Trabajo práctico número 2 de Astrometría

R.C. Martin[1]

[1] *Observatorio Astronómico de Córdoba, UNC, Argentina*

Received: 16/10/2024 / Accepted: ...

Resumen / Este trabajo consiste en 2 partes, una primer sección donde utilizamos ADQL para estudiar estadísticamente los exoplanetas conocidos hasta el momento y las técnicas de detección de los mismos. En la segunda parte, demostramos algunos resultados estadísticos y resolvemos algunos problemas de estadística tales como intervalos de confianza e hipótesis nula.

Keywords / Estadística, Python.

1. Introducción

Debido al gran aumento en el número de datos, se hizo necesario una manera de almacenar y acceder a los datos de manera eficiente. La manera más común es mediante Bases de Datos Relacionales, en estas los datos se organizan en tablas que se relacionan entre sí según una propiedad de los datos. Cada una de estas tablas es un conjunto de registros. Ver Figura 1.

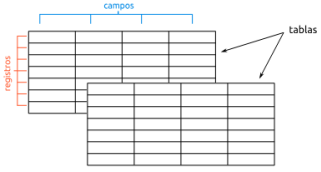

Figura 1: Elementos de una Base de Datos Relacional

Estos grandes volúmenes de datos nos permiten hacer estudios estadísticos. Si queremos ajustar un modelo a un conjunto de datos, vamos a querer maximizar las probabilidades de los parámetros dado un modelo y un conjunto de datos; esto es:

$$max\{P(\phi|d,m)\} = max\{\frac{P(d|\phi,m)P(\phi|m)}{p(d|m)}\} \quad (1)$$

$$max\{P(\phi|d,m)\} = max\{\frac{Likehood * Prior}{\int_\phi P(d|\phi,m)P(\phi|m)d\phi}\} (2)$$

Siendo ϕ los parámetros del modelo, m el modelo y d los datos. La integral del denominador se la conoce como evidencia. Si tomamos un modelo m fijo, y asumimos que todos los datos provienen de un mismo experimento, la expresión se reduce a $max\{P(d|\phi)P(\phi)\}$, luego si asumimos un bias uniforme, se reduce al método de máximo Likehood/Verosimilitud: $max\{P(d|\phi)\}$. Con esto se puede deducir el método de cuadrados mínimos.

2. Exoplanetas

Usaremos la base de datos exoplanet.eu (Schneider et al. (1995)). Cabe destacar que esta base considera a los objetos "exoplanetas" hasta $60M_J$ por lo que el estudio incluirá enanas marrones. Para importar los datos, utilizamos el paquete de python "pyvo". Todo lo que procede a continuación fue hecho mediante el script exoplanetas.py.

2.1. Masas y períodos: Sesgos

Primero, separamos los planetas por método de detección y estudiamos la distribución de masas y períodos orbitales (Ver Figura 2). También vemos las distribuciones individuales de cada parámetro en las Figuras 3 y 4.

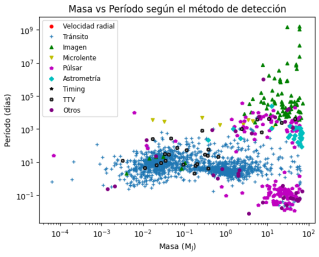

Figura 2: Período orbital vs Masa para todos los exoplanetas detectados. En color referenciamos el primer método de detección mediante el cual fueron encontrados.

Donde rápidamente notamos un gran sesgo hacia planetas de corto período. Esto está vinculado con la técnica de tránsito, que es la más común y es mucho más sensible a planetas cercanos a sus estrellas, dado que eclipsan mayor luz. También la técnica de velocidades radiales es más sensible a planetas cercanos a sus estrellas. Los planetas más alejados se encuentran por la técnica de imagen directa, pues al estar más lejos, la

✉ *Contacto* / romanciromartin@mi.unc.edu.ar

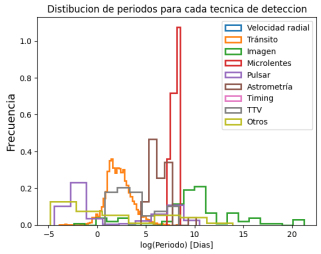

Figura 3: Distribución del período orbital de los exoplanetas hallados con cada técnica.

