International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Investigating the Least Sample Size for Convergence to Normality from Quantile Measure of Continuous Distributions with R

Soumyadip Das[1], Arjun Dutta[2]

[1, 2]Department of Statistics, Kalyani University, India

Abstract: ***For testing of hypothesis, the knowledge of the distribution of the test statistic is necessary to find the cut-off point or the p-value. But in most of the cases, the distribution of quantile measures of samples is not known or not standard or complicated. Thus, for testing the value of quantile measures the common practice is to use the asymptotic distribution of the statistic which is normal in general. But for this asymptotic distribution to be accurate the sample size must be large. Now the question is how large the sample size should be to ensure the convergence of the statistic to anormal distribution. This paper proposes a procedure to find the least Sample size required for some selected continuous distributions to converge to normality using the Shapiro-Wilk's Test of Normality. Simulation and Visualisation are done in the R programming language.***

Keywords: Continuous Distributions, Central Limit Theorem, Convergence, Least Sample Size, Shapiro Wilk's test,R

1. Introduction

The **Central Limit Theorem** states that for most commonly studied scenario, when independent random variables are added their sum tends toward normal distribution even if the variables are not normally distributed. Now our questions are "*What represents a Large Sample*?", "*How many measures we have to compare in order to claim the sample is large*?" Each measure will help in detecting the sample size. Using the programming language "R" we generate r random samples each of size n and find the values of the quantile measures in all the cases. Thus we have r values of a measure. Then we perform **Shapiro-Wilk's Test** of normality to check whether those r values of the statistic can be considered to be a random sample from a normal distribution with suitable mean and variance.We start a loop with n=2 and continue increasing the value of n by 1 in each loop until the test for normality is accepted. The step where the test is accepted, the loop is broken and that value of n is our target. **R** is an Open Source Programming Language and Environment for Statistical Computing and Graphics. It is quite similar to the S language and environment which was developed at Bells Laboratory (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues, but much code written for S runs unaltered under R. It has numerous applications in the field of data analysis and widely used by statisticians and data miners. Along with a command line interface, it has several graphic front-ends. R is extensible through functions, extensions and packages, contributed by the global R community.As of 2016, 10874 additional packages are available for installation. **R-Studio** is a powerful and productive **Integrated Development Environment (IDE)** for R. The software is written in C++ programming and uses Qt framework for graphical user interface. It supports direct code execution as well as tools for statistical analysis, debugging and workspace management. It can manage multiple working Directories and also have an extensive package development environment. The paper covers the **Related Works** of apre-published research paper related to sample size required for Convergence of Normality. **Objective** covers what the problem is about. **Methodology** covers about a procedure of finding the least sample size. **Simulation** describes the R code used to find the least sample size; this section also contains related graphs. **Result** describes the finding of least sample size for different continuous distributions.

2. Related Works

Author[**Carsten Schröder**] and [**Shlomo Yitzhaki**] [1] has proposed a procedure based on the properties of Gini's mean difference (**GMD**) by **Gini** (1914, 1921). The GMD is a Variability measure which derives the Gini Coefficient and asymmetric correlation associated with it. They have used the property of Gini correlation to test for convergence to normality because if the convergence to normality occurred then Gini correlation should be equal and they have used this property to extract the reasonable sample size for convergence to normality.

3. Objective

Here our objective is to find out the least sample size required for the distribution of standard quantile measures for examples,

- **Median** (For location)
- **Quartile deviation** (For dispersion)
- **Coefficient of QuartileDeviation** (For relative dispersion)
- **Bowley's Measure of Skewness**
- **The Kp Measure**

To converge to anormal distribution with some mean and variance by simulation we start with repeated random samples from few standard distributions and find the required value of the sample size for the distributions of the above statistics to converge to normality, if at all.

Here we select one continuous probability distribution. It is Exponential Distributionand we test for the Median, Quartile

Deviation, Coefficient of Quartile Deviation, Bowley's Measure of Skewness and Kurtosis and Kp measure.Here we generate "r" random sample each of size "n" and find the values of the quantile measures in all the class. Then we perform Shapiro-Wilk's test of normality to check whether "r" values of the statistic can be considered to be arandom sample from a Normal distribution with suitable mean and variance.

We show that for six Continuous Distribution.

