Looking for correlations in censored data: the Kendall $\tau$ test

Laura Venuti

Rencontres d’Astrostatistique
14/11/2014
Model of Spectral Energy Distribution (SED) for a young star with (upper panel) and without (lower panel) accretion disk.
The context

$u$-$r$ colors and $r$-band magnitudes for accreting (red circles) and non-accreting (blue triangles) members of the star-forming region NGC 2264 (3 Myr old).

http://caravan.astro.wisc.edu/protostars/

The problem

- 240 objects
- Sample complete in the mass range probed
- 77.5% detections, 22.5% upper limits
- Upper limits are mass-dependent

The problem

- 240 objects
- Sample complete in the mass range probed
- 77.5% detections, 22.5% upper limits
- Upper limits are mass-dependent

(How) does the mass accretion rate correlate with mass?

The issue

How to properly account for censored data?

- Restrict the statistical analysis to detections
- Take upper limits as actual detections

Least-squares fit
The issue

How to properly account for censored data?

• Restrict the statistical analysis to detections
• Take upper limits as actual detections

bypassing the actual information contained in upper limits
The issue

Test of correlation for a synthetic, flat distribution of mass accretion rates randomly assigned to a sample of objects with the same mass and detection limits as the population observed:

*an artificial correlation in the data, merely driven by mass dependence in the censoring effect, is assessed to a probability of >99% when either discarding upper limits or considering them as actual detections.*
The issue

How to properly account for censored data?

- Restrict the statistical analysis to detections
- Take upper limits as actual detections

bypassing the actual information contained in upper limits

- Perform the statistical analysis over the whole sample, weighing the different types of information conveyed by actual detections and upper limits
Kendall’s $\tau$ test for correlation

generalization to the case of censored data

- Helsel, D. R. 2012, Statistics for Censored Environmental Data, Wiley

- deals with multiple censoring (i.e., both on the x-axis and the y-axis):
  - ties (points at the same abscissa)
  - indeterminate relationships (upper limits)

- censored data lower the likelihood of finding a significant correlation, if this is present among the data; they do not “positively” affect the yes/no answer
Kendall’s $\tau$ test for correlation: how it works

Correlation: concordant/discordant variations on x and y

N points

$\frac{N (N-1)}{2}$ pairs

How many times do we find a positive/negative slope?
Kendall’s $\tau$ test for correlation: how it works

4 counters:

- $n_C = $ number of times we measure a positive slope among the sample of pairs
- $n_D = $ number of times we find a negative slope
- $n_{\text{ties}_X} = $ number of times we find a tie on $X$
- $n_{\text{ind}_Y} = $ number of times we have an indeterminate relationship due to upper limits
Kendall’s $\tau$ test for correlation: how it works

1° case
Pair of detections, $x_1 \neq x_2$

$$n_C = n_C + 1$$
$$n_D = n_D + 1$$
Kendall’s $\tau$ test for correlation: how it works

1° case
Pair of detections, $x_1 \neq x_2$

2° case
Pair of points, $x_1 = x_2$

$n_{\text{ties}_X} = n_{\text{ties}_X} + 1$
Kendall’s $\tau$ test for correlation: how it works

1° case
Pair of detections, $x_1 \neq x_2$

2° case
Pair of points, $x_1 = x_2$

3° case
1 detect, 1 upper lim, $x_1 \neq x_2$

$n_D = n_D + 1$
$n_C = n_C + 1$
$n_{\text{ind}_Y} = n_{\text{ind}_Y} + 1$
Kendall’s $\tau$ test for correlation: how it works

1° case
Pair of detections, $x_1 \neq x_2$

2° case
Pair of points, $x_1 = x_2$

3° case
1 detect, 1 upper lim, $x_1 \neq x_2$

4° case
Pair of upper limits, $x_1 \neq x_2$

$n_{\text{ind}_Y} = n_{\text{ind}_Y} + 1$
Kendall’s $\tau$ test for correlation: how it works

Correlation coefficient: $\tau = \frac{n_C - n_D}{n_{tot}}$

where $n_{tot}$ is the total number of pairs:

- $n_{tot} = 0.5 \ N \ (N-1)$

  or

- $n_{tot} = \sqrt{ (0.5 \ N \ (N-1) - n_{ties_X})* (0.5 \ N \ (N-1) - n_{ind_Y}) }$
Kendall’s $\tau$ test for correlation: how it works

In the null hypothesis of no correlation, $\tau$ follows the normal distribution centered on zero and with variance

$$\sigma^2 = \frac{2 (2N + 5)}{9N (N - 1)}$$

(to be adapted to take into account different groups of ties)

$$Z = \frac{\tau}{\sigma}$$ expresses the significance of the correlation result
The results

First question: does $M_{acc}$ correlate with $M_*$?
Yes ($\tau = 0.28$, $\sigma = 0.04$, $z > 6$)
Tested against purposely generated flat distributions of $M_{acc}$ values in the same $M_{acc}$ range and with the same upper limits distribution – no correlation is statistically detected in these cases

Second question: what slope?
Akritas-Theil-Sen non-parametric regression (Feigelson & Babu 2012):
- Take a first guess for the slope and explore a range around this value
- For each test slope, subtract the $Y=mX$ trend from the observed distribution and repeat the Kendall’s $\tau$ test
- The best value of slope $m$ is the one that, subtracted to the distribution, produces $\tau = 0$, while the $m$ values range corresponding to $\tau = \pm n\sigma$ will provide an $n\sigma$ error bar on the slope

For the case in exam, a slope of 1.5 +/- 0.2 is obtained.