Supplementary material for the article

ESTIMATING EFFECT SIZES IN GENOME-WIDE ASSOCIATION STUDIES

József Bukszár and Edwin J. C. G. van den Oord

Derivation of (5):

In our article ([1]) we showed that an asymptotic approximation to Pearson's statistic on a $2\times\upsilon$ contingency table is given by

$$Z_1^2+...+Z_\upsilon^2, \tag{1}$$

where $Z_1,...,Z_\upsilon$ are independent normal random variables with

$$E\left(Z_i\right)=\mathbf{a}_i\mu \ \text{ and } Var\left(Z_i\right)=\lambda_i, \tag{2}$$

where

$$\mu^T=\left(\frac{\left(p_1-q_1\right)\sqrt{\gamma\delta n}}{\sqrt{\gamma p_1+\delta q_1}},...,\frac{\left(p_m-q_m\right)\sqrt{\gamma\delta n}}{\sqrt{\gamma p_m+\delta q_m}}\right),$$

and $\mathbf{a}_i$, $i=1,...,\upsilon$, are the unit length eigenvectors of matrix J with corresponding eigenvalues λ_i, and the entries of J are given by

$$J_{ij}=\begin{cases} -\frac{1}{\sqrt{p_i\gamma+q_i\delta}\sqrt{p_j\gamma+q_j\delta}}\left[\left(1+\frac{(p_i-q_i)\delta}{2(p_i\gamma+q_i\delta)}\right)\left(1+\frac{(p_j-q_j)\delta}{2(p_j\gamma+q_j\delta)}\right)\gamma q_iq_j+\right.\\ \qquad\left.\left(1-\frac{(p_i-q_i)\gamma}{2(p_i\gamma+q_i\delta)}\right)\left(1-\frac{(p_j-q_j)\gamma}{2(p_j\gamma+q_j\delta)}\right)\delta p_ip_j\right] \qquad \text{if } i\neq j\\ \frac{1}{p_i\gamma+q_i\delta}\left[\left(1+\frac{(p_i-q_i)\delta}{2(p_i\gamma+q_i\delta)}\right)^2\gamma q_i\left(1-q_i\right)+\left(1-\frac{(p_i-q_i)\gamma}{2(p_i\gamma+q_i\delta)}\right)^2\delta p_i\left(1-p_i\right)\right] \quad \text{if } i=j.\end{cases} \tag{3}$$

It can be verified that under the usual null hypothesis, i.e. when $p_i=q_i$ for all $i=1,...,\upsilon$, the sum in (1) follows the chi-square distribution with $\upsilon-1$ degree of freedom. In fact, under the null hypothesis $\mu=\mathbf{0}$ and J reduces to J^0 with $J_{ij}^0=-\sqrt{p_ip_j}$ if $i\neq j$ and $J_{ij}^0=1-p_i$ if $i=j$. Thus, J^0 has two eigenvalues, 1 with multiplicity $\upsilon-1$, and 0 with multiplicity 1, which implies that $\upsilon-1$ of the Z's in (1) has standard normal distribution and the remaining one is 0 with probability 1.

The approximation in (1) is the asymptotic equivalent ([1]), which is at least as accurate as the central chi-square with $\upsilon-1$ degrees of freedom for approximating the distribution of Pearson's statistic under the null hypothesis. However, there are too many parameters involved in J. Therefore, we now give another approximation that involves just one parameter and will subsequently be used to estimate p_0. Note that all small fractions in parentheses in (3) are close to 0. By deleting these fractions, we obtain matrix G whose entries are

$$G_{ij}=\begin{cases} -\frac{1}{\sqrt{p_i\gamma+q_i\delta}\sqrt{p_j\gamma+q_j\delta}}\left[\gamma q_iq_j+\delta p_ip_j\right] \ \text{if } i\neq j\\ \frac{1}{p_i\gamma+q_i\delta}\left[\gamma q_i\left(1-q_i\right)+\delta p_i\left(1-p_i\right)\right] \ \text{if } i=j.\end{cases} \tag{4}$$

We show that the approximation based on matrix G rather than J has the form given in (**??**) if $\gamma=\delta$. If $\gamma=\delta$, then the entries of G equal

$$G_{ij}=\left\{\begin{array}{l} -\frac{1}{\sqrt{p_i+q_i}\sqrt{p_j+q_j}}\left[q_iq_j+p_ip_j\right] \text{ if } i\neq j\\ \frac{1}{p_i+q_i}\left[q_i\left(1-q_i\right)+p_i\left(1-p_i\right)\right] \text{ if } i=j. \end{array}\right. \tag{5}$$

