Group 2: Kush Parmar, Jo-Anne Rivera, Simerjeet Mudhar

I pledge my honor that I have abided by the stevens honor system- Kush, Jo-Anne, Simerjeet

Calculation

Problem 1.

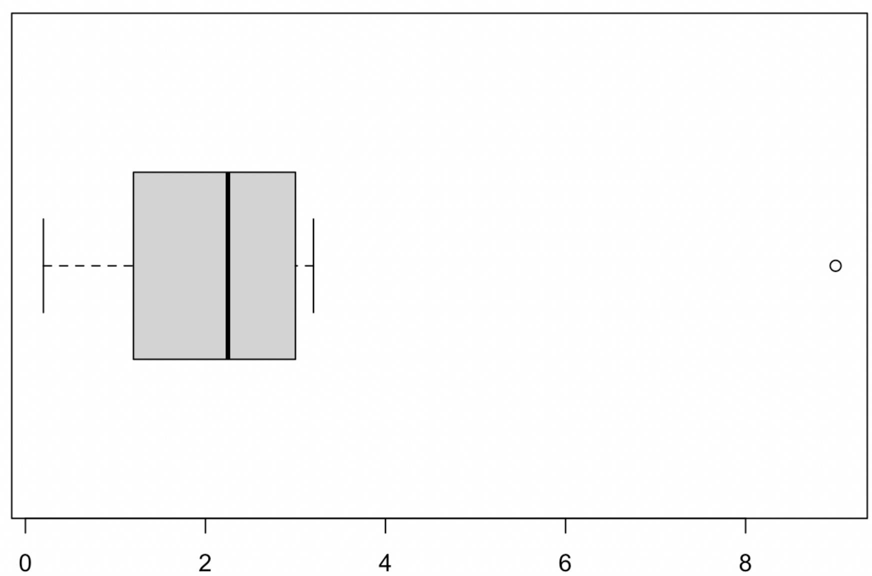
(0.2,1.1), (1.2,2.3), (0.9,1.1), (2.2,3.6), (3.2,0.1), (0.3,1.0), (1.7,6.9), (3.1,4.8), (2.3,6.5), (1.5,7.8), (2.5,5.8), (3.0,8.0), (2.6,9.4), (9.0,9.8).

i. **x** = c(0.2, 1.2, 0.9, 2.2, 3.2, 0.3, 1.7, 3.1, 2.3, 1.5, 2.5, 3.0, 2.6, 9.0)
fivenum(x) = 0.20 1.20 2.25 3.00 9.00
or **summary(x)** =

Minimum	1st Quarter	Median	Mean	3rd Quarter	Maximum
0.200	1.275	2.250	2.407	2.900	9.00

Sample Variance
var(x) = 4.568407

Boxplot for X_i



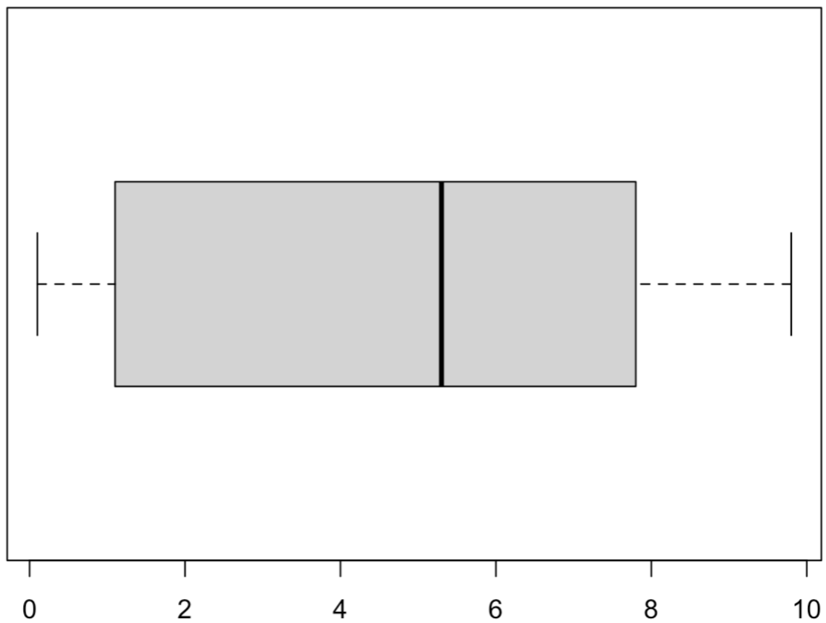
Skewness: Skewed to the right.
Outlier: 9.0

y = c(1.1, 2.3, 1.1, 3.6, 0.1, 1.0, 6.9, 4.8, 6.5, 7.8, 5.8, 8.0, 9.4, 9.8)
fivenum(y) = 0.10 1.10 5.30 7.80 9.80
or **summary(y)** =

