How to Perform Data Analysis Using the Free Software EZR. Part 1: Single Continuous Dependent Variable

Liaquat Roopesh Johnson¹, Usha Devi Karunakaran²

**Key Message:** EZR is based on the free, open-source software R. It is available as a plugin for the package R Commander from the Comprehensive R Archives Network (CRAN). We will limit our discussion to the most common procedures, while giving a general overview of the program. In this article we describe the analysis of continuous dependent variables. No previous experience with statistical software is assumed.

**Key words:** EZR, Data Analysis, Single Continuous Dependent Variable

INTRODUCTION TO EZR

EZR is based on the free, open-source software R. It is available as a plugin for the package R Commander from the Comprehensive R Archives Network (CRAN). It is also available as standalone software from the website of its creator, Dr. Kanda (URL: http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmedEN.html).

Although R Commander has a suite of statistical functions, the layout is not ‘novice friendly’. Besides, the native functionality of R Commander is restricted by the limited number of biostatistical tools available in it. EZR enhances the user experience while adding considerably to the suite of available biostatistical functions. The list of statistical functions available in EZR is available at http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmedEN.html.

Installation of EZR in Windows Operating Systems follows the standard procedure for software installation in such systems. A screenshot tutorial on how to install EZR is available here: URL.

Data Analysis in EZR

As in the previous article of this series, we will limit our discussion to the most common procedures, while giving a general overview of the program. In this article we describe the analysis of continuous dependent variables. No previous experience with statistical software is assumed.

Double clicking the shortcut icon on the desktop causes two windows to open in succession. The first window is the R console, while the second is the EZR on R Commander Graphical User Interface (GUI) window (henceforth referred to as EZR) (Fig.4). EZR operates out of the latter window. The R console offers all the functionality of R, and can be used by typing commands in it. Those unfamiliar with the R programming language are advised to minimize the R console and work in the EZR window exclusively. Closing the R console will result in exiting both R console and EZR.

## Please refer to supplementary file (http://goo.gl/u8yHKm) for figures

¹Assistant Professor, Department of Community Medicine, Sree Mookambika Institute of Medical Sciences, Kulasekharam
²Professor, Department of Community Medicine, Sree Mookambika Institute of Medical Sciences, Kulasekharam
Correspondence to Dr Liaquat Roopesh Johnson (liaquat99@gmail.com)
For the purpose of this tutorial series, we have created a fictitious data set that can be downloaded from: [http://goo.gl/IMn20o]

The data set contains most types of variables one might encounter in practice. Users are encouraged to explore the various procedures available in EZR using the sample data set. This is critical, and will not only increase familiarity with the program, but also boost confidence in performing data analysis.

Figure 3 is a Screenshot of EZR after importation of Excel data set.

Figure 4 shows the tests available for continuous (normally distributed) variables in EZR.

The Kolmogorov-Smirnov Test for normal distribution includes the Shapiro-Wilk test, and tests for skewness and kurtosis. Bartlett’s test checks for homogeneity of variances between groups. Many statisticians recommend testing the data for normality before proceeding with parametric tests of significance. One is expected to check for homogeneity of variances before performing Analysis of Variance (ANOVA).

Figure 5 is a screenshot showing the list of available nonparametric tests in EZR.

Figure 6 is a screenshot explaining the output of ‘Numerical summaries’.

Since we are performing confirmatory data analysis, we propose a hypothesis before performing each test of significance described here.

Procedure: Tests for normality. Shapiro-Wilk Test (performed with the Kolmogorov Smirnov Test)

H₀: The distribution of values follows normal distribution.
H₁ (2-tailed): The distribution of values does not follow normal distribution

Obtaining a p value ≤0.05 indicates deviation from normality. Therefore, we wish to obtain a p value >0.05.

