suma+ 72
Artículos
Marzo 2013
pp. 17-30

Introducción a la Estadística con *R*

José M.ª Barriuso, Virgilio Gómez,
M.ª José Haro y Francisco Parreño

Es remarcable que una ciencia que comenzó con el estudio sobre oportunidades de ganar en juegos de azar se haya convertido en el objeto más importante del conocimiento humano..., las preguntas más importantes sobre la vida son, en su mayor parte, en realidad, sólo problemas de probabilidad. (Pierre-Simon, Marqués de Laplace)

Generalmente, los estudiantes de bachillerato y universitarios tienen dificultades para comprender los conceptos más elementales de probabilidad y estadística. La presentación de conceptos abstractos de una forma visual y dinámica puede ayudar a comprenderlos mejor. La simulación de experimentos aleatorios ayudará a conseguirlo. Presentamos a continuación algunas de las actividades preparadas para ello.

Palabras clave: Probabilidad, Estadística, *software* libre, simulación.

Introduction to Statistics with *R*

High school and university students often have difficulties to understand basic ideas in probability and statistics. Introducing these new concepts in a visual and dinamic way can help to their better understanding. The simulation of random experiments will certainly help to achieve this goal. In this paper we present some exercises prepared for this purpose.

Key words: Probability, Statistics, Free software, Simulation.

A lo largo de nuestra experiencia como docentes en institutos de secundaria y universidad hemos podido comprobar que cuando los estudiantes han de aprender conceptos de Estadística y de Probabilidad adquieren la habilidad suficiente para saber utilizar adecuadamente los algoritmos de cálculo apropiados y para aplicar el modelo correspondiente de forma conveniente. Por ejemplo, saben cómo actuar cuando se encuentran delante de un determinado modelo de distribución de probabilidad, saben cuándo pueden aproximar la distribución binomial mediante la normal o cuándo han de aplicar un determinado tipo de contraste de hipótesis,…

Pero también hemos comprobado que no captan ni comprenden, en muchos casos, el sentido de lo que aprenden. Por ejemplo, ¿qué caracteriza a los valores que se distribuyen de una determinada forma?, ¿cuál es el significado del teorema central del límite y cuáles son sus implicaciones?, ¿cuál es la relación que existe entre el estadístico del contraste y el p-valor?…

Ver desde otro ángulo las ideas que existen detrás de cada concepto o procedimiento y manipularlas puede favorecer la comprensión de las mismas, promoviendo además el descubrimiento de relaciones entre ellas que permitan captar mejor su verdadero

Artículo recibido en *Suma* en marzo de 2012 y aceptado en enero de 2013

significado. La tecnología y determinado software hacen posible este hecho, al permitir la simulación de experimentos aleatorios con toda rapidez y fiabilidad. Si, además, ese software es gratuito los estudiantes podrán experimentar con él en cualquier lugar y momento y no será necesario limitarse al entorno del aula ni a la hora de clase. En este documento queremos presentar algunas de las actividades preparadas para trabajar con estudiantes de Bachillerato y Universidad. Para su diseño nos hemos apoyado en el material citado en la bibliografía al final de este documento.

El trabajo se ha realizado con el paquete estadístico *R*. Se descarga gratuitamente desde:

http://cran.r-project.org

R es un conjunto integrado de programas que permite trabajar con datos, cálculos y gráficos estadísticos. Incluye comandos para manejar y almacenar conjuntos de datos, operadores para desarrollar cálculos con vectores y matrices y comandos para análisis de datos y para representaciones gráficas. Además, permite trabajar con datos procedentes de diferentes sistemas de bases de datos. Inicialmente el programa consta de una serie de paquetes básicos que se pueden ampliar descargándolos desde la página web a la que hemos accedido para descargar el programa. Ello hace que la potencia y posibilidades del mismo sean muy considerables. Además, el lenguaje de programación es relativamente sencillo.

Comparacion de resultados al agrupar valores de una variable continua en diferente número de intervalos

Enunciado

Genera 100 valores correspondientes a una distribución normal *N(0,1)*. Represéntalos gráficamente mediante un histograma, agrupando los datos en un diferente número de intervalos de acuerdo a las reglas de Sturges, Scott y Freedman-Diaconis. ¿Qué observas? Haz lo mismo generando 1 000, 10 000, …

Objetivos

- Reconocer la importancia de las representaciones gráficas.
- Reflexionar sobre el papel que en un histograma desempeña el tamaño de los intervalos en la fiabilidad de la información ofrecida por la imagen visual obtenida.
- Comparar y comprender la potencia de cada una de las fórmulas más habituales utilizadas para obtener el número de intervalos en histogramas.

Metodología

Los estudiantes generan datos al azar correspondientes a una distribución normal de media 0 y desviación típica 1 con la orden:

```
x=rnorm(10000)
```

A continuación representan el histograma utilizando tres métodos diferentes: Sturges, Scott y Freedman-Diaconis. Con el primero de ellos calculan el número de intervalos, mientras que con los otros dos calculan la amplitud de los mismos. Para ello utilizarán las tres órdenes siguientes:

```
hist(x,main="Histograma según Sturges")
hist(x,breaks="Scott",main="Histograma según Scott")
hist(x,breaks="FD",main="Histograma según Freedman-Diaconis")
```

Inicialmente, se incluirá la instrucción:

```
set.seed(111)
```

con la finalidad de que todos los estudiantes obtengan los mismos datos y se pueda discutir sobre idénticas representaciones gráficas.

Después de hacer una primera discusión y reflexión conjunta obtendrán nuevos conjuntos de datos aumentando la cantidad de los mismos y se llevará a cabo una reflexión por parejas. Las conclusiones se

recogerán en el cuaderno de prácticas entregado por cada estudiante.

