How to analyze many contingency tables simultaneously ?

Thorsten Dickhaus

Humboldt-Universität zu Berlin

Beuth Hochschule für Technik
Berlin, 31.10.2012

Outline

Motivation: Genetic association studies

Statistical setup

Refined statistical inference methods

Real data example

Reference:

Dickhaus, T., Straßburger, K., Schunk, D., Morcillo, C., Illig, T., and Navarro, A. (2012): How to analyze many contingency tables simultaneously in genetic association studies. *SAGMB 11, Article 12.*

What is a SNP (single nucleotide polymorphism) ?

Bi-allelic SNPs: Exactly two possible alleles

Locus	1	2	3	4	...	i	...	M

What is a SNP (single nucleotide polymorphism) ?

Bi-allelic SNPs: Exactly two possible alleles

Locus	1	2	3	4	...	i	...	M
Tom	A	A	G	T	...	A	...	G

What is a SNP (single nucleotide polymorphism) ?

Bi-allelic SNPs: Exactly two possible alleles

Locus	1	2	3	4	...	i	...	M
Tom	A	A	G	T	...	A	...	G
Andrew	A	A	G	C	...	A	...	C

What is a SNP (single nucleotide polymorphism) ?

Bi-allelic SNPs: Exactly two possible alleles

Locus	1	2	3	4	...	i	...	M
Tom	A	A	G	T	...	A	...	G
Andrew	A	A	G	C	...	A	...	C
Rachel	A	A	G	C	...	G	...	G

What is a SNP (single nucleotide polymorphism) ?

Bi-allelic SNPs: Exactly two possible alleles

Locus	1	2	3	4	...	i	...	M
Tom (m)	A	A	G	T	...	A	...	G
Tom (p)	A	A	G	T	...	A	...	C
Andrew	A	A	G	C	...	A	...	C
Rachel	A	A	G	C	...	G	...	G

What is a SNP (single nucleotide polymorphism) ?

Bi-allelic SNPs: Exactly two possible alleles

Locus	1	2	3	4	...	i	...	M
Tom (m)	A	A	G	T	...	A	...	G
Tom (p)	A	A	G	T	...	A	...	C
Andrew	A	A	G	C	...	A	...	C
	A	A	G	C	...	G	...	C
Rachel	A	A	G	C	...	G	...	G
	A	A	G	T	...	G	...	G

Contingency table layout in association studies

Assume a bi-allelic marker (SNP) at a particular locus and a binary phenotype of interest, e. g., a disease status.

Genotype	A_1A_1	A_1A_2	A_2A_2	Σ
Phenotype 1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$n_{1.}$
Phenotype 0	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$n_{2.}$
Absolute count	$n_{.1}$	$n_{.2}$	$n_{.3}$	N

In case of allelic tests:

Genotype	A_1	A_2	Σ
Phenotype 1	$x_{1,1}$	$x_{1,2}$	$n_{1.}$
Phenotype 0	$x_{2,1}$	$x_{2,2}$	$n_{2.}$
Absolute count	$n_{.1}$	$n_{.2}$	N

Formalized association test problem

Multiple test problem with system of hypotheses $\mathcal{H} = (H_j : 1 \leq j \leq M)$, where H_j : Genotype$_j \perp$ Phenotype with two-sided alternatives K_j.

Formalized association test problem

Multiple test problem with system of hypotheses $\mathcal{H} = (H_j : 1 \leq j \leq M)$, where H_j : Genotype$_j \perp$ Phenotype with two-sided alternatives K_j.

Abbreviated notation (one particular position):

$$\mathbf{n} = (n_{1.}, n_{2.}, n_{.1}, n_{.2}, n_{.3}) \in \mathbb{N}^5 \quad \text{resp. } \mathbf{n} = (n_{1.}, n_{2.}, n_{.1}, n_{.2}) \in \mathbb{N}^4 \,,$$

$$\mathbf{x} = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{pmatrix} \in \mathbb{N}^{2\times 3} \text{ resp. } \mathbf{x} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \in \mathbb{N}^{2\times 2}.$$

In both cases, the probability of observing $\mathbf{x}$ given $\mathbf{n}$ is under the null given by

$$f(\mathbf{x}|\mathbf{n}) = \frac{\prod_{n\in\mathbf{n}} n!}{N! \prod_{x\in\mathbf{x}} x!}.$$

