POLITECNICO DI TORINO
Repository ISTITUZIONALE

Experimental data modeling: issues in empirical identification of distribution

(Article begins on next page)

25 April 2024

Experimental data modeling: issues in empirical identification of distribution

G. Barbato, G. Genta, R. Levi*

Abstract

Identification of distribution underlying experimental data sets, lacking established mechanistic models, calls for an inferential process involving necessarily uncertainty of some sort. While components pertaining to parameter estimation are routinely taken into account, those related to selection of distribution form often are not; their awkward theoretical evaluation may explain why the issue tends to be conveniently ignored. Such an attitude may however lead to severe underestimation of overall uncertainty, since the component accounting for identification of distribution form often exceeds those concerning estimates of parameters. A pragmatic approach is presented, relying upon numerical simulation, allowing realistic evaluation of uncertainty inherent in empirical identification of form in a straightforward way. Application to an actual case is presented, and issues concerning identification procedure for parameters of auxiliary empirical distributions are discussed.

1. Introductory remarks

Fitting models to experimental data sets affected by scatter - all are - often involve tentative identification of underlying distribution. While standard errors of parameter estimates are routinely assessed, and joint confidence regions mapped as required, substantial information is seldom provided concerning uncertainty related to distribution identification, a practice which may be well explained by unwillingness to embark on a thankless job. Whatever the motivation, such a fairly widespread attitude has far reaching implications. While the bulk of most data sets sits comfortably on the friendly rump of normal distribution, tails may not comply as readily, owing to common causes such as skewness and/or existence of finite bounds. And, disregarding the component pertaining to distribution identification may lead to substantial underestimation in the uncertainty budget, since that component may well exceed those concerning estimates of parameters.

2. Statement of problem

Identification of form of underlying distribution(s) underlying a given experimental data set is a frequent requirement, e.g. to support models for stochastic simulation and related purposes. Normal distribution is ordinarily first selected, owing to its desirable properties as well as its wide applicability; other well-known forms are also considered, such as exponential, lognormal, gamma and beta. While sometimes providing adequate approximation, the usefulness of such distributions is somehow curtailed by inherent lack of flexibility, underlined by interdependence between third and fourth order moments. On the skewness-kurtosis plane, the fields of existence of the first two distributions listed above shrinks down to points only, with coordinates respectively 0, 3 and 4, 9; the third roughly corresponds to the straight line $\beta_2 \approx 3 + 1.95\,\beta_1$, only the last covers a sizable fraction of possible area [1-2]. Should symmetry apply ($\beta_1 = 0$), when $\beta_2 > 3$ Student's t distribution is also a candidate if variance is related to kurtosis according to:

$$\sigma = 2 - \frac{3}{\beta_2} \qquad (1)$$

Empirical representation of a broad variety of data is on the other hand afforded by a number of methods, flexible enough to afford complete coverage of all possible combinations of skewness and kurtosis. Several procedures were developed, mainly based upon transformations of normal distribution in the wake of Pearson's seminal essay [3], providing a systematic, general approach to analytical description of a comprehensive range of distributions. Fitting a distribution to data requires identification of a suitable family, and parameter estimation; both steps entail some uncertainty, to be taken into account if a meaningful uncertainty budget is required. But for the first order one, sample moments are biased estimators of those of populations, and so are ratios such as sample skewness b_1 and kurtosis b_2:

$$b_1 = \frac{m_3^2}{m_2^3}, \quad b_2 = \frac{m_4}{m_2^2} \qquad (2)$$

Other forms currently in use may however reduce bias substantially in some circumstances only, and fail to identify correctly even the sign of population parameters if sample size is not very large.

