Procesamiento y Análisis de Datos Astronómicos

11.- Detección y Sesgos

R. Gil-Hutton

Marzo 2020

Actividades:

Práctica 10:

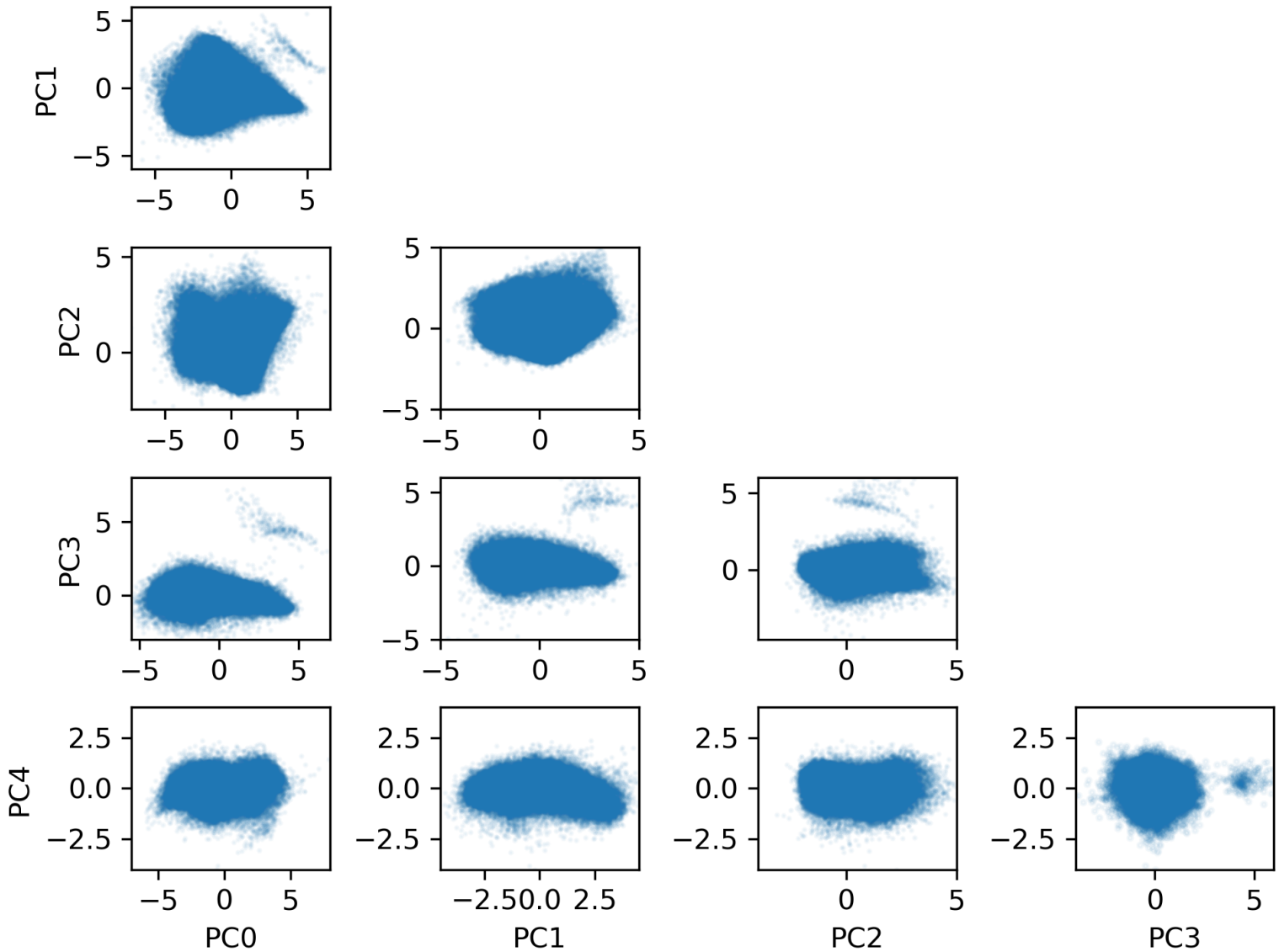
- Hacer un Análisis de Componentes Principales con todas las variables del archivo de datos que está usando para los ejercicios del curso.

- Buscar correlaciones entre variables aleatorias.
- Encuentra las direcciones de las correlaciones.
- Hace una transformación del sistema de coordenadas.
- Proyecta las correlaciones sobre los ejes del nuevo sistema.
- Cuanto más relevante sea la correlación original mayor será el autovalor correspondiente.

Actividades:

- En el nuevo sistema se deben buscar grupos.
- Hay un conjunto de planos entre pares de componentes que es más importante para analizar.
- En general, se utiliza para correlacionar conjuntos de datos físicos.
- Para agrupar datos se puede utilizar agrupamiento jerárquico porque la métrica es uniforme:
  - busco $min(|\mathbf{r_i} - \mathbf{r_j}|)$ para $i \neq j$.
  - reemplazo $\mathbf{r_i}$ y $\mathbf{r_j}$ con $(\mathbf{r_i} + \mathbf{r_j})/2$ y repito la operación.
  - se debe determinar cuál es el valor máximo de agrupamiento.

