Situaciones problemáticas en R Studio: Instrumento de aprendizaje para la prueba de Bondad de Ajuste de Pearson

Marisa Battisti Arduin[1], Nanci Odetti[2], Flavio Lovatto[3].

[1] Maestranda en Estadística Aplicada, Facultad de Ingeniería – Universidad Nacional de Entre Ríos, Oro Verde, Argentina, [2] Bioingeniera, Facultad de Ingeniería – Universidad Nacional de Entre Ríos, Oro Verde, Argentina, [3] Estudiante avanzado de Bioingeniería, Facultad de Ingeniería – Universidad Nacional de Entre Ríos, Oro Verde, Argentina

marisa.battisti@uner.edu.ar, nanci.odetti@uner.edu.ar, flavio.lovatto@uner.edu.ar

Asignaturas: Probabilidad y Estadística

Nombre del eje: Uso de herramientas tecnológicas aplicadas a la educación

Resumen: La asignatura Probabilidad y Estadística pertenece al segundo año de la carrera de Bioingeniería de la Facultad de Ingeniería (Universidad Nacional de Entre Ríos). En la segunda etapa del dictado de la asignatura se abordan fundamentos de la estimación de parámetros y la prueba de hipótesis, incluyendo la denominada Prueba de Bondad de Ajuste de Pearson. Este contraste sirve para evaluar qué tan bien se ajustan los datos experimentales a una distribución de probabilidad en particular.

Al ser dicho contenido un tópico cargado de significación dentro de la disciplina estadística, el equipo de cátedra considera que debe ser explotado aún más mostrando al grupo de estudiantes las posibilidades de aplicación del mismo en problemas inherentes a su futuro campo de actuación. En efecto, se pensó en una actividad de aplicación del contenido en cuestión y que tenga al alumnado como protagonista en la resolución de la misma.

La presente propuesta consiste entonces en utilizar las ventajas de la simulación del software libre R Studio, en su versión en línea, para crear un contexto de enseñanza-aprendizaje que permita aplicar la temática Bondad de Ajuste de Pearson en situaciones problemáticas propias del campo de la Bioingeniería, brindando la posibilidad al alumnado de poner en práctica lo asimilado.

Se pretende seguir incorporando actividades de estas características, pensando en nuevas propuestas didácticas para el abordaje de los contenidos, con entornos motivadores y facilitadores de los procesos de enseñanza y aprendizaje, propiciando la metacognición de saberes.

Palabras clave: Estadística; Problema de aplicación; Instrumento de aprendizaje; R studio; Simulación; Prueba de bondad de ajuste; Distribución chi-cuadrado

1. Introducción

Probabilidad y Estadística es una asignatura de carácter cuatrimestral y correspondiente al segundo año de la carrera de Bioingeniería, impartida en la Facultad de Ingeniería de la Universidad Nacional de Entre Ríos, Argentina. Durante el dictado de la materia, se comienza, en una primera

etapa, abordando los conceptos y las principales leyes que garantizan el cálculo de probabilidades, luego se estudian las variables aleatorias discretas y continuas, las funciones de distribución de probabilidad o de densidad de probabilidad, además de distribuciones teóricas como la Binomial, Poisson, Normal y Exponencial, entre otras. En una segunda instancia se desarrollan los fundamentos de la estimación y la prueba de hipótesis, concluyendo con las denominada Prueba de Bondad de Ajuste de Pearson.

Este contenido no sólo tiene su ubicación al final del programa por su grado de complejidad, sino porque articula un gran número de conceptualizaciones para su tratamiento: desde los cálculos sencillos de probabilidades, pasando por las distribuciones teóricas ya mencionadas, hasta todo el bagaje necesario para efectuar un contraste estadístico, incluyendo la formulación de las hipótesis correspondientes, el cómputo de un estadístico de prueba a partir de los datos muestrales, la obtención de un valor P o p-value y la decisión final con su debida conclusión respecto al interrogante originado.

