Introducing an Online Application for Estimation Plots

https://pcjapan.shinyapps.io/estimation_plots/

Paul Collett — collett@shimonoseki-cu.ac.jp

Rationale

- In light of ongoing critiques of the use of statistical significance testing, present an alternative way to interpret outcomes of quantitative research.
- Estimate effect sizes and their uncertainty rather than a dichotomous “accept-reject” perspective
- Provide standard measures researchers should understand alongside alternatives informed by recent commentary in quantitative methodology.
- Provide graphical data analysis tools for both exploratory and explanatory purposes.
- Deliver a user-friendly application built using modern open-source statistical analysis resources.

t-tests

- Used to assess differences between two means.
- Multiple variants exist. Be aware of the one you are using.
- Delacre, Lakens & Leys (2017) argue that Welch's *t*-test should be used when dealing with data that is non-homogeneous and is non-normal.
- Field & Wilcox (2017) put forward a case for Yuen's *t*-test, a **robust** alternative. Delacre, et al. suggest this test fails under certain circumstances.
- Both tests are seen as better than Student's *t*-test due to its strict assumptions.
- Suggested approach: Run both Welch & Yuen's test and report outcomes concurrently.

Statistical Significance Testing

What's the problem?

- Doesn't tell you what many think it tells you: A *p*-value from a statistical test tells you the probability of the observed data given that the null hypothesis is true.
- Null hypothesis is always false.
- Lack of power of tests.
- Issues with assumptions.
- Focussing on just *p*-value doesn't give details of effect size.

Robust Statistics

- Computer-intensive methods that adjust for issues in samples which violate assumptions of standard statistical tests.
- Alternative to non-parametric tests with more power and reliability.
- Common methods are trimmed means and bootstrapping.

Statistics | QQ Plots | Violin Plot | Estimation Plot

Basic Statistics

```
                V1     V2
median       12.500 16.000
mean         12.958 15.583
SE.mean       0.703  1.177
CI.mean.0.95  1.547  2.591
var           5.930 16.629
std.dev       2.435  4.078
coef.var      0.188  0.262
skewness      0.163  0.269
skew.2SE      0.128  0.211
kurtosis     -1.485 -1.268
kurt.2SE     -0.603 -0.515
normtest.W    0.954  0.949
normtest.p    0.701  0.617

95% Confidence interval for the mean:
Group 1: 11.411, 14.505
Group 2: 12.992, 18.174
```

t-test

- This gives a robust t-test with either 10% or 20% trimmed mean, and x number of bootstapped samples. Select the levels of the trimmed mean and bootstrapping below.
- The second t-test is a Welch test.

Trimmed Mean

Select the trimmed mean %

0.2

Bootstrap Samples

Set the number of bootstrap samples

200 5,000

200 700 1,200 1,700 2,200 2,700 3,200 3,700 4,200 4,700 5,000

Robust t-test and effect size

```
Yuen bootstrapped t-test for independent samples
(Robust test w/ 5000 bootstrapped samples)

t = -1.516, df = NA, p-value = 0.125
95% confidence interval:
-5.516  0.766
Mean difference for 20% trimmed means: -2.375
Effect size = -0.662, 95% confidence interval for effect size -1.838  0.136
```

Welch t-test and effect size

```
        Welch Two Sample t-test

data:  df$Response by df$Grouping
t = -1.9145, df = 17.96, p-value = 0.07162
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.5060144  0.2560144
sample estimates:
mean in group 1 mean in group 2
       12.95833        15.58333

Effect size: 0.782
```

Exploratory Analysis

- Larson-Hall (2017) : data accountable graphics (graphics showing all the relevant details of the dataset)

- data rich graphics (those that illustrate the data distribution and provide a large amount of information) to augment statistical results:

- "*...such graphics will firmly establish the credibility of the statistical arguments.*" (p. 244).

