Descriptive Multivariate Analysis

Multivariate descriptive statistics involves analysing relationships between more than two variables. Descriptive statistics provide simple summaries of (large amounts of) information (or data). These summaries are quantitative (e.g. means, correlations) or displayed visually (in graphs, scatterplots, etc.).

Multivariate descriptive statistics involves analysing relationships between more than two variables.

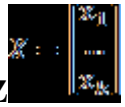
Descriptive statistics provide simple summaries of (large amounts of) information (or data). These summaries are quantitative (e.g. means, correlations) or displayed visually (in graphs, scatterplots, etc.).

Descriptive statistics can be "univariate" (involving one variable), "bivariate" (comparing two variables to determine whether there are any relationships between them), or "multivariate" (analysing whether there are relationships between more than two variables). For multivariate descriptions, the effect of one factor or variable is isolated from others to avoid distorting conclusions.

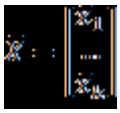
Multivariate statistics employs vectors of statistics (mean, variance, etc.), which can be considered an extension of the descriptive statistics described in univariate Descriptive Statistics.

Mean, Variance, and Standard Deviation Vectors

Definition 1: Given k random variables $x_1, \ldots, x_k$ and a sample of size n for each variable x_j of the form $x_{ij}, \ldots, x_{nj}$. We can define the $k \times 1$ column vector X (also known as a **random vector**) as

Z 

(also written more simply as $X = [x_j]$) and then define the **sample mean (vector)** of X to be



and similarly for the sample variance, standard deviation and other statistics. Also if the μ_j are the population means of the x_j then the **population mean** (**vector**) of X is defined to be

and similarly for population variance, standard deviation, etc. We can also define row vector versions of these.

Example 1: Figure 1 shows the following statistics for each of the EU countries: gross national product (GDP) per capita (measured in the purchasing power parity with thousands of US dollars), accumulated public debt (as a percentage of GDP), current annual public deficit (as a percentage of GDP), current annual inflation rate and percentage of the population that is unemployed. Find the sample mean vector.

	A	B	C	D	E	F
3	Country	GDP/capita	Public Debt	Deficit	Inflation	Unemployment
4	Austria	39.8	72.3	-4.6	1.7	3.9
5	Belgium	36.3	96.8	-4.1	2.3	6.7
6	Bulgaria	12.9	16.2	-3.2	3.0	11.9
7	Cyprus	29.0	60.8	-5.3	2.6	7.8
8	Czech Republic	25.0	38.5	-4.7	1.2	6.6
9	Denmark	36.4	43.6	-2.7	2.2	7.1
10	Estonia	18.5	6.6	0.1	2.7	12.8
11	Finland	34.9	48.4	-2.5	1.7	7.8
12	France	33.9	81.7	-7.0	1.7	9.9
13	Germany	36.1	83.2	-3.3	1.2	5.8
14	Greece	28.5	142.8	-10.5	4.7	16.7
15	Hungary	18.8	80.2	-4.2	4.7	9.9
16	Ireland	39.5	96.2	-32.4	-1.6	14.2
17	Italy	29.5	119.0	-4.6	1.6	8.3
18	Latvia	14.5	44.7	-7.7	-1.2	16.2
19	Lithuania	17.2	33.7	-7.1	1.2	10.3
20	Luxembourg	81.5	18.4	-1.7	2.8	4.8
21	Malta	24.8	68.0	-3.6	2.0	6.6
22	Netherlands	41.0	62.7	-5.4	0.9	4.5
23	Poland	19.0	55.0	-7.9	2.7	9.4
24	Portugal	23.3	93.0	-9.1	1.4	12.5
25	Romania	11.9	30.8	-6.4	6.1	7.5
26	Slovakia	22.2	41.0	-7.9	0.7	13.5
27	Slovenia	28.1	38.0	-5.6	2.1	8.0
28	Spain	29.8	60.1	-9.2	2.0	22.6
29	Sweden	38.2	39.8	0.0	1.9	7.2
30	United Kingdom	35.1	80.0	-10.4	3.3	8.0
31						
32	Mean	29.8	61.2	-6.3	2.1	9.6
33	Var	183.4	1019.9	35.2	2.5	18.3
34	Stdev	13.5	31.9	5.9	1.6	4.3

Figure 1 – Data for Example 1

The sample means row vector (range B32:F32) is [29.8, 61.2, -6.3, 2.1, 9.6], and similarly for variance and standard deviation. We can also look at column vector versions of these statistics. E.g. the sample variance column vector is

183.4
1019.9
35.2
2.5
18.3

Covariance and Correlation Matrices

Definition 2: Given a k × 1 column vector of random variables X = [xj] and samples of size n for each variable xj of the form xij, …, xnj. We can define the k × k sample variance-covariance matrix (or simply the sample covariance matrix) S as [sij] where sij = cov(xi, xj). Since cov(xj, xj) = var(xj) = s_j^2 and cov(xj, xj) = cov(xj, xi), the covariance matrix is symmetric with the main diagonal consisting of the sample variances.

Similarly, we can define the **population variance-covariance matrix** (or simply the **population covariance matrix**) Σ as above where the covariances are population covariances.

The sample and population **correlation matrices** can be defined as $[r_{ij}]$ where

$$r_{ij} = \frac{cov(x_i, x_j)}{stdev(x_i) \cdot stdev(x_j)}$$

Since

$$r_{ii} = \frac{cov(x_i, x_i)}{stdev(x_i) \cdot stdev(x_i)} = \frac{var(x_i)}{var(x_i)} = 1$$

it follows that the main diagonal of this matrix consists only of 1's.

Matrix Equations

By Property o of Least Squares in Multiple Regression, the sample covariance matrix can be expressed by the matrix equation

$$S = \frac{1}{n-1}(X - \bar{X})^T(X - \bar{X})$$

where $\bar{X}$ is the $1 \times k$ row vector of sample means. Also, the correlation matrix can be expressed as

$$\frac{1}{D^T D}S$$

where D = the $1 \times k$ row vector of sample standard deviations.