IE6200: Engineernig Probability and Statistics

LAB 06: Distribution Fitting

Prof. Mohammad Dehghani

Contents

1 Probability Distribution Fitting

Probability distribution fitting is the procedure of selecting a statistical distribution that best fits to a data set generated by some random processes. The aim of distribution fitting is to predict the probability or to forecast the frequency of occurrence of the magnitude of the phenomenon in a certain interval.

Importance of Probability Distribution Fitting

- Probability distributions can be viewed as a tool for dealing with uncertainty. Distributions can be used to perform specific calculations and apply the results to make well-grounded business decisions.
- In many industries, the use of incorrect models can have serious consequences such as inability to complete tasks or projects in time leading to substantial time and money loss, wrong engineering design resulting in damage of expensive equipment, etc.
- Probability Distribution fitting helps to develop valid models of random processes that can help in avoiding potential time and money loss, and helps to make better business decisions.

Selection of Distribution Fitting

The selection of the appropriate distribution depends on the presence or absence of symmetry of the data set with respect to the mean value.

- **Symmetrical Distributions:** When the data are symmetrically distributed around the mean, one may select the normal distribution, the logistic distribution, or the Student's t-distribution. The Student's t-distribution, with one degree of freedom, has "heavier tails" which means that the values farther away from the mean occur relatively more often (i.e. the kurtosis is higher).
- **Positive Skewness:** When the larger values tend to be farther away from the mean than the smaller values, data is said to have positive skewness (i.e. a skew distribution to the right). One may select the log-normal distribution (i.e. the log values of the data are normally distributed), the log-logistic distribution (i.e. the log values of the data follow a logistic distribution), gamma distribution or the exponential distribution,
- **Negative Skewness:** When the smaller values tend to be farther away from the mean than the larger values, data is said to have positive skewness (i.e. one has a skew distribution to the left). One may select the square-normal distribution (i.e. the normal distribution applied to the square of the data values), the inverted (mirrored) Gumbel distribution, the Dagum distribution (mirrored Burr distribution), or the Gompertz distribution.

After the distributions are fitted, it is necessary to determine how well the distributions you selected fit to your data. It helps to select the most valid model describing your data. This can be done by using:

1. Data Visualization
2. Goodness-of-fit tests

1. Data Visualization:

It is a good practice to visualize the data to get an idea of what distributions are more likely to fit the data as compared to others. This can be achieved by performing explanatory data analysis. This includes:

1. **Descriptive statistics:** It is the analysis of the data that helps to describe, show or summarize the data in a meaningful way such that patterns might emerge from the data. Some of the statistics used to describe data are: mean, median, range, variance, standard deviation, etc.
2. **Graphical methods:** The distribution graphs helps to visually assess the goodness-of-fit of a certain distribution and compare several fitted models. Some of the methods which can be used are:

(a) Frequency distribution histograms
(b) Normal probability plots: PP and QQ plots

2. Goodness-of-fit tests:

The goodness-of-fit tests can be used to determine whether a certain distribution is a good fit. Calculating the goodness-of-fit statistics also helps to order the fitted distributions accordingly to how good they fit to your data. This feature is very helpful for comparing the fitted models. The most commonly used goodness-of-fit tests are *Kolmogorov-Smirnov, Anderson-Darling, Cramer-von Mises*, and *Chi-Squared.* The Kolmogorov-Smirnov test can be considered the most widely used goodness-of-fit test for continuous data, while Chi-Squared is used for discrete data.

Computing these tests are advanced topics that will not be covered in this class.

2 Case Studies

2.1 Continuous Data

2.1.1 Combined Cycle Power Plant

The dataset contains 9,568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011). A combined cycle power plant (CCPP) is composed of gas turbines (GT), steam turbines (ST), and heat recovery steam generators. In a CCPP, the electricity is generated by gas and steam turbines, which are combined in one cycle, and is transferred from one turbine to another. While the Vacuum is colected from and has effect on the Steam Turbine, the other three of the ambient variables effect the GT performance. (Source: https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant)

2.1.2 Attributes

- **Temperature (T)** in the range 1.81°C and 37.11°C
- **Ambient Pressure (AP)** in the range 992.89-1,033.30 milibar
- **Relative Humidity (RH)** in the range 25.56% to 100.16%
- **Exhaust Vacuum (V)** in teh range 25.36-81.56 cm Hg
- **Net hourly electrical energy output (EP)** 420.26-495.76 MW

All the above attributes are continuous variables.