Actividades:



Detección:

La noción de detección requiere un cuidadoso análisis para poder definirla:

- Cuando se habla de detección se está haciendo referencia a cuestiones en el límite instrumental que se impone.
- Además se ve afectada por ciertas propiedades físicas de los objetos bajo estudio.
- Frecuentemente se define estudiando casos de no detección que fijan el límite práctico en una población bajo estudio.
- Es un concepto crítico en la definición de una muestra, un catálogo o survey.
- Se aplican siempre criterios estadísticos.

Detección:

- El proceso de detección siempre implica un ajuste a un modelo predeterminado.
- Es de fundamental importancia la manera en que se manejan los errores.
- Desde un punto de vista clásico, el proceso de detección dependerá de un ajuste a ciertos parámetros $\mathbf{a}$ a partir de una función de verosimilitud $\mathcal{L} = p(datos|\mathbf{a})$.
- El ajuste siempre se verá afectado por la completitud y la fiabilidad en el proceso de detección.

Detección:

- La completitud es la probabilidad de que la medición de una fuente, s, se encuentre por arriba de un cierto límite s_{lim} fijado previamente:

$$\mathcal{C}(datos, s_{lim}, s) = p(datos \geq s_{lim} | a = s)$$

- La fiabilidad es $1 - \mathcal{F}$ donde $\mathcal{F}$, la tasa de falsas alarmas, es la probabilidad de que el ruido produzca mediciones por arriba de un cierto límite fijado previamente:

$$\mathcal{F}(datos, s_{lim}) = p(datos \geq s_{lim} | a = 0)$$

- Usualmente, lo que se pretende es fijar el límite de tal modo de maximizar la completitud y minimizar las falsas alarmas.
- Lamenteblemente, una alta completitud incrementa siempre el número de falsas alarmas.

Detección:

- Al ser las probabilidades indicadas condicionales se sugiere fuertemente un planteo bayesiano.
- A medida que se analizan fuentes cada vez más débiles puede aparecer solapamiento y varios objetos pueden contribuir al total de la señal que se quiere medir.
- Esto fuerza a definir un límite de confusión que separa dos regímenes diferentes: el de fuentes únicas y el de fuentes múltiples.
- Usualmente el concepto de detección está muy afectado por algunos sesgos.

Sesgos de selección:

La mayoría de las mediciones astronómicas están afectadas por distancia y las cantidades que se obtienen siempre son aparentes. Por ejemplo:

- movimientos propios.
- magnitudes y/o luminosidad.
- formas y tamaños.
- perfiles de luminosidad.
- otros parámetros afectados por seeing.

En general, el valor real de la medición es una función más o menos complicada de su valor aparente y de la distancia.

Sesgos de selección:

Supongamos que se observa una cierta densidad de flujo S y se infiere una luminosidad $L = SR^2$, donde R es la distancia.

- El menor valor de S que vamos a considerar es s_{lim} y cualquier medición por debajo de este límite no será considerada.
- Se asume que los objetos a observar tienen una función de luminosidad $\rho(l)$ que nos da el número promedio de objetos con luminosidades próximas a l (L teórica) en la unidad de volumen.
- Usando los valores observados, $L_1, L_2, \cdots$, no es posible reproducir ρ.

Sesgos de selección:

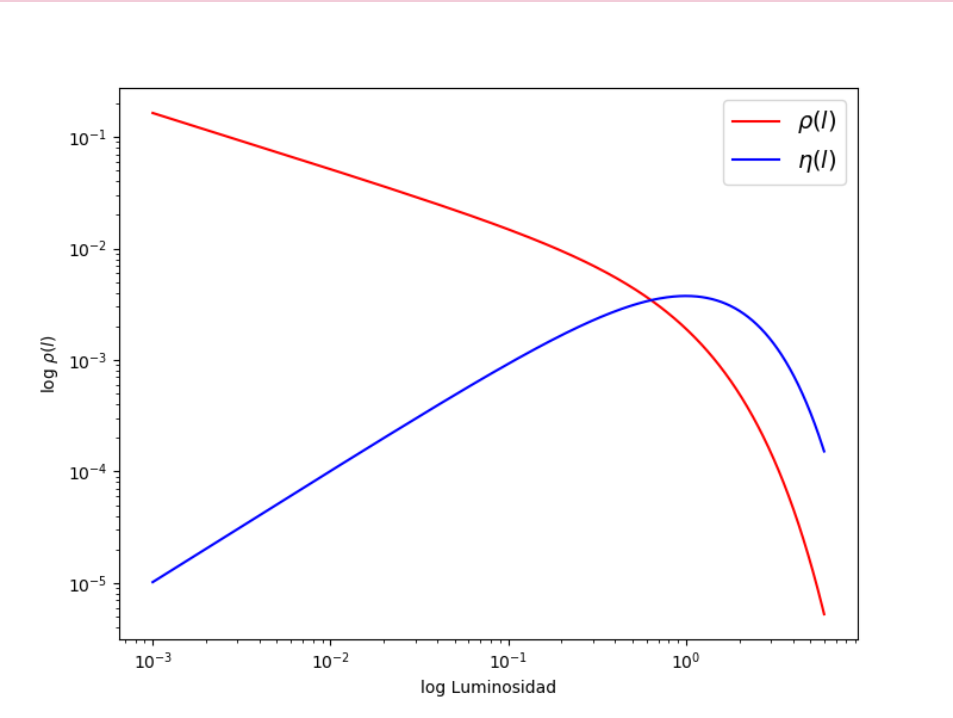
- Para ello es necesario trabajar con la distribución de luminosidades $\eta(l) \propto \rho(l)V(l)$, donde $V(l)$ es el volumen en el cual es necesario buscar para que fuentes con luminosidades intrínsecas l sean consideradas como detectadas.
- Entonces tenemos que:

