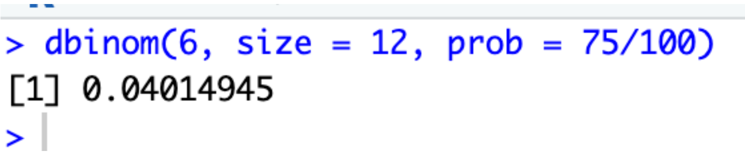
Homework 3 **Name:** Ivan Wang

(Part 6; 30 pts)

1. Suppose that when a person buys their first vehicle from a given dealership, the probability that they will buy their next vehicle from that same dealership is 75%. In March 2019, 12 customers are searching for their second vehicle, having bought their first vehicle (years earlier) from JC Motors. Assume all of them will make a purchase before the end of the month. Let Y be a random variable that counts the number of vehicles purchased from JC Motors by this set of customers during March. Obtain the following binomial probabilities. (8 pts)

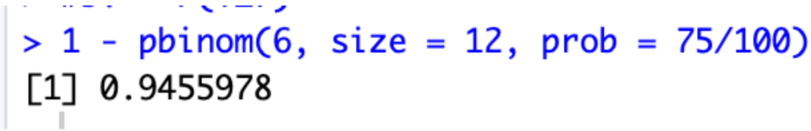
a) $P(Y = 6)$



```
> dbinom(6, size = 12, prob = 75/100)
[1] 0.04014945
> 
```

Use dbionom to find probability of exactly 6 customers choose JC Motors

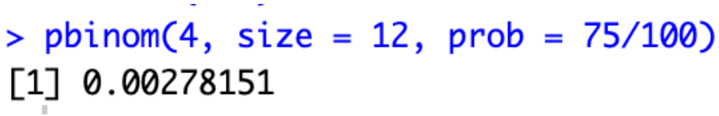
b) $P(Y \geq 7)$



```
> 1 - pbinom(6, size = 12, prob = 75/100)
[1] 0.9455978
```

Use pbinom with given arguments to find probability of getting 6 or fewer customers choosing JC Motors. Subtract this by 1 to get probability of getting 7 or more customers choosing JC Motors.

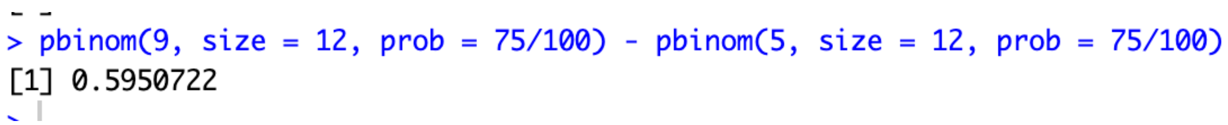
c) $P(Y < 5)$



```
> pbinom(4, size = 12, prob = 75/100)
[1] 0.00278151
```

Use pbinom with given arguments to find probability of getting less then 5 customers choosing JC motors.

