**In The Name of God**

**Statistical Inference**
**Homework #2**

***Amin Asadi Sarijaloo***
***810196410***

---

# Problem 1

a. False, because the coin doesn't have memory, and the tosses are independent. Thus the chance for every toss to become head or tail is equal to 50%.

b. False, because if two random variables A and B are independent, we have that P(A | B) = P(A) which means that knowing B, gives no information about A, but for mutual exclusiveness, we must have P(A ∩ B) = 0 which means knowing that B has happened, A can not happen, hence they are dependent(except when at least one of the events is impossible).

c. False, the general additions rule states that: The probability of either X or Y occurring is: P(X U Y) = P(X) + P(Y) - P(A ∩ B)

d. False, we have that :
   - P(A | B) = P(A ∩ B) / P(B)
   - P(B | A) = P(A ∩ B) / P(A)

   So in general → P(A | B) ≠ P(B | A)

e. True

f. False, P(A and B) = 0 means that A and B are mutually exclusive. (One of them having zero probability is not necessarily true).

g. False, the mean and variance of Poisson Distributions are both equal to lambda which is the expected number of the occurrences of a particular event in unit measurement.

h. True

i. True

---

## Problem 2

X: being infected to COVID-19
Y: speaking other languages than French

P(X) = 14.6%
P(Y) = 20.7%
P(X ∩ Y) = 4.2%

a. No they aren't, because P(X and Y) = 4.2% ≠ 0

b. P(X ∩ ~Y) = P(X) - P(X ∩ Y) = 14.6% - 4.2% = 12.4%

c. P(X U Y) = P(X) + P(Y) - P(X ∩ Y) = 14.6% + 20.7% - 4.2% = 31.1%

d. P(~X ∩ ~Y) = 100% - P(X U Y) = 68.9%

e.
- P(X ∩ Y) = 4.2%
- P(X) * P(Y) = 14.6% * 20.7% = 3.02%
- → P(X ∩ Y) ≠ P(X) * P(Y)

because the necessary condition for independence doesn't hold, they are not independent.

---

## Problem 3

We denote the random variable of points he won in each game of round i with xi

**Calculating $E(x_1)$:**
He predicts the correct winner with probability ½ , that is +1 with probability ½ and +0 with probability ½ , so

$$E(x_1) = (1 * ½ + 0 * 1/2 ) = ½$$

**Calculating $E(x_i)$:**
He predicts the correct winner with probability $(½)^i$ because the predicted team must have won in all the (i-1) previous rounds, so because he earns $2^{i-1}$ for each correct prediction:

$$E(x_i) = (2^{i-1} * (½)^i + 0) = ½$$

Thus the expected value of the total points he wins is =

$$(32+16+8+4+2+1) * ½ = 63 * ½ = 31.5$$

---

## Problem 4

$$P(L^C) = 1 - P(L) = 0.9$$
$$P(M_1 \mid L^C) = 1 - P(M_1{}^C \mid L^C) = 0.1$$

a.

$$P(L|M_1) = \frac{P(M_1 \mid L)\, P(L)}{P(M_1 \mid L)\, P(L) + P(M_1 \mid L^C)\, P(L^C)} = \frac{0.9 * 0.1}{0.9 * 0.1 + 0.1 * 0.9} = 0.5$$

b. $P(L|M_1 \cap M_2) = \frac{P(M_1 \cap M_2|\, L)\, P(L)}{P(M_1 \cap M_2|\, L)\, P(L) + P(M_1 \cap M_2|\, L^C)\, P(L^C)} =$

$$\frac{P(M_1|\, L)\, P(M_2|\, L)\, P(L)}{P(M_1|\, L)\, P(M_2|\, L)\, P(L) + P(M_1|\, L)\, P(M_2|\, L^C)\, P(L^C)} = \frac{0.9^2 * 0.1}{0.9^2 * 0.1 + 0.1^2 * 0.9} = 0.9$$

c. Yes, they are equivalent. We can use both methods for updating probabilities and they give the same results. This is because in both methods we use the same evidence to update our probabilities. The order of applying evidence does not matter.

---

## Problem 5

P(a pill is damaged) = 0.005
P(a pill is not damaged) = 1 - 0.005 = 0.995

a.

This has a binomial distribution.
Each pill of 100 pills in a box is healthy independent from other pills with a probability of 0.995 and damaged with a probability of 0.005.

$$P(box\ having\ no\ damaged\ pills) =$$

$$choose(100,\ 0)\ \times 0.005^0 \times 0.995^{100} \times 100\% = 60.5\,\%$$

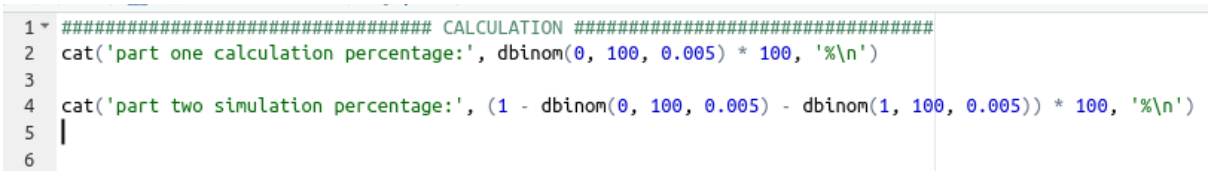
b.

$$P(Having\ 2\ or\ more\ damaged\ pills) =$$

$$1 - choose(100,\ 0)\ \times 0.005^0 \times 0.995^{100} -$$
$$choose(100,\ 1)\ \times 0.005^1 \times 0.995^{99} = 8.9\,\%$$

c. In the image below the result of the R-script is the same as the result of parts a and b:

- Calculation with “dbinom” function of R:



```
1  ################################## CALCULATION ##################################
2  cat('part one calculation percentage:', dbinom(0, 100, 0.005) * 100, '%\n')
3
4  cat('part two simulation percentage:', (1 - dbinom(0, 100, 0.005) - dbinom(1, 100, 0.005)) * 100, '%\n')
5
6
```

- Simulation: I created 1000 boxes with 100 pills in each. Each pill was damaged with a probability of 0.5%. Then I calculated the average number of boxes having no damaged pills and the average number of boxes having at least 2 damaged pills among all 1000 boxes:

```r
################################## SIMULATION ################################
create_one_box <- function() {
  boxes = list(sample(c(0, 1), size = 100, replace = TRUE, prob = c(0.005, 0.995)));
  return(boxes)
}

create_boxes <- function() {
  num_boxes <- 1000;
  boxes <- c()
  for (i in 1:num_boxes)
    boxes <-c(boxes, create_one_box())
  return(boxes)
}

#### PART a : Finding percentage of boxes which have no damaged pill
num_epoches <- 100
all_healthy_percentages <- c()
for (i in 1:num_epoches) {
  boxes <- create_boxes()
  num_healthy_boxes = sum(sapply(boxes, function (x) sum(unlist(x)) == 100))
  percentage = num_healthy_boxes / num_boxes * 100
  all_healthy_percentages <- c(all_healthy_percentages, percentage)
}

cat('part one simulation percentage:', mean(all_healthy_percentages), '\n')


atleast_two_damaged_percentages <- c()
for (i in 1:num_epoches) {
  boxes <- create_boxes()
  num_healthy_boxes = sum(sapply(boxes, function (x) sum(unlist(x)) <= 98))
  percentage = num_healthy_boxes / num_boxes * 100
  atleast_two_damaged_percentages <- c(atleast_two_damaged_percentages, percentage)
}

cat('part two simulation percentage:', mean(atleast_two_damaged_percentages), '\n')
```

Result of R codes:

As we can see the result of both the calculation code and simulation code are equal to what we found in parts a and b:

```
> source('~/Desktop/uni/Spring 00/SI/hws/hw2/p5.r')
part one calculation percentage: 60.57704 %
part two simulation percentage: 8.982231 %
part one simulation percentage: 60.589 %
part two simulation percentage: 8.941 %
>
```

## Problem 6

Since every passenger has two possibilities(showing up or not), each of them have a Bernoulli distribution.