$$\eta(l) \propto \rho(l) \left(\frac{l}{s_{lim}} \right)^{3/2}$$

indicando que η esta sesgada hacia valores de luminosidad intrínseca mayor que lo que esta ρ. Este sesgo es usual en astronomía y se suele denominar sesgo de Malmquist.

Sesgos de selección:

Ejemplo: Utilizando la función de luminosidad de Schechter, $\rho(l) \propto \left(\frac{l}{l^*}\right)^{\alpha} \exp(-l/l^*)$, con $l^* = 1$ y $\alpha = -0{,}5$, se obtiene la distribución de luminosidades $\eta(l)$ para un programa limitado por flujo multiplicando por $(l/l^*)^{3/2}$.



Sesgos de selección:

- El sesgo de Malmquist es un problema serio en muchas ramas de la astronomía.
- El impacto del sesgo depende de la función ρ, la cual no es bien conocida.
- El sesgo también puede estar presente en el caso de objetos cuyas propiedades están correlacionadas con otro parámetro que también esté sesgado. Ejemplo: la luminosidad de las regiones HII están correlacionadas con la luminosidad de la galaxia huesped.
- El sesgo de Malmquist aparece porque los objetos que son intrínsicamente brillantes pueden ser detectados dentro de volúmenes proporcionalmente más grandes ya que están menos afectados por distancia.

Sesgos de selección:

- En las muestras que están limitadas por flujo las propiedades intrínsecas se correlacionan con la distancia produciendo un sesgo.
- Esta correlación distancia-luminosidad está presente con mucha frecuencia, no es fácil detectarla y es muy difícil de remover.
- Consecuencia de este tipo de sesgo es que dos propiedades intrínsecas no relacionadas entre sí parecerán correlacionadas por su dependencia con la distancia.
- Si se encuentra una correlación es necesario hacer un profundo análisis para comprobar su validez, análisis que requerirá un detallado modelado del proceso de detección.

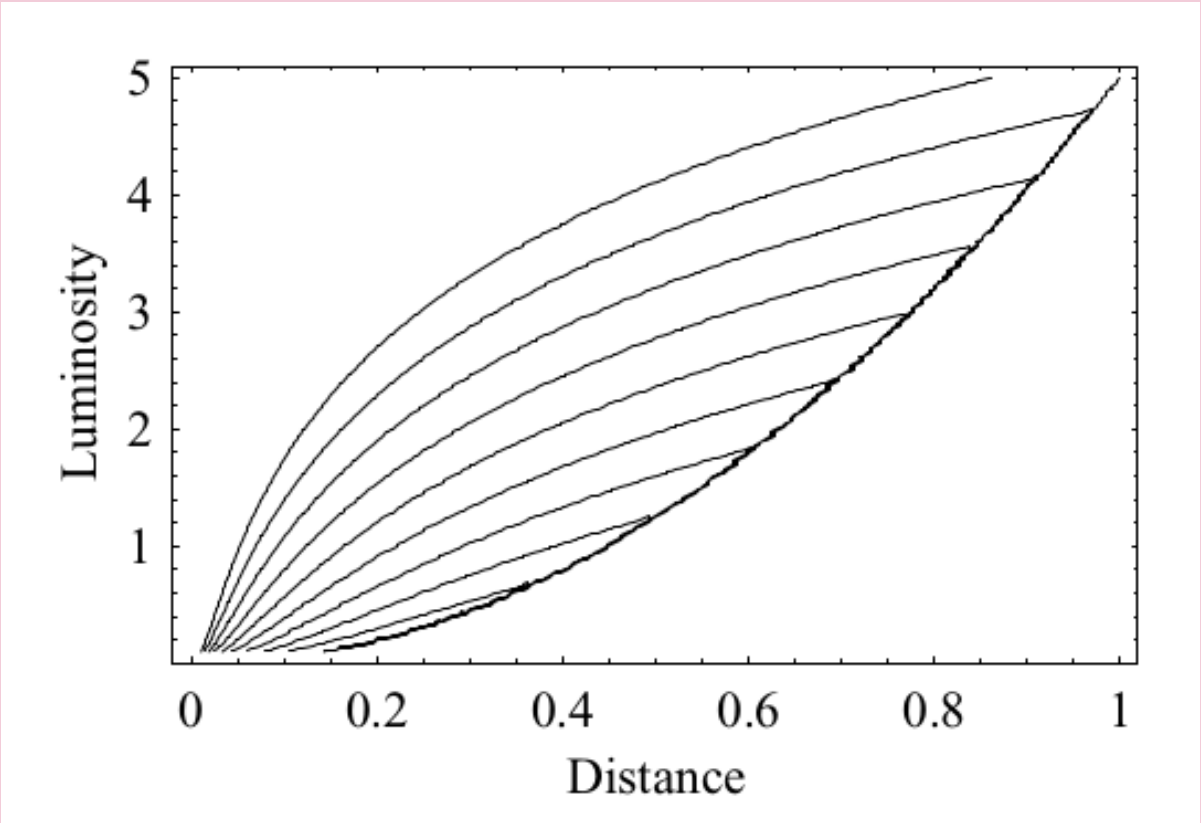
Sesgos de selección:

Ejemplo: Tomemos una muestra de objetos con las siguientes características:

- La probabilidad de que un objeto tenga luminosidad l es proporcional a la función de luminosidad de Schechter, $\rho(l) \propto \left(\frac{l}{l^*}\right)^{\alpha} \exp(-l/l^*)$ (asumamos $l^* = 10$ y $\alpha = 1{,}0$).
- La probabilidad de que un objeto este a una distancia R es proporcional a R^2.
- La probabilidad que un objeto con luminosidad L y distancia R este en la muestra es 1 si $L < s_{lim} R^2$ y 0 en otro caso.

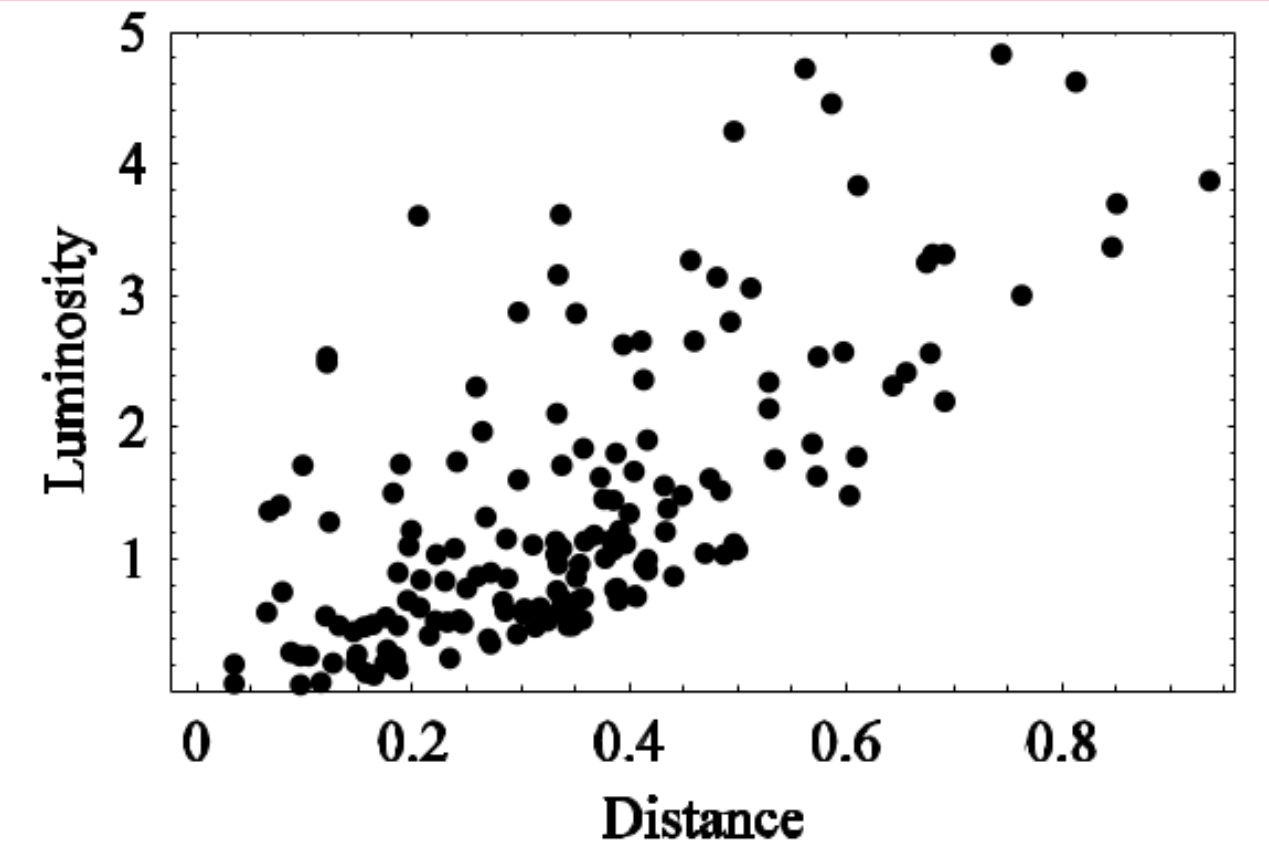
Sesgos de selección:

La probabilidad total es el producto de estas tres probabilidades (los contornos están espaciados con intervalos logarítmicos):



