Fundamentos de Probabilidad y Estadística

Edmond Géraud

Probabilidad

Las bases

La probabilidad es una rama de las matemáticas que se ocupa del estudio de la incertidumbre. Se utiliza para medir la posibilidad de que un evento ocurra. En términos generales, la probabilidad se define como el cociente entre el número de eventos favorables y el número total de posibles resultados.

La probabilidad se puede calcular mediante la fórmula:

$$P(A) = \frac{n(A)}{n(S)}$$

donde $P(A)$ es la probabilidad del evento A, $n(A)$ es el número de resultados favorables para el evento A y $n(S)$ es el número total de resultados posibles.

La probabilidad se mide en una escala que va desde 0 hasta 1. Un evento con una probabilidad de 0 nunca sucederá, mientras que un evento con una probabilidad de 1 siempre sucederá.

La probabilidad también se puede expresar como un porcentaje o una fracción. Por ejemplo, si la probabilidad de que un equipo de fútbol gane un partido es de 0.75, entonces la probabilidad expresada como porcentaje es del 75 $\backslash\%$ y la probabilidad expresada como fracción es de $3/4$.

Existen dos tipos de eventos mutuamente excluyentes: eventos independientes y eventos dependientes. Los eventos independientes son aquellos que no están influenciados por la ocurrencia o no de otros eventos, mientras que los eventos dependientes son aquellos que están influenciados por otros eventos.

La probabilidad de la intersección de dos eventos independientes se calcula mediante la fórmula:

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

Mientras que la probabilidad de la intersección de dos eventos dependientes se calcula mediante la fórmula:

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

donde $P(B|A)$ es la probabilidad condicional de que el evento B ocurra dado que el evento A ya ha ocurrido.

La probabilidad condicional también se puede calcular mediante la fórmula:

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

donde $P(A \cap B)$ es la probabilidad de la intersección de los eventos A y B y $P(A)$ es la probabilidad del evento A.

La probabilidad también se puede calcular mediante la distribución de probabilidad, que es una función que asigna a cada posible valor de una variable aleatoria una probabilidad. La distribución de probabilidad se utiliza para calcular la probabilidad de que una variable aleatoria tome un valor específico o un conjunto de valores.

Las distribuciones

En términos simples, una distribución es una manera de mostrar cómo se reparten los diferentes valores posibles que puede tomar una variable aleatoria. Es decir, una distribución nos permite ver cuántas veces ocurre un valor determinado y cuál es la probabilidad de que ocurra en relación a otros valores posibles.

Por ejemplo, si estamos considerando la variable aleatoria "edad de los estudiantes de una clase", una distribución nos permitiría ver cuántos estudiantes tienen 18 años, cuántos tienen 19 años, cuántos tienen 20 años, etc. Así, la distribución nos muestra cómo se distribuyen los valores posibles y cuál es la probabilidad de que ocurra un valor determinado.

En estadística, una distribución es una función matemática que describe la probabilidad de ocurrencia de cada uno de los posibles valores que puede tomar una variable aleatoria. En otras palabras, una distribución es una forma de representar cómo se distribuyen los valores posibles de una variable aleatoria.

Existen diferentes tipos de distribuciones, cada una de las cuales tiene propiedades y características específicas. Algunos ejemplos comunes de distribuciones incluyen la distribución normal, la distribución binomial, la distribución de Poisson, la distribución exponencial y la distribución uniforme.

La distribución normal es una de las distribuciones más importantes y se utiliza con frecuencia en estadística debido a que muchos fenómenos naturales y sociales se distribuyen de forma aproximadamente normal. Esta distribución se caracteriza por su forma de campana y está completamente determinada por su media y su desviación estándar.

La distribución binomial se utiliza para modelar situaciones en las que una variable aleatoria puede tomar solo dos posibles valores, como por ejemplo éxito o fracaso. Esta distribución se caracteriza por su media y su desviación estándar, que dependen del número de ensayos y la probabilidad de éxito en cada ensayo.

La distribución de Poisson se utiliza para modelar situaciones en las que se está interesado en el número de eventos que ocurren en un intervalo de tiempo o espacio determinado. Esta distribución se caracteriza por su parámetro lambda, que es igual a la media y a la varianza de la distribución.

La distribución exponencial se utiliza para modelar el tiempo que transcurre entre dos eventos consecutivos de un proceso aleatorio. Esta distribución se caracteriza por su parámetro lambda, que determina la tasa a la que ocurren los eventos.

La distribución uniforme se utiliza para modelar situaciones en las que cada valor posible de una variable aleatoria tiene la misma probabilidad de ocurrencia. Esta distribución se caracteriza por su rango, que es la diferencia entre el valor máximo y el valor mínimo que puede tomar la variable aleatoria.

