Summaries of type-1 error rate simulations

2023-05-12

Contents

1 Introduction

- In this document we use our PC to read in the simulation results output by `simnull.R` and prepare a summary that can be included in the paper.
- The "output" of this document is a LaTeX table of estimated type-1 error rates and their SEs in Section 4.1.3.
- Before getting to the summary we read in the data and check that the pedigrees from our pool of pedigrees are being sampled uniformly.
- **Note:** Most other RMarkdown documents in our workflow were written as documentation for a corresponding R script that was intended to be run on the cluster using a SLURM script. This RMarkdown document, like `simaltSummary.Rmd`, is intended to be run, or "knitted" on your PC, and there are no corresponding R and SLURM scripts.

2 Read in simulation results

- `simnull.R` returns files of p-value results.
- Simulations were run on the cluster as an array job of 200 jobs, each containing 10 studies (of three pedigrees).
- Results should be saved on the Compute Canada cluster in the `/project/def-jgraham/FJdata` directory.
- To use this script, you must first copy the results files from the cluster to your PC.
- Several options for transferrin the results files to your PC are discussed in Appendix A.1.4 of the Lab's Workflow document.
- I use the `rsync` method, for which I (i) open a terminal on my Mac and set its working directory to be the directory that contains this .Rmd file, and (ii) run the following command from the terminal:

```
rsync -avz jgraham@cedar.computecanada.ca:project/FJdata/ FJdata
```

- Batches of p-values to estimate the type-1 error rate are in the files `FJdata/pvalnullresi.csv` for i=1,. . . ,200.

- The format of the output files was described in `simnull.Rmd` and this description is repeated in the Appendix of this document.
- We read the results into R in the following code chunk:

```
njobs <- 200
pvalres <- read.csv("FJdata/pvalnullres1.csv") # start with first batch
for(i in 2:njobs){
  pvalres <- rbind(pvalres,read.csv(paste0("FJdata/pvalnullres",i,".csv")))
}
```

3 Summary of simulations: Sampling of pedigrees

- Check that pedigrees from our pool of 55 are being sampled uniformly over the 2000 simulated studies.
- We take the information about which pedigrees were sampled fromthe study information in `pvalres`.

```
pedpool <- scan("FJdata/pedpool/pedpool.txt")
pedpool
```

```
##  [1]  2  3  4  5  8 11 12 13 15 16 18 19 20 22 23 24 25 27 29 30 31 33 34 37 38
## [26] 39 40 41 43 44 45 46 49 50 51 52 53 54 55 56 57 58 59 61 62 63 66 67 69 70
## [51] 73 74 75 76 77
```

```
# Read in the IDs of the pedigrees sampled across the 2000 studies.
peds <- c(pvalres[,"studyped_1"], pvalres[,"studyped_2"],pvalres[,"studyped_3"])
# Check whether every pedigree in the pool of 55 was sampled
all(pedpool %in% peds) # every ped was sampled at least once
```

```
## [1] TRUE
```

```
# See if pedigrees in the pool of 55 have been sampled roughly uniformly.
# There are 6000 study peds (2000 studies containing 3 peds each), so each
# pedigree should be sampled about 6000/55 = 109 times.
table(peds) # some sampled more than others by chance
```

```
## peds
##   2   3   4   5   8  11  12  13  15  16  18  19  20  22  23  24  25  27  29  30
## 122  93 116 107 107 109 107 113 106  94  99 109 115  98 107 113 111 134 145 112
##  31  33  34  37  38  39  40  41  43  44  45  46  49  50  51  52  53  54  55  56
## 109  91 112 118  95 108 118 108 116  89 106 100 105 102 101 106 108 112 111 115
##  57  58  59  61  62  63  66  67  69  70  73  74  75  76  77
## 116 124 120  99 112  99 125  99 104 105 116 106  92 120 116
```

```
sum(table(peds)) # Should be 6000=2000*3.
```

```
## [1] 6000
```

4 Summary for paper: Type-1 error rates

- For the paper we require estimated type-1 error rates for each test.
- The global LR and global transmission tests depend on the value of the carrier probability. Recall that we considered five values of the carrier probability, leading to five global LR and five global transmission tests. In addition there are the three local tests (local LR, RVS and modified RVS) that don't depend on the carrier probability, for a total of 13 tests.

```
true_carrier_prob <- 0.0032
carrier_probs <- true_carrier_prob*c(1/10,1/2,1,2,10)
tests <- c(paste0("globalLR",1:length(carrier_probs)),
           paste0("globaltrans",1:length(carrier_probs)),
           "localLR","RVS","modRVS")
tests
```

```
##  [1] "globalLR1"    "globalLR2"    "globalLR3"    "globalLR4"    "globalLR5"
##  [6] "globaltrans1" "globaltrans2" "globaltrans3" "globaltrans4" "globaltrans5"
## [11] "localLR"      "RVS"          "modRVS"
```

