Probability Models Project|

Analyzing Prices of Personal Computers

Abstract:

To compute the mean prices of computers between the years 1993 – 1995. The data consists of specifications of computers including the likes of processor speed, clock speed, size of head drive, size of RAM, size of Screen etc. The mean price during the 90's era was approx. 2200 $, while ~ 60 % of computers price ranged between 2000 $ and 4000 $. An explicit study applying different statistical techniques were implied to identify the difference in price for a Premium vs Non-Premium brand of computers. The statistics showed the mean price differed for Premium Computers and Non-Premium Computers and of practical use at the same time.

By
Anupreet Gupta (M12823098)
Palash Arora (M12935022)

INDEX

1. INTRODUCTION

Anupreet and Palash (also read as the authors of this report) had bought personal computer back in the time when they were young and dumb (Nowhere close to spelling out Statistical Learning!). But now, when they think back about the time, they feel they could have made a wiser decision if they had known about in details what factors influences the price of the computer and to what extent.

With the zeal to implement and replicate the learning from the class of Probability Models: BANA 7031 taught by Prof. Peng Wang, a dataset *Computers* from the *Ecdat* package in R was chosen as the area of study. The dataset Computers consists of 6259 observations with information including processor speed, clock speed, size of head drive, size of RAM, size of Screen etc. We analyzed various metrics such as price distribution and comparison in order to estimate trends in computer prices between 1993 to 1995. A close look at whether the brand of the computer had a significant impact in price or not was determined and it’s practical significance is determined.

The following analytics techniques have been applied:

- Identifying distribution type
- Empirical CDF
- Bootstrap standard errors and confidence intervals
- MLE and its asymptotic distributions
- Hypothesis testing
- Bayesian analysis

Palash always was inclined towards a bigger screen computer, while Anupreet wanted to go for premium computer brand. A closed look at the study does reveal a correlation coefficient of 0.3 between screen size and price while Anupreet have realized that he would have to shell out approx. 160 $ more to buy a premium brand personal computer.

2. THE PROJECT

Summary of Data:

We have used *Computers* data from *Ecdat* package in R. It is a data frame containing price information for personal computers between years 1993 to 1995 and contains below metrics:

Price: price in US dollars of 486 PCs

speed: clock speed in MHz

hd: size of hard drive in MB

ram: size of Ram in in MB

screen: size of screen in inches

cd: is a CD-ROM present ?

multi: is a multimedia kit (speakers, sound card) included ?

premium: is the manufacturer was a "premium" firm (IBM, COMPAQ) ?

ads: number of 486 price listings for each month

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

Data Exploration:

- Data set contains 6,259 observations with 10 variables
- There are no missing values

S. No.	Column Name	Class Type	Levels	# Missing Values
1.	Price	Numeric		0
2.	Speed	Numeric		0
3.	HD	Numeric		0
4.	Ram	Numeric		0
5.	Screen	Numeric		0
6.	CD	Factor w/2 levels	Yes ; No	0
7.	Multi	Factor w/2 levels	Yes ; No	0
8.	Premium	Factor w/2 levels	Yes ; No	0
9.	ADS	Numeric		0
10.	Trend	Numeric		0

Table 1. Summary of Columns

Data Validation:

The focus of the study is mapping price against different specifications of personal computers. It is of most importance to understand the distribution of price, which would further act as a baseline in carrying out statistical techniques like Parametric Bootstrapping.

Using the ***fitdistrplus*** package from R, function descdist() was applied to check for the distribution. This plotted a *Cullen and Frey graph* as shown below for the price metric for all observations.

