, en R,

```
sum(dbinom(c(0:40, 60:100), 100, .5))
```

```
## [1] 0.05688793
```

o bien, más brevemente,

```
pbinary(40, 100, .5) + 1 - pbinary(59.5, 100, .5)
```

```
## [1] 0.05688793
```

Ese valor así calculado es el llamado p-valor.

Es pura convención adoptar un umbral determinado en el p-valor para *rechazar* o no la hipótesis nula. Pero es habitual (no sin cierta controversia) en muchas disciplinas tomar un valor límite de 0.05; es decir, rechazar la hipótesis cuando los datos son tan *raro*s que solo se observarían en uno de cada veinte experimentos. En otras, sin embargo, se exigen p-valores más pequeños.

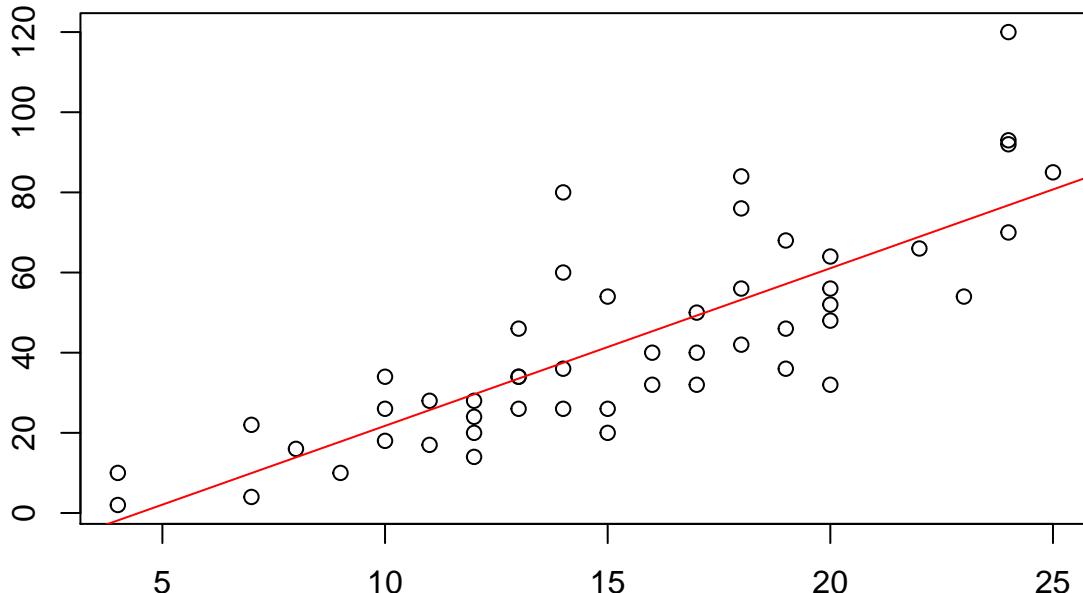
En caso de que se rechace la hipótesis nula, se suele decir que se hace con un nivel de significancia determinado, el $\approx 0,055$ en este caso.

Hay que advertir una cierta incoherencia lógica en este procedimiento: se rechaza la hipótesis nula esgrimiendo una probabilidad que incluye la de eventos no observados (por ejemplo, que haya 63 caras). También hay que tener en cuenta que el test de significancia así planteado no permite *aceptar* o dar por buena la hipótesis nula: solo rechazarla en caso de que la evidencia en su contra sea contundente. El test de significancia tendría su equivalente en un GPS cuyas únicas indicaciones de ruta fuesen del tipo: por donde vas, casi seguro que no es.

Una última advertencia al respecto es que la hipótesis nula, en la práctica totalidad de los contextos, es siempre falsa. Seguro que una moneda no tiene una probabilidad de cara de exactamente el 0,5; tal vez sea, por algún tipo de asimetría, de 0,499998082... y prosiguiendo el experimento con mucha paciencia pudiésemos lograr rechazar la hipótesis nula con un nivel de significancia arbitrariamente bajo.

No obstante estos problemas, el test de significancia se usa muy habitualmente, incluso implícitamente. Por ejemplo, cuando se plantea una regresión lineal, como a continuación:

Velocidad vs distancia de frenado



La regresión lineal construye esa recta (roja en el ejemplo anterior) que mejor se *aproxima* a la nube de puntos. Si extraemos sus coeficientes haciendo

```
summary(lm(dist ~ speed, data = cars))$coefficients
```

```
##             Estimate Std. Error    t value    Pr(>|t|)    
## (Intercept) -17.579095  6.7584402 -2.601058 1.231882e-02
```

```
## speed      3.932409 0.4155128 9.463990 1.489836e-12
```

obtenemos una tabla en la que aparecen la estimación del término independiente (fila superior) y la pendiente (inferior) más una serie de columnas adicionales. La última es el p-valor correspondiente a la prueba de significancia en que la hipótesis nula es que el coeficiente tiene un valor igual a cero. En el caso de la pendiente, el p-valor es muy pequeño (evidentemente, porque los datos tienen una tendencia creciente muy manifiesta y una pendiente nula querría decir que serían aproximadamente planos). Tenemos pues *evidencia* suficiente para rechazar esa hipótesis nula subyacente.

Las pruebas de significancia son, como hemos visto, ubicas en la estadística. Veremos más adelante cómo la prueba de significancia guarda analogías al principio de prueba por contradicción; solo que si los datos muestran una *contradicción* con H_0 , no está siempre clara la alternativa. En el ejemplo anterior de los coeficientes de regresión, una vez *rechazada* la hipótesis de que la pendiente sea cero, la prueba de significancia no aporta mayor información sobre el coeficiente. Otro aspecto controvertido de las pruebas de significancia ya señaladas más arriba es la manera en que construyen el *evento de rechazo*: incluyen en él tanto los datos observados como otros aún más extremos (y, por lo tanto, no observados). Así, en el ejemplo anterior, el p-valor está construido no solo en términos de la probabilidad de haber obtenido 60 caras sino 60 caras o más.

No obstante, la crítica más seria que puede realizarse a las pruebas de significancia están basadas en el hecho de que consideran la expresión $P(D | H_0)$ y no la que sería más lógica, $P(H_0 | D)$, es decir, la probabilidad de la hipótesis nula a la vista de los datos.

10.2. Pruebas de hipótesis

Las **pruebas de hipótesis** de Neyman-Pearson comparan dos hipótesis (a diferencia de la de Fisher, que solo contemplaba una, la nula). El siguiente ejemplo servirá para ilustrarlas. Existen dos monedas, una normal ($p = ,5$) y otra *trucada* ($p = 0,7$). Alguien ha elegido una de ellas (sin que nosotros sepamos cuál) y está dispuesto a realizar todas las tiradas que le indiquemos y a transmitirnos el número de caras. Nuestro objetivo es tomar una decisión sobre cuál es la moneda que se está utilizando realmente.

Por tanto, tenemos dos hipótesis:

- H_0 : ha elegido la moneda normal.
- H_1 : ha elegido la otra.

Para esclarecerlo, vamos a **diseñar un experimento** consistente en solicitar que se tire la moneda N veces y registrar el número n de caras. Después, si $n/N < c$, nos decantaremos por H_0 y si no, por H_1 . El problema consiste ahora en cómo elegir N y c ; para eso es conveniente introducir los llamados **errores de tipo I y II**.

El error de tipo I se comete cuando se cumple H_0 pero optamos por H_1 . A la probabilidad (condicional a que se cumple H_0) de cometer este tipo de error la llamamos α . Por su parte, el error de tipo II es el que se comete cuando la cierta es H_1 pero optamos por H_0 y la probabilidad correspondiente suele llámarsela β . Usando una tabla, los errores de ambos tipos pueden representarse así:

	elijo H_0	elijo H_1
real: H_0	acíerto	error Tipo I
real: H_1	error Tipo II	acíerto

Muchas veces, en la práctica, el tomar una decisión y equivocarse en ella implica unos costes (económicos u otros). Incurrimos pues en costes si cometemos un error, sea de tipo I o II. Además, los costes asociados a ambos tipos de error bien pueden ser distintos. Para minimizar el coste, habría que reducir el tamaño de α y β . Pero eso implica hacer crecer el valor de N (que puede, a su vez, ser costoso).

Esta observación permite introducir el concepto de la **potencia de una prueba**. Una prueba es tanto más potente cuanto menores son las probabilidades de error. Generalmente, la potencia crece al crecer N . Para tener una prueba más potente, por tanto, basta con invertir en N .

El procedimiento habitual, el que recogen casi todos los textos, es el siguiente. Primero, se elige como H_0 la hipótesis cuyo error asociado, el de tipo I, es más caro. Para minimizar el riesgo de incurrir en él, se fija un valor de α pequeño (p.e., 0,01). Entonces quedan dos parámetros libres: N y β .

Si N es grande, podremos reducir el tamaño de β . Y la potencia entonces se define exactamente como $1 - \beta$. A la inversa, si lo que se quiere fijar es la potencia (es decir, β , para reducir el impacto del error de tipo II), se puede deducir el valor de N necesario.

El procedimiento preciso (en nuestro ejemplo) es el siguiente: para un N determinado, se puede seleccionar el umbral c de manera que $P(X \geq c | H_0) \leq \alpha$, es decir, la probabilidad de cometer un error de tipo I es inferior a α . Fijado pues c , $\beta = P(X < c | H_1)$, la probabilidad de cometer un error de tipo II.

TODO: un ejemplo completo

10.3. Zonas grises de las pruebas anteriores y NHST

Ambos tipos de pruebas de hipótesis presentan zonas grises. En la prueba de significancia, no está claro cómo condicionar cuando la hipótesis nula es del tipo $\theta \in [-1, 1]$. También es un debate recurrente y jamás cerrado la determinación del umbral a partir del cual se rechaza H_0 . Finalmente, es un tanto desconcertante, como se ha indicado antes, el hecho de considerar datos que no se han observado para rechazar o no H_0 .

Además, la lógica que subyace a la prueba de significancia se asemeja demasiado al siguiente silogismo espúrio:

- Si alguien juega al fútbol, es muy improbable que juegue en primera división
- Ronaldo juega en primera división
- Por lo que rechazamos la hipótesis de que Ronaldo juegue al fútbol

Las pruebas de hipótesis también presentan ciertos problemas. Por ejemplo, cuando se quiere tomar una decisión entre las hipótesis $H_0: \theta = 0$ y $H_1: \theta \neq 0$. No está en absoluto claro cómo construir $P(D|H_1)$ en ese caso.

Más adelante, veremos la aproximación bayesiana a las pruebas de hipótesis, que vienen a superar estas incongruencias. Sin embargo, la versión de la prueba de hipótesis más corriente, la conocida como NHST (*null hypothesis significance testing*) es una muy controvertida mezcla de las dos anteriores, que consiste en:

- Plantea una hipótesis de trabajo (p.e., la versión B de la página es *mejor* que la A)
- Plantea una hipótesis nula estadística (p.e., que la media del número de conversiones será la misma en ambas versiones de la página)
- Usa el 5% como convención para rechazar esa hipótesis nula
- De rechazarse, acepta la hipótesis de trabajo como cierta

Debe advertirse que, aunque pueda tener sentido en algunas ocasiones, el procedimiento anterior no solo no está justificado teóricamente, sino que, además, facilita el abuso del procedimiento. En efecto, alguien podría plantear una de las llamadas hipótesis nulas *de paja*, una hipótesis nulas fáciles de rechazar, para proceder inmediatamente a *validar* una hipótesis de trabajo. Incluso es posible elegir una hipótesis nula ya rechazada de antemano: en la fase de experimentación se recogen datos de muchos tipos y es frecuentemente posible elegir uno de ellos como la hipótesis nula vinculada a la hipótesis de trabajo de partida.

Que esto es así y que el abuso se produce no se deduce únicamente de las posibles incongruencias teóricas del procedimiento sino que se manifiesta en la conocida como **crisis de reproducibilidad** actual (XXX). La crisis de reproducibilidad solo es posible cuando han fallado los filtros existentes en el mundo académico para evitar que sean publicados y dados por buenos resultados que, después, se han demostrado falsos al tratar de reproducirlos independientemente.

Esta crisis de reproducibilidad planea también sobre un mundo mucho más opaco que el de la ciencia: el del mundo de la empresa privada. En ella suele existir un interés natural en querer *demostrar* que el nuevo algoritmo, el nuevo portal, la nueva imagen corporativa, es *mejor* que la antigua. Existe por tanto un incentivo enorme tanto en *demostrar con números* esa ventaja como en subvertir el procedimiento para embellecer la

realidad. La crisis de reproducibilidad se manifiesta entonces en soluciones, ideas y proyectos que tienden a rendir por debajo de lo previsto.

10.4. El tamaño del efecto y las pruebas S y M

En esta sección plantearemos otro caso muy vinculado con las pruebas de hipótesis: el validar un tratamiento determinado. Existen dos grupos (por ejemplo, de pacientes en un estudio clínico) y uno de ellos recibe un tratamiento tradicional mientras que el otro, uno novedoso. Se puede decir que el segundo está *tratado* e interesa conocer la ventaja que supone (con respecto al anterior).

En el contexto antes descrito es típico que uno de los grupos, el no tratado, reciba un **placebo**, es decir, un tratamiento que no contiene principio médico alguno (agua, azúcar, etc.). En ese caso se estaría comparando el tratamiento contra la falta de tratamiento. Lo adecuado (aunque esto no es opinión unánime), cuando ya existe un tratamiento establecido, es comparar el nuevo contra el anterior por ver si lo supera o no.

En estos casos, puede determinarse a través de una prueba de hipótesis que el nuevo tratamiento es significativamente superior al anterior (con un determinado nivel de significancia) pero que, aun así, no merezca la pena. Aun si la diferencia entre ambos tratamientos es minúscula (e irrelevante), aumentando el tamaño muestral es posible detectar esa diferencia. Eso ha llevado a algunos a acuñar la frase de que la significancia no es (por sí sola) significativa.

Esta cuestión es particularmente acuciante en los tiempos del *big data*, que facilita tamaños muestrales enormes. Y, por lo tanto, facilita el descubrimiento de diferencias significativas pero insignificantes.

En ese sentido, existe una propuesta para reemplazar los errores de tipo I y II por los **errores de tipo S y M**. Un error de tipo S se cometería cuando se afirma que un tratamiento tiene un efecto positivo cuando, en realidad, lo tiene negativo. Es, como puede intuirse, un error muy serio. El segundo, el error de tipo M es el que se cometería al estimar el tamaño del efecto.

10.5. Pruebas de hipótesis e intervalos de confianza

Aunque las técnicas indicadas más arriba son las *de libro* para realizar las llamadas pruebas de hipótesis, en la práctica pueden hacerse y, de hecho, se hacen, de una manera más sencilla y efectiva: utilizando intervalos de confianza.

Si interesa realizar alguna hipótesis sobre un parámetro de una distribución, una estrategia siempre válida y generalizable es la de crear un modelo para los datos en cuestión, realizar la estimación puntual del parámetro de interés y construir los intervalos de confianza asociados al mismo. Eso permite contrastar, p.e., si dicho parámetro es o no igual a cero.

Más arriba habíamos estudiado los coeficientes de una regresión lineal,

```
modelo <- lm(dist ~ speed, data = cars)
summary(modelo)$coefficients

##                 Estimate Std. Error    t value   Pr(>|t|)
## (Intercept) -17.579095  6.7584402 -2.601058 1.231882e-02
## speed         3.932409   0.4155128  9.463990 1.489836e-12
```

que nos proporcionan una prueba de significancia para la hipótesis nula del que los coeficientes sean cero (y que se rechazan en ambos casos al nivel del significancia del 95 %). Como alternativa, se pueden calcular los intervalos de confianza para las anteriores estimaciones puntuales,

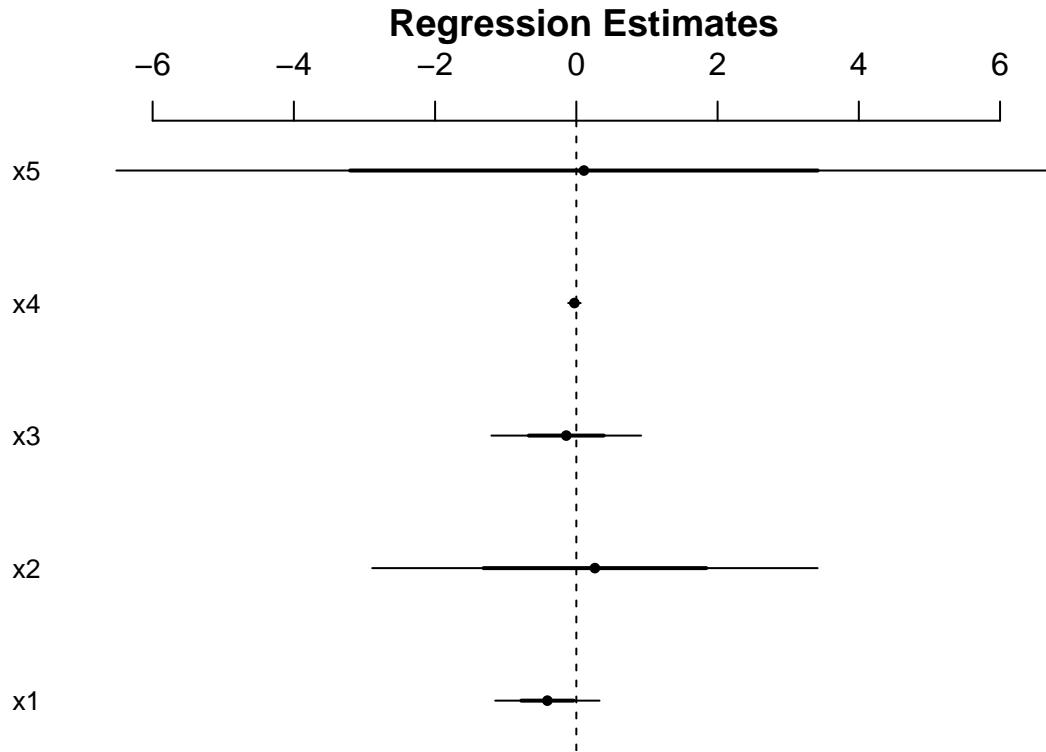
```
confint(modelo)

##             2.5 %    97.5 %
## (Intercept) -31.167850 -3.990340
```

```
## speed      3.096964  4.767853
```

que nos indica que los valores esperados del valor *real* de `speed` se encuentran entre 3,1 y 4,77. Como ese rango no contiene el cero, uno puede *rechazar* la hipótesis de que el coeficiente sea cero. De hecho, existen procedimientos gráficos para analizar visualmente el resultado de estas pruebas de hipótesis implícitas a través de representaciones como esta:

```
## Loading required package: MASS
## Loading required package: Matrix
## Loading required package: lme4
##
## arm (Version 1.10-1, built: 2018-4-12)
## Working directory is /home/carlos/ownCloud/staging/libros/curso_intro_probabilidad_estadistica
```



En ella (adaptada de la ayuda de la función `arm:::coefplot` de R) se ve una variable claramente (significativamente) distinta de cero (los intervalos de confianza contrazo fino son aproximadamente del 95 %) y otras en las que no podría rechazarse la hipótesis.

10.6. Algunas pruebas de libro

En esta sección vamos a repasar algunas pruebas estadísticas *de libro* acompañándolas del preceptivo código en R.

10.6.1. Prueba de la media para poblaciones normales

La prueba de la media para poblaciones normales se usa para comprobar si una población tiene una determinada media. En concreto, cuando existe una variable aleatoria con una distribución normal (o aproximadamente normal) y se quiere comprobar si su media es igual a un valor de referencia μ dado.

La variable aleatoria puede ser la cantidad real de refresco en un envase. Pueden existir pequeñas diferencias debidas al proceso de embotellado e interesa saber si existe una desviación con respecto al volumen anunciado (de, p.e., 33 cl).

