Master en Big Data. Fundamentos matemáticos del análisis de datos.

Tema 4: Variables aleatorias.

Fernando San Segundo

Curso 2019-20. Última actualización: 2019-09-15

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI
DEPARTAMENTO DE MATEMÁTICA APLICADA

1. Variables aleatorias discretas.
2. La distribución binomial.
3. Variables aleatorias continuas.
4. Variables aleatorias normales.
5. Complementos de R: funciones, datos ausentes.

Sección 1

Variables aleatorias discretas.

Modelos teóricos frente a datos empíricos.

- Vamos a proponerte un pequeño experimento mental. Imagínate que lanzamos un dado (honesto, no cargado) un millón de veces y que calculamos las *frecuencias relativas* de cada uno de los valores. ¿Qué números crees que habrá en la segunda fila de esta tabla?

valor del dado	1	2	3	4	5	6
frecuencia relativa	?	?	?	?	?	?

  Esos valores que ves con claridad en tu cabeza son un *modelo* teórico del experimento aleatorio que consiste en lanzar un dado. Y esa es precisamente la idea que trata de captar una variable aleatoria discreta: *un modelo teórico de un experimento aleatorio cuyos resultados son un conjunto discreto de valores.*

- Para describir una variable aleatoria discreta X tenemos por tanto que dar su **tabla (o función) de densidad de probabilidad**: una tabla de valores posibles de X y sus correspondientes probabilidades:

valor de X	x_1	x_2	$\cdots$	x_k
Probabilidad de ese valor $P(X = x_i)$	p_1	p_2	$\cdots$	p_k

  con $p_1 + p_2 + \cdots + p_k = 1$. A veces usaremos *notación funcional* $f(x_i) = P(X = x_i)$.

Ejercicio: usa R para hacer ese experimento y compara los datos empíricos con el modelo.

Media y varianza de distribuciones discretas.

- Una variable aleatoria discreta es un modelo teórico de la distribución de valores en la población. La **media poblacional** o **esperanza** es la media aritmética de dichos valores y se representa con la letra griega μ o con el símbolo $E(X)$. De forma análoga se define la **varianza poblacional** que denotaremos σ^2.
- Nuestro objetivo es utilizar datos muestrales para estimar o inferir los parámetros de una población. Si tenemos una muestra de una variable discreta que toma k valores distintos $x_1, \ldots, x_k$ con frecuencias absolutas $f_1, \ldots, f_k$ respectivamente podemos calcular la *media muestral* haciendo:

$$\bar{x} = \frac{x_1 f_1 + \cdots + x_k f_k}{n} = x_1 fr_1 + \cdots + x_k fr_k$$

donde $fr_1, \ldots, fr_k$ son las *frecuencias relativas* de los valores. Recuerda que las frecuencias relativas son las versiones *empíricas* de las probabilidades *teóricas*. Por eso la media poblacional μ (teórica) se calcula así a partir de la tabla de probabilidades:

$$\boxed{\mu = x_1 p_1 + \cdots + x_k p_k}$$

Una razonamiento similar conduce a esta expresión para la *varianza poblacional*

$$\boxed{\sigma^2 = (x_1 - \mu)^2 p_1 + \cdots + (x_k - \mu)^2 p_k}$$

- **Ejercicio:** usa R para calcular μ y σ^2 para un dado.

Usando `sample` con variables aleatorias discretas.

- **Ejercicio:** Dada esta tabla de densidad de probabilidad de una variable aleatoria X:

valor de X	0	1	2	3
Probabilidad de ese valor $P(X = x_i)$	$\frac{64}{125}$	$\frac{48}{125}$	$\frac{12}{125}$	$\frac{1}{125}$

  usa R para calcular μ, σ^2 y también σ, la desviación típica poblacional.

- Hasta ahora hemos usado `sample` para fabricar muestras en las que todos los elementos del vector eran equiprobables. Pero también podemos simular muestras de una población como la que describe el modelo teórico X usando la opción `prob` así (fíjate en que no hace falta *normalizar las probabilidades*):

```
muestra = sample(0:3, size = 10, replace = TRUE, prob = c(64, 48, 12, 1))
```

- **Ejercicio:**
  (1) Simula una muestra de tamaño 1000 de esta variable. ¿Cuál crees que es la mejor manera de representar gráficamente esa muestra?

  (2) Combina `sample` con `replicate` para simular cien mil muestras de tamaño 10. Estudia la distribución de las medias muestrales como hemos hecho en ejemplos previos.

Operaciones con variables aleatorias.

- Imagina que la variable aleatoria X representa el gasto en seguro del hogar y la variable Y el gasto en seguro del automóvil. Si queremos calcular el gasto total en ambos seguros tenemos que pensar en la *variable suma* $X + Y$. De la misma forma, a veces queremos multiplicar una variable por un número y, en general, vamos a pensar en combinaciones de la forma $a\,X + b\,Y$ donde a y b son coeficientes numéricos.
- La media $E(X + Y)$ de la variable $a\,X + b\,Y$ se calcula a partir de las medias de X e Y usando la misma combinación

$$a\,\mu_X + b\,\mu_Y$$

- Para la varianza las cosas son más complicadas, porque involucran la noción de *independencia*, que discutiremos después. Informalmente, X e Y son independientes si la información sobre el valor de X no afecta a la tabla de probabilidades de los valores de Y. La **covarianza** de X e Y es:

$$\text{cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y))$$

y en general $\sigma^2(a\,X + b\,Y) = a^2\,\sigma_X^2 + b^2\,\sigma_Y^2 + 2\,a\,b\,\text{cov}(X, Y)$

- **Cuando X e Y son independientes** se tiene $\text{cov}(X, Y) = 0$ y por tanto:

$$\sigma^2(a\,X + b\,Y) = a^2\,\sigma_X^2 + b^2\,\sigma_Y^2$$

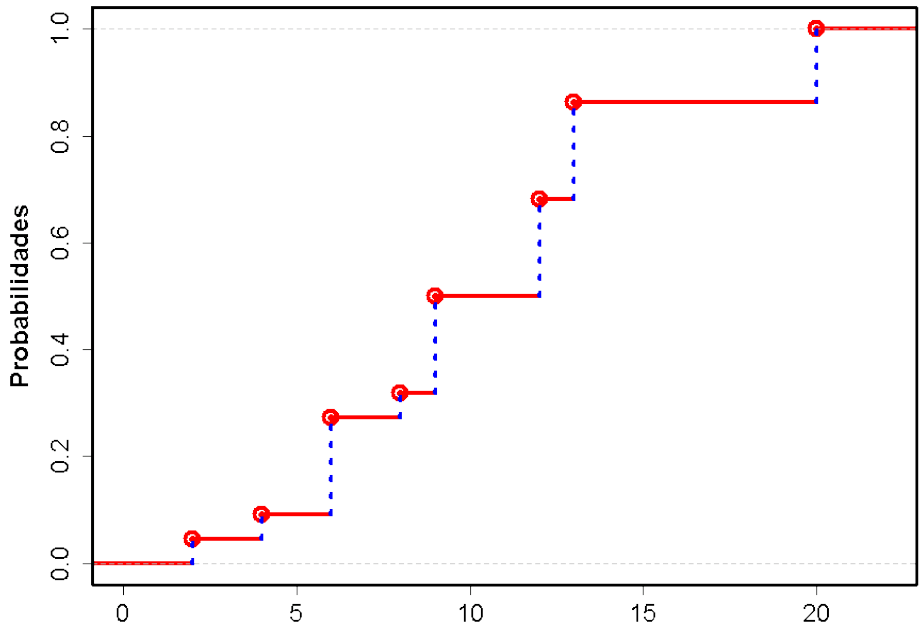
Función de distribución de una variable aleatoria discreta.

