

Resampling-Based Approach for Testing Hardy-Weinberg Proportions in Case- Control Studies

Sanjay Shete, Ph.D.

Department of Epidemiology
M. D. Anderson Cancer Center
December 7, 2009
sshete@mdanderson.org

Hardy-Weinberg

- Consider a simple situation with two alleles, A and a , at a single locus. If the allele frequency of A is p and the allele frequency of a is $(1-p)$, then the expected genotype frequencies of AA , Aa , and aa are p^2 , $2p(1-p)$, and $(1-p)^2$, respectively, assuming HWP in the **population**.
- In a case-control study, the deviation from HWP in **controls**, which is assessed by comparing the difference between observed genotype frequencies and the corresponding expected frequencies, is used to identify potential genotyping errors.

Motivation

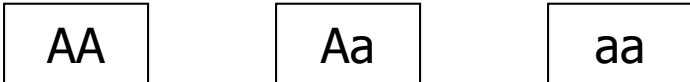
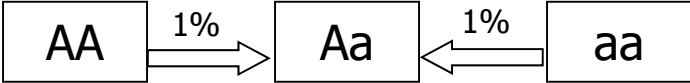
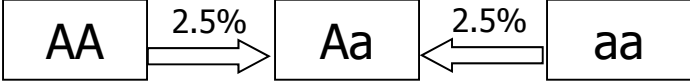
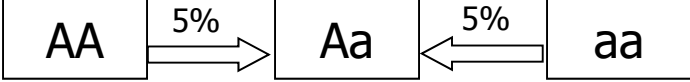
- Assessment of the Hardy-Weinberg proportion (HWP) in controls has been widely used as a quality control measure for identifying genotyping errors in case-control association study.
- However, when the disease of interest is common, controls might not represent the general population. Therefore, using the HWP test in controls would lead to discarding potentially causal SNPs.

Motivation

- Using only controls for HWP assessment is reasonable when assuming a rare disease in the study.
- However, when the disease of interest is common, controls might not represent the general population, as cases account for a relatively large portion of the general population.

Motivation

- The power of HWP test to detect genotyping errors is very poor
 - Consider a diallelic locus, with A as the risk allele and minor allele frequency (MAF) = 0.3. The counts for genotyping are AA = 497, Aa = 418, and aa = 85 when there is no error.

		P value
No error		0.8790
Error rate = 2%		0.6488
Error rate = 5%		0.3655
Error rate = 10%		0.0862

Motivation

- In reality, the genotyping error rates are very small, due to the rapid development of genotyping techniques.

Service	Error Rate (%)	Missing Rate (%)	# of Projects
Human GWAS SNP - Illumina	0.01	0.24	18
Human Linkage SNP	0.008	0.21	43
Human Custom SNP	0.009	0.19	47
Human MHC SNP	0.0	0.17	1
Mouse Linkage SNP	0.003	0.14	11
Mouse Custom SNP	0.0	0.08	5
<hr/>			
*Human Linkage STRP	0.07	3.96	109
*Mouse Linkage STRP	0.06	4.7	24
**Human GWAS SNP - Affymetrix	0.19	0.28	2

