Homework 1

Daniel Detore

2025-01-26

1

i

Using:

```
x <- c(2.7, 4.0, 2.3, 5.4, -5.3, 1.8, -1.3, -2.9, 2.1, 3.9,
       -1.8, 0.4, -4.2, 0.5, -0.1, 1.5, -0.7)
y <- c(1.4, 2.5, 2.6, 5.6, -2.2, 0.4, 0.1, -3.0, 2.2, 0.9,
       -2.4, 1.6, -2.5, 0.1, -9.9, 1.1, -1.7)
```

Five-number summary of x:

```
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -5.3000 -1.3000  0.5000  0.4882  2.3000  5.4000
```

Five-number summary of y:

```
summary(y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -9.9000 -2.2000  0.4000 -0.1882  1.6000  5.6000
```

Sample variance of x:

```
var(x)
```

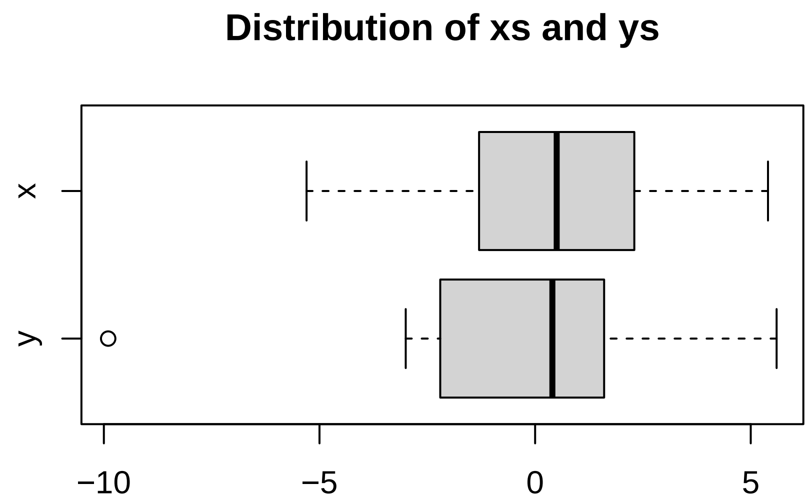
```
## [1] 8.673603
```

Sample variance of y:

```
var(y)
```

```
## [1] 11.37985
```

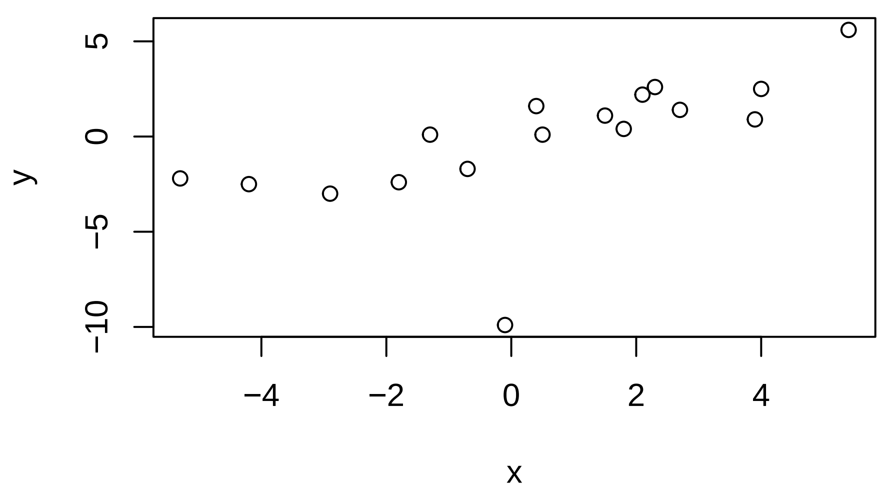
```
boxplot(y,x,
  main = "Distribution of xs and ys",
  names = c("y", "x"),
  horizontal = TRUE
)
```



