

MEMORIA PRÁCTICA – 2:

Probabilidad y distribuciones de probabilidad.



Escuela Politécnica Superior - Universidad Autónoma de Madrid

GRADO EN INGENIERÍA BIOMÉDICA

BIOESTADÍSTICA

Versión del documento número 1

Práctica realizada por: Laura Sánchez Garzón

Fecha: 10/04/2023

Profesorado de la práctica: Mercedes Sotos Prieto y María Téllez Plaza

1. ÍNDICE

Contenido

1. ÍNDICE.....	2
2. INTRODUCCIÓN	3
3. EJERCICIOS GUIADOS	4
3.1. Probabilidad simple y condicionada.....	4
3.2. Distribuciones de probabilidad para variables numéricas discretas y continuas.	6
3.2.1. Código 1.....	6
3.2.2. Código 2.....	9
3.3. Intervalos de confianza y contraste de hipótesis.....	14
4. EJERCICIOS PROPUESTOS	17
4.1. Ejercicio 1	17
4.2. Ejercicio 2	19
4.3. Ejercicio 3	21
4.4. Ejercicio 4	23
5. CONCLUSIONES	25

2. INTRODUCCIÓN

Esta segunda práctica de bioestadística tiene por objetivo trabajar con los conceptos teóricos vistos en clase, para asimilarlos y entender cómo funcionan las probabilidades, distribuciones e intervalos de confianza.

RStudio dispone de todo tipo de funciones y gráficos que nos permiten trabajar con dichos conceptos, y en pocas líneas de código se puede extraer de manera inequívoca, sólidas conclusiones. Además, utilizar R resulta mucho más práctico que realizar los ejercicios a mano, dado que los cálculos son rápidos, y no se debe recurrir a las tablas de distribución (la propia máquina calcula resultados).

3. EJERCICIOS GUIADOS

3.1. Probabilidad simple y condicionada

1. Gran parte de los primeros trabajos en la teoría de la probabilidad trataban de juegos y apuestas. La noción básica es el de una muestra aleatoria: repartiendo un mazo de cartas bien barajado o sacando bolas numeradas de una urna bien agitada.

En R, se puede simular estas situaciones con la función de `sample()`. Si se quiere elegir 5 números al azar del conjunto 1:40, entonces podemos escribir:

#con la función `sample()` se pueden elegir números al azar. En este caso, en un rango del 1 al 40, se eligen 5 números.

```
> sample(x=1:40, size=5)
[1] 28 21 14 15 30
```

Figura 1: Ejemplo de cómo obtener números al azar utilizando la función `sample()`.

El primer argumento (*x*) es un vector de valores a muestrear y el segundo (*size*) es el tamaño de la muestra.

Tener en cuenta que el comportamiento predeterminado de `sample()` es el muestreo sin reemplazo, es decir, las muestras no contendrán el mismo número dos veces, y obviamente, el tamaño no puede ser mayor que la longitud del vector que se va a muestrear.

```
> sample(40, 5)
[1] 33 3 25 8 17
```

Figura 2: Ejemplo de cómo obtener números al azar utilizando la función `sample()`.

Observar que `sample(40,5)` sería suficiente ya que un solo número se interpreta para representar la longitud de una secuencia de números enteros.

2. Si se desea un muestreo con reemplazo, debe agregar el argumento `replace = TRUE`. El muestreo con reemplazo es adecuado para modelar lanzamientos de monedas o de dados. #muestreo con reemplazo. Se asemeja a tirar una moneda. Tenemos como opciones T (tail) y H (head), y se tira 10 veces.

```
> sample(c("H", "T"), 10, replace=T)
[1] "H" "T" "T" "H" "T" "T" "H" "H" "T" "T"
```

Figura 3: Simulación de cómo obtener 10 lanzamientos de monedas con reemplazamiento

3. Volvamos al caso del muestreo sin reposición, concretamente `sample(1:40, 5)`. La probabilidad de obtener un número cualquiera como el primero de la muestra debe ser $1/40$, el siguiente $1/39$ y así sucesivamente. La probabilidad de una muestra cualquiera de 5 elementos sin reemplazamiento debe ser entonces $1/(40 \times 39 \times 38 \times 37 \times 36)$.

```
> 1/prod(40:36)
[1] 1.266449e-08
```

Figura 4: En R, se utiliza la función `prod`, que calcula el producto de un vector de números. En este caso, qué probabilidad hay, del 1 al 40, de obtener un número concreto dentro de 5 tiradas

4. Sin embargo, observe que esta es la probabilidad de obtener números en cualquier orden. Sin embargo, si este fuera un juego tipo "Lotería", entonces se estaría interesado en la probabilidad de adivinar correctamente un conjunto específicos de cinco números en un orden determinado. El conjunto específico de números y orden en sí es irrelevante, ya que todos los conjuntos de cinco números deben tener la misma probabilidad. Así que todo lo que tenemos que hacer es calcular el número de formas para elegir 5 números de entre 40. Esto se denota:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \longrightarrow \binom{40}{5} = \frac{40!}{5!(40-5)!} = 658008$$

```
> 1/choose(40, 5)
[1] 1.519738e-06
```

Figura 5: `choose(k, n)` calcula el número de conjuntos con n elementos que se pueden elegir de un conjunto con k elementos.

5. En este ejemplo se ilustrará como el tipo de muestreo (con o sin reemplazamiento) determina como abordar un cálculo de probabilidad.

Una urna contiene dos bolas blancas y tres rojas. Efectuadas dos extracciones sucesivas se quiere determinar la probabilidad de que la segunda sea roja:

a) en extracciones con reemplazamiento

Al reemplazar la bola extraída en primer lugar, cuando se extrae la segunda la configuración de la urna es la inicial por lo tanto la probabilidad de extraer una roja es $3/5$.

b) en extracciones sin reemplazamiento.

$$P(2^a \text{ Roja}) = P(2^a \text{ Roja} | 1^a \text{ Blanca}) \times P(1^a \text{ Blanca}) + P(2^a \text{ Roja} | 1^a \text{ Roja}) \times P(1^a \text{ Roja}) = \frac{3}{4} \times \frac{2}{5} + \frac{2}{4} \times \frac{3}{5} = \frac{3}{5}$$

Figura 6: Aplicación del teorema de la probabilidad total en caso de no reemplazamiento

3.2. Distribuciones de probabilidad para variables numéricas discretas y continuas.

1. Una de las distribuciones discretas más útiles es la distribución binomial. Se presenta siempre que contemos las veces que ocurre un suceso A al repetir n veces un experimento aleatorio. Por ejemplo, al lanzar un dado sea $A = \{3, 6\}$ (el suceso de obtener múltiplo de tres).

$$P\{X = x\} = \binom{n}{x} \pi^x (1 - \pi)^{n-x} \quad \text{para } x = 0, 1, 2, \dots, n \quad \rightarrow \quad P\{X = x\} = \binom{4}{x} \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{4-x}$$

Figura 7: Ejemplo de distribución de probabilidad: si se lanzan cuatro dados, ¿cuál es la probabilidad de que en x dados se realice el suceso A. $P(A) = \pi$

3.2.1. Código 1

Esta distribución, la binomial, se indica como $X = B(n; \pi)$, y los parámetros que la definen son el número de experimentos (*size*) y la proporción esperada de éxitos (*prob*).

```
> dbinom(2,size=10,prob=0.2)
[1] 0.3019899
```