- La función de distribución F_X de una variable aleatoria X (discreta o continua) se define así para cualquier número k:

$$F_X(k) = P(X \leq k)$$

  Para una variable aleatoria la función de distribución juega un papel similar al de una *frecuencia relativa acumulada*, respondiendo a la pregunta ¿qué probabilidad hay de obtener un valor menor o igual que k?

- La gráfica de la función de distribución de una variable aleatoria discreta típica tiene este aspecto:



Sección 2

La distribución binomial.

Variables aleatorias de Bernouilli.

- Son probablemente las variables aleatorias discretas más sencillas de todas. Una variable aleatoria X es de tipo Bernouilli con parámetro p si su tabla de valores y probabilidades es:

Valor de X:	1	0
Probabilidad de ese valor:	p	$q = 1 - p$

- Por ejemplo, la variable $X =$"número de seises al lanzar un dado una vez" es una variable de tipo Bernouilli, con $p = \frac{1}{6}$, $p = \frac{5}{6}$. Para representar esto decimos que $X \sim Bernouilli(p)$ (el símbolo $\sim$ se lee *"es de tipo..."*).
- Los valores 1 y 0 se denominan, arbitrariamente, **éxito y fracaso** respectivamente.
- La media de una variable $X \sim Bernouilli(p)$ es $\mu = p$
(porque $1 \cdot p + 0 \cdot q = p$)
- Su varianza es $\sigma^2 = p \cdot q$
(porque $(1-p)^2 \cdot p + (0-p)^2 \cdot q = q^2 p + p^2 q = pq(p+q) = pq$).
- Las variables aleatorias de Bernouilli son útiles porque las usaremos como piezas para construir variables más complicadas, como la binomial que vamos a ver a continuación.

Variable aleatoria binomial.

- **Ejemplo:** Lanzamos un dado 11 veces. Es importante entender que *el experimento no es una tirada sino 11 tiradas* del dado. Definimos la variable X así:

$$X = \textit{número de veces que obtenemos un 6 en esas 11 tiradas}$$

- Esta situación tiene las siguientes características:

  (1) Un **experimento básico**, lanzar un dado se **repite** n **veces** (en el ejemplo $n = 11$).

  (2) Las repeticiones del experimento básico son **independientes** entre sí.

  (3) Cada repetición del experimento sólo puede terminar de una de estas dos maneras: en **éxito (success)** (en el ejemplo, sacar un 6) que se representa con el valor 1; o un **fracaso (failure)** (no sacar un 6) que se representa con el valor 0.

  (4) La **probabilidad de éxito** en cada repetición se denomina p y la de fracaso es $q = 1 - p$. En el ejemplo $p = 1/6, q = 5/6$.

  (5) . La variable X es la **suma del número de éxitos en las** n **repeticiones independientes**.

- **Definición de la variable aleatoria binomial.**
  Una variable aleatoria discreta X que reúne esas características es una variable aleatoria binomial de parámetros n y p, y escribiremos $X \sim B(n, p)$.

Ejemplo:

- Vamos a ver un ejemplo de variable binomial. Para ello usaremos la variable `prevalentHyp` de la tabla `fhs` que hemos usado en sesiones previas. Esa variable vale 1 si el paciente hipertenso y 0 en caso contrario. Para insistir en la arbitrariedad de la elección mantenemos esos valores y definimos como *éxito* el hecho de que el paciente sea hipertenso.
- **Ejercicio:** carga esa tabla de valores y comprueba que si elegimos un paciente al azar, la probabilidad de éxito (de que sea hipertenso) es $p \approx 0.3106$.
- Para definir una variable binomial vamos a elegir al azar $n = 7$ pacientes y nos preguntamos por el número X de hipertensos que hay entre esos siete.
- **Ejercicio:**

  (a) ¿Qué valores puede tomar X?

  (b) Escribe código en R para extraer una muestra de 7 pacientes (con remplazamiento) y contar cuántos de ellos son hipertensos (es decir, para calcular X en esa muestra).

  (c) Usa replicate para fabricar 50000 de esas 7-muestras. Llama `X` al vector de 50000 valores de la variable que se obtiene y haz una tabla de frecuencias relativas de valores de `X`.

Densidad de probabilidad en la binomial.

- Las frecuencias relativas de la última tabla son aproximaciones empíricas a las probabilidades teóricas de la binomial que vamos a calcular a continuación.
- Dada una variable $X = B(n, p)$ la probabilidad de obtener k éxitos es:

$$P(X = k) = \binom{n}{k} p^k \, q^{(n-k)}$$

donde el *número combinatorio* es:

$$\binom{n}{k} = \frac{\overbrace{n\,(n-1)\,(n-2)\cdots(n-k+1)}^{k \text{ factores}}}{k!}$$

y $k! = k \cdot (k-1) \cdot (k-2) \cdot \cdots \cdot 2 \cdot 1$ es el factorial de k.

- **Media y varianza de una variable binomial.** Una variable $X \sim B(n, p)$ es la suma de n variables de Bernouilli independientes (recuerda, que toman valores 0 o 1). Usando los resultados generales sobre variables aleatorias se obtiene:

$$\text{Si } X \sim B(n, p) \text{ entonces } \mu = n\,p, \quad \sigma^2 = n\,p\,q.$$

La binomial con R.

- Para calcular probabilidades concretas como $P(X = 3)$ en R usamos la función `dbinom` (suponiendo que ya has definido p):

```
dbinom(x = 3, size = 7, prob = p)
```

```
## [1] 0.2369079
```

Con `dbinom` podemos calcular a la vez *todas* las probabilidades de la variable binomial (mostramos tres cifras significativas):

```
signif(dbinom(x = 0:7, size = 7, prob = p), digits = 3)
```

```
## [1] 0.074000 0.233000 0.315000 0.237000 0.107000 0.028900
## [7] 0.004330 0.000279
```

Compara estos valores, que son predicciones teóricas, con las frecuencias relativas empíricas que hemos obtenido tomando muestras.

- La función de distribución $F(k) = P(X \leq k)$ de una binomial se calcula en R con:

```
signif(pbinom(q = 0:7, size = 7, prob = p), digits = 3)
```

```
## [1] 0.074 0.307 0.623 0.860 0.967 0.995 1.000 1.000
```

- Además R permite simular valores aleatorios de la variable binomial mediante:

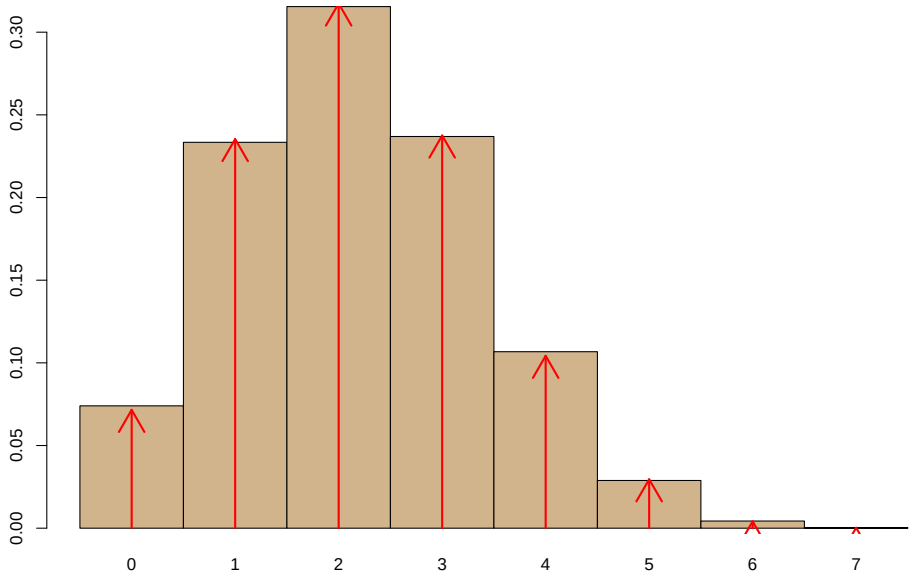
```
rbinom(n = 25, size = 7, prob = p)
```

```
##  [1] 2 5 2 1 2 3 1 1 1 4 3 3 1 1 3 3 3 3 4 0 3 3 2 3 2
```

Representación gráfica de la variable binomial.