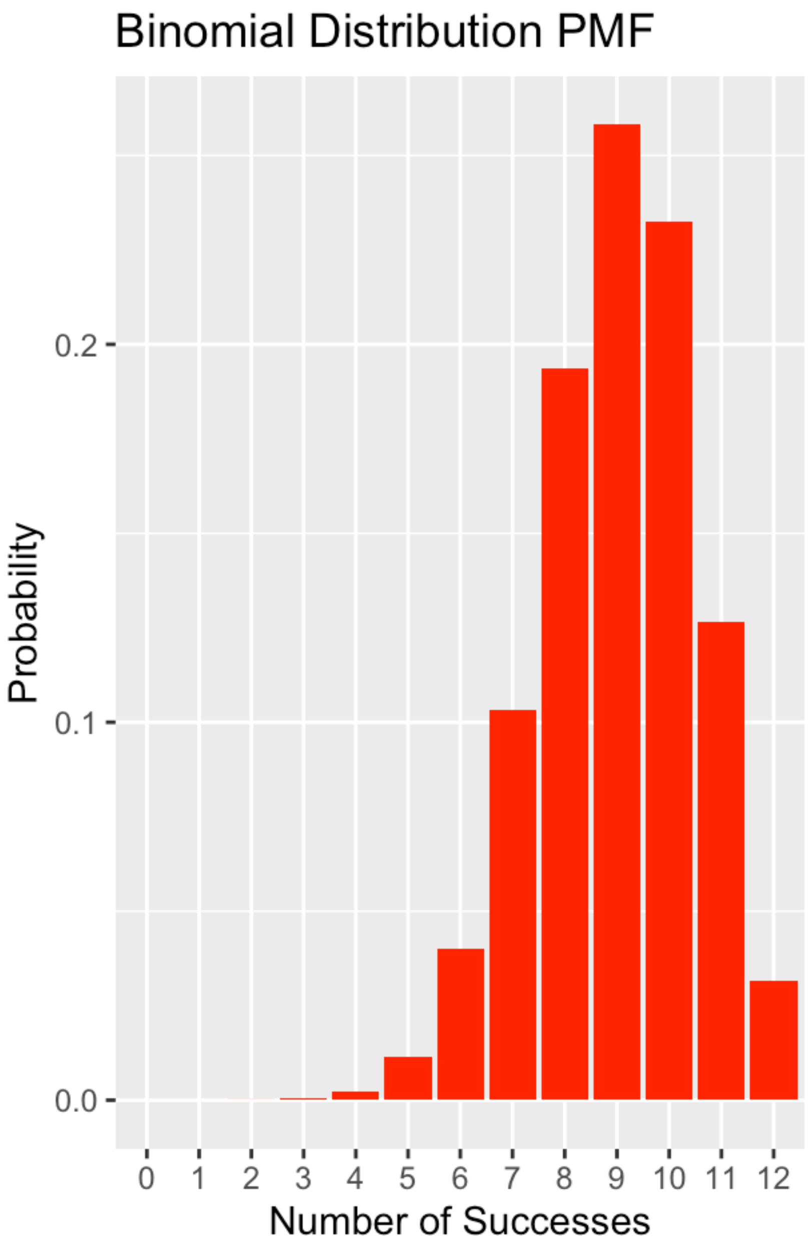
d) $P(6 \leq Y \leq 9)$

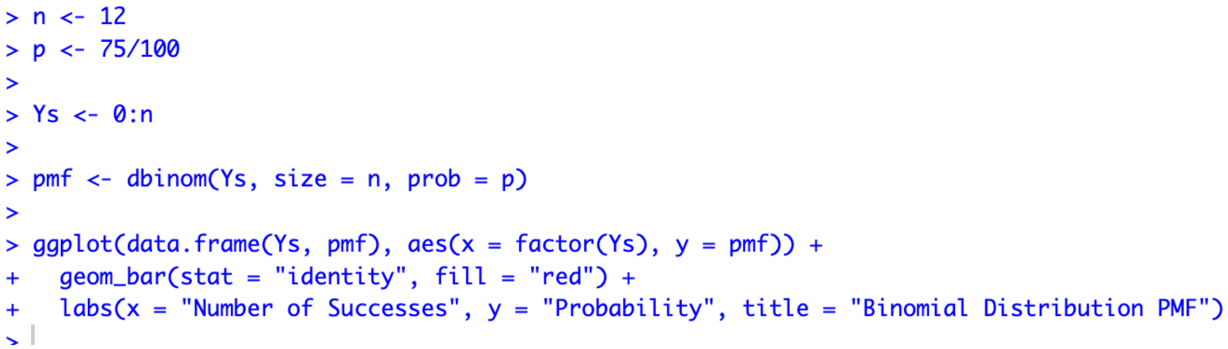


```
> pbinom(9, size = 12, prob = 75/100) - pbinom(5, size = 12, prob = 75/100)
[1] 0.5950722
```

Use pbinom with given arguments to find probability of getting exactly or more than 6 customers and less then 9 customers.

2. Plot the theoretical probability mass function of Y, the random variable from the previous question. Label both axes. (4 pts)

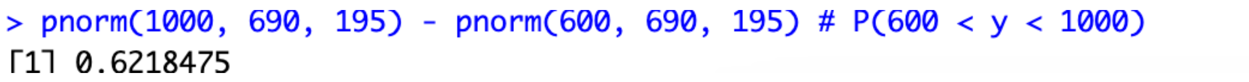




```
> n <- 12
> p <- 75/100
>
> Ys <- 0:n
>
> pmf <- dbinom(Ys, size = n, prob = p)
>
> ggplot(data.frame(Ys, pmf), aes(x = factor(Ys), y = pmf)) +
+   geom_bar(stat = "identity", fill = "red") +
+   labs(x = "Number of Successes", y = "Probability", title = "Binomial Distribution PMF")
>
```

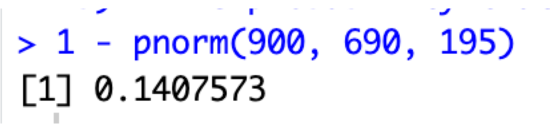
3. Suppose that levels of a blood biomarker follow a normal distribution with mean 690 and standard deviation 195. Obtain each of the following: (12 pts)

a) The probability that an individual biomarker reading is between 600 and 1000.