Figura 8: Probabilidad que una binomial(10,0.2) tome el valor 2.
dbinom(x, size, prob)

```
> pbinom(2,size=10,prob=0.2)
[1] 0.6777995
```

Figura 9: Probabilidad de que $X \sim B(10,0.2)$ tome un valor inferior a 2.
pbinom(q, size, prob, lower.tail=TRUE)

```
> qbinom(0.9,size=10,prob=0.2)
[1] 4
```

Figura 10: Qué valor de $X \sim B(10,0.2)$ presenta una probabilidad acumulada de 0.9.
qbinom(p, size, prob, lower.tail=TRUE)

```
> rbinom(20,size=10,prob=0.2)
[1] 1 2 2 3 0 0 1 1 3 3 0 2 1 1 1 0 0 3 3 2
```

Figura 11: Generación de 20 valores aleatorios de una distribución binomial(10,0.2).
rbinom(n, size, prob)

```

> x<-seq(0,10,by=1)
> data.frame(x,p=dbinom(x,size=10,prob=0.2),F=pbinom(x,size=10,prob=0.2))
  x      p      F
1 0 0.1073741824 0.1073742
2 1 0.2684354560 0.3758096
3 2 0.3019898880 0.6777995
4 3 0.2013265920 0.8791261
5 4 0.0880803840 0.9672065
6 5 0.0264241152 0.9936306
7 6 0.0055050240 0.9991356
8 7 0.0007864320 0.9999221
9 8 0.0000737280 0.9999958
10 9 0.0000040960 0.9999999
11 10 0.0000001024 1.0000000

```

Figura 12: Distribución de probabilidad y función de distribución acumulada de una binomial(10,0.2).

Observar cómo, al ser una distribución discreta, la distribución acumulada consiste en la suma de las probabilidades puntuales \rightarrow para cada $X < x$

```

> x<-rbinom(20,size=1,prob=0.2)
> x
[1] 1 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0
> x<-factor(x)
> x
[1] 1 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0
Levels: 0 1

> levels(x)<-c("TAIL","HEAD")
> table(x)
x
TAIL HEAD
 17     3

```

Figura 13: Simulación de una binomial con 20 lanzamientos de un experimento en el que una moneda está trucada con probabilidad de obtener "cara"=0.20

2. Para modelar datos continuos, necesitamos definir variables aleatorias que pueden obtener el valor de cualquier número real. No existe, por tanto, tal cosa como una probabilidad puntual como para las variables aleatorias discretas. En cambio, tenemos el concepto de “densidad”.

En ciencias de la salud muchos datos surgen de mediciones en una escala esencialmente continua (por ejemplo, el colesterol plasmático o la presión arterial). Típicamente estas se agrupan alrededor de un valor central, siendo las desviaciones grandes del centro más raras que las desviaciones más pequeñas. Ésta es una distribución ampliamente utilizada que se denomina distribución “normal” o “gaussiana”. Formalmente, la función de densidad normal vendría dada por la expresión:

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

La distribución normal por tanto depende de la media (μ) y de la desviación estándar (σ), y cuando una variable aleatoria continua sigue una distribución normal se denota $X = N(\mu, \sigma)$. La función de distribución acumulada se puede definir también para las variables aleatorias continuas mediante la relación:

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

Las secuencias de números aleatorios que sepamos generar artificialmente nos permiten aproximarnos al funcionamiento de probabilidad y fundamentar conjeturas sobre sus propiedades.

3.2.2. Código 2

A continuación utilizaremos simulación para comprobar empíricamente que la distribución normal tiene una forma de campana característica y modificando sus parámetros μ y σ simplemente desplaza y amplía la distribución.

La distribución normal se usa comúnmente para describir el error de un modelo estadístico, y aparece como una distribución aproximada en varios contextos; por ejemplo, la distribución binomial para tamaños de muestra grandes puede estar bien aproximado por una distribución normal adecuadamente escalada.

```
x <- seq(-5,5,0.1)
```

```
> x
[1] -5.0 -4.9 -4.8 -4.7 -4.6 -4.5 -4.4 -4.3 -4.2 -4.1 -4.0 -3.9 -3.8 -3.7 -3.6 -3.5 -3.4 -3.3 -3.2 -3.1
[21] -3.0 -2.9 -2.8 -2.7 -2.6 -2.5 -2.4 -2.3 -2.2 -2.1 -2.0 -1.9 -1.8 -1.7 -1.6 -1.5 -1.4 -1.3 -1.2 -1.1
[41] -1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1  0.0  0.1  0.2  0.3  0.4  0.5  0.6  0.7  0.8  0.9
[61]  1.0  1.1  1.2  1.3  1.4  1.5  1.6  1.7  1.8  1.9  2.0  2.1  2.2  2.3  2.4  2.5  2.6  2.7  2.8  2.9
[81]  3.0  3.1  3.2  3.3  3.4  3.5  3.6  3.7  3.8  3.9  4.0  4.1  4.2  4.3  4.4  4.5  4.6  4.7  4.8  4.9
[101] 5.0
```

Figura 14: La función `seq()` se usa para generar valores equidistantes, aquí de -5 a 5 en pasos de $0,1$; es decir, $(-5.0, -4.9, -4.8, \dots, 4.9, 5.0)$

La función de densidad es probablemente el tipo de función en R menos usado en la práctica para la distribución normal. Pero si por ejemplo se desea dibujar la conocida curva de campana de la distribución normal, entonces se puede hacer así:

```
plot(x, dnorm(x), type="l", col=2, ylim=c(0,1), ylab="dnorm(x, mean, sd)")
```

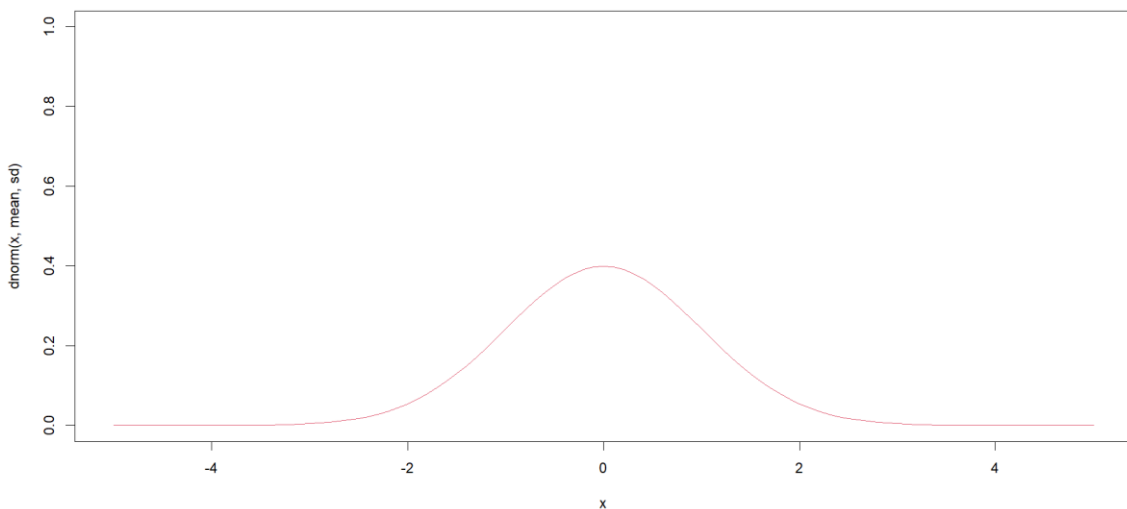


Figura 15: Muestra de cómo representar una campana de distribución normal: `plot(x, dnorm(x))`.