- Para visualizar la tabla de densidad de probabilidad de una variable binomial con n moderado lo mejor es utilizar un diagrama de barras como este que muestra como se *distribuye* la probabilidad sobre los valores de 0 a n (hemos reducido a 0 es espacio entre barras por razones que pronto quedarán claras).

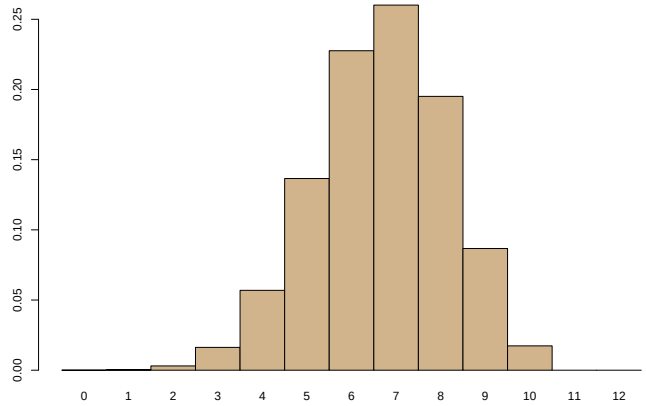
```
probabilidades = dbinom(x = 0:7, size = 7, prob = p)
bp = barplot(probabilidades, space = 0, col="tan", names.arg = 0:7)
```



Las flechas rojas representan las frecuencias relativas (¡empíricas!) de la muestra de miles de valores de X que hemos construido antes. Como puedes ver el acuerdo entre las predicciones de la teoría que representa el modelo de la variable binomial y los valores empíricos de la muestra es muy alto.

El zoo de las binomiales.

- Vamos a fijarnos en la forma de las distribuciones binomiales para distintos valores de n y p. Empezaremos por pensar en valores moderados de n (como 10) y de p (ni cerca de 0, ni cerca de 1). La siguiente figura muestra, a modo de ejemplo la distribución binomial $B(12, \frac{2}{3})$.



En general nos vamos a encontrar con **tres tipos de distribuciones binomiales**:

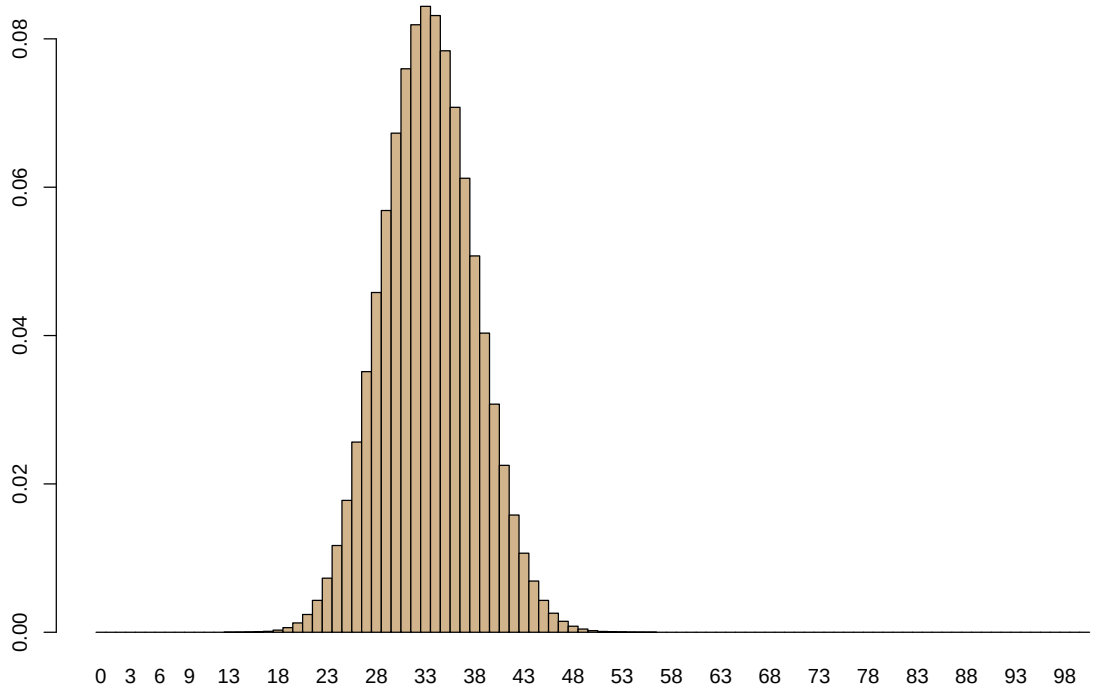
(1) Binomiales con n **pequeño**, como la de la anterior figura. En esos casos usamos la binomial directamente para calcular probabilidades.

(2) Binomiales con n **grande y** p **moderado (ni cerca de 0, ni cerca de 1)**. De estas hablaremos en el resto de este tema.

(3) Binomiales con n **grande y** p **no moderado (cerca de 0 o cerca de 1)**. Hablaremos de ellas más adelante al discutir la *Distribución de Poisson*.

Binomiales con n grande y p moderado.

- Vamos a ver ahora lo que sucede cuando n es grande y mantenemos p moderado (sin acercarlo al 0 o al 1). La siguiente figura muestra un diagrama de barras para la binomial $B(100, 1/3)$:



  Es un diagrama de barras. Pero es evidente que empieza a adivinarse una curva.

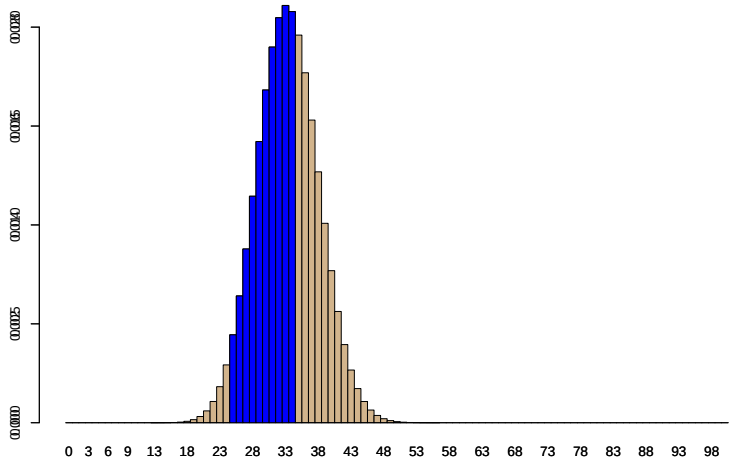
- Y no es una curva cualquiera. Abraham de Moivre descubrió que se trata de la misma curva normal que nos hemos encontrado ya al hablar de la distribución de las medias muestrales. ¿Por qué es útil esa curva?

