

## Surveys - LF likelihood, censorship and confusion

Maarten Schmidt  
1929-



*Maarten Schmidt arrived at [CalTech](#) in 1959, where at first he continued working on the mass distribution and dynamics of the Galaxy. When [Rudolph Minkowski](#) retired, Schmidt took over his project of taking spectra of objects which had been found to be radio emitters. In 1963 he identified the redshift of the first [quasar](#), 3C273, from the redshifted Balmer lines of hydrogen. In 1968 he (and Rowan-Robinson independently) discovered the V/Vmax test, and showed that 3CR radio galaxies evolved strongly; they were not uniformly distributed in space. Past President of the AAS; many scientific awards.*

# In the last lecture ...

We considered **Malmquist bias**, the effect by which distant objects can only be luminous ones in a flux-limited survey - there is a **forced correlation** between distance and intrinsic properties.

Observed luminosities in a flux-limited survey form the **the actual distribution of luminosity**, not a **luminosity function**. This luminosity distribution is **heavily biased to high luminosities**, visible in the large volumes at the distance limits of surveys.

**Malmquist bias** means that any intrinsic properties in flux-limited surveys will appear to be **correlated** - it's widespread, insidious, and hard to remove.

**Eddington bias** is the second serious problem in flux limited surveys - with noise errors and a steep source count, many more faint sources are pushed above the survey limit than bright sources depressed below the limit. The count is distorted to show a false excess of sources at lower flux densities.

To find the **true luminosity function**,  **$1/V_{\max}$  method** is excellent. It requires computation of the maximum volume in which each object can be seen above a given survey limit. However, it fails badly if cosmic evolution is present, ie generally if redshift  $> 0.2$ .

# How to work out combined lum fn + evolution?

In such cases:

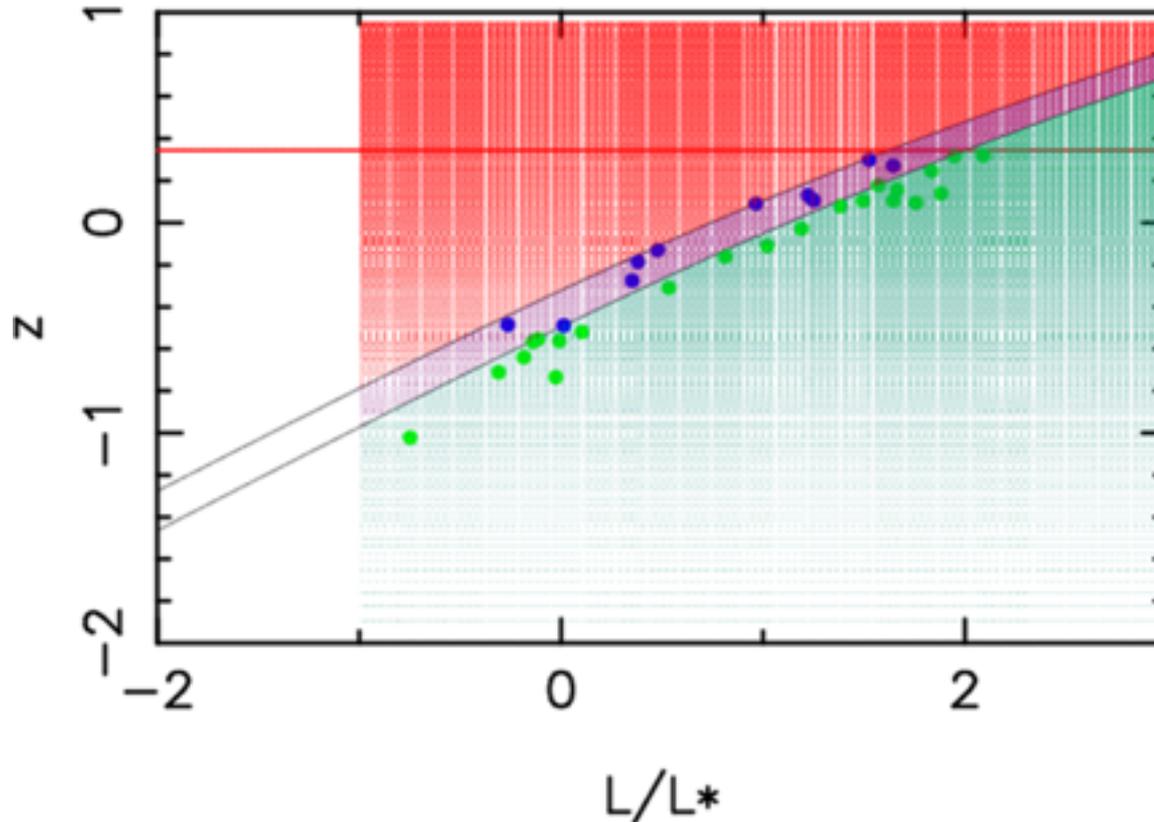
- (1) work out luminosity function by  $1/V_{\max}$  in **shells** of small  $\Delta z$ .
- (2) assume a factorized lum fn  $\rho(l) = \rho(l)(z=0) \cdot \varphi(z, l, \dots)$  and adopt **C- method**, **lumimnosity-distance method**, or any one of 100 more.
- (3) try **free-form evolution**, fitting all the data known on complete samples ( $N(z)$ ,  $N(S)$ , luminosity distributions etc) with polynomial surfaces in  $L$  and  $z$ .

There's a particularly elegant method of doing this using a Bayesian approach that I describe briefly and in a somewhat concentrated way:

# Likelihood and Space Density

## Marshall et al 1983 ApJ 269, 35 (X-ray sample of QSOs)

- 32 X-ray QSOs, two complete samples, 10 Q to  $B=19.20$  mag in  $1.72''$ ,  
22 to  $B=18.25$  in  $37.2''$
- redshift limit  $z = 2.2$ ; incomplete beyond this.
- $L^* = 10^{30}$  ergs  $s^{-1}$   $hz^{-1}$  (optical)
- $L > 0.18 L^*$  to avoid Seyfert galaxy contamination.



