Prashant Sharma

Grading: 20 marks- Complete submission of worksheet13
40 marks- Problem 3 and 40 marks- Problem 5

Problem:1

Somadev finds that his weight in kgs during each month of year to be -
75 76 73 75 74 73 73 76 73 79 77 75

Problem:1(a)

We have to write a function called zcinf that takes in the weights above in a vector x, assumes a known standard deviation of 1.5 and produces default 95% confidence interval.
For which R-code is:

```
> zcinf = function(x, alpha = 0.95, sigma = 1.5){
+ z = qnorm( (1-alpha)/2, lower.tail=FALSE)
+ lower_lim = mean(x) - z*sqrt(1/length(x))*sigma
+ upper_lim = mean(x) + z*sqrt(1/length(x))*sigma
+ cat(lower_lim,',',upper_lim)
+ }
> x= c(75,76,73,75,74,73,73,76,73,79,77,75)
> zcinf(x)
```

```
74.06798 , 75.76536
```

Therefore the default 95% confidence interval for mean weight is $(74.06798, 75.76536)$,when standard deviation is known.

Problem:1(b)

Now we have to write a function called tcinf that takes in the weights above as a vector x, assumes that variance is unknown and produces default 95% confidence interval.
For which R code is:

```
> tcinf = function(x, alpha = 0.95){
+ t = qt(p = (1-alpha)/2,df = length(x)-1 , lower.tail=FALSE)
+ lower_lim = mean(x) - t*sqrt(1/length(x))*sd(x)
+ upper_lim = mean(x) + t*sqrt(1/length(x))*sd(x)
+ cat(lower_lim,',',upper_lim)
+ }
> tcinf(x)
```

```
73.72158 , 76.11175
```

Therefore the default 95% confidence interval for mean weight is $(73.72158, 76.11175)$, when population variance is unknown.

Problem:1(c)

Now, we will use the command t.test() on the vector x.

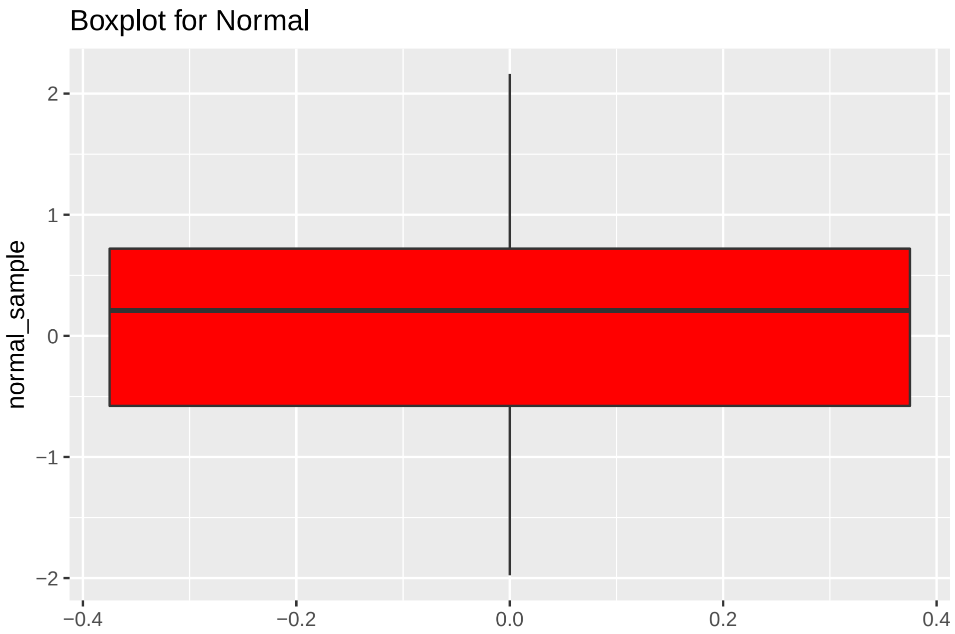
```
> t.test(x)


        One Sample t-test

data:  x
t = 137.97, df = 11, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 73.72158 76.11175
sample estimates:
mean of x
 74.91667
```