Cada distribución tiene sus propias propiedades y características, y se utilizan para modelar diferentes tipos de fenómenos y situaciones. La elección de la distribución adecuada depende del tipo de datos que se están analizando y del objetivo del análisis estadístico.

Las distribuciones discretas y continuas se diferencian en la naturaleza de las variables que representan y cómo se pueden representar.

Una distribución discreta se refiere a una variable aleatoria que solo puede tomar valores enteros o contables. Por ejemplo, el número de hijos en una familia o el número de veces que se gana una partida en un juego. En este tipo de distribución, la probabilidad de cada valor se puede calcular individualmente. La distribución se puede representar mediante una función de probabilidad discreta, que asigna una probabilidad a cada valor posible de la variable aleatoria.

Por otro lado, una distribución continua se refiere a una variable aleatoria que puede tomar cualquier valor en un intervalo determinado. Por ejemplo, la altura de las personas o la velocidad del viento. En este tipo de distribución, la probabilidad de cualquier valor individual es cero, ya que hay infinitos valores posibles en un intervalo continuo. En cambio, la probabilidad se mide mediante la densidad de probabilidad continua, que indica la probabilidad de que la variable aleatoria caiga dentro de un intervalo específico.

Otra diferencia importante es que las distribuciones discretas se representan mediante una función de probabilidad discreta, mientras que las distribuciones continuas se representan mediante una función de densidad de probabilidad continua.

Además, las distribuciones discretas tienen una serie de valores aislados que se pueden contar, mientras que las distribuciones continuas se caracterizan por tener una distribución suave y continua en todo el rango de valores posibles.

Tanto las distribuciones discretas como las continuas tienen una función de densidad de probabilidad (en el caso continuo) o de probabilidad (en el caso discreto) y una función de distribución acumulativa.

En el caso de una distribución continua, la función de densidad de probabilidad f(x) está definida de tal manera que la probabilidad de que la variable aleatoria X caiga en el intervalo [a,b] es la integral de f(x) en el intervalo [a,b]:

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

Mientras tanto, la función de distribución acumulativa $F(x)$ para una distribución continua se define como la integral de $f(x)$ desde menos infinito hasta x:

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

En el caso de una distribución discreta, la función de probabilidad $p(x)$ asigna la probabilidad a cada valor posible de la variable aleatoria X, y la probabilidad de que la variable aleatoria X tome un valor en el conjunto A se define como la suma de las probabilidades de todos los elementos de A:

$P(X \in A) = \sum_{x \in A} p(x)$

Mientras tanto, la función de distribución acumulativa $F(x)$ para una distribución discreta se define como la probabilidad acumulada de que la variable aleatoria X tome un valor menor o igual a x:

$F(x) = P(X \leq x) = \sum_{k \leq x} p(k)$

Para una distribución continua:

La función de densidad de probabilidad (FDP) está relacionada con la función de distribución acumulativa (FDA) de la siguiente manera:

- FDP es la derivada de la FDA:

  $f(x) = \frac{dF(x)}{dx}$
- FDA se puede obtener integrando la FDP:

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

Para una distribución discreta:

La función de distribución acumulativa (FDA) se define como la probabilidad acumulada de que la variable aleatoria X tome un valor menor o igual a x:

$F(x) = P(X \leq x) = \sum_{k \leq x} p(k)$

donde p(k) es la probabilidad de que la variable aleatoria X tome el valor k.

La relación entre la función de probabilidad (FP) y la función de distribución acumulativa (FDA) es que la FP se puede obtener a partir de la FDA como la diferencia entre dos valores consecutivos de la FDA:

$p(x) = F(x) - F(x-1)$

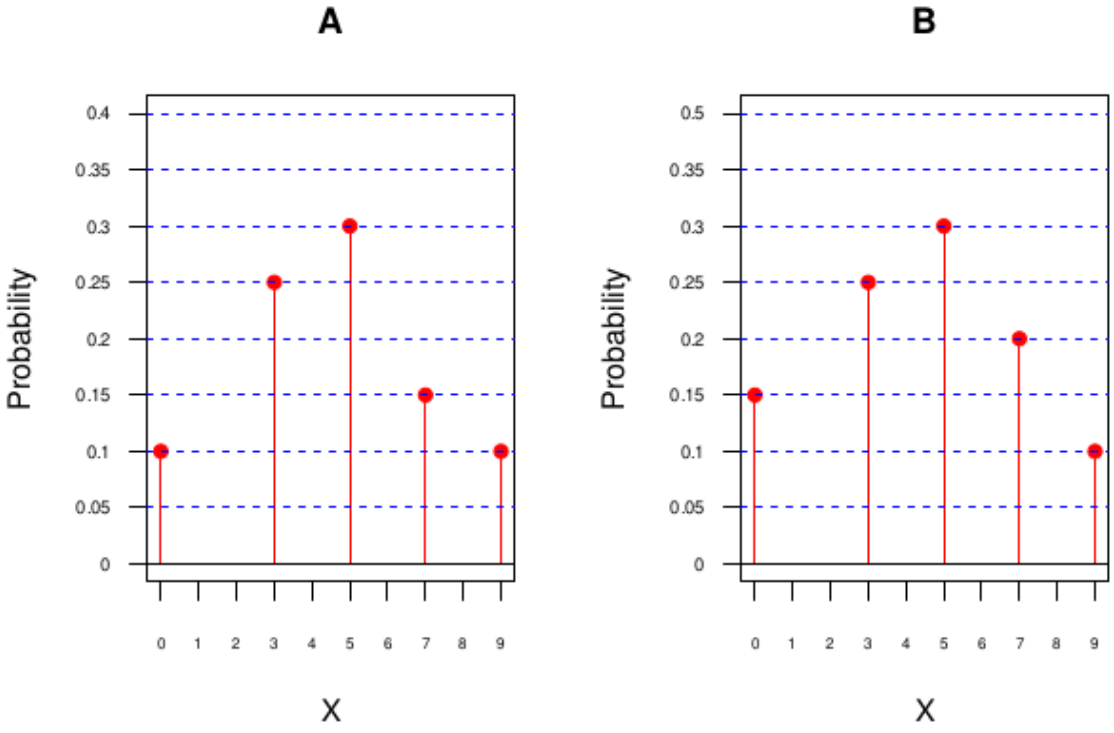
Por otro lado, la FDA se puede obtener a partir de la función de probabilidad acumulada (FPA):

$F(x) = \sum_{k=-\infty}^{x} p(k)$

Ejercicios Probabilidades

Ejercicio 1

Las probabilidades de una variable discreta X están representadas mediante uno de los siguientes gráficos:



Sólo uno de estos gráficos es posible, identifica Cuál es y explica porqué.

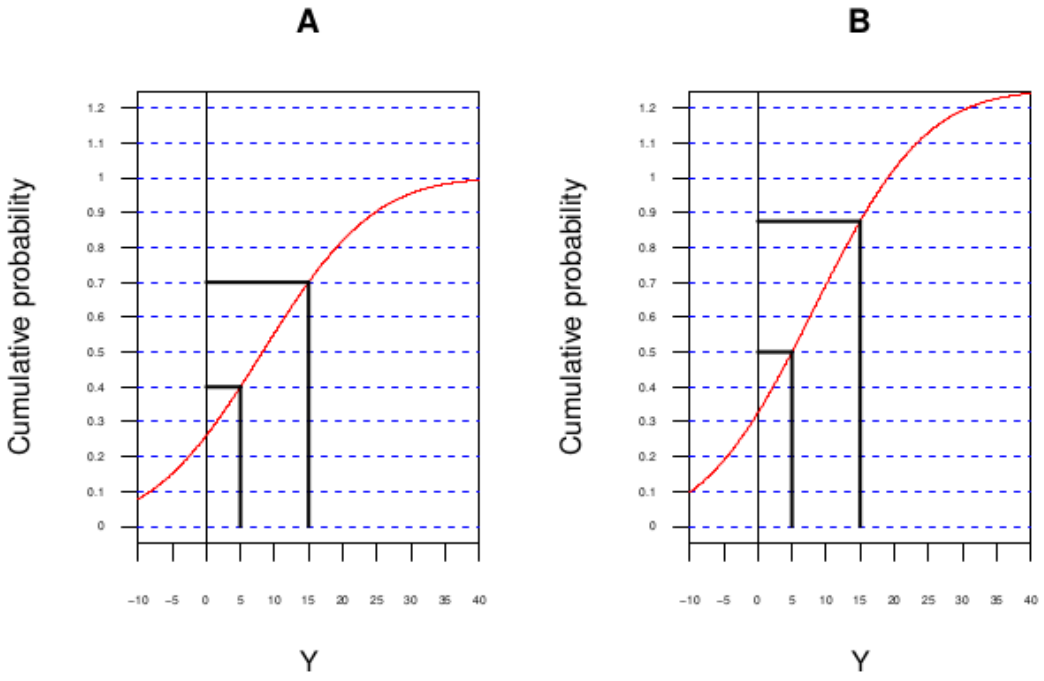
a) $P(X = 7)$

b) $P(X \geq 5)$

c) $P(5 \leq X < 8)$

d) $P(X = 7|X \geq 5)$

b) Uno de los siguientes gráficos representa la función de distribución de una variable continua Y :