El uso de `type="l"` como argumento para trazar hace que la función dibuje líneas entre los puntos en lugar de trazar los puntos mismos. `col=2` sirve para colorear la línea de rojo. `ylim= c(0,1)` sirve para etiquetar el eje y del 0 al 1. `ylab` sirve para etiquetar el eje y.

Si no se especifica, el programa asume media centrada en 0 y desviación estándar de 1

`lines(x, dnorm(x, mean=0, sd=0.50), col =3)`

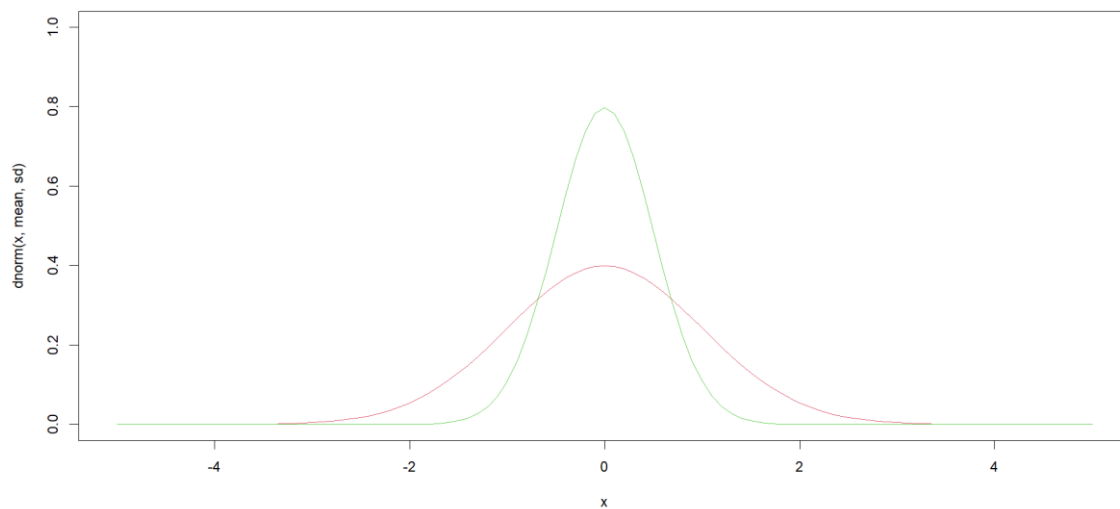


Figura 16: Media en 0 y desviación típica (sd)=0.5 (se empequeñece el ancho de la campana).

`lines(x, dnorm(x, mean=0, sd=2), col =4)`

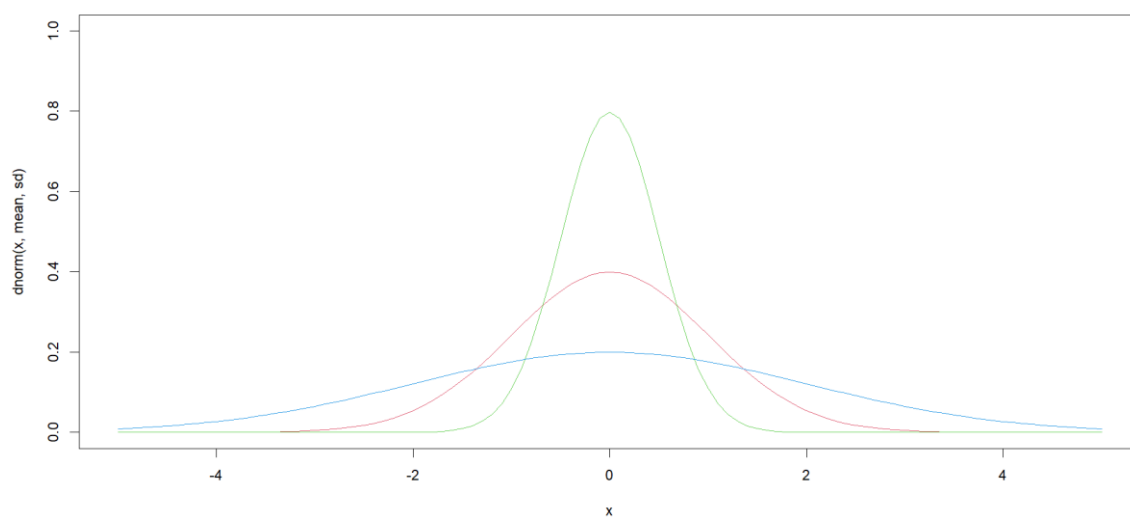


Figura 17: Desviación típica > 0 (se agranda el ancho de la campana de distribución)

```
lines(x, dnorm(x, mean=-2, sd=1), col=6)
```

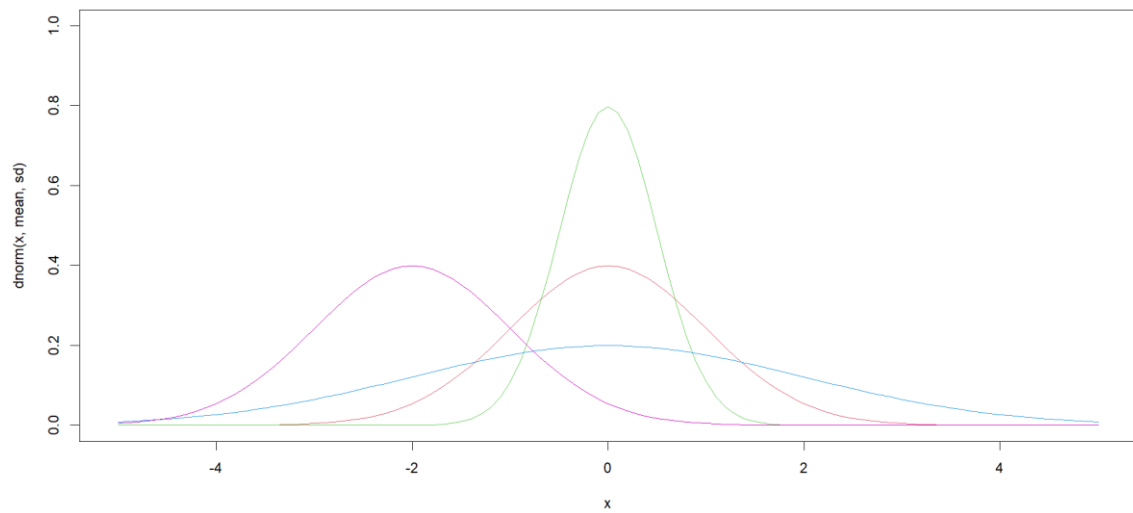


Figura 17: Media desplazada a -2 y desviación estándar de 1

```
legend("topright", c("mu=0, sigma=1", "mu=0, sigma=0.5", "mu=0, sigma=2", "mu=-2, sigma=1"), lty=1, lwd=1, col=c(2,3,4,6))
```