Minimum	1st Quarter	Median	Mean	3rd Quarter	Maximum
0.100	1.400	5.300	4.871	7.575	9.800

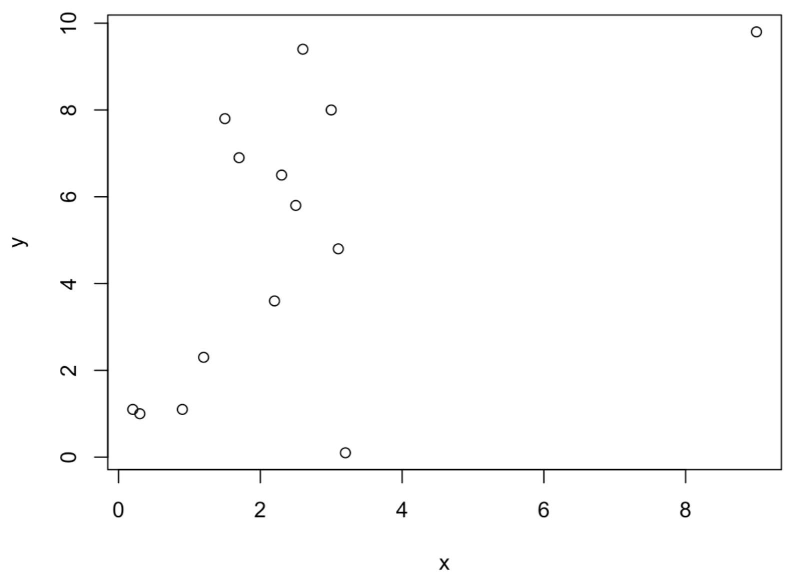
Sample Variance
var(y) = 11.17143

Boxplot for Y_i



Skewness: Skewed to the left.
Outlier: N/A

ii. **plot(x,y)**



Correlation Coefficient = cor(x, y) = 0.5679153
Qualitative description of linear association:

- (x_i, y_i) is more closely centered around a straight line.
- Larger y_is correspond to larger x_is.

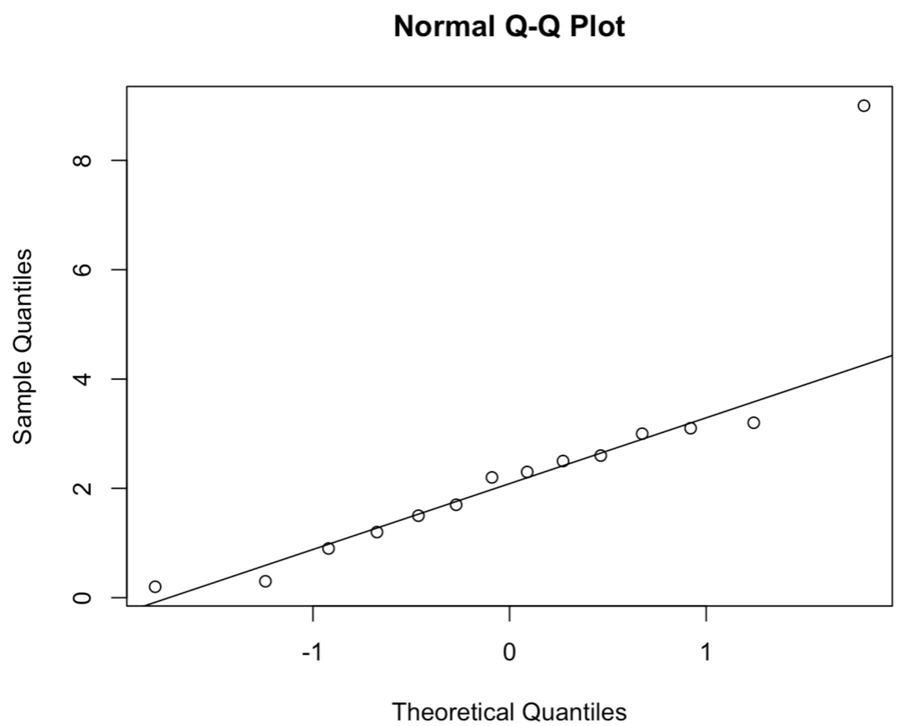
iii. The x_i coordinator 9.0 is an outlier, rendering the paired observation (9.0, 9.8) as an outlier. Must remove (9.0, 9.8).