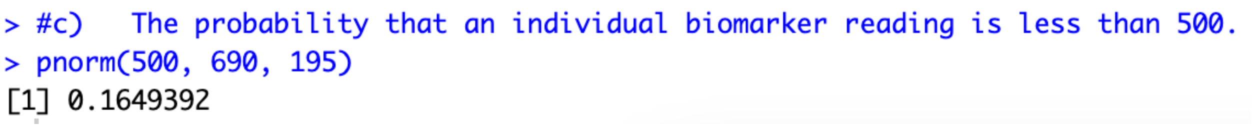
```
> pnorm(1000, 690, 195) - pnorm(600, 690, 195) # P(600 < y < 1000)
[1] 0.6218475
```

b) The probability that an individual biomarker reading is higher than 900.



```
> 1 - pnorm(900, 690, 195)
[1] 0.1407573
```

c) The probability that an individual biomarker reading is less than 500.

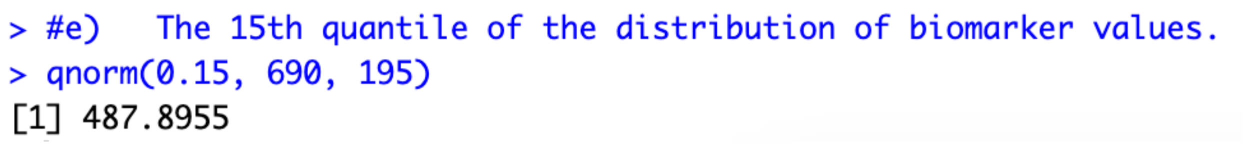


```
> #c)   The probability that an individual biomarker reading is less than 500.
> pnorm(500, 690, 195)
[1] 0.1649392
```

d) The probability that an individual biomarker reading is equal to 690.

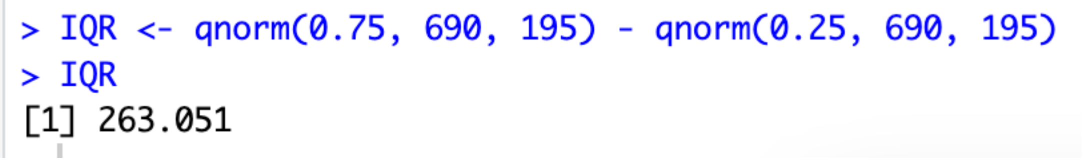
P(y = 0) = 0 because the probability of a single value in a continuous distribution is very very very small.

e) The 15th quantile of the distribution of biomarker values.



```
> #e)   The 15th quantile of the distribution of biomarker values.
> qnorm(0.15, 690, 195)
[1] 487.8955
```