Sólo uno de estos gráficos es posible, identifica cuál es y explica porqué.

a) $P(Y = 0)$

c) $P(5 < Y < 15)$

d) $P(Y > 5|Y < 15)$

Ejercicio 2

Dos tratamientos A y B fracasan en curar una determinada enfermedad en el 20% y 30% de los casos, respectivamente. Suponiendo que ambos actúan de modo independiente.

Según el enunciado:

$P(fracaso/TratammientoA) = 0.20portantoP(curar/TratamientoA) = P(C/A) = 0.80$

Según el enunciado :

$P(fracaso/TratammientoB) = 0.30portantoP(curar/TratamientoB) = P(C/B) = 0.70$

a) ¿Cuál es la probabilidad de curar si se aplican ambos tratamientos a la vez?

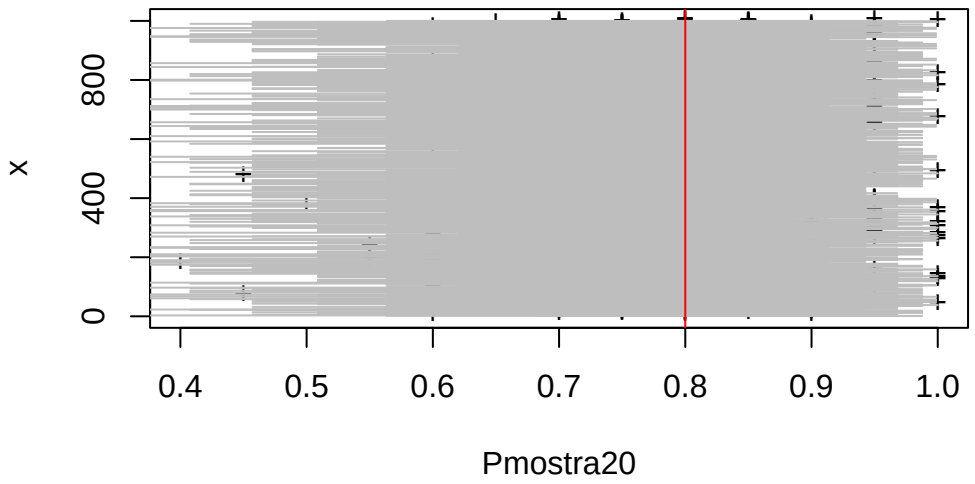
b) ¿Cuál es la probabilidad de curar si se aplica primero el tratamiento A, y el tratamiento B se aplica sólo a los individuos que fallaron el tratamiento A?

c) ¿Cuál es la probabilidad de curarse en esta situación?

d)¿Cuál es la probabilidad de que si un sujeto este curado sea debido altratamiento A?

e) Cual es la probabilidad de que se curen 16 o más sujetos?

```
set.seed(666666666)
mostra20<-rbinom(1000,20,.80)
# Genera las 1000 muestras de tamaño 20 con probabilidad de exito .80 y
Pmostra20<- mostra20/20
# Calcula la proporción de éxitos en cada muestra
eePmostra20<-sqrt(Pmostra20*(1-Pmostra20)/20) # Calcula el error estàndar de cada muestra
liPmostra20<-NULL
# Calcula el límite inferior de cada intervalo en un bucle a partir del test binomial
for (i in 1:length(mostra20)) {
liPmostra20[i] <- binom.test(mostra20[i], 20)$conf.int[1]
}
lsPmostra20<-NULL
# Calcula el límit superior de cada intervalo en un bucle a partir del test binomial
for (i in 1:length(mostra20)) {
lsPmostra20[i] <- binom.test(mostra20[i], 20)$conf.int[2]
}
x<-1:length(mostra20) # Genera un vector índice del 1 al número de intervalos
plot(Pmostra20,x,pch="+") # Dibuja los ejes
segments(liPmostra20,x,lsPmostra20,x,col="grey") # Dibuja cada intervalo
abline(v= .80,col="red")
```



g) Cuantos intervalos esperas que contengan y cuantos verdaderamente contienen el verdadero porciento de curación del tratamiento A

```
tab20<-table(liPmostra20<0.80 & lsPmostra20>0.80) # Genera una tabla de frecuencias
prop.table(tab20) # Calcula porcentages de la tabla anterior
```

```
FALSE  TRUE
0.024 0.976
```