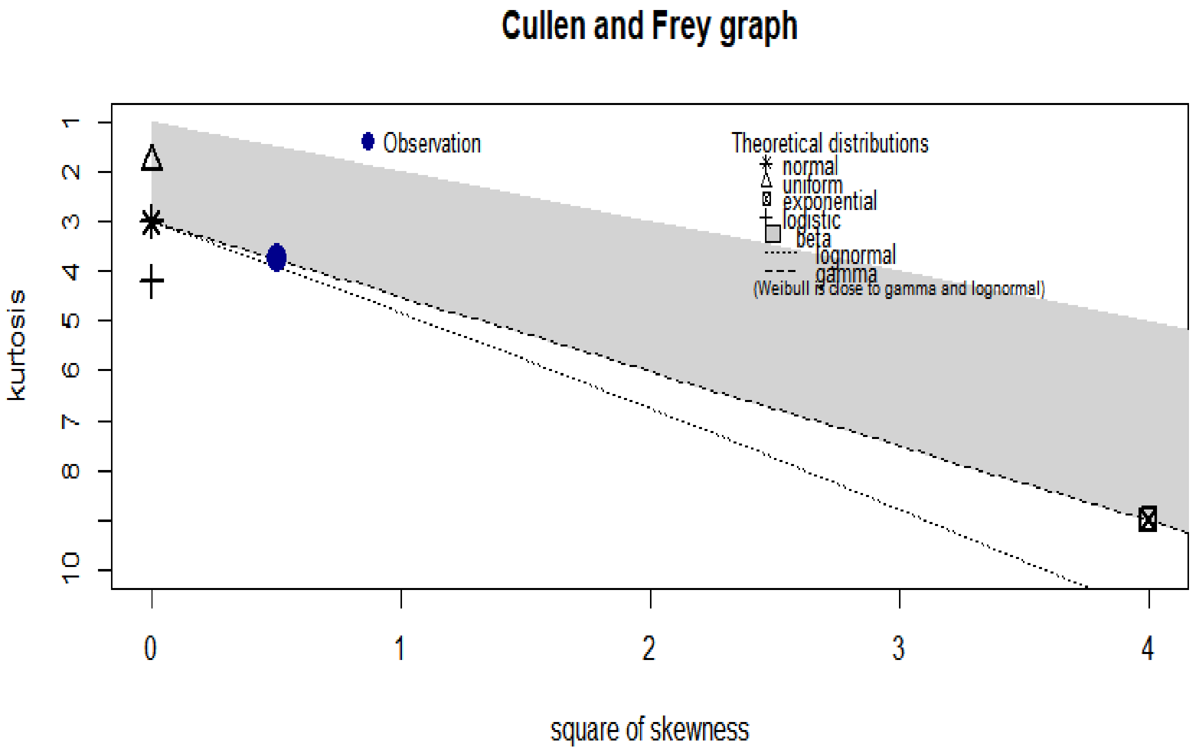

Figure 1. Cullen and Frey Graph for Price

The kurtosis and squared skewness of data is plotted as a blue point named "Observation". It seems that possible distributions follow the Normal distribution.

Also, to re-confirm the results from the above graph, a normal distribution has been plotted above and over the distribution of price to check for any difference. The below graphs confirm that the price distribution here follow Normal Distribution.

