

Una introducción al Análisis Exploratorio de Datos por medio de Google Analytics

Pablo Carranza y Jenny Fuentealba

pcarranza@unrn.edu.ar;

jenny.fuentealba@gmail.com

Universidad Nacional de Río Negro

Resumen

En el presente artículo nos interesamos en las posibilidades de enseñanza de nociones de base del Análisis Exploratorio de Datos en los primeros años de la enseñanza universitaria por medio del software conocido como Google Analytics. Presentamos primeramente una introducción general, seguida de una somera síntesis del marco epistemológico de referencia (Análisis Exploratorio de Datos), continuando con una descripción de la herramienta conocida como Google Analytics, para luego desarrollar algunos tópicos de interés sobre una experiencia llevada a cabo con estudiantes universitarios.

Palabras clave: análisis de datos, software, didáctica de la estadística.

Keywords: data analysis, software, didactics of statistics.

Abstract

In this paper we are interesting in the possibilities of teaching of principles of Data Exploratory Analysis in early years of university education using Google Analytics software. In a first time we present a general introduction of the epistemological frame (Data Exploratory Analysis) continuing with a description of tool known as Google Analytics and then we develop some topics concerning an experimentation carried out with students at university level.

1. Introducción

La enseñanza de la estadística a nivel de grado en Argentina se encuentra fuertemente influenciada por la corriente inferencial conocida como clásica o «frecuentista». Así, es usual encontrar en los programas de esta disciplina y para la gran mayoría de las carreras universitarias, temas tales como probabilidad (frecuentista), tests de hipótesis, intervalos de confianza, regresión lineal, tests χ^2 , etcétera.

Esta corriente estadística no está exenta de críticas y esto por varias razones, entre otras podemos citar por ejemplo, la que se refiere al modelo de descubrimiento científico subyacente en sus tests de hipótesis (Carver, 1978; Cox, 1946; Cumming, 2010; Meehl, 1978; Popper, 1959), o aquella que menciona las dificultades recurrentemente observadas en las interpretaciones, producto de la inadecuación de las respuestas suministradas con respecto a las preguntas que originalmente motivan el estudio estadístico (Carranza, 2009; Jaynes, 1984; Lecoutre, 2005; Régnier & Oriol, 2001; Vallecillos, 1997).

Otra fuente de cuestionamiento hacia la inferencia clásica se refiere a las limitaciones de su aplicación a ciertas prácticas profesionales. En fin, éstas y otras cuestiones resultaron de interés para los docentes de la cátedra de Estadística de las carreras de Licenciatura en Comercio Exterior y Licenciatura en Marketing de la Universidad Nacional de Río Negro (UNRN, Argentina), cátedra a cargo de uno de los autores del presente artículo. Así, nuestra intención no fue la de proponer una exclusividad de la enseñanza de la inferencia clásica sino la de una articulación con otras corrientes, tales como la inferencia bayesiana o el Análisis Exploratorio de Datos (AED).

Además de la mencionada respecto a la complementación entre enfoques estadísticos, otra motivación marcó el cursado de la cátedra, esta vez una de orden didáctico, y se refiere a nuestra intención de abordar temas estadísticos desde contextos próximos a la actividad profesional de

las carreras (Licenciatura en Comercio Exterior y Licenciatura en Marketing).

La confluencia de estas dos motivaciones nos llevó a proponer a los estudiantes la realización de un estudio estadístico sobre las visitas efectuadas a sitios web. Este tipo de contexto está fuertemente relacionado con la futura actividad profesional de los estudiantes (e–comercio, e–marketing, etc.) y es relativamente compatible con los principios del AED que deseábamos discutir con los estudiantes. En efecto, la creciente difusión del comercio electrónico y la creación de sitios web implican una producción importante y continua de datos multivariados de interés para la actividad comercial. Además por sus características generativas entre otras razones, estos datos resultan factibles de ser tratados desde el enfoque del AED, como veremos más adelante.

A modo ilustrativo, si una empresa regional desea exportar sus productos, una vía posible para ello es su difusión por medio de un sitio web. Así, no solo las consultas directamente efectuadas a través del sitio son de utilidad, sino que toda información referida a las visitas al sitio resulta de interés. Esto es, tener conocimiento de la tendencia de los lugares geográficos de las visitas, del tiempo de permanencia tanto en el sitio como en las páginas de los productos, e incluso saber si los visitantes son recurrentes o no, son cuestiones de suma utilidad a toda empresa que desee incursionar en actividades comerciales vía la web. Esta información permite por ejemplo el diseño de campañas de difusión así como la determinación de la aceptabilidad de sus productos a lo largo de diferentes regiones.

Estos datos no necesariamente responden a los requisitos planteados, por ejemplo, para un test de hipótesis. Por un lado, porque es muy probable que se carezca de hipótesis previas al estudio y por el otro lado, porque en muchos casos los datos preexisten a la intención del mismo.

El presente artículo propone entonces algunas

reflexiones sobre una experiencia llevada a cabo durante el primer cuatrimestre del año 2010 con veintisiete estudiantes de segundo año (edades entre 19 y 40 años) de las carreras de Comercio Exterior y de Licenciatura en Marketing de la UNRN, donde los estudiantes realizaron un estudio estadístico de las visitas a sitios web de la región. Comenzaremos presentando un resumen del marco epistemológico de referencia.

2. Marco epistemológico de referencia: el AED

Considerando que en todo proceso de enseñanza se produce una transposición didáctica (Chevallard, 1985), identificamos que, en nuestro caso, la misma se realiza en el seno de la cátedra. Esto es, el objeto epistemológico de referencia sufre transformaciones y/o recortes inevitables: se convierte, primero en objeto reconocido como necesario a ser enseñado, para luego devenir en objeto a enseñar.

