OPTIMAL STRATEGIES FOR ESTIMATING COSMOLOGICAL PARAMETERS

by

MARTIN HENDRY

Thesis

submitted to the

University of Glasgow

for the degree of Ph.D.

Department of Physics and Astronomy,

The University,

Glasgow G12 8QQ

January 1992

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

CONTENTS

for Uncle George

"God Measuring the Universe"

13th Century Manuscript, Austrian National Library - Vien

(Unfortunately these results remain unpublished.)

"So numerous and so powerful are the causes which serve to give a false bias to the judgement that we on many occasions see wise men on the wrong as well as on the right side of questions of the first magnitude. The circumstance, if duly attended to, would furnish a lesson of moderation to those who are ever so much persuaded of their being in the right on any controversy."

Alexander Hamilton (The Federalist Papers)

PREFACE

The work of this thesis concerns the estimation of galaxy distances using methods which are independent of redshift. Such observational techniques have traditionally been used in studies of the velocity field, to test the linearity of the Hubble Law and detect possible anisotropies in the Hubble flow. A source of considerable debate in the literature over the past decade has been the impact of observational selection effects on studies of the Hubble flow. The presence of a magnitude cutoff in one's galaxy sample will in general introduce Malmquist bias. This bias arises because, as we sample the galaxy distribution at greater distances, only intrinsically more luminous galaxies can still be observed - which leads to a distance dependence in the mean luminosity of observable galaxies. A failure to account for this effect would result in one systematically underestimating the distances to remote galaxies, and Malmquist bias has been cited by many authors as responsible for the controversy over the global value of H_0 and the detection of large-scale streaming motions.

In the past few years the study of the peculiar velocity field has taken on - quite literally - a whole new dimension. Sophisticated techniques have been developed to recover in a self-consistent manner the full 3-dimensional velocity and density fields. One method in particular requires redshift-independent distance estimates to galaxies, and will clearly yield a more effective recovery if those distance estimates are made more reliable. It is this task which we set out to achieve in this thesis, by studying the properties of different methods

used to infer galaxy distances and examining how each is affected by Malmquist biasing. Our aim is to identify methods of removing this bias, and thus determining an 'optimal' choice of distance estimator.

This work was carried out while the author was a research student in the Department of Physics and Astronomy, University of Glasgow, while in possession of a Carnegie research scholarship. I would like to thank the Carnegie Trust for their generous support during this time - both in the provision of a stipend and in the award of several travel grants which allowed participation in cosmology conferences, both at home and overseas. I would like to sincerely thank my supervisor, Dr. J.F.L. Simmons, for a constant supply of valuable advice and friendly support throughout this time - and for not objecting too much when I finished his sentences off for him! A special word of thanks also to Daphne for solving many sudden logistical crises, while never compromising the punctuality of tea and coffee, particularly during my 'transitory' phase over the last few months. I think it's safe to remove my name from Rm 412 at last. A general thank you is extended to everyone in the Department - time does not permit me (quite literally!) to particularise; I hope that everyone realises how much I have enjoyed my time working at Glasgow, and the contribution which you have all made to that enjoyment. I think it is safe to say that, even in an infinite and open universe - as appears to be somewhat favoured by current observations - the students and staff of the Glasgow astronomy department are certainly unique.

Thank you to all my family and friends for their great personal support over the past three years: firstly to my Mum and Dad for making all this possible and, together with my sister, Anne, providing a source of great encouragement and support - and showing admirable tolerance of my eccentric working habits and somewhat nomadic existence. A big thank you also to the rest of my family, brothers and sisters, nieces and nephews, and friends in East Kilbride, Glasgow and across the globe who have all helped in various ways to keep me in touch with the real world; I hope they've succeeded.

Martin Hendry

SUMMARY

In this thesis we study the effects of observational selection bias on the estimation of galaxy distances in cosmology. Although the presence of systematic bias in magnitude-limited surveys has long been recognised there remains disagreement in the literature as to precisely how best to reduce or eliminate its effects from redshift-independent distance estimates. The aim of this thesis is to develop a statistically rigorous formulation of the problem of distance estimation, so as to resolve some of the issues which have clouded past discussion and allow one to determine strategies for obtaining optimal distance estimators.

Redshift-independent distance estimates, when combined with the measured redshift, provide an estimate of a galaxy's peculiar velocity. The study of the large-scale peculiar velocity field has been a very active and contentious subject in recent years, following a number of independent reports of coherent structure and velocity flows on very large scales which pose serious problems for popular theories of structure formation. In chapter (1) we present a brief overview of our current picture of the local universe and summarise the basic features of theoretical models for the formation and evolution of structure. We compare in detail two different analytical techniques which have been developed to recover the full peculiar velocity and density fields from redshift surveys: the POTENT method (Bertschinger and Dekel, 1989) and the 'IRAS' method (Strauss *et al*, 1990). The former method requires redshift-independent distance estimates and we consider the effects of sparse and noisy sampling on the recovered density and velocity fields, demonstrating the advantages for POTENT

of removing the effects of selection bias from distance estimates.

Chapter (2) presents a detailed description of distance indicators currently used in cosmology. We review previous analyses of distance estimation and biasing problems and discuss the limitations of the 'Minimum Bias Subset', an early method proposed to remove them. We examine the different linear regression techniques used to calibrate indicators such as the Tully-Fisher relation and address the question of which method is 'best'. In particular, we consider a scheme, proposed by Schechter (1980), for obtaining unbiased distance estimates provided that one's sample is subject only to luminosity selection. This scheme has not been universally endorsed in the literature and many authors prefer other calibration methods. This disagreement is one of the main issues which we aim to clarify and resolve in this thesis.

In chapter (3) we introduce a formulation for defining and investigating the properties of distance estimators in a statistically rigorous fashion. We firstly consider the case where distances are estimated using only measurements of apparent magnitude. Assuming a Gaussian luminosity function we derive expressions for the conditional distribution of observable galaxies at a given (though in general unknown) true distance and use this distribution to define a number of different distance estimators and compare their distributions, bias and mean squared error or *risk* as a function of true distance. This simple case is used to illustrate useful criteria by which we can identify which estimator is 'best'. In this chapter we also describe a procedure for constructing confidence intervals for the true distance of a galaxy.

In chapter (4) we extend our analysis to the case where

distance estimates are made from measurements of two or more observables, accounting for the effects of selection bias. We show that the different methods of regression used to calibrate these relations correspond to simple distance estimators which arise naturally from our rigorous formulation. We also define a 'maximum likelihood' distance estimator and, following the method introduced in chapter (3), we compute the distribution, bias and risk of all of these estimators as a function of true distance. These results allow us to test the validity of the 'Schechter' scheme and identify situations where the corresponding estimator is a poor choice. Finally, we extend our procedure for constructing confidence intervals to this two-observable case.

In chapter (5) we consider distance estimators constructed from a linear combination of three observables - again including the effects of selection. By computing the distribution, bias and risk of these estimators we determine, in particular, whether one may still define unbiased distance estimates in this case by adapting the 'Schechter' scheme. Furthermore, we examine quantitatively the extent to which the addition of a given third observable improves distance estimates obtained from measurements of only two. We consider the importance of these results for e.g. the D_n-σ relation, for which potentially useful third observables exist.

In chapter (6) we summarise the main qualtitative results of this thesis and explore a number of possible avenues for future work; in particular a study of the consequences of our results for the analysis of redshift surveys, by reconstruction methods such as POTENT.

1. NEW METHODS FOR THE ANALYSIS OF REDSHIFT SURVEYS

1.1 Introduction: Setting the Scene

In recent years a startling new picture of how matter is distributed in the universe has begun to emerge. The high degree of smoothness of the microwave background radiation - as repeatedly detected in progressively more accurate experiments - provides stronger than ever evidence of the isotropy and homogeneity of the cosmos on the very largest scales. On scales of a few tens of Mpc, on the other hand, the universe which we survey today appears far from uniform. Observations indicate a rich and varied degree of structure on these scales: dense clusters and superclusters of galaxies embedded in a complex network of intersecting filaments and sheets, enclosing vast underdense regions, or 'voids' which appear to contain almost no luminous material and may be as much as 100Mpc in diameter. These observations have transformed our earlier view of a smooth, homogeneous galaxy distribution and the existence of such large-scale structure has presented serious difficulties for theories of structure formation.

A major factor in allowing this dramatic new picture to emerge has been the introduction to cosmology of powerful new observational techniques and instrumentation. For example, the automatic scanning, measurement and cataloguing of photographic plates has significantly improved exisiting data on the distribution of galaxies and clusters projected on the sky. A project initiated with the Cambridge Automatic Plate Measuring Machine has produced the

APM galaxy survey (Maddox *et al*, 1990) which consists of a catalogue of over two million galaxies - with measured apparent magnitudes, diameters and orientations - brighter than a limiting magnitude of ≈20.5. This catalogue is more than double the size of the earlier Lick catalogue (Shane and Wirtanen, 1967) and probes to an effective depth which is more than twice that of the Lick catalogue. Moreover, the APM survey is also rendered somewhat 'cleaner' than its earlier counterpart by a series of algorithms designed to automatically distinguish foreground stars from galaxies and to correct for plate-to-plate variations in sensitivity, sky background and other contaminating factors. Figure (1.1) shows a projected map of the distribution of galaxies in the APM survey with apparent magnitudes between 17 and 20.5. The impression of complex structure - rich clusters, filaments and voids - is quite clear from a purely visual inspection, without the need to apply any quantitative statistical analysis to the catalogue.

Perhaps an even more significant observational advance for our prospects of understanding cosmic structure, however, has been the advent of fast quantum electronic detectors which make it possible to measure the redshifts of large numbers of galaxies - and thus infer a first estimate of their distance - in a realistic period of time.

In the first systematic survey of galaxy radial velocities by Hubble (1929) - from the results of which Hubble proposed a linear velocity distance law and interpreted this as a general expansion of the universe - the measurement of a single redshift would typically require an entire night of observing time, using the best photgraphic

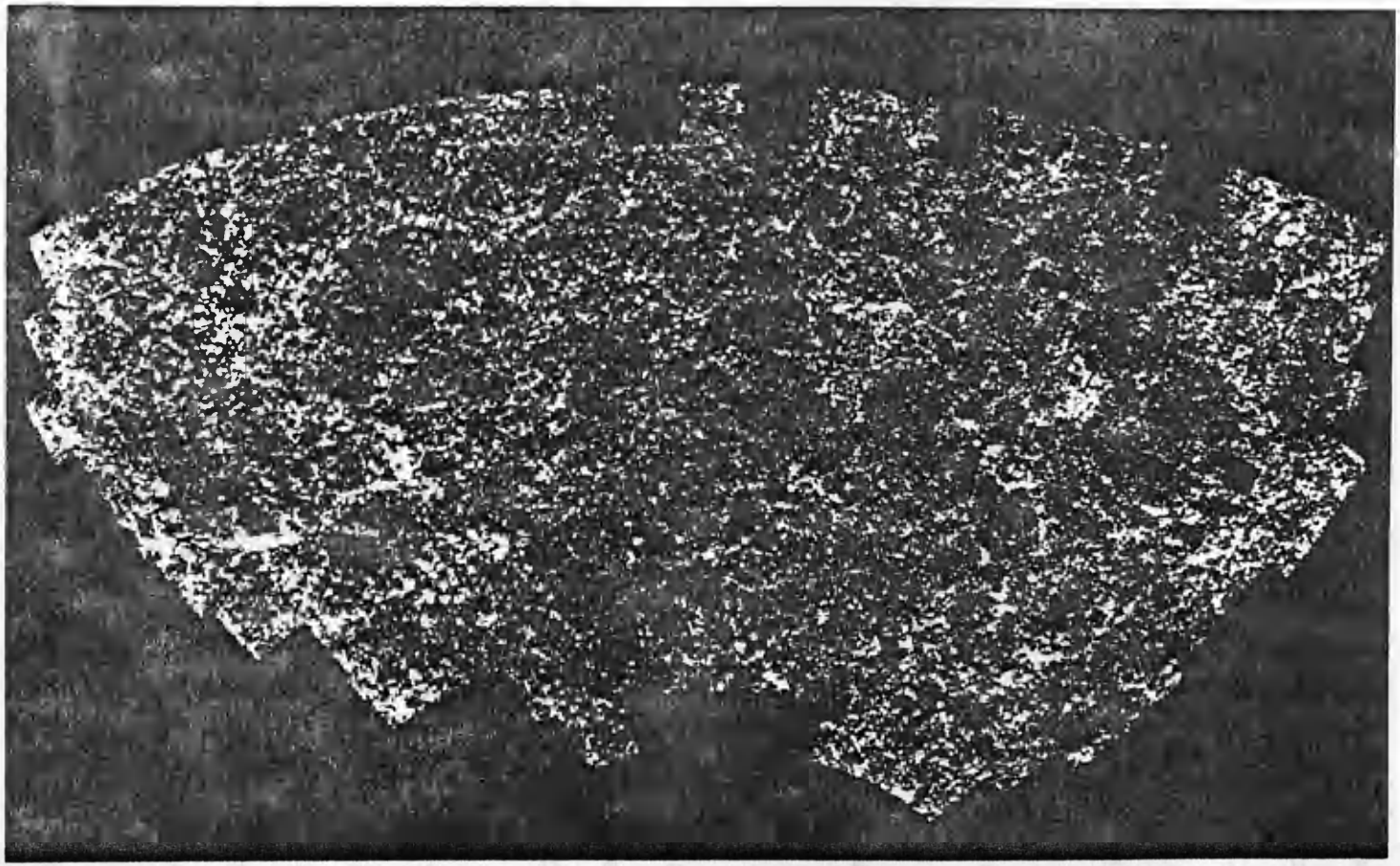

Figure (1.1)

The distribution of more than two million galaxies in the APM survey, constructed from scans of 185 UK Schmidt plates shown in equal area projection centred on the South Galactic pole (Maddox *et al*, 1990)

materials available. By contrast, present-day observations using CCDs can record the redshifts of typical galaxies in less than half an hour. Of course, this is still a considerable length of time for practical purposes, and the number of galaxy redshifts measured currently stands at only a few tens of thousands; a very much smaller total than the number of objects in the APM survey, for example. Nevertheless, the efficiency of redshift measurements can be greatly improved by using multi-channel fibre optics to record simultaneously the redshifts of many galaxies in the same field; such techniques are expected to become routine in the near future, and the redshift database will grow rapidly in size. Indeed, groups in Princeton and Chicago have already begun a 10-year long program to obtain a redshift survey of 1 million northern sky galaxies.

In fact, the existing redshift database is already sufficient in number to give a strong impression of how galaxies are distributed in the line of sight direction, and observations support the existence of the structure detected in projected surveys. The largest single redshift survey carried out to date is the Center for Astrophysics (CfA) Harvard survey which contains around 9000 objects. The 'CfA Slices' obtained from this survey and presented in de Lapparent *et al* (1988) have been very influential in promoting the picture of a local universe rich in structure and provide strong evidence for a bubble-like galaxy distribution in three dimensions. Figure (1.2) shows an example of recent work (de Lapparent *et al*, 1991) extending the original CfA slices to slightly higher redshift.

As a result of these technological developments, therefore,

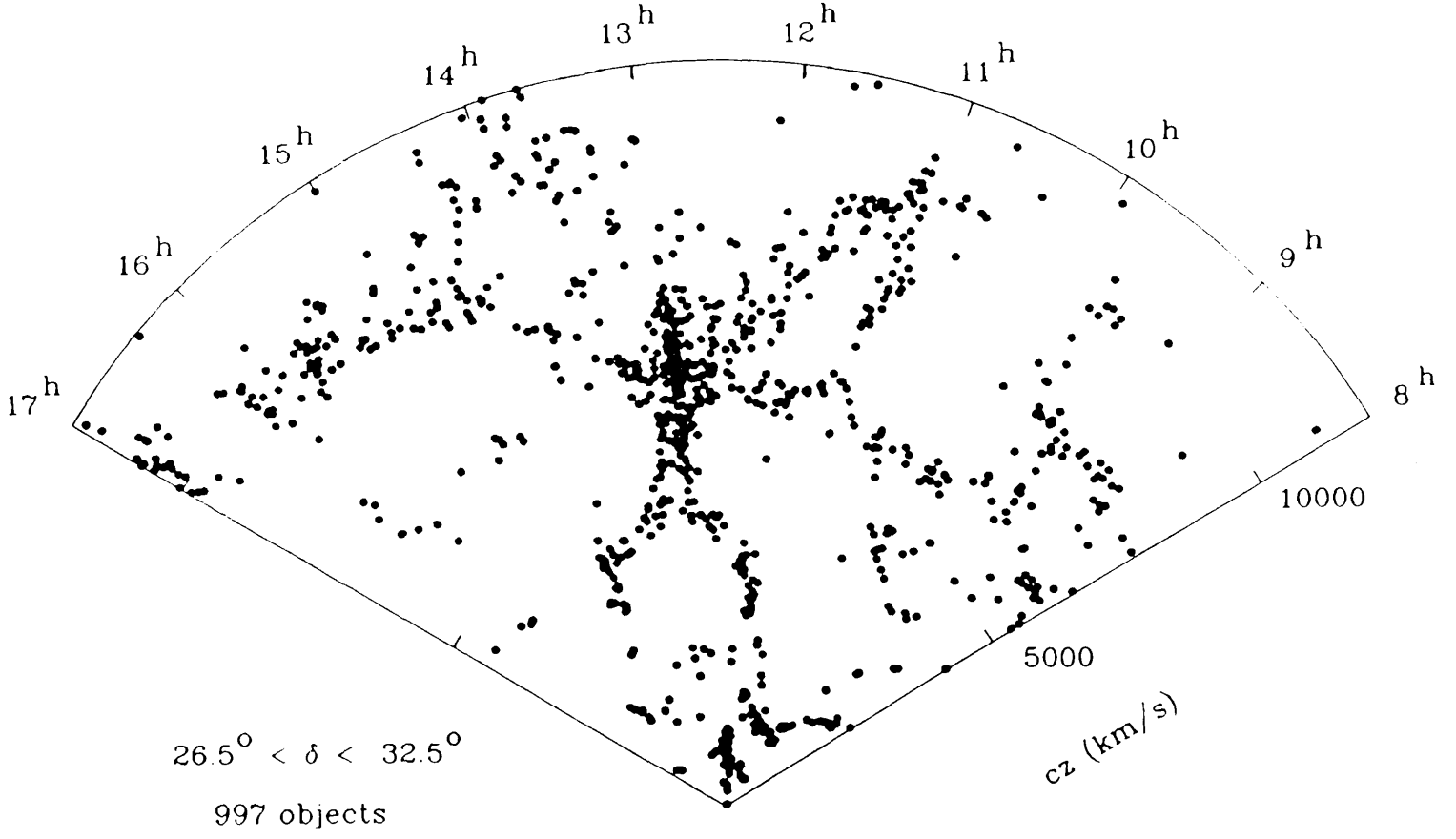

Figure (1.2)

Distribution in velocity and right ascension of sampled galaxies, with m ⩽ 15.5 and v ⩽ 12000kms^{-1}, in a slice of the CfA redshift survey extension (de Lapparent *et al*, 1991)

cosmologists find themselves for the first time able to form a picture of the 3-dimensional distribution of galaxies and, consequently, redshift surveys have been the subject of a great deal of detailed research in the past few years.

A strong further motivation for this current interest has been a desire to better understand the *dynamical* properties of the galaxy distribution. Indeed, one source of concern in considering redshift data has always been that if the measured radial velocities of galaxies display a significant deviation from their predicted Hubble velocity due to random or systematic 'peculiar' motions then the distribution of galaxies in *redshift* space (or, equivalently, velocity space) may be appreciably distorted from the true spatial distribution. This would tend to elongate and exaggerate structure along the line of sight in the CfA slice shown above, and has been termed the 'Finger of God' effect. Over the past fifteen years a number of studies (details of which we will shortly consider) have suggested that such large 'non-Hubble' velocities are indeed present, indicating that not only is there coherent structure on scales in excess of 50Mpc, but there is also coherent velocity streaming on such scales.

Initially such claims were certainly regarded with suspicion, not least because of the trouble which they spell for theories of structure formation, and a focus of the controversy was the possibility that the detected streaming motions could be an artefact of uncorrected statistical biases in the data. At any rate, it was certainly acknowledged that - whatever the status of the streaming motions - a better understanding of the dynamical properties of the galaxy

distribution was crucially important if the mysteries of how the observed structure formed were to be unravelled.

To this end, therefore, sophisticated new analytical techniques have recently been developed in order to reconstruct from redshift data an estimate of both the spatial density distribution *and* the galaxy peculiar velocity field. In this chapter we will describe these methods and discuss some of the assumptions and limitations to which each is subject.

One of the techniques in particular, the POTENT method introduced in Bertschinger and Dekel (1989, hereafter BD), requires galaxy distance estimates which are independent of redshift, and consequently most of the results of this thesis concerning the optimal estimation of distances are of particular relevance to POTENT.

1.2 Modelling Large-Scale Structure

Before we consider the analysis of redshift surveys in more detail, we will briefly sketch the standard mathematical description with which the evolution of structure is modelled. We present here no more than a short synopsis of the key features which are relevant to understanding recent techniques developed to analyse redshift surveys. The modelling of structure formation and evolution is discussed in considerable detail in a large number of textbooks, review articles and papers (for a thorough treatment see, for example, Zel'dovich (1970), Weinberg (1972), Peebles (1980) and Kolb and Turner

(1990), and references therein). Also of particular relevance are Jones and van de Weygaert (1990) and Jones (1991), which give a concise and clear discussion of the topic in the present context of redshift surveys.

Models for the evolution of cosmic structure are developed within the framework of the standard Friedmann Robertson Walker (FRW) model, which is more or less the cornerstone of theoretical cosmology. In the FRW model the universe is considered as an expanding perfect fluid which is spatially homogeneous and isotropic. The matter density in the FRW universe is therefore a constant, $\rho_b(t)$ say, at each epoch, t. The amount of expansion at any given epoch is determined by the scale factor, a(t), which is usually defined as the measured length of some chosen 'yardstick' at time, t, divided by its length measured at the present time, t_0. The firmest evidence for the validity of the FRW model comes from the observed isotropy of the microwave background, as referred to previously. Temperature fluctuations in the background radiation over a range of angular scales from 1' to 180° are of the order of 10^{-4} or less, and this figure gives a direct measure of the corresponding fluctuations in the gravitational potential due to deviations from uniform density.

The basic assumption of most formation scenarios is that the structure which we observe today initially formed from the growth of small-amplitude density fluctuations in a FRW universe, under the mechanism of gravitational instability. The evolutionary behaviour of structure can therefore be studied by considering a weakly perturbed FRW solution to the Einstein field equations. The analysis is further

simplified by the fact that, for perturbations of scale length smaller than the horizon size, a simple Newtonian treatment of the expanding fluid is an adequate approximation.

The evolution of the density field, $\rho(\mathbf{x},t)$, peculiar velocity field, $\mathbf{v}(\mathbf{x},t)$, and gravitational potential, $\Phi(\mathbf{x},t)$, are determined to first order by the following equations (assuming zero pressure):-

$$\frac{\partial \rho}{\partial t} + 3H\rho + \frac{1}{a}\nabla.(\rho \mathbf{v}) = 0 \qquad (1.1)$$

$$\frac{\partial \mathbf{v}}{\partial t} + H\mathbf{v} = -\frac{1}{a}\nabla\Phi \qquad (1.2)$$

$$\nabla^2\Phi = 4\pi G a^2(\rho - \rho_b) \qquad (1.3)$$

These are respectively the continuity equation, Poisson equation and Newtonian Euler equation of motion for the expanding fluid expressed in comoving coordinates (i.e. comoving with the uniform background). Here $H \equiv \dot{a}/a$ is the expansion function which measures the expansion rate of the background model and whose present value we know as the Hubble constant, H_0.

Fluctuations in the density field are usually measured in terms of the dimensionless density contrast, ε, defined as the fractional difference between the actual density and the density of the uniform background model, viz:-

$$\varepsilon(\underline{x},t) = (\rho(\underline{x},t) - \rho_b(t)) \,/\, \rho_b(t) \qquad (1.4)$$

It is also generally assumed that the density fluctuation field was

initially Gaussian, i.e. completely specified by its power spectrum, and the form of the power spectrum is given by theoretical considerations based on, e.g., the assumption of an inflationary phase in the very early universe and on the adopted nature of cosmological dark matter. A number of different dark matter candidates have been proposed: e.g. baryonic and non-baryonic; "hot" or "cold" - each of which carries a characteristic signature for the timescale and spatial configuration of structure formation which may be compared with observations. The cold dark matter (CDM) model in particular has difficulties in explaining the large-scale coherent structure and velocity flows which appear to have been observed.

In the linear regime the peculiar velocity field, $\mathbf{v}$, generated by the density contrast field, δ, is proportional to the gravitational acceleration, $\mathbf{g}$, defined by:-

$$\mathbf{g} = -\frac{\nabla\Phi}{a(t)} \tag{1.5}$$

Thus we find:-

$$\mathbf{v} = -\frac{f(\Omega)}{3H\Omega}\,\mathbf{g} \tag{1.6}$$

where Ω is the density of the background model in units of the critical density ρ_c for a flat universe. The function $f(\Omega)$ can be well approximated by (c.f. Peebles, 1980):-

$$f(\Omega) = \Omega^{0.6} \tag{1.7}$$

Solving the Poisson equation (1.2), we can express $\mathbf{v}$ (this

time in physical coordinates) in terms of the density contrast, δ, viz:-

$$\mathbf{v} = \frac{2}{3}\frac{f(\Omega)}{H\Omega} G.a(t).\rho_b(t) \int \frac{(\mathbf{x}' - \mathbf{x})\delta(\mathbf{x}',t)d^3x'}{|\mathbf{x}' - \mathbf{x}|^3} \qquad (1.8)$$

Conversely we can write down an expression for the density contrast in terms of the peculiar velocity field. In the linear regime this takes the simple form:-

$$\delta = -\frac{\nabla.\mathbf{v}}{Hf(\Omega)} \qquad (1.9)$$

The evolution of the density and velocity fields in the non-linear regime can be followed by numerical simulations. A reasonable approximate solution for quasi-linear perturbations may also be obtained by applying Zel'dovich's formalism (Zel'dovich, 1970) to describe the displacement of particles from the positions which they *would* have had in a homogeneous universe (see section 1.4.2 for more details).

1.3 Detections of Large-Scale Streaming Motions

In a pioneering study, Rubin *et al* (1973, 1976) used a sample of distant spiral galaxies - with recessional velocities in the range 3500 to 6500kms^{-1} - as a means of testing the isotropy of the Hubble Flow. The observed apparent magnitude of each galaxy was used to estimate its distance, which could then be compared with the inferred redshift distance to obtain a measure of the galaxy's peculiar velocity with respect to the Local Group. The authors detected a systematic

variation in the apparent magnitude of the sampled galaxies with position on the sky, which they interpreted as being due to a Local Group velocity of 450 ± 125kms^{-1} towards l = 163°, b = -11°, with respect to the sample of spiral galaxies. This velocity was much larger than expected; earlier studies (c.f. Sandage and Tammann, 1975a,b) had generally indicated a basically smooth and quiet Hubble Flow, and, moreover, later studies using other galaxy samples (Hart and Davies, 1982; de Vaucouleurs and Peters, 1984) failed to confirm the Rubin result. More recently Collins *et al* (1986) and Peterson and Baumgart (1986) have re-evaluated the Rubin data set using a better calibration of the galaxy distances, and again inferred a large Local Group velocity, although in a later paper Collins *et al* (1991) have shown that it is possible to wrongly infer streaming motions as a result of apparent magnitude selection effects present in the Rubin data set, as had been suggested by several authors some years before (c.f. Fall and Jones, 1976; see also chapter (2) for further discussion).

Further evidence for the existence of large peculiar velocities was found from analysis of the microwave background radiation. Measurements indicate a dipole anisotropy in the temperature distribution of the radiation on the sky which is consistent with a Local Group Motion towards l = 269°, b = 28° with respect to the background radiation (Lubin *et al*, 1983; Fixsten *et al*, 1983). Recent results from the COBE satellite report the magnitude of the dipole to be 550kms^{-1} (Smoot *et al*, 1991). Significantly the magnitude and direction of this velocity are different from those inferred from the Rubin galaxies. Hence, given that the streaming detected by Rubin is a real effect, one still requires a more complex model of the velocity

field to adequately explain the Local Group peculiar velocity inferred from the CMBR dipole. The inclusion of an "infall" velocity of ≈ 300kms^{-1} towards the Virgo cluster, for example (Aaronson *et al*, 1982), is insufficient to resolve the overall discrepancy. The Local Group velocity points not towards Virgo but rather in the direction of the Hydra-Centaurus supercluster (c.f. Tammann and Sandage, 1985; Lilje *et al*, 1986; Staveley-Smith and Davies, 1987) and the Virgo "infall" represents only a partial contribution towards the CMBR dipole. Recent studies have indicated the existence of a significant enhancement in the density field in the direction of Centaurus (c.f. da Costa *et al*, 1986) and suggest the possibility of a bulk motion on the scale of the entire Local Group and the Virgo cluster towards this region. Perhaps the most notable of these studies has been the work of Lynden-Bell *et al* (1988), which comprises a survey of more than 400 elliptical galaxies out to a redshift of ≈ 8000kms^{-1} for which photometric data were also available. Redshift-independent distance estimates were obtained to the galaxies via the D_n-σ relation (Terlevich *et al*, 1981) - an empirical relationship between the diameter and central velocity dispersion of ellipticals which is essentially a refinement of the earlier Faber-Jackson relation (Faber and Jackson, 1976) between luminosity and velocity dispersion. The basic principle behind the use of this relation is to infer an estimate of the intrinsic diameter of a galaxy from its measured velocity dispersion and then combine this with its observed angular diameter to infer its distance. Together with the Tully-Fisher relation (Tully and Fisher, 1977), which derives distances in an analogous manner from the relationship between the intrinsic luminosity of spiral galaxies and the width of their 21cm line profiles, the D_n-σ relation represents a powerful tool for obtaining

redshift-independent distance estimates to galaxies and clusters. The effects of selection bias on these relations have been a source of considerable controversy in the literature: it is a clarification of some important aspects of this debate which forms the central theme of this thesis. We will leave further discussion of the statistical issues involved until chapter (2).

To analyse their data Lynden-Bell *et al* constructed a specific dynamical model for the velocity field and then by combining their measured redshifts and D_n-σ distances they obtained maximum-likelihood estimates for the parameters of their model. They obtained the following results: the peculiar motions of the galaxies were 'best-fitted' by a velocity inflow model towards a "Great Attractor" centred on l = 307°, b = 9° in the direction of Centaurus at a distance of 4300 ± 350kms^{-1} in the Hubble Flow - in good agreement with the observations by da Costa *et al*. The excess mass of this concentration was calculated to be of the order of 5 × 10^{16} solar masses - comparable to the largest superclusters - in order to generate the inferred streaming motion of 570 ± 60kms^{-1} at the Local Group.

The details of the Lynden-Bell *et al* results are dependent on the parametric form of the model chosen for the velocity flow in the vicinity of the mass concentration. The model described above gave a better fit to the data than a earlier, simpler, model of a uniform streaming motion over and above the cosmological expansion (Dressler *et al*, 1987) which was in turn a better fit than unperturbed Hubble Flow, but clearly a wide variety of different parametric models

(several "attractors" at discrete locations, for example) are possible. In order to make much further progress in understanding the "Great Attractor" region one requires methods which are not tied to specific parametric models of the density and velocity fields. We will now discuss two such methods which set out - albeit with rather different approaches - to reconstruct the full density and peculiar velocity fields in a self-consistent manner.

1.4 Reconstructing the Density and Velocity Fields

1.4.1 'IRAS' Based Studies.

The first of these reconstruction methods has been adopted in the analysis of two redshift surveys of galaxies in the IRAS catalogue, and is appropriate when one does not have redshift-independent distance estimates. The two surveys are of similar size but are subject to rather different selection criteria: Strauss *et al* (1990) have measured redshifts of all non-stellar objects in the IRAS catalogue with 60μm flux greater than 1.9 Jansky. This gives a survey of over 2500 objects with redshifts less than $\simeq$ 3000kms^{-1}. Efstathiou *et al* (1990) - known as the 'QDOT' survey - probe rather deeper, down to a flux limit of 0.6Jy, but sample randomly only one galaxy in six, yielding a survey of around 2000 galaxies out to a redshift of $\simeq$ 7000kms^{-1}.

The principle of this method is to use the measured redshifts of the sampled galaxies to give a first indication of their distance (i.e.

assuming a 'quiet' Hubble flow) and thus obtain an initial estimate of the density contrast, $\delta(\mathbf{x},t)$, from smoothing the inferred 3-d distribution of IRAS galaxies. The peculiar velocity field is then calculated by assuming linear perturbation theory, so that equation (1.8) will hold, viz:-

$$\mathbf{v} = \frac{2}{3}\frac{f(\Omega)}{H\Omega} G.a(t).\rho_b(t) \int \frac{(\mathbf{x}' - \mathbf{x})\delta(\mathbf{x}',t)d^3x'}{|\mathbf{x}' - \mathbf{x}|^3} \qquad (1.10)$$

The inferred peculiar velocities can then be used to correct for the non-Hubble component in the measured redshift of each galaxy, and hence to obtain a better distance estimate. One can now repeat this procedure iteratively; i.e. compute a new estimate of the density field, use this to redetermine the peculiar velocity field, modify distance estimates again, and so on.

Recent results from the QDOT survey are reported in Kaiser *et al* (1991), in which the recovered density field is presented in the form of isodensity contours of δ. Examples of these results are shown in figure (1.3) taken directly from Kaiser *et al*. The "Great Attractor" density enhancement in Hydra-Centaurus is clearly detected, and is the dominant feature at the $\delta = 1.0$ level (Kaiser *et al* use the notation Δ instead of δ to denote the density contrast). Other notable features are the Virgo and Fornax clusters, at $\delta = 2.0$, and the Perseus-Pisces supercluster at $\delta = 0.7$.

There are several important points about the IRAS analysis. Firstly, the integral expression of equation (1.10) for the peculiar

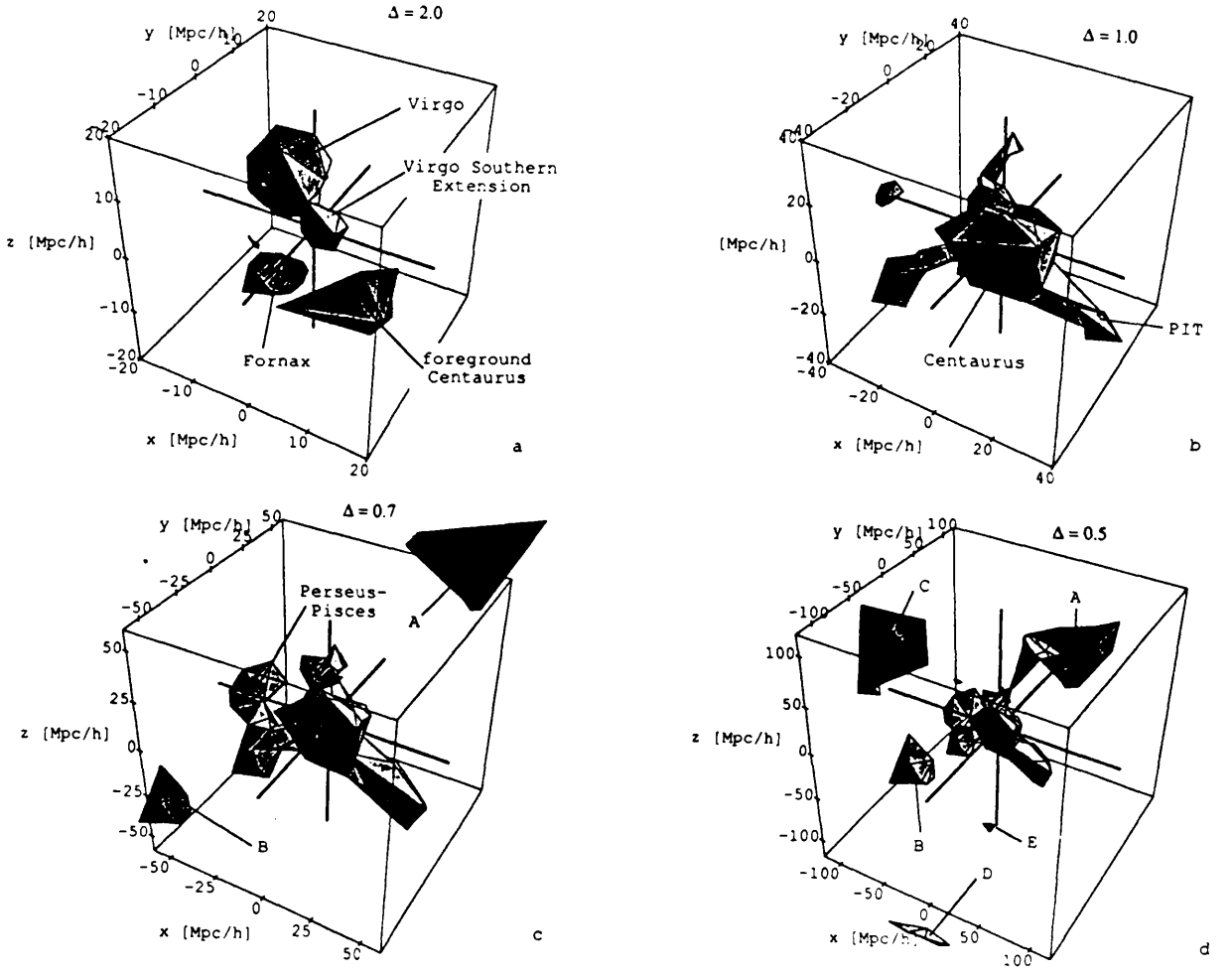

Figure (1.3)

Isodensity contour maps of the density contrast, Δ, recovered from analysis of the IRAS galaxies in the QDOT survey (from Kaiser *et al*, 1991). Prominent features are identified in the text above.

velocity should be taken over all space, but in practise is obviously limited to the volume sampled by the redshift survey. Furthermore, the sampling of galaxies will be increasingly sparse - and increasingly subject to systematic errors - close to the edge of the survey volume. This will serve to undermine the accuracy of the recovered velocity field.

Another, perhaps more serious, problem arises from the fact that one does not directly observe the density contrast, δ, but only the distribution of *luminous* matter (and, moreover, not the component of the light distribution due to ellipticals and early-type galaxies because of the infra-red nature of the IRAS survey). To resolve this problem one must make some assumption about how light traces mass. The standard approach - and that adopted in both IRAS surveys - has been to introduce a constant bias parameter, b, which relates the observed deviations in the galaxy number density to the underlying mass density contrast. Thus b is defined by:-

$$\left[\frac{\delta n}{n} \right]_{lum} = b \frac{\delta \rho}{\rho} \equiv b\delta \qquad (1.11)$$

This parameter is used to modify the density contrast in equation (1.10). Instead of assuming an *a priori* value for b, it can be determined self-consistently (or, more precisely, the product $b/f(\Omega)$ may be determined) as part of the iteration procedure. Latest results obtained from the QDOT survey (Kaiser *et al*, 1991) report a value of $b/f(\Omega) \simeq 1.16 \pm 0.21$, in general agreement with earlier results obtained from analysis of the Local Group Motion (Rowan-Robinson *et al*, 1991)

It would obviously be useful to avoid the assumption of a constant bias parameter - or at least to have some means of testing the validity of such an assumption. The POTENT method offers such a possibility.

1.4.2. The POTENT Method

A basic limitation in recovering the full peculiar velocity field is the fact that - even with reliable redshift-independent distance estimates - one can, in general, infer from the measured redshift of a galaxy only the *radial* component of each galaxy's peculiar velocity. The second reconstruction method which we now consider, labelled POTENT in BD, offers a neat resolution of this problem by making the fundamental assumption that the galaxy peculiar velocity field, $\mathbf{v}$, has zero vorticity and can therefore be expressed as the gradient of a scalar velocity potential, Φ, i.e.:-

$$\mathbf{v} = -\nabla\Phi \qquad (1.12)$$

It follows immediately from this assumption that the potential, Φ, at any point, $\mathbf{r}$, will be given in terms of a line integral of $\mathbf{v}$ from the observer to $\mathbf{r}$ - and that the value of this integral will be independent of the path taken to $\mathbf{r}$. In particular, therefore, we can evaluate Φ by computing the integral along a radial path to $\mathbf{r}$, thus requiring knowledge only of the radial component of $\mathbf{v}$. By differentiating Φ in the transverse directions we may then recover the other two components.

The assumption that the velocity field is irrotational - while not directly testable using POTENT alone - would appear to be reasonable. If the growth of density perturbations is indeed due to gravitational instability then it may be shown (see e.g. Peebles, 1980) that in the linear regime the only growing mode in the primordial peculiar velocity field is irrotational, while the rotational component decays as 1/a(t) as a consequence of angular momentum conservation. Certainly by the end of the linear regime, therefore, the growing mode would dominate and the velocity field would be a potential flow, with the velocity potential proportional to the gravitational potential. Kelvin's circulation theorem then ensures that the velocity field would remain irrotational after density perturbations become non-linear provided that the fluid trajectories do not cross. Even when such orbit mixing does occur, BD show that a suitably smoothed velocity field will remain irrotational to a good approximation well into the non-linear regime.

The main obstacles to the practical implementation of POTENT lie not with the validity of the fundamental assumptions, therefore, but with the limitations of the available data. In order to reconstruct the complete velocity potential one requires to know the radial component of the peculiar velocity field at every point; in practice one can estimate this only at the locations of a sparse sample of galaxies - and, moreover, the individual distance estimates to each galaxy have root mean square errors of typically 15 to 20%.

Given the radial velocities of a sparse and noisy sample, therefore, POTENT first smooths and interpolates the data onto a

spherical grid using a tensor window function to produce a smoothed radial velocity field, $v^*_r(r)$. The velocity potential at each point is then calculated from the integral:-

$$\Phi(r) = -\int_0^r v^*_r(r',\theta,\phi)dr' \qquad (1.13)$$

from which the reconstructed smoothed peculiar velocity field, **v**, is then recovered by differentiation, applying (1.12).

In the linear regime **v** can then be used to infer the density contrast via the simple relation given by equation (1.9), viz:-

$$\delta = -\frac{\nabla.\mathbf{v}}{Hf(\Omega)} \qquad (1.14)$$

In order to allow an effective recovery of δ into the non-linear regime POTENT uses the Zel'dovich formalism (Zel'dovich, 1970) which gives a good approximation to the evolution of mildly non-linear perturbations. As a result the actual application of POTENT - from inputted redshifts and distances through to the recovered density field - is rendered somewhat more complex than the fairly simple scheme outlined thus far.

The Zel'dovich formalism determines approximately the final (Eulerian) comoving position, **x**, at time, t, of a particle moving in the fluid perturbed by the density contrast, δ, by describing the displacement of **x** from the initial (Lagrangian) comoving position, **q**, which the particle would have had (and still would have!) in the absence of any density perturbations. The relationship between **x** and

q is assumed to take the form:-

$$\mathbf{x}(\mathbf{q},t) = \mathbf{q} + D(t)\Psi(\mathbf{q}) \qquad (1.15)$$

The nature of the approximation lies in the assumption that the displacement can be written in separable form as the product of a purely spatial perturbation function, $\Psi(\mathbf{q})$, and a purely time-dependent growth factor, D(t). Beginning from this equation, similar relations can be determined between the Lagrangian and Eulerian peculiar velocity and density contrast, subject to the Zel'dovich approximation, and one can define an inverse mapping (again after smoothing, if necessary, to overcome orbit mixing) relating the Eulerian values of these fields to their initial Lagrangian values.

In short, therefore, POTENT assumes that the Lagrangian peculiar velocity field is irrotational, so that Φ and **v** can be recovered using equation (1.12) and (1.13) expressed in Lagrangian coordinates. The Zel'dovich approximation is used to move from the observed, Eulerian, radial velocities to their Lagrangian counterparts and then - after reconstruction of the velocity and density fields - back to Eulerian coordinates.

The accuracy of the Zel'dovich approximation has been tested against the true density contrast in the quasi-linear regime evaluated numerically for a series of CDM N-body simulations (Dekel *et al*, 1990, hereafter DBF; Dekel, 1991). The rms percentage error in δ was found to be less than 10% over the range $-0.8 \leqslant \delta \leqslant 4.5$, so that the usefulness of the approximation would seem to be clear. The Zel'dovich formalism can also be applied to make non-linear corrections to the

density field estimated by the IRAS method, thus improving the prediction of the peculiar velocity field by that method (Dekel *et al*, 1992)

The results obtained by POTENT are striking. Figure (1.4) shows the recovered density field in the supergalactic plane constructed from a sample of 973 objects; 544 ellipticals and 429 spirals (Bertschinger *et al*, 1990). The dominant recovered feature is the extended 'hump' on the left of this diagram, in the Hydra-Centaurus region, which broadly confirms the "Great Attractor" detection of Lynden-Bell *et al* (1988). The peak density contrast in this region is given by $\delta \simeq 1.2 \pm 0.4$ with a Gaussian smoothing window of radius $\simeq 1400 kms^{-1}$ - consistent with the QDOT results. (This result is for $\Omega = 1$ and scales approximately with $\Omega^{7/4}$). Both the Virgo cluster and Local Group are found to be falling toward the bottom of the Great Attractor potential well with peculiar velocities exceeding $500 kms^{-1}$. Several large regions of below average mass density are also recovered which match known voids in the galaxy distribution.

One of the most encouraging aspects of these results is the fact that the same broad-based features are recovered as those derived from the IRAS, and earlier, surveys - but this time without the need for an *a priori* assumption about how mass traces light. In POTENT galaxies are used as tracers of the large-scale velocity field and not the density field. Consequently, in the long term POTENT offers a means of directly determining the relationship between the distribution of luminous 'tracer' galaxies and the underlying mass

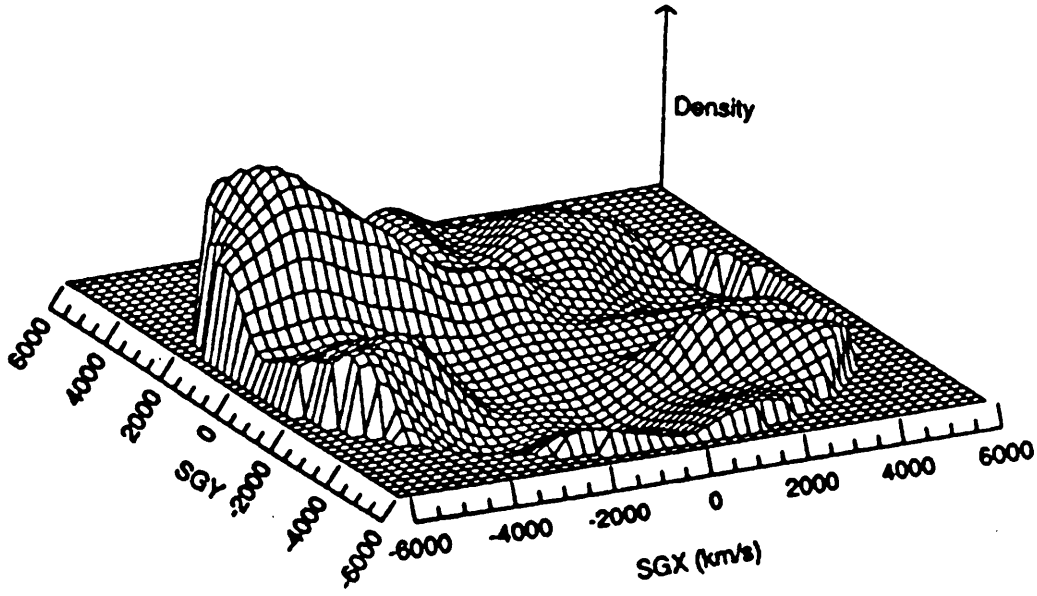

Figure (1.4)

Density contrast recovered by POTENT in the supergalactic plane from a sample of 973 galaxies (taken from Bertschinger *et al*, 1990)

density field, so as to test the validity of a universal linear bias parameter. Preliminary estimates of b from a comparison of POTENT results with optical data seem to favour a value of order unity (Dekel, 1991) consistent with the QDOT results, although no firm conclusions have yet been reached.

Another notable feature of POTENT compared with the IRAS studies is the fact that the recovery within a given volume is not so adversely affected by sparse and noisy sampling outside of that volume: for POTENT the effects of poor sampling are essentially local so that the better the data coverage is within a given region, the better will be the reconstruction in that region. A good example of this is the Perseus-Pisces supercluster which, although clearly identified in the QDOT results, is practically invisible in the density recovery presented in DBF. This region is very sparsely sampled, however, in the data set of 973 galaxies used in DBF. Dekel (1991) reports that the addition of a new sample of spirals covering the Pisces region results in a reconstructed density contrast of $\delta \simeq 1.0 \pm 0.4$ in that region, although the recovery in the as yet poorly sampled Perseus region remains dominated by noise.

In DBF great care is taken over the treatment of sampling errors. The authors recognise that their smoothing procedure can result in a "sampling gradient bias", in which the sampled velocity field from regions of high density pollutes low density regions within the same smoothing volume and thus artificially enhances the density in those regions. They set out to minimise this effect by adopting a system of volume weighting and adjustable smoothing radii which takes

account of the local density in defining the smoothing window.

The sparseness of the sample, combined with significant errors on individual distance estimates are still regarded as the major sources of error, however. Dekel (1991) summarises the status of errors in the recovered density and velocity fields - as assessed from Monte-Carlo simulations. The distance errors were modelled as normally distributed with standard deviation of order 15%, which is the typical size of Tully-Fisher and D_n-σ errors. In the well sampled regions of the POTENT data set the rms errors are less than 250kms^{-1} in the recovered velocity field and less than 0.2 in the density contrast; in more poorly sampled regions these errors exceed 1000kms^{-1} and 1.0 respectively. This clearly indicates the penalties of sparse and noisy sampling.

In providing a brief overview of our current picture of the local universe it is no accident that we have described in some detail the POTENT method for reconstructing the density and velocity fields. POTENT has given a new relevance to determining redshift independent distance estimates to galaxies; moreover the opportunity to improve the quality of the early POTENT results provides a strong motivation for finding ways to improve those distance estimates and remove or reduce systematic errors. It is precisely such an improvement which is the goal of this thesis.

2. DISTANCE INDICATORS AND SELECTION BIAS

2.1 Introduction

In this chapter we will discuss various galaxy distance indicators which have been used in the literature and examine the attempts which have been made to understand and remove the systematic bias in these indicators introduced by observational selection effects.

Methods of measuring the distance to a galaxy which do not make use of the observed redshift have traditionally been classified into 2 groups, denoted as *primary* and *secondary* distance indicators. Primary indicators are methods which can be calibrated from purely theoretical considerations or from distance measurements made within our own galaxy; to calibrate secondary indicators requires first determining the distances to a representative sample of nearby galaxies by some other means (e.g. using primary indicators). An overview of the historical development of different indicators, and the astrophysical principles on which they are based, is given in, e.g., Rowan-Robinson, (1985). Examples of primary indicators are Cepheid variables (c.f. Martin *et al*, 1979) and supernovae (c.f. Kirschner and Kwan, 1974). Secondary indicators include the Tully-Fisher and D_n-σ relations introduced in chapter (1), galaxy HII regions (c.f. Sandage and Tammann, 1974) and various colour-luminosity relations (c.f. Tully *et al*, 1982; Michard, 1979).

The supernovae method has the greatest potential of the

primary indicators because supernovae are in principle observable to enormous distances ($\simeq$ 1000Mpc), although in practice there are still considerable difficulties in the theoretical understanding of their characteristics (c.f. Rees and Stoneham, 1982). With the exception of supernovae, however, as the distance scale has expanded in recent years we have come to rely more and more on secondary indicators as the 'first line of attack' in determining galaxy distances; moreover, of such indicators the Tully-Fisher and D_n-σ relations have proved the most commonly used, as is evidenced by the POTENT data set discussed in chapter (1). The prevalence of secondary indicators is unlikely to change in the near future. The observable limit for Cepheids, for example, is only about 5Mpc from ground based instruments; hence this primary indicator - despite being probably the most securely calibrated - cannot be used on larger scales. It is hoped that the Hubble Space Telescope will push back this limit to around 20Mpc, but estimates for more distant galaxies will still require secondary methods. (Cepheid observations with HST should, nevertheless, susbstantially improve the calibration of secondary indicators.)

Another useful means by which to discriminate between different indicators is the number of observable properties of a galaxy on which each depends. To explain what we mean by this consider, for example, the Tully-Fisher relation. As we remarked in chapter (1), one uses this relation to infer an estimate of the absolute magnitude, M, of a spiral galaxy from its measured 21cm line width, which has been found to be well-correlated with M. One then combines this estimate with the measured apparent magnitude, m, to obtain a distance

estimate by simply inverting the usual magnitude-distance relation, viz:-

$$m = M + 5\log r + 25 \qquad (2.1)$$

where the distance, r, is measured in Mpc. In fact it is more common in the literature to determine an estimate of the distance *modulus*, m-M, which we can see from equation (2.1) is essentially an estimate of logr. Moreover, previous discussions of the statistical properties of distance indicators have also dealt predominantly with estimates of the distance modulus. In this thesis it will frequently be unnecessary to discriminate between the two terms and, at such times, we will use them interchangeably. As we will see in chapters (3) and (4), however, there are situations when a distinction becomes important, and when this is the case we will we clearly indicate whether referring to distances or distance moduli.

Thus, a distance estimate constructed from the Tully-Fisher relation is a function of *two* observables; line width and apparent magnitude. Similarly distances derived from the D_n-σ relation are a function of observed velocity dispersion and apparent angular diameter.

As an aside one should note that in equation (2.1) we assume no absorption either within our own galaxy or internally, within the observed galaxy. Such an assumption is often unreasonable so that to arrive at equation (2.1) one must first make careful corrections to both m and M. Indicators which are less badly affected by obscuration have clear advantages; indeed it was precisely to avoid such extinction

corrections that prompted Aaronson *et al* (1980) to propose that the Tully-Fisher relation should be calibrated using infra-red magnitudes. Not only does this virtually eliminate extinction from the measurements for each galaxy, but it also allows greater sky coverage for galaxy surveys since one can still make useful IR observations at low galactic latitudes.

Later in this chapter we will return to the Tully-Fisher and D_n-σ relations and examine the importance of observational selection effects for their calibration and use, as previously discussed in the literature. As a preliminary to this, however, it is useful first to consider the effects of selection bias on indicators which are functions of only *one* observable: more specifically, distance estimates obtained from only the observed apparent magnitude of a galaxy, having adopted *a priori* a fiducial value for its absolute magnitude.

Such indicators assume the existence of 'standard candles' - i.e. a class of galaxies all of which have identical intrinsic luminosity - and their use was the approach generally adopted in early studies of galaxies and the Hubble flow in the 1970s (c.f. Rubin *et al*, 1976; Sandage and Tammann, 1975a,b). While the standard candle assumption is obviously an idealisation, a number of specific galaxy types have been proposed as good candidates for standard candles because of the small scatter in their intrinsic luminosity; these include ScI galaxies (Sandage and Tammann, 1975b) and first ranked cluster ellipticals (Sandage and Hardy, 1973). We will comment further on the properties of these specific galaxy types in due course.

2.2 Standard Candles and Malmquist Bias

In obtaining distance estimates - from whichever indicator - the cosmologist would ideally wish to use a *volume* limited galaxy sample; i.e. one in which every object within a given volume is observed. In practice, however, galaxy samples are more generally *magnitude* limited; i.e. one observes only those galaxies with apparent magnitude brighter than some limiting magnitude, m_L. It has long been recognised that *any* spread in the luminosity of standard candles would introduce Malmquist bias in a magnitude limited sample. this bias arises because galaxies of different intrinsic luminosities are sampled within different volumes: at greater distances only progressively more luminous galaxies are observable as fainter objects 'fade out'. Consequently the mean luminosity of observable galaxies increases with distance, and a failure to account for this effect results in galaxy distances being systematically underestimated. Eddington (1914) was one of the first authors to study the statistical biases which arise when objects are selected using a distance dependent observable such as apparent magnitude. Indeed the effect described above bears his name, the Eddington correction, in some references (c.f. Feast, 1987). Malmquist (1920) gave a classical discussion of the effect in a stellar context and the bias is most commonly referred to by his name.

Malmquist showed that if the luminosity distribution of standard candle galaxies were a gaussian, with mean M_0 and dispersion σ, then the mean absolute magnitude, M_*, of *observable* galaxies, assuming a constant spatial number density, in a magnitude limited sample is:-

$$M_* = M_0 - 1.38\sigma^2 \qquad (2.2)$$

Many later authors have rederived this result and extended the analysis of Malmquist in a cosmological context. One analysis of note is in Teerikorpi (1975), in which the author showed that - subject to the same assumptions - the mean absolute magnitude of galaxies in a shell at distance, r, is given by:-

$$M(r) = M_0 - \frac{\sigma}{\sqrt{2\pi}} \cdot \frac{\exp(-\frac{1}{2}(m_L - 5\log r - M_0 - 25/\sigma)^2)}{\Phi(m_L - 5\log r - M_0 - 25/\sigma)} \qquad (2.3)$$

where Φ denotes the cumulative normal distribution of mean zero and unit variance. Figure (2.1) shows a graph of M(r) for various values of σ and assuming $M_0 = -20$. We can see from these graphs that the Malmquist bias is negligible at very small distances but increases steadily with increasing r beyond a certain distance. The slope of the bias curves is similar for each value of σ; however, we see that as σ increases the distance beyond which the bias is non-negligible becomes progressively smaller. Consequently the bias at any *given* distance more severe for larger values of σ. For example, for r = 100Mpc and $\sigma = 1$ the mean magnitude is more than 1.2 magnitudes brighter than M_0, while the bias is less 0.4 mag when $\sigma = 0.3$.

