

INFORME DEL CURSO:¹

BIOESTADÍSTICA – FACULTAD DE CIENCIAS – SEGUNDO SEMESTRE 2016

Información general

El presente informe resume las características del curso de Bioestadística dictado en la Facultad de Ciencias en el segundo semestre de 2016. Lo más destacable respecto de ediciones anteriores es la introducción de la enseñanza de software estadístico R, su inclusión en la evaluación, y la modalidad de aprobación que incluyó la posibilidad de exoneración total del examen final.

Se planeó realizar este informe con el fin de presentar en los distintos institutos las innovaciones realizadas y los resultados obtenidos.

Docentes

Los docentes del curso fueron:

- Teóricos: Ernesto Mordecki (responsable) y Juan Kalemkerian,
- Prácticos: Andrés Sosa (coordinador), Verónica Rumbo y Gabriel Illanes.

Contenidos del Curso

El curso se estructuró en dos grandes temas: probabilidad y estadística. Más en detalle, contó con el dictado de 10 temas, organizados en 10 repartidos de práctico:

PROBABILIDAD:

1. Conteo y probabilidad.
2. Probabilidad condicional e independencia.
3. Variables aleatorias discretas.
4. Variables aleatorias continuas.
5. Esperanza y varianza en variables aleatorias discretas.
6. Esperanza y varianza en variables aleatorias continuas.

ESTADÍSTICA:

7. Intervalos de confianza.
8. Tests paramétricos.
9. Boxplot y test de ajuste.
10. Tests de aleatoriedad, independencia y Chi cuadrado.

Carga horaria y Modalidad

La carga horaria total del curso fue de 90 horas presenciales, divididas en dos clases teóricas y dos clases prácticas semanales de 90 minutos de duración.

La dinámica de las clases teóricas consistió principalmente en clases expositivas donde el objetivo era motivar, explicar y desarrollar los conceptos detallados en el cronograma del curso. En general, se omitieron las demostraciones, salvo aquellas más accesibles. En cambio, las clases prácticas consistían en ejercicios discutidos en el pizarrón, y en tiempo para trabajar en pequeños grupos de estudiantes donde el docente intentaba participar en la elaboración del razonamiento. Creemos que es importante contar con un plantel docente en la materia que apoye y estimule dichas dinámicas de clase.

¹Diciembre de 2016. Preparado por el equipo docente.

Plataforma Eva

La Plataforma Eva es una herramienta de mucha utilidad para fomentar en los estudiantes el seguimiento del curso fuera del aula.

Las principales razones por las que se optó usar la Plataforma Eva son las siguientes:

- El material (bibliografía y prácticos, por ejemplo) subido a la Plataforma es fácil de ordenar para los docentes (fecha, tema, etc), y por lo tanto, de fácil acceso para los estudiantes.
- La información sobre el temario dictado cada clase es fácil de actualizar para los docentes, y accesible para los estudiantes.
- Los foros de la Plataforma son muy convenientes para hacer notificaciones a todos los estudiantes por parte de los docentes. Asimismo, los estudiantes los aprovechan para realizar consultas, las cuales son respondidas por los docentes, o por los mismos estudiantes.
- Tiene opciones para implementar evaluaciones del curso y de los docentes por parte de los alumnos.

Número de estudiantes

El curso se dicta para distintas licenciaturas dentro de la facultad de Ciencias: Biología, Bioquímica, Biología Humana, Geografía y Geología. También admite estudiantes de la Facultad de Medicina.

Según las actas del curso se inscribieron:

- 136 estudiantes de biología y bioquímica
- 7 estudiantes de biología humana
- 5 estudiantes de geografía
- 25 estudiantes de medicina

De estos 173 estudiantes inscriptos, sólo 112 se presentaron a alguna de las cuatro instancias de evaluación. Cabe destacar que entre los 112 sólo 2 eran de Facultad de Medicina. Tenemos entonces de los 148 inscriptos de Facultad de Ciencias, la participación efectiva se redujo a 110 estudiantes de nuestra facultad.