Comentarios

Para $n = 100$ se obtendrán gráficas similares a las que se muestran en la figura 1 y los estudiantes podrán observar cómo aumentar el número de intervalos da una información más detallada, pero también puede que no aporte nada nuevo debido a que el número de elementos en cada intervalo es demasiado pequeño. Hay que tener en cuenta también que Sturges sólo es verdaderamente fiable cuando el número de datos es menor que 200 y que en su fórmula sólo se tienen en cuenta el número total de datos:

$$\text{nº. de intervalos} = 1 + \frac{\log(n)}{\log(2)}$$

En cambio, las otras dos fórmulas tienen en cuenta otras características, como la cuasidesviación típica o el rango intercuartílico:

(Scott: amplitud=(3.5 s)/$\sqrt[3]{n}$; FD:amplitud=(2·IQR)/$\sqrt[3]{n}$)

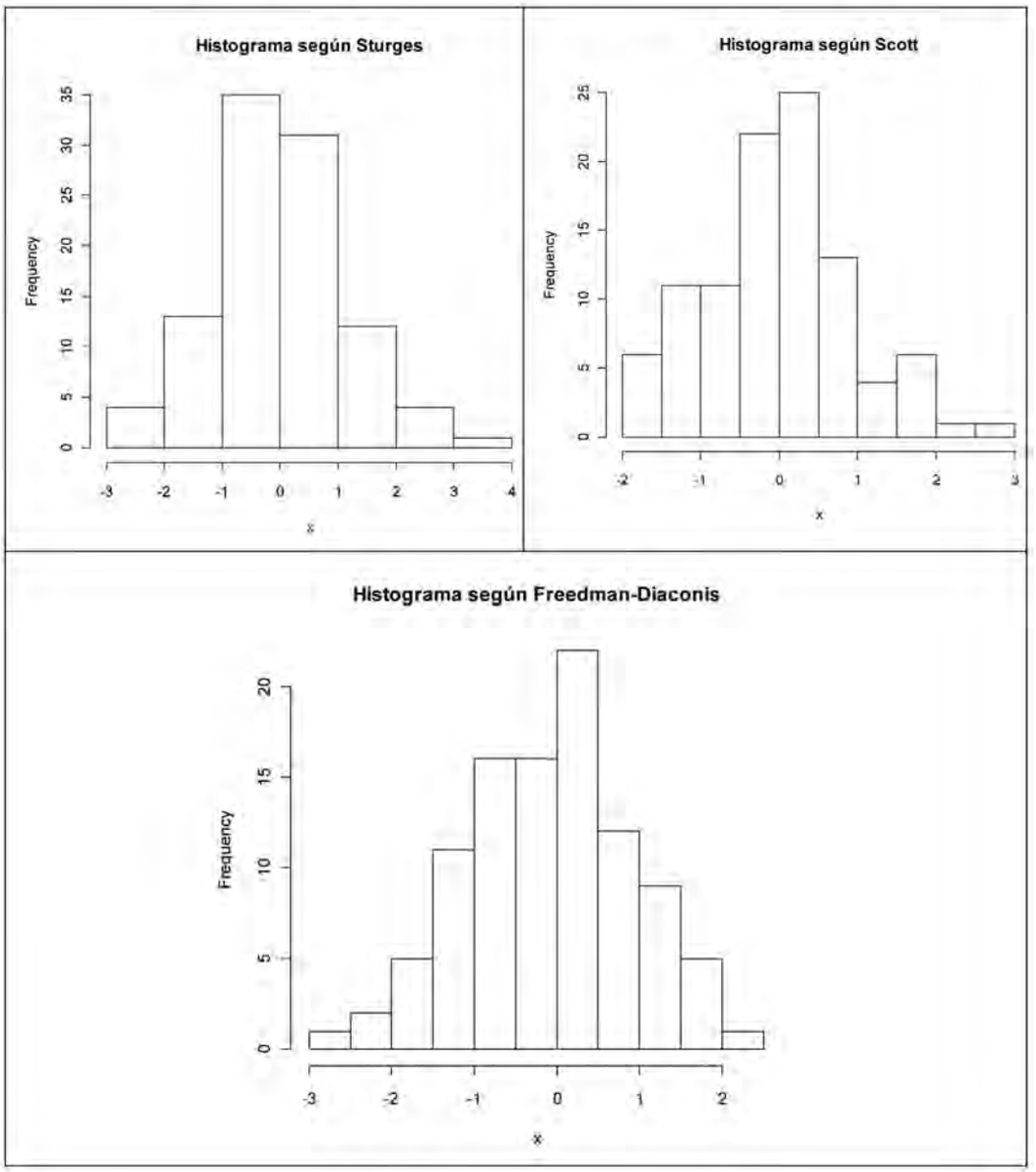

Figura 1

Simulación del lanzamiento de dos dados y cálculo de la suma de puntos

Enunciado

Utilizando los comandos apropiados simula el lanzamiento de un dado dos veces o de dos dados, y suma los puntos obtenidos. Representa gráficamente los resultados que obtendrías si realizaras el experimento 1 000 veces. ¿Cómo se distribuye la suma de puntos?

Simula otro experimento en el que lances 200 dados y sumes los puntos. Haz lo mismo que para el caso de los dos dados. ¿A qué conclusiones llegas?

Objetivos

- Razonar sobre el uso de comandos con el fin de cumplir tareas sencillas.
- Reproducir experimentos cotidianos en estadística.
- Aplicar los conceptos de experimento aleatorio, variable aleatoria y distribución de probabilidad a situaciones experimentales concretas.
- Aclarar los conceptos mencionados en el objetivo anterior y darles significado.

Metodología

Puesto que los estudiantes ya tienen conocimiento de los conceptos y procedimientos presentes en la actividad, así como de los comandos necesarios para llevarla a cabo, la actividad se presenta sin más preámbulos y se deja que trabajen por parejas. El profesor permanece atento para resolver las dudas que surjan y orientar a los que lo necesiten.

Comentarios

Para simular el lanzamiento de un dado un determinado número de veces podemos proceder de las dos formas siguientes. Una es:

```
sample(a:b,c,rep=T)
```

Con esta sencilla instrucción se generan al azar y con reemplazamiento *c* valores de entre los números enteros comprendidos entre *a* y *b* (si no se desea que haya reemplazamiento basta con omitir la opción *rep=T*). Por ejemplo:

```
sample(1:10,10,rep=T)
5 4 7 3 2 10 3 7 9 9
```

Con esta instrucción hemos generado esos 10 valores comprendidos entre 1 y 10.