Como decíamos, en nuestro caso esta transposición didáctica ocurre en el seno de la cátedra. Tal cual hemos mencionado en la introducción, el AED no forma parte del conjunto de objetos reconocidos como «necesarios a ser enseñados». Bien al contrario y tal cual lo comentamos anteriormente, en Argentina existe una fuerte dominancia de la inferencia clásica en la formación estadística de grado.

Los docentes de la cátedra (un titular y dos ayudantes) debimos entonces retener los elementos del AED a ser trabajados con los estudiantes. En esta selección consideramos el escaso o nulo contacto que los estudiantes podrían haber tenido con Estadística, ya sea en otras cátedras o en sus estudios secundarios. Resumiremos a continuación los elementos del AED considerados para la propuesta. Podría afirmarse que el AED es un enfoque estadístico relativamente reciente, al menos comparado con otros como la inferencia clásica o incluso la bayesiana. Un referente del mismo es sin dudas John Tukey

(1977) aunque existen precursores, sobre todo si se consideran los trabajos de Francis Galton sobre los estadísticos de orden (Pearson, 1930).

Caracterizar el AED no es una tarea sencilla, y esto por las especificidades mismas del enfoque. No responde a una estructura procedimental precisa, como puede ser el caso de la inferencia clásica y en particular con sus tests confirmatorios, o de la inferencia bayesiana con el uso del Teorema de Bayes como modelo normativo de resignificación de probabilidades (Gärdenfors *et al.*, 1988; Hacking, 2002; Jaynes, 1980). Bien al contrario, este enfoque se caracteriza por la diversidad de procedimientos a convocar en un estudio estadístico.

A pesar de las diferencias entre el AED y los otros enfoques mencionados, algunos autores (Good, 1983) no dudan en aproximarlos más a la corriente bayesiana que a la frecuentista, y esto por un aspecto común a las dos primeras y que se refiere a las probabilidades de las hipótesis.

Si bien esta cuestión relativa a las hipótesis está presente tanto en el enfoque bayesiano como en el AED, existen diferencias significativas en el lugar que ocupa en cada uno. En particular en el primero de ellos, la atención está centrada en la posibilidad de asignar probabilidades a las hipótesis, así el Teorema de Bayes normaliza el procedimiento abductivo (Carranza, 2012; Hacking, 2002; Peirce, 1932) de re-cálculo de los valores de probabilidades en función de la evidencia constatada. En el AED, en cambio, la atención se centra en la generación de hipótesis a partir del análisis de los datos (Diaconis, 1985; Good, 1983; Mast & Kemper, 2009; Tukey, 1977). Así, los métodos gráficos, la actitud exploratoria por parte del estadístico y el conocimiento que se tenga del contexto del problema son de fundamental importancia en la búsqueda de esas hipótesis o relaciones causales.

Otro aspecto importante del AED se refiere a la manera en que han sido obtenidos los datos. Si es bastante admitido en la corriente inferencial clásica, que los datos deben ser el fruto de un proceso de muestreo relativamente formal, en el AED suele

ocurrir que ellos preexistan a la intención del estudio estadístico e incluso es probable que continúen produciéndose durante el mismo. De ese modo, los datos no necesariamente responden a una intención predeterminada a estudiar sino que es el estadístico quien debe adaptarse a ellos, explorándolos y aplicando estrategias no siempre idénticas a fin de obtener información de los mismos.

La Figura 1 representa esquemáticamente los aspectos antes mencionados del AED. Este conjunto de características constituyó nuestra referencia epistemológica de base. Y si bien estos aspectos no fueron presentados aquí más que de manera resumida, ellos reflejan en esencia la filosofía del AED propuesta a los estudiantes de las Licenciatura en Comercio Exterior y de Marketing.

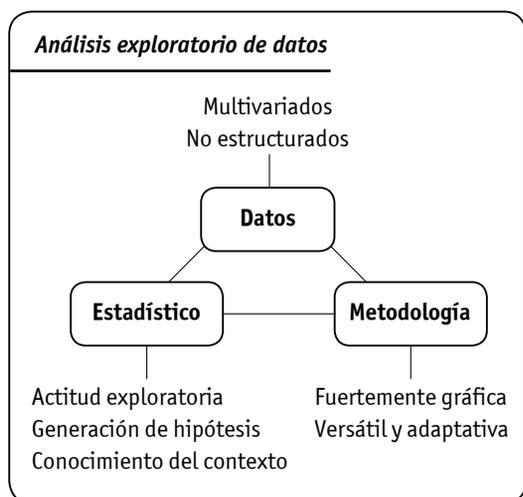


Figura 1.

Aspectos del AED retenidos para el estudio estadístico.

Nuestra referencia epistemológica estuvo compuesta entonces de tres polos. Uno de ellos se refiere a los datos. Estos suelen ser multivariados y obtenidos muy probablemente de manera no estructurada. El segundo polo se refiere a ciertas actitudes del estadístico: él debe estar dispuesto a explorar los datos desde puntos de vista diferentes y en general, basándose en representaciones

gráficas. Entre sus intenciones, se encuentra la de proponer explicaciones a los fenómenos identificados: el conocimiento del contexto debe poder contribuir a la generación de hipótesis explicativas. Por último, el tercer polo retiene elementos metodológicos. La exploración de los datos se apoya fuertemente en herramientas gráficas, esto implica una cierta experiencia sobre el uso de software de tratamiento y representación de datos.

Estos tres polos de nuestra referencia epistemológica serán recurrentemente citados en las secciones siguientes, cuando nos consagremos a la presentación del resumen de la experiencia llevada a cabo con los estudiantes. A continuación presentaremos el contexto del estudio estadístico acompañado de la herramienta metodológica de tratamiento de los datos.

3. Contexto general y software utilizado: análisis de visitas a sitios web por medio de Google Analytics

Dentro de un conjunto de posibilidades, retuvimos la del estudio de visitas a sitios web por medio del software conocido como Google Analytics. Dos principales razones motivaron esta elección; una se refiere a la pertinencia de la temática respecto a las profesiones de Comercio Exterior y de Marketing. La otra trata de características de la interface Google Analytics, ella es gratuita⁽¹⁾ y su uso requiriere escasos conocimientos específicos.