3. Evaluation of confidence region

Empirical distribution fitting may be performed in terms of the first four moments or their combinations - μ, σ, β_1, β_2 - or more frequently of their sample estimates, *faute de mieux*. In the latter case, substantial uncertainty and bias may affect estimates of skewness and kurtosis, particularly when sample size is not very large, as typical of many experimental data sets [4]. Systems of frequency curves capable of modeling a broad range of distributions were developed by Pearson [3] and Charlier [5]; nowadays Johnson's distributions [6] are usually preferred, along with generalizations of Tukey's lambda distribution [7], owing to their adequate coverage of β_1 - β_2 plane and comparatively straightforward fitting. Among Johnson's distributions, one out of three families - S_B, S_U or S_L (lognormal) - is selected, according to whether the estimate of β_2 is less than, exceeds, or comes close to approx. $3 + 1.95\ \beta_1$.

Substantial bias and scatter typical of sample estimates of skewness and kurtosis, not to mention higher order moments, may suggest fitting tentatively more than one family of distributions to data, and then perform selection in terms e.g. of goodness of fit [8].

Starting from a continuous random variable X, whose unknown distribution is estimated in terms of sample values, an approximation to underlying distribution may be inferred by identifying a transformation of x to a standard normal random variable z, according to the method of translation first introduced by Edgeworth [9], further extended by Kapteyn and Van Uven [10] and Rietz [11], see also Hahn and Shapiro [1].

Johnson's empirical distributions are based upon a transformation having the general form.

$$z = \gamma + \delta\tau(x;\xi,\lambda),$$
$$\delta > 0, -\infty < \gamma < +\infty, \lambda > 0, -\infty < \xi < +\infty \qquad (3)$$

where γ and δ are shape parameters, λ is a scale parameter, ξ is a location parameter, and τ is an arbitrary function, taking the following alternate functions

$$\tau_1(x;\xi,\lambda) = \ln\left(\frac{x-\xi}{\lambda}\right), \quad x \geq \xi \qquad (4)$$

$$\tau_2(x;\xi,\lambda) = \ln\left(\frac{x-\xi}{\lambda+\xi-x}\right), \quad \xi \leq x \leq \xi+\lambda \qquad (5)$$

$$\tau_3(x;\xi,\lambda) = \sinh^{-1}\left(\frac{x-\xi}{\lambda}\right), \quad -\infty < x < +\infty \qquad (6)$$

Taking

$$\gamma^* = \gamma - \delta\ln\lambda \qquad (7)$$

the first leads to

$$f_1(x) = \frac{\delta}{\sqrt{2\pi}(x-\xi)}\exp\left[-\frac{1}{2}\delta^2\left(\frac{\gamma^*}{\delta} + \ln(x-\xi)\right)^2\right], \qquad (8)$$
$$x \geq \xi, \delta > 0, -\infty < \gamma^* < +\infty, -\infty < \xi < +\infty,$$

the three-parameter log-normal distribution also designated as Johnson S_L family. Furthermore, for the second we may write

$$f_2(x) = \frac{\delta}{\sqrt{2\pi}}\frac{\lambda}{(x-\xi)(\lambda-x+\xi)}\exp\left[-\frac{1}{2}\left(\gamma + \delta\ln\left(\frac{x-\xi}{\lambda-x+\xi}\right)\right)^2\right], (9)$$
$$\xi \leq x \leq \xi+\lambda, \delta > 0, -\infty < \gamma < +\infty, \lambda > 0, -\infty < \xi < +\infty$$

defining the four-parameter Johnson S_B family of distributions. Eventually, the third takes the form

$$f_3(x) = \frac{\delta}{\sqrt{2\pi}} \frac{1}{\sqrt{(x-\xi)^2 + \lambda^2}} \times$$

$$\times \exp\left[-\frac{1}{2}\left(\gamma + \delta \ln\left(\frac{x-\xi}{\lambda} + \left(\left(\frac{x-\xi}{\lambda}\right)^2 + 1 \right)^{\frac{1}{2}} \right) \right)^2 \right] \quad (10)$$

$$-\infty \le x \le +\infty, \delta > 0, -\infty < \gamma < +\infty, \lambda > 0, -\infty < \xi < +\infty$$

corresponding to the four-parameter Johnson S_U family of distributions. Advantages of this system of distributions for simulation input modeling over other distribution families such as triangular, beta, and normal were examined by DeBrota *et al.* [12].

