Mon. Not. R. astr. Soc. (1984) **210**, *Short Communication*, 19P–23P

A bootstrap resampling analysis of galaxy clustering

John D. Barrow, Suketu P. Bhavsar* and
D. H. Sonoda *Astronomy Centre, University of Sussex, Falmer, Brighton BN1 9QH*

Accepted 1984 July 4. Received 1984 July 4; in original form 1984 June 22

Summary. We demonstrate how statistics and standard errors can be associated with parameters of single data sets using bootstrap resampling methods. These techniques allow significance levels to be associated with any parameter derived from a data set and measure the robustness of the data set to various known and unknown errors and biases. We apply the method to the two-point angular correlation function of the Zwicky 14 mag catalogue of galaxies. If the two-point function has the power-law form $\omega(\theta) = (\theta_0/\theta)^\gamma$ then standard errors of $\sigma(\theta_0) = 0.01$ and $\sigma(\gamma) = 0.13$ are found with means of $\bar{\gamma} = 0.80$ and $\bar{\theta}_0 = 0.06$ radians. These are much larger than the formal errors quoted in previous analyses. Various consequences of these results and further applications of the method employed are discussed.

1 Introduction

In this communication we illustrate the use of the bootstrap resampling technique for determining standard errors. This is done by applying it to the calculation of the galaxy–galaxy correlation function in the Zwicky catalogue (Zwicky *et al.* 1961–68). Besides giving specific significance levels for the parameters of the calculated two-point correlation function we aim to display how this simple technique can be useful in assigning significance levels to single parameters derived from non-repeatable observations.

2 Bootstrap resampling method

The bootstrap method (Efron 1979, 1982; Freedman 1981; Rey 1980; Singh 1981) begins with a particular data set, in our case a galaxy catalogue, and creates from it a large number of comparable 'pseudo-data' sets. Any property of the original data set can now be calculated for all the pseudo-data sets and the resulting mean and variance calculated over the

* Raman Research Institute, Bangalore 560 080, India.

entire ensemble.* A typical pseudo-data set might, for example, be related to the true data set by replacing a single point of the true data set by a random point. This resampling procedure could be repeated many times to generate an entire ensemble of pseudo-data sets, each differing from the true data by one point. By this means the sensitivity of any particular analysis or conclusion regarding the structure of the true data set can be assessed internally. In effect, the resampling method allows us to detect the robustness of any calculated or measured property of the true data set with respect to a variety of known or unknown sources of error or biases and uncertainties in the observational method. This technique is likely to be of importance as a means of assessing the significance of various geometrical statistics, like percolation, that have been used to measure the presence or absence of filamentary patterns in the redshift surveys of galaxies (Bhavsar & Barrow 1983; Dekkel & West 1984; Einasto *et al.* 1982).

We apply this approach to the Zwicky 14 mag catalogue of galaxies. We choose all galaxies in the region $\delta > 0°$, $b > 40°$; this consists of 1091 galaxies within a region of 1.83 steradians. The resolution limit of the Zwicky catalogue is 0°.025 and this defines a lower limit for the extent to which we can determine the two-point angular correlation function (Limber 1953). The Zwicky catalogue lists galaxies to a magnitude of 15.7. There has been some debate concerning the accuracy of Zwicky's magnitudes (Shane 1975) but they should be accurate enough to define a cut-off at 14.0 mag. The two-point angular correlation function, $\omega(\theta)$, is determined in the usual way (Totsuji & Kihara 1969; Fall 1979; Peebles 1980): the angular separation between pairs of observed galaxies is found and binned to give $N_{oo}(\theta)$, the number of pairs observed to be separated by θ. A Poisson sample of 1090 galaxies is then generated with $\delta > 0°$, $b > 40°$ and the pairwise separations between each of the 1091 real galaxies and 1090 Poisson galaxies are found; this gives $N_{op}(\theta)$, the number of pairwise separations at angle θ between observed and Poisson galaxies. The correlation function, $\omega(\theta)$, is just

$$\omega(\theta) = \frac{N_{oo}(\theta) - N_{op}(\theta)}{N_{op}(\theta)} . \tag{1}$$

