An Online Application for Exploratory and Explanatory Data Analysis

Paul Collett
Shimonoseki City University

Reference Data:

Collett, P. (2022). An online application for exploratory and explanatory data analysis. In P. Ferguson, & R. Derrah (Eds.), *Reflections and New Perspectives*. JALT. https://doi.org/10.37546/JALTPCP2021-26

This paper introduces an online application developed to assist with statistical data analysis. While currently limited in scope to *t* tests, correlation, and regression analysis, the application moves beyond standard implementation of these tests by augmentation with bootstrapped statistics and several exploratory and explanatory graphical plots. Measures of effect size and confidence intervals are also calculated. The application is designed to be easy to use, requiring only input of correctly formatted data for the analyses to be run. After providing a brief introduction and rationale for the application, issues related to quantitative data analysis underpinning its analysis are outlined. Guidance on the functionality of the application is also provided. Finally, some caveats regarding its usage are addressed. Whilst aimed primarily at researchers with limited experience in quantitative data analysis, it is hoped the application may also be useful for a broad range of users.

本論文では、統計データ解析を支援するために開発されたオンラインアプリケーションを紹介する。現在のところ、t検定、相関、回帰分析に限定されているが、ブートストラップや探索的、説明的なグラフプロットで拡張することで、これらの検定の標準的な実装を超えるアプリケーションである。該アプリケーションを用れば、効果量と信頼区間も計算され、使いやすく設計されているために正しくフォーマットされたデータを入力するだけで分析が実行される。本論文では、アプリケーションの簡単な紹介と理論的根拠を示した後、解析の基礎となる定量的データ解析に関する問題を概説する。また、アプリケーションの機能に関するガイダンスと、使用の際の注意事項を提供する。定量データ解析の経験が浅い研究者を主な対象としているが、幅広いユーザーにとって有用なアプリケーションとなることを願っている。

This paper introduces *R-pc.net*, a new online web-based application currently in development that aims to make quantitative data analysis more accessible to language teachers and researchers. The application, accessible via http://r-pc.net/shiny/rstudio/, has been designed following contemporary recommendations for statistical testing and reporting. Specifically, this involves providing alternative measures of statistical outcomes to the *p* value calculated and reported in most quantitative research in foreign language learning and teaching research fields. These alternatives include data-rich graphical plots, confidence intervals (CIs), effect sizes, and bootstrapped statistics. While R-pc.net is meant to be easy to use even for researchers with limited experience in statistical analysis, a basic understanding of statistical concepts is required. In this respect, development of the application was motivated not just to provide a tool for analysis, but also to help researchers better understand possible approaches to statistics.

Improving Statistical Literacy

Statistical literacy is an important area of knowledge for practitioners and consumers of research across language learning and teaching fields. Despite the trend toward increased use of inferential statistics in L2 (SLA) research (Gass et al., 2021), several studies suggest deficiencies in this area. For example, Khany and Tazik (2015) concluded that statistical use in applied linguistics research remains largely at a basic level. Brown (2016) identified a few basic research errors, including failure to report test assumptions and misinterpreting outcomes of statistical significance measurements. Al-Hoorie and Vitta (2019) noted incomplete reporting of important testing indices. Amongst SLA doctoral students, Gonulal (2020) reported relatively poor levels of statistical literacy. Given these problems, improved quantitative research practises into SLA are needed, including more advanced education, training, and methodologies for statistical analysis (Gass et al.; Larson-Hall & Herrington, 2010). Mizumoto and Plonsky (2016) recommended embracing *R*, the open-source statistical programming language (https://www.R-project.org) as the *lingua franca* in applied linguistic quantitative research. They

argued that R could contribute to a better understanding of statistical procedures and results, adoption of modern statistical techniques such as improved data visualisation[1], and more accurate reporting of statistical results. To improve statistical literacy, Mizumoto and Plonsky developed *Langtest*, an online web application for statistical analysis (https://langtest.jp). This application was created using the *R Shiny* package, an add-on to the R framework that allows for the creation of self-contained web-based applications. As such, langtest.jp offers various inferential statistical testing options without requiring the user to do any complex coding. However, in some areas, the Langtest system is showing its age. This is certainly not meant as a criticism, as Langtest still has valuable functionality. However, new developments in R mean that enhanced features are constantly being made available. In the spirit of Mizumoto and Plonsky (2016), and to extend the R language, R-pc.net was developed.

R-pc.net Basic Features

Like langtest.jp, the R-pc.net application is built using R packages freely available online. It is offered at no cost to interested users, and no financial gain is intended from its utilisation. Use of R-pc.net does not require any knowledge of the R programming language itself, as all the functionality needed to carry out various statistical tests has been programmed into the application. It is a web-based system that allows the user to enter data and then run several statistical tests, including the following:

- Various *t* tests, for investigating the differences between two groups;
- Correlation, when measuring association between variables;
- Linear regression, for understanding how outcomes on a measure for two or more groups are related, and to predict outcomes based on data trends.

User-configurable options allow for the selection of different test types and statistics. Figure 1 shows a view of the data input pages for different tests.