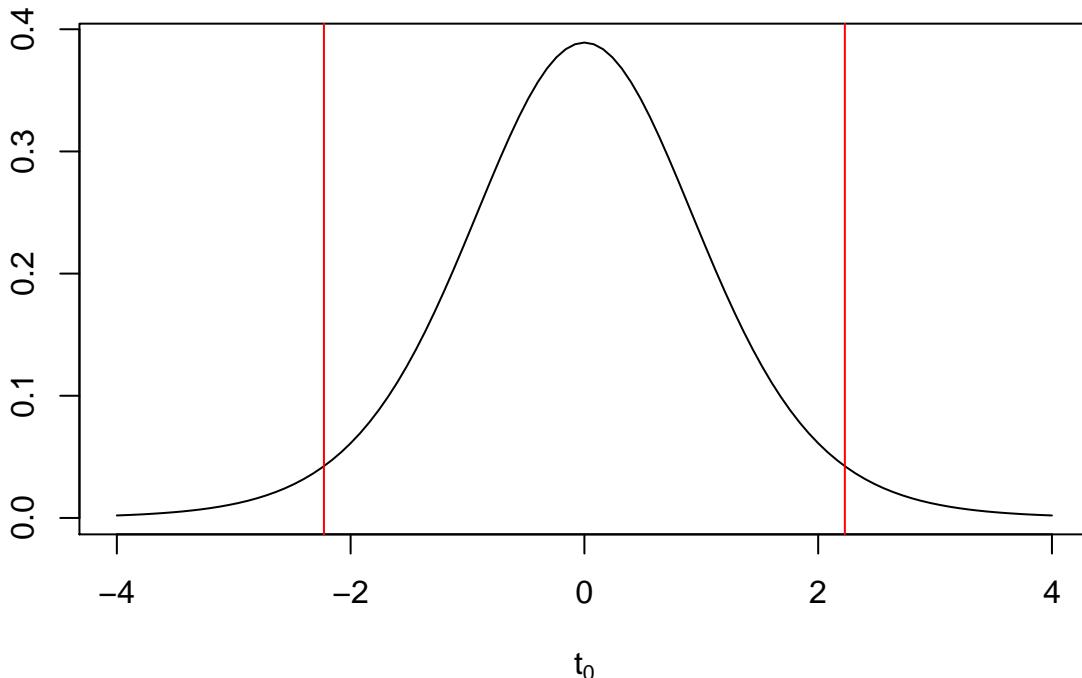
La prueba se basa en que bajo la hipótesis nula de que los datos tienen una distribución $X \sim N(\mu, \sigma)$, donde σ es una constante irrelevante para la cuestión, la expresión

$$t_0 = \frac{1/n \sum_i (x_i - \mu)}{s/\sqrt{n}},$$

donde s es la desviación estándar de la población, sigue una distribución t de Student con $n - 1$ grados de libertad. Esa expresión será pequeña si la media es próxima a μ y grande en caso contrario. Como conocemos su distribución, podemos calcular $P(|t| > |t_0| \mid H_0)$, es decir, la probabilidad de que t_0 esté alejado de su valor esperado, que debería ser próximo a 0.

Gráficamente, bajo la hipótesis nula, la distribución de t_0 (supuestos 10 grados de libertad) sería

distribución esperada de t_0 bajo H_0



En la gráfica anterior aparece la distribución esperada de t_0 bajo la hipótesis nula más dos líneas verticales tales que la integral de la curva entre ellas es 0.95. Es decir, lo que queda fuera de esa franja tiene probabilidad de 0.05. Si el valor obtenido de t_0 queda fuera de ella, puede rechazarse H_0 con el nivel 0.05 de significancia.

Esa es la justificación de la prueba, aunque en R es fácil (tal vez demasiado) aplicarla, como ilustra el siguiente ejemplo:

```
daily.intake <- c(5260, 5470, 5640, 6180, 6390, 6515,
  6805, 7515, 7515, 8230, 8770)

t.test(daily.intake, mu = 7725)

##
##  One Sample t-test
##
## data: daily.intake
```

```

## t = -2.8208, df = 10, p-value = 0.01814
## alternative hypothesis: true mean is not equal to 7725
## 95 percent confidence interval:
## 5986.348 7520.925
## sample estimates:
## mean of x
## 6753.636

```

La salida que proporciona R incluye un p-valor por debajo de 0.05, por lo que se podría rechazar la hipótesis de que la media fuese igual a 7725.

Una variante de la prueba permite *testear* si la media de la población está por debajo o por encima de un nivel dado de referencia (y no solo si es igual a él):

```

daily.intake <- c(5260, 5470, 5640, 6180, 6390, 6515,
 6805, 7515, 7515, 8230, 8770)

t.test(daily.intake, mu = 7725, alternative = "less")

##
## One Sample t-test
##
## data: daily.intake
## t = -2.8208, df = 10, p-value = 0.009069
## alternative hypothesis: true mean is less than 7725
## 95 percent confidence interval:
##      -Inf 7377.781
## sample estimates:
## mean of x
## 6753.636

```

En el caso anterior hemos cambiado la hipótesis nula. Ahora es que la media de la población está por encima del nivel de referencia. El p-valor es muy inferior en este caso.

La función `t.test` también proporciona un **intervalo de confianza**: un rango de valores que incluye nuestra media dentro del cual el valor real *podría estar* con una confianza, en este caso, del 95 %. Cuando este intervalo no contiene al de referencia, la hipótesis se rechaza (al nivel complementario del 5 %).

El p-valor y el intervalo de confianza se construyen de manera distinta y son ambos aproximaciones, por lo que es posible encontrar casos límitrofes en los que el intervalo contiene el valor de referencia pero el p-valor está por debajo del 0.05 y a la inversa.

Hay que recordar que la distribución t de Student para grados de libertad no particularmente altos (es decir, para muestras no demasiado grandes) es prácticamente indistinguible de una distribución normal, que es algo que ocurre con frecuencia. De modo que una manera alternativa de realizar esta prueba cuando la muestra es razonablemente grande consiste en calcular los cuantiles correspondiente a la media muestral de una distribución $N(\mu, s/n)$.

```

x <- rnorm(100, mean = 0.2, sd = 2)
t.test(x, mu = 0, alternative = "greater")$p.value

## [1] 0.1141934
1 - pnorm(mean(x), mean = 0, sd = sd(x) / sqrt(length(x)))

## [1] 0.112751

```

10.6.2. Remuestreos

Cuando se realiza una prueba estadística como la anterior, caben dos opciones: acudir a la literatura, encontrar la prueba correspondiente y aplicarla (con suerte, además, está disponible en R), o recurrir a los remuestreos, que son una técnica general que requiere ordenadores, unos instrumentos no disponibles cuando se desarrollaron aquellos resultados teóricos.

Siguiendo con el ejemplo de la sección anterior, lo que en el fondo queremos comprobar es si la media de nuestras observaciones está cerca o lejos del valor de referencia $\mu = 7725$. Nuestros datos tienen una distribución desconocida X (hemos supuesto que son normales con una media desconocida, pero ni siquiera eso, como veremos, es estrictamente necesario) y lo que queremos comprobar es si su media está cerca o lejos de μ . Idealmente, queríamos obtener muchas muestras de X , pero solo tenemos una. Los remuestreos (o *bootstrap*) permiten simular muchas muestras de X a partir de una dada realizando muestras (con reemplazamiento y del mismo tamaño) de la muestra original.

En concreto, cada muestra es, en R,

```
sample(daily.intake, length(daily.intake), replace = T)
```

```
## [1] 5260 6805 7515 5640 6805 6390 6515 6515 8770 6180 6180
```

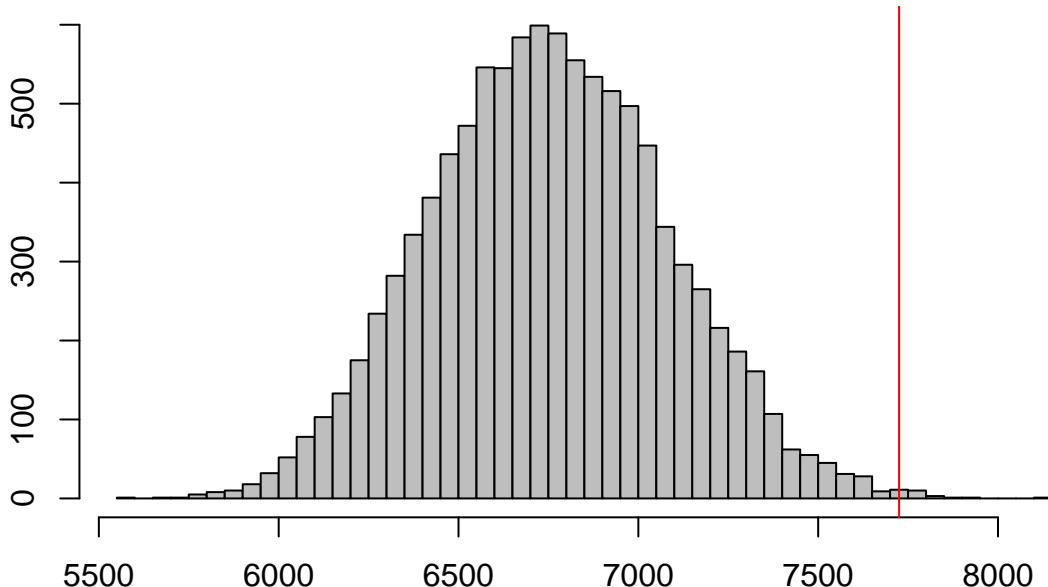
y podemos construir una aproximación a la distribución de la media de X haciendo

```
muestra.media <- replicate(10000, mean(sample(daily.intake, length(daily.intake), replace = T)))
```

El siguiente gráfico muestra la distribución de los valores de esa media junto con una línea vertical roja en el valor de referencia, 7725:

```
hist(muestra.media, breaks = 50, col = "gray",
      xlab = "", ylab = "",
      main = "histograma de las réplicas por\nnremuestreo de la media")
abline(v = 7725, col = "red")
```

**histograma de las réplicas por
remuestreo de la media**



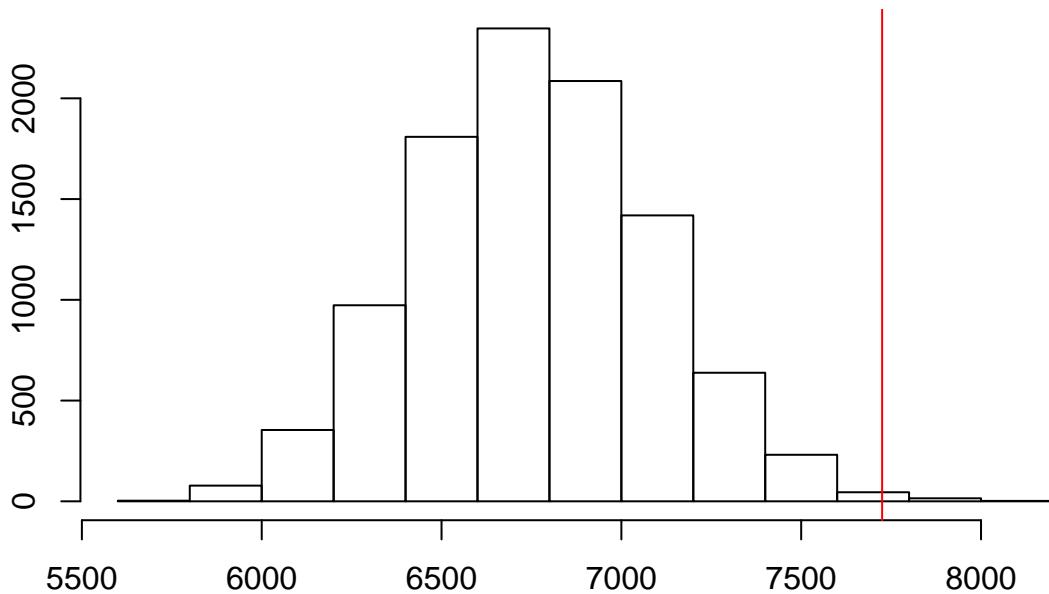
El p-valor correspondiente calculado de esta manera es el porcentaje de observaciones situados a la derecha de la barra

```

muestras <- replicate(10000, mean(sample(daily.intake, length(daily.intake), replace = T)))
hist(muestras, main = "Posibles valores estimados de la media", xlab = "", ylab = "")
abline(v = 7725, col = "red")

```

Posibles valores estimados de la media



- Prueba de Student (una y dos muestras)
- Prueba de los signos de Wilcoxon
- Prueba de igualdad de varianzas
- Prueba de igualdad de proporciones
- Prueba de normalidad de Kolmogorov-Smirnov

```
wilcox.test(daily.intake, mu = 7725)
```

```

## Warning in wilcox.test.default(daily.intake, mu = 7725): cannot compute
## exact p-value with ties

##
##   Wilcoxon signed rank test with continuity correction
##
## data: daily.intake
## V = 8, p-value = 0.0293
## alternative hypothesis: true location is not equal to 7725

```

10.7. Referencias

- Perezgonzalez, J. D., Fisher, *Neyman-Pearson or NHST? A Tutorial for Teaching Data Testing*, enlace
- Gigerenzer, G. *Mindless statistics*, enlace
- XKCD, *Significant*, enlace
- XXX

10.8. Ejercicios

10.8.0.1. Ejercicio

11. Datos tabulares

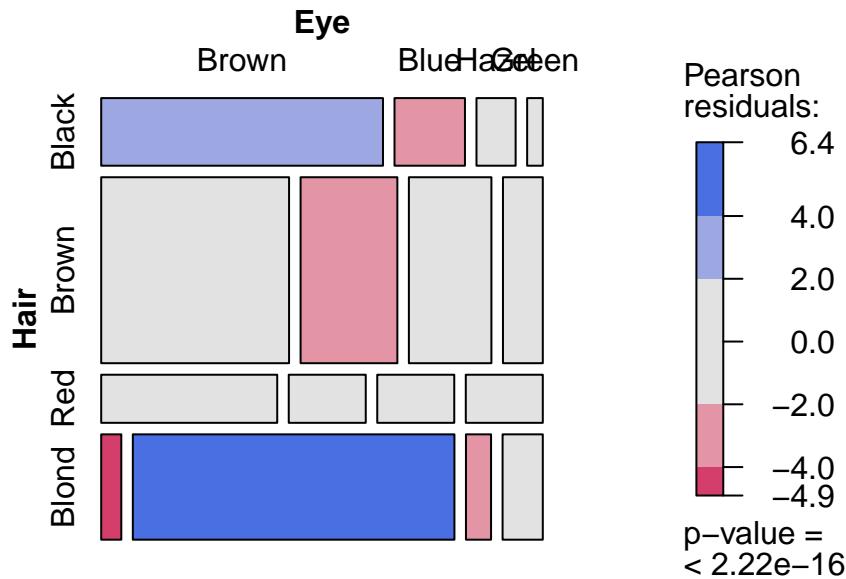
Vamos a estudiar datos de la forma

	Brown	Blue	Hazel	Green
Black	36	9	5	2
Brown	66	34	29	14
Red	16	7	7	7
Blond	4	64	5	8

La primera pregunta que puede formularse sobre este tipo de datos es si las variables involucradas son independientes; luego, en caso negativo, cómo interactúan entre sí.

11.0.1. Mosaicos e independencia

Es conveniente comenzar el análisis de datos tabulares con una representación gráfica. Existen muchas alternativas (p.e., los variogramas), pero no es habitual encontrar **mosaicos**:



Existen muchas versiones de los gráficos de mosaico. Además, existen extensiones para tablas de dimensiones superior a dos (no tratadas aquí). En cualquier caso, todas ellas se pueden interpretar como gráficos de barras (apiladas) con una determinada reparametrización.

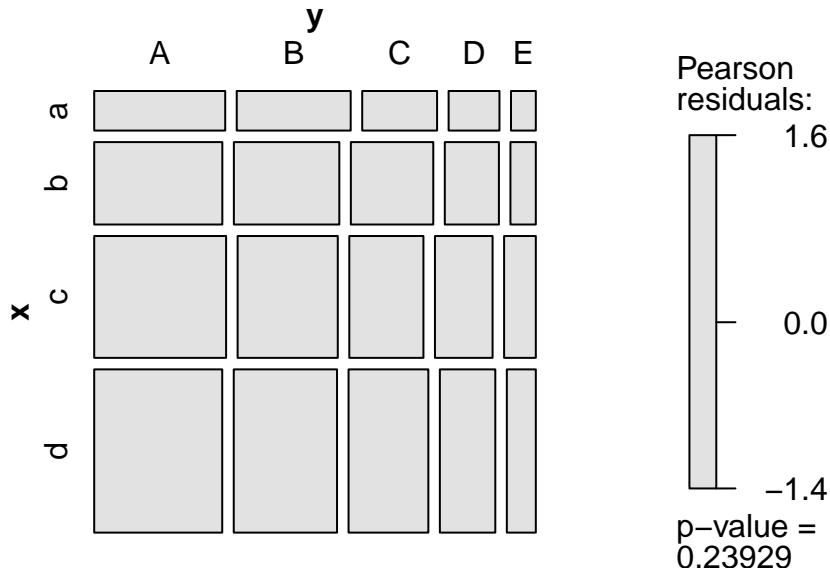
Los mosaicos muestran el tamaño relativo de cada celda de la tabla. Dependiendo de si la variable aparece en las filas o en las columnas, es más fácil leer su probabilidad marginal (altura de los rectángulos en el mosaico de ejemplo) o la condicional (su anchura, dentro de los niveles indicados por la otra variable).

Si filas y columnas fuesen independientes, las probabilidades condicionales serían iguales por lo que la anchura de los rectángulos vendría a ser similar. Que no es lo que ocurre más arriba.

En la versión de los mosaicos que he usado en este capítulo, el color (y la escala adjunta) indican la magnitud (y signo) de los *residuos de Pearson*. Los rectángulos *excesivos* (más) anchos de lo que les correspondería, están marcados en tonos azules; los más estrechos, de rojos.

A diferencia de la anterior, la siguiente es una tabla en la que, por construcción, filas y columnas son independientes. Las oscilaciones en el tamaño de las celdas se debe exclusivamente al azar.

```
set.seed(125)
n <- 10000
xy <- data.frame(x = sample(letters[1:4], n, replace = T, prob = (1:4) / sum(1:4)),
                   y = sample(LETTERS[1:5], n, replace = T, prob = (5:1) / sum(1:5)))
mosaic(table(xy), shade=TRUE, legend=TRUE)
```



Una prueba de independencia común (aunque existen alternativas y variantes) es la de la χ^2 , que en R puede realizarse así:

```
chisq.test(mi.tabla)
```

```
## Warning in chisq.test(mi.tabla): Chi-squared approximation may be incorrect
##
## Pearson's Chi-squared test
##
## data: mi.tabla
## X-squared = 106.66, df = 9, p-value < 2.2e-16
```

En este caso, el p-valor, muy pequeño, indica que hay indicios para pensar que los datos no son independientes entre sí. El color del pelo y de los ojos parecen guardar correlación. De todos modos, la prueba lanza un aviso (*warning*) de que puede resultar inadecuada: la prueba está basado en un resultado asintótico y cuando el número de observaciones en las casillas es demasiado bajo, podrían no darse las condiciones necesarias para que la aproximación fuese válida.

