BANA 7031 - Probability Models
Probabilistic analysis of Computer Prices

By

Syed Imad Husain (M12958531)

Abstract

This report discusses the steps involved in approximating a parametric distribution to real life data. We have used the Computers dataset in R and tried to determine the distribution for Computer Prices. Our conclusion shows us that the Prices follow a Lognormal distribution with Meanlog = 7.67 and Sdlog = 0.257. We also tested if the prices of Computers with different types of RAM is the same using Permutation test and it seems that is not the case. Finally, we also discussed the Bayesian Posterior Distribution of the Prices towards the end of the report.

Contents

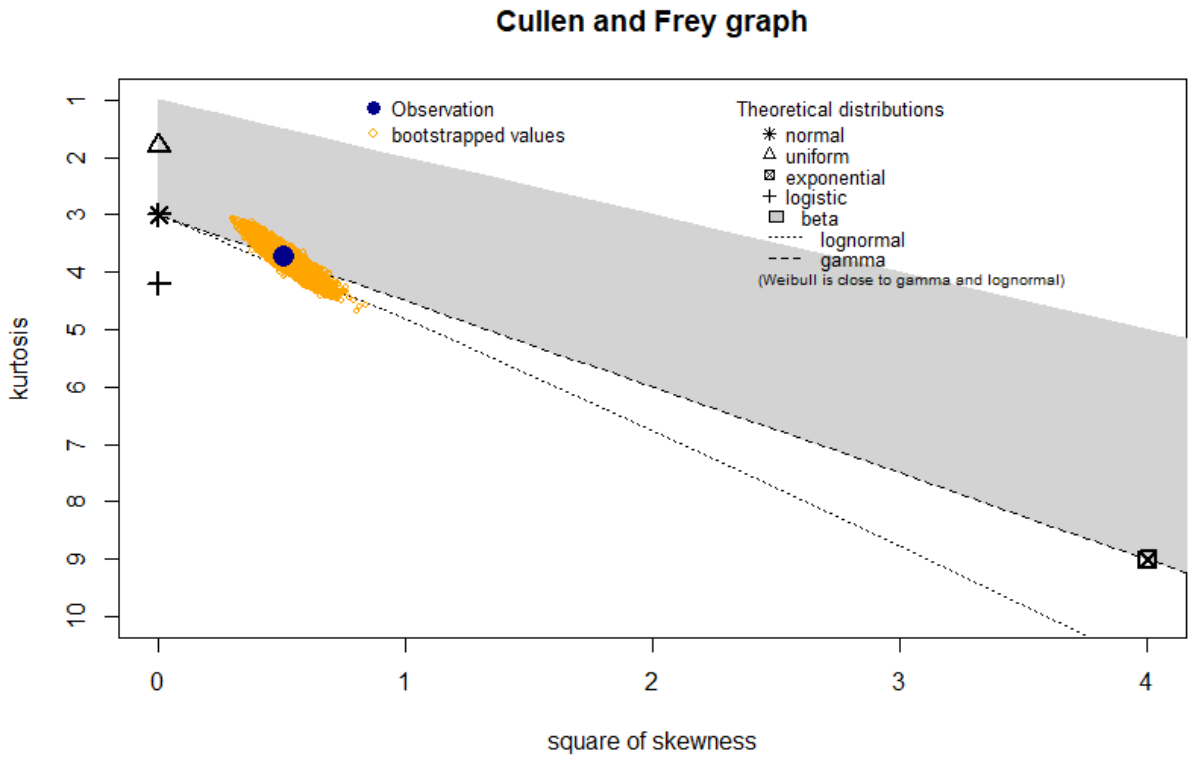
Introduction

This artefact summarizes various types of analyses performed on Prices of Computers. The dataset being used is "Computers"[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.

Exploring Candidate Models

Bootstrap C&F

There are several techniques for model selection. We will start our research by Moment-approach where we will try to evaluate the higher order (skewness², kurtosis) space to narrow down on the set of probable distributions. The best way to visualize this analysis is the Cullen and Frey graph[2](C&F Graph). It is only in our interest that we bootstrap the estimated kurtosis and skewness. Below C&F graph shows the result with 10,000 Bootstrapped values



The nearness of our observations(blue/yellow dots) to different theoretical distributions help us narrow the search space for candidate models. We conclude that Weibull, Normal, Lognormal and even Gamma are some of the probable distributions.

