Un nouveau test statistique pour la comparaison de proportions

Éric TAILLARD[1], Philippe WAELTI[1], Jacques ZUBER[1]

1. EIVD, Route de Cheseaux 1, CP, CH-1401 Yverdon-les-Bains, Suisse

{eric.taillard, philippe.waelti, jacques.zuber} [at] eivd.ch

Mots-clefs : Test statistique, comparaison d'heuristiques

Qui n'a pas été une fois perplexe en lisant, dans un article comparant deux méthodes d'optimisation, des résultats numériques présentés sous la forme suivante : «Nous avons testé notre méthode A sur un jeu de n exemples de problèmes de la littérature et nous avons réussi à résoudre a de ces exemples. La méthode concurrente B n'a réussi à en résoudre que b, mais elle n'a été testée que sur m exemples de problèmes.»

Un tel résultat peut laisser le lecteur perplexe, car il ne dispose pas immédiatement de la réponse à la question de base qu'il doit légitimement se poser : «Est-ce qu'un taux de succès de a/n est significativement supérieur à un taux de succès de b/m ?» Ce que nous appelons «succès» doit naturellement être défini par l'usager : il peut s'agir de l'obtention d'une solution optimale ou de qualité donnée, ou encore simplement d'une solution admissible. Bien souvent, la réponse à cette question centrale ne peut pas se trouver en appliquant un test statistique standard (basé sur le théorème central-limite) car la taille des échantillons (n et m) est trop petite. En se tournant vers des tests non paramétriques (c.f. [1]), il pourra trouver une variante du test du signe, connue sous le nom de test de Mc Nemar, mais qu'il ne pourra appliquer qu'à des données appariées (i.e. un problème doit être résolu par les deux méthodes).

En optimisation combinatoire, il est cependant fréquent de considérer des jeux de problèmes relativement petits et des tableaux de résultats incomplets (une méthode ou l'autre n'a pas été testée sur certains problèmes). Toutefois, le lecteur sera parfaitement convaincu — à juste titre d'ailleurs — qu'une méthode résolvant les 10 exemples d'un jeu de problèmes est meilleure qu'une méthode ne réussissant à en résoudre aucun. Mais qu'en est-il si le jeu ne comporte que 3 exemples ? On peut montrer qu'un taux de succès de 3/3 est supérieur à un taux de 0/3, avec un seuil de confiance supérieur à 98%. Un des buts de cet article est de le montrer.

Soit p_a (respectivement p_b) la probabilité de succès de la méthode A (respectivement, de la méthode B) et n la taille de l'échantillon (nombre d'exécutions) pour la méthode A (respectivement, m pour la méthode B). Pour réaliser un test statistique, on fera l'hypothèse nulle que $p_a = p_b = p$. La valeur de p restant inconnue. Sous cette hypothèse, la probabilité $S(p, a, n, b, m)$ d'observer a succès ou plus pour la méthode A sur n exécutions et b succès ou moins pour la méthode B sur m exécutions est donnée par le produit de deux distributions binomiales :

$$S(p, a, n, b, m) = (\sum_{i=a}^{n} \frac{n!}{i! \cdot (n-i)!} \cdot p^i \cdot (1-p)^{n-i}) \cdot (\sum_{j=0}^{b} \frac{m!}{j! \cdot (m-j)!} \cdot p^j \cdot (1-p)^{m-j})$$

Si $\max_{0<p<1} S(p, a, n, b, m) \leq 1-\alpha$, on peut rejeter l'hypothèse nulle avec un niveau de confiance α et accepter l'hypothèse alternative $p_a > p_b$ (i.e. un taux de succès de a/n est plus grand qu'un taux de succès de b/m).

Malheureusement, la valeur de p qui maximise $S(p, a, n, b, m)$ est en général pour le moins difficile à exprimer analytiquement. C'est pourquoi le tableau 1 donne, pour différentes valeurs de n et m et pour un niveau de confiance de 95% les couples (a, b) pour lesquels un taux de succès de a/n est plus élevé qu'un taux de b/m.

Le calcul des niveaux de confiance peut se réaliser *en ligne* à l'URL :

http ://ina.eivd.ch/projects/stamp/.