La prueba no hace nada particularmente sofisticado: compara los conteos reales de cada casilla con los que se obtendrían en caso de independencia, es decir, $E_{ij} = Np_i p_j$ (bajo independencia, la probabilidad de que una observación caiga en la casilla ij es $p_i p_j$; si hay N observaciones, el número esperado de conteos es $Np_i p_j$). Es decir, está basado en las diferencias

$$n_{ij} - E_{ij}.$$

En concreto, la prueba estudia la expresión

$$\sum_{ij} \frac{(n_{ij} - E_{ij})^2}{E_{ij}},$$

que, bajo ciertas condiciones (p.e., que haya un número mínimo de observaciones, que en la práctica no es particularmente grande) tiene una distribución aproximada χ^2 con $(r - 1)(c - 1)$ grados de libertad. No obstante, no es estrictamente necesario conocer este último resultado teórico: siempre es posible simular la distribución del estadístico anterior mediante remuestreos.

```

mi.tabla <- HairEyeColor[, , 2]
N <- sum(mi.tabla)
marginal.rows <- rowSums(mi.tabla) / N
marginal.cols <- colSums(mi.tabla) / N
probs.indep <- outer(marginal.rows, marginal.cols)

Eij <- N * probs.indep
Eij <- Eij[sort(rownames(Eij)), sort(colnames(Eij))]

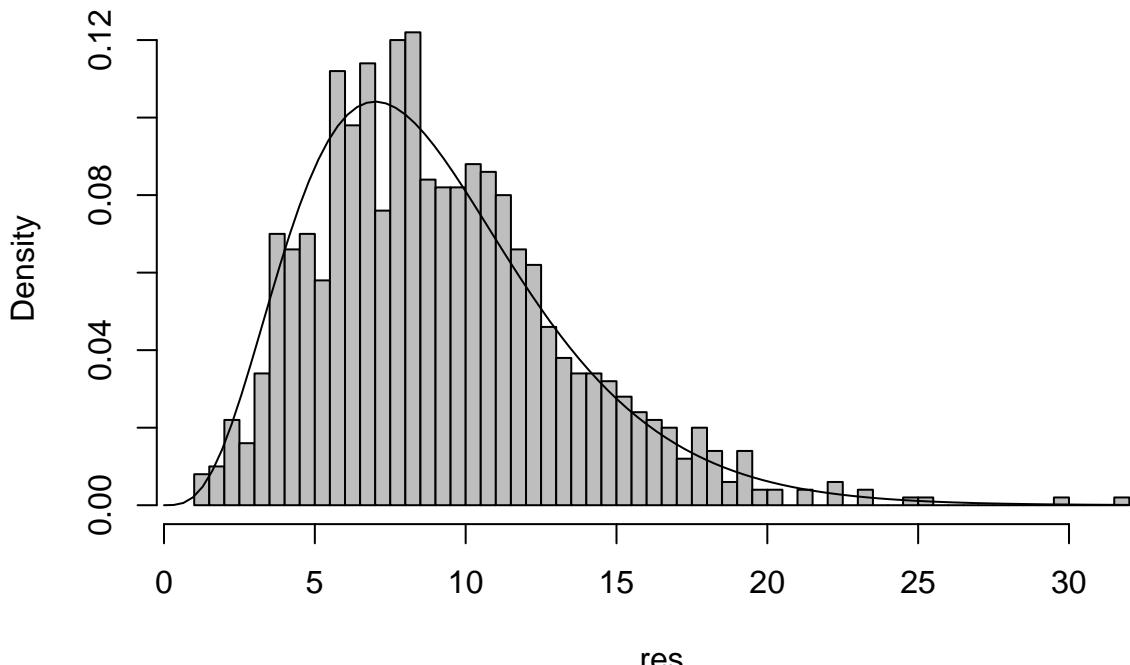
cols.sim <- rep(colnames(mi.tabla), times = colSums(mi.tabla))
rows.sim <- rep(rownames(mi.tabla), times = rowSums(mi.tabla))

res <- replicate(1000, {
  Oij <- table(data.frame(rows.sim, sample(cols.sim)))
  sum((Oij - Eij)^2 / Eij)
})

hist(res, freq = FALSE, col = "gray", breaks = 50,
     main = "Distribución del estimador bajo la\nhipótesis nula")
curve(dchisq(x, df = 9), from = 0, to = max(res), add = TRUE)

```

Distribución del estimador bajo la hipótesis nula

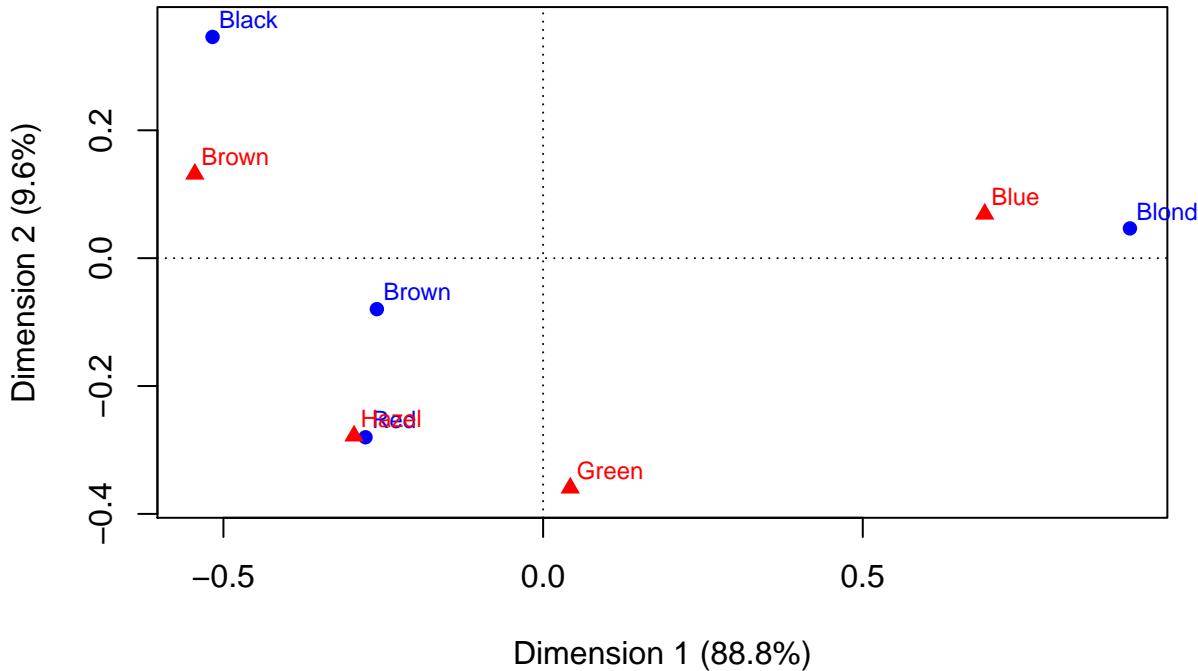


El código anterior permite simular la distribución de la tabla bajo la hipótesis de independencia. La representa en gris junto a la aproximación teórica, la de la χ^2 , sobreimpresa. El p-valor obtenido antes, mayor que 100, queda muy escorado con respecto a la distribución representada más arriba, lo que justifica su pequeño valor.

La simulación se ha realizado reordenando al azar la población de sujetos asignando a cada uno de ellos, independientemente de su color del pelo, un color de ojos (que es la hipótesis de independencia).

11.1. Biplots y dependencias

La independencia, estudiada en la sección anterior, es una propiedad poco interesante en la práctica. Normalmente, nuestros datos van a mostrar relaciones de dependencia manifiestas que querremos estudiar.



Una manera cualitativa de describir estas relaciones de dependencia es mediante los **biplots**, que permiten representar gráficamente (una transformación de) la matriz $O_{ij} - E_{ij}$, es decir, la diferencia entre los valores observados en la tabla y los que se obtendrían bajo condiciones de independencia.

La gráfica anterior muestra dos *dimensiones*. La primera es la más importante porque recoge el 88.8 % de la variabilidad de los datos contra el 9.6 % de la segunda. La principal lección que se extrae de la primera dimensión (en horizontal) es cómo el pelo rubio está muy vinculado a los ojos azules, a diferencia del resto de las combinaciones. La segunda muestra una asociación del pelo negro a los ojos marrones y del pelo rojo a los ojos de color castaño y verde.

11.2. Relaciones de dependencia y modelos loglineales

Existen otros métodos no tan cualitativos como los de la sección anterior de analizar las relaciones de dependencia en tablas. Como, por ejemplo, los **modelos loglineales**, que estudian la distribución del número de observaciones en la celda ij de la tabla, n_{ij} , en función de las variables involucradas. Si las probabilidades marginales de estas variables son p_{i+} y p_{+j} , entonces, de ser independientes, se tendría

$$p_{ij} = p_{i+}p_{+j}$$

y, como consecuencia,

$$n_{ij} \approx np_i + p_{+j},$$

o bien,

$$\log n_{ij} \approx \log n + \log p_{i+} + \log p_{+j}.$$

Más adelante, veremos cómo se pueden crear modelos (GLMs, en concreto) de la forma `log_n ~ x1 + x2`. Ese tipo de modelos permiten para modelar ese tipo de relaciones y estudiar si son adecuados (en cuyo caso podría defenderse la hipótesis de independencia de las variables) o son necesarios otros más complejos de la forma `log_n ~ x1 * x2`, que incluyen las **interacciones** entre las variables para modelarlos adecuadamente. Estos últimos modelos, los que incluyen interacciones, proporcionan una descripción más cuantitativa de la relación de dependencia entre variables. Además, este tipo de modelos pueden usarse en tablas de dimensiones superiores, en las que hay más de dos variables implicadas.

En esta sección, sin embargo, no vamos a estudiar este tipo de modelos. Ni en este libro a ocuparnos de las interacciones. Nos limitaremos a señalar que extensiones de algunos de los contenidos que vendrán a continuación pueden usarse para el estudio de este tipo de datos; de hecho, podría decirse que subsumen lo contado más arriba (p.e., el estudio de la independencia).

11.3. Referencias

- Gil Bellosta, C.J., *Una feliz conjunción estadístico-algebraica*, (partes primera y segunda), en la que se explican los fundamentos estadísticos y algebraicos de los *biplots*.
- XXX algo sobre modelos loglineales
- XXX Mencionar el tochazo de Agresti

12. Introducción a la modelización estadística

Un *modelo* es una representación simplificada de la realidad. Un *modelo probabilístico* es una representación simplificada de un sistema, habitualmente real, que genera datos: una moneda que se lanza al aire, economías que tienen un determinado PIB, tasa de paro, etc., centrales eólicas que producen cierta cantidad de energía eléctrica bajo determinadas condiciones atmosféricas, etc.

Un modelo probabilístico para el evento de obtener una cara al lanzar una moneda al aire podría ser

$$Y \sim \text{Bernoulli}(1/2).$$

Una vez dado por bueno, pueden plantearse sobre él problemas como, por ejemplo, el de calcular la probabilidad de obtener un número par de caras en n tiradas. Los problemas pueden ir de los más triviales a los más sofisticados, pero son siempre de naturaleza deductiva, i.e., las propiedades del sistema se derivan todas deductivamente de su formulación inicial.

Los modelos probabilísticos son infrecuentes en la práctica. Aparecen en los juegos de azar, algunos modelos físicos teóricos y libros de teoría de la probabilidad y poco más.

Un *modelo estadístico* es un modelo probabilístico incompleto: solo se conoce (o justifica) parte de su formulación. En el modelo probabilístico anterior, el parámetro de interés, la probabilidad de obtener cara, era conocido e igual a 1/2. Un modelo estadístico relacionado es

$$Y \sim \text{Bernoulli}(p),$$

donde únicamente se especifica la familia a la que pertenece la variable de interés, Y . Entonces, el problema fundamental consiste en encontrar un valor razonable para p a partir, típicamente, de un histórico de lanzamientos.

Este no es un problema de naturaleza deductiva sino inductiva. Es uno de los llamados *problemas inversos*: a partir de información sobre el funcionamiento de un sistema, encontrar un mecanismo generador plausible¹.

En el caso anterior, muy simple, existe un único modelo natural, uno de Bernoulli. En general, un mismo sistema puede ser modelado estadísticamente de muchas maneras distintas. En tales casos nuestro conocimiento sobre el modelo se extiende más allá de identificar un parámetro razonable, como en el ejemplo, y alcanza a su formulación misma. Pueden proponerse modelos alternativos, con sus méritos y sus deméritos, y parte del trabajo del estadístico consiste en elegir el más adecuado para un fin determinado. Hay que tener siempre presente que²:

Todos los modelos son incorrectos, aunque algunos resultan útiles.

Existe un conjunto de modelos *de libro* que se aplican igualmente en casos *de libro*. Sin embargo, muchos de los que se usan en la práctica son variaciones de ellos.

12.1. El modelo lineal

El modelo lineal se usa cuando existe una variable aleatoria objetivo Y que depende de otras variables aleatorias X_1, \dots, X_n a través del siguiente modelo estadístico:

$$Y | X_1, \dots, X_n \sim N\left(a_0 + \sum_i a_i X_i, \sigma^2\right)$$

La variable aleatoria Y es *condicionalmente normal* y su media depende de las variables X_1, \dots, X_n , i.e., tiene una media distinta para cada combinación de valores de X_1, \dots, X_n . De hecho, lo hace, y de ahí su nombre, a través de una función lineal de estas, $a_0 + \sum_i a_i X_i$. Por su parte, la varianza es fija e igual a σ^2 .

Esa varianza mide el *error irreducible* del modelo, i.e., el efecto sobre Y de otras variables no tenidas en cuenta, errores de medición, etc. Por ejemplo, si se busca modelar un fenómeno social, por mucho que se recojan las variables de los individuos (e.g., edad, sexo, estado civil, etc.) que más podrían determinar el comportamiento de los individuos, siempre quedarán al margen otros factores condicionantes de su voluntad. El modelo asume que su efecto, junto con otros, queda recogido en un *error* con distribución normal. Por otro lado, en el estudio de un fenómeno físico, aunque se recojan y consideran todas las variables que la teoría indica que tienen un efecto sobre el mismo, hay que tener en cuenta el error de los instrumentos de medida, que se manifiesta en forma de error irreductible. Ni que decir tiene que lo más habitual es que los errores de los modelos de fenómenos físicos sean mucho menores que los empleados en ciencias sociales.

Dados unos datos, el problema fundamental consiste en identificar un valor *adecuado* para los parámetros a_0, \dots, a_n . Se trata de un problema de estimación puntual como los estudiados anteriormente y que se resuelve canónicamente usando el método de la máxima verosimilitud.

12.1.1. Modelo lineal simple

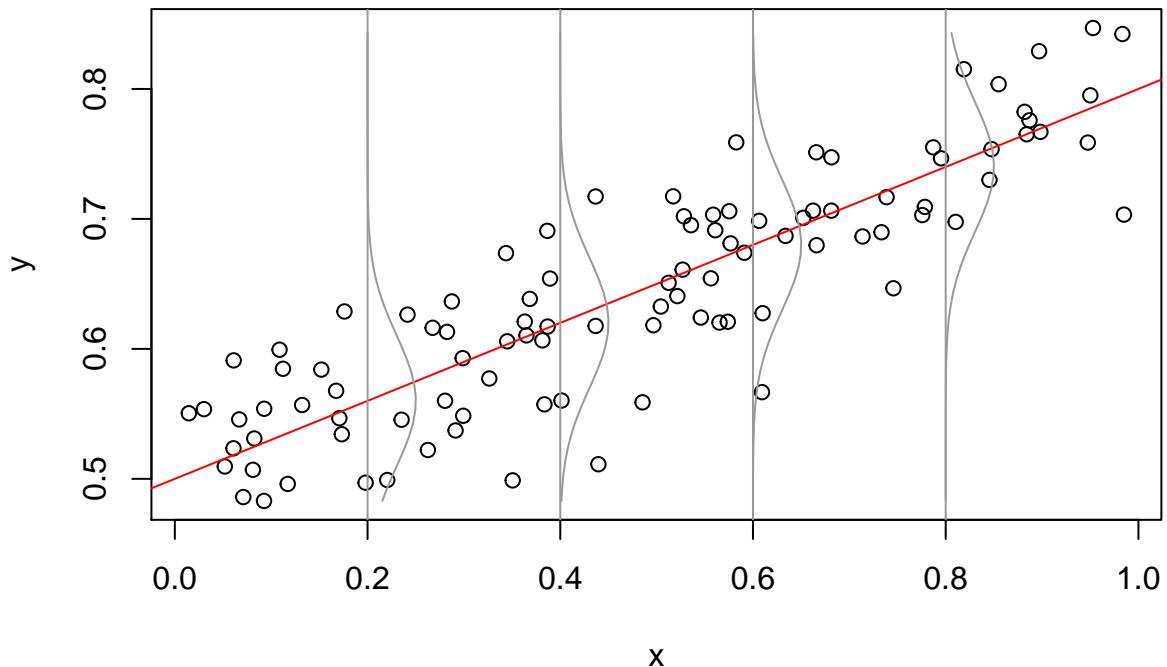
El **modelo lineal simple** es una versión del modelo lineal en el que Y depende de una única variable X . Concretamente,

$$Y | X = x \sim N(a + bx, \sigma^2)$$

¹Hay quien habría de identificar las causas de un efecto; sin embargo, ignoraremos el problema, sumamente esquivo, de la causalidad.

²Véase esto

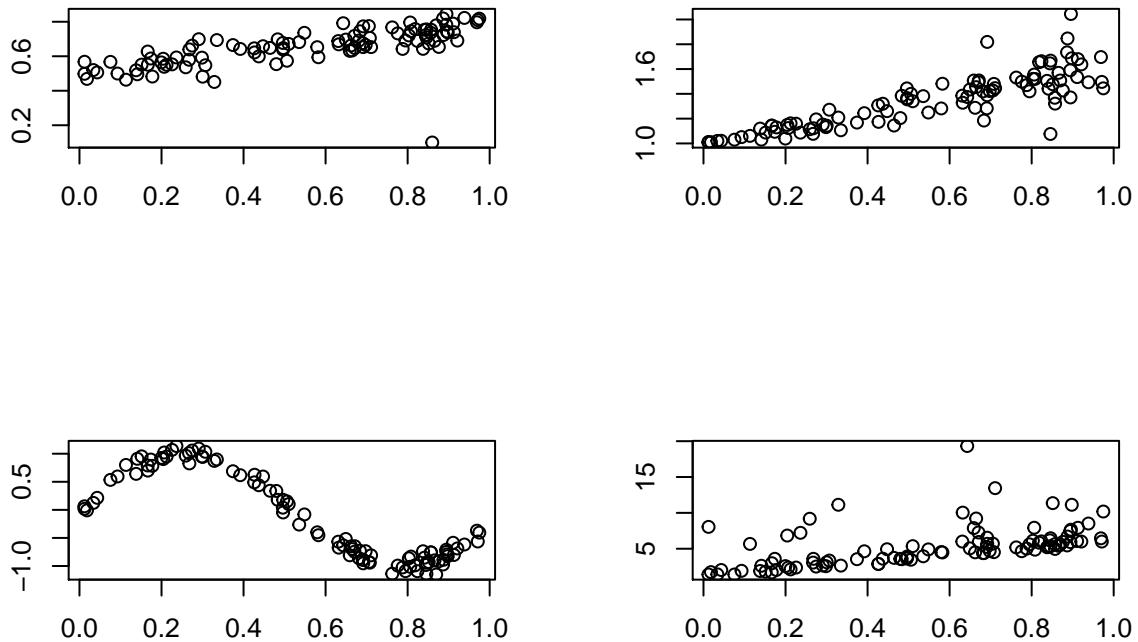
El principal interés del modelo lineal simple es pedagógica: se puede describir e interpretar gráficamente.



En el gráfico anterior se han representado datos con distribución $Y | X = x \sim N(0,5 + 0,3x, 0,05)$. Se ha representado también en rojo la recta de las medias condicionales. Finalmente, en algunas secciones verticales se han representado las correspondientes funciones de densidad condicionadas de Y .

Por definición, en el caso de la regresión lineal simple, si los datos se representan mediante un diagrama de dispersión, tienen que estar contenidos en una franja de anchura fija alrededor de una recta. La recta es el lugar geométrico de las medias condicionales y está descrita por los coeficientes a y b . La anchura de la franja, que viene a ser el error del modelo, está determinada por σ .

En el siguiente panel aparecen cuatro gráficos en los que se representan configuraciones de datos que no cumplen las condiciones anteriores.



