AMS597_HW2_Solution_Spring2024

Question 1(a)

```
set.seed(123)
(mult_sample <- sample(LETTERS[1:4], 200, replace = T, p = c(0.15, 0.2, 0.35, 0.3)))
```

```
##   [1] "C" "B" "D" "A" "A" "C" "D" "A" "D" "D" "A" "D" "B" "D" "C" "A" "C" "C"
##  [19] "C" "A" "A" "B" "D" "A" "B" "B" "D" "D" "C" "C" "A" "A" "B" "B" "C" "D"
##  [37] "B" "C" "C" "C" "C" "D" "D" "D" "C" "C" "C" "D" "C" "A" "C" "D" "B" "C"
##  [55] "D" "C" "C" "B" "A" "D" "B" "C" "D" "C" "B" "D" "B" "B" "B" "D" "B" "D"
##  [73] "B" "C" "D" "C" "D" "D" "D" "C" "C" "B" "D" "B" "C" "D" "A" "A" "A" "C"
##  [91] "C" "B" "C" "B" "C" "C" "B" "C" "D" "D" "D" "C" "D" "A" "D" "A" "A" "D"
## [109] "D" "C" "A" "C" "C" "A" "B" "C" "D" "A" "D" "D" "D" "C" "C" "C" "D" "A"
## [127] "C" "C" "C" "B" "D" "A" "B" "B" "D" "B" "B" "B" "A" "D" "C" "D" "C" "C"
## [145] "B" "C" "C" "C" "C" "B" "B" "D" "D" "C" "C" "D" "D" "C" "D" "C" "D" "D"
## [163] "D" "D" "D" "D" "B" "C" "D" "C" "D" "C" "A" "B" "B" "D" "D" "D" "A" "D"
## [181] "B" "C" "B" "C" "D" "D" "C" "D" "A" "A" "C" "C" "A" "D" "A" "D" "D" "B"
## [199] "C" "D"
```

```
table(mult_sample)
```

```
## mult_sample
##  A  B  C  D
## 30 38 65 67
```

Question 1(b)

```
rand_vec <- runif(200)
## Partition the values so that it matches the desired probaiblities
(rand_samp <- ifelse(rand_vec <= 0.15,"A",
                     ifelse(rand_vec <= 0.35, "B",
                            ifelse(rand_vec <= 0.7, "C", "D"))))
```

