University Of Cyprus, M.Sc. In Data Science
DSC 530: Probabilities and Statistics Project 2

Nikolas Petrou
petrou.p.nikolas@ucy.ac.cy

Rafail Agathokleous
ragath01@ucy.ac.cy

Fotis Kyriakou
fkyria01@ucy.ac.cy

Problem 1

Part A

It is given that $X_1, \dots, X_n$ is a random sample from a Bernoulli distribution with parameter (probability of success) p. The parameter of interest is the log-odds parameter $\theta = \log \frac{p}{1-p}$. An estimator must be suggested for θ.

For Bernoulli Distribution, it is known that $\hat{p}$ is the Maximum Likelihood Estimator (MLE) of p. The MLE of p is calculated as follows:

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

$$l(p) = \log L(p) = \log p^{\sum_{i=1}^{n} x_i} + \log((1-p)^{n-\sum_{i=1}^{n} x_i})$$

$$\frac{\partial l(p)}{\partial p} = \frac{\sum_{i=1}^{n} x_i}{p} + (n - \sum_{i=1}^{n} x_i)\frac{-1}{1-p} = 0$$

$$\Leftrightarrow p = \frac{\sum_{i=1}^{n} x_i}{n} \Rightarrow \hat{p} = \bar{X}$$

Hence, $\hat{p} = \bar{X}$ is the MLE of p.

Let $g(p) = \theta = \log \frac{p}{1-p}$. Based on the *equivariance* property of MLE, since $\hat{p}$ is the MLE estimator of p, then $g(\hat{p})$ is the MLE of $g(p)$. Therefore, the suggester estimator is the following:

$$\hat{\theta} = g(\hat{p}) = \log \frac{\hat{p}}{1-\hat{p}}$$

In order to derive the asymptotic distribution of the estimator and calculate a Confidence Interval (CI) for the estimator, as well as the estimated Standard Error $\hat{se}(\hat{\theta})$ must be handed out.

To quantify the uncertainty about $g(\hat{p})$, the *Delta Method* [1] was utilized in order to quantify the uncertainty by using what is known about the uncertainty of $\hat{p}$ itself. Based on *Delta Method* $g'(p) \neq 0$, then $\hat{se}(\hat{\theta}) = |g'(\hat{p})|\hat{se}(\hat{p})$.

$$g'(p) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$

Since $p \in [0, 1]$, $g'(p) \neq 0$. Therefore, indeed $\hat{se}(\hat{\theta}) = |g'(\hat{p})|\hat{se}(\hat{p})$.

From theory, it is known that $\hat{se}(\hat{p}) = \sqrt{\frac{1}{I_n(\hat{p})}}$

$$I_n(p) = nI(p) = \frac{n}{p(1-p)}$$

$$\Rightarrow I_n(\hat{p}) = \frac{n}{\hat{p}(1-\hat{p})} \Rightarrow \hat{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Therefore,

$$\begin{aligned}\hat{se}(\hat{\theta}) &= |g'(\hat{p})|\, s\hat{e}(\hat{p}) \\ &= \left|\frac{1}{\hat{p}(1-\hat{p})}\right| \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ &= \frac{1}{\sqrt{n\hat{p}(1-\hat{p})}}, \quad \text{for } \hat{p} \in [0,1]\end{aligned}$$

Based on the Central Limit Theorem and the Delta Method the Asymptotic Distribution in `R` was shown, (approximates standard Normal for large sample sizes n):

$$\hat{\theta} \rightarrow_{n\rightarrow\infty} N\left(\log\left(\frac{p}{1-p}\right), \frac{1}{n\hat{p}(1-\hat{p})}\right)$$

$$\Rightarrow \frac{\hat{\theta}-\theta}{\hat{se}(\hat{\theta})} \rightarrow_{n\rightarrow\infty} N(0,1)$$

The Normal-based CI for $\hat{\theta}$, based on the properties of Normal distribution, for a $(1-a)$% is the following:

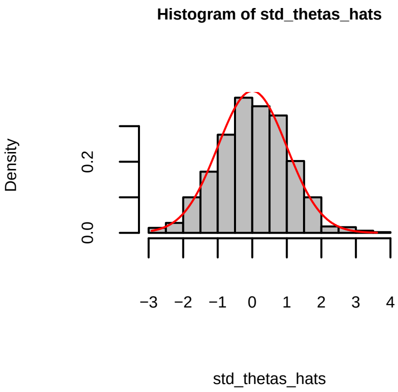
$$\hat{\theta} \pm z_{\alpha/2}\hat{se}(\hat{\theta}) \tag{1}$$

First, a function `simulate.mle()` which returns the MLE $\hat{\theta}$ and Standard Error $\hat{se}(\hat{\theta})$ estimators as well as the observations of the simulation was implemented in order to reduce code repetition:

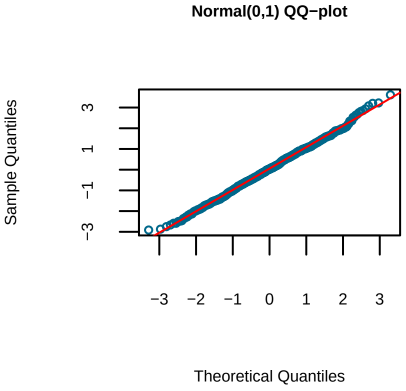
```r
set.seed(420) # Reproducibility
# Function to return log-odds for given prob p
get.log.odds <- function(p){
  log(p/(1-p))
}
# Function which returns the MLE, SE estimators
# as well as the observations of the simulation
simulate.mle <- function(n, p){
  # Generate Bernoulli RVs. Size=1,
  # since interested in  successes or failures only
  X <- rbinom(n=n, size=1, prob=p)
  # Calculate the X_bar which is the theta_hat
  p_hat <- mean(X)
  # Get MLE for the random variables
  theta_hat <- get.log.odds(p_hat)
  # Get Standard Error estimator
  se_hat <- 1 /sqrt(n*p_hat*(1-p_hat))
  # Return MLE, Standard Error estimators, and X
  return(c(theta_hat, se_hat, X))
}
```

Next, $M = 1000$ simulations of $n = 10000$ samples were performed with a randomly set value of $p \in [0, 1]$ in order to confirm the Asymptotic distribution:

```r
# Setting values of p and n
p <- runif(n=1)
n <- 10000
# Simulate for M times to confirm Normality of thetas
M <- 1000
simulated.values <- replicate(M, simulate.mle(n, p))
# Extract the theta_hats
theta_hats <- simulated.values[1,]
# Extract the SE_hats
se_hats <- simulated.values [2,]
# Confirm Normality of theta_hats, based on Delta Method
std_thetas_hats<-(theta_hats-get.log.odds(p))/se_hats
hist(std_thetas_hats, prob=TRUE, col = 'grey', cex=0.5,
cex.lab=0.5, cex.axis=0.5, cex.main=0.5, cex.sub=0.5)
x <- seq(min(std_thetas_hats),
max(std_thetas_hats),0.001)
lines(x, dnorm(x), col='red')
```



```
qqnorm(std_thetas_hats, main='Normal(0,1) QQ-plot', c
cex=0.5, cex.lab=0.5, cex.axis=0.5, cex.main=0.5, cex
qqline(std_thetas_hats, col='red',
cex=0.5, cex.lab=0.5, cex.axis=0.5, cex.main=0.5, cex
```



Indeed, the plotted histogram, QQ-plot show the Normality of the $\hat{\theta}$.

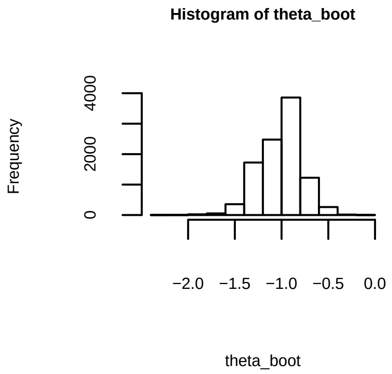
Part B

For this part, different values for p and sample sizes n were sued in order to compare the Normal Based Interval of Equation 1 and the Normal Interval $T \pm Z_{\alpha/2}\hat{se}_{\text{boot}}$, $\hat{se}_{\text{boot}} = \sqrt{v_{\text{boot}}}$ by using Bootstrap (sampling with replacement).

