The use of R language in the teaching of statistical inference

Wai Kwong Cheang

8th Southern Hemisphere Conference on Teaching and Learning Undergraduate Mathematics and Statistics, Rotorua, New Zealand, 27 November - 2 December 2011 (pp. 44 – 52)

This document may be used for private study or research purpose only. This document or any part of it may not be duplicated and/or distributed without permission of the copyright owner.

The Singapore Copyright Act applies to the use of this document.

Citation:

This document was archived with permission from the copyright holder.
The Use of R Language in the Teaching of Statistical Inference

Wai Kwong Cheang

Mathematics and Mathematics Education Academic Group
National Institute of Education, Nanyang Technological University, Singapore
waikwong.cheang@nie.edu.sg

Statistical software packages like Minitab are commonly used in the teaching of statistics. The strength of these packages lies in the ease of performing standard statistical procedures such as hypothesis testing. However, such an ease can also be a hindrance to the understanding of the statistical principles underlying the procedures. Furthermore, these packages are not free and may not be cost-effective to implement. The R language is a powerful software for data analysis within which many statistical procedures have been implemented. The strength of R derives from its many capabilities besides being a data analysis tool. In this paper, we explore the simulation and graphing capabilities of R in teaching statistical inference, so as to enhance students’ conceptual understanding of the inference methodology.

Keywords: R language; statistical inference; simulation

Introduction

The R language is a powerful software for data analysis within which many statistical procedures have been implemented. R was initially written by Robert Gentleman and Ross Ihaka of the Statistics Department of the University of Auckland, and can be considered as a “free” implementation of the S language, which was originally developed at Bell Laboratories. R is an official part of the Free Software Foundation’s GNU Project, and is distributed as Free Software under the terms of the GNU General Public License. Precompiled R binaries as well as the R source code can be downloaded from http://www.r-project.org/.

As R is highly flexible and extensible, the use of R has not been limited to data analysis and statistical research. Coupled with its open-source nature and free availability, R is now commonly used in the teaching of statistics. For example, Dalgaard (2008) and Verzani (2005) considered using R for introductory statistics. Using a series of case studies and activities, Horton, Brown and Qian (2004) described how R can be used in a mathematical statistics course as a toolbox for exploration. In emphasizing that data analysis cannot be learned without actually doing it, Faraway (2005) has chosen to use R to teach linear models due to its “versatility, interactivity, freedom and popularity”. Exploiting the versatility of R as both a statistical software and a programming language, Cheang (2007) discussed how R can be used to enhance students’ understanding of the principles in regression modelling, by providing insight into the “black box” which generates the regression output.

An attractive feature of R is its simulation capability. We can easily simulate observations from common distributions like normal and exponential. For example, the R command
\[ rnorm(30, 0, 1) \]
generates a random sample of size \( n = 30 \) from \( \mathcal{N}(0, 1) \) distribution. In discussing the use of R in teaching financial mathematics and statistics, Stander and Eales (2011) showed how the simulation capability of R can be employed to illustrate features of the Monte Carlo methodology in estimating option price.

Another attractive feature of R is its graphing capability. For example, in teaching continuous bivariate distribution, the author has used R to draw perspective plots of a
joint density surface from different viewing angles. This approach has allowed students to “see” the probability (as volume under the surface) over a certain region of the x-y plane. R has been adopted by Keen (2010) in discussing graphics for statistics and data analysis “because its graphing capabilities are state-of-the-art”.

Using the flexible graphing capability of R, simulation results can easily be presented in publication-quality plots according to user’s need. In teaching the Central Limit Theorem (CLT), Cheang (2009) explored how the simulation and graphing capabilities of R can be used to provide students with better insight into the issues encountered. For example, how large should the sample size n be for CLT to hold? Is the rule of thumb $n \geq 30$ adequate? In this paper, we explore how these two capabilities of R can be employed to implement some teaching strategies for statistical inference. Based on the author’s experience, these strategies help to enhance students’ conceptual understanding of the inference methodology.

**Use of R in Teaching Confidence Intervals**

In statistical inference, one routine procedure is the construction of confidence intervals to estimate a population parameter. One way to interpret a confidence interval, say 90%, is that if we repeatedly draw samples from the population, then the calculated interval will contain the parameter for 90% of the samples drawn. We can impress this interpretation upon students through simulation of intervals. Also, students often (wrongly) think that the population parameter is a random quantity (because it is unknown), while the interval is fixed (because it is calculated). The simulation would help to eradicate such misconception.