**a.**

Let T be the random variable indicating the total number of people who show up. The objective is to find the probability of T being more than 300(overbook). T has binomial distribution:

$$m = E(T) = Np = 324 \times 0.9 = 291.6$$
$$\sigma = \sigma(T) = \sqrt{Np(1-p)} = \sqrt{324 \times 0.9 \times 0.1 = 5.4}$$
$$P(T \geq 301) = P(T \geq 301 - 0.5) = P(\frac{T-m}{\sigma} \geq \frac{301-0.5-m}{\sigma})$$
$$1 - \phi(\frac{301-0.5-324\times0.9}{\sqrt{324\times0.9\times0.1}}) = 0.04$$

**b.**

There are 324/2=162 pairs. Each pair will travel with probability of 0.9*0.9 = 0.81 independent of other pairs.
mean of the pairs will be 162*0.81=131.22
mean of the people will be $2 \times 162 \times 0.81 = 262.44$
Std of pairs will be $\sqrt{162 \times 0.81 \times (1 - 0.81)}$
Std of people will be $\sqrt{2 \times 162 \times 0.81 \times (1 - 0.81)} = 7.06$

So the mean is decreased a lot but std is increased a little. So the probability is decreased comparing to the previous part.

# Problem 7

a.

Shorthand for exams = N(462, 119)
Shorthand for projects = N(584, 151)

```
1  library(ggplot2)
2
3  #a
4  exam_mean <- 462;
5  exam_std <- 119;
6  project_mean <-584;
7  project_std <- 151
8  # Shorthand for exams = N(462, 119)
9  # Shorthand for projects = N(584, 151)
10
11 exam_score <- 620;
12 project_score <- 670;
13
```

---

b.

```
14 #b
15 exam_z_score = (exam_score - exam_mean) / exam_std;
16 project_z_score = (project_score - project_mean) / project_std;
17
18 cat('exam z-score is:', exam_z_score, '\n')
19 cat('project z-score is', project_z_score, '\n')
```

Result:

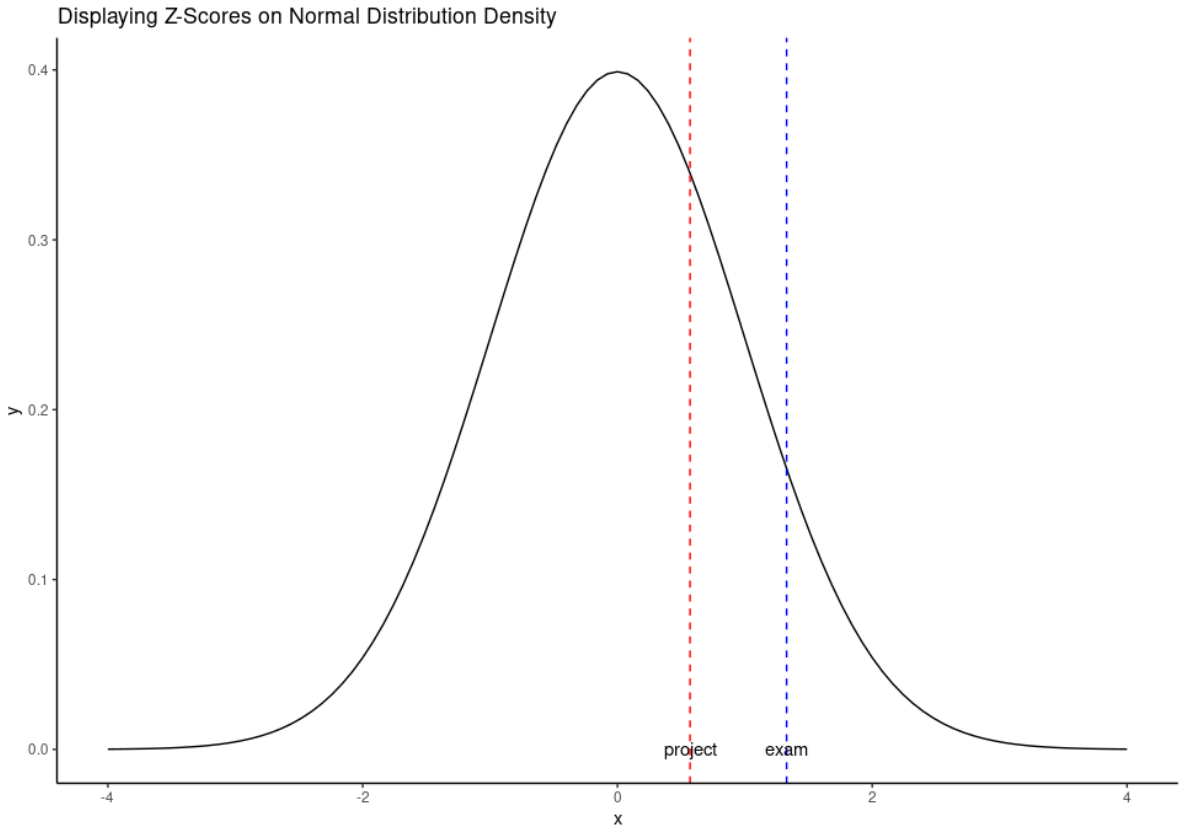
```
exam z-score is: 1.327731
project z-score is 0.5695364
```

---

c.

```r
#c
p1 <- ggplot(data.frame(x = seq(-4, 4, length=100)), aes(x = x)) +

        stat_function(fun = dnorm) +

        geom_vline(xintercept = c(exam_z_score, project_z_score),
                                    colour=c('blue', 'red'),
                                    linetype = "dashed") +
        geom_text(aes(x = c(exam_z_score, project_z_score),
                        y = 0,
                        label = c('exam', 'project')),
                    data=data.frame(c(exam_z_score, project_z_score)))
print(p1)

```

Result:




---

d.

a z-score tells us that how far is a datapoint from the mean, i.e. It measures the number of standard deviations an observation falls below or above the mean. So her exam score is 1.32 standard deviations far from the mean and her project score is 0.56 standard deviations far from the mean.

---

e.

z-score of her exam is more than z-score of her project so she performed better in exam.

---

f.

```r
#f
exam_percentile_score <- pnorm(exam_z_score, 0, 1)
project_percentile_score <- pnorm(project_z_score, 0, 1)
cat('percentile score of exam:', exam_percentile_score , '\n')
cat('percentile score of project ', project_percentile_score, '\n')

```

Result:

```
percentile score of exam: 0.9078665
percentile score of project  0.7155039
```

---

g.

```r
#g
cat(100 - exam_percentile_score * 100, "students performed better that her in exam\n")

```

Result:

```
9.213348  % of students performed better that her in exam
```

---

h.

```r
#h
cat(100 - project_percentile_score * 100, "students performed better that her in project\n")

