Cafe Louis - Group 2 Project

Analysing the data

2023-05-11

Contents

Analysing the data

For analysing the data we decided to use package "distfit" for finding and plotting the "best-fit" distribution for our data. Additionally we used a python function provided by Dr Binh to perform chi-square goodness of fit tests. Other python packages we used were pandas, numpy, matplotlib and scipy.

Chi-Square Goodness of fit function

data analysis with chi-square goodness of fit

```
def obs_cts(n, data):
""" given: the data and number of bins
    returns: the observed values and the bin edges as lists"""

  events, edges = np.histogram(data, n)
  return events.tolist() , edges.tolist()
```

```
def exp_cts(n, data):
""" given: the data and number of bins
    returns: the expected values and prob over each of the bins with
    the necessary modification of the first and last bins"""
  L=[]
  P_bins =[]
  for x in obs_cts(n,data)[1]:
      L.append(rv.cdf(x))
  P_bins.append(L[1])
  for i in range(1,len(L)-2):
      P_bins.append(L[i+1]-L[i])
  P_bins.append(1-L[-2])
  exp_cnt = [x * len(data) for x in P_bins]
  return exp_cnt, P_bins
```

```
def ind_bins_to_reduce(f_exp):
""" given: a list
    returns: the indexes of the elements < 5 """
```

```
NC_to_red =[index for index,value in enumerate(f_exp) if value < 5]
return NC_to_red
```

```
def one_reduce(f_exp, f_obs, f_edge):
""" given: lists of exp, obs, edges returns: new lists with one reduced bin with value < 5 """
BTR = ind_bins_to_reduce(f_exp)
if (len(BTR)>1 or (len(BTR)==1 and BTR[0]!=0)):
    f_exp[BTR[-1]-1] = f_exp[BTR[-1]-1]+f_exp[BTR[-1]]
    f_obs[BTR[-1]-1] = f_obs[BTR[-1]-1]+f_obs[BTR[-1]]
    del(f_edge[BTR[-1]])
    del(f_obs[BTR[-1]])
    del(f_exp[BTR[-1]])
else:
    if BTR[0]==0:
        f_exp[1]= f_exp[1]+f_exp[0]
        f_obs[1]= f_obs[1]+f_obs[0]
        del(f_edge[1])
        del(f_obs[0])
        del(f_exp[0])

f_expN = f_exp
f_obsN = f_obs
f_edgeN = f_edge
BTRN = ind_bins_to_reduce(f_expN)
return f_expN, f_obsN, f_edgeN, BTRN
```

```
def all_reduce(f_expF, f_obsF, f_edgeF, BTRF):
""" finalizes the bin reduction """
while BTRF !=[]:
    u = one_reduce(f_expF, f_obsF, f_edgeF)
    f_expF = u[0]
    f_obsF = u[1]
    f_edgeF = u[2]
    BTRF = u[3]
return f_expF, f_obsF, f_edgeF, BTRF
```

```
def model(data, n, dof):
    """ given data, the number of bins (n) and the number of estimated parameters
    (dof)
    produces the value of the chi-squate test statistics and the p-value"""

    ## final expected count and final observed count after amalgamating bins
    exp, obs = all_reduce(exp_cts(n, data)[0],obs_cts(n, data)[0],
            obs_cts(n, data)[1], ind_bins_to_reduce(exp_cts(n, data)[0]))[0:2]

    # build in chi-gof test, the last argument is the adjustment to the dof
    result = ss.chisquare( np.asarray(obs), np.asarray(exp), dof)
    return result
```

Execution

```
fit_k,fit_loc,fit_beta = ss.erlang.fit(service)

rv = ss.erlang(fit_k,fit_loc,fit_beta)
model(service,100,0)
```

Result

```
Power_divergenceResult(statistic=26.49218594835637, pvalue=0.5460099412205021)
```

Distfit

We used the following process in package disfit to find and plot the best fit distribution for our data:

```
data = pd.read_csv("Cafe_Louis_Data.csv")

arrivals = data["Inter-Arrival Time"]

dfit = distfit(distr = ['expon','erlang', 'gamma'])

dfit.fit_transform(arrivals)

dfit.plot()

plt.savefig('distfit.png')
```