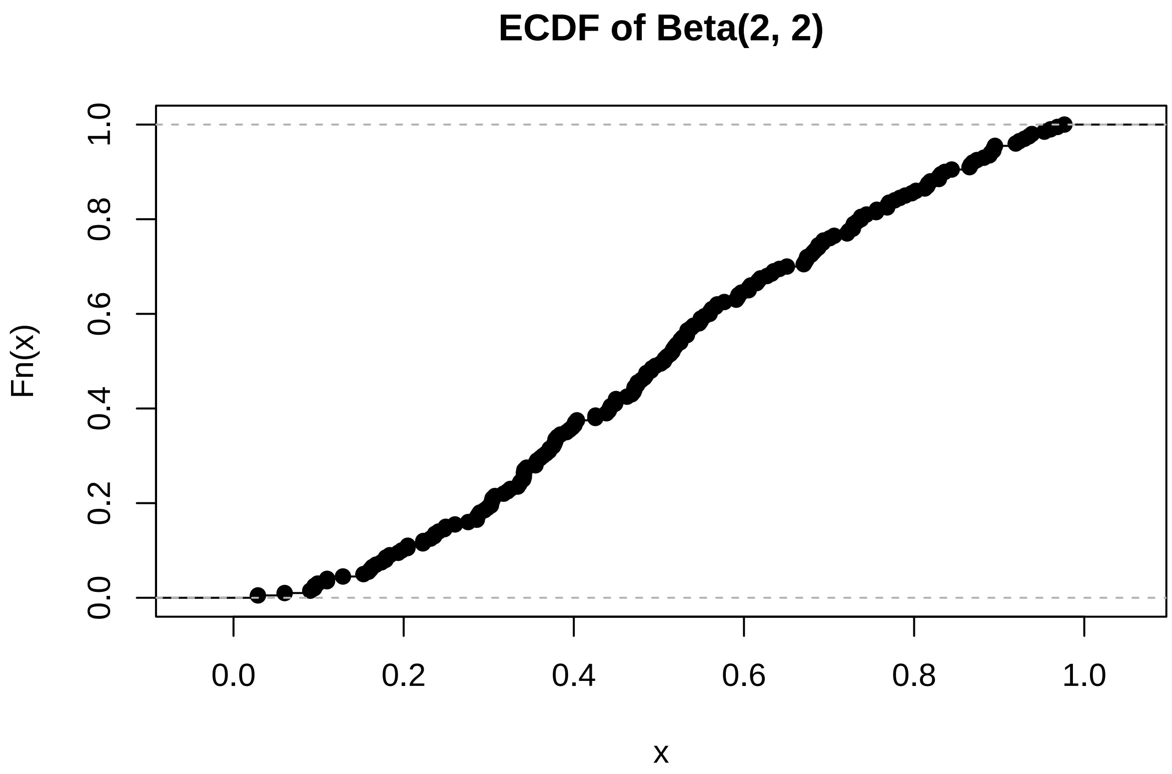
```
##   [1] "B" "D" "C" "C" "C" "D" "C" "B" "B" "B" "C" "B" "B" "C" "A" "D" "C" "C"
##  [19] "D" "D" "B" "D" "D" "C" "A" "C" "C" "C" "C" "D" "C" "C" "C" "A" "B" "C"
##  [37] "B" "D" "B" "D" "C" "C" "B" "C" "B" "D" "C" "D" "D" "D" "B" "B" "C" "B"
##  [55] "C" "D" "B" "C" "C" "D" "D" "D" "C" "D" "C" "C" "B" "B" "A" "C" "D" "A"
##  [73] "A" "B" "D" "D" "D" "C" "A" "C" "D" "A" "C" "B" "A" "C" "A" "B" "A" "C"
##  [91] "B" "A" "A" "D" "D" "D" "D" "A" "A" "D" "D" "A" "D" "D" "C" "C" "B" "A"
## [109] "C" "C" "C" "C" "D" "A" "C" "D" "D" "B" "C" "D" "D" "B" "A" "D" "A" "A"
## [127] "D" "A" "B" "D" "C" "B" "D" "D" "B" "C" "C" "C" "C" "D" "C" "A" "C" "B"
## [145] "C" "C" "D" "C" "B" "C" "A" "D" "C" "B" "C" "D" "C" "C" "B" "D" "A" "B"
## [163] "D" "B" "C" "D" "B" "A" "C" "C" "C" "B" "D" "C" "C" "D" "D" "B" "B" "D"
## [181] "C" "D" "C" "D" "C" "D" "B" "B" "A" "B" "D" "B" "D" "D" "A" "B" "B" "A"
## [199] "A" "D"
```

```
table(rand_samp)
```

```
## rand_samp
##  A  B  C  D
## 30 44 66 60
```

Question 2(a)

