THE ASTROPHYSICAL JOURNAL, 431:147–155, 1994 August 10

REDSHIFT DATA AND STATISTICAL INFERENCE

WILLIAM I. NEWMAN,[1] MARTHA P. HAYNES,[2,3] AND YERVANT TERZIAN[2,3]
Received 1993 September 7; accepted 1994 February 4

ABSTRACT

Frequency histograms and the "power spectrum analysis" (PSA) method, the latter developed by Yu & Peebles (1969), have been widely employed as techniques for establishing the existence of periodicities. We provide a formal analysis of these two classes of methods, including controlled numerical experiments, to better understand their proper use and application. In particular, we note that typical published applications of frequency histograms commonly employ far greater numbers of class intervals or bins than is advisable by statistical theory sometimes giving rise to the appearance of spurious patterns. The PSA method generates a sequence of random numbers from observational data which, it was claimed, is exponentially distributed with unit mean and variance, essentially *independent* of the distribution of the original data. We show that the derived random processes is nonstationary and produces a small but systematic bias in the usual estimate of the mean and variance. Although the derived variable may be reasonably described by an exponential distribution, the tail of the distribution is far removed from that of an exponential, thereby rendering statistical inference and confidence testing based on the tail of the distribution completely unreliable. Finally, we examine a number of astronomical examples wherein these methods have been used giving rise to widespread acceptance of statistically unconfirmed conclusions.

Subject headings: galaxies: distances and redshifts — methods: statistical

1. INTRODUCTION

The analysis of data and the identification of possible patterns remains one of the fundamental objectives of observational science. Many phenomena in the physical sciences lend themselves to the conjecture that there exists some underlying structure or periodicity in different kinds of data. The development of unbiased tests becomes critical to the task of identifying such patterns. Astronomy, especially, has had a long history of claim of pattern and form which ultimately proved to be false. Sheehan (1988), a psychiatrist and psychologist, has chronicled the evolution of planetary astronomy and how observational information of a largely qualitative sort became prone to misinterpretation, a consequence of the way the eye and brain function together. The observational claims by Schiaparelli and Lowell of "canals" on Mars is a classic example of this phenomenon. Only recently have psychologists developed an appreciation for how the eye seemingly finds pattern where none exists. The pioneering work of researchers such as Julesz (1981) have succeeded in *quantifying* how images, consisting of high densities of points (cf. astronomical photographic plates), could suggest to viewers the presence of pattern that high-order correlation statistical methods then showed to be nonexistent. Barrow & Bhavsar (1987) explored the role that filamentary structures and our perception of them could contribute to their misinterpretation. The development of other quantitative measures of pattern and, ultimately, of hypothesis testing (Fukunaga 1990) has become a fundamental objective of the observational sciences. Despite the advent of qualitative improvements in astronomical data, the outcome of any experiment is only as reliable as the statistical methods employed—hence, a qualitative refinement of the statistical tools employed is also necessary.

Data samples are particularly prone to misidentification of pattern or of clustering. Recently, we (Newman, Haynes, & Terzian 1989) showed that the inappropriate application of statistical methods to Gaussian *random noise* could provide an illusion of pattern, a phenomenon we identified with the *statistics of small numbers*. (This effect is particularly prominent in descriptive statistics, such as frequency histograms, which are particularly prone to the inappropriate selection of class interval size or "binning.")

A landmark development in astronomical investigations of clustering is the celebrated "power spectrum analysis" (PSA) method of Yu & Peebles (1969) which generates a sequence of random numbers from observational data that, it was claimed, is exponentially distributed with unit mean and variance, *essentially independent of the distribution of the original data.* Unlike other methods in widespread use at that time, the PSA method was designed to be independent of any form of binning, and hence would produce conclusions that are independent of the chosen class interval or bin size. The name of the method, however, is somewhat misleading as its intent is to identify clustering phenomena in a set of discrete measurements. (We will see later how this name naturally emerged.) The purpose of the PSA technique is to determine whether observational data are "smoothly" distributed in some sense, or possesses a regular, possibly periodic clustering of samples.

The power spectrum analysis method has been widely employed in the investigation of astronomical data for over two decades. In their original paper, Yu and Peebles analyzed the distribution in the sky of the "rich," "compact" clusters of galaxies in an attempt to find an independent test for the existence of superclusters in these data. Yu and Peebles employed

[1] Departments of Earth and Space Sciences, Astronomy, and Mathematics, University of California, Los Angeles, CA 90024.

[2] Department of Astronomy, Cornell University, Ithaca, NY 14853 and National Astronomy and Ionosphere Center, Cornell University, Ithaca, NY 14853.

[3] The National Astronomy and Ionosphere Center is operated by Cornell University under a cooperative agreement with the National Science Foundation.

cataloged sequences of angular positions on the sky to assess the reality of superclustering, and concluded that there was a possibly significant component of nonrandomness in the data. Not long thereafter, Burbidge & O'Dell (1972) employed the method to explore some controversial questions pertaining to the distribution of redshifts of quasi-stellar objects and related emission-line objects. Tifft (1973) in an unusual paper employed the PSA method, in addition to some unquantifiable ad hoc pattern recognition schemes, claiming that it validated his hypothesis that the Coma Cluster of galaxies displays evidence of some form of "quantization" with respect to redshift. Sharp (1984, 1990) explored the effect of "binning" in Coma Cluster investigations, establishing that the evidence for anomalous behavior was not conclusive. (Newman, Haynes, & Terzian 1992 have explored the statistics of a refined set of Coma Cluster redshift data without finding any evidence for quantization, although other data could imply the presence of subclustering—see Beers's comments following Newman et al. 1992.) More recently, Guthrie & Napier (1990) examined data from the Virgo Cluster seeking evidence without success for redshift quantization. Broadhurst et al. (1990) used the method of power spectrum analysis and claimed to find clustering of galaxies near the Galactic poles on extraordinarily large distance scales. It is possible, however, that the apparent clustering noted by Broadhurst et al. was an artifact of the pencil-beam survey method that they employed. Bahcall (1991) has shown that these results could be the outcome of a network structure of large supercluster surfaces surrounding low-density regions, and that the apparent periodicity would diminish when averaged over different directions. Kaiser & Peacock (1991) have also cautioned against the conclusion of a significant signature of very large-scale structure in the Broadhurst et al. (1990) data because of possible aliasing of the high spatial frequencies (typical of small scale clustering). Scott (1991) has explored the question of periodicity in quasar redshifts, concluding that there is no evidence for such a periodic structure. The claimed appearance of clustering in some of the above situations seemingly contradicts long-accepted results in relativistic astrophysics and cosmology.

