ISSN: 2710-0170

Experimental Project Design of Statistics for Non Statistical Majors under the Background of "New Economy and Management"

Anning Ye*, Min Zhang

School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, China

*yeanning@sohu.com

Abstract

Under the development strategy of "new economy and management", many financial institutions focus on the deep integration of the new generation of information technology and education and teaching, and information technology teaching has been paid attention to. The new training program requires students to be able to analyze large-scale data and have strong data analysis ability on the basis of mastering a programming language. The establishment of cases in line with the integration of theoretical teaching and programming teaching can cultivate students' interest in learning statistics and improve the teaching quality of statistics. In this paper, the experimental teaching is divided into three parts: The first part is to familiarize students with basic commands.The second part is to simplify the calculation. The third part is to let students understand the basic theory of statistics, which not only helps students improve Gauss programming ability, but also helps students understand the depth and breadth of statistical knowledge.

Keywords

New Economy and Management; Statistics; Programming Ability; Experimental Project.

1. Problems in "Statistics" Experimental Teaching under the New Economic Management Strategy

Under the development strategy of "new economy and management", it is a necessary ability for students to master a programming language. At present, colleges and universities of finance and economics focus on the deep integration of the new generation of information technology with economic society and education, and put forward higher requirements for information technology teaching. The whole course group of Anhui University of Finance and Economics has built a number of computer courses, and the credit of information technology has been increased from 5 points to 12 points. The overall idea of these courses is that on the basis of students' gradual mastery of computer basic knowledge and operation, taking EXCEL data analysis as the starting point, they are familiar with SQL language, and on the basis of mastering a programming language, they can analyze large-scale data, and have a strong ability of students' data analysis [1].

Researchers have different views on the status of software in statistics teaching. Contents related to computers or computer software are generally not included in textbooks [2]. The compilation of statistics textbooks should be combined with the application of modern teaching methods and statistical software packages, and the specific operation methods should be listed after the corresponding chapters [3]. Due to the limitations of class hours, teachers and other conditions, many computer applications in statistics have been omitted in the statistics teaching of non-statistical majors in finance and economics [4].

ISSN: 2710-0170

In order to meet the needs of statistics teaching in the new situation, our school has now opened a programming language course to learn big data courses. However, there is a phenomenon that programming courses are disconnected from the teaching of statistics theory in the curriculum, which has affected the teaching effect. Part of the course teaches the basic operation of data, which is more repetitive with the statistics course, and less teaches the statistical theory behind the operation. Another part of the software course includes supervised learning and unsupervised learning [5]. Students generally report that it is difficult to connect with the previous teaching content, difficult to accept, and poor learning effect.

Previous experimental research often focused on single or specific experiments, such as the distribution limit form of sample mean [6], hypothesis test, regression analysis, etc. [7]. Because of the single content of the experiment, these experiments have limited effects on comprehensively mastering the basic theory of statistics or improving students' statistical application ability. Of course, there are also some special statistical experimental course materials on the textbook market, but they are often experimental courses for students majoring in statistics, which are not suitable for the teaching of non-statistical students [8-9]. Moreover, most of the textbooks use EXCEL or SPSS software, and use menu-based operation, which is not flexible enough.

Based on the above documents, it can be seen that many achievements have been made in the research and teaching of software language courses in colleges and universities, but there are the following shortcomings. First, experimental research is often designed according to a specific problem, such as the sampling distribution of sample mean, hypothesis testing, etc. Secondly, the programming course offered is not fully connected with the previous statistical teaching content. Thirdly, most of the statistical experiment courses use EXCEL or SPSS software, which is inflexible. Therefore, according to the experimental design system of statistics teaching, this paper will establish a case that is consistent with the integration of theoretical teaching and programming teaching, aiming at solving the problem of the disconnection between the programming course and the theoretical course teaching offered by our school. This will not only help students improve the ability of GAUSS programming, but also help students understand the depth and breadth of statistical knowledge.

