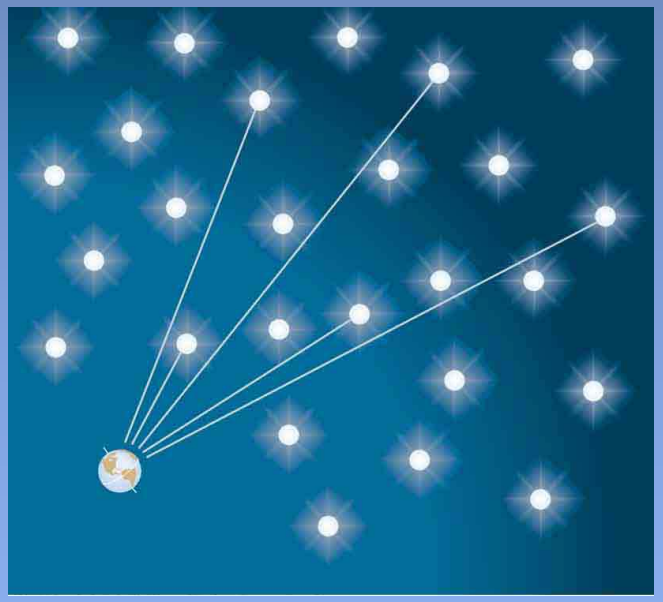
Censoring and truncation in astronomical surveys

Eric Feigelson

2nd East Asian Workshops on Astrostatistics
Nanjing & Guiyang
July 2018

Historical background: star counts
Statistical treatment of censoring
Statistical treatment of truncation

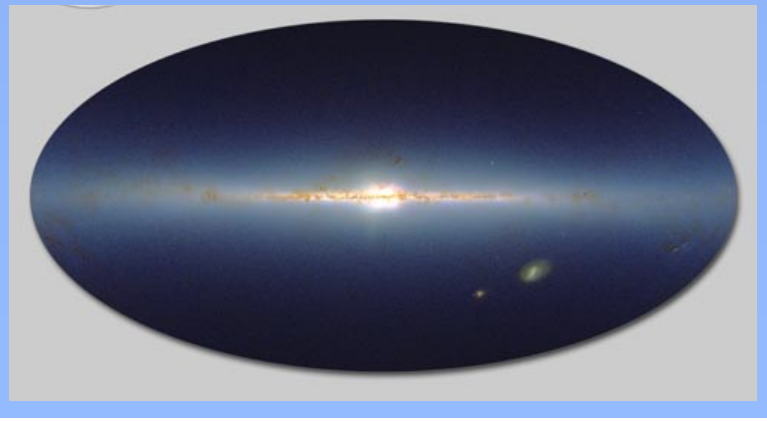
Star counts: The first flux limited surveys



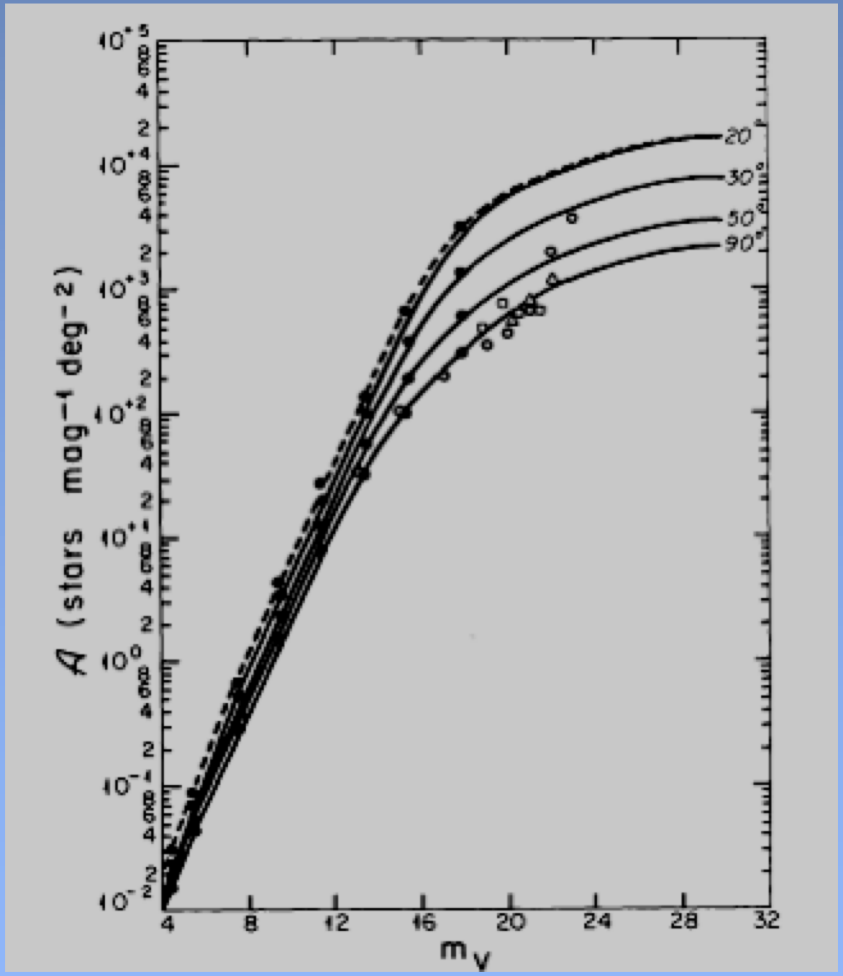
For a uniform population of objects distributed randomly in transparent space:

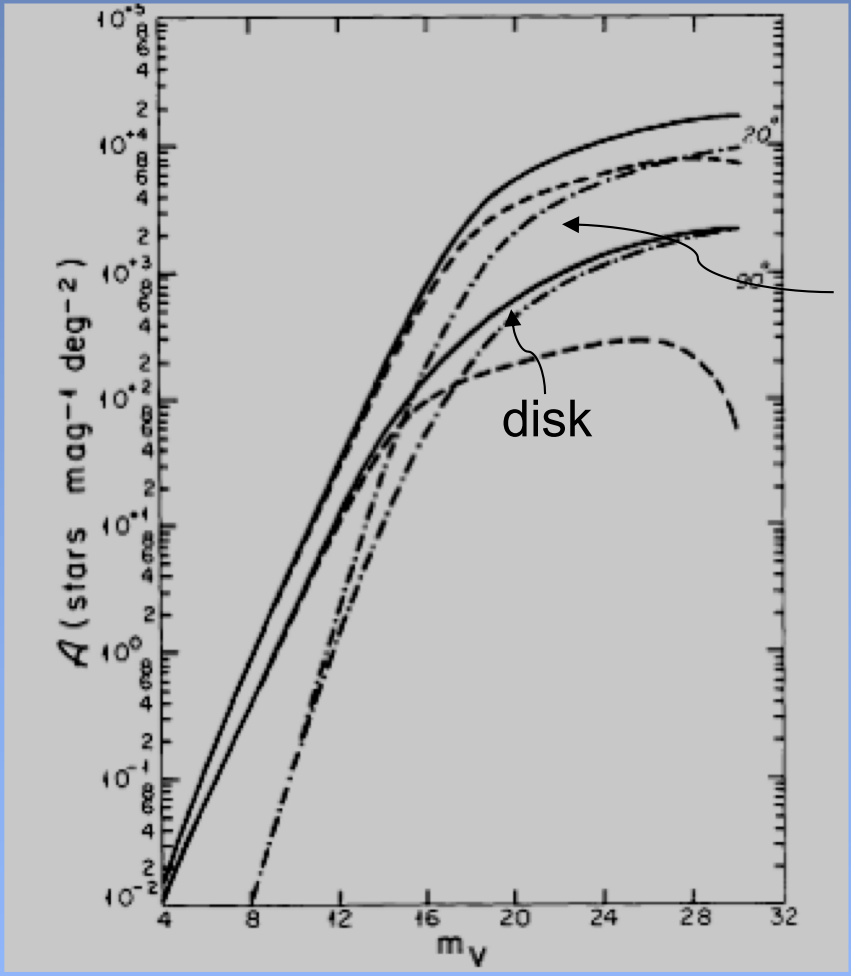
$$S = L / 4 \pi D^2$$
$$V = 4/3 \pi D^3$$
$$N \sim V \sim S^{-3/2}$$

William Herschel (1785) used deviations from this prediction to infer that the Universe (now known as our Milky Way Galaxy) is limited in extent (~1 kpc) and is elongated in shape.



Star counts at different Galactic latitudes





The Galaxy has multiple components with different spatial distributions producing strong deviations from the logN~-1.5 logS law.

Bahcall & Soneira 1980 The universe at faint magnitudes. I - Models for the galaxy and the predicted star counts

But, near the Galactic plane, space is very opaque due to interstellar dust (Trumpler 1930). Thus, Herschel's galaxy is ~20 times too small.

Big effort in 1910-50s to address the "Fundamental equation of stellar statistics" to simultaneously establish the distribution of stellar luminosities and their spatial distribution in the Galactic disk. This effort essentially failed in the disk due to the patchy absorption, but functions reasonably well towards the Galactic poles ($b=90^{o}$). It is described in the monograph "Statistical Astronomy" (Trumpler & Weaver 1953).

Selection bias in flux-limited surveys

Astronomers often struggle to detect faint characteristics of celestial populations and fail. Many surveys are flux-limited ($F=L/4\pi d^2$), and are limited to detecting the closer and/or more luminous members of a population. This leads to biased samples: at large distances, high-luminosity objects are over-represented (e.g., the majority of V<2 stars are giants and supergiants). Two types of bias in flux-limited surveys:

1.A `blind' astronomical survey of a portion of the sky is thus truncated at the sensitivity limit, where **truncation** indicates that the undetected objects, even the number of undetected objects, are entirely missing from the dataset.

1.In a `supervised' astronomical survey where a particular property (e.g., IR emission with Herschel, HCO^+ line emission with ALMA, redshifted Lyα emission with HETDEX) of a previously defined sample of objects is sought, some objects in the sample may be too faint to detect. This gives **upper limits** or **left-censored** datapoints.

Statistical challenges of censoring & truncation

In a truncated or censored sample, neither the same mean nor the median converge to the population values.

The sample distribution (e.g. a Pareto luminosity function $N(L) \sim L^{-\alpha}$) will not converge to the population distribution because faint objects are underrepresented.

Relationships among variables (e.g. $L_{opt} \sim L_{radio}^{-\alpha}$) may be less affected, but sample correlations will be biased unless nondetections are adequately treated.

The situation is much better for censored samples
than for truncated samples,
as the number and censored values of
undetected objects are known …

Survival analysis

A large field of applied statistics called **survival analysis** developed during 1950-80s to treat right-censoring in several applications:

1. **Life insurance** To calculate annuities, Edmund Halley (1692) constructed `life tables' from birth/death records in a city. But some people leave the city; this leads to right-censoring in their survival time. (Need: Univariate distribution function)

2. **Industrial reliability** A company manufactures Widget Mark IV. To find improvement over Mark III, operate 100 widgets until 20% fail. From failure times, compare lifetime distributions. (Need: 2-sample test)

3. **Biometrics** Dangerous Tobacco Co. wants to test the effect of smoking on cancer rates. To compare longevities, 100 rats are given smoke at 0-5 cigarette packs/day. After 1 year, the experiment is stopped but some rats are still alive with right-censored survivals. (Need: Regression)

4. **Astronomy** The VLA seeks 21cm hydrogen line emission from starburst galaxies. Due to low star formation rate and/or large distance, half are not detected. Compare to LINERs and infrared dust emission. (Need: distribution function, 2-sample test, regression)

Concepts of survival analysis

Censoring: The sample is known and all objects are observed, but some are undetected in the desired property. They can be displayed on graphs with `arrows' rather than `points'. Synonyms: Upper limits = Nondetections = Left censoring. Lower limits = right censoring.

Truncation: An unknown number of objects are missing from the sample due to nondetection. They cannot be displayed on graphs.

Survival and hazard functions: Univariate functions closely related to the e.d.f. and p.d.f. widely used in survival analysis.

Proportional hazard model and Cox regression: A mathematically convenient form of the dependence of the hazard function on uncensored variables.

Statistical foundations of survival analysis

The **survival function S(x)** gives the probability that an object has a value of X above a specified value x (the inverse of the e.d.f.):

$$\begin{aligned} S(x) &= Prob(X > x) = \frac{\#observation \geq x}{n} \\ &= 1 - F(x) = 1 - \int_0^x f(s)ds. \end{aligned}$$

The **hazard rate** gives the probability that an object will have a specified value x:

$$h(x) = \frac{f(x)}{S(x)} = -\frac{d\ lnS(x)}{dx}$$

(Note the p.d.f. is the product of the survival function and hazard rate.)

Example: Pareto distribution

$$f(x) = \frac{\theta\lambda^\theta}{x^{\theta+1}}$$

$$\begin{aligned} S(x) &= \frac{\lambda^\theta}{x^\theta} \\ h(x) &= \frac{\theta}{x}. \end{aligned}$$

Kaplan-Meier estimator

For a randomly censored univariate X, the Kaplan-Meier estimator is the unique unbiased nonparametric maximum likelihood estimator of the survival function is

$$\hat{S}_{KM}(x) = \Pi_{x_i \geq x} \left(1 - \frac{d_i}{N_i}\right)$$

where d_i is the number of occurrences at x_i (d_i=1 if no ties are present) and N_i is the number of `at risk' objects left in the sample. An intuitive procedure: construct the e.d.f. of detected points, but increase the step size at low values by redistributing the upper limits to the left.