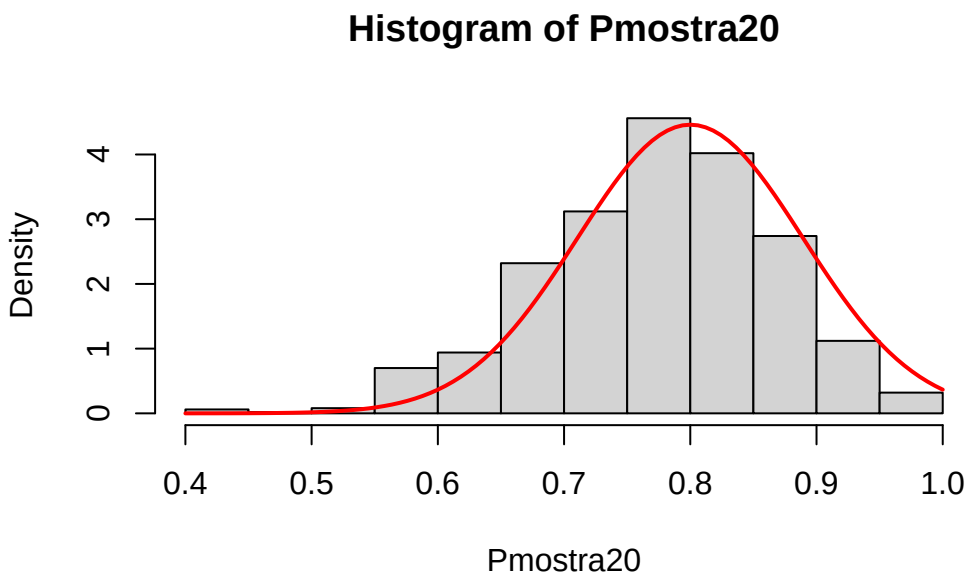
h) Dibuja un histograma con los valores observados y superpón la curva normal teórica que se esperaría según el teorema central del límite.

Para dibujar el histograma sólo hay que pedírselo al R. La línea normal hay que hacerla a mano. Primero se genera un vector entre el valor mínimo y màximo de las proporciones observadas. Según el Teorema central de límite (TCL)las proporciones observadas se han de comportar con una normal con media la proporción poblacional $p = .80$ y desviación típica el error estándar calculado a partir del verdadero valor. . Calculamos el error y por cada proporción calculamos el valor de la densidad de la normal con la función dnorm. Se observa en este caso que al efectuar la aproximación normal obtenemos valores mayores de 1 lo que es imposible y no deberiamos utilizarla.

```
eePop=sqrt(.80*(1-.80)/20) # Calculamos el error estàndar de la normal en la población seg
cat("Se calcula para cada proporción estimada cual sería el valor de la densidad normal se
```

```
Se calcula para cada proporción estimada cual sería el valor de la densidad normal según el
```

```
## Se calcula para cada proporción estimada cual sería el valor de la densidad normal segú
xfit<-seq(min(Pmostra20),max(Pmostra20),length=100) # Genera una secuencia de 1000 valores
xnorm<-dnorm(xfit,.80,eePop) # Genera los valores de la normal del TCL
hist(Pmostra20,freq=FALSE) # Dibuja el histograma
lines(xfit,xnorm,col="red",lwd=2) # Superpone la normal
```

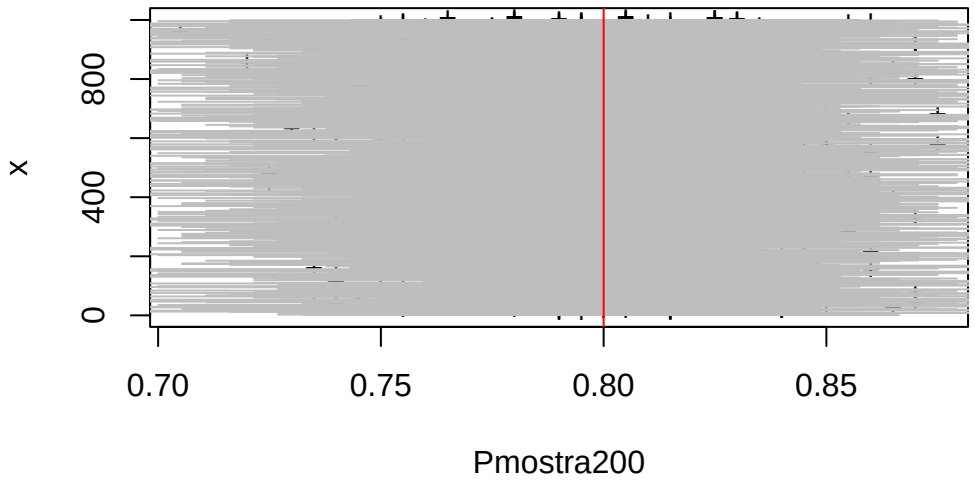


j) Si en lugar de ser la muestra de 20 fuera de 200 que esperarias que ocurriera con respecto a los apartados anteriores (no es necesario que presentes los cálculos)