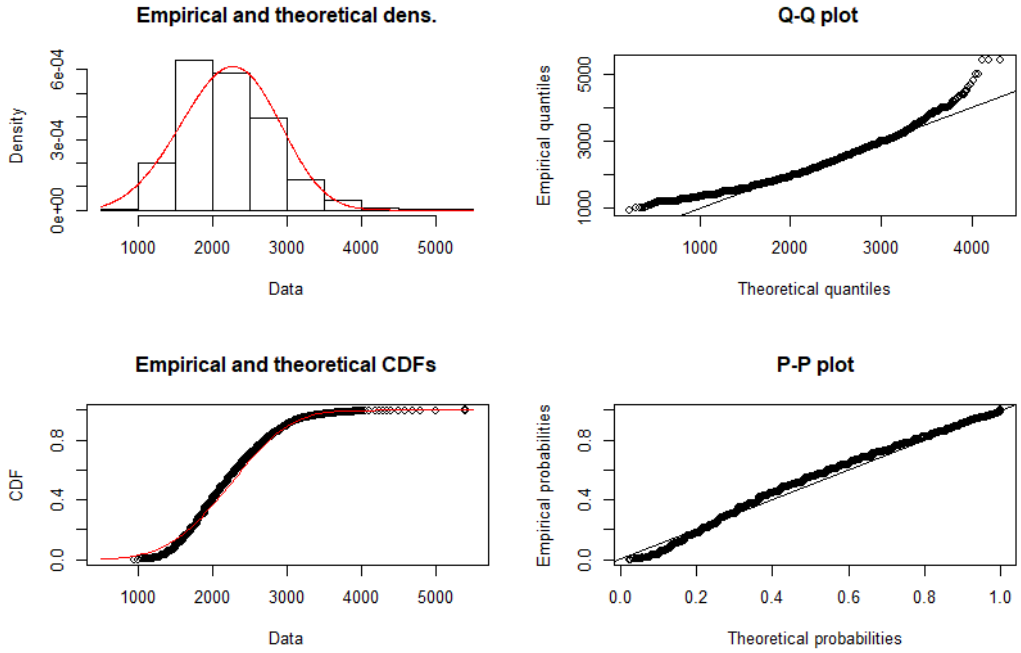
Theoretical vs Empirical graphs

In order to further narrow our analysis, we will look at the following graphs comparing the "Theoretical vs Empirical" to visualize the fitting of the distribution to the data:

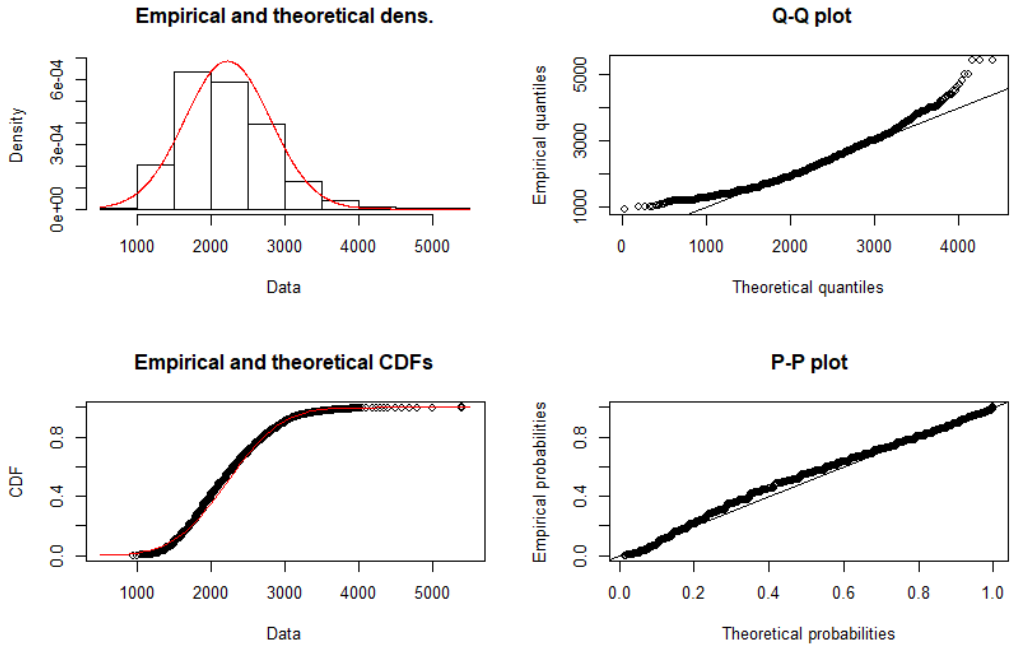
- Empirical PDF
- Empirical CDF
- Quantile Plots
- Percentile Plots