In order to better connect the experimental teaching and statistical theory teaching, the whole experimental design is divided into three parts: The first part is to familiarize students with basic commands.The second part is to simplify the calculation. The third part is to let students understand the basic theory of statistics. The examples in this paper are taken from the textbook [10]. Due to the space limitation, most of the experimental items in this paper do not list the specific GAUSS program.

2. Arrangement of Experiment Content

2.1. Be Familiar with Basic Programming Commands

In GAUSS software, there are more than 1000 built-in functions, and beginners do not need to understand each of them. In this kind of experiment, there are about 10 commands commonly used in data collation and data feature description. With these simple commands, students can realize the interaction between textbooks and computers. They not only master these gauss commands, but also deepen their understanding of data collation, data feature description, etc. in statistics.

Data sorting specifically includes importing data, sorting, grouping, summarizing, drawing stem and leaf graph, and data output. Import data by direct input, "load" command and "xlsreadm" command. Sorting adopts "sortc" command mode; Grouping can be completed by nesting “for” loop statements and “if” conditional statements. The "for" loop statement, "if" conditional statement, "ftocv" command and "strput" command should be used to make stem and leaf

ISSN: 2710-0170

diagrams. Use the "print" command for data output. Take the generation of stem and leaf map as an example, and the procedure is as follows: (page 27 of the textbook)

```
Data={117,122,124,129,139,107,130,125,139,119,108,131,125,122,133,126,118,108,128,124
,110,123,126,133,134,120,123,118,112,121,112,134,120,127,135,137,114,124,115,128};
n=rows(data);
nleaf1=0;
nleaf2=0;
nleaf3=0;
nleaf4=0;
nleafother=0;
leaf1=0;
leaf2=0;
leaf3=0;
leaf4=0;
leafother=0;
fori(1,n,1);
   if  (100<=data[i]) and (data[i]<=109);
      nleaf1=nleaf1+1;
      leaf11=data[i]-100;
      leaf1=leaf1|leaf11;
   elseif (110<=data[i]) and (data[i]<=119);
      nleaf2=nleaf2+1;
      leaf21=data[i]-110;
      leaf2=leaf2|leaf21;
   elseif (120<=data[i]) and (data[i]<=129);
      nleaf3=nleaf3+1;
      leaf31=data[i]-120;
      leaf3=leaf3|leaf31;
   elseif (130<=data[i]) and (data[i]<=139);
      nleaf4=nleaf4+1;
      leaf41=data[i]-130;
      leaf4=leaf4|leaf41;
   else;
      nleafother=nleafother+1;
   endif;
endfor;
leaf1=leaf1[2:nleaf1+1];
leaf2=leaf2[2:nleaf2+1];
leaf3=leaf3[2:nleaf3+1];
leaf4=leaf4[2:nleaf4+1];
leaf1=ftocv((sortc(leaf1,1)),1,0);
leaf2=ftocv((sortc(leaf2,1)),1,0);
leaf3=ftocv((sortc(leaf3,1)),1,0);
leaf4=ftocv((sortc(leaf4,1)),1,0);
```

ISSN: 2710-0170

```
lleaf1="";
fori(1,nleaf1,1);
   lleaf1=strput(leaf1[i],lleaf1,i);
endfor;
   print $lleaf1;
   lleaf2="";
fori(1,nleaf2,1);
   lleaf2=strput(leaf2[i],lleaf2,i);
endfor;
   print $lleaf2;
   lleaf3="";
fori(1,nleaf3,1);
   lleaf3=strput(leaf3[i],lleaf3,i);
endfor;
   print $lleaf3;
   lleaf4="";
fori(1,nleaf4,1);
   lleaf4=strput(leaf4[i],lleaf4,i);
endfor;
   print $lleaf4;
```

The print result is:

788
022457889
001223344455667889
0133445799

Data feature description includes mean, standard deviation, median, mode, skewness and kurtosis. Some of these descriptive statistics have direct commands, such as arithmetic mean "meanc" and standard deviation "stdc". The mode and median needs to be calculated by programming. The programming results are compared with the "mode" and "quartile" commands in EXCEL. Indicators such as skewness and kurtosis also need to be programmed and calculated, and the calculation results and data distribution patterns should be compared.

