Hypothesis testing in genetic (association) studies

25.10.2005
GE02: day 4 part 2

Yurii Aulchenko
Erasmus MC Rotterdam

Testing HWE, using χ^2

$$\sum_{i=1}^{\#genotypes} \frac{(O_i - E_i)^2}{E_i} \infty \chi^2_{\#genotypes-\#alleles}$$

- Example

genotype	No.	Expected	Expected	(o-e)2/e	
DD	2	0,01	1,1	0,73	
ND	17	0,19	18,8	0,17	
NN	81	0,8	80,1	0,01	P-value
	100		chi2(1)=	0,91	0,3396

A genetic case-control study

General (genotypic) 2 x 3 table

	AA	AB	BB	Σ
Cases	R_{AA}	R_{AB}	R_{BB}	R
Controls	S_{AA}	S_{AB}	S_{BB}	S
Σ	n_{AA}	n_{AB}	n_{BB}	N

Decompositions to 2x2 table

- Allele B is dominant

	AA	B-
Cases	R_{AA}	$(R_{AB}+R_{BB})$
Controls	S_{AA}	$(S_{AB}+S_{BB})$

Allele B is recessive

	A-	BB
Cases	$(R_{AA} + R_{AB})$	R_{BB}
Controls	$(S_{AA} + S_{AB})$	S_{BB}

- Allelic table (number of alleles in cases and controls)

	A	B
Cases	$2 \cdot R_{AA} + R_{AB}$	$R_{AB} + 2 \cdot R_{BB}$
Controls	$2 \cdot S_{AA} + S_{AB}$	$S_{AB} + 2 \cdot S_{BB}$

Z test for 2x2 table

	N	D	Σ
Controls	U_N	U_D	N_1
Cases	A_N	A_D	N_2
Σ	N_N	N_D	N

- Frequency in controls: $p_1 = U_D / n_1$
- Frequency in cases: $p_2 = A_D / n_2$
- Overall frequency: $p = N_D / N$

$$Z = \frac{p_1 - p_2}{\sqrt{p \cdot (1-p) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \infty \Phi_{\mu=0,\sigma=1}$$

Genotypic (2x3) table

General (genotypic) 2 x 3 table

	AA	AB	BB	Σ
Cases	R_0	R_1	R_2	R
Controls	S_0	S_1	S_2	S
Σ	n_0	n_1	n_2	N

Tests for 2x3 table

- Armitage's trend (score) test follows χ^2 with 1 df

$$X^2_{A,i} = \frac{N\left[N\left(r_1 + 2r_2\right) - R\left(n_1 + 2n_2\right)\right]^2}{R\left(N - R\right)\left[N\left(n_1 + 4n_2\right) - \left(n_1 + 2n_2\right)^2\right]}$$

- Chi-squared test (on 2 d.f. = (2-1) x (3-1))

$$\sum_{i=1}^{6} \frac{\left(O_i - E_i\right)^2}{E_i} \infty \chi^2_2$$

Multiple testing

- Null hypothesis is true
  - $\alpha = 0.05$ (1 test in 20 is "significant")
  - Z test has a threshold of 1.96
- We do two independent tests
  - What is the chance we will obtain $|Z| \geq 1.96$ in any or both?
  - I. e. what is type 1 error (α) at the threshold of 1.96?
    $\alpha = 1 - (1 - 0.05)^2 = 0.0975$
- To keep $\alpha = 0.05$
  Solve $1 - (1 - x)^2 = 0.05$
  Nominal significance x should be 0.02532
  Threshold is 2.24

Šidak and Bonferroni corrections

If N tests are done, to keep type 1 error of α nominal threshold significance should be x

$\alpha = 1 - (1 - x)^N$

Šidak correction is given by solution of this equation

Bonferroni correction

When $x \rightarrow 0$

$(1 - x)^N \approx 1 - N \cdot x$

then $x = \alpha/N$

Already for N=5 and α=0.05

Šidak x = 0.0102

Bonferroni x = 0.05/5 = 0.01

Estimating GW significance

- Bonferroni correction
  - GW α = 0.05 corresponds to nominal P = 0.05/(# SNPs)
- FDR procedures
  - Less conservative compared to Bonferroni
  - Benjamini & Hochberg 1995: R library "GenABEL"
  - Storey 2006: R library "qvalue"
- SNPs are not independent (LD), therefore Bonferroni and FDR are still conservative
- CURRENT STANDARD: Empirical (permutation) tests
- POSSIBLE STANDARD: $P \leq 5\ 10^{-7}$

GW significance threshold(s)

- Fixed threshold of $P \leq 5\ 10^{-7}$
  - χ^2 test with 1 d.f. 25.26 ($Z = \sqrt{(25.26)} = 5.03$)
  - χ^2 test with 2 d.f. 29.02 ($Z = \sqrt{(29.02)} = 5.39$)

- Conservative Bonferroni threshold
  - $\alpha = 0.05$, 500K SNP array
  - nominal $P = 0.05/500{,}000 = 0.0000001 = 10^{-7}$
  - χ^2 test with 1 d.f. 28.37 ($Z = \sqrt{(28.37)} = 5.33$)
  - χ^2 test with 2 d.f. 32.34 ($Z = \sqrt{(32.34)} = 5.69$)

Useful excel functions

- Cumulative binomial $P_{n,p}(x \leq k)$

  =binomdist(k,n,p,1)

- Cumulative standard normal $\Phi(x \leq k)$

  =normdist(k,0,1,1)

- Poisson $P_{\lambda}(x \leq k)$

  =posisson(k,λ,1)

- Chi-squared with m d.f., $\chi_m^2(x \geq k)$

  =chidist(k,m)