Examination of skewness and kurtosis [generated automatically when performing the KS test]

H₀: The values are close to zero (indicates normal distribution).
H₁ (2-tailed): The values are far from zero (indicates non-normal distribution)

Figure 7 is a screenshot showing how to perform Kolmogorov Smirnov (Shapiro-Wilk) Test

Figure 8 is a screenshot explaining the output of Kolmogorov Smirnov (Shapiro-Wilk) Test

Figure 9 is a Histogram of Systolic BP 2 hours after drug administration.

Examination of measures of central tendency (mean, median)

H₀: The values are close to each other (indicates normal distribution).
H₁ (2-tailed): The values are far from each other (indicates non-normal distribution)

Figure 10 is a screenshot of numerical summaries with values of Systolic BP 2 hours after drug administration highlighted

Table 1 is the summary table of results of Tests of Normality

<table>
<thead>
<tr>
<th>Test name</th>
<th>Result</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk Test</td>
<td>P&lt;0.05</td>
<td>Not normal</td>
</tr>
<tr>
<td>Mean, Median</td>
<td>132.57, 133.00</td>
<td>Probably normal</td>
</tr>
<tr>
<td>Histogram</td>
<td>Bell shaped curve</td>
<td>Probably normal</td>
</tr>
<tr>
<td>Skewness</td>
<td>0.09 (close to zero)</td>
<td>Probably normal</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>0.14 (close to zero)</td>
<td>Probably normal</td>
</tr>
</tbody>
</table>

#please refer to supplementary file ([http://goo.gl/u8yHKm](http://goo.gl/u8yHKm)) for figures
The test output indicates that there is no significant difference in mean systolic BP between male and female subjects at baseline. Therefore, we fail to reject the null hypothesis at 5% significance level.

Note: The investigator can either reject the null hypothesis (H0), or fail to reject it. The H0 can never be accepted. This is due to the fact that even though we have failed to reject the H0, other investigators may well do so later. Summarily accepting the H0 deprives other investigators of that possibility/ opportunity.

Procedure: Bartlett’s Test for Homogeneity of variances

H0: There is no difference in variance of height between the various Socio Economic Status (SES) groups (Low/ Middle/ High)

H1 (2-tailed): There is a difference in variance of height between the various Socio Economic Status (SES) groups

One of the assumptions of ANOVA is that the variances are equal. Therefore, before performing ANOVA we must first establish that this is true. The outcome of Bartlett’s test will inform us if the variances are equal between the various groups. If the p value is ≥0.05 we can conclude that the variances are equal, and perform ANOVA.

Figure 13† is a Screenshot explaining how to perform Bartlett’s Test

Figure 14† is a Screenshot explaining the output of Bartlett’s Test

Note: R deals with very large/ very small numbers by truncating them into the general form xe^y, where x is the numeric value of the result; e^y = 10^y. Thus, if the actual value is 10,23,000 R will report it as 1.023e+05. This simply means “move the decimal 5 places to the right”. Similarly, if the actual value is 0.000000001 R will report it as 1.0e-11. This simply means “move the decimal 11 places to the left”.

Procedure: One-way Analysis Of Variance (One-way ANOVA)

H0: There is no difference in mean height between the various Socio Economic Status (SES) groups (Low/ Middle/ High)

H1 (2-tailed): There is a difference in mean height between the various Socio Economic Status (SES) groups

Figure 15† is a Screenshot explaining how to perform Analysis of Variance (ANOVA)

Figure 16† is a Screenshot explaining output of ANOVA (Interpretation provided by the authors)

The ANOVA table indicates that there is a significant difference in height between (at least two) SES groups. In order to determine between which two groups the difference is statistically significant, we examine the result of the post hoc test (Tukey’s Honestly Significant Difference (HSD) Test). We observe that the pairwise comparison has generated significant p values for two pairs: Middle- Low; and High-Low SES groups.

We conclude that there is a significant difference in height between Middle-Low and High-Low SES groups, and reject the null hypothesis at 5% significance level.