The Distribution we take into consideration are:-

a) Exponential
b) Lognormal
c) Cauchy
d) Beta
e) Normal
f) Rectangular

4. Methodology

- Using R-Studio script we generate "r" random sample of size n & find the values of quantile measures of all the cases.
- We perform Shapiro-Wilk'stest of normality to check whether these "r" values of the statistic can be considered to be a random sample.
- We start with loop n=2 and continue increasing the value of "n" by 1 in each loop until the test of normality is accepted i.e. (p_value < α). The step where the test is accepted or where the condition above is satisfied the loop is broken and that value of "n"is our target.
- Here we take, r =1000 and the level of significance for the Shapiro-Wilk's test to be α=0.05
- Here in most of the cases, it's obvious that for any sequence of random variables Tn which variance exists, (Tn - E (Tn))/V (Tn) N (0, 1) ..... {By law}.

Now the main theorems and results used for the Methods of finding least sample size of convergence are given below with short description:-

- **Convergence in law:**
  Let {Xn} be a sequence of random variable with Fn(X) as the c.d.f of Xn (n=1, 2, 3). Suppose further that Fn(X) be the c.d.f of a random variable X.
  If limFn(X) = F(X) at all continuity points X of F(X) then {Xn} is said to be convergence in law or the convergence in distribution to X. The c.d.f F is called the asymptotic distribution or the limiting distribution of sequence {Xn}.
- **Shapiro-Wilk's test:**
  It has recently been extended to cope with samples of size up to 2000(Royston, 1982a). The purpose of the present algorithm is to enable the calculation of the W and the significance level of any sample size between 3 and 2000. The full description of the theory behind this algorithm is given by Royston (1982a). Using Monte Carlo simulation Royston shows that the transformation is.....

$$y = (1-W)^{\lambda}$$

The mean mu and standard deviation σ_y of the transform y were calculated using the smoothed λ and their logarithms were themselves smoothed with polynomials in ln (n)-d. Given the value of the W, therefore, the significance level of the calculating by referring the quantity

$$z = [(1-W)^{\lambda} - \mu]/ \sigma_y$$

To the upper tail of the standard normal distribution since alarge value of z indicates non-normality of the original sample.

The significance level of W for n =3 is exact and for the 4<n<6 is calculated by adapting Table 1 of Wilk and Shapiro (1968). Full details of all procedure are given by Royston.

The Shapiro-Wilks test for normality is one of three general normality tests designed to detect all departures from normality. It is comparable in power to the other two tests.

The test rejects the hypothesis of normality when the p-value is less than or equal to 0.05. Failing the normality test allows you to state with 95% confidence the data does not fit the normal distribution. Passing the normality test only allows you to state no significant departure from normality was found.

The Shapiro-Wilks test is not as affected by ties as the Anderson-Darling test, but is still affected. The Skewness-Kurtosis All test is not affected by ties and thus the default test.

Quartiles: The quantiles which divide the whole frequency distribution into FOUR equal parts are called quartiles.
$Z_{.25} = Q_1$, $Z_{.5} = Q_2$, $Z_{.75} = Q_3$

Bowley's measure of skewness:
((q3-q2)-(q2-q1))/(2*QD)
Measure of kurtosis (Kp):
$Kp = \{(P_{75}-P_{25})/2\}/ \{P_{90}-P_{10}\}$ Or $(Q_3-Q_1)/ \{2(P_{90}-P_{10})\}$
Kp = .263 for Mesokurtic Distribution
>.263 for Platykurtic Distribution
<.263 for Leptokurtic Distribution

5. Simulation

Let's start with some step by step execution of code below. We start by choosing the distribution (in here we choose Exponential Distribution with rate 1) and then take a dive by calculating the least sample size for different Quantile measure.