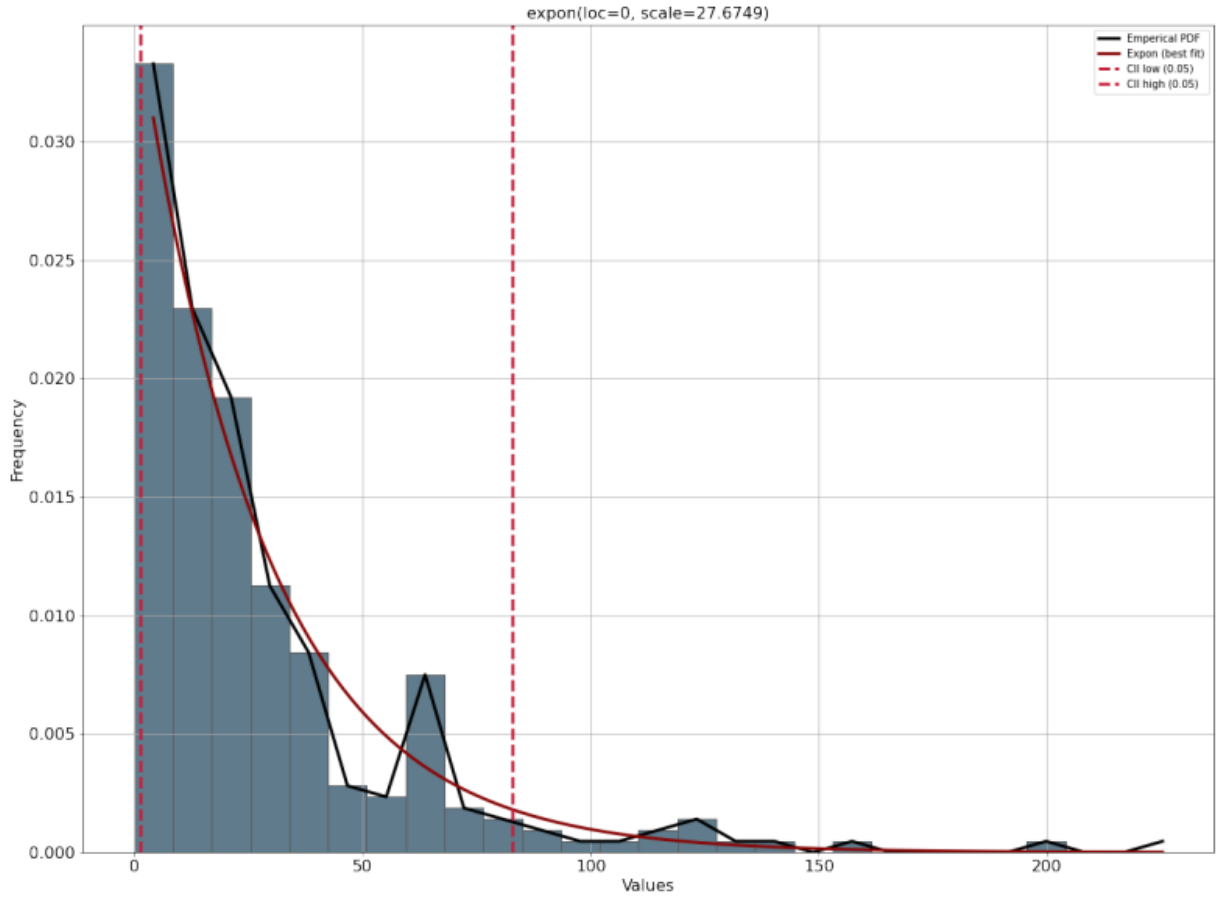
Inter-Arrival Times

$$H_0 : \text{The sample is exponentially distributed}$$

$$H_1 : \text{The sample is not exponentially distributed}$$

Using the distfit package, we determined that the most suitable probability distribution for modeling interarrival times is the Exponential distribution, compared to the Erlang and Gamma distributions. However, we conducted a chi-squared goodness-of-fit test with the null hypothesis that the sample data follows an Exponential distribution and the alternative hypothesis that it does not. The test resulted in a chi-square statistic of 64.836 with a corresponding p-value of 2.184e-05. These results suggest that there is not enough evidence to support the null hypothesis that the sample is exponentially distributed. Therefore, we cannot conclude that the interarrival times are Exponential distributed. Further investigation may be necessary to identify a more appropriate distribution for modeling the interarrival times.

```
knitr::include_graphics("arrivals.png")
```



