Malmquist bias and Eddington bias Luminosity functions

Maurice G. Kendall
1907-1983

of Kendall and Stuart, "Advanced Theory of Statistics"; Kendall published the first two volumes alone, in 1943 and 1946.

Yet another Fellow of St Johns, Cambridge, along with Yule, whose textbook was the basic stats book for the first half of the 1900's. In 1937, Yule's book became Yule and Kendall, and it ran to 14 editions by 1950. Chair of Statistics, LSE, 1949-1961. Many papers on time series, rank order. Resigned to become Chairman of SciCon, an early computer company. 'Retired' at age 65, and asked by the UN to take on Directorship of the World Fertility Survey. Knighted for 'services to statistics' in 1974; awarded UN Peace Medal in 1980.

Way back last Tuesday?

We considered detection - it is a data-modelling process

- as a working definition we adopted location, and confident measurement, of some sort of feature in a fixed region of an image or spectrum.

- non-detections are also important; we can incorporate the information these contain with statistical measures. Two types of these – upper limits, and the confusion continuum.

- clear model required; we can't make it up as we go along.

- we considered concepts of completeness and reliability; these are mutually exclusive at some point, but satisfactory compromise can be reached.

- we began with classical detection via likelihood, but as conditional probabilities are involved at the start, why not go Bayesian? A couple of Bayesian examples....

- we went on to introduce luminosity distributions and luminosity functions.....

Detection summary, again (1)

1. The **simple idea of a detection**, making a measurement of something that is really there, **applies when signal-to-noise is high** and individual objects can be isolated from the general signal. At low s/n, measurements can constrain population properties, with the notion of “detection" disappearing, in two senses: we drop into the **confusion level**, and/or we deal with **censored data**.

2. **Detection is a modelling process**:
   - it depends on what we are looking for,
   - how the answer is expressed depends on what we want to do with it next.

3. There is competition between **completeness** and **reliability,** even in the classical sense; Bayesian analysis can help to sort this out.

Detection summary, again (2)

4. **In fact, Bayesian treatment of detection gives a direct result**; we may read off a suitable flux limit that will give the desired probability of detection.

5. At the 'limits' of detection – the **confusion, and confusion limit** - images or spectral lines **crowd together** **and** **overlap** as we reach fainter. Even if only one 'object' is present, with a steep ***N(s)*** it will be more likely that the flux results from a **faint source plus a large upward noise excursion**, rather than vice-versa. This effect plays havoc with N(S), the source count, and in the limit it may mean that NO source plus an upward noise deflection is more likely than a source detection. The effect that bad flux estimates have on N(S) is known as Eddington bias (Eddington 1911).

6. **Eddington bias** makes almost as much of a mess of surveys as the better-known **Malmquist bias,** which I now discuss.

Catalogues, Selection Effects (1)

Typically, a body of astronomical detections is published in a **catalogue**.
Objects are in this catalogue on the basis of **some clear criteria**.

Most astronomical measurements are **affected by the distance to the object**.
e.g. **proper motion**, **apparent intensity**, **ellipticity**.

We measure an **apparent** quantity ***X*** and infer an **intrinsic** quantity by a relationship ***Y=f(X,R)*** where ***R*** is the **distance** to the object in question. The function ***f*** may be complicated, for reasons of both observation and relativistic geometry.

E.g. observe a flux density S and infer a luminosity l given by $l = S R^2$. The smallest value of S we are prepared to believe is S_{lim}; if $S < S_{lim}$ the object is not in our catalogue or sample.