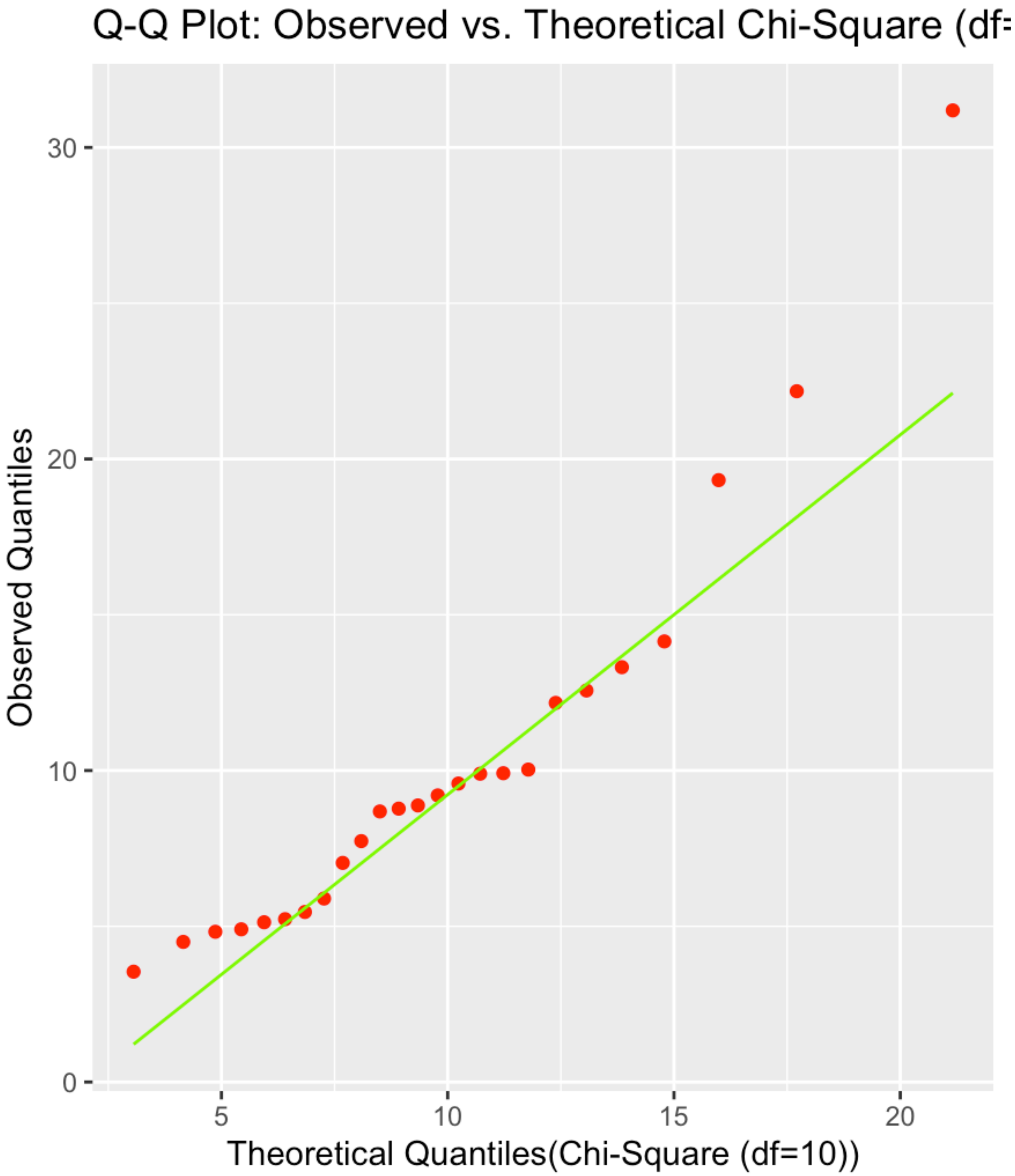
f) The IQR.



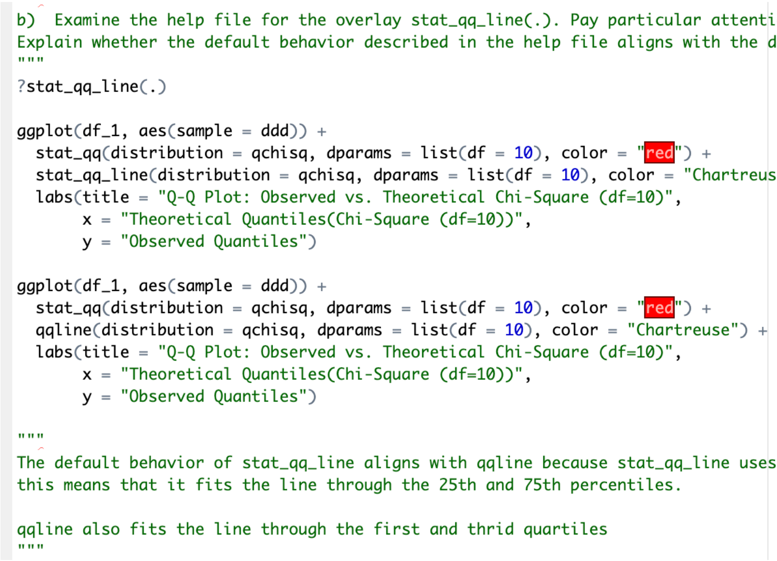
```
> IQR <- qnorm(0.75, 690, 195) - qnorm(0.25, 690, 195)
> IQR
[1] 263.051
```