Tests for association of marker and phenotype

(i) Chi-squared test

$$Q(\mathbf{x}) = \sum_r \sum_s \frac{(x_{rs} - e_{rs})^2}{e_{rs}}, \quad \text{where} \quad e_{rs} = n_{r.} n_{.s} / N.$$

Resulting "exact" (non-asymptotic) p-value:

$$p_Q(\mathbf{x}) = \sum_{\tilde{\mathbf{x}}} f(\tilde{\mathbf{x}}|\mathbf{n}), \text{ with}$$

summation over all $\tilde{\mathbf{x}}$ with marginals $\mathbf{n}$ such that $Q(\tilde{\mathbf{x}}) \geq Q(\mathbf{x})$.

(Local) level α test: $\varphi_Q(\mathbf{x}) = \mathbf{1}_{p_Q(\mathbf{x}) \leq \alpha}$

Tests for association of marker and phenotype

(ii) Tests of Fisher-type

$$p_{\text{Fisher}}(\mathbf{x}) = \sum_{\tilde{\mathbf{x}}} f(\tilde{\mathbf{x}}|\mathbf{n}), \text{ with}$$

summation over all $\tilde{\mathbf{x}}$ with marginals $\mathbf{n}$ such that $f(\tilde{\mathbf{x}}|\mathbf{n}) \leq f(\mathbf{x}|\mathbf{n})$.

Corresponding level α test: $\varphi_{\text{Fisher}}(\mathbf{x}) = \mathbf{1}_{p_{\text{Fisher}}(\mathbf{x}) \leq \alpha}$

Tests for association of marker and phenotype

(ii) Tests of Fisher-type

$$p_{\text{Fisher}}(\mathbf{x}) = \sum_{\tilde{\mathbf{x}}} f(\tilde{\mathbf{x}}|\mathbf{n}), \text{ with}$$

summation over all $\tilde{\mathbf{x}}$ with marginals $\mathbf{n}$ such that $f(\tilde{\mathbf{x}}|\mathbf{n}) \leq f(\mathbf{x}|\mathbf{n})$.

Corresponding level α test: $\varphi_{\text{Fisher}}(\mathbf{x}) = \mathbf{1}_{p_{\text{Fisher}}(\mathbf{x}) \leq \alpha}$