Simple correlation coefficient measures the relationship between two variables and is a powerful tool for studying the relationship between phenomena. The formula of the correlation coefficient has several variations, which should be kept in mind. The corresponding gauss command is "corrx". Note that the output result is 2 × 2 matrix, correlation coefficient is non-diagonal element.

After you are familiar with some basic GAUSS commands, you can perform some more complex calculations in statistics. This complexity is mainly for beginners, especially non-statistical students, because it is generally difficult for beginners to fully understand the meaning of formulas.

2.2. Simplified Calculation

An important role of using statistical software is to calculate. Specifically, it includes descriptive statistics of grouping data, parameter estimation, hypothesis test, analysis of variance, regression analysis (univariate), traditional time series analysis, and index analysis. Most of these calculation programs are relatively simple. As long as you understand the formula in the

ISSN: 2710-0170

book, you can write programs according to the formula, which is not very difficult for students to master.

The descriptive statistics of grouping data are not provided by general commercial software, and they need to be programmed and calculated by us. The difficulty of programming is not much higher than that of non-grouping data. The parameter estimation of the population mainly estimates the mean and standard deviation of the population in the form of point estimation and interval estimation. Pay attention to the difference between the two when programming. Hypothesis test and interval estimation are closely related. We should pay attention to the connection and difference between these two parts. The procedure of variance analysis is relatively complex, because it involves data grouping, intra-group variance and inter-group variance. It is not easy for beginners to write such a procedure.

The calculation of univariate regression analysis mainly includes parameter estimation, hypothesis test and prediction. Parameter estimation is to estimate the intercept and slope of the sample regression line, and estimate the slope first and then the intercept. The hypothesis test part of the regression model includes the calculation of R^2, t test and F test. Prediction includes point prediction and interval prediction. As a beginner, it is recommended to write the procedure of point estimation, R^2 and point prediction for parameter estimation. The interval estimation, t test, F test and prediction interval estimation of regression model parameters are only required for students who have spare time.

The factor decomposition method decomposes the time series into trend factors, seasonal factors, cyclical changes and irregular changes. In our syllabus, cyclic change is a difficult part, and students are not required to master it. In calculating the average index, the commonly used methods are the level method and the equation method, while the equation method needs to solve the higher order equation. The following example is to solve the average development speed, which is taken from the textbook (Page 185).

```
use gpe2;
data1={413030.3,452429.9,487976.2,525835.4,564194.4,603212.1};
t=rows(data1)-1;
fn f(data,x)=(x^1+x^2+x^3+x^4+x^5-sumc(data1[2:t+1])/data1[1])^2;
call reset;
_nplot=0;
_iter=1000;
_b=1.1;
call estimate(&f,0);
end;
```

The result is 1.0787, which is consistent with the calculation result using EXCEL in the textbook.

Index analysis is a part that students can easily accept. Its content mainly includes comprehensive index, average index and factor decomposition of index. Compared with other parts of statistics, the calculation of this part is easier to achieve. The calculation of the price index and volume index needs to select the same measurement factor. The period of the same measurement factor of the price index is the base period, while the period of the same measurement factor of the volume index is the reporting period. Factor decomposition of index includes both absolute factor decomposition and relative factor decomposition.

ISSN: 2710-0170

3. Computer Simulation to Help Students Understand the Basic Theory of Statistics

3.1. Type of Frequency Distribution

The types of frequency distribution include bell distribution, U distribution and J distribution, which can be generated by using random numbers of various distributions. The bell distribution can be generated by using normal distribution random number, and the U-shaped distribution and J-shaped distribution can be generated by using beta distribution. These randomly generated number distributions can be used to intuitively observe the three distribution types.

