Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 48 (4) 711-716 (2016)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

A New Way to Identify the Family of Continuous Probability Distributions and Estimation of Population Parameter(s)

M. M. IQBAL[++], H. KHURRAM, S. AFZAL

Department of Statistics, Bahauddin Zakariya University, Multan, Panjab

Received 21st April 2016 and Revised 20th November 2016

Abstract: Identification of the distribution for a sample data in literature is a subjective approach based on the pre-assumptions of the distribution of that data and hence difficult for non/new statistician. The main objective of the current research is to provide a new methodology that is not exactly based upon subjective approach. By adjusting the width of the bars of the histogram we find the ordinates which show the height of each bar and provide base for our work. The true relative variances of the ordinates of some continuous probability distributions are calculated and then the pattern of these true relative variances of the ordinates is examined after dividing them into groups. Each group shows some specific distribution. Each group consist of one or more than one ranges of true relative variances of the ordinates. A procedure to compute the sample relative variances of the ordinates is defined. A comparison of true relative variances of the ordinates with sample relative variances of the ordinates provided the basis for identification of the family of probability distribution and also the members of family of distribution.

Keywords: Continuous probability distribution, histogram, ordinates, relative variance, Subset sample space

1. INTRODUCTION

The behaviour of random variable is not purely random in the sense that it does follow some *system* which is governed by certain rules that are characterized by underlying assumptions. The generic behaviour of the random variable is classified to belong to a whole lot of possibilities which could further been specialized for more specific assumptions. The class of generic behaviour of random variable is technically termed as family of distribution. So, each probability distribution is said to be a family. The value of parameter(s) in the parameter space of a particular family defines member of that family. Each member is the specific variant of the family and is identified by fixing the value(s) of the parameter(s) from amongst the admissible values. If, for example, $X \sim N(\mu, \sigma)$ and if $-\infty < \mu < +\infty$, $\sigma \geq 0$ defines a family of normal distributions then $X \sim N(5,2)$defines a member of the normal family with $\mu = 5$ and $\sigma = 2$.

For identifying the family of the distribution we usually rely heavily upon goodness of fit tests e.g. Mayooran and Laheetharan (2013). In this method we do not actually identify the family rather we compare the data if its distribution is compatible in terms of family as well as membership with a hypothetical one. Thus it is not identification rather testing. It means that there exist no reliable objective ways of identifying parent family of distributions for the situation. The only exception can be offered to the theoretical background which does not always reliably available.

For estimation of the members of the families some usually used methods that are more common are, Ordinary Least Square estimation, Maximum likelihood estimation, method of moments, etc. Some less common are categorized as graphical estimation methods i.e. Quantile Quantile (Q-Q) Plot and Probability Plot Correlation Coefficient (PPCC) Plot. Su and Chou (2007) presented a neural network-based approach for probability distribution recognition. Back propagation and learning vector quantization were used in categorizing normal, exponential, Weibull, Uniform, Chi-square, t, F, and Lognormal distributions. Experimental results illustrated that their proposed approach performed better than the traditional statistical approach.

2. PROPOSED APPROACH

If the probability distribution is drawn graphically and bounded by horizontal axis, then it covers a unit area. For continuous random variable more specifically the area which has been defined equal to one square unit is distributed over the entire range as per a scheme defined by the probability distribution itself. If the entire sample space is distributed in several smaller bars of same width the corresponding area cover within each bar under the curve can be used to assess the pattern of the distribution. If we keep on decreasing the width of the interval then hypothetically the width reduces to single point having no width at all. In the absence of the width of the bar, the height alone will be able to show the pattern of distribution. That is if the ordinates of a

[++]Corresponding author's email: mutahiriqbal@bzu.edu.pk, Hariskhurram2@gmail.com saimaafzalbzu@bzu.edu.pk

distribution are used as the 'bar' then the pattern of the distribution can be accessed through the pattern of the corresponding ordinate, thus if we have enough ordinate to observe then distribution can be observed precisely. As the pattern of the distribution for a family of the distribution is entirely based upon the value of the entire distribution thus the members of the family can be identified, if the pattern of the ordinates matches with true pattern. In an over specified situation where our family of distribution involves single parameter then the identification of the member will become clearer. By identification of the member of family we actually mean the determination of the value of parameter within the value of parameter space. In simple words it means, estimating the parameter if it is not already known (Webb, 2003)

2.1 Summarizing the Ordinates

We need some statistic that enables us to summarize the ordinates. The major problem is that, in some of the scenarios we have very large or small values of ordinates. The variance of the ordinates is an appropriate statistic in this situation to summarize the ordinates as it is more sensitive and somehow indicates the distance between the ordinates.