Al ser dicho contenido un tópico dotado de complejidad, pero a su vez cargado de significación dentro de la disciplina Estadística, el equipo de cátedra considera que debe ser explotado aún más mostrando al grupo de estudiantes las posibilidades de aplicación del mismo en problemas inherentes a su futuro campo de actuación.

Se coincide en la idea de que el estudiante como tal logra mejores aprendizajes cuando alcanza a dimensionar la aplicación de un nuevo contenido en el ámbito de la carrera que eligió llevar a cabo. Motivo por el cual, se pensó en una actividad de aplicación del contenido en cuestión y que tenga al alumnado como protagonista en la resolución de la misma.

Habitualmente los ejercicios que se afrontan en clases prácticas sobre el tema implican evaluar mediante el contraste de Bondad de Ajuste si, por ejemplo, la puntuación obtenida al lanzar un dado es una variable aleatoria con comportamiento probabilístico Binomial a partir de una serie de resultados de 50 tiradas de un dado balanceado, o bien, comprobar que la distribución de los grupos sanguíneos en una cierta región es uniforme, recurriendo a los datos de una muestra de 200 sujetos.

Existen numerosos ejemplos más en los materiales bibliográficos pero no necesariamente se observa que sean actividades estrechamente relacionadas con la Bioingeniería. Además, se detectan ciertas distancias entre contar con una serie de datos recabados y organizados en una tabla para poder hacer los cálculos necesarios, frente a simular los datos, organizarlos y a partir de allí encarar la resolución del problema. Es decir, se ponen muchas más capacidades en juego cuando se tienen los datos en crudo, originados por una simulación mediante software por ejemplo.

Habitualmente y coincidiendo con el enfoque de Grima y Riba (1995), los problemas de cálculo de probabilidades que aparecen en la mayoría de los cursos de estadística de carreras universitarias, involucran interrogantes que no sólo requieren de un ávido y claro dominio de conceptualizaciones teóricas sino de una cierta experticia práctica y en muchas ocasiones de ideas simples que orienten su resolución.

Desde la interpelación mencionada, los docentes de la asignatura se posicionan en un acercamiento hacia un proceso de innovación didáctica emergente, adhiriendo a las palabras de Libedinsky (2016), buscando la intersección entre lo que se desea alcanzar desde el equipo de

cátedra (la aplicación de un contenido específico) y algo que puede ser de interés para el estudiante de Bioingeniería (la factible utilidad del contenido en su contexto).

La presente propuesta consiste entonces en utilizar las ventajas de la simulación del software libre R Studio, en su versión en línea (https://rstudio.cloud/), para crear un contexto de enseñanza-aprendizaje que permita aplicar la temática Bondad de Ajuste de Pearson en situaciones problemáticas propias del campo de la Bioingeniería, brindando la posibilidad al alumnado de poner en práctica lo asimilado.

Si bien en los útlimos cuatrimestres se ha aplicado el software antes nombrado, se cree que la utilización de un recurso libre de estas características brinda la posibilidad de acceder a él desde cualquier lugar y en cualquier momento, pues no requiere siquiera descarga e instalación del mismo en la computadora que se emplee. Esta cualidad es otra clara faceta de las garantías que aporta el uso de las Tecnologías de la Información (TIC) en cuanto a la mejora de la educación (Libedinsky, 2016).

La inclusión de R Studio en la asignatura, y no sólo para este contenido, pretende contribuir a la comprensión de algunos conceptos de difícil asimilación.

2. Desarrollo de la propuesta

La propuesta consiste en una actividad que el alumnado deberá resolver de forma extra-áulica, reunidos de a pares, y luego de haber hecho el adecuado abordaje teórico del contenido envuelto y una serie de aplicaciones prácticas del mismo. Para fomentar la implicancia de los jóvenes en la situación de aprendizaje planteada se propone brindar un puntaje adicional de 0 a 10 puntos en la calificación del segundo parcial teórico-práctico, según los docentes evaluen la correcta resolución paso a paso y la justificiación de los resultados hallados.

Esta tarea está pensada para implementarse antes de culminar el primer cuatrimestre del año académico en curso.

