INFERENCIA ESTADÍSTICA

MÁSTER BIOINFORMÁTICA Y BIOESTADÍSTICA

PRUEBA DE EVALUACIÓN CONTÍNUA 1

Universitat Oberta de Catalunya – Universidad de Barcelona

Santiago Royuela Samit

Abril de 2022

Ejercicio 1

Identifica qué distribución considerarías más apropiada para cada una de las medidas siguientes? Indica los parámetros necesarios para calcular las probabilidades.

a. Se está interesado en conocer si en una muestra de anatomía patológica de un tumor es maligno o no

A) Estamos ante un proceso en donde una muestra tumoral se analiza y obtenemos si es maligno o no, es decir, éxito o fracaso, según se mire y se defina previamente. Bien el éxito pueda ser el "ser maligno", o bien pueda convenirse que el éxito corresponda a un tumor "no maligno". Supongamos que nuestro objetivo es detectar tumores malignos, así que tomaremos como éxito cuando topamos con uno de ellos y como fracaso, cuando el tumor no es maligno. Lo convenimos así sin perder generalidad alguna.

Cada vez que analizamos una muestra para ver si es maligna o no, claramente estamos ante un **ensayo de Bernoulli**, que es aquel en el que solo se pueden obtener **dos resultados** (Maligno/No Maligno)

El ensayo de Bernoulli está modelado por una variable aleatoria que puede tomar sólo dos valores: 0, 1. Utilizaremos el 1 para el "éxito"/"maligno" y el 0 para "fracaso"/"benigno".

En este caso, en base a datos empíricos, estadísticas y otros estudios, debería existir un parámetro "p" que indica la probabilidad de éxito (maligno), siendo q=1-p el fracaso (benigno) en una muestra de anatomía patológica de un tumor.

De esta manera, la distribución apropiada que describirá este proceso será una distribución de Bernoulli de parámetro p: B(p). También, una Binomial n=1, p, Bin(1,p) será idéntica a la distribución de Bernoulli.

La función de masa de probabilidad de la distribución de Bernoulli es:

$$P(X = x) = p^x(1-p)^{1-x}\,, x = 0{,}1$$

$$x = 1\ (tumor\ maligno); x = 0\ (tumor\ benigno)$$

Su esperanza y varianza serán:

$$E(X) = p\ ;\ \sigma^2 = Var(X) = p(1-p)$$

b. Se dispone de 20 pacientes con un tumor y se desea cuantificar cuántos de ellos tiene un tumor maligno

Tenemos n=20 pacientes con tumor, pero no sabemos cuántos de ellos son malignos. Por la respuesta anterior sabemos el parámetro "p" que indica la probabilidad de que, al analizar un tumor, éste sea maligno o no. De esta manera, podemos proceder razonando.

Sabemos, por el apartado anterior que, en una muestra de anatomía patológica de un tumor, existe una probabilidad "p" de que éste sea maligno. En n=20 pacientes, sin importar su orden, podemos tener de 0 a 20 tumores malignos. Si hubiesen 0<=k<=20 tumores, tendríamos diferentes posibilidades de llegar a ese resultado, dado por el número combinatorio:

$$\binom{n = 20}{0 \le k \le 20} = \frac{n!}{k!\,(n-k)!}$$

La distribución binomial o distribución binómica es una distribución de probabilidad discreta que cuenta el número de éxitos en una secuencia de n ensayos de Bernoulli independientes entre sí, con una probabilidad fija "p" de ocurrencia de éxito (maligno) entre los ensayos. Así que éste será nuestro caso o paradigma estadístico de estudio.

En este caso, solo podemos esperar calcular una **esperanza muestral** con n=20 sujetos con tumores, que pueden ser o no malignos. En dicho caso, podemos utilizar la distribución de una variable aleatoria Bin(20,p) y calcular su esperanza.

Para calcular la esperanza de la Bin(20,p) podemos proceder de diferentes maneras que se detallan en la bibliografía del curso de la asignatura, bien directamente sobre su expresión de la densidad o masa de probabilidad, bien a través del hecho de que las 20 variables independientes de Bernoulli B(p) $\{X_i\}$, si se suman, siguen una binomial Bin(20,p). En este caso, la esperanza y la varianza de esta distribución de sumas de V.A. B(p) será:

$$E\left(\sum_{i=1}^{20} X_i\right) = E(x_1) + \cdots + E(x_{20}) = \boldsymbol{np} = \boldsymbol{20 \cdot p}$$

Sabiendo que la esperanza de una distribución de Bernoulli viene dada por: $E(x_i) = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = p$

De entre los n=20 pacientes con tumor esperamos encontrar "**n·p**" tumores malignos con una varianza (medida de dispersión) que vendrá dada por:

$$Var(X) = Var\left(\left(\sum_{i=1}^{20} X_i\right)\right) = Var(X_i) + \cdots + Var(X_n) = \boldsymbol{n \cdot p \cdot (1 - p)}$$
$$= \boldsymbol{20 \cdot p \cdot (1 - p)}$$

Podemos usar el **teorema de De Moivre-Laplace**, caso particular del **teorema del límite central**, en donde la distribución normal puede ser una aproximación de la distribución binomial. Conformen n→ ∞, esta aproximación es buena si **np ≥ 5** y **n(1-p) ≥ 5**. En particular, el teorema muestra que la función de masa de probabilidad del número aleatorio de

"éxitos" en una serie de n ensayos de Bernoulli independientes, cada uno con probabilidad de éxito p (una distribución binomial con n intentos), converge a la función de densidad de probabilidad de la distribución normal con media $\mu = np = 20 \cdot p$ y desviación estándar $\sigma = \sqrt{np(1-p)} = \sqrt{20 \cdot p(1-p)}$, si n es suficientemente grande y asumiendo que p no es 1 o 0 (Wikipedia: https://es.wikipedia.org/wiki/Teorema_de_De_Moivre-Laplace).

c. Se conoce que un virus genera 4 mutaciones en promedio en un gen cada 12 horas. Se está interesado cuantificar el número de mutaciones diarias que genera el virus.

Un promedio de 4 mutaciones en un gen cada 12 horas, extrapolando linealmente, son 8 mutaciones en 24 horas. 24 horas son 1.440 minutos, unidad de tiempo que utilizaremos. Podemos establecer una tasa de mutación y calcular el parámetro de la distribución de Poisson a utilizar:

$$\lambda = Tasa\ x\ Unidad\ Tiempo = \frac{4\ mutaciones}{720\ minutos} \cdot 1440\ minutos = 8\ mutaciones$$

Teniendo esta tasa en el tiempo de mutaciones, la Variable Aleatoria discreta que modelará la cuantificación diaria será una distribución de Poisson con el parámetro calculado. Esta distribución sirve para modelar el número de veces (0,1,2,3,4,…) que sucede un suceso (mutación, en nuestro caso) en un intervalo de tiempo (24 horas) en base a una tasa. La función de masa de probabilidad viene dada por:

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

Para estimar el número de mutaciones diarias que genera el virus, tomaremos la esperanza de esta distribución de probabilidad, que nos dará una idea de la media esperada, junto con su varianza o desviación típica, que nos indicará la dispersión que podemos esperar:

$$Nº\ mutaciones\ diarias = E(X)$$
$$= \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} \cdot k = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} k = e^{-\lambda} \cdot \lambda \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda = 8$$

Calculamos la varianza:

$Var(X) = E[X^2] - (E[X])^2 = E[X^2] - \lambda^2 = \lambda = 8$ *(Nota: para el cálculo de la varianza he utilizado el programa Maple). Siendo la desviación estándar* $\boldsymbol{\sigma = \sqrt{Var(X)} = 2,828}$*, y que nos dará una idea de las desviaciones entorno a la media esperada.*

d. Se está interesado en conocer la distribución del volumen tumoral en centímetros cúbicos. Se conoce la media y la desviación típica del volumen tumoral

En este caso, podemos suponer que los volúmenes tumorales se distribuirán según una distribución normal, hecho que se suele dar en muchos fenómenos naturales y que, bajo su observancia experimental, los estadísticos llegaron a postular en el Teorema del Límite Central. Sin adentrarnos en la demostración matemática, parece lógico concluir que hay ciertos fenómenos, sobre todo vinculados a procesos naturales, en donde los datos o medidas se distribuirán en torno a una media y con una varianza propias de un modelo de distribución normal, que presenta una simetría entorno a su media, dando cuenta que las desviaciones de ésta, positivas o negativas, son igual de probables y no hay sesgo, siendo la desviación de la media una consecuencia de la deriva o estocasticidad propia de la realidad a modelar.