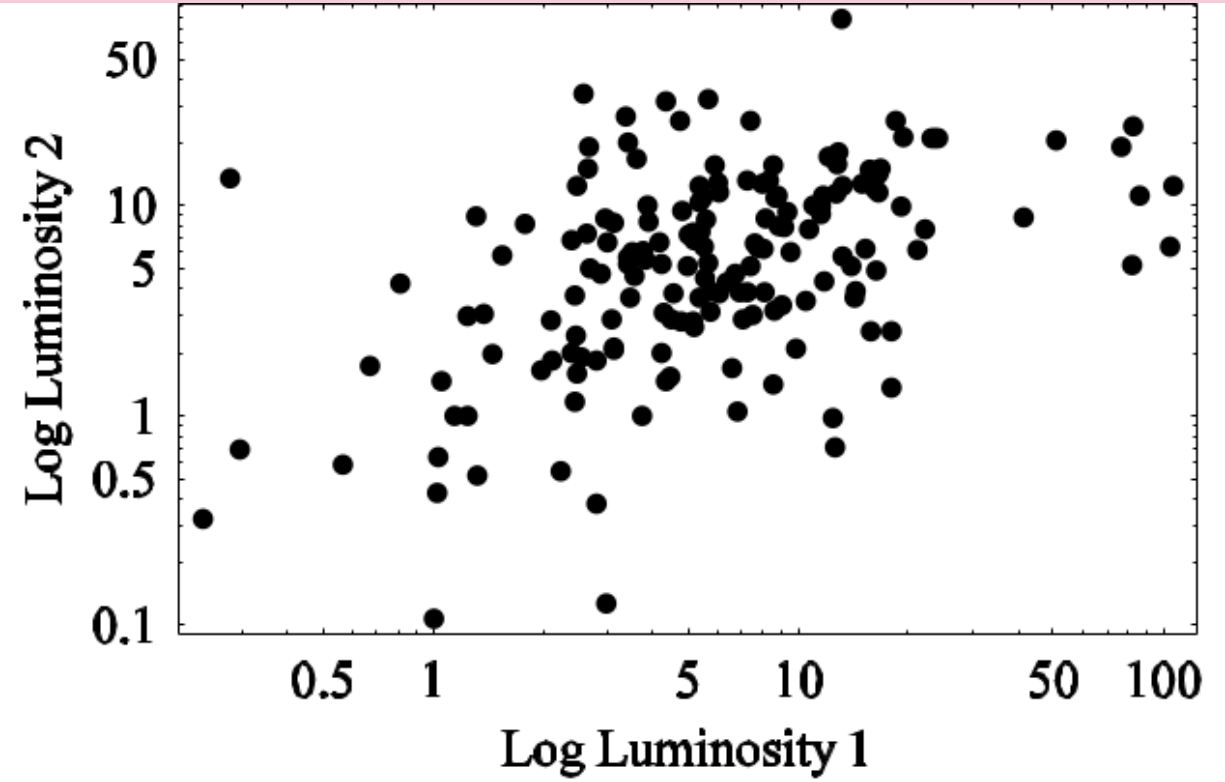
Sesgos de selección:

Si hacemos una simulación de la muestra en base a estos criterios obtenemos:



Sesgos de selección:

Si se asigna a cada objeto dos luminosidades que se supone son estadísticamente independientes, al comparar ambas luminosidades aparece una correlación por la dependencia con la distancia:



Sesgos de selección:

Una forma posible de determinar $\rho(l)$ para una cierta variable de una muestra es utilizar el método del volumen máximo:

- Se define $V_{max}(L_i)$ como el máximo volumen en el cual es posible encontrar un objeto de luminosidad L_i que pertenezca a la muestra.
- V_{max} depende de los límites fijados, la distribución de objetos en el espacio y la forma en que la detectabilidad depende de la distancia.
- Si se asume una distribución uniforme en el espacio y se conoce $V_{max}(L_i)$, una estimación para la función de luminosidad en un cierto intervalo es:

$$\hat{\rho}(B_{j-1} < l \leq B_j) = \sum_{B_{j-1} < L_i \leq B_j} \frac{1}{V_{max}(L_i)}$$

Sesgos de selección:

- El principal punto es la determinación de V_{max}, lo cual es posible hacer fijando el radio del volumen como la distancia del objeto más lejano con L_i observable dentro de la muestra.
- El error fraccional en cada intervalo es $1/N_j^{1/2}$, donde N_j es el número de objetos en ese intervalo.
- si se define a V como el volumen correspondiente a un radio igual a la distancia al objeto, la distribución de V/V_{max} es útil para determinar el límite de la muestra: si se fijó el límite correcto en el flujo cuando se calculó V_{max}, es de esperar que V/V_{max} se distribuya en forma uniforme entre 0 y 1.
- Un sumario sobre este método se puede encontrar en Willmer (1997, Astron. J. **114**, 898).

