Number of inversions in multiset permutations

Janosch Ortmann, Zhaleh Safikhani, Petr Smirnov, Ian Smith

December 2018

1 Introduction

More context: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4314453/ (Kang et al., 2015)

The concordance index is a non-parametric metric for comparison between two orderings on a set. In practice, this metric can be used to determine whether a candidate biomarker is informative of a clinical outcome, like sensitivity to anti-cancer drugs. The metric is applied two a numeric response vector and a numeric predictor vector; both take real numbers and can be tied. For example, the response vector might be the area-above-dose response curve (AAC) of cancer cell lines to a particular anti-cancer drug (e.g. lapatinib), and the candidate biomarker might be the gene expression of EGFR in TPM from RNA-Seq data. The bioinformatic question is then to compute the predictive value of the candidate biomarker for the response vector, assign significance and quantify effect size, and test for generalizability in other data sets. Concordance index is a metric for testing significance, though it also has an effect size interpretation. Concordance index is also analgoous to the area under the ROC curve if the predictor vector or candidate biomarker is considered as a classifier or regressor for the response vector. Concordance index - in its vanilla form - is also essentially equivalent to (a linear transformation of) Kendall's Tau.

As discussed in Obuchowski (2006), the area under the ROC curve is an obvious metric for identifying predictors for binary response variables. However, many response variables are continuous - for instance IC50 or AAC for drug response data, or survival time for clinical trials. Continuous-valued data can be binarized by imposing a threshold, but this choice is often arbitrary. Obuchowski points out through simulation that binarization creates problems and incorrectly estimates the association between the two variables.

Insert more context here.

As in reference https://papers.nips.cc/paper/3375-on-ranking-in-survival-analysis-bounds-on-the-concordance-index.pdf, formally, we define concordance index as:

$$CI(x,y) \equiv \frac{1}{E}\sum_{ij} 1(x_i < x_j, y_i < y_j) \tag{1}$$

The best way to think about permuting elements with ties between them is as follows: Let the set of distinct elements to be permuted be denoted $E = \{e_1, ..., e_n\}$ and let $a_j \in \mathbb{Z}_{>0}$ denote the multiplicity of element e_j, that is how often it appears. Thus there are $\alpha := \sum_{j=1}^{n} \alpha_j$ elements in total. We denote by $M = \{e_1^{\alpha_1}, ..., e_n^{\alpha_n}\}$ the multiset containing all elements (with ties).

2 Exact expressions

2.1 Explicit formula

In Margolius (2001) we have the following result for *sets*, that is $\alpha_j = 1$ for all j (and hence $\alpha = n$): let $I_n(k)$ denote the number of inversions of S with k inversions then

$$\Phi_n(x) := \sum_{j=1}^{\binom{n}{2}} I_n(x) x^k = \prod_{j=1}^{n} \sum_{k=1}^{j-1} x^k. \tag{2}$$

It turns out that an analogous result can be obtained for multisets. The original reference is to a paper from 1915 – see [Mac15] in Remmel and Wilson (2015) – but it's easier to read in modern notation. Let $\mathrm{inv}(\sigma)$ denote the number of inversions of a permutation of the multiset (set with ties) S. In Remmel and Wilson (2015) the *distribution* of inv is shown to be

$$D_M(x) = \sum_{\sigma \in S_M} x^{\mathrm{inv}(\sigma)} = \begin{bmatrix} \alpha \\ \alpha_1 ... \alpha_n \end{bmatrix}_x = \frac{\alpha!_x}{\alpha_1!_x .. \alpha_n!_x} \tag{3}$$

with the *q-factorial* being defined by

$$m!_x = \prod_{k=1}^{r} \left(1 + x + ... + x^{k-1}\right) \tag{4}$$

(The expression on the right-hand side of (3) is also called the *q-multinomial coefficient*. Observe that, by splitting the sum over S_M according to the number of inversions,

$$D_M(x) = \sum_{k \geq 0} \sum_{\substack{\sigma \in S_M \\ \mathrm{inv}(\sigma) = k}} x^{\mathrm{inv}\,\sigma} = \sum_{k \geq 0} \sum_{\substack{\sigma \in S_M \\ \mathrm{inv}(\sigma) = k}} x^k = \sum_{k \geq 0} I_M(k) x^k \tag{5}$$

where $I_M(k)$ denotes the number of permutations of the multiset M with k inversions. Thus, (3) is the exact multiset analogue of (2).

2.2 Recursive formula

The paper Margolius (2001) also has a recursion formula, expressing $I_n(k)$ in terms of the $I_{n-1}(j)$: in terms of the generating function this reads

$$\Phi_n(x) = \left(\sum_{k=0}^{n-1} x^k \right) \Phi_{n-1}(x). \tag{6}$$

The proof proceeds by looking at permutations of the first $n-1$ elements and then inserting the last element at all possible position. By keeping track of how many extra inversions this insertion introduces, we arrive at (6).

This argument extends rather well to the case with ties: let M be the multiset as described in the introduction and denote by M^- the set obtained from M by removing one occurrence of e_n. That is, if $M = e_1^{\alpha_1}, .., e_n^{\alpha_n}$, then

$$M^- = e_1^{\alpha_1}, e_2^{\alpha_2}, \ldots, e_{n-1}^{\alpha_{n-1}}, e_n^{\alpha_n - 1}, \tag{7}$$

and in particular if $|M| = n$ then $|M^-| = n - 1$. We can give the following analogue of (6):

$$D_M(x) = \begin{bmatrix} \alpha \\ \alpha_n \end{bmatrix}_x D_{M^-}(x) = \frac{\alpha!_x}{(\alpha - \alpha_n)!_x \alpha_n!_x} D_{M^-}(x), \tag{8}$$

with $m!_x$ defined in (4).

3 Gaussian approximation

In Margolius (2001), asymptotics are also discussed. It seems like there is also a Gaussian approximation result for the inversions in the multiset case, see Conger and Viswanath (2007).

References

Mark Conger and D. Viswanath. Normal Approximations for Descents and Inversions of Permutations of Multisets arXiv : math / 0508242v2 [math . PR] 29 Sep 2006. *Journal of Theoretical Probability*, 20(2):1–16, may 2007. ISSN 0894-9840. doi: 10.1007/s10959-007-0070-5. URL http://link.springer.com/10.1007/s10959-007-0070-5.

Le Kang, Weijie Chen, Nicholas A Petrick, and Brandon D Gallas. Comparing two correlated C indices with right-censored survival outcome: A one-shot nonparametric approach. *Statistics in Medicine*, 34(4):685–703, feb 2015. ISSN 10970258. doi: 10.1002/sim.6370. URL http://www.ncbi.nlm.nih.gov/pubmed/25399736 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4314453.

Barbara H Margolius. Permutations with inversions. *Journal of Integer Sequences*, 4(2), 2001. ISSN 15307638. URL https://cs.uwaterloo.ca/journals/JIS/VOL4/MARGOLIUS/inversions.pdf.

Nancy A. Obuchowski. An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale. *Statistics in Medicine*, 25(3):481–493, feb 2006. ISSN 02776715. doi: 10.1002/sim.2228. URL http://www.ncbi.nlm.nih.gov/pubmed/16287217 http://doi.wiley.com/10.1002/sim.2228.

Jeffrey B. Remmel and Andrew Timothy Wilson. An extension of MacMahon's equidistribution theorem to ordered set partitions. *Journal of Combinatorial Theory. Series A*, 134:242–277, aug 2015. ISSN 10960899. doi: 10.1016/j.jcta.2015.03.012. URL https://linkinghub.elsevier.com/retrieve/pii/S0097316515000497.