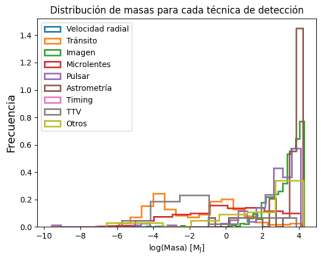

Figura 4: Distribución de masa para los exoplanetas hallados con cada técnica.

estrella no los encandila. Además, se detectan principalmente planetas entre $[0.01, 10]M_J$ lo cual se corresponde con planetas masivos. Este otro sesgo está vinculado también con la técnica de tránsito, pues los planetas más masivos tienden a ser más grandes y, por lo tanto, a ocultar más luz de sus estrellas. Los planetas detectados por técnicas astrométricas tienden a ser de los más masivos dado que perturban más el centro de masas del sistema. Otros sesgos observados son:

- Técnica de imagen directa: Detecta planetas masivos pues su mayor tamaño permite reflejar más luz.
- Microlente: Detecta planetas exclusivamente en la zona de microlente, entre $[1, 10]UA$ de su estrella anfitriona.
- Planeta similares a la Tierra: $1M_T \approx 0.003M_J$: las técnicas son poco sensibles en ese rango de masas, lo cual dificulta la detección de planetas de este tipo, tal y como se ve en el gráfico. Es más, un planeta como la Tierra además de tener una masa poco sensible, posee una distancia a la Estrella incómoda (1 UA), pues no es sensible para imagen directa, está al límite de los microlentes y es poco sensible en tránsito. Por lo tanto existe un bias con este tipo de planetas en estrellas como el Sol.

2.2. Distancia a las estrellas con exoplanetas: sesgo por distancia

En la Figura 5 vemos cómo claramente hay un enorme sesgo hacia estrellas cercanas, dadas las limitaciones en sensibilidad de las técnicas. La técnica de microlentes es la única que permite observar fuera de la vecindad solar; por lo tanto, es esperable este comportamiento.

Vamos a comprobar mediante un test de Kolmogorov-Smirnov que esta distribución no es Gaussiana. El método consiste en calcular la distancia

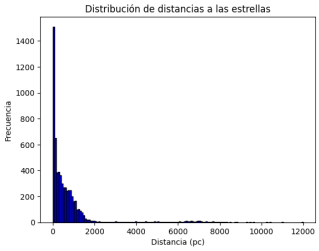

Figura 5: Distancia a las estrellas con exoplanetas.

máxima entre la función acumulada empírica de la distribución de datos, con la teórica. A este estadístico de prueba lo llamaremos D. Luego, si tomamos una significancia del 5 % obtenemos que la condición para que ambas distribuciones sean comparables es $D\sqrt{N_{datos}} < 1.358$. Para una distribución normal centrada en el promedio de los datos y con una varianza estimada con los datos, obtenemos $D\sqrt{N_{datos}} = 75.028$, por lo tanto no es posible que esta sea una distribución Gaussiana. El mismo test con una distribución exponencial arroja un $D\sqrt{N_{datos}} = 76.87$, por lo tanto no es de este tipo tampoco.

2.3. Relación Masa-Radio

Ahora, intentaremos ajustar un modelo para la masa M en función del radio R. En este inciso vamos a filtrar para masas menores a 13 masas de júpiter pues las enanas marrones poseen un mecanismo de estabilidad muy diferente al de los planetas. Luego, en la figura 6. Realizaremos 2 ajustes, uno hasta 1 radio de Júpiter y otro para los radios superiores. Esto lo hacemos dado que a partir de estos tamaños la distribución se vuelve constante por la transición a interiores sostenidos por presión de gas degenerado. En escala logarítmica tenemos $y = log(M)$ y $x = log(R)$, además la densidad central posee forma de un polinomio de orden impar, por lo que propongo el modelo $y = ax^3 + bx^2 + c$ y lo ajusto con Numpy, usando el método de cuadrados mínimos (Ver Figura 7).