Fitting a Johnson distribution to data involves first selecting the proper family, and then obtaining estimates of the four parameters γ, ξ, δ, λ, in terms of the combination considered of values for the mean μ, standard error σ, skewness β_1 and kurtosis β_2, typically assessed in terms of their sample estimates. Closed-form expressions for parameter estimates based on the method of moment matching are not available, close approximations may however be obtained using iterative procedures [13], made substantially easier by dedicated software packages [8]. Three main methods are in general used, namely based upon selected percentile points, evaluation of moments, and maximum likelihood, mentioned in order of ease of application.

Since a Johnson S_B variate ranges between ξ and $\xi+\lambda$, when both end points are known estimation of only γ and η is required, readily performed in terms of selected percentile points. Equating two percentiles from the sample at hand - obtained by ranking and interpolation as required - with those corresponding for normal distribution lead to equations in terms of $\alpha \cdot 100^{\text{th}}$ and of $(1-\alpha) \cdot 100^{\text{th}}$ percentiles of sample and of standard normal distribution respectively which may be solved in a straightforward way. When only the lower bound ξ is known, an additional equation is obtained by matching data median with that of standard normal distribution, that is zero. Fitting is somehow more laborious when neither end points are known, a situation seldom occurring in practice [1].

When precious few sample data are available, visual interactive fitting relying also on subjective information may also be performed [12]. For a Johnson S_U distribution parameters γ and δ of may be estimated expeditiously in terms of skewness and kurtosis e.g. from Johnson's abac; ξ and λ are then determined from the first sample moments [6]. Fitting by moments is not always a desirable procedure, according to an elegant understatement [13]. It may be resorted to in some instances, e.g. as a first step towards evaluation of a maximum likelihood solution, or in the course of theoretical investigations where uncertainties connected to sampling carry little or no weight. A broad range of algorithms and computer aided procedures for distribution fitting, recently collected with a comprehensive, up to date survey of the subject [14], cater for easier selection of candidate distribution and fitting procedure.

Exact evaluation of a joint confidence region in the $\mu - \sigma - \beta_1 - \beta_2$ space entails substantial labor, hardly justified in view of the inherently large scatter in the sample estimates of third and higher order moments. For the purpose at hand expeditious determination of coordinates of apexes of some sort of prism built around the confidence region is quite adequate. Projection on the $\mu - \sigma$ plane yields a trapezium, readily marked off in terms of confidence intervals for mean and standard deviation. A rectangle on the $\beta_1 - \beta_2$ plane provides a rough approximation of a joint confidence region for skewness and kurtosis, shrinking down to a segment on the ordinate in case of symmetry. In the latter case, assuming independence among estimates of mean, variance and kurtosis leads to a three dimensional prism in the $\mu - \sigma - \beta_2$ space, whose eight apexes identify as many limit Johnson distributions still compatible with the data set at hand [15], see Fig. 1.
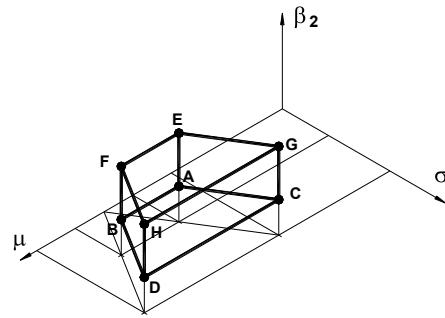

Fig. 1. Prism enclosing a joint confidence region in the μ-σ-β_2 space for location, spread and peakedness of distributions compatible with data set [15].