2.1.3 Required Packages

For this session, we will make use of tidyverse and fitdistrplus packages.

- fitdistrplus: The package fitdistrplus provides functions for fitting univariate distributions to different types of data (continuous data and discrete data) and allowing different estimation methods (maximum likelihood, moment matching, quantile matching and maximum goodness-of-fit estimation). This package also provides various functions to compare the fit of several distributions to a same data set.

```
library(tidyverse)
library(fitdistrplus)
```

You might need to install these packages if they are not installed already.

2.1.4 Importing the dataset

```
ccpp <- read.csv('ccpp.csv', header = TRUE, sep = ',')

head(ccpp) # structure of the dataset
```

```
  ï..AT     V      AP    RH     PE
1  8.34 40.77 1010.84 90.01 480.48
2 23.64 58.49 1011.40 74.20 445.75
```

```
3 29.74 56.90 1007.15 41.91 438.76
4 19.07 49.69 1007.22 76.79 453.09
5 11.80 40.66 1017.13 97.20 464.43
6 13.97 39.16 1016.05 84.60 470.96
```

The aim is to identify how the various attributes are distributed. Using one of the features as an example and functions from the fitdistrplus package, we will see which distribution best models the data.

2.1.5 Visualizing Data

Before fitting one or more distributions to a dataset, it is a generally a good practice to visualize the data in order to get an idea of what distributions are more likely to fit the data as compared to others.

```
ggplot(ccpp, aes(RH)) +
  geom_histogram(bins = 50, color = 'Black', fill = 'steelblue')
```

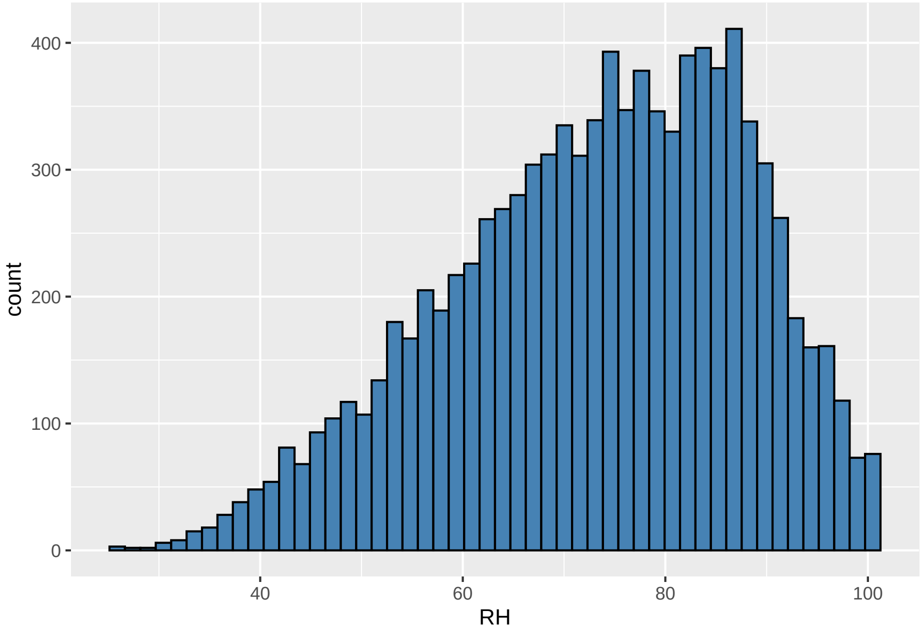


Figure 1: Histogram for Relative Humidity (RH)