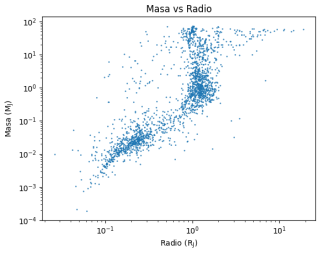

Figura 6: Masa vs Radio para los exoplanetas encontrados, escala logarítmica hasta 1 radio de Júpiter.

Encontramos los valores $a = 0.6704286877825062$, $b = 1.6917850183307248$, $c = 2.8599497249372887$, $d = -0.14258497175397497$ y además un $\chi^2 = 285.00844541972515$. No obstante, en la figura 8 ajustamos un modelo lineal y obtenemos un $\chi^2 = 288.42154520732146$. Considerando la pequeña diferen-

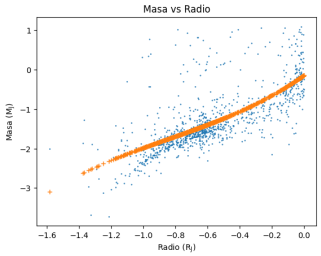

Figura 7: Modelo ajustado a los datos de la Figura 6

cia en χ^2 es preferible este modelo por ser más simple; además nos permite despejar $M(R)$. Tenemos entonces $\log M = A \log R + B$, de lo que sale:

$$M = 10^B R^A \tag{3}$$

Con A=1.7794565866366931 y B=-0.2734600454379645.

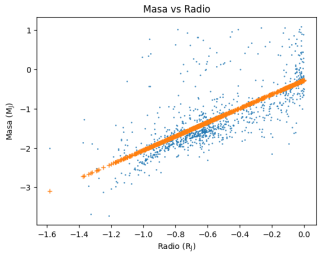

Figura 8: Modelo lineal ajustado a los datos de la Figura 6

Por otra parte, para radios mayores a 1 radio de Júpiter ajustamos un modelo constante (Ver figura 9). Obtenemos $\log(M) = 0.12339458142271102$ con $\chi^2 = 130.62373856995424$. Es un ajuste poco confiable, pero tiene sentido considerando que los puntos poseen una gran dispersión y no aparentan tener una relación clara.

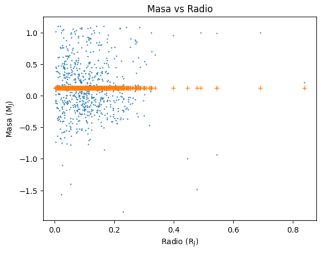

Figura 9: Modelo constante ajustado a los datos de la cola de la Figura 6

2.4. Relación Edad de estrella anfitriona vs Masa planetaria

Finalmente, estudiaremos la relación entre la edad de las estrellas anfitrionas y la masa de los exoplanetas (Ver Figura 10). Se vuelve a observar el sesgo masivo: conocemos más cuerpos masivos por ser más fáciles de detectar. Por otra parte, observamos que las estrellas más viejas, $Edad \in [10, 14]Gy$, poseen menos exoplanetas, lo cual puede estar vinculado a la violenta evolución estelar.

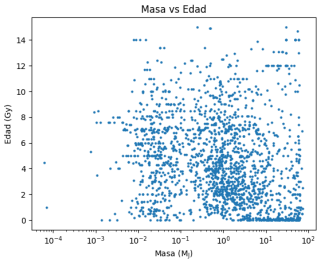

Figura 10: Edad de la estrella anfitriona vs Masa del planeta, escala logarítmica en las masas.

