Example 1: Assessing the Robustness of the One-Sample t-test

Sarah C. Anoke, Nicholas J. Horton*, Yuting Zhao
Department of Mathematics and Statistics
Smith College

July 9, 2012

Contents

1 Introduction

Many scientific computations can be sped up by dividing them into smaller tasks and distributing the computations to multiple systems for simultaneous processing. Such a process is referred to as *parallel computing*. When performed on existing grids of computers, this method can dramatically decrease computation time. Several solutions exist to facilitate this type of computation within R, and we describe one such solution here, that involves using the Apple Xgrid (Apple, 2009), a parallel computing environment.

We created the **xgrid** package to provide a simple interface to this distributed computing system (Anoke et al., 2012). The package facilitates use of an Apple Xgrid for distributed processing of a job with many independent repetitions, by simplifying task submission (or *gridstuffing*) and collation of results. We demonstrate use of our package in the context of a real, although relatively simple, statistical problem.

*Corresponding author: nhorton@smith.edu

2 One-sample t-test

The t-test is remarkedly robust to violations of its underlying assumptions (Sawiloswky and Blair, 1992). However, as Hesterberg (2008) argues, not only is it possible for the total non-coverage to exceed α, the asymmetry of the test statistic causes one tail to account for more than its share of the overall α level. Hesterberg found that sample sizes in the thousands were needed to get symmetric tails.

In this example, we demonstrate how to utilize an Apple Xgrid cluster to investigate the robustness of the one-sample t-test, by looking at how the α level is split between the two tails. When the number of simulations is small ($< 100,000$), this study runs very quickly as a loop in R. However here we provide a study consisting of 10^6 simulations, and compare the results and computation time to the same study run on a local machine.

3 Using the Grid for a Simulation Study

3.1 Directory Structure

Our first step is to set up an appropriate directory structure for our simulation (Figure 1).

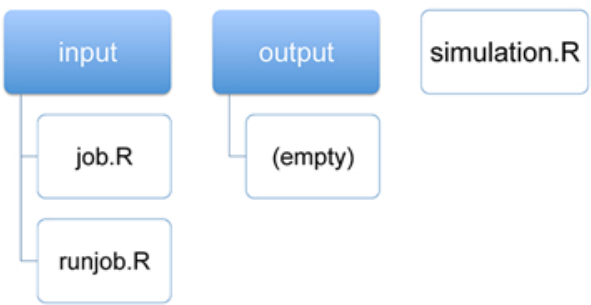

Figure 1: File structure to access the grid

The first item is the directory 'input', which contains two files that will be run on the remote agents. The first of these files, 'job.R', defines the code to run a particular job (Figure 2).

For this example, the `job()` function begins by generating a sample of `param` exponential random variables with mean 1. A one-sample t-test is conducted on this sample, and logical (`TRUE/FALSE`) values denoting whether the test rejected in that tail are saved in the vectors `leftreject` and `rightreject`. This process is repeated `ntask` times, after which the function `job()` returns a data frame with the rejection results and the corresponding sample size.

The folder 'input' also contains 'runjob.R', which retrieves and stores command line arguments from the controller, and passes them to `job()` (Figure 3). The results from the completed job are saved as `res0`, which is subsequently saved to the 'output' folder.

The folder 'input' may also contain other files needed for the simulation. In this example, no additional files are needed.

```
> # Assess the robustness of the one-sample
> # t-test when underlying data are exponential
> # this function returns a dataframe with
> # number of rows equal to the value of "ntask"
> # the option "param" specifies the sample size
> job <- function(ntask, param) {
   alpha <- 0.05         # how often to reject under null
         leftreject <- logical(ntask)   # placeholder
         rightreject <- logical(ntask)  # for results
         for (i in 1:ntask) {
                 dat <- rexp(param)   # generate skewed data
                 left <- t.test(dat, mu=1,
                                          alternative="less")
                 leftreject[i] <- left$p.value <= alpha/2
                 right <- t.test(dat, mu=1,
                                           alternative = "greater")
                 rightreject[i] <- right$p.value <= alpha/2
         }
         return(data.frame(leftreject, rightreject,
                                         n=rep(param, ntask)))
 }
```

Figure 2: Contents of 'job.R'

```
> source("job.R")
> # commandArgs() is expecting three arguments:
> # 1) number of tasks to run within this job
> # 2) parameter to pass to the function
> # 3) place to stash the results when finished
> args     <- commandArgs(trailingOnly = TRUE)
> ntask1   <- as.numeric(args[1])
> param1   <- args[2]
> resfile  <- args[3]
> res0     <- job(ntask = ntask1, param = param1)
> # stash the results
> saveRDS(res0, file = resfile)
```

Figure 3: Contents of 'runjob.R'

The next item in the directory structure is the folder 'output', which will contain results from the simulations. If it doesn't exist, the **xgrid** package will create it. The 'output' directory has a complete listing of the individual results as well as the R output from the remote agents. This can be useful for debugging in case of problems.

The final item in the directory structure is 'simulation.R', which contains an R script to be run on the client machine that calls `xgrid()` (Figure 4). This function submits the simulations to the grid for calculation. Results from all jobs are returned as one object, `res`. The call to `with()` summarizes all results in a table and prints them to the console.

```
> library(xgrid)
> # run the simulation
> res = xgrid(grid="Burton-303-iMac", Rcmd="runjob.R", param=30,
    numsim=10^6, ntask=5*10^4)
> # analyze the results
> with(res, table(leftreject,rightreject))
```

Figure 4: Contents of 'simulation.R'

Here we specify a total of 10^6 simulations, to be split into 20 jobs of 5×10^4 simulations each. Note that the number of jobs is calculated as the total number of simulations (`numsim`) divided by the number of tasks per job (`ntask`). Each simulation has a sample size of `param`.