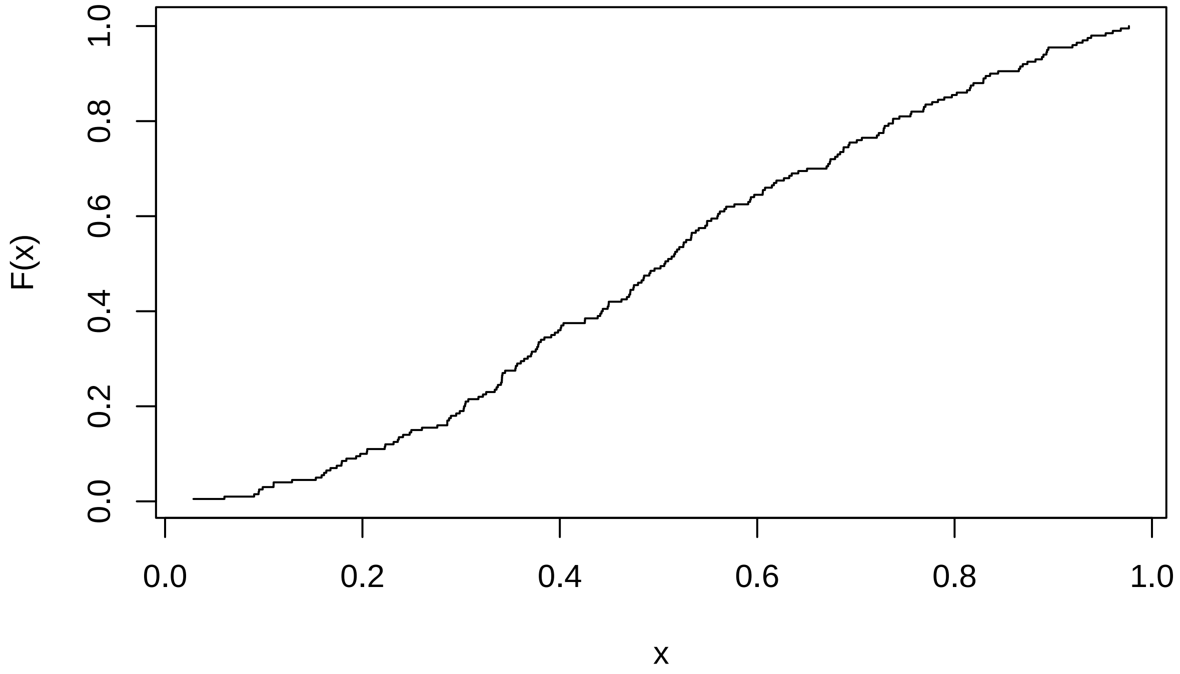
```
rand_beta <- rbeta(200, 2, 2)
plot(ecdf(rand_beta), main = "ECDF of Beta(2, 2)")
```



Question 2(b)

```r
my.plot.ecdf <- function(x) {
  n <- length(x)
  sorted_x <- sort(x)
  ecdf_vals <- 1:n / n
  plot(sorted_x, ecdf_vals, type = "s", xlab = "x", ylab = "F(x)")
}

my.plot.ecdf(rand_beta)
```



Question 3(a)

For the following parts we

```r
df <- read.table("http://www.ams.sunysb.edu/~pfkuan/Teaching/AMS597/Data/TwentyTreatments.txt",
                 header = T)
control <- df$Control
t.test(control, mu = -1)
```

```
##
##  One Sample t-test
##
## data:  control
## t = 1.6548, df = 39, p-value = 0.106
## alternative hypothesis: true mean is not equal to -1
## 95 percent confidence interval:
##  -1.1114718  0.1143957
## sample estimates:
##  mean of x
## -0.4985381
```

Since the p-value is greater than 0.05, we fail to reject the null hypothesis. We do not have evidence to support the claim that the mean of the gene for the control group differs from -1.

Question 3(b)

```r
group20 <- df$Treatment20
wilcox.test(group20, mu = -1)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  group20
## V = 483, p-value = 0.3336
## alternative hypothesis: true location is not equal to -1
```

Since the p-value is greater than 0.05, we fail to reject the null hypothesis. We do not have evidence to support the claim that the mean of the gene for treatment group 20 differs from -1.

Question 3(c)

```r
group10 <- df$Treatment10
group15 <- df$Treatment15
var.test(group10, group15)
```

```
##
##  F test to compare two variances
##
## data:  group10 and group15
## F = 1.4053, num df = 39, denom df = 39, p-value = 0.2923
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7432772 2.6570808
## sample estimates:
## ratio of variances
##           1.405328
```

Since the p-value for the F-test is greater than 0.05, we do not reject the null hypothesis that the variances are equal. We do not have evidence to support the claim that treatment 10 and treatment 15 have different variances.

