

Summary of selected topics from G. J. Babu's Research

G. J. Babu has diverse research interests in both statistics and probability, and their applications to problems in biomedical research, Population Genetics, astronomy and astrophysics. His research interests have varied over the years. He has contributed extensively to probabilistic number theory, resampling methods, nonparametric methods and asymptotic theory. He published **five books** and over 125 research papers in standard journals. A sixth book on Statistical Inference is under preparation.

Inter-disciplinary research and activities:

Since late 1980's, G. J. Babu and his colleague Dr. E. Feigelson from Astronomy & Astrophysics Department, have led the efforts to bring advanced statistical methods to serve the research needs of observational astronomy. Their work in astrostatistics includes multivariate methods for satellite data on Gamma-ray bursts, quantitative comparison of source properties to compare data with astrophysical theories and regression methods to calibrate steps in the cosmic distance ladder. The last one has important application in determining the size of the universe. Analyzed the Third Catalog of Gamma-Ray bursts data from BATSE on board the Compton Gamma Ray Observatory, using multivariate analysis. Babu, Feigelson along with a group of astronomers and statisticians identified a third cluster from the data and it was clearly characterized. These are among the few articles they published in *The Astrophysical Journal*. They have published an inter-disciplinary book **Astrostatistics**.

To enhance the dialogue between astronomers and statisticians on important research issues Babu and Feigelson have organized three highly successful international conferences *Statistical Challenges in Modern Astronomy* I, II, III and IV in August 1991, June 1996, July 2001 and June 2006. The conferences were well attended by leading astronomers and statisticians. Babu has also organized a well attended special invited paper session on Statistics in Astronomy at the joint statistical meetings at San Francisco, in August 1993, a session on statistics in astronomy, on behalf of International Astronomical Union, at the 51st session of the International Statistical Institute at Istanbul in August 1997, and a session on 'Astrostatistics' (Track: Emerging Science: Transforming the Next Generation) at AAAS Annual meeting and science innovation exposition in Philadelphia, in February 1998. He is on the program committees for the conferences, 'Astronomical Data Analysis' (part of the International Society for Optical Engineering's (SPIE) Symposium on Optical Science and Technology) in San Diego, in August 2001, and 'Astronomical Data Analysis II' (part of SPIE's Symposium on Astronomical Telescopes and Instrumentation) in Waikoloa, Hawaii, in August 2002.

Babu is the leading statistician of the Grist (Grid Data Mining for Astronomy) team. In addition, Babu is the director of the Center for Astrostatistics, which serves as a crossroads where researchers at the interfaces between statistics, data analysis, astronomy, space and observational physics collaborate, develop and share methodologies, and together prepare the next generation of researchers.

The National Virtual Observatory:

Babu's current inter-disciplinary research efforts focus on statistical methodology for the National Virtual Observatory (NVO). The NVO initiative has recently emerged, in response to a top priority recommendation of the NAS Taylor/McKee Decadal Report on astronomy for 2000-2010, to federate numerous large digital sky archives and develop tools to explore and understand these massive volumes of data. The effective use of such integrated massive datasets present a variety of new challenging statistical and algorithmic problems that require methodological advances. Recognizing its importance, the National Science Foundation awarded a \$1 million three-year research grant to G. J. Babu to address some of the critically important statistical challenges raised by the NVO. Babu is leading an interdisciplinary team of astronomers, computer scientists and statisticians from Penn State, California Institute of Technology, and Carnegie Mellon University, to develop statistical and computational methodology for the NVO. This effort covers several areas of statistics including the fields of multivariate analysis, nonparametrics, Bayesian analysis, spatial point processes, density estimation and data mining. Specific approaches being investigated for the NVO project include: low-storage percentile estimation for large datasets, multi-resolutional multi-dimensional trees for clustering and outlier detection, and multi-dimensional goodness-of-fit tests for comparison of multivariate astronomical

datasets with astrophysical models. The cross-disciplinary team brings advances in these fields into the toolbox of observational astronomy. The project seeks not only to formulate effective techniques to address NVO problems, but to code these methods into statistical toolkits within NVO software environments for the entire astronomical community. The team has developed ‘VOStat’; it is a prototype knowledge-based statistical toolkit implemented within the VO paradigm for the entire astronomical community. VOSTat consists of an easily extensible distributed web services-based framework.