2.2 Scheme to Identify the Family of a Distribution

Justifying the idea as discussed above, for the identification of the family of a probability distribution the ordinates of each distribution for a defined range of a random variable are computed. These true computed ordinates are based on different values of parameters defined from the parameter space. In the next step the variances of true ordinates are obtained from a specific distribution at a specific value of parameter(s) on defined range of random variable(s).

After finding the large sets of the variances of the ordinates, these sets of values are categorized on the basis of different distributions. For a specific distribution one or more than one intervals are defined. Each interval has some specific portion of the variance of the ordinate at known value of true population parameter. On the basis of this proportion we can assess the importance/significance of that interval in identifying the family of distribution. Now, working with a data set with unknown family, one has to find the sample variance of ordinates and then identify its distribution by just locating this value in pre-defined intervals obtained after the manipulation of true population ordinates. The estimated variance of ordinates is assumed to be distributed as that true probability distribution, for which the interval(s) of variance of ordinates contain estimated variance of ordinates.

2.3 Scheme to Identify the Member of the Family of Distributions

After finding the estimated variance of ordinates and identifying the family of distribution the next step is to check which of the true variance of ordinates have minimum distance with that of estimated in that interval. After locating that true variance of ordinates the true parameter(s) at which those ordinates have been obtained is assumed to be the member of that family in the parameter(s) space of that sample.

3. METHODOLOGY

For the computational purpose, different distributions are used to obtain the true variances of ordinates. A distribution having continuous random variable is used in our work. The selected distributions are: Weibull distribution, F-distribution, Exponential distribution, t-distribution, Beta distribution, Chi-square distribution and Normal distribution. All these distributions belong to the class of continuous probability distributions.

The reason to select these distributions is the computational convenience. As the computational work is done with R language and for these common distributions, R has built-in functions.

3.1 Choosing the Range of Random Variables

Now the major concern is the range of random variable used to find the ordinates of a particular distribution. For different distributions the range of the random variables are shown in **(TABLE 1).**

Table 1: Ranges of Random Variables Used for Obtaining Ordinates

	Random Variables Range		
Distributions	Minimum Range	maximum range	step size
Weibull	0.000	500.000	0.500
F	0.100	1000.000	0.500
Exponential	0.000	1000.000	0.100
t	-1000.000	1000.000	0.100
Beta	0.001	0.990	0.001
Chi-square	0.000	1000.000	0.200
Normal	-1000.000	1000.000	0.500

The defined range as shown in **(Table 1)** is chosen from the actual range. The major reason behind the choice of this range is that it contains maximum informative values of the entire range. If we exceed from this range, it does not affect the results because the ordinates get much closer to zero and their contribution seems to be insignificant in the variance of ordinates. But this range is not limited as defined above. One can increase this range of random variable.

3.2 Choosing the Range of Parameter Space

A wide range of parameter(s) from the parameter space can be considered to find the ordinates of the distributions. These may be location, scale, shape or their combination(s), defined on the basis of particular family of a distribution. The used ranges of the parameter(s) for different probability distributions are shown in **(Table 2)**.

Table 2: Ranges of True Parameters

		Range of parameter		
Distributions	Parameters	Minimum	Maximum range	Step Size
Weibull	Scale	1.0	100	0.9
	Shape	1.0	100	0.9
F	df1	1	200	2
	df2	1	200	2
Exponential	Mean	0.1	500	0.3
t	df	1	1000	1
Beta	shape 1	0.1	150	1.5
	shape 2	0.1	150	1.5
Chi-square	df	1	1000	1
Normal	Mean	0.0	0	0.0
	Variance	0.1	1000	0.5

It can be observed from the above table that different ranges from defined parametric space are defined for different parameter(s) of a probability distribution.

We used the range of variance from 0.1 to 1000 for normal distribution by taking a step size of 0.1 but for mean the range is not defined. We used a constant value (mean = 0) to identify the location because the change of location does not affect the ordinates at a specific range of random variable. Especially for chi-square distribution when the value of random $variable = 0$ and df is 1 then the value of ordinate is infinity because of asymptotic curve. So this point is ignorable. As F, Beta and Weibull distributions involve two parameters so we used all possible combinations (with replacement) for both values of parameters. The results after increasing the interval of the range of the parameters will not effect on the results.