The ys are skewed right while the $xs have no skew. The outlier is the y-value -9.9 from the point $(-0.1, -9.9)$.

ii

```r
plot(x, y)
```



Correlation coefficient:

```r
cor(x, y)
```

```
## [1] 0.6289777
```

Which means x and y are moderately linearly correlated.

iii

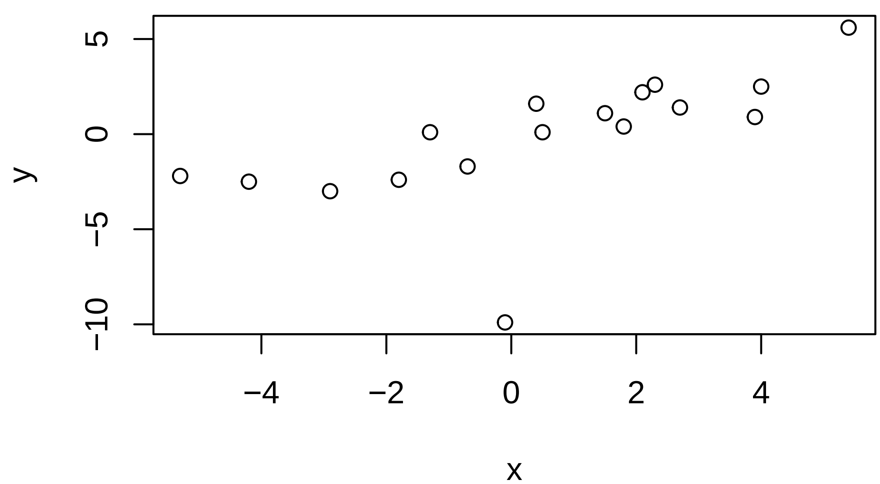
Yes; $(-0.1, -9.9)$ is an outlier.

```
x2 <- c(2.7, 4.0, 2.3, 5.4, -5.3, 1.8, -1.3, -2.9, 2.1, 3.9,
        -1.8, 0.4, -4.2, 0.5, 1.5, -0.7)
y2 <- c(1.4, 2.5, 2.6, 5.6, -2.2, 0.4, 0.1, -3.0, 2.2, 0.9,
        -2.4, 1.6, -2.5, 0.1, 1.1, -1.7)
cor(x2, y2)
```

```
## [1] 0.8822511
```

iv

```
plot(x, y)
```



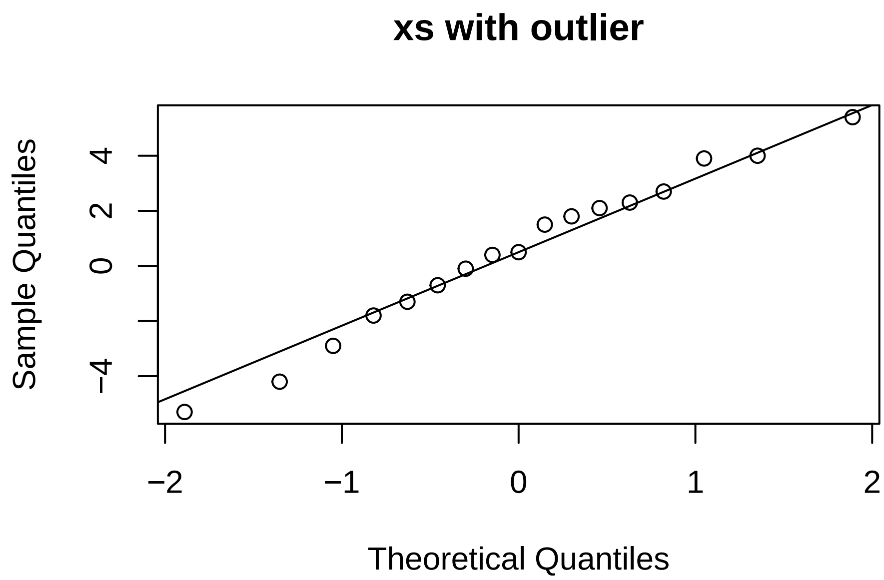
I can see the outlier $(-0.1, -9.9)$ at the bottom-center of the graph.