Esta pregunta se puede responder sin efecgtuar cálculos. Esperamos que los intervalos sean más precisos, pero el número de intervalos esperados que contentan el verdadero valor sera el 95% y el observado no deberia ser mucho menor. En este caso observaremos como la aproximación normal es mejor y no tenemos el problema de tener valores por encima de 100.

```
set.seed(666666666)
mostra200<-rbinom(1000,200,.80)
# Genera las 1000 muestras de tamaño 200 con probabilidad de exito .80
Pmostra200<- mostra200/200
# Calcula la proporción de éxitos en cada muestra
eePmostra200<-sqrt(Pmostra200*(1-Pmostra200)/200) # Calcula el error estàndar de cada mues
liPmostra200<-NULL
# Calcula el límite inferior de cada intervalo en un bucle a partir del test binomial
for (i in 1:length(mostra200)) {
liPmostra200[i] <- binom.test(mostra200[i], 200)$conf.int[1]
}
lsPmostra200<-NULL
# Calcula el límit superior de cada intervalo en un bucle a partir del test binomial
for (i in 1:length(mostra200)) {
lsPmostra200[i] <- binom.test(mostra200[i], 200)$conf.int[2]
}
x<-1:length(mostra200) # Genera un vector índice del 1 al número de intervalos
```

```r
plot(Pmostra200,x,pch="+") # Dibuja los ejes
segments(liPmostra200,x,lsPmostra200,x,col="grey")
abline(v= .80,col="red")
```



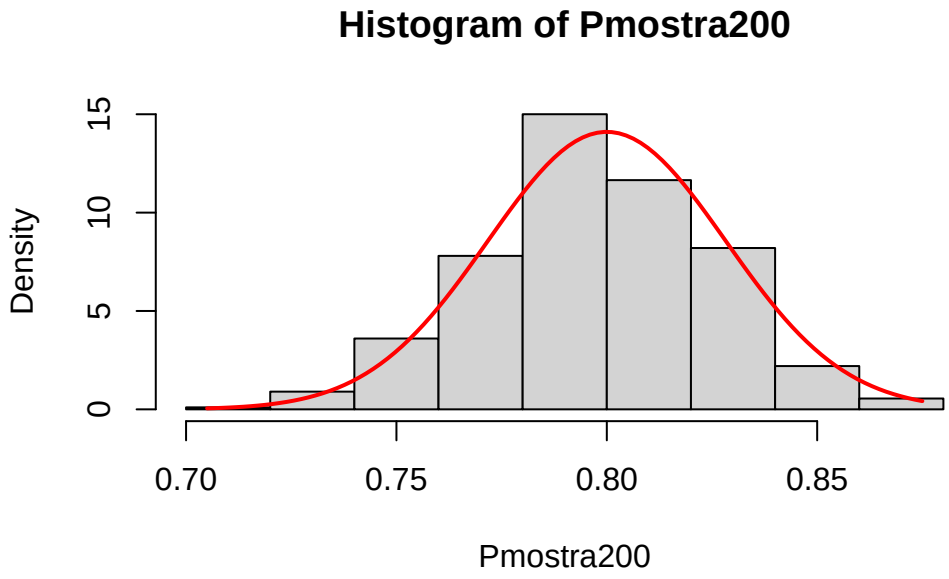

```r
tab200<-table(liPmostra200<0.80 & lsPmostra200>0.80) # Genera una tabla de frecuencias
prop.table(tab200) # Calcula porcentajes de la tabla anterio
```

```
FALSE  TRUE
0.038 0.962
```

```r
eePop=sqrt(.80*(1-.80)/200) # Calculamos el error estàndar de la normal en la población se
cat("Se calcula para cada proporción estimada cual sería el valor de la densidad normal se
```

```
Se calcula para cada proporción estimada cual sería el valor de la densidad normal según el
```

