



Galaxia NGC 1156. Fuente: ESA/Hubble, NASA, R. Jansen.

FIGURAS REVISTA ACADÉMICA DE INVESTIGACIÓN

ISSN 2683-2917

Vol. 1, núm. 1, noviembre 2019-febrero 2020

<https://doi.org/10.22201/fesa.figuras.2019.1.1>



Esta obra está bajo una licencia  
Creative Commons Atribución-NoComercial-  
CompartirIgual 4.0 Internacional

## ¿Qué nos ofrece la Astroestadística?

 Elizabeth Martínez Gómez

<https://doi.org/10.22201/fesa.figuras.2019.1.1.90>

La investigación astronómica moderna enfrenta una amplia variedad de retos estadísticos, que van desde la reducción de bases grandes de datos que caracterizan a los objetos celestes, hasta la verificación de teoría astrofísica. En la mayoría de los casos, los astrónomos hacen mal uso de las herramientas estadísticas que tienen disponibles, o bien, abusan de ellas. Así que, en un intento por unificarlas, se ha propuesto crear el campo de la Astroestadística.

### Presentación

Desde la observación astronómica donde se tienen problemas de calibración hasta la interpretación de los datos obtenidos, surgen varios y diversos problemas tanto estadísticos como computacionales.

A manera de ejemplificar esto se pueden mencionar la selección e identificación de las fuentes astronómicas de interés, la clasificación de objetos, el procesamiento de señales, entre otras. Para abordarlos se emplean técnicas de muestreo (básicamente el Muestreo Aleatorio Simple, MAS), análisis multivariado (componentes principales, análisis de factores, análisis de discriminantes, correlación canónica, escalamiento multidimensional, análisis de correspondencias, modelos de ecuaciones estructurales, clasificación y *clustering*, entre otras), análisis de regresión (regresión lineal simple y múltiple, regresión no lineal, modelos lineales generalizados y análisis de varianza, por mencionar algunas), análisis de series de tiempo, estadística bayesiana (principalmente inferencia), y otras más.

Es esta diversidad de temas estadísticos a los que se enfrenta la astronomía hoy en día, lo que llevó a la creación del área multidisciplinaria conocida como *Astroestadística* en 2003 para facilitar el desarrollo de herramientas útiles y la promulgación de conocimientos estadísticos.

A partir de entonces, diferentes grupos de investigación tanto en Astrofísica como en Estadística han dedicado sus esfuerzos para desarrollar e implementar adecuadamente las metodologías tradicionales y nuevas.

### Tipos de datos astronómicos

Típicamente un astrónomo observa una parte específica del cielo a través de telescopios y con ello obtiene imágenes y/o mediciones del brillo, temperatura, o de cualquier otra propiedad física de un objeto (fig. 1).

Frecuentemente los datos observados pueden constar de conjuntos de valores para una o varias característica(s) determinada(s), imágenes, gráficas de flujos de energía (conocidas como espectros), y series temporales.

Una de las características que hace que los datos astronómicos sean únicos es la *heteroscedasticidad*, esto es, que los errores de medición varían entre observa-

ciones para una misma muestra. Esto se debe a que las condiciones observacionales son distintas, a los tiempos de exposición y al brillo de las fuentes astronómicas.



Figura 1. Algunos ejemplos de instrumentación astronómica empleados para la obtención de datos científicos. Los astrónomos utilizan desde telescopios (o arreglos de telescopios) ubicados sobre la superficie terrestre hasta complejos telescopios espaciales para investigar el Universo.

En general, antes de proponer algún modelo astrofísico para explicar el fenómeno de interés, se deben analizar los datos. Dentro de esto último, lo más común es:

1. Remover *outliers* de la señal mediante un ajuste polinomial o suavizamiento (*splines*). Es importante destacar que éstos son métodos numéricos ampliamente utilizados en el análisis (estadístico) descriptivo de datos.
2. Calibrar adecuadamente. En Astronomía debe distinguirse claramente entre la calibración de imágenes, la instrumental y la de parámetros físicos. Precisamente en la tercera se necesita de las técnicas estadísticas.
3. En caso de tener componentes faltantes al extraer la información de un espectro o imagen, los métodos numéricos conocidos como interpolación son ampliamente recomendados.
4. Determinar las posiciones exactas de los objetos (ajustes por mínimos cuadrados) basados en estrellas de calibración (astrometría).
5. Corregir las imágenes mediante un análisis de discriminantes.

6. Estimar los parámetros de una distribución (por ejemplo de energía).

A continuación, se resume un problema típico en Astronomía: la búsqueda de posibles asociaciones entre variables. En otras palabras, asociaciones entre propiedades o características de fuentes astronómicas.