Particularmente la última de estas características nos resultó un aspecto no menor. En investigaciones anteriores (Carranza, 2009) hemos podido comprobar el gran esfuerzo que representa para los estudiantes articular tres dimensiones fundamentales para este tipo de estudios estadísticos, la semántica, la operatoria y la tecnológica. La primera se refiere al significado que porta un concepto, la segunda a los procedimientos o algoritmos que reglan las transformaciones numéricas en los que

se ve involucrado el mismo (cálculos, procesos para su representación gráfica, etc.). La tercera a la apropiación de las particularidades del software. De esta manera, el trabajo del control semántico, operatorio y tecnológico al mismo tiempo suele ser una tarea ardua para los estudiantes y en muchos casos fuente de muchos errores.

En el caso de integrar un software entonces, emerge una nueva dificultad. Si bien el control sobre los aspectos calculatorios se ven simplificados en apariencia, surgen dificultades referidas a la apropiación del software y a su instrumentalización (Engestrom, 1991; Rabardel, 1995). En este sentido, Google Analytics nos pareció más ergonómico que otros tales como la hoja de cálculo por ejemplo. Con la elección de este software, esperábamos entonces una menor incidencia de las dificultades vinculadas a la apropiación de las especificidades del software.

El principio de funcionamiento de Google Analytics es bastante elemental. Cada visita efectuada a un sitio web es registrada en una base de datos de Google Analytics. Esta visita queda registrada por medio de un conjunto de variables como por

ejemplo la zona geográfica (dirección IP), la hora de ingreso al sitio, el origen de la visita (buscador, sitio de referencia, etc.), el orden de las páginas visitadas, el tiempo de permanencia en cada una de ellas, las palabras claves utilizadas en caso de haber accedido a través de un buscador, etc.

Así, el administrador del sitio web accede a esos datos por medio de las herramientas disponibles en Google Analytics (gráficos, tablas, valores medios y/o porcentuales, etc.). En general entonces, el acceso es indirecto, bajo la forma de tablas de contingencias, indicadores o gráficos.

La Figura 2 representa uno de los gráficos obtenidos con Google Analytics. Este se refiere a la evolución de las visitas registradas durante un mes (18/10/11–17/11/11) al sitio web de la UNRN. En el gráfico se realiza además una comparación entre los usuarios nuevos y recurrentes.

La Figura 3 muestra otra representación de Google Analytics, siempre sobre datos de visitas registradas al el sitio web de la UNRN. En este caso, se trata de la distribución de las visitas a cada una de las principales páginas según las diferentes regiones (Período 18/10/11–17/11/11).

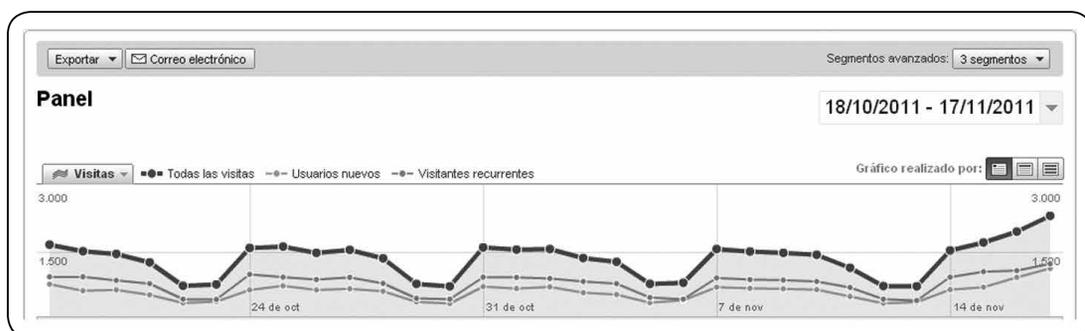


Figura 2.

Visitas a sitio web de la UNRN según tipos de usuarios.

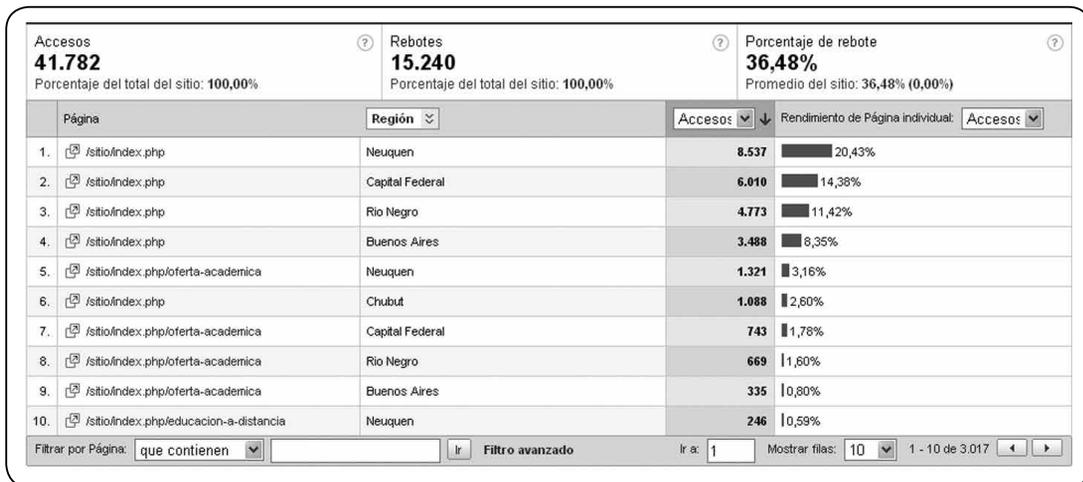


Figura 3.
Visitas de páginas según regiones.

Los anteriores son algunos ejemplos de representaciones posibles en Google Analytics. Invitamos al lector a interiorizarse de mayores detalles del software a través de su sitio web.