Conocemos la media μ y la desviación típica σ, por lo que podríamos pensar que la distribución del volumen tumoral seguirá una normal N(μ, σ). Sin embargo, no necesariamente los volúmenes tumorales deben proceder de una normal, pues tan solo conocemos la media y la varianza.

En tanto que el enunciado dice que se conoce la media y la varianza, hemos de suponer que han sido estimadas a partir de datos experimentales en muestreos. Aplicando el Teorema del Límite Central, la media muestral será la media del modelo, y la distribución será una normal con media la de la variable de interés μ, y desviación típica igual: $\frac{\sigma}{\sqrt{n}}$, siendo n el tamaño de

la muestra considerada. Por tanto, el volumen tumoral seguirá una distribución:

$\boldsymbol{V \sim N(\mu, \frac{\sigma}{\sqrt{n}})}$ –> Que corresponde a la distribución del volumen medio de los tumores.

Ejercicio 2

Para estudiar la regulación hormonal de una línea metabólica, inyectan a ratas albinas un fármaco A que inhibe la síntesis de proteínas del organismo. En general, 4 de cada 20 ratas mueren a causa del fármaco antes de que el experimento haya concluido. Si se utiliza un fármaco B el resultado es que mueren 6 de cada 20 ratas a causa del fármaco antes de concluir el experimento. En un nuevo experimento se dispone de 15 ratas inyectadas con un fármaco A y 10 ratas con un fármaco B.

a) **¿Cuál es la probabilidad de que una rata de las 25 se muera antes de concluir el experimento?**

Sea X = {Vivo, Muerto} = {V, M} es espacio muestral dado el experimento. Tenemos, según el enunciado, que:

P_A(X=Muerte)=4/20=2/10=1/5=0,2 → Para 1 rata con el fármaco A en un experimento concreto.

P_B(X=Muerte)=6/20=3/10=0,3→ Para 1 rata con el fármaco B en un experimento concreto.

Vamos a dar **dos soluciones según entendamos la pregunta**, bien que tan **solo muera 1 rata y ninguna más**, o bien que **muera 1 o más ratas**. Calcularemos ambas probabilidades ante mi duda del enunciado.

Calculamos primera la probabilidad de que, de las 25 ratas, solo muera 1 antes de concluir el experimento y ninguna más. Las probabilidades que acabamos de calcular han sido extraídas desde frecuencias observadas en experimentos independientes, en donde a todas las ratas se les suministraba el mismo fármaco. En cada experimento, todas las ratas que morían lo eran a causa del mismo fármaco, siendo que el hecho de morir no venía condicionado a la pertenencia de algún grupo o categoría, es decir, se sabía que la causa era el fármaco suministrado, lo que supone una **información previa**.

Podríamos estar tentados a pensar que, al tener 25 ratas a las que se ha suministrado el fármaco A o B, poseemos dos experimentos independientes, siendo que, si ha de morir solo una rata y nada más que una (supongo que interpreto así el enunciado), bien deba ser el caso que muere 1 sólo por el fármaco A y ninguna por el fármaco B, o bien que muera 1 sólo por el fármaco B y ninguna por el fármaco A. Y estaríamos tentados a plantear la siguiente ecuación:

*P(muere solo 1 rata)= P(muere 1 rata A)*P(no mueren ratas B) + P(no mueren ratas A)*P(muere 1 rata B).*

Pero en este caso, la probabilidad de muerte de una rata está condicionada al fármaco suministrado, como veremos. Por tanto, no estamos ante el producto de 2 distribuciones Binomiales, Bin(0,2 ; 15) y Bin(0,3 ; 10), y no **podemos hacer lo que voy a poner a continuación**:

$$P(muere\ solo\ 1\ A) = \binom{15}{1} \cdot \left(\frac{1}{5}\right) \cdot \left(\frac{4}{5}\right)^{14} = 0.13194$$

$$P(No\ muere\ ninguna\ A) = \left(\frac{4}{5}\right)^{15} = 0.03518$$

$$P(muere\ solo\ 1\ B) = \binom{10}{1} \cdot \left(\frac{3}{10}\right) \cdot \left(\frac{7}{10}\right)^{9} = 0.12106$$

$$P(No\ muere\ ninguna\ B) = \left(\frac{7}{10}\right)^{10} = 0.02825$$

$$\begin{aligned} &P(muere\ 1\ Rata\ solo) \\ &\quad = P(muere\ solo\ 1\ A) \cdot P(No\ muere\ ninguna\ B) \\ &\quad + P(No\ muere\ ninguna\ A) \cdot P(muere\ solo\ 1\ B) \\ &\quad = (0.13194) \cdot (0.02825) + (0.03518) \cdot (0.12106) = 0.007986 \end{aligned}$$

Podemos calcularla de otra manera, que **es la correcta**. Hemos de tener en cuenta que, cuando se observa una muerte en el experimento, no se sabe si es a causa del fármaco A o del fármaco B –tenemos una falta de información con respecto a los experimentos separados-, y que el hecho de "morir una paloma" depende de 2 factores, los tratamientos A y B. La causa de la muerte en sí, si es independiente entre palomas de grupos distintos, a causa de cada fármaco, pero en nuestro nuevo experimento se desconoce la causa cuando actúa el hecho de morir, lo que supone falta de información.

Calculamos la probabilidad de morir en nuestro experimento. Definimos los sucesos: A={ser paloma tratada con fármaco A}, B={ser

paloma tratada con fármaco B} y M={morir durante el experimento}. Desde una teoría conjuntista de la probabilidad, podemos escribir:

$$P(Muerte) = P(A \cap M) + P(B \cap M) = P(A) \cdot P(M|A) + P(B) \cdot P(M|B) = \frac{15}{25} \cdot 0{,}2 + \frac{10}{25} \cdot 0{,}3 = 0{,}24$$

Ya tenemos la probabilidad de que una rata muera en el experimento, bien a causa del fármaco A o del B, es decir, el parámetro de una Binomial de n=25 que utilizaremos para calcular la probabilidad de que sólo muera 1 paloma:

$$P(X = 1) = \binom{25}{1} (0{,}24)^1 \cdot (0{,}76)^{24} = 0.008273275092$$

Ahora respondemos a la cuestión considerando la probabilidad de que muera 1 o más ratas antes de concluir el experimento. Para este caso, vamos a considerar dos distribuciones de probabilidad binomiales:

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{25}{0} (0{,}24)^0 \cdot (0{,}76)^{25} = 0.9989520518$$

b) ¿Cuál es la probabilidad de que mueran 2 ratas tratadas con el fármaco A?

Definimos la variable aleatoria Y: Nº de ratas tratadas con el fármaco A que mueren de un grupo de 15. Debemos observar que, a diferencia que el apartado a, aquí estamos fiando la condición de que debe ser una rata tratada con un fármaco concreto, el A. Sin embargo, en la cuestión del apartado a, ante una muerte, hay una falta de información, la del fármaco causante. En este caso, al tratarse

$$Y \sim B(n = 15, p = 0.2)$$

La función de probabilidad es

$$P(Y = k) = \binom{n}{k} \cdot p^k \cdot q^{n-k} \qquad q = 1 - p$$

La probabilidad que nos piden es

$$P(Y = 2) = \binom{15}{2} \cdot 0.2^2 \cdot 0.8^{13} = 0.2308974418$$

c) ¿Cuál es la probabilidad de que si una rata está muerta haya sido tratada con el fármaco B?

Veremos **dos maneras de razonar la respuesta**, una un poco más larga que la otra, que es **más directa**.