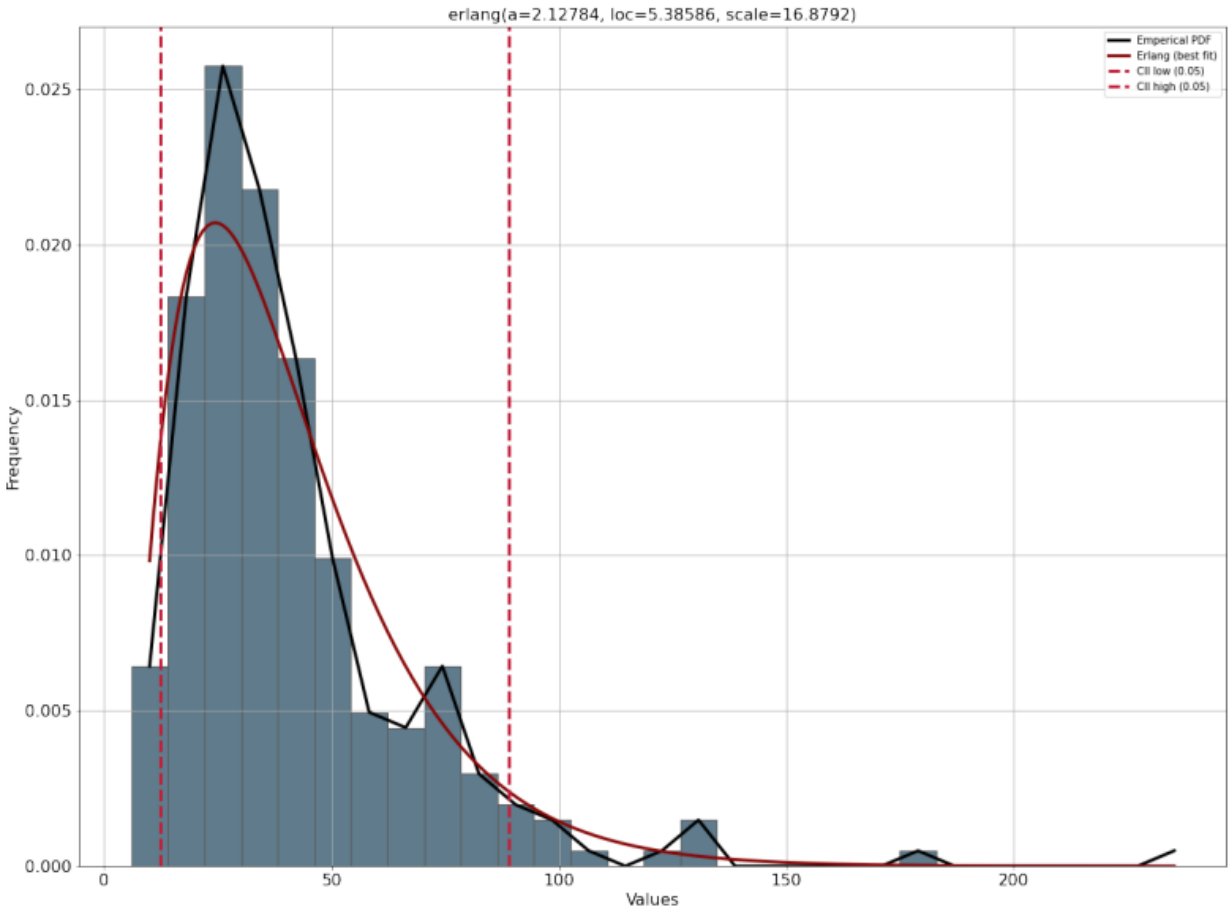
Service Times

H_0 : The sample is erlang distributed

H_1 : The sample is not erlang distributed

Using the distfit package, we identified the most suitable probability distribution for modeling service time as the Erlang distribution, after comparing it to the Exponential and Gamma distributions. We then conducted a chi-squared goodness-of-fit test with the null hypothesis that the sample data follows an Erlang distribution, and the alternative hypothesis that it does not. The resulting chi-square statistic was 26.492 with a corresponding p-value of 0.546. Therefore we conclude that there is sufficient evidence to accept the null hypothesis and conclude that the service times are Erlang distributed. This implies that the Erlang distribution provides a good fit for modeling the service times in our study.

```
knitr::include_graphics("service.png")
```


erlang(a=2.12784, loc=5.38586, scale=16.8792)
Emperical PDF
Erlang (best fit)
CII low (0.05)
CII high (0.05)
Frequency
Values
0.000
0.005
0.010
0.015
0.020
0.025
0
50
100
150
200