When normal distribution may be assumed to a first approximation for the case at hand, as not unusual, a confidence interval for β_2 straddles the nominal value of three, entailing that four out of the eight Johnson distributions belong to S_B family and four to S_U, that is four are bounded and four are not. Since each of these distributions corresponds to the combination of mean − standard deviation − kurtosis estimates of the relevant apex, the envelope of their plots drawn e.g. as cumulative probability graphs identifies a reasonable approximation of the overall confidence region, inclusive of *type A* contribution due to both identification of form, and estimation of parameters of model assumed [4]. In the general case, a four

3

dimensional region would apply. A joint confidence region for skewness and kurtosis is conveniently assessed by numerical simulation; rectangular shape may not apply for rather small sample sizes, where estimates of skewness and kurtosis are strongly correlated [15]. A confidence region for the distribution underlying data at hand is eventually obtained as the envelope of sixteen distributions identified by as many combinations of estimates of μ, σ, β_1, β_2 consistent with the confidence level selected

4. Example

Let us consider a set of data concerning a laboratory verification of micrometer calibration with gauge blocks according to a relevant standard [16], obtained at Politecnico di Torino. The distribution of absolute values of differences between readings and reference values, rounded off to 0.1 μm, is shown in Fig. 2; three mavericks – identified as outliers according to established criteria [17] - observed on the right tail were previously rejected at 95% confidence level.
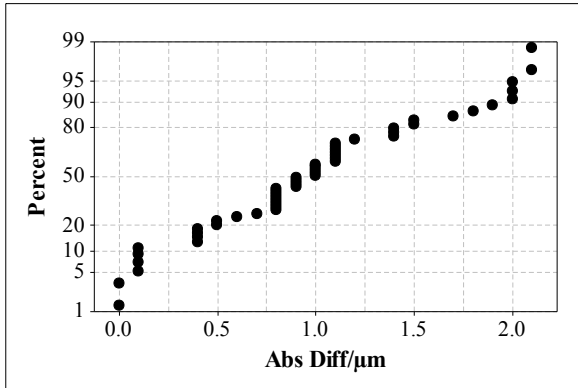

Fig. 2. Normal probability plot of absolute values of differences between micrometer readings and reference values.

At exploratory level sample distribution appears symmetrical, slightly platykurtic, and bounded at either tail, albeit owing to different causes. Lower bound is zero, no negative values being possible; the existence of an upper bound slightly exceeding 2 μm appears to be justified by widespread acceptance as natural for the instrument considered of deviations of the order of 1 μm, and reluctance to consider as legitimate values twice as large. Suspiciously large results suggest careful replication, leading almost invariably to discard as a maverick the offending value.
Confidence intervals for population parameters were readily assessed at 95% level, namely $0.47 \leq \sigma \leq 0.70$ and $0.84 \leq \mu \leq 1.12$ or respectively $0.78 \leq \mu \leq 1.18$. A confidence interval at the same level for kurtosis is

obtained numerically as $2.1 \leq \beta_2 \leq 4.3$, hypothesis of symmetry of the underlying distribution being substantially consistent with sample data. Distribution of kurtosis is closely approximated by a three-parameter lognormal, but for minor departures concerning either tail covering less than 0.1% of population, see Fig. 3.
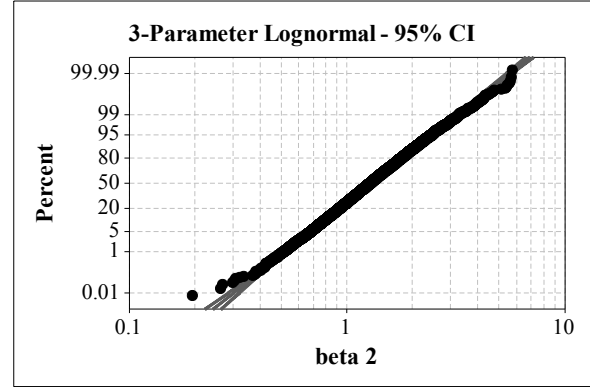

Fig. 3. Lognormal probability plot of 10^4 sample estimates of β_2 ($n = 52$), with 95% confidence bands.

Coordinates of apexes of prism bounding a confidence region for population parameters in the $\mu - \sigma - \beta_2$ space are listed in Table 1, with the relevant terms of the corresponding Johnson distributions. Their envelope is shown in Fig. 4, along with confidence limits pertaining to normal approximation of sample data. Finally, shapes of probability density functions relevant to the eight Johnson distributions are shown in Appendix A.