- Quality control statistics from Center for Inherited Disease Research website

Motivation

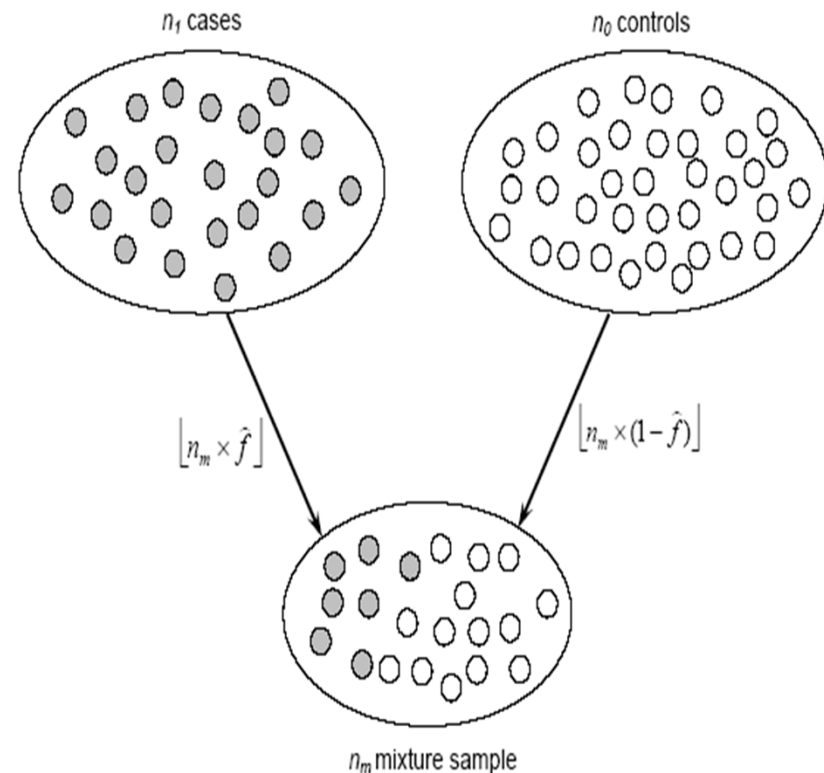
- Using the HWP test in controls might lead to discarding important single-nucleotide polymorphisms (SNPs) that could potentially be **causal SNPs of the disease**.
- Consider a complete penetrance recessive model. All cases are AA and all controls are AG or GG. Let us assume $P(A) = 0.3$ and $P(G) = 0.7$ then $p\text{-value} < 0.000001$

Approach

- We proposed an improved HWP test, called the mixture HWP (mHWP) exact test.
- The mHWP test estimates HWP in a mixture sample that is a combination of cases and controls, which mimics the general population.
- The number of cases in the mixture sample was proportional to the prevalence of the disease.
- We implemented a re-sampling procedure to obtain empirical p values.
- We compared the mHWP approach to the traditional HWP exact test and a likelihood-based test proposed by Li and Li.

mHWP Exact Test

- Construct a mixture sample to represent the general population.
 - Consider a case-control study with n individuals, $n = n_0 + n_1$, where n_0 is the number of controls and n_1 is the number of cases.
 - Let f be the estimated prevalence of disease in the general population.
 - n_m is the mixture sample size.
 - $n_m = \min(n_1/f, n_0/(1-f))$.
 - Randomly sampled $n_m * f$ individuals from cases, and $n_m * (1-f)$ individuals from controls.



mHWP Exact Test

- Re-sampling procedure
 - Repeated the procedure to obtain L mixture samples.
 - Applied the exact HWP test to the mixture samples and obtained L HWP exact p values.
 - Constructed the empirical distribution-based non-parametric density based on L p values (kernel density estimation).
 - Obtained the MLE of this empirical distribution as the final estimate of p value for the mHWP.
- Simulations were conducted to decide the number of mixture samples L .
 - In our study, we selected $L = 500$.

Kernel Density Estimation

- Given a random sample of p values p_1, p_2, \dots, p_N from some density g , the kernel density estimate of g is defined by

$$\hat{g}_h(x) = \frac{1}{N} \sum_{i=1}^N \frac{1}{h} K\left(\frac{x - p_i}{h}\right)$$

- where $K(\cdot)$ is a kernel function and $h > 0$ is a smoothing parameter called bandwidth.
- $K(\cdot)$ is usually taken to be a symmetric unimodal density centered at zero, such as the standard normal density.
- The optimal bandwidth parameter $h = (4/(3N))^{1/5} \hat{\sigma}$,
 - where $\hat{\sigma}$ is the median absolute deviation estimator defined as $\hat{\sigma} = \text{median}\{|p_i - \hat{u}|\} / 0.6745$, where \hat{u} denotes the median of the sample.

Relative Rejection Probability (RRP)

- Relative rejection probability (RRP) measures the relative probability of one approach rejecting SNPs (in HWP), compared to the other approach at a given significance level.
- Consider two approaches of HWP test M_1 and M_2 , then RRP of M_1 compared to M_2 at a given significance level is

$$\text{RRP} = \frac{P(\text{reject HWP hypothesis using } M_1 \text{ at } \alpha \mid \text{HWP}) - P(\text{reject HWP hypothesis using } M_2 \text{ at } \alpha \mid \text{HWP})}{P(\text{reject HWP hypothesis using } M_2 \text{ at } \alpha \mid \text{HWP})}$$

- If RRP is positive, using M_1 is more likely to result in rejection of SNPs than using M_2 , when the SNPs are in HWP.