3.2. Descriptive Statistics

In descriptive statistics, there are three parts that are difficult to understand: skewness, kurtosis, and the relationship between the three concentration trends (the relationship between the median, mode, and arithmetic mean). In order to study skewness, first assume that the first parameter of the two parameters of the beta distribution is 1, the second parameter gradually changes from 1 to 100, and each step is 1. We simulate these situations, and find that the skewness coefficient is getting bigger and bigger. It is verified that the skewness coefficient calculated is getting bigger as the density function is skewed to the right. As for kurtosis, take the beta distribution as an example. Assume that the two parameters are the same, gradually changing from 0.1 to 100, with each step of 0.1, and the density function gradually changing from concave to convex. We simulated these situations and found that the kurtosis coefficient is getting larger and larger. It is verified that the steeper the kurtosis, the larger the kurtosis coefficient. The arithmetic mean, median and mode can be calculated by using the beta distribution on the right. From the simulation results: $\bar{x} - M_o \approx 3(\bar{x} - M_e)$. Here,$\bar{x}$ is the mean, M_o is the mode, and M_e is the median.

3.3. Limit Distribution of Sample Mean

In computer simulation, we can assume that the overall distribution is uniform, v distribution, L distribution, sample size n=2, 5, 30, and sample number 10000. From the simulation results, when n=2, the sample mean does not present a normal distribution. When n=30, the sample mean is approximately normal distribution.

3.4. Expected Value and Variance of Sample Mean

In the theory of sampling distribution, there are the following conclusions: $E(\bar{X}) = \mu$, $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$.

Assume that population={1,2,3,4}, N=4,sample size n=2, through all samples, there are 16 samples in total. It can be calculated that μ =2.5, σ^2 =1.25, $E(\bar{X}) = 2.5$, $\sigma_{\bar{x}}^2 = E(s^2)$ =0.625 $= \frac{1.25}{2} = \frac{\sigma^2}{n}$.

3.5. Simple Random Sampling, Stratified Sampling and Cluster Sampling

Assume that population = {1, 2, 3, 5, 6, 7}, sample size n = 4, we compare the error of simple random sampling and stratified sampling. Under the condition of data grouping, The total variance can be decomposed into inter-group variance and intra-group variance, that is written as $\sigma^2 = \delta^2 + \overline{\sigma_i^2}$. The population is divided into two sub-populations, Sub-population 1 = {1, 2, 3},Sub-population 2 = {5, 6, 7}. The overall variance is σ^2 =4.667, sampling error of simple random sampling $\sigma_{\bar{x}}^2 = 1.1667 = \frac{\sigma^2}{n} = \frac{4.667}{4}$, the error of stratified sampling $\sigma_{\bar{x}\,str}^2 = \frac{\overline{\sigma_i^2}}{n} = \frac{0.667}{4} =$ 0.1667. The error of stratified sampling is small.

ISSN: 2710-0170

The ideal division of layers in stratified sampling. Assume that population={1, 2, 3, 5, 6, 7}. Take any three as one group, and the remaining three as another group, there are 10 types of division ($C_6^3 / 2$). For each division, the inter-group variance and intra-group variance can be calculated, and the results are shown in Table 1. The layer division of No. 1 is the ideal division of stratified sampling. This division divides those with relatively close characteristics into one layer and separates those with large differences.

The ideal division of the group in cluster sampling. Assume that population = {1, 2, 3, 5, 6, 7}, Divided into 3 groups, there are 15 types of division ($C_6^2 C_4^2 / (3*2*1)$) .That the division {{2, 6}, {1, 7}, {3, 5}} is one of them with the mean values of the three groups are the same. Take two groups (r=2) from three groups (R=3) to form a sample, and the inter-group variance is 0.According to the error formula of cluster sampling, the sampling error is also 0. This division is the most ideal one. In cluster sampling, the intra-group variance should be expanded as much as possible and the inter-group square should be reduced.