Figure 1A

Data Input Page of R-pc.net: t test

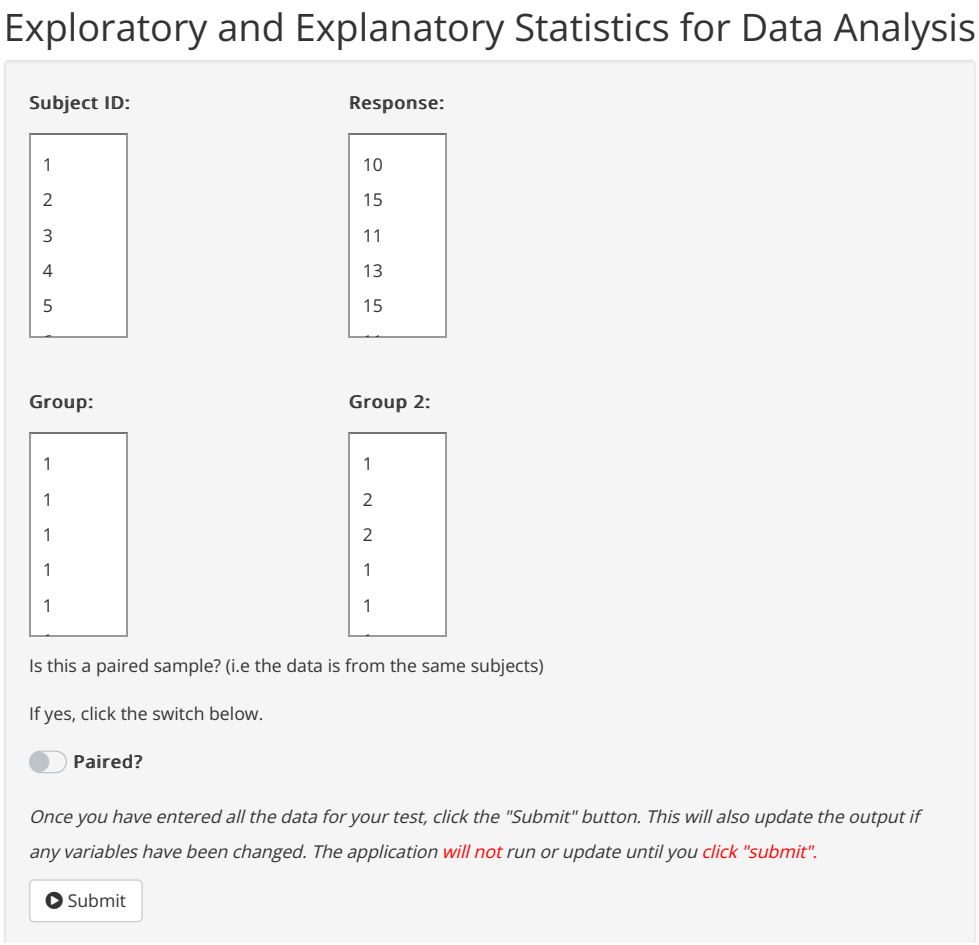


Figure 1B

Data Input Page of R-pc.net: Regression / correlation

Regression Analysis

Upload Data | Output

Uploading Data

Choose CSV File

Browse... sample-data.csv

Upload complete

The variables below control how the application reads your data. Please set as required.

If your data table has a header row, make sure the checkbox below is checked.

☑ Header

How are the columns separated? Select the correct option.

Separator

○ Comma
○ Semicolon
◉ Tab

Are the table values quoted (e.g.; "1","2",...)? Indicate below.

Quote

○ None
◉ Double Quote
○ Single Quote

Subject	Group	Scale.1	Test.Outcome
1	d	57	309
2	c	37	214
3	c	47	323
4	c	44	288
5	d	59	268
6	a	88	245
7	b	27	183
8	b	29	260
9	b	28	256
10	d	53	300
11	c	43	293
12	b	27	182
13	b	32	234
14	d	51	252
15	b	32	272
16	b	33	210
17	c	47	279
18	d	48	263
19	b	36	291
20	c	44	269
21	c	44	261
22	c	37	269
23	c	40	306
24	c	45	264
25	d	65	269
26	d	53	280
27	a	31	168

Note. 1A shows the initial input screen for running a *t* test. In the current version of the application, the data for this test is entered manually. Provision is made for a 2nd grouping variable as well as the primary variable to be tested.
1B shows the input screen for a regression analysis or correlation test. In this case, data is uploaded from a file on the user's computer, and the variables to be tested are selected by the user.

The tests themselves are all run automatically, and the user needs only to enter data, select variables to be tested, and choose from a list of output options. Results are output in both graphical and numerical form. As much as possible, the output is annotated to aid with understanding. Figure 2 shows one view of the result of an analysis.