Simulation Studies

- Specific parameters for simulation studies.

Factors	Coefficients of logistic model	Prevalence
Intercept	-3.4/-2.5/-1.9	
SNP ₁	0.4055 (OR=1.5)	10%/30%/50%
SNP ₂	0 (OR=1)	10%/30%/50%
Sex	0.6931 (OR = 2)	50% (Male)
Ethnicity	0.4055 (OR=1.5)	75% (Caucasian)
Physical activity	-0.4055 (OR = 0.67)	50% (Yes)
Age		
0-30	0.4055 (OR for additive model=1.5)	36%
31-50		39%

Simulation Studies

- Considered two independent SNPs: causal SNP₁ and non-causal SNP₂.
- Studied minor allele frequencies of 10%, 30% and 50% for both SNPs.
- Considered different levels of prevalence, ranging from 19% to 36%, which can represent different common diseases.
- Studied three genetic models: dominant, additive and recessive.
- Simulated 10000 replicates, each with 1000 cases and 1000 controls.

Error Models

- GLHO genotyping error model

Observed genotype	True genotype		
	11	12	22
11	$(1 - \varepsilon_1)^2$	$\varepsilon_2 (1 - \varepsilon_1)$	ε_2^2
12	$2\varepsilon_1 (1 - \varepsilon_1)$	$\varepsilon_1 \varepsilon_2 + (1 - \varepsilon_1)(1 - \varepsilon_2)$	$2\varepsilon_2 (1 - \varepsilon_2)$
22	ε_1^2	$\varepsilon_1 (1 - \varepsilon_2)$	$(1 - \varepsilon_2)^2$

- Assumed $\varepsilon_1 = \varepsilon_2 = \varepsilon$.
 - The expected genotyping error rate is $2\varepsilon - \varepsilon^2(1 + 2p - 2p^2)$.
 - Genotyping error rate was assumed to be 1% or 5%.
- 'empirical' error model
 - Based on a real GWA data in which errors were estimated based on re-sequencing.
 - Genotyping error rate is about 12%.

Simulation Results

- Estimated type I error probability and RRP of causal SNP₁ at 0.05 significance level in simulation studies based on 10,000 replicates.
- The type I error rates for the traditional approach (using controls only) were inflated dramatically as MAFs and prevalence of disease increased, when the dominant or recessive model was assumed.
- The likelihood-based approach and the mHWP exact test could control type I errors.
- However, the mHWP exact test is more likely to keep the causal SNPs for further analysis.

Model	MAF	Prev	Type I Errors			Relative Rejection Probability Likelihood vs mHWP
			Controls only	Likelihood-based	mHWP	
Dominant	0.1	19.56%	0.051	0.057	0.039	0.470
		29.64%	0.058	0.052	0.029	0.808
	0.3	21.58%	0.077	0.053	0.044	0.203
		32.20%	0.111	0.050	0.033	0.545
	0.5	23.09%	0.095	0.053	0.047	0.146
		34.12%	0.152	0.054	0.035	0.551
Additive	0.1	19.64%	0.036	0.053	0.033	0.613
		29.73%	0.041	0.052	0.026	1.031
	0.3	22.25%	0.047	0.049	0.041	0.197
		32.99%	0.049	0.050	0.031	0.604
	0.5	24.96%	0.056	0.054	0.045	0.205
		36.31%	0.056	0.049	0.032	0.553
Recessive	0.1	18.43%	0.041	0.052	0.037	0.403
		28.20%	0.041	0.051	0.030	0.678
	0.3	18.93%	0.080	0.052	0.046	0.144
		28.84%	0.115	0.051	0.037	0.393
	0.5	19.94%	0.095	0.048	0.045	0.060
		31.12%	0.150	0.049	0.038	0.284

Simulation Results

- Estimated type I error probability and RRP of non-causal SNP₂ at 0.05 significance level in simulation studies based on 10,000 replicates.
- All three approaches control type I error well.

