A Case Study of Statistics Teaching under the Background of Big Data --Based on Spatial and Temporal Characteristics of Air Quality Index

Xiaolan Lu

School of Business, Jianghan University, Wuhan, China.

luxiaolan@jhun.edu.cn

Abstract

Under the background of big data, the data structure, data source and teaching focus in Statistics are different. Statistics teaching needs to be adjusted and supplemented in data types, data collection methods, data processing and presentation methods, data analysis methods and statistical software applications. The case design of Statistics teaching should reflect the big data collection method, statistical theory knowledge and statistical software operation ability to the greatest extent. In this paper, the spatial and temporal characteristics of air quality index of 379 cities were studied. *Stata* software was used for data collection, pretreatment, basic descriptive statistical analysis, data chart presentation, single-sample mean test, paired sample hypothesis test, variance analysis, correlation analysis, regression analysis and cluster analysis. This teaching case integrates the main knowledge points of Statistics, through which students can apply theory to practice and greatly improve their practical skills of big data.

Keywords

Teaching Cases; Big Data Collection; Data Analysis Method; Stata Application.

1. Introduction

Traditional statistics teaching usually takes questionnaire as the carrier and collects structured data as the starting point to sort out, analyze and infer structured data. The arrival of the era of big data requires students of business administration to collect, clean, mine, process and analyze big data. Under the background of big data, what are the differences in Statistics teaching, how to reflect the characteristics of the era of big data in the design of teaching cases, and how to improve students' big data processing ability are all issues that need urgent attention in the Statistics teaching process.

2. The influence of big data background on Statistics teaching

2.1 Different data structures

The data collected, sorted, analyzed and applied in the traditional statistics teaching is basically structured data. Even if the original data is not structured data, we should either try to convert it into structured data or discard it. Under the background of big data, a large amount of data mainly comes from the network, and it is more likely to be unstructured data such as sound and image.

2.2 Different data sources

The data used in traditional statistics teaching are mainly from primary data obtained from statistical surveys or secondary data from statistical yearbooks and databases. In the context of big data, data is mainly sourced from the network and acquired through web crawlers and other means.

2.3 Different teaching focus

Traditional statistical data are mainly derived from samples, and the learning focus is mainly on how to infer the population with small sample statistics. Therefore, parameter estimation and hypothesis testing are the main contents of learning. In the context of big data, data sets are basically seen as originating from the population rather than samples. Therefore, the focus of learning is no longer

sampling inference and hypothesis testing, but should be shifted to the deep mining of data to discover the rules and characteristics of the development of things.

3. Adjustment of Statistics teaching content in the context of big data

3.1 Unstructured data types

In traditional statistics, data types are classified into qualitative or quantitative data, and measurement scales are classified into four categories, namely, classification, order, distance and ratio. In the background of big data, data types and measurement scales are different, so data should be divided into different types according to different measurement scales, sources and space-time.

3.2 Data collection methods

In traditional statistics, data collection methods mainly include statistical investigation and second-hand data collection. In the era of big data, data are generated in different ways, and the collection of big data requires the use of corresponding software to compile web crawlers and other methods. This part is relatively weak in the traditional Statistics teaching, so it is in urgent need of strengthening.

3.3 Data processing and presentation methods

In traditional Statistics, data processing and presentation are mainly achieved through simple statistical grouping to obtain statistical tables or graphs. Faced with massive unstructured data, it is difficult to process and display the data with traditional statistical grouping, tables or graphs. More complex data mining or data perspective technology teaching should be supplemented, such as exploratory visual description tools, Tableau and lexicographical visualization tools, etc.

3.4 Data analysis method

In traditional Statistics teaching, data collection, data collation, data analysis and other modules are generally included. Among them, data analysis includes sampling inference, hypothesis testing, variance analysis, correlation and regression analysis, cluster analysis and factor analysis, etc. In the background of big data, descriptive statistics should be emphasized, but this part should emphasize the learning of data cleaning and data mining. Sampling inference and hypothesis testing can be weakened. The principles of cluster analysis, correlation analysis, decision tree analysis and artificial neural network are more applicable in the context of big data, and can be supplemented as the main content of statistics teaching.