In our calculations $N_{op}(\theta)$ and $N_{oo}(\theta)$ are binned so that a Poisson data set is expected to have an equal number of galaxy pairs in each bin (that is, the binning procedure is decided upon before performing the analysis). The maximum angular separation binned is 0.12 radians and the minimum separation is then fixed by the binning assumption. We find for the Zwicky data set:

$$\omega(\theta) = \left(\frac{0.06}{\theta}\right)^{0.77} . \tag{2}$$

This can be compared with other determinations, using a variety of samples, by Totsuji & Kihara (1969), Peebles (1974), Davis & Geller (1973), Davis, Geller & Hochra (1978) and Peebles (1980).

In order to place a significance level on the determination (2) we generate 180 pseudo-data sets from the original Zwicky catalogue of 1091 galaxies. Each pseudo-data set is con-

* Note, this is not a technique for eliminating errors. It simply gives an estimate of the internal variance *given* the original data set. It says nothing about how large or small the errors may be or what their source might be. Traditional approaches to data sets differ from this by assuming that errors are normally distributed by invoking the Central Limit Theorem.

structed by selecting randomly, *with replacements*,† 1091 galaxies from the original catalogue. Therefore, each of the 180 pseudo-data sets will contain duplicates of some galaxies and will not contain all the galaxies in the original Zwicky sample. The angular correlation function, $\omega(\theta)$, is calculated for each of the 180 pseudo-data sets and fitted by a power law of the form

$$\omega(\theta) = \left(\frac{\theta_0}{\theta}\right)^{\gamma}. \tag{3}$$

In Fig. 1(a) and (b) we display the frequency distribution of the constants γ and θ_0 which arise from the 180 determinations. The means and standard errors are found to be

$$\bar{\gamma} = 0.80, \qquad \sigma(\gamma) = 0.13 \tag{4}$$

$$\bar{\theta}_0 = 0.06, \qquad \sigma(\theta_0) = 0.01. \tag{5}$$

The bootstrap technique is only designed to give estimators for $\sigma(\theta_0)$ and $\sigma(\gamma)$; the mean values, $\bar{\theta}_0$ and $\bar{\gamma}$, are not expected to be good estimators of the true θ_0 and γ values.

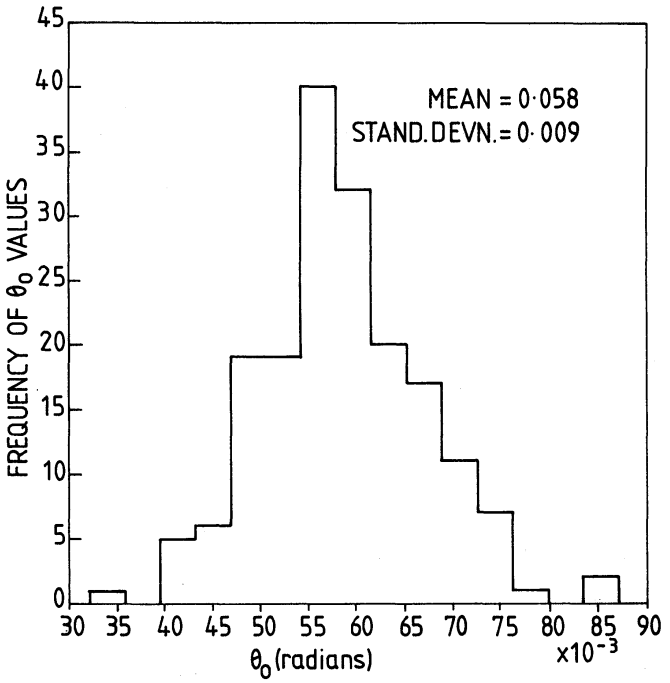
It is instructive to compare these errors with the much smaller formal errors quoted in many discussions of the observed two-point correlations (Groth & Peebles 1977; Davis & Peebles 1983). Also, we note the results of numerical simulations of gravitational clustering using 4000 mass points by Gott, Turner & Aarseth (1979) who found a standard error of $\sigma(\gamma) \simeq 0.15$ arising from different realizations of initial data with the same statistical distribution and total density. Also they found in their numerical experiments a variation of γ with cosmological density parameter Ω_0 and the power index of the initial density inhomogeneities, n, where the density contrast $\delta\rho/\rho$ varies with mass scale as $\delta\rho/\rho \propto M^{-1/2-n/6}$, (Fall 1979). This is fitted by