v

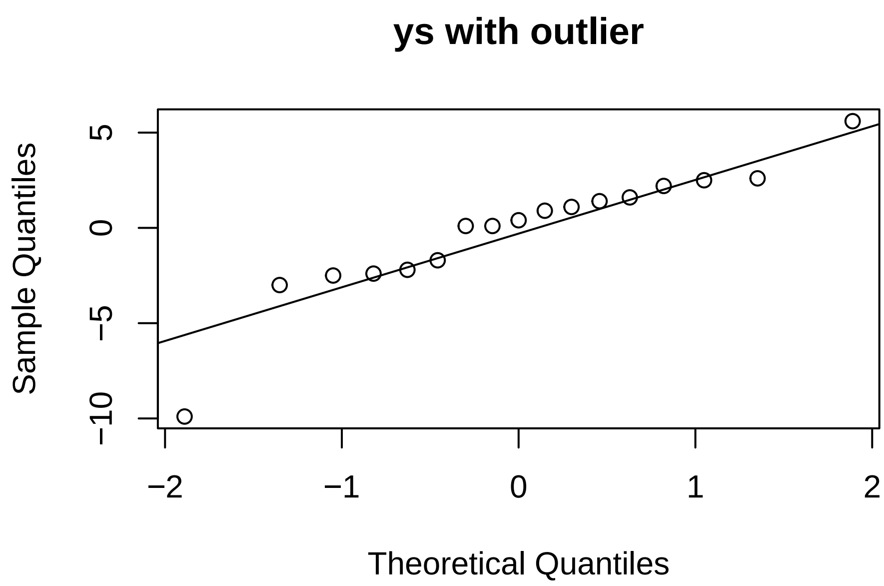
The sample correlation coefficient in iii is much higher than the one in ii.

vi

```
qqnorm(x, main="xs with outlier"); qqline(x)
```

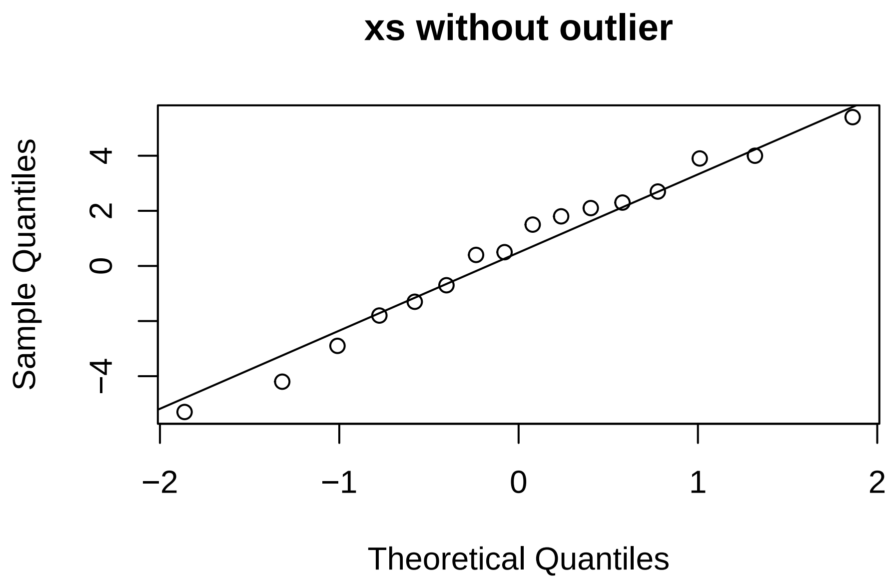



```r
qqnorm(y, main="ys with outlier"); qqline(y)
```

