Statistical Methodology

The structure of the AKI 2016 reflects the same format of the AKI 2015 with its six composite indices: Pre-University Education; Technical Vocational Education and Training (TVET); Higher Education; Economy; Information and Communications Technology (ICT); and Research, Development and Innovation (RDI). As outlined in the previous chapters, some changes, such as adding/removing variables and updating data, have been made to better measure the state of knowledge systems in the region. The scope of the statistical analysis was expanded to: ensure consistency; select variables; determine weights; discover outliers, severe skewness and severe kurtosis; and ensure data adequacy for the accurate calculation of each index. The following sections provide a detailed review of the statistical steps taken in building the AKI 2016.

Variable selection

The constituent variables of the six indices of the AKI 2016 are similar to those of the 2015 edition; some new variables were introduced while others were removed, with most changes being made to the TVET Index and the RDI Index.

To ensure the consistency of the selected variables and their classification into the various pillars and sub-pillars, the principal components analysis and Cronbach's alpha coefficients were employed. In most cases, the explained variance ratio exceeded 50 percent[1] and Cronbach's alpha coefficient exceeded 0.7.[2]

Furthermore, the results of the correlation analysis confirmed the validity of the selection and classification of the variables. The correlation matrix for normalized variables was analysed to ensure they follow the same trend as the composite index, and confirmed the need to include variables that have high correlation coefficients (above 0.9) with the other variables.

Data used

The 432 variables incorporated into the AKI 2016 were obtained from external sources including United Nations Educational, Scientific and Cultural Organization (UNESCO) and other United Nations agencies; the World Bank; the International Telecommunication Union (ITU); the European Union (EU); the Organisation for Economic Co-operation and Development (OECD) and others.[3]

The team reviewed the data more than once to ensure no errors had occurred during data entry. Consequently, data was processed on the assumption that it was error-free. In the cases of variables that were linked to other factors – such as population or GDP – results were recalculated after adjusting for the effect of the size.

For the sake of transparency, simplicity and the possibility of replicating the results, no attempts were made to estimate missing values. The use of the arithmetic mean in computing the index is equivalent to estimating each of the missing values of the variable by the mean value. As usual in such cases, the missing values were not entered into the composite indices, which were calculated using the available data for each country.[4]

Data quality

The data employed in building the six sectoral indices should meet certain statistical criteria. In particular, data should be free from outliers, severe skewness and severe kurtosis, which might lead to biased index values. Therefore, the team had to ensure these criteria were met before calculating each index. In cases where such criteria were not met, data was prepared properly to avoid bias. The following section will explain the methods used to identify and treat outliers, severe skewness and severe kurtosis.

Outliers

The value of a variable is considered an outlier if it falls outside the range of the data fence, i.e. an interval with lower and upper bounds calculated based on data location measures (first and third quartiles) and data dispersion measures (interquartile range) as follows:

*Lower bound = first quartile – 1.5 * interquartile range*
*Upper bound = third quartile + 1.5 * interquartile range*

Outliers are treated by replacing them with the highest value lying within the range of the data fence in the case of high values, and the lowest value within the range of data fence in case of low values.

Skewness and Kurtosis

According to international literature, a variable has severe skewness if its absolute skewness coefficient is above 2. An absolute kurtosis coefficient above 3.5 indicates that the variable has severe kurtosis. Variables that are characterized by severe skewness and/or severe kurtosis require statistical treatment before they may be used for calculating the six sectoral indices. The logarithm transformation is among the well-known transformations used in this respect.

By applying the rules for identifying outliers, severe skewness and/or severe kurtosis in the data of the AKI 2016 variables, the team found 66 variables displaying such phenomena. Table 7 shows the distribution of these variables across the six AKI indices.

The maximum number of outliers for any variable was 2, and the treatment of outliers resolved the problem of severe skewness and/or severe kurtosis in all cases without the need for transformation.

Normalization

The rescaling or "maximum–minimum" method was used for normalization, in which the maximum and minimum indicate the largest and smallest of the available variable values. The values of variables were normalized in the range of 1–100, in which the higher values indicated better results. The normalization criterion depends on whether the variable is good (i.e. has a positive relation with the composite index) or bad (i.e. has a negative relation with the composite index).

The good variables were normalized using the following formula:

$$\text{Normalised variable value of the country} = 99 \times \left(\frac{\text{raw variable value of the country} - \text{raw minimum value of the variable across countries}}{\text{raw maximum value of the variable across countries} - \text{raw minimum value of the variable across countries}} \right) + 1$$

In the case of the bad variables (i.e. those with an inversely correlated relation) this formula is adjusted as follows:

$$\text{Normalised variable value of the country} = 99 \times \left(\frac{\text{raw maximum value of the variable across countries} - \text{raw variable value of the country}}{\text{raw maximum value of the variable across countries} - \text{raw minimum value of the variable across countries}} \right) + 1$$

Weights

In general, the AKI 2016 adopted the same methods for estimating weights as the 2015 edition, which range from equal weighting and budget allocation to the factor analysis method. However, some of the weights were modified as a result of changes to the overall structure of the sectoral index due to the addition and/or deletion of variables.

The weights were also statistically estimated for each variable using factor analysis by: first, using the values of one factor for each individual

Table 7:

Frequency distribution of variables with outliers, severe skewness and/or severe kurtosis

Index	Number of variables with outliers, severe skewness and/or severe kurtosis
Economy	10
Higher Education	15
ICT	8
Pre-University Education	12
RDI	17
TVET	4

variable proposed to measure the relevant index; and second, using the values of three factors – rather than one – for the suggested individual variables in order to propose several alternative weights to help researchers determine the ultimate weights of the various variables.

Sectoral index calculation

The AKI 2016 used the most recent and credible data for each of the 22 Arab countries. It applied a series of successive aggregations, starting with the more detailed-level variables and ending with the overall sectoral index.

Owing to the failure to obtain data for all the main pillars for each country, and in light of the desire to maintain an adequate level of accuracy, the sectoral indices were calculated only in cases where data was available for at least two of the main pillars. This applies to all six AKI indices and for all countries. In cases where data for variables was not available for at least three countries, the results of the sub-pillars were not presented.

The arithmetic aggregation formula was used to calculate each composite index of the AKI. Each composite index (CI) is calculated by aggregating its main pillars (SI_j) as follows:

$$CI = \sum_{j=1}^{n} w_j \times SI_j$$

Endnotes

1 Hair et al., 2015.
2 Tavakol and Dennick, 2011.
3 For more information about the data sources of the AKI, refer to the Annex.
4 Cornell University et al., 2015.