2.2.1 The Hubble Diagram and the Minimum Bias Subset

A useful way to illustrate how Malmquist bias arises is via the Hubble diagram - a plot of log(redshift) versus the apparent magnitude of a survey of galaxies - which has traditionally been used

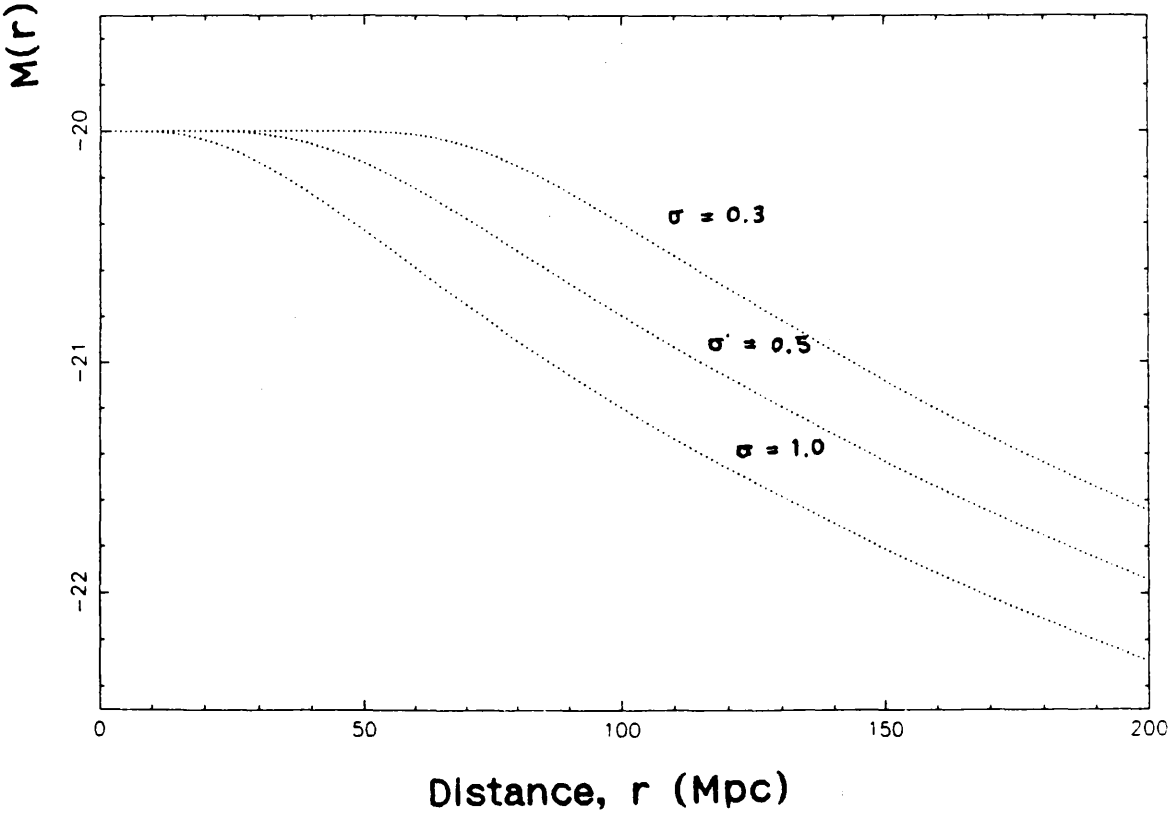

Figure (2.1)

Mean absolute magnitude, M(r), at distance,r, for observable galaxies in a sample 'cut-off' by an apparent magnitude limit, m_L, shown for different values of σ. (M_0 = -20).

to estimate Hubble's constant and test for isotropy in the Hubble flow. In principle the Hubble diagram may also be used to determine the deceleration parameter, q_0, although such attempts have generally proved unsuccessful (c.f. Kristian *et al*, 1978).

Figure (2.2) shows a schematic example of a Hubble diagram for a sample with a sharp upper magnitude limit. If one assumes that the Hubble Flow is uniform, then the observed recessional velocity and apparent magnitude of each galaxy in the sample will be related by the following equation (neglecting absorption):-

$$\log V = 0.2m + \log H_0 - 0.2M_* - 5 \qquad (2.4)$$

In this ideal case, therefore, a galaxy of a given absolute magnitude, M_* say, will lie on the straight line of gradient 0.2 which intercepts the logV axis at $\log V = \log H_0 - 0.2M_* - 5$. If the luminosity function of the galaxies is a gaussian, with mean M_0 and dispersion σ, then one would expect that 99% of galaxies would lie between the two bold diagonal lines in figure (2.1), which correspond to absolute magnitudes of $M_0-3\sigma$ and $M_0+3\sigma$ respectively. *Observable* galaxies sampled at a given distance (i.e. at a given recessional velocity) would therefore lie along a horizontal line in the Hubble diagram, between the bold diagonals and to the left of the apparent magnitude limit, m_L.

We can easily see from figure (2.2) that at larger recessional velocities it will not be possible to observe all galaxies with absolute magnitudes down to $M_0+3\sigma$, as a progressively larger part of the luminosity function will lie in the 'unobservable' region. Consequently, the mean observed magnitude of galaxies in the region denoted 'biased' will *not*

equal the mean, M_0, of the luminosity function, but will decrease monotonically in this region with increasing distance - the precise distance dependence being given by equation (2.3).

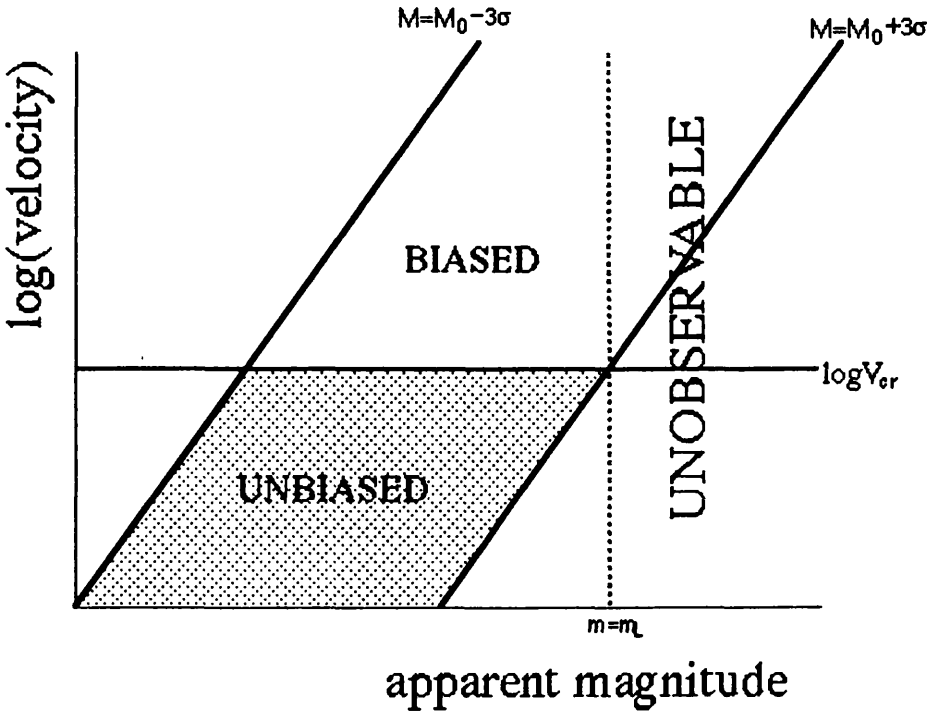

Figure (2.2)

Schematic Hubble diagram for a galaxy sample with a sharp upper apparent magnitude limit, m_L, demonstrating the presence of Malmquist bias and the existence of an unbiased region - the 'Minimum Bias Subset'.

In a study of the local velocity field using nearby spiral galaxies, Sandage and Tammann (1975b) illustrate clearly a simple method for dealing with Malmquist bias, by identifying the region of the Hubble diagram where it is not significant and restricting their sample to only those galaxies lying in this region; thus constructing what has been termed in the literature as a 'minimum bias subset' (MBS) or 'bias free subset'.

We can see from figure (2.2) that there will be a critical distance r_{cr}, (corresponding to the bold horizontal line at $\log V_{cr}$) at which a galaxy of absolute magnitude $M_0+3\sigma$ would be observed at the magnitude limit, m_L. One would expect to be able to observe 99% of all galaxies which lie at distances less than r_{cr}, and so the shaded 'unbiased' region of the Hubble diagram will be 99% complete. It represents the region in which the Malmquist Bias is negligible; i.e. the 'minimum bias subset'. In practice, even without knowledge of the mean absolute magnitude or variance, one may still construct an estimate of the MBS geometrically by drawing lines of gradient 0.2 through each galaxy's position on the Hubble diagram and finding the *minimum* value of logV at which these lines intersect the line of the magnitude limit, $m=m_L$. Expressing this algebraically we have:-

$$\log V_{cr} = \min \{ \log V_i + 0.2(m_L - m_i)\} \qquad (2.5)$$

where the minimum is taken over all the galaxies in the sample. Having thus constructed an MBS one may then reasonably assign the mean absolute magnitude, M_0, to all the galaxies in the MBS. It then follows that $m_i - M_0$ will be an unbiased estimate of the distance modulus of the ith galaxy in the MBS. The approach adopted by Sandage and

Tammann (1975b) is now to define an estimate of $\log H_0$ as follows:-

$$\log H_0 = \frac{1}{n} \Sigma(\log V_i - 0.2m_i) + 0.2M_0 + 5 \qquad (2.6)$$

from which an estimate of H_0 is then obtained in the obvious way. The summation is over all the galaxies in the MBS. Sandage and Tammann point out that equation (2.6) defines an unbiased estimate of $\log H_0$, and observe that consequently the corresponding estimate of H_0 will, in fact, be biased - although they regard this bias as a small effect.

2.2.2 The MBS With a Narrow Magnitude Window

It is frequently the case that galaxy samples are selected not just by an upper but also a *lower* magnitude limit. For example, the data set of ScI galaxies used by Rubin *et al* (1973, 1976) was selected from the Zwicky catalogues (Zwicky *et al*, 1961-68) to have apparent magnitudes between 14 and 15. Recall from chapter (1) that a Local Group peculiar velocity of $450 kms^{-1}$ was claimed with respect to this sample.

The effect of Malmquist bias on a Hubble diagram constructed with both an upper and lower magnitude limit is somewhat different from that of an upper limit only: specifically one cannot define a complete unbiased sample out to some limiting distance since there is a *positive* Malmquist bias (i.e. the mean absolute magnitude is fainter than M_0) at small distances due to the exclusion of over luminous galaxies. In such a case one can still identify a range of redshifts

(again assuming 'quiet' Hubble flow) within which the luminosity function is best sampled; in this way Rubin *et al* defined an MBS for their data in the redshift range 3500kms^{-1} to 6500kms^{-1}, and inferred the Local group motion by 'best-fitting' a dipole to the 96 galaxies in their sample with redshifts between these limits.

Several authors (c.f. Fall and Jones, 1976; Sandage and Tammann, 1975b; Collins *et al*, 1991) have discussed specific problems which arise when the magnitude selection window is narrow, such as is the case for the Rubin data. In particular, Collins *et al* have shown via a Monte-Carlo analysis that a narrow selection window will result in the mean absolute magnitude of observable galaxies being strongly dependent on redshift; we can see this qualitatively from the schematic Hubble diagram shown in figure (2.3), which has a narrow magnitude selection window between m_1 and m_2. The shaded region represents an MBS for the sample in the sense that, within this velocity range - V_1 to V_2 as shown - the magnitude window samples galaxies with absolute magnitudes close to M_0. Note, however, that the mean absolute magnitude of observable galaxies at $\log V_1$ is substantially different from that at $\log V_2$.

Consequently if the redshift distribution of the sampled galaxies were strongly correlated with direction - due to clustering in one or more regions of the sky, for example - then there would be a corresponding systematic correlation between the absolute magnitude and direction of the sampled galaxies. This may lead to an apparent systematic anisotropy in the inferred peculiar velocity distribution which would be suggestive of a streaming motion of our galaxy with

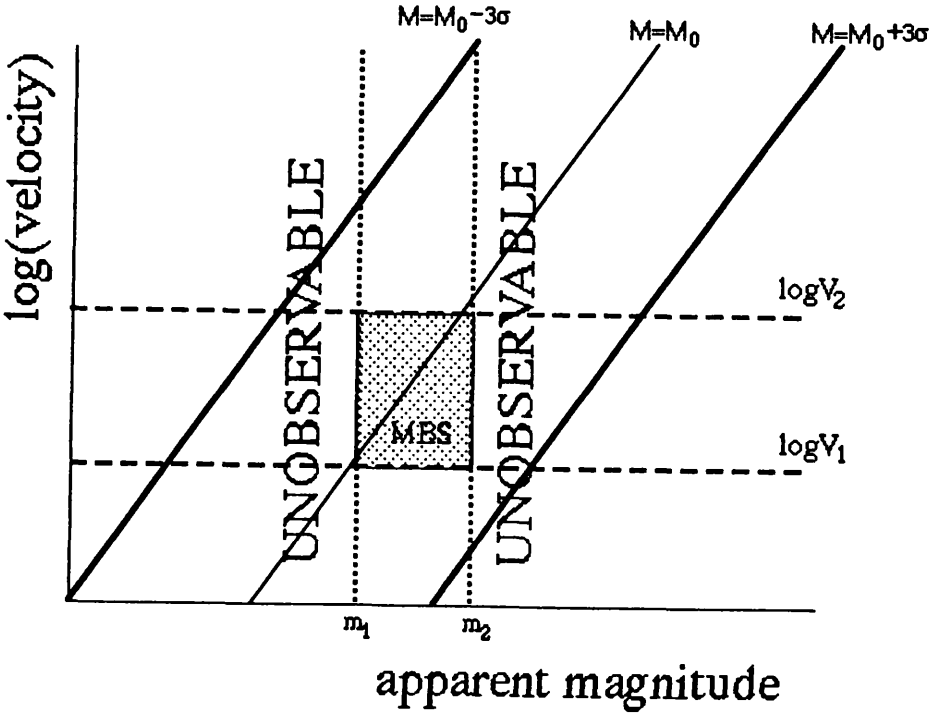

Figure (2.3)

Schematic Hubble diagram for a galaxy sample with a narrow apparent magnitude selection window, demonstrating the correlation between mean absolute magnitude and redshift - even within the MBS.

respect to the sample, but which would be entirely due to the selection procedure. Thus, there seemed good reason to believe that the MBS used by Rubin was not free from selection bias, and hence the evidence for a Local Group motion was not conclusive. In order to avoid this problem Collins *et al* stress the importance of using wide apparent magnitude selection criteria in obtaining samples, in which case it is at least possible to define a subset of galaxies which is complete to some limiting redshift, although the sample will be affected by Malmquist bias beyond this limit.

Nevertheless a more fundamental drawback in the use of the MBS will still exist even in this case: the assumption inherent in its definition of 'quiet' Hubble flow. Deviations from uniform Hubble Flow will distort the linear form of the magnitude - logV relationship on the Hubble diagram and will cause more distant galaxies to be misplaced into the 'unbiased' region and vice versa, so that the true unbiased nature of the MBS will be compromised. Even if one were to accept that such peculiar motions will be less significant at higher redshifts as the Hubble Flow becomes more uniform, these more distant galaxies would be excluded from the MBS because their redshift places them in the 'biased' region of the Hubble diagram. These limitations in the use of the MBS are of particular concern since it precisely to identify systematic peculiar motions that the MBS has often been used.

A more promising approach to dealing with the Malmquist bias of distance estimates derived from magnitudes alone would be to adopt a corrected fiducial value for the absolute magnitude: this would at least attempt to address the variation of mean absolute magnitude with

distance, as described by equation (2.3), and would allow more distant galaxies - which would be excluded from the MBS - to be used effectively. We will explore this approach in chapter (3), as a means of introducing our statistical formulation for studying the properties of distance estimators.

2.3 Distances Derived From Two Observables

In the past fifteen years a considerable observational and theoretical effort has led to the identification of a number of relationships between different intrinsic physical characteristics of galaxies which have proved extremely valuable for the determination of galaxy distances. To provide some background on this work we now list the most closely studied of these relations, with appropriate references which describe them in more detail.

2.3.1 Correlations Useful as Distance Indicators

(1) Tully-Fisher Relation

This is a correlation discovered by Tully and Fisher (1977) between the absolute magnitude of spiral galaxies and the width of their radio emission line at 21cm due to neutral atomic hydrogen, a quantity readily observable in external galaxies using large radio telescopes. The correlation found by Tully and Fisher took the form:-

$$M_{pg} = A\log(W_0/\sin i) + B \qquad (2.7)$$

where M_{pg} is the absolute photgraphic magnitude corrected for inclination and extinction effects, W_0 is the 21cm line width expressed in kms^{-1}, i is the inclination angle between the normal to the plane of the galaxy and the line of sight, and A and B are constants.

As we have already remarked, Aaronson *et al*, (1980) have examined this correlation using infrared magnitudes measured at 1.6μm and find the same linear form. Several authors (c.f. Roberts, 1978; Rubin, 1983) have suggested that the slope of the relation derived from photographic magnitudes is quite sensitive to the spiral galaxy type. Aaronson *et al* on the other hand claim that the infra-red form shows no such dependence, although this has been challenged in Burstein (1982).

(2) Colour-Luminosity Relation

This correlation was first proposed by Baum (1959), who observed that more luminous early-type systems appeared redder. A more precise treatment was carried out by Sandage (1972) which established a linear relation between absolute V magnitude and U-B colour for early-type systems in the Virgo and Coma clusters, viz:-

$$U-B = AV + B \qquad (2.8)$$

where A and B are constants (though, of course different from those of equation 2.7). Visvanathan and Sandage (1977) extended the analysis to other colors and found similar linear relationships.

(3) Faber-Jackson Relation

This correlation, first identified in Faber and Jackson (1976), takes the form of a power law between the luminosity and velocity dispersion, determined from the doppler broadening of optical line profiles, of elliptical galaxies which, when expressed in terms of magnitudes is again linear in form:-

$$M = A\log\sigma + B \qquad (2.9)$$

where σ is the velocity dispersion in kms^{-1}. de Vaucouleurs and Olsen (1982) have also derived a linear relation between M and σ for lenticular galaxies.

(4) D_n-σ Relation

Terlevich *et al*, (1981) deduced from a sample of 24 ellipticals this correlation between the apparent angular diameter and the central velocity dispersion. The form derived was again a power law, as for Faber-Jackson, but represented a considerable improvement over the latter, with a factor of two less scatter. The derived relation may be written as:-

$$\log D_n = A\log\sigma + B - \log r \qquad (2.10)$$

where D_n is the angular diameter, defined precisely and objectively to the same isophote for galaxies at all distances after applying absorption and redshift K corrections, and r is the distance of the galaxy. Note that equation (2.10) essentially expresses a power law relation between σ and the intrinsic physical diameter, d, since we have $D_n = d/r$. Recent evidence (c.f. Giuricin *et al*, 1989) comparing

clusters at different distances indicates that the power law constant of proportionality, B, is indeed a universal constant to within the limits testable by current data.

(5) Diameter-Luminosity Relation

Holmberg (1969) determined the correlation between the intrinsic physical diameter, d, and absolute photographic magnitude, M, of different classes of galaxies. A linear relation was again found, viz:-

$$\log d = AM + B \qquad (2.11)$$

This relation was refined by later analysis (c.f. Paturel, 1979) employing careful corrections to those diameters which had been determined according to different isophotal measurement systems.

(6) HII Regions

Sandage and Tammann (1974) describe a correlation between the absolute photographic magnitude of late-type spirals and irregulars and the size of their HII regions. they derived the following specific relation:-

$$\log(D_H, D_C) = AM + B \qquad (2.12)$$

where (D_H, D_C) is the average of the core and halo diameters of the three brightest HII regions in each galaxy. The procedure adopted by Sandage and Tammann involved a rather subjective method for measuring the diameters on a photographic plate; Kennicutt (1979) has repeated the analysis using objective isophotal diameters.

2.3.2 Calibration by Linear Regression: Malmquist Bias

Determining the constants, A and B, in each of the above relations has been the subject of intense study - and considerable debate - in the literature. The straight line is generally 'best-fitted' by performing a linear regression on a calibrating sample of galaxies (e.g. a nearby cluster) whose distances - and hence absolute magnitudes or diameters - can be determined by some other method. A key issue in discussions has been the question of which linear regression is most appropriate for determining A and B.

Consider for example the Tully-Fisher relation of equation (2.7). In their original presentation of this indicator Tully and Fisher (1977) derived a best-fit straight line by regressing absolute magnitudes on line widths, i.e. assuming line width to be error-free. (In fact Tully and Fisher achieved this fit purely by 'eyeballing' the data and adjusting their line to minimise visually the magnitude residuals; of course this line may be found less subjectively using formulae given in any elementary statistics textbook.)

This regression line determines the mean absolute magnitude at a given line width, W_0, as a linear function of $\log W_0$ (after inclination corrections) viz:-

$$E(M|W_0) = A\log W_0 + B \qquad (2.13)$$

The important issue is that the slope, A, and zero point, B, derived from *other* regression lines are different from those determined by this regression. Consequently the distances inferred by each indicator

depend on one's choice of regression line.

The question of which regression line is 'best' is non-trivial, particularly when one must take account of observational selection effects. When these are present, in general a Malmquist bias is introduced, analogous to the Malmquist bias already discussed for standard candles, which results in systematic errors in distances estimated from each indicator. It is important to note that this bias is distinct from other systematic effects to which the calibration process may be prone: in particular zero-point errors introduced when the distances of the calibrating galaxies are inaccurately determined (c.f. Tammann, 1987). The existence of Malmquist bias means that even for a perfect calibrating sample, distances estimated to more remote galaxies may still be systematically in error if one uses an inappropriate regression on the calibrators.

Four different regression lines have been considered in the literature: these are (defined as for the Tully-Fisher relation, with the obvious corresponding definitions for other indicators):-

(1) Regression of magnitudes on line widths.

(2) Regression of line widths on magnitude; i.e. treating *magnitude* as an error-free variable (c.f. Schechter, 1980)

(3) Orthogonal regression, accounting for errors on both observables (c.f. Giraud, 1987)

(4) 'Bisector' regression; i.e. the line which bisects regression lines (1) and (2) (c.f. Pierce and Tully, 1988)

As a typical example, figure (2.4) shows the Tully-Fisher relation illustrated for a combined sample of spirals from the Virgo and Ursa Major clusters (Tully, 1988). The dashed line denotes the regression of luminosity on line width, i.e. (1) above, while the steeper solid line denotes regression (2), line width on luminosity. The orthogonal and bisector regression lines, if drawn, would lie between these two. Although we can see that the difference in the slope of these lines is fairly small for these data, which are quite well correlated with correlation coefficient, $\rho = -0.8$, the difference in slope increases sharply as the intrinsic scatter in the relation increases. (Note that ρ is negative since an increase in luminosity corresponds to a *decrease* in absolute magnitude.)

We can illustrate this Malmquist bias - again using the Tully-Fisher relation as an example - in a manner similar to the Hubble diagram representation described in section (2.2). This schematic picture is similar to the treatment given in e.g. Tully, 1988 and Lynden-Bell *et al*, 1988).

Figure (2.5a) shows a schematic plot of absolute magnitude versus log line width (assumed corrected for inclination). The shaded area indicates the set of possible values of M and W_0, given the intrinsic scatter of the Tully-Fisher relation. (More precisely, the shaded area indicates the region within which 99%, say, of galaxies would be *expected* to fall - analogous to the diagonal lines which

bounded a 99% confidence region of the Hubble diagram in the previous section.)

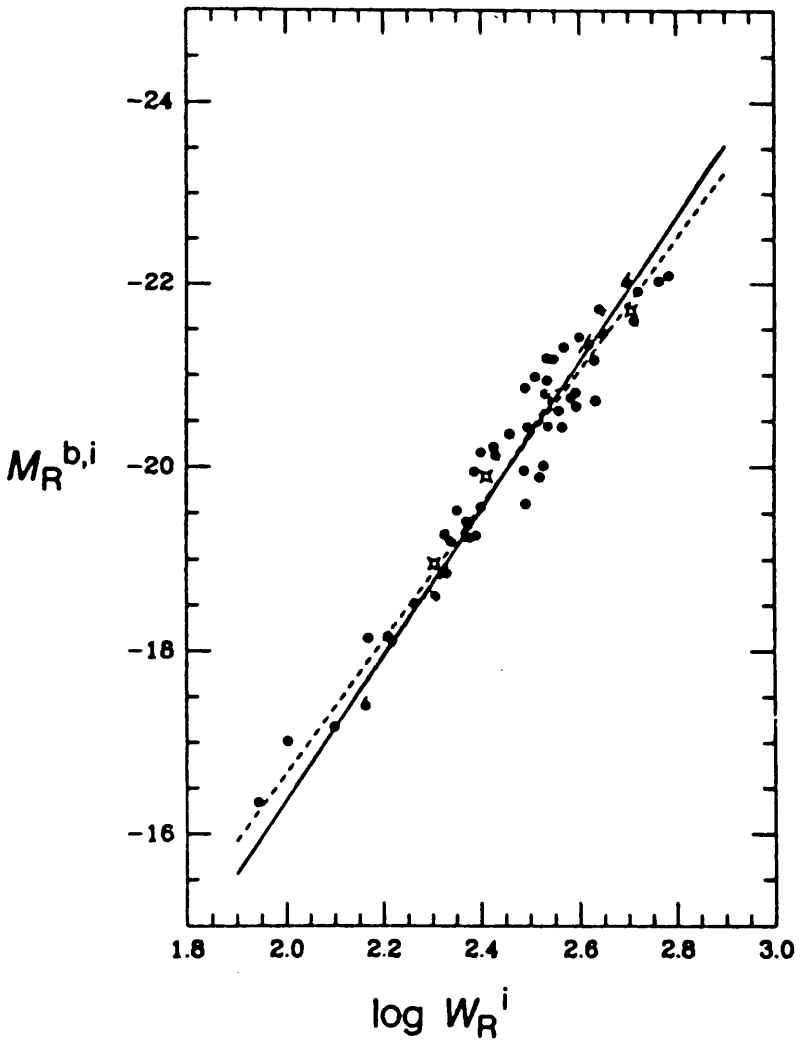

Figure (2.4)

Example of the Tully-Fisher relation derived for a sample of spirals from the Virgo and Ursa Major clusters. The dashed line is obtained by regressing luminosities on line widths; the solid line by regressing line widths on luminosities. Correlation coefficient for the data, $\rho \simeq -0.8$

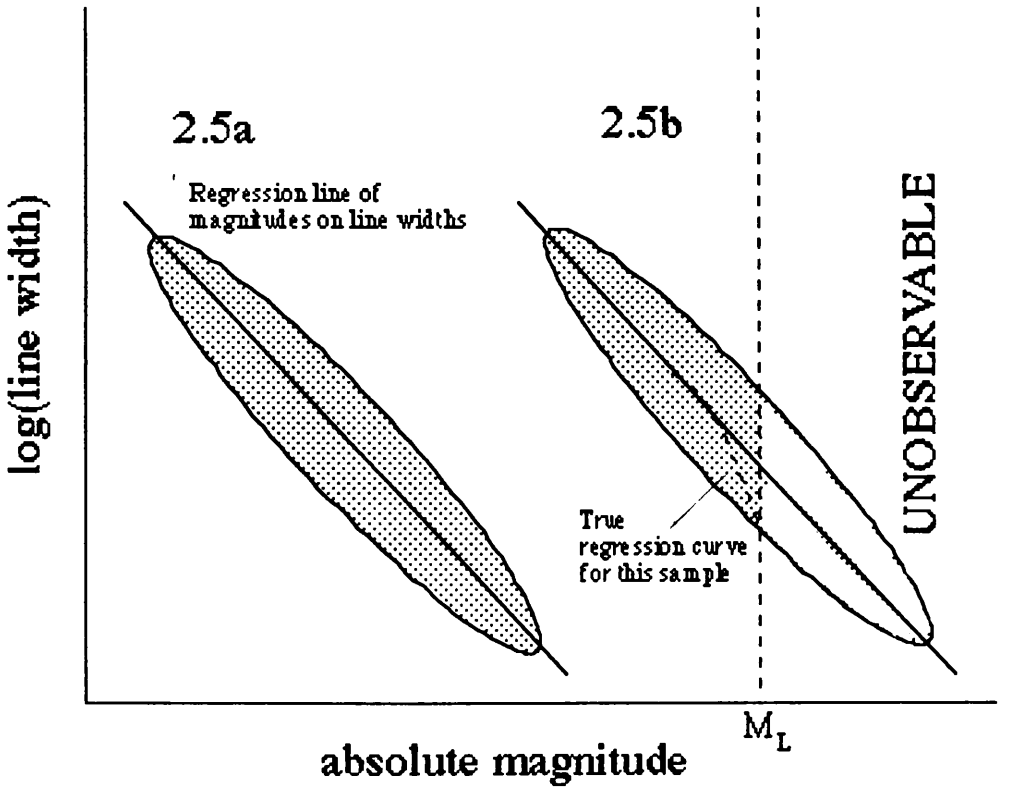

Figure (2.5)

Schematic Tully-Fisher diagrams demonstrating the effects of Malmquist bias on the calibration of the relation by linear regression. The diagrams are explained in the text.

In other words, this region represents the range of values of M and W_0 which one could expect to observe in, for example, a completely sampled nearby cluster as could be used to calibrate the relation. Thus, the straight line shown in figure (2.5a) is the regression line, $E(M|W_0)$, obtained by regressing magnitudes on line widths which may then be used to estimate the luminosity of a more distant galaxy by reading off the point of intersection with the regression line, as shown.

In figure (2.5b), on the other hand, we introduce an apparent magnitude limit, m_L, and demonstrate its effect on the Tully-Fisher relation. In this figure the shaded region indicates the range of *observable* values of M and W_0 for galaxies at some given distance, r, after accounting for the magnitude limit. (It is convenient to consider galaxies at the same distance since the apparent magnitude limit then translates directly to a limiting absolute magnitude, $M_L = m_L - 5\log r - 25$.)

At this point it is important to note that we can regard the shaded region of figure (2.5b) in two different - but entirely equivalent - ways: either as representing the *actual* spread in observed values of M and W_0 assumed by an idealised group of galaxies in, e.g., a cluster at distance, r, or equivalently as representing the underlying *distribution* of values of M and W_0 at distance, r, from which the values taken by each individual galaxy are drawn. This means that we make no distinction between the effects of selection bias on groups of galaxies in a cluster and on isolated field galaxies. This is contrary to e.g. Lynden-Bell *et al*, (1988) or Teerikorpi (1984, 1987) in which selection effects on field and cluster

galaxies are treated separately. It seems to us, however, that no such distinction is necessary.

Now it is clear from figure (2.5b) that for lower values of $\log W_0$ one cannot sample M completely since less luminous galaxies are 'cut-off' by the magnitude limit. This means that the regression curve, $E(M|W_0)$, in this second case is no longer given by the straight line of figure (2.5a) - also shown in figure (2.5b) for comparison - but deviates from this line for small W_0 and is described rather by the dotted curve, as shown. Consequently, if one derives the Tully-Fisher constants by regressing magnitude on line width in a complete calibrating sample, and then uses this regression line to estimate the luminosity of a more remote galaxy from its measured line width, W_0, then the inferred absolute magnitude will *not* equal the mean value, $E(M|W_0)$, at that line width for observable galaxies at the greater distance. As we can see from the deviation of the dotted curve, the value of M estimated from the straight line will be *greater* than the true mean value; i.e. the luminosity will be systematically underestimated, leading to a negative Malmquist bias in the inferred distance of the galaxy.

This same effect will occur when this regression line is used to calibrate other indicators which involve luminosity. Moreover, diameter selection effects (i.e. the exclusion of galaxies with angular diameter less than some limiting value) will affect the D_n-σ relation in an analogous fashion.

The Malmquist bias inherent in the use of this regression has

been recognised by a large number of authors; furthermore, it has been shown that Malmquist bias will also arise when one calibrates using the orthogonal or bisector regression lines (c.f. Giraud, 1987; Bottinelli *et al*, 1986). Several authors have studied the bias in quantitative detail: in particular Teerikorpi (1984) has assumed a bivariate normal for the distribution of M and $\log W_0$ and has determined that, in this case, the Malmquist bias, $E(M|W_0) - M_0$, of the absolute magnitude at given line width has the same distance dependence as the bias for standard candles - as given by equation (2.3). The amplitude of the bias at a given distance is smaller than for standard candles, however, just as the bias amplitude was reduced for smaller values of σ in figure (2.1). This means that - at least from the point of view of Malmquist bias - using the Tully-Fisher relation with this regression is equivalent to using a 'better' standard candle; i.e. one of smaller intrinsic spread in luminosity. Intuitively this would seem to make sense: we can think of the measured line width of a galaxy as providing additional information which 'narrows down' the range of probable values for its luminosity.

It is the fact that the amplitude of the Malmquist bias depends on the *actual* distance of a galaxy, just as was the case for estimates derived from apparent magnitude alone, which makes its removal non-trivial. Any exact correction would require the true galaxy distance to be known - which would render somewhat redundant any attempts to estimate it!

2.3.3 Schechter's Scheme for Defining Bias-Free Distances

An important contribution to the debate over how best to calibrate relations of the Tully-Fisher type was provided in Schechter (1980). This paper was addressed primarily to the Faber-Jackson relation, although the author recognised its wider ramifications. Schechter observed that the mean log velocity dispersion, $E(\log\sigma|M)$, at given absolute magnitude, M, is no different in a magnitude limited sample than in a volume limited sample. Thus, the slope of the line obtained by regressing $\log\sigma$ on M is *unchanged* in a sample subject to magnitude selection; in other words this regression is unaffected by Malmquist bias. We can see that Schechter's observations are correct from figure (2.5b), for the Tully-Fisher case. For observable galaxies (i.e. $M \lessdot M_L$) we can still sample the full range of line widths at a given magnitude, even for magnitudes very close to the limit. Hence the mean log line width at given magnitude, $M \lessdot M_L$, is precisely equal to the mean value at the same magnitude in the completely sampled case of figure (2.5a). This means that we obtain the same regression line of $\log W_0$ on M in both cases. Specifically that regression line takes the form:-

$$E(\log W_0|M) = aM + b \qquad (2.14)$$

where a and b are constants. We can use this equation to estimate M from the measured line width of a galaxy by determining the value of M for which the observed log(line width) is equal to its *expected* value at that magnitude.

Schechter also pointed out that the immunity of this

regression line to Malmquist bias holds regardless of the *form* of the magnitude selection effects (e.g. a narrow magnitude window or a 'fuzzy' cut-off limit). The only condition required is that the selection effects depend on apparent magnitude alone; i.e. that the line widths are selection-free. Similar remarks obviously apply to other indicators. Hence for the D_n-σ relation we require that the observed galaxies are selected by apparent diameter alone. Lynden-Bell (1991) discusses this assumption and recognises that it holds true - to a good approximation - in the data set of ellipticals studied in Dressler *et al*, (1987) and Lynden-Bell *et al*, (1988). Indeed, Lynden-Bell points out that selection by diameter alone is consistent with good observational procedure: measurements of velocity dispersions require relatively long exposure times and large telescopes and are, therefore, costly. By contrast, measurements of galaxy diameters can be made from already existing photographic surveys. Thus, it is quite common to select galaxies for observation on the basis of their angular diameter alone while the velocity dispersion are measured later for the selected objects. Tully (1988) makes a similar point with respect to the Tully-Fisher relation.

The assumption of a selection-free observable is, nevertheless, clearly crucial to the unbiasedness of the Schechter regression line. If there *is* selection on line width then this regression line will exhibit Malmquist bias in precisely the same way as the regression of M on $\log W_0$ since the regression curve $E(\log W_0|M)$ for observable galaxies in an incomplete sample will now deviate from the straight line obtained for a complete sample. A number of authors (c.f. Aaronson *et al*, 1982; Teerikorpi, 1984; Kraan-Korteweg *et al*, 1986; Tully, 1988) have re-affirmed the results of Schechter and have thus

advocated calibration of the Tully-Fisher relation by regressing line widths on magnitudes so as to obtain bias-free distance estimates.

2.4 Summary and Concluding Remarks

In this chapter we have seen how galaxy distance estimates will in general be affected by Malmquist bias when one's galaxy sample is subject to observational selection effects. We have considered firstly how this bias affect distances inferred using only the apparent magnitude of a galaxy, and have reviewed the Minimum Bias Subset method for dealing with the bias, as proposed in early papers. We have discussed the limitations of the MBS: the fact that its use with a narrow magnitude selection window may lead to the detection of spurious streaming motions, and the more fundamental limitation that the MBS rejects those galaxies at higher redshifts for which the assumption of a 'quiet' Hubble flow - inherent in defining the MBS - might be considered more reasonable.

We have gone on to consider how Malmquist bias also affects distance indicators which depend on two observables, such as the Tully-Fisher relation; in particular how the bias may arise when this relation is derived by regressing magnitudes on line widths. By contrast we have considered the scheme proposed by Schechter (1980), whereby one prefers to regress line widths on magnitudes thus obtaining a straight line which is unaffected by Mamlquist bias, provided only that the selection effects are confined to the apparent magnitude alone.

The adoption of this scheme has not been universally endorsed and a number of authors (Sandage and Tammann, 1990; Bottinelli *et al*, 1986; Lynden-Bell *et al*, 1988) continue to favour a regression of magnitude on line width (or its equivalent for other indicators).

Of course the fact that the Schechter regression line is unbiased does not automatically qualify that line as the 'best' choice. In the first instance we must clarify precisely what we mean by 'best'; i.e. based on what criteria do we make our choice. It seems to us that this basic question has not been adequately addressed in the literature - a fact which obviously does nothing to help resolve the disagreements over which method is truly 'best'. Clearly the absence of Malmquist bias is a desireable property and if one adopts this as the *only* criterion then the Schechter scheme would indeed represent the best method of calibration. It is, however, straightforward to envisage a situation where the Schechter regression line is wholly inappropriate: the case where M and W_0 are completely uncorrelated. In this pathological case the mean value of $\log W_0$ at given M is a constant, independent of M, i.e. $E(\log W_0|M) \equiv E(\log W_0)$. Consequently, we obtain no information at all about M from measuring the line width of a galaxy and so the method of using the Schechter line to infer the magnitude, and hence the distance, breaks down completely - notwithstanding the fact that the line is *still* free from Malmquist bias!

In the next chapter we will introduce a rigorous statistical formulation, based on the analytical techniques of risk theory (c.f. Hogg and Craig, 1978; Mood and Graybill, 1974), which will allow us to

tackle more effectively the question of which method of distance estimation is truly 'best'. Later in the thesis we will return to the Schechter scheme so as to verify its validity within our statistical framework - but more importantly to identify under what circumstances the method is a *poor* choice, as was clearly the case in the pathological example considered above. In chapter (3), however, we will first of all restrict ourselves to the simpler case of estimates which depend only on apparent magnitude.

3. ESTIMATION OF DISTANCE USING APPARENT MAGNITUDE

3.1 Introduction

In the preceding chapters we have seen that the luminosity selection effects introduced in a magnitude limited sample of galaxies may lead to systematic errors in the estimation of the distances to those galaxies. In order to understand more precisely the form of these systematic effects, and to explore methods of reducing or eliminating them, we require to formulate the problem of distance estimation in a statistically rigorous manner. Following the standard statistical methods of risk theory, we will introduce a technique for defining different distance estimators and comparing their distributions and properties, so as to identify the estimator most appropriate to a given problem.

We will consider in this chapter the case where distances are estimated using only measurements of apparent magnitude - the most readily understood distance indicator. This simple approach will illustrate clearly the statistical principles involved and provide a framework for subsequent, more detailed, analysis. In particular, the extension of our analysis to include other distance indicators - such as apparent diameter or secondary indicators derived from e.g. the Tully-Fisher relation - is relatively straightforward. We will consider such an extension to estimates derived from two and three observables in the chapters which follow.

The analysis which we adopt here is similar to Simmons and

Stewart (1985) in their discussion of polarimetric estimators. Hereafter we shall denote an estimator of a parameter by a caret - e.g. an estimator of distance, r, will be written $\hat{r}$. Furthermore we adopt the usual statistical convention, wherever possible, of denoting random variables by bold type characters.

We will derive expressions for the apparent magnitude distribution of observable galaxies at given (though in general unknown) true distance, taking into account the effects of luminosity selection, and use this distribution to define various estimators of that distance. Of course in certain astrophysical problems it may be expedient to estimate not the distance, r, but rather some function of r such as logr or r^{-1} (c.f. Feast, 1987). Indeed, we have already seen in chapter (2) that cosmologists make frequent use of the *distance modulus*, which is essentially an estimator of logr. It does not follow that a good estimator of logr, for example, will necessarily be a good estimator of r or vice versa, and we will discuss some of the implications of this in due course. Our analysis may easily be adapted to the estimation of functions of r, as we will see later, but it is instructive to consider firstly only estimates of r itself.

3.2 The Observed Distribution of Apparent Magnitude

Let the absolute magnitude, **M**, and position, $\underline{\mathbf{r}}$, of a galaxy be random variables with some intrinsic joint distribution function. Let $N(\mathbf{M},\underline{\mathbf{r}})d\mathbf{M}d\mathbf{V}$ denote the number of galaxies in volume dV at position $\underline{\mathbf{r}}$ that have absolute magnitudes in the range **M** to **M**+d**M**. Further, let

$n(\underline{r})$ denote the number density of galaxies (of all magnitudes) at position $\underline{r}$. Suppose now that **M** and $\underline{r}$ are uncorrelated so that we may write:-

$$N(\mathbf{M},\underline{r})d\mathbf{V}d\mathbf{M} = \Psi(\mathbf{M})n(\underline{r})d\mathbf{V}d\mathbf{M} \qquad (3.1)$$

i.e. we assume the existence of a luminosity function (hereafter LF), $\Psi(\mathbf{M})$, which describes the magnitude distribution of galaxies, independent of their position. The LF has frequently been defined as a number density of galaxies per unit range of absolute magnitude (c.f. Felten, 1985); clearly such a description cannot be independent of position since there will be an immediate dichotomy between field and cluster samples. In equation (3.1), therefore, we prefer to define the LF as a probability density - a definition which has found increasing favour in recent years and which is identical to that used in stellar statistics (c.f. Kirshner *et al*, 1979; Mihalas and Binney, 1981). By defining $\Psi(\mathbf{M})$ in this way one need make no assumptions about the uniformity of $n(\underline{r})$. Even with this definition, however, the LF of galaxies of all Hubble types (usually referred to as the *general* or *universal* LF) will not be independent of position, since the relative frequencies of the different morphological types depend strongly on the local density (c.f. Dressler, 1980; Postman and Geller, 1984). It has also been demonstrated (c.f. Hamilton, 1988; Einasto, 1990) that the brightest galaxies tend to lie preferentially in groups and clusters, probably as a result of dynamical evolution. Apart from these galaxies, however, recent reviews (c.f. Binggelli *et al*, 1988) report no evidence that the LF for a specific Hubble type depends on the local density. It seems, therefore, that the separability of $\Psi(\mathbf{M})$ and $n(\underline{r})$ assumed in equation (3.1) is a valid approximation for any single Hubble type.

Consider now the joint distribution, $\rho(M,\underline{r})$, of position and absolute magnitude for *observable* galaxies in a sample subject to luminosity selection effects, as described by a selection function, $S(M,\underline{r})$. This function measures the probability that a galaxy of magnitude M and at position $\underline{r}$ will be observable. Thus, it follows that:-

$$\rho(M,\underline{r}) = \frac{\Psi(M)n(\underline{r})S(M,\underline{r})dMdV}{\iint \Psi(M)n(\underline{r})S(M,\underline{r})dMdV} \qquad (3.2)$$

Note that the selection function, $S(M,\underline{r})$, does not determine the probability that a galaxy would actually be *observed*; clearly this would depend on the local number density, $n(\underline{r})$, which will in general be unknown. The definition of $S(M,\underline{r})$ which we adopt here will be independent of $n(\underline{r})$ and, moreover, will be independent of direction provided one may correct for the directional dependence of galactic absorption. A number of standard observational methods exist for carrying out these corrections. (c.f. Sandage and Tammann, 1981; Burstein and Heiles, 1982)

We now consider the conditional distribution, $\xi(M|r_0)$, of absolute magnitude at a given distance, r_0, for observable galaxies. Substituting from equation (3.2), this is given by:-

$$\xi(M|r_0) = \frac{\Psi(M)S(M,r_0)}{\int \Psi(M)S(M,r_0)dM} \qquad (3.3)$$

Note that this conditional distribution is independent of the galaxy number density, $n(\underline{r})$. In other words, the shape of the observed magnitude distribution at distance r will not change with the

density of the local environment. This fact proves very useful in considering the effects of bias on distance estimation, but it seems to us that it has not always been fully appreciated in the literature.

Finally, we change variables in equation (3.3) to obtain the conditional distribution, $\varsigma(m|r_0)$ of apparent magnitude, m, at distance, r_0, for observable galaxies. Formally, m is a function of the random variable, M, and so is itself a random variable. m and M are related in the usual way. viz:-

$$m = M + 5\log r_0 + 25 \tag{3.4}$$

where the parameter r_0, the true distance of the galaxy, is measured in Mpc and we have assumed that the effects of absorption in both our own galaxy and in the observed galaxy have been removed or neglected. It follows that $\varsigma(m|r_0)$ is given by:-

$$\varsigma(m|r_0) = \frac{\Psi(m-5\log r_0-25)S(m)}{\int \Psi(m-5\log r_0-25)S(m)dm} \tag{3.5}$$

To proceed further we must specify the form of the LF, $\Psi(M)$, and the selection function, $S(m)$. In this chapter, we will choose $S(m)$ to be a Heaviside step function at some magnitude limit, m_L. i.e.:-

$$S(m) = \begin{cases} 1 & \text{if } m < m_L \\ 0 & \text{if } m \geq m_L \end{cases} \tag{3.6}$$

We will also consider the case where the LF, $\Psi(M)$, is normally distributed with mean M_0 and variance σ^2. Our primary reason for

these choices is mathematical expediency, since the algebraic expressions for the estimators and their distributions are greatly simplified. Nevertheless, the Gaussian model seems to be a reasonable one for certain specific Hubble types. Various studies of first-ranked cluster E galaxies (c.f. Sandage and Hardy, 1973; Schneider *et al*, 1983) have indicated that their LF is near Gaussian with a dispersion of $\sigma \simeq 0.3$ mag. Sandage *et al* (1985) have modelled the LF of all spirals in the Virgo cluster by a Gaussian of mean absolute magnitude -18.4 and dispersion $\sigma \simeq 1.5$ mag, and the dispersion decreases if one considers only certain sub-classes of spirals. In particular, the LF of ScI spirals has been modelled by a Gaussian of dispersion $\sigma \simeq 0.7$ mag (Sandage and Tammann, 1975).

Substituting for $\Psi(M)$ and $S(m)$ we find that equation (3.5) reduces to:-

$$\varsigma(m|r_0) = \begin{cases} \dfrac{\exp\left(-1/2\sigma^2 (m-5\log r_0 - 25 - M_0)^2\right)}{\sqrt{2\pi}\,\sigma\,\Phi(m_L - 5\log r_0 - 25 - M_0/\sigma)} & m < m_L \\ 0 & m \geq m_L \end{cases} \qquad (3.7)$$

Φ denotes the cumulative standard normal distribution, of mean zero and unit variance.

Figure (3.1) shows examples of $\varsigma(m|r_0)$ for various different true distances and for typical parameter values of $\sigma = 1$, $M_0 = -20$ and $m_L = 15$. It can be seen that at distances greater than 50 MPc the observed distribution quickly becomes highly non-Gaussian in shape.

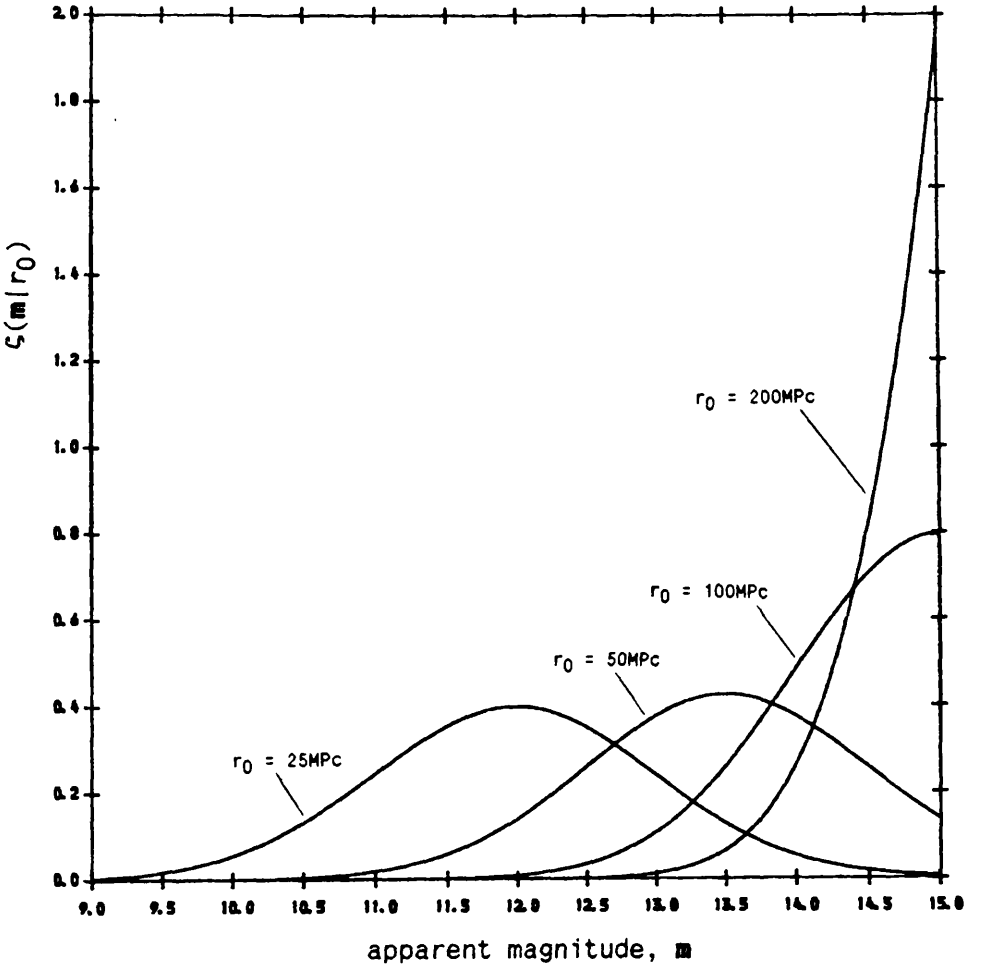

Figure (3.1)

Probability density function, $\varsigma(m|r_0)$, for observable galaxies at true distance r_0 = 25MPc, 50MPc, 100MPc and 200MPc
(σ = 1, M_0 = -20, m_L = 15)

We will now use this distribution to define different estimators of the distance, r_0, and investigate their properties.

3.3 Definitions of Distance Estimators

There are a number of estimators which present themselves as obvious candidates for estimating distance. One may define a set of simple estimators by solving for r_0 in equation (3.4), setting **m** equal to the observed apparent magnitude, m, and assuming a particular value for the absolute magnitude, M_* say, to obtain in general:-

$$\hat{r} = 10^{0.2(m - M_* - 25)} \qquad (3.8)$$

Choosing an appropriate value for M_* is clearly important in order to define a good estimator of distance and reduce the biasing effects of selection. In chapter (6) we will address in some detail the question of how one might choose the 'best' value of M_* for a given sample of galaxies. For the present, however, we will consider three estimators of this type which correspond to specific elementary choices of M_*. The definitions of these estimators now follow.

3.3.1 'Naive' : $\hat{r}_N$

For this estimator we simply choose $M_* = M_0$, the mean absolute magnitude of the intrinsic LF. Thus, we take no account of the bias due to selection or the dispersion of the luminosity function and the estimator is simply given by:-

$$\hat{r}_N = 10^{0.2(m - M_0 - 25)} \qquad (3.9)$$

3.3.2 'Malmquist' : $\hat{r}_{MAL}$

Equation (3.8) may be rewritten as :-

$$\hat{r} = 10^{0.2m} R \qquad (3.8a)$$

where $R = 10^{-0.2(M + 25)}$

Let $\langle R \rangle$ be the mean value of R for observable galaxies in a magnitude limited sample, assuming a uniform spatial distribution, viz (Malmquist, 1920):-

$$\langle R \rangle = 10^{-0.2(M_0 - 1.61\sigma^2 + 25)} \qquad (3.10)$$

We thus define the 'Malmquist' estimator as $\hat{r}_{MAL} = 10^{0.2m} \langle R \rangle$, which is equivalent to adopting $M_* = M_0 - 1.61\sigma^2$ in equation (3.8). Note that, as we have seen in chapter (2), the mean absolute magnitude of observable galaxies in a magnitude limited sample (again assuming a uniform density) is given by $M_0 - 1.38\sigma^2$ (Malmquist, 1920). One could therefore define a second 'Malmquist' estimator by setting M_* equal to this value. Various authors have considered this point (c.f. Feast, 1987) and conclude that a distance estimator derived from the mean distance of galaxies is a better choice than one derived from the mean magnitude. The fact that these estimators are not equivalent is a consequence of the non-linear relationship between magnitude and distance.

3.3.3 'Proximal' : $\hat{r}_P$

In this case we adopt $M = M_0 + k\sigma^2$, where the constant, k, is chosen so that the percentage bias of $\hat{r}_P$ tends to zero as the true distance

tends to zero. The required value of $k \simeq 0.23$ may be determined from equation (3.29) as we will see later. This estimator should be most accurate for nearby galaxies, hence the name 'proximal'.

All of the above estimators might be considered worthy of the title 'naive' since the 'proximal' and 'Malmquist' estimators differ from the former only by a constant correction term. Intuitively it seems clear that such a simple correction will not be adequate to completely remove the effects of selection. Furthermore these estimators make little or no use in their definitions of the observed magnitude distribution. We now consider four further estimators, which are derived directly from the distribution function, $\varsigma(\mathbf{m}|r_0)$.

3.3.4 'Mode' : $\hat{r}_M$

We define this estimator as the value of r_0 for which the observed apparent magnitude is the modal value of $\varsigma(\mathbf{m}|r_0)$ - i.e. the value for which m_{obs} maximises $\varsigma(\mathbf{m}|r_0)$ with respect to $\mathbf{m}$. Hence $\hat{r}_M$ satisfies:-

$$\frac{\partial}{\partial \mathbf{m}} \varsigma(\mathbf{m}|r_0=\hat{r}_M)\Big|_{\mathbf{m}=m_{obs}} = 0 \qquad (3.11)$$

3.3.5 'Median' : $\hat{r}_{MED}$

This estimator is defined as the value of r_0 for which the observed apparent magnitude is the median of $\varsigma(\mathbf{m}|r_0)$. Substituting from equation (3.7) it follows that $\hat{r}_{MED}$ is the solution of:-

$$\frac{\Phi(m_{obs}-5\log\hat{r}_{MED}-25-M_0/\sigma)}{\Phi(m_L-5\log\hat{r}_{MED}-25-M_0/\sigma)} = 0.5 \qquad (3.12)$$

3.3.6 'Mean' : $\hat{r}_{ME}$

This is defined as the value of r_0 for which the observed apparent magnitude, m_{obs}, is equal to the mean of $\varsigma(m|r_0)$. Thus $\hat{r}_{ME}$ satisfies:-

$$\int m\, \varsigma(m|r=\hat{r}_{ME})dm = m_{obs} \qquad (3.13)$$

Substituting from equation (3.7) and carrying out the integration we find that the equation defining $\hat{r}_{ME}$ is:-

$$m_{obs} = M_0 + 25 + 5\log\hat{r}_{ME} - \frac{\sigma \exp\left(-1/2\sigma^2 (m_L - 5\log\hat{r}_{ME} - 25 - M_0)^2\right)}{\sqrt{2\pi}\,\Phi(m_L - 5\log\hat{r}_{ME} - 25 - M_0/\sigma)} \qquad (3.14)$$

3.3.7 'Maximum Likelihood' : $\hat{r}_{ML}$

This estimator is defined as the value of r_0 which maximises the probability, with respect to r_0, of obtaining the observed apparent magnitude. Therefore $\hat{r}_{ML}$ satisfies:-

$$\frac{\partial}{\partial r_0}\varsigma(m=m_{obs}|r_0)\Big|_{r_0=\hat{r}_{ML}} = 0 \qquad (3.15)$$

Upon substitution for $\varsigma(m|r_0)$ from equation (3.7) we find that $\hat{r}_{ML}$ satisfies equation (3.14); $\hat{r}_{ML}$ and $\hat{r}_{ME}$ are, therefore, identical for a Gaussian LF and we will no longer differentiate between them. Both are also equal to the 'Teerikorpi estimator' (Teerikorpi, 1975) which is defined by first computing the mean absolute magnitude of observable galaxies at distance, r_0; (c.f. equation 2.3) - a definition which is clearly equivalent to $\hat{r}_{ME}$.

We also found that $\hat{r}_N$ and $\hat{r}_M$ are identical for a Gaussian LF. That several of our estimators should be equivalent is not too surprising since it may easily be shown that, in the absence of selection effects, the 'naive', 'mean', 'median', 'mode' and 'maximum likelihood' estimators are all identical for a Gaussian LF (c.f. Graybill, 1961).

At this point we introduce a convenient unit of distance which we will refer to as the *limiting* distance, r_L, defined by:-

$$5\log r_L = m_L - M_0 - 25 \tag{3.16}$$

Thus, in these scaled units, a galaxy of absolute magnitude M_0 and at unit distance would be observed to have the limiting apparent magnitude, m_L. By scaling distances in this way we may investigate the properties of our estimators without specifying an explicit value of M_0 and m_L. Throughout the remainder of this - and subsequent - chapters we will, therefore, use the notation $x = r/r_L$ and $\hat{x} = \hat{r}/r_L$, to denote true distances and estimated distances respectively. A typical value for r_L, corresponding to $M_0 = -20$ and $m_L = 15$, would be 100 Mpc. In these scaled units, and for a Gaussian LF, the conditional distribution ς may be written as:-

$$\varsigma(m|x_0) = \begin{cases} \dfrac{\exp\left(-1/2\sigma^2 (m - m_L - 5\log x_0)^2\right)}{\sqrt{2\pi}\,\sigma\,\Phi(-5\log x_0/\sigma)} & m < m_L \\ 0 & m \geq m_L \end{cases} \tag{3.17}$$

It is useful to represent the estimators in terms of the

following implicit equations:-

$$m - m_L = \begin{cases} 5\log\hat{x}_N \\ 5\log\hat{x}_{MAL} - 1.61\sigma^2 \\ 5\log\hat{x}_P + 0.23\sigma^2 \\ 5\log\hat{x}_{ML} - \dfrac{\sigma \exp\left(-1/2\sigma^2(5\log\hat{x}_{ML})^2\right)}{\sqrt{2\pi}\ \Phi(-5\log\hat{x}_{ML}/\sigma)} \end{cases} \qquad (3.18)$$

The solutions of these equations are shown in figures (3.2) and (3.3), for $\sigma = 0.5$ and $\sigma = 1$ respectively. Thus, in geometrical terms, the distances inferred by each of the estimators may be determined from the graphs by drawing a vertical line from the observed value of $m_L - m$ and finding the points of intersection with the appropriate estimator curve. In the interests of clarity we have not plotted the curve of the 'median' estimator, $\hat{x}_{MED}$. Although not identically equal to $\hat{x}_{ML}$, calculations show that the median estimator differs from the former by no more than a few percent over the domain shown. For the purposes of this study, therefore, we will regard $\hat{x}_{MED}$ and $\hat{x}_{ML}$ as equivalent.