Question 3(d)

```
control <- df$Control

## Treatment is a df that does not include the Control column
treatment <- df[, -1]
## Perform the two-sample t-tests
results <- lapply(treatment, function(x) {
  var.equal <- var.test(control, x)$p.value > 0.05
  t.test(control, x, var.equal=var.equal)
}
)
## Extract the p-values
(p.values <- sapply(results, function(x) x$p.value))
```

```
##   Treatment1   Treatment2   Treatment3   Treatment4   Treatment5   Treatment6
## 5.641986e-01 2.609130e-07 1.388434e-01 1.480807e-01 7.648649e-05 9.940690e-03
##   Treatment7   Treatment8   Treatment9  Treatment10  Treatment11  Treatment12
## 2.508631e-02 5.075132e-01 5.209345e-01 7.497424e-04 1.491421e-01 3.336929e-01
##  Treatment13  Treatment14  Treatment15  Treatment16  Treatment17  Treatment18
## 4.904117e-02 9.783195e-01 7.844351e-06 8.012802e-02 6.370440e-03 1.565212e-01
##  Treatment19  Treatment20
## 8.302484e-01 4.856996e-01
```

```
## List the treatments that are significantly different
names(p.values)[p.values<=0.05]
```

```
## [1] "Treatment2"  "Treatment5"  "Treatment6"  "Treatment7"  "Treatment10"
## [6] "Treatment13" "Treatment15" "Treatment17"
```

For treatment groups 2, 5, 6, 7, 10, 13, 15, and 17, we reject the null hypothesis and we have evidence to support the claim that there is any significant difference between the mean of gene for the control group and each of these treatments groups at $\alpha = 0.05$ per comparison rate.

As for the other treatment groups, we fail to reject the null hypothesis and have no evidence to support the claim that there is any significant difference between the mean of gene for the control group and each of these treatments groups at $\alpha = 0.05$ per comparison rate.

Question 3(e)

```
## Perform pairwise Wilcox tests between control group and each treatment group
results_wilcox <- lapply(treatment, function(x) {
  wilcox.test(control, x, alternative = "two.sided", exact = FALSE)
}
)

## Extract the p-values
(p.values_wilcox <- sapply(results_wilcox, function(x) x$p.value))
```

```
##   Treatment1   Treatment2   Treatment3   Treatment4   Treatment5   Treatment6
## 7.038794e-01 1.042609e-06 1.826247e-01 1.763875e-01 7.813876e-05 1.286743e-02
##   Treatment7   Treatment8   Treatment9  Treatment10  Treatment11  Treatment12
## 3.305746e-02 5.800621e-01 5.160029e-01 1.245064e-03 1.890240e-01 3.631777e-01
##  Treatment13  Treatment14  Treatment15  Treatment16  Treatment17  Treatment18
## 6.260949e-02 1.000000e+00 1.661799e-05 7.907013e-02 5.032282e-03 2.092114e-01
##  Treatment19  Treatment20
## 7.691504e-01 5.036460e-01
```

```r
## List the treatments that are significantly different
names(p.values_wilcox)[p.values_wilcox<=0.05]
```

```
## [1] "Treatment2"  "Treatment5"  "Treatment6"  "Treatment7"  "Treatment10"
## [6] "Treatment15" "Treatment17"
```

For treatment groups 2, 5, 6, 7, 10, 15, and 17, we reject the null hypothesis and we have evidence to support the claim that there is any significant difference between the mean of gene for the control group and each of these treatments groups at $\alpha = 0.05$ per comparison rate.

This is similar to using the t-test but treatment group 13 is not significant anymore using the Wilcox test.

