BANA 7031 - Probability Models

Probabilistic Analysis of Depression Levels

By Mohammed Nifaullah Sailappai (M13433497)

*Abstract*

*This report involves the approximation of parametric distribution to real life data. We have used the Ginzberg dataset in R, which contains numerically transformed data for psychiatric patients hospitalized for depression, and tried to determine the distribution for adjusted depression. Based on the results obtained we conclude that adjusted depression follows a Gamma distribution with (shape) α = 4.04 and (rate) β = 4.04. We also tested if adjusted simplicity has any effect on adjusted depression, and from the results it seems that there is a linear correlation between the two variables.*

Contents

Introduction ................................................................................................................................................ 2

Candidate Exploration……….................................................................................................................... 2

Cullen & Frey Graph…………...................................................................................................… 2

Fit Analysis…………… ............................................................................................................... 3

AIC Test...........................................................................................................................................5

Kolmogorov-Smirnov Test ……………...................................................................................................... 5

Empirical Cumulative Distribution Function................................................................................................ 5

Maximum Likelihood Parameter Estimation................................................................................................ 5

Bootstrapping for Confidence Interval ……………………........................................................................ 6

Effect of Adjusted Simplicity on Adjusted Depression……........................................................................ 7

Comparison of two group means via Permutation test…................................................................ 7

Comparison of two group medians via Wilcoxon test…................................................................. 8

Simple Linear Regression Analysis……......................................................................................... 8

Appendix ...................................................................................................................................................... 8

References…………………………………………………………………………………………………. 9

Code ............................................................................................................................................................. 9

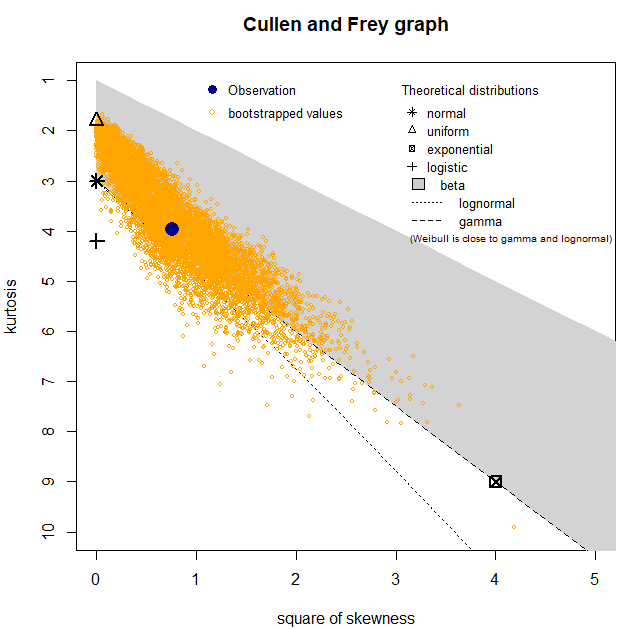
Introduction

This report summarizes various types of analysis performed on quantified and adjusted depression levels of various patients who were hospitalized for depression. The dataset being used is “Ginzberg” [1]. The goal of this project is to come up with a probability distribution for computer prices. This is a parametric approach and all statistical work has been accomplished in R.

Candidate Exploration

Cullen & Frey Graph

First we try to evaluate the distribution of the target variable by using the Cullen & Frey Graph [2], as it would help us narrow down the distributions by a huge margin. We use the Moment approach to evaluate the higher order moments such as skewness² and kurtosis, which will be used to visualize the Cullen & Frey Graph, which in turn was used to narrow down the list of candidate distributions. Below Cullen & Frey graph shows the result with 10,000 Bootstrapped values