National Radio Astronomy Observatory:

The fields of radio and microwave astronomy are experiencing huge advances in instrumentation. Innovations in broadband receiver, fast correlator and other technologies are leading to order-of-magnitude improvements in sensitivities and throughput from meter through submillimeter wavelengths. The NSF-funded National Radio Astronomy Observatory (NRAO) is now constructing the Extended Very Large Array (EVLA). NRAO is joining European and Asian countries in building the Atacama Large Millimeter Array (ALMA) in Chile. The most common scientific product of this new generation of radio telescopes is a 3-dimensional ‘datacube’ (or ‘hyperspectral image’) giving brightness as a function of location in the two-dimensional sky at many channels of frequency. The new EVLA and ALMA correlators will typically produce $\sim 10 - 100$ GBy datacubes with images of $\sim 1 - 10$ million pixels at $\sim 1 - 100$ thousand frequency channels. Petabytes of these datacubes will flow from these telescopes during the 2013-2020 period.

Scientific goals for radio datacube analysis include obtaining a list of reliable continuum and line sources to faint flux levels, and estimation of source properties (total flux, size, morphology, and continuum *vs.* line emission, with error analysis). The statistical and computational problems encountered in the quantitative analysis of radio datacubes are diverse and challenging. The EVLA and ALMA datacubes will exhibit heteroscedastic (i.e., spatially varying), non-Gaussian, and spatially correlated noise. At the invitation of the Director of NRAO, Babu has recently started collaborating with their algorithm group at Pete Domenici Science Operations Center, Socorro, NM. Over the next few years, several steps will be addressed: radio frequency interference removal, quantile functions for noise characterization, faint signal detection with false detection proportion control, local regression for mapping noise spatial correlation; g-inverses; and source characterization with maximum likelihood modeling. Some of this work will be based on Babu’s recent work on faint source detection in multi-epoch data.

Bootstrap method:

Babu along with his student Kesar Singh gave theoretical support to Efron’s resampling method called ‘bootstrap’. This work in early 1980’s on the asymptotic theory for the bootstrap method resulted in establishing the superiority of the bootstrap approximation for a wide class of statistics. This is achieved using Edgeworth expansions. This laid the foundation for subsequent work on second order approximations of the bootstrap method.

Recently G. J. Babu, along with P. K. Pathak and C. R. Rao has proposed a sequential approach called Poisson bootstrap, in which resampling is carried out until a certain proportion of distinct observations are sampled from the original data. Using conditional Edgeworth expansions, they have established the second order correctness (skewness correction) for a wide class of statistics as in the case of classical bootstrap. One of the main advantages of the sequential approach over the fixed sample size bootstrap is, to prevent too many repeated observations in a bootstrap sample that may lead to a degenerate value for the statistic under consideration. Thus Poisson bootstrap avoids zero value for variance estimator.

Practically there is no literature on statistics which are asymptotically distributed as linear combinations of Chi-squares exists. To study such statistics a modification of bootstrap statistic was suggested and was shown that their distributions are very close to each other. This gives a much needed practical method to obtain confidence intervals for such statistics. This method is recently applied by various authors to study the so called U-statistics.

Bootstrap estimation of variance of sample quantiles were studied and exploited the proof to obtain a method of estimation of density quantile function. Simulation studies showed that the bootstrap method

gives better results for studentized statistics. These observations are explained by the theoretical studies. A method to obtain confidence intervals is suggested using the bootstrap method. It is shown to perform better than the percentile method of Efron in the one sided case. The method is extended to autoregressive models.

Subsample and half sample methods are closely related to bootstrap and jackknife method. Babu investigated the large sample performance of this method. It is shown that the half-sample method is robust in estimating the parameters of a linear regression model when the errors are heterogeneous. Theoretical and computational aspects of the bootstrap methodology were reviewed in Babu and Rao (1993). It has an extensive bibliography.

To assess the accuracy of the estimators thus obtained, one needs to estimate the variance of the estimators. There are several options, including bootstrap and halfsample methods (Babu 1992). Bootstrap and jackknife are two widely used methods to estimate variance of a statistic. For non-smooth statistics such as sample median, it is well known that jackknife method fails. In this connection, in 1999, Babu examined the bootstrap method using the notion of breakdown point in robustness, in the context of estimation of the variance of an estimator and of confidence intervals. Even when the estimators are robust, bootstrap estimator of the variance is strongly influenced by a single outlier, as the bootstrap utilizes all the data points. On the other hand halfsample method has high breakdown point, sometimes as high as $1/4$, in the case of estimator of the variance of sample median. This phenomenon is being explored for the M-estimators, and linear order statistics.