Catalogues, Selection Effects (1)

Our objects (="galaxies") are assumed to be drawn from a **luminosity function *ρ(l)***, **the number of objects within *Δl* about l per unit volume**. Using only our catalogue set of measurements l_1, l_2, … however, we will **not** be able to reproduce ***ρ***. Instead, we will get the **luminosity distribution *η***, where

$$\eta(l) \propto \rho(l) V(l).$$

V(l) is the volume within which sources of intrinsic brightness l will be near enough to find their way into our catalogue. We get

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

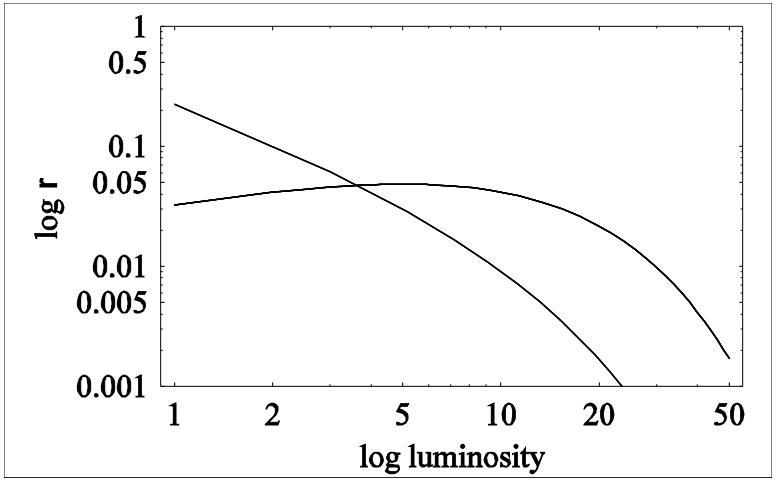
Obviously ***η*** will be biased to higher values of luminosity than ***ρ***. This sort of bias occurs in a multitude of cases in astronomy, and is called **Malmquist bias**.

Malmquist Bias + Example

Example 1: The luminosity function of field galaxies is well approximated by the Schechter function

$$\rho(l) \propto \left(\frac{l}{l_*}\right)^{\gamma} \exp(-\frac{l}{l_*}),$$

in which we take ***γ* = *1*** and $l_* = 10$ for illustration. To obtain the form of the luminosity distribution in a flux-limited survey, we multiply the Schechter function by $l^{3/2}$. The differences between the luminosity function and luminosity distribution are shown.



The luminosity function ρ (steep curve) and the (flat-space) luminosity distribution, plotted for the Schechter form of the luminosity function.

Malmquist bias – a serious issue in survey astronomy. The bias depends on the (probably unknown) **form of the luminosity function**. It passes on all kinds of unwelcome information.

Malmquist Bias and the Luminosity-Distance Correlation

Malmquist bias: intrinsically luminous objects can be seen within proportionately much greater volumes than small ones. Thus in flux-limited samples the luminous objects of the sample will tend to be further away than the faint ones - **there is an in-built distance-luminosity correlation.**

The **luminosity - distance correlation is widespread, insidious and very difficult to unravel**. It means that **for flux-limited samples, intrinsic properties correlate with distance**.

Two unrelated intrinsic properties will appear to correlate because of their mutual correlation with distance. Plotting intrinsic properties -- say, X-ray and radio luminosity -- against each other will be very misleading.

Malmquist Bias and the Luminosity-Distance Correlation

Example 2: Consider measuring the ellipticity of galaxies with direct optical images from the ground.

Because they appear fainter (smaller disks), distant galaxies will look rounder, due to the seeing - the smearing or enlarging of the optical image through atmospheric turbulence.

We are on course for deducing that round galaxies are more luminous (or v-v). There is the possibility of course that this is true. It is far more likely that we have fallen into a Malmqvist trap and made a totally erroneous deduction.

The Luminosity-Distance Correlation - Example 3

Let us adopt a Schechter function with γ **= *1*** and l_* **= *10*** for illustration.