First we show that the eigenvalues of G are 1 with multiplicity $m-2$, 0 with multiplicity 1 and $2\sum_{i=1}^{m}\frac{p_iq_i}{p_i+q_i}$ with multiplicity 1. Matrix G can be written in the form

$$G=I-D\left(\mathbf{p}\mathbf{p}^T+\mathbf{q}\mathbf{q}^T\right)D,$$

where $D=diag\left(\frac{1}{\sqrt{p_i+q_i}}\right)$, $\mathbf{p}=\left(p_1,...,p_\upsilon\right)^T$ and $\mathbf{q}=\left(q_1,...,q_\upsilon\right)$. First we verify that vector $\mathbf{x}=D^{-1}\mathbf{1}=\left(\sqrt{p_1+q_1},...,\sqrt{p_m+q_m}\right)$ is an eigenvector of G with eigenvalue 0 by

$$G\mathbf{x}=\mathbf{x}-D\left(\mathbf{p}\mathbf{p}^T+\mathbf{q}\mathbf{q}^T\right)\mathbf{1}=\mathbf{x}-D\left(\mathbf{p}\left(\mathbf{p}^T\mathbf{1}\right)+\mathbf{q}\left(\mathbf{q}^T\mathbf{1}\right)\right)=\mathbf{x}-D\left(\mathbf{p}+\mathbf{q}\right)=\mathbf{x}-\mathbf{x}=\mathbf{0},$$

where $\mathbf{1}$ is a υ-dimensional vector with components 1. Furthermore, if $\mathbf{x}=D^{-1}\mathbf{z}$, where $\mathbf{z}$ is orthogonal to both $\mathbf{p}$ and $\mathbf{q}$, i.e. $\mathbf{p}^T\mathbf{z}=\mathbf{q}^T\mathbf{z}=0$, then $\mathbf{x}$ is an eigenvector of G with eigenvalue 1, because

$$G\mathbf{x}=\mathbf{x}-D\left(\mathbf{p}\mathbf{p}^T+\mathbf{q}\mathbf{q}^T\right)\mathbf{z}=\mathbf{x}-D\left(\mathbf{p}\mathbf{p}^T\mathbf{z}+\mathbf{q}\mathbf{q}^T\mathbf{z}\right)=\mathbf{x}.$$

Consequently, the eigenvectors of G with eigenvalue 1 span a $(\upsilon-2)$-dimensional eigenspace if $\mathbf{p}\neq\mathbf{q}$, and a $(\upsilon-1)$-dimensional eigenspace if $\mathbf{p}=\mathbf{q}$. Thus, if $\mathbf{p}=\mathbf{q}$, then there is no other eigenvector of G, and if $\mathbf{p}\neq\mathbf{q}$, then there is one more left, particularly $D\left(\mathbf{p}-\mathbf{q}\right)$. This is verified by

$$\begin{aligned}&\left\{I-D\left(\mathbf{p}\mathbf{p}^T+\mathbf{q}\mathbf{q}^T\right)D\right\}D\left(\mathbf{p}-\mathbf{q}\right)=D\left\{\left(\mathbf{p}-\mathbf{q}\right)-\left(\mathbf{p}\mathbf{p}^T+\mathbf{q}\mathbf{q}^T\right)D^2\left(\mathbf{p}-\mathbf{q}\right)\right\}\overset{*}{=}\\ &D\left\{\left(\mathbf{p}-\mathbf{q}\right)+\left(\mathbf{p}-\mathbf{q}\right)\mathbf{q}^TD^2\left(\mathbf{p}-\mathbf{q}\right)\right\}=D\left(\mathbf{p}-\mathbf{q}\right)\left\{1+\mathbf{q}^TD^2\left(\mathbf{p}-\mathbf{q}\right)\right\},\end{aligned} \tag{6}$$

where at $*$ we used that $\mathbf{p}^TD^2\left(\mathbf{p}-\mathbf{q}\right)=-\mathbf{q}^TD^2\left(\mathbf{p}-\mathbf{q}\right)$, which holds because of

$$\mathbf{p}^TD^2\left(\mathbf{p}-\mathbf{q}\right)+\mathbf{q}^TD^2\left(\mathbf{p}-\mathbf{q}\right)=\left(\mathbf{p}+\mathbf{q}\right)^TD^2\left(\mathbf{p}-\mathbf{q}\right)=\mathbf{1}^T\left(\mathbf{p}-\mathbf{q}\right)=0.$$