# Likelihood and Space Density: Example

## Marshall et al 1983 ApJ 269, 35

We initially follow Marshall (1983); but part of the derivation there is written backwards.

Consider the sample as a single homogeneous set of  $i$  objects, for which  $\rho(z, L)(\partial V/\partial z)dzdL$  is the number in volume element  $(\partial V/\partial z)dz$  in luminosity element  $dL$ . The sky fraction accessible to each object is  $\Omega_i(z, L)$ .

The  $\mathcal{L}$ (likelihood) function for the  $i^{\text{th}}$  object is the product of (prob of observing 1 objects in its  $(dz, dL)$  element)  $\times$  (prob of observing zero objects in all other  $(dz, dL)$  elements accessible to it). The Poisson model is the obvious one for the likelihood:

$$f(x : \mu) = \frac{e^{-\mu} \mu^x}{x!}, \quad (1)$$

where  $\mu$  is the expected number. If  $x = 1$ , the function is  $\mu e^{-\mu}$ , if  $x = 0$  it is  $e^{-\mu}$ .

With  $\rho(z, L)$  as the full description of space density, then

$$\mu = \lambda(z, L) dz dL, \quad \text{with } \lambda = \rho(z, L)\Omega(z, L)(\partial V/\partial z). \quad (2)$$

Hence

$$\mathcal{L} = \prod_i^N \lambda(z_i, L_i) dz dL e^{-\lambda(z_i, L_i) dz dL} \times \prod_{j \neq i}^N e^{-\lambda(z_j, L_j) dz dL} \quad (3)$$

from which, if  $S = -2 \ln \mathcal{L}$  then

$$S = -2 \sum_i^N \ln \rho(z_i, L_i) + \int_z \int_L \rho(z, L)\Omega \frac{\partial V}{\partial z} dz dL + \text{constant}. \quad (4)$$

# Example: Marshall et al. 1983

Consider simple factorizable power-law density evolution of the form  $\rho(L, z) = \rho_0 \cdot \phi(z)$ .

For this formulation take

$$\frac{dN}{dL} = \rho(L, z) = \frac{\rho_c}{L_*} \phi(z) \left( \frac{L}{L_*} \right)^{-\alpha} \quad (5)$$

With  $L/L_* = l$  we have the local luminosity function as  $\rho_0 = \rho_c l^{-\alpha}$ .

Adopt 'power-law density evolution'  $\phi(z) = (1+z)^k$ .

Substituting this into equation 4, and setting the differential wrt  $\rho_c$  to zero yields a maximum likelihood estimate for  $\rho_c$

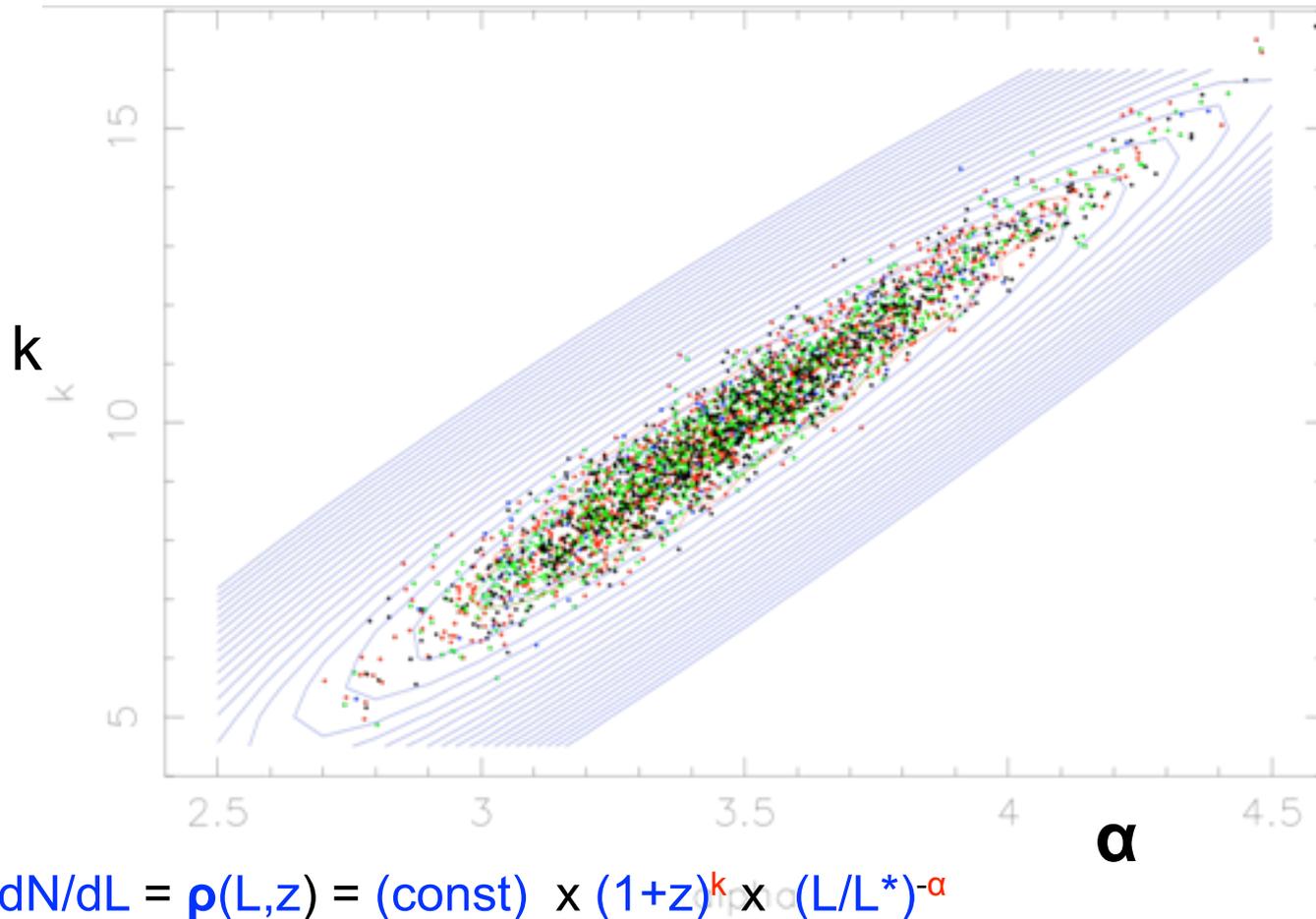
$$\rho_c = \frac{N}{\int_z \int_l (1+z) l^{-\alpha} \Omega(z, l) (\partial V / \partial z) dz L_* dl} \quad (6)$$