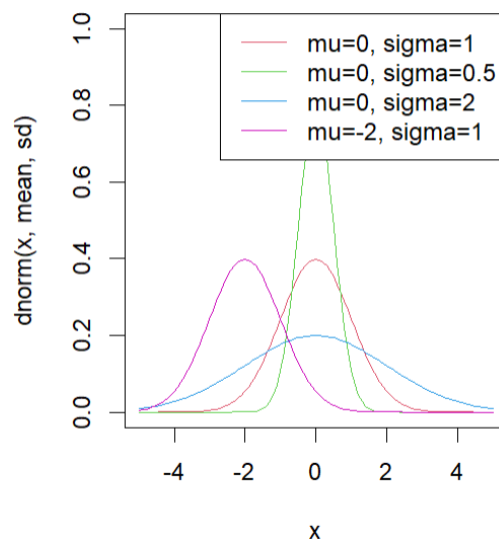


Figura 18: Inserción de leyenda

Ahora comprobaremos empíricamente cómo una distribución discreta como la binomial se puede aproximar a la distribución normal con un tamaño muestral suficientemente grande

```
x <- seq(0,20, 1)
```

```
> x
[1] 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
```

Figura 19: Secuencia de 20 números, en saltos de 1 en 1

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

```
plot(x, dbinom(x, 0.10, size=10), type="h", col=2,
```

```
ylim=c(0,0.40), ylab="P(X=x)", xlab="x (pi=0.10, n=10)")
```

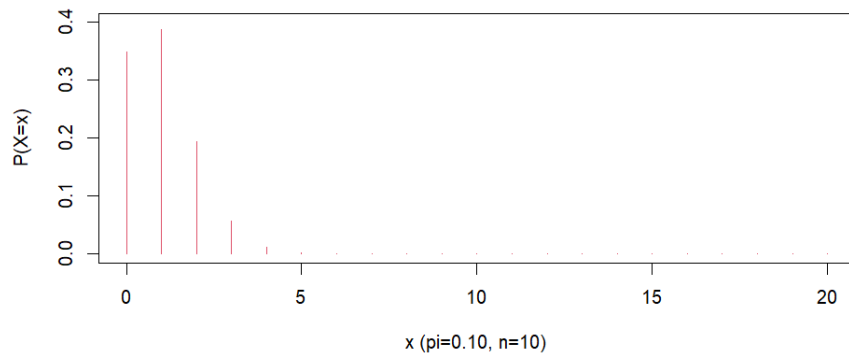


Figura 20: Distribución binomial.

La longitud de muestras es tan reducida, que apenas se distingue una forma definida. La función `type="h"` genera líneas verticales (en vez de puntos), como las que se observan en la Figura.

```
plot(x, dbinom(x, 0.10, size=25),type="h", col=2,
```

```
ylim=c(0,0.40), ylab="P(X=x)", xlab="x (pi=0.10, n=25)")
```

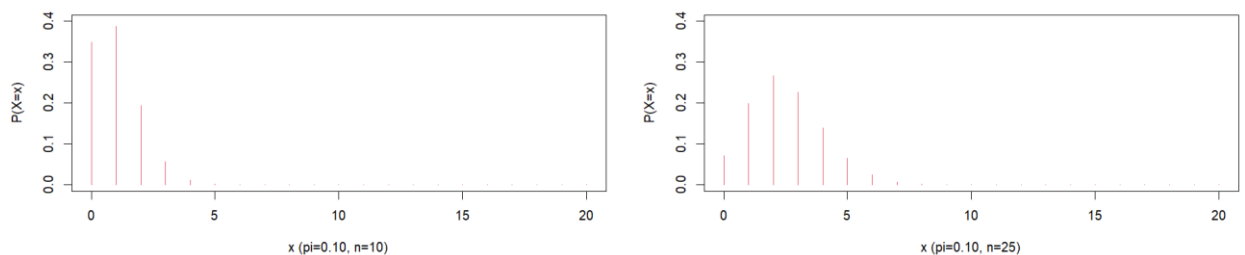


Figura 21: Comienza a distinguirse un esbozo de forma de campana al tener 25 muestras (aún no se puede considerar distribución normal)

```
plot(x, dbinom(x, 0.10, size=50),type="h", col=2,
```

```
ylim=c(0,0.40), ylab="P(X=x)", xlab="x (pi=0.10, n=50)")
```

```
plot(x, dbinom(x, 0.10, size=100),type="h", col=2,
```

```
ylim=c(0,0.40), ylab="P(X=x)", xlab="x (pi=0.10, n=100)")
```

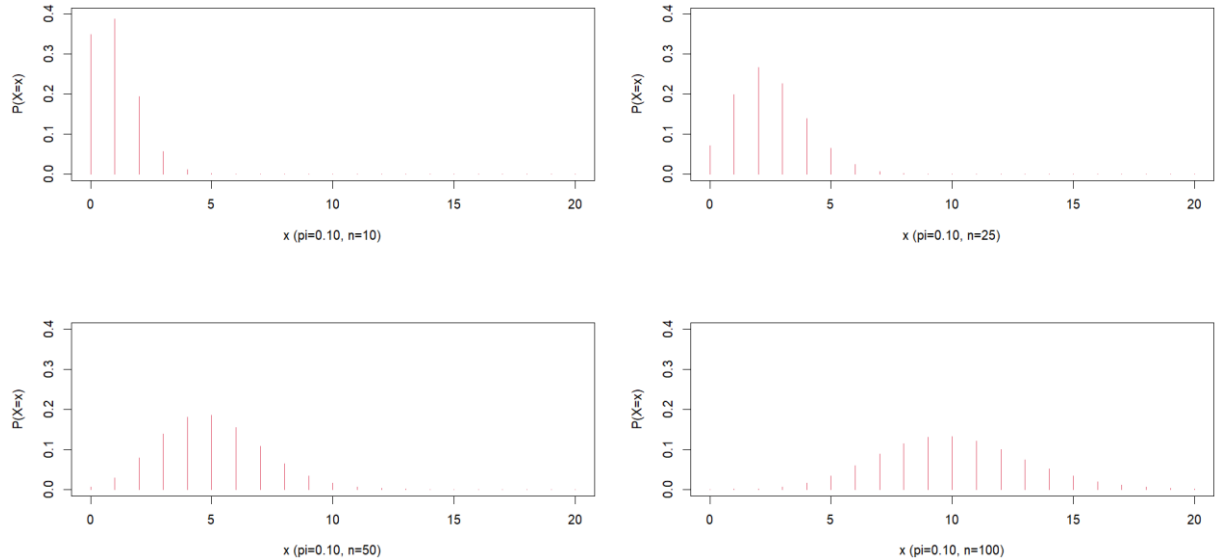


Figura 22: Con 50 y 100 muestras se distingue una forma de campana, que se asemeja más a la campana Gaussiana *conocida de forma teórica*.

```
lines(x, dnorm(x, 10, sqrt(9)), col =1)
```

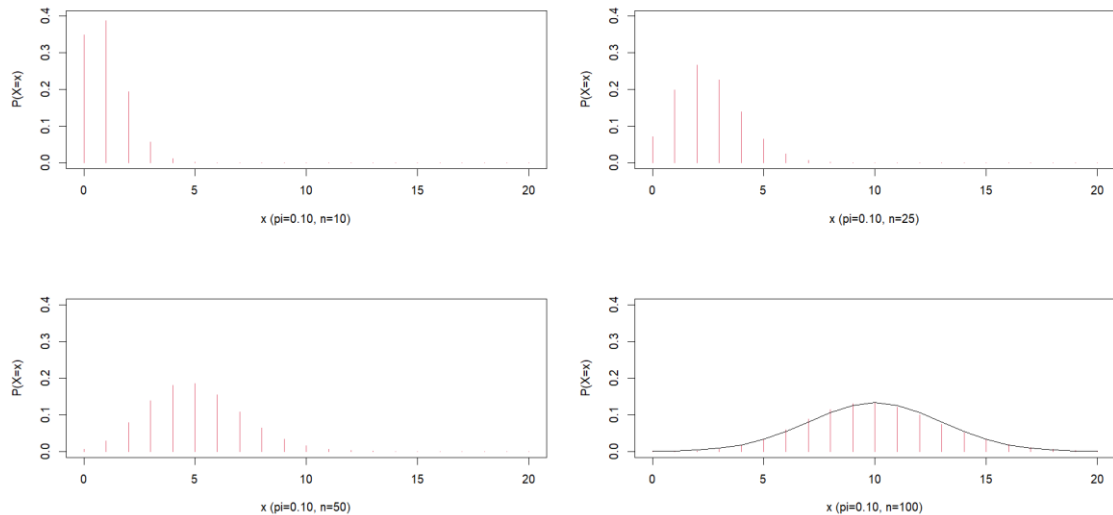


Figura 23: Se pinta una línea sobre el panel con $N=100$, y se obtiene un gráfico en el que se refleja a la perfección la estructura de campana de Gauss (valores centrales muy frecuentes y extremos muy poco frecuentes).