Cálculos de probabilidad en binomiales con n muy grande.

- La binomial que aparece en la anterior figura es $X \sim B(n = 100, p = 1/3)$. Vamos a suponer que queremos calcular esta probabilidad:

$$P(25 \leq X \leq 35) = P(X = 25) + P(X = 26) + \cdots + P(X = 34) + P(X = 35)$$

Calcular la probabilidad de ese intervalo equivale a calcular el área sombreada.



Para calcular esa suma de términos hay que calcular, por ejemplo, el término:

$$P(X = 29) = \binom{100}{29} \left(\frac{1}{3}\right)^{29} \left(\frac{2}{3}\right)^{71}$$

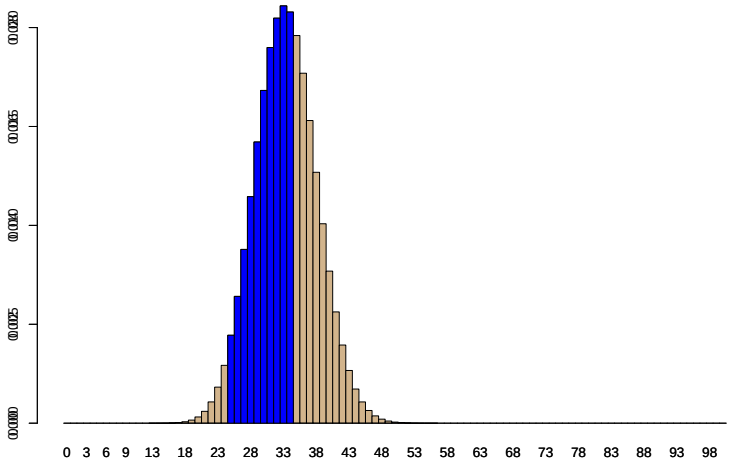
Pero

$$\binom{100}{29} = \frac{100!}{29!\ 71!} = \frac{100 \cdot 99 \cdot 98 \cdots 73 \cdot 72}{29 \cdot 28 \cdot 27 \cdots 2 \cdot 1} = 1917353200780443050763600$$

¡Y esto es solo uno de los términos! La curva normal ofrece una alternativa.

Otra vez la discusión "*discreto frente a continuo*".

- Si volvemos a pensar en la anterior figura del cálculo $P(25 \leq X \leq 35)$ en una $X \sim B(n = 100, p = 1/3))$

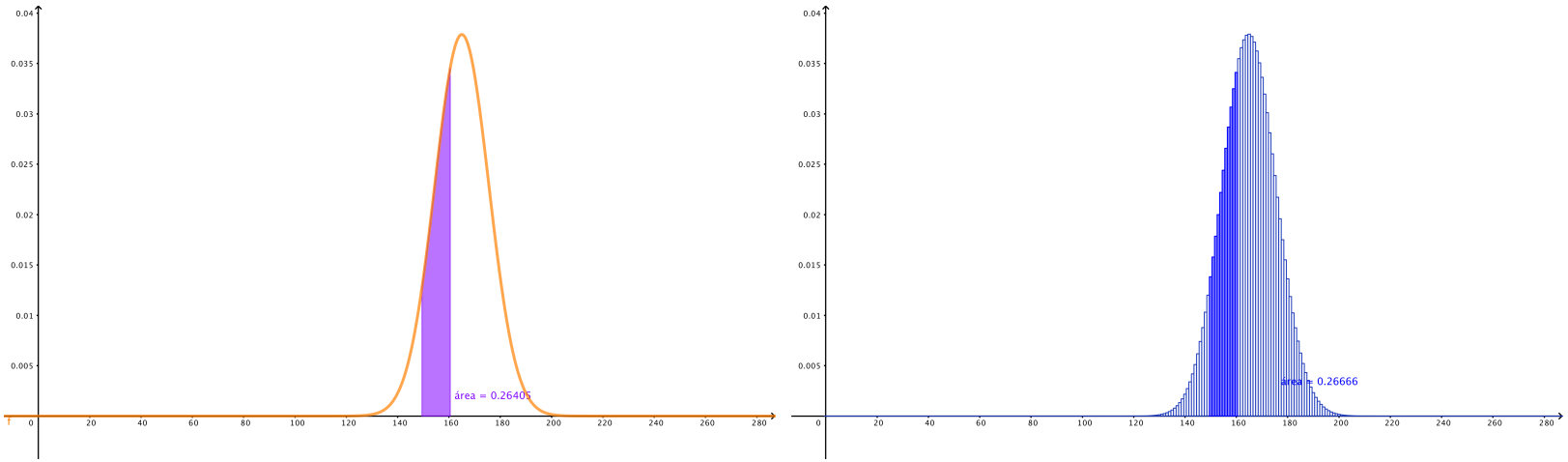


veremos que cada uno de los valores, cada una de las barras que forman ese gráfico, tiene un *peso individual* muy pequeño. Lo que importa es el *área conjunta*. Porque si la variable X puede tomar valores desde 0 hasta 100 entonces en la mayoría de las aplicaciones la diferencia entre $X = 65$ y $X = 66$ será *muy poco relevante*.

- Esta discusión recuerda a la que ya tuvimos al distinguir entre variables discretas y continuas. Cuando una variable toma muchos valores distintos y la diferencia entre valores individuales no es relevante, muchas veces es mejor considerarla continua. **No nos interesa la probabilidad de un valor concreto, sino la de un intervalo**.

Una solución alternativa.

- La curva normal describe muy aproximadamente el perfil de la distribución binomial. Así que para calcular la probabilidad de un intervalo, que es la suma de las áreas de los rectángulos sobre ese intervalo, podemos **aproximarla por el área bajo la curva normal en ese mismo intervalo**.



Las dos figuras muestran las dos formas de trabajar para calcular $P(a \leq X \leq b)$:

- a la izquierda calculamos (de forma exacta) $\sum_{k=a}^{b} P(X = k) = \sum_{k=a}^{b} \binom{n}{k} p^k \, q^{(n-k)}$.
- a la derecha, si la curva normal es $y = f(x)$, *aproximamos* esa probabilidad mediante

$$P(a \leq X \leq b) \approx \text{área bajo la gráfica de } f = \int_a^b f(x)\, dx.$$

- Si crees que integrar es *complicado*, ¡recuerda cómo son los números combinatorios!

Sección 3

Variables aleatorias continuas.

Función de densidad de probabilidad continua.

- Vamos a profundizar en esa idea de usar la integral de una función para calcular la probabilidad de un intervalo. No nos sirve cualquier función, pero basta con que se cumplan dos condiciones.
- Una función $f(x)$ es una **función de densidad continua** si posee estas características:
  - Es no negativa: $f(x) \geq 0$ para todo x; es decir, f no toma valores negativos.
  - El área total bajo la gráfica de f es 1:

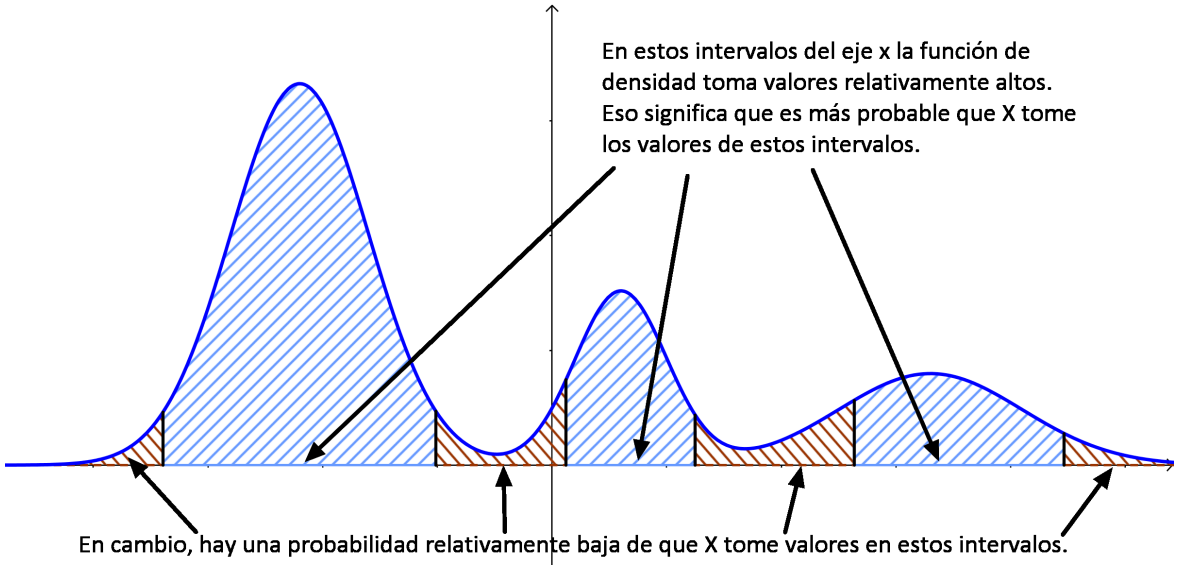
$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

- Si tenemos una función $f(x)$ con las propiedades que acabamos de ver, entonces diremos que f define una **variable aleatoria continua** X con función de densidad f.
- En tal caso la probabilidad de que el valor de X pertenezca a cualquier intervalo (a, b) se **define así**

$$P(a \leq X \leq b) = \text{área bajo la gráfica de } f = \int_a^b f(x)\, dx.$$

Interpretación de la función de densidad.

- La siguiente figura muestra una función de densidad y la forma de interpretar sus valores. Recuerda que los valores de la función no son probabilidades. Las probabilidades son *áreas*.



  Por eso decimos que una de estas funciones define una *distribución* (una forma de repartir) la probabilidad. Las variables discretas tienen una tabla de valores y probabilidades. Ahora tenemos la función f para hacer el mismo trabajo. Es la versión teórica de las curvas de densidad que aprendimos a dibujar para describir los datos de una muestra.

- Otra observación importante y que al principio resulta paradójica es que sea cual sea x_0 se cumple $P(X = x_0) = 0$.

Media y varianza de una variable aleatoria continua.

- Recuerda que para un variable aleatoria discreta con valores $x_1, \ldots, x_k$ y probabilidades $p_1, \ldots, p_k$ era:

$$\mu = E(X) = \sum_{i=1}^{k} x_i \cdot p_i$$

$$\sigma^2 = Var(X) = \sum_{i=1}^{k} (x_i - \mu)^2 \cdot p_i$$

- Para una variable aleatoria continua con densidad $f(x)$ se tiene:

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

$$\sigma^2 = Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)\, dx$$

- El paso de discreto a continuo se consigue cambiando el sumatorio por una integral y la probabilidad p_i por el *diferencial de probabilidad* $dp = f(x)\, dx$ (ver la Sección 5.4.2 de (San Segundo and Marvá 2016)).

La distribución uniforme.

- La *distribución uniforme* en el intervalo $[a, b]$ es un ejemplo sencillo pero muy importante de variable aleatoria continua. Se usa cuando ninguna parte del intervalo es más probable que otra del mismo tamaño.
- Su función de densidad es constante en el intervalo $[a, b]$ y vale 0 fuera:

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

  A veces se dice que en esta distribución *todos los puntos de* $[a, b]$ *son igual de probables*. Pero en cualquier distribución continua, uniforme o no, la probabilidad de un punto es 0

- La media de la variable uniforme es como cabía esperar $\mu = \dfrac{a+b}{2}$

  y su desviación típica es $\sigma^2 = \dfrac{(b-a)^2}{12}$
- En R usaremos la función `runif` para generar puntos aleatorios con distribución uniforme.
- **Ejercicio:** ejecuta varias veces `runif(10, min = 5, max = 15)` para ver como funciona.

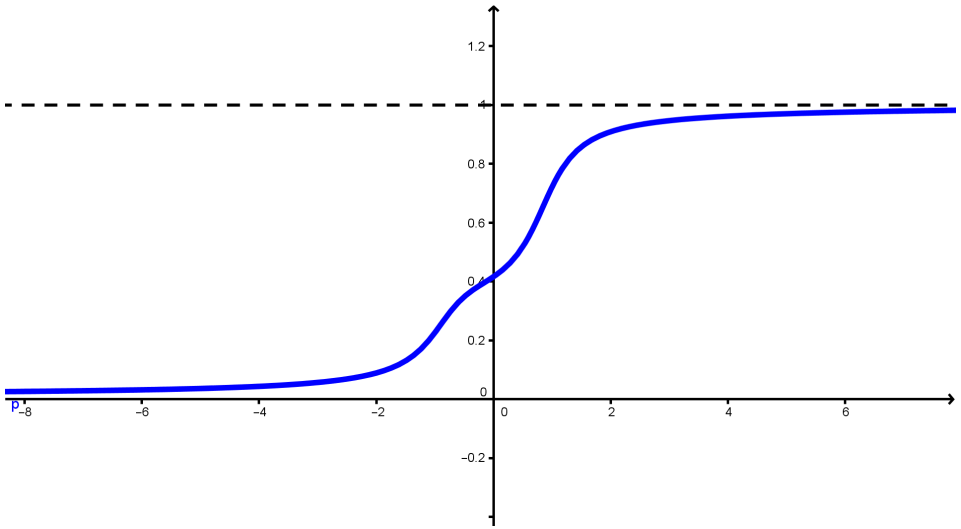
Función de distribución.

- La función de distribución de una variable aleatoria X (¡discreta o continua!) es:

$$F(k) = P(X \leq k)$$

Para una variable continua esto se traduce en $F(k) = \int_{-\infty}^{k} f(x)\,dx$

- Vimos que la gráfica típica de la función de distribución de una variable discreta tiene forma de escalera. Para una variable continua la gráfica típica de la función de distribución es una rampa como esta:



- Lo que hace que F sea a menudo más útil que f es esta **propiedad** :

$$P(a < X < b) = F(b) - F(a)$$

Sección 4

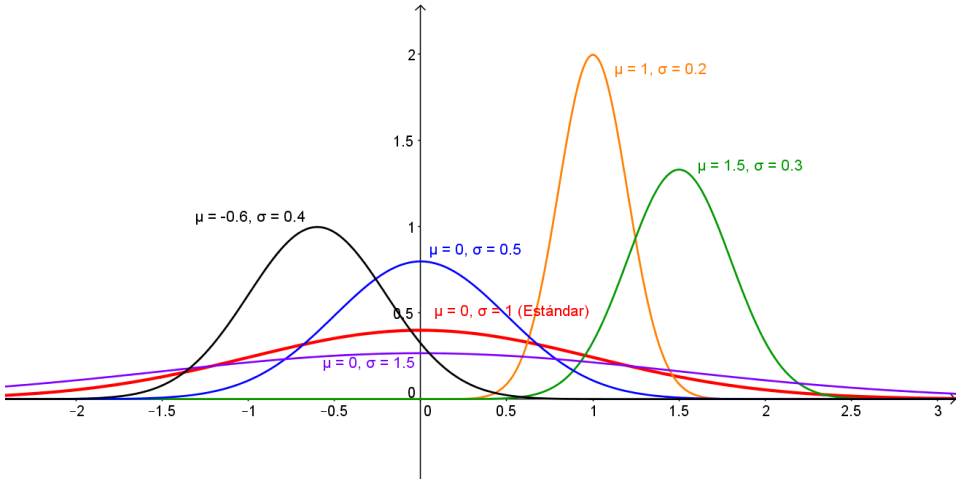
Variables aleatorias normales.