Recomputed correlation coefficient: 0.4586256

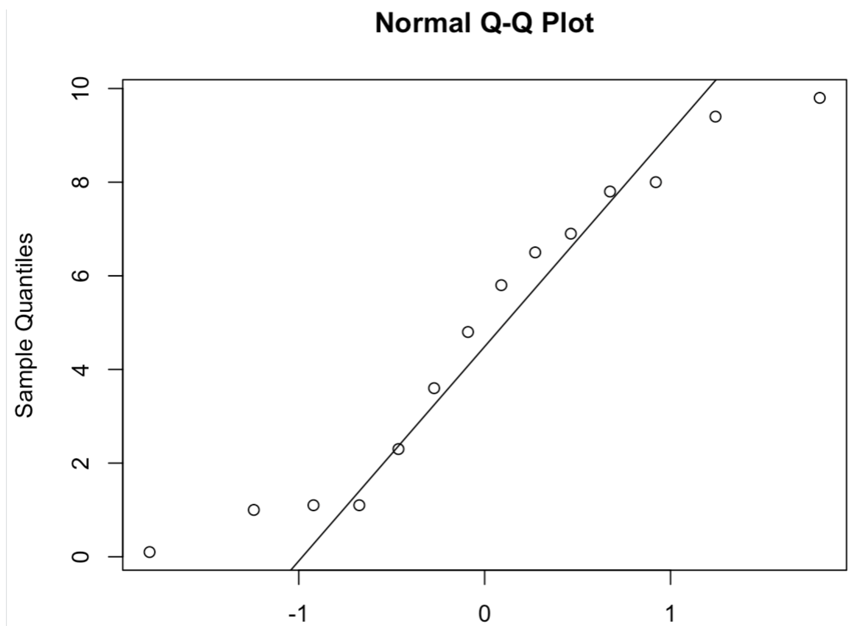
iv. The sample correlation coefficient decreased from the observation obtained in (ii. 0.5679153) to (iii. 0.4586256). Thuse, the paired observations are less closely centered around a straight line compared to before.

v. QQ Plots With Outliers

qqnorm(x), qqline(x)



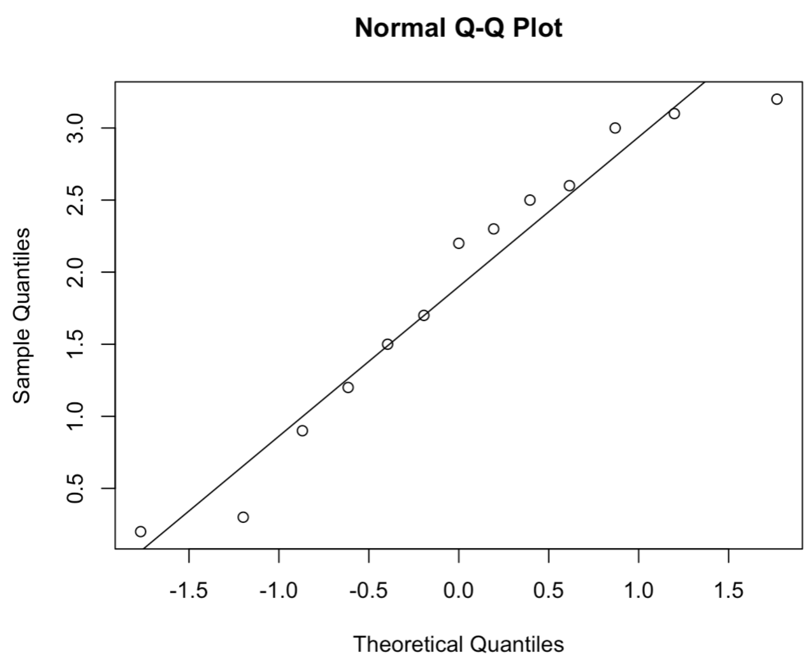
qqnorm(y), qqline(y)