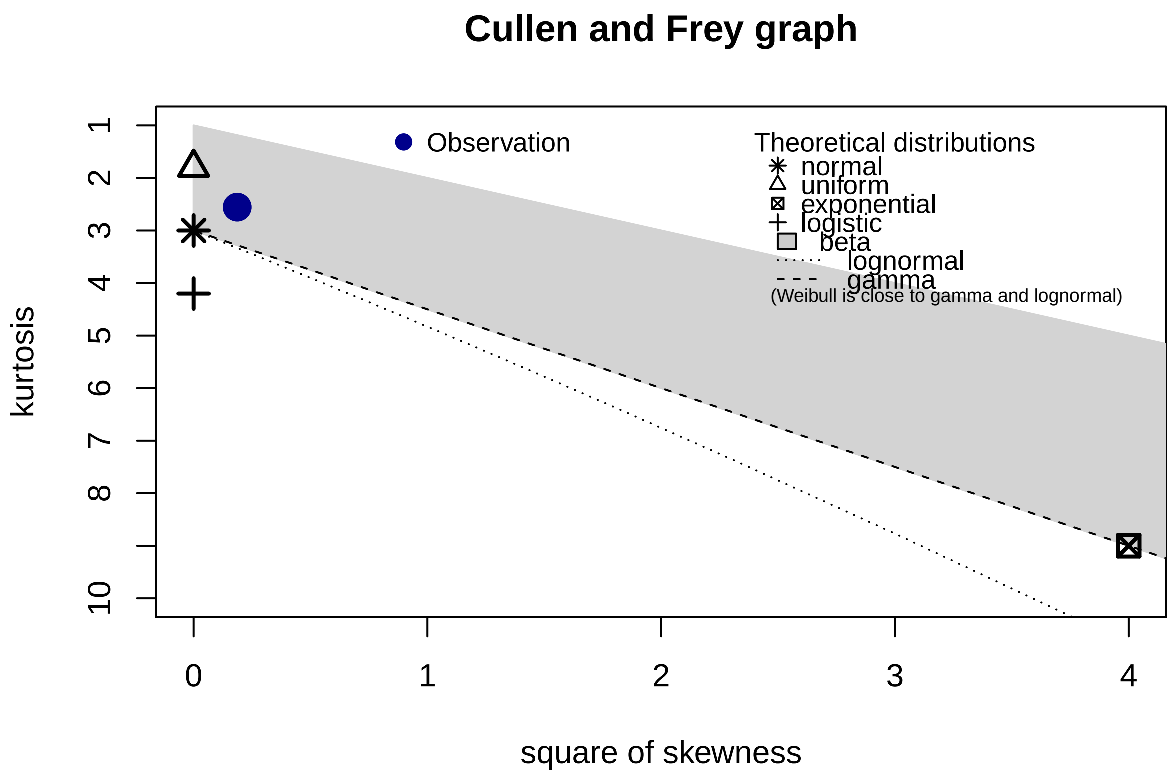
From obesrvation, it seems that RH has a left tail with a normal like distribution . Using other functions from the fitdistrplus package, it is possible to find out if normal distribution is the best fit for this data or some other distribution describes this data better.

2.1.6 Descriptive Statisitics

The descdist() function computes classical descriptive statistics (minimum, maximum, median, mean, standard deviation), skewness and kurtosis. It requires the data for which fit is to be estimated as a numerical vector input. Skewness and Kurtosis linked to the third and fourth moments, are especially important here, **as a non-zero skewness reveals a lack of symmetry, while the kurtosis value quantifies the weight of tails in comparison to the normal distribution for which the kurtosis equals 3**.

The function also plots a skewness-kurtosis plot. On this plot, values for common distributions are displayed in order to help the choice of distributions to fit to data. For some distributions (normal, uniform, logistic, exponential), there is only one possible value for the skewness and the kurtosis. Thus, these distributions are represented by a single symbol/point on the plot. For other distributions, areas of possible values are represented, consisting in lines (as for gamma and lognormal distributions), or larger areas (as for beta distribution).

```
descdist(ccpp$RH)
```



```
summary statistics
------
min:  25.56   max:  100.16
median:  74.975
mean:  73.30898
estimated sd:  14.60027
estimated skewness:  -0.4318387
estimated kurtosis:  2.555474
```

As can be observed from the plot, normal distribution as well as lognormal distribution are the closest to the distribution for *RH*.