Figure 2a
Sample R-pc.net Output

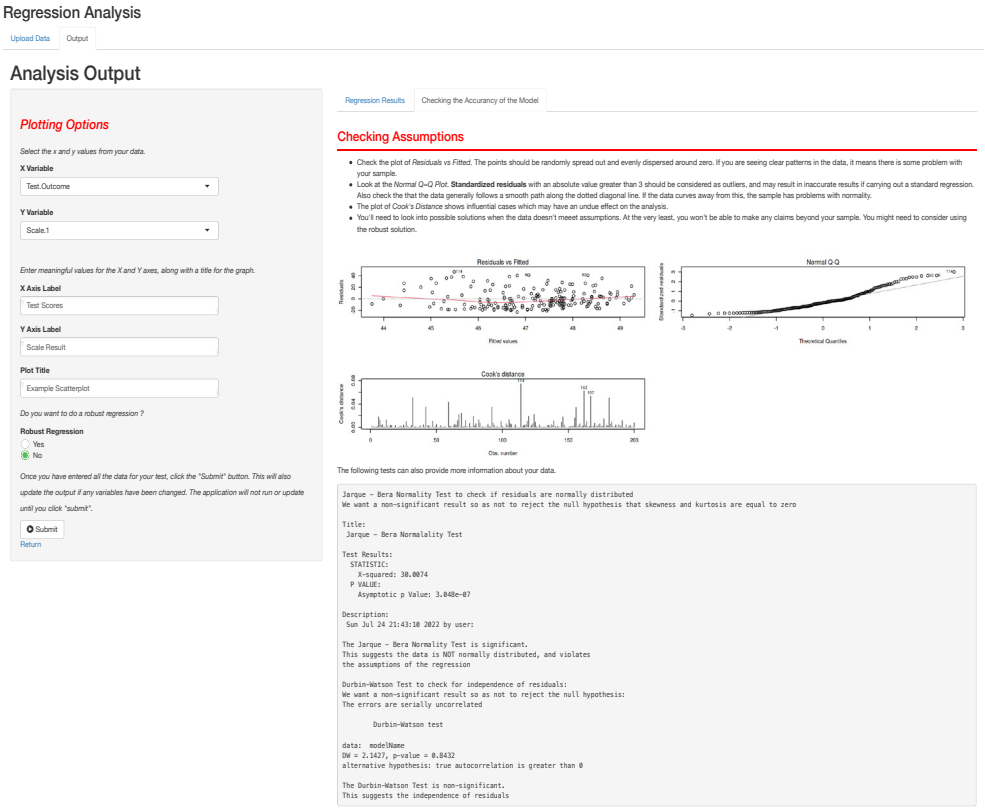
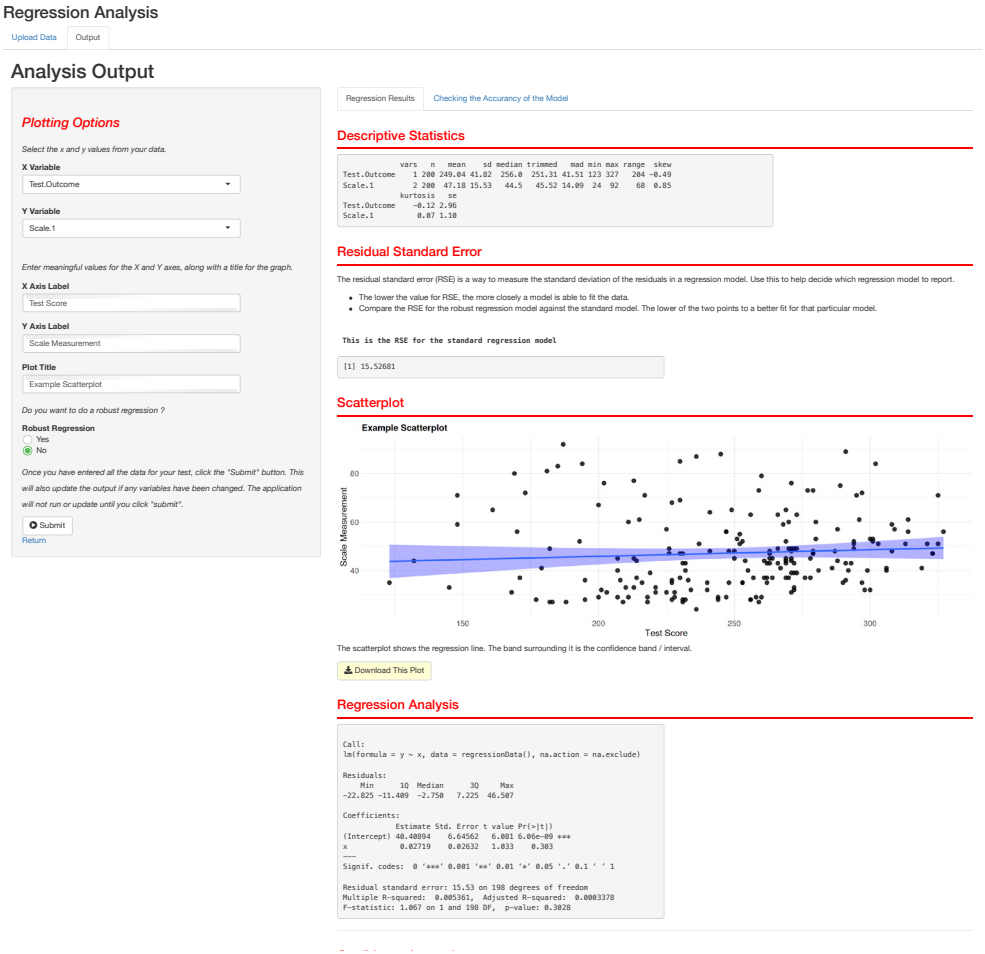