2.1 Enunciado de la actividad

Un bioquímico ha preparado un plato de Petri con una muestra de bacterias. El medio de cultivo empleado ha sido LB-Agar, dejando la placa en un lugar oscuro y cálido durante seis días.

Transcurrido ese plazo el bioquímico ha observado con un microscopio la placa notando que los microorganismos se han reproducido en gran medida.

Para reproducir una situación similar a la experimentada, le pedimos que ejecute la presente sintaxis en R Studio Cloud.

```
#Simulación de una colonia de bacterias en un plato de Petri
plot(x=c(-2,2),y=c(-2,2),type ="n", xlab="",ylab="",asp =1,axes=FALSE)
#dibuja el plato de Petri
curve(sqrt(4-x^2),from=-2,to=2,col="darkgray",lwd=4,add =TRUE)
curve(-sqrt(4-x^2),from=-2,to=2,col="darkgray",lwd=4,add=TRUE)
#resalta los cuadros internos
```

```
rect(xleft=-1,ytop=1.5,xright=1,ybottom=-1.5,border=NA,col="khaki")
rect(xleft=-1.5,ytop=1,xright=1.5,ybottom=-1,border=NA,col="khaki")
#dibuja el cuadriculado
abline(v=seq(from=-2,to=2,by=0.5),lty=3,lwd=1.5,col="steelblue")
abline(h=seq(from=-2,to=2,by=0.5),lty=3,lwd=1.5,col="steelblue")
#dibuja las bacterias
N<-100
x<- runif(n=N*2,min=-2,max=2)
y<- runif(n=N*2,min=-2,max=2)
xd<-x[which(x^2+y^2 <= 3.8)]
yd<-y[which(x^2+y^2 <= 3.8)]
points (xd,yd,pch=19,col="darkolivegreen",cex=rnorm(n=length(xd),mean=0.6,sd=0.1))
#cuenta las bacterias
cuenta<- function(x0,x1,y0,y1) {cnt<- sum(as.numeric((xd>=x0 & xd<x1 & yd>=y0 & yd<y1)))
text(x=mean(c(x0,x1)),y=mean(c(y0,y1)),as.character
(cnt),col="black",cex=1.2,family="mono",font=2)
}
for(x in c(-1,-0.5,0,0.5))
for(y in c(1,0.5,0,-0.5,-1,-1.5))
cuenta(x,x+0.5,y,y+0.5)
for(y in c(0.5,0,-0.5,-1)) {
cuenta(-1.5,-1,y,y+0.5)
cuenta(1,1.5,y,y+0.5)
}
```

El profesional está interesado en estudiar el número de bacterias que se distribuyen en cada uno de los 32 cuadrantes de igual superficie de su placa y supone que el mismo es una variable aleatoria con distribución de Poisson. A continuación, les pedimos:

a. Empleando la imagen que ha observado el bioquímico en su microscopio y que han simulado en R Studio Cloud, efectúe el test de hipótesis correspondiente para probar que efectivamente el número de bacterias por recuadro es una variable con distribución probabilística de Poisson.

b. En caso de que el profesional esté en lo correcto, brinde una estimación del número promedio de bacterias por cuadrante registradas al cabo de seis días de cultivo.