Es posible dibujarla porque se ha calculado, por un lado, que como la función de densidad: $N \cdot \pi = 100 \cdot 0.10 = 10$ (media = 10), y por el otro lado la varianza: $N \cdot \pi(1-\pi) = 100 \cdot 0.1 \cdot 0.90 = 9$ (desviación típica = $\sqrt{9}$)

3.3. Intervalos de confianza y contraste de hipótesis.

A continuación, veremos cómo las distribuciones de probabilidad pueden ser útiles para evaluar cómo de extremo/esperable es el resultado de un experimento/observación.

1. La función inversa de distribución de probabilidad en R (`qbinom()`, `qnorm()`), representa el valor de X con la propiedad de que existe una probabilidad p de obtener un valor menor o igual que él. La mediana es por definición el cuantil 50%. Una instancia de la distribución normal es la normal “tipificada” que tiene $\mu = 0$ y $\sigma = 1$. Las probabilidades bajo la función de densidad normal estandarizada, están tabuladas y son fácilmente accesibles. Las tablas de distribuciones estadísticas casi siempre se dan en términos de cuantiles.

Para un conjunto fijo de probabilidades, la tabla muestra el límite que debe cruzar un estadístico para que una prueba se considere significativa a cierto nivel. En R se puede estimar la p exactamente sin necesidad de consultar tablas. Por ejemplo, si tenemos n observaciones distribuidas normalmente con la misma media μ y desviación estándar σ , entonces se sabe (Teorema del límite central) que el promedio \bar{x} se distribuye normalmente alrededor de μ con desviación estándar σ/\sqrt{n} . El intervalo de confianza al 95% para μ se puede obtener con la expresión:

$$\bar{x} - \frac{\sigma}{\sqrt{n}} * N_{0.025} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}} * N_{0.975}$$

Donde $N_{0.025}$ es el cuantil 2.5% y $N_{0.975}$ es el cuantil 97.5% de la distribución normal.

```
> xbar <- 83; sigma <- 12; n <- 5
> sem <- sigma/sqrt(n)
> sem
[1] 5.366563
> xbar + sem * qnorm(0.025)
[1] 72.48173
> xbar + sem * qnorm(0.975)
[1] 93.51827
```

Figura 24: Desviación típica (σ) = 12, $n = 5$ personas con una media muestral de $\bar{x} = 83$.

Se ha encontrado de esta manera un intervalo de confianza al 95% para la media poblacional (μ) que va del 72.48 al 93.52.

La propiedad fundamental que presenta en intervalo de confianza al $100*(1-\alpha)\%$ es que consiste en un procedimiento que al realizarlo muchas veces contiene al verdadero parámetro de la distribución que estamos buscando con una probabilidad $(1-\alpha)$. Nótese que este cálculo asume que la desviación típica poblacional σ es conocida. Sin embargo, el caso más común es estimar σ a partir de la desviación típica muestral s lo que implica que la distribución de comparación sea la distribución t con $n-1$ grados de libertad en lugar de la normal tipificada

2. Cualquier observación de x sobre la línea base de la curva normal se puede estandarizar como el número de unidades de desviaciones típicas, mediante la expresión: $z = (x - \mu) / \sigma$. En general, la tipificación permite obtener de las tablas la función de distribución, es decir la probabilidad de que la variable normal estandarizada tome un valor igual o inferior al estadístico z . Estandarizando una observación x a su valor z , podemos relacionarlo con las propiedades de las curvas normales para realizar contrastes de hipótesis.

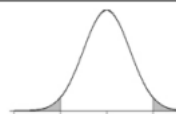
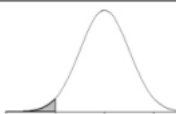
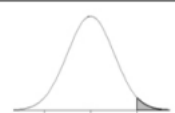
Hipotesis Nula	$H_0: \mu = \mu_0$	$H_0: \mu \geq \mu_0$	$H_0: \mu \leq \mu_0$
Hipotesis Alternativa	$H_1: \mu \neq \mu_0$	$H_1: \mu < \mu_0$	$H_1: \mu > \mu_0$
Representacion grafica			

Figura 25: Representación de los tipos de hipótesis que se comparan: la hipótesis nula y la alternativa.

Si se quiere saber si un medicamento funciona igual en hombres o mujeres, en un contexto u otro, ciudades distintas, países distintos, la hipótesis alternativa en esos casos es la igualdad. El contraste sería bilateral o unilateral según el problema, y también la tradición o la convención de los contrastes que se haya hecho antes con estudios iguales o similares.

El contraste de la hipótesis nula H_0 y alternativa H_1 consiste en plantear una prueba que nos arroje evidencia en favor de una u otra hipótesis. Esta hipótesis nula se aceptará si los datos muestrales no aportan suficiente evidencia en contra de ésta. Por el contrario, si se cuenta con pruebas suficientes para contradecir la hipótesis nula, ésta se rechazará en favor de una hipótesis alternativa, que corresponde generalmente a la negación de la nula.

Por poner un ejemplo, en España la media de altura en hombres es 176,6 cm. Se tomó al azar una muestra de 6 hombres cuyas alturas fueron 176, 168, 173, 167, 181, y 167. ¿Es esta media muestral suficientemente extrema como para rechazar la hipótesis alternativa bilateral de desigualdad de medias?

Para muestras pequeñas o con desviación típica poblacional desconocida se utiliza el estadístico de contraste $t = (\bar{y} - \mu) / (s / \sqrt{n})$ que sigue una distribución t con $n-1$ grados de libertad (df). Generalmente en los contrastes bilaterales se selecciona el p valor de la cola más pequeña de la distribución y se multiplica por dos.

```

> muestra <- c(176,168,173,167,181,167); mu=176.6; media = mean(muestra)
> media
[1] 172
> s = sqrt(var(muestra))
> s
[1] 5.727128
> t = (media - mu)/(s/sqrt(length(muestra)))
> t
[1] -1.967418
> pt(t, df=5, lower.tail=TRUE)
[1] 0.05313725
> 2*pt(t, df=5, lower.tail=TRUE)
[1] 0.1062745

```

Figura 26: En este caso la hipótesis nula es $H_0 \equiv \mu = 176.6$, y para los datos $y = 172.0$, $y s = 5.727$; por lo que el estadístico de contraste es $t = (y - \mu) / (s / \sqrt{n}) = -1.967418$.