If these claims of quantization are correct, then a profound change will become necessary in our picture of the large-scale structure of the universe. Before we abandon our contemporary view of physics, it is essential that we ascertain that we have not been led astray by some subtle *mathematical* artifice associated with, for example, the method of power spectrum analysis. In this paper, we will carefully examine the derivation of this widely used method of astronomical data analysis. In so doing, we will show that the derived random process preserves a subtle imprint of the original distribution, rendering the derived process nonstationary and producing a small but *systematic* bias in the usual estimate of the mean and variance. Further, as the method's derivation employs assumptions derived from the central limit theorem (Dudley 1989; Feller 1968, chap. 8), the *asymptotic rate of convergence* of the random variables to an exponential distribution becomes critically important. Although the derived variable may be reasonably described by an exponential distribution over much of its range, the tail of the distribution is far removed from the behavior expected of a unit mean, unit variance exponential. Consequently, statistical inference and confidence testing *based on the tail of the distribution*, which was performed in some of the more speculative articles cited above, is completely unreliable.

We believe that the power spectrum analysis method incorporates some interesting ideas in producing a statistical test for clustering, but subtle mathematical details render the method inapplicable to situations presented by most available sample sizes. We will conclude by considering some analytic and simulation examples which illustrate these issues and provide a better measure of the utility and limitations of this useful method of statistical analysis.

2. SOME STATISTICAL APPROACHES IN ASTRONOMICAL USE

Observational astronomers employ a battery of statistical methods to develop an understanding of their data. Beginning with descriptive statistics in the form of histograms or sample frequency distributions, they ultimately turn to methods that permit the testing of hypotheses and the assignment of confidence limits. We will begin our discussion by reviewing the history of frequency histogram analysis and provide some illustrations of its use (and abuse) using artificially generated data whose statistical properties (including correlation and periodicity) is a priori known. Then, we will turn our attention to the method of Yu & Peebles (1969) and, by providing a mathematically rigorous analysis, show that the random variable it employs is nonstationary, is a biased estimator for the exponential distribution, and finally exhibits weak convergence to an exponential distribution. These technical subtleties are critical since it is widely believed that the method is in some sense *exact*—we will show that the *approximations* it implicitly contains render it unreliable, particularly in the context of looking for anomalies in redshift data, and any *inference* made for quantization or clustering is unsubstantiated. (It is important to stress that this does not mean that the hypothesis is necessarily falsified; it simply means that the deficits present in the statistic renders these tests inconclusive.) We will again provide numerical illustrations of how these subtle deficits in the method will manifest.

2.1. *Frequency Histogram Analysis*

Descriptive statistics do not permit the assignment of probabilities or hypothesis testing, but useful insight can be obtained if the methods are properly applied. However, descriptive statistics are more prone to misuse that virtually any other statistical technique. Books, such as those by Huff (1954), Kimble (1978), Runyon (1981), and Tufte (1983, chaps. 2–3) describe the many pitfalls that come from using descriptive statistics and, especially, histograms. This point is often overlooked in contemporary textbooks, although some older textbooks such as Hald's (1952, p. 49) exposes the dangers of histogram binning, i.e., its sensitivity to the length of the class interval, and make the case for employing cumulative frequency polygons instead. Regrettably, few current textbooks make this point. More recently, the statistical literature that is oriented toward statistics education have begun to reemphasize this point. For example the paper by Gentlemen (1977) uses interactive graphics in a classroom setting to show among other things how the appearance of histograms is sensitive to the "arbitrariness of (class) interval boundaries."

The application of computers to the statistical analysis of data has resulted in the appearance of a variety of authoritative books devoted to descriptive statistics as a method of indentifying trends in a manner preliminary to a proper statistical treatment. Tukey (1977, p. 125), although oriented toward the needs of social and other nonquantitative scientists, presents in this classic reference a methodology to help avoid the many pitfalls common to descriptive statistics. Although Tukey does not mention histograms or frequency by name, he deals extensively with frequency plotting and the need to smooth data via a Hanning filter to minimize misleading fluctuations. Schmid (1983, p. 68) has an extensive discussion on how histograms with too many class intervals gives misinformation, showing extensive comparisons, similar in character to those of Newman et al. (1989). Schmid discussed problems of frequency polygons and of smoothing distributions. The widely cited volumes by Chambers (1977, pp. 222–226) and Chambers et al. (1983, chap. 2) consider in some detail the question of non-smoothness in histograms, the use of cumulative distributions and of quantiles, "smearing" of distributions by the application of box-car functions, as well as Sturges's (1926) and Pearson's (1936) rules for estimation of frequency class interval.

Although the statistics community does not speak with a single voice on how to deal with histograms, it does offer three routes. The first is to simply avoid histograms and employ, instead, cumulative distribution functions and/or quantiles, measures of skew and kurtosis, *g*- and *h*-distributions, etc. Recent references to this approach can be found in Hoaglin, Mosteller, & Tukey (1985, p. 345), particularly in the individual articles by Hoaglin and Tukey ("Checking the Shape of Discrete Distributions"), and by Hoaglin ("Using Quantiles to Study Shape" and "Summarizing Shape Numerically"). Wilk & Gnanadesikan (1968) discuss the relative fragility of histograms or frequency distributions and advocate instead the use of empirical cumulative distribution functions, including the quantile, percent and hybrid plots for comparing two populations (the same formal basis as that employed in non parametric tests such as the Kolmogorov-Smirnov test). The second approach is to use histograms, but to smooth them by employing box-car or Hanning filters, as discussed in Chambers et al. (1983), Schmid (1983), and Tukey (1977). When properly applied, this approach can be highly informative and, in the limit of infinite data, converges uniformly to the true distribution if sufficient care is exercised. The third of these approaches, which is somewhat related to the smoothing process, is to use a combination of Sturges's (1926) rule and Pearson's (1936) rule (discussed below). More recently, Doane (1976) has developed a scheme for combining the two latter methods of assigning frequency class intervals, i.e., of determining the width of the histogram bins. Other ad hoc schemes have been proposed, e.g., Dixon & Kronmal (1965), but they are neither widely used nor accepted.

During the past decade, Scott (1979) and Freedman & Diaconis (1981a, b) have used other probabilistic considerations in developing yet another class of rules for class interval selection. Despite the apparent plethora of rules for producing histograms, they are all in qualitative agreement in terms of bin size selection for the sizes of data sets frequently encountered in astronomy—and in marked departure from those often selected by astronomers.

Sturges (1926) developed an ad hoc formula for the class interval in applications to frequency distributions, especially those that are near Normality:

> It is based on the principle that the proper distribution into classes is given, for all numbers which are powers of 2, by a series of binomial coefficients. For example, 16 items would be divided normally into 5 classes, with class frequencies 1, 4, 6, 4, 1. Thus if a statistical series had 64 items ..., it should be divided into 6 plus 1 or 7 classes. ...If the formula gives 9, 10 may be chosen, but if the formula indicates 7 or 8, the one actually used should generally be the next lower convenient class interval, 5.