Set the parameters,level of significance and number of samples to be taken

```
> m=1              # Set the parameters
> alpha= 0.05    # Set the level of significance
> r= 1000       # Set number of samples to be taken
> n=2
> pv=0
```

- **Computing the Median and plotting the Graph**

```
> while(pv<alpha)
+ {
+ med=vector()
+ for(i in 1:r)
+   med[i]=median(rexp(n,m))
+ pv=as.numeric(shapiro.test(med)[2])
+ hist((med-mean(med))/sd(med),freq=F,main=paste("median,n=",n))
+ curve(dnorm,add=T)
+ n=n+1
+ }
```

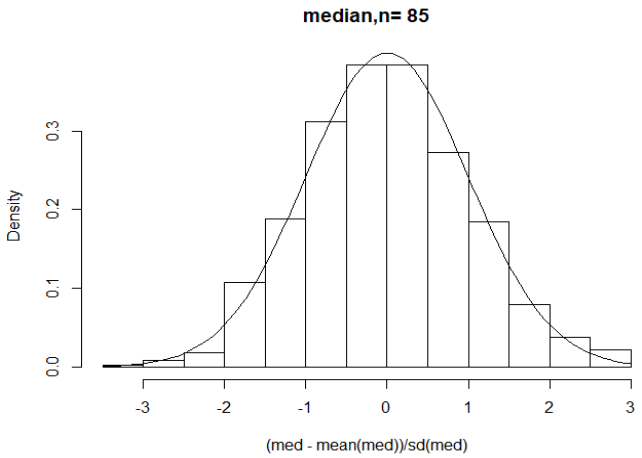

Figure 1: Median Histogram and Fitted Normal Density Curve

Setting the parameters again,Computing the Quartile Deviation and plotting the Graph

```
> r= 1000    # Set number of samples to be taken
> n=2
> pv=0
> while(pv<alpha)
+ {
+   qd=vector()
+   for(i in 1:r)
+   {
+     x=rexp(n,m)
+     qd[i]=(quantile(x,0.75)-quantile(x,0.25))/2
+   }
+   pv=as.numeric(shapiro.test(qd)[2])
+   hist((qd-mean(qd))/sd(qd),freq=F,main=paste("QD,n=",n))
+   curve(dnorm,add=T)
+   n=n+1
+ }
```

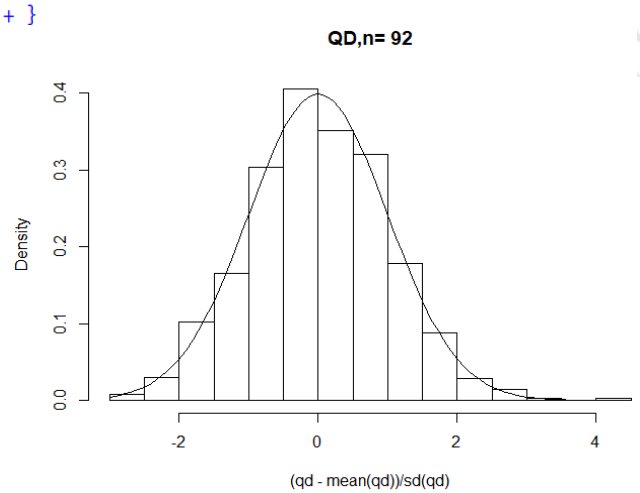

Figure 2: Quartile Deviation Histogram and Fitted Normal Density Curve

Same for Coefficient of Quartile Deviation.

```
> m=1                    # Set the parameters
> alpha= 0.05            # Set the level of significance
> r= 1000     # Set number of samples to be taken
> n=2
> pv=0
> while(pv<alpha)
+ {
+   cqd=vector()
+   for(i in 1:r)
+   {
+     x=rexp(n,m)
+     cqd[i]=as.numeric(((quantile(x,0.75)-quantile(x,0.25))/2)/median(x))
+   }
+   pv=as.numeric(shapiro.test(cqd)[2])
+   hist((cqd-mean(cqd))/sd(cqd),freq=F,main=paste("Coeff. of QD,n=",n))
+   curve(dnorm,add=T)
+   n=n+1
+ }
>
```

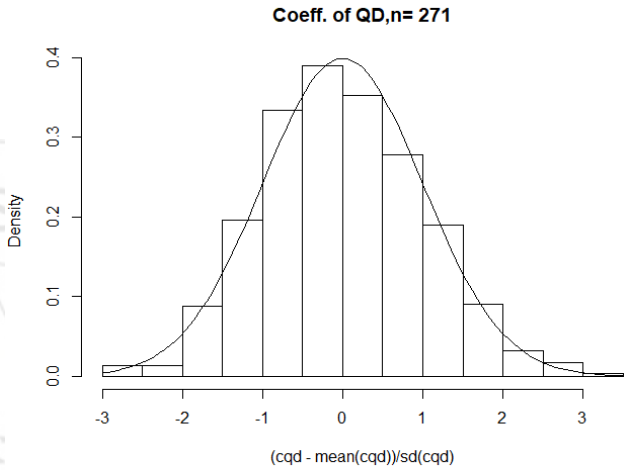

Figure 3: The Coefficient of Quartile Deviation Histogram and Fitted Normal Density Curve