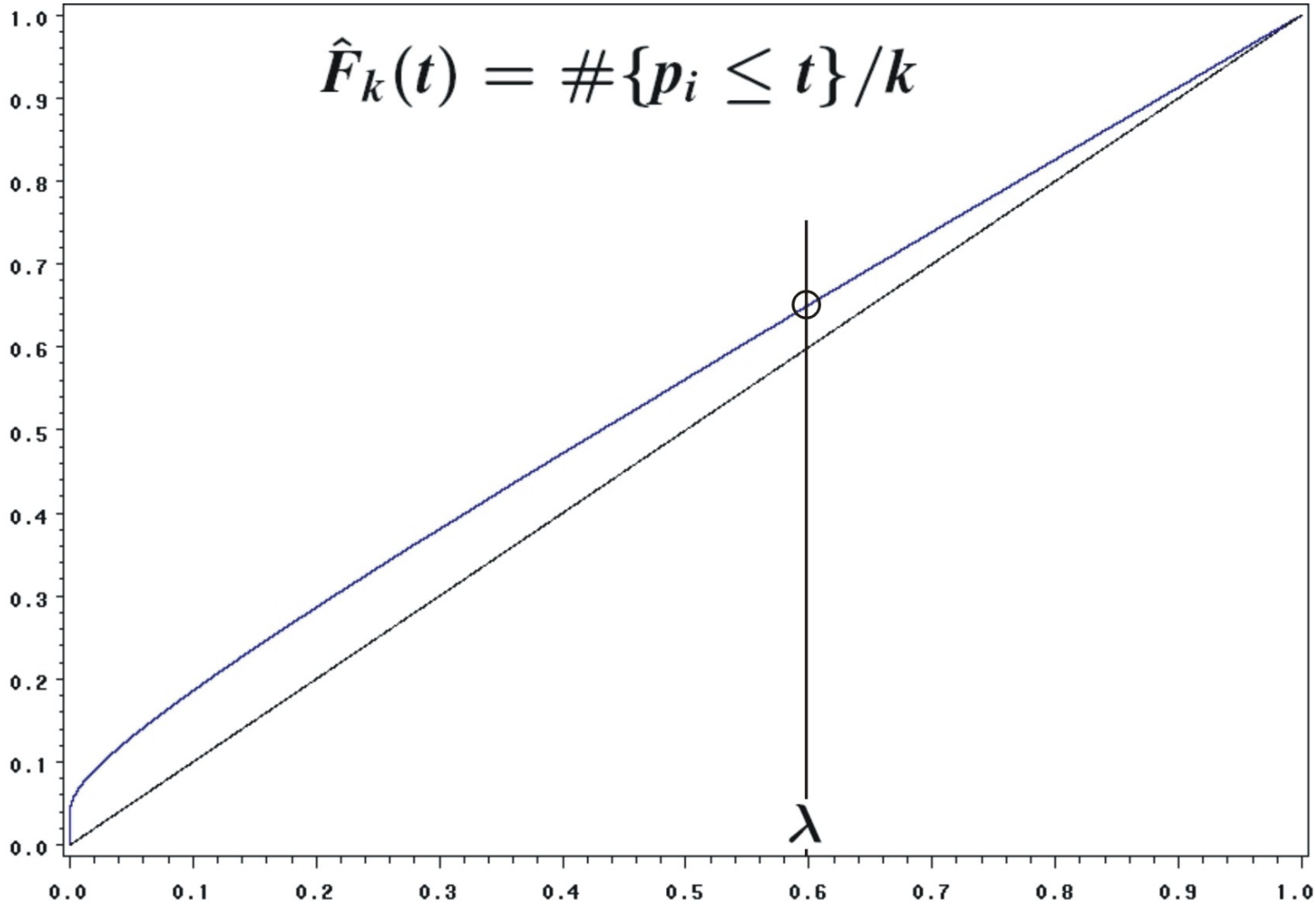
$\varphi_Q(\mathbf{x})$ and $\varphi_{\text{Fisher}}(\mathbf{x})$ keep the (local) significance level α conservatively for any sample size N.

In other words:

$p_Q(\mathbf{X}) \succ U$ and $p_{\text{Fisher}}(\mathbf{X}) \succ U$ under the null, $U \sim UNI[0, 1]$.

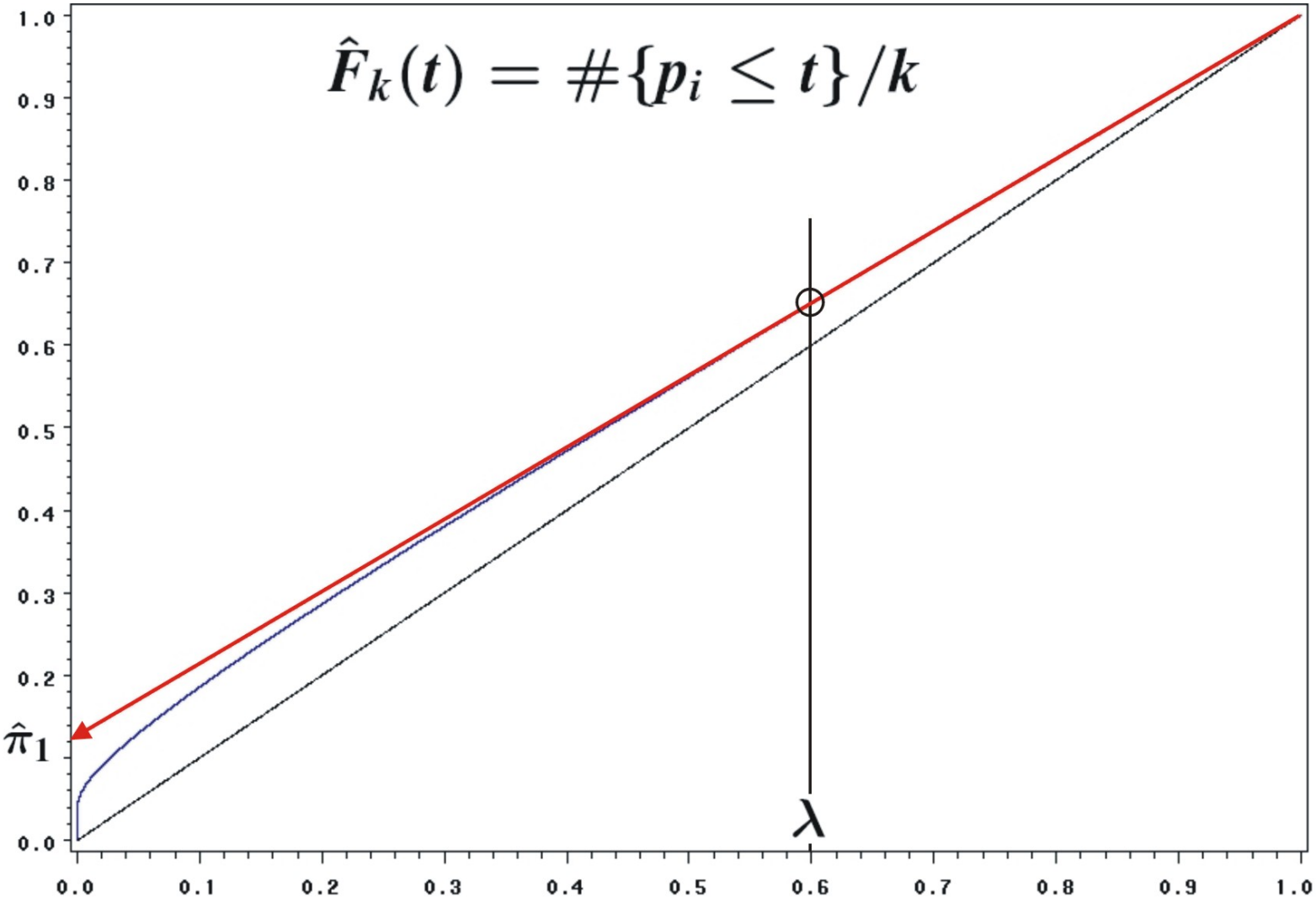
Estimating the proportion of informative SNPs