Accordingly, given N items, Sturges's rule would allocate $1 + \log_2 N$ class intervals, generally reduced to a multiple of 5. By employing Sturges's rule, we avoid the problem of introducing artificial fluctuations associated with excessive numbers of class intervals. In situations with significant skew or kurtosis, it was observed in the decade following Sturges's work that his scheme had to be modified to accommodate departures from symmetry and from a Normal distribution. This is accomplished in Pearson's (1936) well-known $(b_1)^{1/2}$ criterion, which is a dimensionless rendering of the skew relative to the variance of the population. These considerations are particularly relevant to astronomical situations, such as redshift distributions, which often seem to manifest a Gaussian character.

Some refinement of these criteria has taken place over the last five decades, but the state of the art is summed up in Doane's (1976) paper which introduces some ideas from information theory and the theory of cooding into the problem. Doane asked questions such as

> How do you teach a computer to look at a set of sample observations on one variable and make a frequency classification with the "right" number of classes, "nice" class limits, and "round" interval widths....

He repeats Sturges's rule for classifying a series of N items,

> The optimal number of classes, in general, is $K = 1 + \log_2 (N)$.

Doane goes on to show how Sturges's rule should be modified in the light of Pearson's criterion. In the case of a symmetric distribution, Doane shows that no modification is necessary. Moreover, as the sample size becomes larger, the number of class intervals increases but at a decreasing rate. Work cited earlier, e.g., Chambers (1977), give essentially equivalent formulae for the number of class intervals to be employed. In all cases, the appearance of any asymmetry has only a modest effect.

Scott (1979) as well as Freedman & Diaconis (1981a, b), employing other probabilistic considerations, have produced another class of rules for class interval selection. In essence, the arguments presented by the above authors call for selection of the number of intervals *in proportion* to $N^{1/3}$ where N is the cumulative number of samples. (It is important to note that Freedman and Diaconis developed asymptotic scaling properties according to minimizing fluctuations in the histogram, i.e., the $N^{1/3}$ scaling rule, but avoided the question of the selection of the multiplication coefficient. Although they offer a suggestion based on the *interquartile* range, it has no rigorous basis and was chosen on essentially esthetic grounds to give the right order of magnitude.) For sample sizes ranging from $\approx$10–100 points, the estimated number of class intervals to be employed by this latter scheme is quite comparable with that provided by Sturges's rule.

The application of "smoothing" methods involving a box-car or Hanning filter as described above also requires a selection criterion for the filter width in order to obtain an appropriate measure of the class interval. Sturges's rule and Pearson's criterion can readily be adopted to this purpose, but the basic considerations are the same as those just discussed. The first class of methods of characterizing shape are still more conservative and would claim much less detail about any data set than the schemes just described. Thus, as the procedure employed by many astronomers' is closest conceptually to the third class of methods, we now evaluate the degree to which their analysis conforms with this standard.

In order to explore the possibly anomalous behavior that will appear in frequency histograms, we generated Gaussian random data with zero mean and unit variance using the well-known routine "gasdev" in conjunction with the routine "ran2" (Press et al. 1992). (These routines are assured by its authors to have no built-in correlations for 2.3×10^{18} numbers—they are so certain of the integrity of the method that they have offered to pay $1000 to the first reader who convinces them otherwise.) In order to preserve the character of the astronomical problem, 300 "velocities" were generated (scaled to unit variance and zero mean). Using Sturges's or the other class interval selection rules described earlier, we would expect to use $\approx$8–10 bins—we employed 10 class intervals in Figure 1, possibly erring on the side of including too many class intervals. The histogram clearly looks Gaussian with fluctuations that are, as we might expect, the order of the square root of the number of counts in a particular column. In order to provide a sense of what occurs when an inordinate number of class intervals are employed, Figure 2 uses 72, which is comparable to the number employed by a number of authors. The fluctuations persist to be of the order of the square root of the number of counts in a particular column. However, the markedly reduced height of each column has made this effect strong—and the illusion of substantial pattern is striking. By simply using different (but overly large) numbers of bins and their origin chosen at a computer terminal (Gentlemen 1977), we can readily develop a surreal collection of histograms from the same data! (See Newman, Haynes, & Terzian 1989, 1992[4] for further illustrations and discussion.)

Ten years after the discovery of pulsars, about 150 of them had been detected. Manchester & Taylor (1977, p. 9) discussed the histogram of their period distributions (with approximately 20 class intervals) which suggested that pulsars are of two classes, those with short and those with long periods. However, at present more than 500 pulsars have been detected and the binary nature of their period distribution has disappeared!

An example of overbinning can be seen in the histograms of the nearly same data set presented as Figure 2 of Schneider & Salpeter (1992) and then as Figure 1 of Cocke (1992). The latter "histogram" clearly violates the rules presented here, employing more than 50 "bins" for 134 data points. On the basis of the above rules for class interval selection and given the size of astronomical data sets (with generally fewer than 1000 data points), the use of more than 10 class intervals is thoroughly unjustified.

It also should be apparent from these computational examples that, had there been genuine quantization present, a frequency histogram plotted according to these accepted rules would fail to exhibit such properties. Thus, frequency histograms are useful only inasmuch as they could demonstrate some gross anomally, e.g., highly skewed symmetry, but not for showing any detail within the underlying distribution given the quantity of data typically available in astronomical practice. Hence, we turn our attention now to a novel methodology that overcomes the deficiencies of class interval selection or binning.

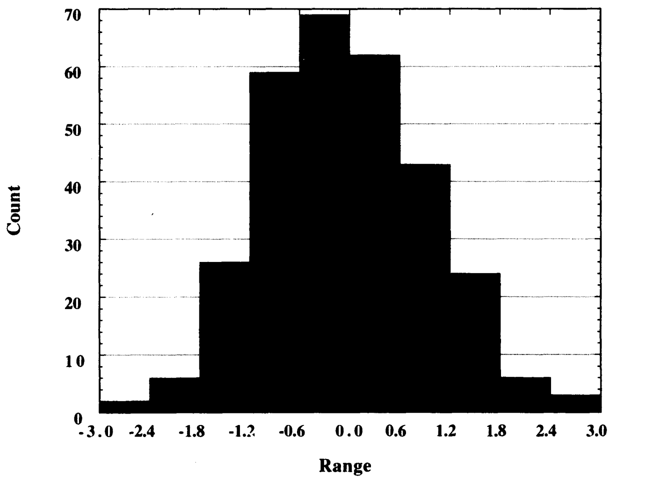

FIG. 1.—Plot of frequency histograms for Gaussian random deviates. 300 data points shown with 10 class intervals (following Sturges's rule).