Table 1. Inter-group variance and intra-group variance of stratified sampling

Layered serial number	Layered	σ^2	δ^2	$\overline{\sigma_i^2}$
1	{1,2,3},{5,6,7}	4.667	4.000	0.667
2	{1,2,5},{3,6,7}	4.667	1.778	2.889
3	{1,2,6},{3,5,7}	4.667	1.000	3.667
4	{1,2,7},{3,5,6}	4.667	0.444	4.222
5	{1,3,5},{2,6,7}	4.667	1.000	3.667
6	{1,3,6},{2,5,7}	4.667	0.444	4.222
7	{1,3,7},{2,5,6}	4.667	0.111	4.556
8	{1,5,6},{2,3,7}	4.667	0.000	4.667
9	{1,5,7},{2,3,6}	4.667	0.111	4.556
10	{1,6,7},{2,3,5}	4.667	0.444	4.222

3.6. Hypothesis Test and Interval Estimation

A simulation that makes the first type of error. Generate 10000 * 20 random numbers with standard normal distribution (sample size is 20).Original assumption H0: μ =0. Calculate 10000 t values respectively. If $|t|>t_{a/2}$, then make the mistake of abandoning the truth (making the first kind of mistake). The simulation results show that there are 488 samples $|t|>t_{a/2}$. If a=0.05 is changed to a=0.01, the result is 85. That is, the smaller the confidence level a, the smaller the probability of making the first type of error. It is similar to the simulation of making type 2 errors and the simulation of the confidence interval of the total mean value containing 0.

3.7. Hypothesis Test of Correlation Coefficient

There are two hypothesis tests for correlation coefficient. The first is to test whether the correlation coefficient is 0. Assume that x and y are random variables (20 elements), randomly select n=10 pairs of data, calculate the correlation coefficient, and see the distribution of the correlation coefficients. At this time, it can be observed that r approximately obeys the bell distribution, constructs t(n-2) distribution, and can test H0: r=0. The second test is whether the correlation coefficient is r_0 ($r_0 \neq 0$). Assume x={1,2,...,20},y=0.5x + random variable, randomly select n=10 pairs of data, calculate the correlation coefficient, and see the distribution of the correlation coefficient. From its distribution, it is leftward distribution, and it is converted into 0.5ln((1+r)/(1-r)),It is approximately normal distribution. At this time, construct a standard normal distribution, which can test H0: $r=r_0$ ($r_0 \neq 0$).

ISSN: 2710-0170

3.8. Gauss-Markov Theorem in Regression

Use software to simulate Gauss-Markov theorem, that is, the least squares estimate is the minimum unbiased estimator.Assume that the actual model is $Y_i = 7 + 0.6X_i + \varepsilon_i$. When the sample size n=20, 1000 samples are randomly generated to obtain $\hat{a}$ and $\hat{b}$, the corresponding histogram is shown in Figure 1 and Figure 2, and the unbiased can be verified. Similarly, the minimum variance and asymptotic behavior can also be simulated.

