One and Two Samples Using Only an R Function

Ying-Ying Zhang [1] Yi Wei [1]

[1] Chongqing University

Abstract

We create an R function one_two_sample() which deals with one and two (normal) samples. For one normal sample x, the function reports descriptive statistics, plot, interval estimations and hypothesis testings of the means and variances of x. For one abnormal sample x, the function reports descriptive statistics, plot, two sided interval estimation of the mean of x. For two normal samples x and y, the function reports descriptive statistics, plot, interval estimations and hypothesis testings of the means and variances of x and y, respectively. It also reports interval estimations and hypothesis testings of the difference of the means of x and y and the ratio of the variances of x and y, tests whether x and y are from the same population, finds the correlation coefficient of x and y if they have the same length. The function is in a tailor-made R package **OneTwoSamples** which is available on CRAN.

Keywords: one and two samples, interval estimation, hypothesis testing, mean, variance, R.

1. Introduction

R software (R Development Core Team 2013) has become more and more popular among researchers due to its freeness, handy and powerful programming language, coherent statistical analysis tools, superior statistical charting and many other advantages. The extensive data from industrial productions, financial economics, medical experiments and many other aspects, may result in one or two samples. First, we are interested in whether it is or they are normal. For one normal sample x, we are further interested in its descriptive statistics, plots (the histogram, the empirical cumulative distribution function (ECDF), the QQ plot), interval estimations and hypothesis testings of the means and variances of x. For two normal samples x and y, except for the descriptive statistics, plots, interval estimations and hypothesis testings of the means and variances of x and y, respectively. We are also interested in interval estimations and hypothesis testings of the difference of the means of x and y and the ratio of the variances of x and y, whether x and y are from the same population, and the correlation coefficient of x and y if they have the same length. All these interested information can be obtained by implementing one R function one_two_sample(), which is in a created R package **OneTwoSamples** available on CRAN (Zhang 2013).

Statistical inferences are main contents of mathematical statistics. Parametric estimation and hypothesis testing are two classical methods widely used in statistical inferences. They are treated in many statistics textbooks

(Casella and Berger 2002; DeCoursey 2003; Freedman et al. 2007; McClave et al. 2008; Ross 2009; Soong 2004; Walpole et al. 2011; Xue and Chen 2007; Yang et al. 2004). It is well known that the R built-in function t.test() can implement the interval estimation and hypothesis testing of one and two normal populations' mean. However, t.test() can neither accomplish those of the one normal population's mean when the population's variance is known, nor accomplish those of the two normal populations' mean when the populations' variances are known. Another R built-in function, var.test(), can implement the interval estimation and hypothesis testing of two normal populations' variances. However, var.test() can neither accomplish those of the one normal population's variance, nor accomplish those of the two normal populations' variances when the populations' means are known. Xue and Chen (2007) write twelve functions to implement all the interval estimations and hypothesis testings of the means and variances of one and two normal populations. See Table 1. In the table, the functions with blue text are superior to others since they still work when mu or sigma is known. '√' denotes the function can handle this case, while 'X' indicates it can not. Most of the functions can compute both one and two sided interval estimation and hypothesis testing except for those marked with 'two sided'. The functions listed above are applicable for normal sample(s). As for an abnormal sample, interval_estimate3() can implement the two sided interval estimation of the mean no matter the variance is known or not.

Table 1: The functions used in interval estimations and hypothesis testings of the means and variances of one and two normal samples.

	one sample		
mu	functions	sigma known	sigma unknown
Interval estimation	interval_estimate1() (two sided)	√	√
	interval_estimate4()	√	√
	t.test()	X	√
Hypothesis testing	mean_test1()	√	√
	t.test()	X	√
sigma	functions	mu known	mu unknown
Interval estimation	interval_var1() (two sided)	√	√
	interval_var3()	√	√
Hypothesis testing	var_test1()	√	√

	two samples			
mu	functions	sigma1, sigma2 known	sigma1= sigma2 unknown	sigma1!= sigma2 unknown
Interval estimation	interval_estimate2() (two sided)	√	√	√
	interval_estimate5()	√	√	√
	t.test()	X	√	√
Hypothesis testing	mean_test2()	√	√	√
	t.test()	X	√	√
sigma	functions	mu1 & mu2 known	mu1 or mu2 unknown	
Interval estimation	interval_var2() (two sided)	√	√	
	interval_var4()	√	√	
	var.test()	X	√	
Hypothesis testing	var_test2()	√	√	
	var.test()	X	√	

However, it is burdensome to remember and apply the functions in Table 1 in a flexible way. Zhang and Wei (2013) integrate them into one R function IntervalEstimate_TestOfHypothesis(). Users only need to input the sample(s) and set the parameters (test, mu, sigma, var.equal, ratio, side, alpha) as needed. It is convenient for users who merely care about the interval estimation and hypothesis testing of the mean or variance. The function one_two_sample() differs from IntervalEstimate_TestOfHypothesis() in many ways.