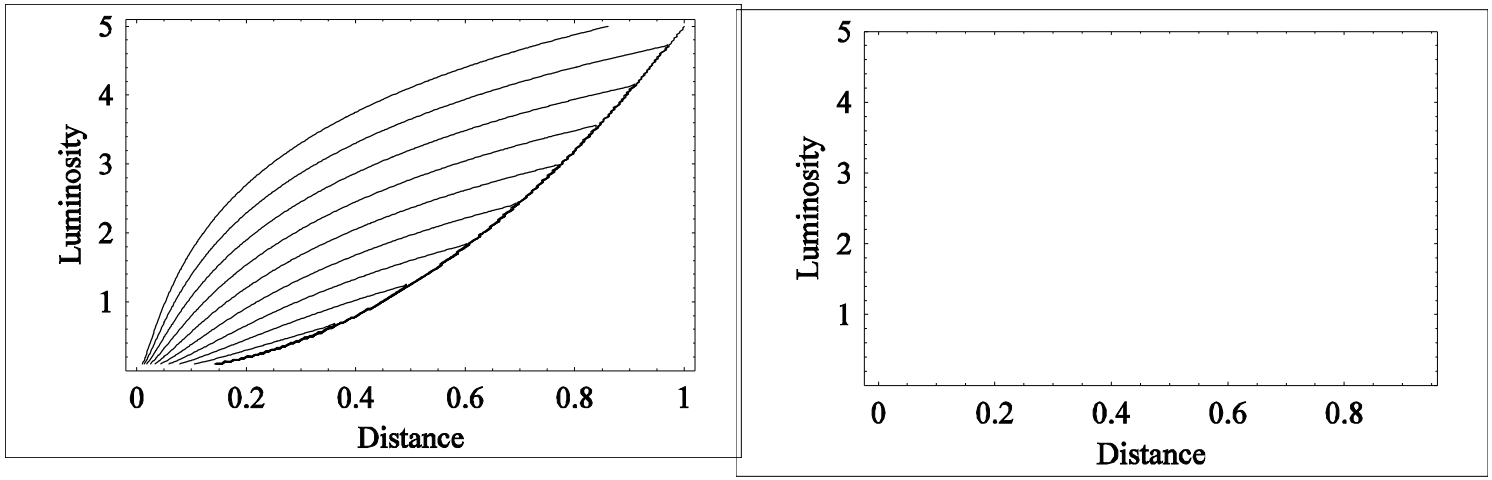
The **probability** of a galaxy being at distance ***R*** is proportional to R^2, in flat space.
The **probability** of it being of brightness l is proportional to the **Schechter function**.
The **probability** of of a galaxy of luminosity l at distance ***R*** being in our sample is

prob (in sample) = 1 $l < S_{lim} R^2$
= 0 ***otherwise***

The product is what we want: **the bivariate distribution *prob(l,r)***, the probability of a galaxy of brightness l and distance ***r*** being in our sample. What's this look like?

The Luminosity-Distance Correlation - Example 3

The bivariate distribution *prob(l,r)*, the probability of a galaxy of brightness l and distance ***r*** being in our sample looks unfortunately like this.



Left: Contour plots of the bivariate prob(l,r). The contours are at logarithmic intervals; galaxies tend to bunch up against the selection line, leading to a bogus correlation between luminosity and distance. Right: Results of a simulation of a flux-limited survey of galaxies drawn from a Schechter function.

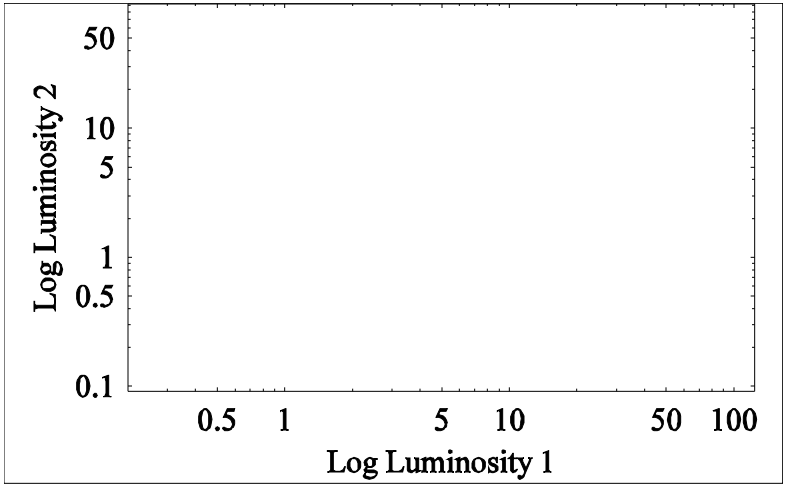
There is a clear correlation between distance and luminosity. It is NOT real.