Cabe destacar que al menos en la versión utilizada, no existió la posibilidad de obtener los datos originales. Ellos estuvieron sólo disponibles bajo formas sintetizadas (tablas de frecuencias, representaciones gráficas, etc.). Además, y vinculado a esta característica, existen en Google Analytics herramientas predeterminadas de tratamiento de los datos, como por ejemplo la selección por período de tiempo o por región, etc. En otras palabras, el espacio de posibilidades de tratamiento de los datos está restringido a las herramientas preexistentes en el software.

A pesar de estos condicionamientos, el tipo de análisis que este software permite, nos pareció *a priori* de interés para introducir a los estudiantes en los principios del AED. En efecto, algunas de sus características nos resultaron coincidentes con las del AED. Por ejemplo, los datos recolectados por el software son multivariados. A su vez, estos pueden considerarse como no estructurados pues su generación no responde a ningún protocolo predeterminado de muestreo ni selección específica. Por otro lado, las herramientas de tratamiento de los datos son relativamente variadas y basadas

fuertemente en representaciones gráficas. Estas características junto a la relativa facilidad de apropiación del software nos permitieron suponer que un estudio con Google Analytics de visitas a sitios web, podría favorecer en los estudiantes la actitud exploratoria y de generación de hipótesis, elementos típicos de un AED (Figura 1). A continuación, presentaremos una síntesis de la experiencia llevada a cabo con los estudiantes.

4. Resumen de la experimentación

Al inicio del cuatrimestre se les presentó la propuesta a los estudiantes, acompañada de un ejemplo. El caso en cuestión correspondió al análisis de visitas del sitio web de la UNRN. Las consignas dadas fueron generales y se acordaron las pautas de trabajo respecto al estudio:

Los veintisiete estudiantes conformarían grupos de 3 ó 4 integrantes. Cada grupo debería procurarse un sitio web a analizar. Algunas características de los sitios web fueron presentadas como necesarias (posibilidad de acceso al código fuente, sitio web público, etc.) y otras como recomendables (sitios con pocas o nulas animaciones Flash, varias páginas por sitio, etc.).

El trabajo se dividiría en dos partes, una primera donde los estudiantes presentarían semanalmente a la clase los avances del estudio estadístico. Cada presentación sería breve (menos de 10 minutos) y estaría acompañada de una devolución colectiva. En la segunda parte las presentaciones disminuirían en tiempo y se acompañarían de informes escritos sobre el análisis de los sitios. Así, una buena parte del estudio lo realizarían los estudiantes fuera de la clase y las presentaciones permitirían tanto compartir la evolución de los equipos como las dificultades encontradas.

La sala de clase era pequeña para el número de estudiantes y contaba con 8 computadoras y un video proyector.

4.1. Los sitios web analizados

Al cabo de dos semanas de presentada la actividad, los grupos tenían definidos los sitios web a analizar e iniciado en la mayoría de los casos la generación de datos. Estos sitios correspondieron a empresas productoras de la región, periódicos en línea, asociaciones sin fines de lucro, empresas relacionadas con el comercio, el turismo, etcétera. En general se trató de emprendimientos pequeños o medianos. La Tabla 1 muestra un listado de los diferentes grupos y las temáticas abordadas en cada uno de los sitios analizados.

Denominación del grupo	Temática del sitio web analizado
G1	Viajes y turismo
G2	Insumos y productos agropecuarios
G3	Repuestos y ventas de automotores usados
G4	Óptica
G5	Periódico local I
G6	Periódico local II
G7	Periódico regional
G8	Teatro

Tabla 1.

Temáticas abordadas por cada grupo de estudiantes.

En la mayoría de los casos, los propietarios cedieron a los estudiantes el acceso irrestricto al software y por ende a la información disponible. En uno solo, dada la envergadura de la empresa y el lugar que ya ocupaba Google Analytics en el desarrollo de estrategias de sus ventas, los estudiantes tuvieron acceso de manera parcial al software. En todos los casos, hubo un compromiso ante las empresas a preservar sus identidades y datos; así como a brindarles un informe final del estudio realizado.

4.2. Las presentaciones e informes de los estudiantes

Las presentaciones orales de los grupos se fueron sucediendo durante 5 semanas. Transcurrida esa primera etapa, los grupos combinaron presentaciones más acotadas en tiempo con informes escritos (3 semanas más). En la síntesis propuesta en este artículo nos hemos basado en los Power Points de la primera etapa, en los informes de la segunda etapa y en notas tomadas durante las respectivas clases.

Analizando el material disponible, hemos distinguido dos categorías, una caracterizada por un trabajo centrado en la descripción («grupos descriptores»); la otra donde se profundiza el análisis y se abordan cuestiones referidas a la búsqueda de explicaciones y/o posibles hipótesis a los fenómenos encontrados («grupos exploradores»).

En la primera categoría («descriptores»), los estudiantes centraron su atención en una «lectura» directa de los datos, propusieron gráficos y tablas con el principal fin de ilustrar la información obtenida. Así hemos podido observar en sus presentaciones e informes un importante uso de este tipo de elementos de síntesis de la información (tablas, valores porcentuales, etc.) pero sin lograr incursionar en proposiciones explicativas. A esta categoría hemos asociado inicialmente a los grupos G2 (sitio web de Insumos y productos agropecuarios) y G7 (sitio web de Periódico regio-

nal). El resto de los grupos los hemos reunidos en la categoría «exploradores», y esto dada la presencia de una profundización del trabajo a nivel exploratorio o de «detective» en el sentido de Tukey (1977): G1: Viajes y turismo, G3: Repuestos y ventas de automotores usados, G4: Óptica, G5 y G6: periódicos locales, G8: Teatro.

La Tabla 2 muestra un ejemplo (Grupo 7) de una producción perteneciente a la categoría «descriptores» donde el mencionado grupo presentó en un informe dos tablas, ambas similares y donde no se propusieron posibles explicaciones a estas diferencias, ni tampoco un seguimiento de las variables que permitía el software.