3.5 Application of statistical software

In traditional Statistics teaching, software such as *Excel*, *Eviews* or *SPSS* is used to process data. In the context of big data, other statistical software such as *R*, *Stata*, or *Python* should be the utilized. By learning the statistical software, students can conduct data collection, data cleaning, data presentation, data analysis and model building, etc.

4. Teaching case design requirements in the context of big data

4.1 Using big data

Stata software was used for data collection. In case teaching, the collection method and data characteristics of dynamic unstructured big data can be illustrated by explaining the collection of data of all vehicle models in the auto announcement of China Commercial Vehicle Network and the collection of location and comment data of "*Villager chicken*" restaurant.

Considering the case teaching in the collected data to be used for subsequent learning, this paper takes the space-time characteristics of Chin's 379 urban air quality indexes as an example. The case data are of great concern to governments at all levels, relevant agencies and the general public, and are released dynamically every 2 hours, featuring large data volume, dynamic rolling, and combination of qualitative and quantitative data ,etc. The data set can be well used to study the main knowledge points in Statistics after matching the data of the corresponding city's social and economic indicators or geographical indicators.

4.2 Reflecting the comprehensiveness of knowledge points

The above data can be used for data collection, preprocessing, basic description statistical analysis, data display, single sample mean test, paired sample hypothesis test, analysis of variance, correlation analysis, regression analysis and cluster analysis.

4.3 Reflecting the practicality of software application

Currently, *python* and *Stata* are popular software for the collection, processing and analysis of big data. Based on the actual situation of students, this case uses *Stata* software for data collection, processing and data analysis, and completes all case teaching with *SPSS* software.

5. Teaching case design process

5.1 Data collecting

Taking Beijing's air quality index as an example, through China's air quality index website (*http://www.pm25.in/*), the URL "*http://www.pm25.in/beijing*" is obtained. If you want to get all 379 cities of *AQI, PM25, PM10, CO,* and *O3* index, you must firstly obtain corresponding URL of 379 cities. Therefore, there are two steps in the design of web crawler. The first step is to use the "*copy*" code to copy and save the page as *TXT* format file from the website of the first page. Then, the "*infix*" code is used to read the file into *Stata*. Then, using "*Keep if index (V1, "<A href="/")* " to get the Chinese pinyin names of all cities and save the file as "*Cityname.dta*". The second step is to use the code "*levels of CityName*" and "*foreach v of Local Levels {*" to loop the URLS of all 375 cities, locate and clean the data, and use "*postfile*" to output the obtained data of *AQI, PM25_1h, PM10_1h, CO_1h, NO2_1h, O3_1h, O3_8h* and save them as "*Cityaqi01.dta*" file for further use. Since the air quality index is published every two hours, *Stata* can be used to crawl the data every two hours and save the data as "*Cityaqi02.dta*" and "*Cityaqi03.dta*", and so on.

5.2 Data preprocessing

The data files at different time points above are merged into one file by "*merge*", and the data of 31 provincial capitals or municipalities are retained. Then the air quality index of each provincial capital (municipality) is matched with the regional economic indicators and other data by "*merge*". According to AQI, the air quality of each city is divided into five air quality grades: excellent, good, medium, qualified and unqualified. The quantitative data are centralized or standardized and stored, and the singular value situations that may exist are screened. Prepare for subsequent statistical analysis.

5.3 Data presentation

Frequency analysis is carried out on the acquired indicators such as *AQI, PM25_1h, PM10_1h, CO_1h, NO2_1h, O3_1h, O3_8h,* etc. Histogram, density histogram or boxplot are drawn on the quantitative indicator data using the codes "*graph*", "*histogram*" and "*scatter*" to observe the distribution of data. In addition, using the "*spmap*" code and the map of China, the display is conducted according to five different colors of air quality grading, such as excellent, good, medium, qualified and unqualified, so as to visually observe the air quality changes, spatial differences and space-related conditions at different time points.

