Fitting distributions in R

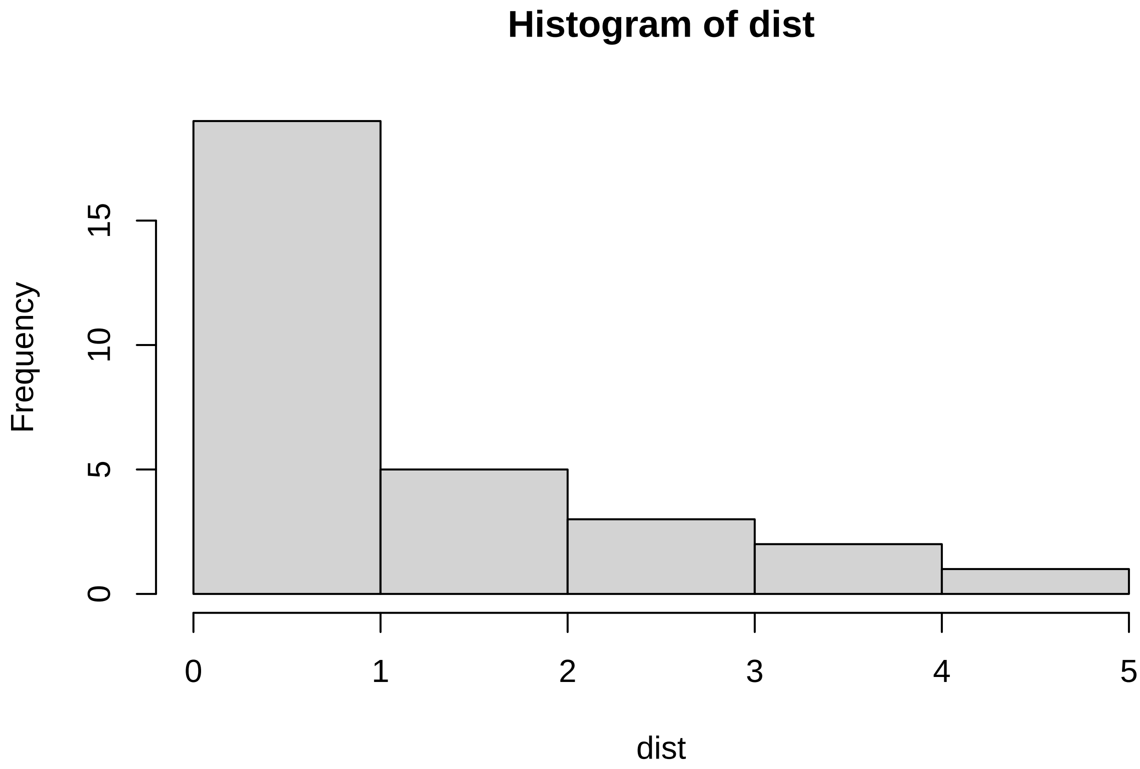
BIshwarup Paul

15/12/2020

Generating data

I am creating a skewed data here for demonstration. In this case, I am generating a distribution with 30 data points from the exponential distribution, and plotting the histogram of the data.

```
dist <- rexp(30, rate=0.75)
hist(dist)
```