Bootstrap procedures for inference on the regression parameters in an inverse Gaussian regression were investigated by Babu and Chaubey (1996). The large sample theory for the ‘pseudo maximum likelihood’ estimators is available in the literature, only when the number of replications increase at a fixed rate. This is inadequate for many practical applications. This paper establishes consistency and derives the asymptotic distribution for the ‘pseudo maximum likelihood’ estimators under very general conditions on the design points. This includes the case where the number of replications do not grow large, as well as the one where there are no replications. The bootstrap procedure for inference on the regression parameters is also investigated.

Babu and Padmanabhan (1995) have proposed a statistic for simultaneous testing for differences in location, scale, symmetry (or skewness) and tailweight between two unknown continuous distribution functions. Simulation studies show that the test based on the statistic tends to be more robust than its competitors. Asymptotic theory presented in the paper justifies the use of the bootstrap method.

Goodness-of-fit:

Many nonparametric goodness-of-fit tests, such as Kolmogorov-Smirnov, and Cramér-von Mises, are based on the empirical distribution function. In the presence of nuisance parameters, the tests are generally constructed by first estimating the nuisance parameters. However, even when the parametric model is specified, the asymptotic null distribution of the test statistic depends in a complex way on the unknown parameters. Babu and Rao used bootstrap methods to estimate the null distribution. They have demonstrated that, under very general conditions, the difference between the empirical process and the population distribution with estimated parameters converges weakly to the same Gaussian process as the corresponding bootstrap version. This result is used to show that the bootstrap method consistently estimates the null distributions of various goodness-of-fit tests. These results hold not only in the univariate case but also in the multivariate setting. For the case when the hypothesis is rejected, using Kullback-Leibler measure of separation, Babu and Rao have developed a resampling method to set confidence bands to the difference of the true and the closest distribution in the specified family. The methods are based on the weak convergence of empirical processes.

Edgeworth expansions:

In 1991, Babu obtained an s -term Edgeworth expansions for a wide class of statistics which are smooth functions of lattice and nonlattice marginals, where $s > 2$. The result is then applied to a statistic similar

to the student's t-statistic, where the scaling factor, the sample standard deviation is replaced by the more robust mean absolute deviation. Edgeworth expansions for the product limit estimator and estimators based on product limit estimator are also obtained in the same year.

Babu and Bai (1992) have established Edgeworth expansions for sums of independent but not identically distributed multivariate random vectors. These results are applied to obtain valid Edgeworth expansions for estimates of regression parameters in linear errors-in-variable models. Using these expansions, the bootstrap distribution is shown to approximate the sampling distribution of the studentized estimators, better than the classical normal approximation. This justifies the use of bootstrap in applying the errors-in-variables regression to the cosmic distance scale, one of the important problems in Astronomy.

Edgeworth expansions play an important role in resampling methods such as bootstrap. Babu and Bai (1993) established expansions for functions of multivariate means under partial Cramér's condition and under minimal moment conditions. Expansions that are local in one coordinate and global in rest of the coordinates were obtained for sums of independent but not identically distributed random vectors in Babu and Bai (1996). The results were then applied to derive Edgeworth expansions for bootstrap distribution, bayesian bootstrap distribution, and for the distributions of statistics based on samples from finite populations. Expansions for sums of non-identically distributed random vectors and of random vectors with lattice and non-lattice coordinates were reviewed in Babu (1993), along with their applications to errors-in-variables models, least absolute deviation estimators etc.

Edgeworth expansions for samples from finite populations:

Edgeworth expansions were obtained for the mean of a simple random sample drawn, from a finite multivariate population, without replacement. These are obtained under very general assumptions, which are easy to verify in practice. Consider the two sample non-parametric statistics of the type $T = (1/n) \sum f(R_i/N)$, where f is a smooth function and R_1, \dots, R_n are the ranks of one of the samples. In our results we need only to assume that f is continuous and monotone in a small interval. It appears that not much is known about the expansions of T , when $\{f(i/N)\}$ takes only two values. This and other related problems in sample surveys can be handled by using our results on the lattice case. As one of the applications we obtain expansions for the univariate statistics which can be expressed in a certain linear plus a quadratic form. A fairly large class of statistics used in sample surveys fall in this category. These results can be used to get bootstrap approximation to various statistics in the finite population case.