Following are the observations

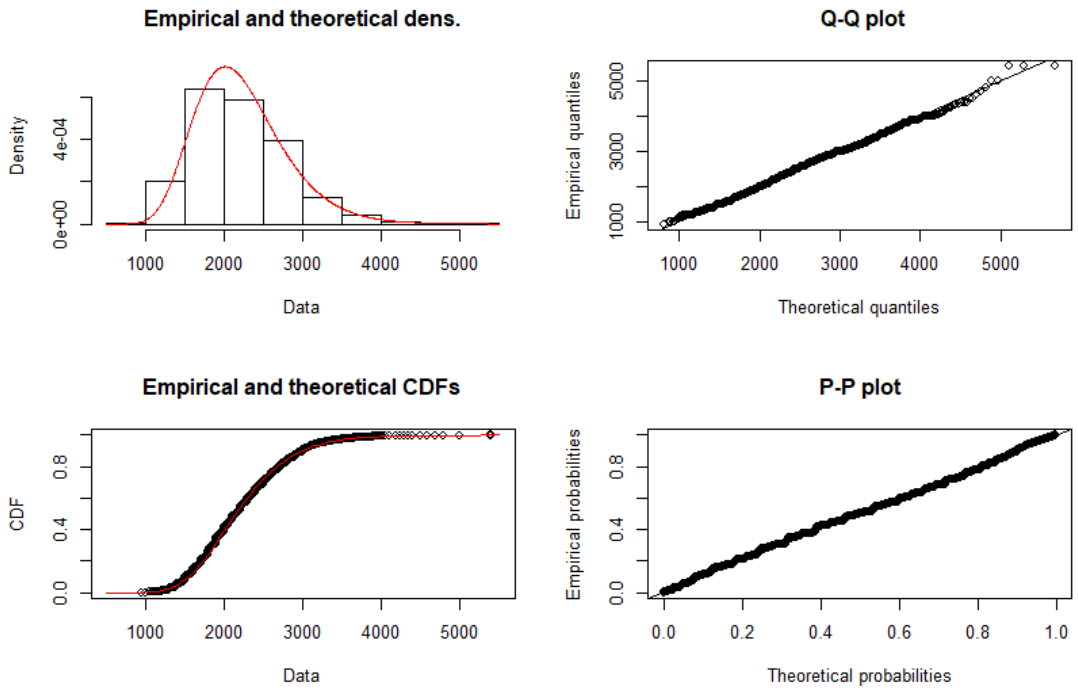
- Weibull Distribution



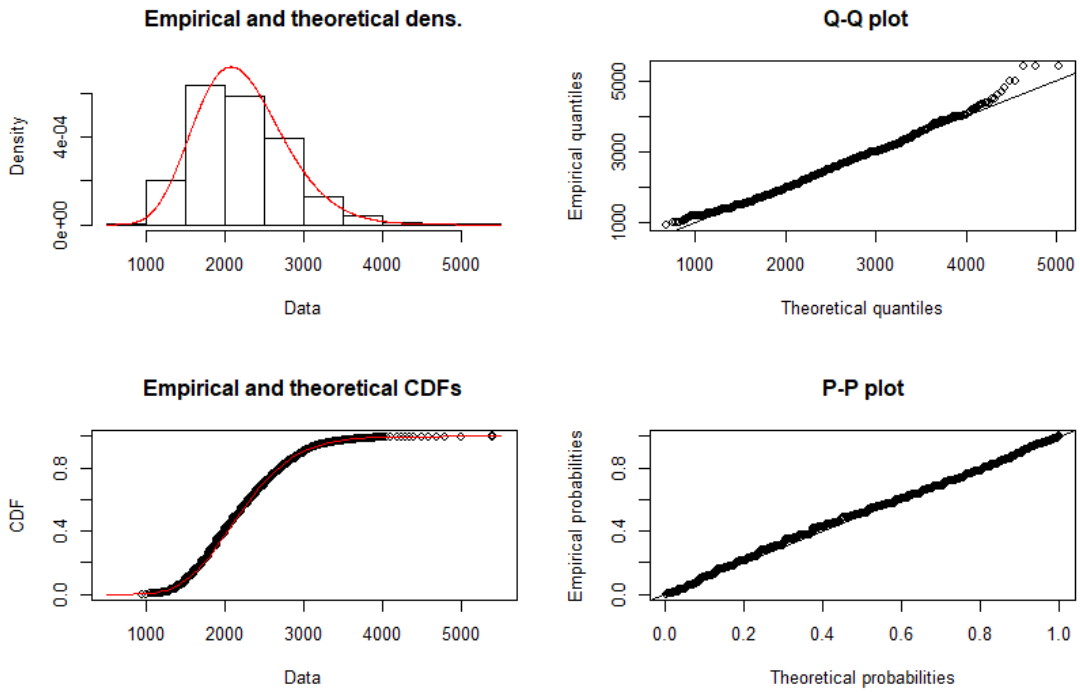
- Normal



- Lognormal



- Gamma



Based on these observations, we finalize **Log-Normal** to be the best candidate distribution. As we can see from above four graphs, the Q-Q plot and P-P plot, the empirical and theoretical dens and Empirical and theoretical CDFs are the best among all four distribution.