Putting this back into equation 4 we get

$$\begin{aligned} S &= -2 \sum_i^N \ln[(1+z_i)^k l_i^{-\alpha}] \\ &\quad + 2N \ln \int_z \int_l (1+z)^k l^{-\alpha} \Omega(l, z) \left( \frac{\partial V}{\partial z} \right) dz dl \\ &\quad + (2N - 2N \ln N). \end{aligned} \quad (7)$$

# Example: Marshall et al. 1983

And look: we have found with merely 32 objects, a simple way of describing the strong evolution of X-ray QSOs as a function of their luminosity (luminosity evolution).



OK. ok there are many more than 32 dots in the picture, but the function form is correct - the more luminous show much more evolution - this was a simulation to look how it might work of r larger samples.

# Survival Analysis: Censored Data - I

We start with a **primary sample of objects** - a series of measurements from which we pick out 'detections'. The results often find their way into **catalogues**, e.g. the New General Catalogue (NGC) or the 3CR Catalogue.

**'Resurvey'** is different, e.g. measuring H $\alpha$  luminosities of NGC galaxies.

- now it is very useful to quote **upper limits**; real objects are there.
- sometimes a resurvey may yield **lower limits**, e.g. measurement of X-ray and radio flux densities for the NGC galaxies would yield both upper limits and lower limits for the radio to X-ray spectral index.

# Survival Analysis: Censored Data -2

Statistics dealing with limits is called **'survival analysis'**, from medical statistics: at end of a study some subjects have survived, some not.

- measurements which are only limits are called **'censored'**.
- introduced into astronomy by Avni, Feigelson, co-workers 1980 >.

Survival analysis offers

- (i) **estimation of intrinsic distributions** (like luminosity functions),
  - (ii) **modelling and parameter estimation**,
  - (iii) **hypothesis testing**
- and (iv) **tests for correlation and statistical independence**, for cases in which some of the available measurements are limits.

# Survival Analysis continued

The key assumption is that the **censoring is random**; this means that the chance of only an upper limit being available for some property is independent of the true value of that property.

Assumption often met for flux-limited samples. For an object of true luminosity  $L$  and distance  $R$ , the condition for censoring is that

$$\frac{L}{R^2} < S_{\text{lim}}$$

the flux limit for the survey. If  $R$  is a random variable, independent of  $L$ , and  $S_{\text{lim}}$  is fixed, then the chance of censoring is independent of  $L$ .

**Careful examination of how a sample was selected is necessary to determine that survival analysis is applicable.**

# Survival Analysis continued continued

**Careful examination of how a sample was selected is necessary to determine that survival analysis is applicable.**

Let's be definite.

Take a survey at **wavelength A**; resurvey the sample at **wavelength B**.

So we get a complete set of  $L_A$ , some **values of  $L_B$**  and some **upper limits  $L_B^U$** .

Our aim – **to construct the normalized distribution** of  $L_B$ , which will be  $\alpha \rho_B$

There are two estimators available, a **recursive relation due to Avni** and the **Kaplan-Meier estimator of the cumulative distribution**. Both are **maximum likelihood**. Both work for upper and lower limits. The former requires binning; the latter does not rely on binning, but being cumulative, **errors are highly correlated** from one point on the estimate to the next.

# Survival Analysis continued even more

Here's the Kaplan-Meier estimator:

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

Lower limits: arrange data in **increasing** order.

$d_i$  is the number of observations of  $L_i$

$n_i$  is the number of observations  $\leq L_i$

$\delta_i = 1$  for detection,  $0$  for lower limit

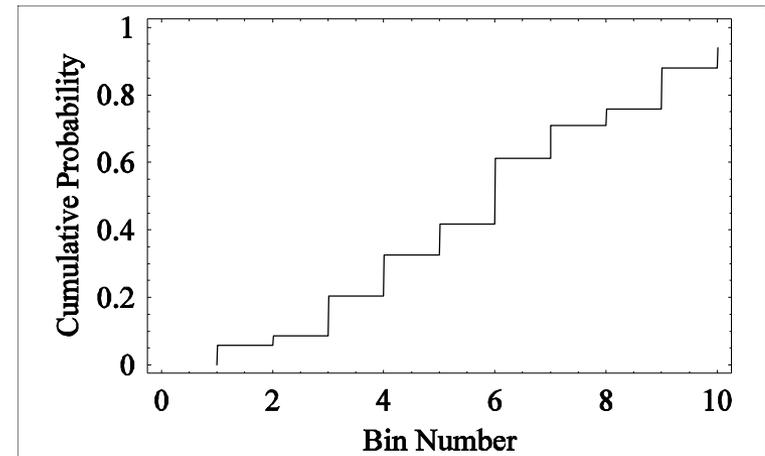
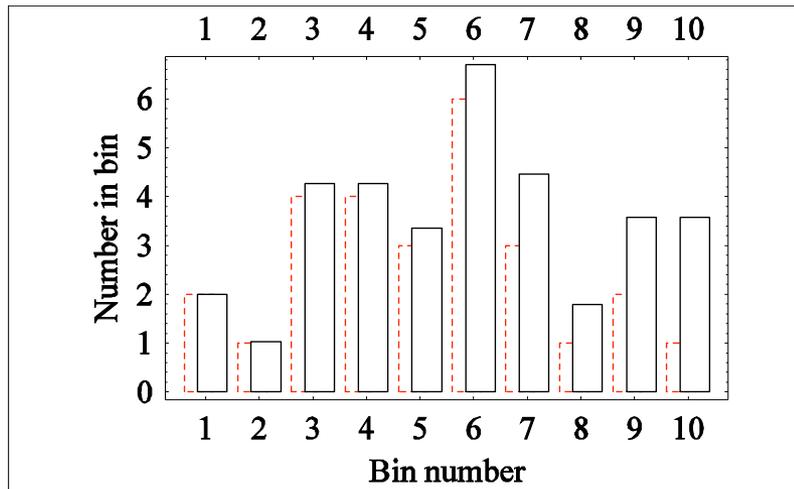
Upper limits: arrange data in **decreasing** order