The Kaplan-Meier estimator is asymptotically normal with variance

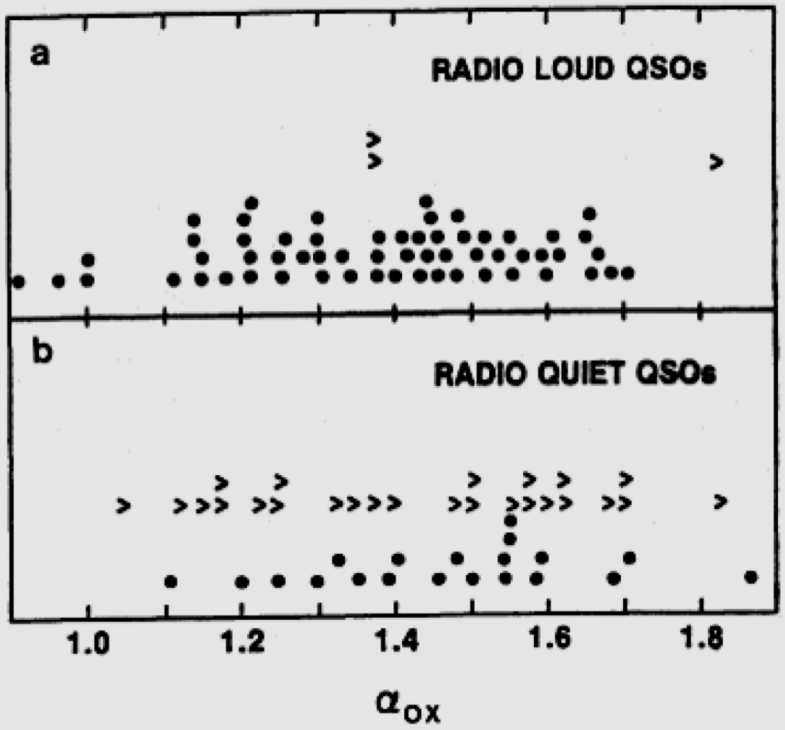
$$\widehat{Var}(\hat{S}_{KM}) = \hat{S}_{KM}^2 \sum_{x_i \geq x} \frac{d_i}{N_i(N_i - d_i)}$$

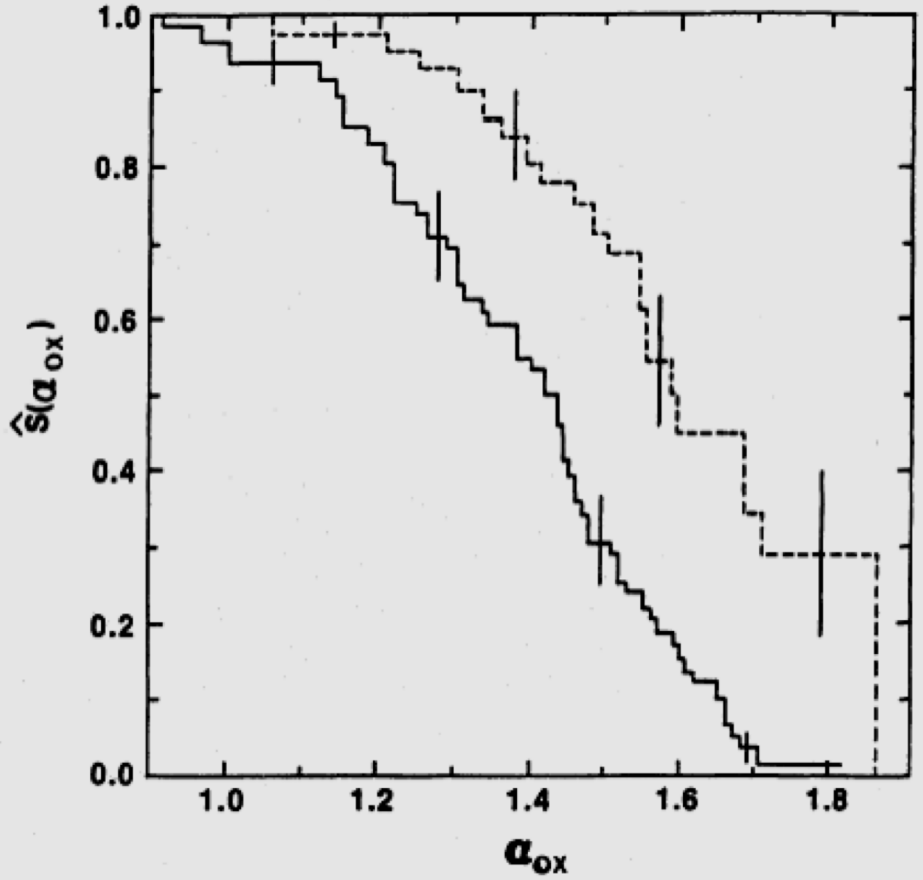
Kaplan, E. L.; Meier, P.: Nonparametric estimation from incomplete observations. J. Amer. Statist. Assn. 53:457–481, 1958.

Greenwood, M. The natural duration of cancer. Rpt Public Health (London) 33:1, 1926

Example of univariate survival analysis (Kaplan-Meier)

Redistribute-to-the-right algorithm (Efron)





Feigelson & Nelson 1986

Two-sample tests, correlation & regression

Several generalizations of 1930s nonparametric 2-sample tests (e.g. Wilcoxon) were developed in 1960-70s that treat censored values in reasonable fashions:

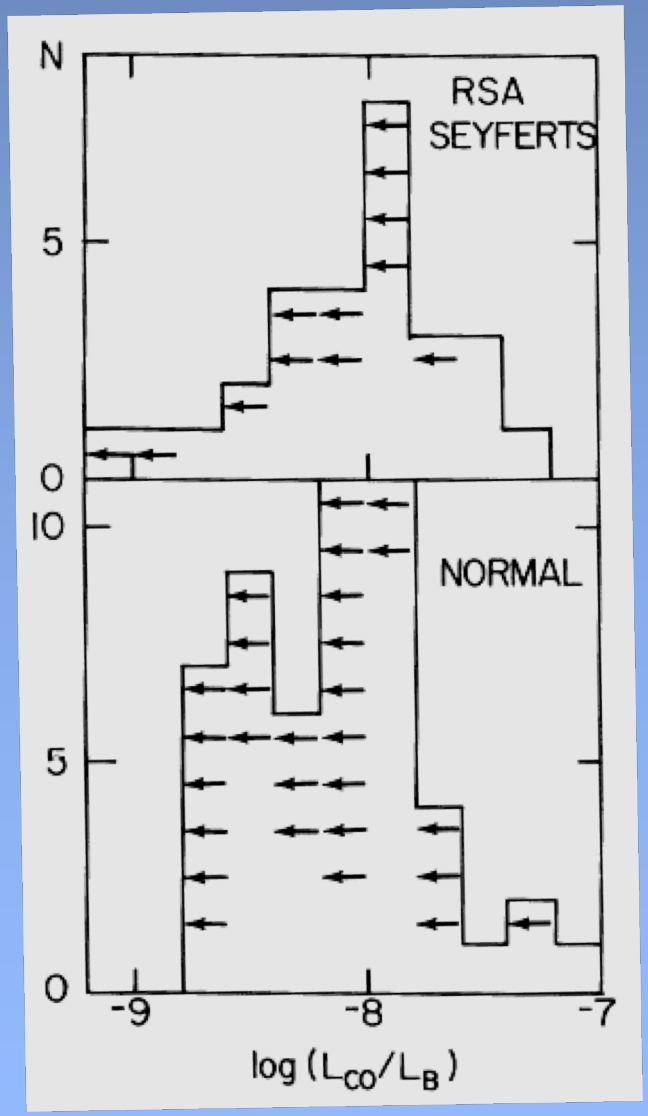
Gehan, logrank, Peto-Peto, Peto-Prentice tests

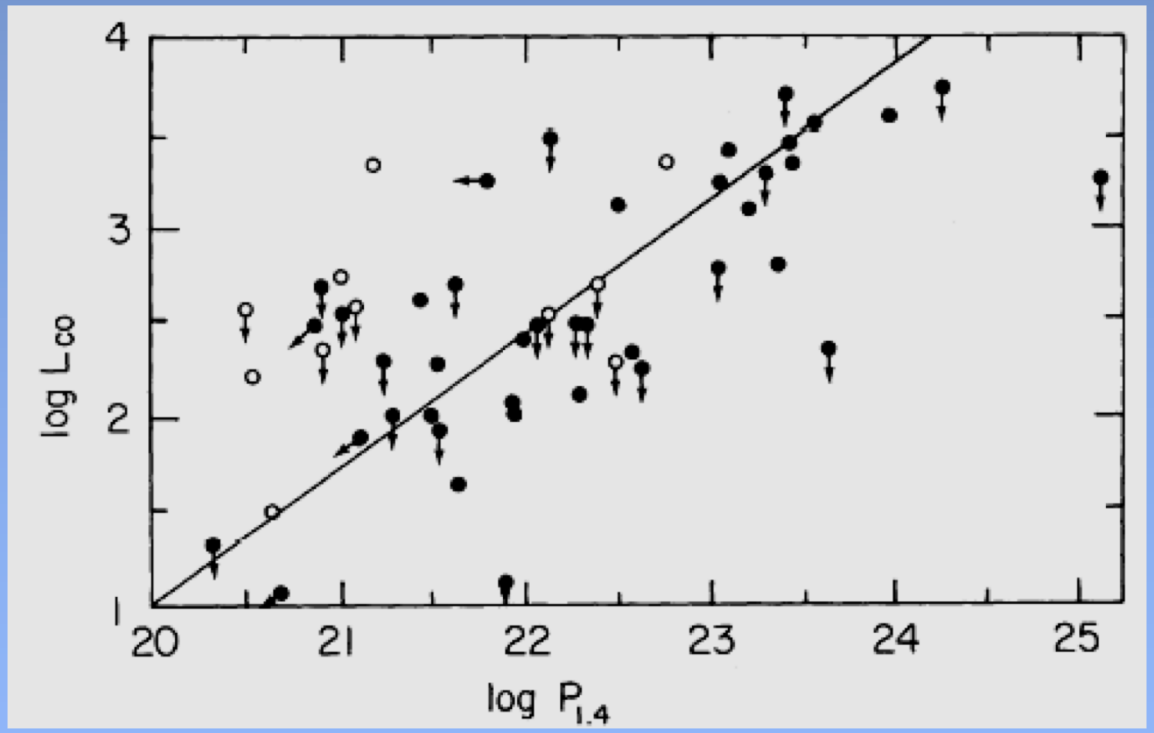
Generalized Kendall's tau correlation coefficients developed in 1970-90s.

Linear regressions developed during 1970-80s:

- Maximum likelihood line assuming Gaussian residuals (using EM Algorithm)
- **Buckley-James line** assuming nonparametric KM residuals
- **Cox regression** for multivariate independent variables
- Akritas-Thiel-Sen semi-parametric line

An example of astronomical censored data





Heckman et al. 1989 A millimeter-wave survey of CO emission in Seyfert galaxies

Gehan's survival 2-sample hypothesis test

What is the chance that two censored datasets do not arise from the same underlying distribution? $H_0 : S_1(x) = S_2(x)$

The Gehan test, developed in 1965 as a generalized Wilcoxon test for survival data, is perhaps the simplest of the censored two-sample tests. For a left-censored Sample 1 with n objects, $x_1^1, x_2^1, \ldots, x_n^1$, and Sample 2 with m objects, $x_1^2, x_2^2, \ldots, x_m^2$, compute the pairwise quantity

$$\begin{aligned} & = +1 \text{ if } x_i^1 < x_j^2 \text{ (where } \mathrm{x_i^1} \text{ may be censored)} \\ U_{ij} & = -1 \text{ if } x_i^1 > x_j^2 \text{ (where } \mathrm{x_j^2} \text{ may be censored)} \\ & = 0 \text{ if } x_i^1 = x_j^2 \text{ or if the relationship is ill} - \text{determined.} \end{aligned} \tag{10.20}$$

Gehan's test statistic is

$$W_{Gehan} = \sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij}. \tag{10.21}$$

W_{Gehan} is asymptotically normal with zero mean. A common estimate of its variance of based on permutations is

$$\widehat{Var}(W_{Gehan}) = \frac{mn \sum_{i=1}^{n+m} U_i^2}{(n+m)(n+m-1)} \tag{10.22}$$

Bivariate correlation for censored data

Consider a nonparametric hypothesis test for correlation. Helsel proposes a generalizing Kendall's coefficient based on pairwise comparison of data points, (x_i, y_i)-(x_j, y_j).

$$\tau_H = \frac{n_c - n_d}{\sqrt{\left(\frac{n(n-1)}{2} - n_{t,x}\right)\left(\frac{n(n-1)}{2} - n_{t,y}\right)}}$$

where n_c is the number of pairs with a positive slope in the (x, y) diagram, n_d is the number of pairs with negative slopes, $n_{t;x}$ and $n_{t;y}$ are the number of ties or indeterminate relationships in x and y respectively.