3.3 Sample Based Procedure

For a sample data, the histogram is prepared by choosing appropriate width of bar. Then the class boundaries are formed and frequencies are obtained. Relative frequencies are calculated from these frequencies. These relative frequencies are treated as ordinates of sample(s).

For a sample size greater than or equal to 50 we find some suitable results for our described methodology. The choice of this sample size is arbitrary.

The first step to make the class boundaries is to choose the suitable number of classes. After repeating the procedure, it is suggested that if the number of classes are between 40% to 60% of the sample size it provides better results. But these should not be considered as established cut-off points.

Finally, most of the repetitions support our suggested sample size and number of classes but they can vary for different scenarios.

3.4 Relative Approach for True Variances of Ordinates

For specific variances of ordinates of a distribution, with some specific family member, we divide it by its sum of the variances of ordinates. This transformation does not affect the pattern of the variances of the ordinates but it provides a good accommodation in favour of the range of the variances of the ordinates. After this transformation we are able to compare the sample variance(s) of ordinate(s) with true variance(s) of ordinate(s) and the true variances of ordinates become true relative variances of ordinates.

4. RESULTS

We used Weibill, F, Exponential, t, Beta, Chi-square and Normal distribution to find the true relative variances of the ordinates. For a specific distribution, if we find the true relative variances of the ordinates at some defined member of the family it will provide unique true relative variances of the ordinates. By this our main idea becomes patent that a distribution can be easily identified by comparing the true relative variances of the ordinates by its sample relative variances of the ordinates. Hence the member of a family can be easily identified by this comparison.

The true relative variances of the ordinates along the value of all distributions are put together in a column and then sorted at three levels. At first level the true relative variances of the ordinates are sorted from maximum to minimum. At second level the true relative variances of the ordinates are sorted with respect to first parameter. At third level both of these sorted levels are again sorted with respect to second parameter (this sorting level is for those distributions which have two parameters).

The main purpose of this type of sorting is to identify the groups of the true relative variances of the ordinates. These groups show the range(s) of distributions. As mentioned above these range(s) are denoted by different digit.

4.1 Range(s) of Relative Variances of Ordinates of Distributions

From true relative variance(s) of the ordinates the groups of different distribution are identified. Each group (which may be based on different ranges) showed a particular distribution. When sample relative variances of the ordinates lie in a particular group then the distribution of that sample is the distribution represented by that group. Range(s) of the groups are shown in **(Table 3)** for different distributions.