$n_i$  is the number of observations  $\geq L_i$

$\delta_i = 1$  for detection,  $0$  for upper limit

# Survival Analysis continued even more

## Example 1 :



Left: Avni estimator, binned data. Distribution of spectral indices (optical to X-ray) for a sample of optically-selected QSOs; observed distribution (red dashed), and true distribution after including lower limits.

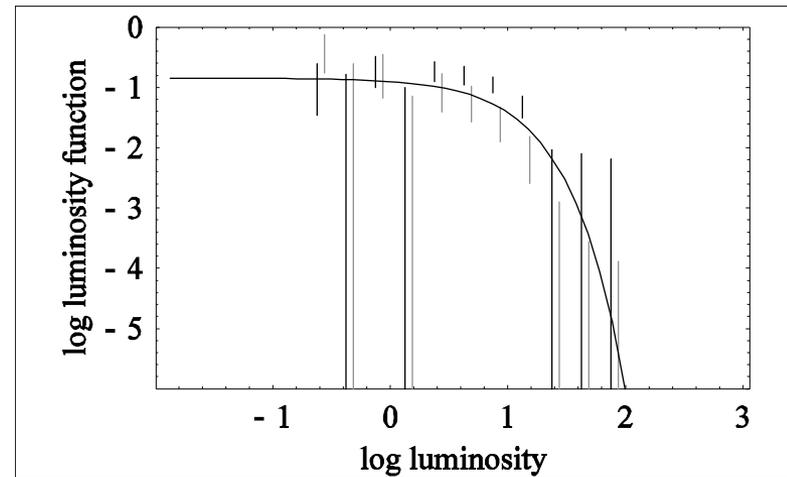
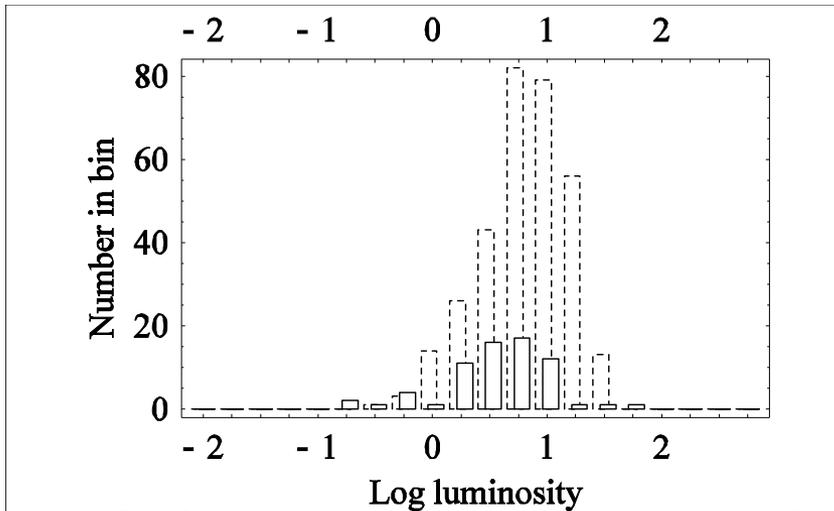
Right: same data, shown cumulatively as dots. Kaplan-Meier estimator is the solid line, following the bins used in the Avni formulation.

# Survival Analysis - Example 2

Example 2 : back to the simulated field-galaxy sample.

This was selected at one wavelength.

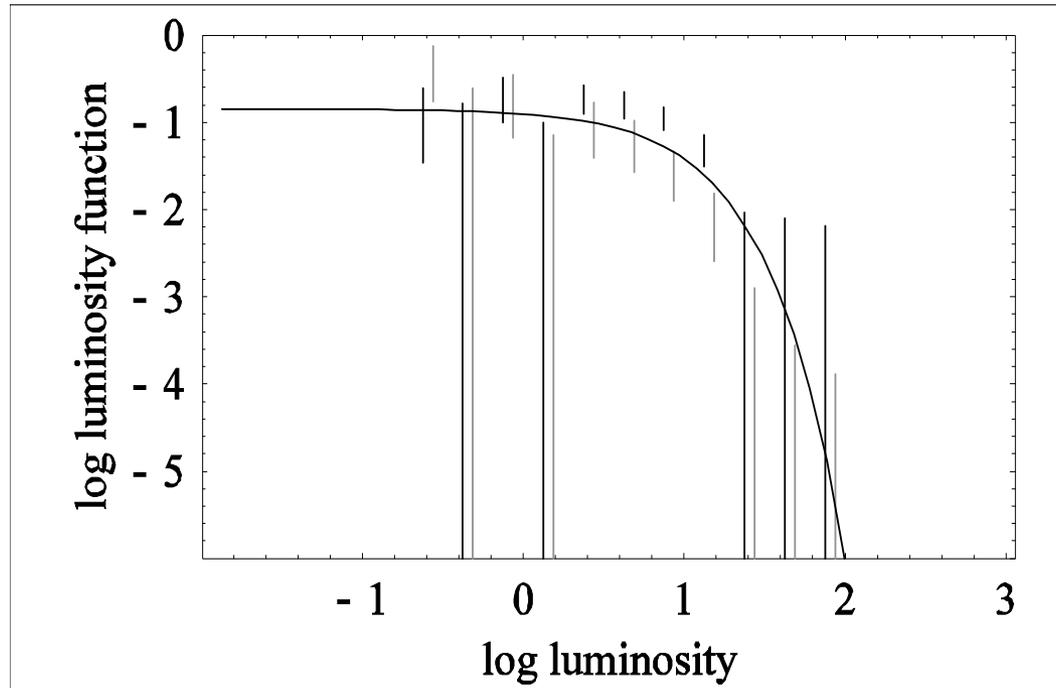
We 'resurvey' at another, and get 67 detections and 317 upper limits, in a simulation allocating luminosities from independent Schechter functions at both wavelengths.