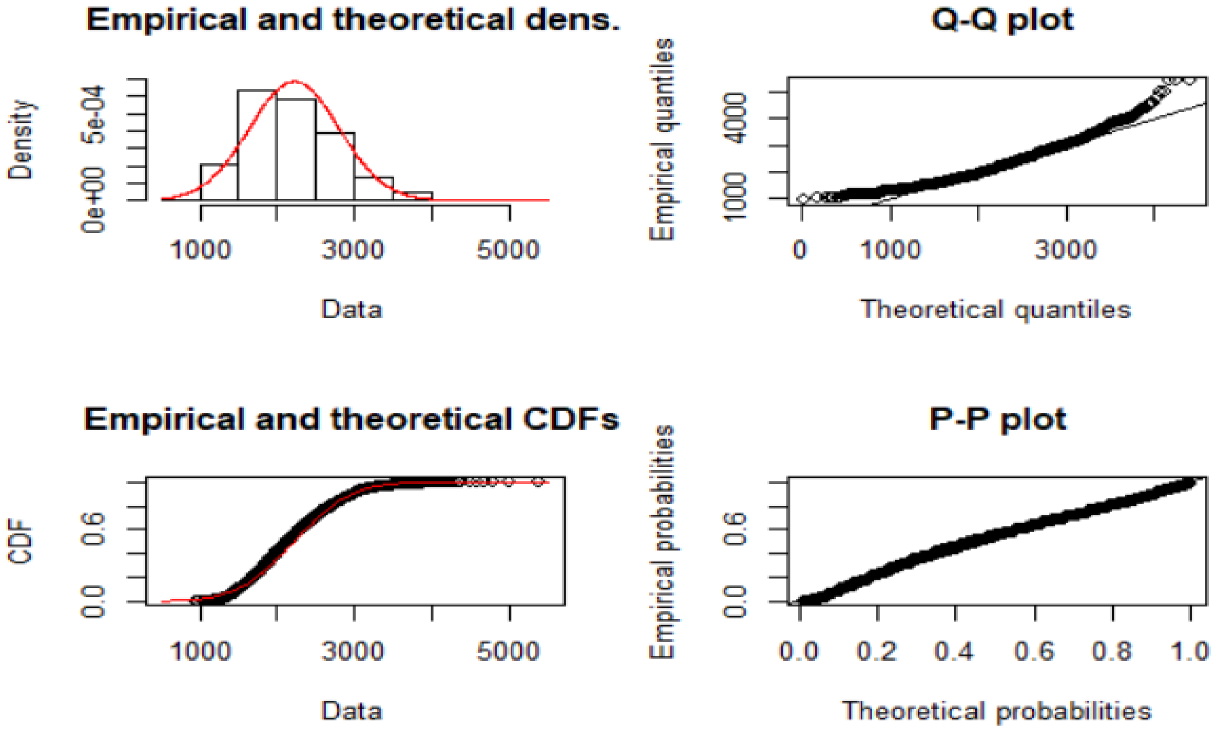

Figure 2. Normal Plot over the Price Distribution

Statistical Analysis:

ECDF |

Finding proportion of personal computers between price range $4000 and $2000

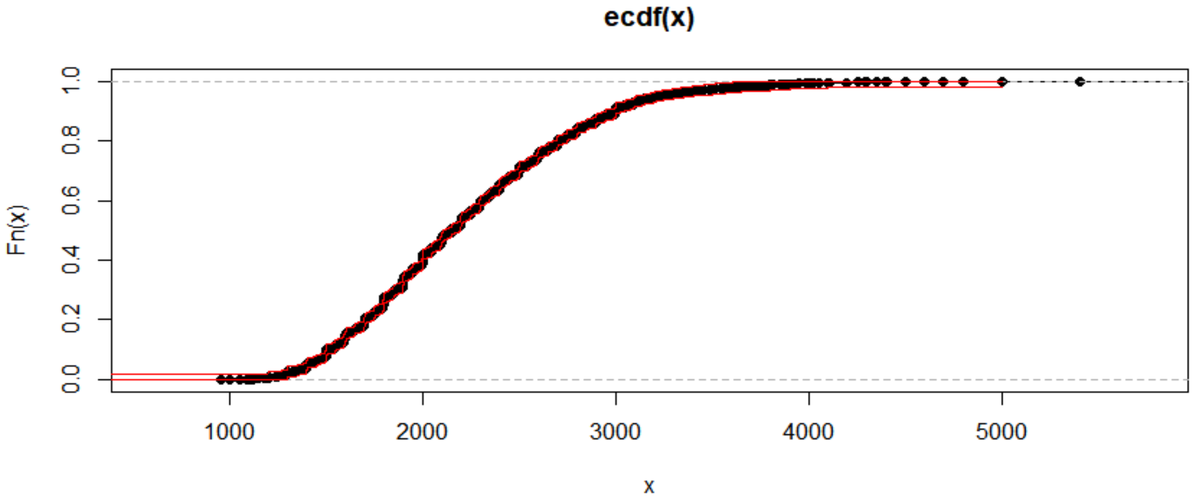

Figure 3. ECDF Plot for Price

The distribution of the price could have been determined and implemented here as Empirical CDF, where every observation has been given equal weightage of 1/6529 and CDF is plotted.