Introducción del Software R

La idea en este curso fue presentar el software R a los estudiantes con el fin de que conozcan las ventajas que presenta trabajar con él (o con otros programas de similares características). En la actualidad, el software R es el estándar para trabajos en estadística a nivel académico. Creemos que es una muy buena herramienta computacional, que puede ser utilizada a nivel académico y profesional.

Como innovación a los prácticos se presentó un tutorial de R, donde se explicaban conceptos básicos del software, y se relacionaban con ejercicios prácticos para aplicar los conocimientos adquiridos. Este tutorial nos permitió hacer la introducción al R en los salones usuales de clase.

Las clases destinadas a R se realizaron (en su mayoría) en los salones 311 y 312 de computadoras. En total se dedicaron una clase de teórico y cinco clases de práctico para trabajar con el programa. En estas clases, la idea era relacionar los conceptos teóricos previamente desarrollados con la manipulación del programa vía los comandos a utilizar. Luego se procedía a que los estudiantes ejecutaran un script preparado por los docentes, o realizaran uno propio.

Se analizaron casos de cursos similares (donde la idea es enseñar probabilidad y estadística usando una herramienta computacional), por lo cual al inicio del curso se consideró que, para generar motivación en los estudiantes en el software, la evaluación del curso debía contemplar el manejo del software R. Por dicha razón es que surgió la idea de evaluar informes de trabajo computacional.

Criterios de evaluación

En esta edición del curso, por primera vez, se concedió el derecho a exonerar la materia por parte de los estudiantes, exigiéndose además un mínimo para aprobar el curso.

Las instancias de evaluación fueron cuatro, divididas en dos informes de 15 puntos cada uno, y dos parciales de 35 puntos cada uno (100 puntos en total). El requerimiento mínimo para la aprobación del curso fue de 40 puntos y el requerimiento mínimo para la exoneración fue de 80 puntos.

Para la realización de los informes, propusimos que se formaran grupos de hasta tres estudiantes. El primer informe debía entregarse en la novena semana del semestre, y el segundo informe en la decimocuarta semana. El primer parcial se realizó en la octava semana del semestre (el sábado 8 de octubre de 9:00 a 11:00), y el segundo parcial en la decimoquinta semana (el sábado 26 de noviembre también de 9 a 11 horas).

PRIMER PARCIAL: Una prueba de dos horas con tres ejercicios, abarcando temas de probabilidad y variables aleatorias. Se presentaron 112 estudiantes y resultó un promedio de 23.8 (en 35) puntos.

PRIMER INFORME: El objetivo fue entregar un informe impreso en el cual se pedía, básicamente, simular e imprimir dos gráficos a partir de datos: el primero de una distribución asignada al grupo (que podía ser exponencial, uniforme o normal), y el segundo verificaba empíricamente el teorema central del límite. Se presentaron 45 informes (107 estudiantes), el promedio de notas fue de 11.4 (en 15).

SEGUNDO INFORME Con características similares al primero, tuvo como objetivo la manipulación de una base de datos por parte de los estudiantes en el software R, y la aplicación de varios test de hipótesis a dichas bases. Se presentaron 46 informes (108 estudiantes), el promedio de notas fue de 12.1.

SEGUNDO PARCIAL: Una prueba de dos horas con tres ejercicios, abarcando temas de estadística. Se presentaron 97 estudiantes y resultó un promedio de 25.3 puntos.

Resultados del Curso

Para el análisis se separaron los estudiantes de acuerdo a su desempeño en el curso. Los tres grupos que se analizarán son:

- Exoneraron el curso (EC);
- Aprobaron el curso (AC);
- No aprobaron el curso (NAC).

El grupo EC está formado por 48 estudiantes. Para acceder a este grupo era necesario realizar las cuatro evaluaciones del curso y obtener más de un 80% del total de puntos. La nota mínima de exoneración era 9. Estas notas se dividieron en 16 estudiantes aprobaron con 9, 17 aprobaron con 10, 10 aprobaron con 11 y 5 aprobaron con 12.

El grupo AC está formado por 45 estudiantes. De estos estudiantes, 37 se presentaron a todas la evaluaciones del curso y en el caso de los restantes estudiantes se verificó que en alguna evaluación no se presentaron, principalmente el segundo informe o el segundo parcial. Una razón de esto es la percepción de no alcanzar los puntos para la exoneración pero si ya encontrarse con la aprobación del curso.