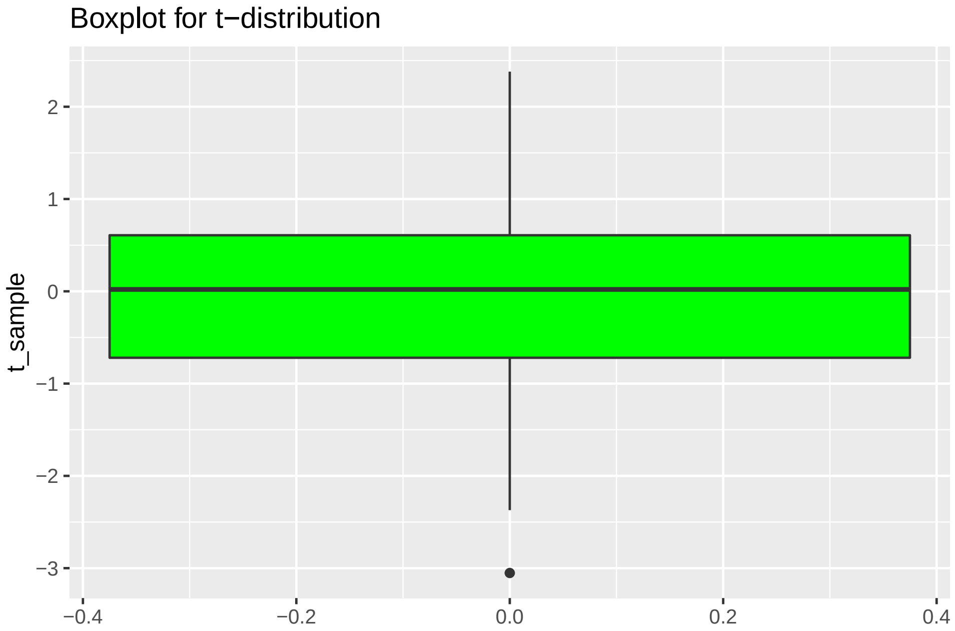
Note that since we are asked to infer output of the command t.test(x) doing that by default sets true mean in null hypothesis to be 0 and performs a one sample t-test to determine whether underlying populations mean is 0 or not.
Inferences are as follows:
The Null Hypothesis (H_0): true mean of x i.e., $\mu = 0$ $vs.$ the alternative Hypothesis (H_1): $\mu \neq 0$
t=137.97 is the observed test statistic i.e. observed $\frac{\sqrt{n}(\bar{x}-\mu)}{s}$; where n is the sample size, here it's length of the vector x, $\bar{X}$ is the sample mean, s is the sample standard deviation, under the null hypothesis, i.e.$\mu = 0$.
Here n=12 so df=12-1=11 and as we know that, p-value is the probability of obtaining the test result as extreme the result is actually observed under the assumption null hypothesis is true. Also, from output we can see that for this test p-value$< 2.2e{-}16 < 0.05$. Therefore, we can reject the default null hypothesis that is the true mean is equal to 0 at 5% level of significance. It also gives the 95% confidence interval for the population or true mean which is $(73.72158, 76.11175)$ and finally, it gives the sample estimates of true mean i.e. sample mean which is $= 74.91667$.