57.4% of computers in 1993 to 1995 costed between $2000 to $4000.

Woah! That does translate to the fact that both Anupreet and Palash had to be prepared to shell out (with 60 % of the times probability) somewhere between 2000 $ and 4000 $ to buy a personal computer.

Bootstrap Standard Errors and Confidence Intervals |

Estimating correlation between computer price and screen size.

Focusing on Palash's priority to buy a personal computer and getting things clear for him, about how are prices and screen size are related, we implemented the Non-Parametric bootstrap.

Since the sample data for Computers was having only 6259 rows of which the distribution was unknown to us (if not for the Cullen and Frey graph), thus we applied non-parametric bootstrap for estimating correlation between screen-size and price of computers measuring those properties when sampling from an approximating distribution. The data was resampled for 3200 samples.

Standard errors and confidence intervals were determined to understand the 95 % probability of where the correlation coefficient would lie if we repeated this survey data experiment.

Confidence Interval Type	Range 2.5 %	Range 97.5 %
Normal	*0.271042*	*0.321041*
Pivotal	*0.2710582*	*0.3209274*
Quantile	*0.2711555*	*0.3210247*

Table 1. Summary of Confidence Intervals for Correlation using Non-Parametric Bootstrap

We are 95% confident that correlation between Computer prices and screen size is between 0.27 to 0.32.

The rate of standard error obtained is *0.01249974.*

Hence, approximately price and screen size follow a positive correlation of 0.3 which is not strong correlation but still a positive.

An interesting fact (based on small sample set), this trend has reversed and people buying laptops in 21st century are required to pay for smaller screen laptop (say 14' inch) in comparison to larger screen (say 17' inch), given every other specification remains same.

MLE and its Asymptotic Distributions |

Estimating difference of means between Premium Vs Non-Premium Computers

Coming to the next problem statement, is to establish and estimate the difference in price for a premium against a non-premium brand of personal computers.

A premium brand personal computer here is one that is manufactured by Intel or IBM.

The data was further analyzed for obtaining the Maximum likelihood estimator for the difference between the mean price for premium against non-premium personal computer.

MLE attempts to find the parameter values that maximize the likelihood function, given the observations. The resulting estimate is called a maximum likelihood estimate, which is also abbreviated as MLE. (*Source: Wikipedia*)

MLE estimate for the mean difference in price is obtained as ***-157.7862 $*** i.e mean price for a Non-premium computer is higher than a Premium brand (Surprising! Need to double sure of this fact!)

Standard error for the difference in price is obtained as 27.4414.

Hence, we are 95 % confident that premium bands computers cost approx. 103 to 212 $ less than the non-premium brand computer.

Parametric Bootstrap |

Estimating difference in average price difference of premium Vs non-premium computers

While MLE had provided an accurate idea about the difference in mean price for a premium against a non-premium brand of personal computer, Anupreet wasn’t sure and somewhere was convinced with the conclusion that how could a premium brand of personal computer cost lesser than a non-premium brand. (though back in mind, he liked the idea of getting a branded personal computer for a relatively lesser price)

Since, the data set was not convincingly big, Anupreet started to question the sanity of the data and decided to go for Parametric bootstrapping, where on the assumptions and proven above that the price distribution follow a Normal distribution, 3200 times the process was repeated; everytime taking samples equal to the original dataset for both premium and non-premium data set and estimating the mean price for them.

Standard error of the distribution of the difference in mean price is 26.98486 and we are 95 % confident that this difference in price lies between (-211.7560, -103.8165)

Wald Test and Wilcox Test |

To compare means/median of Premium Vs Non-Premium Computers

While, it has been confirmed twice that the premium brands of personal computers cost approx. 160 $ less than non-premium brand of personal computer. It does no harm to Anupreet to apply Wald Test and Wilcoxon Test and re-affirm his findings and conclusions that the prices isn’t same for both the types of computers.