It can be seen from figures (3.2) and (3.3) that all of the estimators have similar asymptotic behaviour for $m \ll m_L$; their behaviour close to the limiting magnitude is, however, markedly divergent. Each of the first three estimators, $\hat{x}_N$, $\hat{x}_{MAL}$ and $\hat{x}_P$, tends to a finite limit as m tends to m_L. Hence there will be an upper limit to the distances which may be inferred by each of these estimators. Indeed it follows that any estimator of the type shown in equation (3.8) will be defined only to some upper limit, r_*, given by:-

$$r_* = 10^{\,0.2(m_L - M_* - 25)} \qquad (3.19)$$

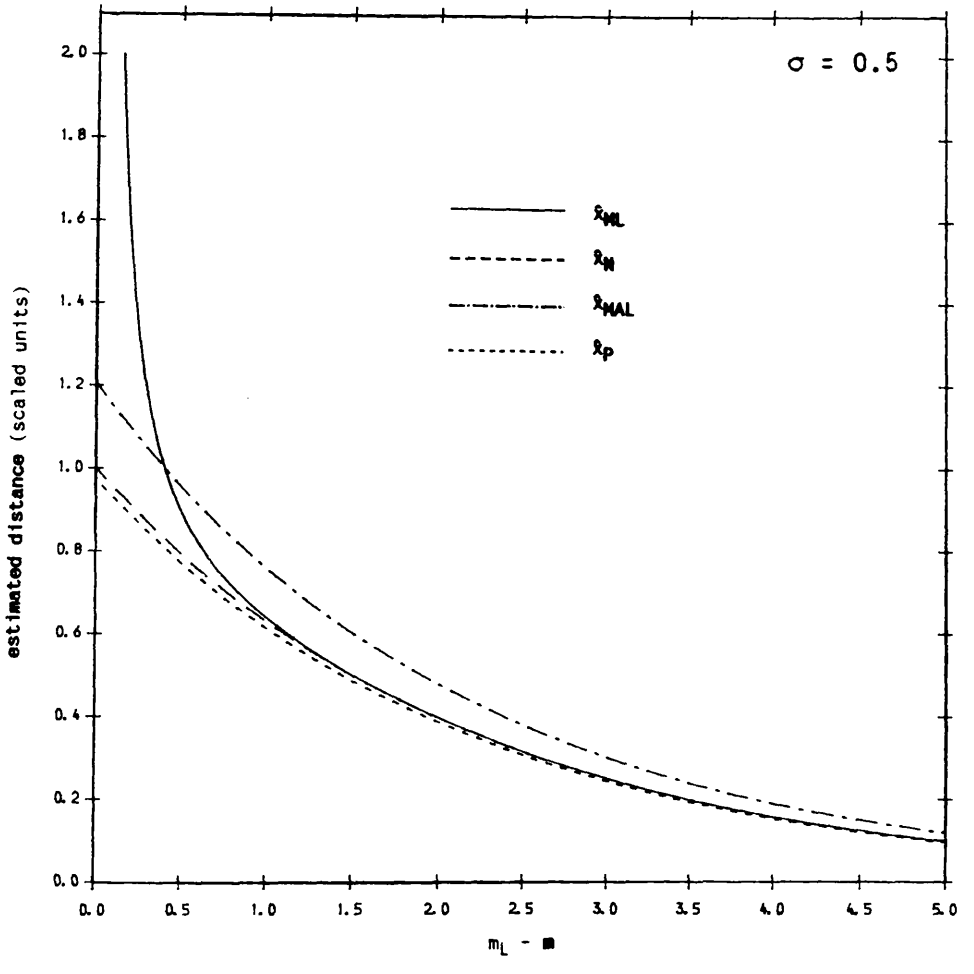

Figure (3.2)

Distance estimator curves for $\hat{x}_{ML}$, $\hat{x}_N$, $\hat{x}_{MAL}$ and $\hat{x}_P$ ($\sigma = 0.5$)

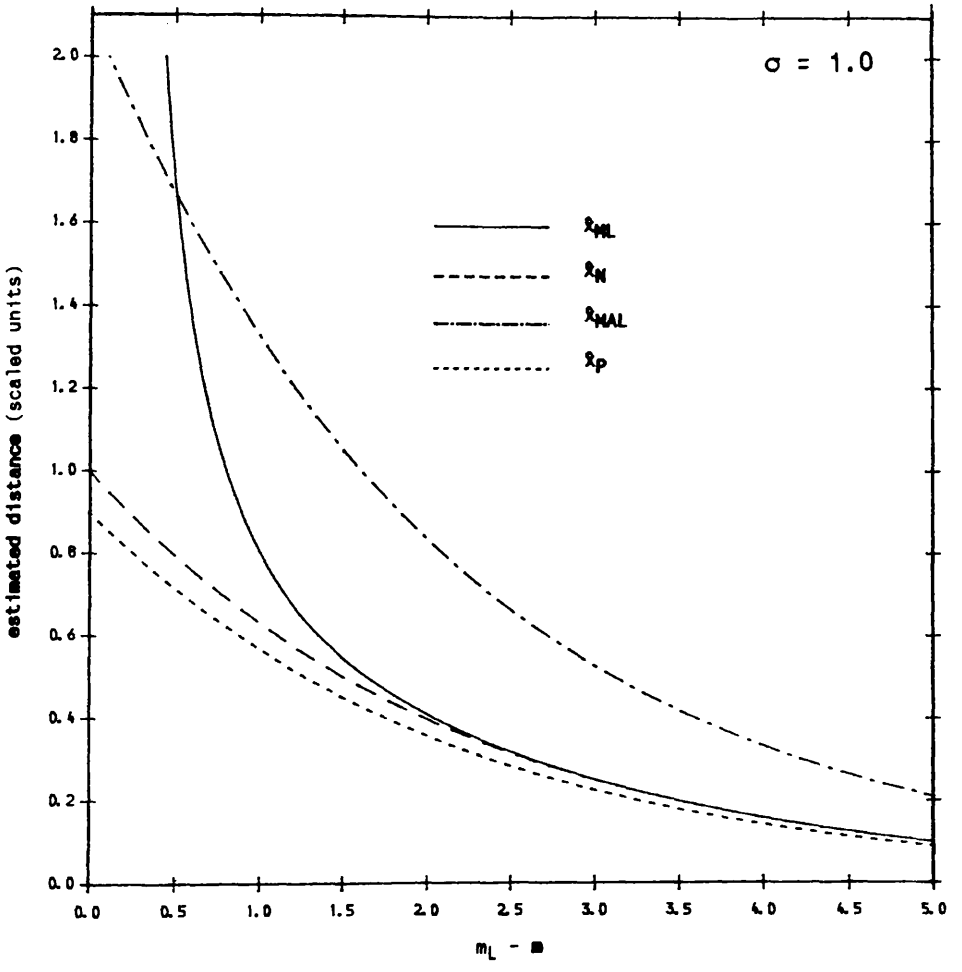

Figure (3.3)

Distance estimator curves for $\hat{x}_{ML}$, $\hat{x}_N$, $\hat{x}_{MAL}$ and $\hat{x}_P$ ($\sigma = 1$)

or, in scaled units, to x_* given by:-

$$x_* = 10^{-0.2(M_* - M_0)} \tag{3.19a}$$

Hence if the true distance of an observed galaxy is greater then x_* then, regardless of the galaxy's apparent magnitude, its distance will be systematically underestimated.

In both figures (3.2) and (3.3) we see that the estimator curve for $\hat{x}_{ML}$, on the other hand, does not intercept the vertical axis. In fact, it follows from equations (3.14) that $\hat{x}_{ML}$ can take arbitrarily large values and $\hat{x}_{ML} \to \infty$ as $m \to m_L$. Intuitively, therefore, it would seem that $\hat{x}_{ML}$ should be the more reliable estimator for very distant galaxies. Before further comment is possible, however, we must first consider the distribution of each estimator.

3.4 Distributions of Distance Estimators

To derive the probability density function of any estimator $\hat{x}$ we note that $\hat{x}$ is a function of the random variable, $\mathbf{m}$: i.e. $\hat{x} = \hat{x}(\mathbf{m})$. The distribution function of $\hat{x}$ may then be written down in terms of the distribution of $\mathbf{m}$, $\varsigma(\mathbf{m}|x_0)$, and so will also depend on the parameter, x_0, the true distance. Thus the distribution, $X(\hat{x}|x_0)$, of an estimator, $\hat{x}$, for galaxies sampled at true distance, x_0 is given by:-

$$X(\hat{x}|x_0) = \varsigma(\mathbf{m}(\hat{x})|x_0) \left|{}^{d\mathbf{m}}/_{d\hat{x}}\right| \tag{3.20}$$

For each estimator $X(\hat{x}|x_0)$ may, therefore, be computed from

equations (3.17) and (3.18). As an example, figures (3.4) and (3.5) show graphs of the distribution functions of $\hat{x}_N$ and $\hat{x}_{ML}$ respectively, for different true distances and for $\sigma = 1$.

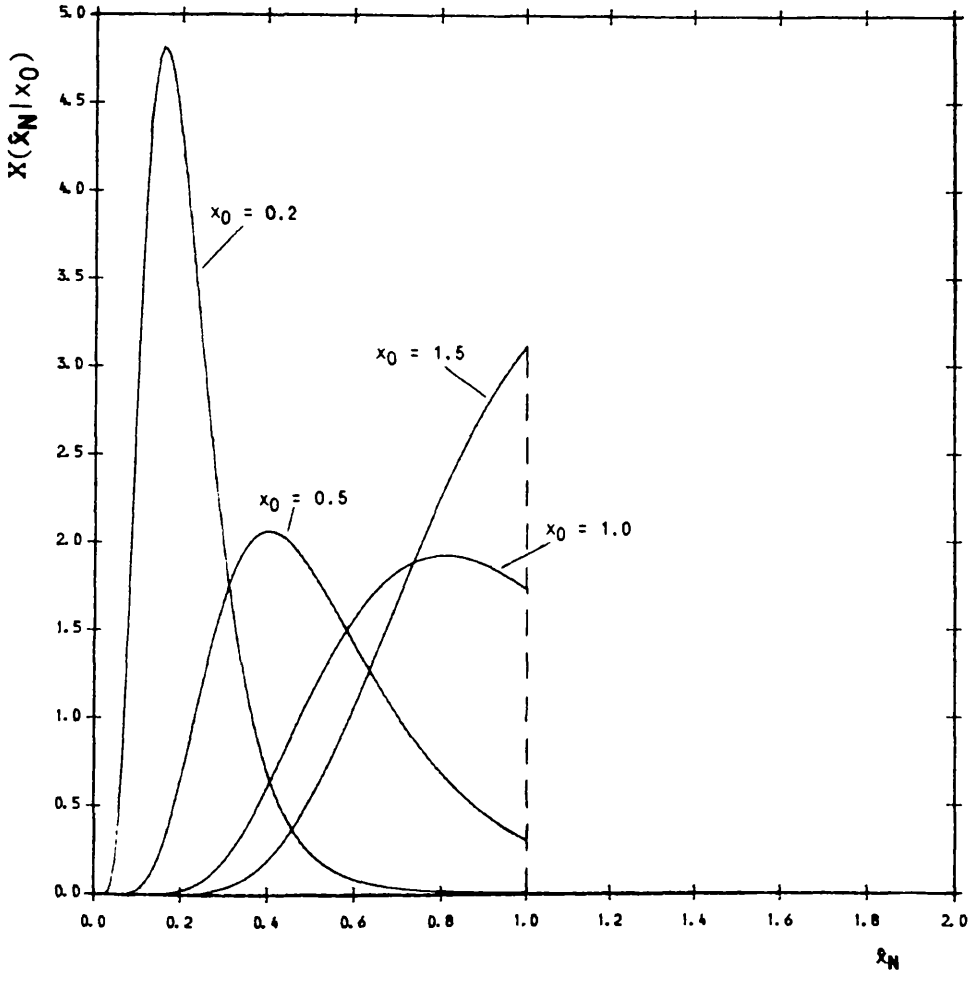

Figure (3.4)

Probability density function, $X(\hat{x}_N|x_0)$, of the estimator $\hat{x}_N$ at true distance $x_0 = 0.2$, 0.5, 1.0 and 1.5 ($\sigma = 1$)

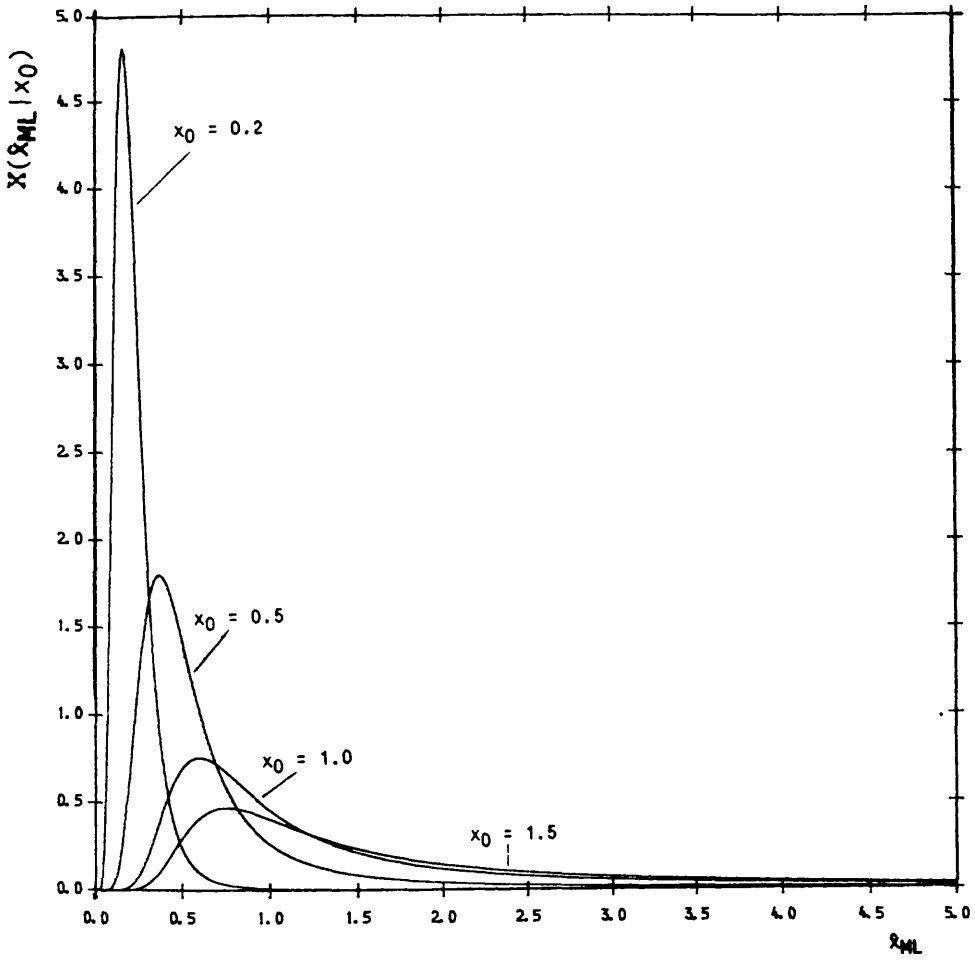

Figure (3.5)

Probability density function, $X(\hat{x}_{ML}|x_0)$, of the estimator $\hat{x}_{ML}$ at true distance x_0 = 0.2, 0.5, 1.0 and 1.5 (σ = 1)

There are a number of different statistical descriptors which may be used in order to compare these distributions. For example, in figure (3.6) we plot the modal value, x_{mode}, of $\hat{x}_N$ and $\hat{x}_{ML}$ as a function of true distance, x_0, and for $\sigma = 1$. For both estimators we see that $x_{mode} < x_0$, for all x_0; i.e. the mode of both distributions is negatively biased. At large distances the bias becomes particularly severe for $\hat{x}_N$ since this estimator is defined only for $\hat{x}_N \leq 1$, so that we must necessarily have $x_{mode} \leq 1$, even for $x_0 >> 1$. Similar behaviour is found for $\hat{x}_{MAL}$ and $\hat{x}_P$.

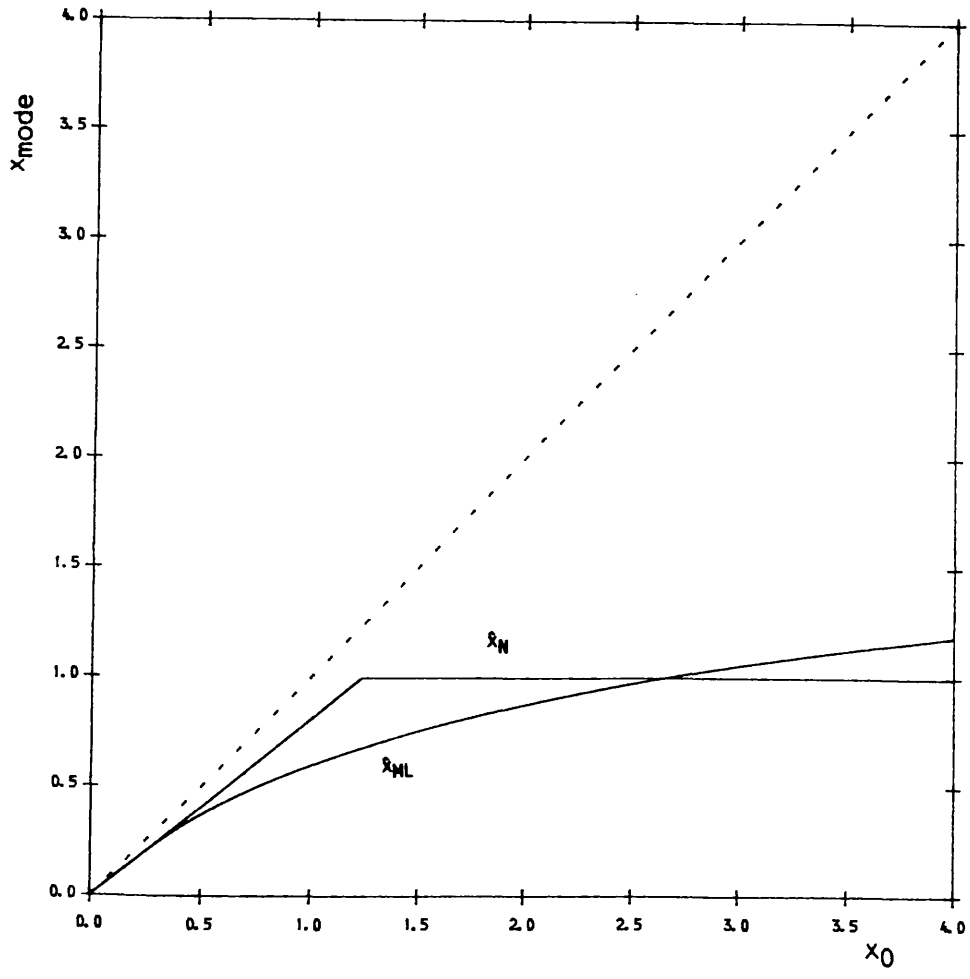

Figure (3.6)

Modal value, x_{mode}, of $\hat{x}_N$ and $\hat{x}_{ML}$ as a function of true distance, x_0 ($\sigma = 1$)

Another common descriptor is the median value, $x_{.5}$, defined implicitly by:-

$$\int_0^{x_{.5}} X(z|x_0)\, dz = 0.5 \qquad (3.21)$$

Clearly $x_{.5}$ will be a function of the true distance, x_0. For any x_0 we may compute the median percentage bias, $PB(x_0)$, of an estimator given by:-

$$PB(x_0) = \frac{x_{.5} - x_0}{x_0} .100\% \qquad (3.22)$$

It is useful to consider the percentage bias, and not just the bias, of an estimator since the former provides a measure of the systematic error which may be directly compared at different true distances. Figure (3.7) shows the median percentage bias of $\hat{x}_N$ and $\hat{x}_{ML}$, as a function of true distance, x_0, and for $\sigma = 1$. It can be seen that $x_{.5}$ is severely biased at large distances for both estimators, although, unlike the mode, the median bias is positive for $\hat{x}_{ML}$.

In most estimation problems it is generally the moments, however - and in particular the expected value and variance - of an estimator which are most frequently used to describe its properties. To be consistent with our previous notation we will denote the expected value of an estimator, $\hat{x}$, at true distance, x_0, by $E(\hat{x}|x_0)$. It follows that $E(\hat{x}|x_0)$ is given by:-

$$E(\hat{x}|x_0) = \int \hat{x}\, X(\hat{x}|x_0) d\hat{x} \qquad (3.23)$$

We thus define the mean bias, $B(\hat{x},x_0)$, of an estimator, $\hat{x}$, at true

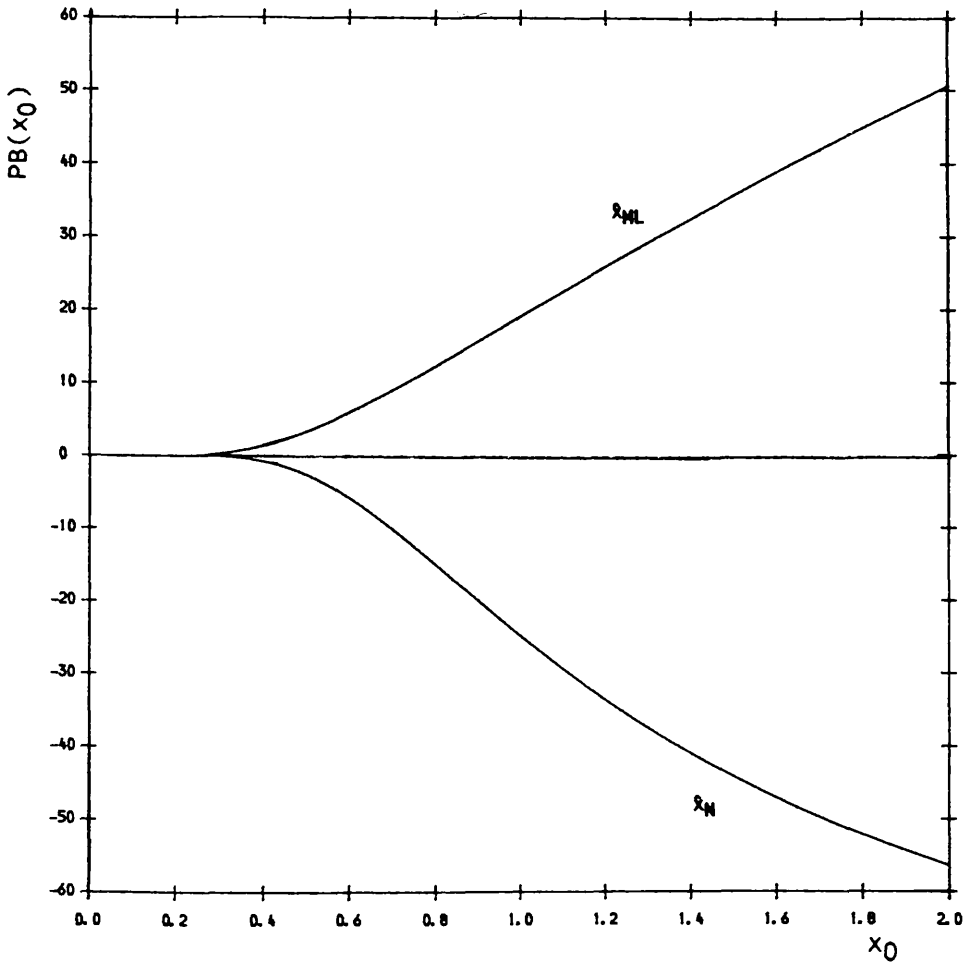

Figure (3.7)

Percentage median bias, $PB(x_0)$, of $\hat{x}_N$ and $\hat{x}_{ML}$ as a function of true distance, x_0 ($\sigma = 1$)

distance, x_0, by:-

$$B(\hat{x},x_0) = E(\hat{x}|x_0) - x_0 \qquad (3.24)$$

i.e.

$$B(\hat{x},x_0) = \int (\hat{x} - x_0)\, X(\hat{x}|x_0)d\hat{x} \qquad (3.24a)$$

Although the bias of the median and mode are recognised as possible indicators of the 'goodness' of an estimator, it is the mean bias as defined in equation (3.24) which is most often considered in the statistics literature. Henceforth we will adopt this latter definition of bias.

We also define here the mean square error (hereafter MSE), $\epsilon(\hat{x},x_0)$, of an estimator as follows:-

$$\epsilon(\hat{x},x_0) = \int (\hat{x} - x_0)^2\, X(\hat{x}|x_0)d\hat{x} \qquad (3.25)$$

Compare this with the definition of variance, viz:-

$$V(\hat{x},x_0) = \int (\hat{x} - E(\hat{x}|x_0))^2\, X(\hat{x}|x_0)d\hat{x} \qquad (3.26)$$

For an unbiased estimator MSE and variance are equivalent; when one is considering biased estimators MSE is the more relevant quantity.

The bracketed quantity $(\hat{x} - x_0)^2$ in equation (3.25) is an example of what is referred to in risk theory as the *loss* or *penalty* function (c.f. Hogg and Craig, 1978). This is defined as a non-negative number, $L(\hat{x},x_0)$, which measures the 'loss' involved in adopting as the

true distance the value taken by the estimator, $\hat{x}$, when that true distance is actually equal to x_0. Another natural choice for the loss function is $L(\hat{x},x_0) = |\hat{x} - x_0|$, i.e. the absolute difference between $\hat{x}$ and x_0. For any general loss function, $L(\hat{x},x_0)$, we may define the *risk* function $R(\hat{x},x_0)$ by:-

$$R(\hat{x},x_0) = \int L(\hat{x},x_0)\, X(\hat{x}|x_0)d\hat{x} \qquad (3.27)$$

i.e. the risk function is just the expected value of the loss function, and may be used as a criterion for determining whether $\hat{x}$ is a good estimator of distance. Furthermore, an estimator for which the risk function is a minimum might be considered in some sense to be 'best'. The most appropriate loss function to use will ultimately depend on the context of the problem. For example, if large estimation errors were considered especially problematic then $L(\hat{x},x_0) = (\hat{x} - x_0)^2$ would be more suitable than $L(\hat{x},x_0) = |\hat{x} - x_0|$.

The loss function $L(\hat{x},x_0) = (\hat{x} - x_0)^2$ appears very frequently in the statistics literature and is often assumed when not stated explicitly. In the interests of keeping our terminology as simple as possible we will follow this convention and, for the most part, will adopt the term *risk* to refer exclusively to the risk function computed with $L(\hat{x},x_0) = (\hat{x} - x_0)^2$, as in equation (3.25). Thus, unless we specifically make a distinction, we will use the term risk, $R(\hat{x},x_0)$, synonymously with the MSE, $\epsilon(\hat{x},x_0)$

The computation of bias and risk is straightforward for $\hat{x}_N$, $\hat{x}_P$ and $\hat{x}_{MAL}$. Indeed, one may readily establish an expression for the

nth moment of an estimator, $\hat{x}$, of the form:-

$$\hat{x} = 10^{\,0.2(m - m_L - \Delta\sigma^2)} \tag{3.28}$$

for any constant, Δ, viz:-

$$E(\hat{x}^n|x_0) = \frac{x_0^{\,n}\,\Phi(-5\log x_0 - \Delta n\sigma^2/\sigma)\,\exp(\kappa n\sigma^2(\Delta - \tfrac{1}{2}\kappa n))}{\Phi(-5\log x_0/\sigma)} \tag{3.29}$$

where $\kappa = 0.2/\ln 10 \simeq 0.46$

It follows from equation (3.29) that $E(\hat{x}|x_0) \to x_0$ as $x_0 \to 0$, if $\Delta = \tfrac{1}{2}\kappa$. This is the condition which defines the 'proximal' estimator, $\hat{x}_P$, for which $\Delta \simeq 0.23$.

Figures (3.8) to (3.11) show graphs of the percentage bias and risk of $\hat{x}_N$, $\hat{x}_{MAL}$ and $\hat{x}_P$ calculated as a function of true distance, and for $\sigma = 0.5$ and $\sigma = 1$.

Note that both the bias and risk of each estimator are highly sensitive to the true distance, x_0 - which is unknown. All of the estimators are negatively biased at large x_0; this is directly due to the effects of luminosity selection and would result in the familiar systematic underestimation of distance. For example, if the true distance were equal to 1.5 then the expected values of $\hat{x}_{ML}$, $\hat{x}_N$, $\hat{x}_{MAL}$ and $\hat{x}_P$ would be 1.05, 0.68, 1.35 and 0.6 respectively ($\sigma = 1$).

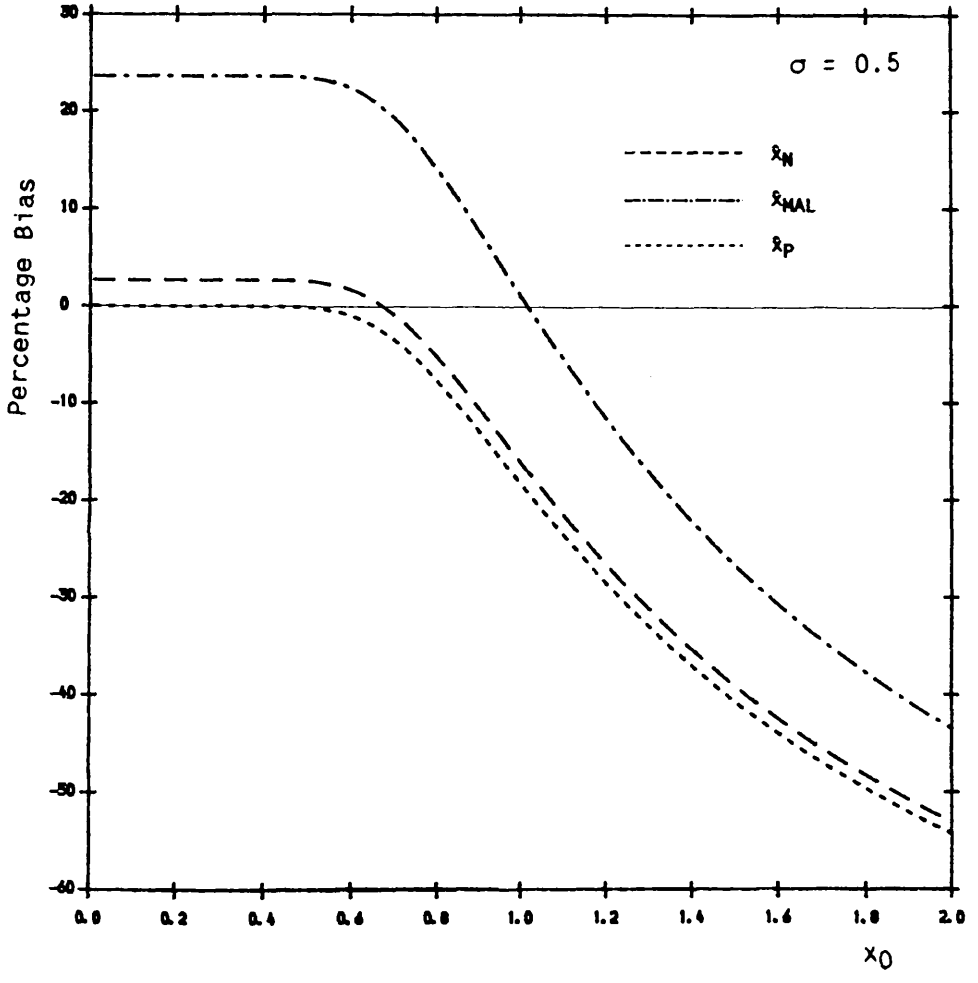

Figure (3.8)

Percentage bias of $\hat{x}_N$, $\hat{x}_{MAL}$ and $\hat{x}_P$ as a function of true distance, x_0 ($\sigma = 0.5$)

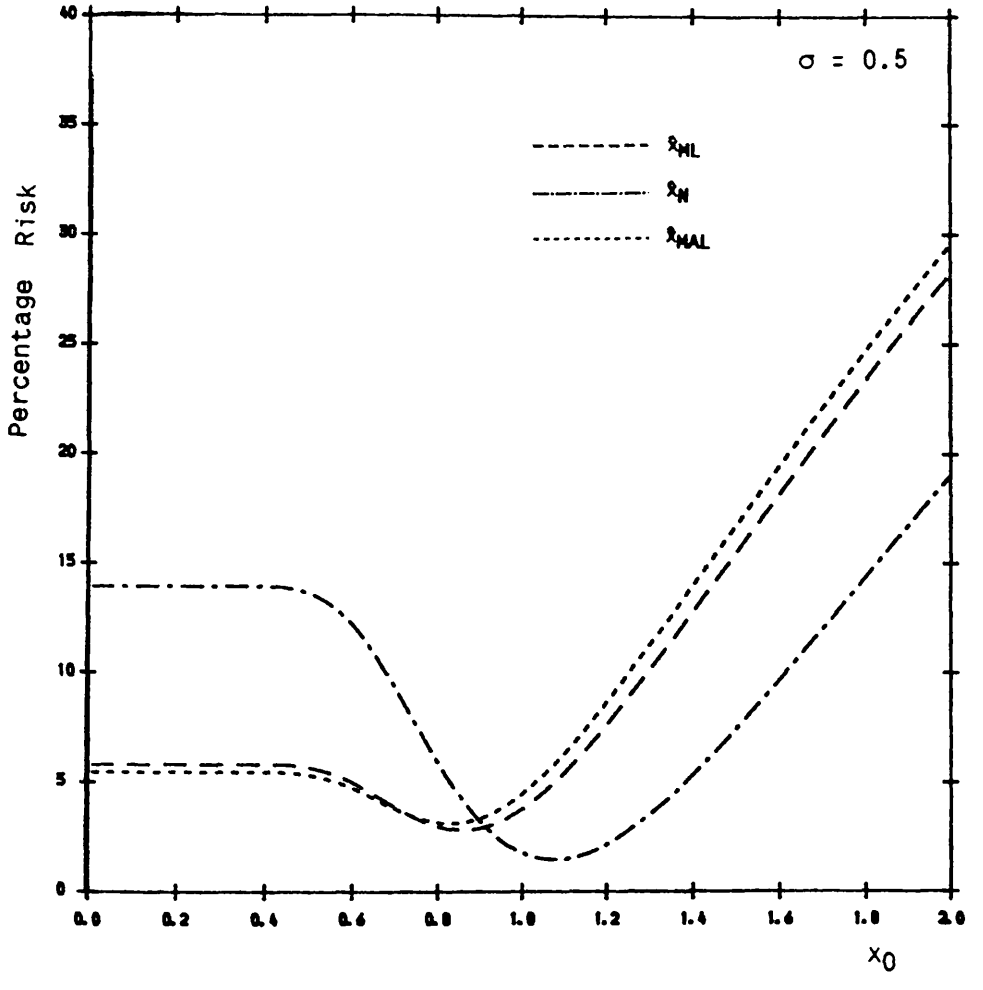

Figure (3.9)

Percentage risk of $\hat{x}_N$, $\hat{x}_{MAL}$ and $\hat{x}_P$ as a function of true distance, x_0 ($\sigma = 0.5$)

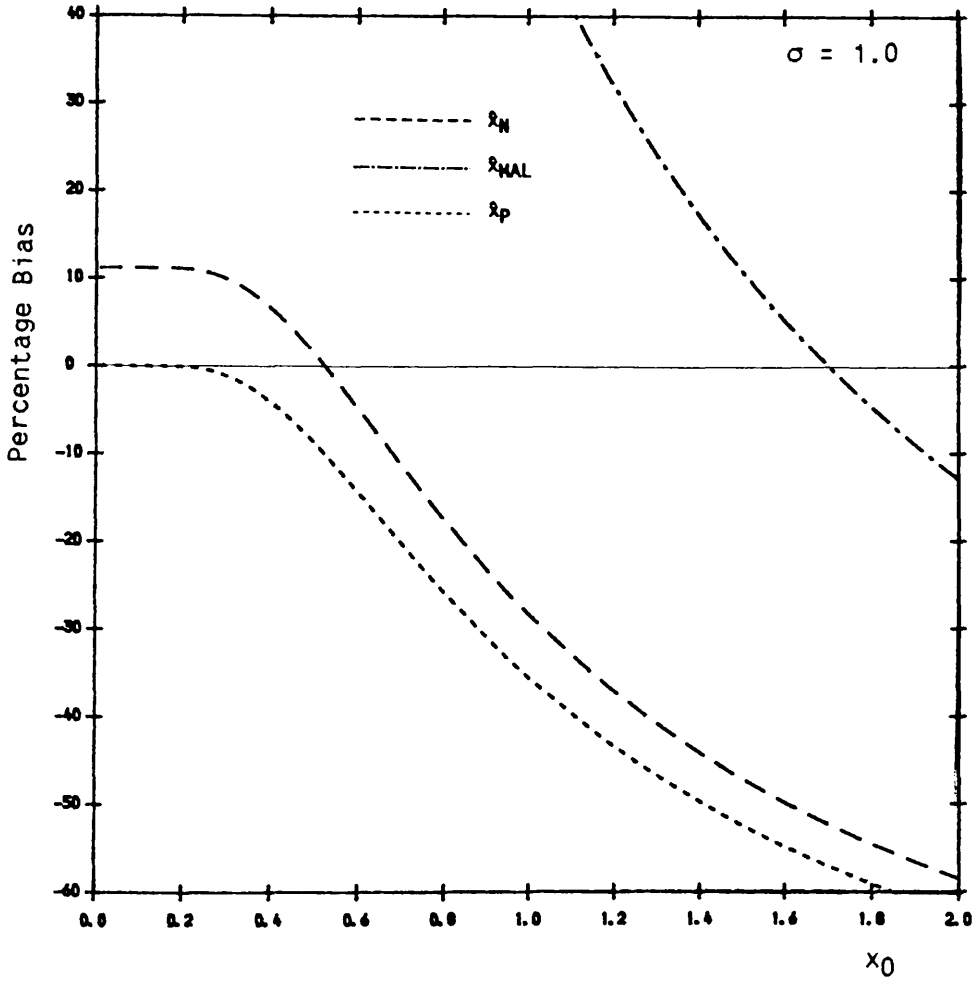

Figure (3.10)

Percentage bias of $\hat{x}_N$, $\hat{x}_{MAL}$ and $\hat{x}_P$ as a function of true distance, x_0 ($\sigma = 1$)

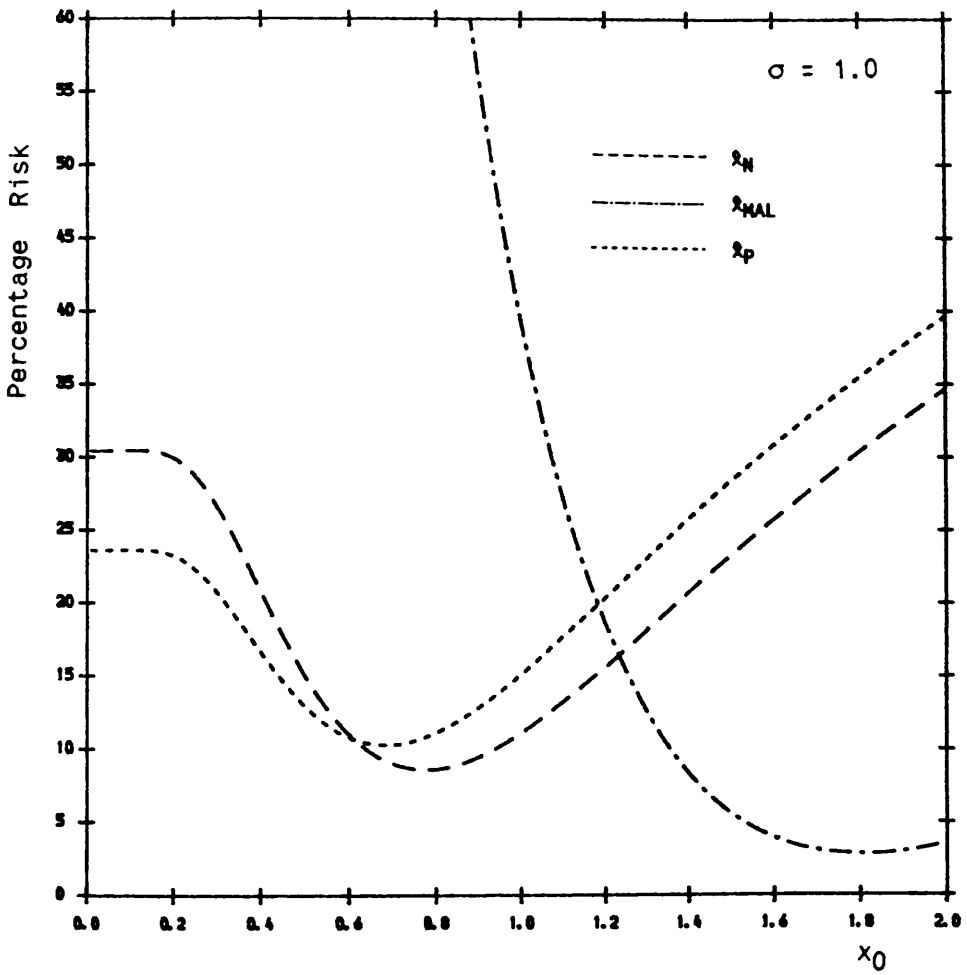

Figure (3.11)

Percentage risk of $\hat{x}_N$, $\hat{x}_{MAL}$ and $\hat{x}_P$ as a function of true distance, x_0 ($\sigma = 1$)

On the other hand, the *positive* percentage bias of the 'naive' estimator at small x_0 is not due to selection but rather to the non-linear relationship between distance and magnitude. This positive bias as x_0 tends to zero disappears if we consider estimates not of distance, but of *log* distance (i.e. the distance modulus). Thus, a 'naive' estimator of $\log x_0$, defined in the obvious way, is unbiased as $x_0 \to 0$; see section (4.5.1). This will no longer hold, however, if there is luminosity selection at small x_0 (c.f. section 4.5.1).

We can see, nonetheless, that the percentage bias of $\hat{x}_P$ does tend to zero as x_0 tends to zero, as expected, and remains essentially zero for all $x_0 < 0.2$ ($\sigma = 1$) or for all $x_0 < 0.4$ ($\sigma = 0.5$). Note also, however, that $\hat{x}_P$ has the worst percentage bias and risk at large distances.

The calculations of bias and risk for the maximum likelihood estimator introduce a significant problem. The integral expression for the expected value of $\hat{x}_{ML}$, as given in equation (3.23), does not converge for any value of x_0. $E(\hat{x}_{ML}|x_0)$ is, therefore, infinite for all true distances, so that this estimator necessarily has infinite percentage bias and risk. Thus, although $\hat{x}_{ML}$ is well-defined for all $m < m_L$, the form of its distribution function close to the magnitude limit renders bias and risk meaningless as a method of comparison with other estimators. This is not an uncommon problem in statistics where many distributions, such as the Cauchy distribution for example, have theoretical moments which are all infinite (c.f. Hoel, 1962). Indeed, Cauchy-type distributions are often used in astrophysical modelling of e.g. the profiles of emission lines in stellar atmospheres (c.f. Rybicki

and Lightman, 1979).

It is possible to modify the definition of the maximum likelihood estimator so that its moments are defined and take on finite values. There are several ways to achieve this; for example, one can define a new estimator, $\hat{x}_{ML}^*$, with finite bias and risk as follows:-

$$\hat{x}_{ML}^* = \begin{cases} \hat{x}_{ML} & \text{for } \hat{x}_{ML} < \alpha \\ \alpha & \text{for } \hat{x}_{ML} \geqslant \alpha \end{cases} \qquad (3.30)$$

Thus whenever $m_L - m$ is less than some prescribed value, ϵ say, the distance inferred by $\hat{x}_{ML}^*$ is simply put equal to some fixed constant, α, which is the distance inferred by the *unmodified* maximum likelihood estimator when $m_L - m$ is equal to ϵ.

The value of α chosen will determine how closely to the magnitude limit observations are allowed to be taken before the modified form of the estimator is used. One would therefore expect the bias and risk of $\hat{x}_{ML}^*$ to be dependent on α. Figures (3.12) and (3.13) show the percentage bias and risk of $\hat{x}_{ML}^*$ for different values of α and for $\sigma = 1$. It is clear from these figures that the choice of α will greatly affect the range of true distances for which the percentage bias and risk of this estimator are small. In particular, we can see that the percentage risk takes its minimum value at $x_0 \simeq \alpha$ but increases sharply at smaller distances. Similarly, the percentage bias is small only within a narrow range of distances around $x_0 \simeq 0.8\alpha$. It would seem, therefore, that at least some order of magnitude knowledge of the true distance is very important before an appropriate choice of α can be made, since a poor choice may result in

a large systematic error.

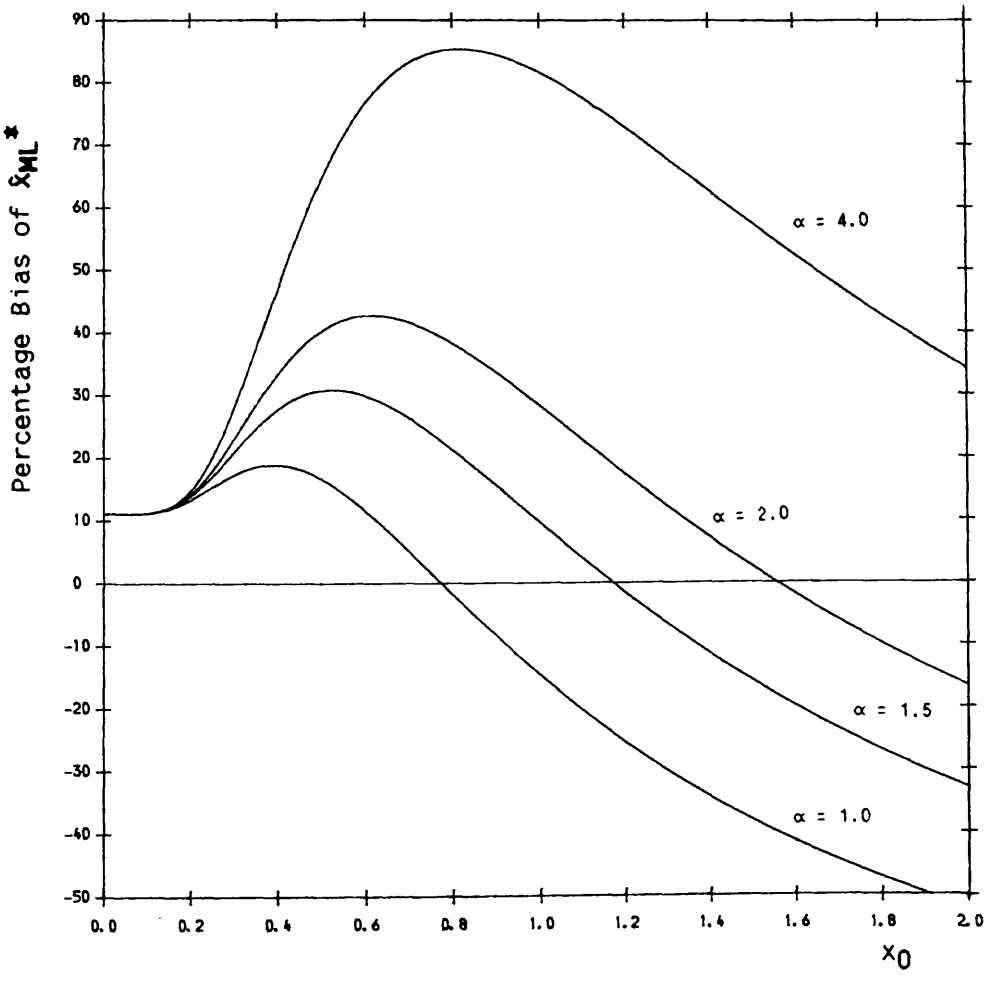

Figure (3.12)

Percentage bias as a function of true distance, x_0, of the maximum likelihood estimator, $\hat{x}_{ML}^*$, modified by an upper limit, α, shown for values of α = 1.0, 1.5, 2.0 and 4.0 (σ = 1)

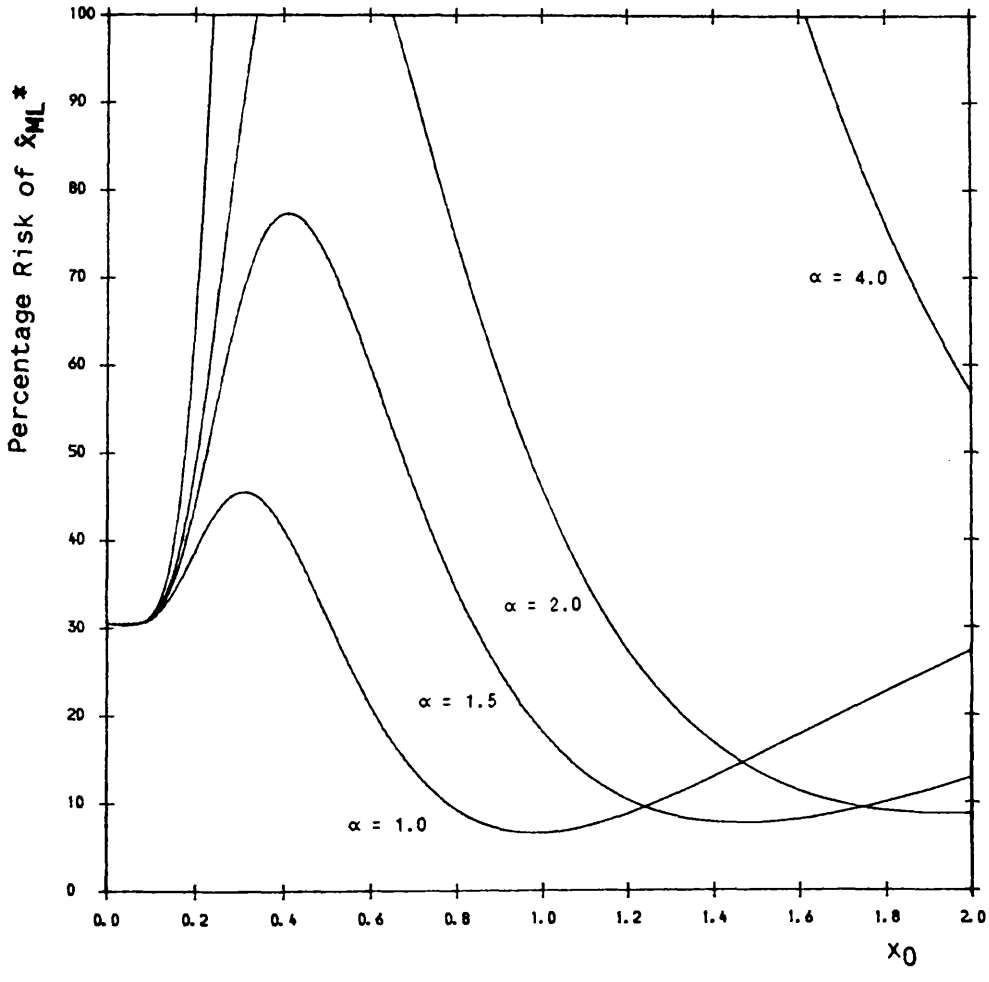

Figure (3.13)

Percentage risk as a function of true distance, x_0, of the maximum likelihood estimator, $\hat{x}_{ML}^*$, modified by an upper limit, α, shown for values of α = 1.0, 1.5, 2.0 and 4.0 (σ = 1)

Another method of modifying $\hat{x}_{ML}$, and indeed any estimator, is to reject galaxies whose magnitudes lie close to m_L (within Δm of m_L, say). The conditional distribution, $\varsigma(m|x_0)$, would then be redefined as non-zero only on the interval $(-\infty, m_L - \Delta m)$, with the appropriate renormalisation. This will, of course, change the expected value of each estimator, and in the case of $\hat{x}_{ML}$ will yield a finite result. Figures (3.14) to (3.17) compare the percentage bias and risk of all estimators for different values of σ, after rejecting observations closer than 0.5 magnitudes to m_L.

A number of points are clear from these graphs. Note that the choice of estimator with least percentage bias or risk is strongly dependent on the unknown true distance. Generally speaking, the bias and risk of $\hat{x}_P$ are least at small distances but greatest at large distances, while the opposite is true for $\hat{x}_{MAL}$. Furthermore, at any given distance the least-biased estimator may not have the smallest risk. For example, at $x_0 \simeq 0.7$ the bias of $\hat{x}_{ML}$ is smallest but $\hat{x}_N$ has the least percentage risk. Another interesting feature is the fact that the minimum bias and minimum risk of any estimator do not, in general, occur at the same distance. Moreover, the range of distances for which each estimator has small bias and risk depends on the value of σ, and also indirectly on m_L and M_0 since these define the unit of distance.