### Aplicación astronómica: búsqueda de asociaciones entre variables usando la información obtenida de un *survey*

Las grandes bases de datos multidimensionales en la Astronomía moderna que provienen de los *surveys* representan retos tanto computacionales como estadísticos. Por ejemplo, el catálogo *Chandra Deep Field South COMBO-17* proporciona mediciones en muchas variables (propiedades) para miles de galaxias que pertenecen a un amplio rango de corrimientos al rojo. En ese sentido nuestro entendimiento actual sobre la formación y evolución de galaxias se apoya en las relaciones que existen entre las distintas variables; por lo que el detectar y verificar asociaciones (o correlaciones) entre ellas es importante en la investigación astrofísica.

Para este propósito se emplea una nueva medida de asociación: la distancia de correlación propuesta por Székely (2007, 2769-2794; 2009, 1233-1265). Tanto la descripción detallada del catálogo de datos así como los detalles de los cálculos realizados se encuentran en los trabajos de Martínez-Gómez *et al.* (2014, 11) y Richards *et al.* (2014, 5).

### Resultados

Las galaxias contenidas en el catálogo se han clasificado de acuerdo con su corrimiento al rojo, denotado por  $z$ , en cercanas ( $0 \leq z < 0.5$ ), intermedias ( $0.5 \leq z < 1$ ) y lejanas ( $1 \leq z < 2$ ); y además según su brillo en cuatro tipos (Wolf *et al.* 2003, 73-98). A manera de ejemplo, en la fig. 2 se muestran los resultados obtenidos para los cuatro tipos de galaxias cercanas.

En los dos paneles superiores de la fig. 2 se identifican las propiedades físicas con una correlación lineal casi perfecta, lo que implica que en un modelo astrofísico no sería adecuado incluir ambas variables; de hecho, éstas son relaciones ya conocidas por los astrónomos. En contraste, los paneles inferiores corresponden a propiedades que no tienen ninguna relación. Las cuatro gráficas restantes comparan el valor del coeficiente de correlación de Pearson con el de la distancia de correlación para cada una de las parejas de variables en los cuatro tipos de galaxias

cercanas. Lo interesante es el patrón que describe cada una en forma de “V” y que parece ser un efecto geométrico que ya ha sido discutido en otros trabajos.

La Astronomía es quizá la ciencia más antigua y la primera que recolecta datos de manera sistemática para su análisis.