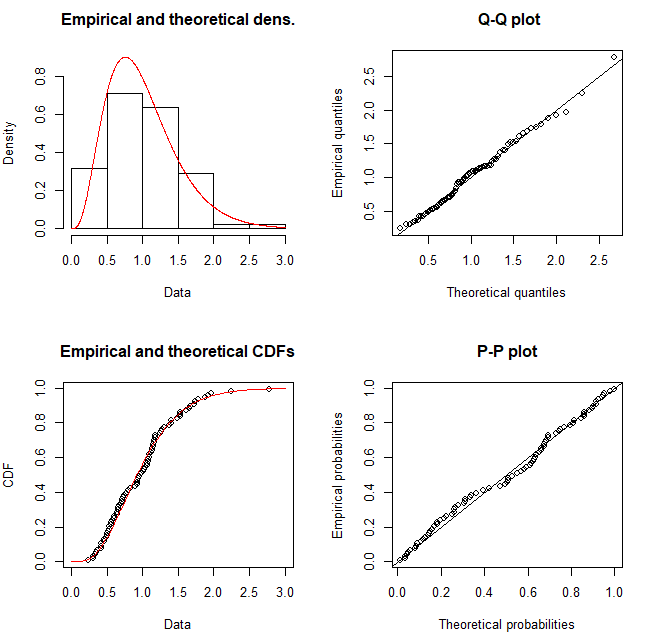


We limit the probable list of candidates to Gamma, Lognormal, Normal and Weibull distributions by selecting only those distribution which are in proximity to the observations.

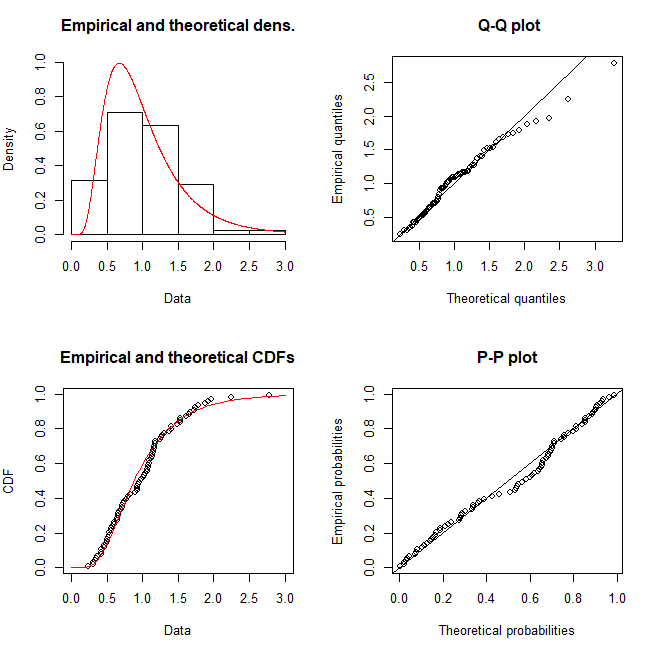
Fit Analysis

We now fit the data to each of the shortlisted distribution and see which distribution fits the data better. We visually compare the empirical vs theoretical density and CDF along with the PP and QQ plots for each candidate.