5.4 Basic descriptive statistics

"*Tabstat*" and "sum" codes were used to basically describe statistical analysis of *AQI, PM25_1h, PM10_1h, CO_1h, NO2_1h, O3_1h, O3_8h* and other indicators of air quality in 379 cities, and the average, median, mode, maximum, minimum and standard difference were obtained, and outliers and singular values were analyzed. The above basic descriptive statistics are reported in the form of statistical tables.

5.5 Hypothesis testing of the mean of single sample

"*Ttest*" is used to test the mean value of air quality indicators such as *AQI, PM25_1h, PM10_1h, CO_1h, NO2_1h, O3_1h, O3_8h* in 379 cities, to test whether the original hypothesis is true.

5.6 Hypothesis testing of paired samples

"Ttest" is used to test whether the air quality index, such as *AQI*, *PM25_1h*, *PM10_1h*, *CO_1h*, *NO2_1h*, *O3_1h*, *O3_8h* and so on, are significantly different from one another different time.

5.7 Correlation analysis

Using "*pwcorr*", the correlation between *AQI*, *PM25_1h*, *PM10_1h*, *CO_1h*, *NO2_1h*, *O3_1h*, *O3_8h* and other air quality indicators in 379 cities and the economic values of each region was analyzed. The corresponding correlation coefficient matrix was obtained and the significance of the correlation was analyzed.

5.8 Regression analysis

Using "*reg*", the linear regression equation was established with per capita GDP as the dependent variable and *AQI*, *PM25_1h*, *PM10_1h*, *CO_1h*, *NO2_1h*, *O3_1h*, *O3_8h* as the independent variable respectively. The multicollinearity, heteroscedasticity and autocorrelation are tested for the obtained model. Then, a more robust regression model is established, and according to the robust regression model, the per capita GDP of a specific region at a specific time is predicted.

5.9 Cluster analysis

According to *AQI*, *PM25_1h*, *PM10_1h*, *CO_1h*, *NO2_1h*, *O3_1h*, *O3_8h* and other air quality indexes, the air quality situation of 379 cities was clustered by the code "*cluster*", mainly to understand which cities can be classified into one category, so as to further analyze the air quality category.

6. Teaching cases regurgitation

This comprehensive teaching case not only reflects the new requirements of statistics teaching in the background of big data, but also runs through the key theoretical and practical knowledge in traditional Statistics teaching, with the characteristics of comprehensive, practical and innovative. Through the study of this case, students can master the basic links such as big data collection, processing, presentation and analysis, laying a solid foundation for further mastering the comprehensive analysis ability of big data.

In addition, after learning the teaching case, students are required to independently design a comprehensive experiment around the experimental purpose and complete a comprehensive statistical experiment. For example, the "*Top200 data of Douban films*" was crawled to obtain the *Douban* score, number of critics, director, leading actor, film type, release year and film evaluation of the Top200 films on *Douban* website and establish a database. Using this data, descriptive statistics, data display, word cloud map, data mining, hypothesis testing, correlation analysis, regression analysis, cluster analysis and box office forecast were conducted.

References

[1] Han Jingshu.2019.Research on Teaching Methods of "Statistics" Course in the Context of Big Data[J], Education and Teaching Forum, No.9.

[2] Liu Surong.2018.Research on the Teaching Reform of Statistics Course in Economics and Management Major under Big Data Thinking[J], Journal of Higher Education, No.10.

[3] Zhu Jianping, Zhang Guijun, Liu Xiaowei. 2014.Analysis of Data Analysis Concept in the Era of Big Data[J], Statistical Research, No.2.

[4] Zhang Ying.2019.Teaching Research of Statistics Courses for Economics and Management Majors in the Era of Big Data[J], Modern Educational Technology, No.4.

[5] Zeng Wuyi. 2019.Interpretation of National Standards for Teaching Quality of Statistics Majors [J]. Chinese University Teaching, No.11.

[6] Sun Xin, Yin Biao. 2017.The Response of Statistics Subject in the Era of Big Data[J]. Statistics and Decision-making, No.6.