La curva normal.

- Hemos hablado ya varias veces de la curva normal. En realidad hay toda una familia de curvas normales, cuya ecuación es

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

  Aunque todas tienen forma acampanada, cada elección de valores de μ y σ produce una curva normal distinta: μ determina el centro de la distribución (que es simétrica) y σ controla cómo de estrecha y alta o ancha y baja es la campana. La figura muestra varias curvas normales para varios valores de μ y σ.



- Una variable aleatoria continua X con función de densidad $f_{\mu,\sigma}(x)$ es una **variable normal** y escribiremos $X \sim N(\mu, \sigma)$. La media de la normal $N(\mu, \sigma)$ es μ y su varianza es σ^2 (algunos libros usan $N(\mu, \sigma^2)$, cuidado).

Distribuciones normales en R. La función pnorm.

- La función pnorm permite calcular en R la función de distribución de una variable normal $X \sim N(\mu, \sigma)$.

$$P(X < b) = \texttt{pnorm(b, mean = mu, sd = sigma)}$$

- Si X es de tipo $N(10, 2)$ y queremos calcular la probabilidad de una **cola izquierda** $P(X < 10.5)$ usaríamos

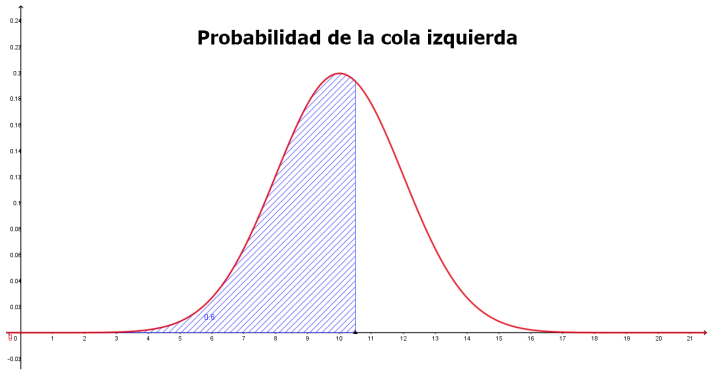
```
pnorm(10.5, mean=10, sd=2)
```

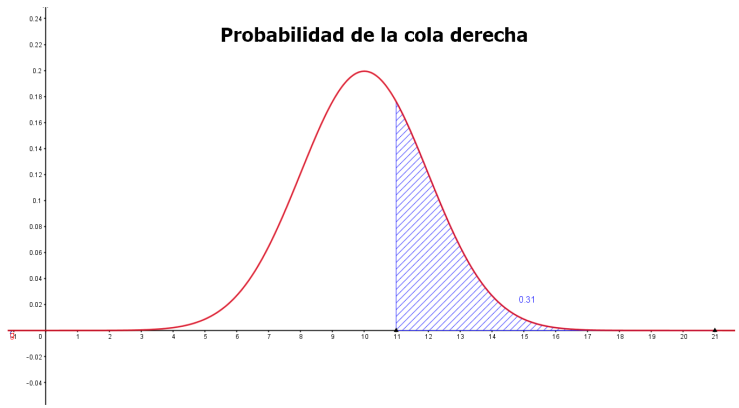
```
## [1] 0.5987063
```

- Si lo que queremos calcular es una **cola derecha** $P(X > 11)$ usaríamos una de estas dos opciones equivalentes:

```
1 - pnorm(11, mean=10, sd=2)
pnorm(11, mean = 10, sd = 2, lower.tail = FALSE)
```

```
## [1] 0.3085375
```



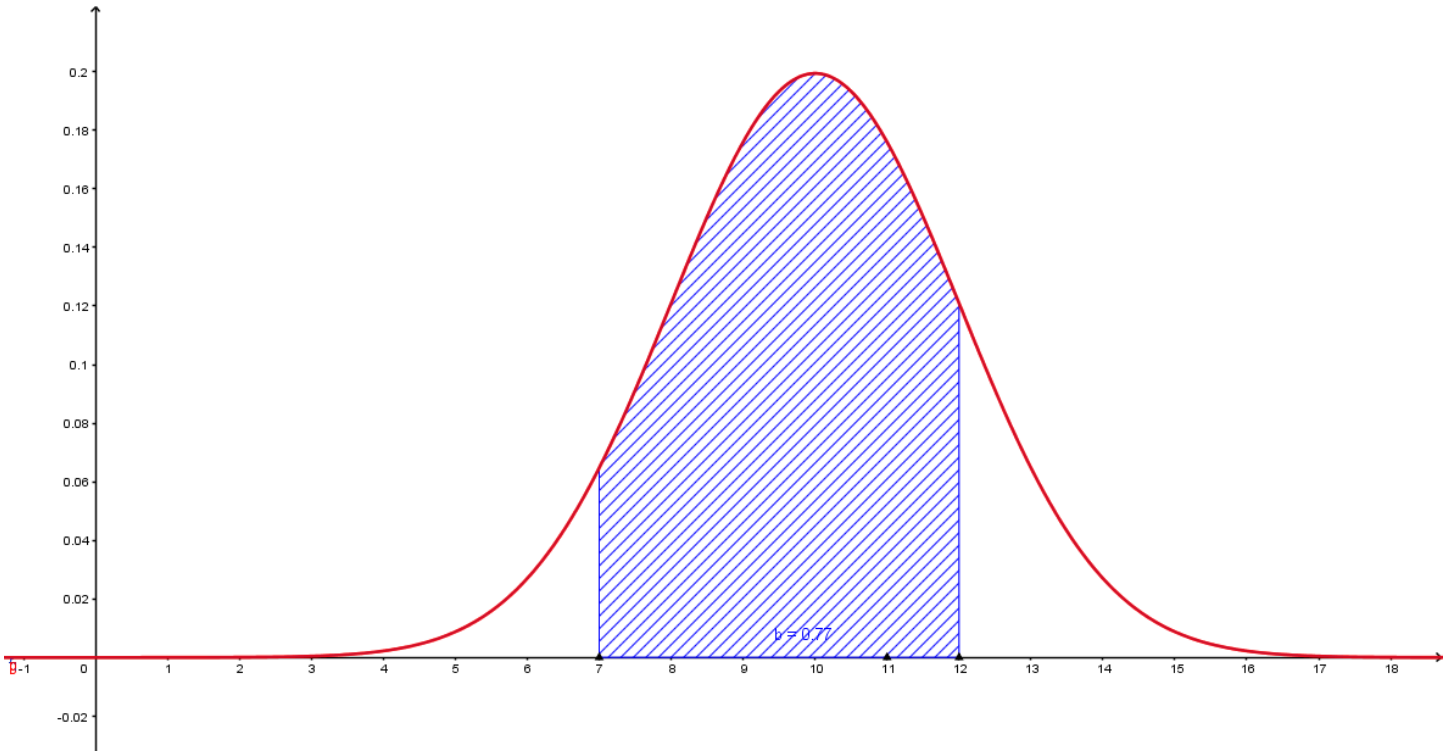


Probabilidad de un intervalo con `pnorm`.