3. Estadística

3.1. Cuadrados mínimos: función lineal

Supongamos que tenemos un conjunto de N datos $\{x_n\}$. Asumimos errores gaussianos en y, es decir:

$$P(y)dy = \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{\frac{-(y_n - y(x_n,\phi))^2}{2\sigma_n^2}} dy \tag{4}$$

Planteamos la verosimilitud:

$$P(datos|modelo) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{\frac{-(y_n - y(x_n,\phi))^2}{2\sigma_n^2}} \tag{5}$$

$$-\log P(d|m) = \sum_{n=1}^{N} \frac{(y_n - y(x_n,\phi))^2}{2\sigma_n^2} + \frac{1}{2}\log 2\pi\sigma_n^2 \tag{6}$$

Entonces, maximizar la verosimilitud, es minimizar esta expresión, por lo que buscamos: $s = min\{\sum_{n=1}^{N} \frac{(y-y(x_n,\phi))^2}{2\sigma_n^2}\}$. Que si asumimos que $\sigma_n = 1$ para todo n, tenemos efectivamente la fórmula de cuadrados mínimos. Para el caso particular de un ajuste lineal, tenemos $y = \phi_0 + \phi_1 x_n$, por lo tanto, la minimización constará en derivar respecto de ambos coeficientes:

$$\frac{\partial s}{\partial \phi_0} = 0 \leftrightarrow \sum_{n=1}^{N} y_n = N\phi_0 + \phi_1 \sum_{n=1}^{N} x_n \tag{7}$$

$$\frac{\partial s}{\partial \phi_1} = 0 \leftrightarrow \sum_{n=1}^{N} y_n x_n = \phi_0 \sum_{n=1}^{N} x_n + \phi_1 \sum_{n=1}^{N} x_n^2 \tag{8}$$

Despejando ϕ_0 de la ecuación 7 y reemplzando en la ecuación 8 despejamos que:

$$\phi_1 = \frac{\sum_{n=1}^{N} y_n x_n - \frac{1}{N}\sum_{n=1}^{N} y_n \sum_{n=1}^{N} x_n}{\sum_{n=1}^{N} x_n - \frac{1}{N}\sum_{n=1}^{N} x_n \sum_{n=1}^{N} x_n} \tag{9}$$

$$\phi_0 = \frac{1}{N}\sum_{n=1}^{N} y_n - \frac{\phi_1}{N}\sum_{n=1}^{N} x_n \tag{10}$$

Si asumimos el error en "x" en lugar de "y", el proceso será análogo: Tenemos la probabilidad:

$$P(x)dx = \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{\frac{-(x_n - x(y_n,\phi))^2}{2\sigma_n^2}} dx \tag{11}$$

Por lo tanto, buscaremos $s = min\{\sum_{n=1}^{N} \frac{(x_n - x(y_n,\phi))^2}{2\sigma_n^2}\}$, con $x = \frac{1}{\phi_0} y_n - \frac{\phi_0}{\phi_1}$. Por lo tanto, si llamamos $A = \frac{1}{\phi_0}$ y $B = \frac{\phi_0}{\phi_1}$, minimizando tenemos:

$$A = \frac{1}{\phi_1} = \frac{\sum_{n=1}^{N} x_n y_n - \frac{1}{N}\sum_{n=1}^{N} x_n \sum_{n=1}^{N} y_n}{\sum_{n=1}^{N} y_n - \frac{1}{N}\sum_{n=1}^{N} y_n \sum_{n=1}^{N} y_n} \quad (12)$$

$$B = -\phi_0 A = \frac{1}{N}\sum_{n=1}^{N} x_n - \frac{1}{N}\sum_{n=1}^{N} y_n A \quad (13)$$

Concluyendo finalmente:

$$\phi_1 = \frac{\sum_{n=1}^{N} y_n - \frac{1}{N}\sum_{n=1}^{N} y_n \sum_{n=1}^{N} y_n}{\sum_{n=1}^{N} x_n y_n - \frac{1}{N}\sum_{n=1}^{N} x_n \sum_{n=1}^{N} y_n} \quad (14)$$

$$\phi_0 = -\frac{\phi_1}{N}\sum_{n=1}^{N} x_n + \frac{1}{N}\sum_{n=1}^{N} y_n \quad (15)$$

Donde encontramos resultados muy similares para ambos casos, una diferencia más notable aparecería si tomamos errores en ambos parámetros al mismo tiempo.