Wald Test

sd_p : Standard deviation for price of Premium brand of Personal Computers.

sd_{np} : Standard deviation for price of Non-Premium brand of Personal Computers.

μ_p : Mean of price for Premium brand of Personal Computers.

μ_{np} : Mean of price for Non-Premium brand of Personal Computers.

$H_o : \mu_p = \mu_{np}$

$H_a : \mu_p \neq \mu_{np}$

mu_hat = $\mu_p - \mu_{np}$

sigma_hat = sd_p /sqrt(n1)+ sd_{np} /sqrt(n2)

z.stat <- (mu_hat-0)/sigma_hat

Using the Z-statistic obtained here or say W; we obtain the corresponding P-value which is less than 0.05.

Hence we can reject null hypothesis and the means of Premium Vs non-premium computers are not equal at level 0.05 %.

Wilcoxon Test

On a similar fashion, we combined the two data set and ranked them. Next we carried out the Wilcoxon test on this to identify if they belong to the same distribution i.e the medians for Premium vs Non-Premium is same or not.

The p-value obtained again is less than 0.05 and is clear indication that at alpha level=0.05, we can reject our null hypothesis of that the two medians are equal.

Based on Wilcoxon test, we are 95 % confident that the difference in medians for the two types of computers under study lies in the range (-181 ,- 80).

Bayesian Analysis |

Finding mean difference between Premium Vs Non-premium Computer Prices
We have a prior belief that the price for premium and non-premium bands of personal computers follow Normal distribution and that of their mean is equal to 1.

Posterior distribution of their difference in mean values is proportional to product of likelihood function and prior.

Hence, we calculate the posterior distribution and estimate the mean of the posterior distribution and 95 % confidence band.

Observation: We are 95% confident that on an average, difference between prices premium Vs non-premium computers lie between $ - 210.6 and $ - 104.

3. CONCLUSIONS

Data:

1. Our metric of concern, Price variable of Computers data is normally distributed
2. Using ECDF we found that 57.4% of computers in 1993 to 1995 costed between $2000 to $4000
3. Using bootstrapping we observed that correlation between Computer prices and screen size is between 0.27 to 0.32

Determining price difference between Premium Vs non-premium computers:

4. According to Wald test and Wilcox tests, the prices of premium and non-premium computers are unequal
5. Using MLE, we determined that premium computers are $160 cheaper than non-premium computers
6. Using parametric bootstrap, we found that premium computers cost less than non-premium computers by about $103.8 to $211.7
7. We then validated this using Bayesian analysis where we found that on an average, difference between prices premium Vs non-premium computers lie between $210.6 and $104

4. BIBLIOGRAPHY

Data:
https://rdrr.io/cran/Ecdat/man/Computers.html
https://cran.r-project.org/web/packages/Ecdat/index.html

Determining distribution of variables:
https://stats.stackexchange.com/questions/132652/how-to-determine-which-distribution-fits-my-data-best

Bayesian Analysis:
https://www.stata.com/features/overview/bayesian-intro/

Packages:
Tidyverse: https://www.tidyverse.org/
Ecdat: https://cran.r-project.org/web/packages/Ecdat/index.html
fitdistrplus : https://cran.r-project.org/web/packages/fitdistrplus/index.html