Países	Cantidad de visitas	Tablas con contenido de visitas, en cuanto a los países de donde provienen las mismas: «Como se puede ver en el cuadro, el país con mayor cantidad de visitas es la República Argentina que cuenta con 1 480 865 de visitas, le sigue Estados Unidos con 18 777 usuarios, España con 15 239, y luego países como Brasil, México, Chile, Colombia, Italia y Perú».
Argentina	1 480 865	
Estados Unidos	18 777	
España	15 239	
Brasil	8164	
México	7186	
Chile	5994	
Colombia	3488	
Italia	2773	
Perú	2410	

Tabla 2.

Visitas mensuales a Periódico regional. Producción de los estudiantes al respecto.

En la presentación de la cual fue parte la Tabla 2 no se registraron propuestas explicativas de tal distribución de datos, ni intentos de complementación con otras variables disponibles tales como tiempo de permanencia y/o tipo páginas visitadas, etc. El foco de atención se restringió a la visualización de la tabla de valores.

En contraposición a esta situación, pudimos observar una presentación de una variable próxima a la presentada en la Tabla 2 y sobre un período de tiempo menor (una semana) hecha por el grupo G5 (categoría «exploradores»).

Nivel de detalle: Ciudad	Vistas	Visitas
1. Neuquen	432	47,26%
2. Buenos Aires	214	23,41%
3. General Roca	176	19,26%
4. Bahía Blanca	25	2,74%
5. Córdoba	15	1,64%
6. La Plata	8	0,88%
7. Rosario	6	0,66%
8. San Jorge	6	0,66%
9. Isidro Casanova	6	0,66%
10. Mar Del Plata	5	0,55%

Tabla 3.

Visitas a Periódico local I. Período: una semana.

A continuación, la explicación brindada por los estudiantes del G5 acerca de los datos observados en la Tabla 3:

Nos ha llamado mucho la atención esta semana que de distintas ciudades de la provincia de Buenos Aires han visitado la página y la permanencia en ella ha sido de 1 a 2 minutos aproximadamente, y observando el sitio nos dimos cuenta que en la tapa del diario (que aparece en la página de inicio) se hace presente una noticia de la visita de la señora presidenta Cristina de Kirchner en diciembre; y creemos que por esa razón hay tantas visitas de esta provincia.

El grupo G5, que tenía a cargo el estudio del sitio web de un periódico local, no se limitó a la descripción de los valores observados en la Tabla 3 sino que propuso posibles explicaciones al importante número de visitas provenientes de la capital del país (Buenos Aires), hecho que llamó la atención de los estudian-

tes dado el perfil del periódico como fuertemente orientado a noticias de una pequeña localidad. Otro aspecto observado en los grupos de la categoría «descriptores» se refiere a la relativa similitud de las presentaciones e informes a lo largo del tiempo. Se observó una tendencia en

las conclusiones a ser reiterativas y de reducido análisis. En particular esto último puede verse en dos presentaciones del grupo G2 distanciadas en 6 semanas (Figura 4), donde una redistribución significativa en el tipo de visitantes no es acompañada de posibles explicaciones:

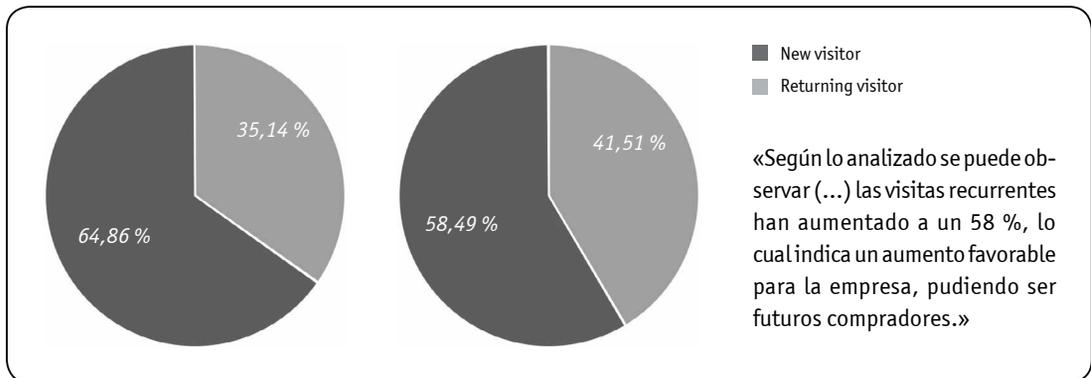


Figura 4.

Evolución de tipo de visitantes en 6 semanas en sitio web Periódico regional I. Producción de los estudiantes del G2.

Este fenómeno no lo hemos observado con tanta nitidez en los trabajos de los otros grupos sino que, bien por el contrario, en las presentaciones e informes de los «grupos exploradores» semanalmente había avances en las presentaciones que refutaban o confirmaban hipótesis acerca del comportamiento de las variables estudiadas,

acompañando con algún nuevo gráfico o tabla y brindando información para el análisis de las hipótesis surgidas con anterioridad. A modo de ejemplo, en las Figuras 5 y 6, presentamos parte de los informes del grupo G5, que se distancian en un lapso de una semana:

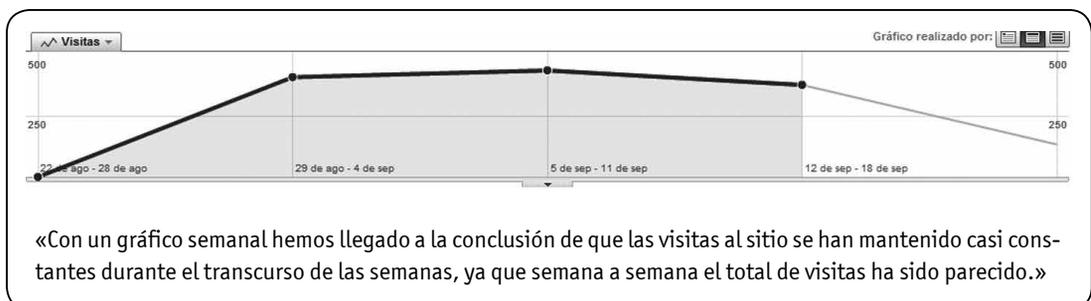


Figura 5.