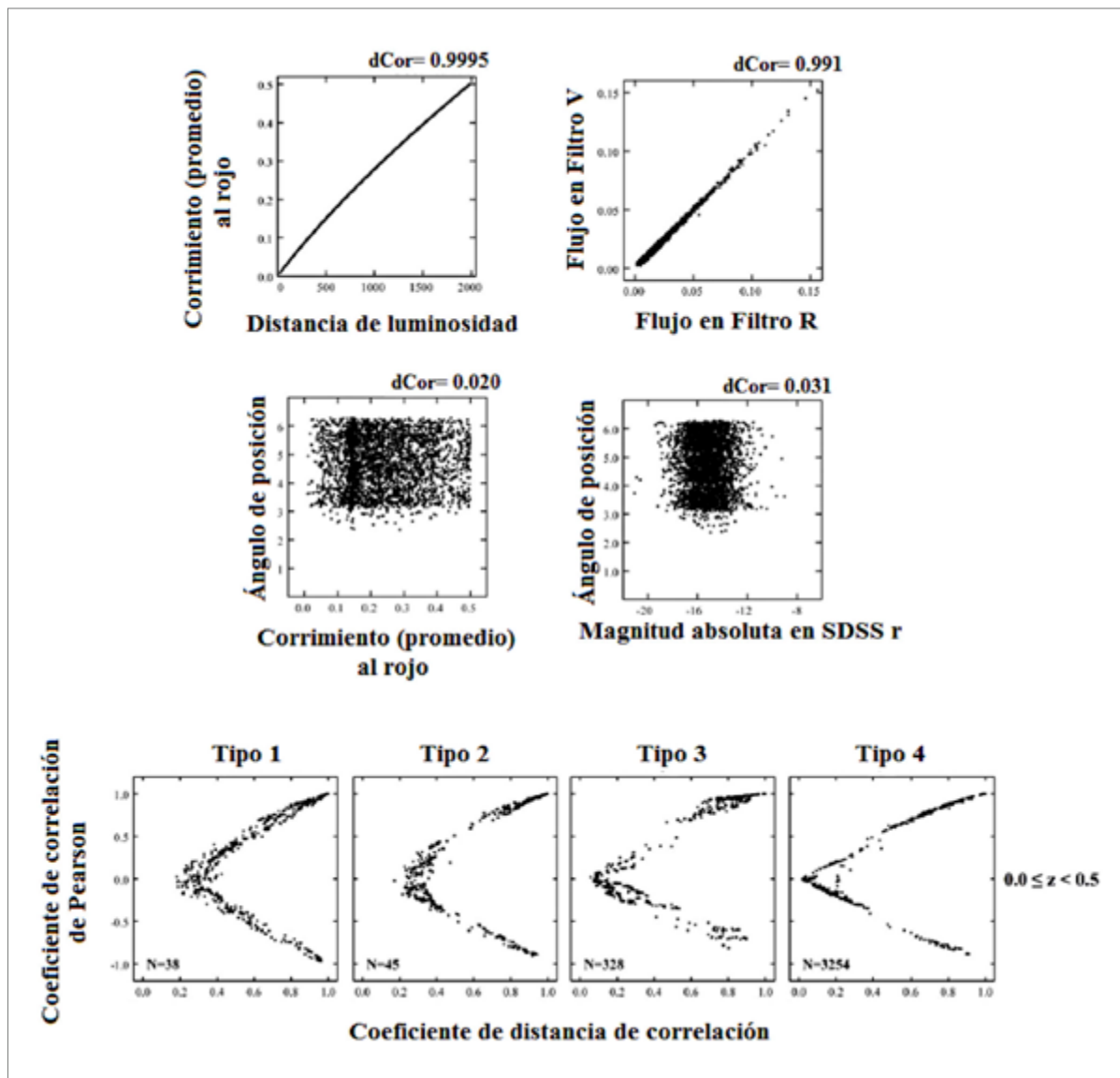


Figura 2. Asociaciones entre variables que representan las características físicas de un conjunto de galaxias observadas con el survey COMBO-17. Se ilustra el caso para los cuatro tipos de galaxias cercanas.

## Para concluir

La Astronomía es quizá la ciencia más antigua y la primera que recolecta datos de manera sistemática para su análisis. Además, los datos astronómicos representan todo un reto para trabajar debido a la existencia de errores (heteroscedásticos) de medición, intervalos desiguales (en el caso de series temporales), datos censurados o truncados (debido a restricciones observacionales o instrumentales), entre otros.

La estadística y probabilidad han ayudado a resolver algunas interrogantes como:

- a) La descripción de las órbitas de los cometas y planetas a través del método de estimación de mínimos cuadrados (desarrollado por Legendre) bajo el supuesto de normalidad en los errores de medición (idea de Gauss).
- b) La existencia y distribución de agrupaciones galácticas se ha estudiado mediante el empleo de técnicas de análisis multivariado (clasificación y *clustering*), conceptos de procesos estocásticos y de inferencia estadística (pruebas de bondad de ajuste).
- c) La aparición de estrellas dobles no es un error óptico sino un hecho astronómico real. Utilizando conceptos probabilísticos, John Mitchell en 1784, descubrió que estas estrellas están realmente muy cerca una de la otra y es por eso por lo que se dificulta su identificación mediante observación directa.
- d) La forma de nuestra galaxia, la Vía Láctea, fue descrita por Herschel gracias a razonamientos inferenciales.
- e) El conocimiento sobre la distribución de materia oscura —materia no observable en el Universo— está siendo posible gracias a la avalancha de datos proveniente de misiones como Planck y a su posterior descripción mediante modelos astrofísicos que han requerido de la estadística bayesiana para su estimación.

La Astroestadística es una nueva área de investigación que ofrece un sinnúmero de atractivos desafíos.

Como se puede notar, la Astroestadística es una nueva área de investigación que ofrece un sinnúmero de atractivos desafíos para los astrofísicos, estadísticos y matemáticos aplicados. Feigelson y Babu (2003, 7) identifican algunos de los retos que estará enfrentando la Astroestadística en las próximas dos décadas:

- a) Análisis Multivariado con errores de medición y observaciones censuradas. Como se ha expuesto antes, las observaciones astronómicas tienen asociados errores de medición, o bien, pueden estar censuradas.<sup>1</sup> La estadística tradicional carece de herramientas para tratar con este tipo de observaciones, por lo que es necesario desarrollarlas para estos casos.
- b) Inferencia estadística y visualización. En la Astronomía moderna es común trabajar con grandes volúmenes de datos; así que deben desarrollarse mejores y más adecuadas técnicas de visualización. Por otro lado, la estadística inferencial no está pensada para tamaños de muestra muy grandes (en donde todo resulta estadísticamente significativo); así que el propósito es desarrollar una “nueva” estadística inferencial.

