Testing and Descriptive Statistics

Covariance and Correlation and more

Week 04

The Covariance

Year	Stock Market Growth (Y)	Economic Growth (X)
Random Year 1	3	1
Random Year 2	-2	-1
Random Year 3	4	2
Random Year 4	6	3
Random Year 5	9	5
Mean	**4**	**2**

$$\operatorname{cov}(X, Y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}.$$

$$\frac{(1-2)*(3-4)+(-1-2)*(-2-4)+(2-2)*(4-4)+(3-2)*(6-4)+(5-2)*(9-4)}{4}$$

Answer is 9

```r
# Covariance

# Input variables
Stock.Market.Growth <- c(3, -2, 4, 6, 9)
Economic.Growth <- c(1, -1, 2, 3, 5)

# Convert to Data Frame

df <- data.frame(Stock.Market.Growth, Economic.Growth)

# Covariance
# cov(x, y)

COV.P <- cov (Stock.Market.Growth, Economic.Growth)
```

The Correlation

Year	Stock Market Growth (Y)	Economic Growth (X)
Random Year 1	3	1
Random Year 2	-2	-1
Random Year 3	4	2
Random Year 4	6	3
Random Year 5	9	5
Mean	**4**	**2**

$$\rho_{r_x, r_y} = \frac{cov(r_x, r_y)}{\sigma_{rx}\sigma_{ry}} \tag{1-1}$$

where
$cov(r_x, r_y)$: the covariance of ranked data r_x and r_y.
σ_{rx} and σ_{ry} are the standard deviations of r_x and r_y.

$$R = \frac{Covariance}{Standard\ Diviation\ of\ X\ * Standard\ Diviation\ of\ Y}$$

Answer is 0.9908674

The Correlation

Correlation	Evaluation
−1～−0.7	Strong Negative Correlation
−0.7～−0.4	Significantly Negative Correlation
−0.4～−0.2	Slightly Negative Correlation
−0.2～0.2	No Correlation
0.2～0.4	Slightly Positive Correlation
0.4～0.7	Significantly Positive Correlation
0.7～1	Strong Positive Correlation

```r
# Covariance

# Input variables
Stock.Market.Growth <- c(3, -2, 4, 6, 9)
Economic.Growth <- c(1, -1, 2, 3, 5)

# Convert to Data Frame

df <- data.frame(Stock.Market.Growth, Economic.Growth)

# Covariance
# cov(x, y)

COV.P <- cov (Stock.Market.Growth, Economic.Growth)
COR.P <- cor (Stock.Market.Growth, Economic.Growth, method = "pearson")
```

The Correlation matrix

```
install.packages("Hmisc")
# Hmisc displays Correlation matrix with significance levels (p-value)

Corruption <- c(3, 2, 4, 6, 8)
df1 <- data.frame(Stock.Market.Growth, Economic.Growth, Corruption)

# Default Correlation matrix
cor(df1, method = "pearson")

# Load Sample data... you can use for practice

data("mtcars") #default sample

head(mtcars)
```

The Correlation matrix

```
install.packages("Hmisc")
library("Hmisc")

res2 <- rcorr(as.matrix(my_data))
res2
```

Plot the Correlation Matrix

```
install.packages("corrplot")

library(corrplot)

corrplot(res, type = "upper", order = "hclust",
        tl.col = "black", tl.srt = 45)
```

Descriptive Statistics

Table 21.1. Datasets descriptive statistics

	Mean	Standard deviation	Median	Q_1	Q_3	Skew
Panel A: Descriptive statistics of news database						
CSS_τ	49.93	5.04	50	50	52	−1.2187
ESS_τ	52.90	30.16	50	50	89	−0.0937
Panel B: Descriptive statistics of variables						
r_t	0.0016	0.8938	0.0000	−0.3140	0.3200	−4.0284
$\|r_t\|$	0.5169	0.7291	0.3170	0.1354	0.6473	19.3614
$NN_{f,t}$	0.7900	4.2480	0.0000	0.0000	0.0000	25.2494
$NP_{f,t}$	1.1596	5.5857	0.0000	0.0000	1.0000	15.9065
$NN_{m,t}$	0.1529	0.7607	0.0000	0.0000	0.0000	7.9527
$NP_{m,t}$	0.1865	0.7503	0.0000	0.0000	0.0000	6.2613

Note: This table presents the summary descriptive statistics of all the variables employed in this study. The summary statistics include mean value (*Mean*), median value (*Median*), 25 percentile (Q_1), 75 percentile (Q_3), and Skewness (*Skew*) for each variable. CSS_τ and ESS_τ are the CSS and ESS of each news story, respectively. r_t is the return in percentage at hour t. $|r_t|$ is the absolute percentage return. $NN_{f,t}$ and $NP_{f,t}$ are the number of negative and positive firm-specific news stories, respectively. $NN_{m,t}$ and $NP_{m,t}$ are the number of negative and positive macroeconomic news stories, respectively. The sample period is from January 1, 2001 to December 31, 2013.

Descriptive Statistics

```
install.packages("psych")
#  very simple library for descriptive statistics
# use describe() for the descriptive statistics

library(psych)

describe(res)
```

Measurement

Not everything that can be counted counts, and not everything that counts can be counted.
— William Bruce Cameron, Informal Sociology

Next Week

Please bring your laptop.

Week 5: The Basics and Practice (Measurement)
Examples of data mining that can be used in the social sciences in Africa and Middle Eastern fields is

- Handling Missing Data
- Visualization
- Correlation
- Exercises