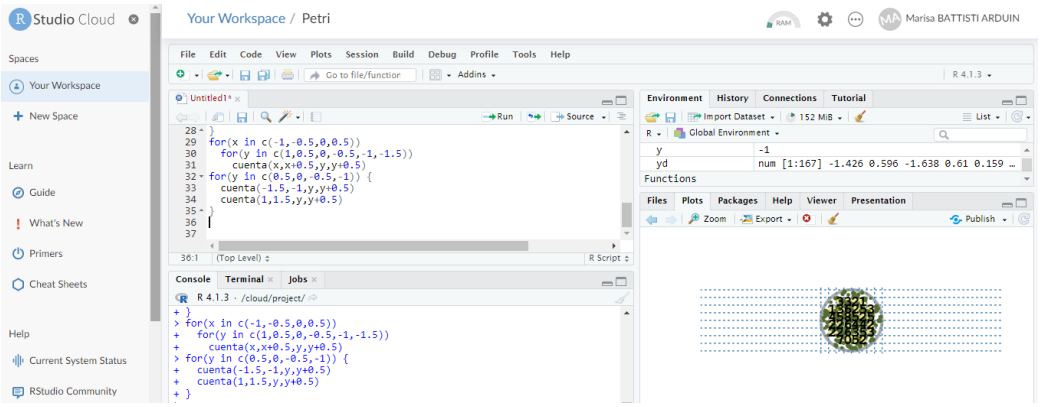
2.2 Resolución de la actividad

Para la resolución de la actividad los estudiantes deberán comenzar por hacer una lectura detallada del enunciado y ejecutar el código facilitado en la aplicación *on line* de R Studio (Figura 2.1).

Efectuando “Zoom” sobre la ventana Plots, podrán observar una imagen con la caja de Petri con 100 bacterias simuladas y distribuidas al azar sobre ella, dividiendo la caja en una serie de compartimientos llamados cuadros, y contabilizando por cada cuadro la cantidad de microorganismos

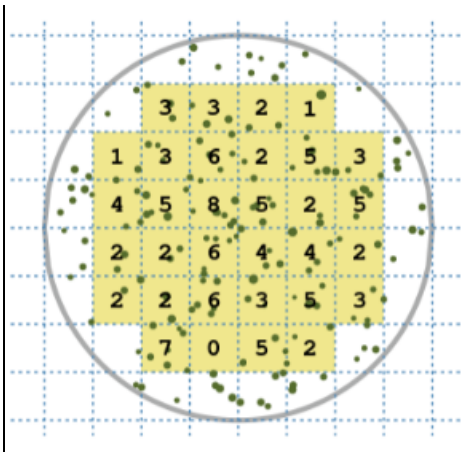
situados en ella (Figura 2.2).

Figura 2.1. Vista del R Studio luego de ejecutada la sintaxis



Fuente: Elaboración propia

Figura 2.2. Vista Zoom de la ventana Plots obtenido en R Studio



Fuente: Elaboración propia

Al ser una simulación con una distribución pseudo aleatoria de las bacterias, cabe aclarar que sucesivas ejecuciones del código generarán resultados similares pero no iguales.

A continuación, el estudiantado deberá definir la variable aleatoria bajo estudio y formular las hipótesis del contraste estadístico adecuado para comprobar la suposición del bioquímico.

Iniciarán definiendo la variable aleatoria discreta X que representa el “número de bacterias por cuadrante de la placa de Petri luego de 6 días de cultivo”. El contraste estadístico adecuado será la prueba de Bondad de Ajuste de Pearson, la cual sirve para evaluar qué tan bien se ajustan los datos experimentales a una distribución de probabilidad en particular. El test se basa en el nivel de ajuste existente entre la frecuencia de ocurrencia de las observaciones en la muestra recabada y las frecuencias esperadas que se obtienen a partir de la distribución hipotética (Walpole et al, 2012). Si la discrepancia entre los valores es de tal magnitud que pudiera deberse al azar, se concluye que la muestra puede haber sido extraída de una población con la distribución supuesta (Daniel, 2002).

Las hipótesis que requerirán establecer para resolver esta situación problemática son las siguientes.

$H_0 : X \sim Poisson(\lambda)$ (es cierta la suposición del bioquímico)

$H_1 : No \quad H_0$ (no es cierta la suposición del bioquímico)

En la cual λ corresponde al parámetro de intensidad de la distribución de Poisson, ya abordada con anterioridad en la asignatura, y de la que saben además que tanto su media como su varianza es igual a dicho parámetro, es decir: $\mu_X = \sigma_x^2 = \lambda$.

A partir de la imagen deberán construir una tabla con las diferentes clases (clase i) y los valores observados correspondientes (o_i). En este sentido, deberán contabilizar y resumir los datos referentes a la cantidad de cuadros que no tienen bacteria alguna, la cantidad de cuadros que tienen una sola bacteria, y así hasta culminar, agrupando convenientemente en la última clase con el número de cuadros que tienen cinco bacterias o más (véase Tabla 2.1).