(References: Schweder and Spjøtvoll (1982), Storey et al., 2004)



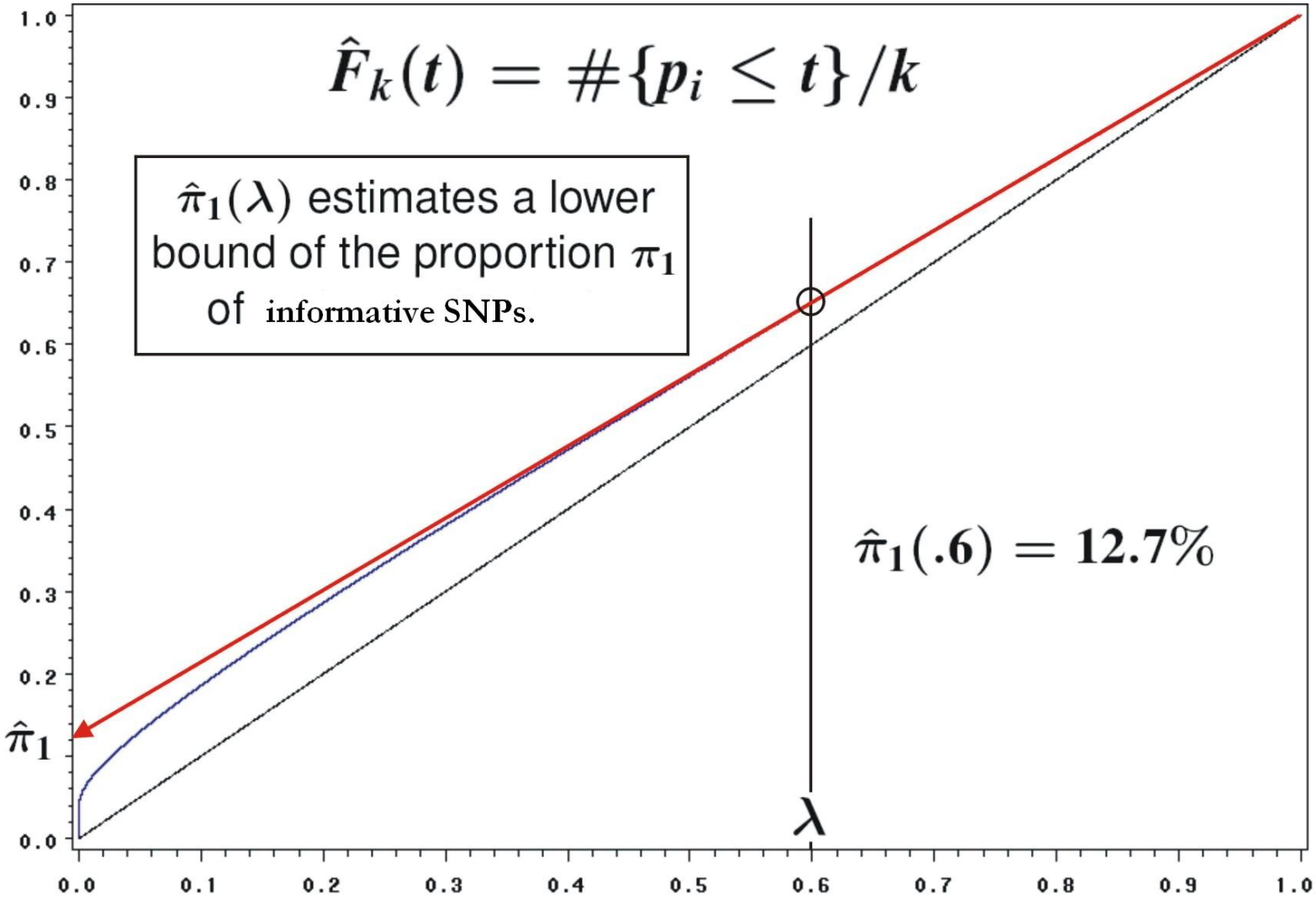
Estimating the proportion of informative SNPs