Conviene elaborar la siguiente tabla en base a los datos que nos proporciona el enunciado y mediante los cuales hemos calculado las probabilidades de morir ante un tratamiento determinado (A o B)

Datos del enunciado (experiencia)	TRATAMIENTO A	TRATAMIENTO B	TOTAL
MUEREN	4	6	**10**
VIVEN	16	14	**30**
TOTAL	**20**	**20**	**40**

Definimos los sucesos siguientes: M={Muerte de rata}, A={Ser rata tratada con Fármaco A}, B={Ser rata tratada con fármaco B}

$$P(B|M) = \frac{P(B \cap M)}{P(M)} = \frac{\frac{6}{40}}{\frac{10}{40}} = \frac{6}{10} = 0{,}60$$

En el nuevo experimento, en base a las probabilidades calculadas y los totales de ratas tratadas con fármaco A o B, tendremos la siguiente tabla que nos indicará el desglose por tratamiento y resultado del experimento nuevo, valores esperados en base a las probabilidades que hemos calculado:

$P(E) = \frac{10\ ratas\ tratadas\ con\ B}{25\ ratas\ tratadas\ en\ total} = 0{,}4$ será la probabilidad de tener una rata tratada con el fármaco B.

Datos del Experimento (valores esperados)	TRATAMIENTO A	TRATAMIENTO B	TOTAL
MUEREN	$0{,}2 \cdot 15 = 3$	$0{,}3 \cdot 10 = 3$	**6**
VIVEN	$0{,}8 \cdot 15 = 12$	$0{,}7 \cdot 10 = 7$	**19**
TOTAL	**15**	**10**	**25**

En base a la tabla que nos da las proporciones de ratas de cada tratamiento, así como las que mueren o no, calculadas en función a las probabilidades experimentales del enunciado, tenemos que:

$P(E|C) = \frac{P(E \cap C)}{P(C)} = \frac{\frac{3}{25}}{\frac{6}{25}} = \frac{3}{6} = 0{,}50$. Resultado que puede deducirse directamente de la tabla al observar que, de 6 ratas que esperamos que mueran, 3 son por el tratamiento A y las otras 3 por el B.

Una forma más directa sería:

Para resolver este apartado vamos a emplear el teorema de Bayes:

$$P(B|M) = \frac{P(B \cap M)}{P(M)} = \frac{P(B) \cdot P(M|B)}{P(M)} = \frac{0.4 \cdot 0.3}{0.24} = 0.5$$

d) ¿Cuál es la probabilidad de que se muera sólo una rata tratada con el fármaco A y sólo una rata tratada con el fármaco B?

Para abordar esta cuestión procederemos al cálculo de forma directa. La probabilidad, tratándose de sucesos independientes los tratamientos A y B en ratas, vendrá dada por la probabilidad de que sólo muera 1 rata de tratamiento A multiplicada por la probabilidad de que sólo muera una rata de tratamiento B.

$$P(muere\ sólo\ 1\ rata\ A\ y\ sólo\ 1\ B) = P(X = 1) \cdot P(Y = 1)$$
$$= \binom{15}{1}(0{,}2)^1 \cdot (0{,}8)^{14} \cdot \binom{10}{1}(0{,}3)^1 \cdot (0{,}7)^9 = 0.015973$$

e) Genera 1 muestra aleatoria de 30 ratas tratadas con el fármaco B utilizando la probabilidad de morir del enunciado (6 de cada 20). Estima el % de ratas que se mueren a partir de la muestra y calcula su intervalo de confianza al 95%.
Interpreta los resultados.
Aplica el siguiente código para poder replicar los resultados si los ejecutas varias veces, ya que si no fijas la semilla(seed), cada vez que ejecutes los comandos te dará un resultado diferente.

set.seed (unnumero) #cambia “unnumero” por un número para que siempre genere la misma muestra rbinom(n, size, prob) # genera n muestras de tamaño “size” con una probabilidad de éxito “prob”

Para generar 1 muestra utilizamos R y escribimos los comandos o instrucciones siguientes:

```
set.seed(8813343)
rbinom(1, 30, 0.3)
```

En rbinom(…) estamos indicando que queremos 1 muestra de una Binomial(p=0,3 ; n=30), pues la probabilidad de morir en el tratamiento del fármaco B es, según el enunciado, p=6/20=0,3

Y obtenemos un resultado en R de: 6 ratas muertas tratadas con el fármaco B. Esto supone un porcentaje de muertes en la muestra de:

$\hat{p} = \frac{6}{30} = 0{,}2 \rightarrow 20\%$ de ratas tratadas con el fármaco B de las 30, que han fallecido en la simulación de 1 muestra de la Bin(p=0,3 ; n=30).

Ahora calculamos su intervalo de confianza al nivel del 95% =(1-α)%, siendo α=0,05. Primero calculamos la aproximación al error estándar que se obtiene de esta muestra simulada:

$$S_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0{,}2(1-0{,}2)}{30}} = 0.07302967433$$

Estamos aproximando la binomial por una distribución normal con media p y varianza n·p(1-p). En nuestro caso conocemos p, pues lo hemos utilizado para la simulación de la muestra, pero decidiré basarnos en el estimador obtenido para evaluar la bondad de la simulación y la aplicación del método; en cierta medida, si nos piden un intervalo de confianza al 95% sabiendo el parámetro a estimar, deja de tener sentido otorgar un intervalo de confianza a un parámetro ya conocido. Como valor crítico tomamos aquel punto que, normalizando y estandarizando, nos otorgue un nivel de confianza del 95%, siendo en este caso:

NOTA: Como condiciones para aproximar una Binomial a una normal debe exigirse que:
n≥30 y n·P≥5 y n(1-p)≥5, en nuestro caso es: n=30, n·p=9 y n·(1-p)=21

$$z_{\alpha/2} = 1{,}96$$

Tenemos que:

$$P\left(-1{,}96 \leq \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \leq 1{,}96\right) = 0{,}95$$

$$P\left(\hat{p} - 1{,}96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + 1{,}96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 0{,}95$$

Y el intervalo buscado, al nivel de confianza del 95%, vendrá dado por:

$$\left(\hat{p} - 1{,}96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + 1{,}96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$
$$= (0.2 - 0.143138\,,\, 0.2 + 0.143138\,)$$
$$= (\,0.056862, 0.343138)$$

Con este nivel de confianza del intervalo del 95% podemos calcular la precisión obtenida en nuestra estimación mediante intervalo:

$$Precisión = z_{0.05/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \ 0.143138 \rightarrow 14{,}3\%$$

La estimación del parámetro en la simulación realizada es $\hat{p} = 20\%$ o $\hat{p} = 0{,}200 \ \pm 0{,}143$

Este resultado de la estimación del parámetro p y su intervalo de confianza al 95% calculado viene a decirnos que, cada vez que procedamos de esta manera ante una muestra de una Bin(1; 0,3) , el 95% de los intervalos de confianza calculados de estas muestras contendrán el verdadero valor del parámetro que se desea estimar, p=0,3 en nuestro caso.

f) Si en lugar de 1 muestra de 30 ratas, repites el experimento y generas 200 muestras de 30 ratas tratadas con el fármaco B. Si en cada una de las 200 muestras calculas % de ratas que se mueren y su intervalo de confianza al 95%. ¿Cuántos de los 200 intervalos de confianza esperas que contengan el verdadero porcentaje de ratas muertas del fármaco B?

Según el método de construcción de los intervalos de confianza, al exigir un 95% de confianza, nos dice que, bajo ese procedimiento, de cada 100 muestras esperamos que 95 intervalos contengan el verdadero valor del parámetro a estimar, p en nuestro caso.

Por tanto, si hemos extraído o simulado 200 muestras de una distribución, al calcular intervalos al nivel de 95% de confianza, el 95% de esos 200 intervalos calculados a partir de muestras contendrá el verdadero valor del parámetro a estimar, "p". En nuestro caso será: $200 \cdot 0.95 = 190$ intervalos que contengan el verdadero valor del parámetro "p"

g) Efectúa la simulación e indica cuántos intervalos contienen verdaderamente el % de ratas muertas.