*Simulation of t-intervals for $\mu$ from normal population with unknown variance*

Suppose we want to estimate the mean $\mu$ of a normal population with unknown variance based on a sample of size $n$. A 100(1−$\alpha$)% confidence interval for $\mu$ is given by

$$\bar{x} \pm t_{n-1}^{(\alpha/2)} \frac{s}{\sqrt{n}},$$

where $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ is the sample variance, and $t_{n-1}^{(\alpha/2)}$ denotes the $(1-\alpha/2)$-quantile of the $t_{n-1}$-distribution.

Wild and Seber (2000, p. 332) presented simulated 95% $t$-intervals for $\mu$ from a normal population. The R code in Appendix A.1 shows how such a simulation can be performed using R. Figure 1 illustrates simulated 90% $t$-intervals for $\mu$ from normal population with $\mu = 50$, when the sample size is $n = 10$. For the 20 samples shown, we see that the intervals based on two of the samples (indicated by “out”) do not contain $\mu$. This is indeed consistent with the interpretation of 90% confidence intervals. In addition, by showing the steps that calculate these intervals, the R code will help students understand the “black box” which generates them in a typical inference output.
Figure 1. Simulation: 90% t-intervals for \( \mu \) from normal population with unknown variance (\( \mu = 50, n = 10 \)).

Simulation of t-intervals for \( \mu_1 - \mu_2 \) from normal populations with equal variances

Suppose we want to compare the means \( \mu_1 \) and \( \mu_2 \) from two normal populations with equal variances based on independent samples of sizes \( n_1 \) and \( n_2 \). A 100(1-\(\alpha\))% confidence interval for \( \mu_1 - \mu_2 \) is given by

\[
\bar{x}_1 - \bar{x}_2 \pm t_{(\alpha/2)} s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},
\]

where \( s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \) is the pooled variance.

Figure 2. Simulation: 90% t-intervals for \( \mu_1 - \mu_2 \) from normal populations with equal variances (\( \mu_1 = \mu_2 = 0, n_1 = 20, n_2 = 10 \)).

Figure 2 displays simulated 90% t-intervals for \( \mu_1 - \mu_2 \) from normal populations with equal variances, when the sample sizes are \( n_1 = 20 \) and \( n_2 = 10 \). The R code to perform this simulation is given in Appendix A.2. Since one usual purpose of constructing an interval for \( \mu_1 - \mu_2 \) is to determine whether \( \mu_1 \) and \( \mu_2 \) are significantly different, we take \( \mu_1 = \mu_2 \) in simulating these intervals. To present both the data points and the intervals on the same plot, we further take \( \mu_1 = \mu_2 = 0 \). For the 20 replications shown, we see that the intervals based on four of the replications (indicated by “out”) do not contain the difference \( \mu_1 - \mu_2 = 0 \). We can verify through simulation of further samples that indeed approximately 10% of the intervals would not contain 0.
Simulation of $\chi^2$-statistic for sample variance

For a normal population with variance $\sigma^2$, we know that the sample variance $S^2$ has a $\chi^2$-distribution given by

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$  

The 100(1−$\alpha$)% confidence interval for $\sigma^2$ constructed based on this distribution is

$$\left[ \frac{(n-1)s^2}{\chi^2_{n-1}(\alpha/2)} , \frac{(n-1)s^2}{\chi^2_{n-1}(1-\alpha/2)} \right],$$

where $\chi^2_{n-1}(\alpha/2)$ denotes the (1−$\alpha$/2)-quantile of the $\chi^2_{n-1}$-distribution. Notice that this interval does not have the usual form of parameter estimate ± critical value × st. dev. of estimate.

The unusual form of the interval is a direct consequence of the unusual form of the statistic $V = (n-1)S^2/\sigma^2$ from which the interval is derived. To impress this “anomaly” upon students, one approach is to convince students that such unusual form is needed to give a $\chi^2$-distribution. The R code in Appendix A.3 shows how the distribution of $V$ can be simulated. In Figure 3(a), the density histogram displays the empirical distribution of $V$ for 10000 samples of size $n = 30$ drawn from a normal population. The $\chi^2_{n-1}$ p.d.f. superimposed indicates that $V$ indeed has a $\chi^2$-distribution. The normal p.d.f. (with the same mean and variance as that of $\chi^2_{n-1}$) superimposed would help to eradicate the common misconception that $V$ has a normal distribution.