Tabla 2.1. Conteo de número de bacterias y número de recuadros (valores observados)

Nro. de bacterias	Nro. de recuadros (o_i)
0	1
1	2
2	9
3	6
4	3
5 o más	11

Fuente: Elaboración propia

El estadístico Chi-cuadrado del test estadístico se define como $\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$ cuya distribución de probabilidad es Chi-cuadrado con $\upsilon = k - r - 1$ grados de libertad y siendo k la cantidad de clases y r la cantidad de parámetros que requieren estimarse. En el cálculo de dicho estadístico, intervienen las denominadas frecuencias esperadas, e_i , que se calculan multiplicando la probabilidad (p_i) de pertenecer a la clase i-ésima de la tabla bajo la premisa planteada en la hipótesis nula, por n , la cantidad de datos de la muestra.

En este problema, estas probabilidades deberán computarlas recurriendo a la distribución de Poisson.

El alumnado recordará que $X \sim Poisson(\lambda) \Leftrightarrow f(x) = \frac{e^{-\lambda}.\lambda^x}{x!}, x = 0,1,2...$, motivo por el cual, en este problema precisarán estimar previamente el valor de λ , ya que es desconocido. Atendiendo a que este valor es la media de la distribución de probabilidad, resultará estimado a partir de los datos de la

muestra, sumanado los productos de multiplicar cada valor de x por su frecuencia y dividiendo el total entre la suma de todas las frecuencias (Daniel, 2002) :

$$\hat{\lambda} = \frac{0.(1)+1.(2)+2.(9)+3.(6)+4.(3)+5.(11)}{32} = 3{,}6563 \quad .$$

En efecto, computarán las probabilidades puntuales con el valor del parámetro estimado y luego obtendrán las frecuencias esperadas.

$$p_1 = P(X=0) = f(0) = \frac{e^{-3{,}6563}.(3{,}6563)^0}{0!} = 0{,}0258 \quad \Rightarrow e_1 = p_1.n = (0{,}0258).(32) = 0{,}8265$$

$$p_2 = P(X=1) = f(1) = \frac{e^{-3{,}6563}.(3{,}6563)^1}{1!} = 0{,}0944 \quad \Rightarrow e_2 = p_2.n = (0{,}0944).(32) = 3{,}0220$$

$$p_3 = P(X=2) = f(2) = \frac{e^{-3{,}6563}.(3{,}6563)^2}{2!} = 0{,}1726 \quad \Rightarrow e_3 = p_3.n = (0{,}1726).(32) = 5{,}5246$$

$$p_4 = P(X=3) = f(3) = \frac{e^{-3{,}6563}.(3{,}6563)^3}{3!} = 0{,}2104 \quad \Rightarrow e_4 = p_4.n = (0{,}2104).(32) = 6{,}7331$$

$$p_5 = P(X=4) = f(4) = \frac{e^{-3{,}6563}.(3{,}6563)^4}{4!} = 0{,}1923 \quad \Rightarrow e_5 = p_5.n = (0{,}1923).(32) = 6{,}1545$$

$$p_6 = 1-(p_1+p_2+p_3+p_4+p_5) = 0{,}3043 \quad \Rightarrow e_6 = p_6.n = (0{,}3043).(32) = 9{,}7392$$

En estos cálculos podrán emplear por ejemplo el software R Studio, con la función “dpois()” que permite obtener probabilidades puntuales en la distribución de Poisson (Figura 2.3). Luego, tendrán la posibilidad de añadir al cuadro anterior una nueva columna, resultando la Tabla 2.2. Recurriendo a las conceptualizaciones teóricas nuevamente, es necesario que se cumpla la condición de que las frecuencias esperadas sean mayores o iguales a 5 para todas las clases. Por esta razón, surgirá la necesidad de que agrupen clases adyacentes pues, en este caso, las clases con 0 y 1 bacterias poseen frecuencias esperadas inferiores a 5 unidades. Reagrupando, resultarán $k = 4$ clases como se resume en la Tabla 2.3.