Figure 2b
Sample R-pc.net Output

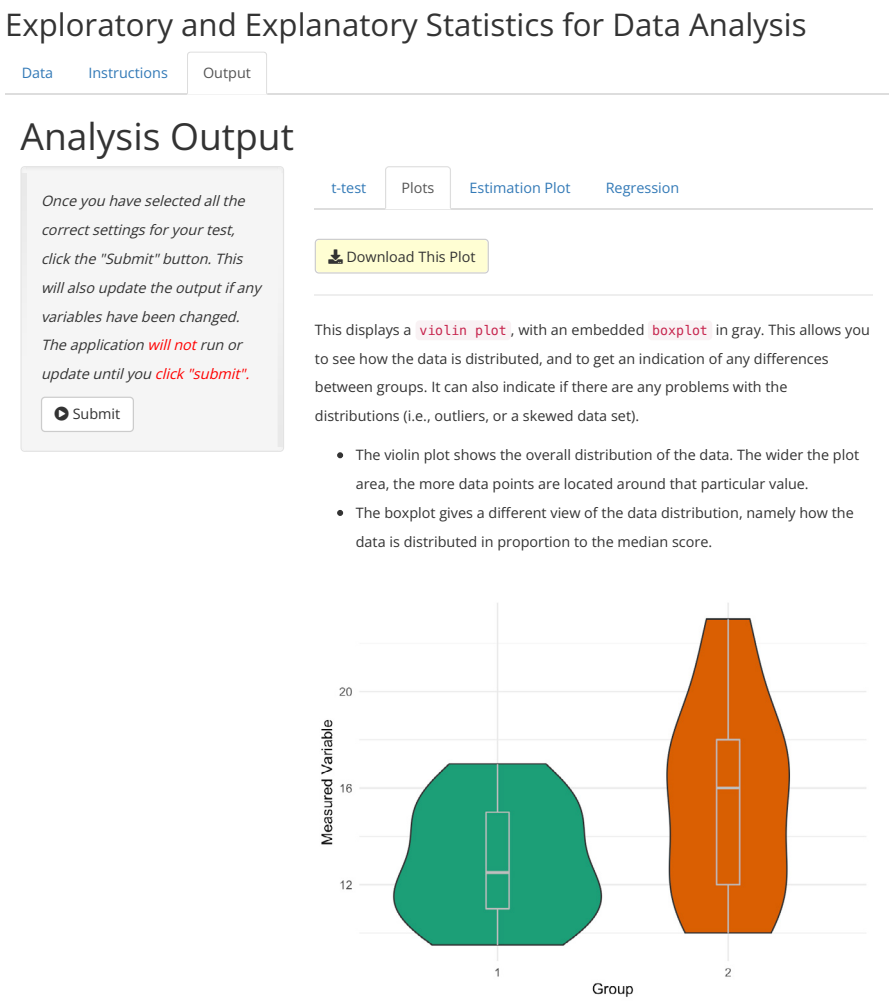


Note. An example of the output for a regression analysis. The application includes details on assumption checks, confidence intervals, and the option to choose a bootstrapped regression.

Where R-pc.net aims to expand on langtest.jp is in its option to calculate bootstrapped versions of statistics, and choices of graphical output. This latter area is something in which R excels and R-pc.net leverages this graphical functionality to produce various kinds of plots useful for understanding and exploring trends in data. This follows the recommendations of Hudson (2015), Larson-Hall (2017), and Larson-Hall and Herrington (2010) for making the graphical presentation of results and data trends easier to interpret. Plots of data are also especially useful when carrying out an initial exploration of a dataset, as visualising the data can indicate potential problems such as deviations from normality or other outliers. Depending on the analytical focus, R-pc.net can provide different sorts of plots. For regression analysis, scatterplots showing correlations and predicting trends are generated. For *t* tests, violin plots are provided. These are a relatively new kind of graph which show how samples are distributed around a central measure, such as the mean or median. In other words, they make the "shape" of the entire dataset readily apparent. Figure 3a provides an example.

Figure 3

Sample Violin Plot



Note. The violin plot incorporates a boxplot to provide a better overview of data distribution. For guidance on how to interpret these graphs, see Collett (2021).

Finally, graphs can provide information about the relationship between samples which transcends the limited information provided by a *p* value from a statistical significance test. Ho et al.'s (2019) implementation of estimation plots is a case in point. These compare the difference between two or more groups (e.g., in situations where a *t* test may be applied), plotting the distribution and relationship of all data points along with effect sizes and their associated CIs. An example is given in Figure 4.