As the censoring fraction increases, fewer points contribute to the numerator of τ_H, but the denominator measuring the number of effective pairs in the sample also decreases. So τ_H depends on the detailed locations of the censored points.

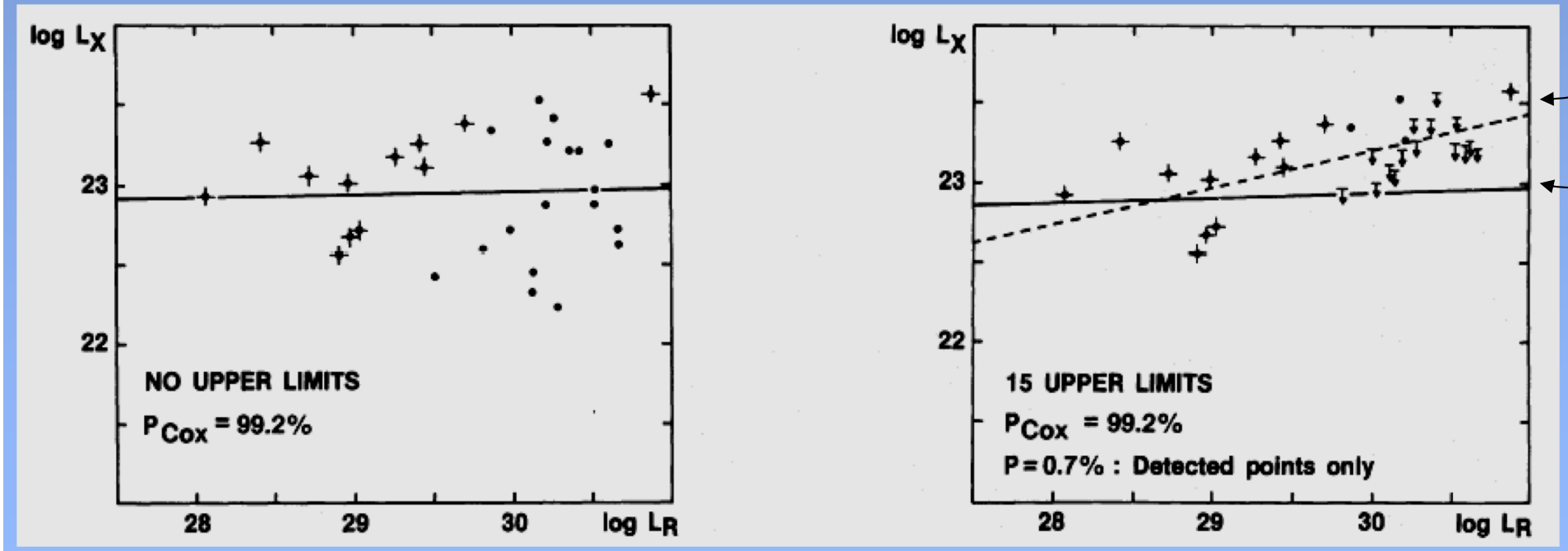
Helsel, D. R. Nondetects and Data Analysis: Statistics for Censored Environmental Data, 2005

Linear regression: Several approaches

1. Iterative least squares, the familiar linear regression model $y = \alpha + \beta \mathbf{x} + \epsilon$ where $\epsilon = N(0, \sigma^2)$. (Industrial reliability)
2. Accelerated failure-time model $\log y = \alpha + \beta \mathbf{x} + \epsilon$. (Industrial reliability)
3. Tobit regression. (Econometrics)
4. Proportional hazards model (Cox regression) where the hazard rate has an exponential dependence on the covariates, $h(y|\mathbf{x}) = h_0(y)e^{\beta \mathbf{x}}$. MLE estimation and inference. (Biometrics)
5. Buckley-James line permiting non-Gaussian residuals around the line. ϵ is estimated in local regions of x using the Kaplan-Meier estimator. (Biometrics)
6. Akritas-Thiel-Sen line for doubly censored data shown in figure above. (Astronomy)

Test of bivariate correlation/regression for simulated flux-limited surveys

Here is a simulation of uncorrelated X-ray and radio luminosities of a hypothetical sample of galaxies. The left panel has infinite sensitivity in the X-ray band, while the right panel shows finite sensitivity giving upper limits. A spurious correlation is found if only detections are considered (dashed line), but no correlation is found if survival methods (e.g., Cox regression) are used.



Biased line using detections only

Unbiased Buckley-James line including nondetections

Isobe, Feigelson & Nelson 1986

Parametric modeling of censored data

Likelihoods can be constructed for censored and truncated samples:

$$L \propto \Pi_{det} f(x_i)\ \Pi_{cens}(1 - S(x_i))\ \Pi_{trunc} f(x_i)/S(x_{trunc})$$

In some cases, the likelihood can be written in closed form.

With the likelihood, the full capabilities of MLE and Bayesian inference are available: parameter estimation with confidence intervals; model selection with penalized likelihoods; marginalization over uninteresting variables; etc.

This approach is not often used in astronomy.

Software for survival calculations

Our group produced a stand-alone Fortran 77 code, Astronomy SURVival analysis (ASURV), which has been widely used. Most of its functionalities are in R/CRAN, and we recommend future work be conducted within R. ASURV implements:

Univariate distribution function Kaplan-Meier estimator with confidence limits & quantiles. Assumes random censoring.

Univariate two-sample tests Gehan, logrank, and Peto-Prentice tests. Can treat unusual censoring patterns.