La otrs forma es utilizar la instrucción:

```
ceiling(runif(c,a,b))
```

Con runif(c,a,b) generaremos *c* números reales al azar entre *a* y *b*. Estos números, con el comando ceiling se aproximarán al entero posterior (a no ser que ya sea entero, en cuyo caso se queda como estaba). Por ejemplo, con:

```
ceiling(runif(10,0,10))
```

hemos generado los 10 valores siguientes:

```
7 5 9 5 5 10 8 4 9 3
```

Si queremos simular el lanzamiento de un dado dos veces, o de dos dados, haríamos:

```
sample(1:6,2,rep=T)
```

o bien:

```
ceiling(runif(2,0,6)).
```

Para sumar los puntos podemos utilizar:

```
sum(sample(1:6,2,rep=T))
```

que genera los números y los suma.

Si lo que pretendemos es simular el mismo experimento 1000 veces utilizamos la siguiente instrucción:

```
t=sapply(1:1000,function(x)
{sum(sample(1:6,2,rep=T))})
```

Para representar gráficamente los datos procedemos con la instrucción:

```
barplot(table(t))
```

Los estudiantes deben reflexionar sobre el tipo de variable aleatoria con la que tra-

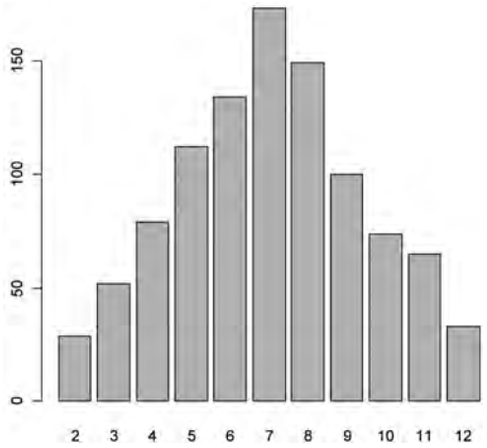

Figura 2

bajan, sobre cuál es la probabilidad de obtener cada uno de los resultados posibles, sobre la media de la distribución y su representatividad, sobre la dispersión de los datos, o sobre la mediana y la moda. La representación gráfica (fig. 2) puede ayudar a determinar la posible simetría de los datos o si la distribución se podría aproximar a una distribución normal o no.

R también permite calcular probabilidades de resultados concretos y valores de parámetros como la media, mediana o varianza, así como del grado y clase de asimetría.

En el segundo caso, en el que hay que lanzar 200 dados y sumar los puntos, al haber un gran número de resultados posibles interesa agrupar como si de una variable continua se tratara y obtener un histograma. Las instrucciones utilizadas son ahora:

```
t=sapply(1:1000,function(x){sum(sample(1:6,200,rep=T))})
hist(t,breaks="Scott")
```

Véase la gráfica obtenida en la figura 3.

Los estudiantes deben analizar los mismos conceptos y comparar resultados. Al igual que antes, sobre el histograma, han de reflexionar sobre si se ajusta a una distribución normal o no. Además de tener en cuenta nuestra impresión visual, podemos ser un poco más precisos y ayudarnos su-

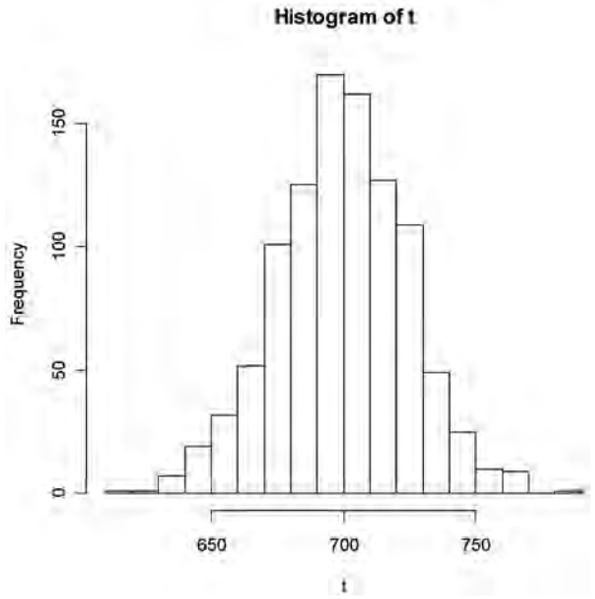

Figura 3

perponiendo a la gráfica una curva normal. Para ello es necesario, además de utilizar la instrucción correspondiente, modificar ligeramente el histograma haciendo que se representen densidades en lugar de frecuencias. Basta con introducir en la instrucción el parámetro *freq=F*. Es decir:

```
hist(t,breaks="Scott",freq=F)
```

Las densidades se calculan a partir de las alturas, considerando las frecuencias relativas equivalentes a las áreas de los rectángulos, con lo cual la densidad correspondiente a cada rectángulo se calculará con la expresión:

$$densidad = \frac{frecuencia\ \ relativa}{amplitud\ \ del\ \ intervalo}$$

Para representar la curva que correspondería a una distribución normal de media y desviación típica correspondientes a la muestra introducimos la orden:

```
curve(dnorm(x,mean(t),sqrt(var(t))),add=T)
```

Con esta instrucción estamos diciendo que se dibuje la campana de Gauss que correspondería a la función de densidad de una distribución normal de media y desviación típica obtenidas a partir de los datos generados por el programa. Con *add=T* indicamos que la curva se superponga al histograma realizado previamente (fig. 4).

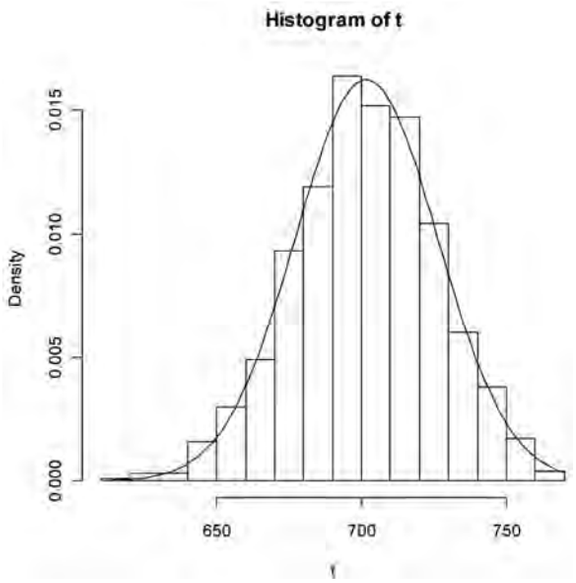

Figura 4

Simulación del lanzamiento de veinte dados (o veinte veces de un dado) y recuento de la frecuencia de un resultado

Enunciado

Simula el lanzamiento de 20 dados y calcula el número de veces que aparece el número 6. Repite el experimento 1000 veces. ¿Cuál crees que será el resultado más probable? ¿Cuál ha sido el resultado más probable? ¿Cuál crees que será el valor medio de la muestra? ¿Cuál es el valor medio de la muestra? ¿Y la desviación típica? ¿Cómo se distribuyen los datos en torno a la media? ¿Cuáles son las características de la distribución que obtienes? ¿Con qué tipo de distribución crees que estás trabajando? ¿Qué ocurriría si lanzaras 100 dados en vez de 20? (Responde las mismas preguntas que para un número de dados igual a 20) ¿Se distribuyen los datos de la misma forma?