Loading required library

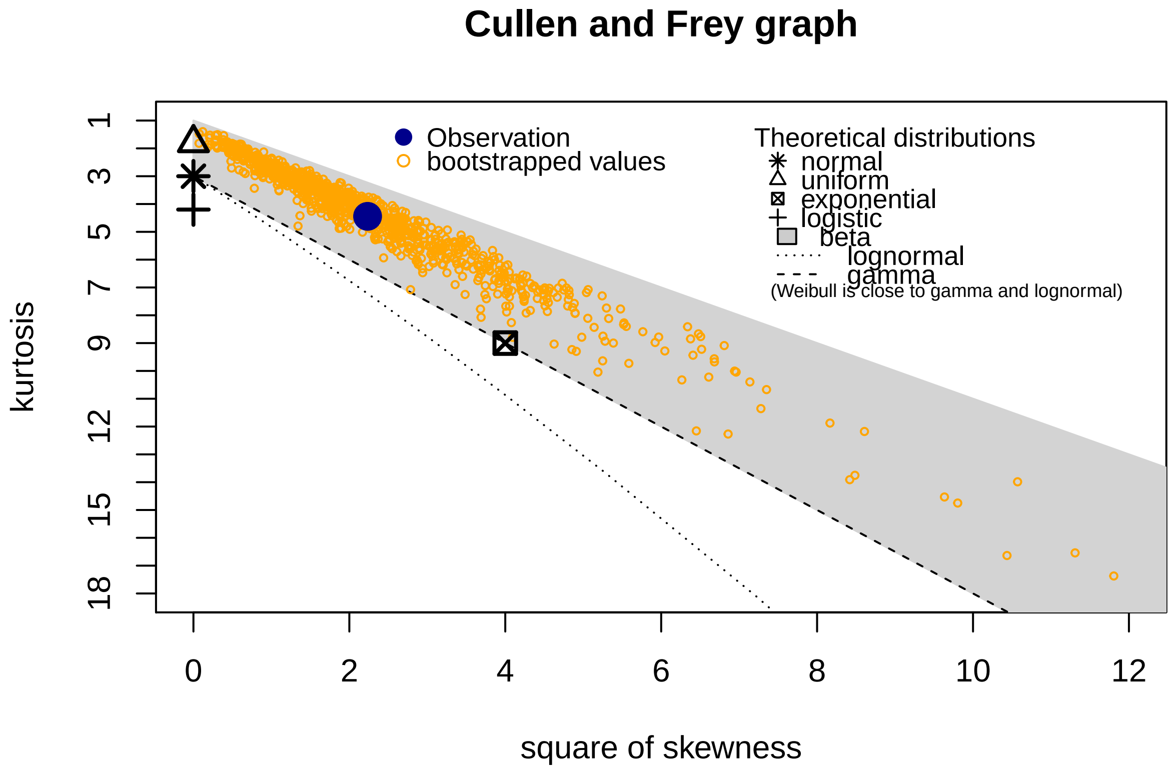
We will be using the package `fitdistrplus`. If you don't have it installed, install it first.

```
library(fitdistrplus)
```

Assessing the distributions - step 1

We will use `descdist` function to assess the distribution. As the data is continuous, we need to provide the argument `discrete=FALSE` in the command (for discrete distributions like Poisson, it will be `TRUE`). As the data size is low, to better assess the distribution, we will ask the command to bootstrap the data to create a large sample size (1000). The command will give us a plot for assessment.

```
descdist(dist, discrete = F, boot=1000)
```



```
## summary statistics
## ------
## min:  0.1122236   max:  4.489916
## median:  0.708646
## mean:  1.204144
## estimated sd:  1.141275
## estimated skewness:  1.494689
## estimated kurtosis:  4.443524
```

In the plot, the blue circle is our data, and the orange circles are bootstrapped points. We see both fall in the grey region, near the dashed line and cross-mark.

The grey region indicates fit to beta distribution - the values of which range between 0-1, thus we won't consider beta distribution here.

The dashed line corresponds to gamma distribution, and the crossed mark corresponds to exponential distribution - a special case of gamma distribution. These two seems to be the viable options.

If the points were centered around the star-mark, then it normal distribution would have been a better fit. If the points were around the dotted line, then lognormal distribution would have been a viable option.

In any case, we will check the fit to all of the options mentioned above.

Assessing the distributions - step 2

Fitting the distributions

We are going to fit the data to the distributions using `fitdist` function.

```r
normal <- fitdist(dist, "norm")
lognormal <- fitdist(dist, "lnorm")
exponential <- fitdist(dist, "exp")
gamma <- fitdist(dist, "gamma")
```

Assessing fit using plots

We are going to plot 4 different types of plots.

1st is a density function plot, providing a density estimation along with histogram of data and theoretical distribution.