Question 4

```r
my.t.test <- function(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0) {

  ## Gather important statistics for the first sample
  n1 <- length(x)
  mu_x <- mean(x)
  sigma_x <- sd(x)

  ## One sample t-test
  if (is.null(y)) {
    t_stat <- (mu_x - mu) / (sigma_x / sqrt(n1))
    df <- n1 - 1
  } else {
    ## Two sample t-test
    ## Gather important statistics for second sample
    n2 <- length(y)
    mu_y <- mean(y)
    sigma_y <- sd(y)

    F_test_pv <- var.test(x,y)$p.value
    ## Pooled Variances
    if (F_test_pv > 0.05) {
      df <- n1 + n2 - 2
      s_p <- sqrt( ((n1-1)*(sigma_x^2) + (n2-1)*(sigma_y^2)) / df  )
      t_stat <- (mu_x - mu_y - mu) / (s_p * sqrt((1/n1 + 1/n2)))
    } else { ## Unpooled Variances
      w1 <- sigma_x^2 / n1
      w2 <- sigma_y^2 / n2
      df <- (w1 + w2)^2 / (w1^2/(n1-1) + w2^2/(n2-1))
      t_stat <- (mu_x - mu_y - mu) / (sqrt(w1 + w2))
    }
  }

  if (alternative == "two.sided") {
    pv <- 2 * pt(abs(t_stat), df = df, lower.tail = FALSE)
    } else if (alternative == "less") {
    pv <- pt(t_stat, df = df, lower.tail = TRUE)
    } else if (alternative == "greater") {
    pv <- pt(t_stat, df = df, lower.tail = FALSE)
    }
  return(list(stat = t_stat, df = df, p.value = pv))
}
```

Check if it produces the same result as t.test().

Checking One sample t-test

```
set.seed(123)
x <- rnorm(10, 2, 3)
(test1 <- t.test(x, mu =3))
```

```
##
##  One Sample t-test
##
## data:  x
## t = -0.85775, df = 9, p-value = 0.4133
## alternative hypothesis: true mean is not equal to 3
## 95 percent confidence interval:
##  0.1769889 4.2707650
## sample estimates:
## mean of x
##  2.223877
```

```
my.t.test(x, alternative = "two.sided", mu = 3)
```

```
## $stat
## [1] -0.8577471
##
## $df
## [1] 9
##
## $p.value
## [1] 0.4132902
```

```
(test2 <- t.test(x, alternative = "greater"))
```

```
##
##  One Sample t-test
```

```
##
## data:  x
## t = 2.4578, df = 9, p-value = 0.01815
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  0.5652049       Inf
## sample estimates:
## mean of x
##  2.223877
```

```
my.t.test(x, alternative = "greater")
```

```
## $stat
## [1] 2.45776
##
## $df
## [1] 9
##
## $p.value
## [1] 0.01814589
```

```
(test3 <- t.test(x, alternative = "less", mu = 1))
```

```
##
##  One Sample t-test
##
## data:  x
## t = 1.3526, df = 9, p-value = 0.8954
## alternative hypothesis: true mean is less than 1
## 95 percent confidence interval:
##      -Inf 3.882549
## sample estimates:
## mean of x
##  2.223877
```

```
my.t.test(x, alternative = "less", mu = 1)
```

```
## $stat
## [1] 1.352591
##
## $df
## [1] 9
##
## $p.value
## [1] 0.895406
```

Checking Two sample t-test (pooled and unpooled)

```r
a <- rnorm(10, 2, 3)
b <- rnorm(10, 1, 3)
c <- rnorm(10, 2, 10)

(test1 <- t.test(a, b, mu =3, var.equal = T))
```

```
##
##  Two Sample t-test
##
## data:  a and b
## t = -0.075947, df = 18, p-value = 0.9403
## alternative hypothesis: true difference in means is not equal to 3
## 95 percent confidence interval:
##  0.120600 5.678485
## sample estimates:
##  mean of x  mean of y
##  2.6258659 -0.2736766
```

```r
my.t.test(a, b, alternative = "two.sided", mu = 3)
```

```
## $stat
## [1] -0.07594737
##
## $df
## [1] 18
##
## $p.value
## [1] 0.9402987
```

```r
(test2 <- t.test(a, b, alternative = "greater", var.equal = T))
```

```
##
##  Two Sample t-test
##
## data:  a and b
## t = 2.1921, df = 18, p-value = 0.02088
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.6058529       Inf
## sample estimates:
##  mean of x  mean of y
##  2.6258659 -0.2736766
```

```r
my.t.test(a, b, alternative = "greater")
```

```
## $stat
## [1] 2.192097
##
## $df
## [1] 18
##
## $p.value
## [1] 0.02088033
```

```r
(test3 <- t.test(a, c, alternative = "less", mu = 1, var.equal = F))
```

```
##
##  Welch Two Sample t-test
##
## data:  a and c
## t = -1.8562, df = 14.597, p-value = 0.04187
## alternative hypothesis: true difference in means is less than 1
## 95 percent confidence interval:
##      -Inf 0.806489
## sample estimates:
## mean of x mean of y
##  2.625866  5.220446
```

```r
my.t.test(a, c, alternative = "less", mu = 1)
```

```
## $stat
## [1] -1.856155
##
## $df
## [1] 18
##
## $p.value
## [1] 0.03993921
```