5. Appendix

```
# Library -----------------------------------------------------------------
library(Ecdat)
library(tidyverse)

# Computers ---------------------------------------------------------------
data(Computers)
?Computers
str(Computers)
dim(Computers)

colSums(is.na(Computers))
hist(Computers$price)

# Checking for normality --------------------------------------------------
install.packages("fitdistrplus")
library(fitdistrplus)

descdist(Computers$price, discrete = FALSE)
normal_dist <- fitdist(Computers$price, distr="norm")
plot(normal_dist)

# https://stats.stackexchange.com/questions/132652/how-to-determine-which-distribution-fits-my-data-best

# ECDF --------------------------------------------------------------------
set.seed(7031)
f <- ecdf(Computers$price)
plot.ecdf(Computers$price)
Alpha=0.05
n=length(Computers$price)
Eps=sqrt(log(2/Alpha)/(2*n))
grid<-seq(0,5000, length.out = 10000)
lines(grid, pmin(f(grid)+Eps,1),col="red")
lines(grid, pmax(f(grid)-Eps,0),col="red")

f(4000) - f(2000)
```

```
# Non-Parametric Bootstrap ------------------------------------------------
set.seed(7031)
(cor.hat <- cor(Computers$price,Computers$screen))

# 0.2960415

theta <- function(x,xdata){
  cor(xdata[x,"price"],xdata[x,"screen"])
}

#Bootstraping
library(bootstrap)
B=3200
cor.boot<-bootstrap(1:nrow(Computers), B, theta,Computers)

(se.boot=sqrt(var(cor.boot$thetastar)))
# 0.01249974

(normal.ci<-c(cor.hat-2*se.boot, cor.hat+2*se.boot))
#  0.271042 0.321041

(pivatol.ci<-c(2*cor.hat-quantile(cor.boot$thetastar,0.975), 2*cor.hat-
quantile(cor.boot$thetastar,0.025)))

# 97.5%      2.5%
# 0.2710582 0.3209274

(quantile.ci<-quantile(cor.boot$thetastar, c(0.025, 0.975)))

# 2.5%     97.5%
# 0.2711555 0.3210247


# MLE ---------------------------------------------------------------------
x=Computers$price[Computers$premium=="yes"]
y=Computers$price[Computers$premium=="no"]

n1=length(x)
mu_hat1=mean(x)
sigma_hat1<-sd(x)
```

```
n2=length(y)
mu_hat2=mean(y)
sigma_hat2<-sd(y)

mu_hat=mean(x)-mean(y)
mu_hat
# -157.7862

sigma_hat<-sqrt(var(x)/n1+var(y)/n2)
sigma_hat
# [1] 27.4414

# Parametric Bootstrap ---------------------------------------------------------

tau.hat_bootstrap=vector()
n_obs=length(data)
for(i in 1:3200){
  X_i=rnorm(n1,mu_hat1,sigma_hat1)
  Y_i=rnorm(n2,mu_hat2,sigma_hat2)
  tau.hat_bootstrap[i]=mean(X_i)-mean(Y_i)
}

##Or use replicate function
tau_hat_bootstrap<-replicate(3200, mean(rnorm(n1,mu_hat1,sigma_hat1))-
mean(rnorm(n2,mu_hat2,sigma_hat2)))

tau.hat_bootstrap_se=sd(tau.hat_bootstrap)
tau.hat_bootstrap_se

#[1]  26.98486

#Confidence Interval
c(mu_hat-2*tau.hat_bootstrap_se,mu_hat+2*tau.hat_bootstrap_se)

#[1] -211.7560 -103.8165
```

```
# Wald Test ------------------------------------------------------------

z.stat<-(mu_hat-0)/sigma_hat
#P(|Z|>|w|)=2*P(Z<-|w|)
p.value=2*(pnorm(-abs(z.stat)))

p.value < 0.05
# TRUE


# Wilcox Test ---------------------------------------------------------
wilcox.test(x, y,conf.int = T,exact=F)

# data:  x and y
# W = 1510700, p-value = 3.091e-07
# alternative hypothesis: true location shift is not equal to 0
# 95 percent confidence interval:
#   -180.99998  -80.00003
# sample estimates:
#   difference in location
# -129

# Posterior ------------------------------------------------------------
#Calculating posterior samples
posterior = rnorm(1000,mean = mu_hat1, sd = sigma_hat1/sqrt(n1)) - rnorm(1000,mean = 
mu_hat2, sd = sigma_hat2/sqrt(n2))

#To find mean
mean(posterior)

#95% confidence interval
quantile(posterior, c(0.025, 0.975))
```