Realizamos la simulación en código R y pasaremos a comentar cada uno de los pasos del código.

set.seed(8813343) Fijamos la semilla.

muestras <- replicate(200,rbinom(1,30,0.3)) Utilizamos la función replicate para generar 200 muestras de una Binomial(30 ; 0,3)

muestras.media <-mean(muestras) Calculamos la media de las 200 muestras.

muestras.media

Resultado de la media: 9.1

hist(muestras,main="Distribución muestral",col="steelblue",
freq=F, breaks=15) Dibujamos el histograma de frecuencias relativas.

lines(density(muestras),col="red") Dibujamos la línea de densidad empírica extrapolada de los 200 datos que nos proporciona esta función de R.

min(muestras) Extraemos el mínimo valor de las muestras: 3

max(muestras) Extraemos el máximo valor de las muestras: 15

sd(muestras) Extraemos la quasi-desviación estándar (divide por n-1 = 199): 2.618066

La siguiente función recibe una lista que contiene las 200 muestras, el tamaño n=30 ratas y el número de muestras tomadas, N=200. La función recorre los valores de la lista de las 200 muestras generadas, la divide por n=30 para tenerlas en porcentaje. Después calcula el error muestral en base al valor del parámetro que pertoca y la precisión según un nivel de significación de 0,05 o nivel de confianza del 95%.

Teniendo la precisión calculada para cada valor del parámetro obtenido en la simulación se calculan los límtes mínimo y máximo del intervalo de confianza al 95% de significación. También, en base a los valores de las 200 muestras iremos calculando la varianza muestral de nuestra simulación.

Una vez hemos calculado el intervalo para un valor del parámetro de las 200 simulaciones, pasamos a mirar si el verdadero valor a estimar, p=0,3, cae dentro de dicho intervalo, si es así, incrementamos una variable que contiene el número de aciertos o intervalos que sí contienen al verdadero valor del parámetro a estimar "p". Por último sacamos pro pantalla resultado para cada muestra generada y los cargamos en un dataframe.

Finalmente la función imprime el número de aciertos, la media de las 200 muestras y su varianza muestral.

```
### Función que estima el parámtero de proporción p y calcula el
# intervalo de confianza para N muestras de tamaño n.
estimacion <- function(muestrass,n,N){
  muestras.media <-mean(muestras)/n
  muestras.varianza <- 0
  aciertos <- 0
  df <- data.frame(pp = double(),
                   intermin = double(),
                   intermax = double(),
                   precisi = double(),
                   dentro = character())

  for (i in 1:N){
    p <- muestrass[i]/n
    errmuestral <- sqrt((p*(1-p)/n))
```

```
    precision <- 1.96*errmuestral
    pmin <- p-precision
    pmax <- p+precision
    interval <- c(pmin,pmax)
    muestras.varianza <- muestras.varianza + (p-muestras.media)**2
    acierta <- (pmin <= 0.3 & 0.3 <=pmax)
    dfprov <- data.frame(pp = p,
                intermin = pmin,
                intermax = pmax,
                precisi = precision,
                dentro = acierta)
    if (acierta){
      aciertos <- aciertos+1
    }

    print (p)
    print(interval)
    print(precision)
    print(acierta)
    df <- rbind(df,dfprov)
   }
   muestras.varianza <- muestras.varianza/N
   print(aciertos)
   print(muestras.media)
   print(muestras.varianza)
   print(df)

   return (list(df,muestras.media,
muestras.varianza,sqrt(muestras.varianza)))
}
```
Final de la función.

Ahora llamamos a la función pasándole las muestras generadas y los parámetros n=30 y N=200.

resultados <- estimacion(muestras,30,200)

resultados En la variable “resultados” obtenemos los cálculos de la función mencionada.

Resultados obtenidos en la simulación efectuada en R:

Media: 0.3033333

Varianza muestral: 0.007577778

Desviación típica muestral: 0.08705043

Nº de aciertos: 192

popo <- as.data.frame(resultados[1]) Cargamos en un dataframe sólo las muestras, sin el resto de resultados, para trabajar sobre ellas.

popo

#str(popo)

mean(popo$pp)

sd(popo$pp) La quasi-desviación típica vale: 0.08726888

#Desviación estándar:

A partir de la quasi-desviación típica calculada por R calculamos la desviación típica y comprobamos que coincide con la calculada por la función

*ssd <-sd(popo$pp)*sqrt(199)/sqrt(200)*

ssd

Desviación típica: 0.08705043

La varianza será la desviación típica al cuadrado.

*ssd**2*

Varianza muestral: 0.007577778

Calculamos la varianza de una Binomial (30 ; 0,3)

var <- 0.3(1-0.3)*30*

var

Resultado de la varianza: 6.3

Frente a este valor modélico calculamos la varianza muestral obtenida en nuestra muestra de 200 ratas:

*sd(muestras/30)**2*

Resultado de la varianza muestral en proporción para el parámetro "p" (dividimos por n=30): 0.007615857

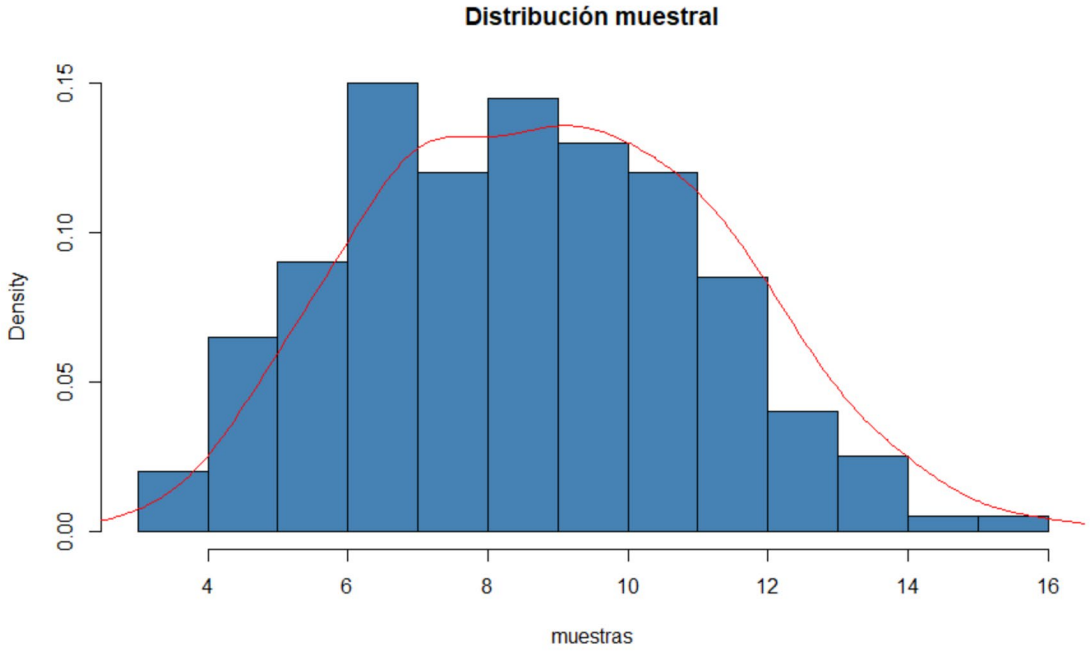
*sd(muestras)**2*

Ahora no dividimos por n=30 y obtenemos la varianza de nuestra muestra simulada en totales: 6.854271 que se aproxima bastante bien a la teórica calculada, así como la media muestral que hemos obtenido de 0.3033333

Dibujamos el histograma de frecuencias relativas de las 200 muestras generadas junto a la función de densidad empírica calculada por R en base a los datos de la simulación. Vemos que, en cierta medida y salvo ligeras desviaciones, la figura recuerda a la de una distribución normal de media 9 y varianza 6,3

```
hist(muestras,main="Distribución muestral",col="steelblue",
    freq=F, breaks=15)
```

lines(density(muestras),col="red")



Ahora vamos a comprar el gráfico de la función de densidad de una distribución de probabilidad normal N(9; 6,3), que es a la distribución a la que tiende una Binomial (30 ; 0,3), con la función de densidad empírica calculada por R en base alos datos de las 200 simulaciones.

En este caso, nuestra función de densidad obtenida de los datos de la simulación experimental, no es más que una aproximación bajo el T.L.C de una suma de N=200 normales $N(\frac{9}{200},\frac{6.3}{200})$ que aproximan 200 muestras binomiales (30 ; 0,3). O estamos sumando binomiales y luego promediando que, según el T.L.C tenderán a una normal.

Tengamos en cuenta que la varianza de 200 binomiales es: n·p(1-p)·N = 30·0,3·(1-0,3)·200, que al promediar en las 200 binomiales quedará como: 6,3.