```

Result:

```
28.44961 % of students performed better that her in project
```

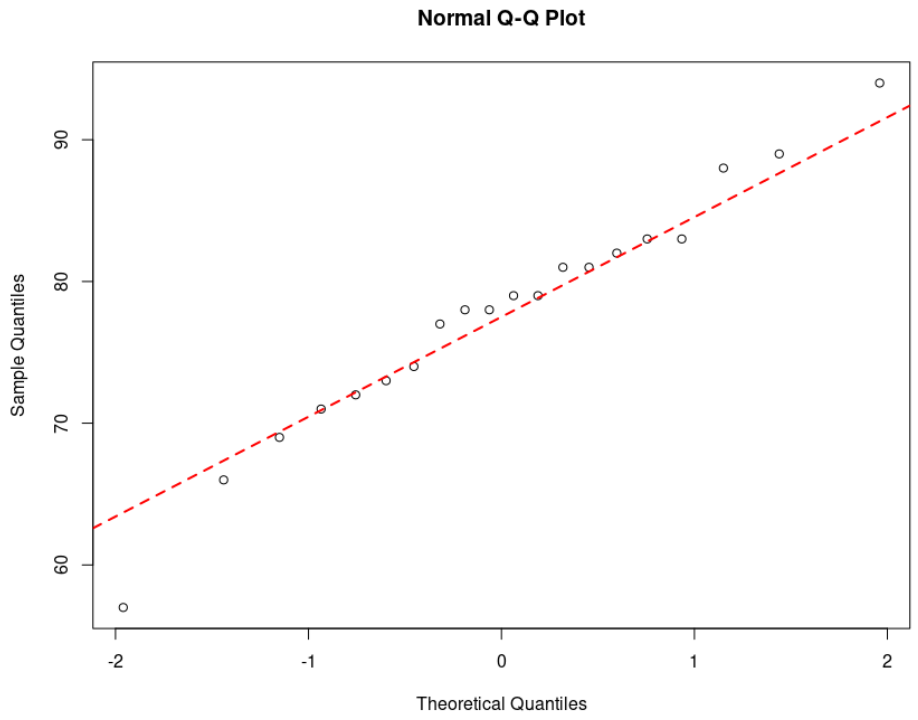
---

i.

```r
#i
scores <- c(57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94)

## QQplot
qqnorm(scores)
qqline(scores, col='red', lwd=2, lty=2)

## Histogram
h <- hist(scores)
xfit <- seq(min(scores), max(scores), by=0.01)
yfit <- dnorm(xfit, mean = mean(scores), sd = sd(scores)) * 5 * length(scores)
lines(xfit, yfit, col="blue")

## Box Plot
boxplot(scores)
title(main = "scores box plot", ylab = "score")
```

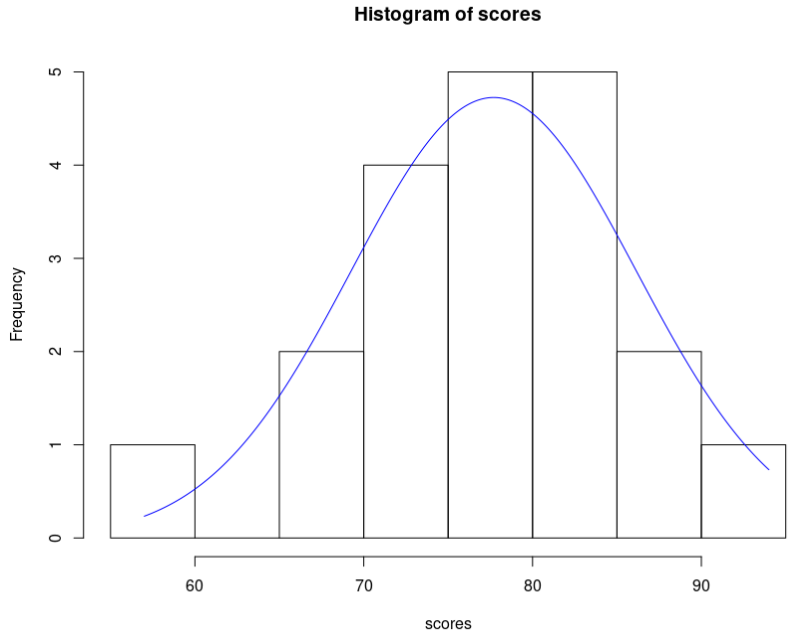
Result:

QQ Plot:
We can see that at some parts specially at first before fist quartile the qqplot of scores is different from qqplot of normal distribution but approximately they are similar.




Histogram:

The histogram and normal curve have similar shapes in general but they are different before first quartile of data.




Box Plot:

Because the mean and median are close to the center of the box, the distribution is close to normal but the box plot is not symmetric.


## Problem 8

a) $P_1$(Nobody gets back his own raincoat and umbrella) =

$1 - P_2$(there is at least one person who gets back both his raincoat and umbrel

$E_i$ : person $i$ gets back his own raincoat and umbrella

$$\rightarrow 1 - P_2 = 1 - P\left(\bigcup_{i=1}^{n} E_i\right) = 1 - \left(\sum_{1}^{n} P(E_i) - \sum_{\substack{i,j=1 \\ j \neq i \\ j > i}}^{n} P(E_i \cap E_j) + \cdots\right)$$

$$= 1 - \left(\binom{n}{1} \times \frac{1}{n^2} - \binom{n}{2} \times \frac{1}{n^2(n-1)^2} + \cdots\right)$$

$$= 1 - \left(\frac{n!}{1!(n-1)!} \times \frac{1}{n^2} - \frac{n!}{2!(n-2)!} \times \frac{1}{n^2(n-1)^2} + \cdots\right)$$

$$= 1 - \left(\frac{1}{n} - \frac{1}{2!n(n-1)} + \frac{1}{3!n(n-1)(n-2)} - \cdots\right)$$

$$= 1 - \frac{1}{n} + \frac{1}{2!n(n-1)} - \frac{1}{3!n(n-1)(n-2)} + \cdots + \frac{(-1)^n}{n!^2} = \sum_{i=0}^{n} \frac{(-1)^i (n-i)!}{i!\,n!}$$

b) $C_i$ = person $i$ gets back neither his raincoat nor his umbrella

$\rightarrow P$(everybody gets back ab least his raincoat or umbrella) $= 1 - P$(there is atleast a person who does not get back his umbrella and raincoat)

c) $(1-0.1)^{\frac{N}{2}} (1-0.2)^{\frac{N}{2}} = (0.9 \times 0.8)^{\frac{N}{2}} = 0.72^{\frac{N}{2}}$

## Problem 9

a.

λ) $V_1$ = P( the tried VPN is the better one)

$V_2$ = P( the " " " " worse one)

$$P(V_1 \mid \text{stablished}) = \frac{P(\text{stablished} \mid V_1) P(V_1)}{P(\text{stablished} \mid V_1) P(V_1) + P(\text{stablished} \mid V_2) P(V_2)}$$

$$= \frac{0.3 \times 0.5}{0.3 \times 0.5 + 0.1 \times 0.5} = \frac{3}{4} = \underline{0.75}$$

<u>9.b.</u>

## Code:

```r
connect_vpn <- function (p) {
  did_stablish <- as.logical(rbinom(1, size=1, prob=p))
  return(did_stablish)
}

select_vpn_randomly <- function() {
  if(as.logical(rbinom(1, size=1, prob=0.5)))
    return(1)
  else
    return(2)
}

experiment <- function() {
  num_try = 0
  while(probabilities[1] < 0.9) {
    if(probabilities[1] < 1e-4)
      return(0)
    i <- select_vpn_randomly()
    stablished_vpn <- connect_vpn(success_rates[i])
    if(stablished_vpn) {
      probabilities[i] <<- (success_rates[i] * probabilities[i]) /
        (success_rates[i] * probabilities[i] + success_rates[3-i] * probabilities[3-i])
    }
    else {
      probabilities[i] <<- ((1-success_rates[i]) * probabilities[i]) /
        ((1-success_rates[i]) * probabilities[i] + (1-success_rates[3-i]) * probabilities[3-i])
    }
    probabilities[3-i] <<- 1 - probabilities[i]

    num_try <- num_try + 1
  }
  return(num_try)
}

success_rates <- c(0.3, 0.1)

tries = c()
for(j in (1:100)) {
  print(j)
  probabilities <- c(0.5, 0.5)
  num_try = experiment()
  tries <- c(tries, num_try)
}
cat("mean_trials to be 90% certain = ", mean(tries))