Left: luminosity distribution and upper limits for the field galaxy simulation; there are 67 detections and 317 upper limits (dashed bins).

Right: luminosity probability distribution (black dots) from the Avni estimator, together with bootstrap error estimates. Solid line - theoretical distribution. Lighter dots are a  $1/V_{\max}$  estimate, displaced slightly in luminosity for clarity.

# Survival Analysis - Example 2 concluded



Survival analysis and  $1/V_{\max}$  results agree – both are MLEs and are based on similar models.

The survival analysis **gives better estimates here of the luminosity function, using more data.**

But the real advantage comes in correlation analyses, or reconstructions of non-distance-dependent distributions (e.g. spectral indices).

# Censored Data and Hypothesis Testing

We can test two distributions of observations against each other, using detections as well as limits.

There are several choices.

- but how were the samples selected? We will expect a problem whenever the variable of interest is correlated with the variable used to define the sample. The Malmquist bias of the defining variable will then be manifest in the other variable.
- If the bias is not the same for the two samples (and it depends on the observational method), a bogus difference will be detected.

# Censored Data and Hypothesis Testing (2)

Variants of W-M-W rank tests are used:

- Intuitively we might expect the ranking procedure to be applicable for data containing limits, as limits should be randomly intermingled.
- Constructing a test statistic depends on the penalty we assign for non-random intermingling, and how we distribute this penalty between detections and limits.
- Feigelson et al (1985) described two variations on this idea, the Gehan and log-rank tests. Asymptotic distributions are known for the statistics, but simulation will be more reliable for small samples.

# Censored Data and Hypothesis Testing, more

The **Gehan test** is simplest and here it is, crammed onto one slide:

- (1) take two samples, labelled **A** and **B**, including both detections and limits.
- (2) Arrange the **detections** in order; **ascending order** for data with **lower** limits, **descending order** for data with **upper** limits.
- (3) Number the observations; this gives each datum a rank. Call the  **$j^{\text{th}}$**  rank for data from sample A  **$r_{jA}$** .
- (4) For the  **$j^{\text{th}}$**  detection from sample A, calculate  **$n_{jA}$** , the number of observations of A which are to the right. By “right” we mean data that are  $\geq$  the  **$j^{\text{th}}$**  observation (in the case of lower limits), or  $\leq$  the  **$j^{\text{th}}$**  observation (in the case of upper limits). Thus this part of the calculation uses the limits. The number of limits from sample A between detection  **$i$**  and detection  **$i+1$**  is  **$m_{iA}$** .
- (5) The Gehan statistic: 
$$\Gamma = \sum_{\text{detections in A}} (n_{iA} - r_{iA}) - r_{iA}m_{iA},$$

This is asymptotically distributed as a Gaussian of mean zero and variance  $\sigma^2 = \sum_{\text{detections}} n_{iA}n_{iB}$

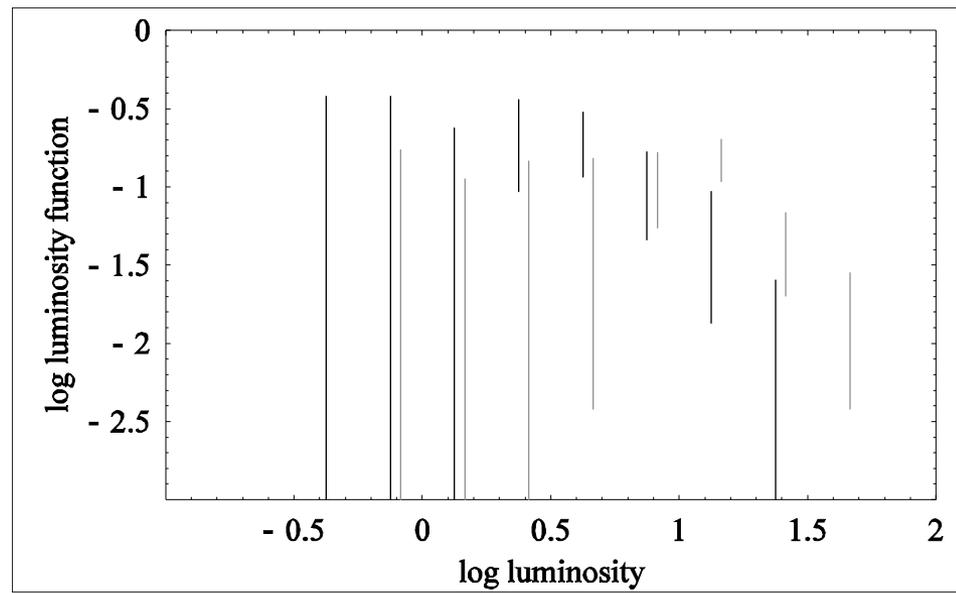
# Hypothesis Testing, Censored Data - Example

Here's an example: We simulated two samples of objects, each drawn from the field-galaxy Schechter function but with different characteristic luminosities:

Sample A: 23 detections and 149 limits,  $L_* = 10$

Sample B: 45 detections and 167 limits,  $L_* = 30$

The estimated luminosity functions (Avni method, bootstrap errors) show an appreciable difference:



The Gehan test gives  $\Gamma/\sigma = 3.3$ , significant at the 0.1~per cent level (if the asymptotic approximation holds for these small numbers, this far out in the wings). 19

# The Confusion Limit

In many cases of astronomical interest, we find that faint objects are much more numerous than bright ones.