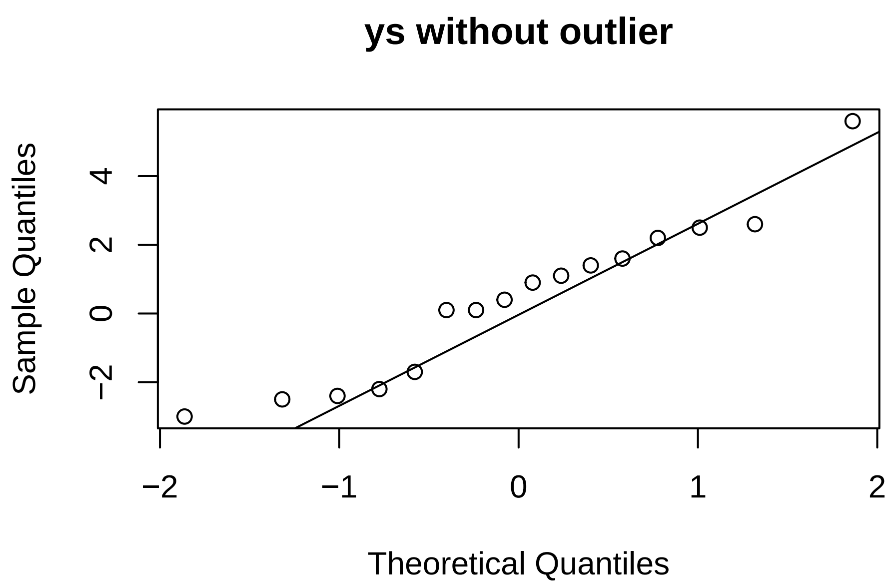


The xs look much closer to normal distribution than the ys.

```r
qqnorm(x2, main="xs without outlier"); qqline(x2)
```




```r
qqnorm(y2, main="ys without outlier"); qqline(y2)
```



The xs still look much closer to normal distribution than the ys.

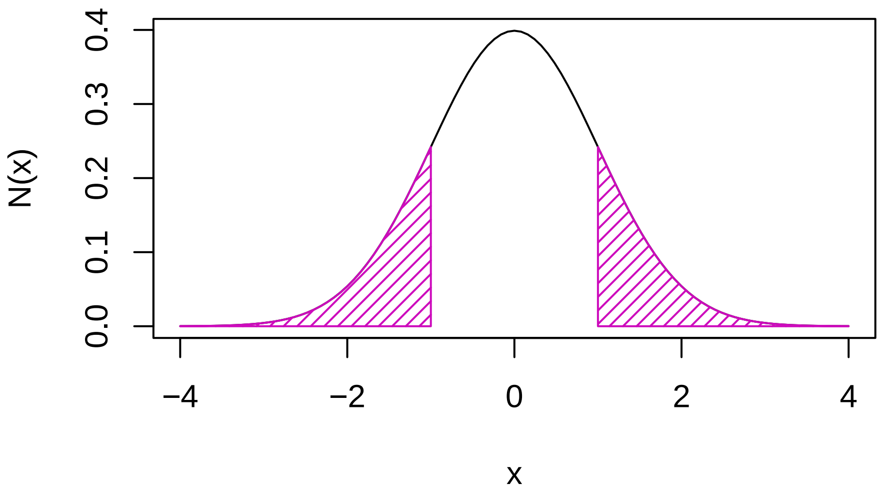
2

$P(|Z| < 1) =$

```r
pnorm(-1) + pnorm(1, lower.tail=FALSE)
```

```
## [1] 0.3173105
```

```r
N <- function(j) dnorm(j)
curve(N, from = -4, to = 4)
colorArea(from=-4, to=-1, dnorm, col=6, dens=20) #P(Z < -1)
colorArea(from=1, to=4, dnorm, col=6, dens=20) #P(Z > 1)
```