```

## Result:

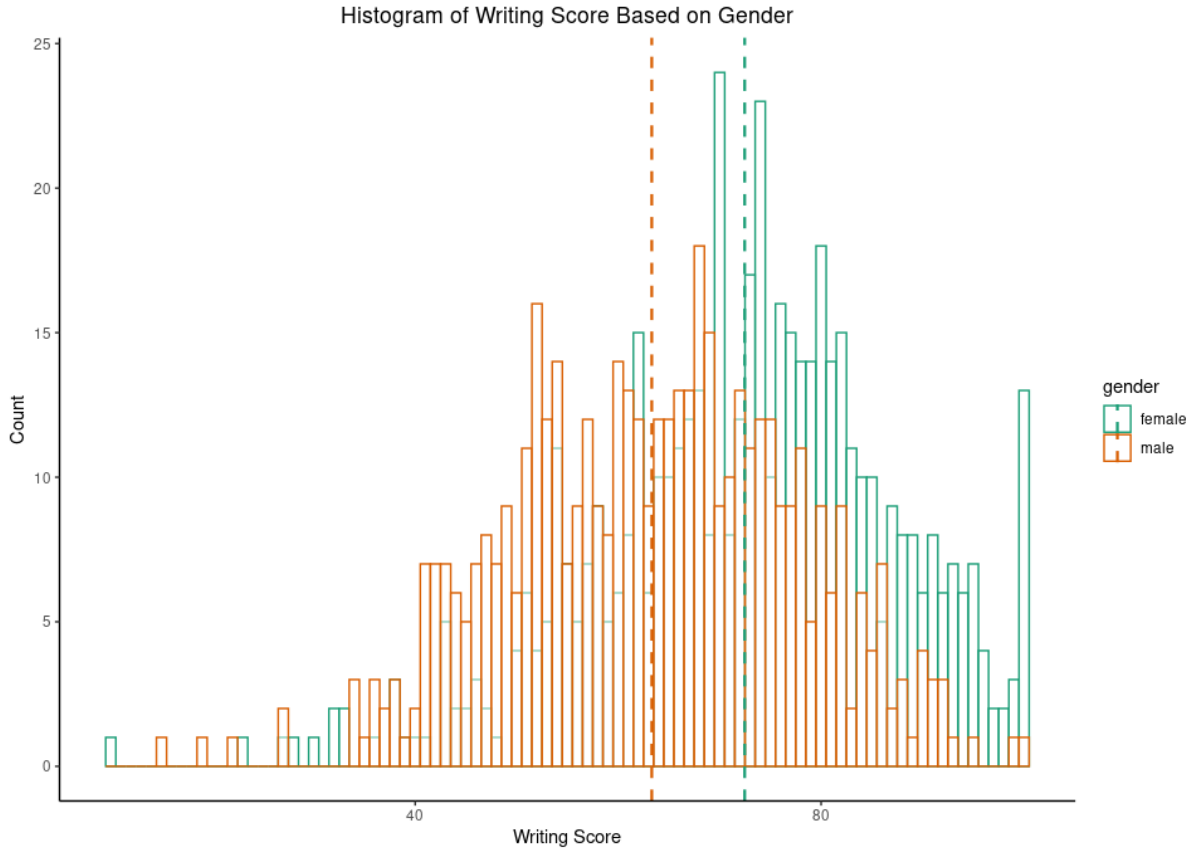
The result changes each time we run the experiments but it is about 50

# Problem 10

## a.

```r
library(ggplot2)

data <- read.csv('StudentsPerformance.csv')

# a

males = data[data["gender"]=='male', ]
females = data[data["gender"]=='female', ]

means <- c(mean(males$writing_score), mean(females$writing_score));

p_a <- ggplot(data,
              aes(x=writing_score,
                  color=gender)) +
  geom_histogram(binwidth=1,
                 fill='white',
                 alpha=0.5,
                 position="identity") +

  geom_vline(data=data.frame(gender=c('male', 'female') , means=means),
             aes(xintercept=means, color=gender),
             linetype="dashed",
             size=0.8) +

  scale_color_brewer(palette="Dark2") +
  labs(title="Histogram of Writing Score Based on Gender", x="Writing Score", y = "Count")+
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks=seq(0, 100, 40))

show(p_a)
```

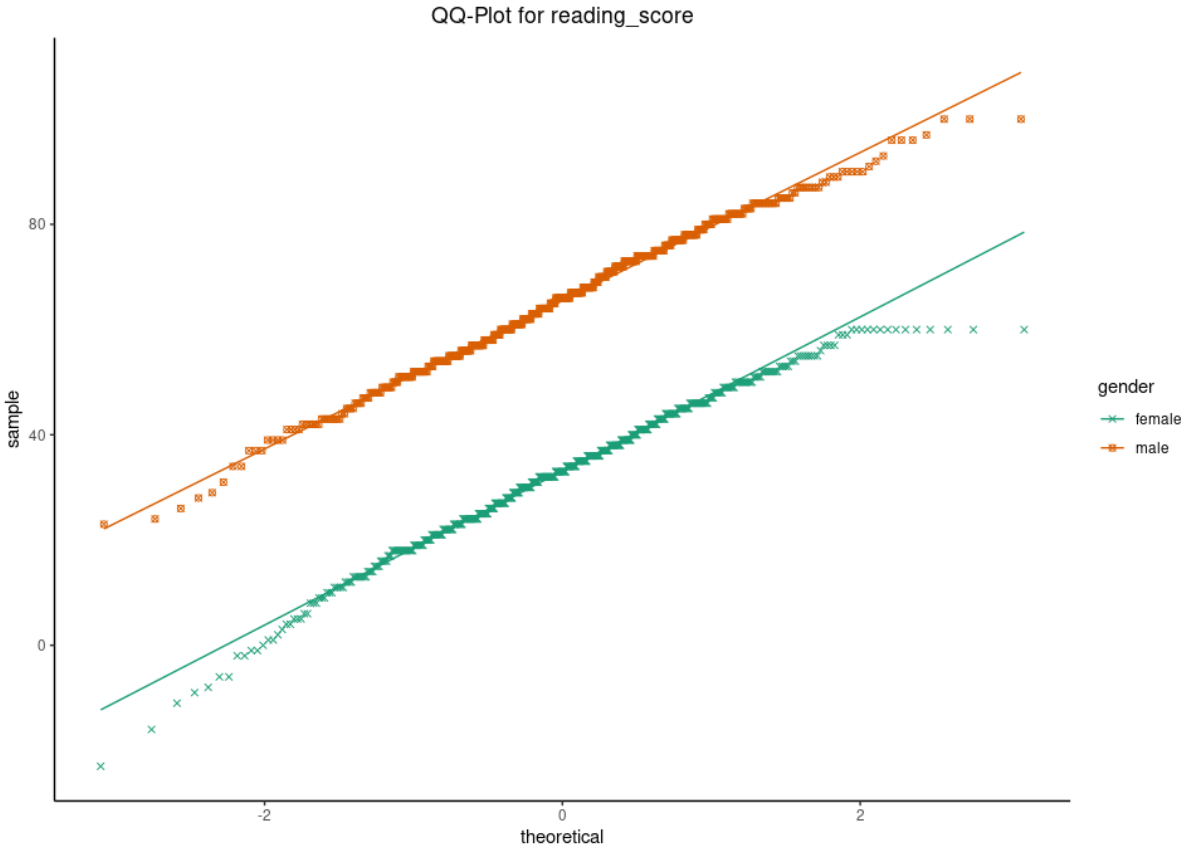
## a. result:

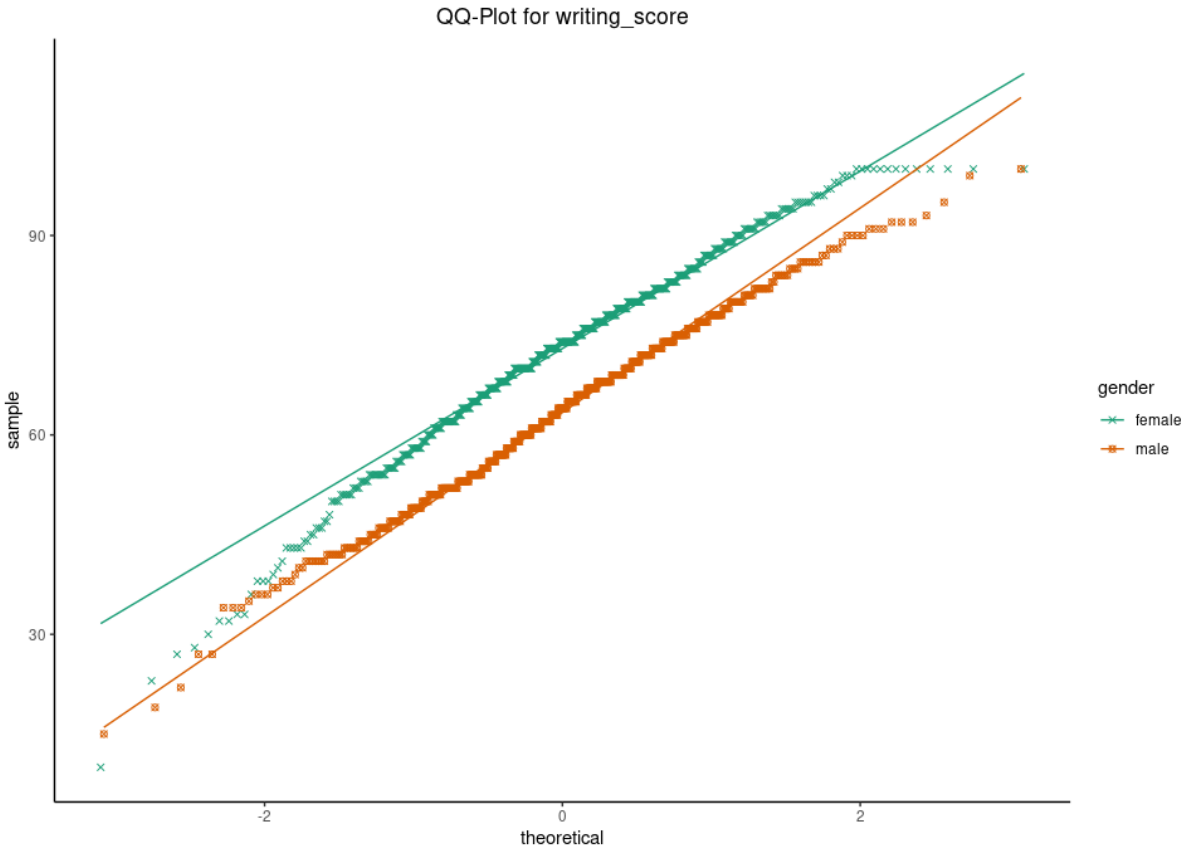

Histogram of Writing Score Based on Gender
Count
25
20
15
10
5
0
40
80
Writing Score
gender
female
male

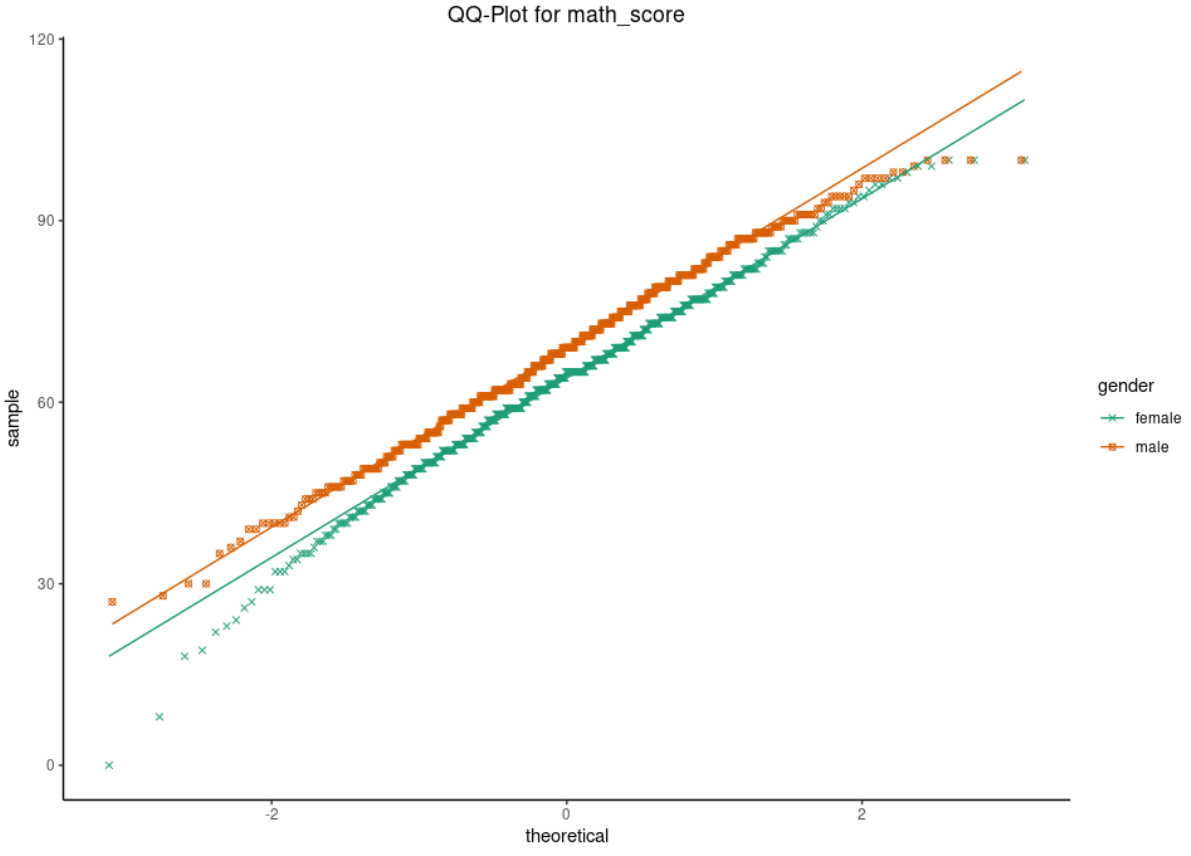
## b.

```r
# b
for (subj in c('reading_score', 'writing_score', 'math_score')) {
    p_b <- ggplot(data,
                aes_string(sample=subj, shape='gender',
                    colour='gender')) +
    stat_qq() +
    stat_qq_line() +
    scale_shape_manual(values=c(4,13)) +
    scale_color_brewer(palette="Dark2") +
    theme_classic() +
    labs(title=paste("QQ-Plot for", subj)) +
    theme(plot.title = element_text(hjust = 0.5))
    show(p_b)
}
```