Question 5

```r
my.wilcox.test <- function(x, y, alternative = "two.sided") {

  n1 <- length(x)
  n2 <- length(y)

  ## Compute W1 and W2 statistics
  r <- rank(c(x, y))
  W1 <- sum(r[1:n1])
  W2 <- sum(r[(n1+1):(n1+n2)])

  ## Test statistics from the Mann-Whitney Test
  U1 <- W1 - n1*(n1+1)/2
  U2 <- W2 - n2*(n2+1)/2

  ## Compute p-value
  if(n1 < 12 || n2 < 12) {

    ## Exact test
    ## Generate all combinations of n1+n2 taken at n1 at a time
    total_com = combn(n1+n2, n1)
    ## Summing the ranks for each ordering
    all_W1 = colSums(total_com)
    ## Calculating the counts less or equal to our statistics
    prob = sum(all_W1<=W1)/length(all_W1)
    pv = 2*prob
    test_type <- "exact test"
  } else {

    ## Normal approximation test
    mean_rank <- n1 * (n1 + n2 + 1) / 2
    var_rank <- n1 * n2 * (n1 + n2 + 1) / 12
    z <- (W1 - mean_rank) / sqrt(var_rank)
    pv <- 2 * pnorm(abs(z), lower.tail = FALSE)
    test_type <- "normal approximation test"
  }

  return(list(W1 = U1, W2 = U2, p.value = pv, test_type = test_type))
}
```

Check if it produces the same result as wilcox.test(). Note that the test statistic for the wilcox.test() is

$$U_1 = W_1 - \frac{n_1(n_1+1)}{2} \; ; U_2 = W_2 - \frac{n_2(n_2+1)}{2}$$

This is from the Mann-Whitney Test.

Checking Exact Test

```
x <- rnorm(11)
y <- rnorm(10, 3, 1)
wilcox.test(x, y, exact = T, correct = F)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  x and y
## W = 1, p-value = 1.134e-05
## alternative hypothesis: true location shift is not equal to 0
```

```
my.wilcox.test(x,y)
```

```
## $W1
## [1] 1
##
## $W2
## [1] 109
##
## $p.value
## [1] 1.134057e-05
##
## $test_type
## [1] "exact test"
```

Checking Normal Approximation

```
x <- rnorm(20)
y <- rnorm(20, 1, 2)
wilcox.test(x, y, exact = F, correct = F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  x and y
## W = 72, p-value = 0.0005354
## alternative hypothesis: true location shift is not equal to 0
```

```
my.wilcox.test(x,y)
```

```
## $W1
## [1] 72
##
## $W2
## [1] 328
##
## $p.value
## [1] 0.0005353582
##
## $test_type
## [1] "normal approximation test"
```

Question 6(a)

Deriving the least-squares estimate of β by minimizing the sum of squared errors:

$$\min_{\beta} SSE = \sum_{i=1}^{n} (Y_i - \beta X_i)^2$$

Taking the derivative and setting it equal to zero:

$$\frac{\partial SSE}{\partial \beta} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta X_i) = 0$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