3.2. Distribución de Poisson: estimador de máxima verosimilitud

Podemos aplicar el mismo procedimiento que para deducir el método de cuadrados mínimos, pero ahora la distribución será de Poisson: Tenemos n datos que siguen la distribución: $P(x_i) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$ Por lo tanto, queremos maximizar:

$$L = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \quad (16)$$

$$\log(L) = \sum_{i=1}^{n} \log(\lambda^{x_i}) + \sum_{i=1}^{n} \log(e^{-\lambda}) - \sum_{i=1}^{n} \log(x_i!) \quad (17)$$

$$\log(L) = \sum_{i=1}^{n} x_i \log(\lambda) + n\lambda - \sum_{i=1}^{n} \log(x_i!) \quad (18)$$

Quiero estimar λ, por lo tanto, maximizo respecto a λ:

$$\frac{\partial \log(L)}{\partial \lambda} = -n + \sum_{i=1}^{n} \frac{x_i}{\lambda} = 0 \quad (19)$$

$$\lambda = \frac{\sum_{i=1}^{n}}{n} = \overline{x} \quad (20)$$

Por lo tanto, el promedio es el mejor estimador para λ.

3.3. Intervalo de confianza

Tenemos una distribución normal con una desviación estándar $S = 30$ y una muestra de $n = 50$ con un promedio $\overline{x} = 1550$. Queremos construir un intervalo de confianza del 95 % para la media. Por lo tanto, tomamos un $\alpha = 0.05$, con lo que la cota para el estadístico de prueba $z = \frac{\overline{x}-\mu}{\frac{S}{\sqrt{n}}}$ es $z_{\frac{\alpha}{2}} = 1.96$. Así tenemos que:

$$-z_{\frac{\alpha}{2}} < z < z_{\frac{\alpha}{2}} \quad (21)$$

$$\overline{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} < \mu < \overline{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \quad (22)$$

En conclusión, tenemos que $\mu \in [1541.68, 1558.32]$.

3.4. Hipótesis nula

Tenemos $\overline{x} = 7$ y $\sigma = 1.2$ para una muestra de $n = 10$. Con un 95 % de significancia ($\alpha = 0.05$) planteamos que:

- H_0:$\mu = 10$
- H_1:$\mu < 10$

Por lo tanto, debemos definir la región de rechazo como $z < z_{\frac{\alpha}{2}}$, siendo $z = \frac{\overline{x}-\mu}{\frac{S}{\sqrt{n}}} = -7.905$ y $z_{\frac{\alpha}{2}} = 1.645$. Por lo tanto queda en evidencia que z pertenece a la región de rechazo y por lo tanto $\mu < 10$.

4. Conclusiones

A lo largo del trabajo aprendimos a gestionar grandes volúmenes de datos: desde la selección de los mismos hasta el posterior análisis. Parte de este trabajo fue aplicar ajustes de cuadrados mínimos, los cuales además demostramos. Para los exoplanetas podemos concluir que hay muchos sesgos por resolver para poder decir algo sobre sus distribuciones. No podemos concluir que hay más planetas masivos que planetas pequeños por los sesgos, lo mismo sucede con los períodos. Tampoco podemos estar seguros de cuántos planetas hay en la galaxia, pues tenemos muestras principalmente en la vecindad solar. Encontramos una relación entre la masa y el radio de estos, pero no resulta del todo convincente, por lo que se requieren más observaciones.

Referencias

Schneider J., et al., 1995, The extrasolar planets encyclopaedia, `http://exoplanet.eu/`. `http://exoplanet.eu/`