Tenemos que:

$$\sum_{i=1}^{200} X_i \sim \sum_{i=1}^{200} N_i(\frac{9}{200},\frac{6.3}{200}) \sim N(9,6.3)$$

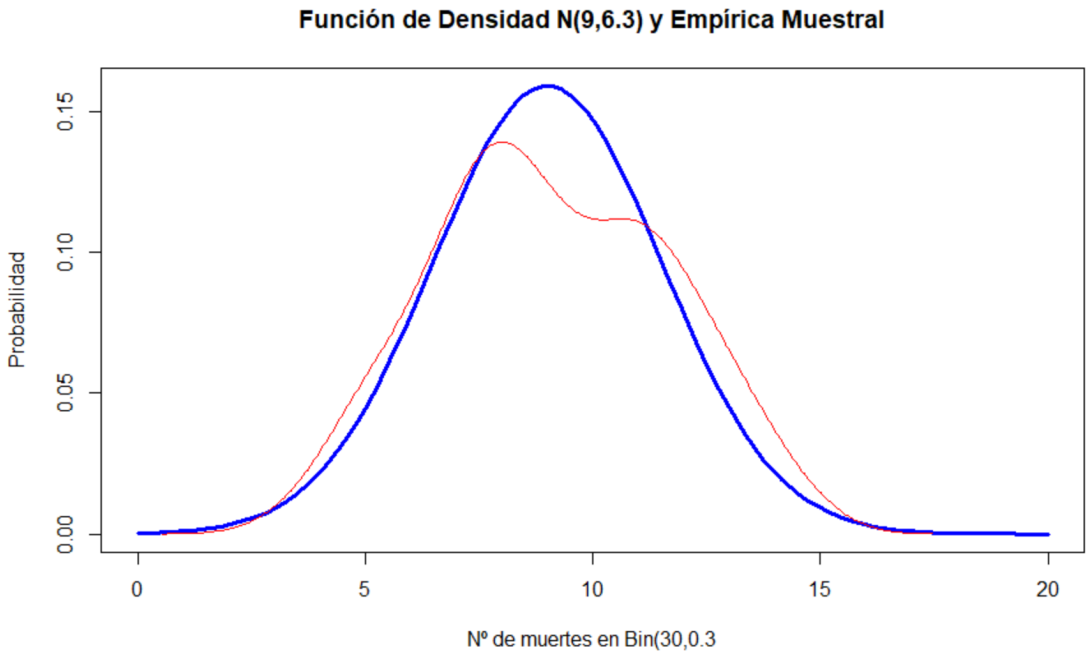
x<-seq(0,20,length=100)

*y<-(1/(sqrt(2*pi)*sqrt(var)))*exp(-(x-(0.3)*30)^2/(2*var))*

plot(x,y,type="l", lwd=3, col="blue", main="Función de Densidad N(9,6.3) y Empírica Muestral",

xlab = "Nº de muertes en Bin(30,0.3", ylab = "Probabilidad")
lines(density(muestras),col="red")

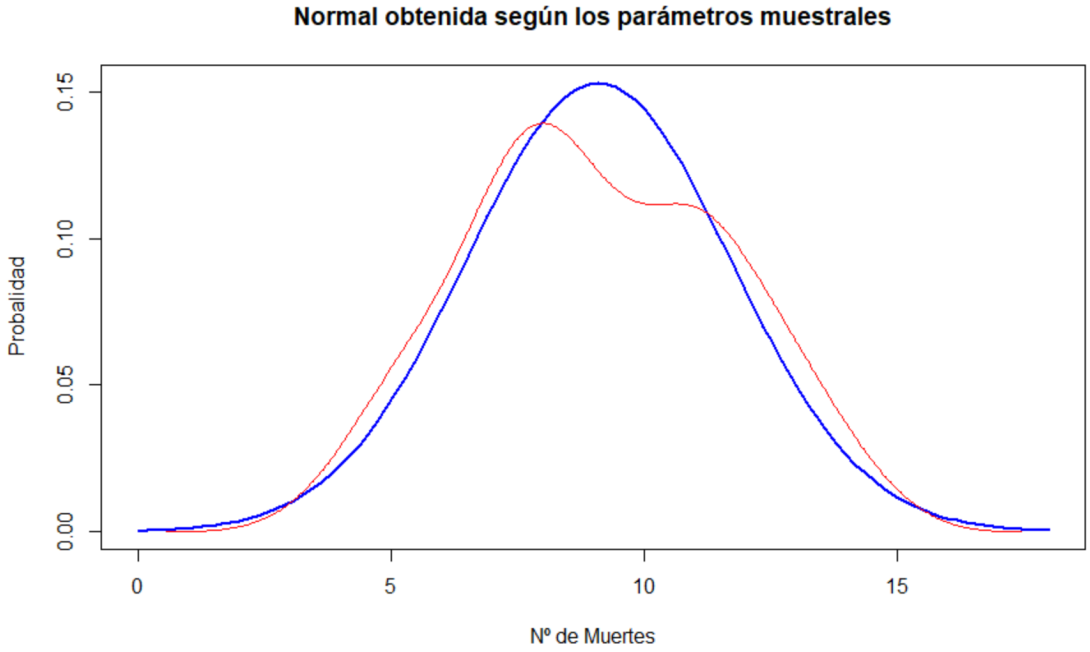
Vemos que la simulación ha generado una distribución muestral empírica que se aproxima bastante bien a la normal esperada según el T.L.C. N(9 ; 6,3):



Ahora graficamos la función de densidad empírica frente a la media y varianza obtenidas en el muestreo de la simulación, no el esperado bajo la tendencia según el T.L.C.: (NOTA, corregimos el hecho de que la función sd en R corresponde a la quasi-desviación típica)

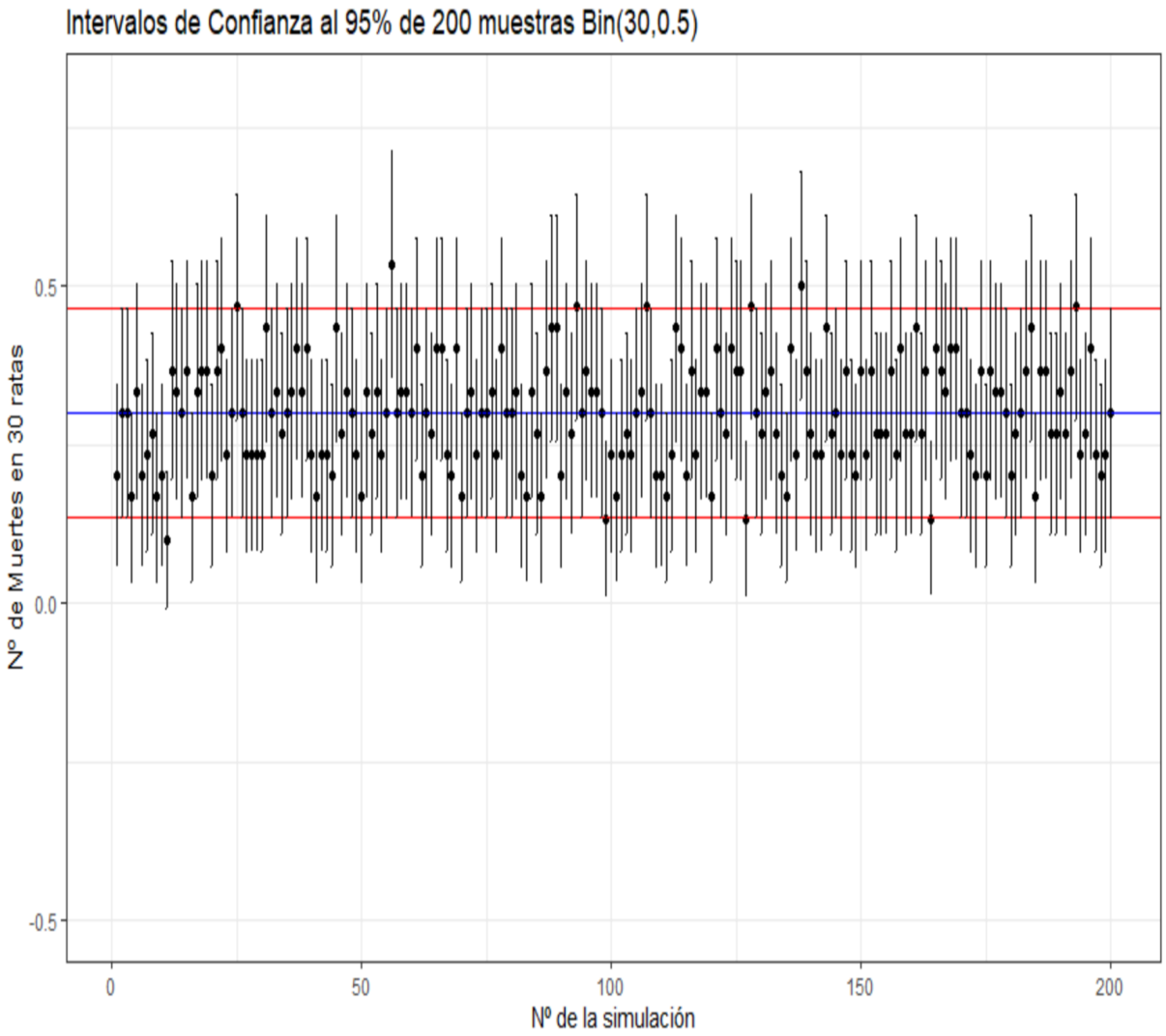
```
x<-seq(0,18,length=100)
f <- dnorm(x, muestras.media, sd(muestras)*sqrt(199)/sqrt(200))

plot(x, f, type = "l", lwd = 2, col = "blue", ylab = "Probalidad",
    xlab = "Nº de Muertes",
    main = "Normal obtenida según los parámetros muestrales")
lines(density(muestras),col="red")
```