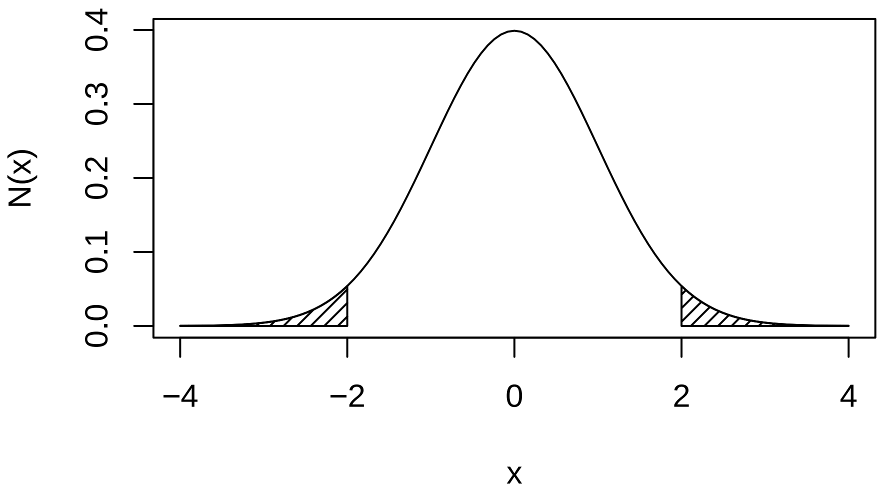


$P(|Z| < 2) =$

```r
pnorm(-2) + pnorm(2, lower.tail=FALSE)
```

```
## [1] 0.04550026
```

```r
curve(N, from = -4, to = 4)
colorArea(from=-4, to=-2, dnorm, col=1, dens=20) #P(Z < -2)
colorArea(from=2, to=4, dnorm, col=1, dens=20) #P(Z > 2)
```

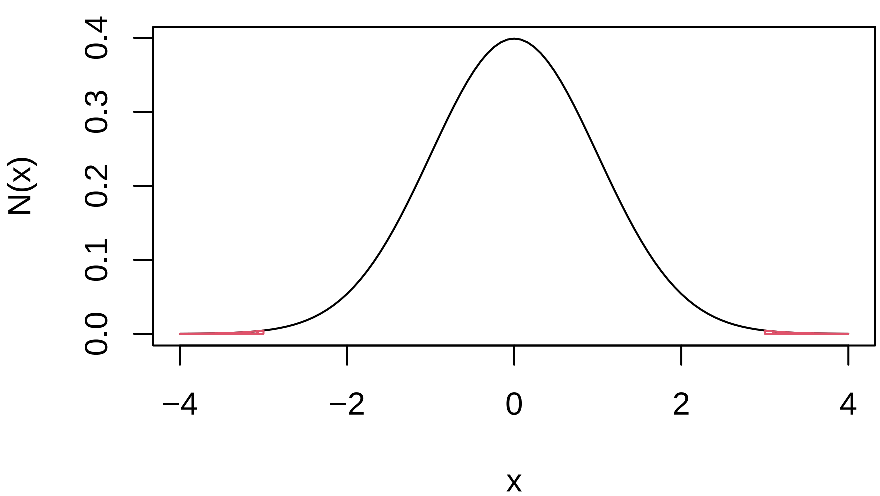


$P(|Z| < 3) =$

```r
pnorm(-3) + pnorm(3, lower.tail=FALSE)
```

```
## [1] 0.002699796
```

```r
curve(N, from = -4, to = 4)
colorArea(from=-4, to=-3, dnorm, col=2, dens=20) #P(Z < -3)
colorArea(from=3, to=4, dnorm, col=2, dens=20) #P(Z > 3)
```