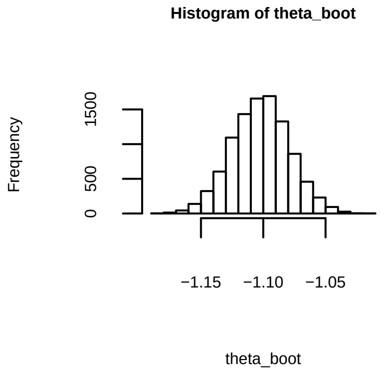
For Bootstrap, $B = 10000$ resampling iterations were used. In addition, a 95% CI was used ($a = 0.05$).

```
# Function that estimates theta with MLE and
# compares the asymptotic distribution with
# Bootstrap for the given n and p values
compare.asymptotic.distribution <- function(n, p, a=0.05){
  cat("Running for p=", p, "and n=", n,"\n")
  # Bootstrap iterations(sampling with replacement)
  B <- 10000
  # Find MLE and Standard Error Estimators for given
  # using our custom defined function
  vals <- simulate.mle(n, p)
  theta_hat <- vals[1]
  se_hat <- vals[2]
  # Rest values are the observations of the
  # simulation see function simulate.mle(n, p)
  X <- vals[3:length(vals)]
  # Print estimated theta and theta
  cat("Estimated theta (theta_hat):", theta_hat, "\n")
  cat("Actual theta:", get.log.odds(p), "\n")
  # Find the CI of the MLE
  ci <- c(theta_hat - qnorm(1-a/2)*se_hat,
          theta_hat + qnorm(1-a/2)*se_hat)
  cat("95% CI for theta_hat:", ci, "\n")
  # Bootstrapping
  theta_boot <- c() # Init empty vector
  for (i in 1:B){
    X_star <- sample(X, size=length(X), replace=TRUE)
    T_boot_i <- get.log.odds(mean(X_star))
    theta_boot <- append(theta_boot, T_boot_i)
  }
  se.theta.boot <- sd(theta_boot)
  ci_boot <- c(theta_hat - qnorm(1-a/2)*se.theta.boot,
               theta_hat + qnorm(1-a/2)*se.theta.boot)
  cat("95% CI using Bootstrap:", ci_boot, "\n")
  # Observe Bootstrap distribution
  hist(theta_boot, cex=0.5, cex.lab=0.5,
       cex.axis=0.5, cex.main=0.5, cex.sub=0.5)
}

# Testing for different p and n values
for (p in c(0.25,  0.6))
  for (n in c(100, 10000))
    compare.asymptotic.distribution(n, p)
```

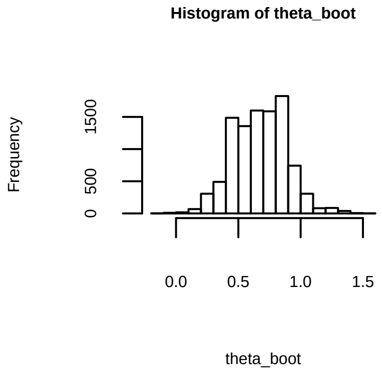
```
## Running for p= 0.25 and n= 100
## Estimated theta (theta_hat): -0.9946226
## Actual theta: -1.098612
## 95% CI for theta_hat: -1.436096 -0.5531489
## 95% CI using Bootstrap: -1.445135 -0.5441098
```



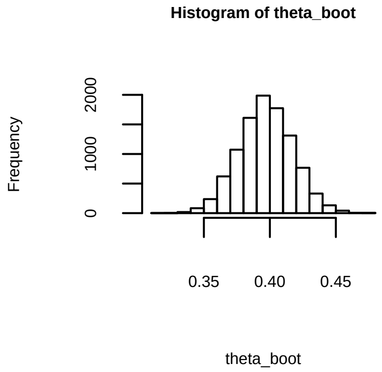
```
## Running for p= 0.25 and n= 10000
## Estimated theta (theta_hat): -1.101815
## Actual theta: -1.098612
## 95% CI for theta_hat: -1.147115 -1.056515
## 95% CI using Bootstrap: -1.147314 -1.056315
```



```
## Running for p= 0.6 and n= 100
## Estimated theta (theta_hat): 0.6632942
## Actual theta: 0.4054651
## 95% CI for theta_hat: 0.2495455 1.077043
## 95% CI using Bootstrap: 0.2389808 1.087608
```



```
## Running for p= 0.6 and n= 10000
## Estimated theta (theta_hat): 0.3963068
## Actual theta: 0.4054651
## 95% CI for theta_hat: 0.3563354 0.4362781
## 95% CI using Bootstrap: 0.3563547 0.4362589
```