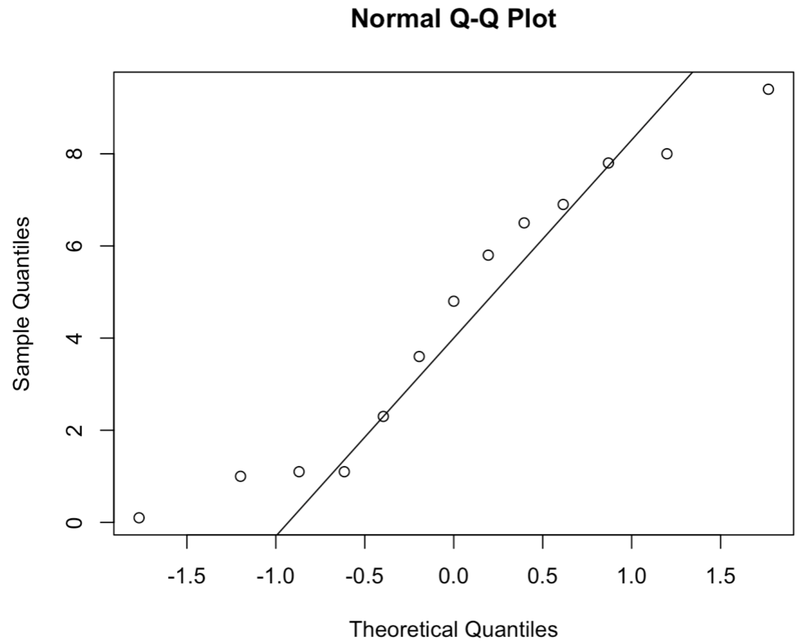
The y_i data set is more likely to be of normal distribution.

QQ Plots Without Outliers

qqnorm(x), qqline(x)



qqnorm(y), qqline(y)



The x_i data set is more likely to be of normal distribution.

Problem 2.

Z ~N(0,1)		
$P(Z \leq z_{0.05/2})$	$P(Z \leq z_{1-0.05/2})$	$P(z_{0.05/2} \leq Z \leq z_{1-0.05/2})$.

With Lower Tail
`> pnorm(1.96,0,1,TRUE)`
`[1] 0.9750021`
	With Lower Tail
	`> pnorm(-1.96,0,1,TRUE)`
	`[1] 0.0249979`
	Without Lower Tail
		With Lower Tail
		`> pnorm(1.96,0,1,TRUE)`
		`[1] 0.9750021`
		`> pnorm(-1.96,0,1,TRUE)`
		`[1] 0.0249979`
		`> 0.0249979-0.9750021`
		`[1] -0.9500042`

Without Lower Tail
`> pnorm(1.96,0,1,FALSE)`
`[1] 0.0249979`
	Without Lower Tail
	`> pnorm(-1.96,0,1,FALSE)`
	`[1] 0.9750021`
		Without Lower Tail
		`> pnorm(1.96,0,1,FALSE)`
		`[1] 0.0249979`
		`> pnorm(-1.96,0,1,FALSE)`
		`[1] 0.9750021`
		`> 0.9750021-0.0249979`
		`[1] 0.9500042`

To us, the more accurate answer is without taking the normal distribution without the lower tail. The issue with taking the lower tail comes from the fact that our resulting probability. When under the bounds of the Z values which includes the lower tail, results in a negative probability.

Problem 3

The following cdf functions utilize the Population Quantile Function. The equation results in the cut-off point below or equals the percentage, for any arbitrary inputted value.

1. $P(X \le F^{-1}(\alpha/2))$ → We are looking at a continuous X value, and this distribution is of the lower quartiles.
    - X → a continuous (possibly a random variable) distribution function, A real value
    - $z_\alpha = F^{-1}(\alpha) \rightarrow z_{\alpha/2} = F^{-1}(\alpha/2)$ →For $\alpha \in (0,1)$ hence the interval of the function follows is $1-\alpha$
    - $P(Z \le z_{\alpha/2}) = \alpha/2$ → finding the z score value for a one-tailed test. From score value
        - $z_\alpha = F^{-1}(\alpha/2)$ → results in a z value
    - $F^{-1}(\alpha/2)$ = inverse of the normal distribution, resulting in a percentage or a decimal between 0 to 1.