Observamos que el ajuste es muy aceptable y que, comparando con la anterior, la distribución muestral empírica se asemeja más a la normal esperada según indica el T.L.C. que no a la Normal con la media y varianza muestrales calculadas procedentes de la simulación.

Ahora realizaremos un gráfico de los 200 intervalos de confianza al 95% calculados en las 200 muestras. A su vez, graficamos la media teórica y el intervalo correspondiente de confianza al 95%. Como hemos visto en los cálculos algorítmicos y podemos ver en el gráfico, se esperaban 190 intervalos que contengan al verdadero valor del parámetro y, en nuestra simulación **hemos obtenido 192 intervalos de confianza al nivel de significación del 95% que sí contienen el verdadero valor del parámetro a estimar**, cosa que concuerda con el porcentaje esperado de los 200 intervalos calculados.



Código en R:

Nota: en popo tenemos los valores estimados del parámetro "p" (pp), el mínimo y máximo del intervalo de confianza, la precisión y una variable booleana que nos dice si el verdadero valor de p=0,3 cae dentro del intervalo en cuestión.

```
#Calculamos los límites del intervalo de confianza al 95% para el verdadero
#valor del parámetro p=0.3
errm <- sqrt((0.3*(1-0.3)/30))
preci <- 1.96*errm
pmini <- p-preci
pmaxi <- p+preci
library(ggplot2)
mm <-c(1:200)
mm
graf1 <- ggplot(data=popo, aes(x=c(1:200), y=pp)) + geom_point() +
  geom_hline(yintercept = 0.3, colour="blue") +
```

```
  geom_hline(yintercept = pmini, colour="red") +
  geom_hline(yintercept = pmaxi, colour="red") +
  ylim(-0.5,0.8)  +  geom_errorbar(aes(ymin=intermin,  ymax=intermax),
width=0.2) +
  xlab("") + ylab("Nº de Muertes en 30 ratas") + xlab("Nº de la simulación")+
  ggtitle("Intervalos de Confianza al 95% de 200 muestras Bin(30,0.5)") +
  theme_bw()
graf1
```

Ejercicio 3

Se supone que la glucemia basal en ayunas en individuos sanos Xs sigue una distribución Normal con media 80mg/100ml y desviación típica 10mg/100ml. En la población diabética la glucemia basal Xd sigue una distribución normal con media 160mg/100ml y una desviación típica de 31.4mg/100ml

a) **¿Qué porcentaje de la población sana tiene un valor de glucemia basal superior a 100mg/100ml en el que se considera a una persona prediabética ?**

Tenemos las siguientes distribuciones de probabilidad para los individuos sanos y diabético:

(NOTA: como todas las unidades implicadas vienen divididas por 100ml, obviaremos este factor divisor para los cálculos, tomando solo la cantidad de mg, luego se divide todo por 100ml)

$$X_s \sim N(80, 10) \quad ; \quad X_d \sim N(160, 31.4)$$

b) **Se considera una persona con diabetes si su glucemia basal está por encima de 200mg/100ml ¿Qué porcentaje de la población diabética no se consideraría diabética siguiendo la consideración anterior?**

Tenemos que:

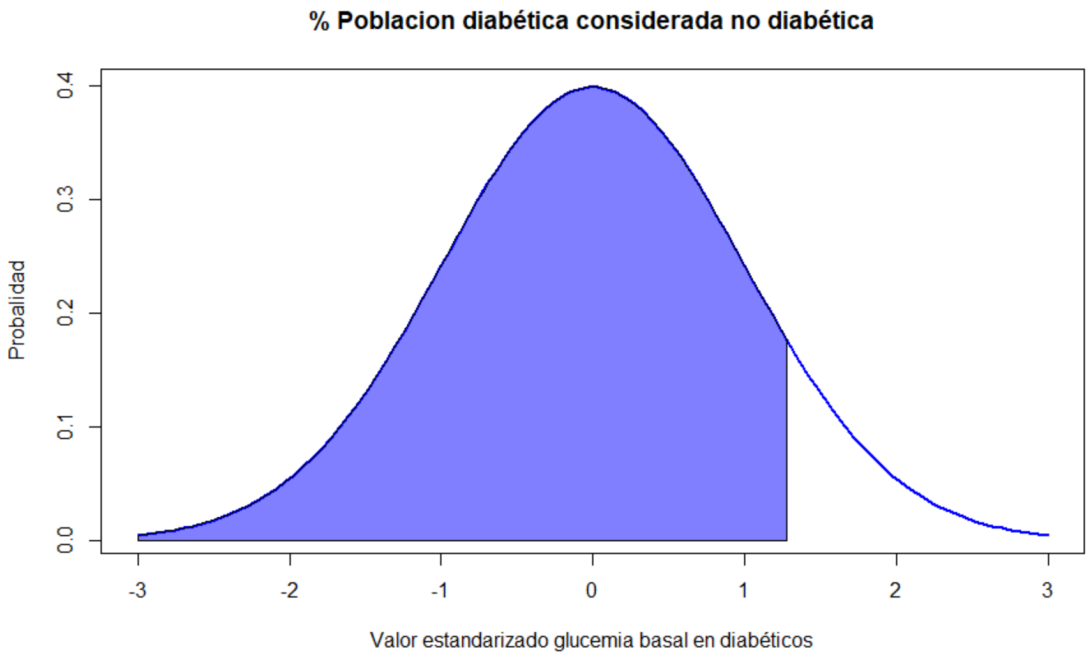
$$X_d \sim N(160, 31.4) \quad ; \quad Z = \frac{X_d - \mu_d}{\sigma_d} \sim N(0, 1)$$

La probabilidad que nos piden es

$$P(X_d < 200) = P\left(Z < \frac{200 - 160}{31.4}\right) = P(Z < 1.2739) = 0.898648$$

En R: pnorm(200, mean = 160, sd = 31.4,lower.tail = TRUE)

Aproximadamente el 89.86% de la población diabética no se consideraría como tal, siguiendo el criterio establecido en el enunciado.



Código en R:

```
x<-seq(-3,3,length=100)
y<-(1/(sqrt(2*pi)))*exp(-(x)^2/2)
plot(x,y,type="l", lwd=3, col="blue", main="Función de Densidad
Normal(0,1)",
    xlab = "Nº de muertes en Bin(30,0.3", ylab = "Probabilidad")
f <- dnorm(x, 0, 1)

plot(x, f, type = "l", lwd = 2, col = "blue", ylab = "Probalidad",
    xlab = "Valor estandarizado glucemia basal en diabéticos",
    main = "% Poblacion diabética considerada no diabética")

### 3.A
lb <- min(x) # Límite inferior
```

```
ub <- 1.2739   # Límite superior
x2 <- seq(lb, ub, length = 80) # Nueva rejilla
yy<-(1/(sqrt(2*pi)))*exp(-x2^2/2)
#y2 <- dnorm(x2, 1, 0) # Densidad
polygon(c(lb, x2, ub), c(0, yy, 0), col = rgb(0, 0, 1, alpha = 0.5))
```

c) Si el 10% de la población es diabética. ¿Qué porcentaje de la población total tiene la glucemia basal por encima de 100mg/100ml?