**Figure 3.** Simulation: Empirical distribution of $V = \frac{(n-1)S^2}{\sigma^2}$ for normal and exponential populations ($n = 30$), with $\chi^2_{n-1}$ (solid curve) and $N(n-1, 2(n-1))$ (dotted curve) p.d.f.’s superimposed.

Furthermore, when learning the interval for $\sigma^2$, students often (wrongly) think that the normality condition is not necessary when $n$ is large (due to their misuse of Central Limit Theorem). Figure 3(b) shows the density histogram obtained for $n = 30$ when the population is exponential. The $\chi^2_{n-1}$ p.d.f. superimposed does not match the empirical distribution of $V$, thus indicating the necessity of a normal population in constructing confidence interval for $\sigma^2$, even when $n$ is large.
Use of R in Teaching Hypothesis Testing

In statistical inference, the procedures of hypothesis testing and confidence interval estimation are closely related. For example, in testing the null hypothesis \( H_0: \mu = \mu_0 \) against a two-sided alternative \( H_1: \mu \neq \mu_0 \), \( H_0 \) is rejected at the \( \alpha \) level of significance if and only if \( \mu_0 \) falls outside the 100(1 - \( \alpha \))% confidence interval for \( \mu \).

The R code in Appendix A.1 can easily be modified to include calculation of the one-sample \( t \)-statistic,

\[
T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}
\]

For 90% \( t \)-intervals in Figure 1 that do not contain \( \mu_0 = 50 \), we can verify that \( |T| > t_{n-1}^{(0.05)} \), where \( n = 10 \) is the sample size. Furthermore, we can demonstrate to students that approximately 10% of the samples generated from a normal population with \( \mu = 50 \) would actually have \( H_0: \mu = 50 \) rejected. This should enhance their understanding of the meaning of \( \alpha \).

Simulation of Welch’s \( t \)-statistic and two-sample \( t \)-statistic

When testing \( H_0: \mu_1 = \mu_2 \) for two normal populations and equality of variances cannot be assumed, one statistic that can be used is the Welch’s \( t \)-statistic \( T_1 \) given by

\[
T_1 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}
\]

\( T_1 \) has a \( t \)-distribution under \( H_0 \), where

\[
\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\frac{s_1^2}{n_1} + \frac{1}{n_2-1}\frac{s_2^2}{n_2}}
\]

The two-sample \( t \)-statistic \( T_2 \) requires the two population variances \( \sigma_1^2 \) and \( \sigma_2^2 \) to be equal, so that

\[
T_2 = \frac{\bar{X}_1 - \bar{X}_2}{\frac{1}{n_1} + \frac{1}{n_2}} \sim t_{n_1+n_2-2} \text{ under } H_0.
\]

As a computer assignment, we can use the simulation capability of R to study the empirical levels of \( T_1 \) and \( T_2 \). To assess the impact of assuming equal variances on the empirical levels, we consider different ratios of \( k = \sigma_1/\sigma_2 \), including \( k = 1 \). The R code to perform this simulation is given in Appendix A.4. Table 1 gives the empirical levels of nominal 5%-level \( t \)-test of \( H_0: \mu_1 = \mu_2 \) versus \( H_1: \mu_1 \neq \mu_2 \) when \( n_1 = 20 \) and \( n_2 = 10 \) (10000 replications), that is, the proportions of replications in which

\[
|T_1| > t_{\nu}^{(0.025)} \quad \text{and} \quad |T_2| > t_{n_1+n_2-2}^{(0.025)}.
\]

These simulation results seem to suggest that for \( n_1 > n_2 \), the empirical level of \( T_2 \) is always less than that of \( T_1 \). Can we then conclude \( T_2 \) is “better” than \( T_1 \) even when the variances are unequal, in the sense that \( T_2 \) has a smaller probability of Type I error? What if \( n_1 < n_2 \) or \( n_1 = n_2 \)? How about the probability of Type II error? With R “freely” available, such exploratory questions can be further investigated by students in project work.
Table 1. Empirical levels of nominal 5%-level $t$-test of $H_0: \mu_1 = \mu_2$ versus $H_1: \mu_1 \neq \mu_2$ ($\mu_1 = \mu_2 = 0$, $n_1 = 20$, $n_2 = 10$), based on 10000 replications.