Objetivos

— Utilizar la simulación de experimentos aleatorios particulares como forma de promover el razonamiento.
— Reconocer las características de los experimentos de Bernoulli en situaciones concretas.
— Identificar los parámetros y las propiedades más relevantes de la distribución binomial.
— Establecer relaciones entre la distribución binomial y la normal.

Metodología

Los estudiantes conocen los comandos e instrucciones que les serán útiles para simular la realización del experimento. Se trata de que reflexionen sobre cómo utilizarlos para lograrlo. Una vez hecho esto han de usar los datos procedentes de la experimentación para responder las preguntas del ejercicio e intentar llegar a conclusiones.

Comentarios

En cuanto a las instrucciones necesarias, la actividad se puede realizar de manera similar a la anterior. Podríamos utilizar:

```
t=sapply(1:1000,function(x)
{sum(sample(1:6,20,rep=T)==6)})
```

Con esta instrucción se simula el lanzamiento de 20 dados 1000 veces, contándose en cada caso el número de veces que en algún dado aparece el resultado 6.

Los resultados se pueden representar gráfiamente (fig. 5) con la instrucción:

```
barplot(table(t))
```

Ayudándose de la gráfica los estudiantes han de reflexionar sobre cuál es el valor

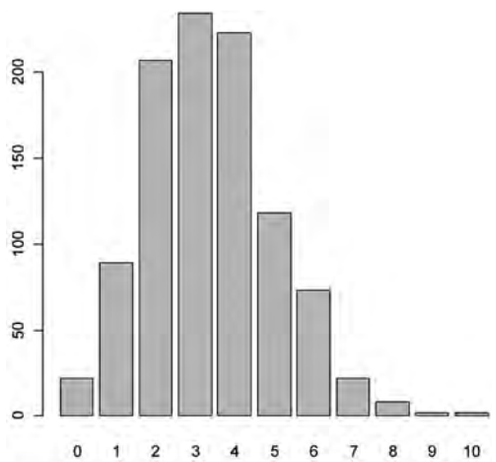

Figura 5

más probable y analizar si coincide con lo que les dicta la lógica. Pueden conjeturar sobre cuál será el valor de la media y considerar si los datos estarán muy dispersos o no, así como sobre la simetría. Pueden confirmar o refutar sus ideas utilizando las instrucciones siguientes. Para la media:

```
mean(t)
```

Para la desviación típica muestral:

```
sqrt(var(t))
```

Para la simetría:

```
skewness(t)
```

El experimento se repetirá simulando el lanzamiento de 100 dados.

Los estudiantes que hayan identificado el experimento con las pruebas de Bernouilli pueden haberse lanzado a utilizar directamente las instrucciones relacionadas con este tipo de pruebas. Para ello han debido identificar el valor de los parámetros n ($n=20$ o $n=100$) y p ($p=1/6$). Las instrucciones a utilizar serían:

```
t=rbinom(100,20,1/6)
barplot(table(t))
```

El resultado es muy similar.

Propiedad reproductiva de determinadas variables

Metodología

Esta actividad consta de diferentes enunciados relacionados con diversas distribuciones. En todos ellos se trata de analizar lo que ocurre cuando se trabaja la suma de variables aleatorias o la media de las mismas, con el fin de analizar el carácter reproductivo de diversas distribuciones y de introducirse en el teorema central del límite. Mediante el programa, los estudiantes generarán valores correspondientes a las diferentes variables, crearán las variables suma o media y observarán lo que ocurre con la distribución correspondiente a las nuevas variables. Apoyándose en la potencia visual de las gráficas generadas se reflexionará sobre los resultados con el fin de llegar a las conclusiones correctas. Este bloque de enunciados se expondrá a los estudiantes antes del desarrollo teórico y de la presentación formal de los contenidos correspondientes, con el fin de que puedan anticipar los resultados por ellos mismos.

Enunciado 1

Considera dos variables aleatorias binomiales $\chi \rightarrowtail B(10, 0.5)$ y $\eta \rightarrowtail B(20, 0.5)$ ¿Cómo se distribuye la variable aleatoria $\chi + \eta$?

Objetivos

Reflexionar sobre lo que ocurre con la suma de variables de distribuciones binomiales de igual parámetro p.

Comentarios

Con las instrucciones:

```
a=rbinom(10000,10,0.5)
b=rbinom(10000,20,0.5)
s=a+b
barplot(table(s))
```

se generarán y representarán valores correspondientes a la suma de las variables aleatorias $\chi+\eta$. A partir de la gráfica y del cálculo de valores como la media se podrá comprobar que la distribución binomial de parámetro p es reproductiva con respecto al número de experimentos de Bernoulli.

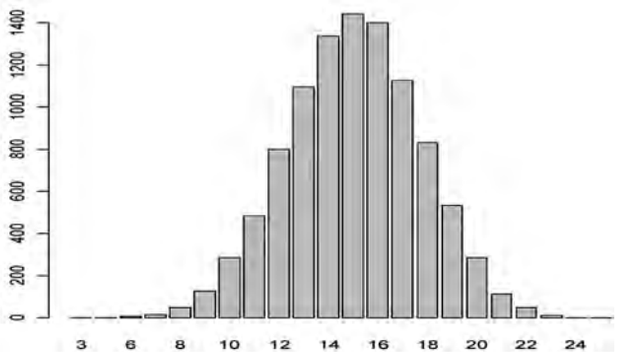

Figura 6

Sobre la misma gráfica (fig. 6), los estudiantes podrán observar que, dada la simetría de la distribución, la media es aproximadamente 15, aunque también se puede obtener la misma con la instrucción:

```
mean(a+b)
```

Al tratarse de una distribución binomial, al dividir la media entre el parámetro n (10+20), obtenemos el valor de p que va a coincidir con 0.5.