2.2. *Power Spectrum Analysis*

The recent acquisition of large-redshift survey data has reintroduced the issue of the significance of characteristic scales in the galaxy distribution as addressed by Yu & Peebles (1969). The PSA and its variants have been applied to such data in a wide variety of contexts ranging from scale determination to redshift quantization. Feldman, Kaiser, & Peacock (1993) summarizes some of the recent applications of PSA to three-dimensional redshift surveys in the search for periodicity or non-Gaussian behavior. Here, we explore the methodology as

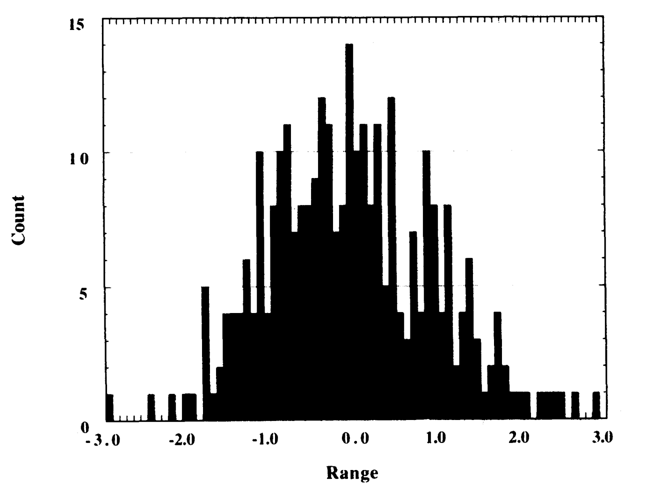

FIG. 2.—Plot of frequency histograms for Gaussian random deviates. 300 data points shown with 72 class intervals (following prevalent astronomical practice).

[4] In the 1992 paper by Newman, Haynes, and Terzian, an editorial oversight resulted in the publication of an earlier draft of the paper which contained some typographical errors. In that paper, we considered a more up-to-date and accurate set of Coma Cluster redshift data and, among other things, employed a Kolmogorov-Smirnov (K-S) test for Normality, acknowledging that this test too could be suspect. When the K-S statistic Q for normality was computed for the updated Coma Cluster data, it was found to be $\approx$0.128, which suggests that there is *a significant possibility of departure from Gaussian behavior*. There are limitations to the K-S statistic; nevertheless, we can be reasonably confident that the Coma Cluster data, if not normally distributed, is *smoothly* distributed.

originally introduced by Yu & Peebles (1969) and point out its limitations.

2.2.1. *Original Definition of Method*

Following Yu & Peebles (1969) (and temporarily employing their equation numbers in square brackets), consider N points x_j distributed in the interval 0 to 2π, and let

$$z_n = N^{-1/2} \sum_{j=1}^{N} e^{inx_j} , \quad [1]$$

where n is an integer. We may regard the z_n as an ancillary series that has the appearance of a Fourier transform. We note that if the measured data x_j were clustered, particularly around uniformly spaced points separated by $\Delta x < 2\pi$, then z_n would be large when $n \approx (\Delta x)^{-1}$. The variable n now has the role of a "frequency" and, in that sense, the power spectrum for the distribution of points is $|z_n|^2, n = 1, 2, \ldots$.

Yu and Peebles go on to say that,

> ... if the points x_j are distributed at random in the interval, the ensemble average of z_n (when $n \neq 0$) is
>
> $$\langle z_n \rangle = N^{-1/2} \sum \langle e^{inx_j} \rangle = N^{-1/2} \sum \int_0^{2\pi} \frac{dx_j}{2\pi} e^{inx_j} = 0 . \quad [2]$$
>
> Similarly, the ensemble average value of the square of the absolute value of z_n is, for a random distribution,
>
> $$\langle |z_n|^2 \rangle = \frac{1}{N} \sum_j \langle |e^{inx_j}|^2 \rangle + \frac{1}{N} \sum_{k \neq j} \langle e^{in(x_k - x_j)} \rangle = 1 . \quad [3]$$
>
> When N is large, and the points are distributed at random, the real and imaginary parts of z_n will have approximately normal distributions. Furthermore, the real and imaginary parts of z_n will be statistically independent. ... Since $|z_n|^2$ is the sum of the squares of two independent variables, each normally distributed, $|z_n|^2$ must have an exponential distribution with the width fixed by equation [3]. Thus we conclude that, when N is large, and the points x_j are distributed at random in the interval $0 \leq x \leq 2\pi$, the a priori probability for finding a value of $|z_n|^2$ greater than x is
>
> $$P(|z_n|^2 > z) = e^{-x} \quad \text{(random)} . \quad [5]$$
>
> It is readily seen ... that the coefficients z_n are statistically independent in the sense that the ensemble average $\langle z_n z_m \rangle$ vanishes when $n \neq m$.

Yu and Peebles then go on to explore what might happen if the data manifested clustering and observed, for the case of "exact" clustering, that is, each point is in a cluster and there are N_c points in a cluster:

$$\langle |z_n|^2 \rangle = N_c . \quad [6]$$

We will return later to the question of the *precision* of their equations [2], [3], and [5], since as we shall show, the outcome of this question has a critical effect on the reliability of hypothesis testing.

It should be noted in passing that, unknown to Yu and Peebles, Bartlett (1963) derived the same approximate formula for the "spectrum" of a point process as a method for identifying clustering. However, Bartlett (1978, § 9.23) observed that this approximation could not be used reliably to estimate the spectrum without an approximate smoothing technique. Put another way, the large excursions in the power spatial amplitude that emerge in this method was known in the statistics community to be artificial and that special measures were necessary to make these effects more tolerable. A variety of methods have emerged in the statistical literature to explore the question of clustering—although none can be identified by the present authors as being particularly relevant to astronomical questions.

Before undertaking a critical analysis of the various approximations that enter into the power spectral analysis method, let us illustrate its performance by using artificially generated data. As before, we employ "gasdev" and "ran2" (Press et al. 1992) and generate 300 samples for two cases. In the first, the data are uniformly distributed from 0 to 2π, while in the second, the data are normally distributed with unit variance and mean of π. In Figure 3, we plot the power spectral amplitude for each. (Note that we use the inverted right-hand scale for the Gaussian distribution-associated data, and employ the usual left-hand scale for the uniform distribution-associated data.) We observe that values of $|z_n|^2$ of 4–6 are not uncommon. (Other simulations we have performed produced excursions in the amplitude of eight or more units.)

Finally, we plot the (natural logarithm of the) corresponding cumulative distribution functions, i.e., $P(|z_n|^2 > x)$. From equation [6] of Yu and Peebles, we expect that the logarithm should vary as $-x$, which we identify as the "ideal situation." Further, we plot the cumulative distribution functions for the uniform data (*solid line*) and the Normal data (*dashed line*)—the figure demonstrates that, as x increases, the error in the corresponding distribution functions systematically increases. It appears that the tail of the distribution functions is in marked disagreement with Yu and Peebles' approximation [6]. We will return shortly to the question of *why* this occurs.