Luminosity-Distance: Example 4 of Forced Correlation

Take the same simulation as before, but now

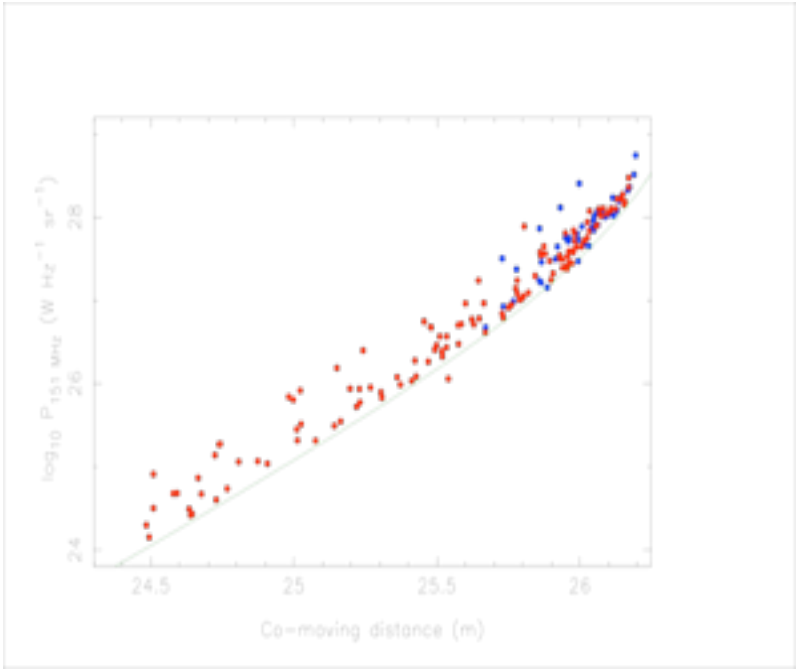
attribute two luminosities to each galaxy, from different Schechter functions.

These might be luminosities in different colour bands, for example, and by definition are statistically independent. If we construct a flux-limited survey in which a galaxy enters the final sample **only if it falls above the flux limit in both bands**, we see in the Fig. that a **bogus but convincing correlation** emerges between the two luminosities.



Results of a simulation of a flux-limited survey of galaxies, in which each galaxy has two statistically independent luminosities associated with it.

Examples 3 and 4 of Malmquist 'Correlations'



Allan Sandages's radio-galaxy - quasar correlation from the Third Cambridge Radio (3CR)Catalogue. Note the survey limit, the faint green line.

Note that baseballs and bricks would probably fit on both these lines as well

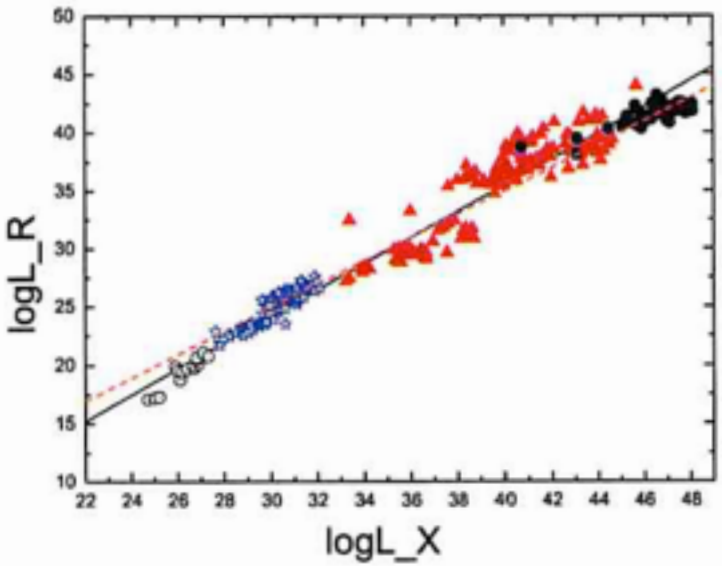


Fig. 4. We give the distribution of the 15 solar flares (black dot circles), 65 cool stars (blue stars), 149 AGNs (red triangles) and 50 GRBs (black solid circles) objects in the $L_R \sim L_X$ plane. The black solid line, $L_R \cong 2.82\times10^{-10} L_X^{1.13}$, is the best fit, and compared to the red dashed line, $L_R = 10^{-5} L_X$. The data of solar and cool stars are taken from Laor & Behar 2008, who tabulated the data of Benz & Güdel 1994. The radio data and the X-ray data of GRBs afterglow are taken from Chandra & Frail 2012.

