Power calculation, Dichotomous traits

Description

Compute an average power of SKAT and SKAT-O for testing association between a genomic region and dichotomous phenotypes from case-control studies with a given disease model.

Usage

```
Power_Logistic(Haplotypes = NULL, SNP.Location = NULL, SubRegion.Length=-1
, Prevalence=0.01, Case.Prop=0.5, Causal.Percent=5, Causal.MAF.Cutoff=0.03
, alpha =c(0.01,10^(-3),10^(-6)), N.Sample.ALL = 500 * (1:10)
, Weight.Param=c(1,25), N.Sim=100, OR.Type = "Log"
, MaxOR=5, Negative.Percent=0)

Power_Logistic_R(Haplotypes = NULL, SNP.Location = NULL, SubRegion.Length=-1
, Prevalence=0.01, Case.Prop=0.5, Causal.Percent=5, Causal.MAF.Cutoff=0.03
, alpha =c(0.01,10^(-3),10^(-6)), N.Sample.ALL = 500 * (1:10)
, Weight.Param=c(1,25), N.Sim=100, OR.Type = "Log"
, MaxOR=5, Negative.Percent=0, r.corr=0)
```

Arguments

`Haplotypes`	a haplotype matrix with each row as a different individual and each column as a separate SNP (default= NULL). Each element of the matrix should be either 0 (major allel) or 1 (minor allele). If NULL, SKAT.haplotype dataset will be used to compute power.
`SNP.Location`	a numeric vector of SNP locations which should be matched with the SNPs in the Haplotype matrix (default= NULL). It is used to obtain subregions. When Haplotype=NULL, it should be NULL.
`SubRegion.Length`	a value of the length of subregions (default= -1). Each subregion will be randomly selected, and then the average power will be calculated by taking the average over the estimated powers of all subregions. If SubRegion.Length=-1 (default), the length of the subregion is the same as the length of the whole region, so there will no random selection of subregions.
`Prevalence`	a value of disease prevalence.
`Case.Prop`	a value of the proportion of cases. For example, Case.Prop=0.5 means 50 % of samples are cases and 50 % of samples are controls.
`Causal.Percent`	a value of the percentage of causal SNPs among rare SNPs (MAF < Causal.MAF.Cutoff)(default= 5).
`Causal.MAF.Cutoff`	a value of MAF cutoff for the causal SNPs. Only SNPs that have MAFs smaller than this are considered as causal SNPs (default= 0.03).
`alpha`	a vector of the significance levels (default= c(0.01,10^(-3),10^(-6))).
`N.Sample.ALL`	a vector of the sample sizes (default= 500 * (1:10)).
`Weight.Param`	a vector of parameters of beta weights (default= c(1,25)).
`N.Sim`	a value of number of causal SNP/SubRegion sets to be generated to compute the average power (default= 100). Power will be computed for each causal SNP/SubRegion set, and then the average power will be obtained by taking mean of the computed powers.
`OR.Type`	a function type of effect sizes (default= “Log”). “Log” indicates that log odds ratio of causal variants equal to *c*|*log10(MAF)*|, and “Fixed” indicates that log odds ratio of all causal variants are the same.
`MaxOR`	a numeric value of the maximum odds ratio (default= 5). When OR.Type=“Log”, the maximum odds ratio is MaxOR (when MAF=0.0001). When OR.Type=“Fixed”, all causal variants have the same odds ratio (= MaxOR). See details
`Negative.Percent`	a numeric value of the percentage of coefficients of causal variants that are negative (default= 0).
`r.corr`	(Power_Logistic_R only) the ρ parameter of new class of kernels with compound symmetric correlation structure for genotype effects (default= 0). See details.

Details

By default it uses the haplotype information in the SKAT.haplotypes dataset. So you can left Haplotypes and SNP.Location as NULL if you want to use the SKAT.haplotypes dataset.

When OR.Type="Log", MaxOR is a odds ratio of the causal SNP at MAF = *10^{-4}* and used to obtain c value in the function \|*log OR = c*\|*log10(MAF)*\|. For example, if MaxOR=5, *c = log(5)/4 = 0.402*. Then a variant with MAF=0.001 has log odds ratio = 1.206 and a variant with MAF=0.01 has log odds ratio = 0.804.

When SubRegion.Length is small such as 3kb or 5kb, it is possible that you can have different estimated power for each run with N.Sim = *50 \sim 100*. Then, please increase N.Sim to *500 \sim 1000* to obtain stable results.

Power_Logistic_R computes the power with new class of kernels with the compound symmetric correlation structure. It uses a slightly different approach, and thus Power_Logistic and Power_Logistic_R can produce slightly different results although r.corr=0.

If you want to computer power of SKAT-O by estimating the optimal r.corr, use r.corr=2. The estimated optimal r.corr is *r.corr = p_1^2 (2p_2-1)^2*, where *p_1* is a proportion of causal variants, and *p_2* is a proportion of negatively associated causal variants among the causal variants.

Value

`Power`	A matrix with each row as a different sample size and each column as a different significance level. Each element of the matrix is the estimated power.
`r.corr`	r.corr value. When r.corr=2 is used, it provides the estimated r.corr value. See details.

Author(s)

Seunggeun Lee

Examples

```
#
#       Calculate the average power of randomly selected 3kb regions
#       with the following conditions.
#
#       Causal percent = 20%
#       Negative percent = 20%
#       Max OR = 7 at MAF = 10^-4
#
#       When you use this function, please increase N.Sim (more than 100)
#
out.b<-Power_Logistic(SubRegion.Length=3000,
Causal.Percent= 20, N.Sim=5 ,MaxOR=7,Negative.Percent=20)

out.b

#
#       Calculate the required sample sizes to achieve 80% power

Get_RequiredSampleSize(out.b, Power=0.8)
```