**Faint objects crowd together**; ultimately they start to be unresolved and our signal becomes a mixture of objects of various intensities, **blended together by the point spread function of our instrument.**

Examples include radio sources in deep surveys, spectral lines in the Lyman- $\alpha$  forest, stars at the cores of globular clusters, and faint galaxies observed in the optical.

The **confusion limit** concept was developed during a bitter controversy amongst radio astronomers and cosmologists in the 1950s, the **source-count/Big Bang/Steady State** controversy.

The root of the problem was **instrumental**, wildly different source counts being obtained at Sydney (Mills Cross; essentially filled aperture) and Cambridge (interferometer).

Scheuer (1957) analyzed the statistics of the source counts and showed that the Cambridge results were seriously affected by **confusion**. The wide beam of the interferometer meant that **many radio sources were contributing to each peak** in the record; these had erroneously been interpreted as discrete sources.

# The Confusion Limit

We touched on this issue early in this lecture.

In many cases of astronomical interest, we find that faint objects are much more numerous than bright ones - galaxies on the sky, etc. This is frequently described as ‘the number count being steep’., i.e.  $N(S)$  where  $S$  is brightness, usually approx a power law, has an exponent of  $< -2$ .

**Faint objects crowd together** under these circumstances; ultimately they start to be unresolved and our signal becomes a mixture of objects of various intensities, **blended together by the point spread function of our instrument.**

Examples include radio sources in deep surveys, spectral lines in the Lyman- $\alpha$  forest, stars at the cores of globular clusters, and faint galaxies observed in the optical.

# The Confusion Limit

The **confusion limit** concept was developed during a bitter controversy amongst radio astronomers and cosmologists in the 1950s, the

## **source-count Big Bang (evolution) versus Steady State**

controversy, main protagonists Ryle (evolution) and Hoyle (Steady State).

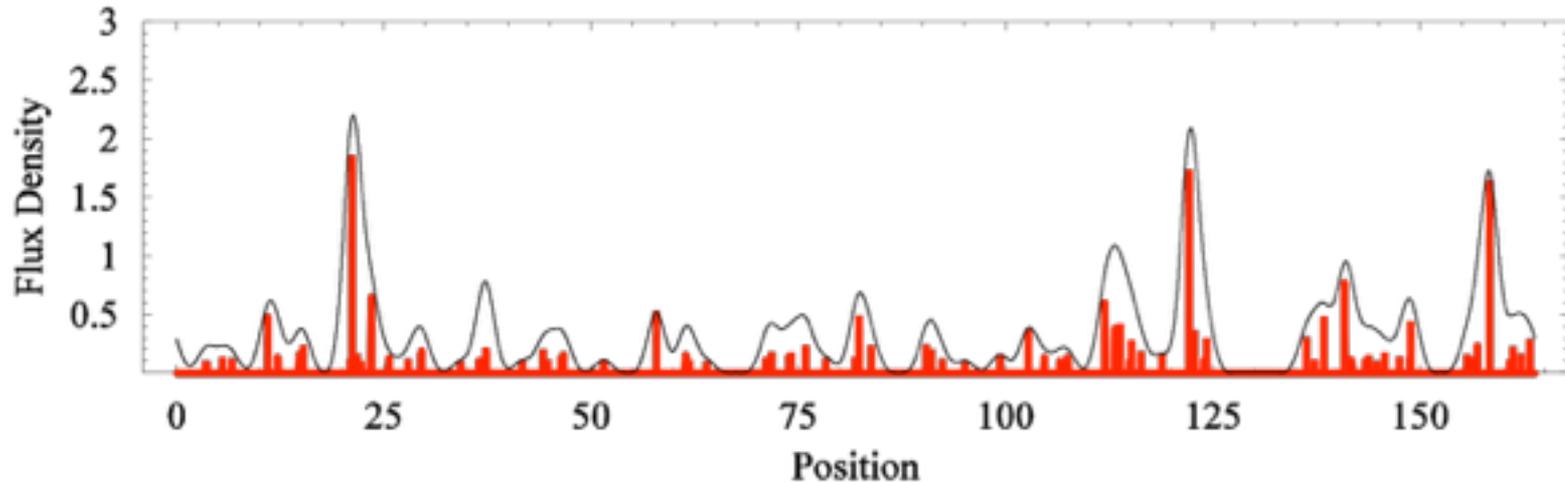
The root of the problem was **instrumental**, wildly different source counts being obtained at Sydney (Mills Cross; essentially filled aperture) and Cambridge (interferometer).

Scheuer (1957) analyzed the statistics of the source counts and showed that the Cambridge results were seriously affected by **confusion**. The wide beam of the interferometer meant that **many radio sources were contributing to each peak** in the record; these had erroneously been interpreted as discrete sources.

>80% of the “sources catalogues in the 2C survey were blends, not real.

# The Confusion Limit - Simulated Example

To show the pronounced effect of confusion, here is a simulation of a 1D scan of sources obeying a Euclidean source count  $N(>S) \propto S^{-3/2}$ . The beam is a simple Gaussian and there is, on average, one source per beam.