(References: Schweder and Spjøtvoll (1982), Storey et al., 2004)

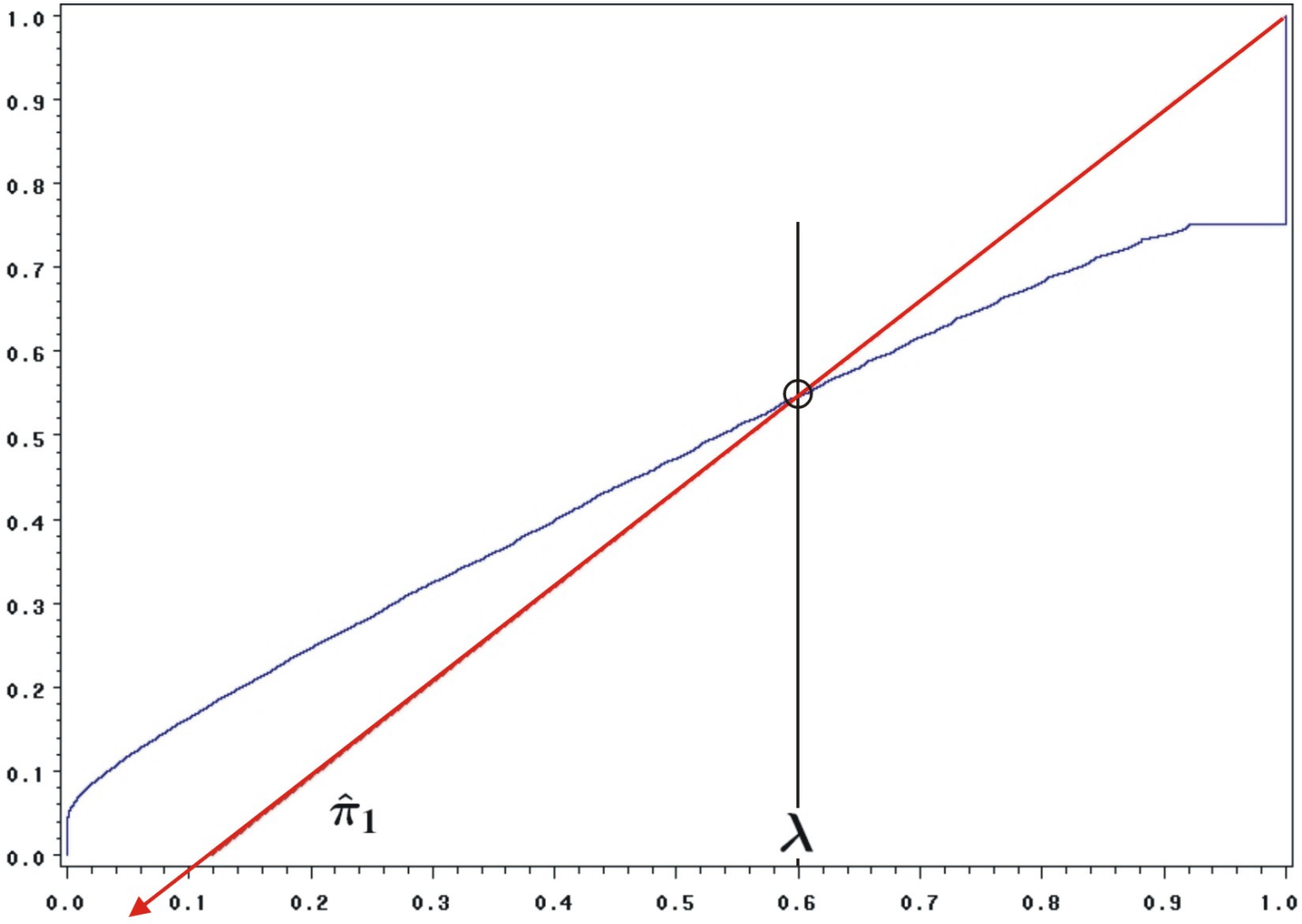


Estimating the proportion of informative SNPs

(References: Schweder and Spjøtvoll (1982), Storey et al., 2004)



Caveat: Storey's method does not work for discrete p-values $p_Q(\mathbf{X})$ and $p_{\text{Fisher}}(\mathbf{X})$



Discreteness: Realized randomized p-values

Definition:

Statistical model $(\Omega, \mathcal{A}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ given

Two-sided test problem $H : \{\vartheta = \vartheta_0\}$ versus $K : \{\vartheta \neq \vartheta_0\}$

Discrete test statistic: $\mathbf{X} \sim \mathbb{P}_\vartheta$ with values in Ω

$U \sim UNI[0, 1]$, stochastically independent of $\mathbf{X}$

A realized randomized p-value for testing H versus K is a measurable mapping $p^r : \Omega \times [0, 1] \to [0, 1]$ with

$$\mathbb{P}_{\vartheta_0}(p^r(\mathbf{X}, U) \leq t) = t \text{ for all } t \in [0, 1].$$

Realized randomized p-values based on $p_Q(\mathbf{X})$ and $p_{\text{Fisher}}(\mathbf{X})$

Lemma:

Based upon the chi-squared and Fisher-type testing strategies, corresponding realized randomized p-values can be calculated as

$$\begin{aligned} p_Q^r(\mathbf{x},u) &= p_Q(\mathbf{x}) - u \sum_{\tilde{\mathbf{x}}:Q(\tilde{\mathbf{x}})=Q(\mathbf{x})} f(\tilde{\mathbf{x}}|\mathbf{n}), \\ p_{\text{Fisher}}^r(\mathbf{x},u) &= p_{\text{Fisher}}(\mathbf{x}) - u\gamma f(\mathbf{x}|\mathbf{n}), \end{aligned}$$

where u denotes the realization of $U \sim UNI[0,1]$, stochastically independent of $\mathbf{X}$ and $\gamma \equiv \gamma(\mathbf{x}) = |\{\tilde{\mathbf{x}} : f(\tilde{\mathbf{x}}|\mathbf{n}) = f(\mathbf{x}|\mathbf{n})\}|$.

We propose realized randomized p-values for estimating π_0.

For final decision making, their non-randomized counterparts should be used (Reproducibility!).

Effective number of tests

A thought experiment

Assume markers indexed by $I = \{1, \ldots, M\}$ can be divided into disjoint groups with indices in subsets $I_g \subset I, g \in \{1, \ldots, G\}$.

Effective number of tests

A thought experiment

Assume markers indexed by $I = \{1, \ldots, M\}$ can be divided into disjoint groups with indices in subsets $I_g \subset I, g \in \{1, \ldots, G\}$.

Let $\varphi = (\varphi_i, i \in I)$ and assume that for each $g \in \{1, \ldots, G\}$ and for any pair $(i,j) \subseteq I_g$ the identity $\{\varphi_i = 1\} = \{\varphi_j = 1\}$ holds.