Some conclusions by interpreting the results, as n (sample sizes) increase Bootstrap's $\hat{se}_{\text{boot}}$ seem to better approximate the Normal distribution. Regarding the CI, as n increases, the interval of the CIs is more precise (the space between the two ”extreme” values is smaller). Finally, as n increases, Bootstrap's CI better agree with the CI of the MLE and vice-versa.

Problem 2

This problem, requires estimating the number of trials N that a hospital will require, in order to acquire the needed number of patients suffering from hemorrhagic strokes ($r = 10$) to start their clinical trial.

To solve the problem, the Probability Mass Function (PMF) of the Negative Binomial distribution was used. More specifically, an alternative version of the formula was used [2]:

$$f(N; r, p) = P(X = N) = \begin{pmatrix} N-1 \\ r-1 \end{pmatrix} p^r (1-p)^{N-r}$$

- **r**: The number of required successful trials
- **p**: The probability of success for each trial
- **N**: Denotes the trial at which the **r**th success occurs where $N = r, r+1, \dots$

The trial **N** must be estimated, where the 10th success would have occurred, for $p = 0.13$ and $r = 10$ thus the likelihood of N;

$$\begin{aligned} L(N) &= P(X = N) = \\ &= \begin{pmatrix} N-1 \\ 10-1 \end{pmatrix} 0.13^{10}(1-0.13)^{N-10} \\ &= \frac{(N-1)!}{(9)!(N-10)!} 0.13^{10}(0.87)^{N-10} \\ \Rightarrow L(N) &= \frac{(N-1)\dots(N-9)}{(9)!} 0.13^{10}(0.87)^{N-10} \end{aligned}$$

From the Likelihood we have

$$\begin{aligned} \frac{L(N+1)}{L(N)} &= \frac{\frac{N\dots(N-8)}{(9)!}0.13^{10}(0.87)^{N-9}}{\frac{(N-1)\dots(N-9)}{(9)!}0.13^{10}(0.87)^{N-10}} \\ &= \frac{N}{N-9}0.87 \end{aligned}$$

The likelihood is decreasing for $L(N+1) < L(N)$ which is equivalent to

$$\frac{N}{N-9}0.87 < 1 \Rightarrow N > 69.23$$

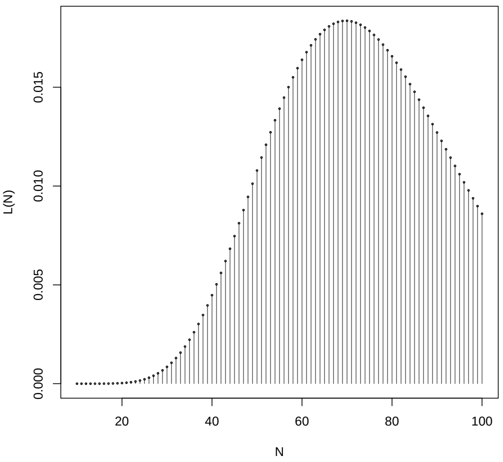
To maximize the likelihood,we want the largest (integer) values of N satisfying this constraint, that is

$$\hat{N} = 70$$

Meaning that, it was estimated that the 10th ”success” (Patient suffering from hemorrhagic strokes) will be reached on the 70th trial.