$$\Delta\gamma = -0.55\,\Delta(\log \Omega_0) + 0.35\,\Delta n. \tag{6}$$

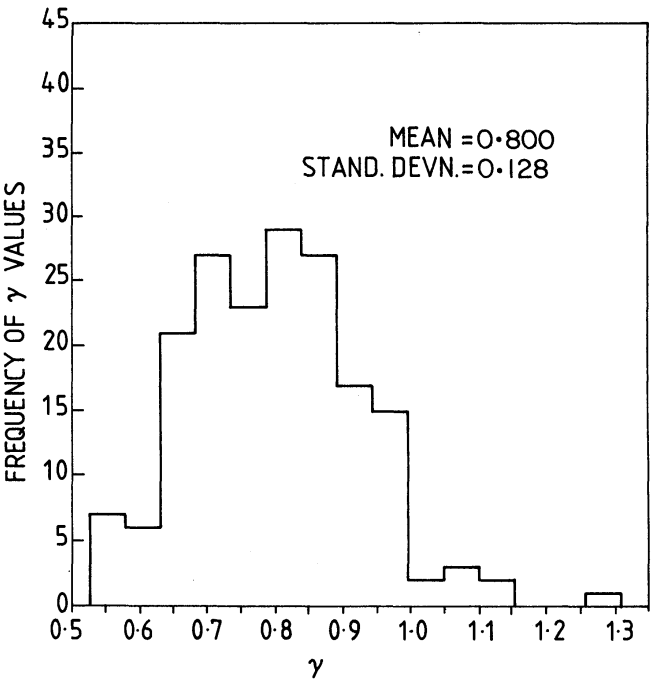
Since the ranges of interest for the parameters are approximately $\Omega_0 \epsilon (10^{-2}, 1)$ and $n \epsilon (-1, 1)$ it is clear from (4) and (5) that even if superior numerical experiments could reduce the experimental scatter around $\bar{\gamma}$, the results (4) and (5) show that the data is unlikely to be robust enough to allow Ω_0 or n to be accurately pinned down by simple determinations of $\omega(\theta)$ (although there may of course be other ways of ascertaining these parameters from the data that are more robust).

The variance generated by the pseudo-data analysis may have other interesting implications. Saslaw (1980) and Saslaw & Hamilton (1984) have developed a description of gravitational clustering based upon a thermodynamic treatment of two-point correlations. This theory predicts that as $t \to \infty$ a system undergoing gravitational clustering should relax to a state of maximal entropy production and clustering efficiency. One of the characteristics of this asymptotic state is that it possesses a two-point correlation function $\omega(\theta) \propto \theta^{-1}$. Despite the disagreement of this prediction with the straightforward determination (2), it may be marginally admissible in view of the standard errors produced by the bootstrap resampling.

† The question of the optimal procedure to follow in generating pseudo-data sets is a topic of detailed study (Rey 1980; Rocke & Downs 1981). We stress that the replacement method we use makes no assumption about the existence, magnitude or source of any errors. If specific sources of systematic error are suspected then they should be tested for with a purpose-built analysis. Our perturbation of the true data to generate a pseudo-data set is moderate and we would expect other more extensive replacement or substitution procedures to produce at least this level of internal variance.