Problem:2(a)

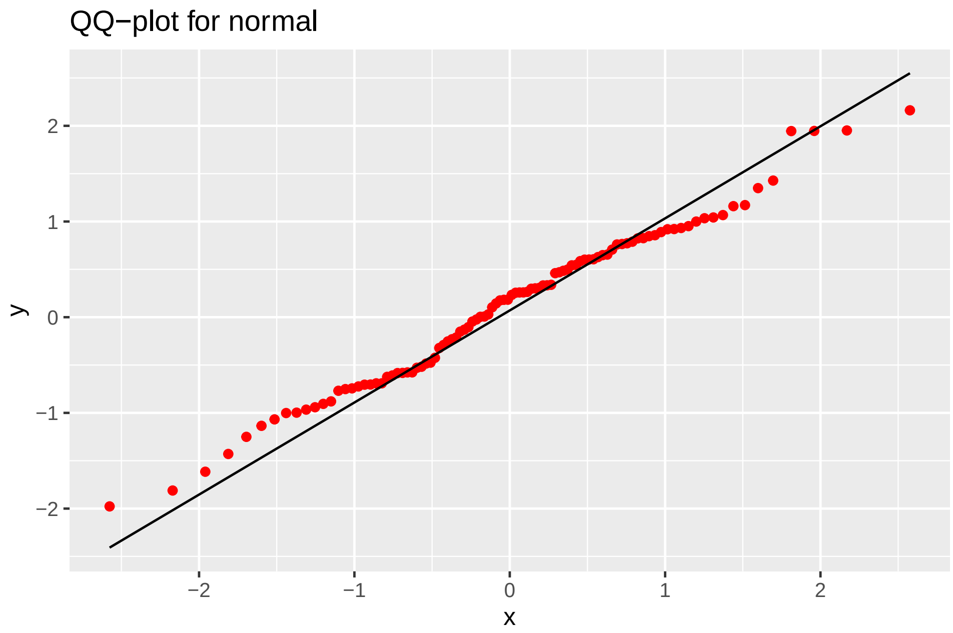
```
> normal_sample <- rnorm(100, mean = 0, sd = 1)
> t_sample <- rt(100, 25)
> library(ggplot2)
> df_normal=data.frame(normal_sample)
> df_t=data.frame(t_sample)
> #boxplot for normal distribution
> box_normal=ggplot(df_normal,aes(y=normal_sample))+geom_boxplot(fill="red")+
+ labs(title="Boxplot for Normal")
```

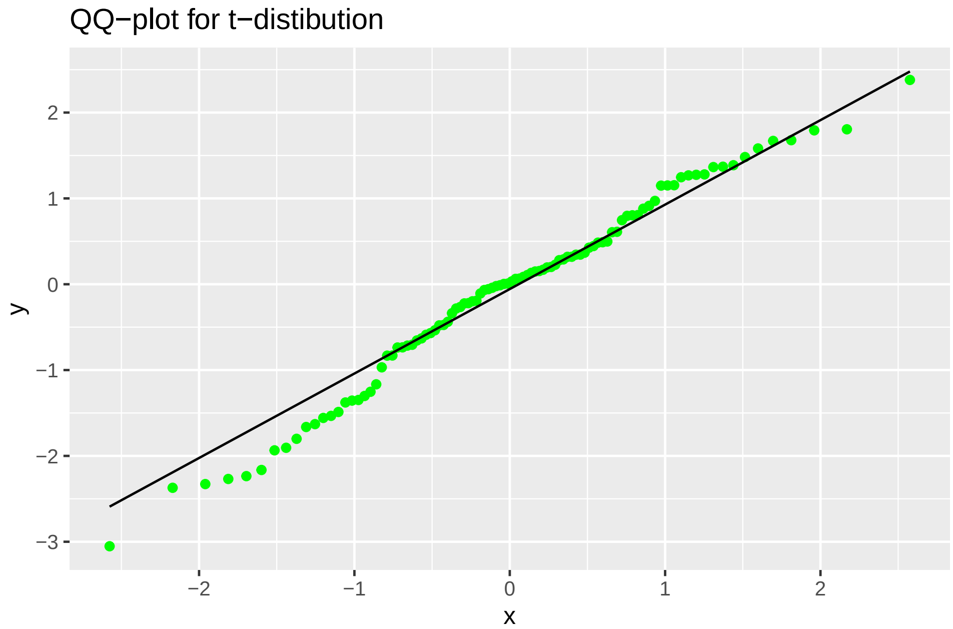
```
> #boxplot for t-distribution
> box_t=ggplot(df_t,aes(y=t_sample))+geom_boxplot(fill="green")+
+ labs(title="Boxplot for t-distribution")
```