Para confirmar esta conjetura los estudiantes podrán generar y representar gráficamente valores correspondientes a una distribución binomial de parámetros $n = 30$ y $p = 0.5$ y ver que es muy semejante a la obtenida para la suma de variables (fig. 7).

```
k=rbinom(10000,30,0.5)
barplot(table(k))
```


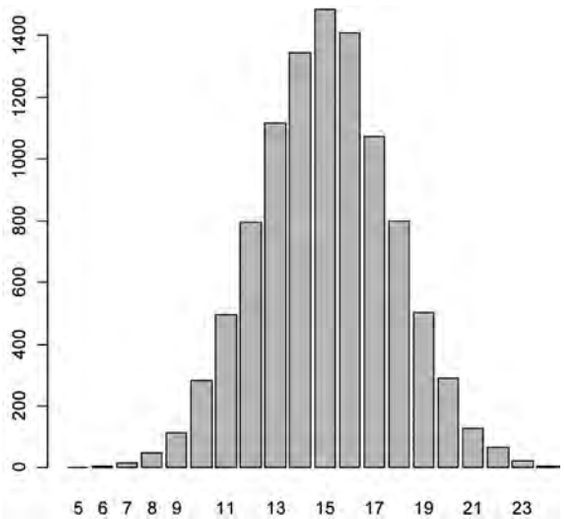

Figura 7

También podrán simular más experimentos cambiando los valores de los parámetros n y p.

Enunciado 2

Considera dos variables aleatorias independientes y con distribución de Poisson de medias 2 y 3. Genera 10 000 valores y represéntalos gráficamente. Calcula la media y la varianza de la suma de dichas variables y obtén una representación gráfica. ¿Cuál crees que es la distribución de la suma?

Objetivos

Conjeturar la propiedad reproductiva de la distribución de Poisson.

Comentarios

Se procederá de manera similar a la anterior, sólo que trabajando con las instrucciones

```
a=rpois(10000,2)
b=rpois(10000,3)
s=a+b
barplot(table(s))
```

Los estudiantes que lleguen a la conclusión de que la suma de variables se distribuye según un modelo de Poisson de media la suma de las medias podrán confirmar sus hipótesis generando los valores de la variable apropiada (fig. 8).

```
t=rpois(10000,5)
barplot(table(t))
```

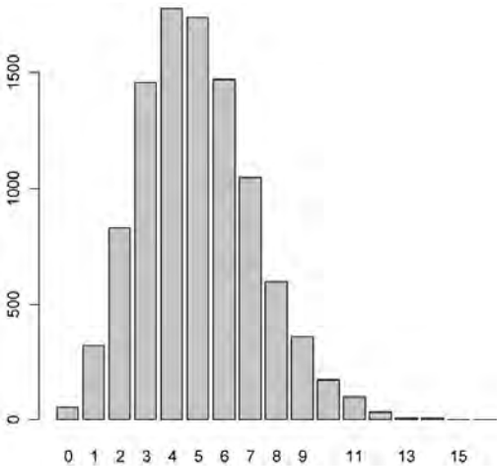

Figura 8

Enunciado 3

Considera tres variables independientes y distribuidas normalmente con media 3 y desviación típica 1, ¿cómo se distribuye la media muestral? ¿Qué pasaría si tuviéramos n variables?

Objetivos

— Estudiar la propiedad reproductiva de la distribución normal.
— Introducir en el estudio de las distribuciones en el muestreo.

Comentarios

Se generarán 10000 valores para cada una de las tres variables aleatorias normales independientes de media 3 y desviación típica 1. Se creará una nueva variable sumando las tres anteriores. Se representarán gráficamente los datos (fig. 9) y se estudiará el tipo de distribución que se obtiene, deteniéndose en la media y en la desviación típica. Así se comprobará la propiedad reproductiva viendo que la media de la variable suma es la suma de las tres medias y la varianza es la suma de las varianzas.

```
a=rnorm(10000,3,1)
b=rnorm(10000,3,1)
c=rnorm(10000,3,1)
s=a+b+c
hist(s,freq=F,breaks=»Scott»)
```

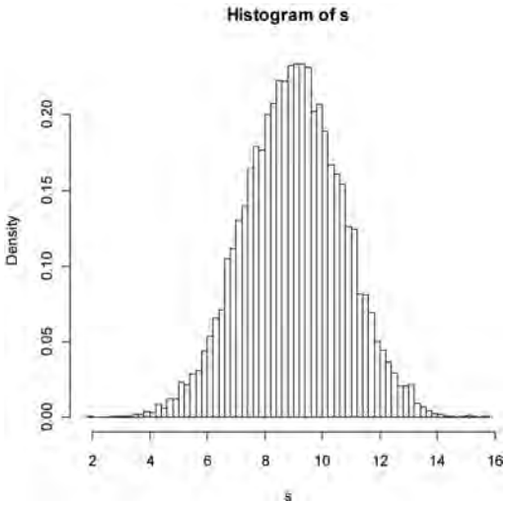

Figura 9

Podrán comprobar sus suposiciones con las instrucciones siguientes:

```
mean(s)
var(s)*10000/9999
```

Se iniciará a continuación el estudio del estadístico media muestral, calculando la media de los valores de las tres variables y representando gráficamente los resultados (fig. 10).

```
a=rnorm(10000,3,1)
b=rnorm(10000,3,1)
c=rnorm(10000,3,1)
m=(a+b+c)/3
hist(m,freq=F,breaks=»Scott»)
```