Confusion simulation at a level of one source per beam area. Input sources - red verticals; solid line is the response when convolved with a Gaussian beam.

A simple count of the peaks in the record gives a maximum-likelihood slope for the source count of -1.8 with standard deviation 0.3, very different from the true (input) value of -1.5:

# The Confusion Limit - **Aside**

As it turned out, both sides were wrong and right.

Firstly, the large redshifts of radio sources require an initial slope much flatter than  $-3/2$ , a point rarely if ever acknowledged in the literature.

Thus Mills “discovery” that the slope was close to  $-3/2$  was hence irrelevant to the argument.

Secondly the Cambridge survey technique was indeed horribly at fault;  $>80\%$  of the sources in the 2C catalogue, the publication at primary issue, were blends, not real sources at all.

Its resultant  $N(S)$  curve was twisted completely out of shape, far too steep initially.

But because Steady State required such a shallow slope:

**Ryle was right, for all the wrong reasons. Strong cosmic evolution prevails, widely recognized in galaxy formation, star-formation rate, and our “concordance” model of the Big-Bang Universe, ‘originating’ in a hot dense phase.**

# The Confusion Limit - the P(D) Technique

The technique developed by Scheuer is known to astronomers as “**P(D)**”, or “**(P)robability of (D)eflection**” - the deflections being of the pen on a chart recorder.

The technique has been used in the radio, the X-ray, infrared, and Lyman- $\alpha$ , at least.

The method **derives the probability distribution of measurements in terms of the underlying source count**, which itself may be recovered by a model-fitting process.

Its benefits are that (a) information is obtained from sources that are much too faint to be “detected” as individuals, and (b) the correct form of the count from the faint levels of the survey is derived in an unbiased way.

# The Confusion Limit - the P(D) Technique

The derivation of the distribution requires conditional probabilities, Poisson statistics, autocorrelation functions, Fourier transforms. It is most elegant bit of analysis by Scheuer.

The result: the FT of the p(D) distribution

$$P(\omega) = \exp(R(\omega) - R(0)) \quad \text{in which} \quad r(s) = \int N \left( \frac{s}{\Omega(x)} \right) \frac{dx}{\Omega(x)}$$

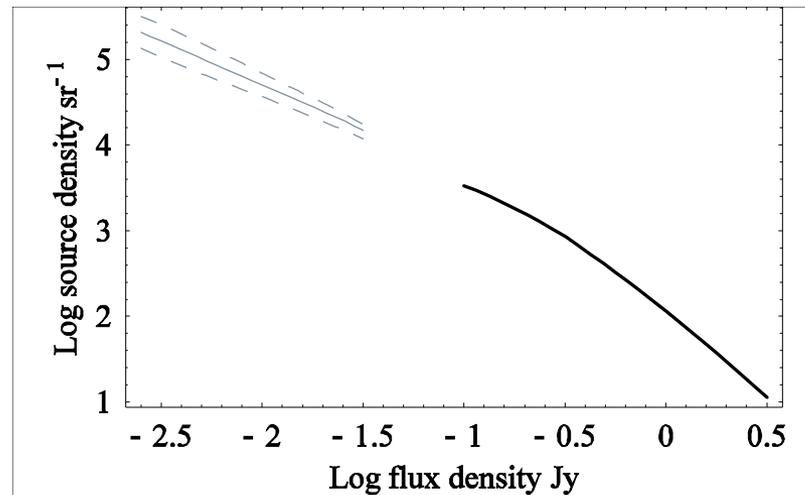
**contains the source count  $N$ .  $R$  is the FT of  $r$ .** Analytic solutions are available when  **$N(S)$**  is a power law, but the inverse transform to get  **$p(D)$**  has to be done numerically.

In real life we need to take account of **differential measurement** techniques in which measurements from two positions are subtracted to avoid baseline errors. And there's always **noise**. A modelling process is needed to recover  $N(S)$ . Vernstrom et al. (2014) have carried this out in detail, using the JVLA to determine radio source counts into the nano-Jy regime.

# Confusion Limit and P(D) - Example 2

The derivation of source counts from  $p(D)$  is another technique in which population characteristics are derived from observations of discrete objects or features without forming an object list or catalogue.

