Permutation Example

Suzanne Dufault

November 6, 2018

Abstract

This paper explores how closely the formula derived by Nick approximates the variation in the permutation distribution for a simple example of $2m = 10$ clusters.

1 Process

1. I set up an example dataset for 10 clusters with variation between the clusters with respect to Cases and OFIs.
2. I then permuted 10 choose 5 (252) unique treatment allocations.
3. I computed the log(OR) for RR = 1 for each of the permuted allocations using 100 cases and 100 controls assigned to clusters according to their current proportions.
4. I computed the estimate of the standard deviation of the log(OR) for each of the permuted allocations using our proposed formula.
5. **Permutation CI** was found by marking the 2.5 and 97.5 percentiles of the log(OR) estimates.
6. **Paper CI** was found by taking 1.959964 times the average estimated standard deviation of the log(OR) according to Nick's formula.

2 Example Dataset

The first 10 columns of the example dataset:

	Cluster	Cases	OFI	Period	tx	tx.1	tx.2	tx.3	tx.4	tx.5
1	1	52	138	1	1	0	0	0	1	0
2	2	74	212	1	1	1	0	1	0	0
3	3	54	125	1	0	0	0	1	0	1
4	4	72	145	1	1	1	1	0	1	0
5	5	46	165	1	0	1	1	0	1	1
6	6	42	194	1	0	1	0	0	1	0
7	7	70	250	1	0	0	1	0	1	1
8	8	50	131	1	1	1	0	1	0	1
9	9	73	229	1	1	0	1	1	0	1
10	10	69	156	1	0	0	1	1	0	0

Table 1: The first few columns of treatment assignments and corresponding Case and OFI numbers.

Example contingency table using the first treatment allocation:

	Cases	OFI
Treated	53	47
Untreated	49	50

Table 2: Example contingency table using the first treatment allocation. The discrepancy for number of untreated individuals (99 instead of 100) is due to rounding. This can be changed, but reflects the way the function for analysis is currently set to work.

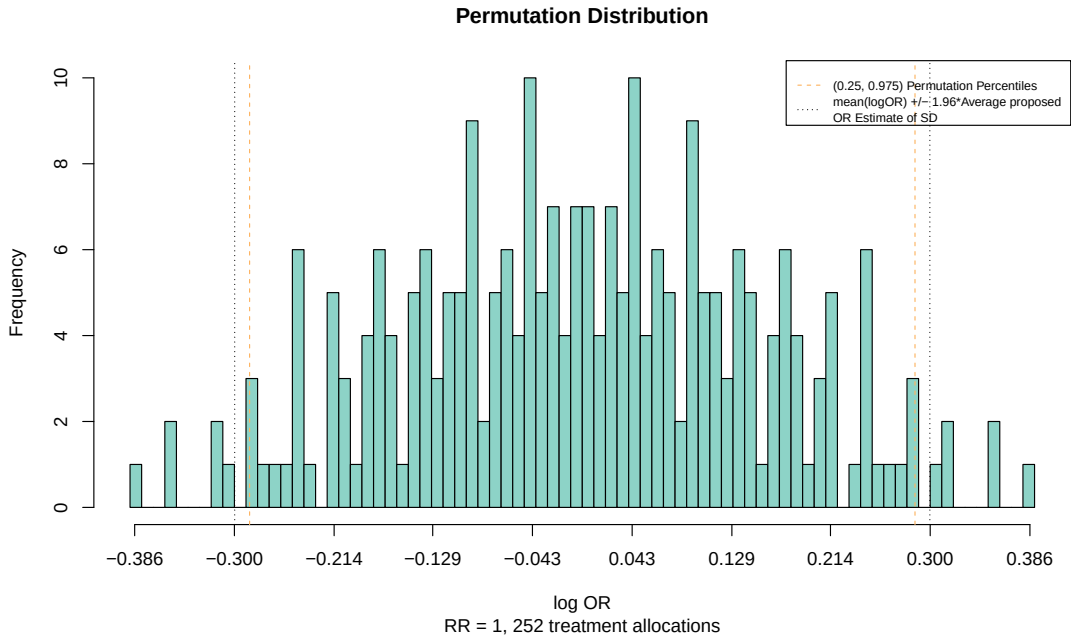
3 Test

```
sdest <- mean(unlist(test$sd_est))
sdest

## [1] 0.1529221
```

The average "paper estimated" standard deviation is 0.1529221.

4 Comparison with Permutation Distribution



	2.5%	97.5%
Permutation 95% CI	-0.2869102696	0.2869102696
Paper 95% CI	-0.2997217290	0.2997217290

Table 3: Comparison of the Wald-style interval from our proposed standard deviation estimates and the .25 and .975 percentiles of the OR permutation distribution.

There is a 2.0446532% discrepancy between our estimate and the truth.

5 Standard Deviation Comparisons

When:

- RR = 1

- nControls = 100 (i.e. ratio of cases to controls = 1)
- nCases = 100

	Standard Deviation Estimate
Permutation	0.15664
Paper	0.15292
GEE	49707014340634.81250
ME	0.28427

Table 4: Comparison of our proposed method (Paper), GEE, and ME standard deviation estimation with the true standard deviation of the permutation distribution (Permutation).