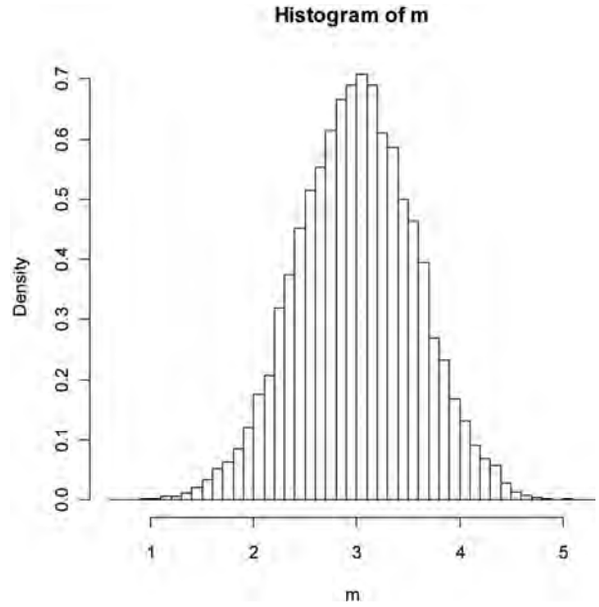

Figura 10

Se puede observar que el valor de la media ha cambiado, siendo la misma de las distribuciones iniciales. También ha cambiado la varianza y se puede obtener su valor, al igual que el de la media, con las mismas instrucciones anteriores. Gráficamente (fig. 11), podemos comparar las dos distribuciones superponiendo curvas normales que nos permitan ver qué ocurre con la media y varianza de la variable media muestral.

```
m=(a+b+c)/3
hist(m,freq=F,breaks=»Scott»)
curve(dnorm(x,3,1),col=»blue»,add=T)
curve(dnorm(x,mean(m),sqrt(var(m)*10000/9999)),
col=»red»,add=T)
```

El ejercicio se debe repetir con más variables. Se puede simular, por ejemplo, la suma de 20 variables. Para ello resulta más cómodo utilizar la instrucción:

```
m=sapply(1:10000,function(x){mean(rnorm(20,3,1))})
```

Esta instrucción generará 10 000 muestras de tamaño 20, calculando la media para cada una de ellas.

Los estudiantes también podrán observar que cuanto mayor es la desviación típica, más aplanada es la curva normal.

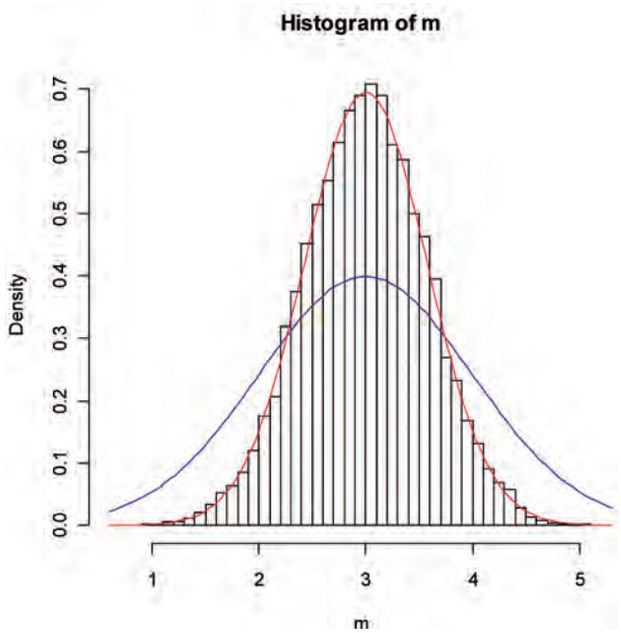

Figura 11

Enunciado 4: Teorema Central del Límite

Analiza el siguiente resultado «La media de *n* variables aleatorias independientes e igualmente distribuidas, se distribuye según una normal cuando *n* tiende a infinito».

Objetivos

- Aplicar los resultados parciales obtenidos en las anteriores actividades a un problema más general y de importantes consecuencias.
- Profundizar en el significado de uno de los teoremas más importantes de la teoría de la probabilidad.
- Profundizar en los conceptos de función de densidad, de distribución, de media y de desviación típica.

Comentarios

Para trabajar sobre la afirmación de este teorema se considerarán poblaciones con distribuciones continuas habituales como, por ejemplo, la uniforme y la exponencial. Posteriormente, se trabajará con otra distribución con función de densidad no nula en intervalos disjuntos con el fin de poder apreciar la potencia de este teorema.

Tomaremos, por ejemplo, 10 000 muestras de tamaño 30 de una distribución uniforme y calcularemos las 10 000 medias:

```
x=sapply(1:10000,function(x){mean(runif(30,
0,1))})
```

Representamos el histograma:

```
hist(x,freq=F,breaks=»Scott»)
```

Representamos la distribución uniforme original:

```
curve(dunif(x),col=”blue”,add=T)
```

Se puede observar en el dibujo que la gráfica es una recta de altura 1 (figura 12).

La curva normal teórica correspondiente al teorema central del límite se calcularía a partir de los valores de la uniforme, es decir, una media igual a μ, y una desviación típica igual a σ:

$$\mu = \int_0^1 x \frac{1}{1-0}\,dx = \left[\frac{x^2}{2}\right]_0^1 = \frac{1}{2}$$

$$\sigma = \frac{\sqrt{\int_0^1 x^2\,dx - \mu^2}}{\sqrt{30}} = \frac{\sqrt{\left[\frac{x^3}{3}\right]_0^1 - \frac{1}{4}}}{\sqrt{30}} =$$

$$= \frac{\sqrt{\frac{1}{3} - \frac{1}{4}}}{\sqrt{30}} = \frac{\sqrt{\frac{1}{12}}}{\sqrt{30}} = \frac{1}{\sqrt{12 \cdot 30}}$$

```
curve(dnorm(x,.5,1/(sqrt(12*30))),col=”red”,
add=T)
```

Podemos simular de nuevo el experimento considerando muestras de mayor tamaño,

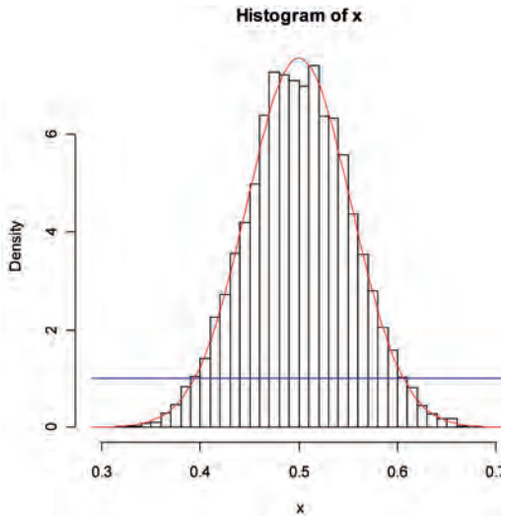

Figura 12

por ejemplo $n = 100$, para observar el comportamiento de la distribución y acercarnos más al significado del teorema.