El grupo NAC está formado por los restantes 19 estudiantes. En este grupo, solamente 2 estudiantes realizaron todas las actividades no alcanzando el nivel mínimo exigido. Los restantes 17 estudiantes no se presentaron (como mínimo) al segundo parcial.

Evaluación del Curso por parte de los estudiantes

Se realizó la encuesta que propone el Centro de Matemática con algunas preguntas nuevas con el fin de obtener información de la perspectiva de los estudiantes en la utilización del R y la aprobación del

curso. Dicha evaluación permaneció activa durante diez días en la Plataforma Eva. Fue de carácter completamente anónima.

Esta evaluación fue completada por 41 estudiantes. Desde el punto de vista de los docentes la evaluación hacia el curso en general y hacia su trabajo individual fue de buena manera y se encuentra a disposición de quien la desee.

Entre los resultados más destacados se encuentra la evaluación positiva de los estudiantes con respecto a la modalidad de aprobación del curso y a la introducción del R en el curso. En ambas preguntas el 68% de los estudiantes lo considera buena o muy buena. Sin embargo, en las preguntas abiertas sobre estos temas se obtuvieron críticas de variados temas (incluso algunas contradictorias entre si) como lo son el carácter no individual de los informes, las exigencias de exoneración y el tiempo destinado a R en el desarrollo del curso.

Resultados del Examen de Diciembre

El primer examen se rindió el viernes 9 de diciembre y constó de tres ejercicios. Se presentaron 28 estudiante, aprobaron 24 y no aprobaron 4. Estos estudiantes se dividen en las siguientes categorías:

- Aprobaron el curso - Aprobaron el Examen: 16 estudiantes
- Aprobaron el curso - No Aprobaron el Examen 2 estudiantes.
- No cursaron en 2016 - Aprobaron el Examen: 8 estudiantes.
- No cursaron en 2016 - No Aprobaron el Examen: 2 estudiantes.

Conclusiones y Aspectos a desarrollar

La evaluación global es que el núcleo de estudiantes que asistieron al curso pudieron aprovecharlo, y un grupo importante de ellos lo aprobó durante el semestre.

1. Creemos que la introducción del R y el método de evaluación fue exitoso. Los estudiantes trabajaron durante el curso, tuvimos buenos porcentajes de exoneración y de aprobación. A su vez, las evaluaciones del curso por parte de los estudiantes son similares a nuestro parecer en cuanto al desarrollo del mismo.
2. El material teórico fue deficiente. Contamos sólo con notas parciales de algunos temas, no teníamos libros de referencia claros para que los estudiantes pudieran recuperar clases o profundizar algún tema.
3. Detectamos que es un problema dictar las clases prácticas antes que las teóricas. En ese sentido entendemos que sería bueno en la mañana dictar primero las clases teóricas y luego las prácticas.
4. Tuvimos limitaciones para ir al laboratorio a explicar el uso del software R, dado la poca disponibilidad de salones con computadoras. Y la asignación de dos salones diferentes para un mismo horario con un solo docentes no nos parece adecuado.

En las páginas siguientes se anexa las letras de los informes y los parciales.

Nombre:	CI:	Carrera
---------	-----	---------

Primer parcial de Bioestadística - 8 de octubre de 2016

Ejercicio 1 (15 puntos):

Un acuicultor tiene un estanque con 7 salmones machos y 5 hembras. La probabilidad de que un macho esté infectado por un virus es de 0,2, mientras que para las hembras es de 0,01.

- (a) Calcular la probabilidad de que un ejemplar extraído al azar esté infectado por el virus.
- (b) Dado que un ejemplar está contaminado por el virus, calcular la probabilidad de que sea macho.
- (c) Calcular la probabilidad de que, al extraer 3 peces simultáneamente, resulten dos hembras y un macho.

Ejercicio 2 (10 puntos):

La probabilidad de éxito de la fecundación in vitro es de 0,3.