To verify these results, the density of the Negative Binomial Distribution was generated and plotted with the parameters of this problem-exercise, in order to estimate the maximum value for the likelihood of N.

```
# The parameter of successes for the Negative
#Bionomial Distribution is 10
r <- 10
# Values of Nrange start from r
Nrange <- r:100
# The default PMF for the Negative Bionomial of R
# express the probability of failures instead of
# the total N, thus use N-r failures
distri <- dnbinom(Nrange-r, size=r, 0.13)
plot(Nrange, distri, xlab = "N",  ylab = "L(N)",
                            type = 'h', lwd=0.1)
points(Nrange, distri, xlab = "N", ylab = "L(N)",
                              pch=1, cex=0.3)
```



```
# The maximum value for the likelihood
Nrange[which.max(distri)]

## [1] 70
```

The results in R, both graphically and numerically, indeed validate the results from the calculations above.(also from the above graph we saw the pmf of N)

Problem 3

Given that $X_1, \ldots, X_n$ is a random sample from the Geometric Distribution with parameter p. It is required, to calculate the Maximum Likelihood Estimator (MLE) of p. The likelihood function is defined as:

$$L(p) = \prod_{i=1}^{n} p(1-p)^{x} = p^{n}(1-p)^{\sum_{i=1}^{n} xi}$$

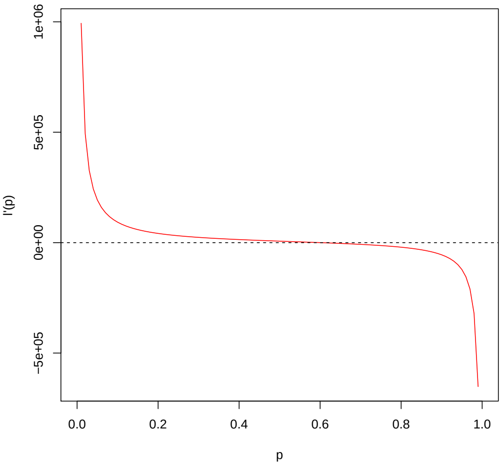
$$l(p) = logL(p) = n \log{(p)} + \sum_{i=1}^{n} x_i \log(1-p)$$

$$\Rightarrow \frac{\partial \ell(p)}{\partial p} = \frac{n}{p} - \sum_{i=1}^{n} x_i \frac{1}{1-p} = 0$$

$$\Rightarrow \hat{p} = \frac{1}{\bar{X}+1}$$

Now $\hat{p}$ can be used as an estimator for p and it will be compared with the MLE obtained through the Newton-Raphson method.

```
set.seed(420)
n <- 10000

# Using random value for actual p
p <- runif(n=1)
x <- rgeom(n, p)
# The l'(p) (score function but for many x_i)
# which was found earlier
func.mle <- function(p){
  n/p - sum(x)/(1-p)
}
# Observe where the root is approximately
curve(func.mle, col='red', ylab="l'(p)",xlab="p")
abline(a=0, b=0, lty=8)
```

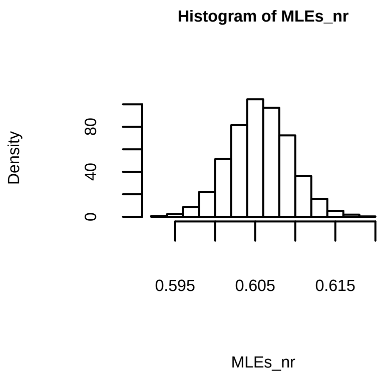


```
# Newton-Raphson method. Using [0, 1] interval
# since the root p is a probability
nr <- uniroot(func.mle, interval=c(0, 1))
# Total iterations of Newton-Raphson Method
nr$iter

## [1] 7

# MLE for p_hat with Newton-Raphson Method
nr$root

## [1] 0.6017165

# Theoretical p_hat with MLE
1/(mean(x)+1)

## [1] 0.6017209

# Actual p
p

## [1] 0.605539
```

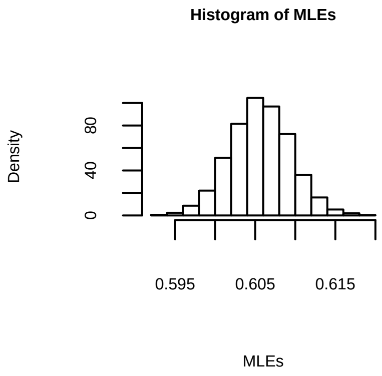
Through the above process, it was observed that Newton Raphson's MLE is approximately identical to the exact answer.