AIC Criterion

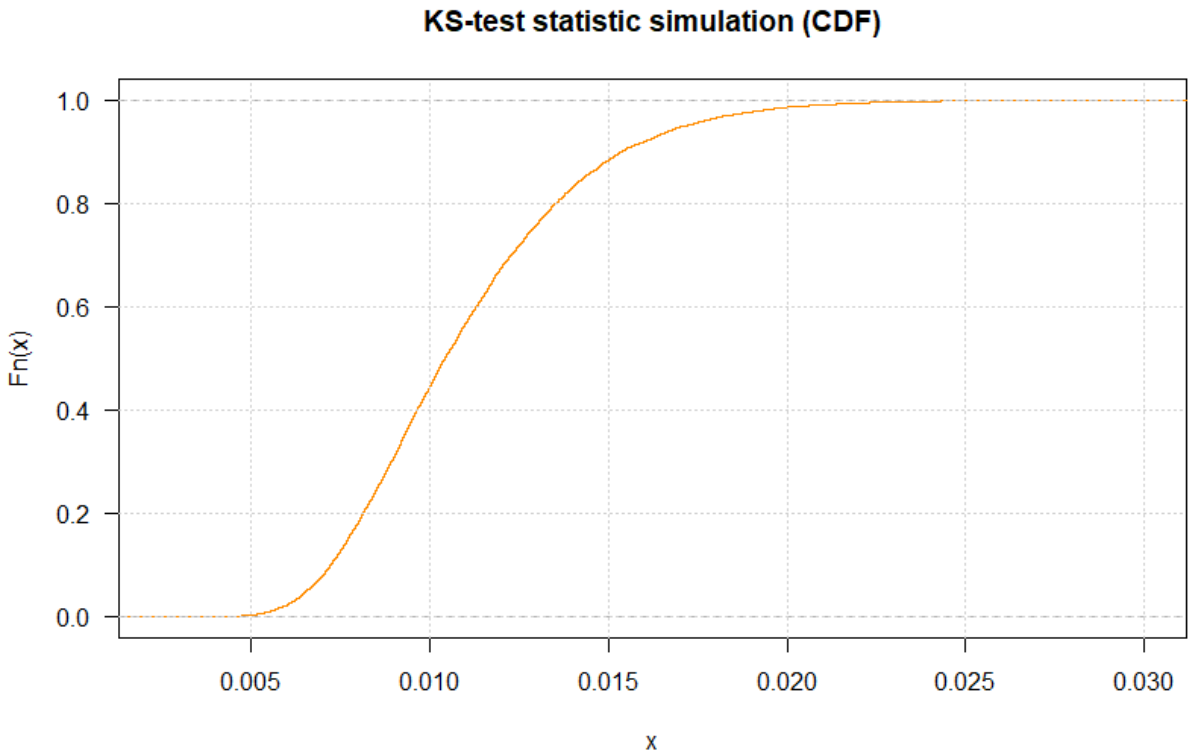
We will further use AIC to ascertain our assumption:

Distribution	AIC
Log-Normal	96841.68
Gamma	96914.93
Normal	97435
Weibull	97768.54

Distribution – Lognormal distribution with parameters Meanlog = 7.67177 Sdlog = 0.2579901. By using AIC, the smaller the better, the above table show us that the Log-Normal has the smallest AIC value.