Population genetics and combinatorial structures:

The Ewens sampling formula gives the distribution of the allelic partition of a sample of genes from the so-called infinitely many neutral alleles model of population genetics. Its appearance in 1972 provided the first rigorous framework for the statistical analysis of allozyme frequency data. In addition, it showed that classical methods for estimating mutation rates from such data were using precisely that part of the data least informative for the parameter of interest, and further that the shape of the distribution of gene frequencies to be expected in neutral samples was exactly the opposite to what the biologists had thought. Though for the past several years the focus has moved away from allozyme data and more towards DNA sequence data, the combinatorial content of the Ewens sampling formula has recently been recognized as central to the study of a broad class of combinatorial structures. Ewens' formula describes the probability structure of *allelic partition*, $\bar{k} = (k_1, \dots, k_n)$, where k_j denotes the number of alleles represented j times in a sample of n genes. The formula is closely associated with independent Poisson random variables.

The recent work of Babu is directed towards approximations for the distribution of general additive functional of (k_1, \dots, k_n) . The Ewens sampling formula can be considered as a measure on the conjugate classes of group of permutations σ of the first n integers. Suppose

$$H_n(\sigma, t) = \frac{1}{\beta(n)} \sum_{j \leq y(t)} h_j(k_j(\sigma)) - A(y(t), n), \quad t \in [0, 1]$$

where $k_j(\sigma)$ denotes the number of cycles of length j of σ , $\beta(n)$ is a scale factor, the location A and ‘time’ $y(t)$ are defined through h_j and β . There is a relation between this and the corresponding process X_n with independent increments defined by

$$X_n(t) = \frac{1}{\beta(n)} \sum_{j \leq y(t)} h_j(1) \xi_j - A(y(t), n), \quad t \in [0, 1],$$

where ξ_j , $1 \leq j \leq n$ are independent Poisson random variables with $\mathbf{E}(\xi_j) = \theta/j$. Babu and Manstavičius have obtained necessary and sufficient conditions for H_n to converge to the Brownian motion. They have also shown that, for a class of infinitely divisible limit processes X in $\mathbf{D}[0, 1]$, H_n converges weakly to X if and only if X_n converges to X . This class of limit processes include stable processes. In sharp contrast to this, the one-dimensional distribution of $H_n(1)$ can converge to a stable distribution, while $X_n(1)$ does not converge to a stable law. This may be due to the contribution of large cycles to $H_n(1)$. In this case H_n is shown to converge to a process with dependent increments. The basic techniques used in this study are borrowed from probabilistic number theory. This framework is extraordinarily powerful. It allows approximation of the distribution of many interesting functionals of the combinatorial structures by the corresponding functionals of simpler processes.

Although the results are motivated by Ewens sampling formula, they have wide range of applications including partitions of an integer in number theory and other combinatorial structures such as, ‘assemblies’. These objects are related to physics (representations and Young diagrams), theoretical computer science (tree-based searching and symbolic processing algorithms based upon forests), cryptology (factorization of polynomials), genetics (Ewens sampling formula), and chemistry (random trees as models for cyclic polymerization).

Probabilistic number theory:

G. J. Babu contributed extensively to *Probabilistic number theory* in early 1970’s, and published a monograph, **Probabilistic Methods in the Theory of Arithmetic Functions**, in 1978. His main contributions in this area include a partial solution to a long standing conjecture of Erdős, and the result that every bounded additive arithmetic function has a singular distribution. He introduced a concept of density of natural numbers, capable of detecting large gaps in a set of natural numbers, and investigated the existence of distribution of values of an arithmetic function under this density. An important consequence of this is that if $\omega(m)$ denotes the number of distinct prime factors of m , then

$$\#\{n < m < n + b(n) : \omega(m) - \log \log m < x \sqrt{\log \log n}\} / b(n) \rightarrow \Phi(x), \quad (*)$$

where $(\log b(n)) \sqrt{\log \log n} / \log n \rightarrow \infty$, and Φ denotes the standard normal distribution function. This is a generalization of the well known Erdős-Kac Theorem and it leads to a better understanding of integers with large number of prime factors. It is shown that (*) fails to hold if $b(n) < (\log n) / (\log \log n)^2$. Given an additive function f , the problem of determining the slowest growing function b so that f has a distribution in the sense of b-density was also studied. Part of this was joint work with late Professor Paul Erdős.