En el primer caso, la distribución de los datos es idéntica a la de nuestro ejemplo anterior con una salvedad: existe un *outlier* bastante evidente, que viola la condición de normalidad de los errores³. En el segundo, la varianza no es constante porque aumenta con el valor de x ; los datos tienen una estructura de *trompeta* que ocurre con cierta frecuencia: a mayor media, mayor error. En el tercero, la media condicional no es lineal sino sinusoidal. Finalmente, en el último, la distribución condicional de Y es manifiestamente no normal: no es simétrica a ambos lados de la media; de hecho, los datos se han simulado utilizando una distribución lognormal, que es positiva y de cola larga.

Existen procedimientos para extender el uso del modelo lineal a alguno de los casos anteriores (desde regresiones robustas hasta determinadas transformaciones de los datos), cuyo estudio excede los objetivos que nos planteamos aquí.

12.1.2. Ajuste del modelo lineal

Una vez que se acepta como razonable que la distribución de los datos es compatible con la que exige el modelo lineal, surge el problema de la determinación de los coeficientes del modelo a partir de una muestra, que típicamente tiene la forma de un vector de valores y_i y una matriz con filas (x_{i1}, \dots, x_{in}) . Se supone que las muestras anteriores son mutuamente independientes⁴.

La estimación de los parámetros desconocidos del modelo, i.e., a_0, \dots, a_n y, habitualmente también, σ (para tener una medida del error) se hace por máxima verosimilitud. La función de verosimilitud, de hecho, es el producto (debido a la independencia) de densidades normales de media $a_0 + \sum_i a_i X_i$ y desviación estándar σ , i.e.,

$$L(a_0, \dots, a_n, \sigma) = \prod_i \frac{1}{(2\sigma^2\pi)^{1/2}} \exp\left(-\frac{(y_i - a_0 - \sum_j a_j x_{ij})^2}{2\sigma^2}\right)$$

Esa expresión, operando ligeramente queda de la forma

$$L(a_0, \dots, a_n, \sigma) = \frac{1}{(2\sigma^2\pi)^{n/2}} \exp\left(-\frac{\sum_i (y_i - a_0 - \sum_j a_j x_{ij})^2}{2\sigma^2}\right)$$

y es evidente cómo maximizar L (al menos, con respecto a los a_i) equivale a minimizar

$$l(a_0, \dots, a_n) = \sum_i \left(y_i - a_0 - \sum_j a_j x_{ij} \right)^2,$$

la pérdida cuadrática. Los detalles de cómo resolver ese problema de minimización, que puede hacerse mediante la aplicación de técnicas muy eficientes de álgebra lineal, pueden consultarse en casi cualquier otro manual⁵.

El hecho de que el ajuste del modelo tal como lo hemos planteado se reduzca a resolver un problema de minimización de una pérdida cuadrática permite que la técnica aquí descrita se extienda a casos en los que no se cumplen estrictamente las condiciones, bastante restrictivas, del modelo. De hecho, en muchos textos se ignora la formulación presentada aquí, en términos de la descripción probabilística del modelo y, simplemente, se sugiere un ajuste mediante ese procedimiento de minimización. Esta línea de razonamiento es muy propia de áreas como, por ejemplo, la econometría, en la que por motivos tal vez históricos, imperan las técnicas de minimización de pérdidas cuadráticas sin que preocupen demasiado las especificaciones probabilísticas subyacentes.

³Recuérdese que, a efectos prácticos, la distribución normal no tiene *outliers*

⁴P.e., que no son muestras de una serie temporal, etc.; véase la discusión al respecto en la sección sobre dependencia e independencia de variables aleatorias.

⁵XXX: añadir una referencia en línea.

Minimización por mínimos cuadrados

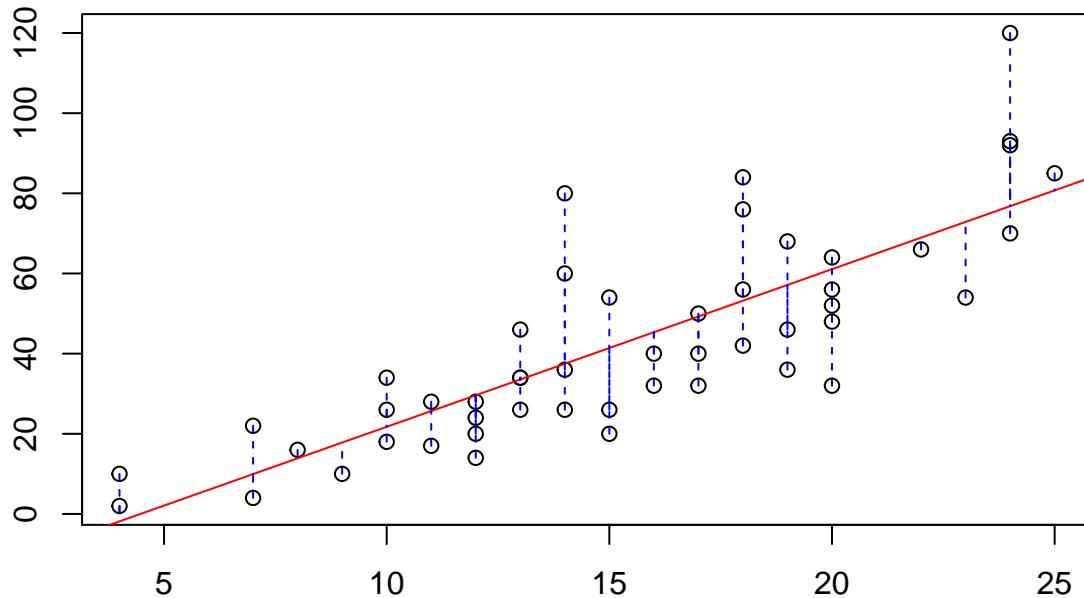
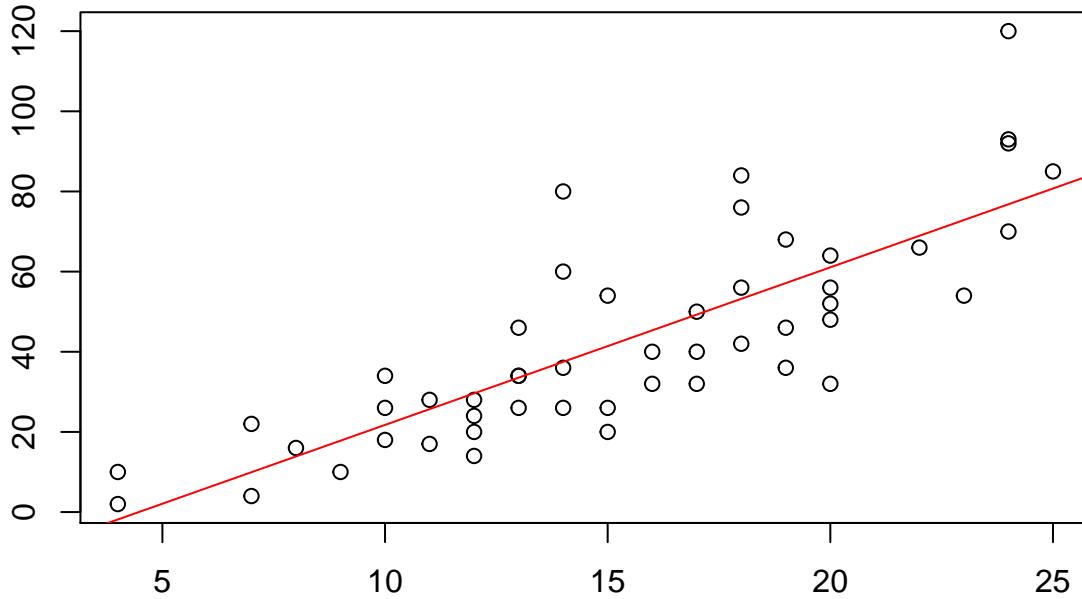


Figura 1: En el gráfico se muestran, como líneas azules a trazos, los segmentos cuyas longitudes (más propiamente, la suma de cuyas longitudes al cuadrado), se quieren minimizar para encontrar la recta óptima.

Desde un punto de vista puramente operacional, el problema puede resolverse así en R:

Velocidad vs distancia de frenado



```
modelo <- lm(dist ~ speed, data = cars)
```

El código anterior construye la regresión lineal de la variable `dist` sobre `speed`. Toda la información sobre el modelo queda resumida en el objeto `modelo` y sus detalles pueden consultarse con `summary`:

```

summary(modelo)

##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##    Min     1Q Median     3Q    Max 
## -29.069 -9.525 -2.272  9.215 43.201 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -17.5791   6.7584  -2.601  0.0123 *  
## speed        3.9324   0.4155   9.464 1.49e-12 *** 
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438 
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12

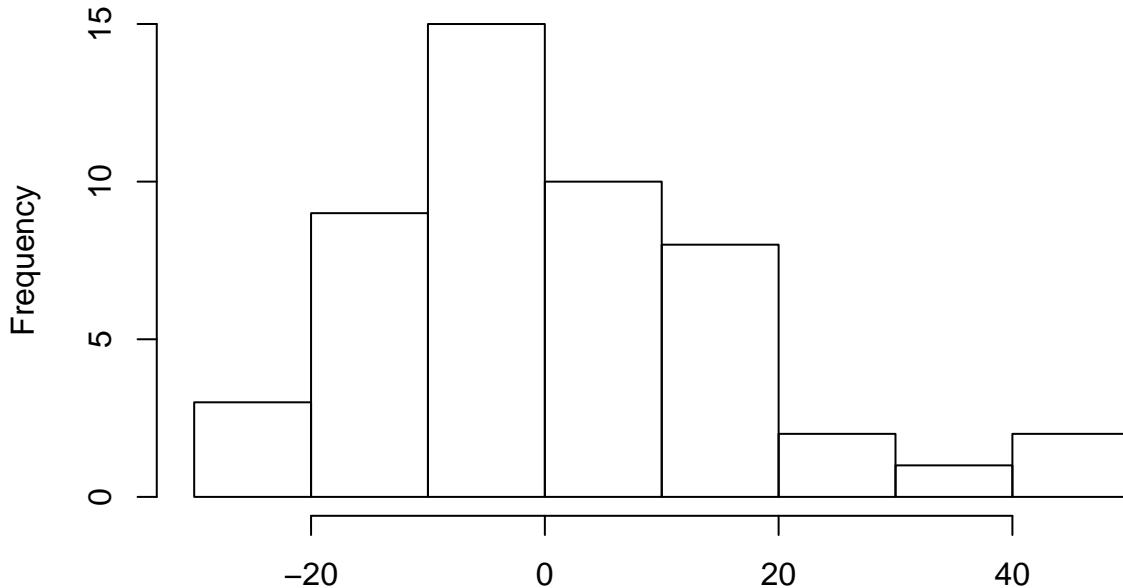
```

El resumen del modelo contiene, en primer lugar, la fórmula usada y unos cuantos estadísticos sobre la distribución de los residuos, i.e., la diferencia entre los valores reales, y_i , y las medias estimadas, $a_0 + \sum_i a_i x_i$. Valores demasiado grandes (o grandes con relación a sus valores típicos) pueden indicar problemas de ajuste, i.e., que los datos no se ajustan a la estructura probabilística subyacente que permite usar el modelo.

Puede obtenerse un histograma de estos residuos haciendo

```
hist(modelo$residuals)
```

Histogram of modelo\$residuals



modelo\$residuals

En él se aprecia una cierta desviación con respecto a la distribución normal. Esto puede ser síntoma de una mala especificación. De hecho, el diagrama de dispersión de los datos originales muestra una cierta estructura de

trompeta: a mayor velocidad, parece haber más dispersión en la distancia.

A continuación se describen los coeficientes. Sus valores corresponden al término independiente y la pendiente de la recta de regresión. Según este modelo, la distancia se incrementa en casi cuatro unidades por cada unidad adicional de velocidad. Además, a velocidad 0 le corresponde una distancia de frenado de -17.6 unidades, que es totalmente irreal: este modelo puede utilizarse, a lo más, en un rango concreto de velocidades.

El resto de las columnas de esa tabla resumen una serie de pruebas de hipótesis implícitas: estudian si el coeficiente es o no cero (en el sentido de Fisher). P-valores tan pequeños como los que arroja el modelo indican que los coeficientes son significativamente distintos de cero (cosa que no debería sorprendernos: en la nube de puntos se aprecia a simple vista una clara estructura creciente). Los intervalos de confianza para estos coeficientes pueden obtenerse así:

```
confint(modelo)
```

```
##             2.5 %    97.5 %
## (Intercept) -31.167850 -3.990340
## speed        3.096964  4.767853
```

Estos intervalos de confianza no contienen el valor 0, en consistencia con la prueba de hipótesis indicada más arriba.

Debajo de esa tabla aparecen tres líneas adicionales que discuten el error del modelo. Lo que llama **residual standard error** no es otra cosa que la estimación de σ , que también puede obtenerse haciendo

```
sigma(modelo)
```

```
## [1] 15.37959
```

Las R^2 , ajustada y sin ajustar, miden el porcentaje de la variabilidad original de los datos que captura el modelo. De alguna manera, si vale la pena plantear el modelo o no. Nuestro modelo captura alrededor del 65 % de la varianza original de los datos. En concreto, la R^2 (sin ajustar) no es otra cosa que

```
1 - var(modelo$residuals) / var(cars$dist)
```

```
## [1] 0.6510794
```

La R^2 suele comunicarse como medida de la bondad del modelo. Curiosamente, el estándar varía sustancialmente entre áreas de aplicación: en las ciencias experimentales se exigen R^2 muy altas ($> 90\%$), mientras que en ciencias sociales, valores alrededor de 30 % o 40 % tienden a darse por buenas.

La prueba de hipótesis que se resume en la última línea estudia si esa reducción del error es o no *significativa*, i.e., si el modelo ayuda a entender *algo* los datos. No es particularmente útil porque indica si el modelo es terriblemente malo (o no); sin embargo, lo que nos interesa saber es si el modelo es suficientemente bueno (o no).

12.1.3. Análisis de la varianza para la comparación de modelos

La prueba de hipótesis que aparece al pie del resumen de un modelo lineal en R es, como se ha indicado, poco informativa: compara el modelo con un no-modelo (o, más propiamente, un modelo trivial, $Y \sim N(\mu, \sigma)$).

Pero, como se indicaba al inicio de esta sección, es habitual considerar y plantear modelos distintos para el análisis de un mismo conjunto de datos. Dos modelos plausibles para los datos anteriores son:

```
modelo.0 <- lm(dist ~ speed, data = cars)
modelo.1 <- lm(dist ~ speed + I(speed^2), data = cars)
```

El primero es el que venimos usando a lo largo de la sección. El segundo es más *realista* desde el punto de vista físico: como la energía cinética de un cuerpo depende del cuadrado de su velocidad, la distancia de frenado dependerá de esa cantidad (que en R se puede indicar como en la fórmula correspondiente al

`modelo.1`, i.e., usando la expresión `I(speed^2)`). Entonces, ¿es (significativamente) mejor el segundo modelo que el primero?

Es mejor, sin duda: puesto que incluye al primero, su error tiene que ser necesariamente menor. ¿Pero lo es tanto como para justificar la inclusión de un término adicional?

Igual que la prueba mencionada al principio de la sección servía para comparar un modelo lineal con un modelo trivial (a través de la comparación de sus residuos), es posible comparar dos modelos en R (al menos, si uno está incluido dentro del otro) así:

```
anova(modelo.0, modelo.1)
```

```
## Analysis of Variance Table
##
## Model 1: dist ~ speed
## Model 2: dist ~ speed + I(speed^2)
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1     48 11354
## 2     47 10825  1    528.81 2.296 0.1364
```

Lo que devuelve `anova` es el resultado de una prueba de significancia que estudia la mejora que el modelo más complejo supone con respecto al primero. La columna `RSS` indica la suma de cuadrados del residuo de ambos modelos (los errores al cuadrado) y la `Res.Df`, los **grados de libertad**. El modelo más complicado sustrae un grado de libertad (porque hay un coeficiente más⁶) para reducir la suma de cuadrados del residuo en unas 529 unidades. La prueba de significancia tiene un p-valor de 0.13, indicando que no encuentra mejora significativa entre ambos modelos. Es decir, que los datos no justifican el uso del modelo más complejo.

Esta técnica puede usarse y, de hecho, se usa, para descartar variables potenciales al diseñar un modelo estadístico de este (y algún otro tipo adicional).

12.1.3.1. Complejidad y generalización: una digresión

En esta sección se han comparado dos modelos distintos. Se ha afirmado, sin demostración, que uno de ellos, el más complejo (o menos *parsimonioso*) tiene un error menor (al menos, si se mide en términos de la R^2). Eso es generalmente cierto. Podría plantearse un modelo con tantos coeficientes que el error fuese cero.

John von Neumann planteó esta cuestión de manera humorística⁷:

Con cuatro parámetros puedo ajustar un elefante y con cinco puedo hacer que mueva la trompa.

El problema de un modelo demasiado complejo es que puede capturar, además de la señal presente en los datos con los que se entrena, el ruido específico que trae consigo. Eso tiene consecuencias indeseables: por ejemplo, si se usa para predecir, puede ser inferior (i.e., dar predicciones más imprecisas) que otro modelo más simple. Se dice entonces que el modelo más complejo tiene problemas de generalización (o que no generaliza bien).

De contar con más datos (p.e., en entornos *big data*), se pueden usar procedimientos alternativos al mostrado en esta sección basados el contraste de los modelos construidos sobre un subconjunto determinado de los datos sobre el resto. La técnica mostrada en esta sección es un estudio de los posibles problemas de generalización de un modelo cuando no se cuenta con datos adicionales basado en resultados teóricos. Que, además, solo se puede aplicar en una clase pequeña de modelos.

Podría decirse que se trata de un residuo teórico de una época preinformática. Hoy en día es más natural utilizar extensiones del modelo lineal considerado aquí como *elastic-net*.

⁶No se entienda esta afirmación como una explicación en sí misma: las razones últimas tienen que ver con la dimensión de ciertas matrices que no se tratan aquí.

⁷Véase esto

12.1.4. El modelo lineal y (algunas) pruebas de hipótesis

Algunas pruebas de hipótesis pueden subsumirse dentro del modelo lineal; otras, veremos, dentro de algunas de sus generalizaciones. El caso de libro es el de la prueba de Student, del que se ocupa esta sección.

Vamos a plantear un pequeño ejemplo en el que compararemos la media de los vectores:

```
x <- rnorm(20)
y <- rnorm(20, mean = 1)
```

Podemos crear dos vectores a partir de ellos: el de los valores observados, `obs` y el `tratamiento`, que distingue los dos grupos. Lo llamaremos así porque, típicamente, los grupos se distinguen por la aplicación a uno de ellos de una determinada intervención cuyo efecto se desea medir.

```
obs <- c(x, y)
tratamiento <- c(rep(0, length(x)), rep(1, length(y)))
```

Podemos usar, como antes, la prueba de Student:

```
t.test(obs ~ tratamiento)
```

```
##
##  Welch Two Sample t-test
##
## data: obs by tratamiento
## t = -2.8337, df = 37.374, p-value = 0.007378
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -1.7853342 -0.2969309
## sample estimates:
## mean in group 0 mean in group 1
## 0.08776541 1.12889796
```

Pero, alternativamente, podemos plantear un modelo lineal:

```
modelo <- lm(obs ~ tratamiento)
```

Así planteado, el coeficiente de la variable `tratamiento` será la diferencia esperada de medias. En efecto,

```
summary(modelo)
```

```
##
## Call:
## lm(formula = obs ~ tratamiento)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.30027 -0.85885 -0.09852  0.56096  2.69898
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.08777  0.25980  0.338  0.73736
## tratamiento 1.04113  0.36741  2.834  0.00733 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.162 on 38 degrees of freedom
## Multiple R-squared:  0.1744, Adjusted R-squared:  0.1527
## F-statistic:  8.03 on 1 and 38 DF,  p-value: 0.007327
```

No debería sorprender que el p-valor de la prueba de confianza asociada al coeficiente `tratamiento` sea idéntico al de la prueba de Student. Además, los intervalos de confianza,

```
confint(modelo)
```

```
##           2.5 %    97.5 %
## (Intercept) -0.4381752 0.613706
## tratamiento   0.2973402 1.784925
```

son similares a los obtenidos con `t.test`. La diferencia radica en que los de `t.test` utilizan la distribución t de Student, mientras que lo que se obtienen como subproducto de `lm` usan la aproximación normal. Por lo que ya sabemos sobre la distribución t, esta diferencia debería decrecer al aumentar el número de observaciones. XXXX: añadir ejercicio.