Procedure: Mann-Whitney U Test

H0: There is no difference in Activity Score between Diseased and Normal subjects

H1: There is a difference in Activity Score between Diseased and Normal subjects

Here, the Activity Score is a fictitious score based on daily physical activity level. ‘Disease’ refers to a representative disease that is affected by physical activity level. We want to see if there really is a difference in terms of physical activity between those who have and do not have disease.

Figure 17† is a Screenshot explaining how to perform Mann-Whitney U Test in EZR.

Figure 18† is a Screenshot explaining the output of Mann-Whitney U Test.

Users might be concerned about the name of test displayed in the output as it is not the Mann-Whitney U Test. However, the Wilcoxon Rank-Sum Test is numerically equivalent to the Mann-Whitney U Test, and yields similar results. The test statistic described (W) is that of the Wilcoxon Rank-Sum Test. We will focus on the p value to determine if the H0 can be rejected. Here, the p value is >0.05. Therefore, we cannot reject the H0 in the above example.

Procedure: Kruskal-Wallis Test

H0: There is no difference in heights of subjects belonging to various SES groups

H1: There is a difference in the heights of subjects belonging to various SES groups

Figure 19 is a Screenshot showing how to perform Kruskal-Wallis Test

Figure 20 is a Screenshot showing how to interpret the output of Kruskal-Wallis Test (Interpretation by the authors)

As in the case of ANOVA, we first examine the result of the Kruskal-Wallis test to check for overall significance. Since the p value is <0.05, we then examine the result of pairwise comparison. The pairwise comparison indicates that there is a statistically significant difference in height between subjects belonging to Low SES and

#please refer to supplementary file (http://goo.gl/u8yHKm) for figures
High SES groups. We conclude that there is a statistically significant difference in height between subjects of Low and High SES groups, and reject the null hypothesis at 5% level.

**Procedure: Pearson’s Correlation**

H₀: There is no correlation between height and weight of subjects

Hₐ (2-tailed): There is a correlation between height and weight of subjects.

*Figure 21* is a Screenshot showing how to perform Pearson’s Correlation in EZR

*Figure 22* is a Screenshot explaining how to interpret the output of Pearson’s Correlation (Interpretation by authors).

The Pearson’s Product-moment Correlation is performed between two continuous variables that are normally distributed. Here, the investigator has examined the correlation between Height and Weight of the study subjects. The output indicates that there is a (weak) positive correlation between height and weight of the study subjects. That is, as the height increases, so does weight also, and vice versa. This is statistically significant, with the p value <0.05. The 95% Confidence Interval is for the test statistic. EZR summarizes the results of the test, giving the correlation coefficient, 95% confidence interval, and the p value.

**DISCUSSION AND CONCLUSION**

The data analysis involving one continuous dependent (outcome) variable can be described in terms of parametric and non-parametric tests of significance.

Although tests of normality are available within EZR, users are advised to exercise caution in the choice of parametric or non-parametric tests based upon those results.

Parametric tests of significance include tests applicable to normally distributed continuous variables. Non-parametric tests have fewer assumptions regarding the data, but are less robust in detecting significance (in normally distributed data) than parametric tests.

The commonest parametric tests are the t tests and ANOVA, while their equivalent tests are the most common non-parametric tests- the Mann-Whitney U Test and the Kruskal-Wallis Test. It is recommended that equality of variances be checked using either Bartlett’s Test or Levene’s Test prior to ANOVA.

Other tests include correlation and regression. Pearson’s product moment correlation is employed for normally distributed continuous variables, while Spearman’s correlation is used when dealing with non-normal continuous variables. Positive correlation implies that as one variable increases in value, so does the other variable (and vice versa). Negative correlation indicates that an increase on value of one variable causes a decrease in value of the other variable. The strength and direction of correlation are indicated by the correlation coefficient, which ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation). The interpretation for Spearman’s correlation is similar to that described above.

We will describe regression in another article later in this series.

The next article in this series will describe data analysis involving one categorical dependent variable.

**REFERENCES**