Considerando el contraste bilateral se dirá que el p-valor es 0.11. La diferencia entre medias no parece muy extrema, pero sí parece que los datos sugieren una tendencia a que la media muestral no sea representativa de la poblacional.

Hay que ser cautos ya que los p-valores obtenidos en los contrastes de hipótesis dependen de factores como el tamaño de la muestra, la magnitud de la diferencia que se estima relevante o la variabilidad de las observaciones. Se puede comprobar empíricamente que al aumentar el número de observaciones el p-valor disminuye. De ahí la importancia de considerar al p-valor como un mero elemento más de juicio que se debe valorar junto con otras consideraciones.

4. EJERCICIOS PROPUESTOS

4.1. Ejercicio 1

Dada la variable aleatoria X cuya distribución de probabilidad es:

x_i	1	2	3	4	5
$P(X = x_i)$	2/8	1/8	2/8	2/8	1/8

realizar la representación gráfica de la función de distribución de X , y calcular las probabilidades $P(1 < X \leq 3.7)$, y $P(1.5 \leq X \leq 3.5)$.

Como datos se obtiene una serie de x_i , cuyas respectivas probabilidades son 2/8, 1/8, 2/8, 2/8 y 1/8. Dado que no se trata de ninguna distribución en concreto, se ha optado por crear un sencillo gráfico de barras que represente en cada valor del eje x , su respectiva probabilidad.

El código propio creado es:

```
x= c(1,2,3,4,5)
```

```
Px= c(2/8,1/8,2/8,2/8,1/8)
```

```
plot(x, Px, type="h", xlab="X", ylab="P(X=x)", col=4, ylim=c(0,0.28), lwd = 5) #type = "h" para  
dibujar líneas verticales
```

```
grid(nx = NA, ny = NULL,lty = 1, col = "gray", lwd = 1) #cuadricular el gráfico
```

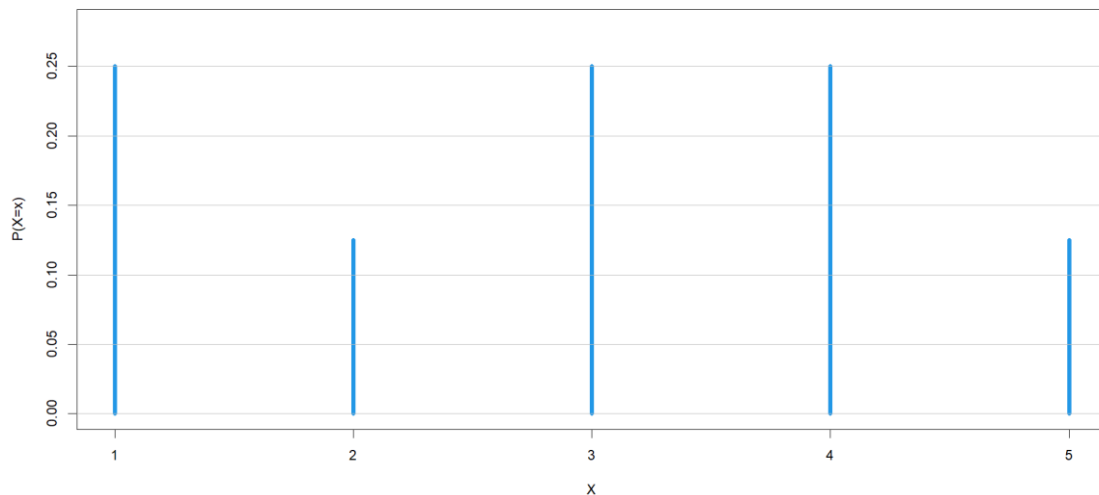


Figura 27: Representación gráfica de la salida del código del ejercicio 1 propuesto. Observar que los valores 1, 3 y 4 tienen la misma probabilidad (una densidad del 0.25), mientras que el 2 y el 4, tienen una densidad de 0.125 (es decir, 1/8).

Por el otro lado, el enunciado pide calcular la probabilidad de $P(1 < X \leq 3.7)$, es y $P(1.5 \leq X \leq 3.5)$. Se han creado un simple código, que agrega líneas verticales, por un lado en 1 y 3.7; y por el otro lado en 1.5 y 3.5, de forma que se pueda observar fácilmente qué valor toma $P(X)$ en cada caso.

```
par(mfrow=c(1,2)) #obtener dos gráficos en la misma figura
```

```
plot(x, Px, type="h", xlab="X", ylab="P(1 < X ≤ 3.7)", col=4, ylim=c(0,0.28), lwd = 5)
```

```
abline(v = c(1,3.7), col=2) #añadir en x=1 y x=3.7 líneas rojas verticales para analizar el contenido entre medias
```

```
plot(x, Px, type="h", xlab="X", ylab="P(1.5 ≤ X ≤ 3.5)", col=4, ylim=c(0,0.28), lwd = 5)
```

```
abline(v = c(1.5,3.5), col=6)
```

```
Pintervalo=Px[2]+Px[3] #0.125+0.25=0.375, que corresponde a  $P(1 < X \leq 3.7)$ , y  $P(1.5 \leq X \leq 3.5)$ 
```

Pintervalo

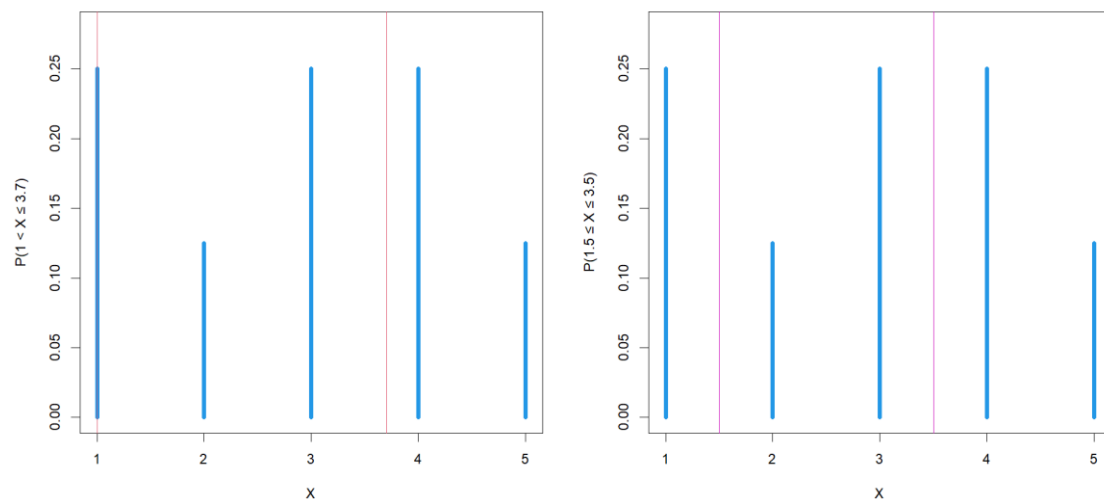


Figura 28: Demostración de en qué intervalos toma $P(X)$ sus valores, en los casos de $P(1 < X \leq 3.7)$, es y $P(1.5 \leq X \leq 3.5)$.

En el primer caso, $P(X) = P(X=2) + P(X=3)$ (no se coje $P(X=1)$, porque el enunciado exige claramente $(P < 1)$). En el segundo caso, puede verse que se obtienen las mismas columnas de probabilidad: $P(X) = P(X=2) + P(X=3)$.