¿Existen motivos para usar una prueba de Student en lugar de un modelo lineal? En el siglo XXI (y siguientes), prácticamente no. La prueba de Student tiene una ventaja notable: exige menos cálculos y se puede aplicar, prácticamente, a mano. Pero esta ventaja ha quedado prácticamente reducida a nada por los ordenadores. Sin embargo, la prueba de Student no permite considerar el efecto de otras variables adicionales, como en el siguiente ejemplo:

```
set.seed(123)
```

```
n <- 30
z <- c(rep(0, 10), rep(1, 30), rep(0, 20))
tratamiento <- c(rep(0, n), rep(1, n))

coef.tratamiento <- 2/3

obs <- 2 * z + coef.tratamiento * tratamiento + rnorm(2*n)
```

Los datos en este pequeño ejemplo están seleccionados de tal modo que, a pesar de que existe un tratamiento que tiene un efecto no nulo, las medias por tratamiento son similares. Es consecuencia del efecto de una variable adicional, `z`, que confunde a la prueba:

```
t.test(obs ~ tratamiento)
```

```
##
##  Welch Two Sample t-test
##
## data: obs by tratamiento
## t = -0.65729, df = 57.952, p-value = 0.5136
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.9120221 0.4611379
## sample estimates:
## mean in group 0 mean in group 1
##          1.286230          1.511672
```

En efecto, dado que las medias en ambos grupos son similares, la prueba de Student, ciega al efecto de `z`, no descubre efecto alguno. Sin embargo, esto no ocurre en el modelo lineal siguiente:

```
modelo <- lm(obs ~ tratamiento + z)
summary(modelo)
```

```
##
## Call:
## lm(formula = obs ~ tratamiento + z)
##
## Residuals:
```

```

##      Min     1Q Median     3Q    Max
## -1.9250 -0.5713 -0.1105  0.5564  1.9961
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.05809   0.23727 -0.245 0.807460
## tratamiento  0.89760   0.25166  3.567 0.000741 ***
## z            2.01648   0.25166  8.013 6.54e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9189 on 57 degrees of freedom
## Multiple R-squared:  0.5332, Adjusted R-squared:  0.5168
## F-statistic: 32.55 on 2 and 57 DF,  p-value: 3.718e-10

```

Y, de hecho, los intervalos de confianza son razonables (aunque, tal vez, demasiado anchos porque la muestra es pequeña):

```
confint(modelo)
```

```

##           2.5 %    97.5 %
## (Intercept) -0.5332089 0.4170246
## tratamiento  0.3936655 1.4015404
## z            1.5125452 2.5204200

```

Esta no es una situación tan atípica. Las pruebas de hipótesis se utilizan normalmente en contextos muy controlados, donde el efecto de las posibles *variables confusoras* se elimina a través de un **diseño experimental** muy preciso. Este diseño experimental aplicado a nuestro conjunto de datos, habría producido otro distinto en el que la presencia de variables con los distintos niveles de *z* estuviese equilibrado en ambos grupos. Sin embargo, generalmente, en estudios observacionales (a diferencia de los experimentales bien planeados) no es habitual que exista ese equilibrio. Por eso es necesario tener en cuenta el efecto de todas las posibles variables confusoras y una manera de hacerlo es mediante el uso de modelos como el de más arriba.

La única gran ventaja de la prueba de Student con respecto a la alternativa aquí planteada es que para muestras pequeñas, al utilizar la distribución de Student, los p-valores y los intervalos de confianza son más fiables (y más conservadores, i.e., más anchos). Sin embargo, a partir de un número no muy elevado de grados de libertad (10 o 15), las distribuciones de Student y la normal son prácticamente indistinguibles. Por eso, salvo para casos cada vez más atípicos, el recurso al test de Student está poco justificado.

Finalmente, hay que recordar que estas consideraciones no se aplican únicamente a la prueba de Student. Existen más pruebas *de libro* que pueden subsumirse también en otros tipos de modelos. Entre ellos, los que veremos a continuación.

12.1.5. El modelo lineal y predicciones

Hablar de *predict* y de la necesidad de dar intervalos de confianza. Relacionarlo con el error irreducible.

12.1.6. Más sobre el modelo lineal

Existen cuestiones adicionales y muy relevantes en la práctica sobre el modelo lineal que hemos omitido. Por ejemplo, la selección de variables o el tratamiento de variables categóricas.

También suelen estudiarse aparte (o como apéndice particular de los modelos lineales) los llamados modelos ANOVA, que tampoco trataremos aquí.

Son importantes también el estudio de transformaciones de los datos. Como por ejemplo, el uso de la transformación logarítmica para transformar un modelo aditivo en otro multiplicativo. O la forma de incorporar variables cuyo efecto no es lineal (por ejemplo, el efecto de la hora de día en la intensidad de la radiación solar).

También queda fuera del estudio de las interacciones entre variables predictoras, con las que se intenta modelar la relación entre diversas variables (un ejemplo hipotético: que el efecto de la edad sea creciente para hombres y decreciente para mujeres).

XXX Nota: un buen libro sobre el modelo lineal. Harrell?

12.2. Regresión logística

La regresión logística sirve para modelizar problemas en los que la variable objetivo tiene una distribución de Bernoulli con una probabilidad *de éxito* p que depende de una serie de variables adicionales. En concreto,

$$Y | X_1, \dots, X_n \sim \text{Bernoulli} \left(p \left(a_0 + \sum_i a_i X_i \right) \right)$$

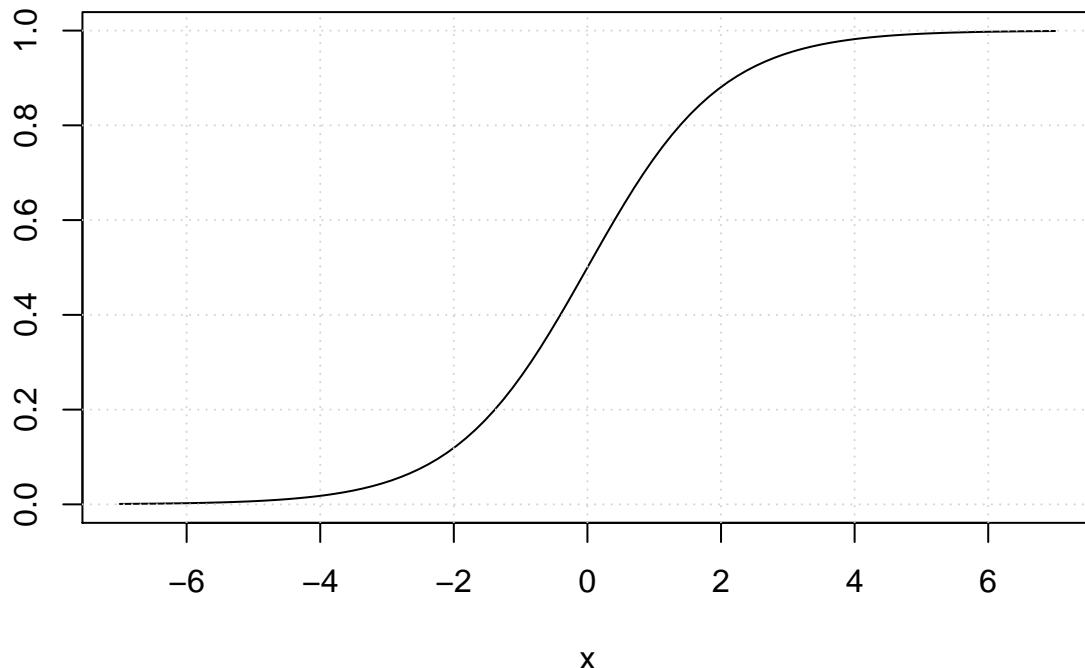
De otra manera:

- Y toma valores 1 o 0, que pueden representar fraude / no fraude, éxito / fracaso, etc.
- La probabilidad de éxito (i.e., $Y = 1$) depende de variables aleatorias X_1, \dots, X_n .
- Esta dependencia se manifiesta a través de una relación lineal, $a_0 + \sum_i a_i X_i$, de ellas.
- Existe una función *de enlace* que vincula esa expresión lineal con la probabilidad, p . Aunque hay otras opciones, típicamente se suele utilizar la función logística,

$$f(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)}$$

La función logística convierte cualquier valor real en un número entre 0 y 1, i.e., en un valor interpretable como una probabilidad.

Función logística



A la vista de la gráfica anterior, cuando la expresión $a_0 + \sum_i a_i X_i$ es positiva, la probabilidad de éxito es más alta que cuando es negativa. Las variables con coeficiente positivo tienden a hacer crecer la probabilidad y las que tienen coeficiente negativo, a hacerla decrecer.

El problema fundamental de la regresión logística consiste en identificar los parámetros a_0, \dots, a_n más adecuados para modelar un conjunto de datos en cuestión. Este es, de nuevo, un problema de estimación puntual que puede resolverse, como con el modelo lineal, mediante la maximización de la verosimilitud.

Si se obtiene un conjunto de datos de observaciones mutuamente independientes y_i con sus correspondientes x_{i1}, \dots, x_{in} , la función de verosimilitud es

$$L(a_0, \dots, a_n) = \prod_{y_i=1} p \left(a_0 + \sum_i a_i X_i \right) \prod_{y_i=0} \left(1 - p \left(a_0 + \sum_i a_i X_i \right) \right)$$

de la que podría obtenerse el mínimo mediante diversos procedimientos. Por ejemplo, usando optimización numérica como en el siguiente ejemplo sencillo.

12.2.0.1. Ejemplo de optimización con optim

Podemos plantear un pequeño ejemplo. Primero vamos a crear un conjunto de datos usando unos parámetros conocidos (preestipicados):

```
set.seed(1234)
n <- 1000
a <- -0.5
b <- 1
x <- runif(n)
```

Merece la pena prestar atención a la creación de la variable objetivo Y , que vamos a construir así:

```

logistic.foo <- function(x) exp(x) / (1 + exp(x))
probs <- logistic.foo(a + b * x)
y <- sapply(probs, function(prob) rbinom(1, 1, prob))

```

Cada valor y_i es un valor 0 o 1, con la probabilidad de 1 igual a la que determina su correspondiente valor x_i .

Ahora podemos construir la función de verosimilitud. En realidad, usaremos el logaritmo de la verosimilitud porque el producto de cantidades pequeñas tiende a producir problemas de estabilidad numérica⁸. Luego usaremos optim para encontrar su máximo:

```

logL <- function(pars){
  a <- pars[1]; b <- pars[2]
  probs <- logistic.foo(a + b * x)
  sum(log(probs[y == 1])) + sum(log(1 - probs[y == 0]))
}

optim(c(0,0), logL, control = list(fnscale = -1))$par
## [1] -0.7119215 1.2164085

```

Esos valores coinciden con los que proporciona R al usar la función con la que se ajusta el modelo logístico, glm:

```

modelo <- glm(y ~ x, family = "binomial")
modelo$coefficients

```

```

## (Intercept)          x
## -0.7122414   1.2168933

```

La función glm ofrece mucho más que el valor de los parámetros. Podemos extraer el resto de la información con summary:

```

summary(modelo)

##
## Call:
## glm(formula = y ~ x, family = "binomial")
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.3975  -1.1115  -0.9092   1.1643   1.4907
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.7122     0.1313  -5.423 5.87e-08 ***
## x            1.2169     0.2237   5.440 5.33e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 1384.2 on 999 degrees of freedom
## Residual deviance: 1353.7 on 998 degrees of freedom
## AIC: 1357.7
##
## Number of Fisher Scoring iterations: 4

```

⁸Trabajar en escala logarítmica es, de hecho, lo que se suele hacer habitualmente para manipular funciones de verosimilitud.

Además de los coeficientes, proporciona el resultado de una prueba de confianza con la hipótesis nula de que el coeficiente es cero. Igual de lo que pasaba con la regresión lineal, una variable con un coeficiente nulo no tendría influencia en el modelo, por lo que podría descartarse en una versión ulterior.

La interpretación del valor de los coeficientes es un poco más complicada que en el caso del modelo lineal. En este último, el valor de un coeficiente a_i se interpreta de la siguiente manera: un incremento de una unidad en X_i *ceteris paribus* produce un incremento (medio) de a_i unidades en Y .

En el caso del modelo logístico, como

$$p = \frac{\exp(a_0 + \sum_i a_i x_i)}{1 + \exp(a_0 + \sum_i a_i x_i)},$$

podemos despejar

$$\exp\left(a_0 + \sum_i a_i x_i\right) = \frac{p}{1-p}$$

y, por lo tanto, si la variable x_i aumenta en una unidad, dividiendo,

$$\exp(a_i) = \frac{p_1/(1-p_1)}{p_0/(1-p_0)},$$

que es una expresión que a la derecha tiene lo que se conoce como *odds ratio*(XXX referencia). Su interpretación no es del todo directa. Lo es por ejemplo, cuando $p_0, p_1 \sim 0$ (y, frecuentemente, la regresión logística se usa en contextos en que las probabilidades son pequeñas); en tal caso, es, aproximadamente, la razón de probabilidades.

12.3. Modelos lineales generalizados

En la sección anterior hemos planteado un pequeño ejemplo en el que hemos maximizado la verosimilitud en el contexto de un modelo logístico usando métodos numéricos. En concreto, los proporcionados por `optim` en R.

Pero, en general, `optim` no es viable. En la práctica, es demasiado lento. Cualquier tipo de método de los usados ya sea en estadística o en ciencia de datos no solo debe plantear un problema de optimización sino, además, proporcionar algún algoritmo, generalmente, *ad hoc*, para resolverlo.

En 1972, Nelder y Wedderburn publicaron un artículo, *Generalized Linear Methods*, en el que proponían una técnica eficiente para maximizar la verosimilitud de modelos del tipo

$$Y | X_1, \dots, X_n \sim D\left(f\left(a_0 + \sum_i a_i X_i\right)\right)$$

para una clase amplia de distribuciones D , las llamadas distribuciones de la familia exponencial. Estas distribuciones incluyen la normal y la binomial, ya tratadas más arriba, así como la Poisson, la gamma y otras.

Como consecuencia, es posible plantear modelos en los que la variable objetivo es, por ejemplo, un conteo usando modelos de Poisson. En tal caso, el parámetro λ depende para cada observación de una serie de variables. Y lo hace a través de una expresión lineal (con coeficientes por determinar) y una función, f , que recibe el nombre de **función de enlace**⁹.

⁹Propiamente, y por motivos históricos, la función de enlace es su inversa.

12.3.1. Ejemplo de una regresión de Poisson

Vamos a tanto a plantear como a resolver un problema de regresión de Poisson. Las dos partes tienen su interés y aunque podríamos haber elegido algún conjunto de datos real, es bueno asentar a través de un ejemplo cómo deberían ser teóricamente los datos para los cuales es válido este tipo de GLM.

Primero, vamos a definir los parámetros del modelo, que son los que trataremos de estimar después:

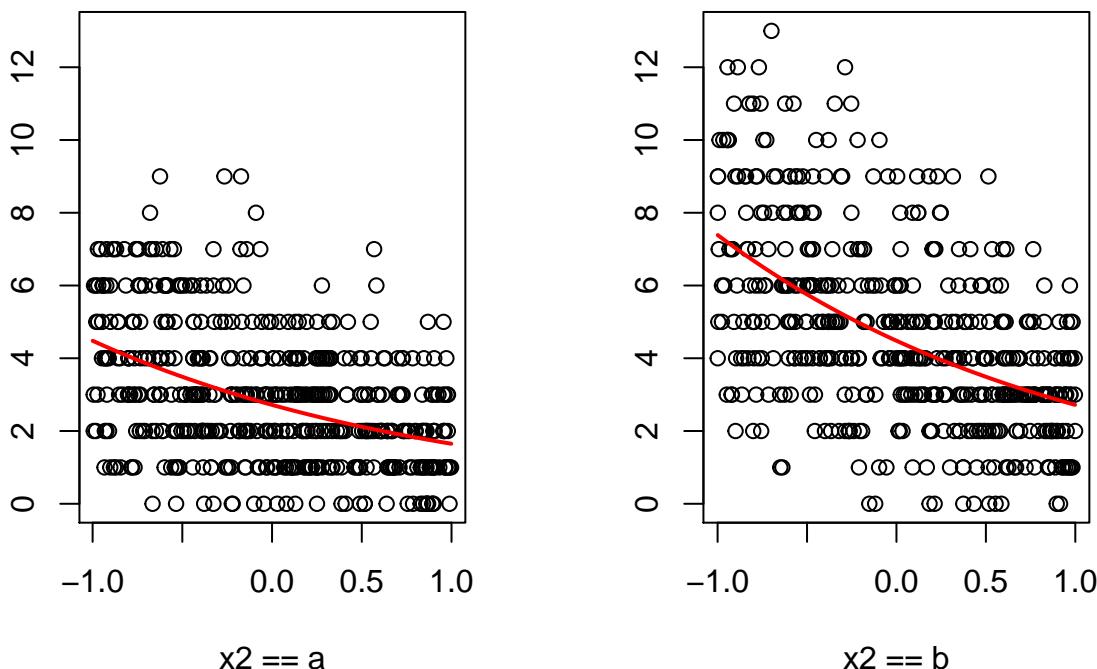
```
n <- 1000
a0 <- 1
a1 <- -0.5
a2 <- 0.5
```

A continuación, vamos a generar datos. En primer lugar, las variables independientes, una continua y otra binaria:

```
set.seed(1234)
x1 <- runif(n, -1, 1)
x2 <- sample(c("a", "b"), n, replace = T)
```

Luego, la variable dependiente. Para ello, construimos la expresión lineal (usando nuestros coeficientes predeterminados). Luego calculamos el λ correspondiente a cada observación. Obviamente, λ es una cantidad positiva, algo que no puede asegurarse de la expresión lineal. De ahí el uso de la función exponencial (i.e., el uso del logaritmo, su inversa, como función de enlace). Finalmente, simulamos una variable objetivo y usando un valor distinto de λ , el inducido por las variables independientes, para cada observación.

```
lin.expr <- a0 + a1 * x1 + a2 * (x2 == "b")
lambdas <- exp(lin.expr)
y <- sapply(lambdas, function(lambda) rpois(1, lambda))
```



```
## Warning in par(mrow = c(1, 1)): "mrow" is not a graphical parameter
```

El gráfico anterior muestra los datos en dos paneles (que corresponden a los dos valores de la variable binaria) de acuerdo con los valores de la continua. Además, se ha sobreimpresionado en color rojo la curva de valores del valor real del parámetro λ en función de la variable x_1 .