Se puede repetir el experimento con otro tipo de distribuciones, incluso un tanto llamativas, como la una población con la siguiente función de densidad:

Figura 13. Función de densidad

En este caso se pide a los estudiantes que calculen la función de distribución y que representen gráficamente ambas, la de función de densidad y la de distribución (figs. 13 y 14). De esta forma pueden observar mejor sus características.

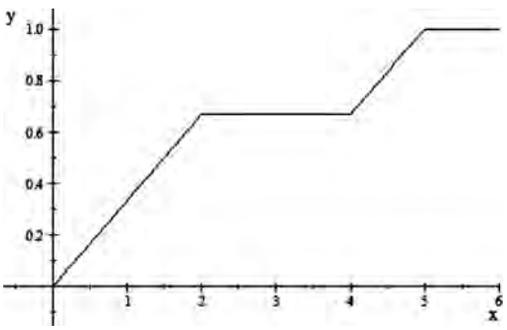

Figura 14. Función de distribución

Simular este tipo de distribución es algo más complejo y los estudiantes han recibido las instrucciones necesarias para poder hacerlo.

```
f=function(x){if(x<2/3)return(3*x)else return(2+3*x)}
b=sapply(1:10000,function(x){mean(sapply(runif(30),f))})
hist(b,freq=F,breaks="Scott")
curve(dnorm(x,(13/6),sqrt(107/(36*30))),add=T)
```

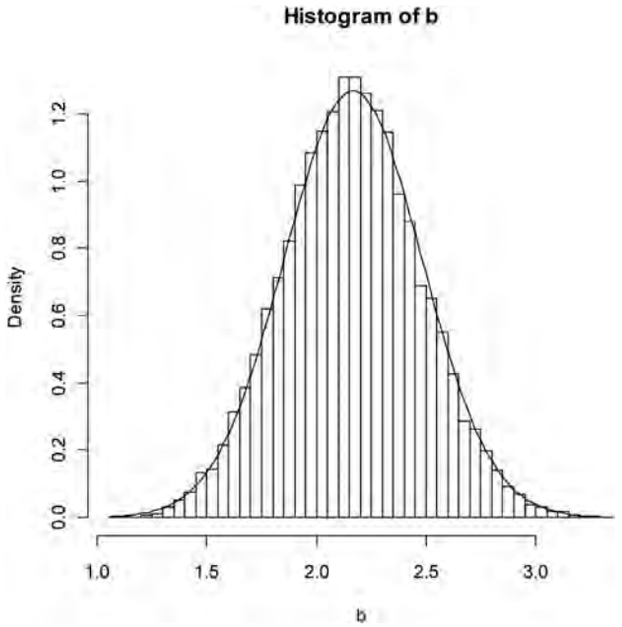

Figura 15

Se puede observar cómo tomando muestras de sólo 30 elementos, el histograma se ajusta muy bien a una curva normal de media la de la distribución y de desviación típica la de la distribución dividida por $\sqrt{n}$ (fig. 15).

Intervalos de confianza

Enunciado

Construir intervalos de confianza al 95 %, significa que de cada 100 intervalos que construimos con el método elegido, en promedio, 95 contendrán la media de la población.

Comprobémoslo para la media de una distribución uniforme en $[0, 1]$ a partir de muestras de tamaño 30 con un nivel de confianza del 95 %.

Objetivos

Profundizar en el significado de intervalo de confianza.

Metodología

Los estudiantes han adquirido en clase los conocimientos necesarios relativos al sentido y utilidad de los intervalos de confianza como entornos que, con una alta fiabilidad, contienen al parámetro poblacional que está siendo estimado. Se trata de que profundicen más en el carácter de los intervalos de confianza analizando el papel de éstos. Los estudiantes procederán de la siguiente forma. Recordando que el intervalo es:

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right)$$

construirán primero la función intervalo que genera los intervalos de confianza (siendo *n* el tamaño de la muestra y *alpha* el nivel de significación):

```
intervalo=function(n,alpha){
```

Generarán un vector con *n* valores aleatorios de la distribución uniforme:

```
x=runif(n)
```

Hallarán su media:

```
m=mean(x)
```

y la estimación del error estándar de la media:

```
et=sd(x)/sqrt(n)
```

La siguiente instrucción devuelve el intervalo calculado a partir de z:

```
c(m-qnorm(1-alpha/2)*et,m+qnorm(1-alpha/2)*et)}
```

A continuación, construirán la base del gráfico, donde representarán la media con una línea roja, utilizando la instrucción

```
plot(c(0.5,0.5),c(1,100),type="l",col="red")
```

Después, con un ciclo *for*, mostrarán 100 intervalos de confianza para la media a alturas crecientes:

```
for(i in 1:100){lines(intervalo(30,0.05),c(i,i),col="blue")}
```

Los intervalos vienen representados por líneas azules horizontales (figura 16) y contando los que contienen a la media poblacional podrán entender mejor este otro aspecto de los intervalos de confianza reflejado en el enunciado.

Comentario

Será conveniente repetir el experimento diversas veces y para diferentes niveles de significación. Esta actividad se debe repetir para obtener intervalos de confianza correspondientes a diversos parámetros, como la proporción, varianza, el cociente de varianzas, diferencia de medias o de proporciones, suponiendo que se trabaja con muestras grandes y pequeñas y que se conoce o desconoce el valor de la varianza o varianzas poblacionales.

También es importante presentar situaciones problemáticas en las que se haya de dar respuesta a problemas reales, próximos al entorno del estudiante.