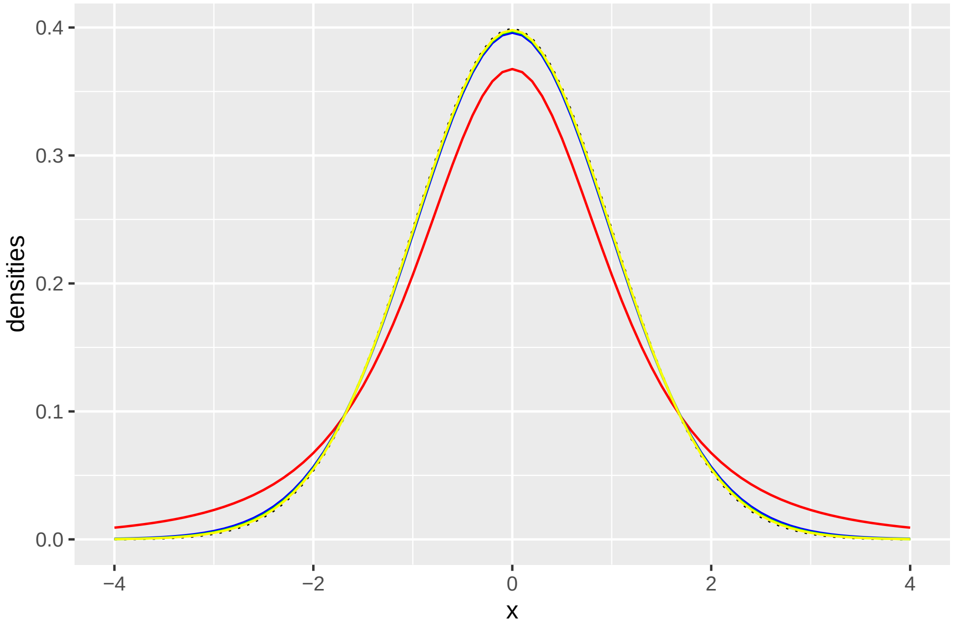
```
> #QQ-plot of normal distribution
> QQ_normal=ggplot(df_normal) + stat_qq(aes(sample = normal_sample), colour = "red")+
+   stat_qq_line(aes(sample = normal_sample))+labs(title="QQ-plot for normal")
> #QQ-plot of t-distribution
> QQ_t=ggplot(df_t) + stat_qq(aes(sample = t_sample), colour = "green")+
+ stat_qq_line(aes(sample = t_sample))+
+ labs(title="QQ-plot for t-distibution")
```





Problem:2(b)

```
> x = seq(-4, 4, by = .1)
> densities = dnorm(x,mean = 0, sd = 1)
> t_3 = dt(x,df = 3)
> t_33 = dt(x,df = 33)
> t_66 = dt(x,df = 66)
> t_99 = dt(x,df = 99)
> data = data.frame(x,densities,t_3,t_33,t_66,t_99)
> plot_nor_t=ggplot(data,aes(x))+ geom_line(aes(y =densities),linetype='dotted')+ geom_line(aes(y = t_3)
+ geom_line(aes(y = t_33),color='blue')+
+ geom_line(aes(y = t_66),color ='green')+
+ geom_line(aes(y = t_99),color = 'yellow')
```



As we can see as we increase the degrees of freedom, the colored smooth lines start to overlap with the dashed line representing normal distribution. Thus, we can say as the degree of freedom increases, t distribution starts behaving like a normal distribution.