Ahora podemos ajustar el modelo:

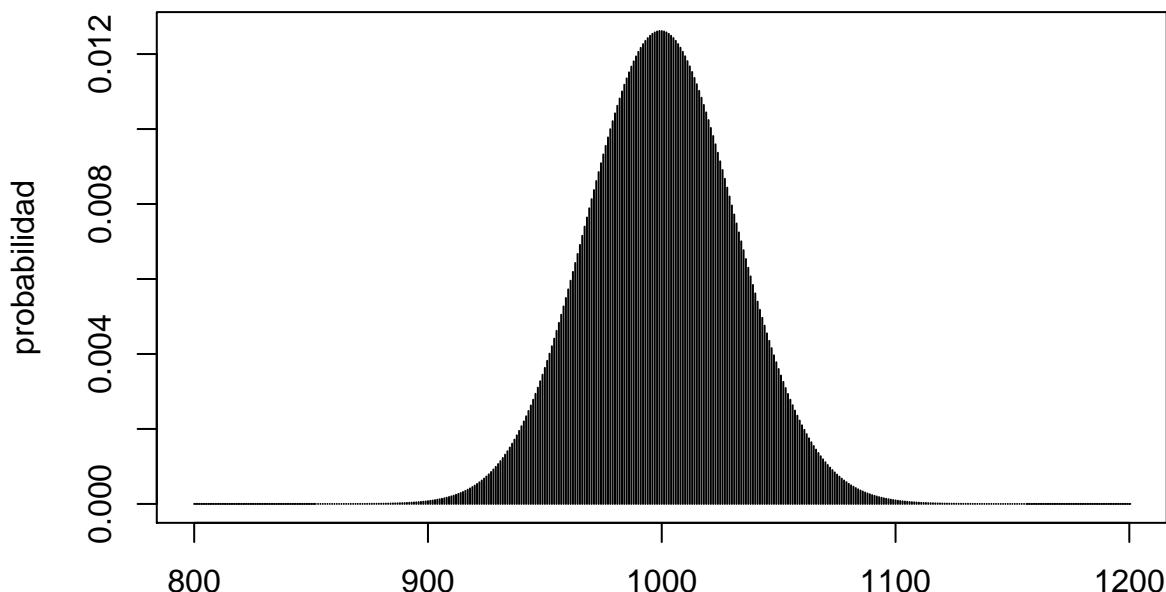
```
modelo <- glm(y ~ x1 + x2, family = "poisson")
summary(modelo)

##
## Call:
## glm(formula = y ~ x1 + x2, family = "poisson")
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -3.12098 -0.74931 -0.06633  0.62743  2.73582
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.03819   0.02609 39.79   <2e-16 ***
## x1          -0.49003   0.02856 -17.16   <2e-16 ***
## x2b         0.47058   0.03325 14.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance: 1548.3  on 999  degrees of freedom
## Residual deviance: 1062.9  on 997  degrees of freedom
## AIC: 4026.6
##
## Number of Fisher Scoring iterations: 5
```

Como puede apreciarse, la estimación de los coeficientes es una aproximación relativamente buena de los fijados más arriba.

No es sorprendente en este caso: los datos son artificiales y están construidos *ad hoc*. En la práctica, es raro encontrar conteos cuya distribución sea exactamente Poisson. El modelo de Poisson es bastante restrictivo: la media tiene que ser igual a la varianza: λ . Pero es frecuente observar conteos con **sobredispersión**, i.e., que tienen una varianza muy superior a la que le asociaría el modelo.

Probabilidad de una distribución de Poisson (lambda = 1000)

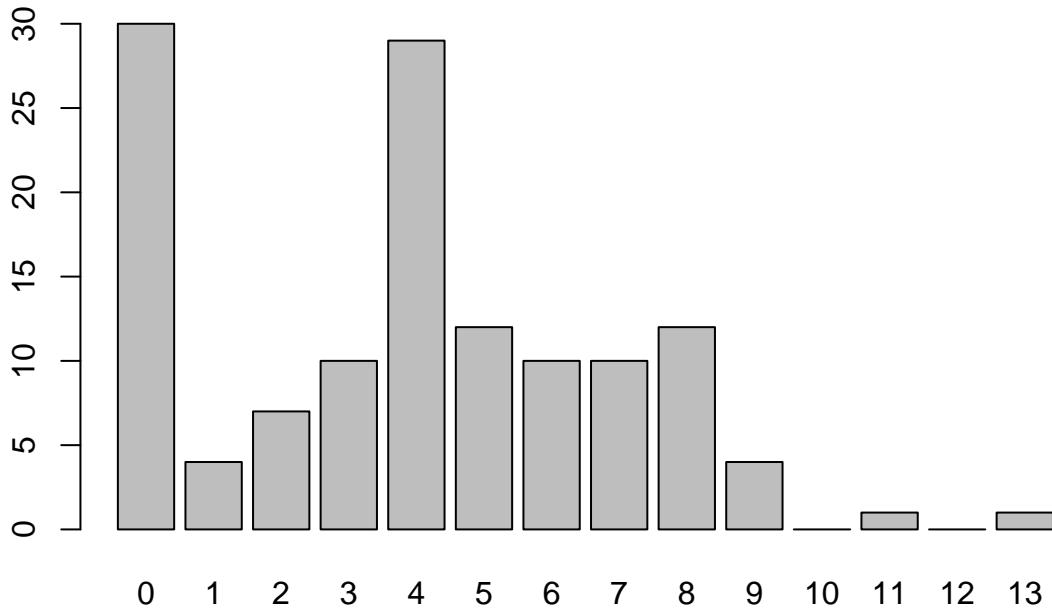


La gráfica anterior muestra la función de densidad de una distribución de Poisson con parámetro $\lambda = 1000$. Está centrada en 1000 y lo relevante es que la mayor parte de la masa está entre 900 y 1200. Sin embargo, es fácil pensar en problemas de conteos en los que la media sigue siendo 1000 pero pueden observarse con cierta frecuencia valores de 500 o 1500, que serían prácticamente imposibles con la formulación de Poisson.

Existe para estos casos un modelo alternativo: usar la distribución binomial negativa en lugar de la de Poisson. Más en general, la sobredispersión puede interpretarse en ocasiones en términos de una mezcla de distribuciones de Poisson: una variable no observada afecta a λ , i.e., dependiendo de ese valor desconocido, λ podría ser una cosa u otra.

Uno de los casos egregios de lo anterior es el de los **modelos de Poisson con inflación de ceros**: los datos parecen seguir una distribución de Poisson, con la salvedad de que parece haber más ceros de la cuenta. Si uno quiere estudiar, por ejemplo, el número de veces que los clientes de un banco usa la tarjeta de crédito en un periodo dado, es posible que los datos puedan modelarse usando un modelo de Poisson. Sin embargo, es fácil que haya muchos más ceros de los esperados porque puede haber clientes que no quieran usar jamás la tarjeta. No es lo mismo que un cliente que la usa, en promedio, dos veces al mes, por azar, no la use ninguna en un determinado mes (comportamiento que estaría *previsto* en la formulación del modelo) y que un cliente con tarjeta no la use nunca porque, p.e., usa la de otra entidad con mejores condiciones.

El aspecto típico de unos datos que tienen una distribución de Poisson con inflación de ceros es:



Existen extensiones específicas del modelo de Poisson para este caso concreto (XXX referencias) aunque es posible también ajustarlos usando modelos bayesianos (como los que veremos más adelante) que tengan en cuenta las características reales (i.e., la inflación de ceros) de la distribución subyacente.

12.3.2. Extensiones

- Lasso, ridge, etc.
- `glmnet`
- GBM...

12.4. Referencias

- Razón de *momios* (*odds ratio*) en la Wikipedia
- Nelder, J.A. y Wedderburn, R.W.M., *Generalized Linear Methods*
- Friedman, J.H, Hastie, T., Tibshirani, R., *Regularization Paths for Generalized Linear Models via Coordinate Descent*

12.5. Ejercicios

dasfsa

13. Introducción a la estadística bayesiana

En las secciones anteriores hemos estudiado expresiones de la forma $P(D | \theta)$, donde D son los datos y θ es un parámetro (o vector de parámetros) de interés. De hecho, tal era la definición del p-valor cuando fijábamos θ para representar la hipótesis nula. O la función de verosimilitud, que hemos usado para realizar estimaciones puntuales; luego, a partir de ese valor de referencia y remitiéndonos a resultados teóricos más o menos razonables, podíamos estimar intervalos de confianza para cuantificar el posible error cometido.

Sin embargo, no nos interesa $P(D | \theta)$. La verdadera expresión de interés es su *inversa*, i.e., $P(\theta | D)$. Esa distribución de probabilidad condicional resume lo que sobre nuestro parámetro de interés, θ , dicen los datos

observados D . Así, θ se convierte en una variable aleatoria propiamente dicha (que nunca dejó de serlo, aunque pretendiésemos lo contrario) cuyo valor es incierto; pero sobre la cual disponemos de la información adicional que nos proporcionan unos datos, D . Esta información nos ayudará a acotar la incertidumbre asociada a θ , determinar las regiones por las que es más probable que se encuentre, etc.

El *producto* de la aproximación bayesiana a la estadística, por tanto, será una distribución de probabilidad: la del parámetro θ , condicionada a (o tras) la observación de los datos D . Por eso se le da el nombre de **distribución a posteriori** (o posterior).

13.1. Teorema de Bayes

La construcción de la distribución *a posteriori* se realiza mediante la aplicación del teorema de Bayes, es decir,

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)},$$

expresión en la que aparecen nuevos términos. El menos interesante es $P(D)$, que es fijo (i.e., no depende de θ) y que se suele obviar. De hecho, es típico encontrar la expresión anterior representada de la forma

$$P(\theta | D) \propto P(D | \theta)P(\theta),$$

que indica que el término de la izquierda es proporcional al producto de la derecha. La constante de proporcionalidad es necesariamente tal que la integral del término de la izquierda sea uno, por lo que es fácilmente identificable.

El término $P(D | \theta)$ es un viejo conocido: la verosimilitud. Finalmente, $P(\theta)$ es la probabilidad del parámetro desconocido θ independientemente de los datos, *antes de la experiencia*, y se la conoce como **probabilidad a priori**.

13.1.1. Un primer ejemplo

Retomamos aquí el ejemplo de la moneda que se lanza 100 veces de las que 60 son caras. El parámetro desconocido es p , la probabilidad de cara. La verosimilitud es conocida,

$$\binom{100}{60} p^{60}(1-p)^{40}$$

y como distribución *a priori* podemos elegir (luego discutiremos la elección) la uniforme en $[0,1]$. La distribución *a posteriori* será

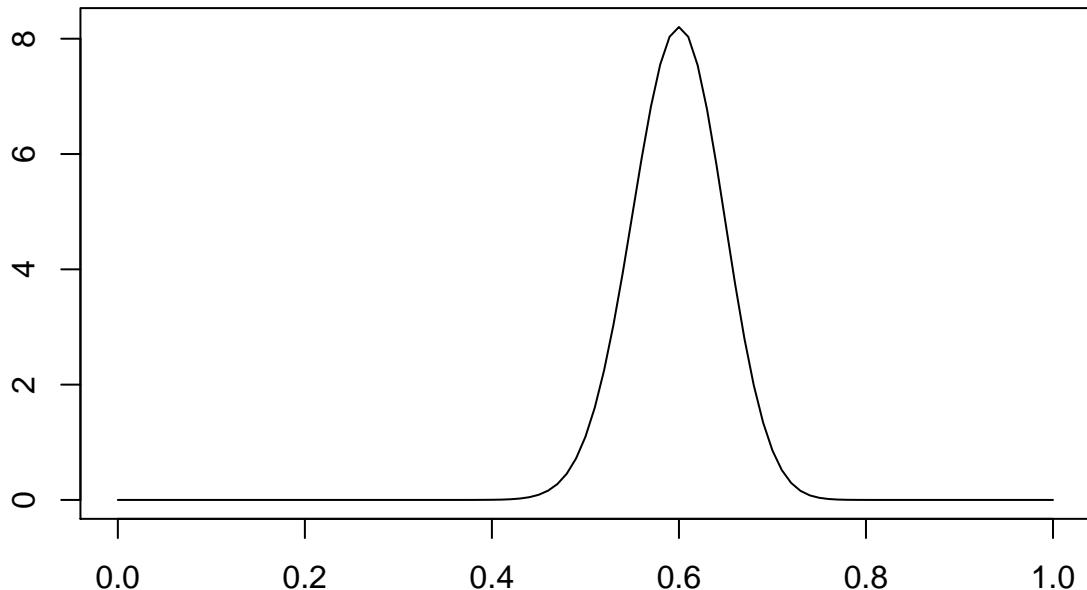
$$P(\theta | D) \propto P(D | \theta) \times P(\theta) = \binom{100}{60} p^{60}(1-p)^{40} \times 1,$$

por lo que, recogiendo los términos que no contienen a p en la constante de proporcionalidad,

$$P(\theta | D) \propto p^{60}(1-p)^{40} \quad (\text{y cero fuera de } [0, 1]).$$

Existe una distribución conocida, de libro, con soporte en $[0, 1]$ y cuya función de densidad es un producto como el anterior: la beta. Podemos reconocer, por tanto, $\theta | D \sim B(61, 41)$.

Distribución a posteriori



Gracias a esta distribución podemos ahora estudiar cuestiones como, por ejemplo, lo probable o no que es que $p = 0,5$ y otras.

En esencia, en eso consiste la aproximación bayesiana a la estadística:

1. Elegir una distribución *a priori* razonable
2. Calcular la distribución *a posteriori*
3. Explotar la distribución *a posteriori* para responder las preguntas relevantes acerca de los parámetros

Estas cuestiones ocuparán el resto de la sección.

13.2. Distribuciones *a priori*

Las distribuciones *a priori* han suscitado un debate innecesario y han sido consideradas por algunos como una puerta abierta a la subjetividad en un asunto, la inferencia estadística que, sostienen, debiera ser tan serio como, por supuesto, objetivo.

En realidad, después de que las ciencias hayan avanzado *una barbaridad*, escasean los experimentos *ex nihilo*. Nada deja de ser una *revisita* y antes de tirar al aire una moneda ya sabemos mucho sobre el resultado esperado. También lo sabemos sobre muchas otras cuestiones de relevancia:

- El peso relativo del hígado con respecto al de una persona
- La variación *normal* del Ibex 35 en una sesión bursátil
- El efecto de una bajada de un cuarto de punto de los tipos de interés en la inflación a medio plazo
- El uso de un determinado abono en el rendimiento por hectárea de un determinado cultivo
- Y un larguísimo etcétera.

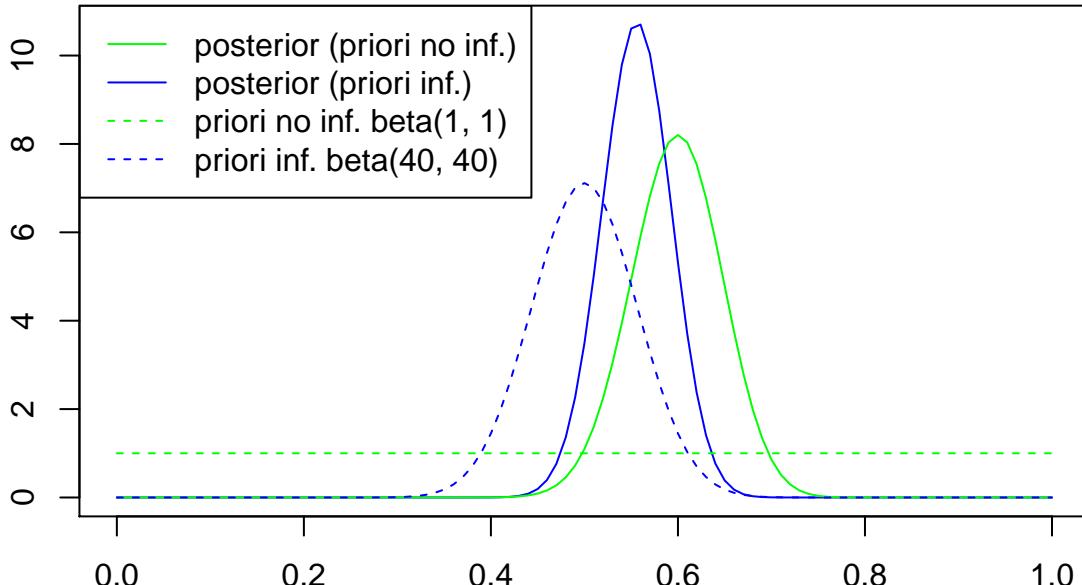
Más concretamente, el autor ha construido modelos en los que el cliente, experto en la materia, le ha dicho: el coeficiente α , para que tenga sentido, debería tener un valor entre 0,5 y 1,5.

La manera de introducir toda esa información en los modelos es mediante distribuciones *a priori* adecuadas.

No es la misma moneda una recién sacada del banco que la de un tahúr en una feria. El grado de evidencia que tenemos que exigir a una y otra para decidir que está trucada es distinto. Ese desigual nivel evidencia se

puede graduar mediante el p-valor en un enfoque no bayesiano o utilizando una distribución *a priori* distinta en el bayesiano.

Distribución a posteriori con priori informativa vs no informativa



El gráfico anterior compara las distribuciones *a posteriori* en dos experimentos similares: el lanzamiento de monedas considerado anteriormente usando dos *prioris* distintas. La primera, una distribución no informativa (una beta(1, 1), que es la distribución uniforme); la segunda, una distribución informativa (beta(40, 40)). Esta segunda *priori* trata de modelar nuestro conocimiento de que la probabilidad de cara es, aproximadamente, un 50 % con un determinado grado de certidumbre (dado por la anchura, o varianza, de la distribución). La *priori* informativa actúa como un atractor y su posterior correspondiente es una combinación de dicha *priori* y la verosimilitud (que coincide en nuestro caso con la posterior correspondiente a la *priori* no informativa).

Precisamente, una de las principales ventajas de la aproximación bayesiana a la estadística es que permite integrar la información *a priori* que se dispone acerca de los parámetros con la extraída de los datos.

A pesar de eso, en los textos habituales, gran parte de la discusión acerca de las distribuciones *a priori* gira alrededor de la construcción de *prioris* no informativas. En el ejemplo que venimos usando, es trivial construir una distribución que represente la absoluta falta de información sobre el parámetro en cuestión: solo puede ser la uniforme. En otros contextos es menos claro y existen procedimientos para identificarlas.

Sin embargo, el verdadero reto de la estadística bayesiana consiste en construir distribuciones *a priori* informativas, distribuciones que efectivamente recojan aquello que ya sabemos sobre los parámetros antes de ver dato alguno.

En el caso concreto expuesto más arriba en el que el experto juzga que un determinado parámetro debería estar comprendido entre los valores 0.5 y 1.5, una distribución *a priori* razonable para el parámetro sería una $N(1, 0.25)$ (en el 95 % de los casos, el parámetro estaría a distancia no mayor de $2\sigma = 0.5$ unidades de distancia de la media, 1).

13.3. Construcción de las distribuciones a posteriori

El segundo gran reto de la estadística bayesiana consiste en la construcción de la distribución *a posteriori*. Su dificultad ha sido, precisamente, el principal motivo por el que la estadística bayesiana haya tenido un desarrollo más lento que la no bayesiana. Esta dificultad, sin embargo, es más computacional que conceptual.

Aunque los problemas computacionales constituían tiempo atrás un escollo insalvable, hoy en día, gracias a los ordenadores, es posible plantear y resolver muchos de manera sencilla. Precisamente, a día de hoy, uno de los grandes ámbitos de investigación (y de aplicación) en estadística es el de los modelos bayesianos en el contexto del *big data*.

En algunos casos tan escasos como interesantes, es posible deducir la distribución a posteriori (así como sus parámetros). Existen parejas de familias de distribuciones A , B tales que *si la priori es A y la verosimilitud es B, entonces la posteriori es A*.

Un ejemplo de esa parejas de distribuciones son las que hemos utilizado en el ejemplo de las monedas:

- La *priori* era beta: sea uniforme ($\text{beta}(1, 1)$) o informativa ($\text{beta}(40, 40)$).
- La verosimilitud era binomial.
- Y la posterior era, de nuevo, beta.

Existen muchas más parejas de estas distribuciones, que reciben el nombre de **distribuciones conjugadas**. Muchas de ellas tienen aplicaciones en la práctica: son lo suficientemente flexibles para representar fidedignamente fenómenos aleatorios complejos. En ocasiones, no obstante, se recurre a ellas precisamente por la facilidad de cómputo que representa la posibilidad de contar con una descripción analítica de la distribución a posteriori.

No obstante, hoy en día es posible explorar la distribución *a posteriori* aun desconociendo su forma analítica. El epígrafe genérico de métodos MCMC (*Markov Chain Monte Carlo*) engloba técnicas como las del muestreo de Gibbs o el algoritmo de Metropolis-Hastings¹⁰ con las que se pueden obtener muestras de la posterior en un problema determinado.

