**Statistical mathematics project work**

Trieu Huynh Ba Nguyen

**Task 1:**

1. Geometric distribution represents the probability of the number of successive failures before the first success. It is calculated with the function:

with k: number of trials (k = 1, 2, 3…)

p: probability of each success

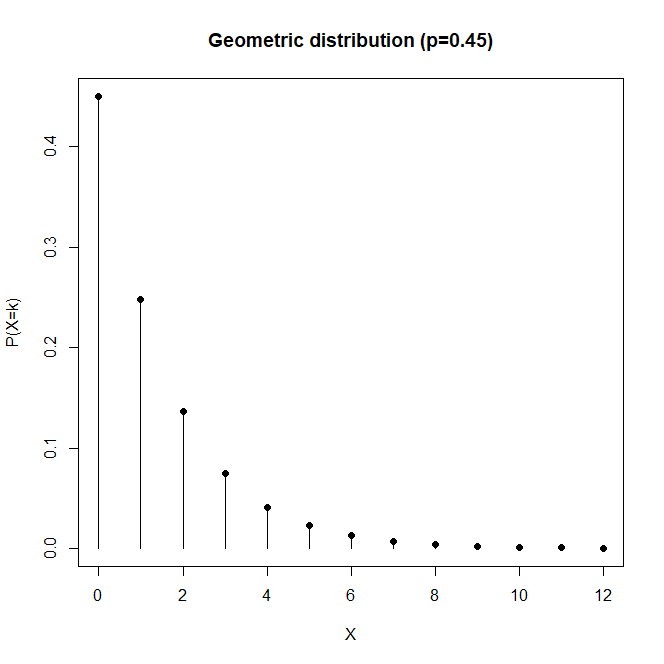
1. R code:

X <- 0:12

Y <- dgeom(X, *p* = 0.45)

plot(X, Y, *type* = "h", *main* = "Geometric distribution (p=0.45)", *ylab* = "P(X=k)")

points(X, Y, *pch* = 16)



**Task 2:**

1. Binominal distribution represents the probability of the number of successes or failures occurring during a number of trials. It is calculated with the function:

with k: number of successes or failures (k = 1, 2, 3…n)

n: total number of trials

p: probability of each success

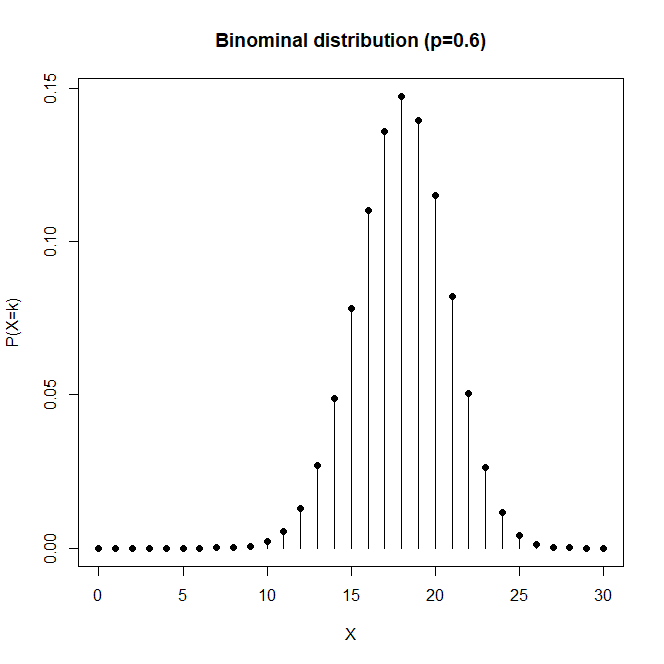
1. R code:

X <- 0:30

Y <- dbinom(X, *size* = 30, *prob* = .6)

plot(X, Y, *type* = "h", *main* = "Binominal distribution (p=0.6)", *ylab* = "P(X=k)")

points(X, Y, *pch* = 16)