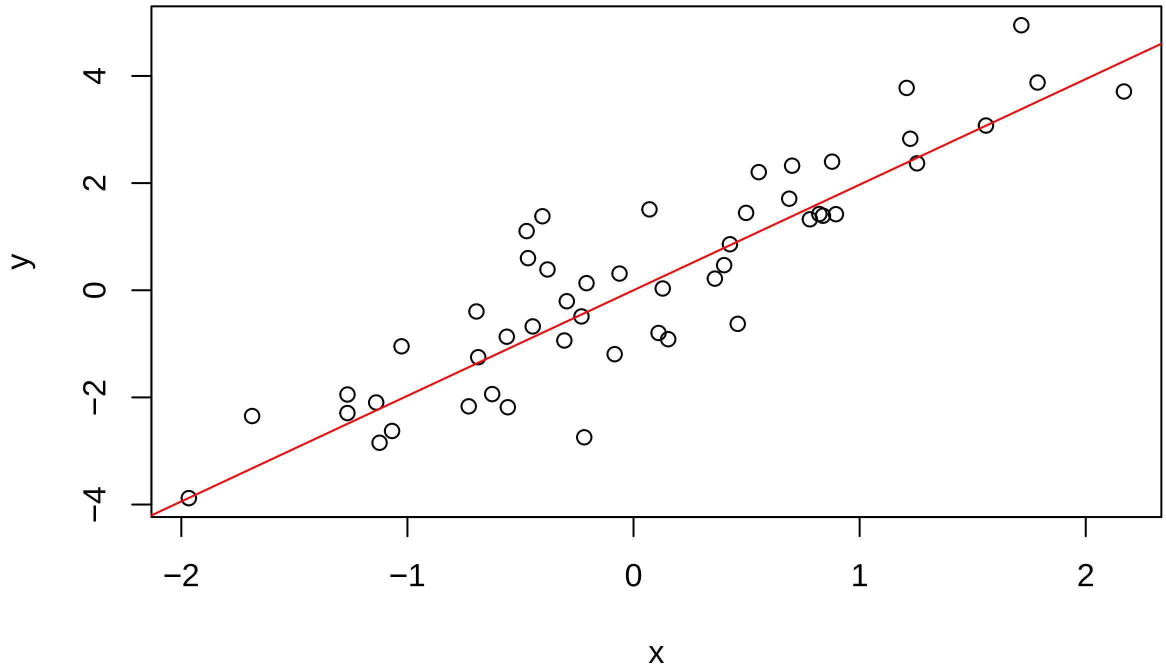
Question 6(b)

```
set.seed(123)
x <- rnorm(50)
y <- 2*x+rnorm(50)
```

```
sum_xy <- sum(x*y)
sum_xx <- sum(x^2)
(beta_hat <- sum_xy / sum_xx)
```

```
## [1] 1.970958
```

```
{plot(x,y)
abline(a = 0, b = beta_hat, col = "red")}
```



The slope from using the lm() function is very close to the value of our estimator we got in part (b).

Question 6(c)

```
fit <- lm(y ~ -1 + x)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ -1 + x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3155 -0.3422  0.1261  0.6394  2.1756
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x   1.9710     0.1414   13.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9171 on 49 degrees of freedom
## Multiple R-squared:  0.7986, Adjusted R-squared:  0.7945
## F-statistic: 194.3 on 1 and 49 DF,  p-value: < 2.2e-16
```

Question 7

```
my.kendall <- function(x, y) {
  num_concordant <- 0
  num_discordant <- 0
  n <- length(x)

  for (i in 1:n) {
    for (j in 1:n) {
      if (i < j) {
        num_concordant <- num_concordant + ((x[i] - x[j]) * (y[i] - y[j]) > 0)
        num_discordant <- num_discordant + ((x[i] - x[j]) * (y[i] - y[j]) < 0)
      }
    }
  }
  tau <- (num_concordant - num_discordant) / choose(n, 2)
  return (list(tau = tau))
}

set.seed(123)
x <- rnorm(50)
y <- 2*x+rnorm(50)
my.kendall(x, y)
```

```
## $tau
## [1] 0.6995918
```

```
cor(x, y, method = "kendall")
```

```
## [1] 0.6995918
```