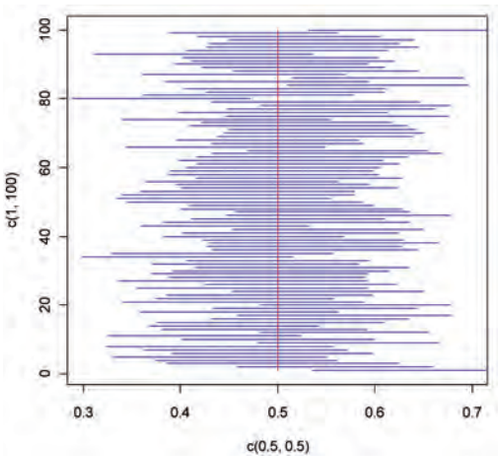

Figura 16

Contraste de hipótesis

Objetivos

- Trabajar con diversos test de hipótesis para comprender su sentido y utilidad.
- Analizar y entender el sentido de los diversos elementos de un test de hipótesis.

— Ver la relación entre test de hipótesis e intervalo de confianza.
— Establecer la relación que existe entre estadístico del contraste y *p*-valor.

Metodología

Después de haber completado en clase el desarrollo teórico de los conceptos y procedimientos más habituales en el contraste de hipótesis se utilizarán las posibilidades gráficas y la potencia de cálculo del programa *R* para profundizar en el sentido de los elementos integrantes de un test de hipótesis. La potencialidad de *R* se utilizará también para realizar contrastes no paramétricos, así como para comprobar si se cumplen las hipótesis necesarias para poder realizar test paramétricos. Mostramos algunos ejemplos.

Enunciado 1

Consideremos una encuesta en la que preguntamos por la calle a 100 personas, elegidas al azar, sobre cierta cuestión a la que responden sí 42 personas. ¿Es este dato compatible con la hipótesis de que la proporción de síes en la población es del 50%? ¿Cuál sería la respuesta si hubiéramos obtenido 420 síes de 1 000 encuestados?

Comentarios

Para realizar este test se utilizará la función *prop.test*, concretamente la instrucción:

```
prop.test(42,100,p=0.5)
```

En ella indicamos que los resultados de la muestra han sido 42 de 100 y que contrastamos la hipótesis nula de que la proporción es 0.5.

La respuesta del programa es la siguiente:

```
1-sample proportions test with continuity correction
data: 42 out of 100, null probability 0.5
X-squared = 2.25, df = 1, p-value = 0.1336
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.3233236 0.5228954
sample estimates:
p
0.42
```

La primera línea de la respuesta nos indica que es una prueba para una proporción. La segunda línea contiene los datos que hemos introducido. De la tercera línea nos interesa el *p*-valor que es el mínimo valor de significación que podríamos tomar para rechazar H_0, es decir, el *p*-valor, nos indica que si H_0 fuera verdadera, la probabilidad de extraer una muestra cuya proporción esté tan lejos de H_0 como el valor observado de 42, es el *p*-valor, en este caso es 0.1336.

Este valor del *p*-valor, nos está diciendo que con los valores de significación habituales no deberíamos rechazar la hipótesis nula.

La cuarta línea nos indica cuál es la hipótesis alternativa.

La quinta línea nos dice que el nivel de significación ha sido de 0,05. Si hubiéramos querido tomar otro nivel de significación, deberíamos haber introducido la orden:

```
conf.level=1-α
```

La sexta línea nos da un intervalo de confianza para la proporción y podemos ver que el valor 0,5 está dentro de él, lo que nos confirma la decisión de aceptar la hipótesis nula. Las dos últimas líneas nos dan el porcentaje obtenido en la muestra.

Para responder la siguiente pregunta la instrucción sería:

```
prop.test(420,1000,p=0.5)
```

y los resultados proporcionados por el programa:

```
1-sample proportions test with continuity correction
data:  420 out of 1000, null probability 0.5
X-squared = 25.281, df = 1, p-value = 4.956e-07
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3892796 0.4513427
sample estimates:
p
0.42
```

Obsérvese que aunque la proporción en la muestra es la misma (42%) el p-valor ha descendido mucho (0,0000004956) a causa del mayor tamaño muestral. En este caso, los datos ofrecen evidencia en contra de H_0, con los niveles de significación habituales.

Enunciado 2

Si otro encuestador realizó 200 encuestas y obtuvo 110 síes en otra población, ¿es la proporción de síes la misma en las dos poblaciones?

Comentario

Se trata de un contraste de dos proporciones. La instrucción que debemos utilizar es:

```
prop.test(c(42,110),c(100,200))
```


Enunciado 3

Un investigador desea comprobar que la cantidad de ingesta diaria de fibra en una determinada población es inferior a la cantidad habitualmente recomendada de 20 g diarios. El investigador sabe que la desviación estándar de dicha variable es aproximadamente de 10 g y considera que una ingesta inferior en 5 g a la cantidad recomendada es clínicamente relevante. ¿Cuál será el mínimo número de individuos objeto de estudio que garantice una potencia del 80% en la detección de las diferencias deseadas con una prueba bilateral y un nivel de significación del 5%?

Comentarios

Utilizaremos la instrucción:

```
power.t.test(delta=5,power=0.8,sig.level=0.0
5,sd=10,type=»one.sample»,alternative=»tw
o.sided»)
```

Se puede observar que hemos introducido la potencia del contraste esperando recoger el número mínimo de elementos en la muestra. Los resultados son:

```
One-sample t test power calculation
n = 33.36720
delta = 5
sd = 10
sig.level = 0.05
power = 0.8
alternative = two.sided
```

Concluimos que el número mínimo de individuos debe ser 34.

Referencias bibliográficas

García Pérez, A. (2005), *Métodos Avanzados de Estadística Aplicada: Métodos Robustos y de remuestreo*, UNED, Madrid.

Peña, D. (2008), *Fundamentos de Estadística*, Alianza Editorial, Madrid.

Ugarte, M. D., y A. F. Militino (2002), *Estadística Aplicada con R*, Universidad Pública de Navarra, Pamplona.

Walpole, R. E., R. H. Myers, S. L. Myers y K. Ye (2007), *Probabilidad y Estadística para Ingeniería y Ciencias*, Pearson Educación, México.

José M.ª Barriuso Rodriguez
IES Bachiller Sabuco (Albacete)

Virgilio Gómez Rubio
Escuela de Ingenieros Industriales UCLM (Albacete)

M.ª José Haro Delicado
IES Al-Basit
Escuela de Ingeniería Informática UCLM (Albacete)
<mariajose.haro@uclm.es>

Francisco Parreño Torres
Escuela de Ingeniería Informática UCLM (Albacete)