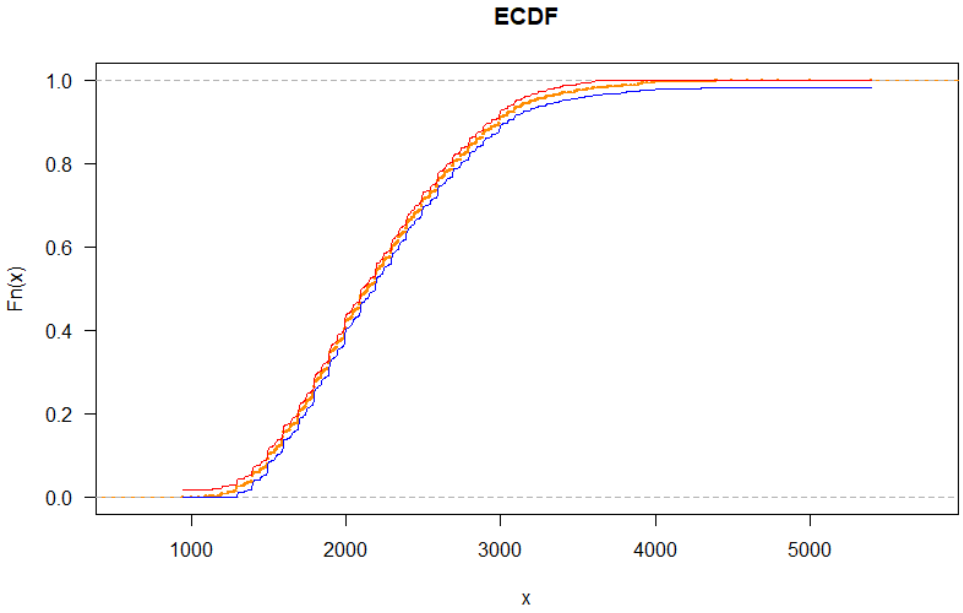
Kolmogorov-Smirnov Simulation Test

After finalizing the distribution, we would like to check if this distribution is indeed a good fit. For this, we will simulate the KS statistic under the null-hypothesis that the parameters are 0. The ECDF for simulated KS statistic is plotted below and the p-value for the test is very less, hence we conclude that we have arrived at the correct distribution



Empirical Cumulative Distribution Function

Since we now have a parametric distribution for the approximation, we will check its performance against an empirical CDF. Plot for the ECDF with 95% level of significance. The coverage of ECDF is 91.52%. This indicates that the ECDF confidence band was able to successfully contain the true value 91.52% times.



Maximum Likelihood Parameter Estimation

The MLE for Lognormal distribution is given as follows, when calculated from data,

- Mean = 7.67177
- Standard Deviation = 0.2579901

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \ln(X_i)}{n} \text{ and}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(\ln(X_i) - \frac{\sum_{i=1}^{n} \ln(X_i)}{n} \right)^2}{n}.$$

