A formula for the correlation between random-set enrichment scores

Michael A. Newton[1]

January 2007

UW Madison Statistics Department Technical Report #1134

[1]Departments of Statistics and of Biostatistics and Medical Informatics (Dept of Statistics; 1300 University Avenue; Madison, WI 53706; newton@stat.wisc.edu), University of Wiscsonsin-Madison, Madison, Wisconsin.

Abstract

Given constants $s_1, s_2, \ldots, s_G$, we consider variables $X = \sum s_g 1[g \in A]$ and $Y = \sum s_g 1[g \in B]$ as random variables with a joint distribution determined by taking (A, B) as uniformly random among set pairs in which A has fixed cardinality m, B has fixed cardinality n, and the intersection $A \cap B$ has fixed size q. Arguments about without-replacement sampling give the marginal mean and variance of X and Y separately. Here we extend those calculations and compute the correlation between X and Y. The joint distribution so determined is helpful in gene set enrichment analysis.

KEYWORDS: gene set enrichment .

1 Problem

The penultimate section in Newton et al. (2006) presents without proof a formula for the correlation between two Z scores, these being the standardized enrichment scores for two possibly overlapping gene sets. Here we show how to derive the correlation formula.

Start with gene-level scores $s_1, s_2, \ldots, s_G$ for G genes. These may be measures of differential expression, or some related quantity, but the important thing is that they are treated as fixed [i.e. we condition on them]. WLOG suppose they are normalized so $\sum_g s_g/G = 0$ and $\sum_g s_g^2/G = 1$.

Unstandardized enrichment scores X and Y are defined

$$X = \sum_{g \in A} s_g \qquad Y = \sum_{g \in B} s_g$$

where A and B are random subsets of $\{1, 2, \ldots, G\}$ that are constrained so that $\#A = m$, $\#B = n$, and $\#(A \cap B) = q$. These set sizes $m, n,$ and q are considered

fixed, and we consider the pair (A, B) to arise uniformly at random from the collection $\mathcal{S}$ of all possible *networks* of set pairs satisfying the size constraints. Note that $q \leq m$, $q \leq n$ and $m + n - q \leq G$ else we do not have legitimate sets. The problem is to compute the correlation between X and Y owing to randomness in the set pair, noting that gene-level scores are fixed.

Claim:

$$\begin{aligned} \text{corr}(X, Y) &= \frac{Gq - mn}{\sqrt{m(G-m)n(G-n)}} \\ &= \frac{q}{\sqrt{mn}} + O\left(\frac{1}{G}\right). \end{aligned}$$

2 Solution

Observe first that the cardinality of $\mathcal{S}$ is

$$\#\mathcal{S} = \frac{G!}{q!\,(m-q)!\,(n-q)!\,(G-m-n+q)!}. \tag{1}$$

This follows by making a correspondence between the four components of our two overlapping sets: i.e. $A \cap B$, $A \cap B^c$, $B \cap A^c$ and $(A \cup B)^c$ and fixed sized subsets of $\{1, 2, \ldots G\}$, as in multinomial sampling. Thus the probability to realize a particular (A, B) is $1/\#\mathcal{S}$.

The marginal means and variances X and Y are known from without-replacement sampling (Newton *et al.* 2006). With the s_g's centered, $E(X) = E(Y) = 0$, and the variances are

$$\text{var}(X) = \frac{m(G-m)}{G-1} \qquad \text{var}(Y) = \frac{n(G-n)}{G-1}. \tag{2}$$