Pour de petits échantillons, le nouveau test proposé est plus puissant que le test de Mc Nemar, lorsque ce dernier peut s'appliquer : la probabilité de rejet de l'hypothèse nulle sera plus élevée avec notre test qu'avec le test de Mc Nemar [2].

m \ n	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2		(3,0)	(4,0)	(5,0)	(5,0)	(6,0)	(7,0)	(7,0)	(8,0)	(9,0)	(9,0)	(10,0)	(11,0)	(11,0)
													(14,1)	(15,1)
3	(2,0)	(3,0)	(3,0)	(4,0)	(4,0)	(5,0)	(5,0)	(6,0)	(6,0)	(7,0)	(7,0)	(8,0)	(9,0)	(9,0)
				(5,1)	(6,1)	(7,1)	(8,1)	(8,1)	(9,1)	(10,1)	(11,1)	(12,1)	(12,1)	(13,1)
4	(2,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)	(6,0)	(6,0)	(7,0)	(7,0)	(8,0)
		(3,1)	(4,1)	(5,1)	(5,1)	(6,1)	(7,1)	(7,1)	(8,1)	(9,1)	(9,1)	(10,1)	(11,1)	(11,1)
					(6,2)	(7,2)	(8,2)	(9,2)	(10,2)	(11,2)	(12,2)	(12,2)	(13,2)	(14,2)
5	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)	(6,0)	(6,0)	(7,0)
		(3,1)	(4,1)	(4,1)	(5,1)	(5,1)	(6,1)	(6,1)	(7,1)	(8,1)	(8,1)	(9,1)	(9,1)	(10,1)
			(4,2)	(5,2)	(6,2)	(7,2)	(7,2)	(8,2)	(9,2)	(10,2)	(10,2)	(11,2)	(12,2)	(12,2)
							(8,3)	(9,3)	(10,3)	(11,3)	(12,3)	(13,3)	(14,3)	(15,3)
6	(2,0)	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)	(5,0)	(6,0)
	(2,1)	(3,1)	(3,1)	(4,1)	(4,1)	(5,1)	(5,1)	(6,1)	(6,1)	(7,1)	(7,1)	(8,1)	(8,1)	(9,1)
		(3,2)	(4,2)	(5,2)	(5,2)	(6,2)	(7,2)	(7,2)	(8,2)	(9,2)	(9,2)	(10,2)	(11,2)	(11,2)
				(5,3)	(6,3)	(7,3)	(8,3)	(9,3)	(9,3)	(10,3)	(11,3)	(12,3)	(12,3)	(13,3)
									(10,4)	(11,4)	(12,4)	(13,4)	(14,4)	(15,4)
7	(2,0)	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(4,0)	(5,0)	(5,0)	(5,0)
	(2,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(6,1)	(6,1)	(7,1)	(7,1)	(7,1)	(8,1)
		(3,2)	(4,2)	(4,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(8,2)	(8,2)	(9,2)	(10,2)	(10,2)
			(4,3)	(5,3)	(6,3)	(6,3)	(7,3)	(8,3)	(9,3)	(9,3)	(10,3)	(11,3)	(11,3)	(12,3)
					(6,4)	(7,4)	(8,4)	(9,4)	(10,4)	(10,4)	(11,4)	(12,4)	(13,4)	(14,4)
										(11,5)	(12,5)	(13,5)	(14,5)	(15,5)
8	(2,0)	(2,0)	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(4,0)	(5,0)	(5,0)
	(2,1)	(3,1)	(3,1)	(3,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(6,1)	(6,1)	(6,1)	(7,1)	(7,1)
		(3,2)	(4,2)	(4,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(8,2)	(8,2)	(9,2)	(9,2)
		(3,3)	(4,3)	(5,3)	(5,3)	(6,3)	(7,3)	(7,3)	(8,3)	(9,3)	(9,3)	(10,3)	(10,3)	(11,3)
				(5,4)	(6,4)	(7,4)	(8,4)	(8,4)	(9,4)	(10,4)	(10,4)	(11,4)	(12,4)	(13,4)
						(7,5)	(8,5)	(9,5)	(10,5)	(11,5)	(12,5)	(12,5)	(13,5)	(14,5)
												(13,6)	(14,6)	(15,6)
