

Cricket Data Analysis

Sunday, 10 February 2013

R® for Fitting univariate Parametric Distributions to Cricketers Oneday

M.R.L.N. Panchanana (M.Sc Statistics from University of Hyderabad, India)
M.Tech (Computer Science & Technology from Jawaharlal Nehru University, Delhi)

Abstract

Fitting a parametric distribution to a data set is a key step in analyzing the data. In many statistical procedures like ANOVA, several hypothesis tests are performed assuming the data as normal. In this presentation, it is discussed how to perform statistical techniques using R Software, to fit a statistical distribution to oneday cricket Score made by Sachin Ramess Tendulkar. DNB (Does Not Batted) cases are not considered.

Introduction

Few functions used in R® is useful in deciding, which Univariate distribution is suitable to the specified variable from the data set. It provides Univariate discrete and continuous distributions. Before going to perform, goodness of fit test, it is advisable to know the characteristics of the data by descriptive statistics methods like **summary** and **stem** and Density estimation methods like **hist**, **boxplot**, **ecdf**, **density**, **fitdistr**, **ks.test**. Graphical illustrative methods are **qqplot** and **qqline**.

Oneday cricket Runs

The data for illustrative examples is collected from <http://www.howstat.com.au/>. Only the oneday runs made by few famous batsmen are used for the analysis. The data is saved in playername.sas7bdat files. Primarily “Sachin Tendulkar’s oneday match runs are analyzed.

Open R Software and load the package sas7bdat to import sas data files.

```
> library(sas7bdat)
```

Read the data file as follows:

```
SACHIN <- read.sas7bdat("D:/Modeling/sachin.sas7bdat")
```

The following commands invoke the function and computes mean, median, and different percentiles.

```
> summary(SACHIN)
```

RUNS

```
Min.      : 0.00  
1st Qu.   : 8.00  
Median    : 28.50  
Mean      : 40.77  
3rd Qu.   : 63.00  
Max.      : 200.00
```

Sachin's Average is around 40 runs. Here, we considered every match whether he is out or not out. With quantile function, distribution of runs at each percentile can be observed.

```
> quantile(sachin, seq(0, 1, by=.1))
```

```
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%  
0.0 2.0 5.0 11.0 18.4 28.5 39.0 53.0 72.8 100.0 200.0
```

To have a look at the shape of the distribution of the Runs, Stem and leaf plot can be generated with the following statements.

```
> sachin <- as.numeric(SACHIN$RUNS)  
> stem(sachin)
```

```
0 | 00000000000000000000111111111111111122222222222222223333333333+46  
1 | 0000001111111111222223334444444555555566667777778888889999  
2 | 0001111111122223334444555566677777888888999  
3 | 000000111122222344455555666666777788889999999  
4 | 0000011112334444555556777888899  
5 | 001222333334444557777  
6 | 01122223333455555677788999  
7 | 0012347789  
8 | 011122223455678899  
9 | 011333334556778999  
10 | 000000112455  
11 | 001234457788  
12 | 002234778  
13 | 4789  
14 | 0111366  
15 | 2  
16 | 3  
17 | 5  
18 | 6  
19 |  
20 | 0
```

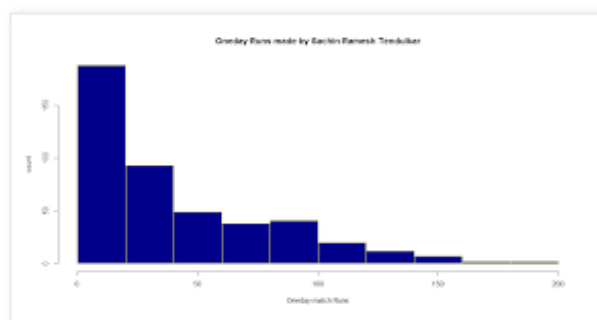
The Graph shows the runs made by Sachin in nice format. For Example, Sachin made 200 runs is shown in the lowest part and he did not make any runs in 190s.