(a)



(b)

Figure 1. (a) Frequency distribution of θ_0 defined by equation (3) found from 180 bootstrap resamplings of the 1091 galaxies in the Zwicky 14 mag catalogue. (b) As (a) but displaying the frequency distribution of γ defined by equation (3).

It is possible that the non-observance of $\omega(\theta) \propto \theta^{-1}$ may be associated with large scale obscuration (Peebles 1980; Seldner & Uson 1983) but this requires the intervening material to possess a particular clustering pattern. Also, the Universe may not yet have relaxed sufficiently close to the predicted asymptotic equilibrium state of maximum entropy to display $\omega(\theta) \propto \theta^{-1}$.

3 Conclusion

In conclusion, we have shown how statistics and standard errors can be generated from single data sets using the bootstrap resampling method. This technique has been applied to the calculation of the two-point correlation function of galaxies in the Zwicky 14 mag catalogue as an example. Various consequences of the results were discussed. These techniques may be essential if any quantitative significance is to be attached to the variety of geometrical statistics currently being employed in the search for a quantitative measure of intrinsic pattern in the three-dimensional distribution of galaxies.

Acknowledgments

D. H. Sonoda was supported by an SERC postgraduate studentship and S. Bhavsar by an SERC visiting fellowship whilst this work was performed. The authors would like to thank R. Ellis, C. Frenk and W. Saslaw for helpful discussions.

References

Bhavsar, S. P. & Barrow, J. D., 1983. *Mon. Not. R. astr. Soc.*, **205**, 61P.
Davis, M. & Geller, M. J., 1973. *Astrophys. J.*, **208**, 13.
Davis, M., Geller, M. J. & Huchra, J., 1978. *Astrophys. J.*, **221**, 1.
Davis, M. & Peebles, P. J. E., 1983. *Astrophys. J.*, **267**, 465.
Dekkel, A. & West, M. J., 1984. *Astrophys. J.*, in press.
Efron, B., 1979. *Ann. Stat.*, 7, 1.
Efron, B., 1982. *CBMS-NSF Regional Conf. Ser. Appl. Math. Monograph 38,* SIAM, Philadelphia.
Einasto, J., Klypin, A. A., Saar, E. & Shandarin, S. F., 1982. *Mon. Not. R. astr. Soc.*, **206**, 529.
Fall, S. M., 1979. *Rev. Mod. Phys.*, **51**, 21.
Freedman, D. A., 1981. *Ann. Stat.*, 9, 1218.
Gott, J. R., Turner, E. L. & Aarseth, S. J., 1979. *Astrophys. J.*, **234**, 13.
Groth, E. J. & Peebles, P. J. E., 1977. *Astrophys. J.*, **217**, 385.
Limber, D. N., 1953. *Astrophys. J.*, **117**, 134.
Peebles, P. J. E., 1974. *Astrophys. J.*, **189**, L251.
Peebles, P. J. E., 1980. *The Large Scale Structure of the Universe,* Princeton University Press, New Jersey.
Rey, W., 1980. *Introduction to Robust and Quasi-Robust Statistical Methods,* Springer Verlag, Berlin.
Rocke, D. M. & Downs, G. W., 1981. *Commun. Stat. Sim. Comput.* **B-10**, 221.
Saslaw, W. C., 1980. *Astrophys. J.*, **235**, 299.
Saslaw, W. C. & Hamilton, A. J. S., 1984. *Astrophys. J.*, **276**, 13.
Seldner, M. & Uson, J. M., 1983. *Astrophys. J.*, **264**, 1.
Shane, C. D., 1975. *Stars and Stellar Systems IX,* ed. Sandage, A., Sandage, M. & Kristian, J., University of Chicago Press, Chicago.
Singh, K., 1981. *Ann. Stat.*, 9, 1187.
Totsuji, H. & Kihara, T., 1969. *Publ. Astr. Soc. Japan,* **21**, 221.
Zwicky, F., Herzog, E., Wild, P., Karpowicz, M. & Kowal, C. T., 1961–68. *Catalogue of Galaxies and Clusters of Galaxies,* California Institute of Technology, Pasadena.