Bootstrap for Standard Error MLE and Confidence Interval

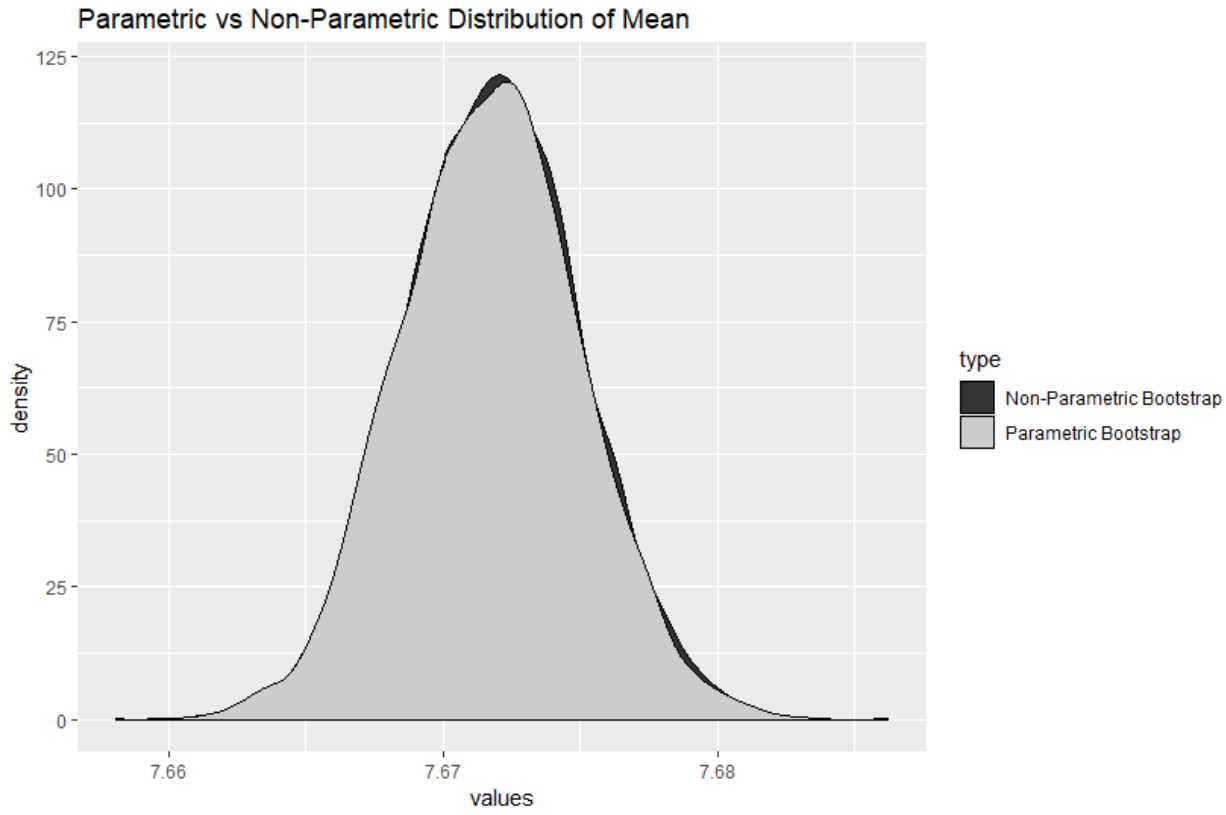
We perform Bootstrap to calculate the confidence interval and standard error for the MLE of the Mean. The Standard error from Bootstrap is

- Parametric = 0.003260
- Non-Parametric = 0.003284

We also calculated the following confidence intervals which all had a lower bound of 7.67 and an upper bound of 7.68 :

- Normal for Parametric
- Normal for Non-Parametric
- Quantile for Non-Parametric
- Pivotal for Non-Parametric

Below is the Bootstrap distribution from the two methods and they are very similar:



Permutation Method to compare Means of two Groups

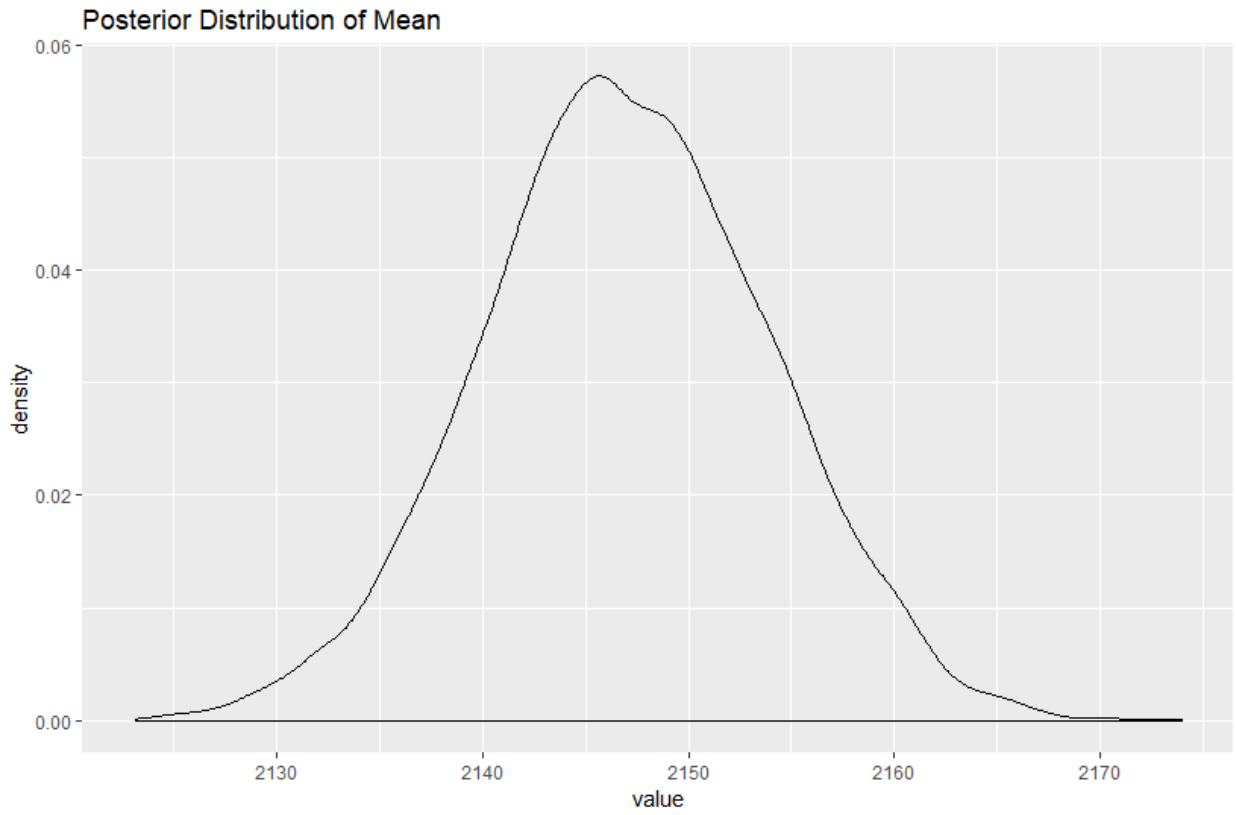
Here we try to check if the prices of computers with 4GB RAM are statistically different from 8GB RAM. In order to test this out, we perform the comparison through Permutation test.