2.2.2. *Extreme Value Statistics*

In light of equation [5] of Yu and Peebles, it seems at face value highly unlikely (probability ≈ 0.002) that excursions of 6 or more in the power spectral amplitude could occur. However, it is essential to note that we are looking at the

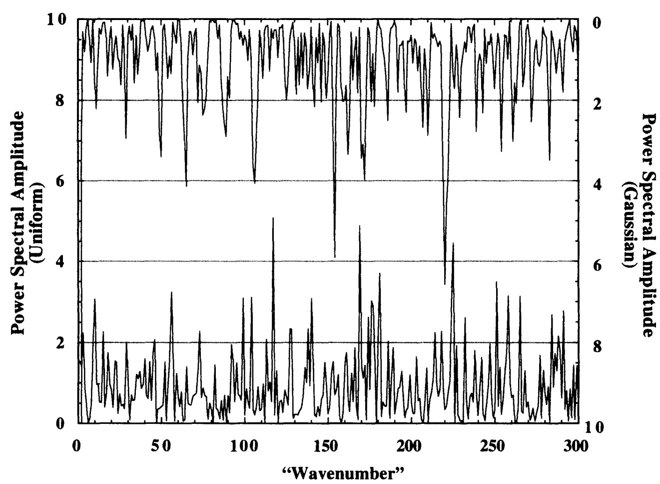

FIG. 3.—Power spectrum analysis plots for uniformly distributed data over the interval [0, 2π] and for Normally distributed data with unit variance and a mean of π; 300 points are generated for each distribution. Note that the second PSA plot, i.e., derived from the Normally distributed data, is inverted and corresponds to the right-hand scale. Also, observe that the first point on this plot is off scale with an amplitude of 106.41.

outcome of a *set* of experiments when we look at such plots, as we are identifying the extremum of a distribution of individual spectral estimates. What is required in looking at the maximum value of the power spectral amplitude or, more conventionally, the maximum value m is that we employ the "statistical theory of extremes" (Galambos 1978). The maximum value that m can obtain from a set of "experiments," i.e., the variation with respect to the reciprocal period n in equation [1] of Yu and Peebles, has a probability distribution that is very different from the statistic P of equation [5]. In particular, if $P(x)$ is the probability that in a given experiment the observed value $\geq x$, then $1 - P(x)$ is the probability that in a given experiment the observed value $\leq x$. Rather than indulge in the subtleties of the theory of the maximum of stochastic processes, we will simply consider that, for a set of N *independent* experiments, the probability that none of the observed results exceeds x is $[1 - P(x)]^N$. Finally, the probability $P_N(x)$ that at least one of the observed N-independent results exceeds x is given by

$$P_N(x) = 1 - [1 - P(x)]^N . \tag{1}$$

(For clarity, we will refer to Yu and Peebles equations in square brackets and refer to our equations in parentheses.) This result is rigorous and general for independent deviates; the definition of $P(x)$ need not be that of equation [5] as derived by Yu and Peebles. For $P(x)$ sufficiently small, it follows that $P_N(x) \approx N \times P(x)$. It is noteworthy that any error in the estimation of $P(x)$ will be amplified by orders of magnitude. See Scott (1991) for a discussion of this equation and its history in an astronomical context.

As a simple illustration of how different extreme value statistics can be from Normal statistics, consider a Gaussian random with zero mean and unit variance. We are well familiar with the fact that 95% of the values realized in this process should lie between -1.96 and 1.96 units, and that 99% of the values should lie between -2.58 and 2.58 units. Similarly, for a single realization of this process, 95% of all values should lie *below* 1.65 units and 99% of all values should lie *below* 2.33 units. (The numbers in this latter situation are of course different, since half of all realized values are negative and we are considering only an *upper* cut off or maximum value.) Suppose now that we looked at the maximum value of N experiments and that these experiments are statistically independent. For a set of 100 experiments, we would expect a largest value of 3.28 with 5% likelihood, and 3.72 with 1% likelihood (Pearson & Hartley 1962, Table 24). Meanwhile, for a set of 1000 experiments, we would expect a largest value of 3.88 with 5% likelihood, and 4.26 with 1% likelihood. The purpose of this demonstration is to show just how different extreme value behavior is from Normal behavior.

It is difficult to assess the number of statistically independent realizations that are present in Figure 4—see Scott (1991)—but in principle we might expect that there are at most 300 (the number of data in the original set). Employing Yu and Peebles' equation [5] as though it were exact, we directly obtain the result that there is a 53% likelihood of seeing at least one excursion in excess of 6 in Figure 4. Therefore, it should come as no surprise that the PSA plot provides excursions as large as those we have provided.

There are, however, examples in the literature of PSA plots where the excursions are very large indeed. As an illustration, Guthrie & Napier (1990) have power spectral amplitudes as great as 20. Could this be the result of a subtle imprint of the approximations employed by Yu and Peebles in deriving the exponential distribution [5]?

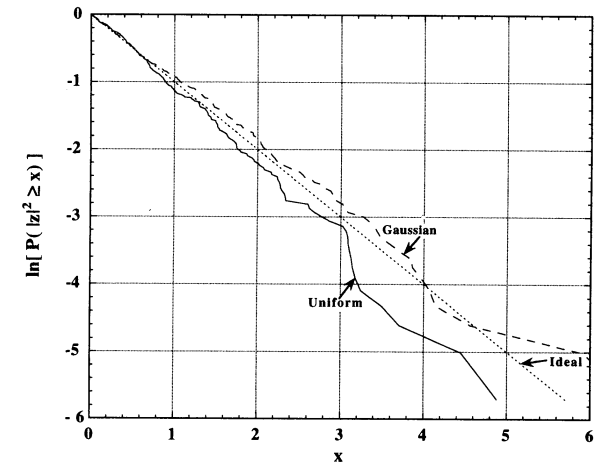

FIG. 4.—Cumulative distribution functions for the probability that an observed value of the Yu & Peebles (1969) statistic $|z|^2$ is greater than x. The ideal case should vary as $\exp(-x)$—and is depicted by a dotted line. Solid and dashed lines are used to show uniformly and Normally distributed cases generated from the previous figure. Observe the systematic increase in the departure from the ideal case—a measure of the bias and the slow asymptotic convergence rate for the Yu and Peebles estimator.

2.2.3. *PSA Approximation Revisited*

In Yu and Peebles equation [2], reprinted earlier, we note that they had implicitly employed a distribution function for x_j that is uniformly distributed over the interval $[0, 2\pi]$ and, as a consequence, their claim of a vanishing mean in z_n for $n \neq 0$ is a consequence of this assumption. (We suspect that they believed their result to be generally correct as a consequence of a "random phase" approximation.) Suppose that x_j is distributed according to some distribution function $\mathscr{P}(x)$ (not necessarily uniform) on a doubly infinite domain. Then, it follows that

$$\langle z_n \rangle = N^{-1/2} \sum \langle e^{inx_j} \rangle$$
$$= N^{1/2} \int_{-\infty}^{\infty} e^{inx}\, d\mathscr{P}(x) \tag{2}$$
$$= N^{1/2} \mathscr{F}(n) , \tag{3}$$

where $\mathscr{F}(n)$ is the *characteristic function* of $\mathscr{P}(x)$ defined by

$$\mathscr{F}(n) = \langle e^{inx} \rangle = \int_{-\infty}^{\infty} e^{inx}\, d\mathscr{P}(x) . \tag{4}$$