Hoy en día tampoco hace falta conocer los detalles de estas técnicas. Existen **DSLs** (*domain specific languages* o, en español, *minilenguajes*) que permiten modelar fenómenos aleatorios condicionados a determinados parámetros de interés y muestrear sus distribuciones *a posteriori*. Algunos de los más conocidos son BUGS (en alguna de sus versiones, WinBUGS u OpenBUGS), JAGS o, más recientemente, Stan, que es el que usaremos en las siguientes secciones.

13.4. Introducción a Stan

Stan es un DSL (o minilenguaje) para la modelización probabilística especialmente indicado para el muestreo de distribuciones a posteriori. A diferencia de muchos otros minilenguajes, no es interpretado: el código desarrollado en Stan, por eficiencia, antes de ejecutarse, se traduce a C++ y se compila. Stan implementa una versión moderna y avanzada de los métodos MCMC, el llamado *Hamiltonian no u-turn sampler* que mejora algunas de las deficiencias de los más clásicos (p.e., el de Gibbs).

Existen extensiones de Stan para su integración con lenguajes como Python y, cómo no, R. En R, el paquete necesario para usar Stan es **rstan**.

Un programa en Stan se compone de una secuencia de bloques de los que, frecuentemente, solo se usan tres:

- El bloque de datos, donde se especifican los datos de entrada y se *instancian* las variables necesarias.
- El bloque de parámetros, donde se especifican los parámetros de interés (p.e., las probabilidades de interés en un modelo de Bernoulli o los coeficientes de una regresión lineal).
- El bloque de código propiamente dicho, donde se describe el modelo probabilístico.

De todos modos, la mejor manera de familiarizarse con Stan es a través de la resolución de algunos problemas ya conocidos.

¹⁰Entre las más clásicas: hoy en día se utilizan otras más eficientes.

13.5. Caso práctico: prueba de la diferencia de medias

El de la diferencia de medias es un problema que hemos planteado desde dos perspectivas en lo que precede: como una prueba de hipótesis clásica y como un modelo lineal.

El problema puede describirse así: hay unos datos que tienen una distribución $N(\mu_0, \sigma)$ y otros con una distribución $N(\mu_1, \sigma)$ e interesa determinar si $\mu_0 = \mu_1$. De otra manera, nos interesa la distribución de la variable aleatoria $\delta = \mu_1 - \mu_0$ para ver si concentra su masa cerca o lejos del valor 0.

Primero plantearemos el problema con unos datos simulados y lo resolveremos usando la prueba de Student tradicional:

```
set.seed(1234)
N1 <- 50
N2 <- 50
mu1 <- 1
mu2 <- -0.5
sig1 <- 1
sig2 <- 1

y1 <- rnorm(N1, mu1, sig1)
y2 <- rnorm(N2, mu2, sig2)

t.test(y1, y2)

##
##  Welch Two Sample t-test
##
## data: y1 and y2
## t = 4.7059, df = 95.633, p-value = 8.522e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5246427 1.2901923
## sample estimates:
## mean of x mean of y
## 0.5469470 -0.3604705
```

A continuación, lo replantearemos con Stan usando el paquete `rstan` de R. En primer lugar, cargaremos el paquete y, por conveniencia, crearemos una lista con los datos que habrá que pasar a Stan:

```
library(rstan)

## Loading required package: StanHeaders

## rstan (Version 2.18.2, GitRev: 2e1f913d3ca3)

## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend calling
## rstan_options(auto_write = TRUE)

##
## Attaching package: 'rstan'

## The following object is masked from 'package:arm':
## 
##     traceplot
```

```

standat <- list(N1 = length(y1),
                 N2 = length(y2),
                 y1 = y1, y2 = y2,
                 mu0 = mean(c(y1,y2)))

```

A continuación, definiremos el problema en Stan. El primer bloque, `data` define los datos necesarios y sus tipos y es, prácticamente, autoexplicativo.

```

data {
    int<lower=1> N1;
    int<lower=1> N2;
    vector[N1] y1;
    vector[N2] y2;
    real mu0;
}

```

En el bloque `parameters` se definen y asignan tipos a los parámetros involucrados en el problema:

- `mu`, la media del primer grupo.
- `diff`, la diferencia entre la media entre los dos grupos; es nuestro principal parámetro de interés.
- `sigma1` y `sigma2` son las varianzas desconocidas de nuestros datos que Stan estimará como subproducto. En la prueba clásica de Student, es necesario que las varianzas de ambos conjuntos de datos sean idénticas (aunque algunas extensiones posteriores relajan esa condición). Con Stan podemos ignorar la restricción dejando que sean distintas.

```

parameters {
    real mu;
    real diff;
    real<lower=0> sigma1;
    real<lower=0> sigma2;
}

```

Si las varianzas son distintas, la distribución subyacente usada en la prueba de Student ya no es una t de Student: a lo más, será una aproximación suya. Con Stan eso no supone un problema porque es capaz de muestrear la distribución *a posteriori* aunque sea desconocida y no figure en los manuales.

```

model {
    // priors
    mu ~ normal(mu0, 10);
    sigma1 ~ cauchy(0, 5);
    sigma2 ~ cauchy(0,5);

    // verosimilitud
    y1 ~ normal(mu, sigma1);
    y2 ~ normal(mu + diff, sigma2);
}

```

El bloque `model` define nuestro modelo. Por un lado, proporcionamos las distribuciones *a priori*. En este caso usamos distribuciones muy poco informativas, i.e, con una varianza grande. Hay que prestar atención a dos detalles:

- `sigma1` y `sigma2` son variables positivas pero su priori es de Cauchy (que admite valores negativos). Esto, sin embargo, no es problemático porque estas desviaciones estándar han sido definidas como positivas en el bloque de parámetros. Esto significa que, en la práctica, se estará usando solo la mitad positiva de la distribución de Cauchy. Sin embargo, si no se establece esa restricción, casi seguro, se producirá un error en la ejecución del código.

- No hay distribución *a priori* para `diff`. Eso equivale, esencialmente, a que cualquier valor real es equiprobable¹¹.

Y finalmente, está el bloque de la verosimilitud. En este caso simplemente le indicamos a Stan que `y1` es un vector con distribución normal con parámetros `mu` y `sigma1` y que `y2` es también normal con parámetros `mu + diff` y `sigma2`. Ahí reside, esencialmente, la sustancia; que no es otra cosa que la especificación de la prueba de Student traducida a código de Stan.

A continuación, juntamos el código en una cadena de texto en R:

```
stanmodelcode <- '
data {
  int<lower=1> N1;
  int<lower=1> N2;
  vector[N1] y1;
  vector[N2] y2;
  real mu0;
}

parameters {
  real mu;
  real diff;
  real<lower=0> sigma1;
  real<lower=0> sigma2;
}

model {
  // prioris
  mu ~ normal(mu0, 10);
  sigma1 ~ cauchy(0, 5);
  sigma2 ~ cauchy(0,5);

  // verosimilitud
  y1 ~ normal(mu, sigma1);
  y2 ~ normal(mu + diff, sigma2);
}
'
```

Finalmente es necesario lanzar el proceso con, por ejemplo:

```
fit <- stan(model_code = stanmodelcode,
            data = standat,
            iter = 12000, warmup = 2000,
            chains = 4, thin = 10)

##
## SAMPLING FOR MODEL '6112d1189e4cf39491023c40ff448a5a' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 5e-06 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 0.05 seconds.
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
```

¹¹No existe ninguna distribución de probabilidad que asigne la misma probabilidad a todos los reales: su integral sería infinita, no uno; sin embargo, es útil poder contar con ese tipo de *variables aleatorias*, que se dice que tienen una **distribución degenerada** y que pueden considerarse casos límite de variables aleatorias no degeneradas: por ejemplo, una sucesión de variables aleatorias normales de varianza creciente.

```

## Chain 1: Iteration: 1 / 12000 [ 0%] (Warmup)
## Chain 1: Iteration: 1200 / 12000 [ 10%] (Warmup)
## Chain 1: Iteration: 2001 / 12000 [ 16%] (Sampling)
## Chain 1: Iteration: 3200 / 12000 [ 26%] (Sampling)
## Chain 1: Iteration: 4400 / 12000 [ 36%] (Sampling)
## Chain 1: Iteration: 5600 / 12000 [ 46%] (Sampling)
## Chain 1: Iteration: 6800 / 12000 [ 56%] (Sampling)
## Chain 1: Iteration: 8000 / 12000 [ 66%] (Sampling)
## Chain 1: Iteration: 9200 / 12000 [ 76%] (Sampling)
## Chain 1: Iteration: 10400 / 12000 [ 86%] (Sampling)
## Chain 1: Iteration: 11600 / 12000 [ 96%] (Sampling)
## Chain 1: Iteration: 12000 / 12000 [100%] (Sampling)
## Chain 1:
## Chain 1: Elapsed Time: 0.027177 seconds (Warm-up)
## Chain 1: 0.134932 seconds (Sampling)
## Chain 1: 0.162109 seconds (Total)
## Chain 1:
## 
## SAMPLING FOR MODEL '6112d1189e4cf39491023c40ff448a5a' NOW (CHAIN 2).
## Chain 2:
## Chain 2: Gradient evaluation took 3e-06 seconds
## Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 0.03 seconds.
## Chain 2: Adjust your expectations accordingly!
## Chain 2:
## Chain 2:
## Chain 2: Iteration: 1 / 12000 [ 0%] (Warmup)
## Chain 2: Iteration: 1200 / 12000 [ 10%] (Warmup)
## Chain 2: Iteration: 2001 / 12000 [ 16%] (Sampling)
## Chain 2: Iteration: 3200 / 12000 [ 26%] (Sampling)
## Chain 2: Iteration: 4400 / 12000 [ 36%] (Sampling)
## Chain 2: Iteration: 5600 / 12000 [ 46%] (Sampling)
## Chain 2: Iteration: 6800 / 12000 [ 56%] (Sampling)
## Chain 2: Iteration: 8000 / 12000 [ 66%] (Sampling)
## Chain 2: Iteration: 9200 / 12000 [ 76%] (Sampling)
## Chain 2: Iteration: 10400 / 12000 [ 86%] (Sampling)
## Chain 2: Iteration: 11600 / 12000 [ 96%] (Sampling)
## Chain 2: Iteration: 12000 / 12000 [100%] (Sampling)
## Chain 2:
## Chain 2: Elapsed Time: 0.026592 seconds (Warm-up)
## Chain 2: 0.125158 seconds (Sampling)
## Chain 2: 0.15175 seconds (Total)
## Chain 2:
## 
## SAMPLING FOR MODEL '6112d1189e4cf39491023c40ff448a5a' NOW (CHAIN 3).
## Chain 3:
## Chain 3: Gradient evaluation took 3e-06 seconds
## Chain 3: 1000 transitions using 10 leapfrog steps per transition would take 0.03 seconds.
## Chain 3: Adjust your expectations accordingly!
## Chain 3:
## Chain 3:
## Chain 3: Iteration: 1 / 12000 [ 0%] (Warmup)
## Chain 3: Iteration: 1200 / 12000 [ 10%] (Warmup)
## Chain 3: Iteration: 2001 / 12000 [ 16%] (Sampling)
## Chain 3: Iteration: 3200 / 12000 [ 26%] (Sampling)

```

```

## Chain 3: Iteration: 4400 / 12000 [ 36%] (Sampling)
## Chain 3: Iteration: 5600 / 12000 [ 46%] (Sampling)
## Chain 3: Iteration: 6800 / 12000 [ 56%] (Sampling)
## Chain 3: Iteration: 8000 / 12000 [ 66%] (Sampling)
## Chain 3: Iteration: 9200 / 12000 [ 76%] (Sampling)
## Chain 3: Iteration: 10400 / 12000 [ 86%] (Sampling)
## Chain 3: Iteration: 11600 / 12000 [ 96%] (Sampling)
## Chain 3: Iteration: 12000 / 12000 [100%] (Sampling)
## Chain 3:
## Chain 3: Elapsed Time: 0.026449 seconds (Warm-up)
## Chain 3:                      0.126119 seconds (Sampling)
## Chain 3:                      0.152568 seconds (Total)
## Chain 3:
## 
## SAMPLING FOR MODEL '6112d1189e4cf39491023c40ff448a5a' NOW (CHAIN 4).
## Chain 4:
## Chain 4: Gradient evaluation took 3e-06 seconds
## Chain 4: 1000 transitions using 10 leapfrog steps per transition would take 0.03 seconds.
## Chain 4: Adjust your expectations accordingly!
## Chain 4:
## Chain 4:
## Chain 4: Iteration: 1 / 12000 [  0%] (Warmup)
## Chain 4: Iteration: 1200 / 12000 [ 10%] (Warmup)
## Chain 4: Iteration: 2001 / 12000 [ 16%] (Sampling)
## Chain 4: Iteration: 3200 / 12000 [ 26%] (Sampling)
## Chain 4: Iteration: 4400 / 12000 [ 36%] (Sampling)
## Chain 4: Iteration: 5600 / 12000 [ 46%] (Sampling)
## Chain 4: Iteration: 6800 / 12000 [ 56%] (Sampling)
## Chain 4: Iteration: 8000 / 12000 [ 66%] (Sampling)
## Chain 4: Iteration: 9200 / 12000 [ 76%] (Sampling)
## Chain 4: Iteration: 10400 / 12000 [ 86%] (Sampling)
## Chain 4: Iteration: 11600 / 12000 [ 96%] (Sampling)
## Chain 4: Iteration: 12000 / 12000 [100%] (Sampling)
## Chain 4:
## Chain 4: Elapsed Time: 0.028069 seconds (Warm-up)
## Chain 4:                      0.157736 seconds (Sampling)
## Chain 4:                      0.185805 seconds (Total)
## Chain 4:

```

El código anterior corre el programa con nuestros datos lanzando cuatro cadenas. Sería posible, además hacerlo en paralelo. Cada cadena realiza 12000 iteraciones, de las cuales, las 2000 primeras se descartan (*warmup*) y finalmente, se elige uno de cada diez valores de los valores restantes para reducir la posible correlación entre observaciones sucesivas.

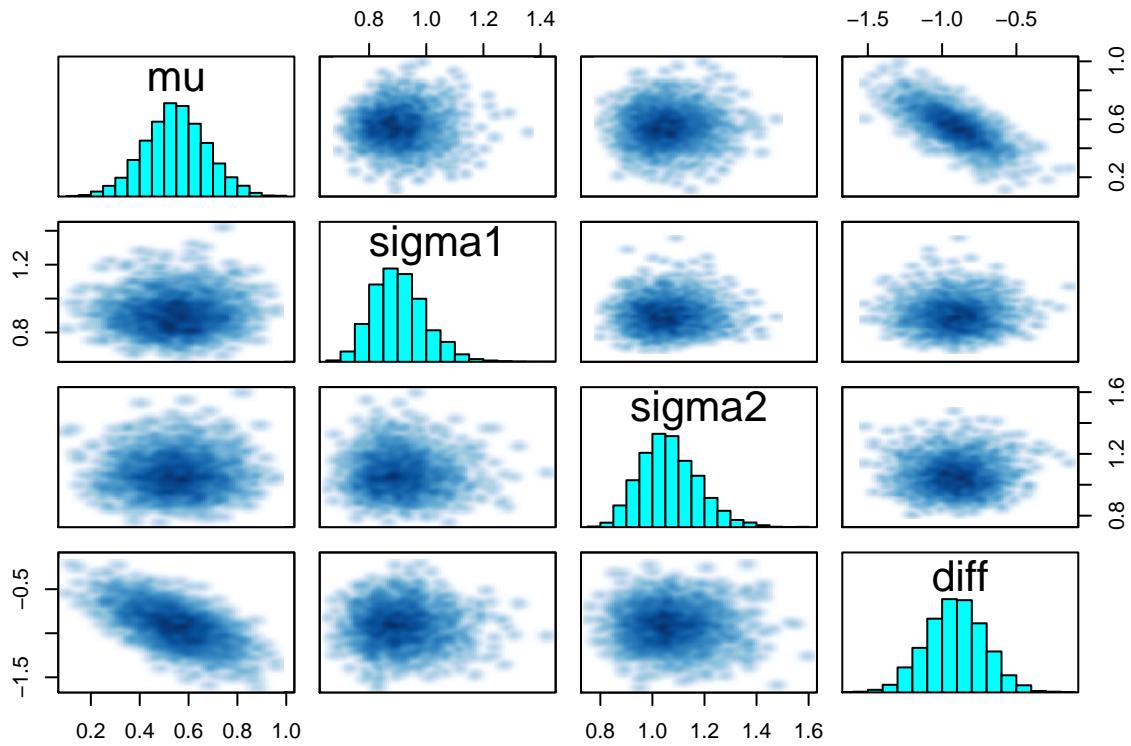
```
summary(fit)$summary
```

	mean	se_mean	sd	2.5%	25%
## mu	0.5475013	0.002049146	0.1291542	0.2892379	0.4639943
## diff	-0.9114074	0.003089509	0.1991153	-1.2920058	-1.0447888
## sigma1	0.9066360	0.001494258	0.0935997	0.7478142	0.8391078
## sigma2	1.0659810	0.001742387	0.1111160	0.8748632	0.9883432
## lp__	-46.9085639	0.022675226	1.4334641	-50.4569468	-47.5839515
	50%	75%	97.5%	n_eff	Rhat
## mu	0.5466361	0.6331784	0.8050759	3972.567	1.0002460
## diff	-0.9096843	-0.7829141	-0.5170305	4153.654	1.0002574

```

## sigma1    0.8987001    0.9625815    1.1071381 3923.718 0.9995454
## sigma2    1.0565903    1.1359470    1.3055842 4066.907 1.0015183
## lp__   -46.6159954 -45.8486394 -45.1213641 3996.413 0.9996956
pairs(fit, pars=c("mu", "sigma1", "sigma2", "diff"))

```



```

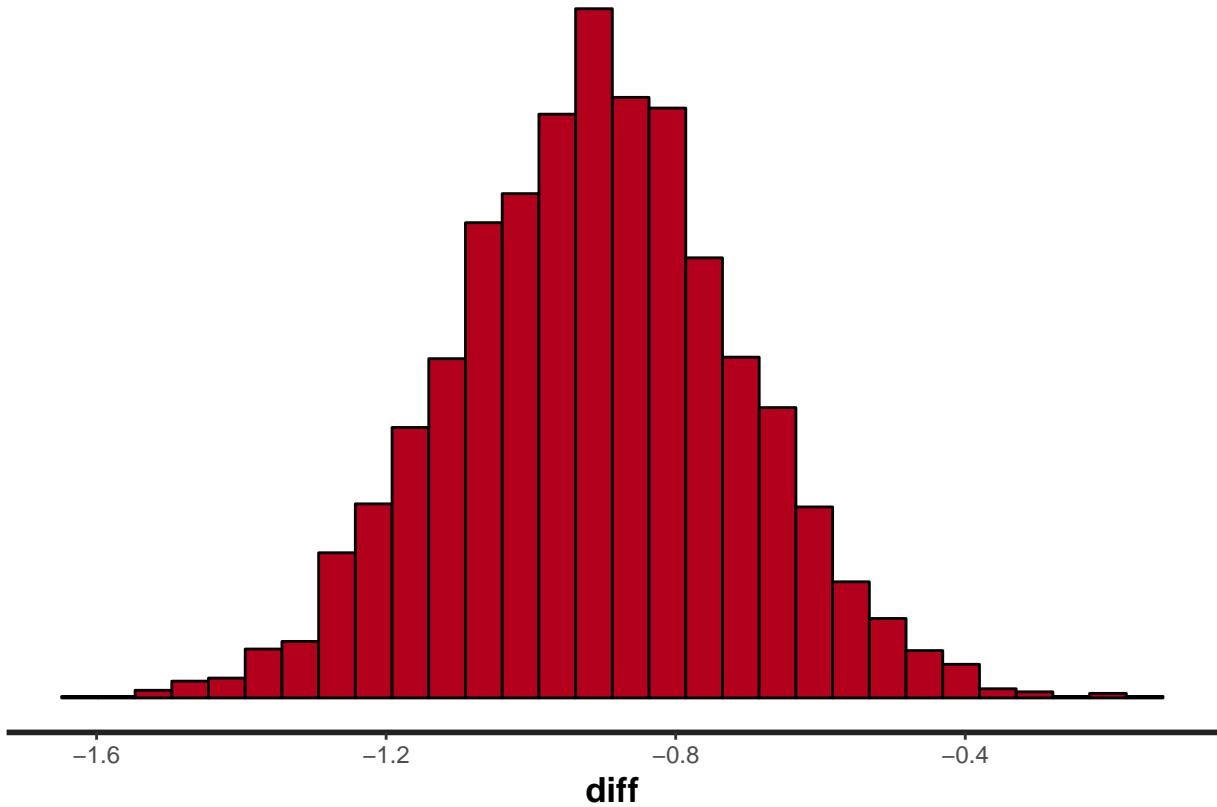
stan_hist(fit, pars = "diff")

```

```

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

```



13.6. Caso práctico: suavizado estadístico

El segundo caso práctico del uso de Stan va a ser mucho más ambicioso, aunque no mucho más complicado: el del suavizado estadístico. En particular, vamos a plantear una versión simplificada del **filtro de Kalman**.