2. $P(X > F^{-1}(1 - \alpha/2))$ → We are looking at a continuous X value, and this distribution is of the upper quartiles.
    - X→ a continuous (possibly a random variable)
    - Symmetric idea $F^{-1}(1 - \alpha/2) = F^{-1}(\alpha/2)$ as established For $\alpha \in (0,1)$, hence the interval of the function follows is $1-\alpha$
      $P(X > F^{-1}(1 - \alpha/2)) = (1 - \alpha/2)$ - finding z value for a one-tailed test, from z score value
        - $z_\alpha = F^{-1}(1 - \alpha/2)$ → results in a z value
    - $F^{-1}(1 - \alpha/2)$= inverse of the normal distribution, resulting in a percentage or a decimal between 0 to 1.

3. $P(F^{-1}(\alpha/2) \le X \le F^{-1}(1 - \alpha/2))$
    - In this function we are looking for the normal distribution Z value between the upper and lower quartiles.
    - X→ a continuous (possibly a random variable)
    - Symmetric idea $F^{-1}(1 - \alpha/2) = F^{-1}(\alpha/2)$ under the bound For $\alpha \in (0,1)$, the value of $(1 - (1 - \alpha/2)) = \alpha/2$
    - $P(F^{-1}(\alpha/2) \le X \le F^{-1}(1 - \alpha/2))$ - finding the Z value for a two-tailed test from the z score value of both bounds.
        - $z_\alpha = F^{-1}(\alpha/2)$ → results in a z value
        - $z_\alpha = F^{-1}(1 - \alpha/2)$ → results in a z value
    - $F^{-1}(\alpha/2)$ and $F^{-1}(1 - \alpha/2)$ are inverses of the normal distribution resulting in a percentage

Problem 4

1. $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

$\sum_{i=1}^{n}(x_i - \bar{x}) = (x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) +...+ (x_n - \bar{x}) = (x_1 + x_2 + x_3 +...+ x_n) - \bar{x}n$

$\Sigma x_i - \bar{x}n$, **Separate equation:** $\bar{x} = \frac{\Sigma x_i}{n} \rightarrow \bar{x}n = \Sigma x_i$

$\bar{x}n - \bar{x}n = 0$

2. $(\sum_{i=1}^{n} x_i)^2 = \sum_{i=1}^{n} x_i^2 + 2 \sum_{1 \le i < j \le n} x_i x_j$

The following equation above is the sum of of squares

$(\sum_{i=1}^{n} x_i)^2 = \sum_{i=1}^{n} x_i^2 + 2 \sum_{1 \le i < j \le n} x_i x_j$ → can be represented as

$(x_i + x_j + x_{j+1}...)^2 = ((x_i)x_i + (2x_i + x_j)x_j + 2((2x_i + x_j)x_j +....)$

The line above shows that the sum of squares can be represented as summands plus the sum of all the double products of the summands in twos.

$\sum_{i=1}^{n} x_i^2 + 2 \sum_{1 \le i < j \le n} x_i x_j$

3. $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$ **i**

$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)$

$\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{n}(x_i^2) - 2\sum_{i=1}^{n}(x_i * \bar{x}) + \sum_{i=1}^{n}(\bar{x}^2)$

$\sum_{i=1}^{n}(x_i^2) - 2\sum_{i=1}^{n}(x_i * \bar{x}) + \sum_{i=1}^{n}(\bar{x}^2)$ → simplify $2\sum_{i=1}^{n}(n\bar{x} * \bar{x}) = 2\sum_{i=1}^{n}(n\bar{x}^2)$

$\sum_{i=1}^{n}(x_i^2 - 2(n\bar{x}^2) + \bar{x}^2) = \sum_{i=1}^{n}(x_i^2 - \bar{x}^2)$

4. $\sum_{i=1}^{n} x_i^2 \ge \frac{1}{n}(\sum_{i=1}^{n} x_i)^2$

$$\frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = \sum_{i=1}^{n}\left(\frac{x_i}{n}\right)^2$$

$$\sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{n}\left(\frac{x_i}{n}\right)^2$$