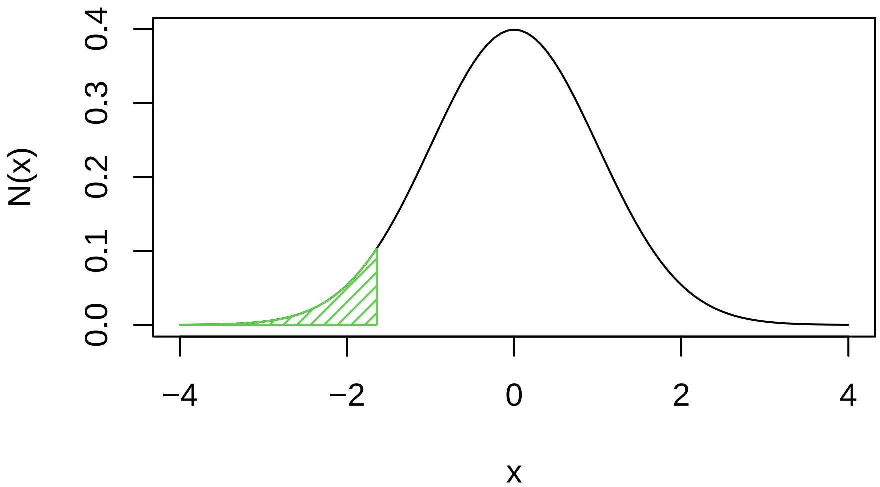


$P(Z \le z_{0.1/2}) =$

```r
pnorm(qnorm(0.1/2))
```

```
## [1] 0.05
```

```r
curve(N, from = -4, to = 4)
colorArea(from = -4, to = qnorm(0.1/2), dnorm, col=3, dens=20)
```

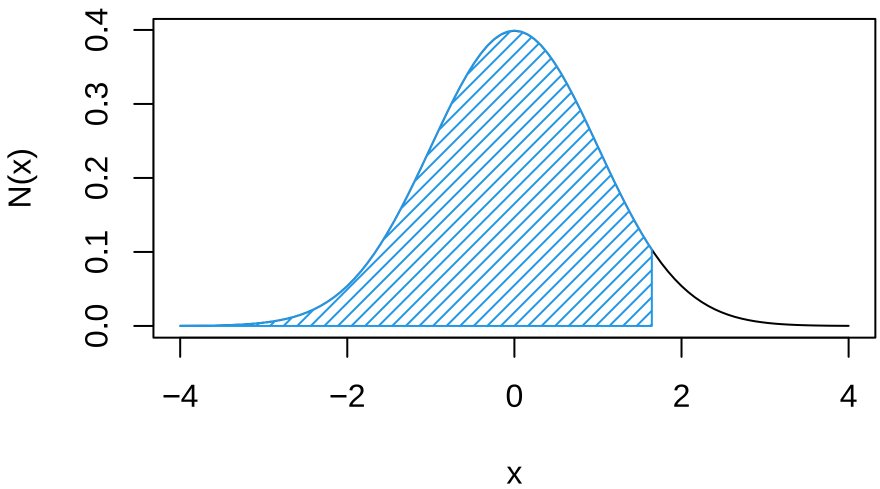


$P(Z \leq z_{1-0.1/2}) =$

```r
pnorm(qnorm(1-0.1/2))
```

```
## [1] 0.95
```

```r
curve(N, from = -4, to = 4)
colorArea(from = -4, to = qnorm(1 - 0.1/2), dnorm, col=4, dens=20)
```