- Orientation

one_two_sample(): Deals with one or two (normal) samples. Reports descriptive statistics, plots, interval estimations and hypothesis testings of the means and variances of one or two normal samples. For two samples, tests whether x and y are from the same population, finds the correlation coefficient of x and y if they have the same length.

IntervalEstimate_TestOfHypothesis(): Implement interval estimation and hypothesis testing of the mean or variance of one or two normal samples.

- Outputs of interval estimation and hypothesis testing

one_two_sample(): For one normal sample, interval estimation and hypothesis testing of mu AND sigma. For two normal samples, interval estimation and hypothesis testing of mu AND sigma of x and y, respectively. Interval estimations and hypothesis testings of the difference of the means of x and y AND the ratio of the variances of x and y.

IntervalEstimate_TestOfHypothesis(): For one normal sample, interval estimation and hypothesis testing of mu OR sigma. For two normal samples, interval estimation and hypothesis testing of the difference of the means of x and y OR the ratio of the variances of x and y.

- Call functions of interval estimation and hypothesis testing

one_two_sample(): Directly call the following functions according to the input parameters:
interval_estimate4(),
interval_estimate5(),
mean_test1(), mean_test2(),
interval_var3(), interval_var4(),
var_test1(), var_test2(),
t.test(), var.test().

IntervalEstimate_TestOfHypothesis(): Make up the following four functions, and call them according to the input parameters:
single_mean(), single_var(),
double_mean(), double_var().

- Availability

one_two_sample(): An R package **OneTwoSamples** available on CRAN.

IntervalEstimate_TestOfHypothesis(): Through email to the author.

2. R function: one_two_sample()

The function one_two_sample() deals with one or two (normal) samples. In this section, we will introduce the usage and practical application of the function in detail.

2.1. Usage

The usage of one_two_sample() is as follows:

one_two_sample(x, y = NULL, mu = c(Inf, Inf), sigma = c(-1, -1), var.equal = FALSE, ratio = 1, side=0, alpha=0.05)

The meanings of the arguments of one_two_sample() can be obtained by typing "?one_two_sample" in the R console.

In Table 2, we further illustrate the usage of one_two_sample() by examples. All the examples are implemented in 'tests_OneTwoSamples.R' in the 'tests' folder of the package **OneTwoSamples**. In the table, each cell is divided into two parts. The upper part illustrates the usage of input parameters, and the lower part lists the functions called by one_two_sample().

Table 2. The usage of one_two_sample().

One normal sample	sigma known	sigma unknown
mu known	Example 1: x, mu =, sigma =, side = 0, alpha = 0.05	Example 3: x, mu =, side = 0, alpha = 0.05
	interval_estimate4(), mean_test1(), interval_var3(), var_test1()	t.test(), interval_var3(), var_test1()
mu unknown	Example 2: x, sigma =, side = 0, alpha = 0.05	Example 4: x, side = 0, alpha = 0.05
	interval_estimate4(), mean_test1(), interval_var3(), var_test1()	t.test(), interval_var3(), var_test1()
One abnormal sample	Example 5: x, sigma =, alpha = 0.05	Example 6: x, alpha = 0.05
	interval_estimate3()	inter-val_estimate3()

Two normal samples	mu1, mu2 known	mu1, mu2 unknown
sigma1, sigma2 known	Example 7: x, y, mu = c(,), sigma = c(,), side = 0, alpha = 0.05	Example 10: x, y, ratio = 1, sigma = c(,), side = 0, alpha = 0.05
	interval_estimate5(), mean_test2(), interval_var4(), var_test2()	interval_estimate5(), mean_test2(), var.test()
sig-ma1= sigma2 unknown	Example 8: x, y, mu = c(,), var.equal = TRUE, side = 0, alpha = 0.05	Example 11: x, y, ratio = 1, var.equal = TRUE, side = 0, alpha = 0.05
	t.test(), interval_var4(), var_test2()	t.test(), var.test()
sig-ma1 != sigma2 unknown	Example 9: x, y, mu = c(,), side = 0, alpha = 0.05	Example 12: x, y, ratio = 1, side = 0, alpha = 0.05
	t.test(), interval_var4(), var_test2()	t.test(), var.test()

2.2. Practical application

As mentioned earlier, one_two_sample() call other functions according to the input parameters. Thus the validity of one_two_sample() replies on those functions. In this section, we apply the function one_two_sample() to a dataset ‘women’ in the **datasets** package. Users are encouraged to apply the function to their own samples.