2.1.7 Fit

The fitdist() function is used to fit univariate (single variable) distributions to data. Numerical results returned by the fitdist() function are:

1. The parameter estimates
2. The estimated standard errors
3. The loglikelihood
4. Akaike and Bayesian information criteria (the so-called AIC and BIC)*
5. The correlation matrix between parameter estimates

AIC and BIC

AIC and BIC are widely used in model selection criteria. AIC means Akaike's Information Criteria and BIC means Bayesian Information Criteria.

Akaike's Information Criteria was formed in 1973 and Bayesian Information Criteria in 1978. Hirotsugu Akaike developed Akaike's Information Criteria whereas Gideon E. Schwarz developed Bayesian information criterion.

The AIC can be termed as a mesaure of the goodness of fit of any estimated statistical model. The BIC is a type of model selection among a class of parametric models with different numbers of parameters.

When comparing the Bayesian Information Criteria and the Akaike's Information Criteria, penalty for additional parameters is more in BIC than AIC. Unlike the AIC, the BIC penalizes free parameters more strongly.

* A summary of AIC and BIC can be read here

Below the fitdist() function is used to fit a normal and lognormal distributions to Relative Humidity (RH) variable. The arguments required for the the function are the the data to be fitted and the distribution for which the parameters are to be estimated.

2.1.7.1 Normal Distribution

```r
# get paramter estimates for normal distribution
fit_n <- fitdist(ccpp$RH, "norm")
summary(fit_n)
```

```
Fitting of the distribution ' norm ' by maximum likelihood
Parameters :
     estimate Std. Error
mean 73.30898  0.1492545
sd   14.59951  0.1055389
Loglikelihood:  -39228.09   AIC:  78460.19   BIC:  78474.52
Correlation matrix:
     mean sd
mean    1  0
sd      0  1
```

2.1.7.2 Log Normal Distribution

```r
# get paramter estimates for lognormal distribution
fit_ln <- fitdist(ccpp$RH, "lnorm")
summary(fit_ln)
```

```
Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters :
         estimate  Std. Error
meanlog 4.2723668 0.002234967
sdlog   0.2186158 0.001580211
Loglikelihood:  -39906.85   AIC:  79817.69   BIC:  79832.03
Correlation matrix:
        meanlog sdlog
meanlog       1     0
sdlog         0     1
```

2.1.8 Goodness-of-Fit Plots

The plot of an object of class "fitdist" provides four classical goodness-of-fit plots:

- A density plot representing the density function of the fitted distribution along with the histogram of the empirical distribution,
- A CDF plot of both the empirical distribution and the fitted distribution,
- A Q-Q plot representing the empirical quantiles (y-axis) against the theoretical quantiles (x-axis)
- A P-P plot representing the empirical distribution function evaluated at each data point (y-axis) against the fitted distribution function (x-axis)

```r
par(mfrow=c(2,2))
plot.legend <- c("normal")
denscomp(list(fit_n), legendtext = plot.legend, xlab = 'Relative Humidity (RH)', xlegend = 'topleft')
cdfcomp (list(fit_n), legendtext = plot.legend, xlab = 'Relative Humidity (RH)')
qqcomp  (list(fit_n), legendtext = plot.legend, xlab = 'Relative Humidity (RH)')
ppcomp  (list(fit_n), legendtext = plot.legend, xlab = 'Relative Humidity (RH)')
```