<table>
<thead>
<tr>
<th>$k = \sigma_1/\sigma_2$</th>
<th>Welch's $t$-statistic $T_1$</th>
<th>Two-sample $t$-statistic $T_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0486</td>
<td>0.0461</td>
</tr>
<tr>
<td>2</td>
<td>0.0492</td>
<td>0.0206</td>
</tr>
<tr>
<td>5</td>
<td>0.0519</td>
<td>0.0101</td>
</tr>
</tbody>
</table>

Conclusion

By exploiting its simulation and other capabilities, R has the potential to be an effective teaching and learning tool for statistics. In a sense, R is an “evolving” language as every user is also a developer. With its ease of adaptability according to user’s need, the R language provides a platform for statistics educators to “freely” explore how technology can be integrated into their teaching to enhance students’ understanding.

References


Appendix: R Codes

In this appendix, we give the R codes for some of the examples discussed. A good introduction to the R language is The R Manual edited by the R Development Core Team and downloadable from http://www.r-project.org/.

A.1: Simulation of $t$-intervals for $\mu$ with unknown variance

```r
set.seed(238)
alpha <- 0.10
n <- 10
mu <- 50
sigma <- 3

# t-interval
cvalue <- qt(1-alpha/2, df=n-1)

nr <- 2 # No. of plots
r <- 10 # No. of samples per plot

for (j in 1:nr)
```

49
\{ x <- matrix(NA,r,n)

\textbf{for} (i in 1:r)
\{ x[i,] <- rnorm(n,mu,sigma)
\}

\textbf{plot}(0,0,\textbf{type}="n",\textbf{bty}="n",\textbf{axes}=F,\textbf{xlab}="",\textbf{ylab}="",\textbf{mgp}=c(2,0.5,0),
\textbf{xlim}=mu+c(-1,1)*4*sigma,\textbf{ylim}=c(0,r+1),\textbf{cex}=0.7)
\textbf{axis}(side=1,\textbf{at}=\textbf{seq}(mu-4*sigma,mu+4*sigma),\textbf{mgp}=c(2,0.5,0),\textbf{cex}=0.7)
\textbf{abline}(v=mu,\textbf{lty}=2)

\textbf{for} (i in 1:r)
\{ \textbf{points}(x[i,],\textbf{rep}(i,n),\textbf{pch}=21,\textbf{cex}=0.7)
\textbf{text}(mu-4*sigma,i,as.character(10*(j-1)+i),\textbf{adj}=0,\textbf{cex}=0.8)
\textbf{xbar} <- \textbf{mean}(x[i,])
\textbf{s} <- \textbf{sd}(x[i,])

\textbf{# sigma unknown, n small}
\textbf{lower} <- \textbf{xbar} - \textbf{cvalue} * s / \textbf{sqrt}(n)
\textbf{upper} <- \textbf{xbar} + \textbf{cvalue} * s / \textbf{sqrt}(n)

\textbf{segments}(\textbf{lower},i+0.15,\textbf{upper},i+0.15,\textbf{lty}=1,\textbf{col}="blue")
\textbf{points}(\textbf{c}(\textbf{lower},\textbf{upper}),\textbf{rep}(i+0.15,2),\textbf{pch}="|",\textbf{col}="blue",\textbf{cex}=0.8)
\textbf{if} (\textbf{lower}>mu | \textbf{upper}<mu)
\textbf{text}(mu+4*sigma,i+0.1,"out",\textbf{col}="blue",\textbf{adj}=1,\textbf{cex}=0.8)
\}
\textbf{text}(mu-4*sigma,r+1,"Sample No.",\textbf{adj}=0,\textbf{cex}=0.8)
\}

\textbf{A.2: Simulation of t-intervals for} \mu_1 - \mu_2 \textbf{with equal variances}

\textbf{set.seed}(238)
\textbf{alpha} <- 0.10
\textbf{n1} <- 20
\textbf{n2} <- 10
\textbf{mu1} <- 0
\textbf{mu2} <- 0
\textbf{sigma} <- 3 # \textbf{sigma1} = \textbf{sigma2} = \textbf{sigma}

\textbf{# t-interval}
\textbf{cvalue} <- \textbf{qt}(1-alpha/2,n1+n2-2)
\textbf{nr} <- 2 # \textbf{No. of plots}
\textbf{r} <- 10 # \textbf{No. of replications per plot}