Conclusion: After extending the samples from stars, AGNs, GBHs to GRBs, we find a uniform correlation between L_R and L_X in these different scale astrophysical objects, which implies a common physical origin behind them.

Umm - total garbage conclusion!

Eddington Bias

A little paper in 1913 MNRAS by Eddington: On a formula for correcting statistics for the effects of a known error of observation

On a Formula for Correcting Statistics for the Effects of a known Probable Error of Observation. By A. S. Eddington, M.A., M.Sc.

In astronomical investigations it often happens that we construct a table to exhibit the number of stars having successive values of a certain measured property, *e.g.* the number of stars of successive magnitudes, or the number between successive limits of proper motion. In general the observations are subject to a probable error, which is at least approximately known. The error is in general small, otherwise our table would not be of much value, but it must have some effect on the numbers of the table. It is not, however, usual to take account of the probable error, probably because it is not generally realised that it can be eliminated and an improved table formed in a very simple way.

To avoid the awkwardness of general terms, suppose we are concerned with counts of stars between given limits of magnitude, and, knowing the average accidental error of our determinations of magnitude, we wish to apply corrections to our counts to eliminate these errors.

Let $u(m)dm$ = observed number of stars between magnitudes m and $m+dm$
$v(m)dm$ = true number.

Let the probable error of the observed magnitudes be $0{\cdot}477/h$, so that the frequency of an error x is proportional to $e^{-h^2x^2}$.

We have $$u(m) = \frac{h}{\sqrt{\pi}}\int_{-\infty}^{\infty} v(m+x)e^{-h^2x^2}dx;$$

for, of the stars having a true magnitude $m+x$, the proportion $\frac{h}{\sqrt{\pi}}e^{-h^2x^2}$ will have an error of measurement $-x$, and will therefore be observed as of magnitude m.

By the symbolic form of Taylor's theorem

$$v(m+x) = e^{x\frac{d}{dm}}\cdot v(m),$$

therefore $$u(m) = \frac{h}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{x\frac{d}{dm}-h^2x^2}\cdot v(m)dx.$$

The integral is of the well-known form

$$\int_{-\infty}^{\infty} e^{-a_2x-a_1x^2}dx = \sqrt{\frac{\pi}{a_1}}\exp\frac{a_2^2}{4a_1},$$

thus integrating

$$u(m) = \exp\left(\frac{1}{4h^2}\frac{d^2}{dm^2}\right)\cdot v(m);$$

therefore $$v(m) = \exp-\left(\frac{1}{4h^2}\frac{d^2}{dm^2}\right)\cdot u(m)$$
$$= u(m) - \frac{1}{4h^2}u''(m) + \frac{1}{2!}\left(\frac{1}{4h^2}\right)^2 u^{iv}(m) - \ldots$$

The result so far is accurate.

For a small accidental error, confining ourselves to the first two terms, the result can be given in a form convenient for computation as follows:—

The tabular second difference for intervals a
$$= u(m+a) + u(m-a) - 2u(m) = a^2u''(m) \text{ approximately.}$$

Also $$\frac{1}{2h} = 1{\cdot}046 \times \text{probable error.}$$

Thus the approximate correction is
$$-\left(\frac{1{\cdot}046 \times \text{probable error}}{\text{tabular interval}}\right)^2 \times \text{tabular second difference.}$$

Note on the Convergency of the Series.—The Astronomer Royal has pointed out to me that the series in some typical cases is divergent, *e.g.* $u(m) = (1+m^2)^{-1}$. The operator $\frac{d^{2n}}{dm^{2n}}$ introduces a factor of the order $2n!$, whilst the divisor is only $n!$. Apparently, however, in these cases the expansion is asymptotic; and it seems certain that the first few terms give the approximate correction quite correctly. The divergence arises from using the Taylor expansion beyond its range of convergence.