- (a) Calcular la probabilidad de que en 10 intentos resulten exitosos a lo sumo dos de ellos.
- (b) Calcular la esperanza del número de intentos necesarios para tener un primer éxito.

Ejercicio 3 (10 puntos):

La medida de un determinado experimento da como resultado una variable aleatoria $X \sim \mathcal{N}(1, 4)$.

- (a) Calcular la probabilidad de que la medida efectuada resulte mayor que 3.
- (b) Determine c para que un intervalo de la forma $(1 - c, 1 + c)$ comprenda el 90 % de los resultados.

PARA USO DOCENTE.			
Ej. 1	Ej. 2	Ej. 3	TOTAL

Soluciones

Ejercicio 1: (15 puntos)

(a) Es

$$\frac{7}{12} \times 0,2 + \frac{5}{12} \times 0,01 = 0,121.$$

(b) Es C contaminado, M , macho:

$$P(M | C) = \frac{P(M | C)P(C)}{P(C)} = \frac{0,2 \times 7/12}{0,121} = 0,964.$$

(c) Es

$$\frac{C_2^5 C_1^7}{C_3^{12}} = 0,318.$$

Ejercicio 2 (10 puntos): La probabilidad de éxito de la fecundación in vitro es de 0,3.

(a) Es

$$\begin{aligned} & C_0^{10} (0,7)^{10} (0,3)^0 + C_1^{10} (0,7)^9 (0,3)^1 + C_2^{10} (0,7)^8 (0,3)^2 \\ &= (0,7)^8 [(0,7)^2 + 10 \times 0,7 \times 0,3 + 45 \times (0,3)^2] = 0,383. \end{aligned}$$

(b) Es $1/p = 1/0,3 = 3,33$.

Ejercicio 3 (10 puntos): La medida de un determinado experimento da como resultado una variable aleatoria $X \sim \mathcal{N}(1, 4)$.

(a) Calcular la probabilidad de que la medida efectuada resulte mayor que 3. Para $Z \sim \mathcal{N}(0, 1)$

$$P(X > 3) = P(X - 1 > 2) = P((X - 1)/2 > 1) = 1 - P(Z < 1) = 1 - 0,84 = 0,16.$$

(b) Determine c para que un intervalo de la forma $(1 - c, 1 + c)$ comprenda el 90 % de los resultados. Para $Z \sim \mathcal{N}(0, 1)$, tenemos

$$P(X \in (1 - c, 1 + c)) = P(X - 1 \in (-c, c)) = P(Z \in (-c/2, c/2)) = 0,90.$$

De la tabla normal $c/2 = 1,64$ entonces $c = 3,28$.

Primer Informe de Bioestadística

Parte 1. Indicar la distribución correspondiente al grupo, su media y su desvío estándar. Denotemos por F esta distribución.

- (a) Generar una muestra aleatoria simple de 1000 datos con distribución F .
- (b) Presentar los datos en un histograma, y en el mismo gráfico y superpuesto, la densidad correspondiente a la distribución F .
- (c) Calcular la media empírica \hat{a} de la muestra, el desvío estándar empírico $\hat{\sigma}$ de la muestra.
- (d) Determinar la media teórica a y el desvío estándar teórico σ . Calcular el valor estandarizado de la media empírica, definido como $(\hat{a} - a)/\sigma$.

Parte 2. Simular ahora 100 muestras de 1000 datos como las de la parte 1.

- (a) Calcular los correspondientes valores estandarizados $(\hat{a}_1 - a)/\sigma, \dots, (\hat{a}_n - a)/\sigma$.
- (b) Presentar los datos de esta segunda muestra en un histograma, y en el mismo gráfico y superpuesto, la densidad normal estándar.

Instrucciones

1. El Informe se entrega impreso.
2. Debe contener las siguientes secciones:
 - Título e integración del equipo (nombre y cédula de cada uno de los integrantes)
 - Introducción: Se describe brevemente el problema
 - Metodología: Se detalla el procedimiento para resolverlo
 - Resultados: Se exponen los resultados, con ayuda de gráficos
 - Conclusiones: Se comentan los resultados, y se plantean nuevas preguntas
 - Bibliografía: Se indica la bibliografía y las páginas de internet utilizadas
 - Anexos: se explican los códigos o comandos utilizados.