**Task 3:**

1. Poisson distribution represents the probability of a given number of events occurring in an interval of time. It is calculated with the function:

with k: the number of occurrences (k = 1, 2, 3…)

e: Euler's number (e=2.71828…)

𝜆: lambda – total number of events divided by the number of units in the data

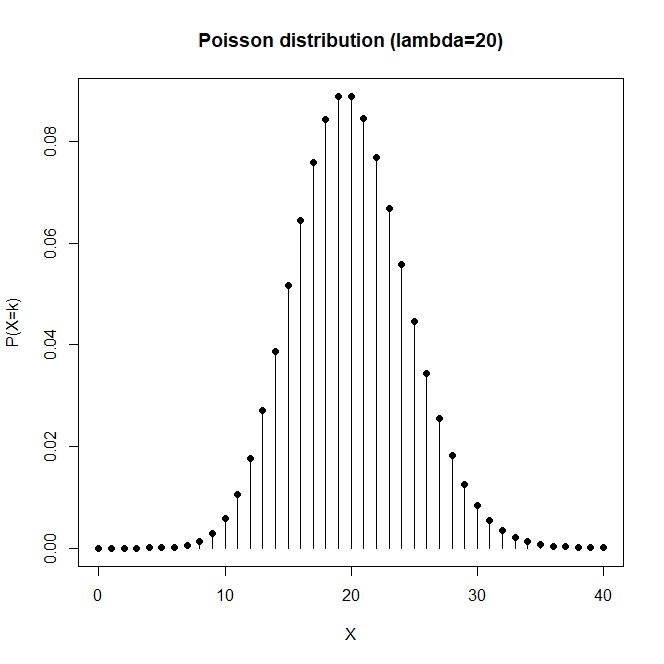
1. R code:

X <- 0:40

Y <- dpois(X, *lambda* = 20)

plot(X, Y, *type* = "h", *main* = "Poisson distribution (lambda=20)", *ylab* = "P(X=k)")

points(X, Y, *pch* = 16)



**Task 4:**

1. There is one column named “Val”.
2. There are 1029 rows.
3. Min = 4.193534

min(data\_set1)

1. Max = 109.379

max(data\_set1)

1. Mean = 50.49665

mean(data\_set1$Val)

1. Median = 50.52415

median(data\_set1$Val)

1. Variance = 218.7175

var(data\_set1$Val)

1. Standard deviation = 14.7891

sd(data\_set1$Val)

**Task 5:**

R code:

library(readr)

data\_set1 <- read\_csv("data\_set1.csv")

X <- 0:100

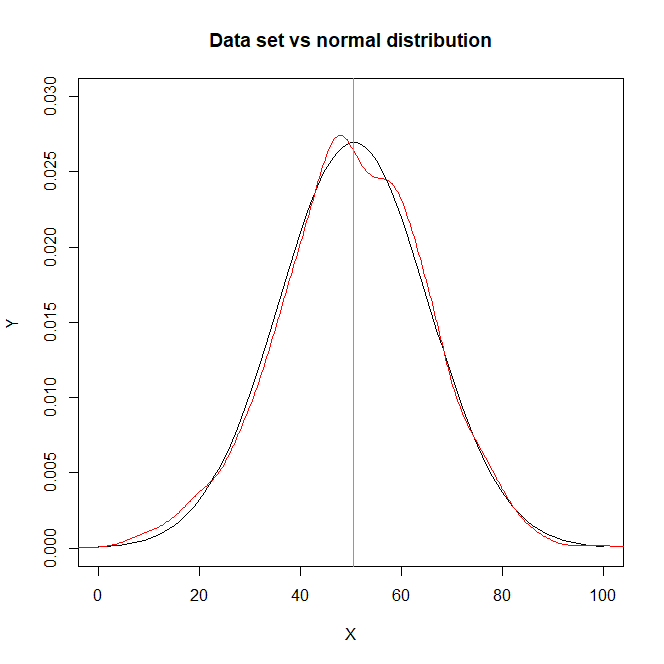
Y <- dnorm(X, *mean* = mean(data\_set1$Val), *sd* = sd(data\_set1$Val))

plot(X, Y, *type* = "l", *ylim* = c(0, 0.03), *main* = "Data set vs normal distribution")

d <- density(data\_set1$Val, *bw* = 3)

points(d, *col* = "red", *type* = "l")

abline(*v* = mean(data\_set1$Val), *col* = "green")