- H_o : Difference of Mean = 0
- H_a : Difference of Mean $\neq$ 0

The p-value comes out to be very small and hence we reject the null hypothesis, thereby concluding that the prices of the two group are significantly different.

Bayesian Posterior Density for Distribution Mean

There is no difference between observing a normal sample and a log-normal sample as one can be turned into the other by a log or exponential transform. We may hence use the ultra-classic Gaussian model framework. With a Prior for the Mean as $f(\mu) = 1$ the Posterior Distribution is Log Normal $(\bar{x}, \sigma^2/n)$. Below is the simulation of Posterior distribution:



Appendix

1. Computers Data set:
    a. Description a cross-section from 1993 to 1995
    b. number of observations: 6259
    c. observation: goods
    d. country: United States
    e. Variables
        i. **price** price in US dollars of 486 PCs
        ii. **speed** clock speed in MHz
        iii. **hd** size of hard drive in MB
        iv. **ram** size of Ram in in MB
        v. **screen** size of screen in inches
        vi. **cd** is a CD-ROM present?
        vii. **multi** is a multimedia kit (speakers, sound card) included?
        viii. **premium** is the manufacturer was a "premium" firm (IBM, COMPAQ)?
        ix. **ads** number of 486 price listings for each month

x. **trend** time trend indicating month starting from January of 1993 to November of 1995.

f. Source Stengos, T. and E. Zacharias (2005) "Intertemporal pricing and price discrimination: a semiparametric hedonic analysis of the personal computer market", Journal of Applied Econometrics, forthcoming

g. References Journal of Applied Econometrics data archive: http://qed.econ.queensu.ca/jae/