Furthermore, the asymptotic distribution of both the actual MLE and Newton Raphson's MLE by simulating $M = 5000$ times was studied.

```
# Studying asymptotic normality by simulation
M <- 5000
MLEs <- c() # Has the simulated theoretical MLEs
MLEs_nr <- c() # Has the simulated MLEs of Newton-Raphson
for (i in 1:M){
  # Simulate samples for every iteration
  x <- rgeom(n, p)
  # Saving the MLEs of both methods for every iteration
  MLEs_nr <- append(MLEs_nr, uniroot(func.mle,
                    interval=c(0, 1))$root)
  MLEs <- append(MLEs, 1/(mean(x)+1))
}

# Histograms will potentially show normality
# Hist for Newton Raphson
hist(MLEs_nr,prob=TRUE,cex=0.5, cex.lab=0.5,
     cex.axis=0.5, cex.main=0.5, cex.sub=0.5)
```



```
# Hist for theoretical MLE
hist(MLEs,prob=TRUE,cex=0.5, cex.lab=0.5,
     cex.axis=0.5, cex.main=0.5, cex.sub=0.5)
```



Through the above code it was concluded that both histograms approximate the normal distribution and are similar to each other. This is reasonable as both of these approaches for the MLE gave very similar results, and their histograms are symmetric around the actual p of interest.

Problem 4

Supposing that George is an average student, preparing his application for a Msc degree program. One of the requirements to be accepted in the program is to have proof for the knowledge of the English language.

Problem statement: Will George meet the English Language conditions?

George has decided to take the TOEFL exams, which provide candidates with certificates of the English language. For George pass the requirements of the specific program, he must get a TOEFL(IBT) score greater or equal than 104 (out of 120). The question of the problem is, *"Is the required scored easily obtainable?"*. George wants a fair idea about his chance for a particular degree.

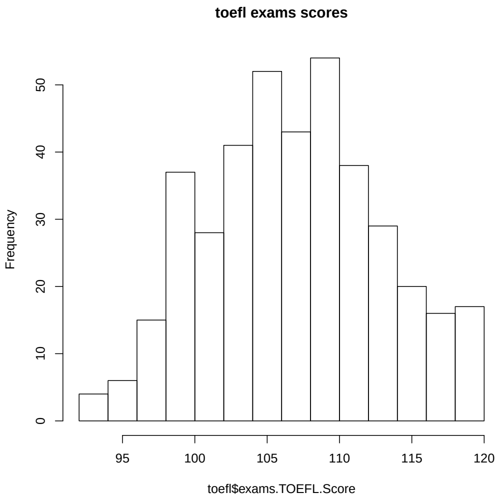
The dataset *AdmissionPredict.csv* [1] contains nine variables corresponding to 400 applicants for master's degrees. The variable which contains the TOEFL exam scores was examined. A portion of the data is given below:

```
exams <- read.csv("Admission_Predict.csv")
exams <- data.frame(exams)
toefl<-data.frame(exams$TOEFL.Score)
```

```
head(toefl,10)

##    exams.TOEFL.Score
## 1                118
## 2                107
## 3                104
## 4                110
## 5                103
## 6                115
## 7                109
## 8                101
## 9                102
## 10               108
```

First, it was confirmed that there were no missing values, and an Exploratory Data Analysis was performed.

```
sum(is.na(toefl$exams.TOEFL.Score))

## [1] 0

hist(toefl$exams.TOEFL.Score,
                main = 'toefl exams scores')
```



A model had to be proposed for the given data to begin addressing the initial question. It was proposed to model the data as IID (Independent and Identically Distributed) draws from the Normal distribution.

Why is this a suitable model?

Firstly, from the above histogram, it was shown that the observations follow a bell-shaped distribution. Moreover, the observations scores were for $n = 400$ students and there are a lot of cases where scores of a large amount of individuals is Normally distributed. The normality is going to be furthered examined during the next steps.

What assumptions are being made?