Problem:3

We wish to test if a coin given to us is fair. We toss it a 100 times and find there are 45 heads. For which R-code is as follows:

```
> prop.test(45,100)



        1-sample proportions test with continuity correction

data:  45 out of 100, null probability 0.5
X-squared = 0.81, df = 1, p-value = 0.3681
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3514281 0.5524574
sample estimates:
   p
0.45
```

Here the prop.test() command tests the default null hypothesis i.e. population proportion(here,coin is fair) H_0 :$p = 0.5$ against the alternative hypothesis H_1:$p \neq 0.5$.
The value of chi-squared test statistic is 0.81.
As we know that p-value is the probability of obtaining the test result as extreme the result is actually observed under the assumption null hypothesis is true.
It also gives the 95% confidence interval for the population proportion and finally, it gives the sample estimates of population proportion i.e. sample proportion.
From the output we can see that for the test $p - value = 0.3681 > 0.05$. Therefore, we fail to reject the default null hypothesis(by definition of p-value) that is the population proportion i.e. the probability of head occurs is equal to 0.5 at 5% level of significance. So, we could agree that the coin given to us is fair.
Also, we get a confidence interval of p (probability of head occurs)is $(0.3514281, 0.5524574)$.

Problem:4

In the previous example if we toss the coin a 10000 times and find that there are 4500 heads and then we have to make conclusion about coin is fair or not.
For which R-code is:

```
> prop.test(4500,10000)


        1-sample proportions test with continuity correction

data:  4500 out of 10000, null probability 0.5
X-squared = 99.8, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4402205 0.4598181
sample estimates:
   p
0.45
```

Here in this case of 10000 tosses our null hypothesis is that the coin is fair i.e. H_0 :$p = 0.5$ but we get that the $p - value$ for this test as $p - value < 2.2e - 16 < 0.05$. So, we would reject the null hypothesis at 5% level of significance.
Therefore, we couldn't agree that the coin given to us is fair in this case.

Problem:5

Suppose Doddapple manufactures claims that their batteries last 25 years. Let, X is the random variable denoting the battery life of a randomly selected battery of Doddapple.
We assume, $X \sim N(\mu, \sigma^2)$ Given, Students from CMI's Data Science programs sample 10 users and find the sample mean time for battery life was 21 with a sample standard deviation of 1.7.
To test,
The Null Hypothesis H_0:$\mu = 25$i.e., true mean of X is 25
against, The Alternative Hypothesis H_1:$\mu \neq 25$ i.e.,true mean of X is not equal to 25.
And, the population standard deviation is unknown.
The t-statistic is defined as $T = \frac{\sqrt{n}(\bar{X}-c)}{s}$, where, c is the value of μ under null hypothesis and s is the sample estimate of standard deviation. Under H_0, the test statistic $\frac{\sqrt{n}(\bar{X}-c)}{s}$ follows t_{n-1} distribution. So, we reject H_0, if $|T_{observed}| > t_{\frac{\alpha}{2},n-1}$ and we take the level of significance $\alpha = 0.05$ Now,

```
> alpha = 0.05
> c = 25
> n=10
> sample_mean = 21
> sample_sd = 1.7
> #calculating observed value
> t_observed = sqrt(n)*(sample_mean - c)/sample_sd
> t_observed
```

```
[1] -7.440653
```

```
> #now, calculating critical value
> t_crit = qt(p=alpha/2, df= 9 ,lower.tail=F)
> t_crit
```

```
[1] 2.262157
```

Here, we get
$n = 10, \bar{X} = 21, c = 25, s = 1.7,$
Value of statistic under Null Hypothesis $= -7.440653$, df = degrees of freedom $= 10 - 1 = 9$ (1 degree is lost as standard deviation is unknown)
the critical value $= t_{0.025,9} = 2.262157$
Therefore, we get the value of $|T_{observed}| = 7.440653 > 2.232157(critical value)$
Thus, we reject the null hypothesis H_0.
Therefore, we can say from the sample study the mean of battery life, differ significantly from claimed mean 25 i.e., the claim of Doddapple is not believable.