Expressed another way, the characteristic function is the Fourier transform of the probability distribution for x, hence the name "power spectra analysis" method. (In order to extend the range of integration from that in [2], it is useful to think of the Yu and Peebles distribution function as having compact support, i.e., it vanishes identically outside of the interval from 0 to 2π.) As a relevant example owing to the observed apparent Normality of data from many different clusters, suppose $\mathscr{P}(x)$ is a Gaussian distribution $\mathscr{N}(\mu, \sigma^2)$ with mean μ and variance σ^2, i.e.,

$$\mathscr{P}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x' - \mu)^2}{2\sigma^2} \right] dx' . \tag{5}$$

Then, the characteristic function $\mathscr{F}(n)$ becomes

$$\mathscr{F}(n) = e^{\imath n\mu} e^{-n^2\sigma^2/2} \tag{6}$$

and, therefore,

$$\langle z_n \rangle = N^{1/2} e^{\imath n\mu} e^{-n^2\sigma^2/2} \neq 0 \tag{7}$$

for our Normally distributed example. (This result also explains the off-scale power spectral amplitude generated for the Gaussian case in Figure 3, i.e., $300 \times e^{-1} = 110.363832351 \approx 106.41$.)

In a similar way, it follows that equation [3] of Yu and Peebles for the variance of z_n should generally be

$$\begin{aligned} \langle |z_n|^2 \rangle &= \frac{1}{N} \sum_j \langle |e^{\imath n x_j}|^2 \rangle + \frac{1}{N} \sum_{j \neq k} \langle e^{\imath n(x_j - x_j)} \rangle \\ &= 1 + (N-1)\mathscr{F}(n)\mathscr{F}^*(n) \geq 1 \end{aligned} \tag{8}$$

and, for our Normally distributed example,

$$\langle |z_n|^2 \rangle = 1 + (N-1)e^{-n^2\sigma^2} . \tag{9}$$

Our equations (3) and (8) show that equations [2] and [3] of Yu and Peebles are valid only for a set of measure zero, the uniform distribution function (which is rarely a good approximation for astronomical data). Furthermore, our equations (7) and (9) show that the random variables z_n are not stationary with respect to n. Moreover, we observe that, for $n \neq m$,

$$\begin{aligned} \langle z_n z_m^* \rangle &= \frac{1}{N} \sum_{j,k=1}^{N} \langle e^{\imath(nx_j - mx_k)} \rangle \\ &= \frac{1}{N} \sum_j \langle e^{\imath(n-m)x_j} \rangle \\ &\quad + \frac{1}{N} \sum_{j \neq k} \langle e^{\imath(nx_j - mx_k)} \rangle \\ &= \mathscr{F}(n-m) + (N-1)\mathscr{F}(n)\mathscr{F}^*(m) \\ &\neq 0 . \end{aligned} \tag{10}$$

This result reduces to equation (8) in the case $n = m$. By inspection we note that the correlation matrix is Toeplitz, and only for a set of measure zero does it reduce to a diagonal one. Thus, we see that the ancillary variables z_n are not individual independent deviates but are correlated and hence are not stationary. This underscores the difficulty in estimating the number of independent random variables N in a power spectrum analysis plot. It is in principle possible to employ this bias to obtain evidence of clustering, a feature anticipated by Yu and Peebles (their eq. [6]). Now we must explore the role of the Central Limit Theorem in this discussion.

2.2.4. *Central Limit Theorem and Distribution Functions*

In particular, the Central Limit Theorem (Dudley 1989; Feller 1968) states that if $s_i \in \boldsymbol{R}$, $i = 1, \ldots, N$ are individual independent deviates drawn from the same population and if their mean μ and variance σ^2 exist, i.e., $|\mu|, \sigma < \infty$, then the variable

$$\mathscr{S} \equiv \sum_{i=1}^{N} \left[\frac{s_i - \mu}{\sqrt{N}\sigma}\right] \tag{11}$$

describes, in the limit $N \to \infty$, a Gaussian zero mean, unit variance process $\mathscr{N}(0, 1)$. Counterexamples to the Central Limit Theorem (Romano & Siegel 1986) do occur, but generally involve some violation of the assumed conditions.

Special attention must be paid to the fact that z_n, as defined in equation (1), is complex valued, and that the properties of both the real and the imaginary parts of z_n must be considered separately. It is easy to show that

$$\begin{aligned} \mu &\equiv \langle z_n \rangle \\ &= \langle \mathscr{R}(z_n) \rangle + \imath \langle \mathscr{I}(z_n) \rangle \\ &= \mu_R + \imath \mu_I \end{aligned} \tag{12}$$

where μ is defined as the population mean of z_n and where μ_R and μ_I denote the real and imaginary parts of μ, respectively. Additionally, we obtain

$$\begin{aligned} \sigma^2 &\equiv \langle |z_n - \mu|^2 \rangle \\ &= \langle |\mathscr{R}(z_n - \mu)|^2 \rangle + \langle |\mathscr{I}(z_n - \mu)|^2 \rangle \\ &= \sigma_R^2 + \sigma_I^2 \end{aligned} \tag{13}$$

where σ^2 is defined as the variance of z_n defined by $\langle |z_n - \mu|^2 \rangle$ and where σ_R^2 and σ_I^2 are the variances of the real and imaginary parts of z_n, respectively. However, it is now easy to show that the Central Limit Theorem can be employed to each of the components of z_n. The preceding results (7) and (8) show that conditions *necessary* for the Central Limit Theorem to hold with respect to equation [1] of Yu and Peebles are not satisfied until n is (asymptotically) large, and equation [1] results in a poor approximation to a Gaussian zero-mean, unit variance process. The one outstanding issue that remains is to determine the distribution function for the real, semi-positive definite variable $|z_n|^2$, as well as to consider issues of accuracy.

Consider now, for any n, the decomposition of z_n into its real and imaginary parts $\mathscr{X}$ and $\mathscr{Y}$, respectively. Let us now define a variable $\mathscr{W}'$ according to

$$\mathscr{W}' = \mathscr{X}^2 + \mathscr{Y}^2 . \tag{14}$$

It should be clear that $\mathscr{W}'$ has the role of $|z_n|^2$. From the preceding discussion of the Central Limit Theorem, we know that the distribution functions for $\mathscr{X}$ and $\mathscr{Y}$ are approximately Gaussian but, for the present time, will assume that they are completely arbitrary, say $\mathscr{P}_x(\mathscr{X})$ and $\mathscr{P}_y(\mathscr{Y})$, respectively. It follows that the characteristic function for $\mathscr{P}_w(\mathscr{W}')$ must satisfy

$$\int e^{\imath k \mathscr{W}'} d\mathscr{P}_w(\mathscr{W}') = \int e^{\imath k(\mathscr{X}^2 + \mathscr{Y}^2)} d\mathscr{P}_x(\mathscr{X})\mathscr{P}_y(\mathscr{Y}) . \tag{15}$$