Evolución de visitas semanal a sitio web Periódico regional I. Producción de los estudiantes G5.

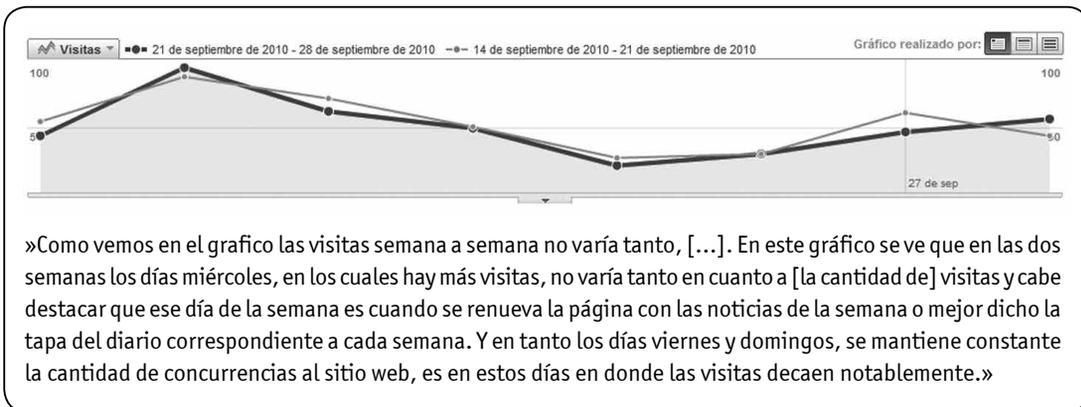


Figura 6.

Comparación de tendencias de visitas por día, durante dos semanas, a sitio web Periódico regional I. Producción de los estudiantes G5.

Hemos podido advertir entonces que los grupos de la categoría «exploradores» pudieron formular hipótesis y trataron de validarlas con información disponible en Google Analytics. Otro ejemplo de

esta categoría puede encontrarse, entre otros, en el grupo G8 (sitio web de un teatro), quienes además de analizar la variable «cantidad de visitas», propusieron hipótesis sobre un fenómeno observado.

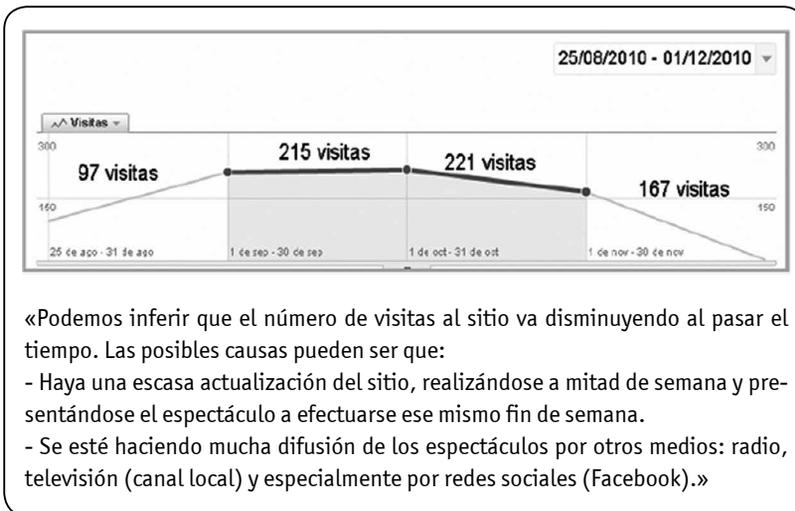


Figura 7.

Evolución de visitas a sitio web de Teatro. Producción de los estudiantes.

Otro fenómeno se refiere a la identificación por parte de los grupos «exploradores» de anomalías o situaciones extrañas en los datos; esta estrategia es

ampliamente sugerida en el AED (Tukey, 1977). Por ejemplo, en un informe del G1 (sitio web de empresas de viajes y turismo) se presenta la Figura 8:

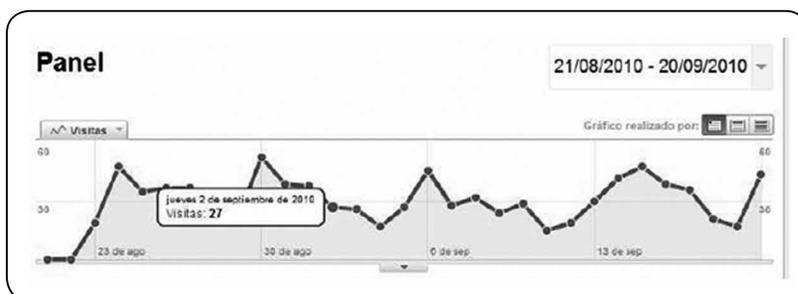


Figura 8.
Evolución de visitas a sitio web de empresa de viajes y turismo.

Y se lee la siguiente reflexión:

se puede apreciar que (...) las visitas aumentan durante los días de semana, es decir hábiles, y decaen los fines de semana; lo que se puede atribuir a que las personas consultan durante los 5 días laborales, y los dos restantes realizan algún viaje o se dedican a otras actividades.

Otro ejemplo de la misma situación puede encontrarse en el análisis presentado por el grupo G3 (sitio web de repuestos y venta de automotores usados):

las visitas no varían notablemente, (...) pero si nos detenemos a mirar detalladamente podemos observar que los días sábados las visitas disminuyen en más cantidad que el resto de la semana. Estuvimos deliberando en el grupo cual podría ser el motivo pero no llegamos a ninguna conclusión. No podemos decir que esta disminución se debe a que es fin de semana porque en varios días domingo no sucede esto. Tratamos de encontrar alguna respuesta revisando el mes anterior pero sucede exactamente lo mismo, el día sábado decaen las visitas.