Clearly, therefore, it is not easy to identify which of these estimators is the most appropriate to use, since this will depend very much on the true distance of the observed galaxy. One might select the 'best' estimator by assuming the true distance to lie within a

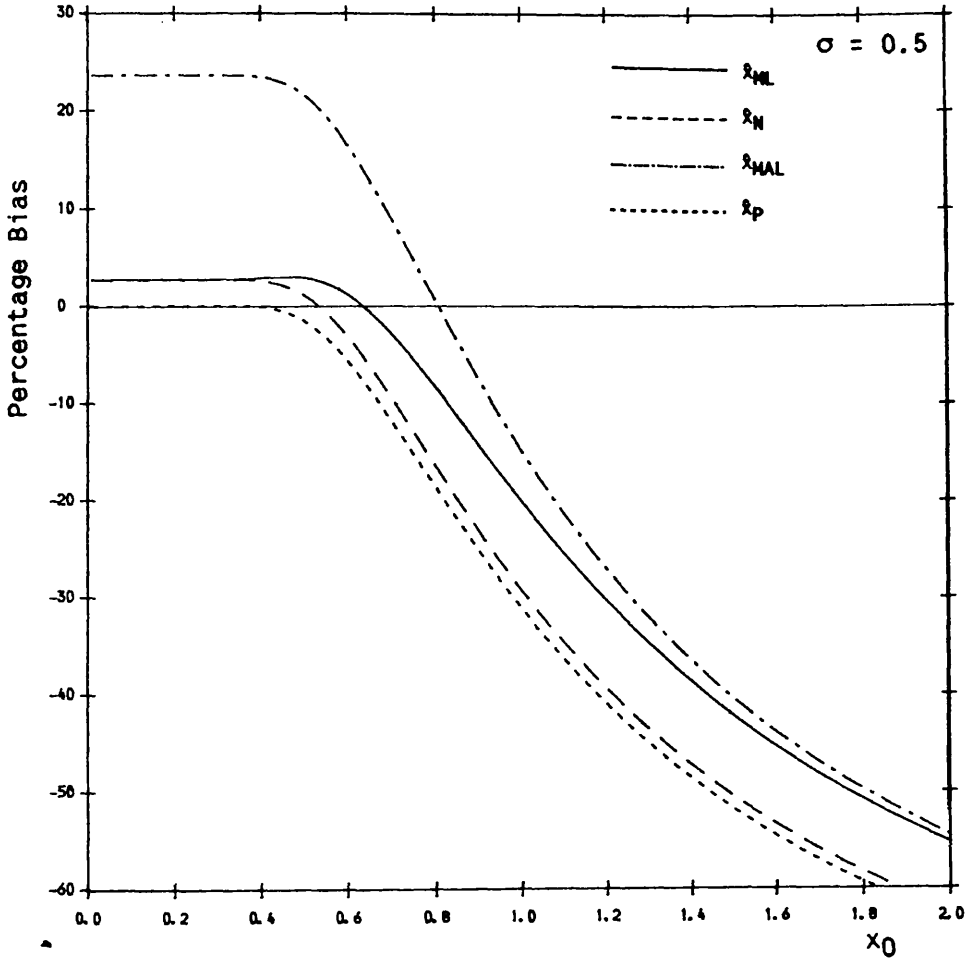

Figure (3.14)

Percentage bias of $\hat{x}_{ML}$, $\hat{x}_N$, $\hat{x}_{MAL}$ and $\hat{x}_P$ as a function of true distance, x_0, after rejecting observations within 0.5 mag of m_L (σ = 0.5)

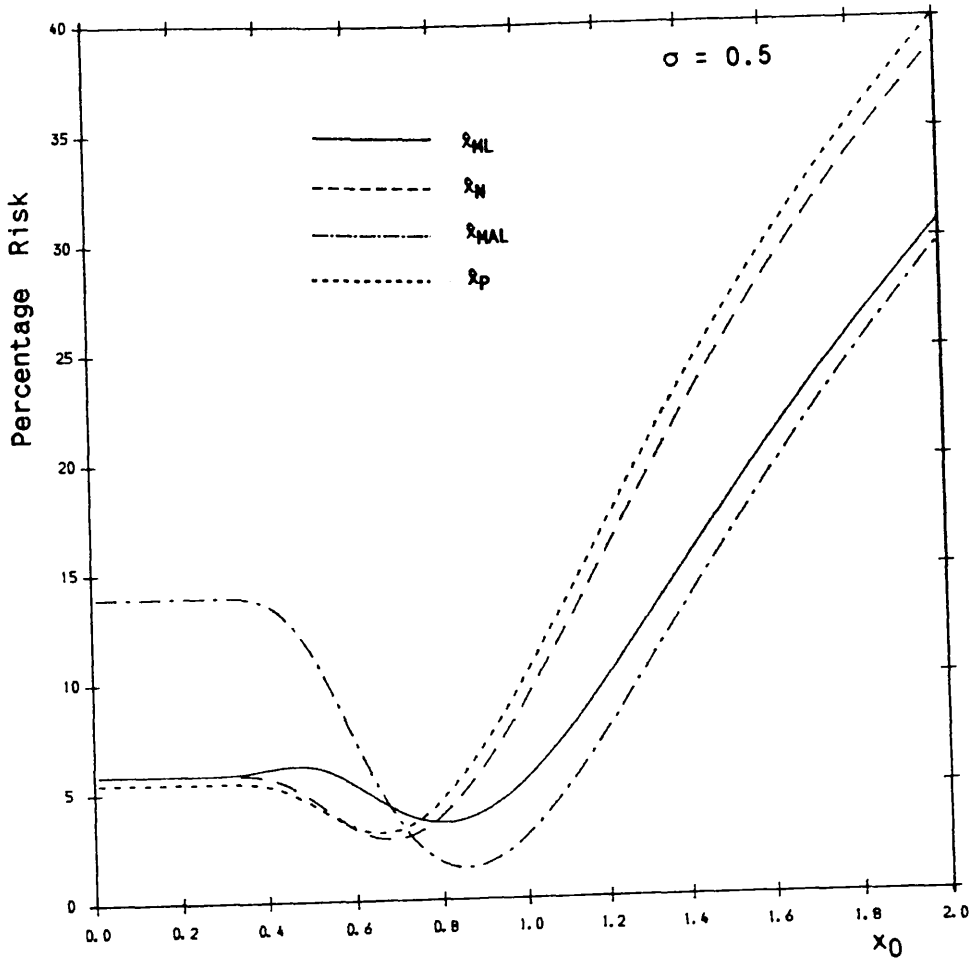

Figure (3.15)

Percentage risk of $\hat{x}_{ML}$, $\hat{x}_{N}$, $\hat{x}_{MAL}$ and $\hat{x}_{P}$ as a function of true distance, x_0, after rejecting observations within 0.5 mag of m_L ($\sigma = 0.5$)

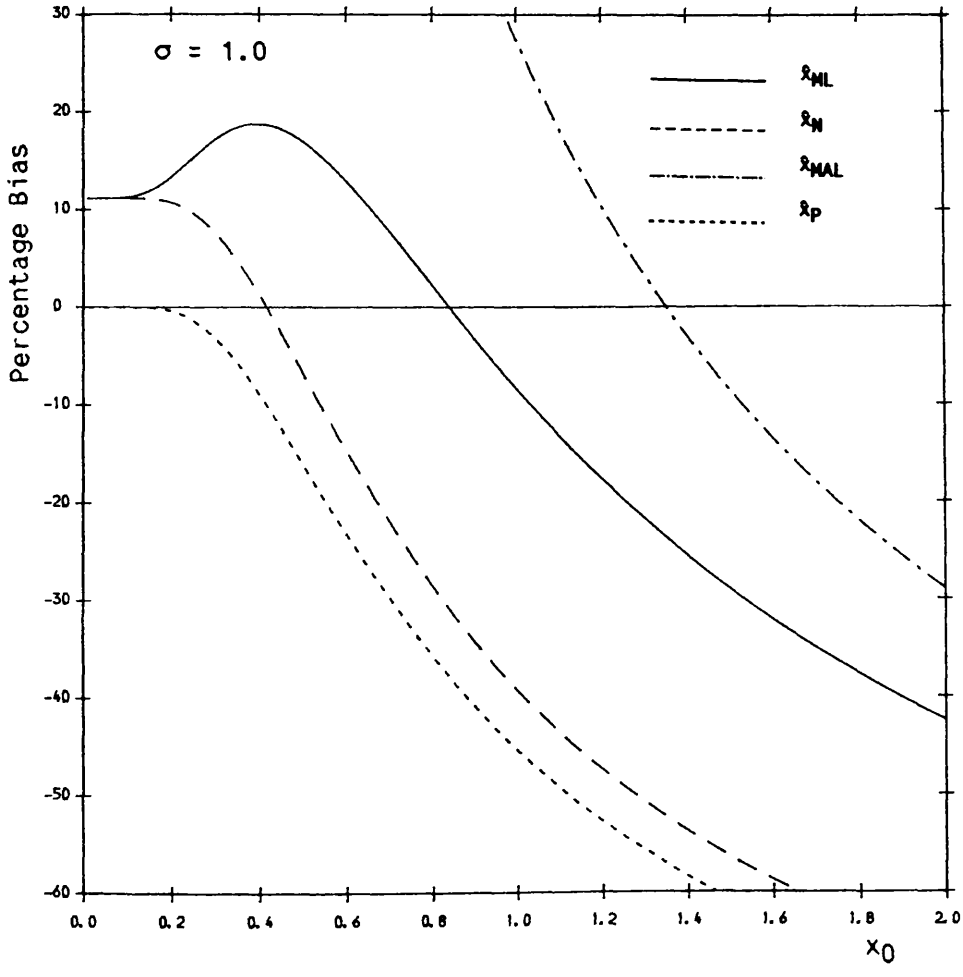

Figure (3.16)

Percentage bias of $\hat{x}_{ML}$, $\hat{x}_N$, $\hat{x}_{MAL}$ and $\hat{x}_P$ as a function of true distance, x_0, after rejecting observations within 0.5 mag of m_L ($\sigma = 1$)

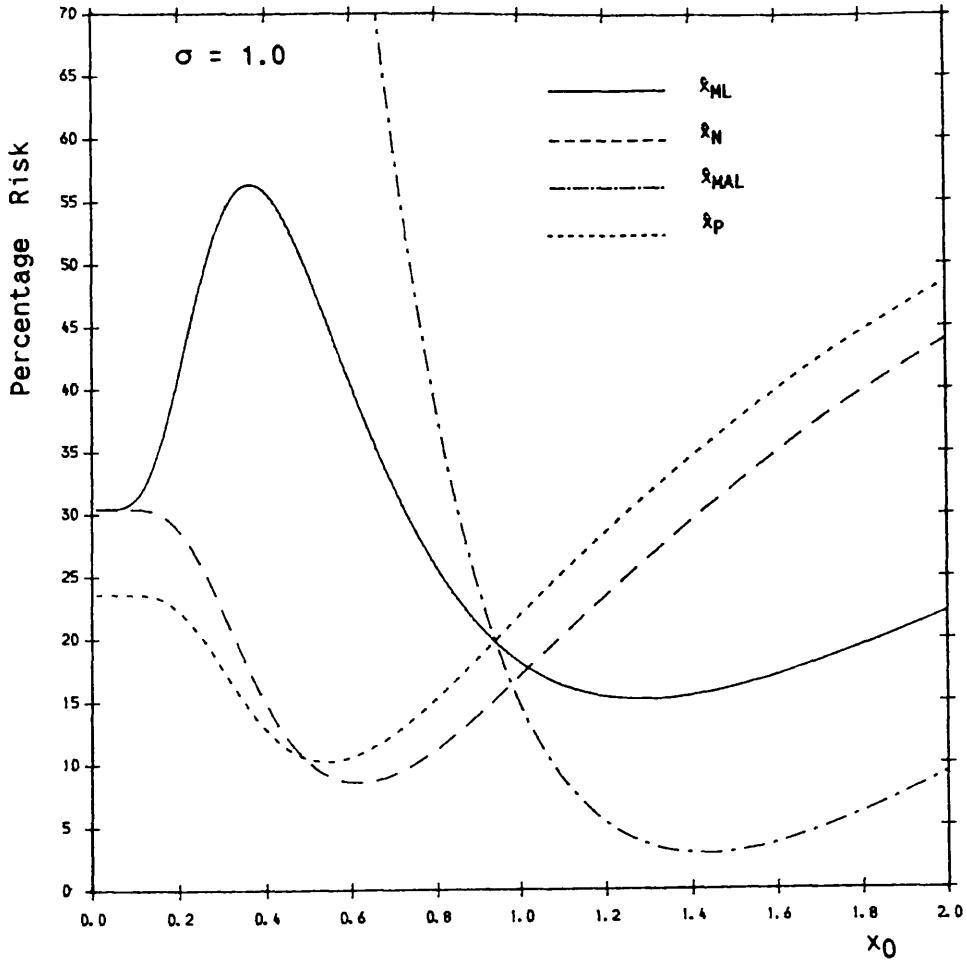

Figure (3.17)

Percentage risk of $\hat{x}_{ML}$, $\hat{x}_N$, $\hat{x}_{MAL}$ and $\hat{x}_P$ as a function of true distance, x_0, after rejecting observations within 0.5 mag of m_L ($\sigma = 1$)

particular range. For example, if we assume that $x_0 < 1.2$ then $\hat{x}_{MAL}$ appears to be the best choice of estimator. If, however, the true distance were considerably less than unity, then using $\hat{x}_{MAL}$ would result in a large systematic error.

3.5 Confidence Intervals for Distance Estimates

If we are to use our estimators in a practical setting, for the analysis of real data, then it is important that we are able to assign some error to the estimates which we obtain. Thus far we have considered only point estimators, for which the risk provides some indication of their precision. One could, therefore, use the risk to assign a distance error to an observation were it not for the fact that the risk is a function of the *true* distance, and so is not known precisely. One possible solution to this problem is to assign an error by computing the risk at the *estimated* distance. However, this may lead to spurious results if the risk changes rapidly with true distance, as is frequently the case with the estimators which we have studied.

A more rigorous approach is to directly obtain an interval estimate for the true distance. We will now describe a procedure for constructing a $(1-\alpha)100\%$ confidence interval for the distance. The formal definition of such an interval states that the probability of the interval containing the *true* distance, x_0, is precisely $(1-\alpha)$ *no matter what that true distance is.*

Consider a random variable, **z**, whose distribution, **X**, is a function only of the unknown parameter, x_0. Clearly any of the point estimators which we have studied are suitable choices for **z**. One simple way in which a confidence interval can be constructed (c.f. Mood and Graybill, 1974; Simmons and Stewart, 1985) is to find z_1 and z_2 *for each* x_0 such that:-

$$\int_0^{z_1} \mathbf{X}(z|x_0)dz = \lambda \qquad (3.31)$$

and

$$\int_0^{z_2} \mathbf{X}(z|x_0)dz = (1-\alpha) + \lambda \qquad (3.32)$$

where $\lambda \in [0,\alpha)$ but is otherwise arbitrary.

We assume that both z_1 and z_2 are monotonic functions of x_0; an example of this is shown in figure (3.18), where we plot z_1 and z_2 as functions of x_0.

For any value of $\mathbf{z} = z_*$, we may draw a line parallel to the x_0 axis and find $x_2 = z_2^{-1}(z_*)$ and $x_1 = z_1^{-1}(z_*)$, the x_0 coordinates of the points of intersection with the curves z_2 and z_1 respectively, as shown in figure (3.18). Now, whatever the true distance, x_0, may be it follows from the construction of z_1 and z_2 that

$$Pr(\, z_1(x_0) < \mathbf{z} < z_2(x_0) \,) = (1-\alpha) \qquad (3.33)$$

However, $z_1(x_0) < z < z_2(x_0)$ if and only if $x_2(z) < x_0 < x_1(z)$, for any possible value of the random variable, z. Thus, the interval (x_2,x_1) will form a $(1-\alpha)100\%$ confidence interval for the distance x_0.

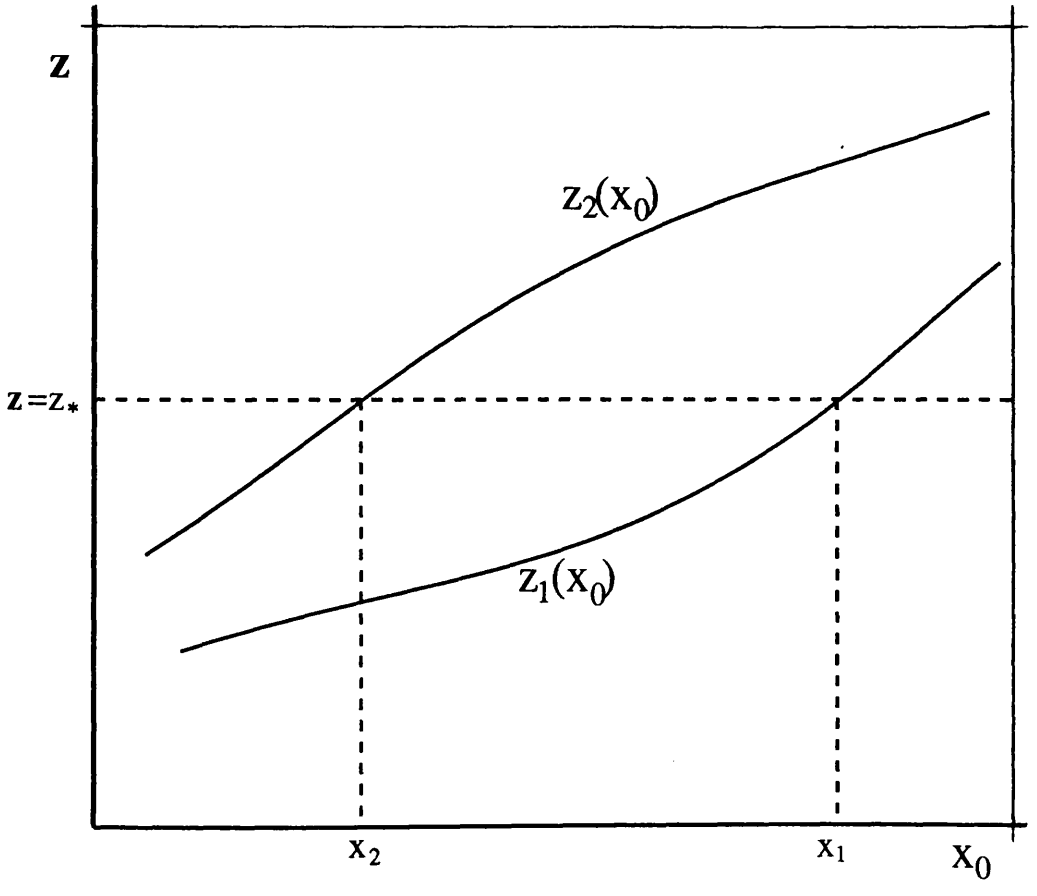

Figure (3.18)

Example of typical upper and lower confidence interval curves, $z_1(x_0)$ and $z_2(x_0)$, as a function of true distance, x_0

As an illustration we now demonstrate how these ideas may be applied using as our random variable the 'naive' estimator, $\hat{x}_N$, the distribution of which depends only on the true distance, x_0, for a given σ. Figure (3.19) shows graphs of $z_1(x_0)$ and $z_2(x_0)$ at the 68% (1σ), 95% (2σ) and 99% (3σ) level, for $\sigma = 1$. The value of λ used in each case was $\frac{1}{2}\alpha$, so that the two tails of the distribution were of equal area. For any estimate, $\hat{x}_N = x_*$, of distance, one may read off a confidence interval for the distance from the points of intersection of the appropriate curves with the line $\hat{x}_N = x_*$. For example, suppose that the apparent magnitude of a galaxy is measured to be 13.6, and that the limiting magnitude is 15. From figure (1a) $\hat{x}_N \simeq 0.5$, so that a 68% confidence interval for the true distance is given by figure (10) to be [0.3, 0.9].

It is easily seen from figure (3.19) that for observations close to the limiting magnitude the confidence intervals will quickly become extremely large. Furthermore, the estimated distance will not lie within the confidence interval for x_0; this is consistent with the large negative bias of the point estimator, $\hat{x}_N$, at large distances.

Clearly, however, the confidence intervals constructed using $\hat{x}_N$ are in no way unique. Any of the estimators defined in this chapter (or, indeed, any suitable random variable, whether a distance estimator or not!) could have been used in place of $\hat{x}_N$, according to the prescription which we have described. In the next chapter we will see further examples of confidence intervals constructed from more reliable distance estimators.

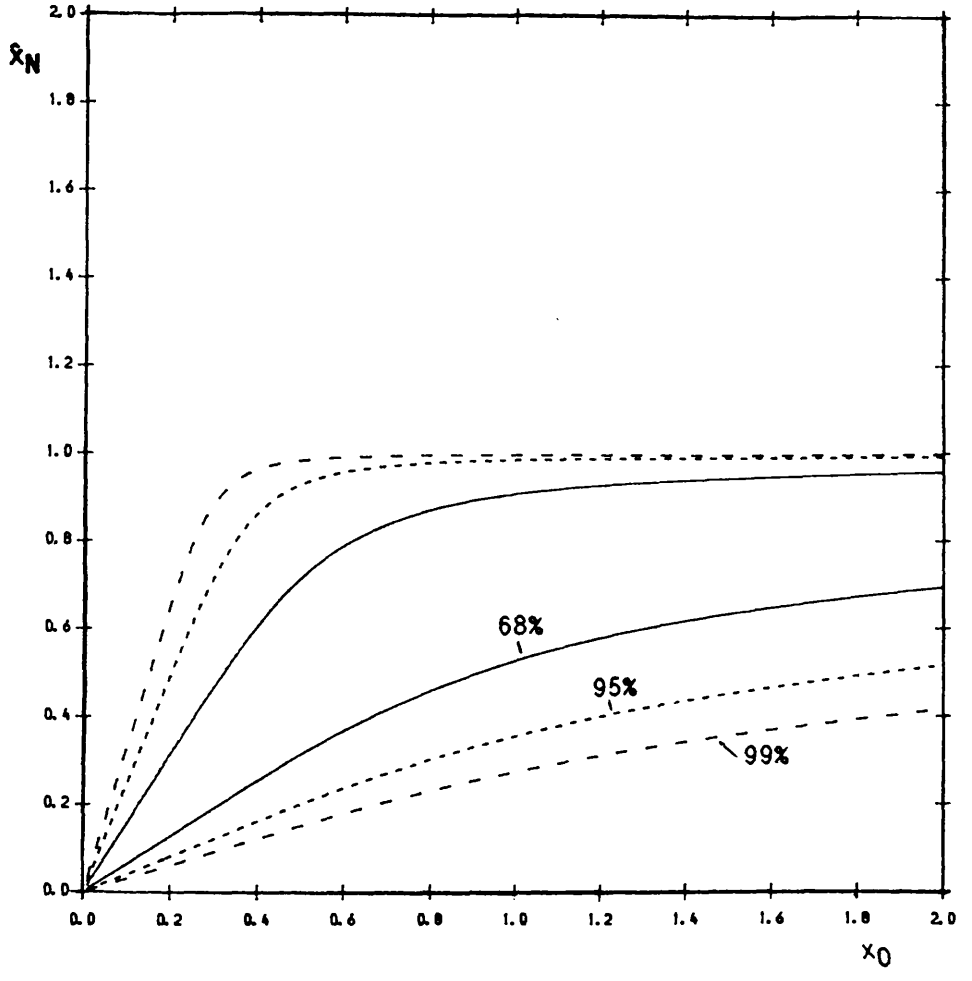

Figure (3.19)

Confidence interval curves for the true distance, x_0, at the 68% (1σ), 95% (2σ) and 99% (3σ) levels, computed from the distribution function of the estimator $\hat{x}_N$ ($\sigma = 1$)

3.6 Summary and Conclusions

In this chapter we have presented a statistically rigorous method for defining point and interval estimators of distance, and for studying their distributions and properties. We have illustrated the application of this method to a simple case: where distances are estimated using only measurements of apparent magnitude, and for a Gaussian LF and Heaviside selection function. Our analysis may easily be adapted to derive estimates of other functions of distance, or to incorporate a different LF or selection function - e.g. a Schechter LF (c.f. Schechter, 1976) or a sigmoid-type selection function (c.f. Teerikorpi, 1975).

We have compared the distributions of a number of different estimators and, in particular, have calculated the percentage bias and risk of each estimator as a function of true distance. Our results demonstrate that the problem of choosing an 'optimal' distance estimator has no straightforward solution. The fundamental difficulty is that the properties of an estimator will, in general, depend on the true distance, which is unknown. The estimator which we select as 'best' - by whichever criterion we choose to adopt (e.g. minimum bias, minimum risk or any other appropriate statistical measure) - may be a poor choice if the true distance of the observed galaxy does not lie within a preferred range.

Using precisely the same statistical methods as we have introduced in this chapter, we will now extend our analysis to the estimation of distance from two observables.

4. ESTIMATION OF DISTANCE USING TWO OBSERVABLES

4.1 Introduction

The aim of this chapter is to extend the analysis introduced in chapter (3) to the case where one defines distance estimators which are functions of two observables: e.g. by combining with the apparent magnitude measurements of another observable quantity such as the apparent diameter or 21cm line width of a galaxy. Following closely the statistical formulation presented in the preceding chapter we will derive expressions for the distribution of two observables at a given true distance, taking into account luminosity selection effects, and use this distribution to define and investigate a number of estimators of that distance.

Intuitively there would seem good reason to suppose that these estimators might be more reliable than those considered in chapter (3) since one is utilising more information about the galaxies which one is observing. We have seen in chapter (2), however, that opinion is divided over precisely how best to combine measurements of several observables. Recall in particular that there has been some disagreement over the choice of linear regression most appropriate for deriving distances from e.g. the Tully-Fisher or D_n-σ relations (c.f. Tully, 1988; Dressler *et al*, 1987). We will show that the different methods of regression used in the literature each correspond to distance estimators which arise naturally from our rigorous formulation. Thus, by computing the distribution, bias and risk of these - and other - estimators we will be able to assess critically the

relative merits of each and clarify the question of how best to estimate distances from measurements of several observables.

4.2 The Observed Distribution of m and P

Let the absolute magnitude, **M**, and position, $\underline{r}$, of a galaxy be random variables. Introduce a third random variable, **P**, which denotes some intrinsic physical characteristic of the galaxy such that the measured value of **P** provides information on the value of **M**; i.e. **M** and **P** are correlated. This notation is used by Teerikorpi (1984). Suppose, however, that neither **M** nor **P** is correlated with $\underline{r}$ so that we may introduce $\Psi(\mathbf{M},\mathbf{P})$, the intrinsic joint distribution of **M** and **P**, which is independent of position. Let $N(\mathbf{M},\mathbf{P},\underline{r})d\mathbf{M}d\mathbf{P}d\mathbf{V}$ denote the actual number of galaxies in volume $d\mathbf{V}$ at position $\underline{r}$ with absolute magnitudes in the range **M** to **M**+$d\mathbf{M}$ and **P** values in the range **P** to **P**+$d\mathbf{P}$. Clearly we may write:-

$$N(\mathbf{M},\mathbf{P},\underline{r})d\mathbf{M}d\mathbf{P}d\mathbf{V} = \Psi(\mathbf{M},\mathbf{P})n(\underline{r})d\mathbf{M}d\mathbf{P}d\mathbf{V} \qquad (4.1)$$

where $n(\underline{r})$ is the number density of galaxies at position $\underline{r}$.

Consider now the joint distribution, $\rho(\mathbf{M},\mathbf{P},\underline{r})$, of **M**, **P** and $\underline{r}$ for observable galaxies in a sample subject to selection effects - as described by a selection function, $S(\mathbf{M},\mathbf{P},\underline{r})$, defined in a similar manner to the function, $S(\mathbf{M},\underline{r})$, of chapter (3). Note that $S(\mathbf{M},\mathbf{P},\underline{r})$ is independent of direction; i.e. we may write $S(\mathbf{M},\mathbf{P},\underline{r}) = S(\mathbf{M},\mathbf{P},|\underline{r}|)$. Substituting from equation (4.1), it follows that $\rho(\mathbf{M},\mathbf{P},\underline{r})$ is given by:-

$$\rho(M,P,\underline{r}) = \frac{\Psi(M,P)n(\underline{r})S(M,P,|\underline{r}|)}{\iiint \Psi(M,P)n(\underline{r})S(M,P,|\underline{r}|)dMdPdV} \qquad (4.2)$$

It now follows from equation (4.2) that the conditional distribution, $\xi(M,P|r_0)$, of absolute magnitude, **M**, and **P** at a given distance, r_0, for observable galaxies is given by:-

$$\xi(M,P|r_0) = \frac{\Psi(M,P)S(M,P,r_0)}{\iint \Psi(M,P)S(M,P,r_0)dMdP} \qquad (4.3)$$

Note that, as for the magnitude-only case, this distribution is independent of the local density, $n(\underline{r})$.

Finally, we change variables in equation (4.3) to apparent magnitude, **m**, and **P**. In general one may be unable to measure **P** directly: for example, if **P** were equal to the absolute diameter of the galaxy then one would observe instead the apparent angular diameter - so that both **P** and **M** would require a change of variable in order to express equation (4.3) in terms of measurable quantities. We will generalise to this case later; for the moment, however, suppose that **P** is measurable directly - as would be the case with, for example, the 21cm line width. Thus, the conditional distribution, $\zeta(m,P|r_0)$, of **m** and **P** at true distance, r_0, is given by:-

$$\Psi(m,P|r_0) = \frac{\Psi(m-5\log r_0-25,P)S(m,P)}{\iint \Psi(m-5\log r_0-25,P)S(m,P)dmdP} \qquad (4.4)$$

We will assume that $\Psi(M,P)$ has a bivariate normal distribution, following e.g. Teerikorpi (1984), viz:-

$$\Psi(\mathbf{M},\mathbf{P}) = (2\pi\sigma_M\sigma_P\sqrt{(1-\rho^2)})^{-1} \exp[-Q(\mathbf{M},\mathbf{P})/2(1-\rho^2)] \qquad (4.5)$$

where the quadratic form, $Q(\mathbf{M},\mathbf{P})$, in the variables **M** and **P** is given by:-

$$Q(\mathbf{M},\mathbf{P}) = (\mathbf{M}-M_0)^2/\sigma_M^2 + (\mathbf{P}-P_0)^2/\sigma_P^2 - 2\rho(\mathbf{M}-M_0)(\mathbf{P}-P_0)/\sigma_M\sigma_P.$$

Here M_0 and P_0 are respectively the mean values of **M** and **P**, σ_M^2 and σ_P^2 are the variances of **M** and **P**, and ρ is the correlation coefficient for **M** and **P**, which lies in the range [-1,1].

We will consider here the case where the selection function depends only on the apparent magnitude; i.e. there is no selection on **P** and one may sample **P** completely at any apparent magnitude. This is the key assumption in the 'bias-free' recipe of Schechter (1980) and later authors, and also leads to an algebraically simpler analysis. Furthermore we will firstly assume that the magnitude selection effects are described by a Heaviside function at some magnitude limit, m_L, as in equation (3.6). Thus, we may write:-

$$S(\mathbf{m},\mathbf{P}) = \begin{cases} 1 & \text{if } \mathbf{m} < m_L \\ & \text{otherwise} \end{cases} \qquad (4.6)$$

Substituting for Ψ and S, it may easily be seen that equation (4.4) reduces to:-

$$\varsigma(\mathbf{m},\mathbf{P}|r_0) = \begin{cases} \dfrac{\exp[-Q(\mathbf{m}-5\log r_0-25,\mathbf{P})/2(1-\rho^2)]}{\sqrt{2\pi(1-\rho^2)}\sigma_M\sigma_P\,\Phi(m_L-5\log r_0-25-M_0/\sigma_M)} & \mathbf{m} < m_L \\ 0 & \text{otherwise} \end{cases} \qquad (4.7)$$

Where Q is as defined for equation (4.5).

Note that, because there is no selection on **P**, the argument of the cumulative normal distribution, Φ, in equation (4.7) is the same as for the magnitude-only case in equation (3.7); i.e. the normalisation of the conditional distribution, ς, is the same for both cases.

We can further simplify equation (4.7) by re-expressing the distribution using the scaled distance unit, r_L, introduced in chapter (3). Upon substitution from equation (3.16) we find that the conditional distribution of observable galaxies at scaled true distance, x_0, may be written as:-

$$\varsigma(\mathbf{m},\mathbf{P}|x_0) = \begin{cases} \dfrac{\exp[-Q(\mathbf{m},\mathbf{P})/2(1-\rho^2)]}{\sqrt{2\pi(1-\rho^2)}\sigma_M\sigma_P\ \Phi(-5\log x_0/\sigma_M)} & \mathbf{m} < m_L \\ 0 & \text{otherwise} \end{cases} \qquad (4.7)$$

where Q is now given by:-

$$Q(\mathbf{m},\mathbf{P}) = (\mathbf{m}-m_L-5\log x_0)^2/\sigma_M^2 + (\mathbf{P}-P_0)^2/\sigma_P^2 - 2\rho(\mathbf{m}-m_L-5\log x_0)(\mathbf{P}-P_0)/\sigma_M\sigma_P$$

Note that if we integrate out equation (4.7) over **P**, to determine the *marginal* distribution of **m** at a given true distance, x_0, we obtain precisely the conditional distribution $\varsigma(\mathbf{m}|x_0)$ of equation (3.17), as derived for the magnitude-only case.

Figures (4.1) to (4.4) show plots of the joint conditional distribution, $\varsigma(\mathbf{m},\mathbf{P}|x_0)$, for different values of the scaled true

distance, x_0. The magnitude limit shown is $m_L = 15$ and the value of ρ is taken to be -0.8; this is a typical value of the correlation coefficient measured for e.g. the Tully-Fisher relation (c.f. Tully, 1988 and figure 2.4 of section 2.3). The surfaces are shown in isometric projection and have been normalised to the same peak height by the graphics routine. The **P** axis scale is in units of σ_P.

Note that in figure (4.1), for $x_0 = 0.25$, the joint conditional distribution is still to a good approximation bivariate normal, since there is no appreciable Malmquist Bias at this distance. It is clear from the other figures, however, that at larger true distances the effects of selection cause considerable distortion to the shape of $\varsigma(\mathbf{m},\mathbf{P}|x_0)$ - although it should be noted that this distortion occurs only in the **m** dimension, and the *conditional* distribution of **P** at any given value of **m** (which corresponds pictorially to a vertical section through the surface parallel to the **P** axis) is still a normal distribution - even at large true distances - as may be easily verified analytically.

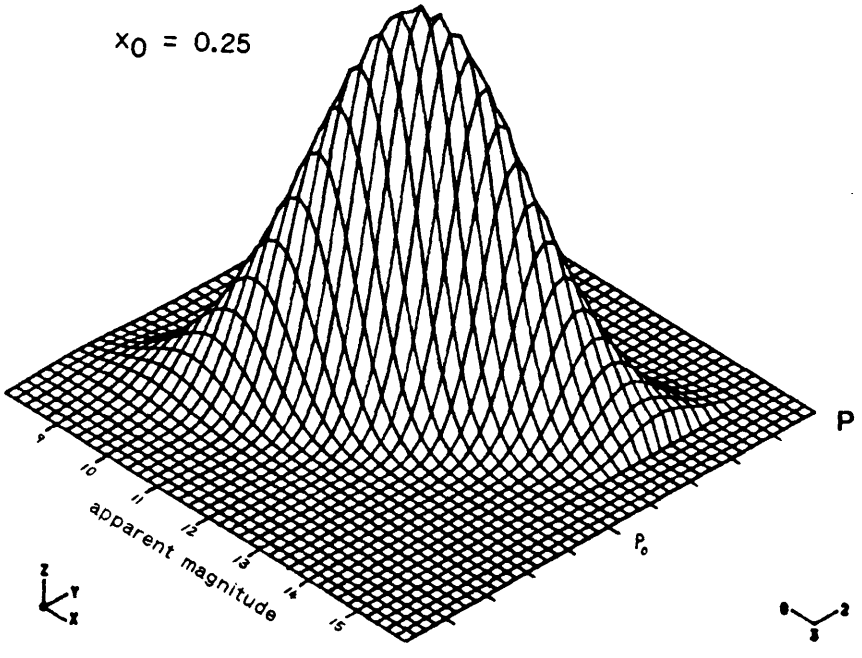

Figure (4.1)

Probability density function, $\varsigma(\mathbf{m},\mathbf{P}|x_0)$, of observable galaxies at (scaled) true distance $x_0 = 0.25$ ($\sigma_M = 1$, $\sigma_P = 0.1$, $\rho = -0.8$, $m_L = 15$)

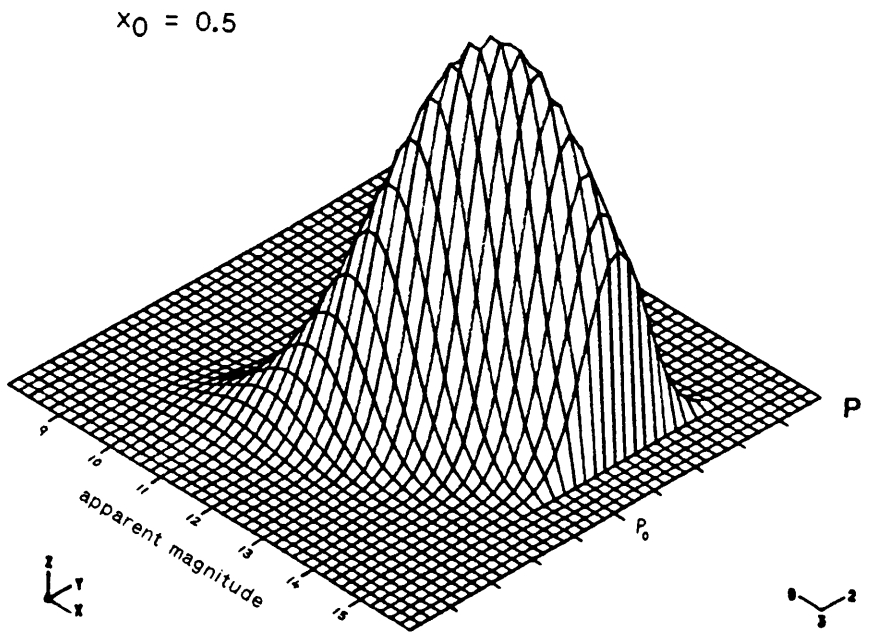

Figure (4.2)

Probability density function, $\varsigma(\mathbf{m},\mathbf{P}|x_0)$, of observable galaxies at (scaled) true distance $x_0 = 0.5$ ($\sigma_M = 1$, $\sigma_P = 0.1$, $\rho = -0.8$, $m_L = 15$)

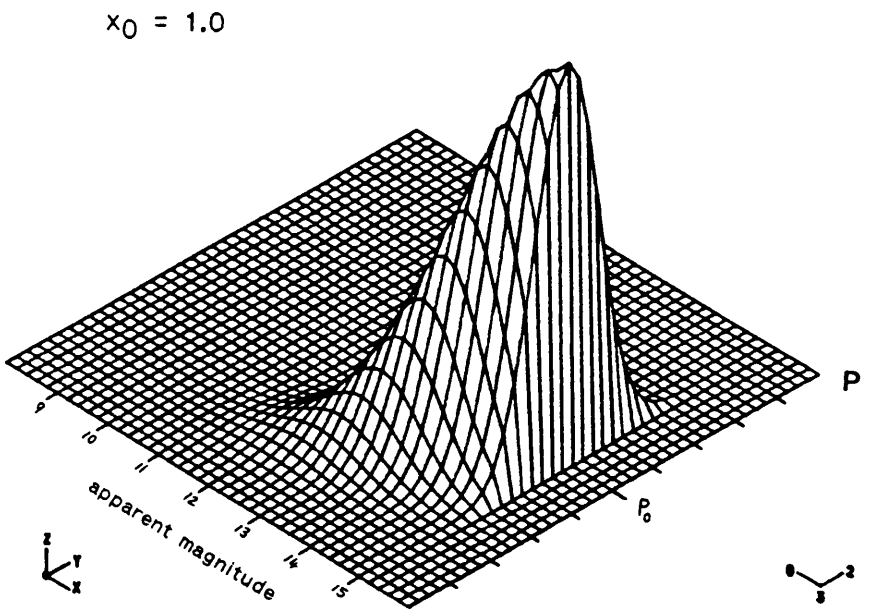

Figure (4.3)

Probability density function, $\varsigma(\mathbf{m},\mathbf{P}|x_0)$, of observable galaxies at (scaled) true distance $x_0 = 1.0$ ($\sigma_M = 1$, $\sigma_P = 0.1$, $\rho = -0.8$, $m_L = 15$)

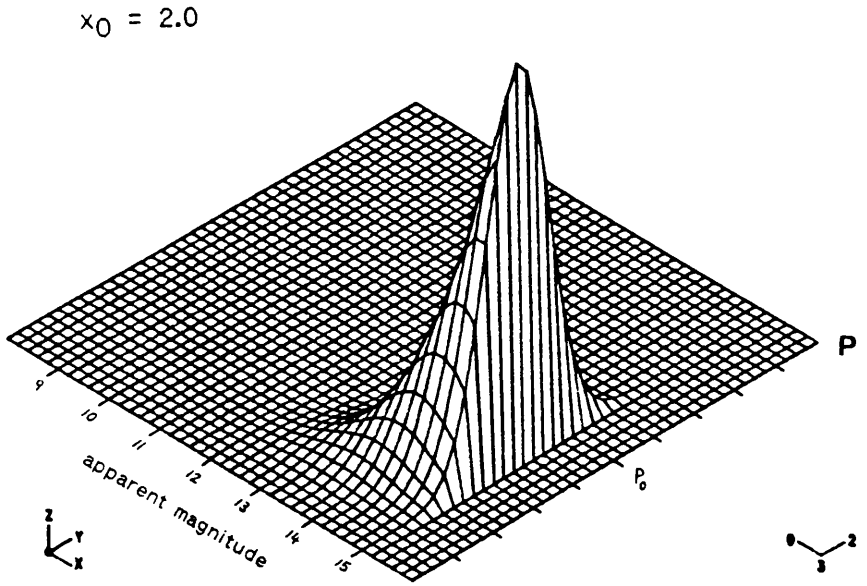

Figure (4.4)

Probability density function, $\varsigma(\mathbf{m},\mathbf{P}|x_0)$, of observable galaxies at (scaled) true distance $x_0 = 2.0$ ($\sigma_M = 1$, $\sigma_P = 0.1$, $\rho = -0.8$, $m_L = 15$)

4.3 Definitions of Distance Estimators

In this section we will introduce a number of different estimators of distance - or, more precisely, estimators of *log* distance, μ, although clearly from any estimator of μ, $\hat{\mu}$ say, one may immediately define an an equivalent distance estimator, $\hat{r}$, viz:-

$$\hat{r} = 10^{\hat{\mu}} \tag{4.8}$$

The distinction between $\hat{r}$ and $\hat{\mu}$ becomes non trivial only when one considers the distribution of the estimators: in particular, an unbiased estimator of log distance will not, in general, correspond in an unbiased estimator of distance. The resulting bias is usually regarded as a small effect, however, (c.f. Sandage and Tammann, 1975) and - as we have already commented - almost all previous discussions in the literature have defined distances to galaxies and discussed the bias of estimators in terms of log distance. We will, therefore, adopt this convention, although remaining aware of the potential source of systematic error which it introduces.

It is convenient to describe log distances using the same scaled distance units which we introduced in chapter (3). Consider a galaxy at true distance, r MPc - i.e. at true distance, x, in scaled units, where $x = r/r_L$. Let ω denote the true log distance of the galaxy in scaled units. Then it follows that:-

$$\omega \equiv \log x = \mu - \log r_L \tag{4.9}$$

Similarly, we introduce the notation $\mathbf{\Omega} = \mathbf{\mu} - \log r_L$ to indicate an estimator of the log distance in scaled units.

4.3.1 'General Linear Estimator' : Ω_{GL}

We may define an estimator of ω simply by taking a linear combination of the observables **m** and **P**. Thus we introduce a set of estimators of the form:-

$$\Omega_{GL} = A(\mathbf{m} - m_L) + B\mathbf{P} + C \qquad (4.10)$$

where A, B and C are constants. This is a natural choice for an estimator of ω if one assumes a linear regression relationship between **M** and **P** (e.g. if one assumes that $E(\mathbf{M}|\mathbf{P}) = a\mathbf{P} + b$, for constants a and b). More specifically, suppose one derives (via linear regression or otherwise) the straight line relation $\mathbf{M} - M_0 = \Delta(\mathbf{P} - P_0)$, where M_0 and P_0 are as defined above and the slope, Δ, is a constant. We have seen in chapter (2) the standard procedure used to infer distances from such a relation (i.e. estimate **M** from the observed value of **P** and then combine with the observed apparent magnitude); it follows from equations (3.8) and (4.8) that this procedure is equivalent to defining an estimator of ω given by:-

$$\Omega = 0.2(\mathbf{m} - m_L) - 0.2\Delta(\mathbf{P} - P_0) \qquad (4.11)$$

which fits the general form of equation (4.10).

Recall from chapter (2) the four types of linear regression

which have appeared in the literature in connection with the calibration of Tully-Fisher type relations: these are (in the notation of this chapter) regression of **M** on **P**; regression of **P** on **M**; 'orthogonal' regression and 'bisector' regression (see section 2.3 for more details). By substituting in equation (4.11) the appropriate value of Δ, the slope of the regression line, we may thus write down the estimator which corresponds to calibration by each particular type of linear regression. Feigelson *et al* (1990) provide algebraic expressions for Δ determined for the four regression lines listed above, as well as a fifth - the 'reduced major axis' regression, known more commonly to astronomers as Stromberg's "impartial" line (Stromberg, 1940; Kermack and Haldane, 1950). Using the results of Feigelson *et al* we define in table (4.1) five general linear estimators, corresponding to the five different regression lines. The first of the these, derived from a regression of **M** on **P**, we will denote as the 'Tully-Fisher' estimator, Ω_{TF}, since this regression was used (albeit purely on the basis of an 'eyeball' fit) to calibrate the Tully-Fisher relation in the original paper which introduced the method (Tully and Fisher, 1977). The second estimator, which is derived from the **P** on **M** regression, we will denote as the 'Schechter' estimator, Ω_S, since a regression of **P** on **M** is the central idea of the 'bias-free' recipe first proposed by Schechter (1980).

Several points are clear from table (4.1). Note that if $\rho = \pm 1$ (i.e. no scatter in the relation) then $\Delta = \pm\sigma_M/\sigma_P$ for each regression line, so that all five estimators are identically equal in this case. As the correlation between the variables decreases the values of Δ for the five regression lines will diverge and the estimators will, in general,

infer different distances from the same observed values of **m** and **P**. One is then, of course, faced with a choice of which estimator to use.

Estimator	Type of Regression	Value of Δ
ω_{TF}	**M** on **P**	$\frac{\sigma_M \rho}{\sigma_P}$
ω_S	**P** on **M**	$\frac{\sigma_M}{\sigma_P \rho}$
ω_B	Bisector	$\frac{\rho}{1-\rho^2}\left[\frac{\sigma_M^2 - \sigma_P^2}{\sigma_M \sigma_P} + \left[\frac{\sigma_P^2}{\sigma_M^2} + \rho^2 + \frac{1}{\rho^2} + \frac{\sigma_M^2}{\sigma_P^2}\right]^{\frac{1}{2}}\right]$
ω_O	Orthogonal	$\frac{1}{2\rho\sigma_M\sigma_P}\left[\sigma_M^2 - \sigma_P^2 + \left[(\sigma_M^2 - \sigma_P^2)^2 + 4\rho^2\sigma_M^2\sigma_P^2\right]^{\frac{1}{2}}\right]$
ω_I	Reduced Major Axis (Impartial)	$\frac{\pm\sigma_M}{\sigma_P}$

Table (4.1)

General linear estimators of log distance, ω, corresponding to calibration of the **M** - **P** relation by different regression lines. Each estimator is of the form $\Omega = 0.2(\mathbf{m} - m_L) - 0.2\Delta(\mathbf{P} - P_0)$, where Δ is the slope of the appropriate regression line.

One might perhaps reject immediately the 'impartial' estimator if there is appreciable scatter in the relation, since the slope of the 'impartial' regression line is not sensitive to the value of ρ. Indeed, for this reason Feigelson *et al* (1989) warn against use of the 'impartial' regression line to calibrate linear relationships. To assess rigorously the relative merits of the other estimators we must compute their distribution - as was the case for the simple distance estimators of chapter (3). Before we do this, however, it is instructive to look at the estimators "in action" by considering some specific numerical examples.

Figures (4.5) to (4.7) show graphs of the distances inferred from our different general linear estimators for the particular case of the Tully-Fisher relation. We plot the estimated distance as a function of observed 21cm line width, for different values of the correlation coefficient, ρ. The distances were calculated from the estimated log distance, Ω, in the obvious way - after first converting from scaled distance units by assuming the values of $M_0 = -20$ and $m_L = 15$. (These give the convenient scaled distance unit of 100 MPc.) The mean log(line width), P_0, was taken to be 2.5 and the dispersions σ_M and σ_P to be 1.0 and 0.1 respectively. The observed value of $\mathbf{m}$ substituted into equation (4.11) was taken to be 15 in all cases; i.e. the galaxy was assumed to be observed at the magnitude limit. In fact, the choice of value of $\mathbf{m}$ is not important for comparative purposes since it follows from equation (4.11) that any change in $\mathbf{m}$ would simply rescale each of the distance estimates by the same constant.

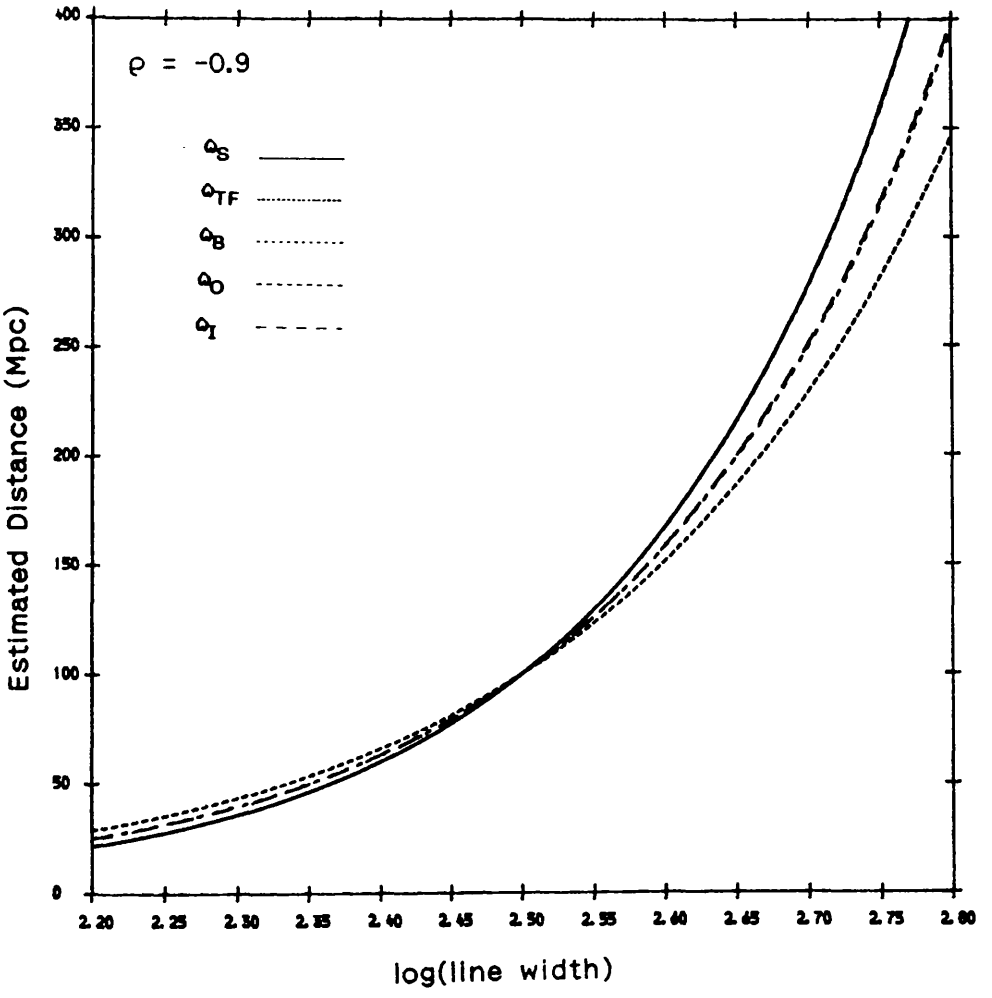

Figure (4.5)

Distances inferred from different general linear estimators of ω, as a function of observed log(line width), **P**, and for $\rho = -0.9$ ($M_0 = -20$, $m = m_L = 15$, $\sigma_M = 1$, $\sigma_P = 0.1$, $P_0 = 2.5$)

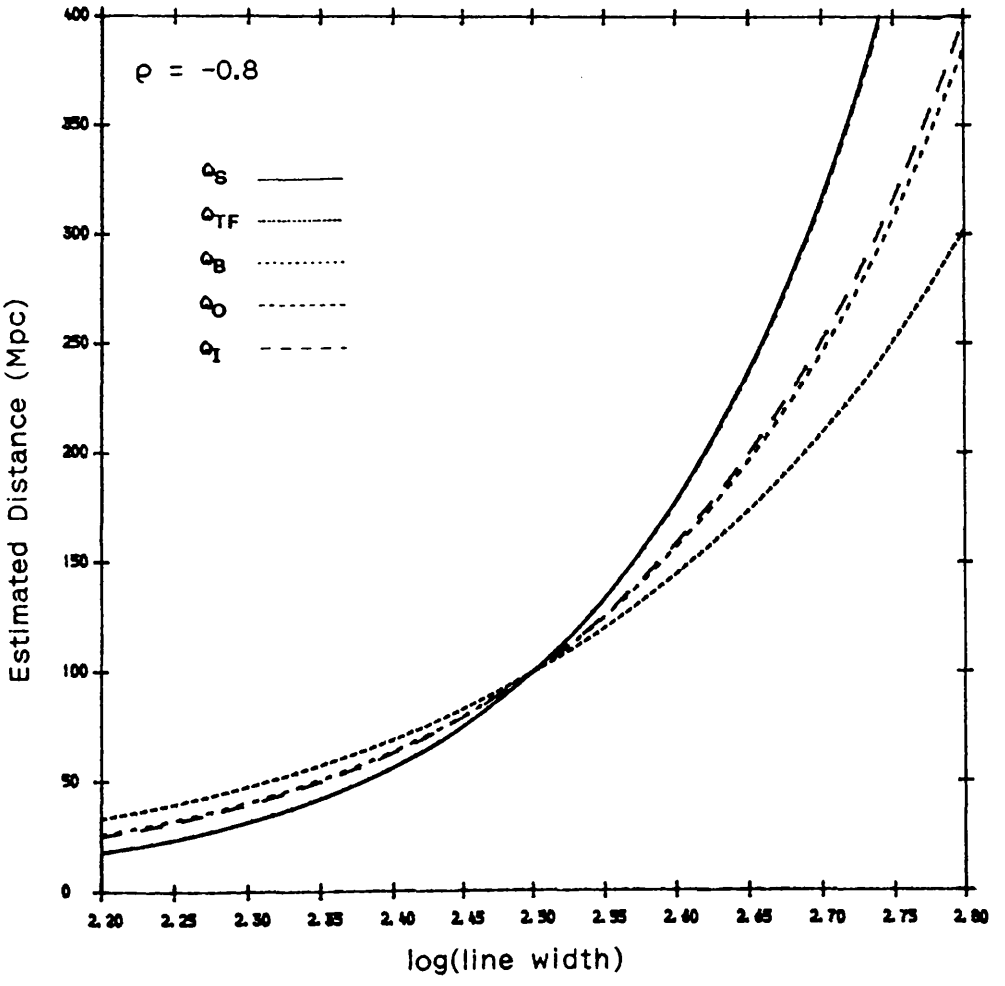

Figure (4.6)

Distances inferred from different general linear estimators of ω, as a function of observed log(line width), P, and for $\rho = -0.8$ ($M_0 = -20$, $m = m_L = 15$, $\sigma_M = 1$, $\sigma_P = 0.1$, $P_0 = 2.5$)

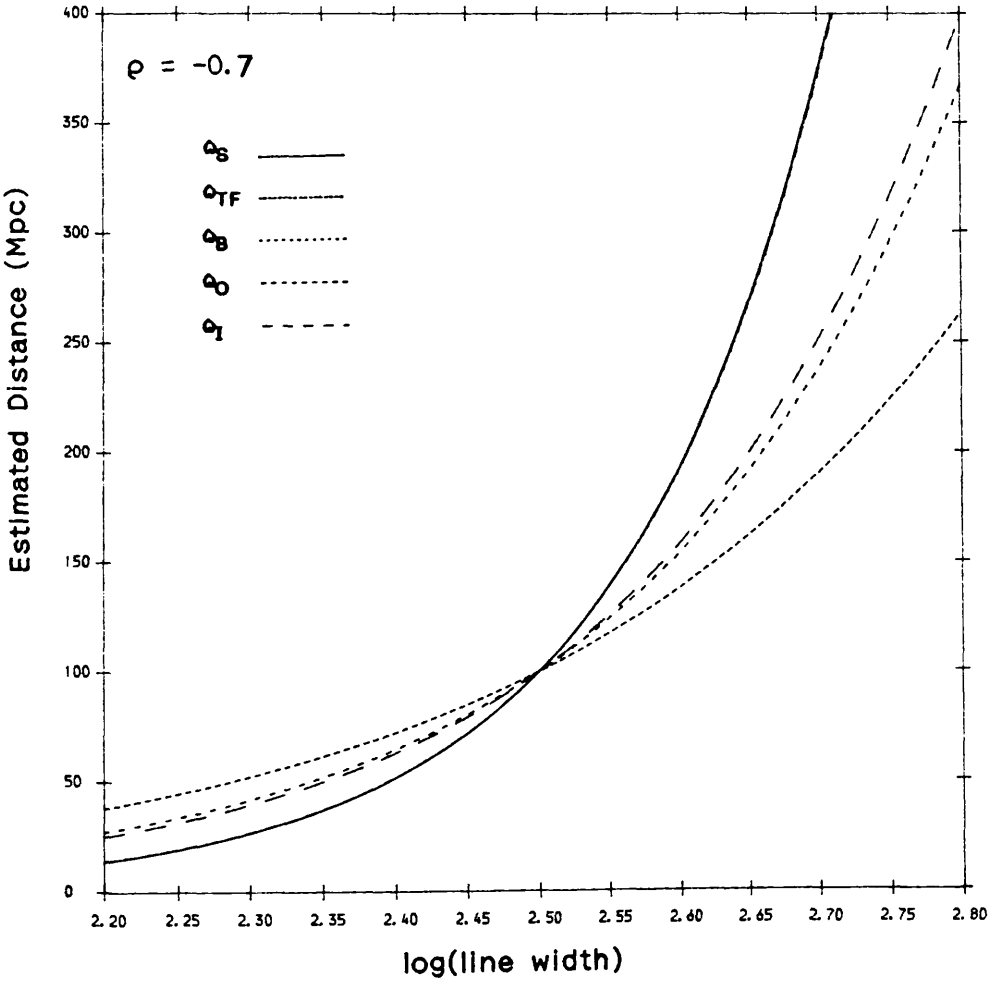

Figure (4.7)

Distances inferred from different general linear estimators of ω, as a function of observed log(line width), **P**, and for $\rho = -0.7$
($M_0 = -20$, $\mathbf{m} = m_L = 15$, $\sigma_M = 1$, $\sigma_P = 0.1$, $P_0 = 2.5$)

A number of properties of the estimators can be seen from these graphs. Firstly, note that when the observed log(line width) lies close to the mean value, P_0 = 2.5, there is little difference in the distance inferred by each of the estimators. At other observed line widths, on the other hand, there is considerable spread in the distance estimates - particularly when $P > P_0$. (Note, however, that the estimates given by the 'orthogonal' estimator, ω_O, are almost indistinguishable from those of the 'Schechter' estimator, ω_S. This is not the case in general but in our present example is a consequence of the fact that the dispersion of **P** is very much smaller than that of **M**. In fact it may easily be shown that as $\sigma_P/\sigma_M \to 0$, $\omega_O \to \omega_S$.)

Suppose, for example, that the observed log(line width) were equal to 2.7 (i.e. $P_0 + 2\sigma_P$). From figure (4.5) we see that, for ρ = -0.9, this would give the following distance estimates (converting back to MPc): 229MPc ($\hat{r}_{TF}$); 278MPc ($\hat{r}_S$); 278MPc ($\hat{r}_O$); 249MPc ($\hat{r}_B$) and 251MPc ($\hat{r}_I$). This spread increases further as the scatter in the **M-P** relation increases. If ρ = -0.7, for example, then we obtain the following distance estimates from figure (4.7): 190MPc ($\hat{r}_{TF}$); 373MPc ($\hat{r}_S$); 370MPc ($\hat{r}_O$); 237MPc ($\hat{r}_B$) and 251MPc ($\hat{r}_I$). Note that in both cases the largest estimate is obtained from $\hat{r}_S$ and the smallest from $\hat{r}_{TF}$, as one would expect from the slopes of the various regression lines.

It is clear from figures (4.5) to (4.7), therefore, that as the observed line width increases beyond the mean value, P_0, the discrepancy between the different estimators will increase and the choice of which is most appropriate becomes non-trivial. We can look at this another way: although we have assumed that there is no

selection in **P** (i.e. that any line width is *observable*) we know that because of magnitude selection we can expect the more distant galaxies which we observe to be intrinsically more luminous. It then follows from the correlation between **M** and **P** that the expected line width of more distant observed galaxies will be large. We can express this idea more rigorously by determining the expected log(line width), $E(\mathbf{P}|x_0)$, at given (scaled) true distance, x_0, . This is easily obtained by integrating equation (4.7) and is given by:-

$$E(\mathbf{P}|x_0) = P_0 - \frac{\sigma_P \rho \exp[-1/2\sigma_m^2(5\log x_0)^2]}{\sqrt{2\pi}\,\Phi(-5\log x_0/\sigma_M)} \qquad (4.12)$$

The form of the right hand side is familiar from equation (3.14), which gives the mean observed apparent magnitude at true distance, x_0, in the magnitude-only case. Substitution of the parameter values used in the above examples confirms that $E(\mathbf{P}|x_0)$ increases with true distance, x_0. We can, therefore, expect the discrepancy between the distance estimators to become more pronounced as the true distance of an observed galaxy increases, so that the choice of which estimator is 'best' becomes particularly important for more distant objects.

It should be mentioned in passing that one might consider using equation (4.12) as a means of defining an estimator of distance: i.e. one defines the distance estimate to be equal to the value of x_0 for which the observed log(line width), P_{obs}, is equal to the mean value at that distance, $E(\mathbf{P}|x_0)$. This is precisely the same idea as used to define the 'mean' estimator of chapter (3). The problem here is that one observes both **m** and **P**, so that the value of x_0 which satisfies $E(\mathbf{P}|x_0) = P_{obs}$ will in general be different from the value

which satisfies $E(\mathbf{m}|x_0) = m_{obs}$; i.e. one cannot give a consistent definition of the estimator. A similar problem arises if one tries to define a 'mode' or 'median' estimator using two or more observables. One may still define a 'maximum likelihood' estimator, on the other hand, as we will now show.

4.3.2 'Maximum Likelihood Estimator': ω_{ML}

This estimator is defined as the value of ω_0 which maximises the probability, with respect to ω_0, of obtaining the observed values of **m** and **P**. Therefore ω_{ML} satisfies:-

$$\frac{\partial}{\partial\omega_0}\, \varsigma(\mathbf{m}=m_{obs}, \mathbf{P}=P_{obs}|x_0)\Big|_{\omega_0=\omega_{ML}} = 0 \qquad (4.13)$$

where $\omega_0 = \log x_0$.

It should be stressed that, because of selection, the 'maximum likelihood' estimator will not, in general, be expressible as a linear function of **m** and **P**. Indeed, substituting from equation (4.7) we find that ω_{ML} satisifes:-

$$\mathbf{m}-m_L - \frac{\rho\sigma_M}{\sigma_P}(\mathbf{P}-P_0) = 5\omega_{ML} - \frac{(1-\rho^2)\sigma_M \exp[-1/2\sigma_m{}^2(5\omega_{ML})^2]}{\sqrt{2\pi}\,\Phi(-5\omega_{ML}/\sigma_M)} \qquad (4.14)$$

which clearly shows the non-linearity of ω_{ML}. Moreover, it is not even possible to express ω_{ML} in closed form as an explicit function of **m** and **P**; the value of the estimator must be determined implicitly from the above equation. Comparing with equation (4.11) it is clear that when $\rho = \pm 1$, ω_{ML} is identically equal to each of the general linear

estimators, as the second term on the right hand side vanishes. Since this term is always negative for $\rho \neq \pm 1$, it follows that $\omega_{ML} \geq \omega_{TF}$, for all observed values of **m** and **P**.

