

# Summaries of type-1 error rate simulations

2023-05-18

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Read in simulation results</b>	<b>1</b>
<b>3</b>	<b>Summary of simulations: Sampling of pedigrees</b>	<b>2</b>
<b>4</b>	<b>Summary for paper: Type-1 error rates</b>	<b>2</b>
4.1	Estimated type-1 error rates . . . . .	4
4.2	Comments on results . . . . .	4
4.3	LaTeX table of results . . . . .	4
<b>A</b>	<b>Appendix: Format of the simulation-output file</b>	<b>5</b>

## 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 transferring 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 from the study information in `pvalres`.

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

## [1] 2 3 4 5 8 11 13 15 16 18 19 20 22 23 24 25 27 29 30 31 33 34 37 38 39
## [26] 40 41 43 44 45 46 49 50 51 52 53 54 55 56 57 58 59 61 62 63 66 67 69 70 73
## [51] 74 75 76 77 78

# 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 13 15 16 18 19 20 22 23 24 25 27 29 30 31
## 120 102 143 116 101 101 107 112 120 100 115 115 117 99 107 105 104 116 110 82
## 33 34 37 38 39 40 41 43 44 45 46 49 50 51 52 53 54 55 56 57
## 110 102 106 97 103 119 102 104 118 91 117 79 118 122 95 119 79 131 103 115
## 58 59 61 62 63 66 67 69 70 73 74 75 76 77 78
## 105 128 124 124 123 118 101 96 112 110 106 110 104 103 114

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 seven values of the carrier probability, leading to seven global LR and seven 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(1/100,1/10,1/2,1,2,10,100)
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] "globalLR6" "globalLR7" "globaltrans1" "globaltrans2" "globaltrans3"
## [11] "globaltrans4" "globaltrans5" "globaltrans6" "globaltrans7" "localLR"
## [16] "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, 2004 candidate cRVs in total were tested.

# number of candidate cRVs in a study is the number of non-missing cRV IDs
cRV_IDs <- pvalres[,c("cRV_1","cRV_2","cRV_3")]
numcRVs <- apply(cRV_IDs,1,function(x) sum(!is.na(x)))
table(numcRVs)

## numcRVs
##      1      2
## 1996      4

• 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	0.037	0.002
## globalLR2	0.037	0.002
## globalLR3	0.036	0.002
## globalLR4	0.036	0.002
## globalLR5	0.036	0.002
## globalLR6	0.033	0.002
## globalLR7	0.007	0.001
## globaltrans1	0.035	0.002
## globaltrans2	0.035	0.002
## globaltrans3	0.035	0.002
## globaltrans4	0.034	0.002
## globaltrans5	0.034	0.002
## globaltrans6	0.033	0.002
## globaltrans7	0.024	0.002
## localLR	0.028	0.002
## RVS	0.041	0.003
## modRVS	0.052	0.003

## 4.2 Comments on results

- Type-1 error appears to be controlled at the 5% level for all tests.
- However, all tests but the modified RVS appear to be conservative.
  - In particular, the LR tests (local LR, global LR and global transmission) and the RVS test have estimated type-1 error rates at 3 or more SEs below 0.05.
- The estimated type-1 error rates of the global tests (LR and transmission) are similar across assumed values of the carrier probability within 10 times or less of the true carrier probability.

## 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}
\caption{Template table}
\begin{tabular}{lcccccc}
& & \multicolumn{6}{c}{Method} \\ \cline{3-8}
& & \multicolumn{3}{c}{Local} & & \multicolumn{2}{c}{Global} \\ \cline{3-5} \cline{7-8}
Factor$~$ & $ & $ & RVS & $ & ModRVS & $ & LR & $ & LR & Transm. \\ \hline
NA & & e & e & e & e & & -- & & -- & \\
1/100 & & -- & -- & -- & -- & & e & & e & \\
1/10 & & -- & -- & -- & -- & & e & & e & \\
1/2 & & -- & -- & -- & -- & & e & & e & \\
1 & & -- & -- & -- & -- & & e & & e & \\
2 & & -- & -- & -- & -- & & e & & e & \\
10 & & -- & -- & -- & -- & & e & & e & \\
100 & & -- & -- & -- & -- & & e & & e & \\ \hline
\multicolumn{8}{l}{Assumed carrier probability is factor times true value $p_c = 0.000320$.}
\end{tabular}
\end{table}
```

Table 1: Estimated type-1 error rate (SE), all tested RVs (2004 in total)

Factor*	Method				
	Local			Global	
	RVS	ModRVS	LR	LR	Transm.
NA	0.041 (0.003)	0.052 (0.003)	0.028 (0.002)	–	–
1/100	–	–	–	0.037 (0.002)	0.035 (0.002)
1/10	–	–	–	0.037 (0.002)	0.035 (0.002)
1/2	–	–	–	0.036 (0.002)	0.035 (0.002)
1	–	–	–	0.036 (0.002)	0.034 (0.002)
2	–	–	–	0.036 (0.002)	0.034 (0.002)
10	–	–	–	0.033 (0.002)	0.033 (0.002)
100	–	–	–	0.007 (0.001)	0.024 (0.002)

\* Assumed carrier probability is factor times true value  $p_c = 0.000320$ .

## 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 polymorphic cRVs, the empty slots for the p-values are encoded as **NA**.