Histograms and Nonparametric density estimation methods

Histogram is a density estimation method for continuous variable, where each bin contains the observations. Bin is an interval, which partitions range of observations into intervals. The following is the input.

```
> hist(sachin, col="darkblue", border = "yellow",  
+ xlab="Oneday match Runs",  
+ ylab="count",  
+ main="Oneday Runs made by Sachin Ramesh Tendulkar")
```

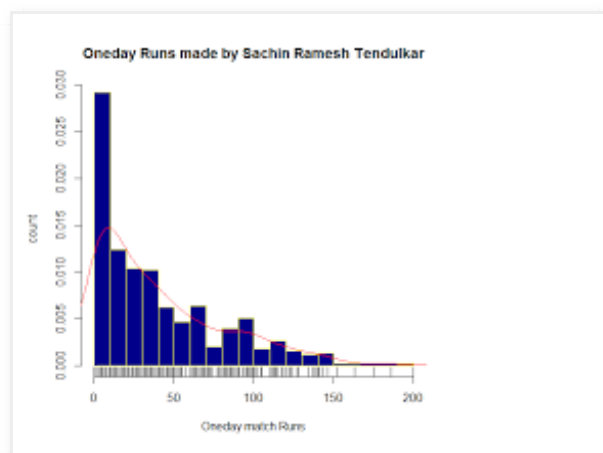
The output is:



Since the variable RUNS is greater than or equal to 0 and the height of the first bar is high and you can observe the gradual decline in height of the bars from second bar onwards. But, it is not forgettable that the histogram depends on the width of the bar and number of bars. If number of bars is increased, histogram is plotted with small intervals by decreasing the smoothness. At this point it is unable to decide what type of function it is. If we decrease the number of intervals to 1, then the function is uniform. R Software gives an option `breaks=` to the user to change the numbers of bars.

```
> bins=c(seq(0,150,by=10),200)
> hist(sachin,breaks=bins,col="darkblue", border = " yellow ", xlab="Oneday match
Runs", ylab="count", main="Oneday Runs made by Sachin Ramesh Tendulkar")
> lines(density(sachin), col= "red")
> rug(sachin)
```

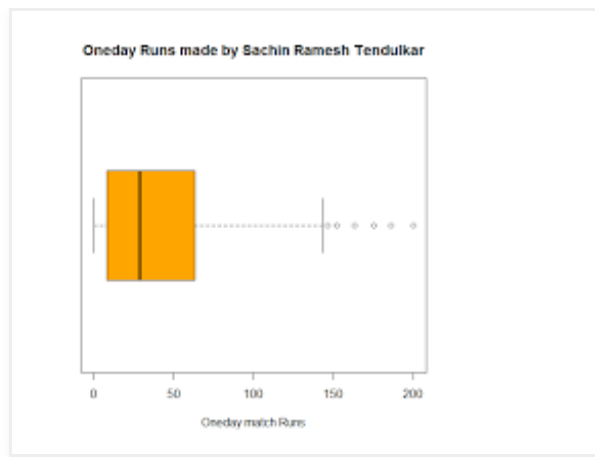
The Output is as follows:



You can observe that there is difference in the shape of the histogram by giving the bins as 0-10, 11-20, ...,150-200. With `rugs()` function you can observe the data points on X-axis. You can clearly identify the runs mad in each interval, particularly 100+ runs.

Before deciding the shape of the density function generate box plot. These are useful in deciding the symmetric and asymmetric shapes of the function of the existing data.

```
> boxplot(sachin, horizontal = TRUE, col = "orange", xlab="Oneday match Runs",main="Oneday Runs
made by Sachin Ramesh Tendulkar")
```



You can observe the Outliers in box plot after 149+ runs.

With the above statistics and graphs, the distribution is not Symmetric. Once again, we can look at summary statistics.