Figure 4

Sample Estimation Plot

Results | Plots | Estimation Plot | Regression

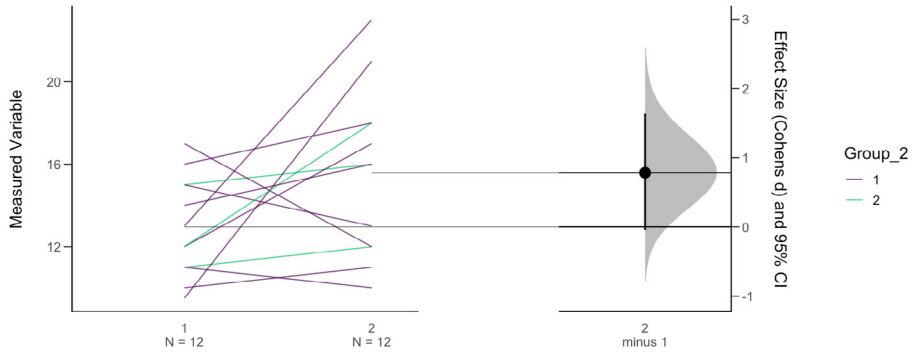
Estimation Plot

The `estimation plot` shows a plot of each data point in the two testing conditions, along with the mean sore for each sample, and the effect size with confidence interval.

- In the case of an **unpaired sample**, the individual points are displayed.
- If a **paired sample**, the points are joined.
- The mean for each test outcome is shown by the two horizontal lines running from the right-hand y-axis.
- The right-hand graph displaying the curve shows the effect size for the analysis and the 95% confidence interval (CI) for the effect size. If the CI crosses zero, then you can comclude that any difference between the two samples is likely not due to the effect being tested for. If the CI does not include zero, then the difference between samples warrents further investigation or consideration.

```
dabestr (Data Analysis with Bootstrap Estimation in R) v0.3.0
=============================================================

Good morning!
The current time is 10:01 AM on Thursday June 02, 2022.

Dataset    :  df
X Variable :  Grouping1
Y Variable :  Measured

Paired Cohen's d of 2 (n = 12) minus 1 (n = 12)
 0.782 [95CI  -0.0447; 1.64]


5000 bootstrap resamples.
All confidence intervals are bias-corrected and accelerated.
```

Download This Plot



This particular kind of plot accords with recommendations for reform in statistical reporting practices advanced over the recent past.

Calls for Reform in Statistics

The use of statistical significance testing has long been the standard method for quantitatively assessing the difference between groups. The aim of this process is to find and report a *p* value that is calculated based on the difference between a measure of central tendency for the groups, such as the mean. However, the process is problematic, and has been under criticism for years (Meehl, 1990). The consensus is that the *p* value found and reported in inferential statistical tests does not tell us much, and often not what many researchers think it does (Cohen, 1994; Wasserstein et al., 2019). Calls for reform in statistics abound. Perhaps the easiest to implement is the reporting of interval estimates and effect sizes, as they provide a measure of the possible size of differences between means of samples. This gives better data on which to frame conclusions rather than the simple accept-or-reject dichotomy of a *p* value. Calls for change notwithstanding, a focus on statistical significance continues to dominate quantitative research methodology in many disciplines. The reasons are manifold. Partly it can be linked to statistical education, textbooks, publications, and editorial practices. Broader historical and philosophical factors related to the development of the social & behavioural sciences as academic disciplines have also played a part (Manicas, 1987; Ziliak & McCloskey, 2008). However, some progress can be seen. For example, if carrying out research with an eye to publication, limiting results to reports of *p* values is becoming less acceptable for many editorial boards. Although there does not seem to be a clear policy on the reporting of statistical results within fields related to L2 research, an adherence to American Psychology Association (APA) conventions is common (Larson-Hall & Plonsky, 2015; Lindstromberg, 2016; Norris, 2015). The APA (American Psychological Association, 2010) recommends that in reporting outcomes of statistical analyses, *p* values should at the very least be supplemented with effect sizes and CIs. Subsequent interpretation of results should then be made with reference to these two indices.

Confidence Intervals (CIs)

One of the clearest definitions of a CI is supplied by Larson-Hall and Plonsky (2015, p. 136): "...a set of two numbers that represent a range of values where, with 95% confidence, you would expect some point estimate...to appear if you repeated the same study again and again." Here, the point estimate is usually the mean, but the median can also be used. 95% is the default level for CIs, but larger or smaller values are possible. The size of the CI gives an indication of the precision of a test—the larger the CI, the less predictive power the test has. The CI can also indicate an effect size. For example, a smaller CI can be a sign of a more powerful effect of the measured intervention. CIs can also be used in place of p-values to help formulate decisions on whether to accept or reject the null hypothesis. If the CI contains zero, then this means zero is one of the possible values for the true mean, i.e., one can accept the null hypothesis of no difference between the means. However, as with many aspects of statistics, there is debate about the usefulness of CIs. Like *p* values, evidence suggests they are often misunderstood (Al-Hoorie & Vitta, 2019; Cumming, 2012). Cautions about their interpretation abound (Greenland, et.al, 2016), with some implying that it is better not to report them (Al-Hoorie & Vitta, 2019). I would argue that interpreted correctly, the evidence suggests they are more useful than the *p*-value from a significance test (Calin-Jageman & Cumming, 2019; Cumming, 2012, 2014; Greenland, et. al, 2016) and should be calculated and reported where possible. As the American Statistical Association statement on statistical significance and *p* values lays out, "good statistical practice emphasizes...a variety of numerical and graphical summaries of data..." (Wasserstein & Lazar, 2016). The options in R-pc.net provide CI output for most tests, and it is up to the researcher to be sure they understand exactly what a CI represents and to correctly express this conceptualization when reporting results.