Table 3: Range(s) of relative variances of ordinates for different distribution

Distribution		LL	UL	% lie in groups
Weibull	1st range	0.000799026642800000	0.000999000999000000	6.730
	2nd range	0.000539181445600000	0.000539929722900000	0.090
	3rd range	0.000513910194700000	0.000514680861500000	0.060
	4th range	0.000501254011200000	0.000502659394700000	0.110
	5th range	0.000499185074600000	0.000499816512600000	0.060
	6th range	0.000492274361500000	0.000496190226200000	0.260
	7th range	0.000485346212600000	0.000485490687400000	0.030
	8th range	0.000483561137500000	0.000483857892200000	0.050
	9th range	0.000451286675500000	0.000474359391600000	1.460
	10th range	0.000424245008200000	0.000442538923300000	1.100
	11th range	0.000420032860400000	0.000420152695200000	0.040
	12th range	0.000411712600100000	0.000416876701300000	0.410
	13th range	0.000386798000200000	0.000402493951500000	1.040
	14th range	0.000363322809600000	0.000383945457200000	1.290
	15th range	0.000315402729000000	0.000357008283800000	2.840
	16th range	0.000301446997000000	0.000310487777000000	0.810
	17th range	0.000254221910700000	0.000297410597600000	4.030
	18th range	0.000214976459800000	0.000251689584500000	4.670
	19th range	0.000100001411700000	0.000213087456500000	31.820
	20th range	0.000099816985300000	0.000099892253400000	0.040
	21st range	0.000099726389300000	0.000099769025100000	0.030
	22nd range	0.000099611684600000	0.000099691247700000	0.050
	23rd range	0.000099511377200000	0.000099594579500000	0.050
	24th range	0.000099410783800000	0.000099484455200000	0.040
	25th range	0.000099308294600000	0.000099390974200000	0.040
	26th range	0.000099205937300000	0.000099280414600000	0.040
	27th range	0.000099147342900000	0.000099193123100000	0.050
	28th range	0.000075305235400000	0.000098917522200000	10.950
	29th range	0.000035184707800000	0.000075103478100000	17.430
	30th range	0.000034669412800000	0.000034693350000000	0.110
	31st range	0.000016595989700000	0.000034464310000000	9.570
	32nd range	0.000014542388500000	0.000014608859900000	0.070
F	1st range	0.000540357003227300	0.000798901548934400	88.710
	2nd range	0.000514711688181300	0.000539038512989100	2.620
	3rd range	0.000502778188561800	0.000513654277850700	1.190
	4th range	0.000499838534476700	0.000501137632631800	0.110
	5th range	0.000496289312814800	0.000498873858541200	0.150
	6th range	0.000485594016245500	0.000492258249467100	0.750
	7th range	0.000484030081367500	0.000485300123224900	0.110
	8th range	0.000474468428038700	0.000483481294438100	1.220
	9th range	0.000442679112129500	0.000450982286529300	1.090
	10th range	0.000420180410988700	0.000424210487080900	0.490
	11th range	0.000417002962342300	0.000419977464319400	0.270
	12th range	0.000402715648655300	0.000411514963287300	1.270
	13th range	0.000383969189995400	0.000386777006183400	0.940
	14th range	0.000357049365253800	0.000363285442766300	1.060
	15th range	0.000310490278993500	0.000315350832622300	0.750
	16th range	0.000297449909211700	0.000301394557427200	1.080
	17th range	0.000251700387004200	0.000254183718752000	0.900
	18th range	0.000213120012479600	0.000214945511730200	0.970
	19th range	0.000075105157371100	0.000075305018132600	0.880
Exponential	1st range	0.000099900000000000	0.000100000000000000	86.320
	2nd range	0.000099800000000000	0.000099800000000000	1.080
	3rd range	0.000099700000000000	0.000099700000000000	0.660
	4th range	0.000099600000000000	0.000099600000000000	0.540
	5th range	0.000099500000000000	0.000099500000000000	0.420
	6th range	0.000099400000000000	0.000099400000000000	0.300
	7th range	0.000099300000000000	0.000099300000000000	0.300
	8th range	0.000099200000000000	0.000099200000000000	0.300
	9th range	0.000099000000000000	0.000099100000000000	0.480
T	1st range	0.000034702062761734	0.000035177332825988	97.800
	2nd range	0.000034472455291941	0.000034665957266473	1.600
Beta	1st range	0.000014614251689762	0.000016593730728163	1.850
	2nd range	0.000001275098246734	0.000014524412882033	101.720
Chi-square	1st range	0.000000231000000000	0.000001260000000000	117.100
Normal	1st range	0.000000001260000000	0.000000231000000000	88.050

4.2 Some Clarifications

There is needed to be some clarification for the above defined groups and there range(s). The percentage of the parameter lies in a specific group or over all groups may be more or less than 100%. The fact behind this issue is the presence of true relative variances of the ordinates of some other distribution in a specific group or range(s). For the Weibull distribution the true relative variance of the ordinates is approximately 4.62% which lie in some other groups.

As we discuss that there are other true relative variances of the ordinates that lie in a group and show some specific distribution. So each group having true relative variances of the ordinates of some other group(s) is treated as the group of that distribution for which the true relative variances of the ordinates are in majority. In other words, true relative variances of the ordinates of other group(s) which are not makes some patterns are ignored to justify that group for a specific distribution.

4.3 Sample based Results

Here we present the sample results calculate by using our sampling procedures. The sample size is taken as 50 because it is not considered as too sample or too large. The number of the classes is taken as 60% of the whole data. The results for some distributions are shown below.

4.4 Sample Results for Pre-Assumed Weibull Distribution

Figures 1(a) and 1(b) shows the sample results for pre-assumed Weibull distributed data. Procedure to define the sample relative variances of ordinates is presented in **Fig-1(A).**

```
> n=50
> sim=1000
> l=round(0.6*n)
> x=rweibull(n,6.4,59.5)
> b=matrix(NA,sim,l)
> ord=matrix(NA,sim,l-1)
> for(i in 1:sim){
+   rn=sample(x,n,T)
+   b[i,]=seq(min(rn),max(rn),length=l)
+   dcut = cut(rn, b[i,])
+   f=as.vector(table(dcut))
+   ord[i,]=(f)/(sum(f))
+ }
> bs=apply(b,2,mean)
> ords=apply(ord,2,mean)
> var(ords)
[1] 0.0004271121
> |
```