Table 1. Apex coordinates of prism shown in Fig. 1 for the case at hand.

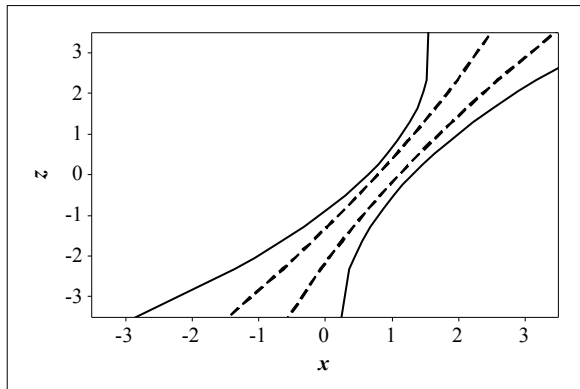| Apex	A	B	C	D	E	F	G	H
μ	0.84	1.12	0.78	1.18	0.84	1.12	0.78	1.18
σ	0.47		0.70		0.47		0.70	
β_2	2.1				4.3			
Type	S_B				S_U			

Fig. 4. Envelope of Johnson distributions corresponding to apexes *A* to *H* of prism shown in Fig. 1 (continuous), along with 95% confidence bands pertaining to normal approximation of sample data (dashed).

5. Discussion

How broad should be a realistic confidence region for distribution underlying sample data is a subject preferably dealt with taking into account prior knowledge, or expert opinion in GUM [18] parlance, if at all possible. Should sample data represent the bulk of relevant information available, a prudent approach would suggest to take into account the uncertainty component due to empirical identification of population, besides that concerning only parameter estimation, routinely considered. On the other hand, when the experimenter enjoys the advantage of substantial accumulated experience on the specific subject, on whose ground the form of distribution is well known, disregarding such knowledge would hardly make sense. Between no *a priori* information, and well established knowledge about distribution, a substantial gap in terms of width of confidence intervals may be observed, which in another instance (for a smaller sample) was found to be even larger [15,19]. The component concerning distribution form to overall uncertainty, rather small in the neighborhood of average, increases dramatically towards either tail, where neglect of that component would lead to ludicrously optimistic predictions.

References

[1]   Hahn, G.J., Shapiro, S.S., 1958. "*Statistical Models in Engineering.*" Wiley, New York.
[2]   Rhind, A., 1909. Tables to facilitate the computation of the probable errors of the chief constants of skew frequency distributions. Biometrika 7 (1/2), pp. 127.
[3]   Pearson, K., 1895. Contributions to the Mathematical Theory of Evolution.-II. Skew Variation in Homogeneous Material. Philosophical Transactions of the Royal Society of London Series A 186, pp. 343.
[4]   Barbato, G, Genta, G, Levi, R., 2011. "Uncertainty component pertaining to empirical identification of distribution", Committee STC "P". 61th CIRP General Assembly, Budapest.
[5]   Charlier, C.V.L., 1906. Über die Darstellung willkürlicher Funktionen. Ark. Mat. Astr. Fys. 2, 20, p. 1.
[6]   Johnson, N.L., 1949. System of Frequency Curves Generated by Methods of Translation. Biometrika 36, p. 149.
[7]   Tukey, J.W., 1960. The Practical Relationship Between the Common Transformations of Percentages of Counts and of Amounts. Technical Report 36, Statistical Techniques Research Group, Princeton University, New Jersey.
[8]   Swain, J.J., Venkatraman, S., Wilson, J.R., 1988. Least-Squares Estimation of Distribution Function in Johnson's Translation System, Journal of Statistical Computation and Simulation 29, p. 271.
[9]   Edgeworth, F.Y., 1898. On the representation of statistics by mathematical formulae, Journal of the Royal Statistical Society 61, p. 670.
[10] Kapteyn, J.C, van Uven, M.J., 1916. Skew Frequency Curves in Biology and Statistics. Astronomical Laboratory, Hoitsema Brothers, Groningen.
[11] Rietz, H.L., 1924. Handbook of Mathematical Statistics. Houghton Mifflin, Boston.
[12] DeBrota, D.J., Dittus, R.S., Roberts S.D., Wilson J.R., 1989. Visual interactive fitting of bounded Johnson distributions. Simulation: Transactions of the Society for Modeling and Simulation International 52(5), p. 199.
[13] Hill, J.D., Hill, R., Holder, R.L., 1976. Fitting Johnson Curves by Moments. Applied Statistics 25, p. 190.
[14] Karian, Z.V., Dudewicz, E.J., 2011. Handbook of Fitting Statistical Distributions with R. CRC Press, Taylor and Francis – Chapman & Hall, New York.
[15] Barbato, G., Genta, G., Germak, A., Levi, R., Vicario, G., 2010. "Approaches to handling discordant observations: an appraisal". Measurement Systems and Process Improvement (MSPI 2010), NPL, Teddington.
[16] UNI 9191:1988. Taratura di micrometri per esterni. Ente Nazionale Italiano di Unificazione, Milano.
[17] Barbato, G., Genta, G., Levi, R., 2010, "Mavericks in metrology". 7th CIRP International Conference on Intelligent Computation in Manufacturing Engineering, Capri.
[18] JCGM 100:2008. Evaluation of measurement data - Guide to the expression of uncertainty in measurement (GUM). Joint Committee for Guides in Metrology, Sèvres.
[19] Natrella, M.G., 1963. Experimental Statistics. NBS Handbook 91, Washington.