4. Download the script hw3_4.R from UBLearns. The script creates a vector of length 25 (called "ddd"), and stores these values in a data frame called "df_1." (6 pts)

a) Use ggplot(.) to produce a Q-Q plot that compares the observed quantiles of the sample ("ddd") to theoretical quantiles from a χ^2_{10} distribution. Include a reference line, also based on χ^2_{10}. (1)



b) Examine the help file for the overlay stat_qq_line(.). Pay particular attention to the argument “line.p.” Explain whether the default behavior described in the help file aligns with the description of qqline(.) in the Part 5 notes. (1)



```
b)  Examine the help file for the overlay stat_qq_line(.). Pay particular attention to the argument “line.p.”
Explain whether the default behavior described in the help file aligns with the description of qqline(.) in the Part 5 notes. (1)
"""
?stat_qq_line(.)

ggplot(df_1, aes(sample = ddd)) +
  stat_qq(distribution = qchisq, dparams = list(df = 10), color = "red") +
  stat_qq_line(distribution = qchisq, dparams = list(df = 10), color = "Chartreuse") +
  labs(title = "Q-Q Plot: Observed vs. Theoretical Chi-Square (df=10)",
       x = "Theoretical Quantiles(Chi-Square (df=10))",
       y = "Observed Quantiles")

ggplot(df_1, aes(sample = ddd)) +
  stat_qq(distribution = qchisq, dparams = list(df = 10), color = "red") +
  qqline(distribution = qchisq, dparams = list(df = 10), color = "Chartreuse") +
  labs(title = "Q-Q Plot: Observed vs. Theoretical Chi-Square (df=10)",
       x = "Theoretical Quantiles(Chi-Square (df=10))",
       y = "Observed Quantiles")

"""
The default behavior of stat_qq_line aligns with qqline because stat_qq_line uses line.p = c(0.25, 0.75),
this means that it fits the line through the 25th and 75th percentiles.

qqline also fits the line through the first and thrid quartiles
"""
```

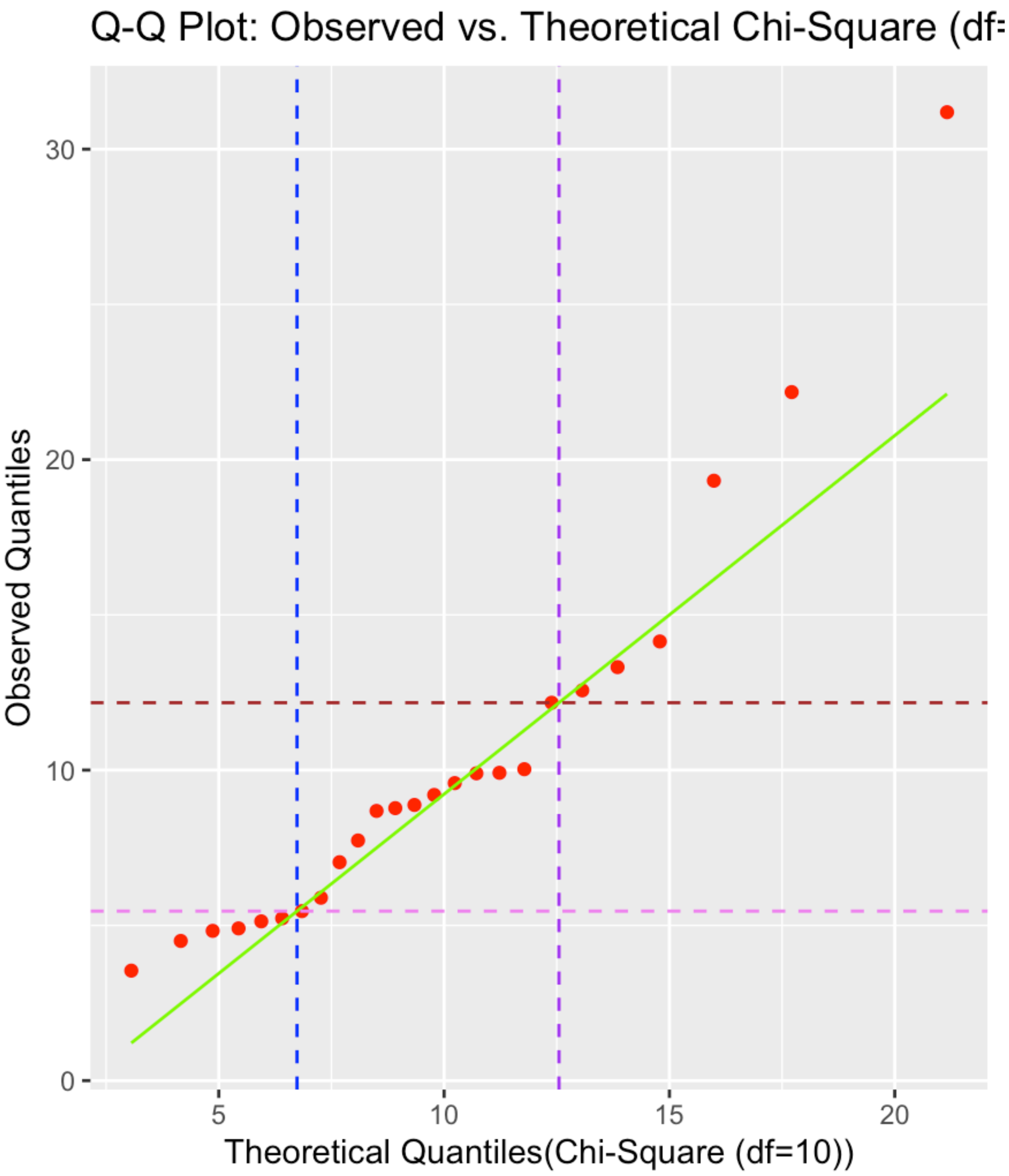
The default behavior of stat_qq_line aligns with qqline because stat_qq_line uses line.p = c(0.25, 0.75).