Figure (4.8) compares distances inferred from ω_{ML} with those from ω_S and ω_{MP}, for the particular case of the Tully-Fisher relation using the same parameter values as figure (4.6). Note that distance estimates obtained from ω_{ML} are greater than those from the general linear estimators - as was the case in chapter (3). One might therefore suspect that ω_{ML} would be more reliable for very distant galaxies; we will investigate this further by studying the distribution of each of the above estimators.

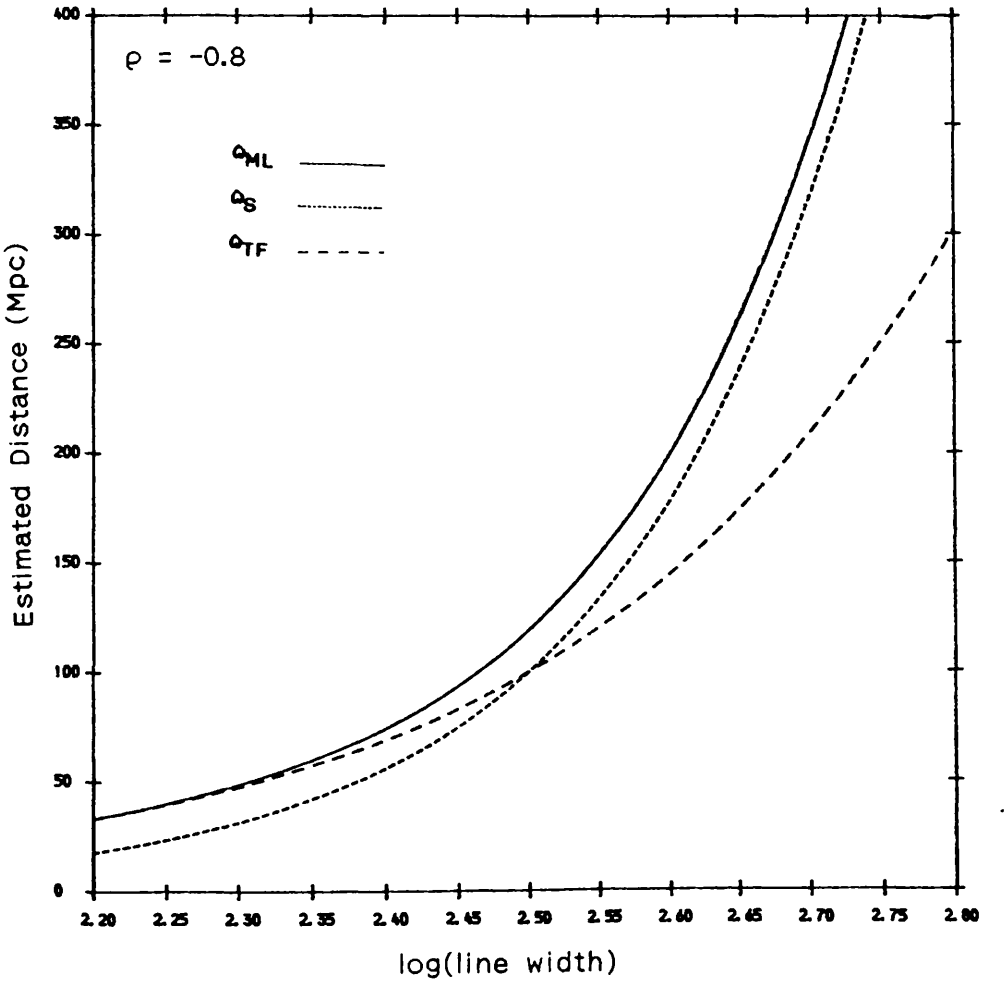

Figure (4.8)

Distances inferred from Ω_{ML}, Ω_{TF} and Ω_S as a function of observed log(line width), P, and for $\rho = -0.8$

($M_0 = -20$, $m = m_L = 15$, $\sigma_M = 1$, $\sigma_P = 0.1$, $P_0 = 2.5$)

4.4 Estimator Distributions

We may derive the probability density function of any of the above estimators by a very similar method to that of section (3.4). First observe that each Ω is a function of the random variables **m** and **P**; i.e. $\Omega = \Omega(\mathbf{m},\mathbf{P})$. Now define a transformation, τ, which maps $(\mathbf{m},\mathbf{P})$ to $(\Omega,\mathbf{P})$. The joint distribution of Ω and **P** may then be written down in terms of the observed distribution of **m** and **P** at true distance, x_0 - as given by equation (4.7) - and will therefore depend on the true log distance, ω_0. viz:-

$$n(\Omega,\mathbf{P}|\omega_0) = \varsigma(\mathbf{m}(\Omega,\mathbf{P}),\mathbf{P}|x_0)\left|\frac{\partial(\mathbf{m},\mathbf{P})}{\partial(\Omega,\mathbf{P})}\right| \qquad (4.15)$$

The jacobian of the transformation is defined in the usual way and may be computed from equation (4.11) for the 'general linear' estimators or from equation (4.14) for the 'maximum likelihood' estimator.

Finally, the distribution of Ω is obtained by integrating equation (4.15) over **P**. Thus we have:-

$$\Omega(\Omega|\omega_0) = \int n(\Omega,\mathbf{P}|\omega_0)d\mathbf{P} \qquad (4.16)$$

For the 'maximum likelihood' estimator this integral must be carried out numerically, but it is possible to determine the distribution of $\Omega_{GL} = A(\mathbf{m} - m_L) + B\mathbf{P} + C$ analytically, expressed in terms of the constants A, B and C and the parameters of the bivariate distribution of **M** and **P**. After some rather messy algebraic manipulation we obtain

the following:-

$$\Omega(\Omega_{GL}|\omega_0) = \frac{\exp[-\tfrac{1}{2}(\tau-\upsilon^2/4\phi)]\ \Phi[\phi^{\frac{1}{2}}(\Omega_{GL}-C)/B+\upsilon/2\phi]}{\sqrt{2\pi}A\sigma_M\sigma_P\sqrt{\phi(1-\rho^2)}\ \Phi(-5\omega_0/\sigma_M)} \qquad (4.17)$$

τ, υ and ϕ appear here simply as a shorthand to make equation (4.17) more compact. The expressions which they stand for are:-

$$\phi = \frac{1}{1-\rho^2}\left[\frac{B^2}{A^2\sigma_M{}^2} + \frac{2\rho B}{A\sigma_M\sigma_P} + \frac{1}{\sigma_P{}^2}\right]$$

$$\upsilon = \frac{-2}{1-\rho^2}\left[\frac{B(\Omega_{GL}-C-5A\omega_0)}{A^2\sigma_M{}^2} + \frac{B\rho P_0 + \rho(\Omega_{GL}-C-5A\omega_0)}{A\sigma_M\sigma_P} + \frac{P_0}{\sigma_P{}^2}\right]$$

$$\tau = \frac{1}{1-\rho^2}\left[\frac{(\Omega_{GL}-C-5A\omega_0)^2}{A^2\sigma_M{}^2} + \frac{2\rho P_0(\Omega_{GL}-C-5A\omega_0)}{A\sigma_M\sigma_P} + \frac{P_0{}^2}{\sigma_P{}^2}\right]$$

Figures (4.9) to (4.11) show examples of the distribution functions of Ω_S, Ω_{TF} and Ω_{ML} respectively, at different true distances.

A number of basic properties of the estimators are illustrated by these graphs. Firstly, we can see that when $x_0 = 0.5$ (or 50MPc, taking our usual conversion from scaled units) each estimator is to a good approximation normally distributed, with modal value coincident with the true log distance, ω_0 (indicated by the dotted line) so that there is no significant bias in the expected value of any of the estimators at this distance. This must be the case since, if there is no magnitude selection at this distance, then the joint distribution of **m** and **P** for observable galaxies will be bivariate normal, and hence any

linear combination of **m** and **P** will also be normally distributed.

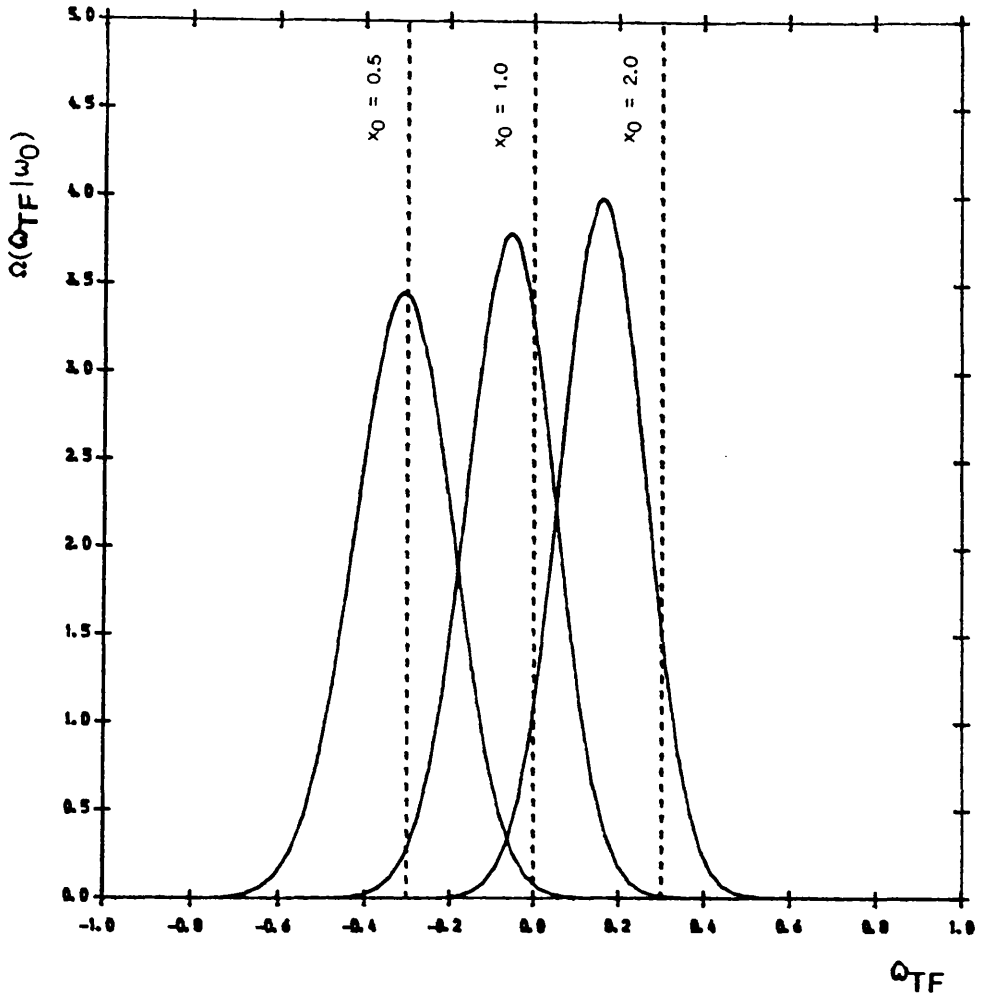

Figure (4.9)

Probability density function, $\Omega(\omega_{TF}|\omega_0)$, of the 'Tully-Fisher' estimator at true distance, x_0 = 0.5, 1.0 and 2.0

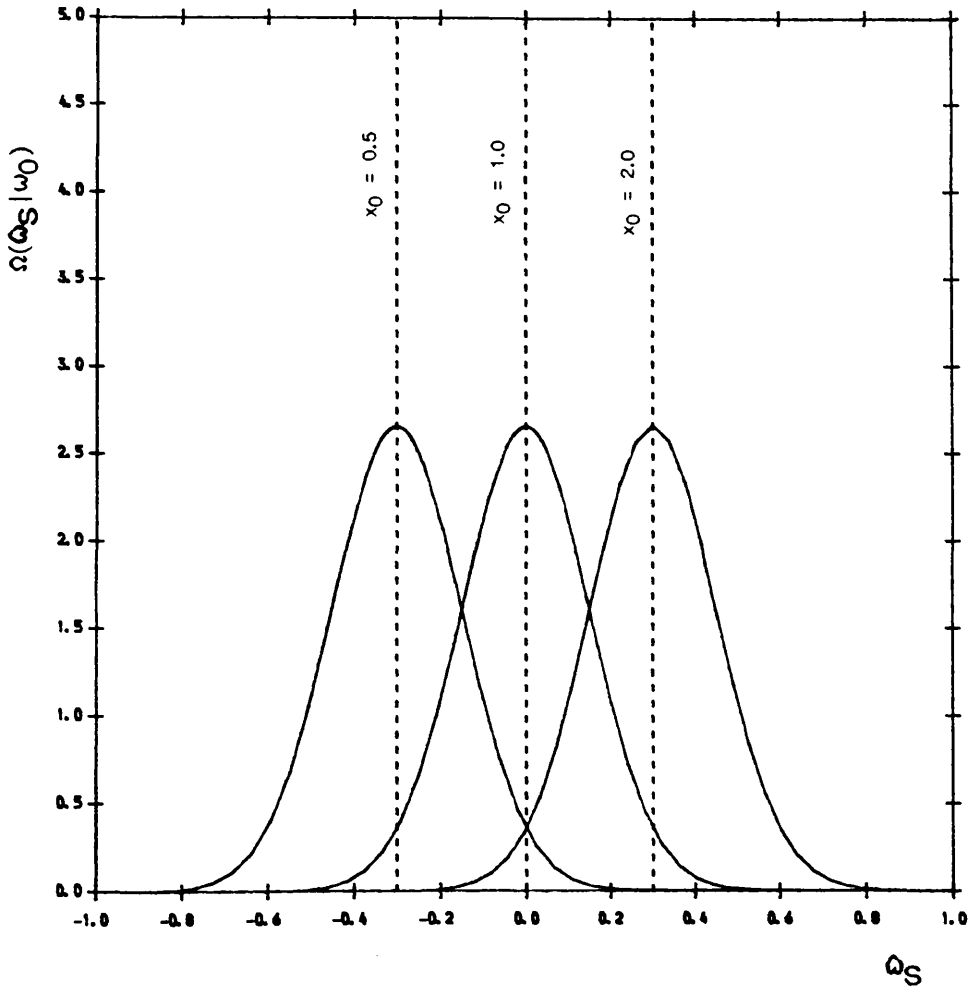

Figure (4.10)

Probability density function, $\Omega(\omega_S|\omega_0)$, of the 'Schechter' estimator at true distance, x_0 = 0.5, 1.0 and 2.0

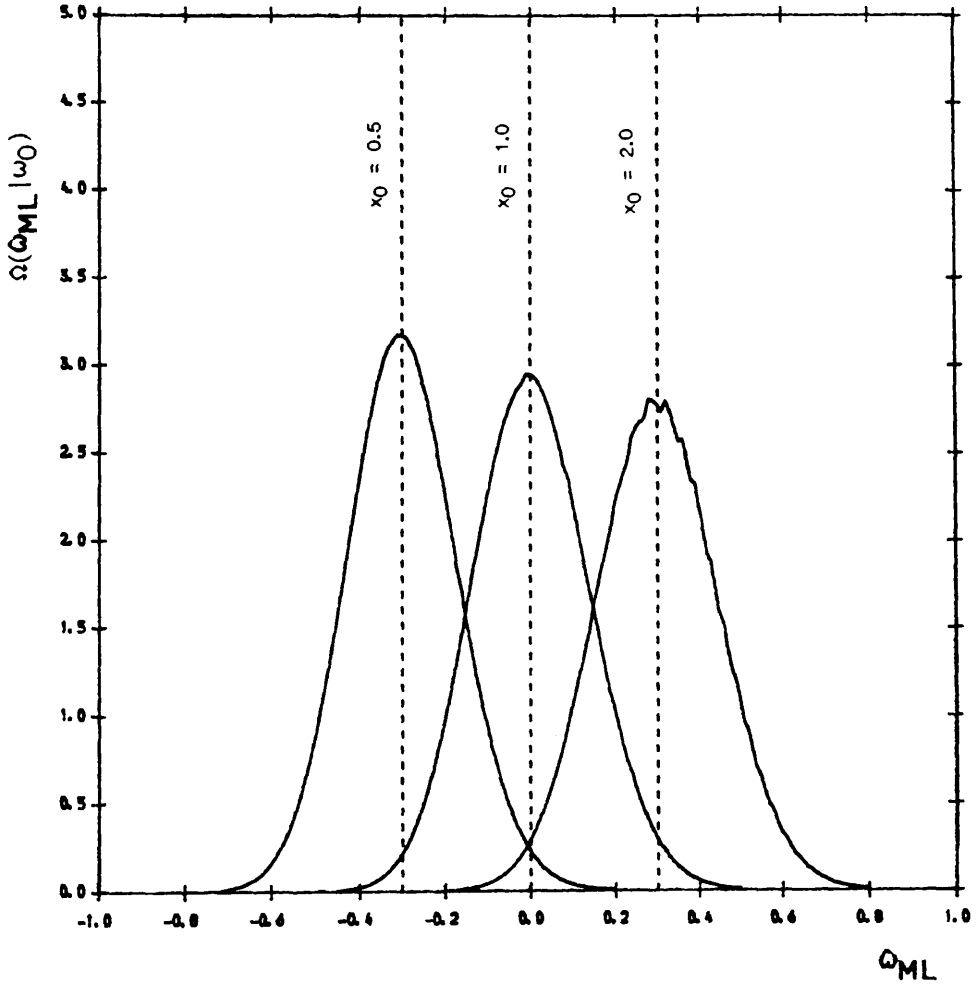

Figure (4.11)

Probability density function, $\Omega(\omega_{ML}|\omega_0)$, of the 'maximum likelihood' estimator at true distance, x_0 = 0.5, 1.0 and 2.0

From figure (4.9), however, we can see that at larger true distances this is no longer the case for ω_{TF}: although the distribution of ω_{TF} remains close to normal for large x_0, it is no longer symmetrical about ω_0 but is displaced progressively leftwards of ω_0. The 'Tully-Fisher' estimator is , therefore, negatively biased at large true distances and its use will result in the systematic underestimation of galaxy distances.

Consider, on the other hand, the 'Schechter' estimator: it is apparent from figure (4.10) that the distribution of ω_S does not change in shape as the true distance increases. In fact, upon substitution of the appropriate values of A, B and C in equation (4.17), we may show that the distribution of ω_S reduces precisely to a normal distribution, with mean value ω_0 and variance given by $0.04\sigma_M^2(1-\rho^2)/\rho^2$, at *all* true distances. This is entirely consistent with the fact that measurements of **P** are free from selection effects, and confirms that the 'Schechter' recipe will indeed give unbiased estimates of the true log distance, ω_0, for all ω_0.

Finally we can see from figure (4.11) that the distribution of the 'maximum likelihood' estimator is also basically gaussian in form at various true distances; moreover the modal value of ω_{ML} appears to be approximately coincident with ω_0, so that this estimator does not display significant bias, even at $x_0 = 2.0$. Furthermore, on closer comparison with figure (4.10) it may be noted that the spread in the distribution of ω_S is at least as large as that of ω_{ML} at the same distance - and indeed is somewhat larger for $x_0 = 0.5$ and $x_0 = 1.0$; i.e. the variance (or, equivalently for unbiased estimators, the risk) of

ω_{ML} is less than or equal to the variance of ω_S. If we recall the definition of risk from equation (3.26), this means that the expected root mean square error on a distance estimate obtained from any single observation will be smaller for ω_{ML} than for ω_S. It may also be seen from figure (4.9) that, at each true distance shown, the variance of ω_{TF} is smaller than that of both ω_{ML} and ω_S - although the latter estimators may in general have smaller risk since it follows from the definitions of risk and variance that the risk of a *biased* estimator is always strictly greater than its variance.

The examples of estimator distributions considered above give some indication of the relative merits of these estimators; in particular the unbiasedness of the 'Schechter' estimator is clearly a desirable property, and when one wishes to estimate distances from a large number of observations ω_S would seem to be the most appropriate choice. However, these graphs also demonstrate that the dual role of both bias and risk in determining a 'best' estimator cannot be overlooked. In particular the risk of ω_S is found to be given by $0.04\sigma_M^2(1-\rho^2)/\rho^2$, which may be very large if **M** and **P** are poorly correlated. If, for example, ρ were as low as -0.5 then the risk would be equal to 0.12, which corresponds to a root mean square percentage distance error of 35%. Thus, although ω_S would still be *unbiased* in this case - so that one would not expect a large number of estimates of ω_0 to display a systematic error - the expected distance errors on a small number of observations would be considerably larger than those found by using ω_{TF} or ω_{ML}.

In the following section we will compute the bias and risk of

each estimator as a function of true distance, in order to provide a more quantitative assessment of which estimator is 'best'.

4.5 Bias and Risk of Estimators

The equations introduced in chapter (3) to define the bias and risk of a distance estimator, $\hat{x}$, have obvious counterparts for an estimator of log distance. Thus we may define the bias, $B(\Omega,\omega_0)$, and risk, $R(\Omega,\omega_0)$, of an estimator, Ω, at true log distance, ω_0, viz:-

$$B(\Omega,\omega_0) = \int (\Omega - \omega_0)\, \Omega(\Omega|\omega_0)d\Omega \qquad (4.18)$$

$$R(\Omega,\omega_0) = \int (\Omega - \omega_0)^2\Omega(\Omega|\omega_0)d\Omega \qquad (4.19)$$

The computation of the bias and risk of the 'maximum likelihood' estimator must be carried out numerically; an analytical treatment is possible, however, for the 'general linear' estimators.

4.5.1 Bias and Risk of Ω_{GL}

The derivation of expressions for the bias and risk of the 'general linear' estimator is relatively straightforward but algebraically tedious. Using equation (4.17) one can, of course, substitute for the distribution, $\Omega(\Omega_{GL}|\omega_0)$, in the above equations and then integrate directly. The analysis is made somewhat more tractable, however, by re-expressing Ω_{GL} and $\Omega(\Omega_{GL}|\omega_0)$ in terms of functions of $\mathbf{m}$ and $\mathbf{P}$,

using equation (4.10), (4.15) and (4.16). The expressions for both the bias and risk are then reduced to a linear combination of integrals over **m** and **P**, each of the form:-

$$I_{\mu,\nu} = \iint (m - m_L)^{\mu} P^{\nu} \zeta(m,P|x_0)\,dm\,dP \qquad (4.20)$$

where the exponents μ and ν lie between zero and two. In this notation the bias, $B(\omega_{GL},\omega_0)$, is therefore given by:-

$$B(\omega_{GL},\omega_0) = A.I_{1,0} + B.I_{0,1} + C - \omega_0 \qquad (4.21)$$

The expression for the risk is somewhat lengthier. This approach is similar to the use of moment generating functions for calculating the bias and variance, which may often result in a simpler analysis of statistical problems (c.f. Hoel, 1962).

Upon performing the necessary integrations and after some further regrouping of the terms we find the following expressions for the bias and risk:-

$$B(\omega_{GL},\omega_0) = (5A-1)\omega_0 - \frac{(A\sigma_M+\rho B\sigma_P)\,\exp[-\tfrac{1}{2}(5\omega_0/\sigma_M)^2]}{\sqrt{2\pi}\,\Phi(-5\omega_0/\sigma_M)} + BP_0 + C \qquad (4.22)$$

$$R(\omega_{GL},\omega_0) = \omega_0^2(5A-1)^2 + \frac{(A\sigma_M+\rho B\sigma_P)^2\,Q[1.5,\tfrac{1}{2}(5\omega_0/\sigma_M)^2]}{2\,\Phi(-5\omega_0/\sigma_M)} - \frac{2(A\sigma_M+\rho B\sigma_P)(\omega_0(5A-1) + BP_0+C)\,\exp[-\tfrac{1}{2}(5\omega_0/\sigma_M)^2]}{\sqrt{2\pi}\,\Phi(-5\omega_0/\sigma_M)} + 2\omega_0(5A-1)(BP_0+C) + (BP_0+C)^2 + B^2\sigma_P^2(1-\rho^2) \qquad (4.23)$$

Here Q(a,x) denotes the incomplete gamma function, defined as follows (c.f. Abramowitz and Stegun, 1968):-

$$Q(a,x) = \frac{1}{\Gamma(a)} \int_x^\infty t^{a-1} e^{-t} dt \qquad (4.24)$$

We can see that each of these expressions involves highly non-linear functions of the true log distance, ω_0, so that in general the bias and risk of $\hat{\omega}_{GL}$ will be strongly dependent on the true distance. The precise behaviour will clearly depend on the values of A, B and C, however, and may be considerably simplified by suitable choice of these constants. In particular, recall equation (4.11) which defines the general linear estimator corresponding to a regression line of slope, Δ; for this special case we found that A = 0.2, B = -0.2Δ and C = $0.2\Delta P_0$. Upon substitution of these values into equations (4.22) and (4.23) we see that most of the terms vanish, regardless of the value of Δ, and the bias and risk are both independent of P_0. In fact, for all of the general linear estimators considered in section (4.3) we may re-express the bias and risk in the simpler form:-

$$B(\hat{\omega}_{GL},\omega_0) = -\frac{0.2(\sigma_M - \Delta\rho\sigma_P)\exp[-\tfrac{1}{2}(5\omega_0/\sigma_M)^2]}{\sqrt{2\pi}\,\Phi(-5\omega_0/\sigma_M)} \qquad (4.25)$$

$$R(\hat{\omega}_{GL},\omega_0) = \frac{0.04(\sigma_M - \Delta\rho\sigma_P)^2 Q[1.5,\tfrac{1}{2}(5\omega_0/\sigma_M)^2]}{2\,\Phi(-5\omega_0/\sigma_M)} + 0.04\Delta^2\sigma_P^2(1-\rho^2) \qquad (4.26)$$

It is now clear from equation (4.25) that by choosing Δ to equal $\sigma_M/\rho\sigma_P$, the bias of $\hat{\omega}_{GL}$ vanishes for all values of ω_0 since the constant term on the numerator is identically zero. This is precisely

the value of Δ which defines the 'Schechter' estimator and confirms that the 'Schechter' scheme will give bias-free estimates of log distance, ω_0, for all values of ω_0. Similarly the non-linear term in the expression for the risk vanishes from equation (4.26), leaving only the constant term which is equal to $0.04\sigma_M^2(1-\rho^2)/\rho^2$, as was previously established in section (4.4). The risk of ω_S is, therefore, constant and independent of the true log distance, which means that the percentage risk of distances estimated using ω_S will be constant.

At this point we should note that the values of A, B and C which define ω_S (and thus make ω_{GL} unbiased) are unique: the 'Schechter' estimator is, therefore, the only estimator of this form which is unbiased for all true log distances. Certainly the 'general linear' estimators corresponding to the other regression lines will all, in general, be biased. Figures (4.12) and (4.13) show the bias and risk of the first four 'general linear' estimators of Table (4.1) - we have omitted the 'impartial' estimator, ω_I, which we have already seen in insensitive to the value of ρ - together with a fifth, which we denote as the 'naive' estimator, ω_N, defined by:-

$$\omega_N = 0.2(m - m_L) \qquad (4.27)$$

Thus, ω_N is a function only of apparent magnitude and is, in fact, the estimator of log distance corresponding directly to the 'naive' estimator of distance, $\hat{x}_N$, defined in equation (3.18). We can compute the bias and risk of ω_N simply by putting A=0.2, B=0 and C=0 in equations (4.22) and (4.23).

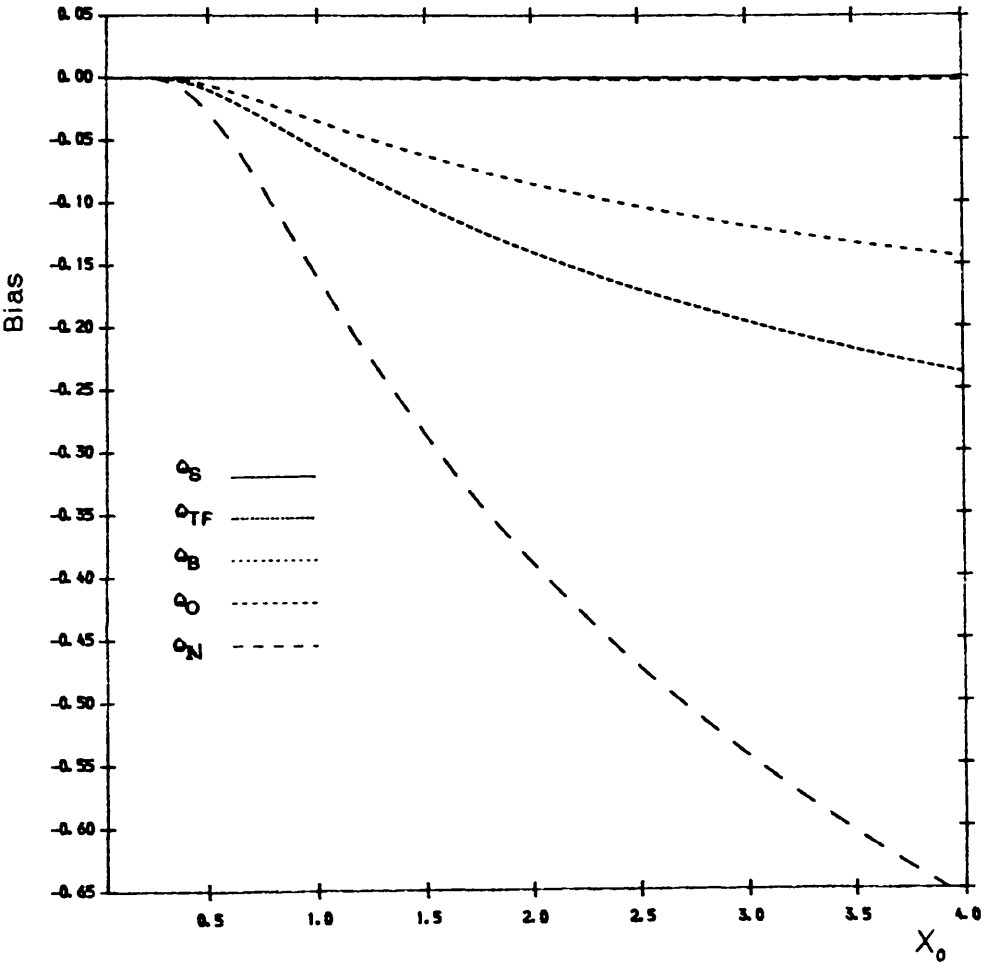

Figure (4.12)

Bias of different 'general linear' estimators of distance modulus as a function of true distance, x_0
($\sigma_M = 1$, $\sigma_p = 0.1$, $\rho = -0.8$)

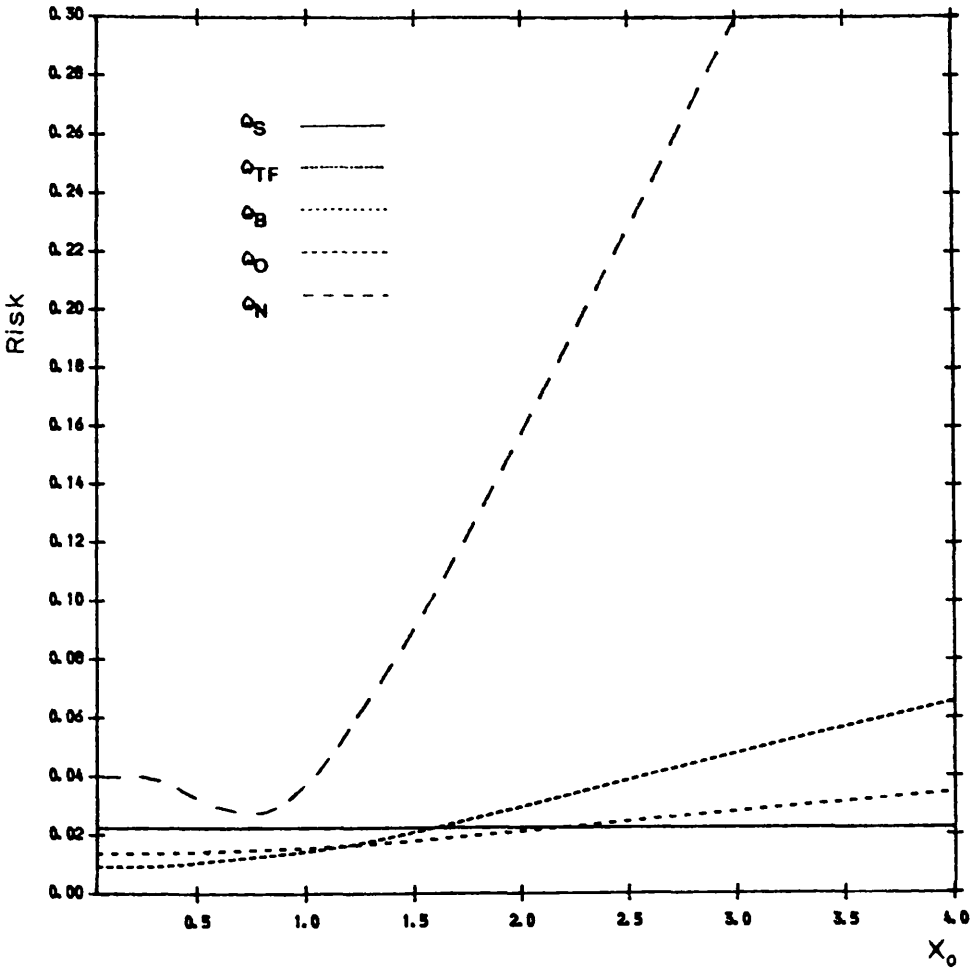

Figure (4.13)

Risk of different 'general linear' estimators of distance modulus as a function of true distance, x_0
($\sigma_M = 1$, $\sigma_p = 0.1$, $\rho = -0.8$)

Note that all of the estimators are unbiased when the true distance is very small - including, ω_N, as was remarked in section (3.4). Note also that the bias of the 'orthogonal' estimator, ω_O, is negligible even at large true distances, and indeed both the bias and risk of ω_O are almost indistiguishable from those of the 'Schechter' estimator, ω_S; this follows from the fact that the dispersion of **P** is very much smaller than that of **M**, as we remarked previously in section (4.3.1) in the context of the actual *values* inferred by these two estimators.

Observe that, for $x_0 \geqslant 2.3$, ω_S has the smallest risk of all of these estimators - and is, of course, unbiased. It seems clear, then, that at very large true distances ω_S is the 'best' of the 'general linear' estimators which we have considered. The picture is not quite so simple at smaller true distances, however; we see that, although ω_{TF} is the most biased of the four estimators which are functions of **m** and **P**, it has the smallest risk for $x_0 \leqslant 1.2$ and consequently might be regarded as the most suitable estimator in this true distance range - particularly if one observes only a small number of galaxies. Against this, of course, one must balance the fact that both the bias and risk of ω_{TF} increase sharply for $x_0 > 1.2$; if one wrongly assumes that the true distance is less than 1.2 when it is, in fact, considerably greater than unity then the use of ω_{TF} would result in a large systematic error. Nevertheless, it is at least evident that both the bias and risk of ω_N are considerably worse than those of ω_{TF}, demonstrating the advantage which may be gained by making use of other observable quantities besides apparent magnitude to estimate distances. The precise behaviour of the bias and risk of all of these estimators will

depend on the value of the correlation coefficient, but the general trends shown by figures (4.12) and (4.13) change little for different values of ρ.

4.5.2 Bias and Risk of ω_{ML}

Calculation of the integral expressions for the bias and risk of the 'maximum likelihood' estimator cannot be performed analytically, but may easily be carried out using standard numerical packages. Figures (4.14) and (4.15) show the bias and risk of ω_{ML} as a function of the true distance, x_0, and for $\sigma_M = 1$, $\sigma_P = 0.1$ and $\rho = -0.8$. Also shown, as a comparison, are the bias and risk of ω_S and ω_{TF}.

We can see from these graphs that the bias of ω_{ML}, although non-zero, is very small - even at large true distances. (It is also interesting to note that the bias of ω_{ML} is, in fact, positive.) Furthermore, the risk of ω_{ML} is less than that of ω_S, for all $x_0 < 4$. Similar results are found for other values of ρ, although the bias of ω_{ML} does become increasingly significant for $|\rho| < 0.4$. On the basis of these criteria, therefore, there would seem to be little to choose between ω_{ML} and ω_S as the 'best' estimator of log distance. Again, the final choice may come down to the number of observed galaxies. If one wishes to estimate the distance of only a few galaxies then the slightly smaller risk of ω_{ML} is a desirable property. If, on the other hand, one has a large sample of galaxies then the removal of a systematic bias - however small - in the distance estimates would be better achieved by the use of ω_S. This would certainly be an important advantage in

obtaining distance estimates for analysis of the velocity field - either by sophisticated techniques such as POTENT, or indeed via simpler, classical, methods such as the Hubble diagram. We will comment further on these points in chapter (6).

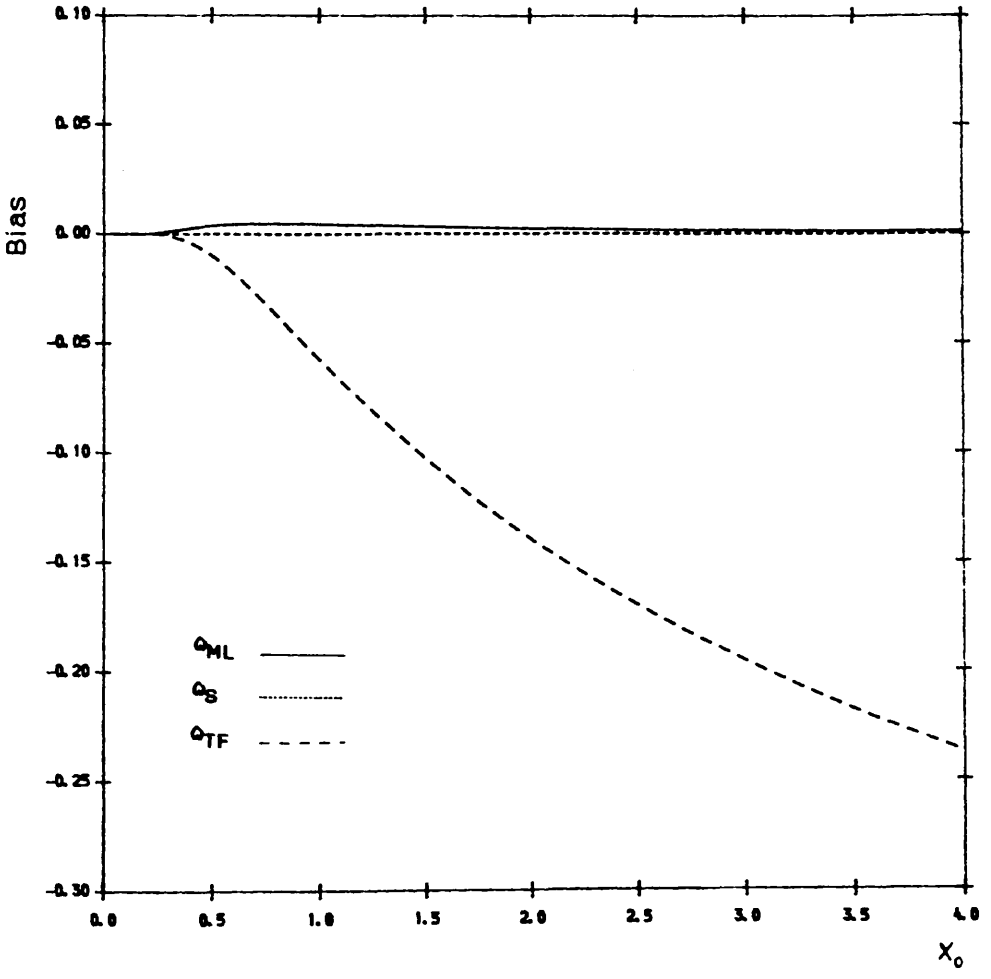

Figure (4.14)

Bias of Ω_{ML}, Ω_{TF} and Ω_S as a function of true distance, x_0
($\sigma_M = 1$, $\sigma_P = 0.1$, $\rho = -0.8$)

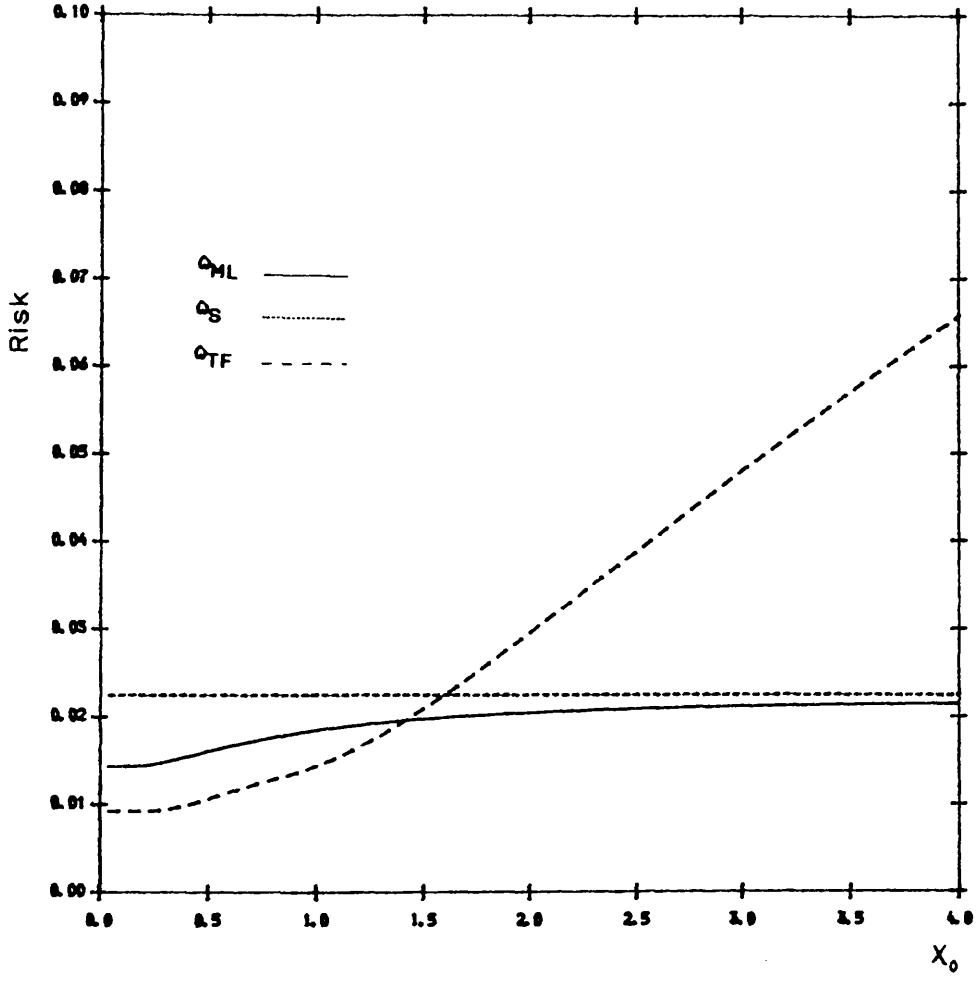

Figure (4.15)

Risk of ω_{ML}, ω_{TF} and ω_S as a function of true distance, x_0
($\sigma_M = 1$, $\sigma_P = 0.1$, $\rho = -0.8$)

4.5.3 Bias of Distance Estimators

At the beginning of section (4.3) we drew attention to the fact that the distance estimator, $\hat{x}$ - defined according to equation (4.8) from an unbiased estimator, Ω, of log distance - will not, in general, be unbiased. Although the resulting bias of $\hat{x}$ has usually been regarded as very small, it is worthwhile verifying this. Suppose we define $\hat{x}_{GL}$, a 'general linear' estimator of distance corresponding to Ω_{GL}, as follows:-

$$\hat{x}_{GL} = 10^{A(m-m_L)} + BP + C \qquad (4.28)$$

By a similar analysis to that of the previous section, we can determine the distribution of this estimator, and hence its bias and risk, as function of true distance, x_0, and for general constants A, B and C. This again involves some rather tedious algebra, however, and in the case of the 'Schechter' estimator, Ω_S, there is a much simpler route to the same result. We know that the distribution of Ω_S is normal with mean value ω_0 and variance equal to $0.04\sigma_M^2(1-\rho^2)/\rho^2$. Using this fact, it is straightforward to derive an expression for the expected value of the equivalent estimator of distance, $\hat{x}_S$, viz:-

$$E(\hat{x}_S|x_0) \equiv E(10^{\Omega_S}|x_0) = x_0\exp[0.02\kappa^2\sigma_M^2(1-\rho^2)/\rho^2] \qquad (4.29)$$

where $\kappa = ln10 \simeq 2.3$. The bias of $\hat{x}_S$ is then given by:-

$$B(\hat{x}_S,x_0) = x_0(\exp[0.02\kappa^2\sigma_M^2(1-\rho^2)/\rho^2] - 1) \qquad (4.30)$$

Thus, the unbiased 'Schechter' estimator of log distance will derive estimates of distance which are positively biased at all true

distances. The effect is fairly small: substituting $\sigma_M = 1$ and $\rho = -0.8$ in equation (4.29), we find that the percentage bias of $\hat{x}_S$ is less than 6%. The bias increases sharply, however, if there is poorer correlation between **M** and **P**.

There would, therefore, seem to be reasonable justification for studying estimators of log distance in order to establish 'best' estimators of distance; the distance bias which is introduced is not too large and the algebra is often simpler. Nevertheless, it follows from equation (4.29) that simply by dividing the value of $\hat{x}_S$ by the constant factor $\exp[0.02\kappa^2\sigma_M^2(1-\rho^2)/\rho^2]$, we can define distance estimates which are completely unbiased at all true distances. (Of course, if the correlation between **M** and **P** is high then this constant is very close to unity and the bias correction is not important.)

4.6 Confidence Intervals

We have already seen in section (4.4) that the estimators discussed so far in this chapter all have a distribution which is a function only of the true log distance, ω_0; each may, therefore, be used to construct confidence intervals for that log distance following the method developed in section (3.5).

Specifically, in order to determine a $(1-\alpha)100\%$ confidence interval for ω_0, using an estimator $\hat{\omega}$ with distribution $\Omega(\hat{\omega}|\omega_0)$, we require to find ω_1 and ω_2 such that:-

$$\int_0^{\omega_1} \Omega(\Omega|\omega_0)d\Omega = \lambda \qquad (4.31)$$

and

$$\int_0^{\omega_2} \Omega(\Omega|\omega_0)d\Omega = (1-\alpha) + \lambda \qquad (4.32)$$

where $\lambda \in [0,\alpha)$

By plotting ω_1 and ω_2 as a function of true log distance, ω_0, a confidence interval for ω_0 may then be found by the simple graphical procedure described in section (3.5). As an illustration of this, figure (4.16) shows graphs of $\omega_1(\omega_0)$ and $\omega_2(\omega_0)$ for 'equal tail' (i.e. with $\lambda = \frac{1}{2}\alpha$) 68% confidence intervals, computed from the distributions of Ω_{TF}, Ω_S and Ω_{ML} and assuming $\sigma_M = 1$, $\sigma_P = 0.1$ and $\rho = -0.8$.

We can see from figure (4.16) that the 68% confidence interval curves for Ω_S are, in fact, parallel lines. The same result is found for any other confidence interval constructed from Ω_S; this is because the distribution of Ω_S is normal with constant risk at all true distances. Consequently, the width of confidence intervals constructed using Ω_S will be independent of the 'observed' value of the estimator (i.e. the value of Ω_S determined from the observed values of **m** and **P**). This is not the case with Ω_{TF} and Ω_{ML}: we can see from figure (4.16) that when the observed value of these estimators is small, both give confidence intervals which are marginally narrower than those obtained from Ω_S, but the intervals become slightly wider as the value of the estimators increases. This is not a large effect, however,

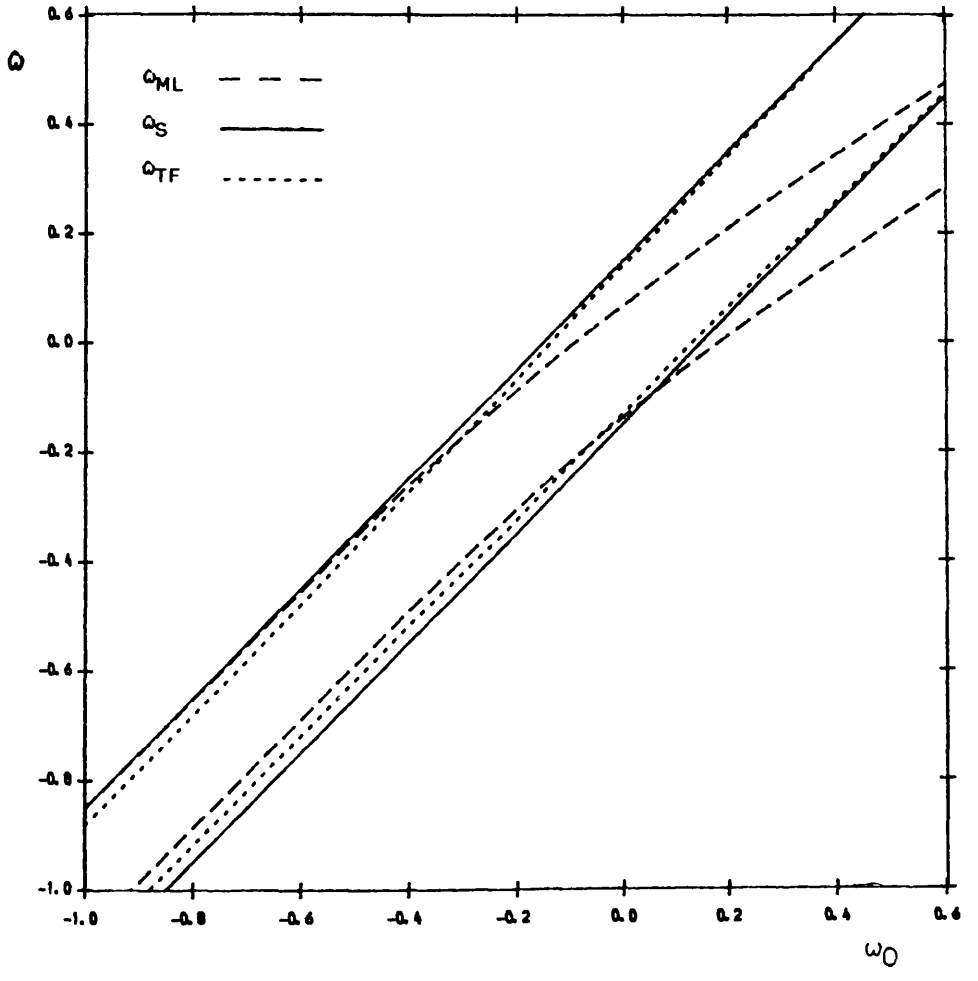

Figure (4.16)

'Equal Tail' (i.e. with $\lambda = \frac{1}{2}\alpha$) 68% confidence interval curves for the true distance modulus, ω_0, computed from the distribution functions of Ω_{TF}, Ω_S and Ω_{ML} ($\sigma_M = 1$, $\sigma_p = 0.1$, $\rho = -0.8$)

and there is basically little difference in the width of the intervals constructed from each of the three estimators. The other immediately noticeable feature of figure (4.16) is the changing slope of the confidence curves for ω_{TF}, and their divergence from those of the other two estimators at large values of ω_{TF}. Consequently, one might suspect that the confidence intervals constructed from ω_{TF} will tend to be shifted towards larger values of ω_0 compared with those constructed from the other estimators. This is not necessarily the case, however: for given observed values of **m** and **P**, ω_{TF} tends to take smaller values than ω_S, because of the negative bias of ω_{TF} at large distances, and this will shift the ω_{TF} confidence interval to the left, and closer to the interval constructed from ω_S at the same observed values of **m** and **P**.

To fix these ideas consider a specific numerical example, for the particular case of the Tully-Fisher relation, as was met previously in section (4.3.1). Let the mean log(line width), P_0 = 2.5, the mean absolute magnitude, M_0 = -20 and the limiting magnitude, m_L = 15. Suppose that we observe a galaxy with a log(line width) of 2.6 and an apparent magnitude of 13.9. It then follows from equations (4.11) and (4.14) that ω_{TF}, ω_S and ω_{ML} are approximately equal to -0.06, 0.03 and 0.0 respectively. We then obtain the following 68% confidence intervals for ω_0 from figure (4.16): [-0.16,0.12] (ω_{TF}); [-0.12,0.18] (ω_S) and [-0.15,0.14] (ω_{ML}). We see, then, that each of the estimators gives a fairly similar confidence interval for ω_0. If we take the narrowest interval, that obtained from ω_{TF}, and translate the upper and lower limits into *distances* we find a 68% confidence interval of approximately 70MPc $< r_0 <$ 130MPc for the true distance, r_0.

It is useful to compare the confidence intervals which we have determined from these estimators with those obtained from the 'naive' estimator, ω_N, of log distance, introduced in section (4.6), which is a function only of the observed apparent magnitude. Figure (4.17) shows the upper and lower confidence curves constructed from the distributions of ω_N and ω_{TF}. (For clarity we have omitted the ω_S and ω_{ML} curves). If we consider again the same numerical example as above, we saw that the values of m_{obs} = 13.9 and P_{obs} = 2.6 led to a 68% confidence interval for ω_0 of [-0.16,0.12] from ω_{TF}. If, on the other hand, we use only the measured apparent magnitude of the galaxy to estimate its distance (so that ω_N takes the value -0.22) then we see from figure (4.17) that the 68% confidence interval for ω_0 constructed from ω_N is [-0.39,0.15]. This is almost twice as large as the interval obtained from ω_{TF}. Translating both confidence intervals into distance ranges we find the following: from ω_N, 40MPc $< r_0 <$ 140MPc as opposed to 70MPc $< r_0 <$ 130MPc from ω_{TF}.

The narrower confidence interval obtained from ω_{TF} as compared with ω_N is consistent with the lower bias and risk of this estimator, and may be regarded as a consequence of having more complete information about the true distance. In other words, by observing the line width of a galaxy we obtain useful information about its absolute magnitude, which allows the likely true distance to be 'narrowed down' more effectively than may be done purely on the basis of the observed apparent magnitude. The precise amount by which the width of the confidence interval is reduced depends on the value of ρ; as the scatter in the M-P relation increases, the width of the interval found from ω_{TF} increases. Indeed, in the limiting case

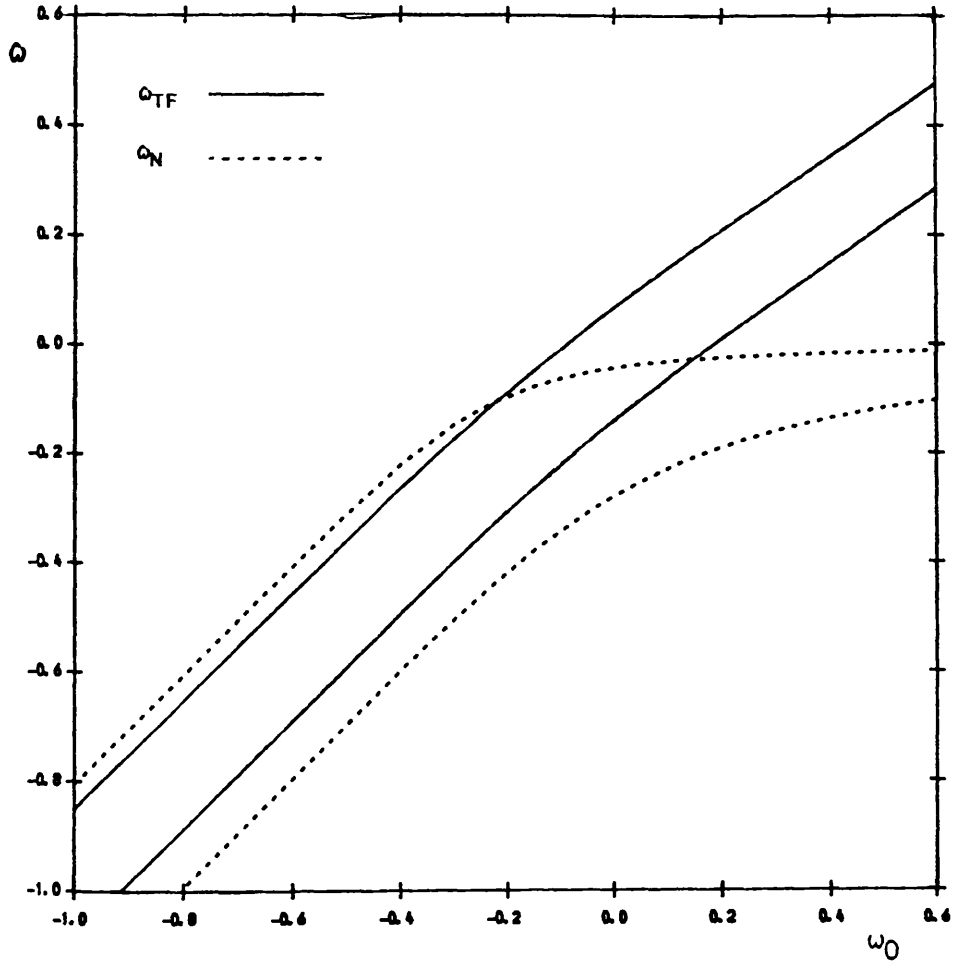

Figure (4.17)

'Equal Tail' (i.e. with $\lambda = \frac{1}{2}\alpha$) 68% confidence interval curves for the true log distance, ω_0, computed from the distribution functions of ω_{TF}, ω_N ($\sigma_M = 1$, $\sigma_P = 0.1$, $\rho = -0.8$)

where **M** and **P** are uncorrelated, ω_{TF} must give precisely the same confidence intervals as ω_N, since as $\rho \to 0$, $\omega_{TF} \to \omega_N$.

The value of the correlation coefficient will also affect the width of confidence intervals constructed from the other estimators. We can easily see this for ω_S in the above example: since ω_S is normally distributed with variance $0.04\sigma_M^2(1-\rho^2)/\rho^2$ at all true distances, it follows immediately that the width of the 68% confidence interval found from this estimator in figure (4.16) is simply twice the dispersion of ω_S, i.e. $0.4\sigma_M\sqrt{(1-\rho^2)}/|\rho|$; the same ρ-dependence is found for confidence intervals of other percentage levels. Thus we see that the intervals found from ω_S can be very narrow if **M** and **P** are sufficiently well correlated, but that if $|\rho|$ is small then the intervals will become very large; wider, indeed, than those obtained from ω_{TF}. This is again consistent with a comparison of the risk of these two estimators.