Definimos los siguientes sucesos:

A: Tener una glucemia basal por encima de 100mg/100ml

D: Ser diabético

A partir del enunciado tenemos los siguientes datos:

$$P(D) = 0.10 \rightarrow P(\overline{D}) = 1 - P(D) = 1 - 0.1 = 0.9$$

La probabilidad de tener una glucemia basal por encima de 100mg/100ml en población sana es:

$$P(A|\overline{D}) = P(X_s > 100) = 0.02275013$$

NOTA. En R:

pnorm(100, mean = 80, sd = 10, lower.tail = FALSE) # 0.02275013

La probabilidad de tener una glucemia basal por encima de 100mg/100ml en población diabética es

$$P(A|D) = P(X_d > 100) = P\left(Z > \frac{100 - 160}{31.4}\right) = P(Z > -1.9108) = 0.9719849$$

En R: *pnorm(-1.9108, mean = 0, sd = 1,lower.tail = FALSE) # 0.9719849*

Pasamos a calcular la probabilidad de tener una glucemia basal por encima de 100mg/100ml para la población total mediante el teorema de la probabilidad total:

$$P(A) = P(A \cap D) + P(A \cap \overline{D}) = P(D) \cdot P(A|D) + P(\overline{D}) \cdot P(A|\overline{D})$$

$$= 0.1 \cdot 0.9719849 + 0.9 \cdot 0.02275013 = 0.1176736$$

Aproximadamente el 11.77% de la población total tiene glucemia basal por encima de 100mg/100ml.

d) Si elegimos a un individuo de toda la población con una glucemia basal superior a 100mg/100ml ¿Qué probabilidad tiene de que realmente sea diabético?

Aplicando el teorema de Bayes tenemos que:

$$P(D|A) = \frac{P(D \cap A)}{P(A)} = \frac{P(D) \cdot P(A|D)}{P(A)} = \frac{0.1 \cdot 0.9719849}{0.1176736} = 0.8260008$$

Lo que significa una probabilidad aproximada del 82,6%

e) Extrae una muestra de 300 diabéticos y estima el valor medio de la glucemia basal. Calcula el intervalo de confianza al 99% e interpreta los resultados. ¿Qué diferencias encontrarías si el tamaño de la muestra fuera de 30 diabéticos.?

Nota.
Aplica el siguiente código para poder replicar los resultados si los ejecutas varias veces,ya que si no fijas la semilla(seed) cada vez que ejecutes los comandos te dará un resultado diferente set.seed (unnumero) cambia "unnumero" por un número para que siempre genere la misma muestra.
Así por ejemplo.
x<-rnorm(n,mean=###, sd=####)
Genera una muestra de tamaño n con las medias y desviaciones típicas por las que sustituyas ####

Vamos a exponer el código en R para responder a la pregunta. En él, veremos 2 funciones diferentes para realizar el mismo cálculo, uno según los preceptos vistos en los apuntes del curso y el otro valiéndonos de las funciones implementadas en R.

Código en R:

Definimos la función **est_media2**. Esta función recibe los parámetros de entrada:
n: tamaño de la muestra a generar
m: media de la distribución normal a emplear en la simulación.
dt: desviación típica de la distribución normal a emplear en la simulación.

La función **est_media2** calculará el intervalo según el método estudiado en la asignatura basado en el **T.L.C** y utilizando el tamaño de la muestra, la desviación típica de la misma y el valor crítico de la distribución normal fijado al 99% de confianza.

Teniendo la función, procederemos a llamarlas bajo los parámetros de la normal indicada a simular, así como el tamaño de la muestra que se desea.

```
#EJERCICIO 3, APARTADO E
Función est_media2:
est_media2 <- function(n,m,dt,nc){
  set.seed(1234)
  muestra<-rnorm(n,m,dt)
  errmuestral <- dt/sqrt(n) Calculamos el error muestral.
  media<-mean(muestra)
  precision <- 2.575*errmuestral Calculamos la precisión multiplicando el
valor crítico por el error muestral
  pmin <- media-precision    Calculamos el mínimo del intervalo.
  pmax <- media+precisión Calculamos el máximo del intervalo.
  amplitud <-pmax-pmin Calculamos la amplitud.
  return(list(muestra,"Media"=media,"Limite          inferior"=pmin,"Limite
superior"=pmax,
          "Amplitud"=amplitud))

}
```

Ahora fijamos los valores de los parámetros de llamada a las funciones según los datos proporcionados por el enunciado.

```
n1=300
n2=30
m=160
dt=31.4
nc=0.99
```

Llamamos a la función citada. Se realizan dos llamadas para n1=300 y n2=30.

res2.1 = est_media2(n=n1,m=m,dt=dt,nc=nc)
res2.2= est_media2(n=n2,m=m,dt=dt,nc=nc)

Sacamos los resultados calculados:

Para n=300:

res2.1

res2.1$Media → 160.4091
res2.1$`Limite inferior` → 155.7409
res2.1$`Limite superior` → 165.0773
res2.1$Amplitud → 9.336331

Para n=30:

res2.2

res2.2$Media → 150.6923
res2.2$`Limite inferior` → 135.9302
res2.2$`Limite superior` → 165.4543
res2.2$Amplitud → 29.52407

Observamos una mayor precisión en la estimación de la media y desviación típica para la muestra de n=300 que para la de n=30, como era de esperar. De otro lado, la precisión en el caso del intervalo calculado para n=300 es mayor, siendo menor la amplitud del intervalo de confianza al 99% que para n=30, que es casi el triple de amplio. Por tanto, hay una mayor precisión en el cálculo del parámetro a estimar, la media de la glucemia basal en diabéticos, para el procedimiento cuando la muestra es mayor, n=300.

Podemos analizar mediante el efecto del tamaño de las muestras simuladas. Tenemos que el Margen de Error viene dado por:

$ME = Z_{0.01/2} \cdot \frac{\sigma}{\sqrt{n}}$ Por tanto, el ME es para:

N=300 → $ME(300) = \frac{80.855}{\sqrt{300}} = 4.668166$

N=30 → $ME(30) = \frac{80.855}{\sqrt{30}} = 14.76204$

$$ME(300) = \frac{ME(30)}{\sqrt{10}} = \frac{ME(30)}{3.162278}$$

El margen de error, y por tanto el intervalo de confianza del estimador al 99%, para n=300 es $\frac{1}{\sqrt{30}}$ $parte$ que el margen de error para n=30.