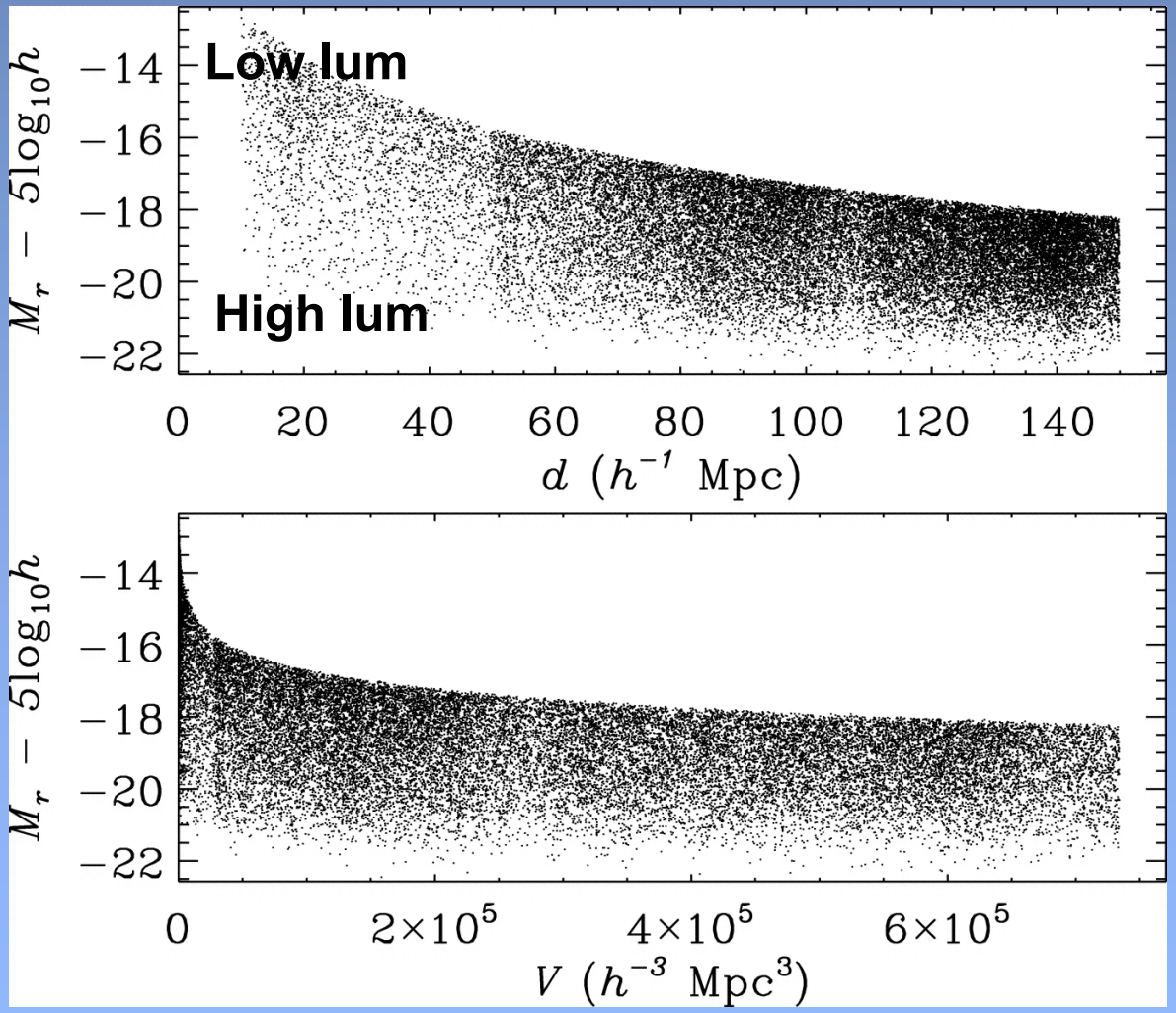
Bivariate correlation coefficient Generalized Kendall's τ. Can treat censoring in both variables.

Bivariate linear regression MLE assuming normal residuals (EM Algorithm), Buckley-James line treating non-normal residuals, Schmitt's binned regression.

But we now encourage use of R/CRAN rather than the old ASURV !!

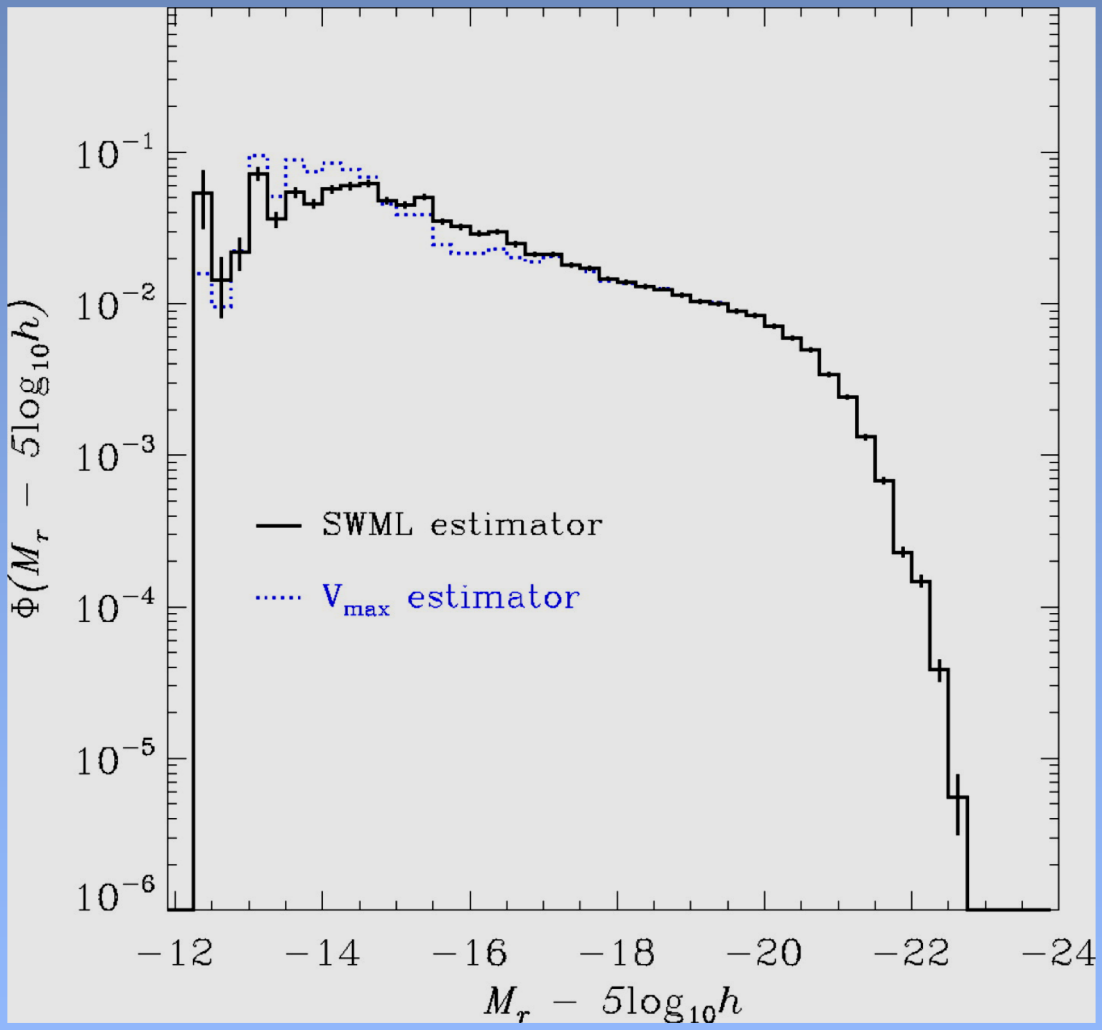
Truncation in astronomical surveys: The case of the galaxy luminosity function

Volume-limited survey of 28K SDSS galaxies with d<150 Mpc



Blanton et al. 2005 The Properties and Luminosity Function of Extremely Low Luminosity Galaxies

Normal galaxy luminosity function



Parametric model:
Schechter function
related to gamma distribution

$$dN/dL \sim \Phi \sim L^{-\alpha} e^{-L/L^*}$$

Note that the Schechter function is just the gamma distribution introduced by Pearson (1901) based on the gamma function of Euler & Lagrange.