Load the library "fBasics" and look at the Statistics:

```
> library(fBasics)
```

```
> basicStats(sachin)
```

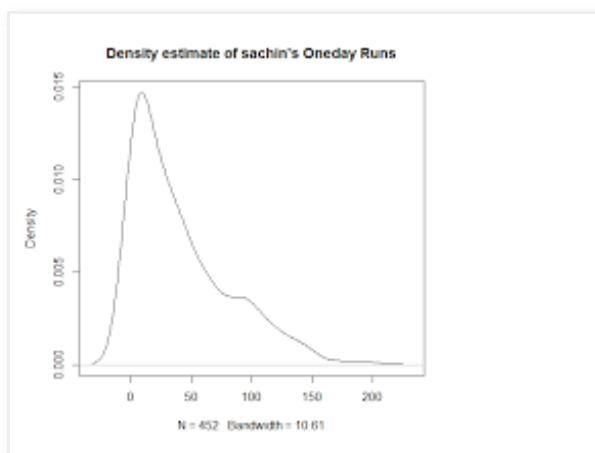
```
      sachin
nobs      452.000000
NAs        0.000000
Minimum     0.000000
Maximum    200.000000
1. Quartile  8.000000
3. Quartile 63.000000
Mean       40.765487
Median     28.500000
Sum       18426.000000
SE Mean    1.883299
LCL Mean   37.064357
UCL Mean   44.466617
Variance   1603.159959
```

Stdev	40.039480
Skewness	1.152907
Kurtosis	0.753945

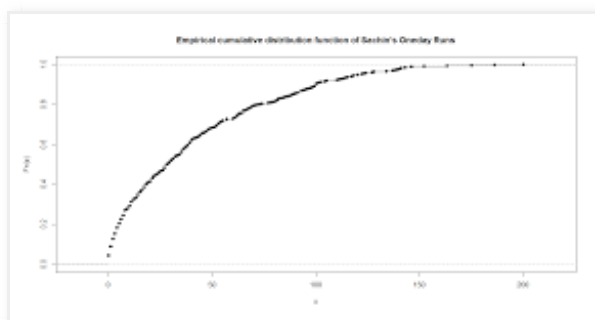
With high variance, mean is almost equal to standard deviation. the distribution is heavily weighted far from the mean. Since skewness > 0 , the distribution is right skewed as shown in the Histogram too, most values are concentrated on left side of the mean, with extreme values to the right. with all the above interpretation, Symmetric distributions will not be a fit to the data.

You can estimate the density function and Cumulative density function as follows:

```
> plot(density(sachin),main="Density estimate of sachin's Oneday Runs")
```



```
> plot(ecdf(sachin),main="Empirical cumulative distribution function of Sachin's Oneday Runs")
```



From all the above, it can be observed that the data is asymmetric and skewed positively. So, You can look at distributions like exponential, weibull, etc.,

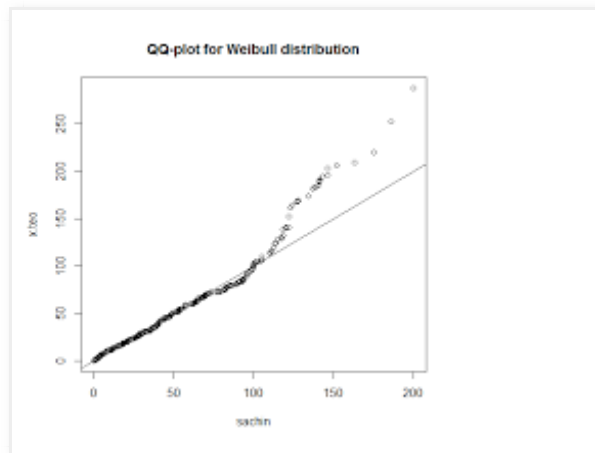
Quantile-Quantile Plots

Before, going to decide the distribution, examine with other graphical procedures, Quantile-Quantile plots and Probability plots. The quantile of a sample is the data point corresponding to a given fraction of the data. A one-sample quantile plot looks like a cumulative sample distribution function. `qqnorm()` is used to test the goodness of fit of a Normal distribution and `qqplot()` for any kind of other distribution. To plot Q-Q plot for variable RUNS, the input is:

```
> x.teo<-rweibull(n=452,shape=0.83908325, scale=37.62877983)
```

```
> qqplot(sachin,x.teo,main="QQ-plot for Weibull distribution")
> abline(0,1)
```