Figura 2.3. Vista de la función “dpois()” y sus resultados en R Studio

```
> dpois(0, lambda=3.6563, log = FALSE)
[1] 0.0258279
> dpois(1, lambda=3.6563, log = FALSE)
[1] 0.09443455
> dpois(2, lambda=3.6563, log = FALSE)
[1] 0.1726405
> dpois(3, lambda=3.6563, log = FALSE)
[1] 0.2104085
> dpois(4, lambda=3.6563, log = FALSE)
[1] 0.1923292
```

Fuente: Elaboración propia

Tabla 2.2. Frecuencias observadas y frecuencias esperadas

Nro. de bacterias	Nro. de recuadros (o_i)	Frecuencias esperadas (e_i)
0	1	0,8265
1	2	3,0220
2	9	5,5246
3	6	6,7331
4	3	6,1545
5 o más	11	9,7392

Fuente: Elaboración propia

Tabla 2.3. Frecuencias observadas y frecuencias esperadas con agrupación de celdas adyacentes

Nro. de bacterias	Nro. de recuadros (o_i)	Frecuencias esperadas (e_i)
0 a 2	**12**	**9,3731**
3	6	6,7331
4	3	6,1545
5 o más	11	9,7392

Fuente: Elaboración propia

De los cálculos anteriores resultará el valor del estadístico de prueba evaluado en los datos observados.

$$\chi^2_{obs} = \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i} = \frac{(12-9{,}3731)^2}{9{,}3731} + \frac{(6-6{,}7331)^2}{6{,}7331} + \frac{(3-6{,}1545)^2}{6{,}1545} + \frac{(11-9{,}7392)^2}{9{,}7392} = 2{,}5961$$

Se sabe que el estadístico sigue la distribución Chi-cuadrado con $\upsilon = k - r - 1 = 4 - 1 - 1 = 2$, entendiendo que el número de parámetros estimados es $r = 1$, ya que debieron estimar el parámetro lambda de la distribución de Poisson. En suma, tendrán que llevar a cabo el cálculo del p-value para esta prueba, el cual podrán obtener empleando la función “pchisq()” de R Studio.

Resultando $p - value = P(\chi^2 \geq \chi^2_{obs}) = P(\chi^2 \geq 2{,}5961) = 0{,}7269$, deberán optar acertadamente por el no rechazo de la hipótesis nula del test, pudiendo concluir que no existen evidencias suficientes para afirmar que el número de bacterias no se distribuye Poisson, y por ende, el profesional está en lo cierto con su suposición.

Para culminar la resolución de la actividad, precisarán resolver el inciso (b) brindando una estimación del número promedio de bacterias por cuadrante del cultivo, obteniendo $\hat{\lambda} = 3{,}6563$.

3. Resultados esperados y conclusiones

El contraste de Bondad de Ajuste es un recurso de elevada relevancia pues gran cantidad de los procedimientos estadísticos utilizados en la praxis, dependen en un sentido teórico de la suposición de que las observaciones reunidas procedan de una distribución de probabilidad específica (Walpole et al, 2012).

Se considera que la importancia del contenido amerita una actividad diferente como cierre de la

asignatura, que fomente la utilización de software en la resolución de un problema con datos propios del contexto en que se desempeña el alumnado y las ventajas que involucra la simulación en estadística.

Se pretende seguir incorporando actividades de estas características, pensando en nuevas propuestas didácticas para el abordaje de los contenidos, con entornos motivadores y facilitadores de los procesos de enseñanza y aprendizaje, propiciando la metacognición de saberes.

Bibliografía

Daniel, W. (2002). *Bioestadística: Base para el análisis de las ciencias de la salud*. Limusa Wiley.

Grima, P. y Riba, A. (1995). La simulación y la enseñanza de la estadística: casos prácticos. *Estadística española, 37* (140), pp. 409-434.

Libedinsky, M. (2016). *La innovación educativa en la era digital*. Paidós.

Walpole, R.; Myers, R.; Myers, S. y Ye, K. (2012). *Probabilidad y estadística para ingeniería y ciencias.* Pearson.