The average SD for GEE is massive, but this is really only due to 16 out of 252 estimates that were massive. If we remove those 16 estimates, the average is 0.1275

When:

- RR = 1
- nControls = 1000 (i.e. ratio of cases to controls is 1)
- nCases = 1000

	Standard Deviation Estimate
Permutation	0.15664
Paper	0.15292
GEE	0.14068
ME	0.14164

Table 5: Comparison of our proposed method (Paper), GEE, and ME standard deviation estimation with the true standard deviation of the permutation distribution (Permutation). This is at the null for the setting where nControls = nCases = 1000.

When $2m = 10$ GEE and ME appear to do a poor job estimating the permutation distribution. However, as the number of cases and controls increase, mixed effect modeling (with random effects at the cluster level) improves with respect to approximating the permutation distribution.

6 Further Investigation of GEE

Subsetting to the treatment allocations causing the problems:

clust	Cases	OFI	tx.28	tx.29	tx.57	tx.60	tx.61	tx.67	tx.68	tx.87	tx.92	tx.106	tx.114	tx.121	tx.127	tx.138	tx.212	tx.236
1	52	138	0	0	1	1	1	1	0	1	0	0	0	1	0	1	0	1
2	74	212	1	0	1	1	0	0	0	0	1	1	1	1	1	0	0	0
3	54	125	1	0	1	1	1	0	1	0	0	1	0	0	1	1	0	0
4	72	145	1	1	0	1	1	1	1	0	0	0	0	0	1	1	0	0
5	46	165	0	1	0	0	0	1	0	1	1	0	1	1	1	0	1	0
6	42	194	0	1	0	0	0	0	1	1	1	1	1	0	0	0	1	1
7	70	250	1	1	0	0	1	0	0	0	1	1	0	1	0	0	1	1
8	50	131	0	0	1	0	0	1	1	1	0	0	1	0	1	1	1	0
9	73	229	0	1	0	0	0	0	0	1	0	1	1	1	0	1	1	1
10	69	156	1	0	1	1	1	1	1	0	1	0	0	0	0	0	0	1

When we examine one particular instance of the above assignments, we see that for each cluster the number of cases and OFIs are very similar. This is logical, as RR = 1

```
g1 <- geeglm(case.status ~ tx, data = errslong, family = binomial(link = "logit"), id = clust, corstr = "exchangeable", scale.fix = TRUE)
summary(g1)

##
## Call:
## geeglm(formula = case.status ~ tx, family = binomial(link = "logit"),
##     data = errslong, id = clust, corstr = "exchangeable", scale.fix = TRUE)
##
##  Coefficients:
##              Estimate   Std.err  Wald Pr(>|W|)
## (Intercept) 1.064e-01 8.493e-02 1.571    0.210
## tx          8.737e+14 1.050e+15 0.693    0.405
```

```
##
## Scale is fixed.
##
## Correlation: Structure = exchangeable  Link = identity
##
## Estimated Correlation Parameters:
##       Estimate   Std.err
## alpha 6.665e+14 6.995e+29
## Number of clusters:   10   Maximum cluster size: 26

# For comparison, a standard logistic model:
stand1 <- glm(case.status ~ tx, data = errslong, family = binomial)
summary(stand1)

##
## Call:
## glm(formula = case.status ~ tx, family = binomial, data = errslong)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.22   -1.15    1.13    1.21    1.21
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.108      0.208    0.52     0.60
## tx            -0.183      0.284   -0.64     0.52
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 275.87  on 198  degrees of freedom
## Residual deviance: 275.45  on 197  degrees of freedom
## AIC: 279.5
##
## Number of Fisher Scoring iterations: 3
```

When I use a different package **gee** (rather than **geepack**), I get a warning that convergence is not achieved and reasonable estimates for the coeffiicent and the standard error.

```
library(gee)
g2 <- gee(case.status ~ tx, data = errslong, id = clust, family = binomial(link = "logit"), corstr = "exchange
```

```
## (Intercept)          tx
##      0.1076     -0.1831
```

```
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate
(Intercept)          tx
    -0.1076      0.2210
Warning messages:
1: In gee(case.status ~ tx, data = errslong, id = clust, family = binomial(link = "logit"),  :
  Maximum number of iterations consumed
2: In gee(case.status ~ tx, data = errslong, id = clust, family = binomial(link = "logit"),  :
  Convergence not achieved; results suspect
3: In gee(case.status ~ tx, data = errslong, id = clust, family = binomial(link = "logit"),  :
  Cgee had an error (code= 104).  Results suspect.
```

```
## gee(formula = case.status ~ tx, id = clust, data = errslong,
##     family = binomial(link = "logit"), corstr = "exchangeable",
##     scale.fix = TRUE)
## $link
## [1] "Logit"
##
## $varfun
## [1] "Binomial"
##
## $corstr
## [1] "Exchangeable"
##             Estimate Naive S.E. Naive z Robust S.E. Robust z
## (Intercept)  0.06465    0.08002  0.8079     0.06662  0.97052
## tx           0.11676    0.13040  0.8954     1.39570  0.08365
## [1] "Number of iterations: " "25"
## [1] "Estimated Scale Parameter: " "1"
```