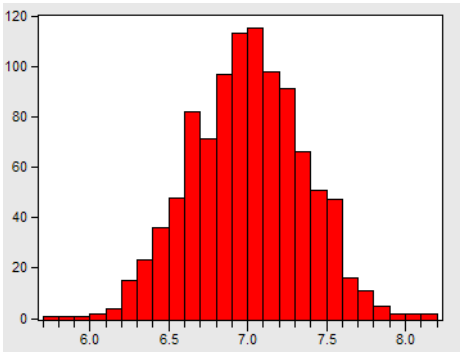

Figure 1. Histogram of $\hat{a}$

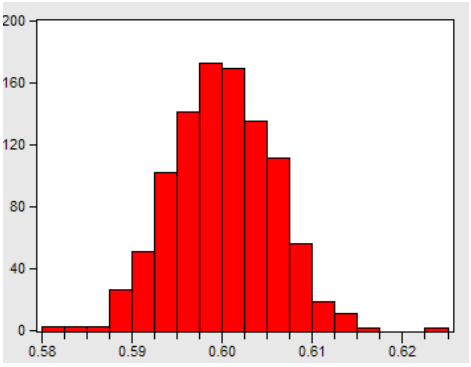

Figure 2. Histogram of $\hat{b}$

3.9. Non-uniqueness of Exponential Decomposition

Understand the non-uniqueness of exponential decomposition. In terms of the period selection of the same measurement factor, there are other options besides the ones specified in the textbook: the period of the same measurement factor of the price index is the reporting period, and the period of the same measurement factor of the volume index is the base period. There are two methods for the index decomposition of the two factors. There are six methods for exponential decomposition of three factors (3!=6).

4. Precautions

4.1. The Program Cannot be too Complex

One of the important principles in programming languages is to simplify the code, use the software built-in functions as much as possible, and ensure the elegance and conciseness of the code. It is not recommended to make the program too complex. On the one hand, it is not conducive to students' understanding of statistical theory. On the other hand, it makes it more difficult to read the program, which easily dampens the enthusiasm of students to learn programming.

4.2. The Content of Experimental Design Should be Classified

The experiment design is divided into the following three parts according to the content: (1) This kind of experiment makes students familiar with some basic GAUSS commands. (2) Simplify calculation operation. (3) Help students understand the basic theory of statistics. The teacher should pay attention to the choice according to the characteristics of the teaching object in the explanation process.

4.3. The Experiment Content Should be Closely Related to the Teaching Material

The experimental content is directly taken from the examples in the textbook (the textbook uses EXCEL calculation). Our programming method will greatly simplify the calculation, which will greatly reduce the calculation work of students, and also help students understand the basic theory of statistics.

ISSN: 2710-0170

4.4. Most of the Contents Can be Mutually Verified by Excel Operation and GAUSS Programming Operation

As a primary statistics, most cases of statistics can be operated with EXCEL. If you can use EXCEL or programming to do it. When the two confirm each other, students will enhance their confidence in programming to solve problems and stimulate their enthusiasm for learning statistical programming.

Acknowledgments

This work is supported by teaching research project of Anhui University of Finance and Economics, "*Design of "Statistics" experiment project for non-statistics majors*" (Grant No: acjyzd2021027).

References

[1] Wu,L., Wang, H., Zhou, J., Duan, A. Discussion on the construction of data analysis course group in finance and economics universities -- Taking the new management strategy of Anhui University of Finance and Economics as an example[J]. Journal of Langfang Normal University (Natural Science Edition), 2020,(03):114-116.

[2] Huang, S. Where is the course of statistics for non-statistical majors in finance and economics [J]. Statistics and Decision, 2001, (12): 16-17.

[3] Xu, J. Discussion on the teaching method of the course "Statistics" for financial majors [J]. Statistical Education, 2001, (03): 24-26.

[4] Hu, R. Discussion on the construction of statistics curriculum for non-statistical majors in finance and economics [J]. Statistical Education, 2002, (03): 25-27.

[5] Wu, X., Liu, M. Introduction to data science - R and Python implementation [M]. Beijing: Higher Education Press, 2019.

[6] Wu, Z., Chen, Y., Liu,Y. Teaching experiment design of "statistical method" course based on R language [J]. The Science Education Article Collects, 2021, (28): 111-114.

[7] Wang, Z., Jin, Z. Application of matlab's statistical toolbox in statistics teaching [J]. Science and Education Guide, 2017, (36): 97-98.

[8] Qiu, Y. Statistical experiment course [M]. Beijing: Beijing University Press, 2012.

[9] Zhang, H., Nie, T. Statistical experiment course [M]. Beijing: Social Science Literature Press, 2013.

[10] Zhang, H. Statistics [M]. Shanghai: Shanghai Jiaotong University Press, 2019.