MSMS - 106

Ananda Biswas

Practical 06

✎ Calculate the correlation coefficient for the following bivariate data.

Y \ X	18	19	20	21
10−20	4	2	2	–
20−30	5	4	6	4
30−40	6	8	10	11
40−50	4	4	6	8
50−60	–	2	4	4
60−70	–	2	3	1

➔ First we consider the midpoints of the intervals.

$f(x, y)$ be the frequency corresponding to (x, y).

$$N = \sum_x \sum_y f(x, y)$$

$$\bar{x} = \frac{1}{N} \sum_x \sum_y x \cdot f(x, y); \qquad \bar{y} = \frac{1}{N} \sum_x \sum_y y \cdot f(x, y)$$

$$\sigma_x^2 = \frac{1}{N} \sum_x \sum_y x^2 \cdot f(x, y) - \bar{x}^2; \qquad \sigma_y^2 = \frac{1}{N} \sum_x \sum_y y^2 \cdot f(x, y) - \bar{y}^2$$

$$cov(x, y) = \frac{1}{N} \sum_x \sum_y xy \cdot f(x, y) - \bar{x} \cdot \bar{y}.$$

```
x <- 18:21
```

```
lower_y <- seq(10, 60, by = 10)
upper_y <- seq(20, 70, by = 10)
y <- (lower_y + upper_y) / 2
```

```
values <- c(4, 2, 2, NA, 5, 4, 6, 4, 6, 8, 10, 11, 4, 4, 6, 8, NA, 2, 4, 4, NA, 2, 3, 1)
```

Here are the frequencies.

```r
freq <- matrix(values, nrow = 6, ncol = 4, byrow = TRUE)
freq
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    4    2    2   NA
## [2,]    5    4    6    4
## [3,]    6    8   10   11
## [4,]    4    4    6    8
## [5,]   NA    2    4    4
## [6,]   NA    2    3    1
```

```r
N <- 0

for (i in 1:dim(freq)[1]) {
  for (j in 1:dim(freq)[2]) {
    if(!is.na(freq[i, j])) N <- N + freq[i, j]
  }
}

N
```

```
## [1] 100
```

Total frequency is 100.

```r
marginal_x <- rep(0, length(x))

for (j in 1:dim(freq)[2]) {
  for (i in 1:dim(freq)[1]) {
    if(!is.na(freq[i, j])) marginal_x[j] <- marginal_x[j] + freq[i, j]
  }
}

marginal_x
```

```
## [1] 19 22 31 28
```

```r
marginal_y <- rep(0, length(y))

for (i in 1:dim(freq)[1]) {
  for (j in 1:dim(freq)[2]) {
    if(!is.na(freq[i, j])) marginal_y[i] <- marginal_y[i] + freq[i, j]
  }
}

marginal_y
```

```
## [1]  8 19 35 22 10  6
```

```r
numerator1 <- 0

for (i in 1:length(x)) {
  numerator1 <- numerator1 + x[i] * marginal_x[i]
}

mean_x <- numerator1 / N
mean_x

## [1] 19.68
```

$\bar{x} = 19.68$.

```r
numerator2 <- 0

for (i in 1:length(y)) {
  numerator2 <- numerator2 + y[i] * marginal_y[i]
}

mean_y <- numerator2 / N
mean_y

## [1] 37.5
```

$\bar{y} = 37.5$.

```r
total1 <- 0

for (i in 1:length(x)) {
  total1 <- total1 + x[i]^2 * marginal_x[i]
}

var_x <- total1 / N - mean_x^2
var_x

## [1] 1.1576
```

$\sigma_x^2 = 1.1576$.

```r
total2 <- 0

for (i in 1:length(y)) {
  total2 <- total2 + y[i]^2 * marginal_y[i]
}

var_y <- total2 / N - mean_y^2
var_y

## [1] 160.75
```

$\sigma_y^2 = 160.75$.

```r
total3 <- 0

for (i in 1:dim(freq)[1]) {
  for (j in 1:dim(freq)[2]) {
    if(!is.na(freq[i, j])) total3 <- total3 + y[i] * x[j] * freq[i, j]
  }
}

cov_xy <- total3 / N - mean_x * mean_y
cov_xy

## [1] 3.5
```

$cov(x, y) = 3.5$.

```r
corr_xy <- cov_xy / sqrt(var_x * var_y)
corr_xy

## [1] 0.2565744
```

$corr(x, y) = 0.2565744$.