Nonparametric methods:
two used here ($1/V_{max}$,
stepwise max. likelihood)

(Many technical issues concerning correction for missing low-surface brightness galaxies, K-correction and evolution-corrections to luminosity, double Schechter fits, etc.)

Nonparametric estimators to LFs

1. Classic estimator:

$$\Phi(L) = N(L) / V$$

where N is the number of stars/galaxies/AGNs in surveyed volume V. A biased estimator.

2. Schmidt estimator (~40 citations/yr):

$$\Phi(L) = \Sigma\ 1 / V_{max}(L_i)$$

where V_{max} is the maximum volume within which an object of the observed flux could have been seen given the survey's sensitivity limit. Unbiased estimator but with high variance. This is typically calculated in arbitrary luminosity bins.

Schmidt 1968 Space Distribution and Luminosity Functions of Quasi-Stellar Radio Sources
Felten 1976 On Schmidt's V_m estimator and other estimators of luminosity function

Stepwise maximum-likelihood estimator (~10 citations/yr):

$$\ln \mathscr{L} = \sum_{i=1}^{N_{\rm obs}} \left\{ \sum_{\ell=1}^{K} W(M_\ell - M_i) \ln \phi_\ell \, . \right.$$
$$\left. - \ln \left[\sum_{\ell=1}^{K} \phi_\ell \, H(M_{\rm lim}(z_i) - M_\ell) \Delta M \right] \right\} . \quad (11)$$

where $$W(M_\ell - M) \equiv \begin{cases} 1 & \text{for } M_\ell - \dfrac{\Delta M}{2} \le M \le M_\ell + \dfrac{\Delta M}{2} , \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

$$H(M_{\rm lim}(z_i) - M) \equiv \begin{cases} 1 & M_{\rm lim}(z_i) - \dfrac{\Delta M}{2} > M , \\ \dfrac{M_{\rm lim}(z_i) - M}{\Delta M} + \dfrac{1}{2} & M_{\rm lim}(z_i) - \dfrac{\Delta M}{2} \le M < M_{\rm lim}(z_i) + \dfrac{\Delta M}{2} , \\ 0 & M_{\rm lim}(z_i) + \dfrac{\Delta M}{2} \le M , \end{cases} \quad (10)$$

Set dL/dϕ = 0 to obtain

$$\phi_k \Delta M = \sum_{i=1}^{N_{\rm obs}} W(M_k - M_i)$$
$$\times \left[\sum_{i=1}^{N_{\rm obs}} \frac{H(M_{\rm lim}(z_i) - M_k)}{\sum_{\ell=1}^{K} \phi_\ell \, H(M_{\rm lim}(z_i) - M_\ell) \Delta M} \right]^{-1} . \quad (13)$$

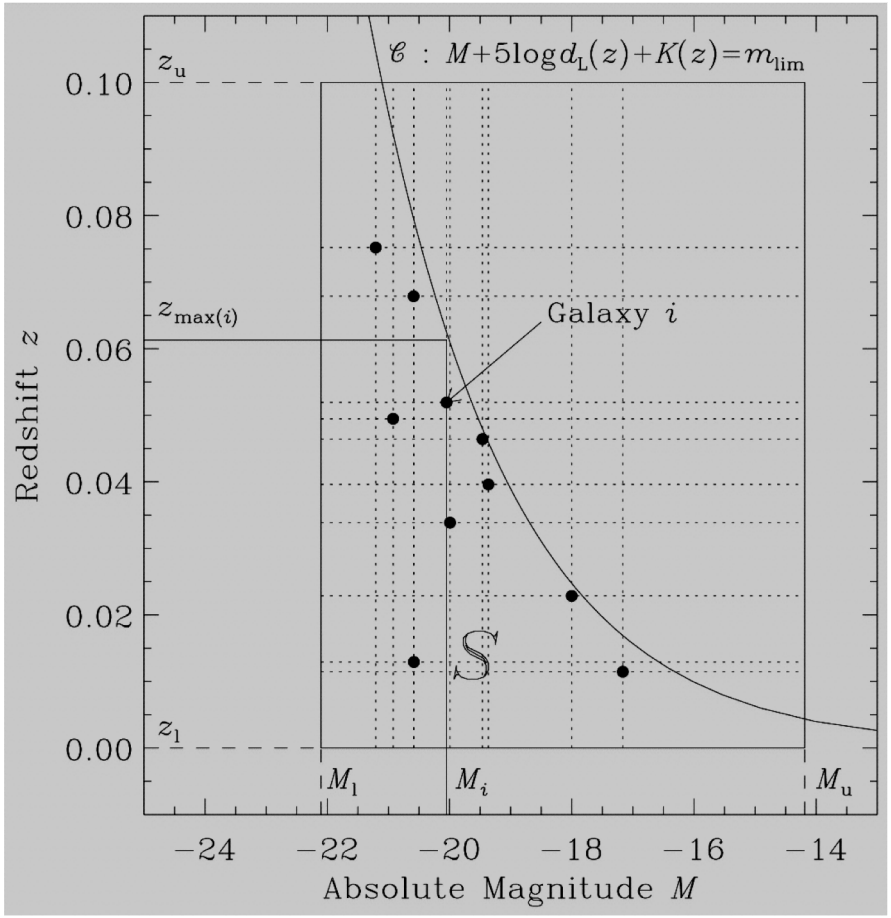
Efstathiou et al. 1988 Analysis of a complete galaxy redshift survey. II - The field-galaxy luminosity function

Lynden-Bell-Woodroofe MLE (~6 citations/yr):

Recursion relation:

$$\Phi(M) \propto \sum_{k=1}^{M_k<M} \psi_k = \psi_1 \prod_{k=1}^{M_k<M} \frac{C_k + 1}{C_k} .$$

where C_k is the number of stars/galaxies/AGNs in the k-th rectangle in the luminosity-distance diagram