A continuación, presentaremos algunas conclusiones de la propuesta llevada a cabo con estudiantes de las carreras de Comercio Exterior y Marketing de la UNRN donde nos propusimos trabajar con ellos algunas nociones elementales del AED.

5. Discusión

Nuestro objetivo principal para el estudio propuesto a los estudiantes consistió en brindarles una introducción al AED, y esto, en un marco contextual próximo a los posibles usos estadísticos asociados a sus futuras carreras profesionales. En particular, nos interesó observar en este estudio estadístico aspectos relacionados al proceso de formulación de hipótesis explicativas a fenómenos que pudieran llamar la atención a los estudiantes. Si bien esta experiencia llevada a cabo durante el primer semestre del año 2010 no es más que introductoria a la problemática de la enseñanza del AED, podemos afirmar que es posible introducir nociones de base de este enfoque estadístico, al menos en los primeros años de la formación universitaria. En este sentido, Google Analytics resultó una herramienta posible para estos fines. En lo que respecta a la generación de hipótesis y como muestran los ejemplos presentados, los estudiantes han podido formular proposiciones explicativas a fenómenos observados y esto de manera progresiva con el correr de los informes y presentaciones. Sin embargo, algunas cuestiones quedan por responder, por ejemplo:

5.1. La relación causal entre los datos y las hipótesis

Hemos visto informes y presentaciones donde la fundamentación de las hipótesis escapaba a la información que se desprendía de los datos.

Al punto que en algunos casos estos últimos no eran de utilidad para fundamentar una hipótesis. En esas situaciones, era el conocimiento general que los estudiantes tenían del contexto lo que les servía como fundamento para sostener esas hipótesis. También, encontramos grupos donde los datos contradecían las hipótesis propuestas, estas hipótesis fueron sostenidas por estar en consonancia con el conocimiento general que tenían (o pensaban tener) del contexto.

Así, hemos notado en algunos grupos que los datos no alcanzaban a cuestionar las hipótesis defendidas. En otras palabras, pareciera haber existido una cierta coherencia o al menos una estabilidad en el sistema de creencias de los estudiantes que impidió que los datos cuestionaran ciertas hipótesis preexistentes al estudio. En este sentido, nos parece, jugó un papel importante la comprensión de la idea de «hipótesis». Pareciera que algunos estudiantes hubieran considerado la noción de hipótesis como «un enunciado o situación desconocida de la que se desea determinar su verdad o falsedad» (Vallecillos, 1997:37), mientras que el AED plantea la necesidad de generar hipótesis como conjeturas, como posibles explicaciones a lo observado mediante los datos.

Si bien consideramos fundamental un mínimo de conocimiento del contexto por parte de los estudiantes para abordar cuestiones referidas al AED (Good, 1983; Mast & Kemper, 2009), es posible la manifestación de conflictos entre los conocimientos previos de los estudiantes y los en construcción a partir del estudio estadístico. Al respecto, observa Schield (2006) la existencia de una «necesidad» de los estudiantes de plantear una relación causa–efecto entre los fenómenos observados en lugar de asimilar nuevos conceptos estadísticos, como podría ser la correlación.

En este sentido, no consideramos estos conflictos como un problema de interpretación de datos por parte de los estudiantes, sino como una consecuencia de tensiones entre el conjunto de creencias preexistentes en los estudiantes y los

nuevos conocimientos que el estudio estadístico aporta. Es decir, el estudio estadístico pareció no haber cuestionado suficientemente el conjunto de creencias de algunos estudiantes. Creemos que, tal vez, esto esté en relación con los mecanismos psicológicos que condicionan, de forma general, el comportamiento de los sujetos en situaciones de incertidumbre (Vallecillos, 1997).

5.2. La cantidad de información y la generación de hipótesis

Los dos grupos de la categoría llamada «descriptores» (G2 y G7) coincidieron en una característica, aunque en posiciones antípodas. El grupo G2 (Insumos y productos agropecuarios) analizó un sitio web con escasas visitas mensuales (alrededor de 100 aproximadamente). Por su parte, el Grupo G7 (Periódico regional) tuvo a su cargo uno donde las visitas eran significativas en número (del orden de los 2 millones mensuales).

En el primer caso, la escasa cantidad de datos desmotivó al grupo en la búsqueda de hipótesis explicativas. Un argumento recurrente en las presentaciones del Grupo G2 fue precisamente que la evidencia constatada, percibida como débil por su escasa cantidad, no les resultaba convincente para sostener hipótesis e incluso para proponerlas.

En el segundo caso vimos un fenómeno diferente, los 2 millones de datos mensualmente registrados en el sitio web del periódico regional necesitaron un importante trabajo de tratamiento para una posible interpretación. Así, los estudiantes concentraron sus esfuerzos en sintetizarlos de diferentes maneras, ya fuera mediante gráficos relativamente apropiados a los fines propuestos, o por medio de medidas de posición y/o dispersión. Sin embargo, no incursionaron en la proposición de hipótesis explicativas. Su mayor motivación consistió en obtener información de esa importante masa de datos. Es probable entonces que contextos donde existe una disponibilidad de datos importante, no resulte el más

apropiado para motivar la generación de hipótesis en los estudiantes, al menos en un período de tiempo de un semestre.

5.3. La motivación de los estudiantes

El tipo de contexto del estudio estadístico efectuado por los estudiantes resultó sin dudas motivador para ellos. Su realismo y su relación con la actividad profesional futura fueron ampliamente apreciados por los diferentes grupos. La posibilidad de seguimiento de las visitas en tiempo casi real fue otro elemento importante en la motivación de los estudiantes. Esto no solamente permitió a los grupos observar la influencia de fenómenos naturales y sociales en las visitas a los sitios web, sino que motivó a algunos a iniciar acciones para incrementarlas. Por ejemplo, el grupo G5 (Periódico local I) y con la autorización de los propietarios del sitio, creó cuentas en redes sociales virtuales tales como Facebook y Twitter a fin de difundir el sitio web. Esto no solamente originó un intercambio profesional considerable entre los estudiantes y los propietarios del sitio, sino que permitió a su vez observar los efectos de tales publicidades sobre las visitas al sitio en cuestión.