Sesgos de selección:

- Otro sesgo importante es producido por el error de selección observacional el cual produce efectos que compiten con el sesgo de Malmquist.
- Como hay muchos más objetos débiles que brillantes, $N(s)$ crece para s pequeño y usualmente se trunca la muestra para un cierto s_{lim}.
- El efecto del error de selección observacional se produce al convolucionar los valores observados de $N(s)$ con la distribución de errores.
- Como hay muchos más objetos débiles la muestra final se ve contaminada con un exceso de estos objetos, produciendo un sesgo notorio en la función de luminosidad para los objetos más débiles.

Análisis de supervivencia:

- Si contamos con una muestra de objetos que fueron seleccionados en base a una de sus propiedades y repetimos la muestra considerando una segunda propiedad es posible que en ambas muestras no se incluyan los mismos objetos debido a los diferentes criterios de detección.
- En estos casos es muy útil fijar cotas inferiores y/o superiores para los objetos no detectados.
- La parte de la estadística que trabaja con cotas se denomina análisis de supervivencia y las mediciones que son solo cotas se denominan mediciones censuradas.

Análisis de supervivencia:

En el caso de que alguna de las mediciones sea una cota, el análisis de supervivencia permite:

- estimar distribuciones intrínsecas.
- modelar y estimar parámetros.
- realizar pruebas de hipótesis.
- probar si hay correlaciones o independencia estadística.

Se asume siempre que la probabilidad de que una cota se encuentre disponible para cierta propiedad es independiente del verdadero valor de esa propiedad.

Normalización de distribuciones:

Supongamos que tenemos una muestra de objetos medidos en la banda A y repetimos el trabajo obteniendo la misma muestra para la banda B. Para algunos objetos se obtiene su luminosidad L_B pero para otros solo se puede fijar una cota superior L_B^U debido a que fueron detectados. El objetivo es reconstruir la distribución de L_B. Hay que tener en cuenta que:

- la distribución de L_B es proporcional a la función de luminosidad ρ_B.
- L_A puede sufrir de sesgo de Malmquist.
- si L_A y L_B están correlacionados, cualquier sesgo de L_A afectará a L_B.

Normalización de distribuciones:

Si es posible determinar intervalos de clase para los datos y las cotas, la probabilidad estimada en cada intervalo de L_B se puede obtener usando la relación recursiva de Avni et al. (1980, Astrophys. J, **238**, 800):

$$\hat{p}_k = \frac{n_k}{M - \sum_{j=1}^{k} \left(\frac{u_j}{1 - \sum_{i=1}^{j-1} \hat{p}_i} \right)} \qquad \hat{p}_1 = \frac{n_1}{M - u_1}$$

donde n_k es el número de objetos detectados en el intervalo k, u_k es el número de cotas en ese intervalo, y M es el total de objetos (detectados y cotas). Para usar la fórmula con cotas superiores hay que numerar los intervalos de mayor a menor según el valor de la propiedad observada y al revés en el caso de cotas inferiores.

Normalización de distribuciones:

Un estimador alternativo de la distribución acumulativa es el estimador de Kaplan-Meier, el cual tiene la ventaja de no requerir intervalos de clase. De todos modos, hay que tener cuidado porque al ser acumulativo los errores se correlacionan desde un valor a la estimación del siguiente:

$$\hat{\mathcal{K}}(L_k) = 1 - \prod_{i=1}^{k-1} (1 - d_i/n_i)^{\delta_i}$$

donde d_i es el número de objetos con valor L_i, δ_i es 1 si es una detección y 0 en el caso de una cota, y:

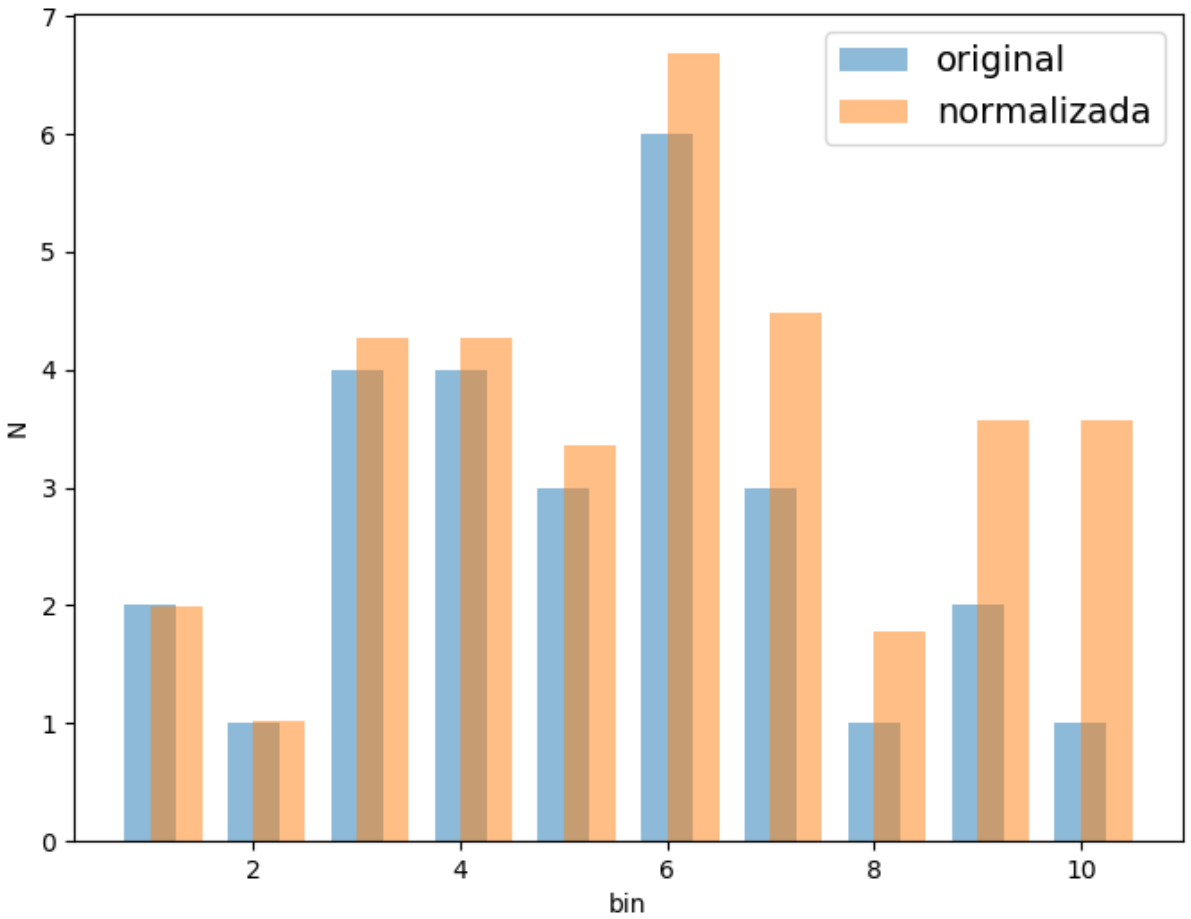
- para cotas inferiores, los datos se ordenan de manera creciente, n_i es el número de objetos con $L \geq L_i$.
- para cotas superiores, los datos se ordenan de manera decreciente, n_i es el número de objetos con $L \leq L_i$.

Normalización de distribuciones:

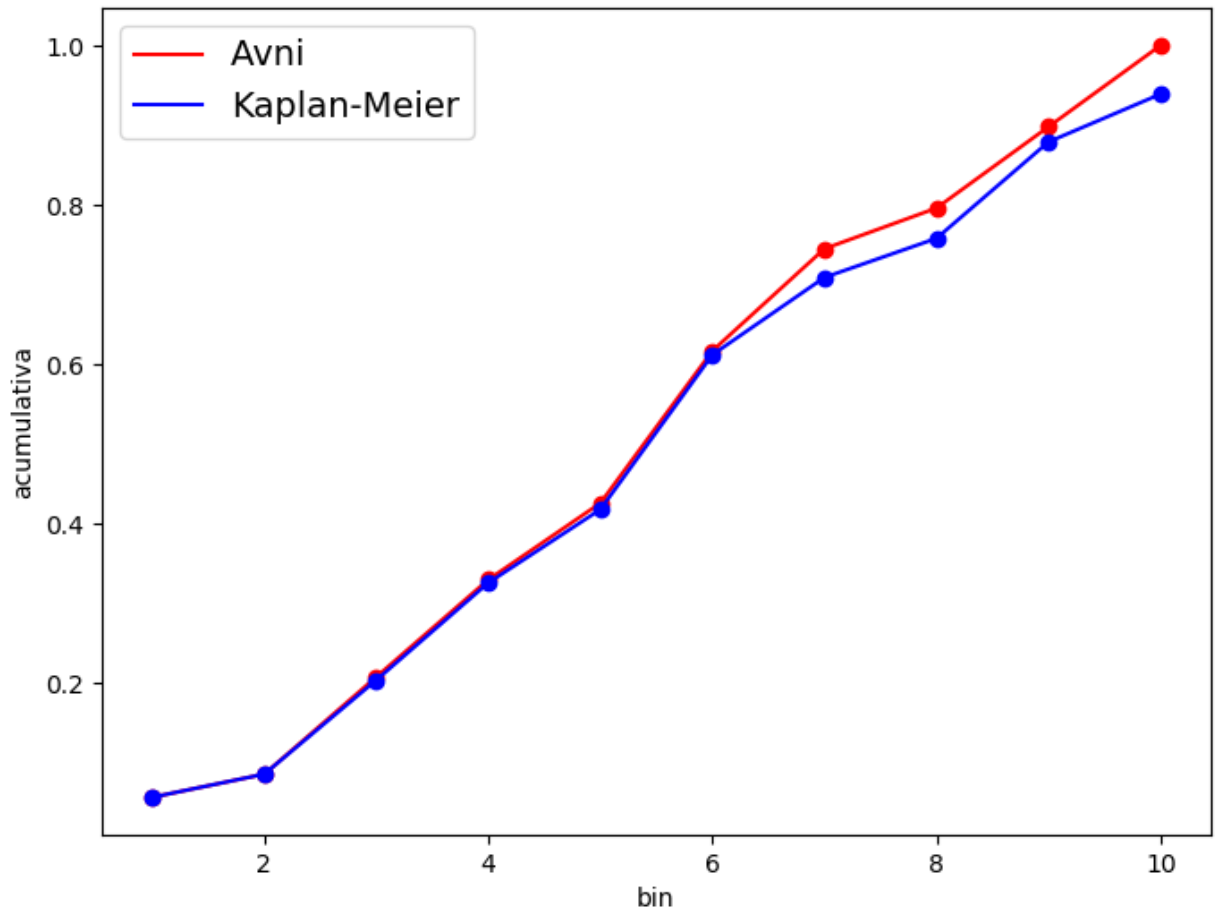
Ejemplo: Usando datos de Avni et al. (1980) de la distribución de índices espectrales para quasares, donde u son cotas inferiores correspondientes a la no detección en rayos X:

k	n	u	$\hat{p}$	$\hat{\mathcal{K}}$
1	2	0	0.057	0.057
2	1	1	0.029	0.086
3	4	1	0.122	0.204
4	4	0	0.122	0.326
5	3	1	0.096	0.418
6	6	0	0.191	0.612
7	3	3	0.128	0.709
8	1	1	0.051	0.758
9	2	0	0.102	0.879
10	1	1	0.102	0.939

Normalización de distribuciones:



Normalización de distribuciones:



Límite de confusión:

- Como los objetos débiles son mucho más numerosos que los brillantes a ciertos niveles de brillo se superponen, no es posible resolverlos y la señal termina siendo una mezcla de objetos de diferente intensidad amalgamados por la PSF del instrumento. Entonces aparece el concepto de límite de confusión.
- La raiz del problema es innegablemente instrumental y producida por la mezcla de señales que pueden ser erróneamente interpretadas como una sola fuente discreta.
- Para recuperar en este caso la distribución de las mediciones en función de las fuentes que contribuyen se debe utilizar un método de bondad de ajuste denominado probabilidad de deflexión.

Límite de confusión:

Los detalles del método se pueden obtener en Scheuer (1974, MNRAS **166**, 329) y Condon (1974, Astrophys. J. **188**, 279). Como ejemplo se considerará un caso unidimensional:

- Una fuente de brillo f se observa bajo una PSF $\Omega(x)$, donde x es una "distancia" al eje central del instrumento (en valores de ángulos, longitud de onda, etc.).
- La intensidad medida es:

$$s(x) = f\Omega(x)$$

- Si las fuentes individuales que contribuyen son $N(f)$ para un cierto brillo f, el aporte total debido a un cierto número de fuentes individuales es:

$$s(x) = \int N(f)\Omega(x)dx$$

Límite de confusión:

- La anterior expresión es la distribución observada de cuentas para un solo objeto y es una probabilidad condicional, de donde se obtiene $p_1(s)$ que es la probabilidad de una intensidad s como resultado de una única fuente.
- Si una cierta medición D proviene de muchas fuentes aportando señal y si las fuentes están distribuídas al azar se espera que sigan una distribución de Poisson. La probabilidad de obtener una medición D será:

$$p(D) = \sum_{k=1}^{\infty} p_k(s) \frac{\mu^k}{k!} e^{-\mu}$$

donde μ es la media del número de fuentes y $p_k(s)$ es la distribución de probabilidad de una intensidad s debido a la k-ésima fuente.

Límite de confusión:

- Como la probabilidad de una suma está dada por la autocorrelación de la distribución de los términos de la suma, asumiendo que se distribuyen de igual forma, tenemos una relación entre sus transformadas de Fourier:

$$P_k(\omega) = P_1(\omega)^k$$

- Entonces, la transformada de Fourier de $p(D)$ es $P(\omega) = \exp(R(\omega) - R(0))$, donde $R(\phi)$ es la transformada de:

$$r(s) = \int N\left(\frac{s}{\Omega(x)}\right)\frac{dx}{\Omega(x)}$$

siendo N el número de fuentes involucradas. En general se asume que $N(f) = kf^{-\gamma}$.

Actividades:

Práctica 11:

- Utilizando la descomposición en Componentes Principales obtenida haga un agrupamiento jerárquico de los objetos. Indique cuál sería un valor máximo para definir grupos.
- Simule una muestra limitada por flujo (luminosidad, magnitud absoluta) de objetos distribuídos en un gran volumen de espacio utilizando para el flujo una distribución en ley de potencias del tipo:

$$\rho(> f) \propto f^{-\gamma}$$

y una distribución volumétrica uniforme para la distancia. Para armar la muestra fije un valor límite para el flujo de manera arbitraria.

Actividades:

Práctica 11:

- Con esta muestra:
  - aplique el método V_{max} y vea si es posible recuperar la distribución de flujo utilizada.
  - produzca errores utilizando bootstrap y compare los resultados con errores basados en $\sqrt{N}$.
  - asigne a cada objeto dos flujos diferentes, detecte y utilice el método de Kaplan-Meier para normalizar la distribución. Aparece alguna correlación?.

Entrega

Para la próxima clase

Por consultas: ricardo.gil-hutton@conicet.gov.ar
Grupo de Ciencias Planetarias - CUIM 2