It remains, therefore, to compute $E(XY)$ in order to obtain the correlation. From the definition,

$$\begin{aligned} E(XY) &= \sum_{(A,B)\in\mathcal{S}} \frac{1}{\#\mathcal{S}} \sum_{g\in A} \sum_{h\in B} s_g s_h \\ &= \frac{1}{\#\mathcal{S}} \sum_{g=1}^{G} \sum_{h=1}^{G} s_g s_h k_{g,h} \end{aligned} \tag{3}$$

where

$$k_{g,h} = \sum_{(A,B)\in\mathcal{S}} 1[g \in A]\, 1[h \in B] \tag{4}$$

The simpler situation to consider has $g = h$. Then $k_{g,g} = \sum_{(A,B)\in\mathcal{S}} 1[g \in A \cap B]$. We are counting set pairs (A, B) that have a fixed gene g in their intersection, which, as defined, is of a fixed size q. Of course if $q = 0$ then $k_{g,g} = 0$. Otherwise it is useful again to make the correspondence between a set pair (A, B) and an allocation of the G genes into four groups of fixed sizes. Presently we are fixing gene g to be in $A \cap B$, so we count ways to allocate the other $G - 1$ genes to groups of sizes $q - 1$ (the rest of $A \cap B$), $m - q$ (stuff in $A \cap B^c$), $n - q$ (stuff in $A^c \cap B$) and $G - m - n + q$ (remainder). Thus,

$$k_{g,g} = \frac{(G-1)!}{(q-1)!\,(m-q)!\,(n-q)!\,(G-m-n+q)!}. \tag{5}$$

Taken against the probability of a set pair,

$$\frac{k_{g,g}}{\#\mathcal{S}} = \frac{q}{G}. \tag{6}$$

When $g \neq h$ in (4), it is useful to consider four subsets of $\mathcal{S}$ (relative to the fixed

g and h), depending on where the two genes land:

$$\begin{aligned}
\mathcal{S}_1 &= \{(A,B) : g,h \in A \cap B\} \\
\mathcal{S}_2 &= \{(A,B) : g \in A \cap B^c \, , h \in A \cap B\} \\
\mathcal{S}_3 &= \{(A,B) : g \in A \cap B \, , h \in A^c \cap B\} \\
\mathcal{S}_4 &= \{(A,B) : g \in A \cap B^c \, , h \in A^c \cap B\}
\end{aligned}$$

These compenents may be empty depending on values $m, n,$ and q; for example $\mathcal{S}_1$ is non-empty only if $q \geq 2$. But importantly $k_{g,h} = \#\mathcal{S}_1 + \#\mathcal{S}_2 + \#\mathcal{S}_3 + \#\mathcal{S}_4$. By the same counting approach to derive (5), we get

$$\#\mathcal{S}_1 = \frac{(G-2)!}{(q-2)!\,(m-q)!\,(n-q)!\,(G-m-n+q)!},$$

$$\#\mathcal{S}_2 = \frac{(G-2)!}{(q-1)!\,(m-q-1)!\,(n-q)!\,(G-m-n+q)!},$$

$$\#\mathcal{S}_3 = \frac{(G-2)!}{(q-1)!\,(m-q)!\,(n-q-1)!\,(G-m-n+q)!},$$

and

$$\#\mathcal{S}_3 = \frac{(G-2)!}{q!\,(m-q-1)!\,(n-q-1)!\,(G-m-n+q)!}.$$

Simplifying in relation to the probability of a set pair, we obtain, for $g \neq h$,

$$\frac{k_{g,h}}{\#\mathcal{S}} = \frac{mn-q}{G(G-1)}. \tag{7}$$

Reconsidering the expectation $E(XY)$ from (3), we take advantage of the fact that $k_{g,h}$ has one value when $g = h$ (6) and one other value when $g \neq h$ (7). We

combine, using the centering assumption $\sum_g s_g = 0$ and the scaling assumption $\sum_g s_g^2 = G$, to get

$$E(XY) = \frac{Gq - mn}{G - 1}$$

which leads to the claimed correlation, noting (2).

References

Newton, M. A., Quintana, F. A., den Boon, J. A., Sengupta, S., and Ahlquist, P. (2006), "Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis," *UW Statistics Department Technical Report*, 1130.