- Si queremos calcular la probabilidad de un **intervalo**, como $P(7 < X < 12)$ lo expresamos como una diferencia:

```
pnorm(12, mean=10, sd=2) - pnorm(7, mean=10, sd=2)
```

```
## [1] 0.7745375
```

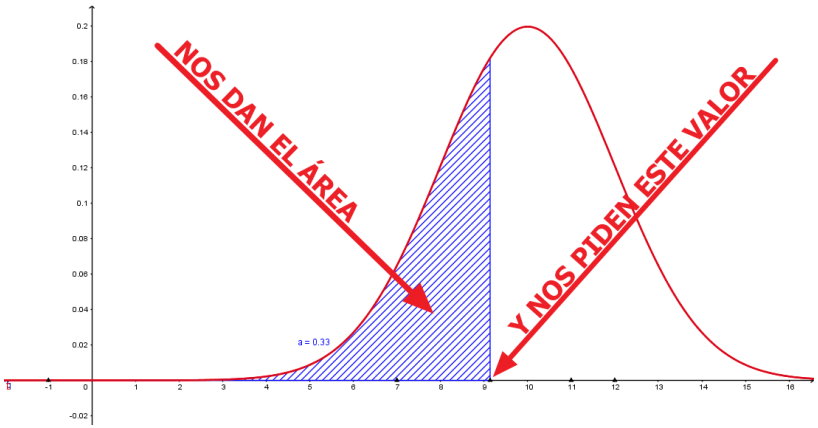


Problema inverso de probabilidad. La función `qnorm`.

- El *problema inverso* de probabilidad es este: dada una probabilidad p ¿cuál es el valor b tal que $P(X < b) = p$? Es un problema *muy importante para la Inferencia Estadística*.
- **Ejemplo** (problema inverso de cola izquierda) dada una distribución normal de tipo $N(10, 2)$ ¿cuál es el valor k para el que se cumple $P(X \leq k) = \frac{1}{3}$? En R lo obtenemos así:

```
qnorm(p = 1/3, mean = 10, sd=2)
```

```
## [1] 9.138545
```



- **Ejercicio importante:** dada una normal $N(0, 1)$, ¿cuál es el valor k para el que se cumple $P(X \geq k) = 0.025$? Volveremos a encontrar esta pregunta cuando hablemos de intervalos de confianza.

Otras funciones para trabajar con normales: `rnorm` y `dnorm`.

- La función `rnorm` es *muy útil para simulaciones*. Sirve para fabricar una muestra con n valores de una variable $X \sim N(\mu, \sigma)$ mediante:

```
muestra = rnorm(n, mean = mu, sd = sigma)
```

  ¡Atención! `mean(muestra)` no es `mu` y `sd(muestra)` no es `sigma`. ¿Ves por qué? Cuando queramos conseguir eso usaremos la función `mvrnorm` de la librería `MASS`.

- **Ejercicio:** genera vectores `x1` e `y1` cada uno con 1000 valores de una normal $N(0, 1)$. Luego ejecuta este código.

```
ggplot(data.frame(x1, y1)) +
  geom_point(mapping = aes(x1, y1), col="red")
```

  Ahora genera `x2` e `y2` cada uno con 1000 valores de una distribución uniforme en $N(0, 1)$ y ejecuta ese código cambiando `x1` e `y1` por `x2` e `y2`. ¿Ves la diferencia?

- La función `dnorm` es la función de densidad de la variable normal. Es decir, su valor es la altura de la curva normal y *no se debe interpretar directamente en términos de probabilidad*. Sirve casi exclusivamente para dibujar esa curva.

- Cuando conozcamos otras distribuciones verás que para todas ellas existen funciones similares a estas. Por ejemplo, para la distribución exponencial existen `pexp`, `qexp`, `rexp`, `dexp`. El sufijo `exp` identifica la distribución y el prefijo `p`, `q`, etc. identifica la función.

Tipificación y normal estándar Z.

- **Tipificación:** Si X es una variable aleatoria normal de tipo $N(\mu, \sigma)$, entonces la variable que se obtiene mediante la transformación de tipificación:

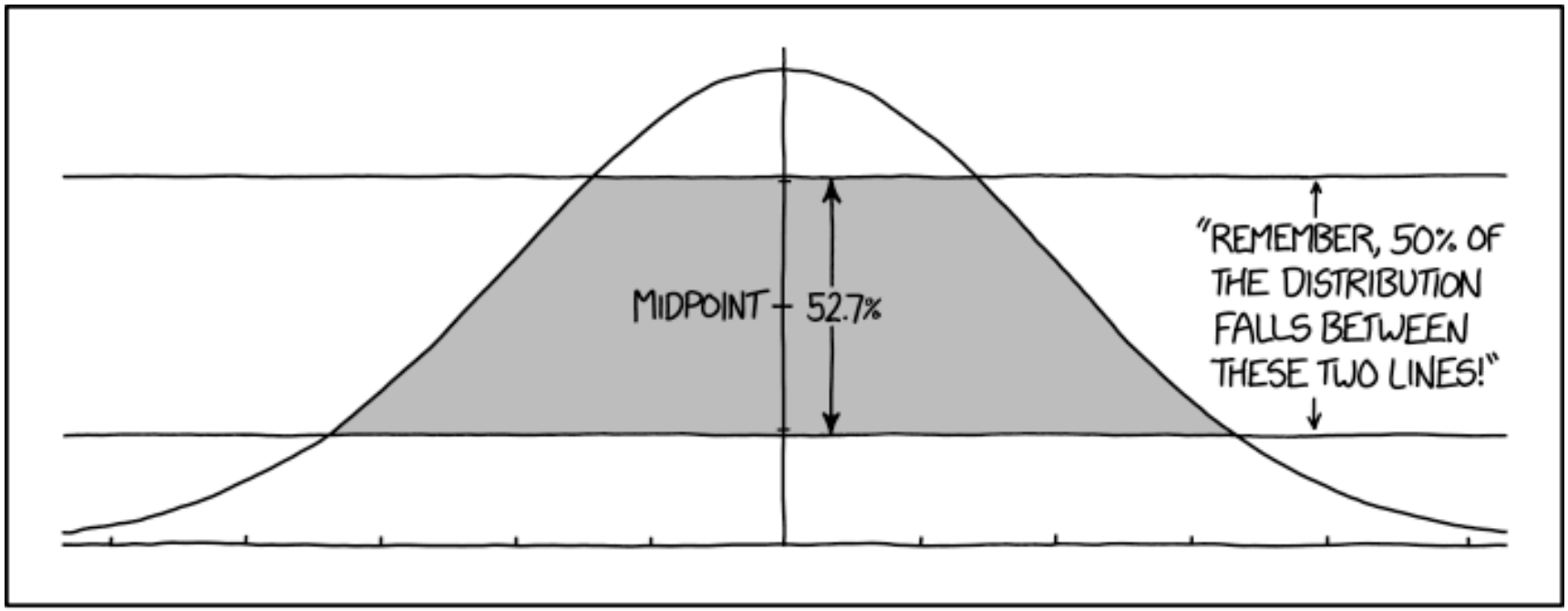
$$Z = \frac{X - \mu}{\sigma}$$

es una variable normal de tipo $N(0, 1)$, la **normal estándar** a la que siempre llamaremos Z. La tipificación permite reducir cualquier observación de una normal $N(\mu, \sigma)$ a una *escala universal* que nos proporciona la distribución Z.