# b.results:



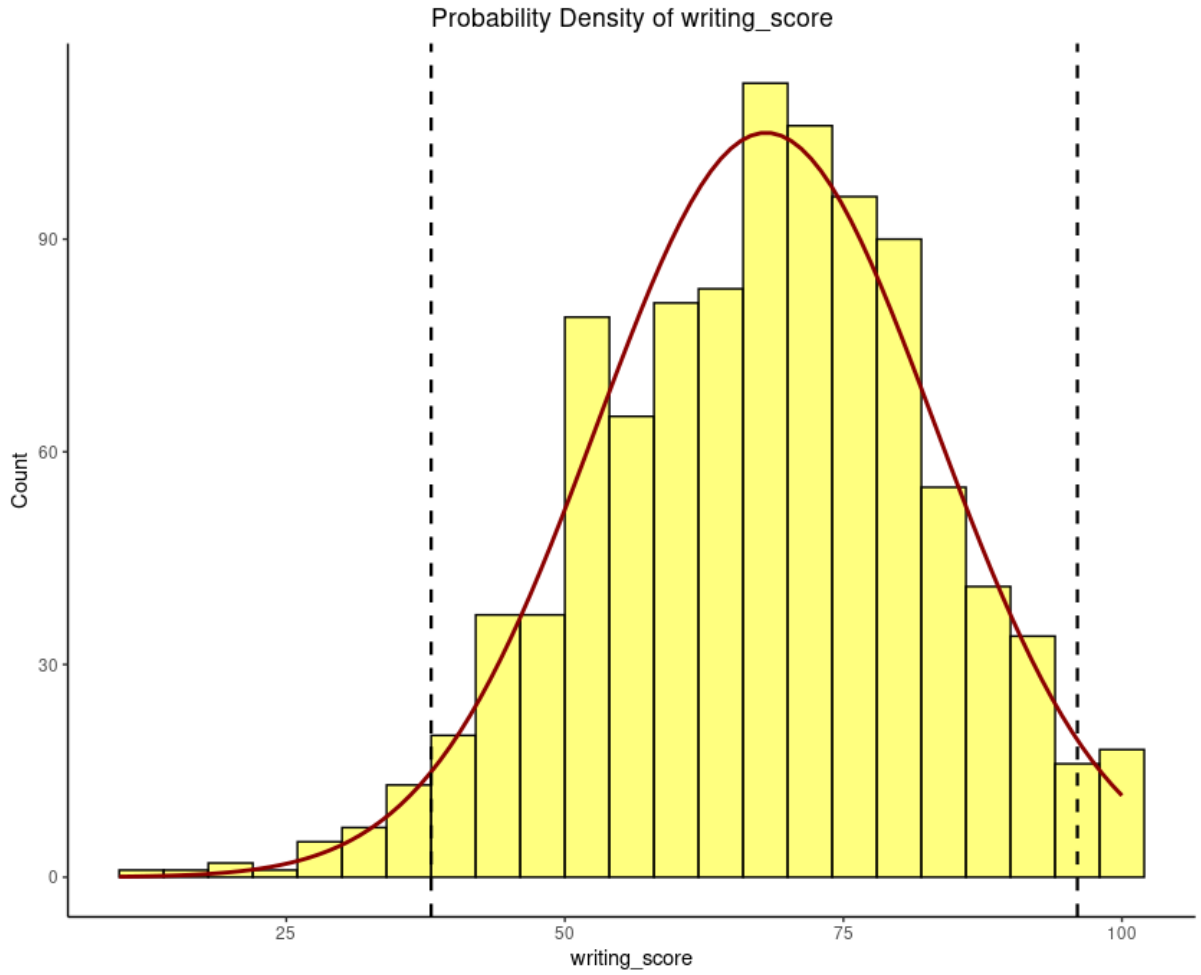

QQ-Plot for writing_score
sample
theoretical
90
60
30
-2
0
2
gender
female
male

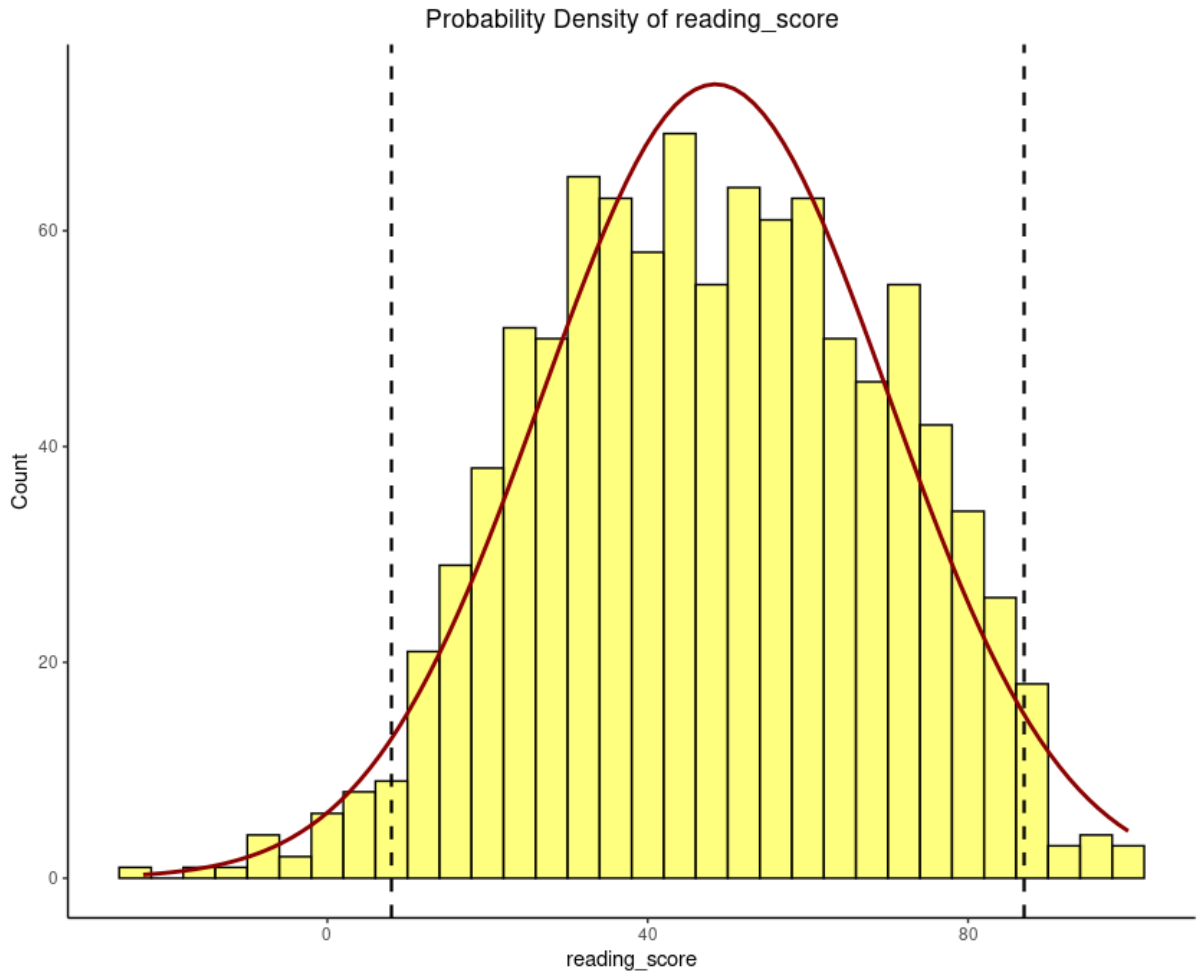

QQ-Plot for math_score
sample
theoretical
120
90
60
30
0
-2
0
2
gender
female
male