Then, "effectively" only one single test is performed in each subgroup.

Effective number of tests

A thought experiment

Assume markers indexed by $I = \{1, \ldots, M\}$ can be divided into disjoint groups with indices in subsets $I_g \subset I, g \in \{1, \ldots, G\}$.

Let $\varphi = (\varphi_i, i \in I)$ and assume that for each $g \in \{1, \ldots, G\}$ and for any pair $(i,j) \subseteq I_g$ the identity $\{\varphi_i = 1\} = \{\varphi_j = 1\}$ holds.

Then, "effectively" only one single test is performed in each subgroup. Denoting $i(g) = \min I_g$ for $g = 1, \ldots, G$, it holds

$$\mathrm{FWER}_\vartheta(\varphi) = \mathbb{P}_\vartheta\left(\bigcup_{g=1}^{G} \bigcup_{i \in I_0 \cap I_g} \{\varphi_i = 1\}\right) \leq \mathbb{P}_\vartheta\left(\bigcup_{g=1}^{G} \{\varphi_{i(g)} = 1\}\right).$$

Consequently, multiplicity correction in this extreme scenario only has to be done with respect to $G << M$.

Bonferroni-type adjustment α/G would be valid!

Effective number of tests

Cheverud-Nyholt method and beyond

$$M_{\text{eff.}} = 1 + \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{M}(1 - r_{ij}^2).$$

The numbers r_{ij} are measures of correlation among markers i and j and can typically be obtained from linkage disequilibrium (LD) matrices.

More sophisticated methods exist in the literature, e. g.:

- simple$\mathcal{M}$ by X. Gao et al. (2008)
- $K_{\text{eff.}}$ by Moskvina and Schmidt (2008)

All rely on the correlation structure reflected by the r_{ij}'s.

Our proposed data analysis workflow

1. Compute realized randomized p-values $p^r(\mathbf{x}_j, u_j)$ and non-randomized versions $p(\mathbf{x}_j), j = 1, \ldots, M$.
2. Estimate the proportion π_0 of uninformative SNPs by $\hat{\pi}_0$.
3. Determine the effective number of tests $M_{\text{eff.}}$ by utilizing correlation values obtained from an appropriate LD matrix of the M SNPs.
4. For a pre-defined FWER level α, determine the list of associated markers by performing the multiple test $\varphi = (\varphi_j, j = 1, \ldots, M)$, where $\varphi_j(\mathbf{x}_j) = \mathbf{1}_{p(\mathbf{x}_j) \leq t^*}$ with $t^* = \alpha / (M_{\text{eff.}} \cdot \hat{\pi}_0)$.

Real data example: Herder et al. (2008)

Replication study

Herder, C. et al. (2008). Variants of the PPARG, IGF2BP2, CDKAL1, HHEX, and TCF7L2 genes confer risk of type 2 diabetes independently of BMI in the German KORA studies. Horm. Metab. Res. 40, 722–726.

Data:

$M = 44$ SNPs on ten different genes
($N \approx 1900$ study participants)

"Results" section:

"...(conservative) Bonferroni correction for 10 genes..."

Authors' claim:

Threshold $t^* = 0.005$ for raw marginal p-values controls the FWER at $\alpha = 5\%$

Herder et al. (2008): Data re-analysis

LD information:

Taken from the HapMap project (population 'CEU')

Estimated effective number of tests:

$M_{\text{eff.}} = 40.63$ (Cheverud-Nyholt method),
$K_{\text{eff.}} = 16.73$ (Moskvina-Schmidt method).

Estimated proportion of uninformative SNPs:

$\hat{\pi}_0 = 0.4545$ (Storey et al., 2004)

Resulting threshold according to our method:

$$t^* = \alpha/(K_{\text{eff.}} \times \hat{\pi}_0) = \alpha/(16.73 \cdot 0.4545) = \alpha/7.604 = 0.0066.$$

In conclusion:

Our proposed method confirms the authors' heuristic argumentation and endorses their scientific claims.

Future research goals

- Effective number of tests for continuous response
- Effective number of tests for FDR control
- Adaptive estimation of effective numbers of tests

- Statistical methodology for confirmatory functional studies (fMRI data)
- Hierarchical multiple testing methods for (auto-) correlated data (time series)