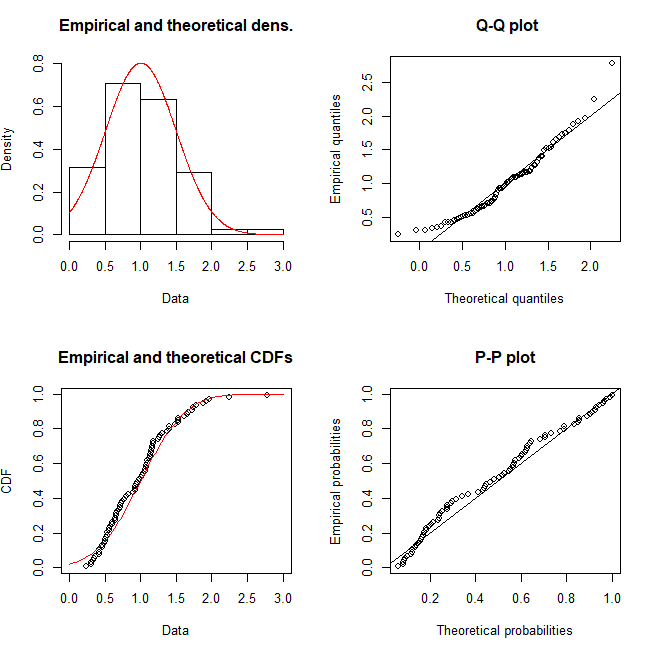
* Gamma



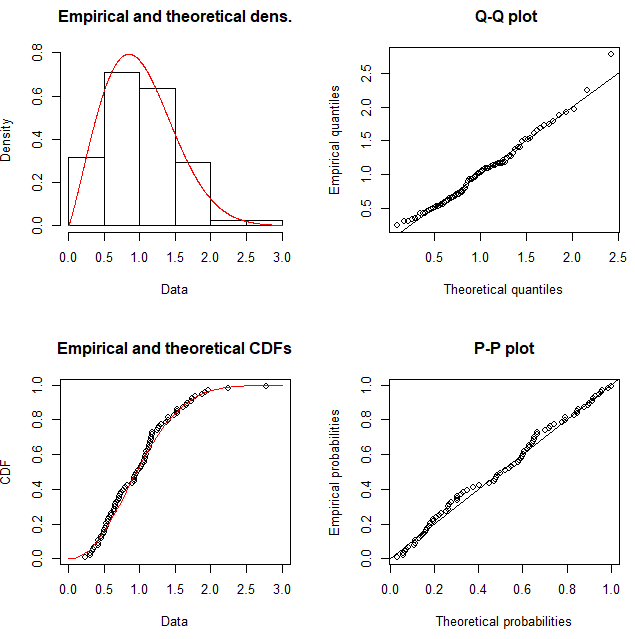
* Lognormal



* Normal



* Weibull



From the visual analysis and comparison of the above graphs, clearly, both Gamma and Weibull distributions fit our observations better.

AIC Test

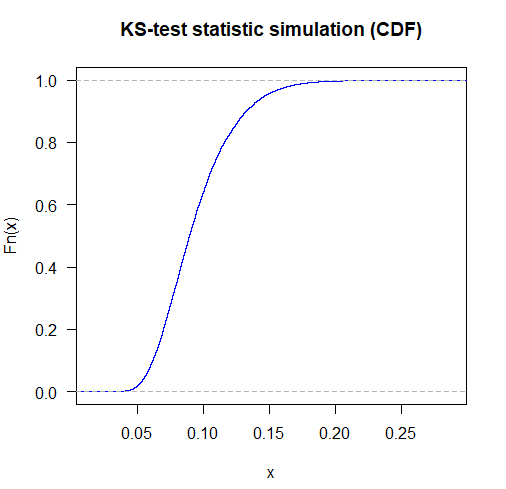
We further use the AIC test to select the appropriate distribution for our data.

|  |  |
| --- | --- |
| Distribution | AIC |
| Gamma | 107.8416 |
| Weibull | 110.9726 |
| Lognormal | 109.5877 |
| Normal | 122.0236 |

Thus based on the visual analysis and the AIC we assert that Gamma distribution with estimated parameters α = 4.04 and β = 4.04 fits our data best among all parametric distributions.

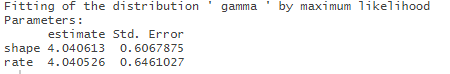
Kolmogorov-Smirnov Simulation Test

Now we run the Kolmogorov-Smirnov Simulation test to check if the selected distribution is indeed a good fit for our data. We simulate the KS Statistic under the null-hypothesis that all sampled values come from the same distribution. We then used the ECDF of the simulated KS statistic (displayed below) to find that p-value to be very large, thus rendering us to the conclusion that indeed the selected distribution is a good fit for the data.