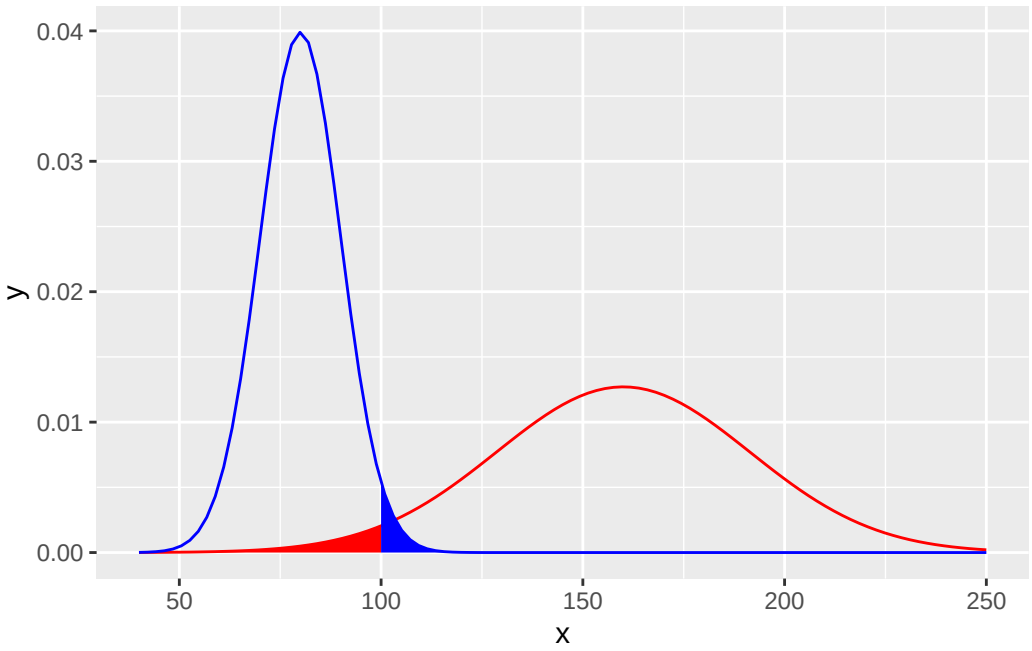
```r
## Se calcula para cada proporción estimada cual sería el valor de la densidad normal segú
xfit<-seq(min(Pmostra200),max(Pmostra200),length=100) # Genera una secuencia de 1000 valor
xnorm<-dnorm(xfit,.80,eePop) # Genera los valores de la normal del TCL
hist(Pmostra200,freq=FALSE) # Dibuja el histograma
lines(xfit,xnorm,col="red",lwd=2) # Superpone la normal
```



Ejercicio 3

Se supone que la glucemia basal en individuos sanos,Xs sigue una distribución $N(\mu = 80, \sigma = 10)$, mientras que en los diabéticos Xd , sigue una distribución $N(\mu = 160, \sigma = 31.4)$ En la gráfica teneis en azul la distribución de la glucemia basal de los indiduos sanos y en rojo la de los diabéticos.

```
library(ggplot2)
ggplot(data.frame(x = c(40, 250)), aes(x)) +
stat_function(fun = dnorm, args = list(mean = 160, sd = 31.4), col='red') +
stat_function(fun = dnorm, args = list(mean =160, sd =31.4), xlim = c(40, 100),
geom = "area", fill = "red")+
stat_function(fun = dnorm, args = list(mean = 80, sd = 10), col='blue') +
stat_function(fun = dnorm, args = list(mean= 80, sd =10), xlim = c(100,250),
geom = "area", fill = "blue")
```



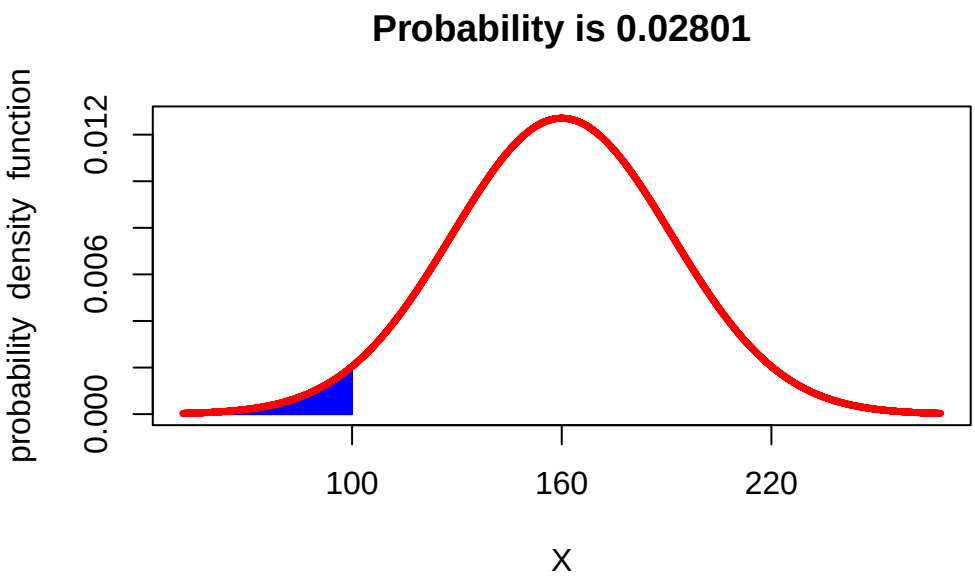
a) Si se conviene en classificar a los individuos como diabéticos por valores de la glucemia basal superiores a 100. ¿Qué porcentaje de la población se clasificaría incorrectamente como sano?