Fig. 1(a): First Sample relative variances of the ordinates for Weibull distribution

The value of sample relative variances of the ordinates 0.0004271121 lie in the group of Weibull distribution in the tenth range. Against these sample relative variances of the ordinates the value of true population parameter is 27.1 and 91 respectively. That is not accurate but provided us an idea about the true population parameters.

```
> n=50
> sim=1000
> l=round(0.6*n)
> x=rweibull(n,6.4,11.5)
> b=matrix(NA,sim,l)
> ord=matrix(NA,sim,l-1)
> for(i in 1:sim){
+   rn=sample(x,n,T)
+   b[i,]=seq(min(rn),max(rn),length=l)
+   dcut = cut(rn, b[i,])
+   f=as.vector(table(dcut))
+   ord[i,]=(f)/(sum(f))
+ }
> bs=apply(b,2,mean)
> ords=apply(ord,2,mean)
> var(ords)
[1] 0.000867367
> |
```

Fig,1(b): Second Sample Relative variances of the ordinates for Weibull distribution

Again one can verify that the value 0.000867367 lies in the group of Weibull distribution in the first range. Against these sample relative variances of the ordinates the value of true population parameter is 9.1 and 55.9 respectively.

4.5 Sample Results for Pre-Assumed F Distribution

Fig. 2(a) and 2(b) shows the sample results for pre-assumed F distributed data. Procedure to define the sample relative variances of ordinates is presented in **Fig.2(A)**.

```
> n=50
> sim=1000
> l=round(0.6*n)
> x=rf(n,77,197)
> b=matrix(NA,sim,l)
> ord=matrix(NA,sim,l-1)
> for(i in 1:sim){
+   rn=sample(x,n,T)
+   b[i,]=seq(min(rn),max(rn),length=l)
+   dcut = cut(rn, b[i,])
+   f=as.vector(table(dcut))
+   ord[i,]=(f)/(sum(f))
+ }
> bs=apply(b,2,mean)
> ords=apply(ord,2,mean)
> var(ords)
[1] 0.000359847
> |
```

Fig. 2(a): First Sample relative variances of the ordinates for F distribution

The value of sample relative variance of the ordinates is 0.000359847 which lie in the fourteenth range of the true parameters. The true relative variances of the ordinates that are close to these sample relative variances of the ordinates are 7 and 99 respectively.

These values provide us an idea about true population parameters.

```
> n=50
> sim=1000
> l=round(0.6*n)
> x=rf(n,173,45)
> b=matrix(NA,sim,l)
> ord=matrix(NA,sim,l-1)
> for(i in 1:sim){
+   rn=sample(x,n,T)
+   b[i,]=seq(min(rn),max(rn),length=l)
+   dcut = cut(rn, b[i,])
+   f=as.vector(table(dcut))
+   ord[i,]=(f)/(sum(f))
+ }
> bs=apply(b,2,mean)
> ords=apply(ord,2,mean)
> var(ords)
[1] 0.0004846664
>
```

Fig.2(b): First Sample relative variances of the ordinates for F distribution

Another sample is also show the sample relative variances of the ordinate is 0.0004846664.It is verified that the sample relative variances of the ordinates lie in the 9th range of the group and the values of true population parameters against the true relative variances of the ordinates are closest to sample relative variances of the ordinates i.e. 119 and 9 respectively.

5. CONCLUSIONS

After applying the proposed approach a new way to identify the family of continuous probability distributions is established which is not exactly based upon subjective approach and do not need statistical expertise, a method to calculate the sample relative variances of the ordinates is developed and the comparison of true relative variances of the ordinates and sample relative variances of the ordinates gives us an idea about the parameters of that distribution.

Future Openings

Further possible extensions the current work are:

- Application of the proposed approach on discrete probability distributions.
- Develop a procedure to estimate the width of class interval for a given sample of a distribution.
- Develop an improved method to get the sample relative variances of the ordinates

REFERENCES:

Mayooran, T. and A. Laheetharan, (2013). The Statistical Distribution of Annual Maximum Rainfall in Colombo District. Sri Lankan Journal of Applied Statistics, 15(2). 1023-1032.

Su, C. T. and C. J. Chou (2007) A Neural Network-Based Approach for Statistical Probability Distribution Recognition. Quality Engneering, 8(3) 293-297.

Webb, A. R., (2003). Statistical pattern recognition. John Wiley & Sons.