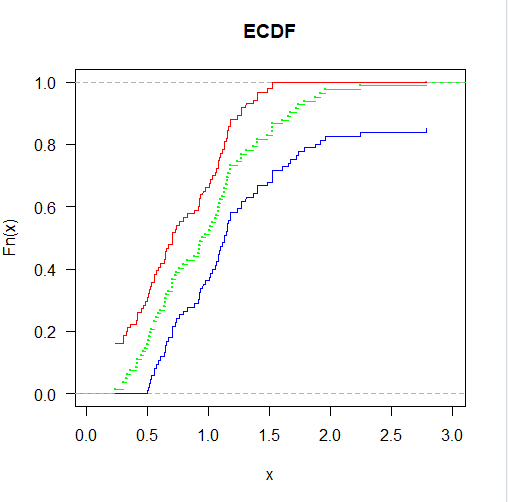
Maximum Likelihood Parameter Estimation

Gamma distribution parameter estimates of shape (α) and rate (β) with the it’s respective standard error is given below



Empirical Cumulative Distribution Function

Now we compare the parametric distribution with the Empirical Cumulative Distribution Function we generated from the data. We plotted the ECDF at 95% significance level to find out that confidence band failed to contain only 100 of the 10000 true values that were generated.



Bootstrapping for Confidence Interval

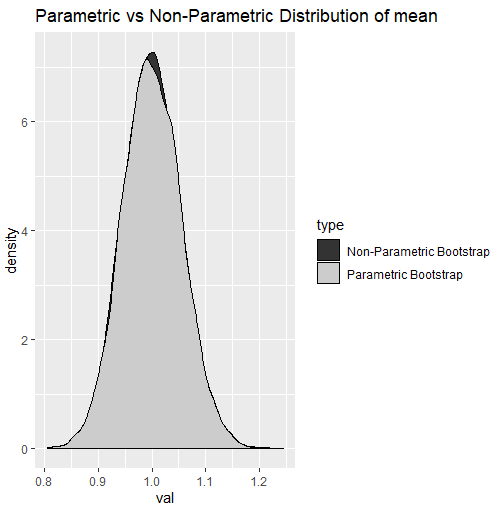
We Bootstrap to calculate the confidence interval and standard error for the estimate of the Mean using the mean function. The Standard error from Bootstrap is

* Parametric = 0.0549
* Non-Parametric = 0.055

We calculated the following confidence intervals which all had a similar Lower Bound (LB) and Upper Bound (UB) of:

* Normal for Parametric: LB = 0.890, UB = 1.110
* Normal for Non-Parametric: LB = 0.889, UB = 1.110
* Quantile for Non-Parametric: LB = 0.890, UB = 1.100
* Pivotal for Non-Parametric: LB = 0.893, UB = 1.109

Bootstrap distributions, given below, from the two methods appear to be are very similar:



Effect of Adjusted Simplicity on Adjusted Depression

We first do the comparison of means using the permutation test and then additionally, run a simple linear regression to find out if adjusted simplicity has some effect on adjusted depression.

Comparison of two group means via permutation test

We first divide depression levels in to 2 groups. One, for which the adjusted simplicity is less than or equal to one and other, for which the adjusted simplicity is greater than one. Adjusted simplicity ranges from 0 to 2.

* Ho: Difference of Mean = 0
* Ha: Difference of Mean ≠ 0

The p-value for the test is very small, hence we reject the null hypothesis and conclude that there is a significant difference in the mean of two groups.

Comparison of two group Medians via Wilcoxon test

Similar to above we first divide depression levels in to 2 groups. One, for which the adjusted simplicity is less than or equal to one and other, for which the adjusted simplicity is greater than one. Adjusted simplicity ranges from 0 to 2. We then run Wilcoxon test, which is a very robust test to compare medians of two groups, with following null and alternative hypothesis.

* Ho: No significant difference in medians
* Ha: Significant difference in medians

The p-value for the test is very small, hence we reject the null hypothesis and conclude that there is a significant difference in between the median of two groups.