Provided that the intrinsic variables, **M** and **P**, are suitably correlated, therefore, these results would seem to provide a further indication that the combination of a second observable with apparent magnitude allows one to estimate the true distance of a galaxy significantly more reliably than by using apparent magnitude alone.

4.7 *Extension to More General Cases*

In concluding this chapter we now consider the fact that the precise form of the expressions obtained in section (4.5) for the bias and risk of Q_{GL} - or indeed any estimator - will certainly depend on the particular assumptions which we have made. For example, we have assumed that **P** is measurable directly and that the joint distribution of **M** and **P** is a bivariate normal, independent of position; we have described the selection effects by a Heaviside step function of **m**, as given by equation (4.6), so that measurements of **P** are assumed to be selection-free: it is important to determine the extent to which the properties of these estimators are dependent on any, or all, of the above assumptions. We will therefore now examine some specific examples of ways in which our treatment may be extended to consider more general cases.

4.7.1 **P** Not Directly Measurable

Suppose that the distance independent quantity, **P**, is not measurable directly but rather that one may observe some other quantity, **p**, which is related to **P** via the true distance, r_0 - analogous to absolute and apparent magnitude. (We will refer to **M** and **P** as *intrinsic* random variables, and to **m** and **p** as *extrinsic* random variables.) Consider, for example, the apparent angular diameter and absolute physical diameter of a galaxy, the latter of which we have seen is well correlated with the intrinsic luminosity for a number of different morphological types (c.f. Holmberg, 1969; Paturel, 1979).

Suppose further, for the sake of argument, that measurements of the angular diameter are selection-free. If we identify **P** with the log of absolute diameter and **p** with the log of apparent angular diameter then neglecting, or correcting for, cosmological effects we have:-

$$\mathbf{P} = \mathbf{p} + \log r_0 \equiv \mathbf{p} + \omega_0 + \log r_L \qquad (4.33)$$

This fits the more general relation:-

$$\mathbf{P} = \alpha\mathbf{p} + \beta\omega_0 + \gamma \qquad (4.34)$$

where α,β and γ are constants. If we again assume a bivariate normal for the intrinsic distribution of **M** and **P** and adopt the selection function of equation (4.6), so that we assume the measurements of **p** to be selection-free, we can determine the joint distribution of the extrinsic variables, **m** and **p**, for observable galaxies and use this to compute the distribution, bias and risk of the estimator, Ω_{GL}, defined by:-

$$\Omega_{GL} = A(\mathbf{m}-m_L) + B\mathbf{p} + C \qquad (4.35)$$

We obtain the following expressions for the bias and risk respectively of Ω_{GL}:-

$$B(\Omega_{GL},\omega_0) = (5A-B\beta/\alpha-1)\omega_0 - (A\sigma_M+\rho\sigma_p B\beta/\alpha)\,\frac{\exp[-\tfrac{1}{2}(5\omega_0/\sigma_M)^2]}{\sqrt{2\pi}\,\Phi(-5\omega_0/\sigma_M)}$$

$$+ B(P_0-\gamma)/\alpha + C \qquad (4.36)$$

$$R(\hat{\omega}_{GL},\omega_0) = \omega_0^2(5A-B\beta/\alpha-1)^2 + \frac{(A\sigma_M+\rho\sigma_P B\beta/\alpha)^2\, Q[1.5,\tfrac{1}{2}(5\omega_0/\sigma_M)^2]}{2\,\Phi(-5\omega_0/\sigma_M)}$$

$$- \frac{2(A\sigma_M+\rho\sigma_P B\beta/\alpha)(\omega_0(5A-B\beta/\alpha-1) + B(P_0-\gamma)/\alpha+C)\exp[-\tfrac{1}{2}(5\omega_0/\sigma_M)^2]}{\sqrt{2\pi}\,\Phi(-5\omega_0/\sigma_M)}$$

$$+ 2\omega_0(5A-B\beta/\alpha-1)(B(P_0-\gamma)/\alpha+C) + (B(P_0-\gamma)/\alpha+C)^2 + (B/\alpha)^2\sigma_P^2(1-\rho^2) \qquad (4.37)$$

For the apparent diameter case, given by equation (4.33), the constants of the **p-P** relation take the values $\alpha = 1$, $\beta = 1$ and $\gamma = \log r_L$, which simplifies the expressions for the bias and risk a little. Upon substitution we see that in order to define an unbiased estimator we require to solve the following equations for the constants A, B and C:-

$$\begin{aligned} 5A - B - 1 &= 0 \\ A\sigma_M + B\rho\sigma_P &= 0 \\ B(P_0 - \log r_L) + C &= 0 \end{aligned} \qquad (4.38)$$

We find the following solution:-

$$\begin{aligned} A &= \frac{\rho\sigma_P}{\sigma_M + 5\rho\sigma_P} \\ B &= \frac{-\sigma_M}{\sigma_M + 5\rho\sigma_P} \\ C &= \frac{\sigma_M}{\sigma_M + 5\rho\sigma_P}(P_0 - \log r_L) \end{aligned} \qquad (4.39)$$

Thus, the estimator defined by:-

$$\omega_{GL}^{*} = (\sigma_M + 5\rho\sigma_p)^{-1}(\rho\sigma_p(\mathbf{m}-m_L) - \sigma_M\mathbf{p} + \sigma_M(P_0-\log r_L)) \quad (4.40)$$

is unbiased at all true distances. Moreover, we can see from equation (4.37) that the risk of this estimator is given by $B^2\sigma_p^2(1-\rho^2)$, which is constant, independent of the true distance and is identically equal to the risk of the 'Schechter' estimator.

As in the case where **P** is measurable directly, the values of A, B and C which make ω_{GL}^{*} unbiased are unique, so that there is only one general linear estimator which is unbiased and of constant risk. By suitable combination of A, B and C, however, it is possible to construct estimators which - although biased - have a smaller risk than ω_{GL}^{*} within a particular range of true distances. One may also define a 'maximum likelihood' estimator from the joint distribution of the extrinsic variables **m** and **p**; this estimator is found to have a very small bias and a somewhat smaller risk than ω_{GL}^{*} over a large range of true distances. (This bias is, of course, dependent on ρ, however, and becomes large if the correlation between **M** and **P** is poor.) Thus, the behaviour of the estimators which one may define in the case where **P** is not directly measurable shows no qualitative differences from the results of our earlier analysis.

4.7.2 Arbitrary Selection Function, S(**m**)

One of the most important points made in Schechter (1980) is the fact that if one has a sample of galaxies for which the measurements of **P** are selection-free, then one may obtain unbiased estimates of log distance regardless of how the apparent magnitudes of

the sample are selected. In other words, provided the selection function, S, depends only on the apparent magnitude, **m**, then the 'Schechter' estimator, Φ_S, will be unbiased at all true distances irrespective of the form of S(**m**). It is instructive to confirm this result by computing the bias and risk of Φ_{GL} for a more general selection function.

It is certainly the case that the Heaviside function which we have thus far assumed will not always be a reasonable approximation to the selection function. For example, the cut-off at faint magnitudes may not be sharp, but instead 'smeared out' close to m_L. A more appropriate form for S(**m**) would then be a sigmoid-type function which changes smoothly from 0 to 1 across the magnitude limit (c.f. Teerikorpi, 1975). Rather than consider a specific alternative selection function such as this, however, we will assume that S(**m**) is a completely arbitrary function of the apparent magnitude. Thus, the joint distribution, $\varsigma(\mathbf{m},\mathbf{P}|x_0)$, of **m** and **P** for observable galaxies at true distance, x_0 (assuming for simplicity that **P** is directly measurable) takes the form:-

$$\varsigma(\mathbf{m},\mathbf{P}|x_0) = \frac{\exp[-Q(\mathbf{m},\mathbf{P})/2(1-\rho^2)]\; S(\mathbf{m})}{\sqrt{2\pi(1-\rho^2)}\sigma_M\sigma_P\; \Sigma(x_0)} \qquad (4.41)$$

where Q is given by:-

$$Q(\mathbf{m},\mathbf{P}) = (\mathbf{m}-m_L-5\log x_0)^2/\sigma_M^2 + (\mathbf{P}-P_0)^2/\sigma_P^2 - 2\rho(\mathbf{m}-m_L-5\log x_0)(\mathbf{P}-P_0)/\sigma_M\sigma_P$$

and the normalisation factor, $\Sigma(x_0)$, ensures that $\int\int\varsigma(\mathbf{m},\mathbf{P}|x_0)d\mathbf{m}d\mathbf{P} = 1$. Of course, $\Sigma(x_0)$ will, in general, be different from the normalisation,

$\Phi(-5\log x_0/\sigma_M)$, of equation (4.7).

If we now calculate the bias and risk of Ω_{GL} we obtain the following expressions:-

$$B(\Omega_{GL},\omega_0) = (5A-1)\omega_0 - \frac{(A\sigma_M+\rho B\sigma_P)G(x_0;1)}{\sqrt{2\pi}\,\Sigma(x_0)} + BP_0 + C \qquad (4.42)$$

$$\begin{aligned} R(\Omega_{GL},\omega_0) = {} & \omega_0^2(5A-1)^2 + \frac{(A\sigma_M+\rho B\sigma_P)^2 G(x_0;2)}{\sqrt{2\pi}\,\Sigma(x_0)} \\ & - \frac{2(A\sigma_M+\rho B\sigma_P)(\omega_0(5A-1) + BP_0+C)G(x_0;1)}{\sqrt{2\pi}\,\Sigma(x_0)} \\ & + 2\omega_0(5A-1)(BP_0+C) + (BP_0+C)^2 + B^2\sigma_P^2(1-\rho^2) \end{aligned} \qquad (4.43)$$

where the function $G(x_0;n)$ is defined as:-

$$G(x_0;n) = \int t^n \exp(-\tfrac{1}{2}t^2)S(\sigma_M t+m_L+5\log x_0)dt \qquad (4.44)$$

Comparing the above equations with equations (4.22) and (4.23) we note that they differ from the latter pair only in the presence of the functions $G(x_0;n)$ and $\Sigma(x_0)$; the terms involving the constants A, B and C are unchanged. (Of course when $S(\mathbf{m})$ is given by a Heaviside function both $G(x_0;n)$ and $\Sigma(x_0)$ reduce to their earlier counterparts.) It follows immediately, therefore, that the 'Schechter' estimator is unbiased and has constant risk at all true distances, regardless of the form of the magnitude selection effects.

The properties of any other estimator, however, will depend

on the precise behaviour of S(m). Suppose, for example, that the selection function takes the form of a magnitude window of width ε, viz:-

$$S(m) = \begin{cases} 1 & m_L - \varepsilon < m < m_L \\ 0 & \text{otherwise} \end{cases} \qquad (4.45)$$

(i.e. bright galaxies are rejected from the sample in addition to the denial of very faint galaxies.) Assuming this selection function, $G(x_0;1)$ is then found to be given by:-

$$G(x_0;1) = \exp[-\tfrac{1}{2}((\varepsilon+5\omega_0)/\sigma_M)^2] - \exp[-\tfrac{1}{2}(5\omega_0/\sigma_M)^2] \qquad (4.46)$$

and the normalisation factor, $\Sigma(x_0)$ is given by:-

$$\Sigma(x_0) = \Phi(-5\omega_0/\sigma_M) - \Phi(-(\varepsilon+5\omega_0)/\sigma_M) \qquad (4.47)$$

Substituting into equation (4.40) we find that the bias of the 'Tully-Fisher' estimator, ω_{TF}, is therefore given by:-

$$B(\omega_{TF},\omega_0) = 0.2\sigma_M(1-\rho^2)\frac{(\exp[-\tfrac{1}{2}((\varepsilon+5\omega_0)/\sigma_M)^2] - \exp[-\tfrac{1}{2}(5\omega_0/\sigma_M)^2])}{\sqrt{2\pi}\,(\Phi(-5\omega_0/\sigma_M) - \Phi(-(\varepsilon+5\omega_0)/\sigma_M))} \qquad (4.48)$$

Figure (4.18) shows the bias of ω_{TF} as a function of true distance, x_0, for a magnitude selection 'window' of width ε = 5 mag. We can see from this graph that ω_{TF} has a significant positive bias for small x_0. If we use equation (4.48) to examine the limiting behaviour of $B(\omega_{TF},\omega_0)$, we find that the bias of ω_{TF} does not tend to zero as the true distance tends to zero, as was the case with only a faint magnitude limit; a similar increase is found in the risk at small x_0. This is simply a consequence of the fact that the mean absolute

magnitude of observable galaxies is no longer equal to M_0 at small true distances, but is in fact *fainter* than M_0 (resulting in the systematic *over*-estimation of distances) because the selection effects remove brighter galaxies from the sample at small true distances.

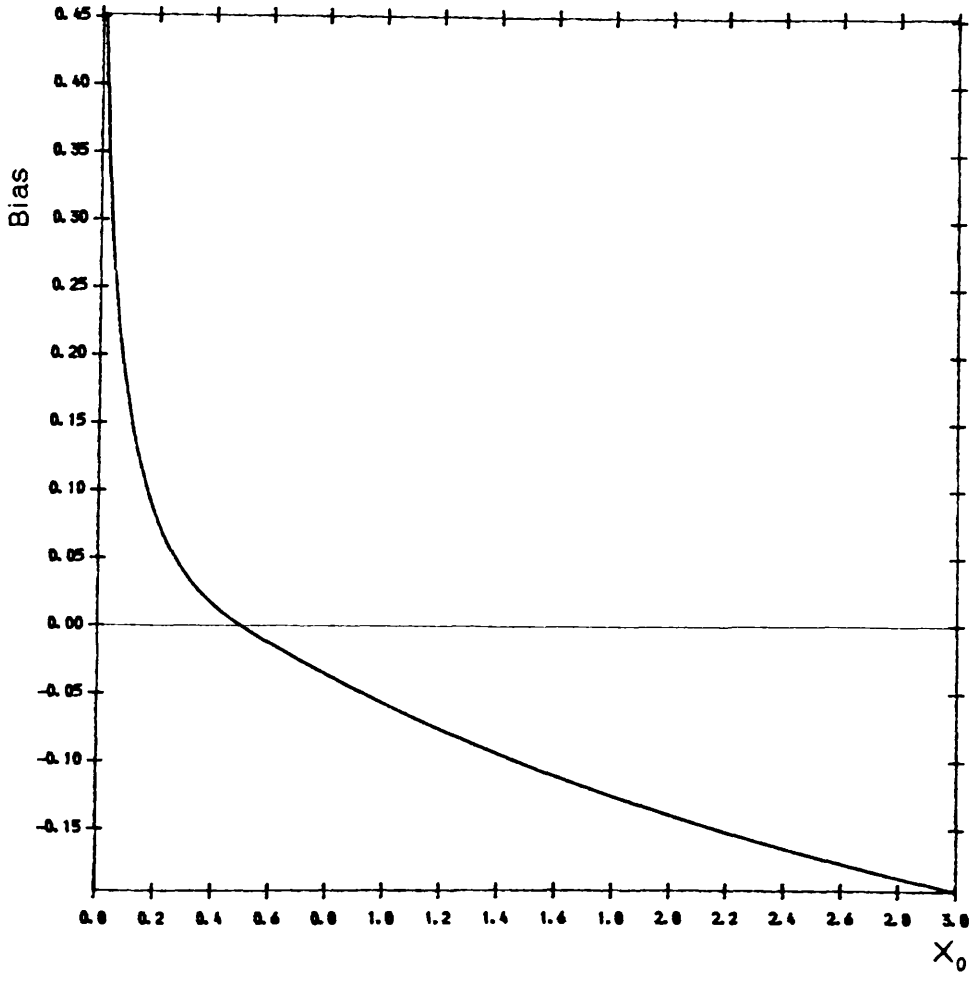

Figure (4.18)

Bias of the 'Tully Fisher' estimator as a function of true distance, x_0, assuming a magnitude selection function given by a 'window' of width, δ = 5 mag. (σ_M = 1, σ_P = 0.1, ρ = -0.8)

A particular problem introduced by a narrow magnitude selection function is the fact that there will be only a narrow range of distances within which the bias of the 'Tully-Fisher' estimator is close to zero, whereas we can see from figure (4.13) that with only a faint magnitude limit the bias of $\mathcal{Q}_{TF}$ is very small for *all* $x_0 < 0.7$; the same problem afflicts to an even greater extent the 'naive' estimator $\mathcal{Q}_N$, defined in equation (4.27). As we have seen in section (2.2) this effect places severe limitations on the usefulness of the Minimum bias Subset for removing selection bias.

4.7.3. Estimators Derived from the D_n-σ Relation

The analysis which we have presented thus far has concerned the estimation of distance from the combination of apparent magnitude with a second observable. It is straightforward to adapt this analysis to deal with the case of the D_n-σ relation, where one estimates distances by combining the central velocity dispersion of a galaxy with its measured apparent angular diameter. To this end, therefore, suppose that the log of absolute diameter and the log of velocity dispersion of a galaxy are random variables - denoted by **D** and **P** respectively - whose intrinsic joint distribution is a bivariate normal. (We use **P** instead of σ to denote the log of velocity dispersion so as to avoid confusion with the dispersions of the bivariate normal distribution, σ_D and σ_P, which characterise the intrinsic scatter in the D_n-σ relation.) Denote by **d** the log of apparent angular diameter which is related to **D**, after cosmological corrections (c.f. Burstein and Heiles, 1982) by:-

$$\mathbf{d} = \mathbf{D} - \mu_0 \qquad (4.49)$$

where μ_0 is the true log distance. Suppose now that we observe a sample of galaxies and measure **d** and **P** for each object. Suppose, further, that the measurements of diameter are subject to selection effects described by a selection function, S(**d**), but that the velocity dispersion measurements are selection-free; this is approximately the case for the selection effects of the Lynden-Bell *et al*, (1988) data set (c.f. section 2.3.3).

We can then derive the joint distribution of **d** and **P**, $\varsigma(\mathbf{d},\mathbf{P}|\mu_0)$, at given true log distance, μ_0, taking into account the sample selection effects; this joint distribution may be used to define estimators of μ_0. Consider for example the 'general linear' estimator, $\hat{\mu}_{GL}$, defined by:-

$$\hat{\mu}_{GL} = A\mathbf{d} + B\mathbf{P} + C \qquad (4.50)$$

To derive expressions for the bias and risk of $\hat{\mu}_{GL}$ we merely require to follow an essentially equivalent analysis to that of section (4.7.2), with the apparent diameter now replacing apparent magnitude as the observable which is subject to selection effects. We obtain the following results:-

$$B(\hat{\mu}_{GL},\mu_0) = -(A+1)\mu_0 + \frac{(A\sigma_D+\rho B\sigma_P)G(\mu_0;1)}{\sqrt{2\pi}\,\Sigma(\mu_0)} + AD_0 + BP_0 + C \qquad (4.51)$$

$$R(\hat{\mu}_{GL},\mu_0) = \mu_0^2(A+1)^2 + \frac{(A\sigma_D+\rho B\sigma_P)^2\, G(\mu_0;2)}{\sqrt{2\pi}\,\Sigma(\mu_0)}$$

$$+ \frac{2(A\sigma_D+\rho B\sigma_P)(AD_0+BP_0+C - \mu_0(A+1))G(\mu_0;1)}{\sqrt{2\pi}\,\Sigma(\mu_0)}$$

$$+ (AD_0+BP_0+C)^2 - 2\mu_0(A+1)(AD_0+BP_0+C) + B^2\sigma_P^2(1-\rho^2) \qquad (4.52)$$

where $G(\mu_0;n)$ is defined as:-

$$G(\mu_0;n) = \int t^n \exp(-\tfrac{1}{2}t^2)S(\sigma_D t+D_0-\mu_0)dt \qquad (4.53)$$

and $\Sigma(\mu_0)$ normalises $\zeta(\mathbf{d},\mathbf{P}|\mu_0)$.

Note the similarity between these expressions and those of equations (4.42) and (4.43): note in particular that it is possible to define an unbiased estimator of μ_0 by choosing the constants $A = -1$, $B = \sigma_D/\rho\sigma_P$ and $C = D_0 - P_0\sigma_D/\rho\sigma_P$. i.e. the estimator, $\hat{\mu}_{GL}^*$, defined by:-

$$\hat{\mu}_{GL}^* = \sigma_D/\rho\sigma_P\mathbf{P} - \mathbf{d} + D_0 - P_0\sigma_D/\rho\sigma_P \qquad (4.54)$$

is an unbiased estimator of the true log distance, μ_0, for all values of μ_0. This is precisely the same estimator as the 'Schechter' estimator, $\hat{\mu}_S$, which one may construct from the D_n-σ relation; i.e. the estimator which corresponds to calibration of the D_n-σ relation by a regression of velocity dispersions on apparent diameters. The risk of $\hat{\mu}_S$ is found to be constant, independent of the true distance, and equal to $\sigma_D^2(1-\rho^2)/\rho^2$, which again corresponds exactly to our previous result for $\hat{\omega}_S$. Moreover, we can see from equations (4.49) and (4.50) that the constants which define $\hat{\mu}_S$ as unbiased are unique: estimators

constructed from any other regression line - or indeed any other arbitrary values of A, B and C - will certainly be biased. Thus, provided that the selection effects may be adequately described by a function only of the apparent diameter, we see that the 'Schechter' scheme for identifying unbiased distances applies equally well to the D_n-σ relation.

4.8 Summary and Concluding Remarks

In this chapter we have investigated the properties of distance estimators which are functions of two observables, as is the case for distances derived from e.g. the Tully-Fisher or D_n-σ relations. Assuming that the intrinsic scatter in each relation is described by a bivariate normal distribution, we have derived expressions for the bias and risk of a 'general linear' estimator (i.e. a linear combination of the two observables) of log distance. We have shown that, provided the measurements of one of the observables are free from selection effects, then it possible to define a 'general linear' estimator which is unbiased at all true distances. This unbiased estimator corresponds exactly to the scheme proposed by Schechter (1980), whereby one calibrates the Tully-Fisher or D_n-σ relation by minimising the residuals on the selection-free variable (in the case of Tully-Fisher, for example, one regresses line widths on magnitudes).

Moreover, we have shown that the risk of this 'Schechter' estimator, ω_S, is a constant independent of the true distance, so that the percentage risk of *distance* estimates obtained from ω_S will be

constant. Similarly, confidence intervals for the true log distance constructed using ω_S, following the method introduced in section (3.5), will be of constant width. We have also shown that even if both observables are extrinsic varaiables (as is the case for e.g. apparent magnitude and apparent angular diameter) then one may still define an estimator which has zero bias and constant risk at all true distances - again dependent only on one of the observables being selection-free.

In each of the cases which we have considered the 'Schechter' estimator is given uniquely in terms of the parameters of the relevant intrinsic joint distribution: e.g. ω_S for the Tully-Fisher relation is defined in terms of ρ, σ_M and σ_P. It is, therefore, very important to use accurate values for these parameters; clearly a failure to do so will result in a biased estimator, and the bias and risk may quickly become non-negligible due to the non-linear form of the expressions which we have obtained. One possible source of error in the estimates of ρ, σ_M and σ_P is the use of a poor calibrating sample; e.g. one that carries a significant zero-point error, or is incomplete in **M** or **P**, or contaminated by foreground galaxies which are not correctly identified. Problems of this kind which arise in the use of Tully-Fisher type relations have been discussed in some detail by several authors (c.f. Teerikorpi, 1989; Tammann, 1987).

Even if these distribution parameters are well-determined, the properties of the 'Schechter' estimator will still depend crucially on the assumption that one observable is free from selection; if this is not the case then it will no longer be possible to define an unbiased estimator at all true distances. Nevertheless, for any *given* selection

function, one may still compute the bias and risk of a 'general linear' estimator at any true distance - although this may no longer be possible analytically. One strategy for identifying a 'best' estimator would then be to find the coefficients A, B and C which minimise the bias or risk (or some chosen combination of both) over a relevant range of true distances. We will comment further on the application of this strategy in chapter (6).

5. ESTIMATION OF DISTANCE USING THREE OBSERVABLES

5.1 Introduction

In this chapter, as a natural extension of the results of chapters (3) and (4), we will consider the properties of distance estimators which are a function of the measured apparent magnitude and of two other observable quantities. We will derive expressions for the joint distribution of three observables at a given true distance, taking into account sample selection effects, and use this distribution to compute the bias and risk of a 'general linear' estimator of the true log distance. We will then show that, in this extended case, it is again possible to derive unbiased distance estimators analogous to the 'Schechter' estimators of the preceding chapter. Moreover we will demonstrate that, as one might expect, it is possible to construct unbiased estimators which have smaller risk than their counterparts in chapter (4).

5.2 The Observed Distribution of m, P and D

Our analysis will follow precisely the same formulation as section (4.2), differing only in the inclusion of an additional observable. Let the absolute magnitude, **M**, and position, $\underline{r}$, of a galaxy be random variables. Further, suppose that **P** and **D** are random variables which denote intrinsic physical characteristics of the galaxy such that **M**, **P** and **D** are correlated. (An example of three such variables - which is suggested by our choice of notation, following

that of the previous chapter - is the absolute magnitude, absolute diameter and velocity dispersion.) Suppose, however, that each of **M**, **P** and **D** is uncorrelated with $\underline{r}$ so that we may introduce $\Psi(\mathbf{M},\mathbf{P},\mathbf{D})$, the intrinsic joint distribution of **M**, **P** and **D**, which is independent of position. Consider now the joint distribution, $\rho(\mathbf{M},\mathbf{P},\mathbf{D},\underline{r})$, of **M**, **P**, **D** and $\underline{r}$ for observable galaxies in a sample subject to selection effects - as described by a selection function, $S(\mathbf{M},\mathbf{P},\mathbf{D},|\underline{r}|)$, defined in the same way as before. It follows that $\rho(\mathbf{M},\mathbf{P},\mathbf{D},\underline{r})$ is given by:-

$$\rho(\mathbf{M},\mathbf{P},\mathbf{D},\underline{r}) = \frac{\Psi(\mathbf{M},\mathbf{P},\mathbf{D})n(\underline{r})S(\mathbf{M},\mathbf{P},\mathbf{D},|\underline{r}|)}{\iiiint \Psi(\mathbf{M},\mathbf{P},\mathbf{D})n(\underline{r})S(\mathbf{M},\mathbf{P},\mathbf{D},|\underline{r}|)d\mathbf{M}d\mathbf{P}d\mathbf{D}dV} \qquad (5.1)$$

where $n(\underline{r})$ is the number density of galaxies (of all values of **M**, **P** and **D**) at position $\underline{r}$. Using this equation we find that the conditional distribution, $\xi(\mathbf{M},\mathbf{P},\mathbf{D}|r_0)$, of the intrinsic variables, **M**, **P** and **D** at a given distance, r_0, for observable galaxies is given by:-

$$\xi(\mathbf{M},\mathbf{P},\mathbf{D}|r_0) = \frac{\Psi(\mathbf{M},\mathbf{P},\mathbf{D})S(\mathbf{M},\mathbf{P},\mathbf{D},r_0)}{\iiint \Psi(\mathbf{M},\mathbf{P},\mathbf{D})S(\mathbf{M},\mathbf{P},r_0)d\mathbf{M}d\mathbf{P}d\mathbf{D}} \qquad (5.2)$$

Note that, as before, this distribution is independent of the local density, $n(\underline{r})$.

We can re-express equation (5.2) in terms of *extrinsic* random variables to allow for the fact that **P** or **D** may not be measurable directly, as we have already noted in the case of e.g. the absolute diameter of a galaxy. Hence, in addition to the apparent magnitude, **m**, we introduce the extrinsic random variables **p** and **d**, related to **P** and **D** as follows (c.f. equation 4.34):-

$$\mathbf{P} = \alpha_1 \mathbf{p} + \beta_1 \omega_0 + \gamma_1$$
$$\mathbf{D} = \alpha_2 \mathbf{d} + \beta_2 \omega_0 + \gamma_2 \qquad (5.3)$$

where α_i, β_i and γ_i are constants and ω_0 is the true log distance in scaled units. (If either **P** or **D** is directly measurable then we will simply have the special case of $\alpha_i = 1$, $\beta_i = 0$, $\gamma_i = 0$.) Thus, by changing variables in equation (5.2) to **m**, **p** and **d**, we may determine the joint distribution of the extrinsic variables for observable galaxies at true log distance, ω_0, given an arbitrary selection function, $S(\mathbf{m},\mathbf{p},\mathbf{d})$, and for an arbitrary joint distribution, $\Psi(\mathbf{M},\mathbf{P},\mathbf{D})$, of the intrinsic variables (provided only that such a distribution exists!). In the present treatment, however, we will consider only two specific cases.

5.3 Case 1: Selection Only on Apparent Magnitude

Suppose, firstly, that the selection function, S, is a function only of the apparent magnitude. i.e. $S = S(\mathbf{m})$. Thus we assume that the measurements of **p** and **d** are free from selection effects. Suppose, further, that the intrinsic joint distribution $\Psi(\mathbf{M},\mathbf{P},\mathbf{D})$ is given by a trivariate normal distribution. Hence Ψ takes the form:-

$$\Psi(\mathbf{M},\mathbf{P},\mathbf{D}) = (2\pi)^{-3/2}|\mathbf{V}|^{-1/2} \exp[-\tfrac{1}{2}\mathbf{Q}] \qquad (5.4)$$

where **Q** is a quadratic form in **M**, **P** and **D** which also involves the elements of the covariance matrix, **V**, of the intrinsic variables. In the trivariate case the covariance matrix may be completely specified by six parameters: the dispersions of **M**, **P** and **D** - denoted σ_M, σ_P and σ_D

respectively - and the three correlation coefficients, denoted ρ_{MP}, ρ_{MD} and ρ_{PD}, which measure the correlation between each pair of the variables. In this notation, **V** is given by:-

$$\mathbf{V} = \begin{bmatrix} \sigma_M^2 & \sigma_M\sigma_P\rho_{MP} & \sigma_M\sigma_D\rho_{MD} \\ \sigma_M\sigma_P\rho_{MP} & \sigma_P^2 & \sigma_P\sigma_D\rho_{PD} \\ \sigma_M\sigma_D\rho_{MD} & \sigma_P\sigma_D\rho_{PD} & \sigma_D^2 \end{bmatrix} \qquad (5.5)$$

The precise form of the trivariate normal may be written down as a special case of the general multivariate normal distribution, which is studied in detail in many standard textbooks on statistics and probability (c.f. Graybill, 1961; Kendall and Stuart, 1963). For *any* trivariate distribution function, however, the following relation will always hold:-

$$\Psi(\mathbf{M},\mathbf{P},\mathbf{D}) = \Psi_1(\mathbf{M})\Psi_2(\mathbf{P},\mathbf{D}|\mathbf{M}) \qquad (5.6)$$

i.e. we may always express Ψ as the product of the marginal distribution, $\Psi_1(\mathbf{M})$, of **M** multiplied by the conditional distribution, $\Psi_2(\mathbf{P},\mathbf{D}|\mathbf{M})$, of the other two variables at a given value of **M**. Given the form which we have assumed for the selection effects, it is useful to write Ψ as a product in this way, since we can make use of the result that when Ψ is multivariate normal, then so too will be Ψ_1 and Ψ_2. A proof of this property may be found in Graybill (1961), which also derives expressions for the means and covariance matrices of the two distributions, Ψ_1 and Ψ_2. Applying the results of Graybill, we find that Ψ may be written as:-

$$\Psi(\mathbf{M},\mathbf{P},\mathbf{D}) = \frac{\exp[-1/2\sigma_M^2(\mathbf{M}-M_0)^2]}{(2\pi)^{\frac{1}{2}}\sigma_M} \cdot \frac{\exp[-\frac{1}{2}\mathbf{Q}_1]}{2\pi\ |\mathbf{V}_1|^{\frac{1}{2}}} \qquad (5.7)$$

where $\mathbf{Q}_1$ is a quadratic form in **M**, **P** and **D**, and $\mathbf{V}_1$ is the covariance matrix of the conditional distribution, $\Psi_2(\mathbf{P},\mathbf{D}|\mathbf{M})$, given by:-

$$\mathbf{V}_1 = \begin{bmatrix} \sigma_P^2(1-\rho_{MP}^2) & \sigma_P\sigma_D(\rho_{PD}-\rho_{MP}\rho_{MD}) \\ \sigma_P\sigma_D(\rho_{PD}-\rho_{MP}\rho_{MD}) & \sigma_D^2(1-\rho_{MD}^2) \end{bmatrix} \qquad (5.8)$$

The main advantage of splitting the trivariate distribution in this way is that the quadratic form, $\mathbf{Q}_1$, may now be compactly expressed in terms of the equations for the regression lines of **P** on **M** and **D** on **M**, which of course we have already seen in the preceding chapter play a crucial role in the definition of an unbiased 'Schechter' estimator. $\mathbf{Q}_1$ takes the following form:-

$$\mathbf{Q}_1 = |\mathbf{V}_1|^{-1}.(\sigma_D^2(1-\rho_{MD}^2)\mathbf{E}_P^2 - 2\sigma_P\sigma_D(\rho_{PD}-\rho_{MP}\rho_{MD})\mathbf{E}_P\mathbf{E}_D + \sigma_P^2(1-\rho_{MP}^2)\mathbf{E}_D^2) \qquad (5.9)$$

and $\mathbf{E}_P$ and $\mathbf{E}_D$ are shorthand for the following:-

$$\mathbf{E} = \mathbf{P} - P_0 - \rho_{MP}\sigma_P/\sigma_M(\mathbf{M} - M_0) \qquad (5.10)$$

$$\mathbf{E}_D = \mathbf{D} - D_0 - \rho_{MD}\sigma_D/\sigma_M(\mathbf{M} - M_0) \qquad (5.11)$$

i.e. $\mathbf{E}_P = 0$ and $\mathbf{E}_D = 0$ define the '**P** on **M**' and '**D** on **M**' regression lines respectively.

If we now substitute in $\mathbf{Q}_1$ for **P** and **D** in terms of the extrinsic variables, **p** and **d**, as given by equations (5.3), and also for

the absolute magnitude, **M**, in terms of **m** using equation (3.12), then we obtain the following expression for the joint distribution, $\varsigma(\mathbf{m},\mathbf{p},\mathbf{d}|\omega_0)$, of the extrinsic variables for observable galaxies at true log distance, ω_0:-

$$\varsigma(\mathbf{m},\mathbf{p},\mathbf{d}|\omega_0) = \frac{\exp[-1/2\sigma_M^2(\mathbf{m}-m_L-5\omega_0)^2 - 1/2Q_1(\mathbf{m},\mathbf{p},\mathbf{d})]\ S(\mathbf{m})}{\alpha_1\alpha_2(2\pi)^{\frac{1}{2}}\sigma_M\ 2\pi|V_1|^{\frac{1}{2}}\ \Sigma(\omega_0)} \qquad (5.12)$$

where $\Sigma(\omega_0)$ normalises ς at each true log distance, ω_0, and α_1 and α_2 are introduced by the transformation from **(P,D)** to **(p,d)**, as in equation (5.3).

5.3.1 Bias and Risk of Ω_{GL}

We define a 'general linear' estimator of ω_0 as follows:-

$$\Omega_{GL} = A(\mathbf{m} - m_L) + B_1\mathbf{p} + B_2\mathbf{d} + C \qquad (5.13)$$

where A, B_1, B_2 and C are constants and the observables **p** and **d** are related to the intrinsic variables **P** and **D** via equations (5.3). Following the same approach as for the bivariate case, we can use the distribution of equation (5.12) to determine the bias and risk of Ω_{GL} for arbitrary constants A, B_1, B_2 and C. We obtain the following:-

$$\begin{aligned} B(\Omega_{GL},\omega_0) &= (5A-B_1\beta_1/\alpha_1-B_2\beta_2/\alpha_2-1)\omega_0 \\ &+ \frac{(A\sigma_M+B_1\rho_{MP}\sigma_P/\alpha_1+B_2\rho_{MD}\sigma_D/\alpha_2)\ G(x_0;1)}{\sqrt{2\pi}\ \Sigma(x_0)} \\ &+ B_1(P_0-\gamma_1)/\alpha_1 + B_2(D_0-\gamma_2)/\alpha_2 + C \end{aligned} \qquad (5.14)$$

$$R(\Omega_{GL},\omega_0) = \omega_0^2(5A-B_1\beta_1/\alpha_1-B_2\beta_2/\alpha_2-1)^2$$

$$+ \frac{(A\sigma_M+B_1\rho_{MP}\sigma_P/\alpha_1+B_2\rho_{MD}\sigma_D/\alpha_2)^2\, G(x_0;2)}{\sqrt{2\pi}\,\Sigma(x_0)}$$

$$+ 2(A\sigma_M+B_1\rho_{MP}\sigma_P/\alpha_1+B_2\rho_{MD}\sigma_D/\alpha_2)\Big\{\omega_0(5A-B_1\beta_1/\alpha_1-B_2\beta_2/\alpha_2-1)$$

$$+ B_1(P_0-\gamma_1)/\alpha_1+ B_2(D_0-\gamma_2)/\alpha_2 + C\Big\}\frac{G(x_0;1)}{\sqrt{2\pi}\,\Sigma(x_0)}$$

$$+ 2\omega_0(5A-B_1\beta_1/\alpha_1-B_2\beta_2/\alpha_2-1)(B_1(P_0-\gamma_1)/\alpha_1+B_2(D_0-\gamma_2)/\alpha_2+C)$$

$$+ (B_1(P_0-\gamma_1)/\alpha_1 + B_2(D_0-\gamma_2)/\alpha_2 + C)^2$$

$$+ (B_1/\alpha_1)^2\sigma_P^2(1-\rho_{MP}^2) + (B_2/\alpha_2)^2\sigma_D^2(1-\rho_{MD}^2)$$

$$+ 2(B_1/\alpha_1)(B_2/\alpha_2)\sigma_P\sigma_D(\rho_{PD}-\rho_{MP}\rho_{MD}) \qquad (5.15)$$

where $G(x_0;n)$ is as defined in equation (4.44)

We can see immediately that these expressions, although somewhat lengthier in form, are very similar to those obtained for the bivariate case. In particular the functions $G(x_0;n)$ and $\Sigma(x_0)$ - which determine the dependence of the bias and risk on the true distance - appear identically in both cases: the additional contribution of B_2**d** to the estimate of ω_0 changes only the *coefficients* of the distance-dependent terms. It follows immediately, therefore, that if we use only the measured values of **m** and **p** to estimate ω_0 (i.e. setting B_2 equal to zero) then the above expressions reduce exactly to those of the bivariate case - as one would expect.

This correspondence with our earlier results certainly indicates that, by suitable choice of the constants A, B_1, B_2 and C, one can define an estimator which is unbiased at all true distances;

the trivial choice of $B_2 = 0$ would reduce the problem simply to that of defining A, B_1 and C for the appropriate bivariate 'Schechter' estimator. Intuitively, however, one would hope that by making use of an additional observable, one would be able to improve upon the 'Schechter' estimators of chapter (4): in other words to find unbiased estimators with lower risk than ω_S.

We can see from equation (5.14) that the constants which define an unbiased estimator must satisfy the following three equations:-

$$\begin{aligned} 5A - \frac{B_1\beta_1}{\alpha_1} - \frac{B_2\beta_2}{\alpha_2} - 1 &= 0 \\ A\sigma_M + \frac{B_1\rho_{MP}\sigma_P}{\alpha_1} + \frac{B_2\rho_{MD}\sigma_D}{\alpha_2} &= 0 \\ \frac{B_1(P_0 - \gamma_1)}{\alpha_1} + \frac{B_2(D_0 - \gamma_2)}{\alpha_2} + C &= 0 \end{aligned} \qquad (5.16)$$

Furthermore, provided these equations hold then it follows that all but the final three terms of equation (5.15) will also vanish - i.e the risk of ω_{GL} will be independent of the true distance and given by:-

$$\begin{aligned} R(\omega_{GL},\omega_0) = {} & (B_1/\alpha_1)^2\sigma_P^2(1-\rho_{MP}^2) + (B_2/\alpha_2)^2\sigma_D^2(1-\rho_{MD}^2) \\ & + 2(B_1/\alpha_1)(B_2/\alpha_2)\sigma_P\sigma_D(\rho_{PD}-\rho_{MP}\rho_{MD}) \end{aligned} \qquad (5.17)$$

If we consider again equations (5.16) the potential value of utilising a third observable becomes apparent: to obtain an unbiased estimator we now must satisfy three equations in the four unknowns

A, B_1, B_2 and C. The unbiased solution for ω_{GL} is, therefore, no longer unique; in particular we may solve for the values of A, B_1, B_2 and C which give an unbiased estimator and at the same time which minimise the risk of ω_{GL}, as given by equation (5.15).

The method of Lagrange multipliers is amenable to finding such a solution. We will not present here a treatment of the general equations (5.14) and (5.15), but rather we will consider the specific example where the intrinsic variables **P** and **D** are both directly measurable (i.e. **p** = **P** and **d** = **D**, so that $\alpha_i = 1$, $\beta_i = 0$ and $\gamma_i = 0$). Thus, we find that we now require to minimise the risk, R, given by:-

$$R = B_1^2\sigma_P^2(1-\rho_{MP}^2) + B_2^2\sigma_D^2(1-\rho_{MD}^2) + 2B_1B_2\sigma_P\sigma_D(\rho_{PD}-\rho_{MP}\rho_{MD}) \qquad (5.18)$$

subject to the constraint equations:-

$$\begin{aligned} A\sigma_M + B_1\sigma_P\rho_{MP} + B_2\sigma_D\rho_{MD} &= 0 \\ 5A - 1 &= 0 \qquad (5.19) \\ B_1P_0 + B_2D_0 + C &= 0 \end{aligned}$$

This is equivalent to minimising the *unconstrained* expression, R*, given by:-

$$R^* = R + \lambda(0.2\sigma_M + B_1\sigma_P\rho_{MP} + B_2\sigma_D\rho_{MD}) \qquad (5.20)$$

where λ is a Lagrange multiplier. Note that A = 0.2 follows immediately from the second constraint equation. Moreover, we need not include the third of the constraint equations since this has no bearing on the minimisation of the risk and can always be satisfied for *any* values of

B_1 and B_2 by taking $C = -B_1P_0 - B_2D_0$.

Taking partial derivatives of equation (5.20) with respect to B_1, B_2 and λ we obtain three equations which, when set equal to zero, define the values of B_1 and B_2 for the unbiased minimum-risk estimator. These are:-

$$\begin{aligned} 2B_1\sigma_P^2(1-\rho_{MP}^2) + 2B_2\sigma_P\sigma_D(\rho_{PD}-\rho_{MP}\rho_{MD}) + \lambda\sigma_P\rho_{MP} &= 0 \\ 2B_2\sigma_D^2(1-\rho_{MD}^2) + 2B_1\sigma_P\sigma_D(\rho_{PD}-\rho_{MP}\rho_{MD}) + \lambda\sigma_D\rho_{MD} &= 0 \\ 0.2\sigma_M + B_1\sigma_P\rho_{MP} + B_2\sigma_D\rho_{MD} &= 0 \end{aligned} \qquad (5.21)$$

Solving these equations we find that the unbiased minimum risk trivariate estimator, ω_T, is therefore defined by the following constants:-

$$A = 0.2$$

$$B_1 = \frac{-0.2\sigma_M}{\sigma_P} \cdot \frac{(\rho_{MP} - \rho_{MD}\rho_{PD})}{(\rho_{MP}^2 - 2\rho_{MP}\rho_{MD}\rho_{PD} + \rho_{MD}^2)}$$

$$B_2 = \frac{-0.2\sigma_M}{\sigma_D} \cdot \frac{(\rho_{MD} - \rho_{MP}\rho_{PD})}{(\rho_{MP}^2 - 2\rho_{MP}\rho_{MD}\rho_{PD} + \rho_{MD}^2)} \qquad (5.22)$$

$$C = \frac{0.2\sigma_M\,(\sigma_D(\rho_{MP} - \rho_{MD}\rho_{PD})P_0 + \sigma_D(\rho_{MD} - \rho_{MP}\rho_{PD})D_0)}{\sigma_D\sigma_P(\rho_{MP}^2 - 2\rho_{MP}\rho_{MD}\rho_{PD} + \rho_{MD}^2)}$$

This solution is only defined if $\rho_{MP}^2 - 2\rho_{MP}\rho_{MD}\rho_{PD} + \rho_{MD}^2 \neq 0$. It can be shown, however, that this expression will always be strictly positive provided $\rho_{MP} \neq 0$ or $\rho_{MD} \neq 0$, and $\rho_{PD} \neq \pm 1$.

There are a number of interesting features about this solution. Firstly, note that the expressions for B_1 and B_2 are identical in form and differ only in the labels of the correlation coefficients and dispersions; this must be the case since both **P** and **D** were treated equivalently in deriving the minimum risk solution. It follows that if $\rho_{MD} = \rho_{MP}$ and $\sigma_P = \sigma_D$ then $B_1 = B_2 = -0.2\sigma_M/2\sigma_P\rho_{MP}$, independent of the value of ρ_{PD}; this is precisely half the value of the coefficient of **P** in the 'Schechter' estimator, ω_S.

In general, the values of the correlation coefficients and dispersions will determine the relative size of B_1 and B_2, and hence the relative contribution of **P** and **D** respectively to ω_T. Note that σ_P and σ_D appear on the denominator of B_1 and B_2; this means that if the intrinsic dispersion of **P** or **D** is large, then that observable will make a proportionately smaller contribution to the distance estimate. Note also that if we have $\rho_{MD} = \rho_{PD} = 0$, then the solution reduces precisely to ω_S. This is exactly as one would expect: if we measure a third observable which is uncorrelated with both **M** and **P**, then the minimum risk unbiased estimator formed from all three observables will be no better than the 'Schechter' estimator defined from **m** and **P** alone.

It is instructive to examine how the values of B_1 and B_2 change as a function of the correlation coefficients upon the addition of a third observable. We can assume that third observable to be **D**, without loss of generality, and from equations (5.22) calculate B_1 and B_2 as a function of ρ_{MD} and ρ_{PD}, for a fixed value of ρ_{MP}. A useful measure of the relative contribution of B_1 and B_2 is then given by the

ratio B_2/B_1. From equations (5.22) we have:-

$$B_2/B_1 = \frac{\sigma_P(\rho_{MD} - \rho_{MP}\rho_{PD})}{\sigma_D(\rho_{MP} - \rho_{MD}\rho_{PD})} \qquad (5.23)$$

Consider now some numerical examples. Suppose that $\sigma_P = \sigma_D$, and that $\rho_{MP} = -0.8$. Figure (5.1) shows an isometric surface plot of the ratio, B_2/B_1, as a function of ρ_{MD} and ρ_{PD}. (The shaded regions of the ρ_{MD}-ρ_{PD} plane are disallowed owing to the fact that the covariance matrix, $\mathbb{V}$, of the trivariate distribution must be positive definite.)

It follows from equation (5.23) that the intersection of this surface with the ρ_{PD}-ρ_{MD} plane is the straight line $\rho_{MD} = \rho_{MP}\rho_{PD}$; i.e. for all points on this line - including, of course, the point $(\rho_{MD},\rho_{PD}) = (0,0)$ - we find that B_2 is equal to zero. Hence $\mathcal{Q}_T$ again reduces precisely to $\mathcal{Q}_S$ and **D** makes no contribution to the distance estimate. This means, in other words, that using the measured value of **D** would not improve upon the 'Schechter' distance estimate obtained from **m** and **P** alone. We will refer to this line as the 'Schechter' line, for obvious reasons, and we will explore in more detail some interesting - and potentially very important - consequences of its existence a little later.

We can see from figure (5.1) that as ρ_{MD} increases in modulus then so also does the value of B_2/B_1. In other words as the correlation between magnitude and the third observable, **D**, increases then so does the relative contribution of $B_2\mathbf{D}$ to the distance estimate.

Suppose, for example, that ρ_{PD} were equal to 0.5. We then

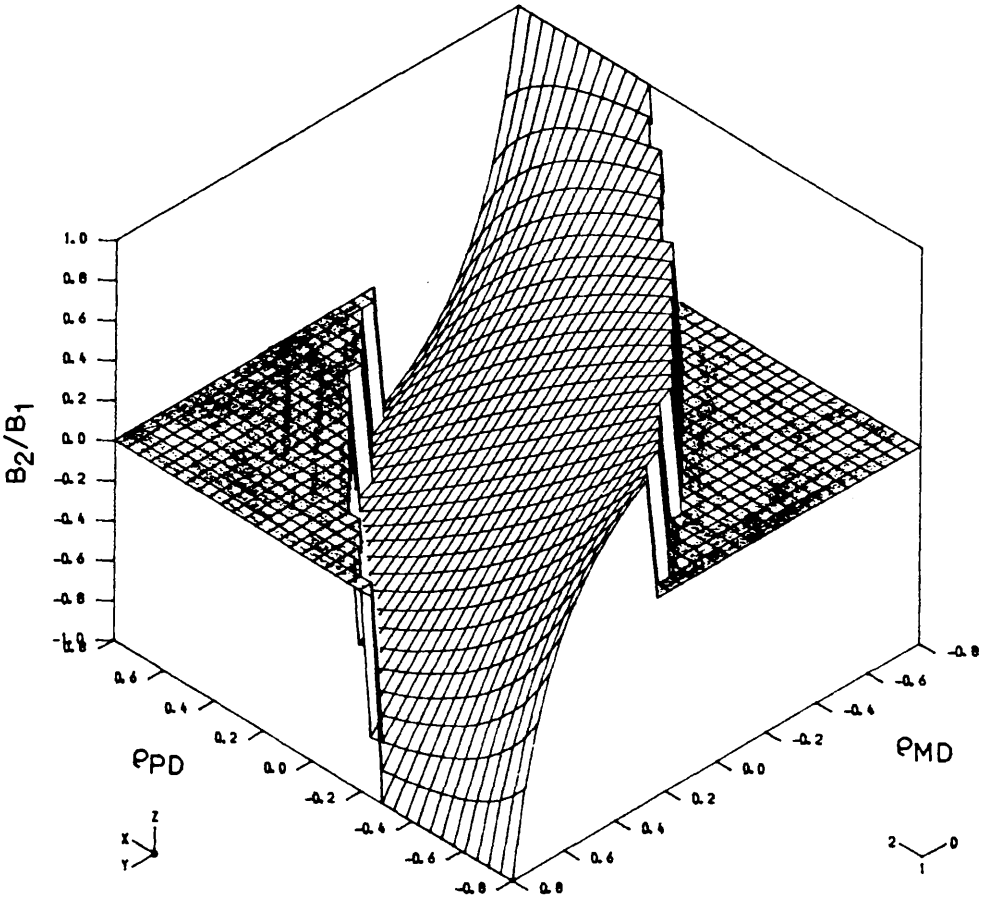

Figure (5.1)

Ratio, B_2/B_1, of the coefficients, B_1 and B_2, of the observables **P** and **D** respectively in the unbiased minimum-risk solution as a function of the correlation coeeficients ρ_{MD} and ρ_{PD} and assuming $\rho_{MP} = -0.8$ and $\sigma_D = \sigma_P$.

find that when $\rho_{MD} = -0.5$, $B_2/B_1 \simeq 0.18$, so that the coeeficient of **P** makes the dominant contribution to ω_T. By contrast, when $\rho_{MD} = -0.7$ then B_2/B_1 increases to 0.67. Moreover, it follows from our earlier remarks that when $\rho_{MD} = -0.8 = \rho_{MP}$, then $B_1 = B_2$; i.e. the coefficients make an equal contribution to ω_T. Indeed, we find that if ρ_{MD} is *greater* in modulus than ρ_{MP} then in general $|B_2/B_1| > 1$. Similar trends in the behaviour of $|B_2/B_1|$ as a function of ρ_{MD} are observed for other values of ρ_{MP} and ρ_{PD}.

Consider now the risk, R_3, of ω_T. Substituting the optimal values given by equations (5.22) into equation (5.18) we find that R_3 reduces to the following expression:-

$$R_3 = 0.04\sigma_M^2 \left[\frac{(1-\rho_{PD}^2)}{(\rho_{MP}^2 - 2\rho_{MP}\rho_{MD}\rho_{PD} + \rho_{MD}^2)} - 1 \right] \qquad (5.24)$$

Since the risk of any estimator is, by definition, positive, the above equation makes sense only if $\rho_{MP}^2 - 2\rho_{MP}\rho_{MD}\rho_{PD} + \rho_{MD}^2 < 1 - \rho_{PD}^2$; however, this is precisely the inequality which we obtain as the condition for the covariance matrix, **V**, of the trivariate distribution to be positive definite (c.f. Graybill, 1961). It follows, therefore, that the above expression for the risk is well-defined for all physically meaningful values of the correlation coefficients. Note that R_3 does not depend explicitly on σ_P or σ_D.

We can compare R_3 with the risk, R_2, of ω_S, viz:-

$$R_2 = 0.04\sigma_M^2\left[\frac{1}{\rho_{MP}^2} - 1\right] \qquad (5.25)$$

The difference between R_2 and R_3 may be written as:-

$$R_2 - R_3 = 0.04\sigma_M^2 \frac{(\rho_{PD} - \rho_{MD}/\rho_{MP})^2}{\rho_{MP}^2 - 2\rho_{MP}\rho_{MD}\rho_{PD} + \rho_{MD}^2} \qquad (5.26)$$

which is always greater than or equal to zero. In other words this confirms that the risk of the optimal trivariate estimator is always less than or equal to the risk of the 'Schechter' estimator formed from **m** and **P** alone.

The precise factor by which the addition of a third observable, **D**, reduces the 'Schechter' risk will depend on the values of the correlation coefficients. We have already noted, for example, that when **D** is uncorrelated with both **m** and **P** then $\mathbf{\Omega}_T$ offers no improvement over $\mathbf{\Omega}_S$. As a further illustration, figure (5.2) shows an isometric surface plot of the ratio, R_3/R_2, as a function of ρ_{MD} and ρ_{PD}, and again assuming that $\rho_{MP} = -0.8$. Note that the locus of points for which $R_3/R_2 = 1$, when projected down onto the ρ_{PD}-ρ_{MD} plane, gives the 'Schechter' line $\rho_{MD} = \rho_{MP}\rho_{PD}$, in agreement with our previous result.

In figure (5.2) the shaded region of the ρ_{PD}-ρ_{MD} plane marks the exterior of the region, $\mathbf{C}$, within which the correlation coefficients are constrained to lie in order that the covariance matrix, **V**, be positive definite. As we referred to previously, the condition which defines this region reduces to the following inequality:-

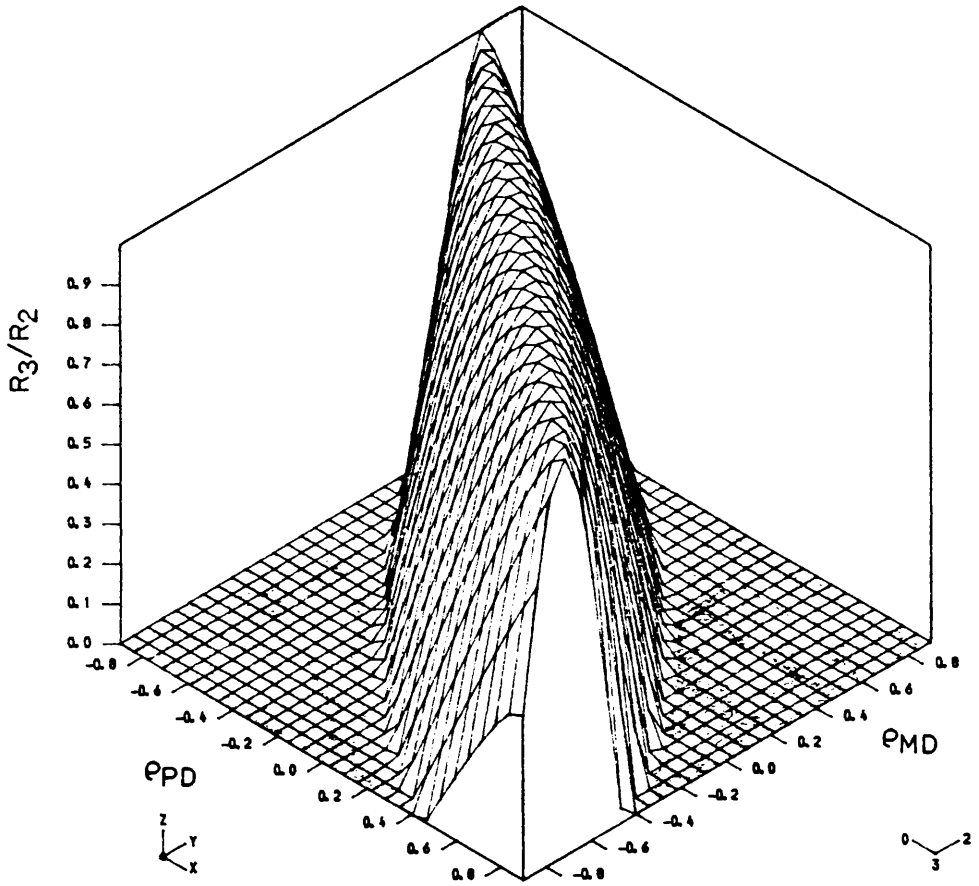

Figure (5.2)

Ratio, R_3/R_2, of the risk of the optimal trivariate estimator, ω_T, and the bivariate 'Schechter' estimator, ω_S, as a function of the correlation coefficients ρ_{MD} and ρ_{PD}, and for $\rho_{MP} = -0.8$.

$$\underline{x} = (\rho_{PD}, \rho_{MD}) \subset \mathbb{C} \iff$$

$$\varepsilon = 1 - (\rho_{MP}^2 + \rho_{MD}^2 + \rho_{PD}^2) + 2\rho_{MP}\rho_{MD}\rho_{PD} > 0 \qquad (5.27)$$

The boundary, $\partial\mathbb{C}$, of this region is defined by $\varepsilon = 0$.

We can see from figure (5.2) that as $\underline{x} \to \partial\mathbb{C}$, then the ratio, R_3/R_2, falls off sharply to zero; i.e. the risk of Ω_T tends to zero as we approach the boundary curve. Now, of course, such a zero-risk estimator could never be defined in pracice since, for physically meaningful values of the correlation coefficients, ε must be strictly positive. Nevertheless the fact that $R_3 \to 0$ as we approach $\partial\mathbb{C}$ means that if we can identify a third observable, **D**, which is correlated with **M** and **P** such that ε is very *close* to zero, (so that $\underline{x}$ will, therefore, lie close to $\partial\mathbb{C}$) then the risk of Ω_T may be considerably smaller than that of Ω_S.

