Practice of Analysis of Variance and Multiple Comparison Using Free Software 'R'

フリーソフトウェア 'R' を使った分散分析と多重比較の実践

川口雄一
Yuuichi KAWAGUCHI

The steps of statistical processing using programming tools are described. In this paper, a programming tool named 'Perl' is used for formatting sample data and a programming tool named 'R' is used for statistical processing. A sample is given as an example and concrete processing steps are shown. Each statistical method has a premise. A way for checking whether the premise is satisfied or not is described.

プログラムツールを利用して統計処理を行う場合の手法を示す。データの前処理には「Perl」を用い、実際の統計処理には「R」を利用する。例としてある標本群が与えられ、これに対する具体的な処理の方法を示す。統計処理としては「分散分析」と「多重比較」を使う。それぞれに前提として満たすべき条件があり、これらを調べる方法も含めて示す。

Key words: analysis of variance (分散分析)
multiple comparison (多重比較)
perl (パール)
R (アール)

Liberal Arts and Sciences, School of Nursing and Nutrition, Tenshi College

(Received 21 January 2008／Accepted 16 May 2008)

I. Introduction

This is a survey of statistical processing by using programming tools. The original data are given as an electronic file in CSV format. For applying those data to a statistical application program, they must be arranged in a format that can be used in the application.

In this paper, a programming tool named 'Perl' [WCS96] is used for formatting data and 'R' [R D07] is used for statistical processing.

Let us assume that there are samples of concentrations of some material in blood. Those samples are statistically processed in this paper. The original data are partly shown in Table 1.

Table 1 Original Data

Num.,	1st.,	2nd.,	3rd.
10,	80.09,	97.14,	65.79
12,	70.84,	72.41,	67.94
16,	72.38,	93.71,	85.12
23,	90.76,	76.07,	81.09
25,	99.81,	73.36,	70.87
…	…	…	…

The first line is a header. From the second line, each line shows a person. A number of each person is shown in the column named 'Num,' and three concentrations are shown in the '1st.,' '2nd.,' and '3rd.' columns.

Each sample consists of data shown in each of the colums named '1st.,' '2nd.' and '3rd.' Thus, there are three samples shown in 1.

In this case, the purpose of statistical processing is to investigate if the averages of the three samples are different. For the purpose, the statistical method 'Analysis of Variance' (AOV) and 'Multiple Comparison' [NY97], [II07] are used.

Use of the method 'AOV' enables determination of whether all samples have the same averages or not, and use of the method 'Multiple Comparison' enables determination of which two samples differ in averages.

II. Cases when AOV and Multiple Comparison are Required

When two samples are compared in arithmetic averages, the t-test is used. When there are three or more samples, iterative applications of the t-test for two samples increase the rate of error [NY97].

To avoid such error, the statistical method 'Analysis of Variance' (AOV) and 'Multiple Comparison' are used. Those two methods compare some samples one time, and therefore do not increase the rate of error.

The use of AOV enables determination of whether there is a significant difference in averages of all samples or not, and the use of Multiple Comparison enables determination of which one sample is different from other samples in averages.

III. Arrangement of Data

'R' is a programming tool for statistical processing. It requires data to be arranged in a format. Each value of concentration must be recorded in one line.

The original data shown in Table 1 are arranged in another format, in which one key ('Num.') and three values ('1st.,' '2nd.' and '3rd.') are recorded in one line.

To meet the requirement, the table must be rearranged. The programming tool 'R' can read data and can arrange the data in the required format.

The author uses a programming tool named 'Perl' [WCS96], which is an easy language to learn and use for formatting texts. The source code for arranging is shown in Figure 1.

The data arranged in the format required by 'R' are partly shown in Table 2.

Those lines are read by 'R' by a procedure '**`read.table.`**' Data are read into a variable

```
$line = <STDIN>;
print "No, Num, Val, Tim\n";
#
$n = 1;
while ($line = <STDIN>) {
chop($line);
@f = split(/,/, $line);
print $n\t$f[0],\t$f[1],\1st\n";
print $n\t$f[0],\t$f[2],\2nd\n";
print $n\t$f[0],\t$f[3],\3rd\n";
}
exit(0);
```

Figure 1 Code for Perl

Table 2 Arranged Data

No,	Num,	Val,	Tim
1,	10,	80.09,	1st
2,	10,	97.14,	2nd
3,	10,	65.79,	3rd
4,	12,	70.84,	1st
5,	12,	72.41,	2nd
…	…	…	…

named '**t**,' by putting a sequence such as

```
> t <- read.table("file.csv",
      header=True);
```

into a command line of 'R.'

The option '**header=True**' tells 'R' that the first line is a header line. This is a default behavior of 'R.'

IV. Premises

To use AOV and 'Multiple Comparison,' there are some premises for data in samples.

1. Quantitative or Qualitative
   If the data are quantitative, parametric methods for the arithmetic average are used, and if the data do not have normality, non-parametric methods for the median are used.
2. Normality
   If the data are quantitative and have normality, parametric methods are used. Otherwise, non-parametric methods are used.
3. Uniformity of variances (*i.e.*, homogeneity)
   This is the same as the case of normality.

1. Quantitative or Qualitative

Samples used in this paper, which are shown in Table 2, are quantitative.