The equality in (6) also shows that the eigenvalue corresponding to eigenvector $D\left(\mathbf{p}-\mathbf{q}\right)$ is

$$1+\mathbf{q}^TD^2\left(\mathbf{p}-\mathbf{q}\right)=1+\sum_j\frac{q_j\left(p_j-q_j\right)}{p_j+q_j}=\sum_j\frac{q_j\left(p_j+q_j\right)}{p_j+q_j}+\sum_j\frac{q_j\left(p_j-q_j\right)}{p_j+q_j}=2\sum_j\frac{p_jq_j}{p_j+q_j}.$$

Finally, note that

$$D\left(\mathbf{p}-\mathbf{q}\right)=\left(\frac{p_1-q_1}{\sqrt{p_1+q_1}},...,\frac{p_\upsilon-q_\upsilon}{\sqrt{p_\upsilon+q_\upsilon}}\right)^T=\sqrt{\frac{2}{n}}\mu.$$

Since G is symmetric, all eigenvectors corresponding to eigenvalue 1 are orthogonal to eigenvector $D\left(\mathbf{p}-\mathbf{q}\right)$, and hence to μ.

The approximation to Pearson's statistic based on matrix G rather than J has the form

$$U_1^2+...+U_\upsilon^2,$$

where $U_1,...,U_\upsilon$ are independent normal random variables with $E\left(U_i\right)=\mathbf{b}_i\mu$ and $Var\left(U_i\right)=\varepsilon_i$, and $\mathbf{b}_i$, $i=1,...,\upsilon$, are the unit length eigenvectors of G with corresponding eigenvalues ε_i. As we have shown above the eigenvalues of G are $\varepsilon_1=...=\varepsilon_{\upsilon-2}=1$, $\varepsilon_{\upsilon-1}=2\sum_{j=1}^{\upsilon}\frac{p_jq_j}{p_j+q_j}$, $\varepsilon_\upsilon=0$ and $\mathbf{b}_1,...,\mathbf{b}_{\upsilon-2}$ are orthogonal to

μ, hence $U_\upsilon = 0$ and $U_1^2 + ... + U_{\upsilon-2}^2$ is a central chi-square random variable with $\upsilon - 2$ degrees of freedom. Furthermore, since

$$\mathbf{b}_{\upsilon-1}^T = \frac{1}{\sqrt{\sum_{j=1}^{\upsilon} \frac{(p_j-q_j)^2}{p_j+q_j}}} \left(\frac{p_1 - q_1}{\sqrt{p_1 + q_1}}, ..., \frac{p_\upsilon - q_\upsilon}{\sqrt{p_\upsilon + q_\upsilon}} \right),$$

we have

$$E\left(U_{\upsilon-1}\right) = \mathbf{b}_{\upsilon-1}^T \mu = \sqrt{\frac{n}{2} \sum_{j=1}^{\upsilon} \frac{\left(p_j - q_j\right)^2}{p_j + q_j}} = \sqrt{n}\Delta \ ,$$

where

$$\Delta = \sqrt{\frac{1}{2} \sum_{j=1}^{\upsilon} \frac{\left(p_j - q_j\right)^2}{p_j + q_j}}.$$

Since

$$1 - \Delta^2 = \frac{1}{2} \sum_{j=1}^{\upsilon} \frac{\left(p_j + q_j\right)^2}{p_j + q_j} - \frac{1}{2} \sum_{j=1}^{\upsilon} \frac{\left(p_j - q_j\right)^2}{p_j + q_j} = 2 \sum_{j=1}^{\upsilon} \frac{p_j q_j}{p_j + q_j},$$

we have

$$Var\left(U_{\upsilon-1}\right) = 1 - \Delta^2.$$

We obtained that the approximation to Pearson's statistic based on G is the one given in (5) if $\gamma = \delta$ in the manuscript.

References

[1] Bukszár J, Van den Oord EJCG. Accurate and efficient power calculations for $2 \times m$ tables in unmatched case-control designs. *Statistics in Medicine.* 2006; **25**:2632-2646.