We can develop this idea more quantitatively - and, in particular, clarify precisely what we mean by 'close' and 'considerably smaller' in the above remarks. Figures (5.3) to (5.6) show contour plots of the ratio, R_3/R_2, as a function of ρ_{MD} and ρ_{PD} and for different, fixed, values of ρ_{MP}. The contours of R_3/R_2 are, in fact, nested ellipses in the ρ_{PD}-ρ_{MD} plane: the zero-level contour is the boundary curve, $\partial\mathbb{C}$, and the major axis of this ellipse lies along the straight line $\rho_{MD} = -\rho_{PD}$. At higher values of R_3/R_2 the contour ellipses become increasingly eccentric and are progressively rotated anti-clockwise. The 1.0 contour (i.e. when $R_3 = R_2$) is the degenerate ellipse given by the 'Schechter' line $\rho_{MD} = \rho_{MP}\rho_{PD}$. To keep the plots as uncluttered as possible we show only four contours: at the 0.0, 0.1,

0.5 and 1.0 levels.

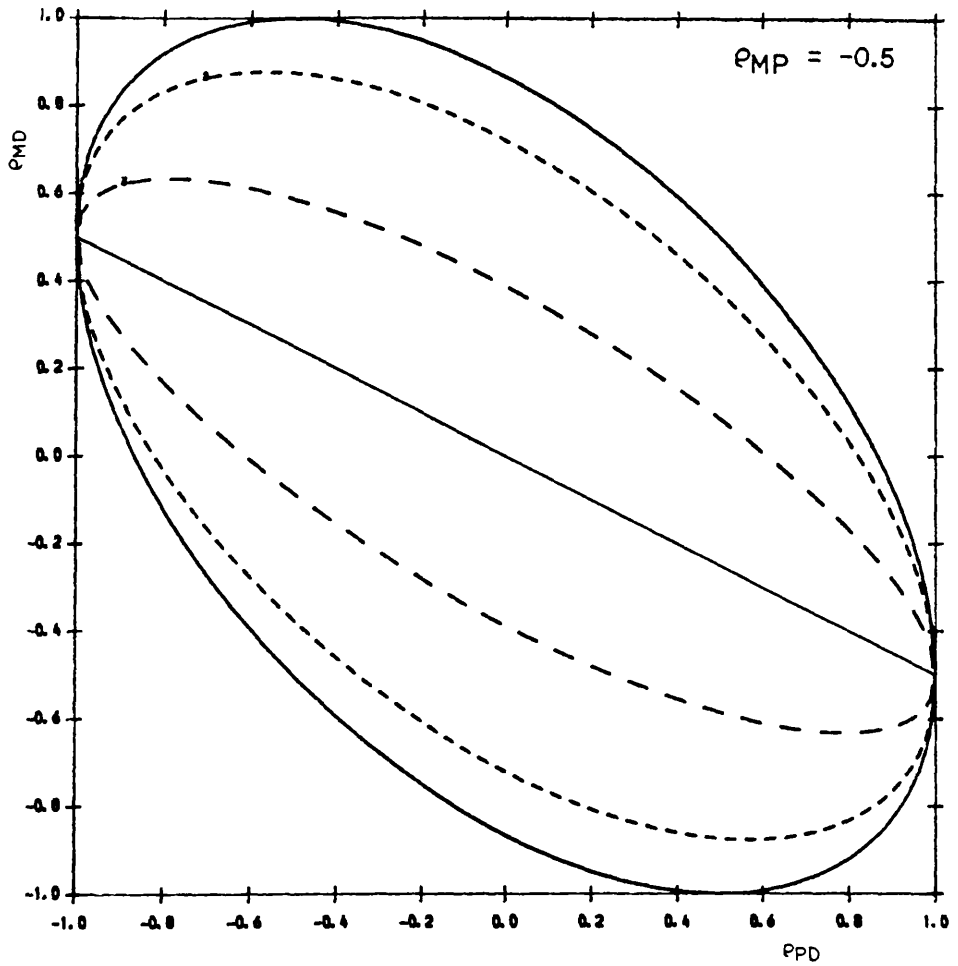

Figure (5.3)

Contour plot of R_3/R_2 as a function of the correlation coefficients ρ_{MD} and ρ_{PD}, and for $\rho_{MP} = -0.5$

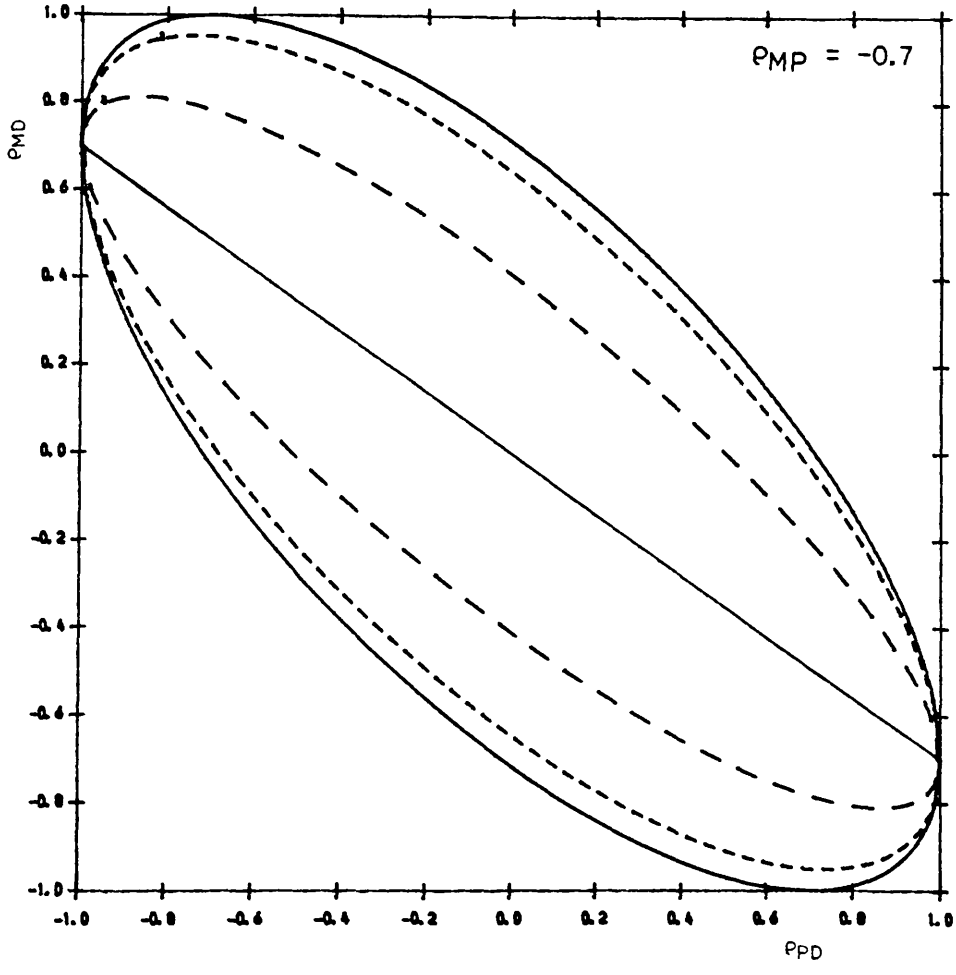

Figure (5.4)

Contour plot of R_3/R_2 as a function of the correlation coefficients ρ_{MD} and ρ_{PD}, and for $\rho_{MP} = -0.7$

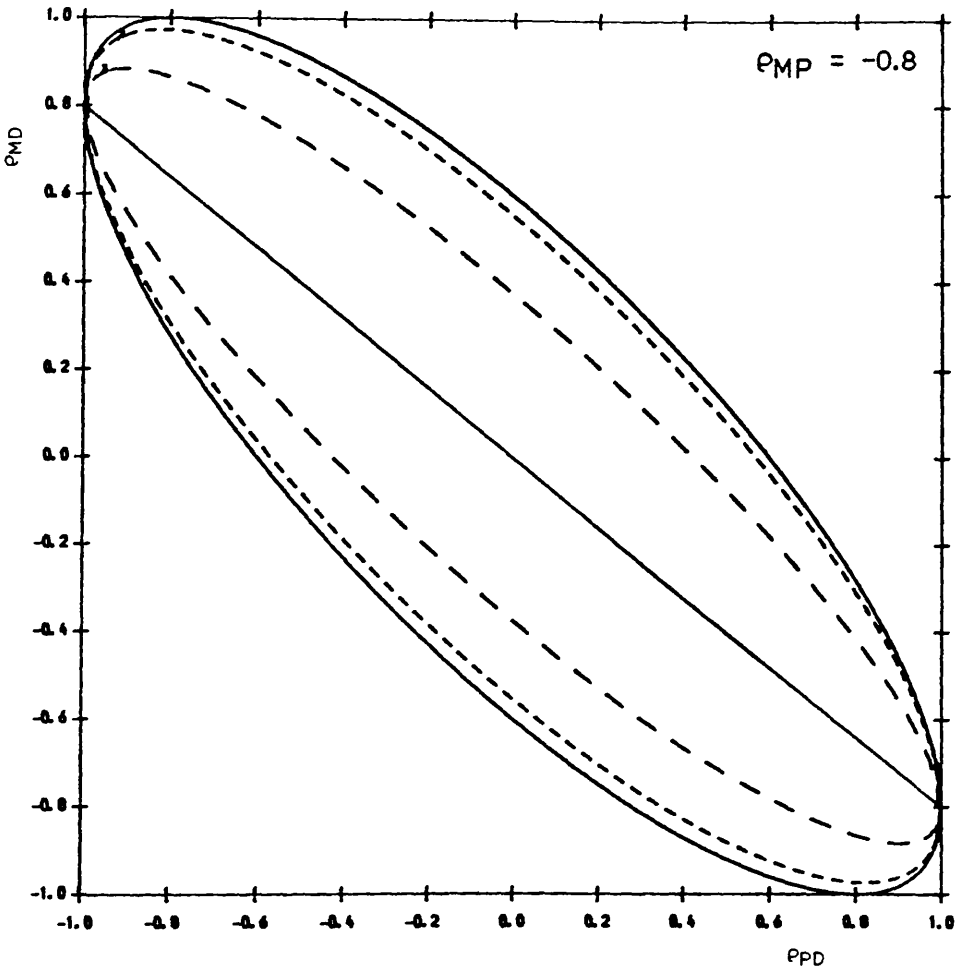

Figure (5.5)

Contour plot of R_3/R_2 as a function of the correlation coefficients ρ_{MD} and ρ_{PD}, and for ρ_{MP} = -0.8

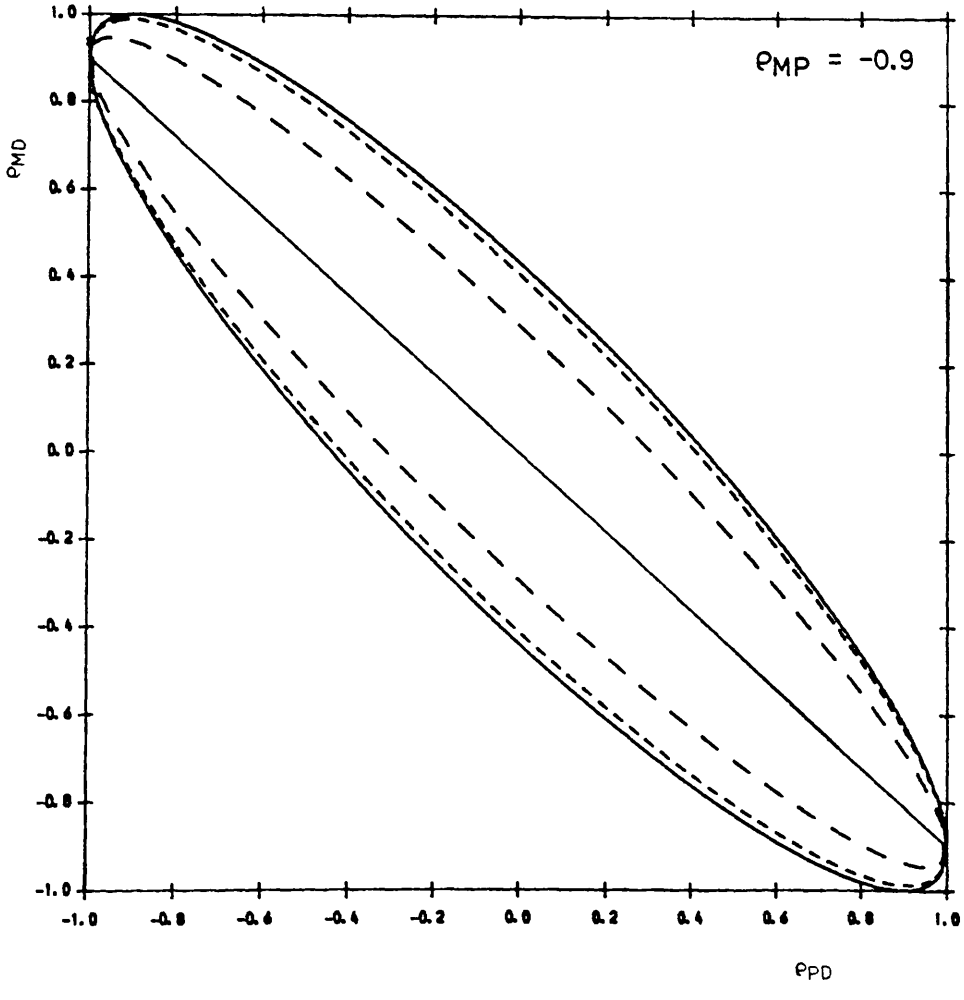

Figure (5.6)

Contour plot of R_3/R_2 as a function of the correlation coefficients ρ_{MD} and ρ_{PD}, and for $\rho_{MP} = -0.9$

A number of qualitative features are clear when we compare the four contour maps. Firstly we see that as $|\rho_{MP}|$ increases then so does the eccentricity of each contour, and in particular the area of the region, $\mathbb{C}$, enclosed by the 0.0 contour decreases. In other words, the higher the correlation between **M** and **P** then the more stringent is the restriction on the allowable values of ρ_{MD} and ρ_{PD}. Consequently, as $|\rho_{MP}|$ increases, the contours become more densely packed: in particular the area exterior to the 0.1 and 0.5 contours becomes smaller (clearly the same is true for any other contour level). This means that as $|\rho_{MP}|$ increases there will be a progressively smaller area of the ρ_{PD}-ρ_{MD} plane for which the risk of Ω_T will be significantly smaller than that of Ω_S. Another way of expressing this is to note that the higher the correlation between **M** and **P** (and hence the smaller the risk of Ω_S) then the more difficult it becomes to identify another observable, **D**, for which Ω_T will represent a significant improvement over Ω_S.

Clearly we can use these contour maps to provide a direct measure of the extent to which the use of a given third observable will reduce the risk of Ω_T. If we find, for example, that **D** is correlated with **M** and **P** such that $\underline{x} = (\rho_{PD},\rho_{MD})$ lies outside of the 0.1 contour for the appropriate value of ρ_{MP}, then we can immediately conclude that the risk of Ω_T formed using **D** is at least a factor of ten smaller than that of Ω_S.

We can now consider some specific numerical examples and at the same time highlight a very interesting - and somewhat surprising - feature of the contour maps: that increasing the correlation between

the intrinsic variables does not always lead to a greater reduction in the risk of ω_T. A simple way to illustrate this is as follows.

Suppose that, for a given value of ρ_{MP}, we also fix the value of $\rho_{MD} = \rho_{MD}^*$, say. We can then determine the dependence of R_3 on the remaining correlation coefficient, ρ_{PD}, simply by drawing a horizontal straight line through ρ_{MD}^* on the appropriate contour map and observing where this line crosses each contour. (Similarly for any fixed value of $\rho_{PD} = \rho_{PD}^*$ we can examine the dependence of R_3 on ρ_{MD} by drawing a vertical line through ρ_{PD}^*.)

Suppose, therefore, that $\rho_{MP} = -0.8$, the value which we have frequently taken as appropriate for considering e.g. the Tully-Fisher relation. Figure (5.7) again shows a contour map of R_3/R_2 for this value of ρ_{MP}. Consider now a third observable, **D**, such that $\rho_{MD} = -0.6$; this is the bold line shown in figure (5.7). We see that this line crosses the 0.0 contour at $\rho_{PD} \simeq 0.0$ and $\rho_{PD} \simeq 0.95$, i.e. ρ_{PD} is constrained to lie between these two values. It is the lower point of intersection which is most interesting: from it we see that the risk of ω_T tends to zero as the correlation bewteen **P** and **D** tends to zero. Furthermore, when ρ_{PD} is close to zero then an increase in ρ_{PD} will, in fact, *increase* R_3.

More specifically, observe that the line $\rho_{MD} = -0.6$ crosses the 0.1 contour at $\rho_{PD} \simeq 0.06$, so that an increase of ρ_{PD} to this value or higher would increase R_3 to at least 10% of the 'Schechter' risk. Similarly if ρ_{PD} increases to 0.3 then $R_3 = \frac{1}{2}R_2$, and if $\rho_{PD} = 0.75$, so that (ρ_{PD},ρ_{MD}) lies on the 'Schechter' line, then $R_3 = R_2$.

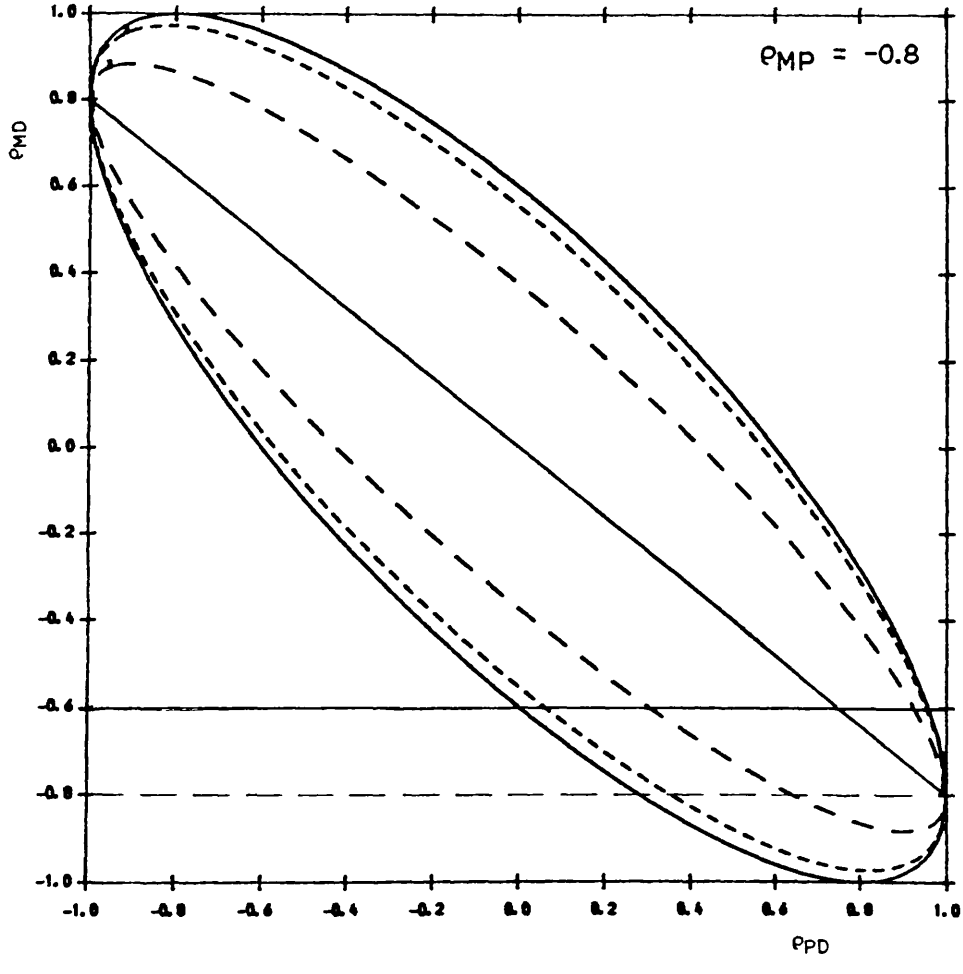

Figure (5.7)

Contour plot of R_3/R_2 as a function of the correlation coefficients ρ_{MD} and ρ_{PD}, and for $\rho_{MP} = -0.8$, displaying the dependence of R_3 on ρ_{PD}, for fixed values of ρ_{MD}.

Only for $\rho_{PD} \geq 0.75$ in this case will a further increase in ρ_{PD} now result in a reduction of R_3. In fact, we see that R_3 now falls off rapidly with increasing ρ_{PD} and tends to zero as ρ_{PD} tends to 0.95.

Similar behaviour is observed at other fixed values of ρ_{MD}. Suppose, for example, that ρ_{MD} were also equal to -0.8, which is a resonable value for the observed correlation between absolute magnitude and absolute diameter (c.f. Holmberg, 1979 and section 2.5); this case is shown as the dashed line in figure (5.7). We now find that ρ_{PD} is constrained to lie in the range (0.3,1.0). Furthermore, we see that $R_3 \rightarrow R_2$ as $\rho_{PD} \rightarrow 1$, from which it follows that R_3 increases monotonically with ρ_{PD}. Hence, in this case, using a third observable which is *less* well correlated with **P** would *always* result in a lower trivariate risk.

This second illustration is an example of the special case where $\rho_{MP} = \rho_{MD}$, for which we can easily show that R_3 simplifies to:-

$$R_3 = 0.04\sigma_M^2 \left[\frac{1 + \rho_{PD}}{2\rho_{MP}^2} - 1 \right] \qquad (5.28)$$

It is clear from this expression that in the limit as $\rho_{PD} \rightarrow 1$, $R_3 \rightarrow R_2$, while for smaller values of ρ_{PD}, R_3 will be considerably smaller than R_2, and will in fact tend to zero as $\rho_{PD} \rightarrow 2\rho_{MP}^2 - 1$.

To summarise, then, we have found that upon the addition of a third observable, **D**, the risk of ω_T thus formed will always be less than or equal to that of ω_S, as one would expect. However, the extent to which **D** reduces the risk depends strongly on the three correlation coefficients. If these take values such that the expression ε given in equation (5.27) is very small then R_3 will be significantly smaller than R_2. If, on the other hand, the correlation coefficients satisfy the relation $\rho_{MD} = \rho_{MP}\rho_{PD}$ (which, for a given value of ρ_{MP}, we refer to as the 'Schechter'line in the ρ_{PD}-ρ_{MD} plane) then we find that $R_3 = R_2$ and, in fact ω_T and ω_S are identical in this case. Thus, if the correlation coefficients lie on or near to the 'Schechter' line then the measurement of **D** adds little or nothing to the 'Schechter' estimator formed from **m** and **P** alone. Furthermore, it will frequently be the case that another observable, $\mathbf{D}^*$ say, which is *less* well correlated with **P** - so that $(\rho_{PD}*,\rho_{MD}*)$ lies further from the 'Schechter' line - would result in a considerably larger reduction in the risk of ω_T. Hence if such an observable could be identified then its use would clearly be preferable to that of **D**.

Another way of understanding the existence of the 'Schechter' line is by considering $\Psi(\mathbf{M}|\mathbf{P},\mathbf{D})$, the conditional distribution of **M** given the values of **P** and **D** for observable galaxies. This distribution is closely related to that of ω_T since, in essence, we use the measured values of **P** and **D** to infer an estimate of **M** and then combine this with the observed apparent magnitude to obtain a distance estimate.

Upon determining this distribution for arbitrary values of

the correlation coefficients, and then substituting the equation of the 'Schechter' line $\rho_{MD} = \rho_{MP}\rho_{PD}$ we find that $\Psi(\mathbf{M}|\mathbf{P},\mathbf{D})$ is identically equal to $\Psi(\mathbf{M}|\mathbf{P})$, the conditional distribution of **M** given only **P**. Hence, for this unique combintaion of the correlation coefficients $\Psi(\mathbf{M}|\mathbf{P},\mathbf{D})$ is independent of the value of **D**. This confirms that **D** will provide no more information about **M** - and hence the true distance - than is given by **P** alone, and it follows that the risk of Ω_T will equal the risk of Ω_S in this case.

It is also worth noting that when ε, as given by equation (5.27), tends to zero then so too does the variance of **M** in the conditional distribution $\Psi(\mathbf{M}|\mathbf{P},\mathbf{D})$. In this limit, in other words, **M** is determined exactly by the measured values of **P** and **D**; this is simply another way of saying that the risk of Ω_T tends to zero as $\varepsilon \to 0$.

*5.4 Case 2: Selection on **m** and **d***

We will now briefly consider a second case where the selection function is a function of both **m** and **d**, i.e. $S = S(\mathbf{m},\mathbf{d})$, but where measurements of **P** are free from selection effects. This case is appropriate for considering an extension of the D_n-σ relation to also take account of the observed apparent magnitude of a galaxy. We can, therefore, identify **d**, **D** and **P** as in section (4.5.3); i.e. log of apparent diameter, log of absolute diameter and log of central velocity dispersion respectively. In section (4.5.3) we saw that when measurements of the velocity dispersion were selection-free, it was possible to define a 'Schechter' estimator of log distance which is

unbiased and of constant risk at all true distances. The current treatment extends this analysis to the case where the galaxy sample may be magnitude-selected as well as diameter-selected, and we will determine if an unbiased estimator may still be defined in this case.

Suppose that the intrinsic joint distribution, $\Psi(\mathbf{M},\mathbf{P},\mathbf{D})$, is again a trivariate normal. In the same way as for Case 1, we will write Ψ as the product of two distributions, viz:-

$$\Psi(\mathbf{M},\mathbf{P},\mathbf{D}) = \Psi_1(\mathbf{M},\mathbf{D})\Psi_2(\mathbf{P}|\mathbf{M},\mathbf{D}) \tag{5.29}$$

i.e. the marginal distribution of **M** and **P** multiplied by the conditional distribution of **P** given **M** and **D**. Both Ψ_1 and Ψ_2 are normally distributed: Ψ_1 is the standard bivariate normal in **M** and **D**, the form of which is given by equation (4.5), while the distribution of Ψ_2 is univariate normal with mean $\mathbf{P}_*(\mathbf{M},\mathbf{D})$ and variance, $\Delta\sigma_P^2$, viz:-

$$\Psi_2(\mathbf{P}|\mathbf{M},\mathbf{D}) = \frac{1}{\sqrt{2\pi}\Delta\sigma_P} \exp(-1/2\Delta\sigma_P^2(\mathbf{P} - \mathbf{P}_*)^2) \tag{5.30}$$

$\mathbf{P}_*$ and the constant, Δ, are found to be (see Graybill, 1961):-

$$\mathbf{P}_* = P_0 + \frac{\sigma_P(\rho_{MP}-\rho_{MD}\rho_{PD})}{\sigma_M(1 - \rho_{MD}^2)}(\mathbf{M} - M_0) + \frac{\sigma_P(\rho_{PD}-\rho_{MP}\rho_{MD})}{\sigma_D(1 - \rho_{MD}^2)}(\mathbf{D} - D_0) \tag{5.31}$$

or, writing this more compactly by introducing constants κ_M and κ_D:-

$$\mathbf{P}_* = P_0 + \kappa_M(\mathbf{M} - M_0) + \kappa_D(\mathbf{D} - D_0) \tag{5.31a}$$

and

$$\Delta = \frac{1 - (\rho_{MP}^2 + \rho_{MD}^2 + \rho_{PD}^2) + 2\rho_{MP}\rho_{MD}\rho_{PD}}{1 - \rho_{MD}^2} \qquad (5.32)$$

Note that the numerator of the right hand side of this equation is precisely ς of equation (5.27). Hence it again follows, from the positive definiteness of $\mathbf{V}$, that $\Delta > 0$.

Substituting Ψ_1 and Ψ_2 into equation (5.12) and changing variables from $(\mathbf{M},\mathbf{D})$ to $(\mathbf{m},\mathbf{d})$ using equations (3.4) and (4.47), we can derive the joint distribution, $\zeta(\mathbf{m},\mathbf{P},\mathbf{d}|\mu_0)$, of the extrinsic variables $\mathbf{m}$, $\mathbf{P}$ and $\mathbf{d}$ for observable galaxies at true log distance, μ_0. We obtain:-

$$\zeta(\mathbf{m},\mathbf{P},\mathbf{d}|\mu_0) = \frac{\Psi_1(\mathbf{m}-5\mu_0-25,\mathbf{d}-\mu_0)S(\mathbf{m},\mathbf{d})\ \exp(-1/2\Delta\sigma_P^2(P - P_*)^2)}{\sqrt{2\pi\Delta}\sigma_P\ \Sigma(\mu_0)} \qquad (5.33)$$

where $\Sigma(\mu_0)$ normalises ζ at each true log distance, μ_0, and $\mathbf{P}_*$ is now re-expressed in terms of $\mathbf{m}$ and $\mathbf{d}$.

5.4.1. Bias and Risk of $\hat{\mu}_{GL}$

We define a 'general linear' estimator of μ_0 as follows (c.f. equation 4.48):-

$$\hat{\mu}_{GL} = A\mathbf{d} + B_1\mathbf{P} + B_2\mathbf{m} + C \qquad (5.34)$$

for constants A, B_1, B_2 and C. Using the distribution of equation (5.31), we can determine an expression for the bias, $B(\hat{\mu}_{GL},\mu_0)$, of $\hat{\mu}_{GL}$ for arbitrary values of the constants, viz:-

$$B(\underline{\mu}_{GL},\mu_0) = (5B_2 - A - 1)\mu_0 + (B_1\kappa_M + B_2)G_1(\mu_0) + (A + B_1\kappa_D)G_2(\mu_0) + B_1(P_0 - \kappa_M M_0 - \kappa_D D_0) + 25B_2 + C \qquad (5.35)$$

where $G_1(\mu_0)$ and $G_2(\mu_0)$ are defined as follows (c.f equation 4.44):-

$$G_1(\mu_0) = \iint \mathbf{M}\,\Psi_1(\mathbf{M},\mathbf{D})S(\mathbf{M}+5\mu_0+25,\mathbf{D}+\mu_0)d\mathbf{M}d\mathbf{D} \qquad (5.36)$$

$$G_2(\mu_0) = \iint \mathbf{D}\,\Psi_1(\mathbf{M},\mathbf{D})S(\mathbf{M}+5\mu_0+25,\mathbf{D}+\mu_0)d\mathbf{M}d\mathbf{D} \qquad (5.37)$$

Note that if there is *no* selection in **m** or in **d** then $G_1(\mu_0)$ or $G_2(\mu_0)$ respectively will be identically zero for all μ_0. Assuming for the moment that this is *not* the case, then it follows from equation (5.35) that in order to define an unbiased 'general linear' estimator, $\underline{\mu}_U$, at all μ_0 we now require to satisfy four equations in the unknowns A, B_1, B_2 and C viz:-

$$\begin{aligned} 5B_2 - A - 1 &= 0 \\ B_1\kappa_M + B_2 &= 0 \\ A + B_1\kappa_D &= 0 \\ B_1(P_0 - \kappa_M M_0 - \kappa_D D_0) + 25B_2 + C &= 0 \end{aligned} \qquad (5.38)$$

After some manipulation we find that there exists the following, unique, solution which defines $\underline{\mu}_U$:-

$$A = -\frac{\sigma_M(\rho_{PD} - \rho_{MP}\rho_{MD})}{(\sigma_M(\rho_{PD} - \rho_{MP}\rho_{MD}) - 5\sigma_D(\rho_{MP} - \rho_{MD}\rho_{PD}))}$$

$$B_1 = \frac{\sigma_M\sigma_D(1 - \rho_{MD}^2)}{\sigma_P(\sigma_M(\rho_{PD} - \rho_{MP}\rho_{MD}) - 5\sigma_D(\rho_{MP} - \rho_{MD}\rho_{PD}))} \qquad (5.39)$$

$$B_2 = -\frac{\sigma_D(\rho_{MP} - \rho_{MD}\rho_{PD})}{(\sigma_M(\rho_{PD} - \rho_{MP}\rho_{MD}) - 5\sigma_D(\rho_{MP} - \rho_{MD}\rho_{PD}))}$$

$$C = -25B_2 - B_1(P_0 - \kappa_M M_0 - \kappa_D D_0)$$

There are several interesting features about this solution. Note firstly that if **M** is uncorrelated with both **P** and **D** then we obtain for μ_U precisely the 'Schechter' estimator, μ_S, of section (4.7.3), viz:-

$$\mu_U = \sigma_D/\rho_{PD}\sigma_P\mathbf{P} - \mathbf{d} + D_0 - P_0\sigma_D/\rho_{PD}\sigma_P \qquad (5.40)$$

This not surprising since in this case any selection of apparent magnitudes has no bearing on the joint distribution of **d** and **P** for observable galaxies; hence, we would expect the results of section (4.7.3) still to apply.

If, on the other hand, **M** is correlated with **D** and **P** then μ_U will, in general, be different from the 'Schechter' estimator. Moreover, μ_S will now be *biased* at all true distances. Again, this should not be too surprising since magnitude selection would now affect the joint distribution of **d** and **P** for observable galaxies so that the form of this distribution used in section (4.7.3) to derive expressions for the bias and risk of μ_S will no longer be valid.

It is also important to note that the solution of equations (5.39) will generally demand that B_2, the coefficient of **m**, is non-zero. In other words, this means that when a sample is subject to both magnitude and diameter selection one *cannot* in general define an unbiased distance estimate using only the measured values of **d** and **P** (nor indeed using only **m** and **P** or **m** and **d**!); only by using the measured values of *all three* observables can one define an unbiased estimator, $\hat{\mu}_U$.

There are two interesting exceptions to this, however, which arise in a similar manner to the 'Schechter' line discussed for case (1). Firstly, if the correlation coefficients satisfy the equation:-

$$\rho_{MP} = \rho_{MD}\rho_{PD} \tag{5.41}$$

then it follows from equations (5.39) that B_2, the coefficient of **m**, is equal to zero and $\hat{\mu}_U$ again reduces precisely to $\hat{\mu}_S$, the 'Schechter' estimator formed from **d** and **P**.

Secondly, if the correlation coefficients satisfy:-

$$\rho_{PD} = \rho_{MP}\rho_{MD} \tag{5.42}$$

then we find that A, the coefficient of **d**, is zero. $\hat{\mu}_U$ now reduces to:-

$$\hat{\mu}_U = 0.2(\mathbf{m} - \sigma_M/\rho_{MP}\sigma_P(\mathbf{P} - P_0) - M_0 - 25) \tag{5.43}$$

On comparison with equation (4.11) we see that in this case $\hat{\mu}_U$ is essentially equal to the bivariate 'Schechter' estimator, $\hat{\omega}_S$, formed from **m** and **P**. The above expression differs from $\hat{\omega}_S$ only by a

constant, which is entirely due to the fact that we have defined μ_U in units of MPc and not the scaled distance units used for ω_S.

Thus we have shown that if the correlation coefficients satisfy either of equations (5.41) and (5.42) - which we can think of as defining two 'Schechter' lines in the ρ_{MP}-ρ_{PD} plane, in a manner similar to case (1) - then we can still define an unbiased estimator using only *two* observables: **d** and **P** or **m** and **P** respectively. Otherwise μ_U can only be defined for all μ_0 by combining the measured values of **m**, **P** and **d**.

We now calculate the risk, $R(\mu_U,\mu_0)$, of μ_U at true log distance, μ_0. We obtain the following expression:-

$$R(\mu_U,\mu_0) = B_1^2\Delta\sigma_P^2 \qquad (5.44)$$

where the constant, Δ, is defined in equation (5.32). Thus, the risk of the unbiased trivariate estimator is again a constant, R_3, independent of the true distance, just as we found for the 'Schechter' estimators of the previous chapter and the unbiased estimators of case (1).

Substituting for B_1 from equations (5.39) we find that R_3 is given by:-

$$R_3 = \frac{\sigma_M^2\sigma_D^2(1 - \rho_{MD}^2)(1 - (\rho_{MP}^2 + \rho_{MD}^2 + \rho_{PD}^2) + 2\rho_{MP}\rho_{MD}\rho_{PD})}{(\sigma_M(\rho_{PD} - \rho_{MP}\rho_{MD}) - 5\sigma_D(\rho_{MP} - \rho_{MD}\rho_{PD}))^2} \qquad (5.45)$$

Note that R_3 is a function of both σ_M and σ_D, i.e. the dispersions of both observables which are subject to selection effects. Contrast this

with the risk of the optimal trivariate estimator of case (1) - for which we considered selection only on apparent magnitudes. Correspondingly, we see from equation (5.23) that R_3 depends only on σ_M.

Note also the appearance of the expression for ε on the numerator of the above equation, from which it follows that $R_3 \to 0$ as $\varepsilon \to 0$. In practical terms, therefore, we see that - as for case (1) - if the intrinsic variables are suitably correlated so that ε is very small, then the risk of μ_U may also be small. (Note, however, the question of precisely how small ε must be in order to obtain a given value of R_3 will, of course, also depend on the size of the denominator in the above expression; if this too is very small then R_3 may still be large.)

Clearly, then, we can use equation (5.45) to compare the risk of the unbiased estimator formed from different sets of observables - so as to identify those which give the lowest risk, and hence the most reliable distance estimates.

Finally, it is worth noting that we can regard case (1) essentially as a corollary of case (2). Suppose, for example, that measurements of **d** are free from selection. It then follows that $G_2(\mu_0)$, as defined in equation (5.37), is identically zero for all μ_0. This would effectively remove the third of the constraint equations (5.38) and, hence, we would now require to satisfy only three equations in the four unknowns, A, B_1, B_2 and C, in order to identify an unbiased estimator. This is, of course, precisely the situation dealt with in case

(1). Similar remarks apply when there is no magnitude selection.

5.5 Summary of Conclusions

We now summarise the main results of this chapter and make some further comments on their significance. We have considered the set of 'general linear' estimators of log distance; i.e. estimators formed by taking a linear combination of three mutually correlated observables - denoted **m**, **p** and **d** - corresponding to the intrinsic variables **M**, **P** and **D** whose intrinsic joint distribution we assume to be trivariate normal. We have compared the properties of these estimators with those of the previous chapter - formed from a linear combination of only two observables (**m** and **p**, say), one of which we assumed to be free from selection effects so that one could define a 'Schechter' estimator which is unbiased and of constant risk at all true distances. In particular, we have considered whether by adding a third observable, **d**, to our analysis it is still possible to define 'Schechter'-type estimators of log distance.

In all cases we find that the answer is *yes*! Two important points emerge, however. If the third observable, **d**, is subject to selection effects, then the unbiased estimator can, in general, only be defined by using *all three* observables, **m**, **p** and **d**. Moreover, this means that the 'Schechter' estimator defined in terms of **m** and **p** alone will no longer be unbiased, since the selection on **d** affects the joint distribution of **m** and **p**. Two exceptions to this, however, are the 'Schechter' lines $\rho_{MP} = \rho_{PD}\rho_{MD}$ and $\rho_{PD} = \rho_{MP}\rho_{MD}$: if the correlation

coefficients satisfy either of these equations then one *can* still define an unbiased estimator using only **d** and **P** or **m** and **P** respectively, and in each case the general unbiased trivariate estimator reduces exactly to the appropriate bivariate 'Schechter' estimator.

If, on the other hand, **d** is free from selection effects then one can still define an unbiased estimator using only **m** and **p** alone. In other words the 'Schechter' result of chapter (4) will still be valid in this case. By using all three observables, however, we have sufficient freedom to define an unbiased trivariate estimator, ω_T, of minimum risk - and we have shown that the risk, R_3, of this optimal estimator is always less than or equal to the risk, R_2, of the 'Schechter' estimator formed from **m** and **p** alone. There exists another 'Schechter' line $\rho_{MD} = \rho_{PD}\rho_{MP}$ for this case, however: if the correlation coefficients satisfy this equation then ω_T reduces precisely to ω_S, so that using the measured value of **d** does not improve the 'Schechter' estimate of log distance.

These results will clearly have an important bearing on the use of, e.g., the D_n-σ relation. Assuming a trivariate normal for the joint intrinsic distribution of absolute magnitude, absolute diameter and velocity dispersion, then it certainly follows that the 'Schechter' estimator defined from the D_n-σ relation (i.e. using only the measured apparent diameter and velocity dispersion of a galaxy - as discussed in section 4.7.3) will still be unbiased provided that one's sample of galaxies is complete in apparent magnitude. Of course in this case it would also follow that by using the observed apparent magnitude in addition to the diameter and velocity dispersion, one could define an

optimal unbiased estimator of *lower* risk than the bivariate 'Schechter' estimator; indeed the risk of the trivariate estimator may be considerably smaller for suitably correlated observables, and we have demonstrated how one may easily determine when this is the case.

In short, then, utilising the measurements of a third observable can certainly offer a means of significantly reducing the risk of unbiased estimators, and thus obtaining more reliable distance estimates. When such an observable is available, therefore, its use would seem to be strongly advised.

It seems clear that one may extend this analysis further to consider distance estimates obtained from four or more observables, with results very much analogous to those which we have obtained in this chapter. We will not attempt such an extension here.

6. CONCLUSIONS AND FUTURE WORK

6.1 Qualitative Overview of Main Results

In this thesis we have developed a statistical framework within which to assess rigorously the properties of different distance estimators by computing their distribution, bias and risk as a function of true distance, after accounting for luminosity selection effects. We have applied this formulation firstly to a number of different estimators which are a function only of apparent magnitude, assuming a gaussian luminosity function and a Heaviside selection function. This simple case illustrates a fundamental problem in removing or reducing selection effects: the question of which estimator is 'best' has, in general, no clear cut answer since both the bias and risk are complicated non-linear functions of the true distance of a galaxy - which is, of course, unknown! The best estimator (in the usual statistical sense of minimum bias or minimum risk or some combination of both) if the galaxy is very remote may be a poor choice if the galaxy is, in fact, nearby.

We have next analysed the properties of distance estimators derived from combining measurements of two observables. These results are relevant to understanding the effects of bias on, e.g., the Tully-Fisher and D_n-σ relations. We have shown that the different linear regressions used in the literature to calibrate such relations each correspond to estimators of log distance defined as different linear combinations of the two observables. Modelling the joint distribution of the intrinsic variables by a bivariate normal, we have

determined the distribution, bias and risk of a linear estimator for general coefficients - into which expressions the values corresponding to each regression line may be substituted - and also of a maximum likelihood estimator of log distance. We find that the estimators have widely different properties. In particular, if there is no selection in (for the Tully-Fisher case) line widths then the estimator corresponding to a regression of line widths on magnitudes is unbiased and has constant risk for all true distances; this confirms the unbiasedness of this regression line, as claimed by Schechter (1980). The 'Tully-Fisher' estimator which corresponds to a regression of magnitudes on line widths, on the other hand, is increasingly biased at large distances, but has a smaller risk than the 'Schechter' estimator for true distances less than, typically, several hundred Mpc, although the risk of the 'Tully-Fisher' estimator increases sharply at greater true distances. The 'Schechter' risk will, in fact, be very large if the magnitude and line width are poorly correlated and becomes infinite for $\rho = 0$. This is consistent with the pathological example which we considered in section (2.4), for which we saw that when line width and magnitude are uncorrelated, the measured line width yields no information about the magnitude of a galaxy, so that the 'Schechter' regression line cannot be used to infer the galaxy distance. The maximum likelihood estimator is found to have very small (though non-zero) bias and smaller risk than the 'Schechter' estimator over a larger range of true distances, although again this result depends on ρ.

These results indicate that some care must be taken before choosing the most appropriate estimator: for example, although the

'Schechter' estimator has the desirable property of unbiasedness for any value of ρ, its large risk for ρ small means that *for any single observation* there will be a higher probability of a large systematic error than with e.g. the 'maximum likelihood' estimator, so that the latter may be more appropriate for a small sample of galaxies. Provided the observables are fairly well correlated, however, (typically $|\rho| > 0.6$) we recommend the use of the 'Schechter' estimator, whenever it may be defined. Our strongest reason for this is the fact that, regardless of the form of the luminosity selection effects, the 'Schechter' estimator is normally distributed, with mean equal to $\log x_0$ and constant variance, at all true distances, x_0. The 'Schechter' estimator is unique in this regard: for no other linear combination of the observables, nor indeed for the maximum likelihood estimator, is the shape of the estimator distribution preserved at all true distances. Not only does the zero bias and constant risk follow immediately from this but consequently, as we have shown, confidence intervals constructed from the 'Schechter' estimator are of constant width. Moreover, because the 'Schechter' estimator is normally distributed, it follows that the joint distribution of the distance estimates obtained for a number of galaxies at different distances will be a multivariate normal. This allows the statistical properties of a large sample of galaxies to be derived very easily.

We have also considered estimators formed from a linear combination of three observables, which may or may not be distance-dependent, and have addressed the question of whether unbiased 'Schechter' estimators may still be defined in this case. Modelling the joint distribution of the intrinsic variables by a

trivariate normal distribution we have determined the bias and risk of a 'general linear' estimator as a function of true distance for two separate cases; in the first instance where two of the three observables are free from selection effects and secondly where only one of the observables is selection-free. We have found that in both cases the definition of an unbiased estimator is indeed still possible. Two important points emerge, however: firstly, in the former case of two selection-free observables the unbiased estimator is no longer unique. We have, therefore, obtained expressions for the unbiased estimator of minimum risk. We have found that, as one might expect, the risk of this optimal trivariate estimator is always less than or equal to that of the 'Schechter' estimator formed from only two observables; it is interesting to note, however, that there are situations, for particular values of the correlation coefficients, when the addition of a third observable does not improve the bivariate 'Schechter' estimate of log distance. For the latter case of only one selection-free observable, we have found that an unbiased estimator can in general only be defined by using all three observables; estimators formed from combinations of only one or two will be biased. There are again exceptions to this, however, for particular values of the correlation coefficients where the trivariate estimator again reduces precisely to the unbiased bivariate 'Schechter' estimator.

6.2 Applications of Multivariate Estimators

The main conclusion which we reach from our analysis of trivariate distance estimators is that, where additional observables are

available for indicators such as the Tully-Fisher or D_n-σ relations, then there use in defining unbiased distance estimators would seem to be well-advised. A number of potentially useful observables exist for such relations. As we have already remarked in chapter (5), one can use the apparent magnitude in addition to the angular diameter and velocity dispersion of ellipticals to extend the D_n-σ relation to three observables. (Indeed, we should remark that if our galaxy sample is diameter limited *and* magnitude limited then the inclusion of magnitudes will, in general, be *essential* in order to define an unbiased estimator.) Dressler *et al* (1987) also consider a line strength indicator, first introduced in Terlevich *et al* (1981), which shows a weak but signifcant correlation with D_n. Clearly this observable could also be included in deriving our unbiased estimator; it would be interesting in the future to attempt this extension to four observables and apply the results to a comparison with the distance estimates obtained by previous authors. Similarly for spiral galaxies we could consider the combination of, e.g., apparent diameter and/or colour with apparent magnitude and line width to explore an extension of the Tully-Fisher indicator; we have seen in chapter (2) that correlations with luminosity have been measured for both of these observables.

Another interesting application of our analysis of trivariate estimators would be to the calibration of the period-luminosity-colour relation which is well-established for Cepheid variables (c.f. Sandage, 1958). This relation takes the form:-

$$\log P = aM_V + b(B-V) + c \qquad (6.1)$$

where P is the period, M_V is the absolute photographic magnitude, B-V

denotes the B-V colour and a, b and c are constants.

Clearly equation (6.1) is equivalent to a trivariate linear estimator of log distance which is a function of P, B-V and apparent magnitude, m. It would be particularly interesting to apply our analysis to this distance indicator since, as we have remarked, it is probably the most securely calibrated - from both theoretical considerations and detailed observations in our galaxy and the Large Magellanic Cloud - so that zero-point errors are less significant (Martin *et al*, 1979). As the Hubble Telescope expands by a factor of four or five the limiting distance to which Cepheid observations will be possible, it will certainly become more important to ensure that the distances inferred are not adversely affected by selection bias. We would hope to study the question of obtaining optimal distance estimates from the Cepheid relation in future work.

6.3 Other Methods for Reducing Bias and Risk

One important area for future study is to consider those situations where the Schechter scheme cannot be successfully applied; i.e. when one does not have available one or more observables the measurements of which are free from selection effects. As we have commented previously, this would be the case if, for example, galaxies are selected for the Tully-Fisher relation by line width as well as by magnitude. A more straightforward example, however, is simply that of estimators which depend on apparent magnitude alone - of the type studied in chapter (3). In that chapter we saw that it was possible, in

general, to define such an estimator which is unbiased for all true distances. Nevertheless, we have seen that our statistical formulation allows us, for any *given* selection function, to compute and bias and risk of any estimator as a function of true distance. This suggests a possible scheme for reducing the bias and risk.

Suppose we have some estimator, $\hat{x}$, of the true distance, x_0, of a galaxy and suppose that we have computed the bias, $B(\hat{x},x_0)$, of $\hat{x}$ for any x_0. We note that $B(\hat{x},x_0)$ may be given as:-

$$E(\hat{x}|x_0) = x_0 + B(\hat{x},x_0) \qquad (6.2)$$

(c.f. equation 3.24)

We can now define a new estimator, $\hat{x}_1$, given by:-

$$\hat{x}_1 = \hat{x} - B(\hat{x},x_0=\hat{x}) \qquad (6.3)$$

i.e. from each value of our first estimator, $\hat{x}$, we subtract the bias of $\hat{x}$ not at the true (but unknown) distance, x_0, but at the *estimated* distance, $\hat{x}$; in other words we assume this first estimate to be equal to the actual distance of the galaxy.

We can now compute the bias and risk of $\hat{x}_1$ as a function of x_0 and one would hope that this bias would be less than that of $\hat{x}$. Moreover, this process may be repeated iteratively, defining next $\hat{x}_2 = \hat{x}_1 - B(\hat{x}_1,x_0=\hat{x}_1)$ and so on, although in general there is no guarantee that the iteration will converge.

In Appendix (1) we decribe this iteration scheme in a little

more detail and demonstrate its application to the naive estimator, Φ_N of log distance (clearly $\hat{x}$ above can also be a *function* of distance) as defined by equation (4.27). The results are very interesting: we find that the iteration procedure does indeed significantly reduce the bias and risk of Φ_N at large true distances; this is, however, achieved only at the expense of a severe *increase* in both the bias and risk of the iterated estimators for large x_0. This means, therefore, that the range of true distances in which the iteration scheme is most effective is strongly dependent on the number of iterations which we perform.

The failure of our scheme to converge to an unbiased estimator for *all* true distances would seem to us to be due largely to the severe non-linear nature of the distribution of Φ_N at large true distances, caused by the magnitude selection effects. It would be instructive, therefore, to investigate whether the scheme is more successful when the selection effects are less severe. (Indeed we already have a partial answer to this question in that we see in appendix (1) that the convergence is more extensive for smaller values of σ_M, as one would expect.) Since it is straightforward to extend the scheme to deal with estimators which are functions of several observables, it would be interesting to apply it to the general linear estimators studied in chapters (4) and (5), but in the case where *no* observable is selection-free so that unbiased Schechter estimators cannot be defined. One would hope, for example, that the addition of a second (albeit incompletely sampled) observable, **P**, would reduce the non-linearity sufficiently to make the iteration scheme converge to an unbiased estimator over a wider range of true distances. As a first step we have already applied the scheme to the 'Tully-Fisher'

estimator, ω_{TF}, in the case where P *is* selection-free. It is encouraging to find that ω_{TF} does indeed appear to converge to the Schechter estimator, ω_S (which, of course, *is* unbiased for all x_0 in this case.) A useful future exercise, then, would be to determine how robust this convergence is to selection on P; i.e. how completely sampled must P be in order that one may still obtain an approximately unbiased estimator after several iterations.

We can envisage another simple strategy by which to reduce the bias and risk of an estimator within a given range of true distances. Again this is most easily illustrated using apparent magnitude-based estimators. Suppose we define the following 'general linear' estimator of $\log x_0$, analogous to those of chapters (4) and (5):-

$$\omega_{GL} = 0.2(\blacksquare - m_L) + C \qquad (6.4)$$

In other words ω_{GL} is just the naive estimator, ω_N, plus some constant correction, C. This is equivalent to the general linear distance estimator defined in equation (3.8) from which the 'Malmquist' and 'Proximal' estimators were derived.

For any given luminosity and selection function we can derive expressions for the bias and risk of ω_{GL} as a function of true log distance, and for an arbitrary value of C. We may then ask the question what value of C should be chosen in order that ω_{GL} be a reliable estimator within some desired range of true distances.

Our idea is to choose the value of C which minimises the bias or risk, or some appropriate combination of both, over that true

distance range. Suppose, for example, we adopt minimum risk as our criterion and determine the optimal value of C for the distance interval (x_1,x_2). If we define:-

$$F(C) = \int_{x_1}^{x_2} R(\Omega_{GL},\omega_0)dx_0 \qquad (6.5)$$

Then our optimal value of C satisfies $\partial F/\partial C = 0$.

Consider the familiar case of a gaussian luminosity function and a Heaviside selection function at magnitude limit, m_L. For this case we find that C is given by:-

$$C = \frac{1}{x_2-x_1}\int_{x_1}^{x_2} \frac{0.2\sigma_M \exp(-\tfrac{1}{2}(5\omega_0/\sigma_M)^2)}{\Phi(-5\omega_0/\sigma_M)}\,dx_0 \qquad (6.6)$$

Figures (6.1) and (6.2) show the bias and risk of Ω_{GL} with values of C determined from this equation chosen to minimise the risk within the true distance intervals (1.0,1.5), (1.5,2.0) and (2.0,2.5) respectively, and for $\sigma_M = 1.0$.

We can see from figure (6.1) that the bias of each Ω_{GL} passes through zero at approximately the midpoint of the appropriate interval. The *slope* of the bias curves does not change with different C, however, as must be the case since the estimators differ only by a constant. This results in each estimator being positively biased at small true distances: the more remote the interval in which Ω_{GL} is optimised, then the more positively biased is Ω_{GL} for small x_0.

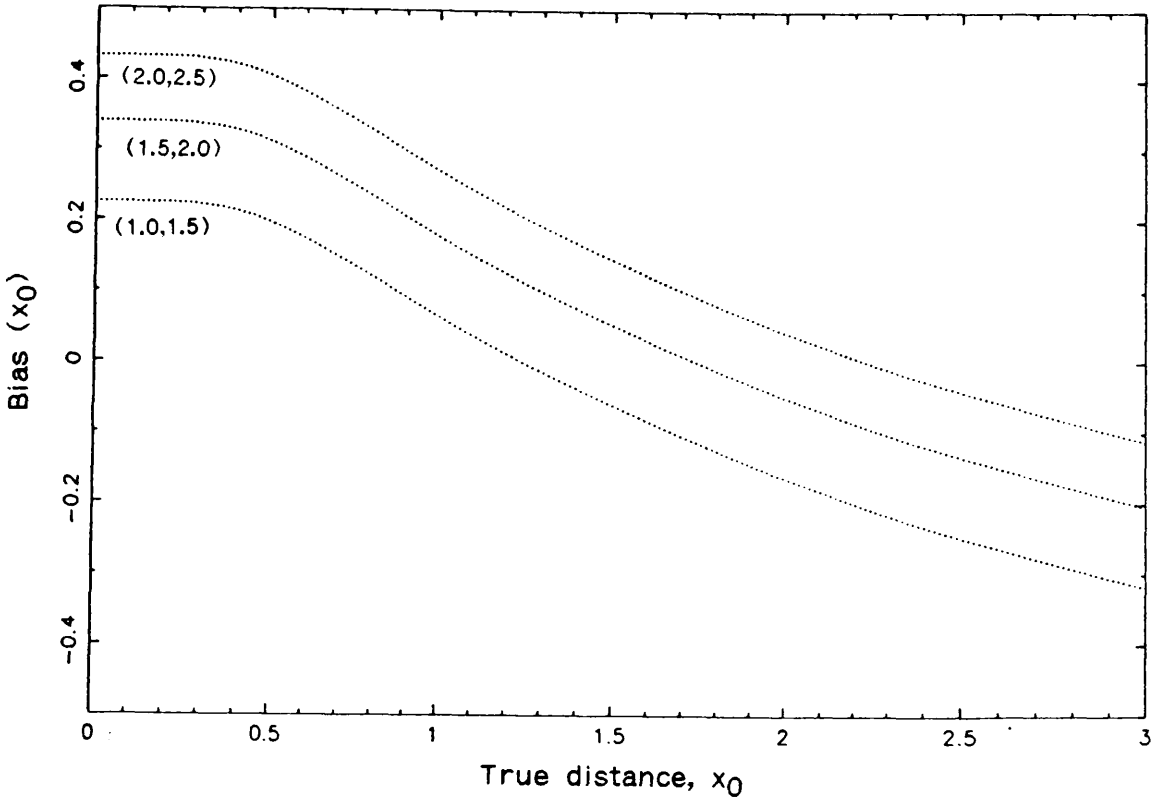

Figure (6.1)

Bias of optimal linear estimators of log distance, defined by minimising the risk integrated in the true distance range (1.0,1.5), (1.5,2.0), (2.0,2.5) respectively (σ = 1)

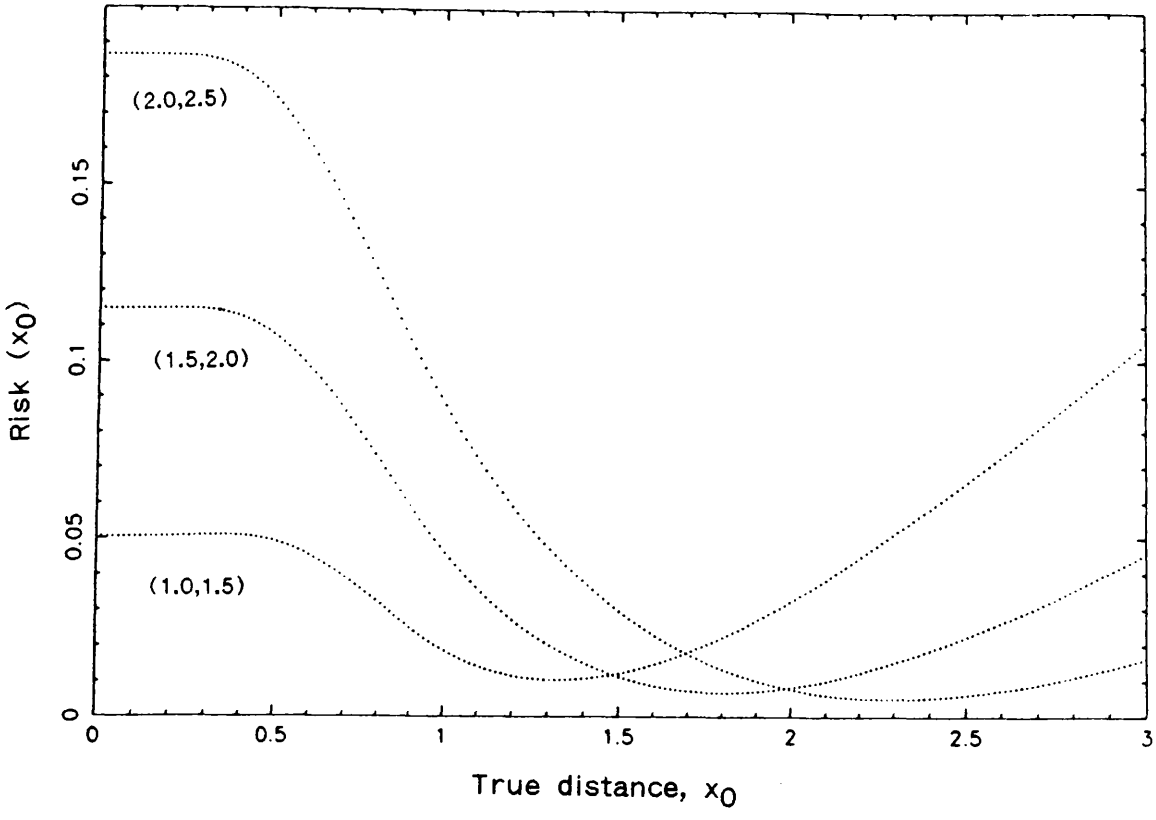

Figure (6.2)

Risk of optimal linear estimators of log distance, defined by minimising the risk integrated in the true distance range (1.0,1.5), (1.5,2.0), (2.0,2.5) respectively (σ = 1)

Moreover, another consequence of the constant slope is that there is a non-negligible positive and negative bias at each end of the chosen distance interval. Clearly if we take progressively wider intervals in which to optimise the estimator then this effect will be steadily magnified, although the bias will still change sign close to the midpoint of the interval. Compare this with our iteration scheme which tends to flatten the slope of the bias curves and thus yields a large range of true distances for which the bias is close to zero.