Sugerencias: Comandos que pueden ser de utilidad

- `runif(n, a, b)` genera una muestra de tamaño n con distribución uniforme en el intervalo $[a, b]$
- `rexp(n, λ)` genera una muestra de tamaño n con distribución exponencial de parámetro λ .
- `rnorm(n, μ, σ)` genera una muestra de tamaño n con distribución normal con parámetros (μ, σ^2)
- `mean(vector)` calcula la media empírica de un vector de datos `vector`
- `sd(vector)` calcula el desvío estándar empírico de un vector de datos `vector`
- `hist(vector, freq=FALSE)` calcula el histograma de un vector de datos `vector`. Para cambiar la cantidad de intervalos definidos por defecto, consultar en internet
- `for` Método iterativo o bucle.

A su vez, recomendamos fuertemente utilizar la ayuda que brinda el R mediante los comandos `help()` y `help.search()` como también la búsqueda en Internet.

Segundo Informe de Bioestadística

La base de datos consta de dos series que se denominan Muestra A y Muestra B de 1000 datos cada una. Los objetivos son investigar propiedades de cada una de las muestras, y estudiar si ambas muestras provienen de la misma distribución. Todos los test son a nivel $\alpha = 0.05$.

Parte 1.

- Calcular el promedio, el desvío estándar, graficar un box-plot y un histograma de cada una de las muestras.
- Establecer tres intervalos de confianza de niveles 90%, 95%, y 99% para la media de la muestra A.
- Testear la independencia de cada una de las muestras.

Parte 2.

- Testear si las muestras son independientes entre sí.
- Realizar un test de diferencia de medias.
- Realizar el test de Kolmogorov-Smirnov para dos muestras.

Instrucciones

- El Informe se entrega impreso. En cualquiera de los prácticos del día jueves 17 de noviembre.
- Debe contener las siguientes secciones:
 - Título e integración del equipo (nombre y cédula de cada uno de los integrantes).
 - Introducción: Se describe brevemente el problema y se detalla el marco teórico necesario.
 - Resultados: Se exponen los resultados, con ayuda de gráficos.
 - Conclusiones: Se comentan los resultados (relacionándolos con el marco teórico), y se plantean nuevas preguntas.
 - Bibliografía: Se indica la bibliografía y las páginas de internet utilizadas
 - Anexos: Se comentan el código escrito y los comandos utilizados.

Sugerencias: Comandos que pueden ser de utilidad

- Para importar datos en RStudio: **Tools** -> **Import Dataset** -> **From Local File...** se debe seleccionar el archivo correspondiente y elegir las opciones adecuadas.
- Para importar datos de manera genérica: `read.table("Grupo X datos.txt", sep = " ", header = TRUE)` comando para cargar una base de datos denominada *Grupo X datos.txt*. Hay que tener cuidado que el archivo se encuentre en el directorio de trabajo.
- `boxplot(vector)` genera un boxplot a partir de los datos de `vector`.
- `rank(vector)` devuelve el vector de rangos de `vector`.
- `cor.test(rangos, 1:n, method = "spearman", alternative = "two.sided")` realiza un test de rangos de Spearman para el vector de rangos `rangos`. El comando retorna, entre otras cosas, el valor del estadístico y el p-valor.
- `cor.test(rangos-a, rangos-b, method = "spearman", alternative = "two.sided")` realiza un test de rangos de Spearman de dos muestras para los vectores de rangos `rangos-a` y `rangos-b`.
- `ks.test(vector-a, vector-b, alternative = "two.sided")` realiza el test de Kolmogorov-Smirnov para los datos `vector-a` y `vector-b`. El comando retorna, entre otras cosas, el valor del estadístico y el p-valor.

A su vez, recomendamos fuertemente utilizar la ayuda que brinda el *R* mediante los comandos `help()` y `help.search()` como también la búsqueda en Internet.

Nombre:	CI:	Carrera
---------	-----	---------

Segundo parcial de Bioestadística - 26 de noviembre de 2016

Realizar todos los tests a nivel $\alpha = 0.05$.