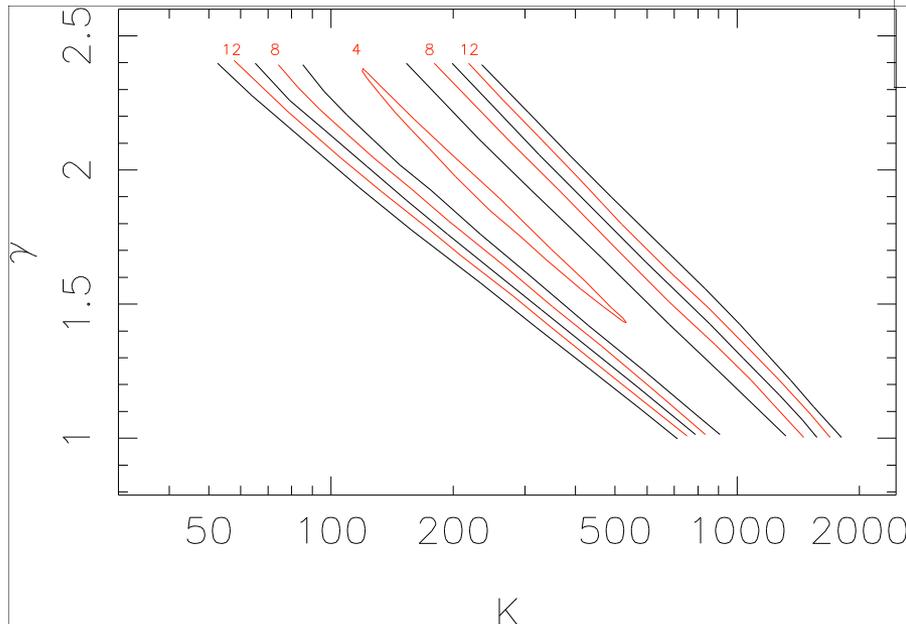
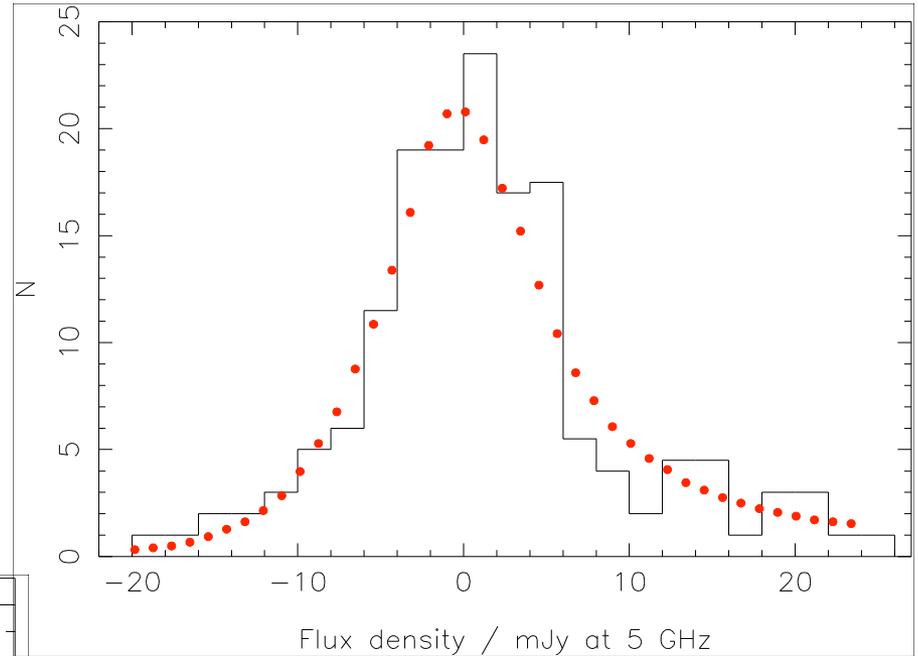
Wall & Cooke (1975) applied the  $p(D)$  technique for filled aperture telescopes to extend the 2.7-GHz radio source counts to much fainter levels than could be achieved by identifying individual sources:



The 2.7-GHz counts from Wall & Cooke (1975): the darker line is derived from ordinary source counts with error bars not much wider than the line, while the  $p(D)$  results are shown in grey, the dashed lines representing one standard deviation of the fitted parameters.

# Minimum Chi-Square method / P(D) - Example 3

**Chi-square testing/modelling:** the object of the experiment was to estimate the surface-density count (the  $N(S)$  relation) of faint radio sources at 5 GHz, assuming a power-law  $N(>S) = KS^{-(\gamma-1)}$ ,  $\gamma$  and  $K$  to be determined from the distribution of background deflections, the **p(D) method**. The histogram of measured deflections is shown right.



The dotted red curve above represents the optimum model from minimizing  $\chi^2$ . Contours of  $\chi^2$  in the  $\gamma - K$  plane are shown left.

With the best-fit model,  $\chi^2 = 4$  for 7 bins, 2 parameters; thus dof = 4. **Right on.**

# Confusion and P(D) - Example 3

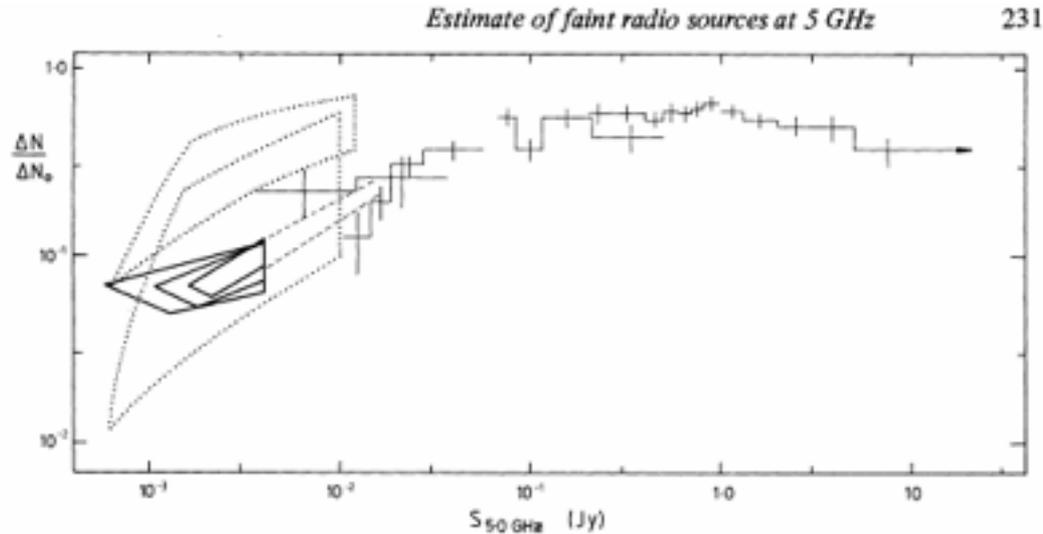


Figure 7. The 5-GHz source count. The points from direct surveys are as in Fig. 6. The region defined by solid lines is that from the present statistical experiment. Of the two regions defined by dotted lines, the larger is from the original analysis by Wall & Cooke (1975), while the smaller (Wall 1978) was obtained from ON-ON observations giving the systematic error discussed in Section 4. The dashed lines suggest likely boundaries for the faint count at 5 GHz.

If only Wall & Cooke (1975) or Wall et al. 1982 had known anything about Bayes and priors.....

END