This means that it fits the line through the 25th and 75th percentiles.

qqline also fits the line through the first and third quartiles.

c) Add four reference lines to your plot from part (a). The two vertical lines (give them a color) should be located at the **true** 25th and 75th percentiles of χ^2_{10}. The two horizontal lines (give them a different color) should be located at the **sample** 25th and 75th percentiles of the data. Explain how the intersections of your reference lines produce the reference line from part (a). (4)

The intersection of the vertical and horizontal lines shows how the sample quantiles align with the theoretical quantiles.


Q-Q Plot: Observed vs. Theoretical Chi-Square (df=
Observed Quantiles
30
20
10
0
5
10
15
20
Theoretical Quantiles(Chi-Square (df=10))

Question 1-3 R code:

```
""

1. Suppose that when a person buys their first vehicle from a given dealership, the probability that they

will buy their next vehicle from that same dealership is 75%. In March 2019, 12 customers are searching

for their second vehicle, having bought their first vehicle (years earlier) from JC Motors. Assume all of

them will make a purchase before the end of the month. Let Y be a random variable that counts the number

of vehicles purchased from JC Motors by this set of customers during March. Obtain the following binomial

probabilities. (8 pts)

""


#a.	P(Y=6)

dbinom(6, size = 12, prob = 75/100)



#b.	P(Y≥7)

1 - pbinom(6, size = 12, prob = 75/100)


#c.	P(Y<5)

pbinom(4, size = 12, prob = 75/100)


#d.	P(6≤Y≤9)


pbinom(9, size = 12, prob = 75/100) - pbinom(5, size = 12, prob = 75/100)
```

"""

2. Plot the theoretical probability mass function of Y, the random variable from the previous question.

Label both axes. (4 pts)

"""

```
install.packages("ggplot2")

library(ggplot2)


n <- 12

p <- 75/100


Ys <- 0:n


pmf <- dbinom(Ys, size = n, prob = p)


ggplot(data.frame(Ys, pmf), aes(x = factor(Ys), y = pmf)) +
  geom_bar(stat = "identity", fill = "red") +
  labs(x = "Number of Successes", y = "Probability", title = "Binomial Distribution PMF")
```

"""

3. Suppose that levels of a blood biomarker follow a normal distribution with mean 690 and standard deviation

195. Obtain each of the following: (12 pts)

"""

```
#a)	The probability that an individual biomarker reading is between 600 and 1000.
```

```
pnorm(1000, 690, 195) - pnorm(600, 690, 195) # P(600 < y < 1000)


#b)	The probability that an individual biomarker reading is higher than 900.
1 - pnorm(900, 690, 195)


#c)	The probability that an individual biomarker reading is less than 500.
pnorm(500, 690, 195)


#d)	The probability that an individual biomarker reading is equal to 690.


#P(y = 0) = 0 because the probability of a single value in a continous distribution is very very
very small.


#e)	The 15th quantile of the distribution of biomarker values.
qnorm(0.15, 690, 195)



#f)	The IQR.


#IQR = Q3 - Q1
IQR <- qnorm(0.75, 690, 195) - qnorm(0.25, 690, 195)
IQR
```