En ambos casos, $P(X) = 0.125 + 0.25 = 0.375$

```
> Pintervalo=Px[2]+Px[3] #0.125+0.25=0.375, que corresponde a  $P(1 < X \leq 3.7)$ , y  $P(1.5 \leq X \leq 3.5)$ 
> Pintervalo
[1] 0.375
```

Figura 29: Salida del código $P(1 < X \leq 3.7)$, y $P(1.5 \leq X \leq 3.5) = 0.375$

4.2. Ejercicio 2

Cinco estudiantes preparan un examen de cierta asignatura en la que la probabilidad de que un alumno apruebe es de $2/5$. Determinar la probabilidad de que:

- a) Todos aprueben
- b) Que aprueben al menos dos
- c) Que sólo aprueben dos
- d) Que al menos apruebe uno.

Dado que en los ejercicios resueltos se ha aprendido a utilizar la distribución binomial (distribución discreta que cuenta el número de éxitos en una secuencia de N ensayos) para calcular probabilidades, en este ejercicio se ha recurrido a dichas funciones de código.

En el primer apartado, que todos aprueben implica (visualizando, por ejemplo, un árbol de probabilidades), que se siga siempre “el camino” de los $2/5$, es decir, se debe multiplicar $2/5$, 5 veces. Manualmente se obtiene 0.01024.

La línea de código utilizada en este caso es:

#a $P(x=5)$

`dbinom(5, size=5, prob=(2/5))` #calcular la probabilidad de que aprueben 5 de 5

```
> #a P(x=5)
> dbinom(5, size=5, prob=(2/5)) #calcular la probabilidad de que aprueben 5 de 5
[1] 0.01024
```

Figura 30: Salida del código referido al apartado a) del ejercicio 2 propuesto. Observar que se obtiene el mismo resultado que calculándolo manualmente.

`dbinom(x, size, prob)` implica una distribución normal que coge el número exacto (x), y calcula su probabilidad ($prob$), entre el número de muestras ($size$).

En el apartado b, se pide que se aprueben al menos dos. Se deben buscar todas las combinaciones que implique que dos o más alumnos aprueben; o viéndolo de otra forma, la probabilidad complementaria a que apruebe 1 o 0 alumnos.

Saber que `pbinom(x, size, prob)`, devuelve la probabilidad inferior a x , por tanto, es lógico pensar que se desea obtener el cálculo contrario a obtener la probabilidad menor a 2; por lo que se realiza el complementario a ello, y se le suma la probabilidad de obtener 2 de manera exacta. Se ha querido comprobar calculando la probabilidad complementaria a que aprueben 0 más la probabilidad de que apruebe 1

Se ha codificado como:

#b $P(x \geq 2)$

`1-pbinom(2, size=5, prob=(2/5)) + dbinom(2, size=5, prob=(2/5))` #calcular la probabilidad de que aprueben 1 - (menores a 2 de 5) + probabilidad de 2 ($x \geq 2$)

`1-(dbinom(0, size=5, prob=(2/5)) + dbinom(1, size=5, prob=(2/5)))` #para comprobar, se calcula la probabilidad complementaria a que aprueben 0 + que apruebe 1

```

> #b P(x>=2)
> 1-pbinom(2, size=5,prob=(2/5)) + dbinom(2, size=5,prob=(2/5)) #c
ben 1 - (menores a 2 de 5) + probabilidad de 2 (x>=2)
[1] 0.66304
> 1-(dbinom(0, size=5,prob=(2/5)) + dbinom(1, size=5,prob=(2/5)))
bilidad complementaria a q aprueben 0 + que apruebe 1
[1] 0.66304

```

Figura 31: Salida del código referido al apartado b) del ejercicio 2 propuesto. Se ha calculado de dos maneras (para verificar que se ha realizado correctamente):
La primera, es calculando la probabilidad de obtener dos o más aprobados.
La segunda, la probabilidad de obtener el contrario a 0 + 1 aprobados.

En el apartado c, se pide calcular la probabilidad de que sólo aprueben 2. Se ha realizado dicho cálculo en el apartado anterior:

#c P(x=2)

dbinom(2, size=5,prob=(2/5)) #calcular de manera binomial que sólo aprueban 2

```

> #c P(x=2)
> dbinom(2, size=5,prob=(2/5))
[1] 0.3456

```

Figura 32: Salida del código referido al apartado c) del ejercicio 2 propuesto.

Por último, se pide calcular la probabilidad de que apruebe al menos uno. Teniendo en cuenta que la función *pbinom(x, size, prob)* calcula la probabilidad a inferior a x, se va a calcular mediante el complementario a la probabilidad de obtener menor a uno (o lo que es lo mismo, el complementario a obtener 0 suspensos).

#d P(x>=1)

1-pbinom(1, size=5,prob=(2/5)) + dbinom(1, size=5,prob=(2/5)) #calcular la probabilidad de que aprueben 1 - (menores a 1 de 5) + probabilidad de 1 (x>=1)

1-dbinom(0, size=5,prob=(2/5)) #para comprobar, se calcula la probabilidad complementaria a q aprueben 0

```

> #d P(x>=1)
> 1-pbinom(1, size=5,prob=(2/5)) + dbinom(1, size=5,prob=(2/5))
[1] 0.92224
>
> 1-dbinom(0, size=5,prob=(2/5))
[1] 0.92224

```

Figura 33: Salida del código referido al apartado c) del ejercicio 2 propuesto. Se ha calculado de dos maneras (para verificar que se ha realizado correctamente):
La primera, es calculando la probabilidad de obtener uno o más aprobados.
La segunda, la probabilidad de obtener el contrario a 0 aprobados.

4.3. Ejercicio 3

Una muestra aleatoria de valores de colesterol total (mg/dL) en menores de 21 años (X), cuya distribución de probabilidad en la población se supone Normal, obtiene los siguientes resultados: 165, 162, 165, 166, 164, 165, 170, 169, y 168. Elaborar un intervalo de confianza al 99% para la media de la población e interprételo.

Nota: para muestras pequeñas, o si la varianza poblacional es desconocida, para la estimación de intervalos de confianza se utiliza la distribución "t de Student" con $n - 1$ grados de libertad.

Antes de proceder a resolver el ejercicio, se va a explicar el razonamiento que se ha tratado de seguir:

Se tiene una serie de 6 muestras de valores de colesterol total tomados en menores de 21 años, y se quiere hallar, con un 99% de confianza, los intervalos en los que debería de poder estar la media real.

Por el momento, se sabe que la media muestral será la media de 165, 162, 165, 166, 164, 165, 170, 169, y 168.

Por el otro lado, para muestras pequeñas, o si la varianza poblacional es desconocida (nos enfrentamos a ambos casos), para la estimación de intervalos de confianza se utiliza la distribución "t de Student" con $n - 1$ grados de libertad, cuya fórmula principal es:

$$z_{\alpha/2} = x_1 - X' / (\sigma / \sqrt{N})$$

Si se despeja x_1 , se obtiene que el intervalo inferior será: $x_1 = X' - z_{\alpha/2} \cdot (\sigma / \sqrt{N})$, y el superior:

$$x_2 = X' + z_{\alpha/2} \cdot (\sigma / \sqrt{N})$$

Entonces:

$$IC = (x_1, x_2) \longrightarrow IC = (X' - \text{error}, X' + \text{error}) \longrightarrow IC = (X' - z_{\alpha/2} \cdot (\sigma / \sqrt{N}), X' + z_{\alpha/2} \cdot (\sigma / \sqrt{N}))$$

Por tanto, sólo hay que restarle y sumarle a la media muestral el error. Debemos calcularlo.