Effect Sizes

An effect size is a statistic that gives a standardised measure of the magnitude of difference between groups under consideration. In other words, it is a theoretical measure of the possible effect of whatever intervention is being tested. It is becoming more common to see mention of effect sizes in research. Here too, there are some issues surrounding their use and interpretation that warrant consideration. Numerous possible measures of effect exist, with different ways of interpretation. A basic measure of effect is the mean difference between two samples, simply calculated as μ1 - μ2. Another common effect size is a correlation coefficient, such as Pearson's *r*. One effect size perhaps most frequently encountered in research is Cohen's *d*, used as a measure of the difference between groups (i.e., when carrying out *t* tests or ANOVA, etc). A recommended overview of issues related to effect sizes is provided by Kelley and Preacher (2012).

When calculating and reporting effect sizes, it is important to consider the aim of the analysis. If this is to apply conclusions from the analysis to a wider population, simply

giving the effect size itself may be somewhat limiting as it does not help on its own for extrapolating conclusions to the population from which the sample is drawn (Kelley & Preacher, 2012). If, for example, the calculated CI around the effect size is very wide, incorporates low values, or crosses zero, this would lead one to question its usefulness. One of the benefits of R-pc.net is that it can be set to produce different effect size indices and perform the calculation of CIs around effect sizes where appropriate.

Restricting reporting of statistical tests to the *p* values from statistical significance tests also reduces the value of research results. It is difficult to use such results in subsequent meta-analyses or replications of research. These are both important parts of ongoing research, necessary for advancing knowledge and theory-creation. Meta-analysis involves statistically comparing outcomes of research into a specific area to generate an estimation of the effect in the population. As such, it requires a measure of effect size for each study if research outcomes are to be comparable. Replication studies, if they are to be meaningful, should aim to show if effect sizes are similar or not for the phenomena being tested. Not including an effect size in the results of research findings effectively means the study cannot contribute to this kind of research and has limited extensibility.

A final reason to report effect sizes is that they provide future researchers with a helpful metric for calculating the power of a statistical test. Here, power refers to the probability that the test will be able to detect an effect when it exists, or "...the probability that it will yield statistically significant results" (Cohen, 1988, p. 1). Power is a combination of the sample size, the effect size, and the type-I error probability (i.e., the *p* value) the researcher is willing to accept. Power should always be calculated (Brown, 2016) ideally before carrying out a study, and reported as part of the results. The point is to ensure an analysis has a large enough sample to detect an effect before carrying out the research. Not having reference to effect sizes from previous studies in related areas means a researcher carrying out a new study would effectively need to guess what effect to expect, potentially over- or under-estimating the true effect and the sample size required to detect it. This could result in the researcher using a too small sample or being put off carrying out a study if the calculated sample size is larger than resources allow for.

Where appropriate, based on the kind of analysis, R-pc.net output includes calculation of CIs and measures of effect size. Both standard and robust statistical tests can be performed. A range of graphical data plots are also generated to aid in explanatory and exploratory data analytic processes, such as when checking the assumptions a dataset needs to meet for the particular statistical test to be applied.

While CIs and effect sizes should be reported, one possible problem that could arise is that like *p* values, they are vulnerable to extreme variability in the data. This is due to the nature of the tests and means that they may not provide appropriate results in all situations. One method that can help overcome this problem is the use of robust statistical methods.

Robust Statistics

Robust statistics are statistical methods developed for cases where collected data to be analysed may violate the assumptions upon which statistical procedures are predicated. The logic underlying the standard inferential statistical tests such as *t*-tests or regression commonly used in L2 research is based on certain assumptions regarding the data being tested. These include random sampling, the data following a normal distribution, and adequate sample sizes. Failure to adhere to these assumptions can lead to misinterpretations of the sampled data or spurious conclusions. Random sampling is important if the aim is to make conclusions that can be extended beyond the test subjects. If the subjects are all drawn from the same class or some other sample of convenience, underlying similarities in the group will influence outcomes and effectively limit applicability of results to just the group under study. Sample sizes need to be large enough to allow for variability in the data to be detected. Non-normally distributed data can skew results in one direction, or potentially cancel out effects. An ongoing discourse on the use of inferential statistics is that standard tests are resilient to violations of assumptions given an appropriate sample size, but most of these arguments appear to apply to limited cases (Field & Wilcox, 2017). One solution when assumptions are violated is the use of nonparametric tests. This is a perfectly acceptable approach, but in the interest of simplicity, R-pc.net does not calculate nonparametric statistics[2] because there are cases where they may not perform well, thus leading to increased likelihood of erroneous results. A better approach is the use of robust statistical tests. Comparison of methodologies suggests using robust statistics will in most situations provide as valid or better statistical outcomes as standard or nonparametric tests (Field & Wilcox, 2017; Wilcox, 2022). Methods used in robust statistics include the use of trimmed means and bootstrapping. The former is a process whereby the top and bottom *x*% of scores in a dataset are dropped from the data, and the statistical analysis is carried out on the modified dataset. Bootstrapping is a more complex process involving resampling from a dataset multiple times to create a new theoretical data distribution generated from the individual data points in the original sample. This new distribution better represents the sample population and approaches normality and is then used as the test data. Within the general applied linguistics or foreign / second language learning and teaching research context, there is so far limited treatment of robust statistics. Larson-Hall and