We multiply both sides of the latter equation by $(1/2\pi)e^{-\imath k \mathscr{W}''}$, and integrate over k from $-\infty$ to ∞. We then integrate over $\mathscr{W}''$ from $\mathscr{W}$ to ∞ to obtain

$$1 - \mathscr{P}_w(\mathscr{W}) = \int \Theta(x^2 + y^2 - \mathscr{W}) d\mathscr{P}_x(\mathscr{X}) d\mathscr{P}_y(\mathscr{Y}) . \tag{16}$$

This equation has a direct intuitive interpretation. The left-hand side represents the probability that $|z_n|^2$ is greater than $\mathscr{W}$, namely $P(|z_n|^2 > \mathscr{W})$ as in equation [5] of Yu and Peebles, while the right-hand side represents the contribution to the probability from all points $\mathscr{X} + \imath\mathscr{Y}$ such that $\mathscr{X}^2 + \mathscr{Y}^2 > \mathscr{W}$. (This is the basis of the "Box-Muller" scheme employed by Press et al. (1992) in their Gaussian random number generator "gasdev.") As an illustrative example, suppose that both $\mathscr{X}$ and

$\mathscr{Y}$ are $\mathscr{N}(0, 1/2)$. The right-hand side then becomes

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{\pi} \Theta(\mathscr{X}^2 + \mathscr{Y}^2 - \mathscr{W}) e^{-(\mathscr{X}^2+\mathscr{Y}^2)} d\mathscr{X}\, d\mathscr{Y} = \int_{r=\sqrt{\mathscr{W}}}^{\infty}\int_{\theta=0}^{2\pi} \frac{1}{\pi} e^{-r^2} r\, dr\, d\theta \quad (17)$$

after making the usual transformation to polar coordinates and then becomes

$$\int_{\sqrt{\mathscr{W}}}^{\infty} d(r^2) e^{-r^2} = e^{-\mathscr{W}} ; \quad (18)$$

combining the left and right sides, we have verified equation [5] of Yu and Peebles. However, our general result (16) will become important later in the discussion. (P. K. Sen discussed the nature of the true distribution for $|z_n|^2$ in Newman et al. 1992.)

2.2.5. *Consequences of Slow Convergence Rates*

Suppose that n is sufficiently great that these concerns regarding the biased estimate of the mean and variance in equations (7) and (8) are not realized. Then, the real and imaginary parts of $\lim_{n\to\infty} z_n$ must become Gaussian zero-mean one-half unit-variance random variables. Then, the mean and variance of the complex variable z_n, as defined in equations (12) and (13), become zero and unity, respectively. The operative question then becomes how *rapidly* does z_n converge to a Gaussian random variable as a function of large N? The formal answer, in the sense of probability (see Dudley 1989, for example), is that the convergence rate goes as $N^{-1/2}$. The practical answer is that convergence is reasonably rapid near the center of the distribution, but is manifestly slower in the tail, depending on the underlying distribution for s_i in equation (11).

As a concrete example of this phenomenon, consider the "Sum of Uniform Deviates" method described in Abramowitz & Stegun (1965) for generating Gaussian random numbers. Extracting from Abramowitz and Stegun, let $U_1, U_2, \ldots, U_n$ be a sequence of n uniform deviates. Then

$$X_n = \left(\sum_{i=1}^{n} U_i - \frac{n}{2}\right)\left(\frac{n}{12}\right)^{-1/2}$$

will be distributed asymptotically as a normal random deviate. When $n = 12$, the maximum errors made in the normal deviate are 0.009 for $|X| < 2$, 0.9 for $2 < |X| < 3$. In other words, random numbers generated in this way that are between two and three standard deviations from the mean can have errors nearly as large as one standard deviation! The error in the normal deviate will be greater still if it is further than three standard deviations from the mean. There is a possibly apocryphal story that IBM used exactly this scheme for generating Gaussian random numbers on their mainframe computers many years ago, but later abandoned it owing to the anomalous behavior in the tail of the generated distribution.

Therein lies the crux of the problem. Hypothesis testing using the method of power spectrum analysis is of particular interest only as $|z_n|^2$ becomes large, i.e., when we are in the tail of the distribution. Owing to the form of equation (16), it follows that the calculation of $P(|z_n|^2 > \mathscr{W})$ depends critically and in a nonlinear way upon the tail of the distribution functions $\mathscr{P}_x(\mathscr{X})$ and $\mathscr{P}_y(\mathscr{Y})$, respectively. Empirically, we observe that the errors in the associated probability density functions for $\mathscr{X}$ and $\mathscr{Y}$ can be as much as an order of magnitude. This error is further amplified by equation (16) in the calculation of $P(|z_n|^2 > \mathscr{W})$, and the relative error between the true distribution and the assumed exponential one can readily exceed an order of magnitude. For example, the data displayed in Figure 4 are manifestly *not* exponentially distributed, and it is in fact easy to produce a random set of data with "spectral peaks" that are very large. Finally, when we apply the extreme value statistic $1 - [1 - P(x)]^N$ in equation (1) to estimate the likelihood that a peak exceeds some threshold, the typically large value of N then dramatically amplifies this error.

In addition to the applications of the PSA method by Guthrie & Napier (1990) and by Broadhurst et al. (1990), Scott (1991) investigated the possible periodicity in quasar redshifts. The claimed periodicity is with respect to $\ln(1 + z) \approx 0.205$, where z is the quasar redshift, and application of the PSA method by Scott revealed no such periodicity from the available data. Previous statistical studies employing various binning methods had indicated a variety of possible periodicities, and Burbidge (1967) in particular had pointed out a peak at $z \approx 1.96$ in the quasar redshift distribution and interpreted this as evidence that z, and hence the quasars, were not at cosmological distances. Scott also suggests that the one-dimensional power spectrum analysis method, when applied to real data, possess some degree of subjectivity in assessing the significance level of any effect.

3. CONCLUSION

We have reviewed two classes of statistical methods in common use today by astronomers seeking to identify periodicities or anomalies in their data. The first methodology, a strictly descriptive one, employs frequency histograms to obtain a sense of the distribution of the variable at hand, usually the redshift velocity. The essential problem here is the selection of the frequency class interval or "bin size." Mathematicians have developed ad hoc rules for this purpose based upon the need to minimize unrealistic fluctuations in the estimate of the distribution function, a goal which is shared by astronomers who need to discern genuine patterns from statistical noise. Generally speaking the methodology developed here introduces class interval sizes that are nearly an order of magnitude larger than those employed by many astronomers in the recent past—this leads us to conclude that any pattern evinced in many contemporary astronomical frequency histograms is likely to be an artifact of the small numbers of data present in a class interval and is, therefore, illusory—see Newman et al. (1989). Moreover, the restriction posed by these class interval selection rules basically exclude the possibility of seeing any pattern within a frequency histogram. Thus, it becomes necessary to develop a class of methods that is not susceptible to bin size selection effects.