- **Regla 68 - 95 - 99.** Una consecuencia de lo anterior es que si X es una variable normal de tipo $N(\mu, \sigma)$ entonces **siempre** se cumplen estas aproximaciones:

$$\begin{cases} P(\mu - \sigma < X < \mu + \sigma) \approx 0.683, \\ P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.955 \\ P(\mu - 3\sigma < X < \mu + 3\sigma) \approx 0.997 \end{cases}$$

- **Ejercicio:**
(a) comprueba estos resultados para, por ejemplo, la normal $N(0, 1)$ y la normal $N(40, 3.6)$.
(b) Tenemos una variable $X \sim N(123, 17)$ y observamos el valor 168. ¿Como de *raro* es este valor? Tipifícalo para responder.
(c) Ejecuta `scale(168, center = 123, scale = 17)`



HOW TO ANNOY A STATISTICIAN

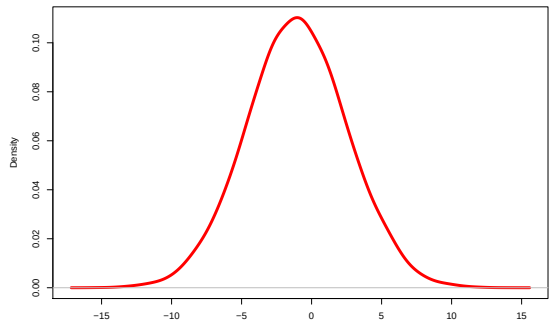
Suma (y mezcla) de normales independientes.

- Si $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$ son variables **normales independientes**, su suma es **de nuevo una variable normal** de tipo

$$N\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right).$$

Insistimos, la novedad es que *la suma de dos normales independientes sigue siendo normal*. Ejecuta este código para ver un ejemplo

```
set.seed(2019)
pob1 = rnorm(30000, mean = -3, sd = 1)
pob2 = rnorm(30000, mean = 2, sd = 0.5)
pobSuma = 3 * pob1 + 4 * pob2
plot(density(pobSuma, adjust = 1.6), main="", lwd=5, col="red", xlab="")
```



Este resultado se generaliza a la suma de k variables normales independientes, que dan como resultado una normal de tipo $N\left(\mu_1 + \cdots + \mu_k, \sqrt{\sigma_1^2 + \cdots + \sigma_k^2}\right)$.

La **mezcla** de variables normales es un proceso completamente distinto. A menudo da como resultado distribuciones bimodales o multimodales.

Sección 5

Complementos de R: funciones, datos ausentes.

Funciones de R.

- Aunque R básico y todas las librerías disponibles nos ofrecen miles de funciones para las más diversas tareas, pronto llegará el día en que necesitarás escribir una función para resolver un problema específico.
- Para escribir una función de R podemos usar este esquema básico

```
nombreFuncion = function(argumento1, argumento2, ...){
  ...
  ...

  líneas de código del cuerpo de la función
  ...
  ...
}
```

Como se ve la función tiene un *nombre*, una lista de *argumentos* y un *cuerpo* que contiene las líneas de código R que se ejecutarán al llamar a la función.

Ejemplo

- Crearemos una función `genPasswd` que genere contraseñas aleatorias. Los argumentos serán la longitud de la contraseña `size` y 3 booleanos `upp`, `low` y `nmb` que sirven para incluir o no respectivamente mayúsculas, minúsculas y números. Todos ellos menos `size` tienen valores por defecto.

```
genPasswd = function(size, upp = TRUE, low = TRUE, nmb = TRUE){

  # El vector pool guarda el juego de caracteres del password
  pool = character()

  # Generamos pool según las opciones
  if(upp) pool = c(pool, LETTERS)
  if(low) pool = c(pool, letters)
  if(nmb) pool = c(pool, 0:9)

  # Sorteamos los símbolos que aparecen en el password
  passwd = sample(pool, size, replace = TRUE)
  # Y lo reducimos a un string con paste
  paste(passwd, sep = "", collapse = "")
}
```

La función se ejecuta como cualquier otra función de R (*pero cuidado:* si tratas de ejecutarla sin darle un valor a `size` habrá un error.):

```
genPasswd(size = 15)
```

```
## [1] "oCbsgxj6zJiDHlV"
```

- **Ejercicio:** lee la ayuda de la función `paste` (y después la de `paste0`). Es una función extremadamente útil para trabajar con texto.

Acceso a las componentes de una función.

- La función `formals` permite acceder la lista de argumentos de cualquier función. Prueba a ejecutar:

```
formals(genPasswd)
```

El resultado es una *lista*. Este es un tipo de estructuras de datos muy importante que aún no hemos tratado, pero que veremos en breve.

- La función `body` permite acceder (¡y modificar!) el cuerpo de la función:

```
body(genPasswd)
```

```
## {
##     pool = character()
##     if (upp)
##         pool = c(pool, LETTERS)
##     if (low)
##         pool = c(pool, letters)
##     if (nmb)
##         pool = c(pool, 0:9)
##     passwd = sample(pool, size, replace = TRUE)
##     paste(passwd, sep = "", collapse = "")
## }
```

Observa el resultado si ahora haces

```
body(genPasswd) = "No me apetece trabajar..."
genPasswd(12)
```

Puedes leer más sobre funciones en el Capítulo 18 de (Boehmke 2016).

Manejo de datos ausentes. La función `is.na`

- Hasta ahora hemos tocado sólo tangencialmente el tema de los datos ausentes, pero es sin duda uno de los quebraderos de cabeza más habituales que te encontrarás al trabajar con un nuevo conjunto de datos.
- En R los datos ausentes se representan con el símbolo `NA`. Y disponemos de varias funciones para detectarlos. La más básica es `is.na`. Por ejemplo:

```
x = c(2, 3, -5, NA, 4, 6, NA)
is.na(x)
```

```
## [1] FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE
```

  Esta función es muy útil cuando se combina con otras como `which` que ya conoces o como `all` y `any`. Estas dos últimas actúan sobre un vector de booleanos y valen `TRUE` si todos o alguno, respectivamente, de los valores del vector son `TRUE`.

- Por ejemplo, podemos saber si `fhs$glucose` tiene algún valor ausente con

```
any(is.na(fhs$glucose))
```

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

Más sobre datos ausentes: `complete.cases` y `na.rm`

- Una función relacionada es `complete.cases`, Aplicada a una tabla (`data.frame`) nos dirá para cada fila si esa fila tiene o no datos ausentes.

```
head(complete.cases(fhs), 17)
```

```
##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [10]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
```

  El primer `FALSE` corresponde a la fila 15 de `fhs` que tiene un valor ausente en la columna `glucose`, como ya sabemos.

- La presencia de datos ausentes puede hacer que muchas funciones produzcan `NA` como resultado (o peor, que no funcionen correctamente). Por ejemplo, una media aritmética:

```
mean(fhs$glucose)
```

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

  Muchas funciones de R disponen de un argumento `na.rm` para excluir los valores `NA` de la operación que se realice:

```
mean(fhs$glucose, na.rm = TRUE)
```

```
## [1] 81.96366
```

  Puedes encontrar más información en la Sección 7.4 de R for Data Science, la Sección 5.12 de (Peng 2015) y el Capítulo 14 de (Boehmke 2016).

Referencias para la sesión

Enlaces

- Código de esta sesión

Bibliografía

Boehmke, B. C. (2016). *Data Wrangling with R* (p. 508). Springer. doi:10.1007/978-3-319-45599-0

Peng, R. D. (2015). *R Programming for Data Science* (p. 132). Leanpub. doi:10.1073/pnas.0703993104

San Segundo, F., & Marvá, M. (2016). *PostData 1.0.* (p. 616). Lulu.com. http://www.lulu.com/shop/fernando-san-segundo-and-marcos-marv{\'{a}}/postdata-10/paperback/product-22855863.html