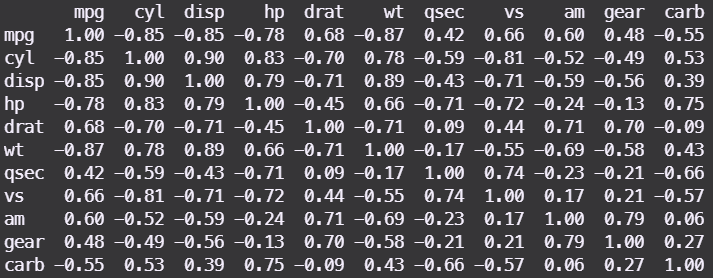
****

**Task 6:**

4 variables most correlated with hp: mpg, cyl, disp, carb.

cars <- mtcars

round(cor(cars), *digits* = 2)



**Task 7:**

R code:

model <- lm(hp ~ cyl + disp + carb + mpg, *data* = mtcars)

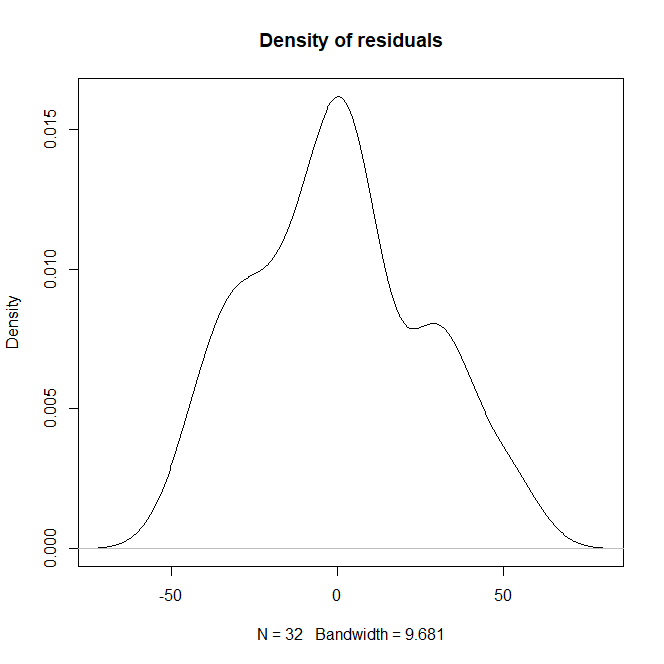
hp\_hat <- predict(model)

residuals <- mtcars$hp - hp\_hat

hpplot <- density(residuals)

plot(hpplot, *main* = "Density of residuals")

summary(model)$r.squared



The plot has a bell shape.

R-squared = 0.8594845

Therefore, the model is correct and accurate.

**Task 8:**

R code:

library(readr)

data\_set2 <- read\_csv("data\_set2.csv")

X <- min(data\_set2):max(data\_set2)