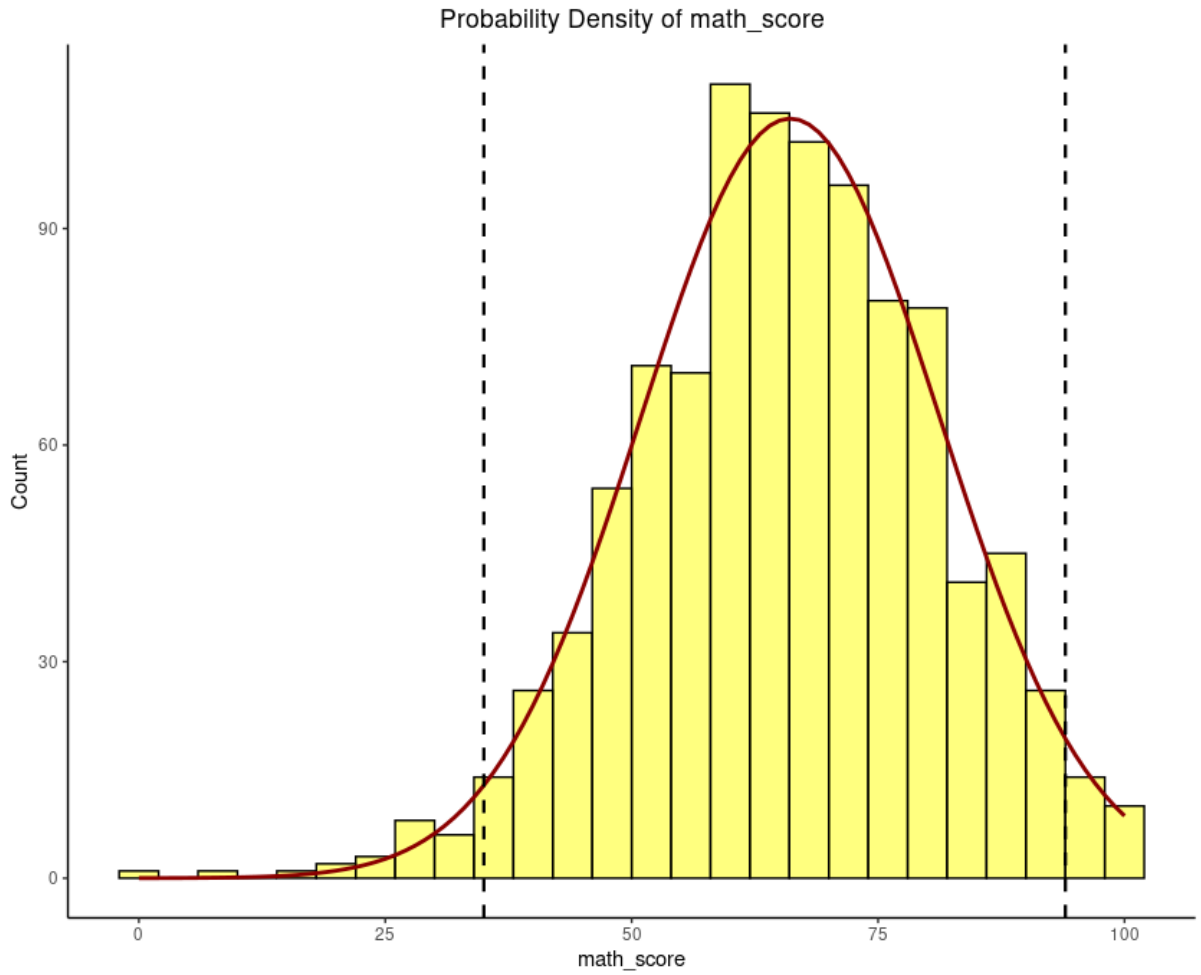
## c.

```r
# c

reading_score_mean <- mean(data$reading_score)
writing_score_mean <- mean(data$writing_score)
math_score_mean <- mean(data$math_score)

reading_score_sd <- sd(data$reading_score)
writing_score_sd <- sd(data$writing_score)
math_score_sd <- sd(data$math_score)

bin_width = 4;
num_samples = nrow(data);

scaled_dnorm = function(x, mean, sd, n) {
  bin_width * num_samples * dnorm(x, mean, sd)
}

for (subj in c('reading_score', 'writing_score', 'math_score')) {
  p_c <- ggplot(data,
                aes_string(x=subj)) +

  geom_histogram(binwidth=bin_width,
                 color='black',
                 fill='yellow',
                 position="identity",
                 alpha=0.5) +
  stat_function(fun=scaled_dnorm,
                args=list(data[[subj]],
                          mean=mean(data[[subj]]),
                          sd = sd(data[[subj]])),
                colour = "darkred", size = 1) +
  geom_vline(xintercept=unname(quantile(data[[subj]], c(0.025, 0.975))),
             linetype="dashed",
             size=0.8) +
  labs(title=paste("Probability Density of", subj), x=subj, y = "Count")+
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))
  show(p_c)
}
```