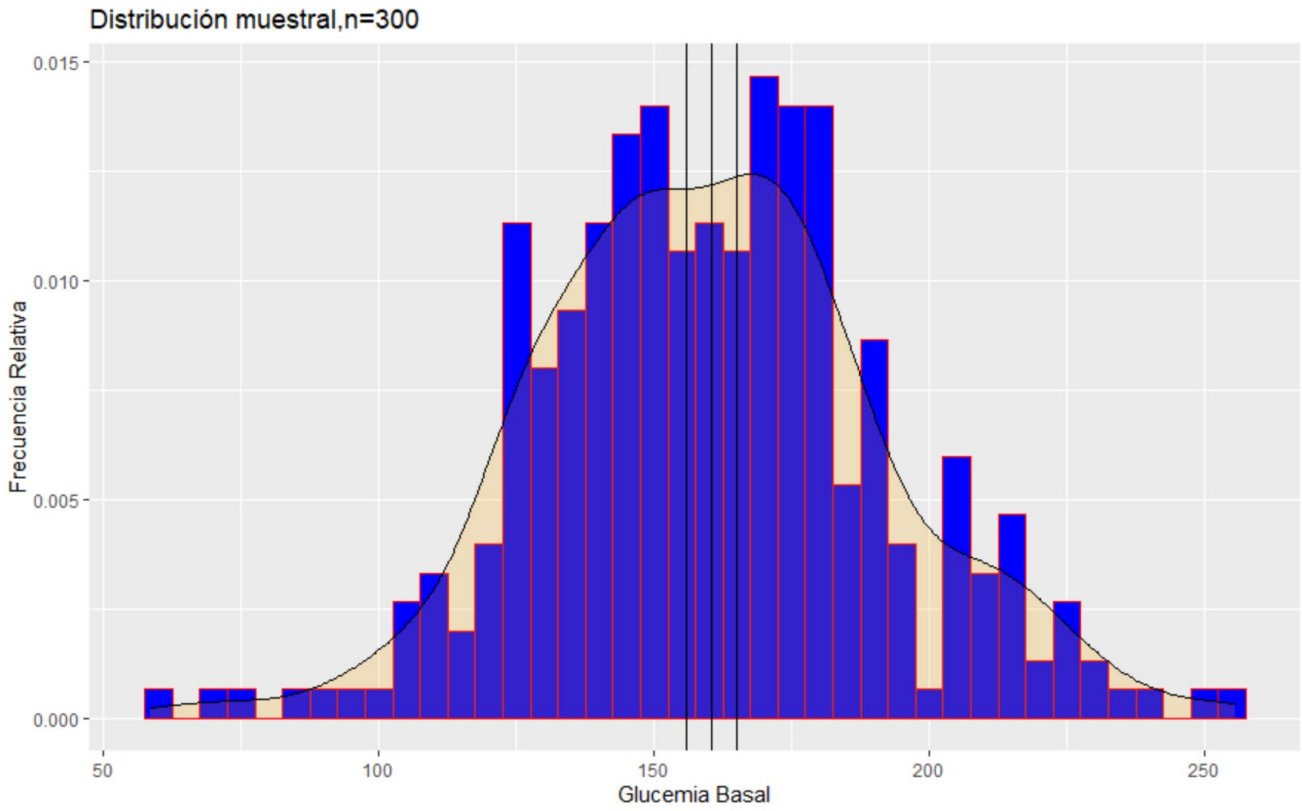
```
me300 <- res2.1$Media
me30 <- res2.2$Media
inf300 <- res2.1$`Limite inferior`
max300 <- res2.1$`Limite superior`
inf30 <- res2.2$`Limite inferior`
max30 <- res2.2$`Limite superior`

res2.1 <- as.data.frame(res2.1[1])
res2.2 <-as.data.frame((res2.2[1]))
colnames(res2.1)[1] <-"dato"
colnames(res2.2)[1] <-"dato"
```

Graficamos los histogramas y las funciones de densidad empírica de las simulaciones con ggplot, así como las líneas que marcan los intervalos de confianza y la estimación:

Código en R:

```
#ggplot
library(ggplot2)
ggplot(data=res2.1,aes(dato)) + geom_histogram( binwidth = 10,fill=I("blue"),col=I("red"),aes(y =..density..))+
  labs(title="Distribución muestral,n=300", x="Glucemia Basal", y="Frecuencia Relativa")+
   geom_density(fill = "orange", alpha = 0.2)+ geom_vline(xintercept = me300)+
 geom_vline(xintercept = inf300) +  geom_vline(xintercept = max300)
```

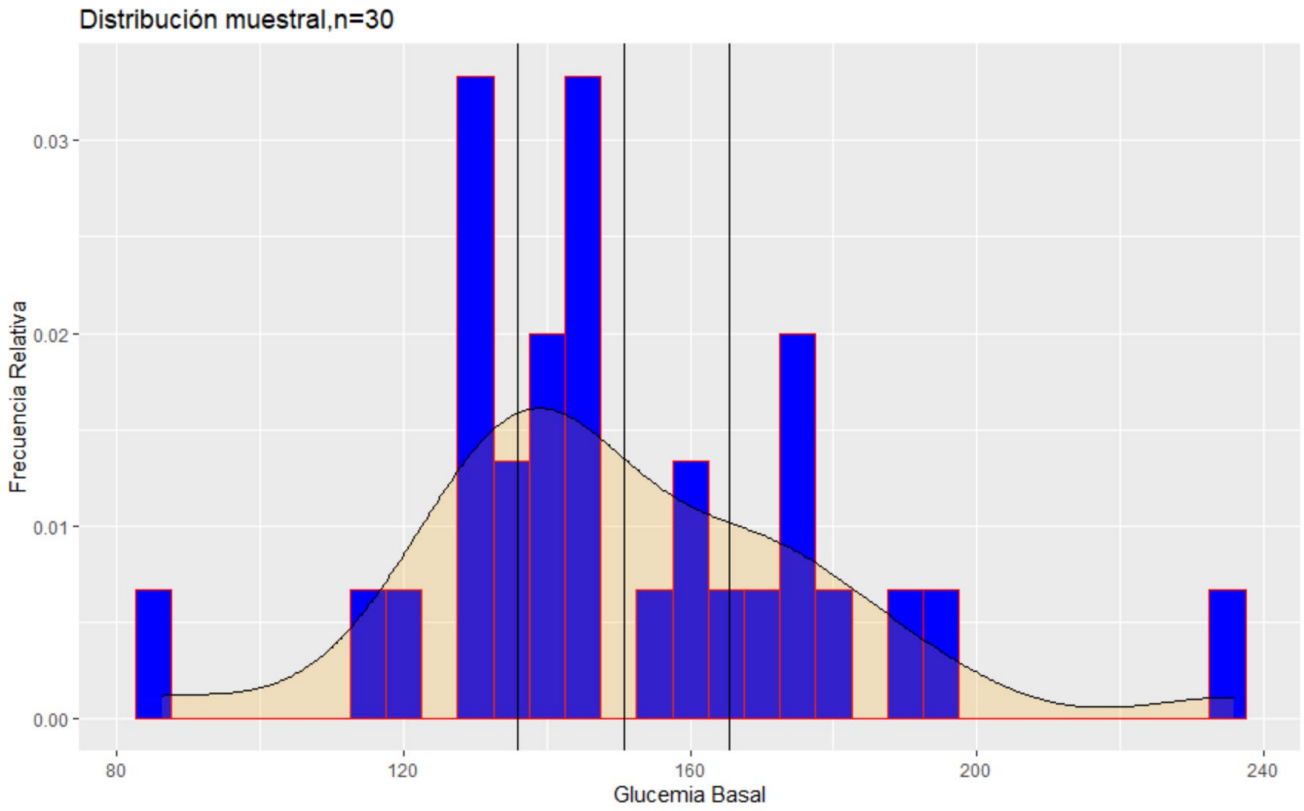


Histograma y curva de densidad para una simulación de 300 sujetos. Se indica en líneas verticales negras el intervalo de confianza calculado.

Para una simulación de 30 individuos tenemos:

Código en R:

```
ggplot(data=res2.2,aes(dato)) + geom_histogram( binwidth = 5,fill=I("blue"),col=I("red"),aes(y =..density..))+
 labs(title="Distribución muestral,n=30", x="Glucemia Basal", y="Frecuencia Relativa")+
 geom_density(fill = "orange", alpha = 0.2)+ geom_vline(xintercept = me30)+
 geom_vline(xintercept = inf30) +  geom_vline(xintercept = max30)
```



Histograma y curva de densidad para una simulación de 30 sujetos. Se indica en líneas verticales negras el intervalo de confianza calculado.

Como hemos razonado en el análisis del Margen de error y su dependencia con $\frac{1}{\sqrt{n}}$, vemos que la precisión para n=300 es mucho mayor y su intervalo menos amplio que para n=30.

BIBLIOGRAFÍA CONSULTADA

- https://stackoverflow.com/questions/22728422/how-do-i-stop-set-seed-after-a-specific-line-of-code
- https://r-coder.com/grafico-densidad-r/
- https://osoramirez.github.io/R_Para_Biologos/distribucion-normal-estandar.html
- https://ggplot2.tidyverse.org/reference/geom_abline.html
- https://es.acervolima.com/convertir-columna-de-dataframe-en-numerico-en-r/#:~:text=La%20conversi%C3%B3n%20se%20puede%20hacer,numeric().
- https://www.rdocumentation.org/packages/ggplot2/versions/0.9.1/topics/geom_vline
- Apuntes asignatura Inferencia Estadística Máster Bioinformática y Bioestadística UB-UOC.
- "Estadística Básica con R." Alfonso García Pérez. UNED, editorial, 2010.