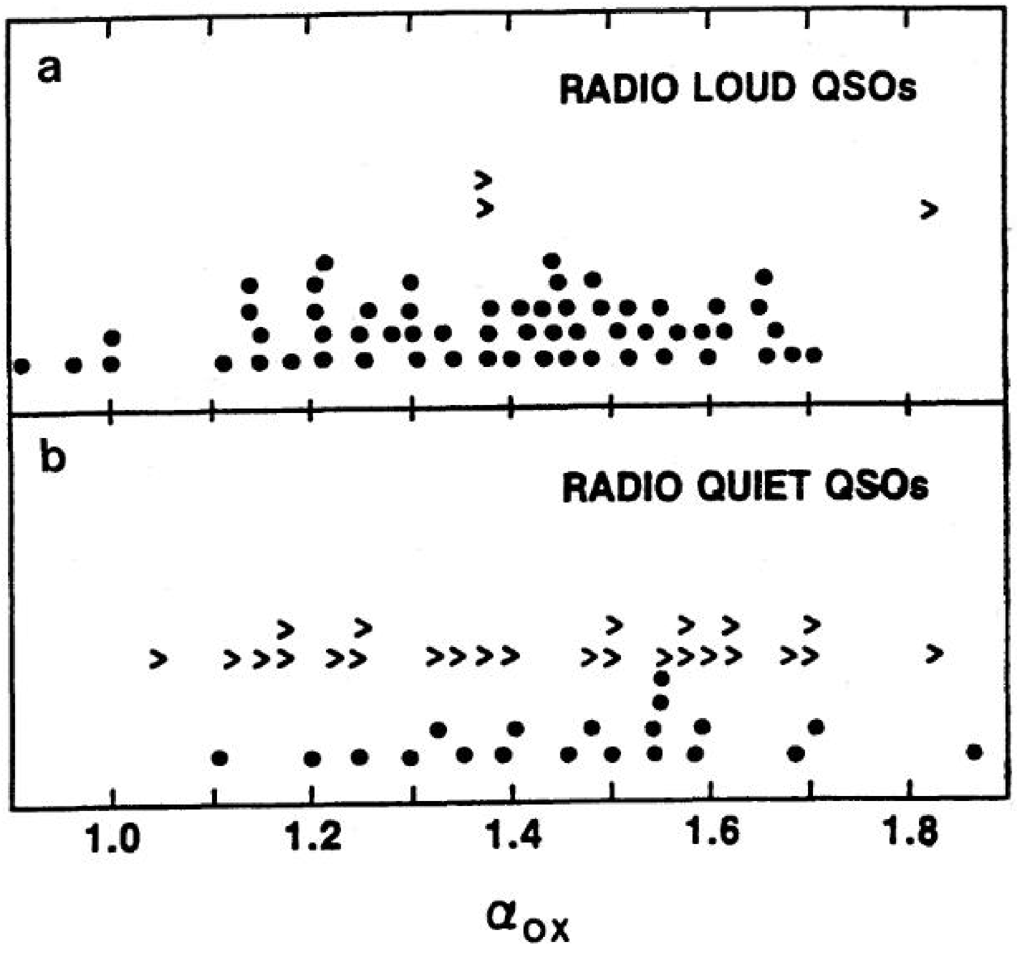
天文探测中上限数据的处理

1984，ApJ，293，192

1986，ApJ，306，490

问题


a
RADIO LOUD QSOs
b
RADIO QUIET QSOs
1.0
1.2
1.4
1.6
1.8
αox

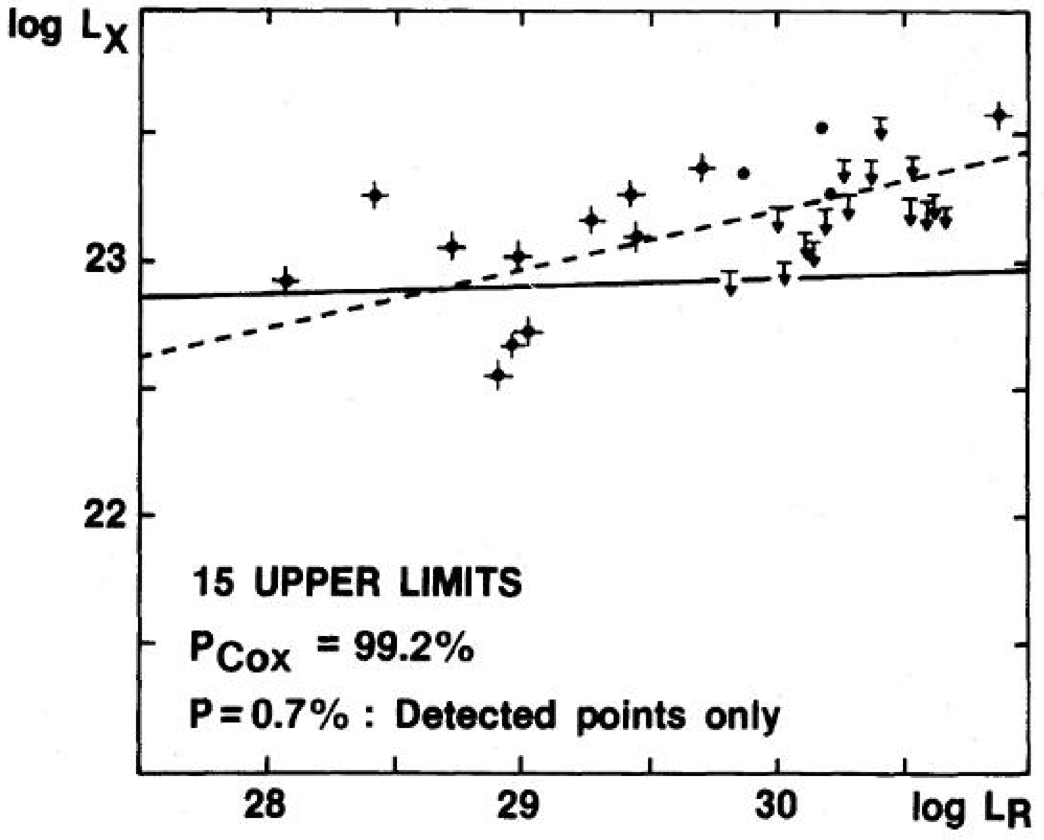

log LX
23
22
15 UPPER LIMITS
PCox = 99.2%
P = 0.7% : Detected points only
28
29
30
log LR

FIG. 2b

Survival analysis

- 天文：censored Data （upper/lower limit)
- 医学：lifetime Data
- https://scikit-survival.readthedocs.io

> Let T denote a continuous non-negative random variable corresponding to a patient's survival time. The survival function $S(t)$ returns the probability of survival beyond time t and is defined as
>
> $$S(t) = P(T > t).$$

KAPLAN-MEIER PRODUCT-LIMIT ESTIMATOR

$$F(t) = P(T \leq t) .$$

$$\begin{aligned} S(t) &\equiv P(T \geq t) \\ &= P(T > t) + P(T = t) \\ &= 1 - F(t) + P(T = t) . \end{aligned}$$

分步积分

$$\begin{aligned} \mu &\equiv \int_0^\infty x dF(x) , \\ &= \int_0^\infty S(x) dx . \end{aligned}$$

x是人的生存期或者调查时间（调查的时候人还没死）

$$x_{(1)} < x_{(2)} < \cdots < x_{(n)} .$$

Set $x_{(0)} \equiv 0$. For $i = 0, 1, 2, \ldots, n-1$, let

$$P_i = P[T \geq x_{(i+1)} \mid T \geq x_{(i)}]$$

$$\begin{aligned} S(x_{(j)}) &= P[T \geq x_{(j)}] , \\ &= \prod_{i=0}^{j-1} P[T \geq x_{(i+1)} \mid T \geq x_{(i)}] , \\ &= \prod_{i=0}^{j-1} P_i . \end{aligned}$$

The P_i values are estimated as follows. For $i \geq 1$, if $x_{(i)}$ is not a censored value, there are $n - i + 1$ "true" values $\{x_{(j)}\}$ at least as large as $x_{(i)}$ of which only one ($x_{(i)}$ itself) is not at least as large as $x_{(i+1)}$. In this case, estimate P_i by