La propuesta del estudio estadístico sobre las vi-

sitas a sitios web resultó abierta a los intereses de los estudiantes. Por ejemplo, el grupo G8 (Teatro) se interesó en el análisis del sitio web de una organización cultural local sin fines de lucro. Para este grupo, el estudio estadístico fue percibido como una oportunidad de contribuir a su localidad por medio de la mejora en la difusión de los eventos culturales de la región.

En fin, creemos que es posible y positiva la experiencia de realizar estudios estadísticos de este tipo para introducir nociones de base del AED. Sin embargo algunas cuestiones plantean desafíos que merecen continuar siendo analizados, como por ejemplo la referida a la institucionalización. En efecto, tanto para docentes como para alumnos no fue evidente en un primer momento la identificación de los objetivos didácticos perseguidos con el estudio, entre ellos por ejemplo, la construcción y justificación de hipótesis explicativas a partir de información disponible en el sitio. Este tipo de objetivos resultó, al menos en un primer momento, difuso para una buena parte de los estudiantes. Y la construcción progresiva de un contrato didáctico (Brousseau, 1988) permitió a los estudiantes ir cerniendo en las diferentes presentaciones las intenciones de los docentes en el estudio propuesto.

Nota

⁽¹⁾ El término «gratuidad» debe considerarse de manera relativa, pues la habilitación del uso de Google Analytics sobre un sitio web implica el permiso del propietario del sitio a la empresa Google para que registre los datos de la visita.

Referencias bibliográficas

- Brousseau, G. (1988).** Le contrat didactique. *Recherche en didactiques des mathématiques*, 9(3), 309–336.
- Carranza, P. (2009).** La dualité de la probabilité et enseignement de la statistique. Une expérience en bts. Paris VII Denis Diderot, *Savoirs Scientifiques: Epistemologie, histoire des sciences, didactique des disciplines*.

- Carranza, P. (2012).** Sensibilisation à l'abduction en statistique. Paper presented at the EMF 2012. *Enseignement des mathématiques et contrat social: enjeux et défis pour le 21e siècle*, Genève.
- Carver, R. (1978).** The case against statistical significance testing. *Harvard Educational Review*, 48(3), 378–399.
- Chevallard, Y. (1991 [1985]).** *La transposition didactique*. Grenoble: La Pensée Sauvage.
- Cox, R. T. (1946).** Probability, frequency, and reasonable expectation. *American Journal of Physics*, 14, 1–13.
- Cumming, G. (2010).** Statistics education in the social and behavioural sciences: From dichotomous thinking to estimation thinking and meta-analytic thinking. Paper presented at the *Data and context in statistics education: towards an evidence-based society*. ICOTS 8, Ljubljana, Slovenia.
- Diaconis, P. (2006 [1985]).** Theories of data analysis: From magical thinking through classical statistics. En Hoaglin, F., Tukey, J. (Eds.), *Exploring data tables, trends, and shapes*. New York: John Wiley & Sons, Inc.
- Engestrom, Y. (1991).** Activity theory and individual and social transformation. *Activity Theory*, 7/8, 6–17.
- Gärdenfors, P., Sahlin, N.-E., Ramsey, F., Luce, D., Raiffa, H., Savage, L., et al. (1988).** *Decision, probability and utility*. Cambridge: Cambridge University Press.
- Good, I. J. (1983).** The philosophy of exploratory data analysis. *The Philosophy of Science*, 50(2), 283–295.
- Hacking, I. (2002).** *L'émergence de la probabilité*. Paris: Seuil.
- Jaynes, E. (1980).** What is the question? En Bernardo, J. M., deGroot, M. H, Lindly, D. V. and Smith A. F. M. (Eds.). *Bayesian statistics*. Valencia: Valencia Univ. Press.
- Jaynes, E. (1984).** The intuitive inadequacy of classical statistics. *Epistemologia*, 7(43), 43–74.
- Lecoutre, B. (2005).** Et si vous étiez un bayésien qui s'ignore? *Modulad*, 1(32), 92–105.
- Mast, J. d., & Kemper, B. P. H. (2009).** Principles of exploratory data analysis in problem solving: What can we learn from a well-known case? *Quality Engineering*, 21, 366–375.
- Meehl, P. (1978).** Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806–834.
- Pearson, K. (1930).** *The life, letters and labours of francis galton*. London: Cambridge University Press.
- Peirce, C. S. (1932).** Collected papers. Hartshorne C. y Weiss, P. (eds.). Vol II: *Elements of Logic*. Cambridge MA: Harvard University Press.
- Popper, K. R. (1959).** *The logic of scientific discovery*. London: Hutchinson.
- Rabardel, P. (1995).** *Les hommes et les technologies, une approche cognitive des instruments contemporains*. Paris: Armand Colin.
- Régnier, J.-C., & Oriol, J.-C. (2001).** Fonctionnement didactique de la simulation en statistique. Paper presented at the *Journées de Statistique Lyon 2003*, Lyon, France.
- Schild, M. (2006).** Statistical literacy survey analysis: Reading graphs and tables of rates and percentages. *Proceedings of the Sixth International Conference on Teaching Statistics*. Ciudad del Cabo: International Statistical Institute and International Association for Statistical Education.
- Tukey, J. W. (1977).** *Exploratory data analysis*. Reading, Massachusetts: Addison-Wesley Publishing Company.
- Vallecillos Jimenez, A., & Batanero Bernabeu, C. (1997).** Conceptos activados en el contraste de hipótesis estadísticas y su comprensión por estudiantes universitarios. *Recherches en didactique des mathématiques*, 17(1), 29–48.