## c.Results:


Probability Density of writing_score
Count
90
60
30
0
25
50
75
100
writing_score

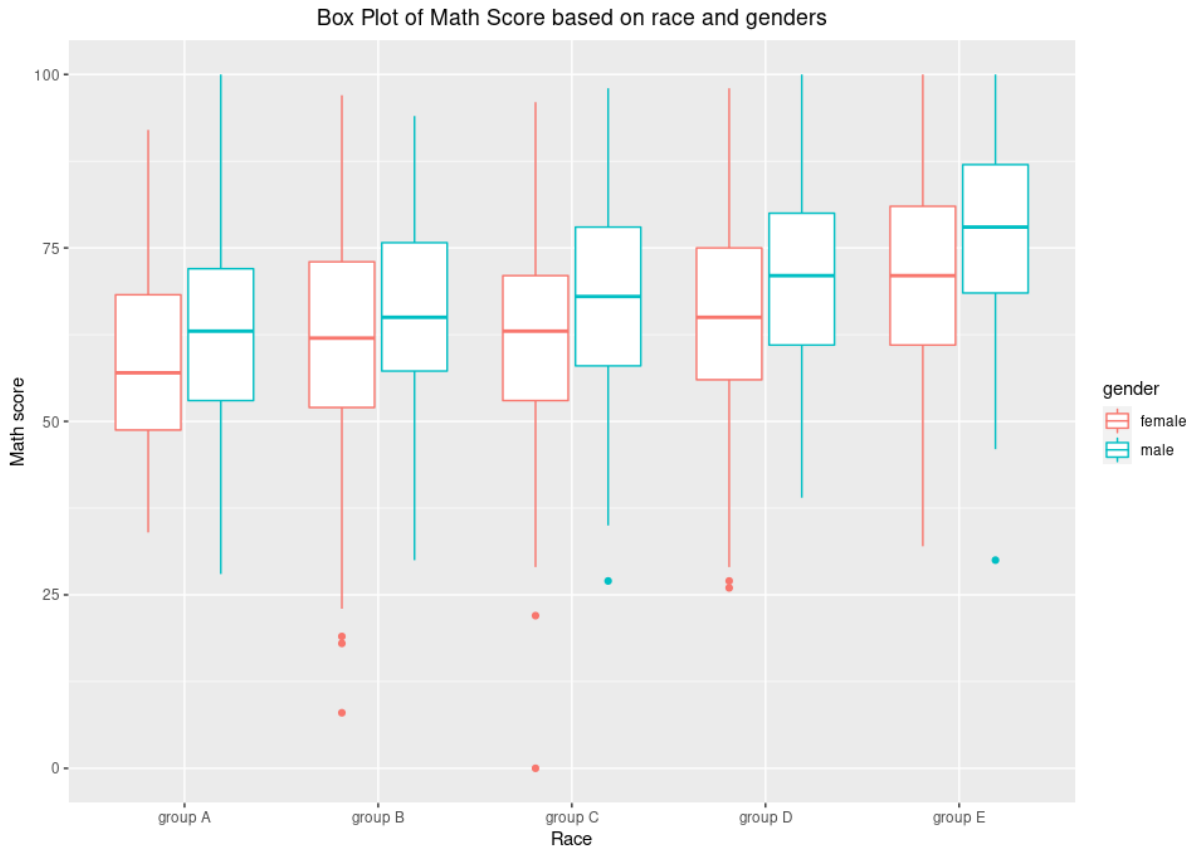

Probability Density of reading_score
Count
60
40
20
0
0
40
80
reading_score




## d.

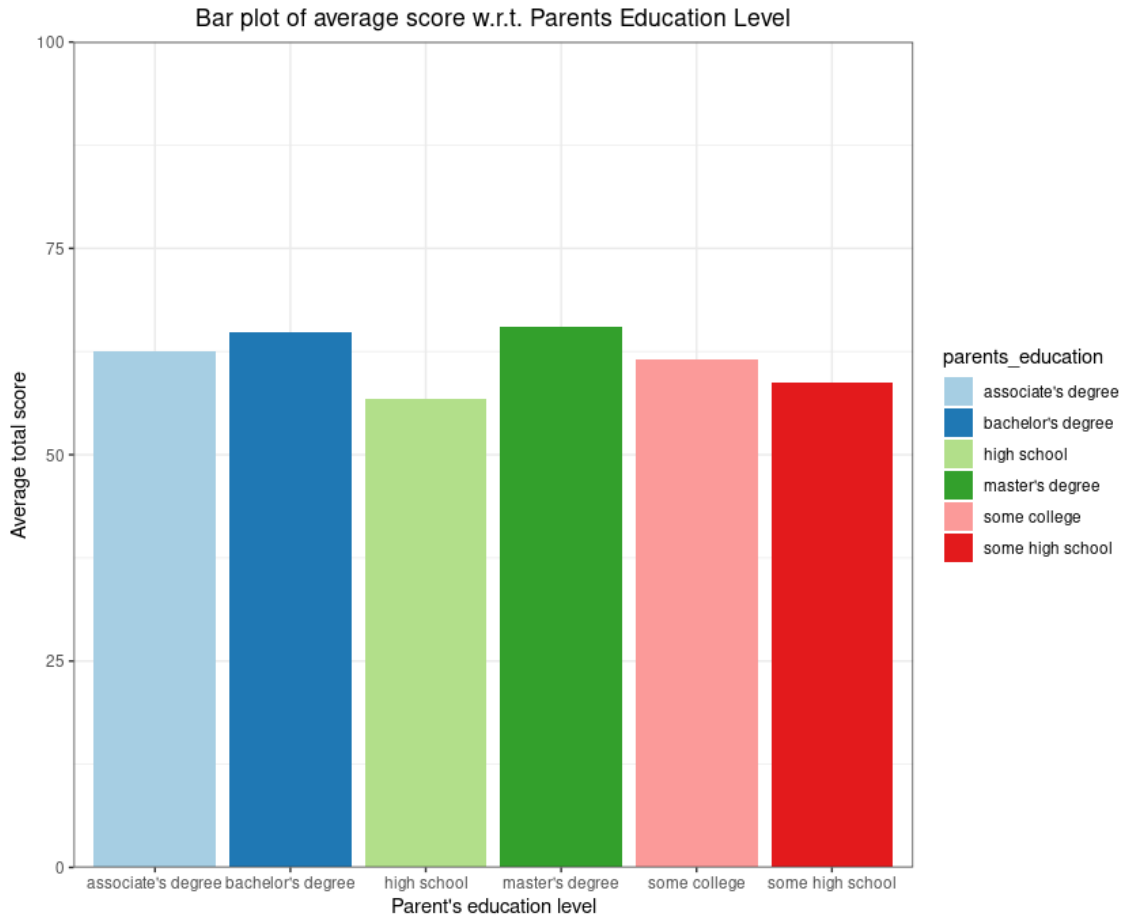
```
89  #d
90  p_d <- ggplot(data, aes(race, math_score)) +
91    geom_boxplot(aes(colour=gender)) +
92    labs(title="Box Plot of Math Score based on race and genders", x="Race", y = "Math score") +
93    theme(plot.title = element_text(hjust = 0.5))
94  show(p_d)
```

# d.Results:

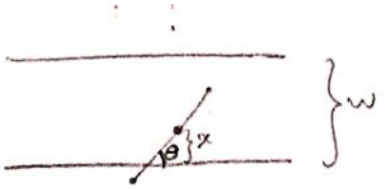

Box Plot of Math Score based on race and genders
Math score
100
75
50
25
0
group A
group B
group C
group D
group E
Race
gender
female
male

## e.

```
# e
data['avg_score'] <- round(rowMeans(data[, c('math_score', 'reading_score', 'writing_score')]), 2);

p_e <- ggplot(data, aes(x=parents_education, y=avg_score, fill=parents_education)) +
  geom_bar(stat="summary", fun="mean") +
  scale_fill_brewer(palette="Paired") +
  theme(axis.text.x=element_blank(),
      axis.ticks.x=element_blank()) +
  scale_y_continuous(expand = c(0, 0), limits = c(0, 100)) +
theme(axis.line.x = element_line(size=1.5, colour="grey")) +
  labs(title="Bar plot of average score w.r.t. Parents Education Level", x="Parent's education level",
       y = "Average total score") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
show(p_e)
```



# Problem 11




$x$: distance between center of knife-trace and the closest line

$\theta$: acute angle between knife trace and closest line.

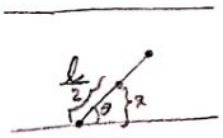
always:

$0 \le x \le \frac{w}{2} \longrightarrow f(x) = \begin{cases} \frac{1}{\frac{w}{2}} = \frac{2}{w} & 0 \le x \le \frac{w}{2} \\ 0 & o.w \end{cases}$

$0 \le \theta \le \frac{\pi}{2} \longrightarrow f(\theta) = \begin{cases} \frac{1}{\frac{\pi}{2}} = \frac{2}{\pi} & 0 \le \theta \le \frac{\pi}{2} \\ o.w \end{cases}$

$\rightarrow$ we know that $x$ and $\theta$ are independent

$\rightarrow f(x,\theta) = f(x)\,f(\theta) = \begin{cases} \frac{4}{\pi w} & 0 \le x \le \frac{w}{2},\ 0 \le \theta \le \frac{\pi}{2} \\ 0 & o.w \end{cases}$




$\longrightarrow$ doesn't cross a line if $\frac{l}{2}\sin\theta \le x \le \frac{w}{2}$

$$\rightarrow P(\text{fall into one strip}) = \int_{\theta=0}^{\frac{\pi}{2}} \int_{\frac{l}{2}\sin\theta}^{\frac{w}{2}} \frac{4}{\pi w}\, dx\, d\theta = \int_0^{\frac{\pi}{2}} \frac{4}{\pi w}\left(\frac{w}{2} - \frac{l}{2}\sin\theta\right) d\theta$$

$$= \frac{4}{\pi w}\left(\frac{\pi}{2} \times \frac{w}{2} + \frac{l}{2}\cos\theta\Big|_0^{\frac{\pi}{2}}\right)$$

$$= \frac{4}{\pi w}\left(\frac{w\pi}{4} - \frac{l}{2}\right) = 1 - \frac{2l}{\pi w}$$

$$\longrightarrow P(\text{3 successive round}) = \left(1 - \frac{2l}{\pi w}\right)^3$$

because rounds are independent