\textbf{for} (\textbf{j} in 1:\textbf{nr})
\{ \textbf{x} <- \textbf{matrix}(NA,r,n1)
\textbf{y} <- \textbf{matrix}(NA,r,n2)

\textbf{for} (\textbf{i} in 1:r)
\{ \textbf{x}[i,] <- \textbf{rnorm}(n1,mu1,sigma)
\textbf{y}[i,] <- \textbf{rnorm}(n2,mu2,sigma)
\}

\textbf{plot}(0,0,\textbf{type}="n",\textbf{bty}="n",\textbf{axes}=F,\textbf{xlab}="",\textbf{ylab}="",\textbf{mgp}=c(2,0.5,0),
\textbf{xlim}=4*sigma*c(-1,1),\textbf{ylim}=c(0,r+1),\textbf{cex}=0.7)
\textbf{axis}(side=1,\textbf{at}=\textbf{seq}(-4*sigma,4*sigma),\textbf{mgp}=c(2,0.5,0),\textbf{cex}=0.7)
\textbf{abline}(v=0,\textbf{lty}=2)

\textbf{for} (\textbf{i} in 1:r)
A.3: Simulation of empirical distribution of sample variance

```r
set.seed(238)
n <- 30
mu <- 2 # mu = sigma for exponential
sigma <- 2
r <- 10000 # No. of samples
Vstat <- rep(NA,r)
for (i in 1:r)
{
x <- rnorm(n,mu,sigma) # Normal population
# x <- rexp(n,rate=1/mu) # Exponential population
s2 <- var(x)
# V-statistic
Vstat[i] <- (n - 1)*s2/sigma^2
}
upper <- 8*sqrt(2*(n-1))
chisqpdf <- dchisq(seq(0,upper,0.1),n-1)
Npdf <- dnorm(seq(0,upper,0.1),sqrt(2*(n-1)),cex=0.7)
# Plot density histogram of V-statistic
hist(Vstat,breaks=seq(min(Vstat),max(Vstat)+1,1),prob=T,right=T,main="",xlim=c(0,upper),ylim=c(0,max(chisqpdf)),mgp=c(2,0.5,0),cex=0.7)
# Plot chisq-pdf
lines(seq(0,upper,0.1),chisqpdf,lty=1,col="red")
# Plot normal pdf
lines(seq(0,upper,0.1),Npdf,lty=2,col="blue")
mtext(side=3,line=1,outer=F,"(a) Normal population",cex=1.0)
# mtext(side=3,line=1,outer=F,"(b) Exponential population",cex=1.0)
mtext(side=3,line=0,outer=F,paste("m = ",mu," , s = ",sigma,sep=""),font=5,cex=1.0)
```
A.4: Simulation of empirical levels of Welch’s t-statistic and two-sample t-statistic for testing $\mu_1 = \mu_2$

```r
set.seed(238)
alpha <- 0.05
n1 <- 20
n2 <- 10
mu1 <- 0
mu2 <- 0
k <- 2
sigma2 <- 1
sigma1 <- k*sigma2

r <- 10000 # No. of replications
tstat <- matrix(NA,r,2)
cvalue <- rep(NA,2)
cvalue[2] <- qt(1-alpha/2,n1+n2-2)
# Count no. of times H0: $\mu_1 = \mu_2$ (vs H1: $\mu_1 \neq \mu_2$) is rejected
count <- rep(0,2)
names(count) <- c("Welch's t","Two-sample t")

for (i in 1:r)
{  
x <- rnorm(n1,mu1,sigma1)
y <- rnorm(n2,mu2,sigma2)
xbar <- mean(x)
ybar <- mean(y)
s1 <- sd(x)
s2 <- sd(y)
sp <- sqrt(((n1-1)*(s1^2) + (n2-1)*(s2^2))/(n1+n2-2))

# Welch's t-statistic
tstat[i,1] <- (xbar - ybar)/sqrt(s1^2/n1 + s2^2/n2)

# Two-sample t-statistic
tstat[i,2] <- (xbar - ybar)/(sp*sqrt(1/n1 + 1/n2))

a <- s1^2/n1
b <- s2^2/n2
da <- ((a + b)^2)/(a^2/(n1-1) + b^2/(n2-1))
nu <- floor(da)
cvalue[1] <- qt(1-alpha/2,nu)

if (abs(tstat[i,1]) > cvalue[1])
{  
}

if (abs(tstat[i,2]) > cvalue[2])
{  
}

level <- count/r
print(level)
```

52