MAF	Prev	Type I Errors			Relative Rejection Probability
		Controls only	Likelihood-based	mHWP	Likelihood vs mHWP
0.1	19.21%	0.041	0.052	0.040	0.296
	29.19%	0.041	0.053	0.033	0.619
0.3	20.92%	0.045	0.050	0.045	0.114
	31.34%	0.046	0.051	0.035	0.449
0.5	22.66%	0.048	0.051	0.045	0.127
	33.85%	0.049	0.052	0.035	0.501

Power to Detect Genotyping Errors

- GLHO genotyping error model
 - Low genotyping error rates: 1% and 5%.
 - All three approaches had low power: 5%~10% power at 0.05 significance level.

When genotyping error rates are small, the observed genotype counts will not be significantly different from the expected genotype counts under HWP.

- 'empirical' error model
 - High genotyping error rate: ~12%
 - All three approaches had almost 100% power.

When genotyping error rates are higher, the genotyping error can generate extreme deviation from HWP.

Sensitivity Analysis to Estimated Prevalence

- The true prevalence of a disease in a population is not known with certainty.
- Assessed the sensitivity of the mHWP exact test to the estimated prevalence of disease f .
- Evaluated the mHWP exact test p values using a range of prevalence centered on real prevalence $[f-2\%, f+2\%]$.
- All the results were very similar to those obtained with the use of the real prevalence.
- The mHWP approach is not sensitive to the estimated prevalence.

Real Disease Application

- Applied the mHWP exact test to the real case-control association study of adult obesity.
- The data is from an association study of a common variant in the FTO gene that is associated with obesity.
- The SNP rs9939609 predisposes individuals to diabetes through an effect on BMI.
- Standard cut-off points can be used to define the cases and controls when using BMI as the outcome variable.
 - Cases (obesity): individuals with a BMI ≥ 30 kg/m²
 - Controls (normal weight): individuals with a BMI < 25 kg/m²
- Considered the association study in the UK type 2 Diabetes Genetics Consortium Collection Cases

Real Disease Application

- The mHWP exact test has more likelihood to keep this SNP rs9939609 for further analysis, compared to the other two approaches.
- Importantly, the investigators kept this SNP in the analyses of BMI because originally this study was for the investigation of type 2 diabetes, and in the type 2 diabetes GCC controls the HWP test gave a p-value of 0.83 for this SNP.

Cohort	Genotype Counts						Prev	p Values		
	Normal Weight			Obese				Controls only	Likelihood-based	mHWP
	TT	AT	AA	TT	AT	AA				
UK T2D GCC Cases	113	174	37	524	818	321	0.52	0.018	0.038	0.054

Discussion

- When the prevalence of the disease is large (ranging from 19% to 36% in our simulation studies), the type I error probability of the traditional approach was inflated for the disease-associated SNPs when either the dominant or recessive model was assumed.
- This range of prevalence is realistic for common disease, such as smoking, obesity and hypertension.
- The mHWP exact test can effectively control the type I error probability in all scenarios examined, including models with causal or non-causal SNPs, different genetic models, and different MAFs and prevalence.
- On average, the mHWP exact test proposed in this paper is more likely than the likelihood-based approach to keep causal SNPs in the analysis when the disease is common.
- Prevalence misspecification would not inflate the type I error rate of our approach.

Discussion

- The relationship between genotyping error and the HWP test has been studied in the literature. These studies suggest that the traditional HWP test in controls has very low power for detecting genotyping errors, especially when the genotyping error rate is low and the MAF is not rare.
- Like the traditional HWP test and the likelihood-based test, our test is not very sensitive for detecting genotyping errors when error rates are low.
- Furthermore, recent study showed that genotyping errors will not increase the false-positive rate for detecting associated variants.
- Therefore, one may also consider a strategy of keeping all SNPs for the association study, performing the HWP test using our proposed approach only among significant SNPs.

Discussion

- In genome-wide association studies, using the improved mHWP exact test in the discovery stage will increase the chance that causal SNPs will be carried over for replication. To achieve a more stringent significance level (e.g. 10^{-5} used in GWAS), more mixture samples will be needed to obtain robust MLE of the empirical mHWP p value.

Conclusion: the improved HWP exact test (mHWP exact test) using a mixture sample is a better HWP test for case-control genetic association studies than the traditional HWP only in controls or the likelihood-based approach. This approach will improve our ability to keep causal SNPs in the case-control genetic association studies.