Ejercicio 1 (7 puntos): Los siguientes datos son una muestra aleatoria simple de una distribución con media μ y varianza σ^2 desconocidos:

3.47 0.69 -4.72 1.48 -0.27 0.24

- (a) Establecer la fórmula del estimador de la media y calcular su valor.
- (b) Establecer la fórmula del estimador de la varianza y calcular su valor.

Ejercicio 2 (14 puntos):

(a) En un experimento se obtiene la siguiente muestra aleatoria simple formada por cincuenta datos:

1 0 0 0 0 1 0 0 0 0 0 0 1 1 1 0 1 0 1 1 1 0 1 0 1 1 0 1 0 1 0 1 1 1 0 0 0 1 0 0 0 1 0 0 0 1 1 0 0 0

Sea p la probabilidad de obtener un 1 en cada dato. Realizar el test de hipótesis: $\begin{cases} H_0 : p \geq 1/2 \\ H_1 : p < 1/2 \end{cases}$.

(b) Se tiene una muestra aleatoria simple de $n = 100$ datos con media μ y varianza σ^2 desconocidos, de la cual se conoce que su promedio $\bar{X}_n = 0.3$ y el estimador de la varianza es $s_n^2 = 1.2$. Realizar el test de hipótesis: $\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu \neq 0 \end{cases}$.

Ejercicio 3 (14 puntos):

(a) Los siguientes datos corresponden a una variable de interés medida antes y después de un tratamiento:

antes	-5.56	2.27	-0.33	0.79	1.82
después	-2.37	-0.44	1.99	0.39	0.21

Realizar un test de hipótesis de Kolmogorov-Smirnov para detectar si las distribuciones de ambas muestras son iguales.

(b) Se quiere saber si la frecuencia de aparición de personas zurdas es igual para hombres y mujeres, para lo cual se recolectan los siguientes datos en una tabla de contingencia.

	derechos	zurdos	totales
hombres	43	9	52
mujeres	44	4	48
totales	87	13	100

Realizar un test de hipótesis para determinar si la frecuencia de zurdos no depende del género.

Soluciones

Ejercicio 1 (7 puntos):

(a) La fórmula es $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$, el resultado es $\bar{x}_6 = 0.15$.

(b) La fórmula es $s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$, el resultado es $s_6^2 = 7.39$.

Ejercicio 2 (14 puntos):

(a) Tenemos $\hat{p} = 21/50 = 0.42$. La región crítica es de la forma $RC = (0, c)$, donde

$$c = \frac{1}{2} - 1.645 \sqrt{\frac{1}{200}} = 0.3836.$$

Como $c < 0.42 = \hat{p}$, no rechazo la hipótesis nula.

(b) Ahora la región crítica es de la forma

$$RC = \left\{ \bar{X}_n \leq -1.96 \times \frac{\sqrt{1.2}}{\sqrt{100}} \right\} \cup \left\{ \bar{X}_n \geq 1.96 \times \frac{\sqrt{1.2}}{\sqrt{100}} \right\}.$$

Como $1.96 \times \frac{\sqrt{1.2}}{\sqrt{100}} = 0.215$ y $\bar{X}_n = 0.3 > 0.215$ se rechaza la hipótesis nula.

Ejercicio 3 (14 puntos):

(a) El estadístico de KS da $D = 0.4$. De la tabla vemos que $D \times 5 \times 5 = 10$ nos da un p-valor mayor que 0.357. No rechazo la hipótesis nula.

(b) Agregamos las frecuencias esperadas en la tabla en azul:

	derechos	zurdos	totales
hombres	43 / 45.24	9 / 6.76	52
mujeres	44 / 41.76	4 / 6.24	48
totales	87	13	100

El estadístico es

$$\chi_1^2 = \frac{(43 - 45.24)^2}{45.24} + \frac{(9 - 6.76)^2}{6.76} + \frac{(44 - 41.76)^2}{41.76} + \frac{(4 - 6.24)^2}{6.24} = 1.77$$

De la tabla de chi cuadrado con un grado de libertad tenemos que la región crítica es $\chi_1^2 > 3.84$. No se rechaza entonces la hipótesis nula.