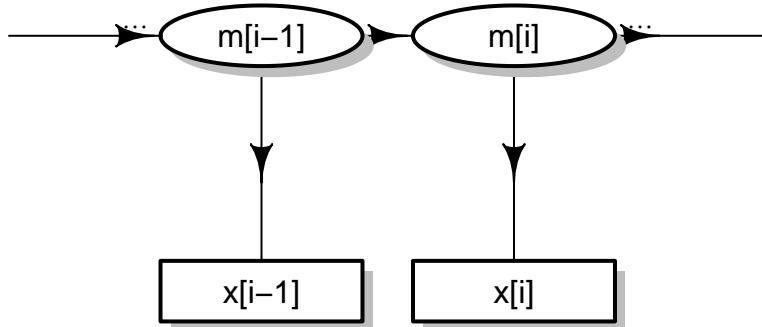
El contexto es el siguiente: observamos un fenómeno subyacente real que evoluciona de la siguiente manera:

$$m_{t+1} = m_t + \epsilon_t,$$

donde $\epsilon_t \sim N(0, \sigma_0)$. Es decir, en cada momento, el proceso va dando saltos normales a partir de su posición anterior. Sin embargo, no observamos directamente los valores m_t sino a través de un instrumento de medición *ruidoso*; es decir, observamos valores x_t que tienen distribución

$$x_t = m_t + \eta_t,$$

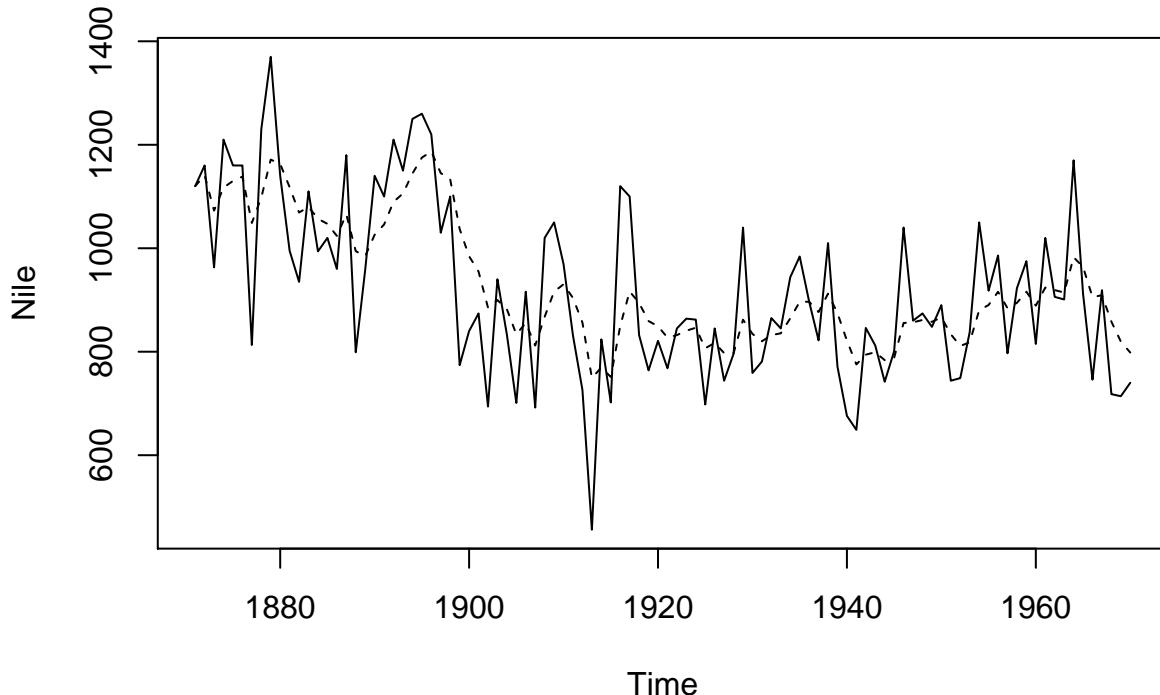
donde $\eta_t \sim N(0, \sigma_1)$.



Unos datos para cuyo análisis se ha usado ese tipo de modelos es el que en R se distribuye con en nombre de `Nile` y que recoge el flujo anual del río Nilo en Asuán desde 1871 a 1970 (véase `?Nile`). En R se pueden usar técnicas *ad hoc* (y avanzadas) para modelar estos datos como a continuación:

```
plot(Nile)

fit <- StructTS(Nile, type = "level")
lines(fitted(fit), lty = 2)
```



pero, alternativamente, se puede implementar el modelo directamente sobre Stan. Esencialmente, todo se reduce a:

```
nilo[1] ~ normal(m[1], sigma1);

for (i in 2:N){
  m[i] ~ normal(m[i-1], sigma0);
  nilo[i] ~ normal(m[i], sigma1);
}
```

En el código anterior, `m[i]` representan los valores reales subyacentes. Cada uno de ellos tiene una distribución normal con media igual a `m[i-1]`. Finalmente, los valores observados, `nilo[i]` tienen una distribución normal con media en `m[i]`.

Lo *inhabitual* del código anterior es que Stan, en una expresión del tipo

```
nilo[i] ~ normal(m[i], sigma1);
```

no calcula el término de la izquierda, `nilo[i]` en función de la expresión de la derecha, `normal(m[i], sigma1)`, sino lo contrario. Los valores `nilo[i]` son conocidos (son nuestras observaciones) y a partir de ellas se exploran posibles valores compatibles de los parámetros `m[i]` y `sigma1`.

El código completo para resolver el problema en Stan es:

```
codigo <- '
data {
  int N;
```

```

    vector[N] nilo;
}

parameters {
  vector[N] m;
  real<lower=0, upper = 1000> sigma0;
  real<lower=0, upper = 1000> sigma1;
}

model {
  for (i in 1:N)
    m[i] ~ normal(1000, 500);

  nilo[1] ~ normal(m[1], sigma1);

  for (i in 2:N){
    m[i] ~ normal(m[i-1], sigma0);
    nilo[i] ~ normal(m[i], sigma1);
  }
}
'

fit <- stan(model_code = codigo,
            data = list(N = length(Nile), nilo = as.vector(Nile)),
            iter = 12000, warmup = 2000,
            chains = 4, thin = 10)

## Warning: There were 4 chains where the estimated Bayesian Fraction of Missing Information was low. See
## http://mc-stan.org/misc/warnings.html#bfmi-low

## Warning: Examine the pairs() plot to diagnose sampling problems

```

La salida contiene 4000 secuencias de variables $m[i]$ (con i entre 1 y N), cada una de las cuales corresponde a una posible trayectoria del fenómeno subyacente.

```

library(reshape2)
tmp <- as.data.frame(fit)
tmp <- tmp[,1:100,]
colnames(tmp) <- 1:100
tmp$id <- 1:nrow(tmp)

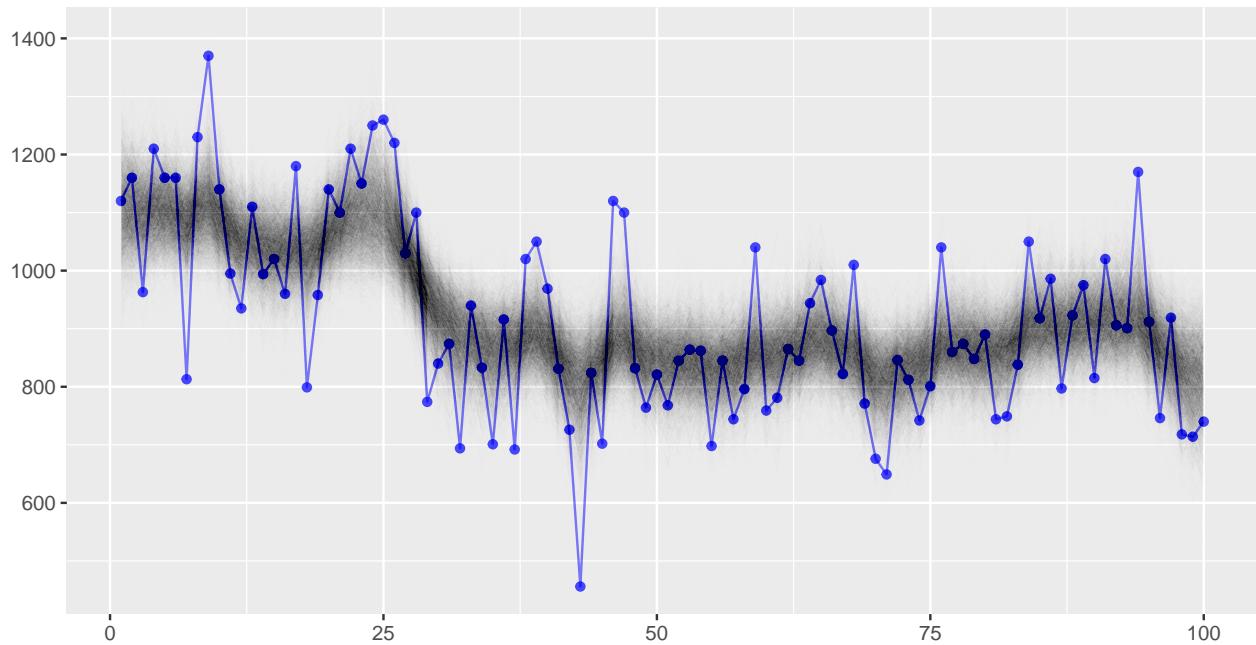
tmp <- melt(tmp, id.vars = "id")
tmp$variable <- as.numeric(tmp$variable)

nilo <- data.frame(x = 1:length(Nile), nilo = as.vector(Nile))

ggplot(nilo, aes(x = x, y = nilo)) + geom_line(col = "blue", alpha = 0.5) + geom_point(col = "blue", alpha = 0.5) +
  geom_line(data = tmp, aes(x = variable, y = value, group = id), alpha = 0.005) +
  xlab("") + ylab("") + ggtitle("4000 trayectorias simuladas")

```

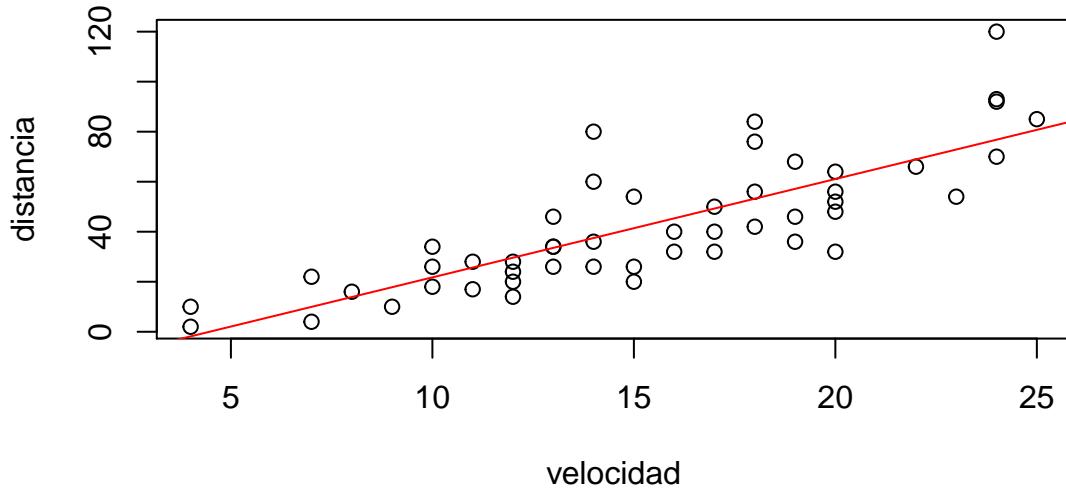
4000 trayectorias simuladas



13.7. Caso práctico: regresión lineal simple

La solución tradicional

```
modelo <- lm(dist ~ speed, data = cars)
plot(cars$speed, cars$dist, xlab = "velocidad", ylab = "distancia", main = "")
abline(modelo, col = "red")
```



La solución tradicional: coeficientes, etc.

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.5000  -0.7215  -0.1250   0.4900 12.5000
```

```

## -29.069 -9.525 -2.272  9.215 43.201
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed       3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12

```

rstan: código

```
stanmodelcode <- '
```

```
data {
  int N;
  vector[N] speed;
  vector[N] dist;
}
```

```
parameters {
  real a0;
  real a1;
  real<lower = 0, upper = 100> sigma;
}
```

```
model {
  // priors
  a0 ~ cauchy(0, 100);
  a1 ~ cauchy(0, 100);
  sigma ~ cauchy(0, 5);

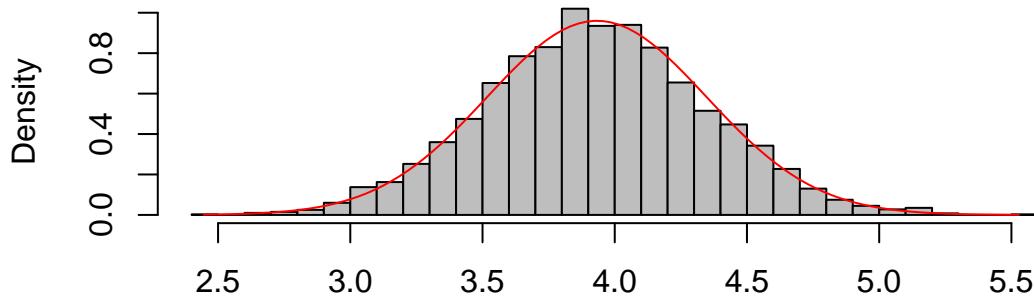
  // verosimilitud
  for (i in 1:N)
    dist[i] ~ normal(a0 + a1 * speed[i], sigma);
}
```

```
fit <- stan(model_code = stanmodelcode,
            data = list(N = nrow(cars), speed = cars$speed, dist = cars$dist),
            iter=12000, warmup=2000,
            chains=4, thin=10)
```

rstan: resultados

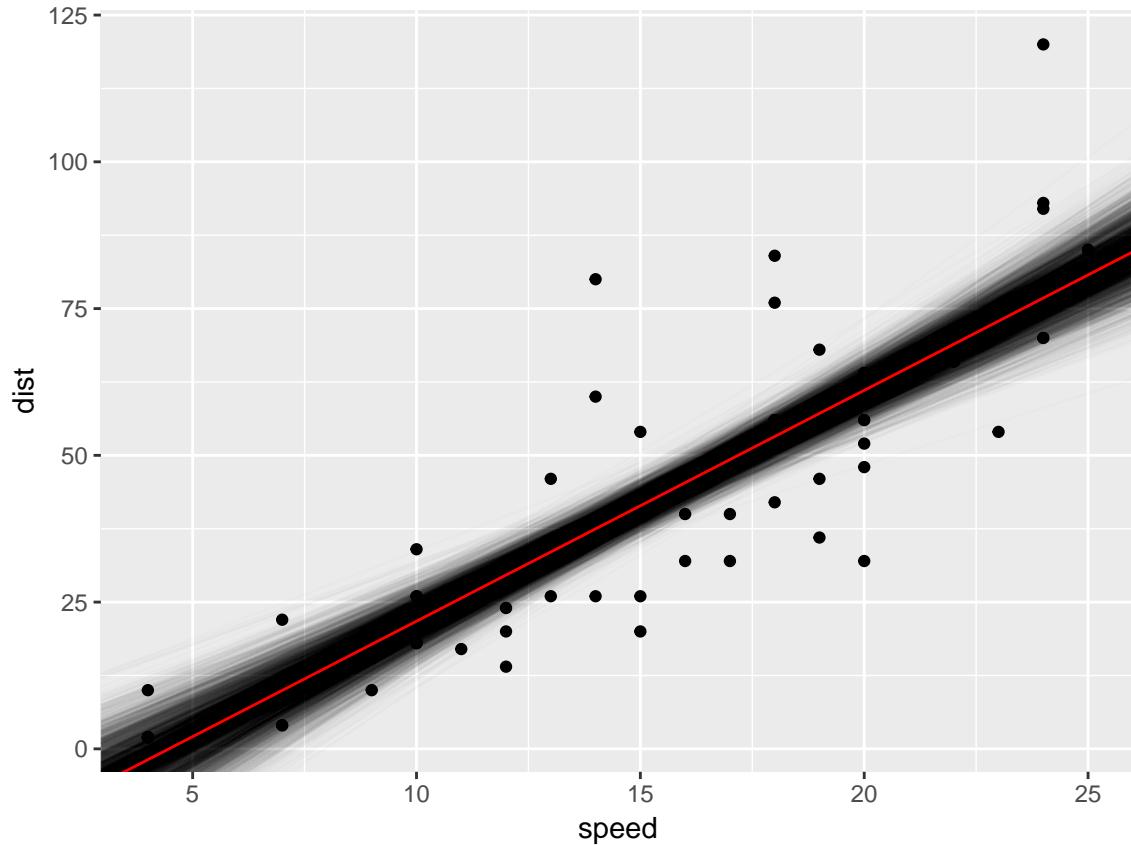
```
tmp <- as.data.frame(fit)$a1
hist(tmp, breaks = 30, col = "gray", xlab = "", freq = FALSE,
      main = "coef speed: posteriori (hist) vs lm (red)")
coefs <- summary(modelo)$coefficients[2,]
curve(dnorm(x, coefs[1], coefs[2]), from = min(tmp), to = max(tmp), col = "red", add = TRUE)
```

coef speed: posteriori (hist) vs lm (red)



rstan: resultados

```
library(ggplot2)
tmp <- as.data.frame(fit)[, c("a0", "a1")]
ggplot(cars, aes(x = speed, y = dist)) + geom_point() +
  geom_abline(data = tmp, aes(slope = a1, intercept = a0), alpha = 0.01) +
  geom_abline(intercept = coef(modelo)[1], slope = coef(modelo)[2], col = "red")
```



rstanarm: rstan para todos

```
library(rstanarm)

post <- stan_lm(dist ~ speed, data = cars,
                 prior = R2(location = 0.5, what = "median"),
                 chains = 4, cores = 4, seed = 1234)
```

- Incluye funciones para versiones *estanizadas* de `lm`, `glm`, `lmer`, etc.
- Son más robustos, según los autores, que los que uno pueda construir *a mano*
- Sobre todo, permiten especificar modelos con la interfaz habitual (con fórmulas)

13.8. Referencias

- *A priori* y *a posteriori*
- Distribuciones conjungadas
- Stan y rstan y rstanarm

13.9. Ejercicios

dafsfs

13.10. TODO

- Añadir más tipos de pruebas de hipótesis (¡comparar dos poblaciones!)
- Pruebas de hipótesis mediante permutaciones
- Ejemplo de maximización de la verosimilitud con optim