Same for Bowley Measure of Skewness

```
> m=1                    # Set the parameters
> alpha= 0.05           # Set the level of significance
> m=1                    # Set the parameters
> alpha= 0.05           # Set the level of significance
> r= 1000      # Set number of samples to be taken
> n=2
> pv=0
> while(pv<alpha)
+ {
+   sk=vector()
+   for(i in 1:r)
+   {
+     x=rexp(n,m)
+     qd=as.numeric((quantile(x,0.75)-quantile(x,0.25))/2)
+     sk[i]=as.numeric((quantile(x,0.75)+quantile(x,0.25)-2*median(x))/(2*qd))
+   }
+   pv=as.numeric(shapiro.test(sk)[2])
+   hist((sk-mean(sk))/sd(sk),freq=F,main=paste("SK,n=",n))
+   curve(dnorm,add=T)
+   n=n+1
+ }
```

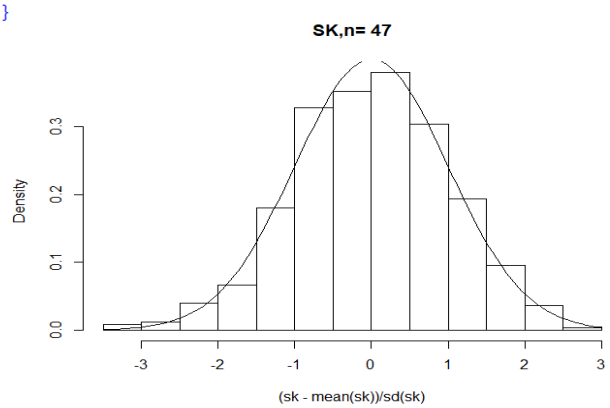

Figure 4: Bowley Measure of Skewness Histogram and Fitted Normal Density Curve.

Same for Percentile Measure of Kurtosis

```
> m=1                          # Set the parameters
> alpha= 0.05                  # Set the level of significance
> r=1000
> n=2
> pv=0
> while(pv<alpha)
+ {
+   kp=vector()
+   for(i in 1:r)
+   {
+     x=rexp(n,m)
+     qd=as.numeric((quantile(x,0.75)-quantile(x,0.25))/2)
+     kp[i]=as.numeric(qd/(quantile(x,0.90)-quantile(x,0.10)))
+   }
+   pv=as.numeric(shapiro.test(kp)[2])
+   hist((kp-mean(kp))/sd(kp),freq=F,main=paste("KP,n=",n))
+   curve(dnorm,add=T)
+   n=n+1
+ }
```

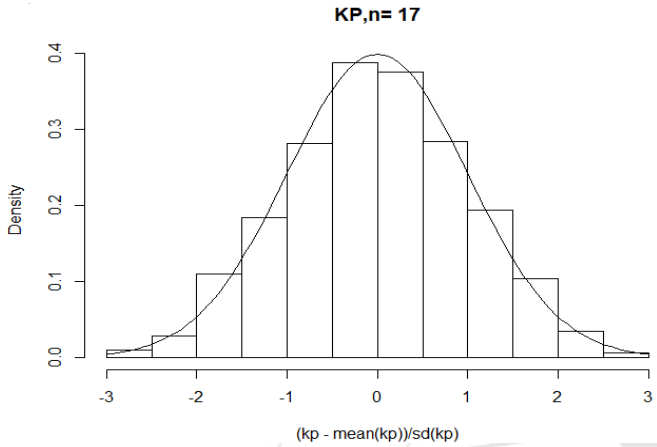

Figure 5: Percentile Measure of Kurtosis Histogram and Fitted Normal Density Curve

In this section, we have only worked for exponential distribution but we have shown the least sample size for different continuous distributions in the Results section.You can download the whole R script from GitHub and execute it line by line in R Studio.