Jobs are submitted to the grid by running 'simulation.R'. Because this script is just an example of how to call `xgrid()` and manipulate the resulting object, it can be run the way one typically submits commands during an R session. After the completion of each job, the results are saved to a file in the 'output' directory.

3.2 Retrieval and Analysis of Results

Figure 5 is an example of what to expect before and after completing all simulations.

At the beginning of Figure 5, we see the expected directory structure. We then call `xgrid()` to send `numsim=10^6` simulations to the grid for calculation. After the completion of the entire simulation study, results from all `numsim` simulations are collated and returned by the `xgrid()` function as the object `res`. This object is also saved as the file 'RESULTS.rda' at the top of the directory structure.

Figure 5 also displays an example of what would be seen in the 'output' folder. As expected, there are ten files of the form 'RESULT-1000#', which contain the results from each individual job. The second set of ten files (e.g. 'runjob.RRESULT-1000#.Rout') contain the code that was run to generate the corresponding result file.

In this example, `res` is a data frame with `numsim` rows (one row for each simulation) and three columns (as defined in the `return` statement of 'job.R'). In Figure 5, we list the dimensions of `res` and summarize the results using `with()`.

In terms of our motivating example, when the underlying data are normally distributed, we could expect to reject the null hypothesis 2.5% of the time on the left and 2.5% on the right. The simulation yielded rejection rates of 6.5% and 0.7% for the left and right, respectively. This confirms Hesterberg's argument regarding lack of robustness for both the overall α level as well as the individual tails. As for computation time, this simulation took 48.2 seconds (standard deviation of 0.675) when run on a heterogeneous mixture of 20 iMacs and Mac Pros. When run locally on a single Quad-Core Intel Xeon Mac Pro computer, this simulation took 592 seconds (standard deviation of 0.365).

```
> list.files()
[1] "directorystructure2.pdf" "example1-source.Rnw"     "example1-source.tex"
[4] "example1.Rnw"            "input"                   "output"
[7] "simulation.R"
> list.files("input")
[1] "job.R"    "runjob.R"
> library(xgrid)
> res = xgrid(grid="Burton-303-iMac", Rcmd="runjob.R", param=30,
    numsim=10^6, ntask=5*10^4)
> dim(res)
[1] 1000000       3
> with(res, table(leftreject, rightreject))
          rightreject
leftreject  FALSE   TRUE
     FALSE 927343   7430
     TRUE   65227      0
> list.files()
 [1] "RESULTS.rds"             "directorystructure2.pdf" "example1-source.Rnw"
 [4] "example1-source.tex"     "example1.Rnw"            "input"
 [7] "job.err"                 "job.out"                 "output"
[10] "simulation.R"
> list.files("output")
 [1] "RESULT-10000"              "RESULT-10001"              "RESULT-10002"
 [4] "RESULT-10003"              "RESULT-10004"              "RESULT-10005"
 [7] "RESULT-10006"              "RESULT-10007"              "RESULT-10008"
[10] "RESULT-10009"              "RESULT-10010"              "RESULT-10011"
[13] "RESULT-10012"              "RESULT-10013"              "RESULT-10014"
[16] "RESULT-10015"              "RESULT-10016"              "RESULT-10017"
[19] "RESULT-10018"              "RESULT-10019"              "runjob.RRESULT-10000.Rout"
[22] "runjob.RRESULT-10001.Rout" "runjob.RRESULT-10002.Rout" "runjob.RRESULT-10003.Rout"
[25] "runjob.RRESULT-10004.Rout" "runjob.RRESULT-10005.Rout" "runjob.RRESULT-10006.Rout"
[28] "runjob.RRESULT-10007.Rout" "runjob.RRESULT-10008.Rout" "runjob.RRESULT-10009.Rout"
[31] "runjob.RRESULT-10010.Rout" "runjob.RRESULT-10011.Rout" "runjob.RRESULT-10012.Rout"
[34] "runjob.RRESULT-10013.Rout" "runjob.RRESULT-10014.Rout" "runjob.RRESULT-10015.Rout"
[37] "runjob.RRESULT-10016.Rout" "runjob.RRESULT-10017.Rout" "runjob.RRESULT-10018.Rout"
[40] "runjob.RRESULT-10019.Rout"
```

Figure 5: Expected console output, before and after calling `xgrid()`

4 Acknowledgements

This material is based in part upon work supported by the National Institute of Mental Health (5R01MH087786-02) and the US National Science Foundation (DUE-0920350, DMS-0721661, and DMS-0602110).

5 Bibliography

Apple. *Mac OS X Server: Xgrid Administration and High Performance Computing (Version 10.6 Snow Leopard)*. Apple Inc, 2009.

T. Hesterberg. It's time to retire the $n \geq 30$ rule. *Proceedings of the Joint Statistical Meetings*, 2008. http://home.comcast.net/~timhesterberg/articles/.

S. C. Anoke, Y. Zhao, H. Jaeger, and N. J. Horton. xgrid and R: Parallel Distributed Processing Using Heterogeneous Groups of Apple Computers. *R Journal*, 4(1):45–55, 2012. http://journal.r-project.org/archive/2012-1/RJournal_2012-1_Anoke~et~al.pdf

S. S. Sawiloswky and R. C. Blair. A more realistic look at the robustness and Type II error properties of the t test to departures from population normality. *Psychological Bulletin*, 111(2): 352–360, 1992.