Herrington (2010) offer an introduction to trimmed means and bootstrapping with the latter covered further in LaFlair et al. (2015). Larson-Hall (2021) provides a positive review of McLean et al.'s (2020) use of bootstrapping in their analysis. Given the general value of robust statistics, it is hoped their use will be embraced by more researchers. The R-pc.net application aims to provide an easy way to calculate them in a number of situations, such as robust vs standard *t* tests and regression analysis.

Robust vs Standard t tests

The *t* test, used for assessing the difference in means of two groups, has multiple forms. Student's *t*, which tends to be the default version of the test used in many studies, is suitable for normally distributed samples with equal variance in the two groups. Welch's version of the test is designed to be used with samples which are normally distributed but have unequal variance. Delarce, Lakens, and Leys (2017) argue that in research involving measured variables or predetermined sampling criteria, Welch's *t* test should be used, as we cannot assume equal variance between groups in these cases. Lindstromberg (2016) points out that Student's test is unlikely to be suitable in cases of small samples ($N < 30$), and that Welch's test should be preferred. However, a better approach is to use a robust bootstrapped version of the test. Field and Wilcox (2017) outline one suitable approach here, in the form of the Yuen *t* test, a two-sample trimmed *t* test that can be used in cases with unequal population variances. However, Delacre et al. (2017) claimed that Yuen's test does not perform as well as Welch's test under several cases where the data is not distributed symmetrically. For the online application, both Welch's *t* test (the default version used in standard R packages) and a bootstrapped version of Yeun's test are calculated. It is recommended that results for both tests, along with all descriptive statistics and graphical plots of the data, are then reported. This will allow the audience to better judge conclusions from the research and how they relate to other findings.

Regression

Regression analysis is used when the aim is to look at correlations between data and how variations in one measure predict outcomes in others. The output for regression in R-pc.net includes tests of assumptions of the data and graphical exploration of the dataset in the form of scatterplots. As well as allowing for the calculation of a standard regression coefficient, there is the option to run a bootstrapped regression in the event of outliers or other problems with data. Here, the weighting of variables is adjusted, and multiple computations generated to "smooth out" any problems with the distribution. Due to the nature of the calculations, this does not provide an effect size measure in the form of an R^2 statistic that is available from a standard regression analysis, but the β value (slope) can serve as an alternative effect size as it represents the amount of change in the dependant variable for every one unit increase in the independent variable.

Research practitioners should be encouraged to utilise modern robust techniques such as bootstrapping (Larson-Hall & Herrington, 2010). Judicious use of graphical data analysis methods should also be promoted. Any approach aimed at helping researchers with statistical analysis should, in addition to providing easy-to-understand output such as descriptive statistics, simultaneously make use of modern techniques to present complex results in a usable fashion. R-pc.net aids here by providing a means for researchers to analyse their results using modern statistical methodology and helping to clearly explicate the results generated.

Some Caveats

There are several caveats regarding R-pc.net that should be noted:

- The application is a "black box" system. The inner workings are not apparent, so one needs to have faith in the process. However, once the data is processed, enough information should be provided to ensure the analysis is error-free.
- Garbage in, garbage out: An analysis is only as valid as the data you use. However, using graphical means to check assumptions and distributions, along with descriptive statistics, should point to any problematic issues with the datasets being tested and results generated.
- The application cannot interpret the analysis for you. It provides output based on user data, accompanied by a suitable level of information to help interpret the results. Beyond that it is up to the user to draw their conclusions on the meaningfulness of the results.
- The application is, in a sense, opinionated. The choices of tests are chosen based on theoretically driven decisions about the "best" options to use. However, there is considerable debate within methodological circles relating to statistical data analysis approaches. It is hoped users of this system will familiarise themselves with the issues brought up in this paper to help better understand the output obtained.

Summary

It is hoped that the preceding explanation will interest researchers enough to investigate R-pc.net. Providing resources to explore and explain data trends in ways that move beyond the limited information provided by a *p* value may be one step towards improving the quality of research reports. While somewhat simple in functionality, the rationale is to prioritise usability over advanced functionality. Researchers in need of more complex analyses will likely have the requisite expertise to do so. Offering a more complex solution entails a more complex user interface. This would make the system harder to work with and perhaps lead to errors in analysis. Going forward, other functionality can and will be added as time permits. For now, however, the preference is to provide a system that will be helpful for anyone who may need to carry out a relatively non-complex statistical analysis, and to provide the results of the analysis in a format which can contribute to substantive conclusions about the data. Mizumoto and Plonsky (2016) hoped that promoting the use of R as a tool for statistical analysis would help contribute to increased statistical literacy. This sentiment also underlies R-pc.net. If it can help contribute to better understanding and reporting of statistical outcomes, it will have accomplished a useful goal.