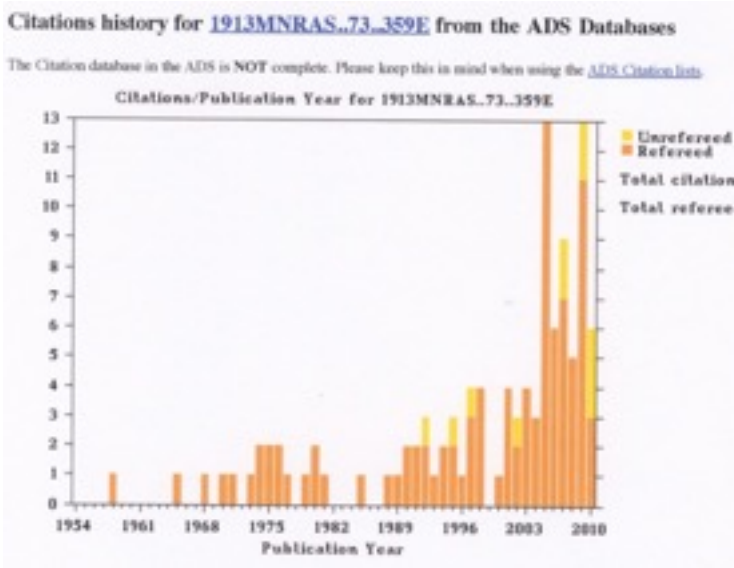
The difficulty does not really arise in practice. In a table with a finite number of entries, we are actually dealing with a polynomial, in which case all the series terminate, and no question of divergence arises. Thus, if a is the tabular interval, the tabular values of $v(m)$ may be represented from $m-na$ to $m+na$ by a polynomial of the $2n^{th}$ degree, say $v_{2n}(m)$. By taking n sufficiently great the whole range of v, which contributes appreciably to $u(m)$, can evidently be included. Beyond the limits $m \pm na$, the divergence between $v_{2n}(m)$ and $v(m)$ will generally increase rapidly; it can be shown, however, that

$$\int e^{2n\frac{d}{dm}-h^2x^2}dx$$

tends to zero as n is increased, and hence the part of $v_{2n}(m)$ beyond the limits $m \pm na$ will (when n is great) make no contribution to $u(m)$.

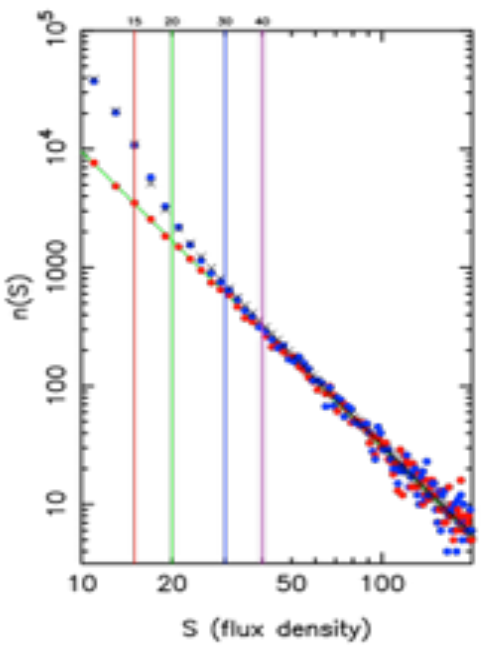
Thus the polynomial and the true function agree for the $2n+1$ tabular entries $m-na$ to $m+na$, and beyond these limits they both make contributions to $u(m)$, which tend to zero. Hence the polynomial can be used for our purpose, and a terminating series results.