The results for the risk are in some respects more encouraging - which is perhaps not surprising since it was with respect to risk that we optimised ω_{GL}. We see that in each of the three cases the risk is significantly reduced over the full range of the relevant distace interval - although again this is at the expense of a prohibitive increase outside of that interval, just as was found with our iteration scheme.

It would be interesting to explore this optimisation method in more detail to study, for example, how wide the optimisation range may be in order to achieve a significant risk reduction over the whole interval. Furthermore it would also be very instructive to apply this technique to multi-observable estimators. As for the iteration scheme, one could reasonably expect that the reduction of bias and risk would be more effective in this case, particularly since one would then be able to optimise with respect to more than one constant coefficient of the estimator.

One should remark that provided one is confident that an

observed galaxy *does* lie within, or close to, the optimised distance range then it is clear from figure (6.2) that Ω_{GL} would indeed be a good choice of distance estimator, since its increased risk outside this range would then be largely irrelevant. Expressing this more precisely, our point is that in defining an optimal distance estimator one should ideally take account of the *true* relative number density of galaxies as a function of distance in order that the estimator be most effective at distance where galaxies are most likely to be observed.

Suppose that $n(x_0)$ denotes the relative number density of galaxies at true distance, x_0. (Note that we can still define $n(x_0)$ for an anisotropic galaxy distribution by simply averaging over all directions - or over some solid angle of interest - at each true distance.) Incorporating $n(x_0)$ as a weighting factor in equation (6.5), viz:-

$$F^*(C) = \int_{x_1}^{x_2} n(x_0) R(\Omega_{GL}, \omega_0) dx_0 \qquad (6.7)$$

we can define a modified optimal value of C by solving $\partial F^*/\partial C = 0$.

Now of course the basic problem which we face with this is the same difficulty as we have faced in earlier chapters: the fact that the true galaxy distance distribution, $n(x_0)$, is unknown. It seems to us, however, that it may be possible to overcome this problem by applying the following scheme. Suppose one adopts some initial choice (prior distribution) for $n(x_0)$ - e.g. one might consider as a neutral choice the case of $n(x_0)$ = constant. With this distribution one determines from equation (6.7) an optimal distance estimator which one

can then use to infer the distances of a sample of a galaxies. These distance estimates can next be used to form a new and improved estimate of $n(x_0)$, which may then be plugged into equation (6.7) to obtain a new optimal distance estimator, and so on.

This scheme is closely based on techniques used in Bayesian analysis (c.f. Mood and Graybill, 1974) and indeed is very much in spirit of not only our previous iteration scheme but also the IRAS reconstruction algorithm discussed in chapter (1). Although its convergence properties are not obvious, it seems to us that this scheme could offer a very useful means of improving distance estimators in a manner which is, in some sense, *consistent* with the true galaxy distribution - we mean by this that if the scheme converges then the estimator thus constructed would be 'best' precisely when $n(x_0)$ is equal to the true galaxy distribution.

It would again be particularly interesting to study this procedure applied to estimators derived from two or more observables. Moreover, the introduction of additional observables might allow one to define iteratively optimal estimators which are consistent not only with the galaxy number density distribution, but also with the luminosity function or selection function. If this were possible then it would provide an important method for verifying our assumed analytic form for these functions - and, more fundamentally, for testing the validity of assuming a universal luminosity function, independent of position, for galaxies of the same morphological type. These ideas are still rather speculative at this stage, but in our opinion certainly deserve further detailed study.

6.4 Applications to Velocity Field Analysis

The results which we have obtained concerning the removal of Malmquist bias by the definition, where possible, of Schechter distance estimators will be of considerable value in studying the properties of the velocity and density field - either via simple, classical, treatments such as the Hubble diagram, or by sophisticated reconstruction techniques such as POTENT. The POTENT error analysis presented in DBF models log distance errors which are normally distributed with constant variance, σ^2 - precisely as we have shown to be the case for the 'Schechter' estimators. Using a mixture of analytic and Monte-Carlo methods, the authors find that the bias and variance of their estimated radial peculiar velocity - interpolated and smoothed after accounting for distance measurement errors and sampling biases - is, to first order, proportional to σ^2 (See DBF, equations A24 and A25). It follows, therefore, that methods which reduce the risk of our distance estimators would improve the estimate of the smoothed peculiar velocity field which is used by POTENT.

It would be useful, nevertheless, to extend this analysis to consider a general distance estimator, $\hat{\mu}$ say, of arbitrary distribution, and determine the precise relationship between the distribution of $\hat{\mu}$ as a function of true distance and the distribution of errors in the density and velocity fields finally recovered by POTENT. In principle one could adopt an approach very similar to our analysis of distance estimators: in other words set out to derive the distribution, bias and

risk of the POTENT estimates of the recovered velocity and density fields as a function of position, given their *true* values. Such a general treatment seems rather ambitious, however; a more reasonable first line of attack might be to consider a specific velocity field model - expanded in terms of spherical harmonics, for example - and determine how the distribution, bias and risk of our distance estimator affect the distribution, bias and risk of estimates of H_0 and multipole components of the velocity field obtained from such a model.

Even for this case it seems likely (although not certain!) that little progress would be possible analytically; however the project would certainly be amenable to Monte-Carlo studies. We have already considered one special case - the estimation of Hubble's constant from 'quiet' Hubble flow - for which an analytic treatment *is* possible. We will now briefly outline this analysis to demonstrate the basic principles of the method.

6.4.1 Optimal Estimation Of H_0

Consider the estimation of H_0 from a given sample of galaxies. Let the position, $\underline{r}$, of the sample galaxies be a random variable and let $n(r)$ denote the probability density function of the true distance, r, of a galaxy, where $r = |\underline{r}|$.

Suppose that $\hat{\mu}$ is our chosen estimator of log distance, as constructed from some given set of observables (e.g. apparent magnitude, diameter etc) and let $M(\hat{\mu}|r_0)$ denote the distribution of $\hat{\mu}$ at

a given true distance, r_0, as may be determined by the methods described in detail in the previous chapters.

Suppose now that we measure the redshift - and hence the radial recessional velocity, v_i - of each galaxy in the sample and then combine these measurements with the estimated log distance, μ_i, to obtain an estimator, H, of H_0 defined as follows:-

$$H = 10^{\tau} \tag{6.8}$$

where τ is given by:-

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \log v_i - \mu_i \tag{6.9}$$

and n is the number of galaxies in the sample. Note that this estimator is of the same form as the standard estimate of H_0 which one would obtain from, e.g., the MBS identified from the Hubble diagram of a sample (c.f. Sandage and Tammann, 1975b; see also equation 2.6) although the estimator, μ_i, in the above case is arbitrary.

Thus we see that our estimate of H_0 will depend on the measured redshifts of the sampled galaxies and also on the properties of our chosen (log) distance estimator. To proceed further we now introduce a number of simplifying assumptions.

Suppose, firstly, the case of 'quiet' Hubble Flow; i.e. where the radial velocity of the ith galaxy is simply proportional to its true

distance, viz:-

$$v_i = H_0 r_i \tag{6.10}$$

In this case we may now write equation (6.9) as:-

$$\tau = \frac{1}{n} \sum_{i=1}^{n} \log H_0 + \mu_i - \hat{\mu}_i \tag{6.11}$$

writing $\mu_i = \log r_i$. Substituting into equation (6.8), therefore, we find that we can write H as follows:-

$$H = H_0 . 10^{\frac{1}{n} \Sigma \mu_i - \hat{\mu}_i} \tag{6.12}$$

Writing H in this way is of no immediate practical value, since on the right hand side both H_0 and the true log distance of each galaxy are, of course, unknown. The advantage of introducing H_0 as a parameter become clear, however, when we consider the bias and risk of H.

We can see from equation (6.12) that H is a function of the random variables $\hat{\mu}_1,....,\hat{\mu}_n$ and $\mu_1,....,\mu_n$ - or, equivalently, $\hat{\mu}_1,....,\hat{\mu}_n$ and $r_1,....,r_n$. Hence, we can determine the expected value of H, given the true value H_0, by integrating equation (6.12) over the sampling distribution, F, of the $\hat{\mu}_i$ and the r_i. Thus we obtain:-

$$E(H|H_0) = H_0 . \int 10^{\frac{1}{n} \Sigma \mu_i - \hat{\mu}_i} \, dF \tag{6.13}$$

where dF is given by:-

$$dF = F(\mu_1,....\mu_n,r_1,....,r_n)d\mu_1...d\mu_n dr_1...dr_n \qquad (6.14)$$

Similarly, the kth moment of H, $E(H^k|H_0)$, is given by:-

$$E(H^k|H_0) = H_0{}^k.\int 10^{\frac{k}{n}\Sigma\,\mu_i - \mu_i}\, dF \qquad (6.15)$$

It is now a straightforward matter to calculate the bias, $B(H,H_0)$, and risk, $R(H,H_0)$, of H from its first and second moments using the standard definitions:-

$$B(H,H_0) = E(H|H_0) - H_0 \qquad (6.16)$$

and

$$R(H,H_0) = E(H^2|H_0) - 2H_0E(H|H_0) + H_0{}^2 \qquad (6.17)$$

We now make the further assumption that the galaxies are sampled independently; more specifically that the random vector (μ_i,r_i) is independently and identically distributed for each galaxy in the sample. This would not seem unreasonable if the galaxies are sampled at well-separated positions on the sky, given that recent measurements of two point angular correlation function for galaxies indicate an amplitude of less than 0.01 on angular scales in excess of 10° (Maddox *et al*, 1990). This simplifies the sampling distribution, F, considerably, viz:-

$$F = \prod_{i=1}^{n} f(\mu_i,r_i) \qquad (6.18)$$

where it is also useful to write the bivariate distribution function, f, as follows:-

$$f(\underline{\mu},\mathbf{r}) = M(\underline{\mu}|\mathbf{r})n(\mathbf{r}) \qquad (6.19)$$

Finally, we approximate the integrand of equation (6.15) by expanding and truncating after terms of second order. Thus we obtain:-

$$10^{\frac{k}{n}\Sigma\,\mu_i - \hat{\mu}_i} \simeq 1 + \frac{\lambda k}{n}\Sigma\mu_i - \hat{\mu}_i + \frac{1}{2}\frac{\lambda^2 k^2}{n^2}\sum_i\sum_j(\mu_i - \hat{\mu}_i)(\mu_j - \hat{\mu}_j) \qquad (6.20)$$

where $\lambda = ln(10) \simeq 2.3$

Substituting equations (6.20) and (6.18) into equation (6.15), and using the independence of the $\underline{\mu}_i$ and $\mathbf{r}_i$, we obtain the following simple expression for the kth moment of H_0:-

$$E(H^k|H_0) \simeq H_0{}^k.\left\{1 + \lambda k\int(\mu - \hat{\mu})M(\underline{\mu}|\mathbf{r})n(\mathbf{r})d\mathbf{r}d\underline{\mu} + \frac{\lambda^2k^2}{2n}\int(\mu - \hat{\mu})^2M(\underline{\mu}|\mathbf{r})n(\mathbf{r})d\mathbf{r}d\underline{\mu}\right\} \qquad (6.21)$$

Moreover we can simplify this expression further by observing that both integrands may be expressed in terms of the bias and risk of $\hat{\mu}$ at true log distance μ, viz:-

$$E(H^k|H_0) \simeq H_0{}^k.\left\{1 - \lambda k\int B(\hat{\mu},\mu)n(\mathbf{r})d\mathbf{r} + \frac{\lambda^2k^2}{2n}\int R(\hat{\mu},\mu)n(\mathbf{r})d\mathbf{r}\right\} \qquad (6.22)$$

The bias and risk of H can now be found by substituting this expression into equations (6.16) and (6.17).

Several important points emerge from the form of equation (6.22). Firstly observe that the integral over the risk of $\underline{\alpha}$ is multiplied by a factor of 1/n, but the term involving the integral over bias has no dependence on n. This means that the contribution of the risk of $\underline{\alpha}$ to the bias and risk of our estimate of H_0 can be effectively reduced simply by taking a larger sample of galaxies; this will not be the case, however, for the bias of $\underline{\alpha}$ which will make the same contribution to the bias and risk of H regardless of the sample size.

This analysis provides a clear reason for favouring the Schechter distance estimator in this context. The reasons for this are threefold:-

Firstly the bias term in equation (6.22) will vanish leaving only the integral over the risk, which can be made progressively less significant by taking a larger sample of galaxies. Hence, even if the risk of our 'Schechter' estimator is significantly larger than that of some other, biased, estimator such as, e.g., the 'Tully-Fisher' estimator at small true distances (see section 4.5.1), we can reduce the size of the 'Schechter' risk term in equation (6.22) by observing a larger sample so that, for sufficiently large n, this term can always be made smaller than the bias term for the 'Tully-Fisher' estimator.

Secondly, in the absence of distance bias the bias of H_0 will be proportional to 1/n, so that for any given sample this residual bias can be further reduced by applying resampling techniques such as the bootstrap or jackknife (c.f. Efron, 1982) which are amenable to

bias problems with this approximate n-dependence. This would again not be possible in the case of a biased estimator.

Thirdly, the fact that the risk of the 'Schechter' estimator is constant means that the bias and risk of **H** will not depend on the true distance distribution of the sampled galaxies; for example whether the sampled galaxies are drawn from the field or from clusters. Environment effects on the determination of Hubble's constant have been regarded as important in the literature (c.f. Tammann, 1987; Giraud, 1987) and indeed for an estimator whose bias and risk are a function of true distance then we see from equation (6.22) that the bias and risk of **H** will indeed depend on $n(\mathbf{r})$; we have demonstrated here, however, that the properties of the 'Schechter' estimators allow us to circumvent this problem.

6.5 Final Remarks

The aim of this thesis has been to compare quantitatively and rigorously the properties of different distance estimators, so as to provide an objective means of identifying a 'best' estimator. If we were to pick out one single conclusion which we have reached through this analysis, it would be the fact that the question of which estimator is 'best' has no straightforward answer. We have seen that there are many factors to be considered in making this choice, including the number of galaxies sampled, the number and type of available observables, and the nature of the correlation between those observables. The situation is somewhat more clear-cut when it is

possible to define 'Schechter' estimators, since we have established their zero bias and constant risk at all true distances. Based on these properties we would strongly advocate the use of 'Schechter' estimators - and, moreover, the inclusion of additional correlated observables when these are available - for the analysis of redshift surveys; we have noted, nevertheless, that when the observables are poorly correlated then the 'Schechter' estimators, although still unbiased, will have a larger risk than other, biased, estimators.

Ultimately the choice of 'best' estimator must also take account of the context in which distances are being used. The cosmologist is less interested in obtaining optimal galaxy distances than in obtaining optimal estimates of the cosmological parameters derived from those distance estimates, and it does not follow immediately that the former will lead directly to the latter. The simple example outlined above, of estimating H_0 from 'quiet' Hubble flow, illustrates this point: we have shown that the an unbiased 'Schechter' estimator is to be preferred to a biased estimator of lower risk - even though the latter could be regarded as more 'accurate'.

In this concluding chapter we have indicated how one may - at least in principle - tackle this problem rigorously by using the same techniques of risk theory which we have presented here to determine the distribution of estimators of cosmological parameters as a function of their true values and of the distributions of the observables from which they are derived. Such a formulation would offer a more complete and rigorous determination of the optimal strategies for estimating cosmological parameters, and would be a

logical extension of the analysis presented in this thesis.

APPENDIX (1): ITERATIVE REDUCTION OF BIAS AND RISK

A1.1 Introduction

We have seen in chapters (4) and (5) that the definition of unbiased 'Schechter'-type estimators of log distance, ω_0, depends on one being able to sample completely at least one observable; if this condition is not met - as for example in the case of the distance estimators studied in chapter (3) - then one cannot in general define an unbiased estimator of ω_0 for all ω_0. In such a case it would be useful to identify methods of reducing the bias and risk of distance estimators so as to at least partially remove the effects of selection. In this appendix we will outline one such method: a simple iterative scheme designed to reduce progressively the bias of a given estimator. We will apply this scheme to the 'naive' estimator, ω_N, as defined in equation (4.27), and assuming the selection function, $S(m)$, and galaxy luminosity function, $\Psi(M)$, used in chapter (3). Thus, we will show that one may substantially reduce the bias of ω_N at large true distances. Although the application of the iteration algorithm is made somewhat easier in this simple case, its formulation for estimators which are functions of two or more observables, and for other functions S and Ψ, nonetheless follows immediately.

A1.2 Definition of the Iteration Algorithm

The algorithm which form the basis of the iteration scheme can be stated quite simply.

STEP 1: Define an initial estimator, Ω_1 say, of log distance; in our case we consider $\Omega_1 = \Omega_N = 0.2(m - m_L)$, which is thus a function only of the apparent magnitude.

STEP 2: Calculate the bias, $B(\Omega_1,\omega_0)$, of Ω_1 as a function of ω_0.

STEP 3: Define a new estimator, Ω_2, as follows. Subtract from $\Omega_1(m)$ the bias of Ω_1 not at the true (but unknown) log distance, ω_0, but at the estimated log distance as given by Ω_1. In other words, assume that this initial estimate for ω_0 is , in fact, the true log distance and thus correct the bias in Ω_1 at that distance. Performing this correction for all values of m, we construct:-

$$\Omega_2(m) = \Omega_1(m) - B(\Omega_1,\omega_0=\Omega_1(m)) \qquad (A.1)$$

with the obvious generalisation to the case where Ω_1 is a function of two or more observables.

Clearly we can now repeat steps 2 and 3 iteratively: i.e. we can compute the bias, $B(\Omega_2,\omega_0)$, of Ω_2 as a function of ω_0 and thence define another estimator, Ω_3, viz:-

$$\Omega_3(m) = \Omega_2(m) - B(\Omega_2,\omega_0=\Omega_2(m)) \qquad (A.2)$$

and so forth for Ω_4, Ω_5 etc.

One would hope that each iteration would produce an estimator progressively less biased than its predecessor. In practice the effectiveness of the algorithm will depend on a number of factors - the severity of the selection effects, the form of the intrinsic

luminosity function and the 'goodness' of one's initial estimator, Ω_1 - and, in general, there is no guarantee that the scheme will always converge to give an unbiased estimator.

To demonstrate this point, it is instructive to consider first the scheme applied in a much simpler setting. Let ω be a random variable normally distributed with mean, ω_0, and unit variance. Now suppose we measure the value of ω and use this to estimate ω_0. Of course, it follows immediately from the normality of ω that it is an unbiased estimate of ω_0; i.e. in our standard notation, the estimator $\Omega_1(\omega) = \omega$ is unbiased for all ω_0.

Suppose, however, that we adopt $\Omega_1 = A_1\omega$ as our estimate of ω_0, where A_1 is a constant. Clearly the bias of Ω_1 is now given by:-

$$B(\Omega_1,\omega_0) = (A_1 - 1)\omega_0 \qquad (A.3)$$

Now of coursesince A_1 is independent of ω_0 this bias can be removed immediately simply by rescaling. Suppose, instead, we apply our iteration scheme to Ω_1. Thus, we define:-

$$\Omega_2(\omega) = A_1\omega - (A_1-1)A_1\omega \equiv A_2\omega \qquad (A.4)$$

where $A_2 = A_1(2-A_1)$. Generalising to the ith iteration, we obtain:-

$$\Omega_{i+1}(\omega) = A_{i+1}\omega \qquad (A.5)$$

and

$$B(\Omega_{i+1},\omega_0) = (A_{i+1} - 1)\omega_0 \qquad (A.6)$$

where $A_{i+1} = A_i(2 - A_i)$.

It is easy to show, by induction or otherwise, that the sequence $\langle A_i \rangle$ converges to the limit 1 if $0 < A_1 < 2$ but diverges if A_1 lies outside this range. In other words our iteration scheme will converge to an unbiased estimator for all ω_0 provided that A_1 is chosen to lie between 0 and 2. Of course this also means that the scheme will fail if our initial estimator is inadequate (i.e. if $A_1 \geqslant 2$). Thus, even in this simple example - in which we have assumed that no selection effects are present - convergence is not guaranteed. We can expect that the non-linear effects introduced by selection will impose further restrictions on the convergence of the scheme. With these notes of caution in mind consider now the iterated estimators which we obtain from ω_N.

A1.3 Application to 'Naive' Estimator

It is easy to calculate analytically the bias of ω_N - indeed we have already done this for figure (4.12) - and thence the form of $\omega_2(m)$. Subsequent iterations cannot be treated analytically but are amenable to numerical calculation. Figures (A.1) and (A.2) show the estimator curves obtained for $\omega_1 = \omega_N$ and for the first four iterated estimators (ω_2 to ω_5) as a function of $m_L - m$ and for $\sigma_M = 0.5$ and 1.0 respectively.

The first feature which is clear from both figures is that the estimators appear to differ only close to the limiting magnitude; at

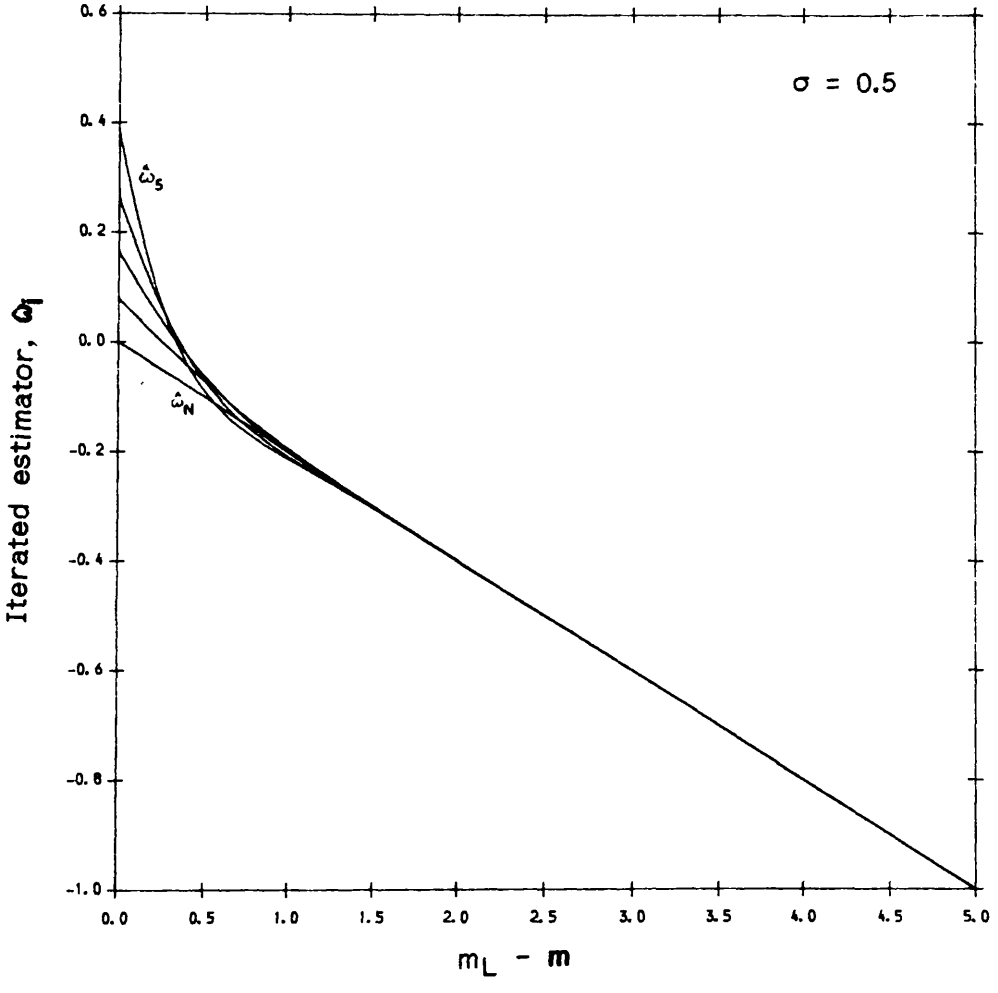

Figure (A.1)

Estimator curves obtained for $\omega_1 = \omega_N$ and for the first four iterated estimators (ω_2 to ω_5) as a function of $m_L - m$, for $\sigma_M = 0.5$

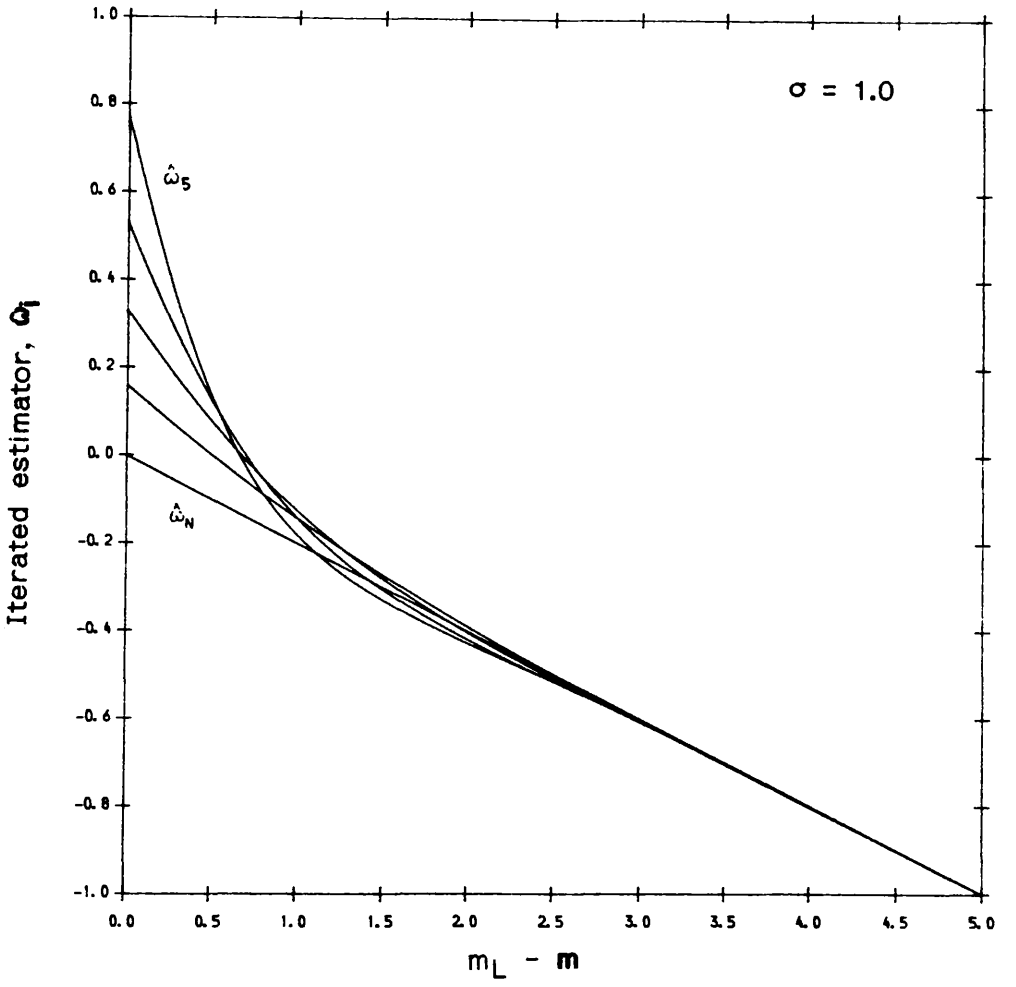

Figure (A.2)

Estimator curves obtained for $\omega_1 = \omega_N$ and for the first four iterated estimators (ω_2 to ω_5) as a function of $m_L - m$, for $\sigma_M = 1.0$

brighter magnitudes the iteration scheme does not appreciably change the linear form of $\hat{\omega}_1$. Close to m_L, however, each subsequent iteration gives a progressively greater estimate of the distance. Note that the increase in the estimator values is greater for $\sigma_M = 1$; this is not surprising since the negative bias of $\hat{\omega}_1$ at large true distances increases with σ_M, resulting in a greater positive correction to $\hat{\omega}_1$ when $\sigma_M = 1.0$. In particular, the *maximum* distance (i.e. when $m = m_L$) which can be inferred by each estimator increases with the order of the iteration. Thus, while the maximum distance which can be inferred by $\hat{\omega}_1$ is 1.0 - so that if the true distance of a galaxy is greater than unity then, no matter what its apparent magnitude, its distance will be systematically underestimated by $\hat{\omega}_1$ - in the case of $\hat{\omega}_4$, for example, this maximum estimable distance is pushed up to 1.8, for $\sigma_M = 0.5$, and to 3.4, for $\sigma_M = 1.0$. From this behaviour close to m_L, therefore, one would expect the iterated estimators to be progressively less biased at larger true distances.

Figures (A.3) to (A.6) show the bias, $B(\hat{\omega}_1,\omega_0)$, and risk, $R(\hat{\omega}_1,\omega_0)$, of $\hat{\omega}_1 = \hat{\omega}_1$ to $\hat{\omega}_5$ as a function of true distance, for $\sigma_M = 0.5$ and 1.0 respectively.

These figures show that the bias and risk of $\hat{\omega}_1$ is indeed progressively reduced at large true distances by successive iterations. Significantly, however, both the bias and risk are *not* reduced at *all* true distances. Consider $B(\hat{\omega}_2,\omega_0)$ for $\sigma_M = 1.0$, for example. We can see from figure (7.4) that $\hat{\omega}_2$ has a small positive bias in the range of true distances $x_0 \simeq 0.1$ to $x_0 \simeq 0.6$, and moreover the bias of $\hat{\omega}_2$ is greater in modulus than that of $\hat{\omega}_1$ for part of this range. This positive 'hump'

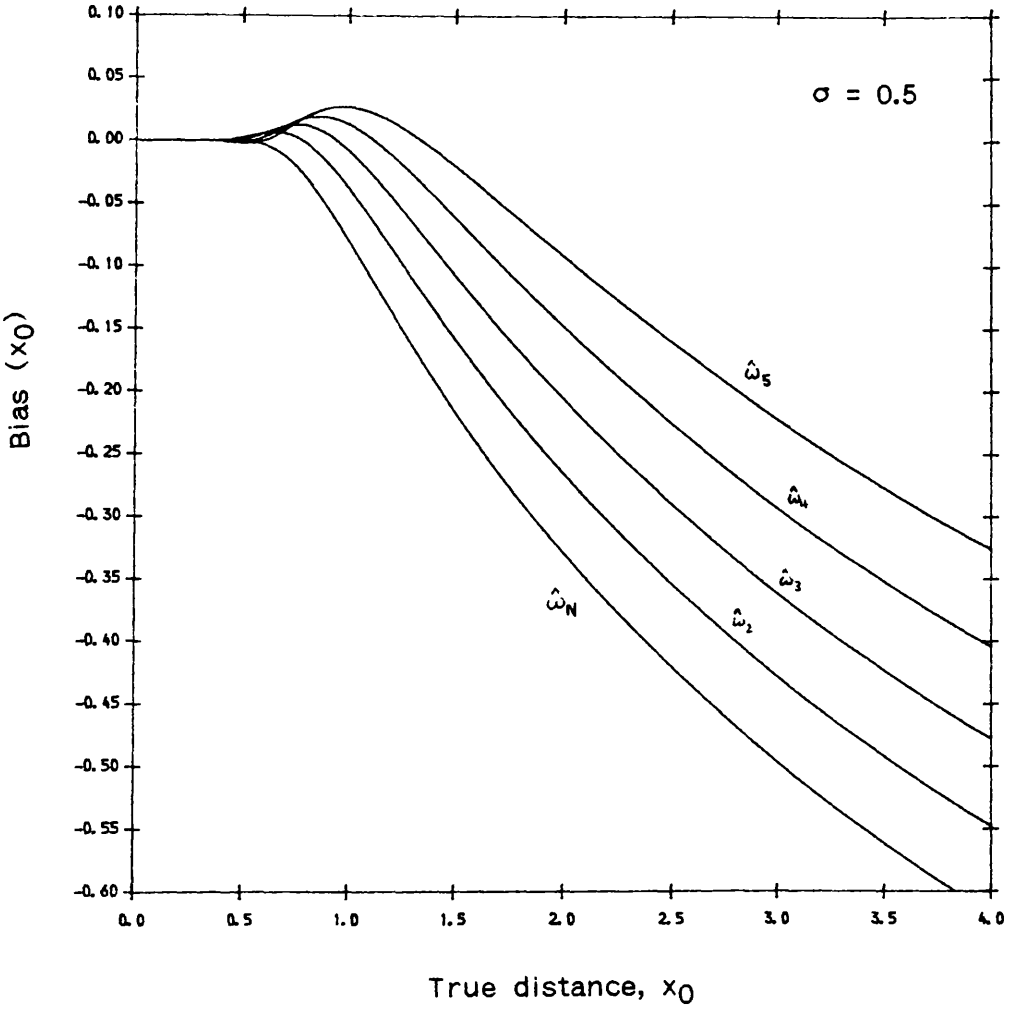

Figure (A.3)

Bias of $\omega_1 = \omega_N$ and of the first four iterated estimators, ω_2 to ω_5, as a function of the true distance, x_0, for $\sigma_M = 0.5$

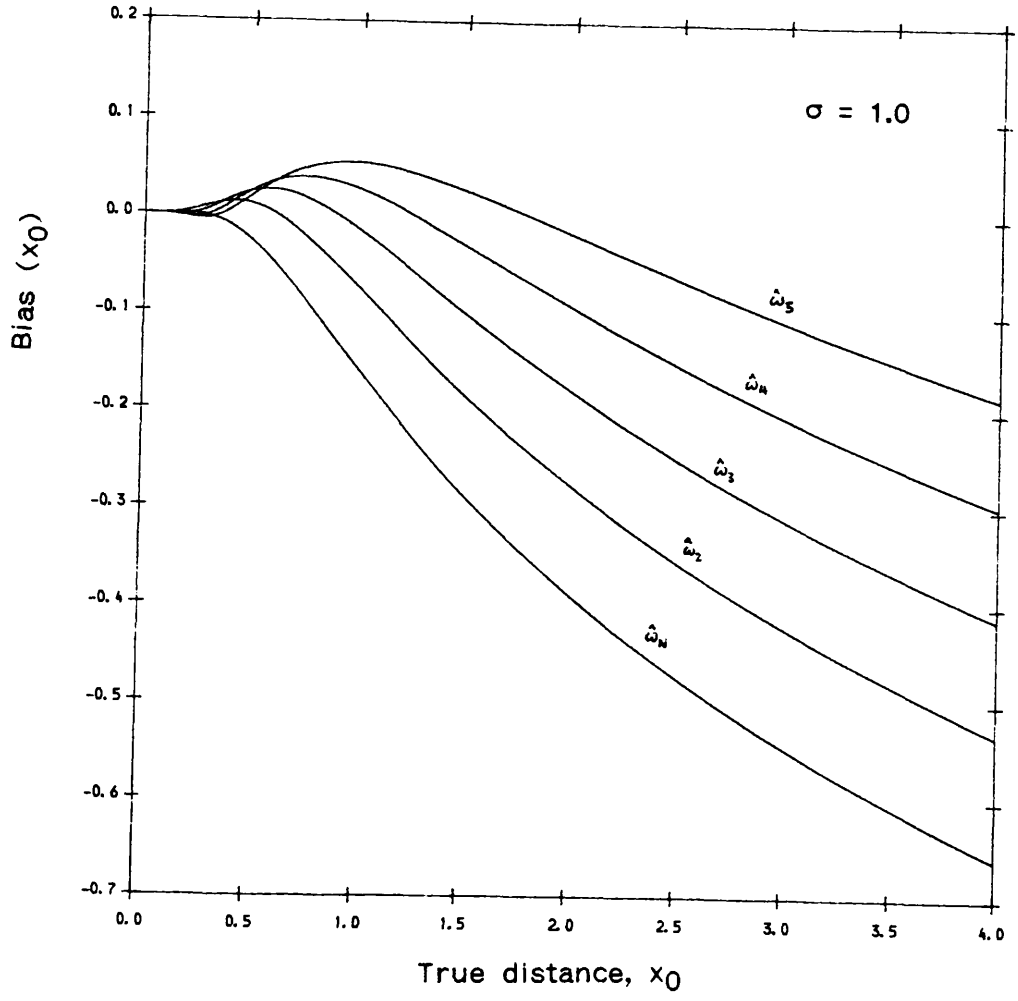

Figure (A.4)

Bias of $\omega_1 = \omega_N$ and of the first four iterated estimators, ω_2 to ω_5, as a function of the true distance, x_0, for $\sigma_M = 1.0$

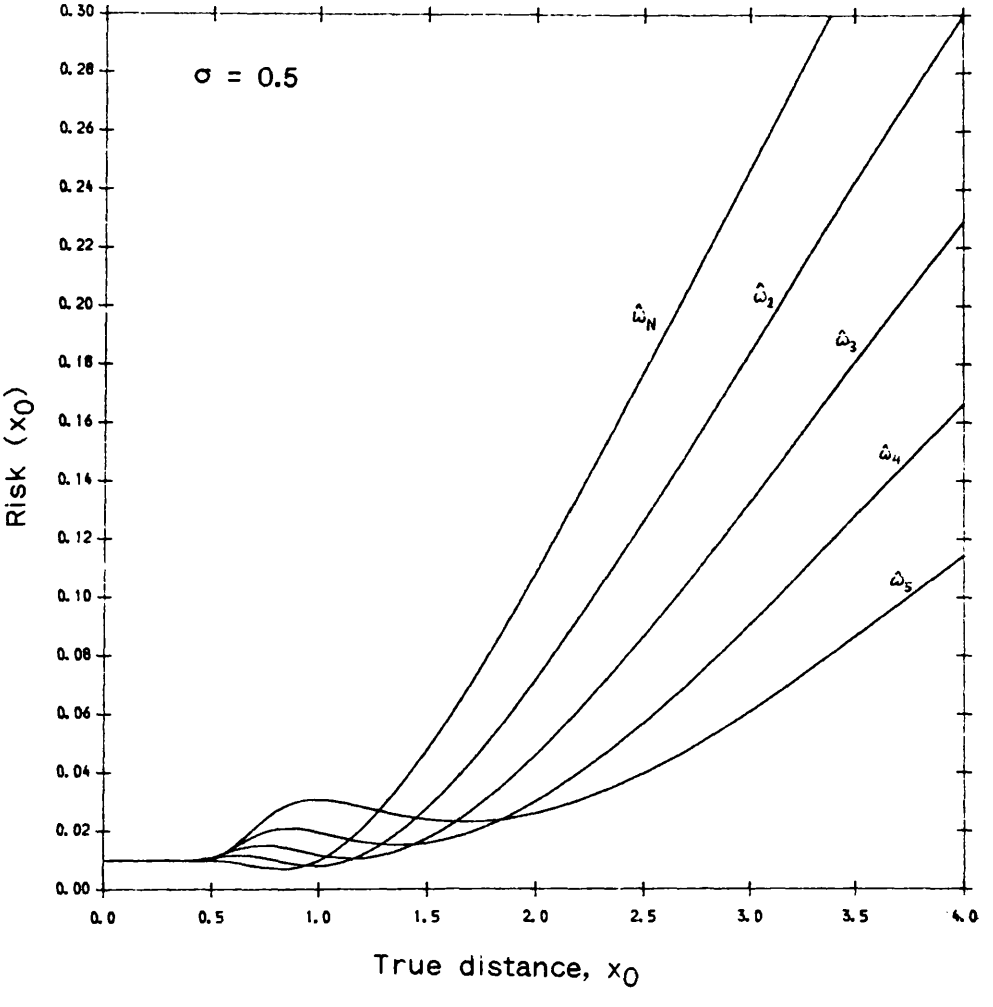

Figure (A.5)

Risk of $\omega_1 = \omega_N$ and of the first four iterated estimators, ω_2 to ω_5, as a function of the true distance, x_0, for $\sigma_M = 0.5$

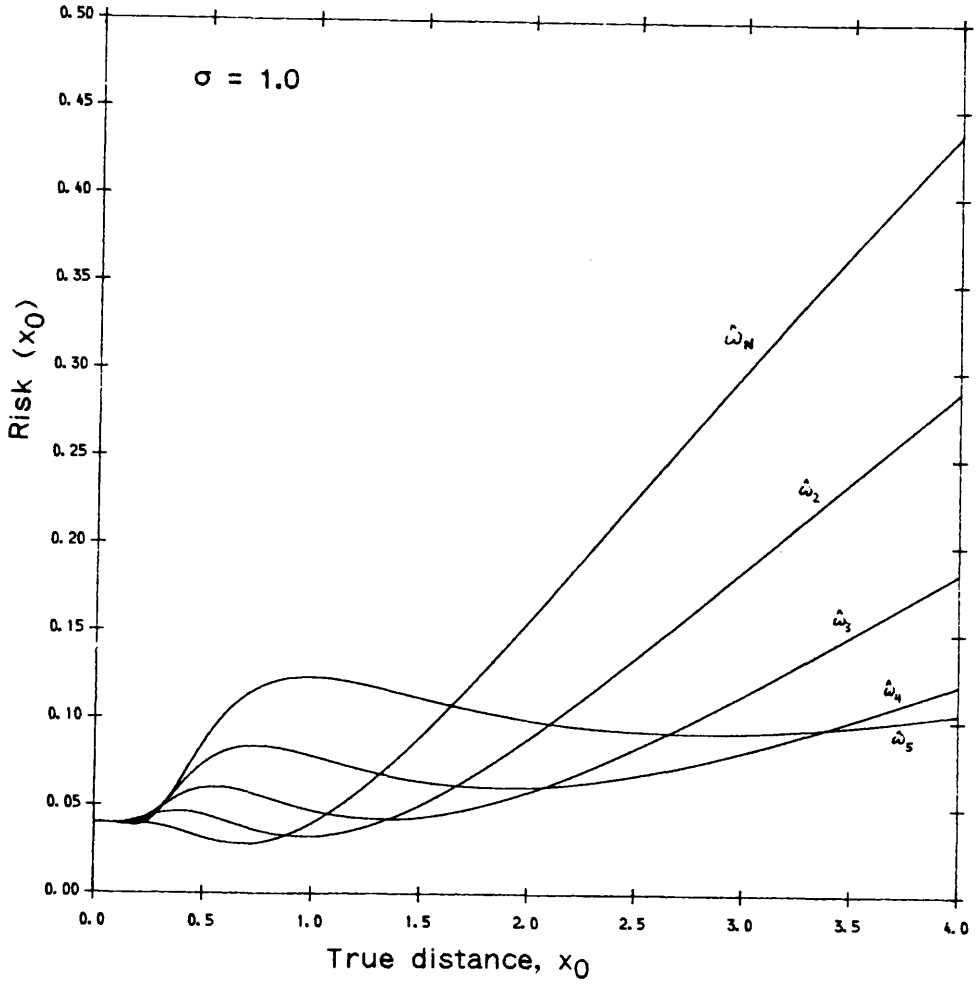

Figure (A.6)

Risk of $\omega_1 = \omega_N$ and of the first four iterated estimators, ω_2 to ω_5, as a function of the true distance, x_0, for $\sigma_M = 1.0$

in the bias curve becomes more pronounced for higher iterations so that, although the bias at large true distances is further reduced, the bias is progressively *increased* at smaller true distances. If we continue iterating beyond $\hat{\omega}_5$ we find that this effect does not disappear ; we conclude, therefore, that - as a result of the non-linear form of the bias of $\hat{\omega}_1$ - the iteration scheme does not converge for all true distances.

A similar 'hump' is seen in the risk curves, so that again the effect of successive iterations is, in fact, to *increase* the risk at small true distances. Indeed the effect on the risk is more severe, in that it is increased over a larger range of true distances. For example, we find that for $\sigma_M = 1.0$ the risk of $\hat{\omega}_2$ is greater than that of $\hat{\omega}_1$ for all true distances in the range $x_0 \simeq 0.0$ to $x_0 \simeq 0.9$. This means that, while a first application of the iteration scheme does reduce the bias of $\hat{\omega}_1$ at almost all true distances, there is a substantial range within which this is only achieved at the expense of an increase in the risk.

This trend continues, and indeed worsens, as we proceed to higher iterations: we can see that the risk curves tend to flatten out, but at a considerably higher value than the risk of $\hat{\omega}_1$ for small x_0. Thus, for example, we find that the risk of $\hat{\omega}_5$ is more than 50% greater than that of $\hat{\omega}_4$ for most of the range $0.1 \leqslant x_0 \leqslant 3.5$ ($\sigma_M = 1.0$). The same qualitative effects are evident for $\sigma_M = 0.5$, although less severe, and the precise range in which the bias and risk are increased after each iteration is also dependent on σ_M.

The fact that the bias is not reduced at all distances and the

damaging 'trade-off' between reduction of bias and increase of the risk both cause the choice of which of the iterated estimators is 'best' to be rendered difficult. Indeed, we are faced with the classic problem that the 'best' choice depends not only on value of σ_M - which one can at least assume is known - but also on the distribution of *true* distances - which will most certainly be unknown! If the galaxies in the sample are very distant ($x_0 \gtrsim 2.0$, say) then $\mathcal{Q}_4$ or $\mathcal{Q}_5$ would be a good choice. On the other hand, if a number of the galaxies are much nearer ($x_0 \lesssim 1.0$) then using $\mathcal{Q}_5$ would give very poor distance estimates for these galaxies, resulting not only in a greater bias than arises from using the original estimator, $\mathcal{Q}_1$, but also incurring as much as a threefold increase in the risk.

Similar results are obtained if we take as our initial choice one of the other estimators studied in chapter (3) (or more precisely the estimator of *log* distance corresponding to one of the distance estimators studied in that chapter.) Moreover, it is easy to show that if our initial estimator differs from $\mathcal{Q}_N$ only by a constant - as is the case if we use the 'Malmquist' or 'Proximal' distance estimator - than after the first iteration all subsequent estimators ($\mathcal{Q}_3$, $\mathcal{Q}_4$, etc) are identically equal to found obtained if we start the scheme with $\mathcal{Q}_N$.

In summary, therefore, the iteration scheme clearly does offer a means of reducing the bias and risk of distance estimators at very large true distances, but its use is limited by the fact that after only a few iterations this is achieved only at the cost of a significant increase in the risk at *smaller* true distances. Hence, if one's sample contains a number of relatively nearby galaxies (typically at $x_0 \lesssim 1.0$)

then performing more than one or two iterations will, in fact, be counter productive and will result in *poorer* distance estimates.

REFERENCES

Aaronson, M.; Huchra, J.; Mould, J.; Schechter, P.; Tully, R.B.
1982 Ap.J. 258, 64

Aaronson, M.; Mould, J.; Huchra, J. 1980 Ap.J. 237, 655

Abramowitz, M.; Stegun, I. *Handbook of Mathematical Functions*
(Dover, 1968)

Baum, W.A. 1959 P.A.S.P. 71, 106

Bertschinger, E.; Dekel, A. (BD) 1989 Ap.J.(Lett) 336, L5

Bertschinger, E.; Dekel, A.; Faber, S.; Dressler, A.; Burstein, D.
1990 Ap.J. 364, 370

Binggelli, B.; Sandage, A.; Tammann, G. 1988 *Ann Rev Astron Astrophys*
26, 509

Bottinelli, L.; Gougenheim, L.; Paturel, G.; Teerikorpi, P.
1986 Astron. Astrophys. 156, 157

Burstein, D. 1982 Ap.J. 253, 539

Burstein, D.; Heiles, C. 1982 Astron. J. 87, 1165

Collins, C.A.; Joseph, R.D.; Roberston, N.A. 1986 *Nature* 320, 506

Collins, C.A.; James, P.A.; Joseph, R.D. 1991 MNRAS 248, 444

da Costa, L.N.; Nunes, M.A.; Pellegrini, P.S.; Willmer, C.
1986 Astron. J. 91, 6

Dekel, A. "Reconstruction Methods in Cosmology", in *Physical Cosmology*
eds A. Blanchard, L. Celnikier (Editions Frontieres, 1991)

Dekel, A.; Bertschinger, E.; Faber, S. (DBF) 1990 Ap.J. 364, 349

Dekel, A.; Bertschinger, E.; Yahil, A.; Strauss, M.S.; Davis, M.
1992 (in preparation)

Dressler, A. 1980 Ap.J. 236, 351

Dressler, A.; Lynden-Bell, D.; Burstein, D.; Davies, R.L.;

Faber, S.; Terlevich, R.; Wegner, G. 1987 Ap.J. 313, 42

Eddington, A. in *Stellar Movements and the Structure of the Universe* (MacMillan, 1915)

Efron, B. *The Jackknife, Bootstrap and other Resampling Plans* (SIAM, 1982)

Efstathiou, G.; Kaiser, N.; Saunders, W.; Lawrence, A.; Rowan-Robinson, M.; Ellis, R.; Frenk, C.S. 1990 MNRAS 247, 10

Einasto, M 1990 MNRAS 242, 56

Faber, S.M.; Jackson, R.E. 1976 Ap.J. 204, 668

Fall, S.M.; Jones, B.J.T. 1976 *Nature* 262, 457

Feast, M. 1987 *Observatory* 107, 1080

Feigelson, E.; Isobe, T.; Akritas, M.G.; Babu, G.J. in *Errors Bias and Uncertainites in Astronomy* (CUP, 1990)

Felten, J. 1985 Comments Astrophys. 11, 53

Fixsen, D.; Cheng, E.; Wilkinson, D. 1983 *Phys Rev Lett* 50, 620

Giraud, E. 1987 Astron. Astrophys. 174, 23

Giuricin, G.; Mardirossan, F.; Mezzetti, M.; Ravella, M.; Pisani, A. 1989 *Astrophys Space Sci.* 156, 153

Graybill, F. *An Introduction to Linear Statistical Models* (McGraw-Hill, 1961)

Hamilton. A.J.S. 1988 Ap.J.(Lett) 331, L59

Hart, L.; Davies, R.D. 1982 *Nature* 297, 191

Hoel, P. *An Introduction to Mathematical Statistics* (Wiley, 1962)

Hogg, R.; Craig, A. *An Introduction to Mathematical Statistics* (MacMillan, 1978)

Holmberg, E. *Arkiv Astronomi* 1969 5, 305

Hubble, E. 1929 Proc. Nat. Acad. Sci. 15, 168

Jones, B.J.T. *Lectures for the IAC Winter School "Observational and Physical Cosmology"* (CUP, 1991)

Jones, B.J.T.; van de Weygaert, R. *Lecture for the XII Autumn School "The Physical Universe"* Lisbon, October 1990

Kaiser, N.; Efstathiou, G.; Ellis, R.; Frenk, C.S.; Lawrence, A.; Rowan-Robinson, M; Saunders, W. 1991 MNRAS 252, 1

Kendall, M.; Stuart, A. *The Advanced Theory of Statistics vol 1* (Haffner Publ Co. NY, 1963)

Kennicutt, R.C. 1979 Ap.J. 228, 696

Kermack, K.; Haldane, J.B. 1950 Biometrika 37, 30

Kirschner, R.P.; Kwan, J. 1974 Ap.J. 193, 27

Kirschner, R.P; Oemler, A.; Schechter, P.L. 1979 Astron. J. 84, 951

Kolb, E.W.; Turner, M.S. *The Early Universe* (Addison-Wesley, 1990)

Kraan-Korteweg, R.C.; Cameron, L.; Tammann, G.A. in *Galaxy Distances and Deviations from Universal Expansion* (eds Madore, B.; Tully, R. Reidel, 1986)

Kristian, J.; Sandage, A.; Westphal, J.A. 1978 Ap.J. 221, 383

de Lapparent, V.; Geller, M.J.; Huchra, J. 1988 Ap.J. 332, 44

de Lapparent, V.; Geller, M.J.; Huchra, J. 1991 Ap.J. 369, 273

Lilje, P.; Yahil, A.; Jones, B.J.T. 1986 Ap.J. 307, 91

Lubin, P.M.; Epstein, G.L.; Smoot, G.F. 1983 *Phys Rev Lett* 50, 616

Lynden-Bell, D 1991 Preprint.

Lynden-Bell, D.; Faber, S.M.; Burstein, D.; Davies, R.; Dressler, A.; Terlevich, R.J.; Wegner, G. 1988 Ap.J. 32, 19

Maddox, S.J.; Sutherland, W.; Efstathiou, G.; Loveday, J. 1990 MNRAS 243, 692

Malmquist, K.G. 1920 Medd Lund Astron Obs, 20

Martin, W.L.; Warren, P.R.; Feast, M.W. 1979 MNRAS 188, 139

Michard, R. 1979 Astron. Astrophys. 74, 206

Mihalas, D.; Binney, J *Galactic Astronomy* (Freemann, 1981)

Mood, A.M.; Graybill, A.F. *Introduction to the Theory of Statistics* (McGraw-Hill, New York, 1974)

Paturel, G 1979 Astron. Astrophys. 71, 19

Peebles, P.J.E. *Large Scale Structure of the Universe* (Princeton, 1980)

Peterson, C.J.; Baumgart, C.W. 1986 Astron.J. 91, 530

Pierce, M.J.; Tully, R.B. 1988 Ap.J. 330, 579

Postman, M.; Geller, M.J. 1984 Ap.J. 281, 95

Rees, M.J.; Stoneham, R.J. *eds "Supernovae: a Survey of Current Research"* (Reidel, 1982)

Roberts, M.S. 1978 Astron.J. 83, 1026

Rowan-Robinson, M. *The Cosmological Distance Ladder* (Freeman, 1985)

Rowan-Robinson, M.; Sauders, W.; Lawrence, A.; Leech, K
1991 MNRAS 253, 485

Rubin, V.C. *in "Internal Kinematics and Dynamics of Galaxies"* (IAU Symposium 100) Ed E. Athanassoula (Reidel, 1983)

Rubin, V.C.; Ford, W.K.; Rubin, J.S. 1973 Ap.J (Lett) 183, L111

Rubin, V.C.; Thonnard, N.; Ford, W.K.; Roberts, M.S.
1976 Astron. J. 81, 719

Rybicki, G.; Lightman, A. *Radiative Processes in Astrophysics* (Wiley, 1979)

Sandage, A. 1972 Ap.J. 176, 21

Sandage, A.; Binggelli, B.; Tammann, G.A. 1985 Astron. J. 90, 1759

Sandage, A.; Hardy, E. 1973 Ap.J. 183, 743

Sandage, A.; Tammann, G.A. 1974 Ap.J. 190, 525

Sandage, A.; Tammann, G.A. 1975a Ap.J. 196, 313

Sandage, A.; Tammann, G.A. 1975b Ap.J. 197, 265

Sandage, A.; Tammann, G.A. 1981 *Revised Shapely-Ames Catalog of Bright Galaxies* (Carnegie Inst. Washington)

Sandage, A.; Tammann, G.A. 1990 Ap.J. 363, 1

Schechter, P. 1976 Ap.J. 203, 297

Schechter, P. 1980 Astron. J. 85, 801

Schneider, D.P.; Gunn, J.E.; Hoessel, J.G. 1983 Ap.J. 264, 337

Shane, C.D.; Wirtanen, C.A. 1967 Publ. Lick Obs. 22, Pt1

Simmons, J.F.L.; Stewart, B.G. 1985 Astron. Astrophys. 142, 100

Smoot, G. *et al* in *"Physical Cosmology"* eds A. Blanchard, L. Celnikier (Editions Frontieres, 1991)

Staveley-Smith, L; Davies, R.D. 1987 MNRAS 224, 953

Strauss, M.; Davis, M; Yahil, A.; Huchra, J.P. 1990 Ap.J. 361, 49

Stromberg, G. 1940 Ap.J. 92, 156

Tammann, G.A. *"The Cosmic Distance Scale"* in Observational Cosmology (IAU Symposium 130) Ed. A. Hewitt (1987)

Tammann, G.A.; Sandage, A. 1985 Ap.J. 294, 81

Teerikorpi, P. 1975 Astron. Astrophys. 45, 117

Teerikorpi, P. 1984 Astron. Astrophys. 141, 407

Teerikorpi, P. 1987 Astron. Astrophys. 173, 39

Teerikorpi, P. 1990 Astron. Astrophys. 234, 1

Terlevich, R.J.; Davies, R.L.; Faber, S.M.; Burstein, D. 1981 MNRAS 196, 381

Tully, R.B. 1988 *Nature* 334, 209

Tully, R.B.; Fisher, J.R. Astron. Astrophys. 54, 661

Tully, R.B.; Mould, J.; Aaronson, M. 1982 Ap.J. 257, 527

de Vaucouleurs, G. 1977 *Nature* 266, 126

de Vaucouleurs, G.; Olsen, D.W. 1982 Ap.J. 256, 346

de Vaucouleurs, G.; Peters, W.L. 1984 Ap.J. 287, 1

Visvanathan, N.; Sandage, A. 1977 Ap.J. 216, 214

Weinberg, S. *Gravitation and Cosmology* (wiley, 1972)

Zel'dovich, Ya B. 1970 Astron. Astrophys. 5, 84

Zwicky, F.; Wield, P.; Herzog, E. Karpowicz, M.; Kowal, C.
1961-68 *Catalogue of Galaxies and Clusters of Galaxies* 6 vols
(Pasadena, Calif. Inst. Tech.)