Appendix A. Shape of Johnson distributions

Numerical approximations of probability density functions pertaining to the eight distributions defined in Table 1 were obtained with a commercial software. Representative shapes of S_B distributions, relevant to apexes *A*, *B*, *C* and *D*, and S_U distributions, relevant to apexes *E*, *F*, *G* and *H*, are shown in Fig. 5.
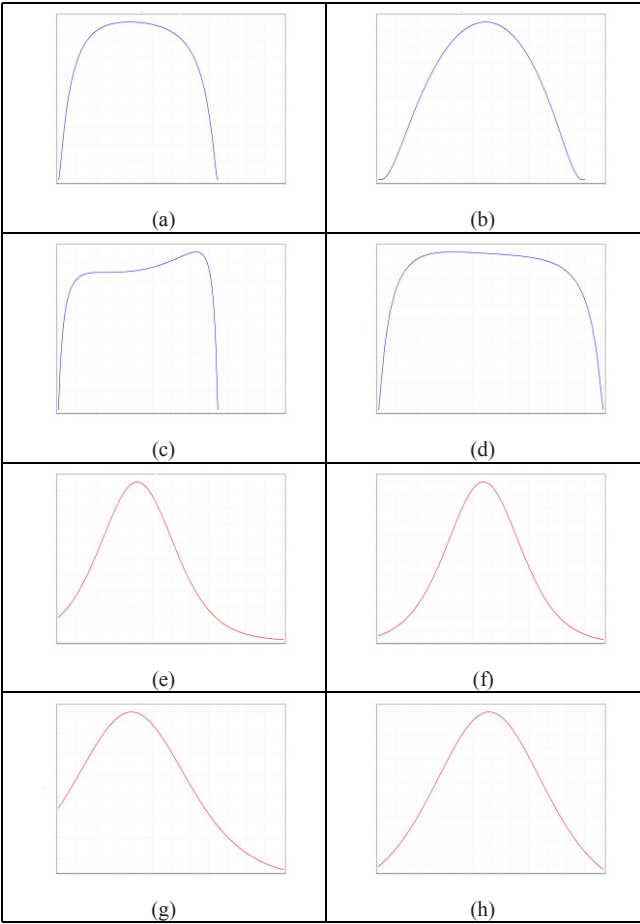
Fig. 5. Representative shapes of probability density functions of S_B distributions relevant to apexes *A*, *B*, *C* and *D*, and S_U distributions relevant to apexes *E*, *F*, *G* and *H*. Abscissas range from 0 to 2.4 in all graphs.