Other research:

Asymptotics for functions of marginal quantiles: While the large-sample properties of marginal sample quantiles in the case of independent vectors were studied by Babu & Rao (1988), large-sample properties of means of functions of marginal quantiles:

$$\frac{1}{n} \sum_{i=1}^n \phi \left(Y_{n:i}^{(1)}, \dots, Y_{n:i}^{(k)} \right),$$

where $Y_{n:i}^{(j)}$ denotes the i th order statistic of $\{Y_1^{(j)}, Y_2^{(j)}, \dots, Y_n^{(j)}\}$, and ϕ is a function, are investigated by Babu in collaboration with colleagues Bai & Choi of National University of Singapore. Here $\{(Y_i^{(1)}, Y_i^{(2)}, \dots, Y_i^{(k)})\}$,

$i = 1, 2, \dots\}$ is a sequence of random vectors such that for each j ($1 \leq j \leq k$), $\{Y_1^{(j)}, Y_2^{(j)}, \dots\}$ forms a sequence of independent and identically distributed random variables. The study includes asymptotic normality, the strong law of large numbers, and functional limit theorems.

Density quantile estimation: Density quantile estimators based on the smoothness properties of the density were constructed and were shown to be asymptotically efficient in the mean square error sense. Unlike density estimators these do not require knowledge of the actual values of the derivatives of the density. Uniformly almost sure bounds for these estimators in an interval were obtained in the dependent case too.

Competing risk models: In a series of papers, a standard competing risk situation with possibility of censoring due to withdrawal or end of study were investigated. Large sample theory is derived and bootstrapping is discussed as a way to estimate the variance. Some of this is joint work with C. R. Rao.

Chaotic processes: In the past few years, Babu has used the Bernstein Polynomials for smooth estimation of multivariate distributions and density functions by taking advantage of the knowledge of the support of the distribution. He has extend these results to multivariate as well as dependent case. He has applied these results to discrete time deterministic chaotic dynamical systems with noise contamination.

Robust estimation: The large sample properties of statistics based on the robust mean absolute deviation from the sample mean as well as sample median are obtained. These are applied to regression context, the procedures are more robust against outliers compared to the usual procedures base on the classical least squares method.

Mixing sequences: The so called \mathbf{r} -quick limit points of empirical distribution functions of mixing processes were characterized. Also obtained an \mathbf{r} -quick version of Bahadur-Kiefer type representation for sample quantiles. These results are applied to linear functions of order statistics.

Moderate deviations in Banach spaces: Probabilities of moderate deviations for i.i.d. sequences taking values in a separable Banach Space under precise necessary and sufficient conditions were obtained jointly with C. M. Deo. These results are not known earlier even for real valued random variables.

Occupation measure of empirical processes: The limit behavior of the occupation distribution

$$L_t(E) = \int_0^1 I_E(\sqrt{(\log \log t)/t} K(s, t)) ds,$$

where K denotes a Kiefer process, was studied. Strong approximation results are then used to derive the law of iterated logarithms in Chung's form for various functions of empirical processes.

Environmental statistics: Trawl surveys are carried out regularly by the Northeast Fisheries Center (NEFC) to assess the fish stocks of various species. The external factors influencing the assessment include the ship used, doors, nets, etc. Occasionally some of these have to be replaced. So there is a need for a conversion factor to neutralize this influence. Major difficulty encountered in this problem is due to the large proportion of zero catches. A method of estimation for the conversion factor is suggested and various asymptotic properties of the estimator studied. The results are illustrated by a data set provided by NEFC.

Honors

G. J. Babu was elected Fellow of the Institute of Mathematical Statistics in 1987 for his work on the asymptotic theory of bootstrap methods, and was elected Member of the International Statistical Institute in 1989. He was elected Fellow of American Statistical Association in 1997, for outstanding and wide-ranging contributions to probability and statistics; for leadership in promoting interdisciplinary activities to bring astronomers and statisticians together; and for service to the statistical profession. He was elected Fellow of American Association for the Advancement of Science in 1997, for research on asymptotic theory, resampling methods, probabilistic number theory, and statistical methods for astronomy and for promoting interdisciplinary activities. He has also received National Research Council's Twinning fellowship for 1997-1999 to initiate collaboration, on Statistical Group Theory and Probabilistic Number Theory, with colleagues from Vilnius University, Lithuania.