Exploratory and Explanatory Data Analysis

Subject ID:

1
2
3
4
5
6
7
8
9
10
11
12
1
2
3
4
5
6
7
8

Response:

10
15
11
13
15
11
9.5
12
12
17
14
16
11
16
12
23
13
10
21
17

Group:

1
1
1
1
1
1
1
1
1
1
1
1
2
2
2
2
2
2
2
2
2

Group 2:

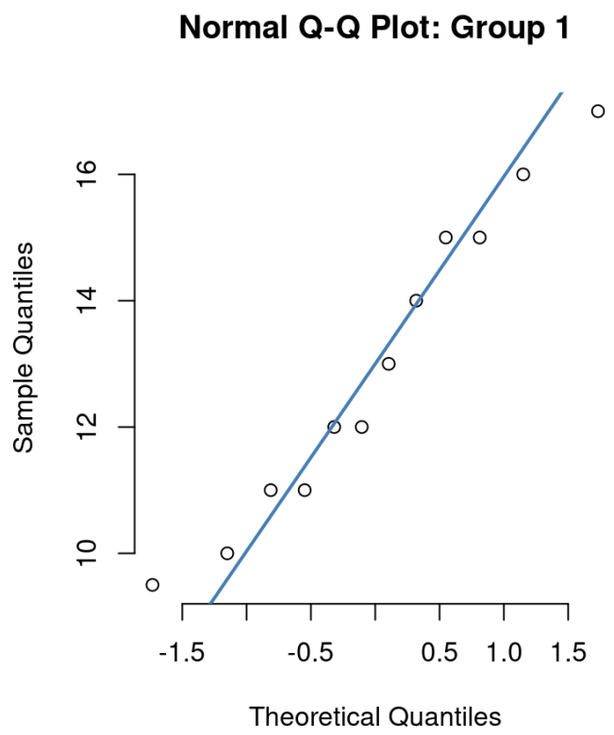
1
2
2
1
1
1
1
1
2
1
1
1
2
2
2
1
2
2
2

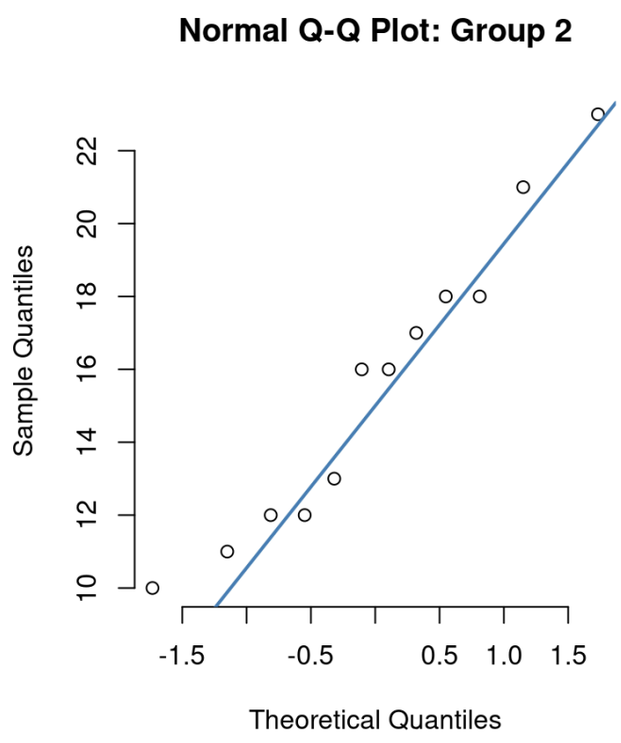
Is this a paired sample? (i.e the data is from the same subjects)

If yes, click the switch below.

Paired?

Statistics | QQ Plots | Violin Plot | Estimation Plot





Exploratory and Explanatory Data Analysis

Subject ID:

1
2
3
4
5
6
7
8
9
10
11
12
1
2
3
4
5
6
7
8

Response:

10
15
11
13
15
11
9.5
12
12
17
14
16
11
16
12
23
13
10
21
17

Group:

1
1
1
1
1
1
1
1
1
1
1
1
1
2
2
2
2
2
2
2
2
2
2
2

Group 2:

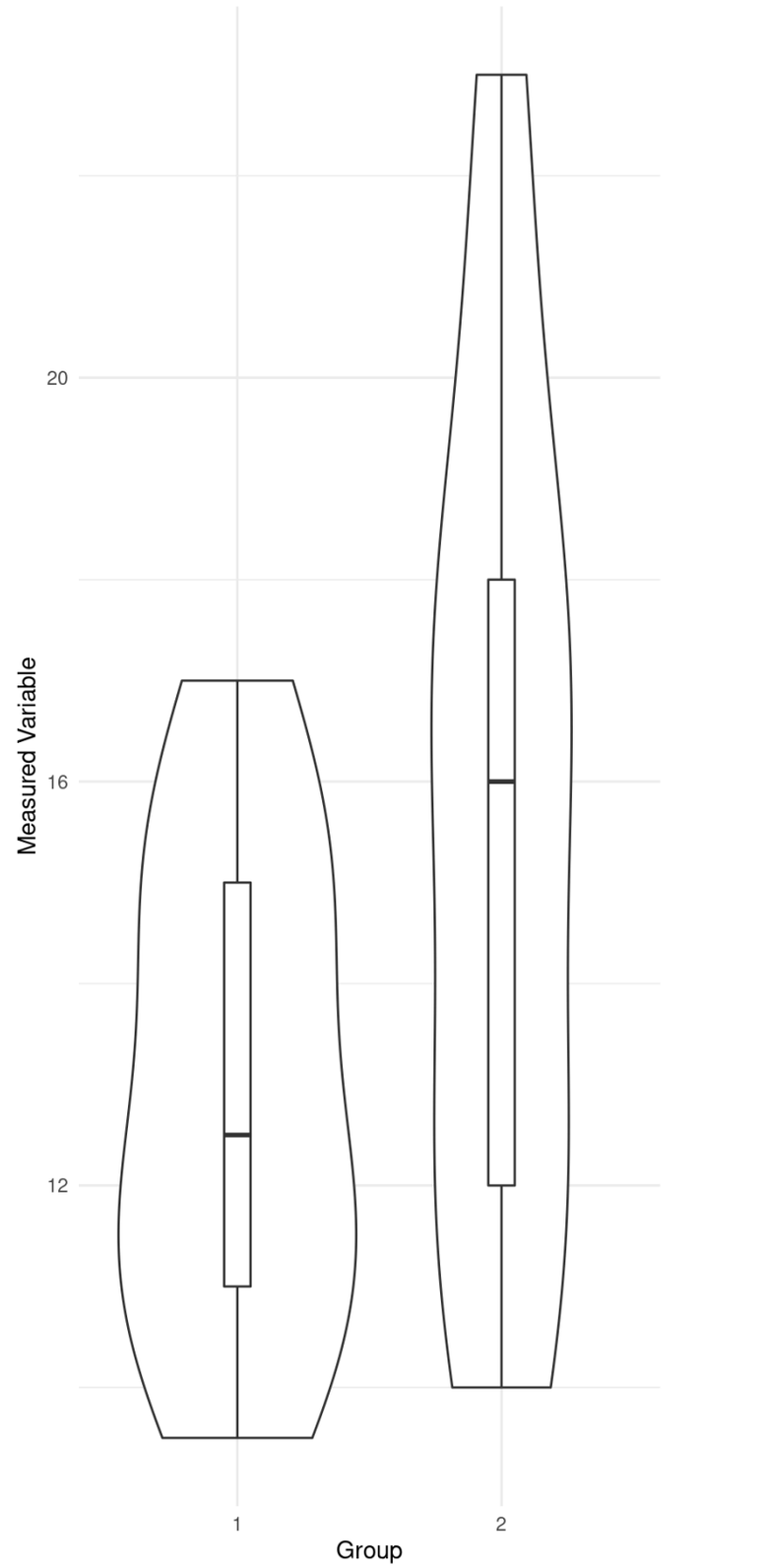
1
2
2
1
1
1
1
1
2
1
1
1
1
2
2
2
1
2
2
2
1

Is this a paired sample? (i.e the data is from the same subjects)

If yes, click the switch below.

Paired?

Statistics | QQ Plots | Violin Plot | Estimation Plot



Confidence Intervals

- CIs show a range within which you can be confident at the *x*% level of probability that the interval contains the true mean of the population.

- Comparing CIs provide as much information and more as that from statistical significance tests.

- Can be calculated for both measures of location (e.g., means) and effect sizes.

Effect Sizes

- Effect size is a indication of the magnitude of the experimental effect. The larger the effect size the stronger the relationship between two variables.
- Cohen (1988) describes effect size as “the degree to which the null hypothesis is false” (p.9).
- Provide a standardised way to report results focussing on the actual likelihood that what you have been testing for had some kind of impact.
- Recommended they should be reported in quantitative research results.
- Lack of effect sizes limits generalisability of research, especially for meta-analysis.
- Like CIs, problematic if you consider the general limits of the logic of statistical significance tests.

Estimation Plots

- Utilizes Ho, et al's (2019) Dabest package in R.
- Follows Cumming's (2012) *New Statistics* principles.
- Plots complete data set, along with effect size and associated CI.