Question 4 R code:

```
install.packages("ggplot2")

library(ggplot2)


""

4.  Download the script hw3_4.R from UBLearns. The script creates a vector of length 25 (called
“ddd”),

and stores these values in a data frame called “df_1.” (6 pts)

""


""

a.        Use ggplot(.) to produce a Q-Q plot that compares the observed quantiles of the sample
(“ddd”) to theoretical quantiles

          from a χ_10^2 distribution. Include a reference line, also based on χ_10^2. (1)

""


ddd <- c(3.542, 8.776, 5.89, 7.034, 9.897, 13.317, 7.734,

       8.879, 8.686, 12.171, 9.914, 5.129, 5.228, 5.459,

       4.499, 9.201, 4.903, 22.175, 12.569, 31.19, 19.321,

       4.824, 14.144, 9.582, 10.03)

df_1 <- data.frame(ddd)


df_1



ggplot(df_1, aes(sample = ddd)) +

  stat_qq(distribution = qchisq, dparams = list(df = 10), color = "red") +

  stat_qq_line(distribution = qchisq, dparams = list(df = 10), color = "Chartreuse") +

  labs(title = "Q-Q Plot: Observed vs. Theoretical Chi-Square (df=10)",
     x = "Theoretical Quantiles(Chi-Square (df=10))",
     y = "Observed Quantiles")
```

"""

b) Examine the help file for the overlay stat_qq_line(.). Pay particular attention to the argument “line.p.”

Explain whether the default behavior described in the help file aligns with the description of qqline(.) in the Part 5 notes. (1)

"""

```
?stat_qq_line(.)


ggplot(df_1, aes(sample = ddd)) +
  stat_qq(distribution = qchisq, dparams = list(df = 10), color = "red") +
  stat_qq_line(distribution = qchisq, dparams = list(df = 10), color = "Chartreuse") +
  labs(title = "Q-Q Plot: Observed vs. Theoretical Chi-Square (df=10)",
     x = "Theoretical Quantiles(Chi-Square (df=10))",
     y = "Observed Quantiles")


ggplot(df_1, aes(sample = ddd)) +
  stat_qq(distribution = qchisq, dparams = list(df = 10), color = "red") +
  qqline(distribution = qchisq, dparams = list(df = 10), color = "Chartreuse") +
  labs(title = "Q-Q Plot: Observed vs. Theoretical Chi-Square (df=10)",
     x = "Theoretical Quantiles(Chi-Square (df=10))",
     y = "Observed Quantiles")
```

"""

The default behavior of stat_qq_line aligns with qqline because stat_qq_line uses line.p = c(0.25, 0.75),

this means that it fits the line through the 25th and 75th percentiles.

qqline also fits the line through the first and thrid quartiles.

""

""

c) Add four reference lines to your plot from part (a). The two vertical lines (give them a color) should be located at

the true 25th and 75th percentiles of χ_10^2. The two horizontal lines (give them a different color) should be located at

the sample 25th and 75th percentiles of the data. Explain how the intersections of your reference lines produce the reference

line from part (a). (4)

""

```
q25 <- qchisq(0.25, df = 10)

q75 <- qchisq(0.75, df = 10)


h25 <- quantile(df_1$ddd, 0.25)

h75 <- quantile(df_1$ddd, 0.75)



ggplot(df_1, aes(sample = ddd)) +

  stat_qq(distribution = qchisq, dparams = list(df = 10), color = "red") +

  stat_qq_line(distribution = qchisq, dparams = list(df = 10), color = "Chartreuse") +

  geom_vline(xintercept = q25, color = "blue", linetype = "dashed") +

  geom_vline(xintercept = q75, color = "purple", linetype = "dashed") +

  geom_hline(yintercept = h25, color = "violet", linetype = "dashed") +
  geom_hline(yintercept = h75, color = "brown", linetype = "dashed") +
  labs(title = "Q-Q Plot: Observed vs. Theoretical Chi-Square (df=10)",
       x = "Theoretical Quantiles(Chi-Square (df=10))",
       y = "Observed Quantiles")
```

#The intersection of the vertical and horizontal lines shows how the sample quantiles align with the theoretical quantiles.