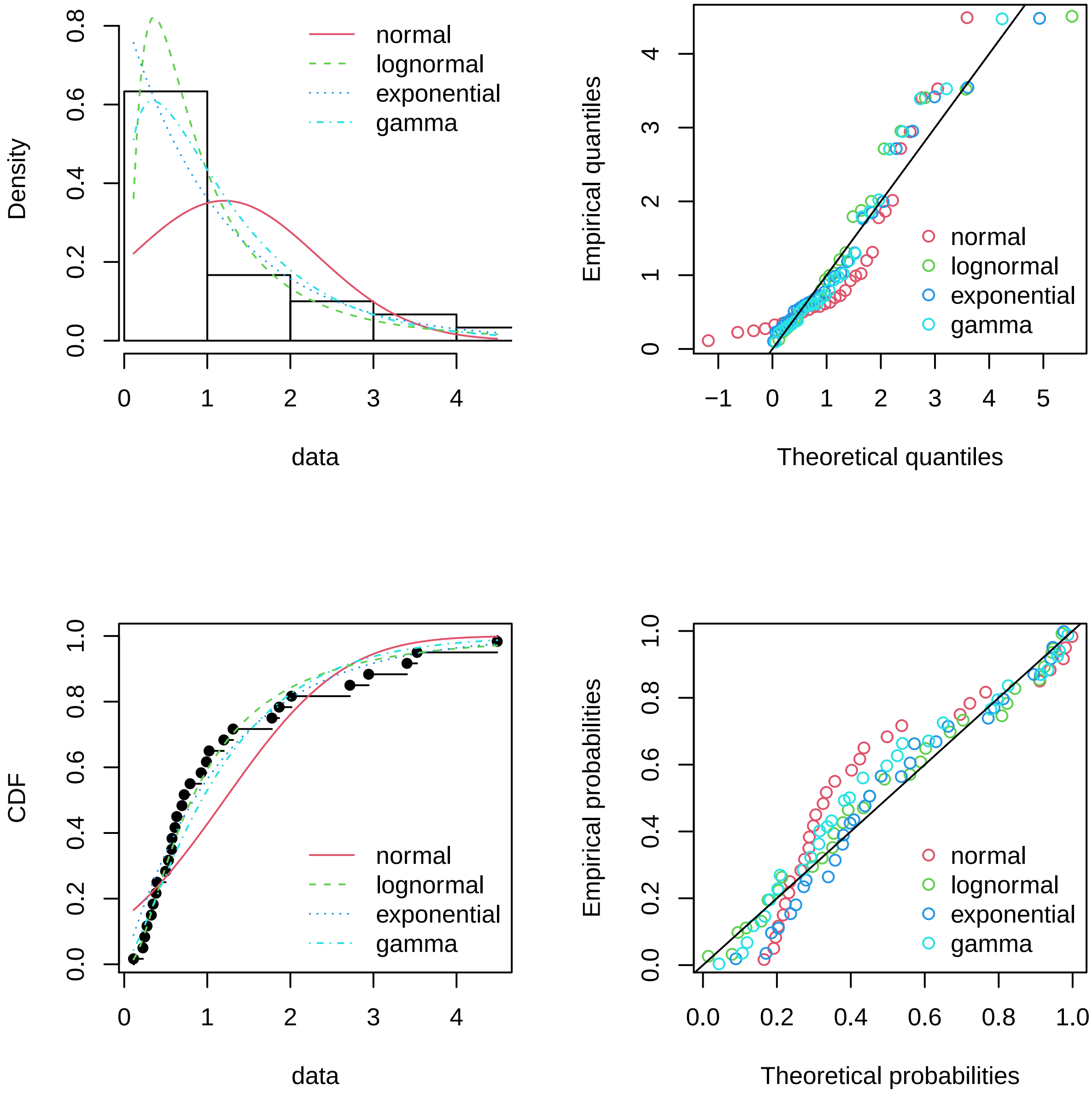
2nd is a Q-Q plot, providing a plot of the quantiles of our distribution along with the quantiles of the theoretical distribution we are comparing with.

3rd is a plot of the cumulative distribution function of the data and theoretical distribution.

4th is a plot of the probabilities of data and theoretical distribution.

```r
plot.legend <- c("normal", "lognormal", "exponential", "gamma")

par(mfrow=c(2,2))
denscomp(list(normal, lognormal, exponential, gamma),
         legendtext = plot.legend, main="")
qqcomp(list(normal, lognormal, exponential, gamma),
       legendtext = plot.legend, main="")
cdfcomp(list(normal, lognormal, exponential, gamma),
        legendtext = plot.legend, main="")
ppcomp(list(normal, lognormal, exponential, gamma),
       legendtext = plot.legend, main="")
```



From the plots above, we see that the exponential and gamma distributions fits the data better, while normal and lognormal deviates to some extent.

Assessing fit using AIC values

```r
aic <- c(normal$aic, lognormal$aic, exponential$aic, gamma$aic)
res <- data.frame(plot.legend, aic)
names(res) <- c("Distribution", "AIC")
```

```
library(knitr)
kable(res, format="simple")
```

Distribution	AIC
normal	96.04803
lognormal	75.27052
exponential	73.14612
gamma	73.28060

From the AIC scores, we see that normal distribution has the worst AIC score. Though the other three are pretty close, exponential has the best AIC, followed by gamma - thus both of them can be said to fit the data well.