El código creado para calcular los intervalos es:

```
muestra=c(165, 162, 165, 166, 164, 165, 170, 169, 168)
```

```
N=9
```

```
media_muestral=mean(muestra) #se calcula la media de la muestra
```

```
sigma= sd(muestra) #se valcula la varianza
```

```
z_a2=qt(0.01/2, df=N-1) #se calcula el cuantil del 99% (es decir, 0.01/2). df=grados de libertad
```

```
error= z_a2*(sigma/sqrt(N)) #se calcula el error
```

```
error
```

```
int_inf=media_muestral - (-error) #intervalo inferior
```

```
int_inf
```

```
int_sup=media_muestral + (-error) #intervalo superior
```

```
int_sup
```

```

> muestra=c(165, 162, 165, 166, 164, 165, 170, 169, 168)
> N=9
> media_muestral=mean(muestra) #se calcula la media de la muestra
> sigma= sd(muestra) #se calcula la varianza
> z_a2=qt(0.01/2, df=N-1) #se calcula el cuantil del 99% (es decir, 0.01/2). df=grados de libertad
> error= z_a2*(sigma/sqrt(N)) #se calcula el error
>
> int_inf=media_muestral - error #intervalo inferior
> int_inf
[1] 168.8515
> int_sup=media_muestral + error #intervalo superior
> int_sup
[1] 163.1485

```

Figura 34: Salida del código del ejercicio 3

Se puede concluir que con una confianza del 99%, según la muestra dada (con una N mayor, hubiéramos acotado el intervalo), la media real de colesterol total para personas menores de 21 años se va a encontrar entre los 163.1485 y los 168.8515 mg/dL.

Se ha querido verificar en Internet, y según la fuente

<https://myhealth.ucsd.edu/Spanish/RelatedItems/90,P04693> , tal y como anuncia el Instituto Nacional del Corazón, los Pulmones y la Sangre, estos son los parámetros que definen el estado de salud (según los niveles de colesterol) en niños y adolescentes (entre 2 y 19 años):

	Colesterol total	Colesterol LDL
Aceptable	Menor a 170 mg/dL	Menor a 110 mg/dL
Límite	De 170 a 199 mg/dL	De 110 a 129 mg/dL
Alto	200 mg/dL o mayor	130 mg/dL o mayor

Figura 35: Figura ofrecida por el Instituto Nacional del Corazón, los Pulmones y la Sangre que califica el estado de salud de niños entre 2 y 19 años según sus niveles de colesterol

La conclusión a la que se ha llegado es que las personas de las que fue tomada la muestra de manera aleatoria correspondían a los niveles aceptables de colesterol, y se ha podido comprobar que es correcto anunciar que, con una confianza del 99%, la supuesta media se encuentra entre los 163.1485 y los 168.8515 mg/dL de colesterol.

4.4. Ejercicio 4

La diferencia entre el peso a los 4 meses de vida y el peso al nacer en 16 recién nacidos de una muestra, sigue una distribución $N(\mu, \sigma)$, y se supone que debe ser al menos de 2.5 kg. En un estudio de 16 niños a los que se le ha evaluado el peso al nacer y a los 4 meses, las diferencias en el peso han dado una media muestral de 2.340 Kg y una varianza muestral igual a 0.36.

¿Hay evidencia para pensar que estos niños no se han desarrollado lo suficiente? Interprete el p valor en el contexto de este contraste de hipótesis.

Nota: para muestras pequeñas, o si la varianza poblacional es desconocida, el estadístico necesario para el contraste de hipótesis se distribuye según una “t de Student” con $n - 1$ grados de libertad.

La hipótesis nula (H_0) de la que se parte es que $\mu \geq \mu_0$, es decir, la media a suponer que un niño crece sano es de 2.5 kg, y la media obtenida al muestrear, es 2.3 (menor a μ). La pregunta entonces es, ¿hay razones, o mejor dicho, confianza suficiente, para pensar que la media obtenida es suficientemente baja respecto a la teórica (¿no se han desarrollado estos niños lo suficiente?).

Nos situamos en un contraste unilateral tipo $\mu \geq \mu_0$, en el que la hipótesis alternativa será que $\mu < \mu_0$.

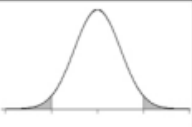
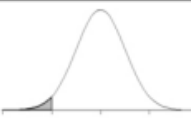
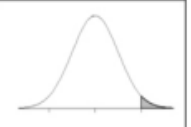
Hipotesis Nula	$H_0: \mu = \mu_0$	$H_0: \mu \geq \mu_0$	$H_0: \mu \leq \mu_0$
Hipotesis Alternativa	$H_1: \mu \neq \mu_0$	$H_1: \mu < \mu_0$	$H_1: \mu > \mu_0$
Representacion grafica			

Figura 36: Representación gráfica de la hipótesis de contraste referida al ejercicio 4 (Hipótesis nula: $\mu \geq \mu_0$)

El código creado es el siguiente:

```
supuesta_media=2.5
```

```
nmuestral=16
```

```
media_muestral=2.340
```

```
desv_muestral= sqrt(0.36)
```

```
t = (media_muestral - supuesta_media)/(desv_muestral/sqrt(16)) #y' - μ/(s/√n)
```

```
p=pt(t, df=15, lower.tail=TRUE) #probabilidad de significación
```

Ahora bien, cuando en los artículos científicos se realiza un contraste de hipótesis, se toma de referencia la probabilidad de significación p. Esta p se puede obtener en cualquier paquete estadístico:

- Si $p < 0.05$, las diferencias son significativas a un nivel de confianza del 95%.
- Si $p < 0.01$, las diferencias serían entonces significativas a un nivel de confianza del 99%.

- Si $p < 0.001$, las diferencias serían significativas a un nivel de confianza del 99.9%.

En esta memoria, se utilizarán los mismos parámetros: se suele utilizar tradicionalmente 0.05 como nivel de significación (construyendo intervalos de confianza al 95%).

```
> supuesta_media=2.5
> nmuestral=16
> media_muestral=2.340
> desv_muestral= sqrt(0.36)
>
> t = (media_muestral - supuesta_media)/(desv_muestral/sqrt(16)) #y'- μ/(s/√n)
> t
[1] -1.066667
> p=pt(t, df=15, lower.tail=TRUE) #probabilidad de significación
> p
[1] 0.1514952
```

Figura 37: Salida del código creado para resolver el ejercicio propuesto 4. Se obtiene como valor estadístico de -1.066667 y una probabilidad de significación de 0.1514952.

Dado que el valor p (no $2*p$ porque es unilateral, no bilateral) es mayor a 0.05 (confianza del 95%), no se disponen de suficientes datos para demostrar, y sospechar que dichos niños estaban poco desarrollados. Quizá con una muestra mayor se hubieran reducido los errores de tipo I y de tipo II. Por tanto, se aceptará la hipótesis alternativa, en este caso, que los niños no están subdesarrollados.

5. CONCLUSIONES

Tal y como era de esperar, RStudio agiliza los cálculos referidos a probabilidad, distribuciones e intervalos de confianza, que hechos a mano pueden resultar engorrosos y puede haber errores de cálculo por aproximaciones.

En esta práctica se han repasado conceptos de la práctica pasada, pero era más importante entender la teoría, que entender cómo funcionaba R (siguiendo los ejercicios guiados, los ejercicios propuestos eran bastante parecidos, y sencillos).