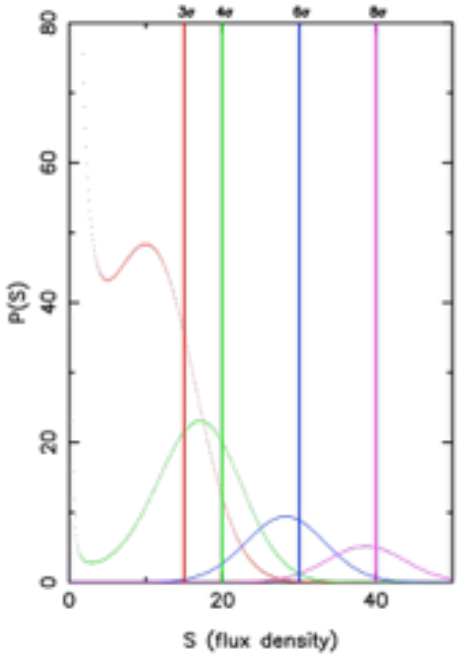
Substituting the polynomial only requires that we should deduce the differential coefficients from the tabular differences—a procedure which we should naturally adopt in any case.



Eddington bias rediscovered by Submm Galaxy researchers

Eddington Bias: a Simulation Example





Left: Red points show a source count (number of sources on the `sky' per interval of 2 units in flux) from a toy Euclidean universe. The green line is the theoretical power law of slope -5/2. The blue points represent the distorted source count resulting from Eddington bias assuming a Gaussian flux measurement error of σ = 5 units. Black crosses represent the analytical calculation $N'(S)=\int N(S)p(S)dS$.

Right: plots of the integrand N(S)p(S) for apparent flux densities of 15, 20, 30 and 40 units (3σ, 4σ, 6σ, 8σ) given an underlying law of slope -5/2 and Gaussian errors of σ= 5 units.

3σ is hopeless; 4σ is still seriously biased

Survey Biases - still not the last warning

(1) Malmquist bias => distance – luminosity correlation, with many insidious effects to do with correlations between unrelated (probably!) variables.

(2) Eddington bias => poorly defined 'complete' samples, serious errors inferring fluxes, intrinsic luminosities and number counts, and hence errors in analysis of space distribution.

Mix these together, as all surveys do =>

Simulations are essential, even after careful thought and analysis.

The Luminosity Function and V_{max}

- assume a catalogue of objects, high reliability and well-defined limits.

- examination of an intrinsic variable l (e.g. luminosity) requires **ρ(l)** .
- measure l_i for all of the objects in some (large and complete) volume.

One of the best methods to determine the space density as a function of luminosity (the **LUMINOSITY FUNCTION**) is the **$1/V_{max}$** estimator.

- **$V_{max}(l_i)$** are the maximum volumes within which the **i^{th}** object could lie, and still be in the catalogue.

- **V_{max}** thus depends on the survey limits, distribution of the objects in space, and the way in which detectability depends on distance.

- simplest case: a uniform distribution in space is assumed. Given the **$V_{max}(l_i)$** , an estimate of the luminosity function is

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

in which its value is computed in bins of luminosity, bounded by the **B_j** .

The Luminosity Function and V_{max}

- a maximum-likelihood estimator, and so has minimum variance for any estimate based on its statistical model.

- errors are uncorrelated from bin to bin and can easily be estimated – the fractional error in each bin is ~ $1/\sqrt{N_j}$, where N_j is the number of objects in each bin. Better error estimates - bootstrap.

- note 'bin bias'. If bins are chosen large/wide enough, there is some error associated with the form of the function across the bin. I.e.if you are plotting the function, where do you place the abscissa in the bin?

- NOT IN THE MIDDLE!

The Luminosity Function and V_{max} continued ....

- determination of V_{max} is the crunch; choosing the flux limit of a survey affects:
  (1) the number of sources that are missed,
  (2) the number of bogus ones included, and
  (3) the extent to which faint sources are over-represented.

=> MC simulations? Rough idea of the luminosity function enables this.

- with **V** the volume defined by the distance to the source as its radius, the distribution of V/V_{max} is very useful in estimating the actual limit of a survey.....

=>

The Luminosity Function and V_{max} continued ....

- if the correct flux limit adopted then we expect V/V_{max} to be uniformly distributed between zero and one. This can be checked by, e.g. a K-S test, and such a process is a model-fitting procedure to estimate survey limit.

- this procedure fails horribly and spectacularly at large ($z > 0.2$) cosmological distances where cosmic evolution dominates.