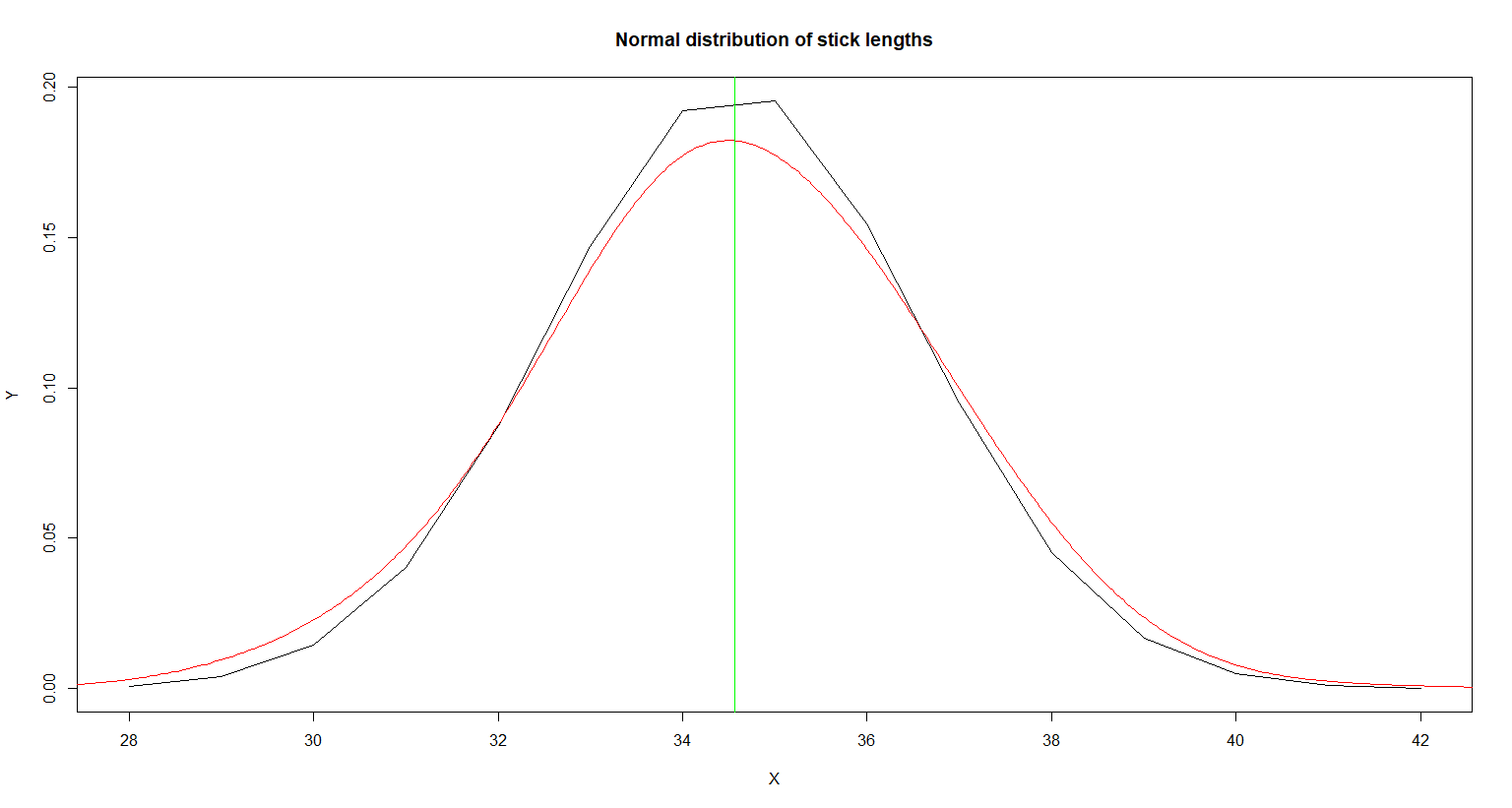
Y <- dnorm(X, *mean* = mean(data\_set2$Val), *sd* = sd(data\_set2$Val))

plot(X, Y, *type* = "l", *main* = "Normal distribution of stick lengths")

d <- density(data\_set2$Val, *bw* = 1)

points(d, *col* = "red", *type* = "l")

abline(*v* = mean(data\_set2$Val), *col* = "green")



The length of the sticks is not acceptable as the mean and most values are much higher than the null hypothesis of µ = 30.