Notes

1. See also Larson-Hall & Plonsky, 2015.
2. See Turner (2014) for more on nonparametric tests.

Bio Data

Paul Collett works at Shimonoseki City University. His interests are in research methodology and epistemology, as well psychological factors related to language learning. <collett@shimonoseki-cu.ac.jp>

References

Al-Hoorie, A. H., & Vitta, J. P. (2019). The seven sins of L2 research: A review of 30 journals' statistical quality and their CiteScore, SJR, SNIP, JCR Impact Factors. *Language Teaching Research*, *23*(6), 727-744. https://doi.org/10.1177/1362168818767191

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). American Psychological Association.

Brown, J. D. (2016). *Statistics corner: Questions and answers about language testing statistics* (1st ed.). JALT Testing and Evaluation Special Interest Group.

Calin-Jageman, R. J., & Cumming, G. (2019). The new statistics for better science: Ask how much, how uncertain, and what else is known. *The American Statistician*, *73*(Suppl. 1), 271-280. https://doi.org/10.1080/00031305.2018.1518266

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Psychology Press.

Cohen, J. (1994). The Earth is round (p < .05). *American Psychologist*, *49*(12), 997-1003. https://doi.org/10.1037/0003-066X.49.12.997

Collett, P. (2021). Graphical data analysis. In P. Clements, R. Derrah, & P. Ferguson (Eds.), *Communities of teachers & learners* (pp. 1-11). JALT. https://doi.org/10.37546/JALTPCP2020-01

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7-29. https://doi.org/10.1177/0956797613504966

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology*, *30*(1), 92-101. https://doi.org/10.5334/irsp.82

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy*, *98*, 19-38. https://doi.org/10.1016/j.brat.2017.05.013

Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*, *54*(2), 245-258. https://doi.org/10.1017/S0261444819000430

Gonulal, T. (2020). Statistical knowledge and training in second language acquisition: The case of doctoral students. *International Journal of Applied Linguistics*, *171*(1), 62-89. https://doi.org/10.1075/itl.18031.gon

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *p* values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*, 337-350. https://doi.org/10.1007/s10654-016-0149-3

Ho, J., Tumkaya, T., Aryal, S., Choi, H, & Claridge-Chang, A. (2019). Moving beyond *p* values: data analysis with estimation graphics. *Nature Methods*, *16*, 565-566. https://doi.org/10.1038/s41592-019-0470-3

Hudson, T. (2015). Presenting quantitative data visually. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 78-105). Routledge/Taylor & Francis.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137-152. https://psycnet.apa.org/doi/10.1037/a0028086

Khany, R. & Tazik, T. (2015). Levels of statistical use in applied linguistics research articles: From 1986 to 2015. *Journal of Quantitative Linguistics*, *26*(1), 48-65. https://doi.org/10.1080/09296174.2017.1421498

LaFlair, G. T., Egbert, J., & Plonsky, L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, t tests, and ANOVAs. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 46-77). Routledge/Taylor & Francis.

Larson-Hall, J. (2017). Moving beyond the bar plot and the line graph to create informative and attractive graphics. *The Modern Language Journal*, *101*(1), 244-270. https://doi.org/10.1111/modl.12386

Larson-Hall, J. (2021). Discussion paper: Using statistics to solve practical vocabulary problems. *Vocabulary Learning and Instruction*, *10*(2), 101-113. https://doi.org/10.7820/vli.v10.2.larson-hall

Larson-Hall, J., & Herrington, R. (2010). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics*, *31*(3), 368-390. https://doi.org/10.1093/applin/amp038

Larson-Hall, J. & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*(Suppl. 1), 127-159. https://doi.org/10.1111/lang.12115

Lindstromberg, S. (2016). Guidelines, recommendations, and supplementary discussion. *Language Teaching Research*, *20*(6). https://journals.sagepub.com/doi/suppl/10.1177/1362168816649979/suppl_file/10.1177_1362168816651895.pdf

Manicas, P. T. (1987). *A history and philosophy of the social sciences*. Basil Blackwell.

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, *37*(3), 389-411. https://doi.org/10.1177/0265532219898380

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195-244. https://doi.org/10.2466/pr0.1990.66.1.195

Mizumoto, A., & Plonsky, L. (2016). R as a lingua franca: Advantages of using R for quantitative research in applied linguistics. *Applied Linguistics*, *37*(2), 284-291. https://doi.org/10.1093/applin/amv025

Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, *65*(Suppl. 1), 97-126. https://doi.org/10.1111/lang.12114

Turner, J.L. (2014). *Using Statistics in Small-Scale Language Education Research: Focus on Non-parametric Data* (1st ed.). Routledge. https://doi.org/10.4324/9780203526927

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose, *The American Statistician*, *70*(2), 129-133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "*p*<0.05". *The American Statistician*, *73*(*Suppl. 1*), 1-19. https://doi.org/10.1080/00031305.2019.1583913

Wilcox. R. R. (2022). *Introduction to Robust Estimation and Hypothesis Testing* (5th ed.) Academic Press.

Ziliak, S. T., & McCloskey, D. N. (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. University of Michigan Press. https://doi.org/10.3998/mpub.186351