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

R Code

```
#install.packages("fitdistrplus")
#install.packages("logspline")
library(fitdistrplus)
library(logspline)
library(Ecdat)
set.seed(007)

x <- Computers$price

descdist(x, discrete = FALSE,boot = 10000)

fit.weibull <- fitdist(x, "weibull")
fit.norm <- fitdist(x, "norm")
fit.lognorm <- fitdist(x, "lnorm")
fit.gamma <- fitdist(x, "gamma")

plot(fit.weibull)
plot(fit.norm)
plot(fit.lognorm)
plot(fit.gamma)

fit.weibull$aic
fit.norm$aic
fit.lognorm$aic
fit.gamma$aic

meanlog.x <- fit.lognorm$estimate[1]
sdlog.x <- fit.lognorm$estimate[2]
n <- length(x)
sim <- 10000
ks.sim <- rep(0,sim)

for(i in 1:sim)
{
  resample.x <- rlnorm(n,meanlog = meanlog.x,sdlog = sdlog.x )
  ks.sim[i] <- as.numeric(ks.test(resample.x,"plnorm",
                        meanlog = meanlog.x,sdlog = sdlog.x)$statistic)
}

plot(ecdf(ks.sim), las = 1,
     main = "KS-test statistic simulation (CDF)", col = "darkorange", lwd =
  1.7)
grid()
```

```r

fit.final <- logspline(ks.sim)

1 - plogspline(ks.test(x,"plnorm"
                       ,meanlog = meanlog.x,sdlog = sdlog.x)$statistic
                       ,fit.final)

x.ecdf <- ecdf(x)
plot(x.ecdf,
     las = 1,
     main = "ECDF",
     col = "darkorange",
     lwd = 1,cex=0.2
       )

Alpha=0.05
Eps=sqrt(log(2/Alpha)/(2*n))
grid<-seq(min(x),max(x), length.out = 10000)
lines(grid, pmin(x.ecdf(grid)+Eps,1),col='red')
lines(grid, pmax(x.ecdf(grid)-Eps,0),col='blue')

sum(plnorm(grid,meanlog.x,sdlog.x) >=pmax(x.ecdf(grid)-Eps,0) &
plnorm(grid,meanlog.x,sdlog.x) <=pmin(x.ecdf(grid)+Eps,1))/100

mu.hat <- sum(log(x))/n
sd.hat <- sqrt(sum((log(x) - mu.hat)^2)/n)

B <- 10000
mu.hat.star.p <- rep(0,B)
mu.hat.star.np <- rep(0,B)
for(i in 1:B)
{
  x.p <- rlnorm(n,mu.hat,sd.hat)
  x.np <- sample(x,size=n,replace=TRUE)
  mu.hat.star.p[i] <- sum(log(x.p))/n
  mu.hat.star.np[i] <- sum(log(x.np))/n
}
se.hat.boot.p <- sd(mu.hat.star.p)
se.hat.boot.np <- sd(mu.hat.star.np)

par(mfrow <- c(2,1))
hist(mu.hat.star.p,main = "Parametric Method",prob=TRUE)
lines(density(mu.hat.star.p),col="red")
hist(mu.hat.star.np,main = "Parametric vs Non-Parametric Method",prob=TRUE)
lines(density(mu.hat.star.np),col="blue")

library("tidyverse")
theta.np <- as.data.frame(mu.hat.star.np)
theta.p <- as.data.frame(mu.hat.star.p)

names(theta.np) <- "values"
names(theta.p) <- "values"
theta.np$type <- 'Parametric Bootstrap'
theta.p$type <- 'Non-Parametric Bootstrap'
dis <- rbind(theta.p,theta.np)
     dis <- as.data.frame(dis)
     ggplot(dis, aes(values, fill = type)) + geom_density(alpha = 1) +
```

```
        scale_fill_grey() +
        ggtitle("Parametric vs Non-Parametric Distribution of Mean")

      #normal ci, at 95% we get z.alpha/2 approx equal to 2
      normal.np<-c(mu.hat-2*se.hat.boot.np, mu.hat+2*se.hat.boot.np)
      #pivotal ci at 95%
      pivotal.np<-c(2*mu.hat-quantile(mu.hat.star.np,0.975),
                    2*mu.hat-quantile(mu.hat.star.np,0.025))
      #quantile ci at 95%
      quantile.np<-quantile(mu.hat.star.np, c(0.025, 0.975))

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

      y.4 <- Computers$price[Computers$ram=="4"]
      y.8 <- Computers$price[Computers$ram=="8"]

      df <- c(y.4,y.8)
      length <- length(y.4)+length(y.8)
      options(expressions=1e5)
      B <- 10000
      perm.matrix <- replicate(B, sample(df,length,replace=F))
      perm.T <- rep(0,B)
      for(i in 1:B){perm.T[i] <- abs(mean(perm.matrix[1:length(y.4),i])-
mean(perm.matrix[length(y.4):length(y.8),i]))}
      print(paste("The p-value for permutation test is",p.value <-
mean(perm.T==0)))

      sim <- 1000
      posterior.x <- as.data.frame(
        rlnorm(sim,meanlog = mu.hat,sdlog = sd.hat/sqrt(n)))
      names(posterior.x) <- "value"
      ggplot(posterior.x, aes(x=value,fill = value)) + geom_density(alpha = 1)
+
        scale_fill_grey() +
        ggtitle("Posterior Distribution of Mean")
```