2. Normality

To determine whether samples have normality or not, the procedure named '**shapiro.test**' is used in 'R.' For example, the sample of '1st' is tested by a sequence

```
> shapiro.test(x[x$Tim=="1st"]).
```

such as

The result for this command is shown in a p-value. For example, let us suppose that the p-value is '**4.713e-05**' and that the level of signifiance α is 0.05 (= 5%). In this case, the p-value is less than the level of significance. The null hypothesis H_0 is that the sample has normality, and H_0 is is rejected in this example. The sample of '1st' does not have normality.

3. Homogeneity

In fact, samples used in this paper are paired, and there is thererfore no need to detect uniformity of variances.

V. Analysis of Variance

In the samples shown in Table 2, data are repeatedly measured three times for one 'Num.' Those are paired. In this case, the method named 'repeated measures AOV' is used. This is the same as a two-way layout AOV. The purpose of this case is to determine whether there is a significant difference for the factor 'Tim.' in averages of all samples or not.

1. Parametric Method

If a sample has normality, then parametric methods are used for AOV. The parametric AOV method is implemented by a procedure named '**aov**' in 'R.' By putting a sequence such as

```
> summary(aov(Val~Num+Tim, data=t))
```

into a command line of 'R,' we get results such as those shown in Table 3.

Table 3 Results of Parametric AOV

```
        ...   F value      Pr(>F)
Num.    ...   2.8563     5.079e-06 ***
Tim.    ...   8.8083     0.0003035 ***
...
```

The sign '***' in Table 3 shows that there is a significant difference in that factor. In this case, the sign is marked for both factors 'Num.' and 'Tim.,' and there is therefore a significant difference for the factor 'Tim.' in averages of all samples.

2. Non-Parametric Method

If a sample does not have normality, then non-parametric methods are used for AOV. The method is implemented in 'R' by a procedure named '**friedman.test.**'

By putting a sequence such as

```
> m <- matrix(c(
       t$Val[t$Tim==‘1st’],
       t$Val[t$Tim==‘2nd’],
       t$Val[t$Tim==‘3rd’]),
ncol=length(t$Val[t$Tim==‘1st’]),
       byrow = T)
> friedman.test(m)
```

into a command line of 'R,' we get results such as those shown in Table 4.

Table 4 Results of Friedman's Test

```
data: m
Friedman chi-squared = ...,
df = ..., p-value = 0.001651
```

The p-value **0.001651** is smaller than the level of significance $\alpha = 0.05$ (= 5%), and there is therefore a significant difference in averages for the factor 'Tim.'

VI. Multiple Comparison

1. Parametric Method

There is a procedure named '**TukeyHSD**' in ' R.' The procedure performs the Tukey's HSD (Honestly Significant Difference) method.

By putting a sequence such as

```
> TukeyHSD(aov(Val~Num+Tim,
                 data=t), "Tim")
```

into a command line of 'R,' we get results such as shown those in Table 5.

Table 5 Results of Tukey's HSD Method

```
 $Tim
          ...      p adj
2nd-1st ... 0.2521688
3rd-1st ... 0.0002006
3rd-2nd ... 0.0315814
```

At the line named '3rd-1st,' the p-value (p adj) **0.0002006** is smaller than $\alpha = 0.05$ (= 5%), and there is therefore a significant difference for the pair of samples '1st' and '3rd.'

2. Non-Parametric Method

There are a few different methods for multiple comparison. In this paper, the method named 'Bonferroni's correction' is used for the non-parametric multiple comparison. When the entire level of significance is α (*e.g.*, = 5%) and the statistical test for two samples is repeated k (*e.g.*, = 3) times, then the level of significance is $\alpha \div k$ (= 0.017 = 1.7%) for each test.

For a statistical test for a pair of samples, we use Wilcoxon's method. The method is implemented in 'R' by a procedure named '**wilcox.test.**' By putting a sequence such as

```
> x <- t$Val[t$Tim==’ 1st’ ]
> y <- t$Val[t$Tim==’ 2nd’ ]
> wilcox.test(x, y, paired = T)
```

into a command line of ‘R,’ we get results such as shown those in Table 6.

Table 6 Results of Wilcoxon’s Test

```
data: x and y
V = …, p-value = 0.1321
```

The p-value **0.1321** is greater than $\alpha \div k=0.017$, and there is therefore not a significant difference for the pair ‘1st’ and ‘2nd.’

VII. Conclusion

In this paper, statistical methods for ‘AOV’ and ‘Multiple Comparison’ by using programming tools are described. A programming tool named ‘Perl’ is used for formatting original data, and ‘R’ is used for statistical processing. ‘Perl’ is efficient for processing textual data. Many statistical methods are implemented in ‘R.’

Samples are given for example, and steps of statistical processing for them are described.

Acknowledgements

The author thanks a colleague T. Yamada, who gave the author the opportunity to study statistical methods, and the author thanks the developers of ‘R’ and ‘Perl,’ who are developing excellent programming tools and distributing them freely. The author also thanks many researchers who publish web pages describing ‘R’ for the Internet. Those pages are very helpful for using ‘R.’

References

[II07] Ishimura, S. and Ishimura, K., Nyumon Hajimeteno Bunsanbunseki to Tajuhikaku. Tokyo Tosho, 2007.

[NY97] Nagata, Y. and Yoshida, M., Tokeiteki Tajuhikakuhou no Kiso. Saientisuto Sya, 1997.

[R D07] R Development Core Team. R: *A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[WCS96] Larry Wall, Tom Christiansen, and Randal L. Schwartz. *Program-ming Perl.* O’Reilly & Associates, Inc., 2nd edition, 1996.