The link is given below:-

https://raw.githubusercontent.com/DuttaArjun/Least-Sample-Size-/master/Least%20Sample%20Size%20for%20Normality/R%20code.R

Another way you can view this is through R shiny application which I have created for the ease of people who don't want to execute the R Script line by line and pretty much focused on the output. To do this open up R Studio or R Gui and execute this three line of commands in the console.

```
>install.packages("shiny")
>library(shiny)
>runGitHub("DuttaArjun/Least-Sample-Size-Shiny",
"app.R", subdir = "LeastSampleSize/app.R")
```

(P.S:- You need a sound internet connection for both of the case above.)

6. Results

By executing the R codes for some fixed or standard values of the parameters of some continuous distributions like Exponential, Log Normal, Cauchy, Normal, Beta, Rectangular Distributions we get the following observations. For all the continuous distributions in most of the cases, the distributions of the statistics have converged to normal distributions for some values of n. The following table shows the least required value of n for them to converge for aparticular execution of the code:

Table 1: Values of n for different Distribution and their Quantile Measure

	Median	Q.D	C.Q.D	B.M.S	Kp
Exp(1)	80	105	234	58	15
B(1,1)	9	20	101	32	25
LN(0,1)	77	216	315	94	26
C(0,1)	26	201	D.N.C	27	52
N(0,1)	3	35	D.N.C	19	7
β(2,5)	19	3	125	21	9
β(0.5,0.5)	27	3	426	26	112
β(2,2)	6	4	110	23	7
U(0,1)	11	32	120	43	27

Footnote: The value in the (i, j)th cell represents the required value of n for convergence of the jth measure of ith distribution. D.N.C. -> Does not converge to normality. Further investigation is needed.

7. Conclusion

The theorem for convergence of the distribution of the pth quantile of arandom sample from some distribution to normal distribution holds good only for the continuous densities. The distribution of the all Quantile measure has converged to N (0, 1) for the definite value of n.

For extremely positively skewed distribution like an exponential, Log-normal distribution the least sample size required the much higher value for all measure & for symmetric distribution like normal or Cauchy, through the least sample size required for all the measurements are low, the distribution of the coefficient of quartile deviation does not converge for Normal and Cauchy distribution. From the graph it is apparent that the coefficient of quartile deviation is converging to a point, still, further investigation into this is necessary.

References

[1] Carsten Schröder and Shlomo Yitzhaki, Reasonable sample sizes for convergence to normality, SOEP — The German Socio-Economic Panel Study at DIW Berlin, https://www.diw.de/documents/publikationen/73/diw_01.c.492474.de/diw_sp0714.pdf
[2] Element of Large Sample Theory by E.L. LEHMANN, http://pointer.esalq.usp.br/departamentos/lce/arquivos/aulas/2011/LCE5866/Springer_-_E.L.Lehmann_-_Elements_of_Large-sample_Theory.pdf
[3] Asymptotic Statistics by A.W. VANDER VAART, https://books.google.co.in/books?hl=en&lr=&id=UEuQE

M5RjWgC&oi=fnd&pg=PR13&dq=asymptotic+statistics+van+der+vaart&ots=mnWJTBf2Px&sig=0F_L_W2W2ubxXzbtgpyB_0TFusc#v=onepage&q=asymptotic%20statistics%20van%20der%20vaart&f=false

[4] Approximation Theorems of Mathematical Statisticshttps://books.google.co.in/books?hl=en&lr=&id=enUouJ4EHzQC&oi=fnd&pg=PP2&dq=Approximation+theorem+of+mathematical+statistics+by+robert+serfling&ots=ehVGyKgbW6&sig=cAe8ezE9tialS5naqI0Py_U6Oc#v=onepage&q=Approximation%20theorem%20of%20mathematical%20statistics%20by%20robert%20serfling&f=false

[5] An Introduction to R by W. N. Venables, D. M. Smith https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

Author Profile

Soumyadip Das received his B.Sc (Hons) in Statistics from St. Xavier CollegeKolkata (2016) and currently pursuing his Masters in Statistics from Kalyani University.

Arjun Dutta received his B.Sc (Hons) in Statistics from Asutosh College Kolkata (2016 and currently pursuing his Masters in Statistics from Kalyani University.