The second methodology, due to Yu & Pebbles (1969), makes significant progress in this direction. Although its basis is analytic, it contains the influence of a number of approximations. Certain of these deleterious effects can be effectively excluded, e.g., the extreme value statistic correctly estimates probabilities for the set of values while the effect of nonstationarity and bias in the estimators is avoided by excluding the first few points of the spectral amplitude, following our equations (3) and (10)—although this could potentially introduce some selection effects. However, the method is based on an *asymptotic* convergence property which is very slow indeed, varying only as $N^{-1/2}$, where N is the length of the observational data set. Thus, even hundreds of data points do not necessarily alleviate this effect. In particular, the tail of the distribution generally will be different (by, possibly, an order of magnitude) from the ideal exponential distribution, and this error is amplified dramatically by the computation of the extreme value statistic. Hence, the ability to do hypothesis testing using this asymptotic result is severely compromised. It must be stressed, however, that while the test may be inconclusive, the original hypothesis need not be false.

Statistical methods are predicated upon the ability to *exclude* hypotheses—they can prove nothing, but disprove anything. And that is the essence of the dilemma we face. It is easy to show that data are *not*, for example, Normally distributed. However, it is difficult to show that data are not "smooth"—because that notion cannot be formulated in a simultaneously mathematically and physically realistic way—which is what we need, for example, to establish clustering. Similarly it is not simple to show that data contain periodicities, unless they too can be defined in an appropriate way, and then we encounter again a very difficult task. The method of Yu and Peebles was an important step in that direction. Unfortunately, the accumulated effect of the approximations built into it renders it incapable of *testing the outliers in the random variable that they have constructed.* This problem will be remedied in small part by the acquisition of more accurate and more abundant data. Nevertheless, the interpretation of any experiment is only as reliable as the statistical methods employed—hence, a qualitative refinement of the statistical tools employed is also necessary.

We wish to thank D. Ylvisaker and H. Jerrison for some useful discussions. One of us (W. I. N.) was partly supported by NASA grant NAGW 3132 to the University of California, Los Angeles and by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University, while M. P. H. was partly supported by NSF grant AST-9023450. W. I. N. wishes to thank the Department of Astronomy at Cornell University for its hospitality where part of this research was performed.

REFERENCES

Abramowitz, M., & Stegun, I. A. 1965, Handbook of Mathematical Functions (New York: Dover)
Bahcall, N. A. 1991, ApJ, 376, 43
Barrow, J. S., & Bhavsar, S. P. 1987, QJRAS, 28, 109
Bartlett, M. S. 1963, J. R. Statist. Soc., B25, 264
———. 1978, An Introduction to Stochastic Processes, 3d ed. (Cambridge: Cambridge Univ. Press)
Broadhurst, T. J., Ellis, R. S., Koo, D. C., & Szalay, A. S. 1990, Nature, 343, 726
Burbidge, G. R. 1967, ApJ, 147, 845
Burbidge, G. R., & O'dell, S. L. 1972, ApJ, 178, 583
Chambers, J. M. 1977, Computational Methods for Data Analysis (New York: Wiley)
Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. 1983, Graphical Methods for Data Analysis (Boston: Duxbury Press)
Cocke, W. J. 1992, ApJ, 393, 59
Dixon, W. J., & Kronmal, R. A. 1965, J. Assoc. Comput. Mach., 12, 259
Doane, D. P. 1976, Amer. Statistician, 30, 181
Dudley, R. M. 1989, Real Analysis and Probability (Pacific Grove, CA: Wadsworth & Brooks/Cole)
Feldman, H. A., Kaiser, N., & Peacock, J. 1993, preprint
Feller, W. 1968, An Introduction to Probability Theory and Its Applications, Vol. 1 (New York: Wiley)
Freedman, D., & Diaconis, P. 1981a, Z. Wahrscheinlichkeitstheorie verw. Gebiete, 57, 453
———. 1981b, Z. Wahrscheinlichkeitstheorie verw. Gebiete, 58, 139
Fukunaga, K. 1990, Introduction to Statistical Pattern Recognition, 2d ed. (Boston: Academic Press Inc.)
Galambos, J. 1978, The Asymptotic Theory of Extreme Order Statistics (New York: Wiley)
Gentleman, J. F. 1977, Amer. Statistician, 31, 166
Guthrie, B. N. G., & Napier, W. M. 1990, MNRAS, 243, 431
Hald, A. 1952, Statistical Theory with Engineering Applications (New York: Wiley)
Hoaglin, D. C., Mosteller, F., & Tukey, J. W., eds. 1985, Exploring Data Tables, Trends, and Shapes (New York: Wiley)
Huff, D. 1954, How to Lie with Statistics (New York: W. W. Norton & Company)
Julesz, B. 1981, Nature, 290, 91
Kaiser, N., & Peacock, J. A. 1991, ApJ, 379, 482
Kimble, G. A. 1978, How to Use (and Misuse) Statistics (Englewood Cliffs, NJ: Prentice-Hall Inc.)
Manchester, R. N., & Taylor, J. H. 1977, Pulsars (San Francisco: Freeman)
Newman, W. I., Haynes, M. P., & Terzian, Y. 1989, ApJ, 344, 111
———. 1992, in Statistical Challenges in Modern Astronomy, ed. E. D. Feigelson & G. J. Babu (New York: Springer), 137
Pearson, E. S. 1936, Biometrika, 28, 306
Pearson, E. S., & Hartley, H. O. 1962, Biometrika Tables for Statisticians (Cambridge: Cambridge Univ. Press)
Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, Numerical Recipes, 2d, ed. (Cambridge: Cambridge Univ. Press)
Romano, J. P., & Siegel, A. F. 1986, Counterexamples in Probability and Statistics (Monterey, CA: Wadsworth & Brooks/Cole)
Runyon, R. P. 1981, How Numbers Lie: A Consumer's Guide to the Fine Art of Numerical Deception (Lexington: The Lewis Publishing Company)
Schneider, S. E., & Salpeter, E. E. 1992, ApJ, 385, 32
Schmid, C. F. 1983, Statistical Graphics: Design Principles and Practices (New York: Wiley)
Scott, D. W. 1979, Biometrika, 66, 605
———. 1991, A&A, 242, 1
Sharp, N. A. 1984, ApJ, 286, 437
———. 1990, ApJ, 354, 418
Sheehan, W. 1988, Planets and Perception: Telescopic Views and Interpretation (Tucson: Univ. Arizona Press)
Sturges, H. A. 1926, J. Am. Stat. Assoc., 21, 65
Tifft, W. G. 1973, ApJ, 179, 29
Tufte, E. R. 1983, The Visual Display of Quantitative Information (Cheshire, CT: Graphics Press)
Tukey, J. W. 1977, Exploratory Data Analysis (Reading, MA: Addison-Wesley)
Wilk, M. B., & Gnanadesikan, R. 1968, Biometrika, 55, 1
Yu, J. T., & Peebles, P. J. E. 1969, ApJ, 158, 103