$$\hat{P}_i = 1 - 1/(n - i + 1) . \quad (5)$$

If, on the other hand, $x_{(i)}$ is a censored value, it is known that all the true values in the set $\{x_{(j)}\}$ which are at least as large as $x_{(i)}$ are also at least as large as $x_{(i+1)}$. Hence, here estimate P_i by

$$\hat{P}_i = 1 . \quad (6)$$

最大似然法：需要假设分布函数f(y|x)

A likelihood function describing a given data set can be defined using the above formulations. Consider a detected point falling in a bin $(z_i, z_i + \Delta z)$. The probability that this occurs is determined by the probability density and is

$$P_D(z_i) \approx f(z_i)\Delta z \ . \tag{10}$$

If an object is right censored at z_i, so that the true location of the point is somewhere between z_i and ∞, the contribution from this point can be written in terms of the survival function

$$P_C(z_j) \approx \int_{z_j}^{\infty} f(t)dt = S(z_j) \ . \tag{11}$$

If there are m detected observations, and n censored observations, the likelihood function is expressed by

$$L = \prod_D^m f(z_i) \cdot \prod_C^n S(z_j)(\Delta z)^m \ ,$$

where $\prod_D^m$ denotes the product over the m detected points, and $\prod_C^n$ denotes the product over the n censored points. Since $(\Delta z)^m$ does not contribute to the maximum, the likelihood can be rescaled to be

$$L = \prod_D^m f(z_i) \prod_C^n S(z_j) \ . \tag{12}$$

Taking the logarithm, we get the log likelihood function

$$l = \sum_D^m \log f(z_i) + \sum_C^n \log S(z_j) \ . \tag{13}$$