An assumption in this case was that a random sample, i.e. each student in the population has an equal chance of being included in the sample. This reduces the chance that differences in materials or conditions gives strongly bias results. Furthermore, statistical independence was also assumed, in the sense that the TOEFL score of one student does not influence the scores of other students.

Are these assumptions reasonable?

Non-random samples introduce bias and can result in incorrect interpretations. In addition, non-independent observations introduce bias and could make statistical tests give too many false results. Before we proceed to examine if our model is correct first we will set our estimators for μ and σ^2

[1] https://www.kaggle.com/mohansacharya/graduate-admissions?select=Admission_Predict.csv

Method of Moments Estimators

For the task, estimators were utilized. The estimators based on the method of moments will be introduced.

Let $X_1, X_2, \ldots, X_n$ be normal variables with μ and σ^2. For this work $n = 400$ (number of students). The method of moments estimators of μ and σ^2 is: The first and the second theoretical moments are

$$E\left(X_i\right) = \mu \quad E\left(X_i^2\right) = \sigma^2 + \mu^2$$

Equating the first theoretical moment with the corresponding sample moment, the method of moments estimator for μ is the sample mean:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

And, substituting the sample mean in for μ in the second equation and solving for σ^2 , then the method of moments estimator for the variance is:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$$

which is an unbiased estimator of the variance. Now based on these estimators we are going to further examination of the normality of the data.

```
median(toefl$exams.TOEFL.Score)

## [1] 107
```

```
median(toefl$exams.TOEFL.Score)-
             mean(toefl$exams.TOEFL.Score)

## [1] -0.41
```

From the above histogram, it is shown that the observations seems to spread symmetrically below the median. The mean and the median is approximately equal.

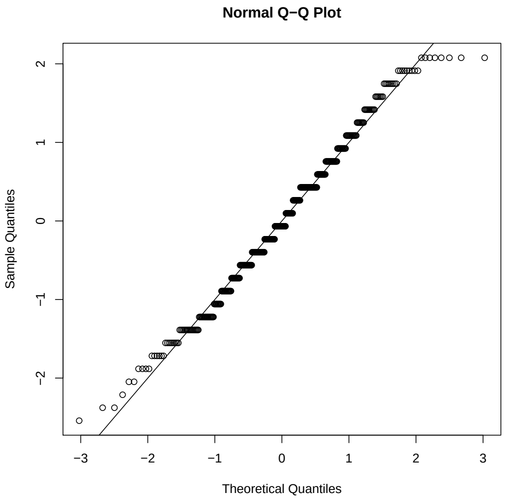
```
toefl<-toefl$exams.TOEFL.Score
n=400
mean.hat<-sum(toefl)/n
mean.hat

## [1] 107.41
```

```
var.hat<-(sum((toefl-mean.hat)^2))/n
var.hat

## [1] 36.7469
```

```
sd.hat<-sqrt(var.hat)
```

The data was manually standardized.

```
stdtoefl<-(toefl-mean.hat)/sd.hat
qqnorm(stdtoefl)
abline(0,1)
```



The fact that the points are approximately on the $y = x$ line, shows that the variable is approximately normal. This further supports that the proposed model is approximately good.

Answering the initial question

George wants a fair idea about his chance for a particular degree.

The theoretical probability, for George, to get a grade in the exam greater or equal to 104 under the assumed model, where $X \sim$ Normal $\left(\hat{\mu}, \hat{\sigma^2}\right)$

$P(X\geq$ 104)=1-P(X < 104)=1-P$\left(\frac{X-\hat{\mu}}{\hat{\sigma}} > \frac{104-\hat{\mu}}{\hat{\sigma}}\right)$

$$= 1 - \Phi\left(\frac{104 - \hat{\mu}}{\hat{\sigma}}\right)$$

```
##P[X>=104]
answer<-1-pnorm((104-mean.hat)/sd.hat)
answer

## [1] 0.7131218
```

Therefore, based on the above said, the probability is approximately equal to 0.7131.

References

[1] G. W. Oehlert. A note on the delta method. *The American Statistician*, 46(1):27–29, 1992.

[2] K. Siegrist. The negative binomial distribution.