Se conviene en clasificar a diabéticos con valores por encima de 100, así aquellos diabéticos que tengan un valor inferior a 100 seràn clasificados incorrectamente como sanos. Se trata del área roja de la gráfica

```
library(fastGraph)# Esta libreria permite dibujar distribuciones
p_errorsano<- pnorm(100,mean=160, sd=31.4,lower.tail=TRUE)
cat("La probabilidad de ser clasificado incorrectamente como sano es ",p_errorsano)
```

```
La probabilidad de ser clasificado incorrectamente como sano es  0.02801334
```

```
## La probabilidad de ser clasificado incorrectamente como sano es0.02801334
shadeDist( 100, "dnorm",
160, 31.4, lower.tail=TRUE, col=c("red","blue") )
```



b) ¿Qué porcentaje se clasificaria como diabético y en realidad está sano.

d) Genera una muestra de 20 pacientes de la población diabética, Estima la media de la muestra con su intervalo de confianza al 95%. Interpreta los resultados

Enunciados

Indica que tipo de análisis o que pruebas estadísticas que utilizarías en cada uno de los y si fuera necesario algún tipo de prueba adicional para llevar a cabo el análisis. Formula la hipótesis a contrastar de acuerdo con las hipótesis seleccionadas

a) Se está interesado en conocer que factores están asociados al riesgo nutricional en ancianos, medidos mediante una escala que clasifica a los ancianos de 1 a 20, donde 20 es un alto riesgo y 1 bajo riesgo. Como posibles factores entre otros se mide el ser hombre o mujer, el comer sólo o acompañado y el beber o no vino en las comidas. Se dispone de 150 ancianos para resolver la situación. Se desea contratar si por separado son factores de riesgo de malnutrición.

b) Se dispone de 45 mujeres obesas de las cuales 22 eran premenopáusicas y 23 postmenopáusicas. Se tomaron 3 medidas antropométricas como el índice de masa corporal, el índice cintura-cadera y cintura-muslo. Además se clasificó a las mujeres según su obesidad fuera androide o ginoide. Se está interesado en evaluar la relación entre la menopausia y cada uno de los índices de obesidad por separado. Asì como si el tipo de obesidad se asocia con tener o no la menopausia.

c) Se efectuó un ensayo aleatorio doble ciego para evaluar el efecto del inhibidor de la enzima conversiva de la angiotensina, trandolapril, sobre la neuropatía diabética, asignando a 23 pacientes el trandolapril y a otros 23 a un placebo. Como respuesta entre otras variables se midió la velocidad de conducción motora del nervio del perineo(m/s). Indica el análisis para comparar el nivel de dicha variable al finalizar el estudio.

d). Para estudiar el posible efecto beneficioso de la meditación trascendental para el tratamiento de pacientes con síntomas de ansiedad, se decide administrar a un grupo de 20 pacientes un protocolo de este tipo de actividad para ser realizado diariamente. Se mide el nivel de ansiedad de cada individuo al inicio del tratamiento y a los seis meses, a través de un test que proporciona una puntuación cuantitativa entre 0 y 100 puntos como valoración de dicho nivel de ansiedad. Se quiere comparar los niveles de ansiedad entre el inicio y a los seis meses para ver si se han producido cambios importantes.

e) Se estudia un grupo de 33 pacientes afectados de Carcinoma hepatocelular, un grupo de 22 afectos únicamente de cirrosis y un grupo control de 31 donantes desangre. Se determina la actividad celular NK frente a la línea celular K562 marcada con CR mediante la prueba de corta duración de citotoxicidad y el número de células CK en los tres grupos. (Nota la media de actividad celular en los hombres es de 39 unidades líticas /10^7 de linfocitos y la mediana es de 28 y la media del número de células es de 178 y la mediana de 163)