---

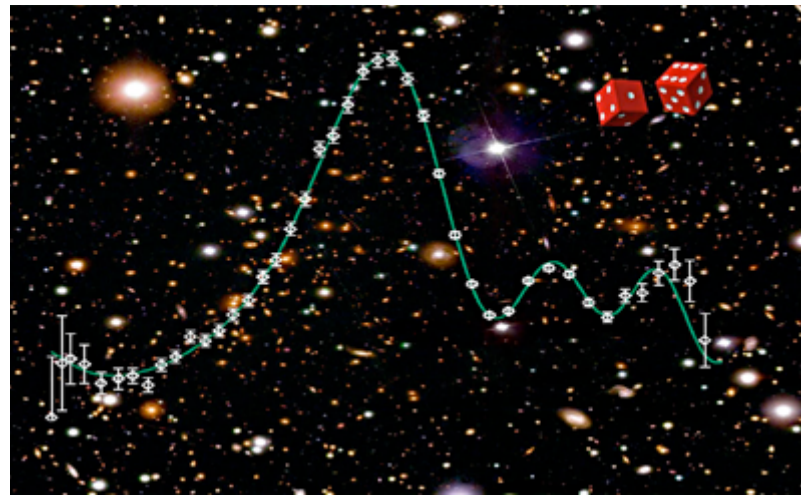
<sup>1</sup> No debe confundirse el término *censura* con el de *truncamiento*. Las observaciones son censuradas si superan cierto umbral (o están en cierto intervalo) y esa información parcial puede usarse cuando se modela estadísticamente el fenómeno. Con el truncamiento, las observaciones quedan descartadas completamente.



- c) Cómputo bayesiano y funciones de verosimilitud. Cuando se emplea el método de máxima verosimilitud (MLE, por sus siglas en inglés) en la estimación de parámetros astrofísicos, es necesario que el astrónomo construya “verosimilitudes” para situaciones paramétricas comunes (por ejemplo, distribuciones de leyes de potencia o un proceso Poisson), lo cual no necesariamente es fácil en la aplicación del MLE, se requieren métodos computacionales más poderosos y de *software* especializado.
- d) Un Observatorio Virtual [ov] (Hernández 2009, 1) es un ambiente federativo internacional virtual de investigación, basado en nuevas tecnologías de la información e Internet, que está abierto a científicos y estudiantes que desean trabajar con conjuntos de datos astronómicos reales. Además de reunir archivos provenientes de distintos proyectos astronómicos, ofrece servicios de análisis de datos. Entre las múltiples tareas que implica mantener un ov están el dar acceso a los diversos catálogos, además de implementar metodologías (numéricas y estadísticas) para ayudar al usuario. De allí la necesidad de colaboración entre matemáticos aplicados y expertos en cómputo.

Vale la pena explorar más allá de las técnicas y métodos tradicionales, y trabajar en conjunto para resolver problemas cada vez más complejos y también para desarrollar nuevas ideas. México cuenta con investigadores de prestigio en ambas disciplinas, ¿por qué no crear un grupo sólido en esta área?

Para motivar al lector se ha presentado de manera muy breve una aplicación concreta al estudio de las posibles asociaciones entre variables usando datos de galaxias. Se calculó la distancia de correlación para todas las posibles parejas de variables y además de verificar asociaciones entre variables astrofísicas ya conocidas, es posible identificar otras nuevas. Esta medida de asociación es más poderosa para detectar independencia entre variables y parece ser una técnica viable para la selección de éstas. —



Astronomía + Estadística + Probabilidad = nuevo reto científico.

## Referencias

- Feigelson, E. and G. Babu. 2003. *PHYSTAT 2003*, SLAC, Stanford, California, (September): 8-11.
- Hernández Cervantes, L., A. J. Santillán González and A. R. González-Ponce. 2009. “Observatorios Virtuales Astrofísicos.” *Revista Digital Universitaria*. 10, no. 10. (octubre): 1-9. <http://www.revista.unam.mx/vol.10/num10/art62/int62.htm>
- Martínez-Gómez E., M. T. Richards, and D. St. P. Richards. 2014. *The Astrophysical Journal Letters* 748, no. 39.
- Richards M. T., D. St. P. Richards, and E. Martínez-Gómez. 2014. *The Astrophysical Journal Letters* 784, L34.
- Szekely G. J., M. L. Rizzo, and N. K. Bakirov. 2007. *Annals of Statistics* 35, 6: 2769-2794.
- Szekely G. J. and M. Rizzo. 2009. *Annals of Applied Statistics* 3, 4: 1233-1265.
- Wolf, C. et al. 2003. *Astronomy and Astrophysics* 401: 73-98.

El texto pertenece al Seminario de Investigación de Ciencia, Tecnología, Ingeniería y Matemáticas. PAIDI/007/18 de la FES Acatlán, UNAM.