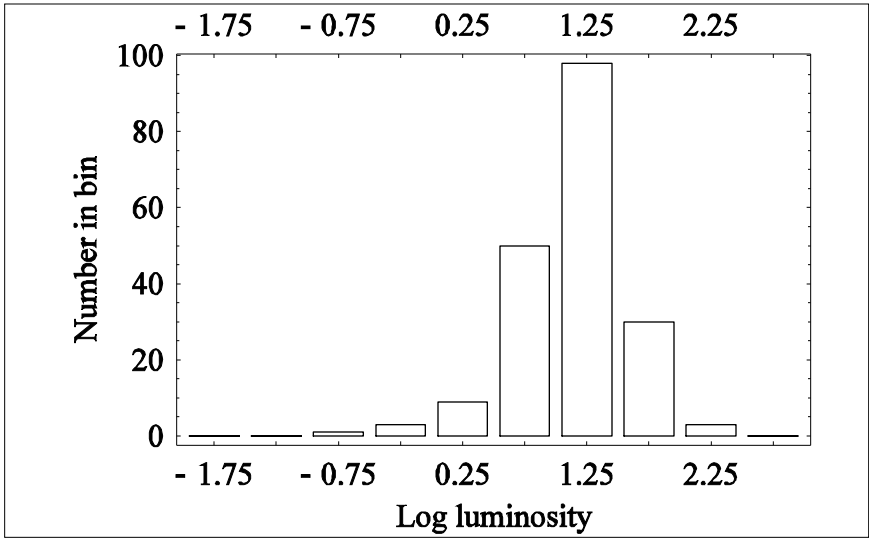
- in fact the derivation of cosmic evolution led Schmidt (1968) to derive the technique.

- there's a vast literature; ~ recent summary: Willmer 1997.

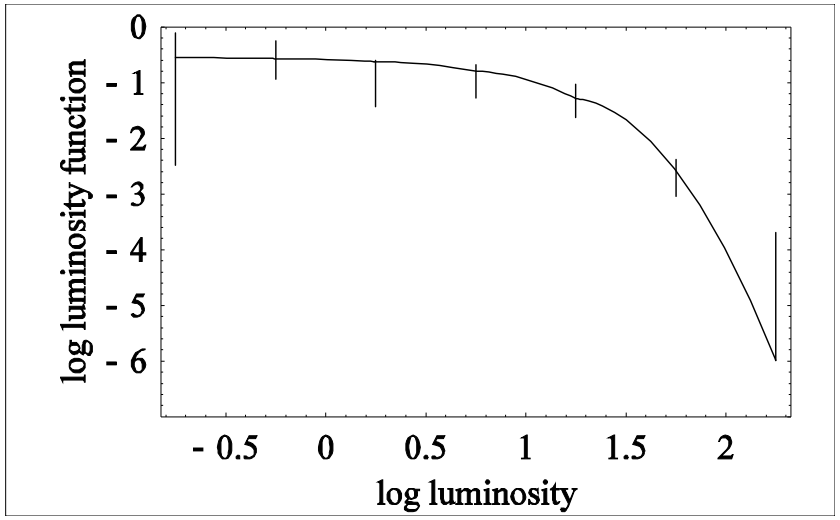
V/V_{max} Luminosity Function - Example

From the previous simulation with a Schechter lum fn, a flux limit of 20 units gives a sample of ~200 objects.

Shown below: the **luminosity distribution** $\eta(\ell)$ shows a strong peak at ~5 units, related to the characteristic luminosity $\ell^* = 10$. Bins are 0.5 dex wide. (Faint sources are greatly under-represented, because they are above the flux limit for only small distances.)



V/V_{max} Luminosity Function - Example concluded



The result of applying the V_{max} method and **bootstrapping** to derive errors. The input lum fn is the solid line and the estimate from the ~200 sample via $1/V_{max}$ is shown as dots and error bars. Because V_{max} is so small for the faint sources, the few faint sources in the sample give a large contribution to $\rho(l)$ although the errors are correspondingly large. For simplicity the luminosity functions were normalized, so giving **luminosity probability distributions**.