To use the function one_two_sample(), we should first: library("OneTwoSamples"). Note: the outputs explanations of a specific function can be obtained through the help page, for example, ‘?data_outline’, ‘?t.test()’.

```
## generate samples x and y
> x = women$height; x
 [1] 58 59 60 61 62 63 64 65 66 67 68
69 70 71 72
> y = women$weight; y
 [1] 115 117 120 123 126 129 132 135
139 142 146 150 154 159 164

## operate on one sample
## one_two_sample(x) is equivalent to
one_sample(x)
> one_two_sample(x)
Outputs are omitted to save space.

## one_two_sample(y) is equivalent to
one_sample(y)
> one_two_sample(y)
Outputs are omitted to save space.
```

Illustration: The outputs of one_two_sample(x) and one_two_sample(y) can be obtained by running the above R code lines. For x, first the function reports descriptive statistics (the quantile of x and the data outline of x). Then in Shapiro-Wilk normality test, p-value = 0.7545 > 0.05, so the data x is from the normal population. After that, the 3 plots show the histogram, the ECDF, and the QQ plot of x. The 3 plots all indicate that the data x is from the normal population, in agree with the result of Shapiro-Wilk normality test. Finally, the function displays interval estimations and hypothesis testings of the means and variances of x. The interval estimation and hypothesis testing of mu call the function t.test(). We find that the 95 percent confidence interval of mu is [62.52341, 67.47659], the p-value < 2.2e-16 < 0.05, so reject H0: mu = 0 and accept H1: mu != 0. The interval estimation of sigma calls the function interval_var3(). We find that the 95 percent confidence interval of sigma is [10.72019, 49.74483]. The hypothesis testing of sigma calls the function var_test1(). We find that the P_value = 0 < 0.05, so reject H0: sigma2 = 1 and accept H1: sigma2 != 1. The explanations of the outputs of one_two_sample(y) are omitted.

```
## operate on two samples
> one_two_sample(x, y)
Outputs are omitted to save space.
```

Illustration: The outputs of one_two_sample(x, y) can be obtained by running the above R code lines. The explanations of the results for one sample x and y are omitted, since they have already been explained before. The interval estimation and hypothesis testing of mu1 - mu2 call the function t.test(). We find that the 95 percent confidence interval of mu1 - mu2 is [-80.54891, -62.91775], the p-value = 6.826e-12 < 0.05, so reject H0: mu1 = mu2 and accept H1: mu1 != mu2. The interval estimation and hypothesis testing of sigma1^2 / sigma2^2 call the function var.test(). We find that the 95 percent confidence interval of sigma1^2 / sigma2^2 is [0.02795306, 0.24799912], the p-value = 3.586e-05 < 0.05, so reject H0: sigma1^2 = sigma2^2 and accept H1: sigma1^2 != sigma2^2. We obtain n1 == n2, i.e., x and y have the same length. Three functions ks.test(), binom.test(), and wilcox.test() are used to test whether x and y are from the same population. Three pvalues are all less than 0.05, so reject H0: x and y are from the same population. The function cor.test(x, y, method = c('pearson', 'kendall', 'spearman')) is used to find the correlation coefficient of x and y. Three p-values are all less than 0.05, so reject H0: rho = 0 (x, y uncorrelated). Thus x and y are correlated. In fact, x and y have nearly 1 correlation.

3. Conclusions

The function one_two_sample() can deal with one and two (normal) samples. The function is in a tailor-made R package **OneTwoSamples** which is available on CRAN. In addition, the usage of arguments of one_two_sample() is straightforward. It will simplify the users' operations of dealing with one and two (normal) samples to a great extent.

4. Acknowledgment

The research was supported by Natural Science Foundation Project of CQ CSTC CSTC2011BB0058.

5. References

[1] Casella G, Berger RL (2002). Statistical Inference. Duxbury, United States. 2nd edition.
[2] DeCoursey W (2003). Statistics and Probability for Engineering Applications with MS Excel. Newnes, New York.
[3] Freedman D, Pisani R, Purves R (2007). Statistics. W. W. Norton & Company, New York. 4th edition.
[4] McClave JT, Benson PG, Sincich T (2008). A First Course in Business Statistics. Prentice Hall, New Jersey. 8th edition.
[5] R Development Core Team (2013). R: A Language and Environment

for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http: //www.R-project.org/.

[6] Ross S (2009). Introduction to Probability and Statistics for Engineers and Scientists. Elsevier Academic Press, USA. 4th edition.

[7] Soong TT (2004). Fundamentals of Probability and Statistics for Engineers. John Wiley & Sons, USA.

[8] Walpole RE, Myers RH, Myers SL, Ye KE (2011). Probability and Statistics for Engineers and Scientists. Pearson, New York. 9th edition.

[9] Xue Y, Chen LP (2007). Statistical Modeling and R Software. Tsinghua University Press, Beijing. This is a Chinese book.

[10] Yang H, Liu QS, Zhong B (2004). Mathematical Statistics. China Higher Education Press, Beijing. This is a Chinese book.

[11] Zhang YY (2013). **OneTwoSamples**: Deal with One and Two (Normal) Samples. R package version 1.0-3, URL http://CRAN.R-project.org/package=OneTwoSamples.

[12] Zhang YY, Wei Y (2013). "Implement All the Interval Estimations and Hypothesis Testings of the Means and Variances of Normal Populations in One R Function." Statistics and Decision. This is a Chinese journal, accepted for publication.