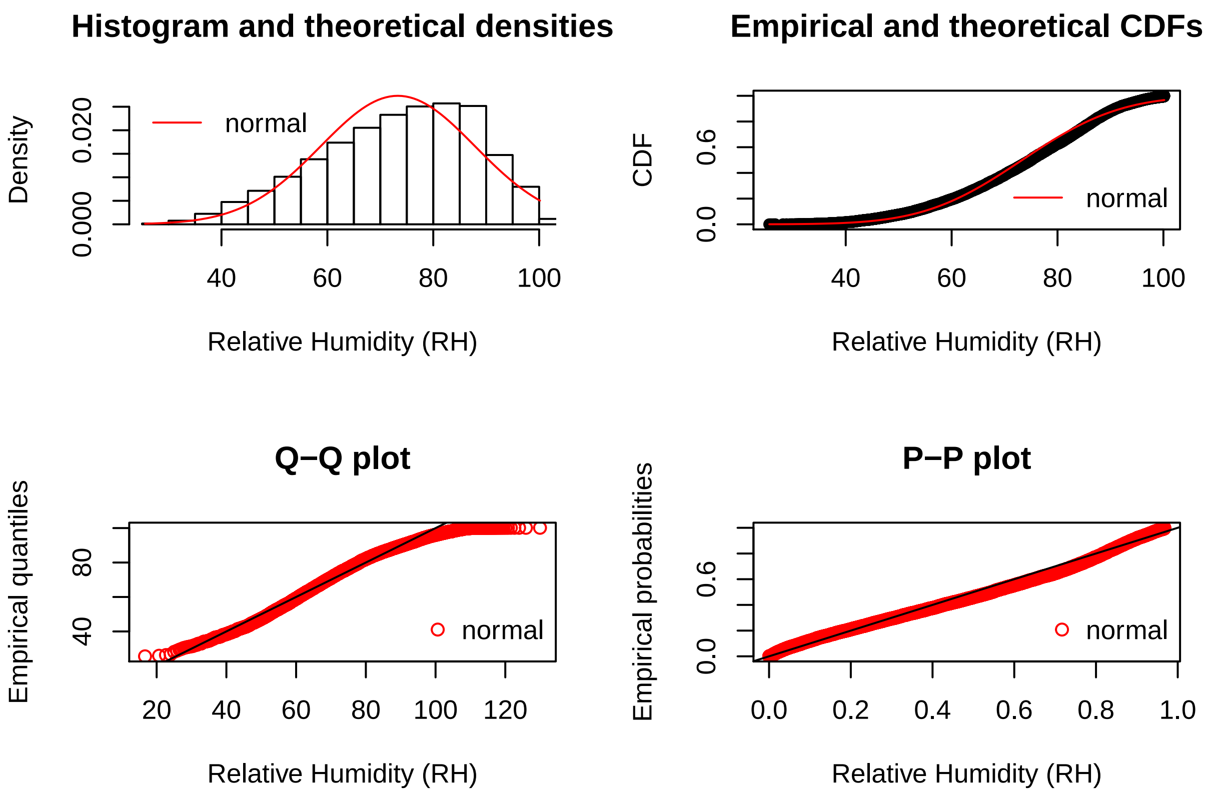


Figure 2: Goodness-of-Fit plots

ϟ Can you create similar plots to check the fitness for the lognormal distributions as well?

From these, and since AIC and BIC values are smaller and Loglikelihood is higher, it is evident that normal distribution is a better fit for the RH feature with a $\mu = 73.30$ and $\sigma = 14.59$.

- $X \sim \mathcal{N}(\mu = 73.30, \, \sigma = 14.59)$

2.2 Discrete Data

The discrete data is a sample of number of patient arrivals at hostpital in a single hour collected over a period of time. The goal is to find out which discrete distribution best fits this data and estimate the parameters.

```
data <- read.csv('patient.csv')
```

2.2.1 Visualizing Data

As before, the first step is to visualize the data

```
ggplot(data, aes(x)) +
  geom_histogram(bins = 50, color = 'Black', fill = 'steelblue')
```