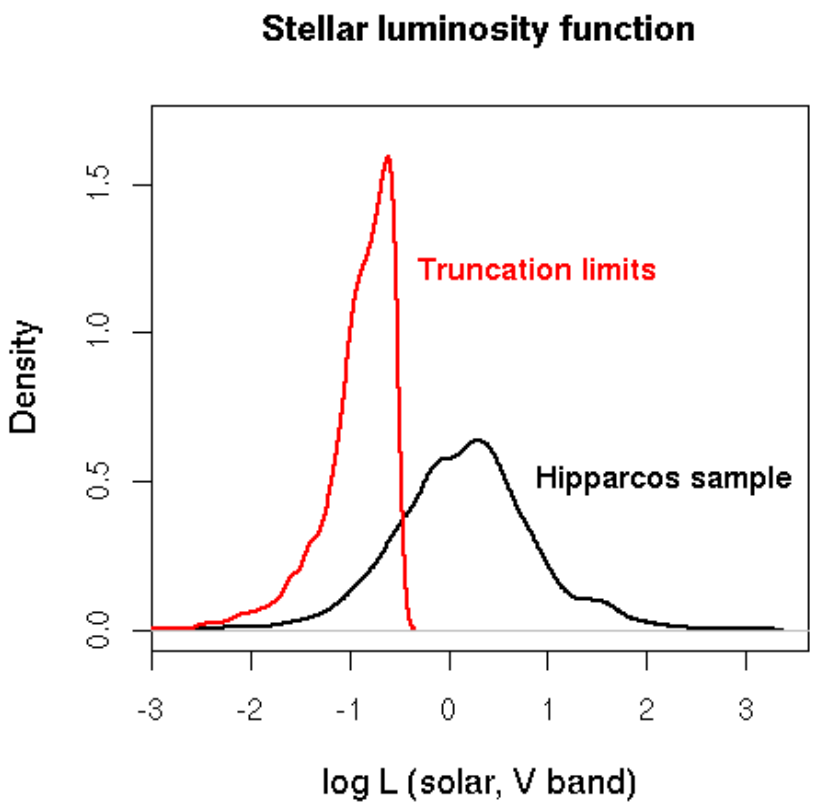
Lynden-Bell, A method of allowing for known observational selection in small samples applied to 3CR quasars, MNRAS 1971
Woodroofe, Estimation of a distribution function with truncated data, Annals Stat 1985

Takeuchi et al. (2000) apply these luminosity function estimators to simulations of small galaxy catalogs (N=100 and 1000). All perform well when spatial distribution is homogeneous. But for spatially clustered distributions, the $1/V_{max}$ estimator is badly affected.

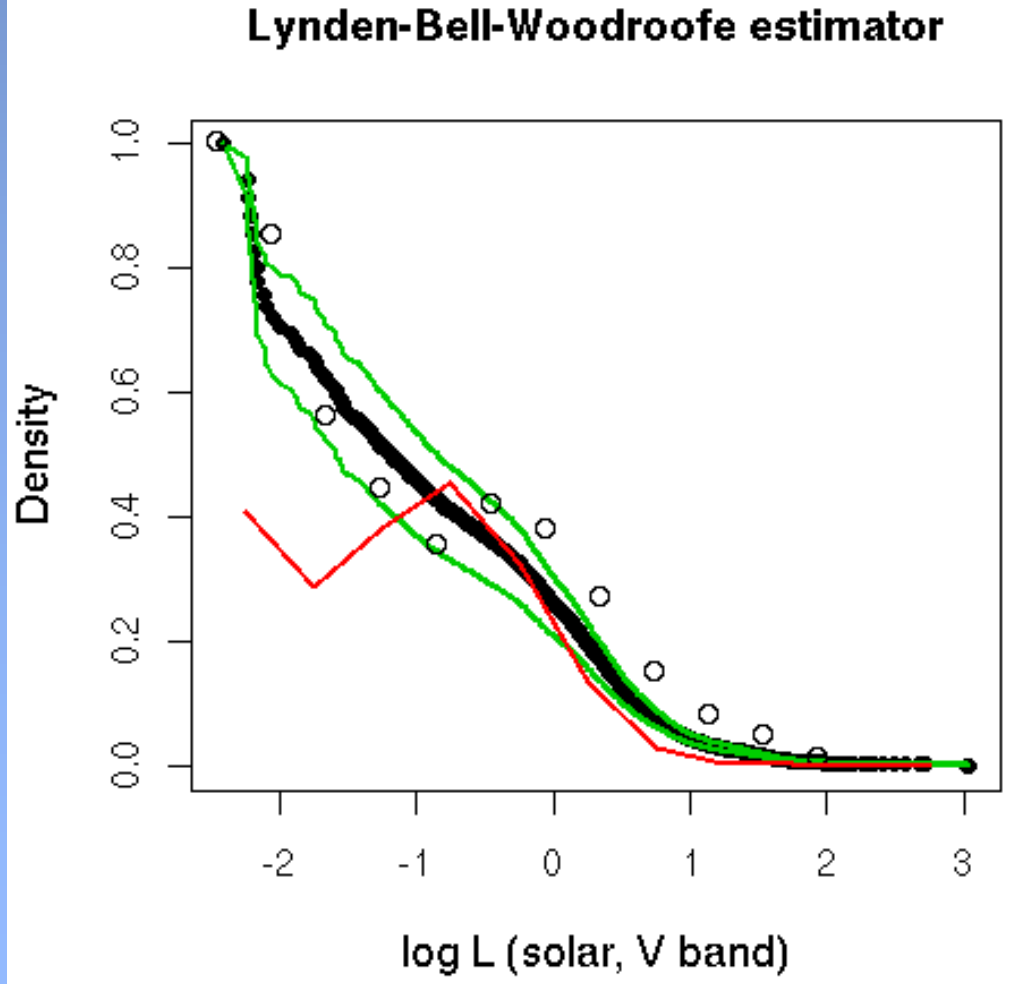
There has been no analytical evaluation of the mathematical properties of these estimators since study of $1/V_{max}$ by Felten (1976).

The Lynden-Bell-Woodroofe estimator for randomly truncated univariate data is mathematically the best choice: unbiased, unbinned, nonparametric maximum likelihood, asymptotically normal.

It is the analog of the Kaplan-Meier estimator for randomly censored univariate data.



Data: (L, L_{min}) for 3,307 Hipparcos stars with V<10.5 and parallactic distance < 75 pc



Open circles: Known stellar LF from volume-limited sample

Black circles with green confidence bands: Lynden-Bell-Woodroofe estimator

Red histogram: Schmidt's $1/V_{max}$ estimator

computed with CRAN package DTDA

References

Lee, E. T. & Wang, J. W. (2013) Statistical Methods for Survival Data Analysis, 4th ed., Wiley

Helsel, D. R. (2012) Statistics for Censored Environmental Data Using Minitab and R, 2nd ed., Wiley

Moore, D. F. (2016) Applied Survival Analysis Using R, Springer

Feigelson, E. D. & Nelson, P. I. (1985) Statistical methods for astronomical data with upper limits. I. Univariate distributions, Astrophys. J. 293, 192-206