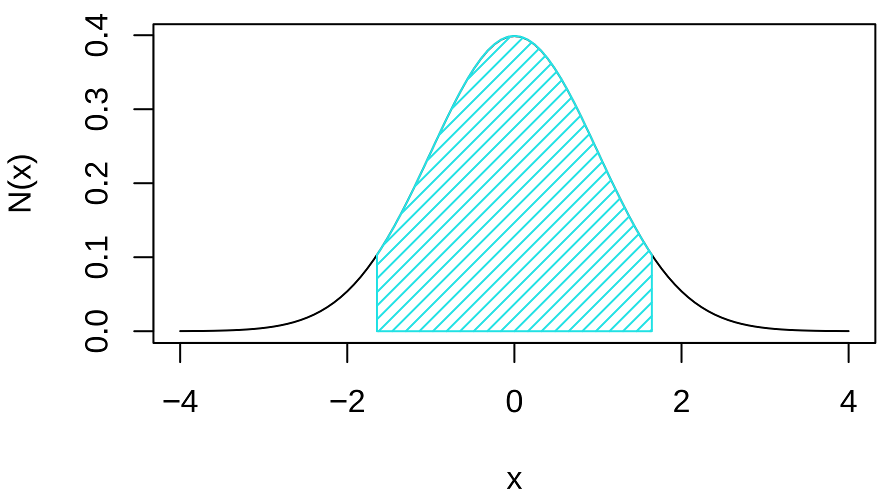


$P(z_{0.1/2} \leq Z \leq z_{1-0.1/2}) =$

```
pnorm(qnorm(0.1/2)) + pnorm(qnorm(1-0.1/2), lower.tail=FALSE)
```

```
## [1] 0.1
```

```
curve(N, from = -4, to = 4)
colorArea(from = qnorm(0.1/2), to = qnorm(1 - 0.1/2), dnorm, col=5, dens=20)
```



3

We know F' is the inverse of cdf $F(x) = P(X \leq x)$, which means $P(X \leq F'(x)) = F(F'(x)) = x$. This means:

- $P(X \leq F^{-1}(\alpha/2)) = \alpha/2$.
  - This probability will decrease with a decrease in α.
- $P(X > F^{-1}(1-\alpha/2)) = 1 - P(X \leq F^{-1}(1-\alpha/2)) = 1 - 1 - \alpha/2 = -\alpha/2$.
  - This probability will increase with a decrease in α.
- $P(F^{-1}(\alpha/2) \leq X \leq F^{-1}(1-\alpha/2)) = P(X \leq F^{-1}(1-\alpha/2)) - P(X \leq F^{-1}(\alpha/2)) = 1-\alpha/2-\alpha/2 = 1-\alpha$.
  - This probability will increase with a decrease in α.

4

i

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

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

$\square$

ii

If $n = 2$, then $(\sum_{i=1}^{n} x_i)^2 = (x_1 + x_2)^2 = x_1^2 + x_2^2 + 2x_1x_2 = \sum_{i=1}^{2} x_i^2 + 2\sum_{1\leq i<j\leq 2} x_ix_j$.

If $n = 3$, then $(\sum_{i=1}^{n} x_i)^2 = (x_1+x_2+x_3)^2 = x_1^2+x_2^2+x_3^2+2x_1x_2+2x_1x_3+2x_2x_3 = \sum_{i=1}^{3} x_i^2+2\sum_{1\leq i<j\leq 3} x_ix_j$.

We can generalize $\sum_{i=1}^{2} x_i + 2\sum_{1\leq i<j\leq 2} x_ix_j$ and $\sum_{i=1}^{3} x_i + 2\sum_{1\leq i<j\leq 3} x_ix_j$ to the form

$\sum_{i=1}^{n} x_i^2 + 2\sum_{1\leq i<j\leq n} x_ix_j$.

$\square$

iii

Note that by convention $n \geq 1$.

$(\sum_{i=1}^{n} x_i)^2 = \sum_{i=1}^{n} x_i^2 + 2\sum_{1\leq i<j\leq n} x_ix_j \implies$

$\frac{1}{n}(\sum_{i=1}^{n} x_i)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 + \frac{2}{n}\sum_{1\leq i<j\leq n} x_ix_j \implies \frac{1}{n}(\sum_{i=1}^{n} x_i)^2 \leq \frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq \sum_{i=1}^{n} x_i^2 \implies$

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

$\square$

iv

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

Recall $\sum_{i=1}^{n} x_i = n\bar{x} \implies \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

$\square$