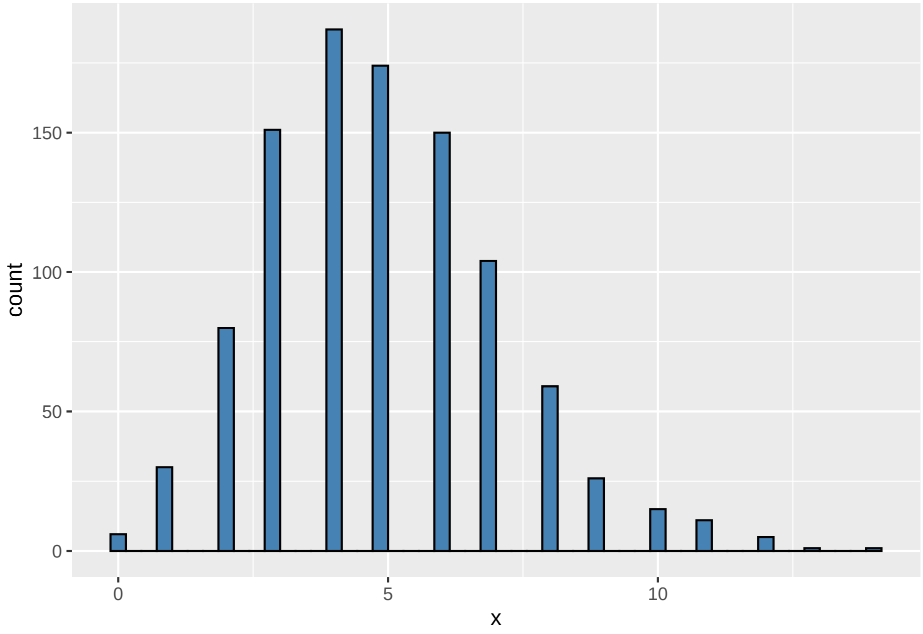


Figure 3: Histogram for Patient Arrivals

ϟ Can you say from the graph which distribution is most likely to fit this data?

2.2.2 Fit

Unlike continous data, descdist() function does not provide many options for estimating which distribution does the data lie closest to. As such, we will try to fit negative binomial and poisson distribution to our data, get the parameter estimates and then run goodness-of-fit tests to determine which one is better.

2.2.2.1 Negative Binomail Distribution

```
# get paramter estimates for negative binomial distribution
fit_nb <- fitdist(data$x, 'nbinom')
summary(fit_nb)
```

```
Fitting of the distribution ' nbinom ' by maximum likelihood
Parameters :
         estimate  Std. Error
size 8.404903e+05 56.64875753
mu   4.952428e+00  0.07036963
Loglikelihood:  -2180.354   AIC:  4364.708   BIC:  4374.524
Correlation matrix:
     size mu
size    1  0
mu      0  1
```

2.2.2.1.1 Poisson Distribution

```
# get paramter estimates for poisson distribution
fit_p <- fitdist(data$x, 'pois')
summary(fit_p)
```

```
Fitting of the distribution ' pois ' by maximum likelihood
Parameters :
       estimate Std. Error
lambda    4.953 0.07037755
Loglikelihood:  -2180.354   AIC:  4362.708   BIC:  4367.616
```

2.2.3 Goodness-of-Fit Tests

The gofstat() function can be used to evaluate distributions using goodness-of-fit statistics. For discrete distributions, the test is Chi-Squared and based on the significance level chosen, one can decide which distribution is a better fit to the data being evaluated.

```
gofstat(list(fit_nb, fit_p))
```

```
Chi-squared statistic:  7.364595 7.362353
Degree of freedom of the Chi-squared distribution:  7 8
Chi-squared p-value:  0.3919302 0.498089
Chi-squared table:
      obscounts theo 1-mle-nbinom theo 2-mle-pois
<= 1         36          42.06165        42.04122
<= 2         80          86.65548        86.62557
<= 3        151         143.05116       143.01881
<= 4        187         177.11222       177.09304
<= 5        174         175.42689       175.42837
<= 6        150         144.79817       144.81612
<= 7        104         102.44334       102.46775
<= 8         59          63.41806        63.44035
<= 10        41          52.17982        52.20591
> 10         18          12.85323        12.86287
```

```
Goodness-of-fit criteria
                               1-mle-nbinom 2-mle-pois
Akaike's Information Criterion     4364.708   4362.708
Bayesian Information Criterion     4374.524   4367.616
```

From this test, we see that both negative binomial and poisson distribution fit the data well. This does not mean that the data actually follows these distributions with the parameters obtained using the fitdist() function, but that it cannot be distinguished from the above distributions with the parameters obtained. Other distributions with different parameters can be consistent with the same data. In such a scenario, it is really important to understand the data, how it has been collected and what is the use case, before deciding on any conclusions.

⚡ Try the gofstat() function to evaluate continuous distributions.

⏸ Some of the functions of the fitdistrplus package require the argument discrete = TRUE when working with discrete data. Use the help to figure out when and where to use it.