Simple Linear Regression Analysis

We then check if the quantified (and adjusted) simplicity is linearly correlated to the quantified (and adjusted) depression levels by running a simple linear regression with adjusted simplicity as the predictor variable and adjusted depression as the target variable. First we define the null and alternative hypothesis and then run a simple regression fit to check if the null hypothesis is true or not.

* Ho: There is no significant correlation between adjusted simplicity and adjusted depression
* Ha: There is some significant correlation between adjusted simplicity and adjusted depression

The p-value for the test is very small, hence we reject the null hypothesis that there is no significant correlation between adjusted simplicity and adjusted depression. There appears to be some significant correlation between the variables.

Appendix

* 1. 1. Ginzberg Data set:
  2. a. Description: The data are for psychiatric patients hospitalized for depression.
  3. b. number of observations: 82
  4. c. country: Canada
  5. d. Source: Department of Mathematics and Statistics, York University

e. Variables

i. **Simplicity:** Measures subject's need to see the world in black and white

ii. **Fatalism:** Fatalism scale.

iii. **Depression:** Depression scale or level

iv. **AdjSimp:** Adjusted simplicity

v. **AdjFatal:** Adjusted fatalism

vi. **AdjDep:** Adjusted depression levels

2. Probabilistic Techniques in Exposure Assessment by **Cullen**, Alison C., **Frey**, H. Christopher

References:

1. Stat 3701 Lecture Notes: Bootstrap <http://www.stat.umn.edu/geyer/3701/notes/bootstrap.pdf>
2. Estimating Gamma Distribution by Thomas P. Minka [<https://tminka.github.io/papers/minka-gamma.pdf>]
3. Cullen & Frey graph [<https://www.r-project.org/conferences/useR-2009/slides/Delignette-Muller+Pouillot+Denis.pdf>]
4. Cullen & Frey graph [ <https://stackoverflow.com/questions/31741742/how-to-identify-the-distribution-of-the-given-data-using-r> ]
5. Gamma Distribution Wikipedia- [<https://en.wikipedia.org/wiki/Gamma_distribution#Characterization_using_shape_%CE%B1_and_rate_%CE%B2>]
6. Gamma Distribution in an SWedge Probabilistic Analysis [<https://www.rocscience.com/help/swedge/swedge/Gamma_Distribution.htm>]

Code:

*library(carData)*

*library(fitdistrplus)*

*library(logspline)*

*library(bootstrap)*

*library("tidyverse")*

*set.seed(29)*

*dep <- Ginzberg$adjdep*

*# Cullen & Frey Graph*

*descdist(dep, discrete = FALSE,boot = 10000)*

*gamma <- fitdist(dep, "gamma")*

*lognormal <- fitdist(dep, "lnorm")*

*normal <- fitdist(dep, "norm")*

*weibull <- fitdist(dep, "weibull")*

*# Fit Analysis*

*plot(gamma)*

*plot(lognormal)*

*plot(normal)*

*plot(weibull)*

*# AIC Test*

*gamma$aic*

*lognormal$aic*

*normal$aic*

*weibull$aic*

*gamma*

*# Kolmogorov-Smirnov Test*

*dep.shape <- gamma$estimate[1]*

*dep.rate <- gamma$estimate[2]*

*n <- length(dep)*

*ks.sim <- rep(0,10000)*

*for(i in 1:10000)*

*{*

*resample.dep <- rgamma(n, shape = dep.shape, rate = dep.rate)*

*ks.sim[i] <- as.numeric(ks.test(resample.dep,"pgamma",*

*shape = dep.shape,rate = dep.rate)$statistic)*

*}*

*kstest.fit <- logspline(ks.sim)*

*1 - plogspline(ks.test(resample.dep,"pgamma",*

*shape = dep.shape,rate = dep.rate)$statistic*

*,kstest.fit)*

*par(mfrow = c(1,1))*

*plot(ecdf(ks.sim),*

*las = 1,main = "KS-test statistic simulation (CDF)", col = "blue")*

*# ECDF and Plotting of Confidence Interval*

*dep.ecdf <- ecdf(dep)*

*plot(dep.ecdf,*

*las = 1,*

*main = "ECDF",*

*col = "green",*

*cex=0.2*

*)*

*Alpha=0.05*

*Eps=sqrt(log(2/Alpha)/(2\*n))*

*x<-seq(min(dep),max(dep), length.out = 10000)*

*lines(x, pmin(dep.ecdf(x)+Eps,1),col='red')*

*lines(x, pmax(dep.ecdf(x)-Eps,0),col='blue')*

*sum(pgamma(x,dep.shape,dep.rate) >=pmax(dep.ecdf(x)-Eps,0) &*

*pgamma(x,dep.shape,dep.rate) <=pmin(dep.ecdf(x)+Eps,1))/100*

*#Parametric and Non Parametric Bootstrappin for parameter(mean) estimation*

*mu.cap <- dep.shape/dep.rate*

*sd.cap <- sd(dep)*

*B <- 10000*

*mu.cap.star.p <- rep(0,B)*

*mu.cap.star.np <- rep(0,B)*

*for(i in 1:B) {*

*x.p <- rgamma(n,shape = dep.shape, rate = dep.rate)*

*pfit <- fitdist(x.p, "gamma")*

*mu.cap.star.p[i] <- pfit$estimate[1] / pfit$estimate[2]*

*x.np <- sample(dep,size=n,replace=TRUE)*

*npfit <- fitdist(x.np, "gamma")*

*mu.cap.star.np[i] <- npfit$estimate[1] / npfit$estimate[2]*

*}*

*se.cap.boot.p <- sd(mu.cap.star.p)*

*se.cap.boot.np <- sd(mu.cap.star.np)*

*theta.np <- as.data.frame(mu.cap.star.np)*

*theta.p <- as.data.frame(mu.cap.star.p)*

*theta.np$type <- 'Parametric Bootstrap'*

*theta.p$type <- 'Non-Parametric Bootstrap'*

*dis <- rbind(theta.p,theta.np)*

*dis <- as.data.frame(dis)*

*#Bootstrap distribution of estimate (mean)*

*ggplot(dis, aes(values, fill = type)) + geom\_density(alpha = 1) +*

*scale\_fill\_grey() +*

*ggtitle("Parametric vs Non-Parametric Distribution of mean")*

*normal.p<-c(mu.cap-2\*se.cap.boot.p, mu.cap+2\*se.cap.boot.p)*

*#normal ci, at 95%*

*normal.np<-c(mu.cap-2\*se.cap.boot.np, mu.cap+2\*se.cap.boot.np)*

*#pivotal ci at 95%*

*pivotal.np<-c(2\*mu.cap-quantile(mu.cap.star.np,0.975),*

*2\*mu.cap-quantile(mu.cap.star.np,0.025))*

*#quantile ci at 95%*

*quantile.np<-quantile(mu.cap.star.np, c(0.025, 0.975))*

*x1 <- Ginzberg$adjdep[Ginzberg$adjsimp <= 1]*

*x2 <- Ginzberg$adjdep[Ginzberg$adjsimp > 1]*

*#permutation test*

*perm.matrix <- t(replicate(500, sample(72)))*

*perm.T<-apply(*

*perm.matrix,1,*

*function(x){*

*abs(mean(Ginzberg$adjdep[x[1:46]])-mean(Ginzberg$adjdep[x[47:72]]))*

*})*

*perm.T*

*p.value<-mean(perm.T>abs(mean(x1)-mean(x2)))*

*p.value*

*#Wilcox test*

*wilcox.test(x1, x2,conf.int = T,exact=F)*

*#Simple Linear Regression*

*lm.fit <- lm(depression~adjsimp, Ginzberg)*

*summary(lm.fit)*