Statistics | QQ Plots | Violin Plot | Estimation Plot

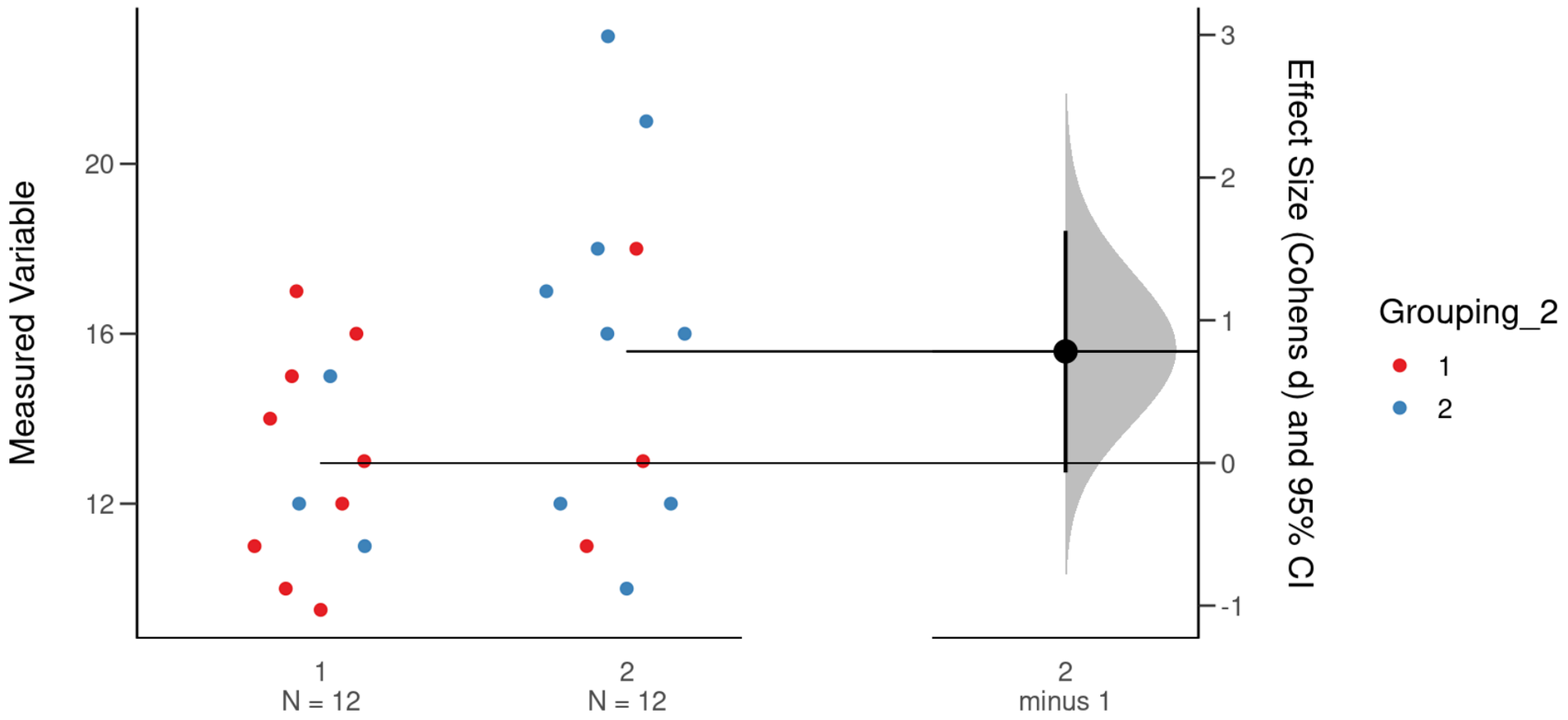
```
dabestr (Data Analysis with Bootstrap Estimation in R) v0.3.0
============================================================

Good evening!
The current time is 19:01 PM on Monday May 10, 2021.

Dataset    :  df
X Variable :  Grouping1
Y Variable :  Measured

Unpaired Cohen's d of 2 (n = 12) minus 1 (n = 12)
 0.782 [95CI  -0.0669; 1.63]


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



Future Aims

- Increase range of tests.
- Upload data from text file (csv).
- More customisation of labels, meta-data.
- Other suggestions?

- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New York: Routledge
- Cumming, G. (2012). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge.
- 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. http://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
- 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
- Larson–Hall, J. (2017). Moving beyond the bar plot and the line graph to create informative and attractive graphics. *The Modern Language Journal 101*, 244-270. https://doi.org/10.1111/modl.12386
- Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods 52,* 464–488. https://doi.org/10.3758/s13428-019-01246-w