- In the code chunk below we loop over the tests and find all the p-values for each one.
- For a given test, the simulation output from `simnull.R` includes p-values for each candidate cRV observed in each study. The different cRVs are distinguished by the suffix `_1` for the first, `_2` for the second and `_3` for the third.
  - Here "first", "second" and "third" refer to the order of the observed candidate cRVs along chromosome 8, not to the size of the p-values.
  - When only two candidate cRVs are observed in a study the third p-values are `NA` and when there is only one candidate cRV observed in a study the second and third p-values are `NA`.
  - Note: Of the 2000 simulated studies, 1996 had only one observed candidate cRV, four had two observed candidate cRVs and none of the simulated studies had three observed candidate cRVs. Thus, there were 2004 candidate cRVs tested in total.
- For a given test, the estimated type-1 error rate is the proportion of the 2004 candidate-cRV tests that reject the null hypothesis.

```
# Initialize matrix to hold all p-values. Rows of the matrix are
# studies and columns are tests.
pall <- matrix(NA,nrow=nrow(pvalres)*3,ncol=length(tests))
for(i in 1:length(tests)) {
  pvalcols <- paste0(tests[i],"_",1:3)
  p <- pvalres[,pvalcols] # p is a data frame of p-values for this test
  pall[,i] <- unlist(p) # vector of all p-values
}
colnames(pall) <- tests
type1errres<- function(pmat){
  ests <- apply(pmat,2,FUN=function(x) mean(x<=0.05,na.rm=TRUE))
  ses <- apply(pmat,2,FUN=se)
  res <- cbind(ests,ses); colnames(res) <- c("Estimates","SEs")
  return(res)
}
se <- function(x) {
  n <- length(x)
  p <- mean(x<=0.05,na.rm=TRUE)
  return(sqrt(p*(1-p)/n))
}
```

4.1 Estimated type-1 error rates

```
round(type1errres(pall),3)
```

```
##              Estimates   SEs
## globalLR1        0.046 0.003
## globalLR2        0.045 0.003
## globalLR3        0.045 0.003
## globalLR4        0.045 0.003
## globalLR5        0.041 0.003
## globaltrans1     0.046 0.003
## globaltrans2     0.046 0.003
## globaltrans3      0.046 0.003
## globaltrans4      0.046 0.003
## globaltrans5      0.043 0.003
## localLR           0.033 0.002
## RVS               0.039 0.003
## modRVS            0.049 0.003
```

4.2 Comments on results

- Type-1 error appears to be controlled at the 5% level for all tests.
- The estimated type-1 error rate of the local LR and RVS methods is more than 2 SEs below 0.05, so these methods appear to be conservative.
- The estimated type-1 error rate of the global tests (LR and transmission) is quite similar across most values of the carrier probability $p_c$, but is more than 2 SEs below 0.05 (i.e., is conservative) when the carrier probability is misspecified at 10-times its true value.

4.3 LaTeX table of results

- I organized the estimated type-1 error rates and their SEs into a LaTeX-formatted table similar to the table in Christina's thesis.
    - Note: The table is *not* automatically generated. It was done by hand from an empty template table (shown below) with the cells of the table filled in by cutting-and-pasting from the R output.

```
\begin{table}
\centering
\caption{Template}
\begin{tabular}{lcccccc}
       & & \multicolumn{6}{c}{Method} \\ \cline{3-8}
       & & \multicolumn{3}{c}{Local} & & \multicolumn{2}{c}{Global} \\ \cline{3-5} \cline{7-8}
Assumed $p_c$ & & RVS & ModRVS & LR &   & LR & Transm. \\ \hline
NA & & e & e & e &    & -- & -- \\
$0.000032$& & -- & -- & -- &  & e & e \\
$0.000160$& & -- & -- & -- &  & e & e \\
$0.000320^*$& & -- & -- & -- &  & e & e \\
$0.000640$& & -- & -- & -- &  & e & e \\
$0.003200$& & -- & -- & -- &  & e & e \\ \hline
\multicolumn{8}{l}{$^*$ True carrier probability is $p_c = 0.000320$}
\end{tabular}
\end{table}
```

Table 1: Estimated type-1 error rate (SE), all tested RVs (1002 in total)

	Method				
	Local			Global	
Assumed $p_c$	RVS	ModRVS	LR	LR	Transm.
NA	0.039 (0.003)	0.049 (0.003)	0.033 (0.002)	–	–
0.000032	–	–	–	0.046 (0.003)	0.046 (0.003)
0.000160	–	–	–	0.045 (0.003)	0.046 (0.003)
0.000320*	–	–	–	0.045 (0.003)	0.046 (0.003)
0.000640	–	–	–	0.045 (0.003)	0.046 (0.003)
0.003200	–	–	–	0.041 (0.003)	0.043 (0.003)

* True carrier probability is $p_c = 0.000320$

A Appendix: Format of the simulation-output file

- In the output files of p-values we include the following columns of information about the study replicate: the replicate number, the IDs of the three study pedigrees, and the IDs of the polymorphic cRVs in each study.

- For n_p carrier probabilities, we record $3 \times (2 \times n_p + 3)$ p-values: for each of the (up to) three cRVs tested we have n_p p-values for the global LR, n_p for the global transmission approaches, and one each for the three local approaches (local LR, RVS and modified RVS).
  - For the i^{th} carrier probability and j^{th} cRV, the column names are `globalLR`*i*`_`*j* for the global likelihood ratio test and `globaltrans`*i*`_`*j* for the global transmission test.
  - For the local tests (which do not depend on the carrier probability), the column names for the j^{th} cRVs are `localLR_`*j*, `RVS_`*j* and `modRVS_`*j*, respectively.
  - If there are fewer than three polymporhphic cRVs, the empty slots for the p-values are encoded as `NA`.