The output is:



fitdistr

Converted 0 RUNS to 0.1; From all the above methods, the variable with positive numbers greater than zero may follow some asymmetric distributions, with more observations in the first intervals and decreasing number of observations in the later intervals.

```
> SACHIN1 <-read.sas7bdat("D:/Modeling/SYSTAT/Cluster/GRAPHS/Rdatastore/sachin1.sas7bdat")
> sachin1 <- as.numeric(SACHIN1$RUNS)
```

We can see, whether any of the weibul or exponential distributions fit to the data with the help of **fitdistr**. Parameters of the distribution can be calculated by maximum likelihood estimation methods. The following statements will fit distributions to the variable RUNS.

```
> library(MASS)
> fitdistr(sachin1,"Exponential")
```

The Output is:

```
rate
0.024527892
(0.001153695)
```

```
> fitdistr(sachin1,"weibull")
```

The Output is:

```
shape    scale
0.83908325 37.62877983
( 0.03253613) ( 2.20929924)
```

Goodness of fit tests

KS test can be conducted for Goodness of fit test.

```
> ks.test(sachin1,"pexp",rate=0.024527892)
```

The output is as follows:

One-sample Kolmogorov-Smirnov test
data: sachin1

D = 0.0917, p-value = 0.0009926

alternative hypothesis: two-sided

```
>ks.test(sachin1,"pweibull",shape=0.83908325 ,scale=37.62877983)
```

The output is as follows:

One-sample Kolmogorov-Smirnov test
data: sachin1
D = 0.0596, p-value = **0.08023**
alternative hypothesis: two-sided

We accept null hypothesis that the data follow a Weibull distribution because the p-value is enough higher than significance levels usually referred. Since shape parameter of weibull is <1, **it is harder to get Sachin out as his score increases.**

In an unpublished work done by me and Prof Traimbakam Krishnan during 2004 to 2007, analysis was done on censored data for several famous players. Not-outs are considered as right censored. The results will be kept in separate Article.

Conclusions

By using function available in R software, we can analyze univariate data very well. Similar kind of Analysis can be done on all famous oneday match players.

References

- [1] Johnson, N. L., Kotz, S., and, Balakrishnan, N. (1994). *Univariate continuous distributions*. Vol. 1, 2nd ed. :New York: John Wiley & Sons.
- [2] Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Univariate continuous distributions*. Vol. 2, 2nd ed. :New York: John Wiley & Sons.
- [3] Johnson, N. L., Kotz, S., and Kemp, A. W. (1993). *Univariate discrete distributions*. 2nd ed.: New York: John Wiley & Sons.
- [4] A. Law and D. Kelton. *Simulation Modelling and Analysis*. McGraw Hill, New York, 1990.
- [5] <http://cran.r-project.org/doc/manuals/R-intro.pdf>

2 comments:



Unknown 18 April 2017 at 11:12

Nice work! How do you handle not-outs if we want to understand probability of failure?

[Reply](#)



Panchanana M R L N  18 April 2017 at 19:24

Not-outs can be considered as right censored and Survival Model will be fitted

[Reply](#)

Enter your comment...

Comment as: Google Account

[Publish](#)

[Preview](#)



[Home](#)



[View web version](#)

About Me



 **Panchanana M R L N**

 [Follow](#) 56

[View my complete profile](#)

Powered by [Blogger](#).