9	(2,0)	(2,0)	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(4,0)	(4,0)
	(2,1)	(2,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(6,1)	(6,1)	(6,1)	(7,1)
	(2,2)	(3,2)	(3,2)	(4,2)	(4,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(8,2)	(8,2)	(9,2)
		(3,3)	(4,3)	(4,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)	(8,3)	(9,3)	(9,3)	(10,3)	(10,3)
			(4,4)	(5,4)	(6,4)	(6,4)	(7,4)	(8,4)	(8,4)	(9,4)	(10,4)	(10,4)	(11,4)	(12,4)
				(5,5)	(6,5)	(7,5)	(8,5)	(9,5)	(9,5)	(10,5)	(11,5)	(12,5)	(12,5)	(13,5)
							(8,6)	(9,6)	(10,6)	(11,6)	(12,6)	(13,6)	(14,6)	(14,6)
														(15,7)
10	(2,0)	(2,0)	(2,0)	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)	(4,0)
	(2,1)	(2,1)	(3,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(6,1)	(6,1)	(6,1)
	(2,2)	(3,2)	(3,2)	(4,2)	(4,2)	(5,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(8,2)	(8,2)
		(3,3)	(4,3)	(4,3)	(5,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)	(8,3)	(9,3)	(9,3)	(10,3)
		(3,4)	(4,4)	(5,4)	(5,4)	(6,4)	(7,4)	(7,4)	(8,4)	(9,4)	(9,4)	(10,4)	(10,4)	(11,4)
			(4,5)	(5,5)	(6,5)	(7,5)	(7,5)	(8,5)	(9,5)	(9,5)	(10,5)	(11,5)	(12,5)	(12,5)
					(6,6)	(7,6)	(8,6)	(9,6)	(10,6)	(10,6)	(11,6)	(12,6)	(13,6)	(14,6)
								(9,7)	(10,7)	(11,7)	(12,7)	(13,7)	(14,7)	(15,7)
11	(2,0)	(2,0)	(2,0)	(2,0)	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)	(4,0)
	(2,1)	(2,1)	(3,1)	(3,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(6,1)	(6,1)
	(2,2)	(3,2)	(3,2)	(4,2)	(4,2)	(4,2)	(5,2)	(5,2)	(5,2)	(6,2)	(6,2)	(7,2)	(7,2)	(7,2)
		(3,3)	(4,3)	(4,3)	(5,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)	(7,3)	(8,3)	(8,3)	(9,3)
		(3,4)	(4,4)	(5,4)	(5,4)	(6,4)	(6,4)	(7,4)	(7,4)	(8,4)	(9,4)	(9,4)	(10,4)	(10,4)
			(4,5)	(5,5)	(6,5)	(6,5)	(7,5)	(8,5)	(8,5)	(9,5)	(10,5)	(10,5)	(11,5)	(12,5)
				(5,6)	(6,6)	(7,6)	(8,6)	(8,6)	(9,6)	(10,6)	(11,6)	(11,6)	(12,6)	(13,6)
						(7,7)	(8,7)	(9,7)	(10,7)	(11,7)	(11,7)	(12,7)	(13,7)	(14,7)
									(10,8)	(11,8)	(12,8)	(13,8)	(14,8)	(15,8)
12	(2,0)	(2,0)	(2,0)	(2,0)	(2,0)	(2,0)	(3,0)	(3,0)	(3,0)	(3,0)	(3,0)	(3,0)	(4,0)	(4,0)
	(2,1)	(2,1)	(3,1)	(3,1)	(3,1)	(3,1)	(4,1)	(4,1)	(4,1)	(4,1)	(5,1)	(5,1)	(5,1)	(5,1)
	(2,2)	(3,2)	(3,2)	(3,2)	(4,2)	(4,2)	(4,2)	(5,2)	(5,2)	(6,2)	(6,2)	(6,2)	(7,2)	(7,2)
	(2,3)	(3,3)	(3,3)	(4,3)	(4,3)	(5,3)	(5,3)	(6,3)	(6,3)	(7,3)	(7,3)	(8,3)	(8,3)	(8,3)
		(3,4)	(4,4)	(4,4)	(5,4)	(5,4)	(6,4)	(7,4)	(7,4)	(8,4)	(8,4)	(9,4)	(9,4)	(10,4)
		(3,5)	(4,5)	(5,5)	(5,5)	(6,5)	(7,5)	(7,5)	(8,5)	(8,5)	(9,5)	(10,5)	(10,5)	(11,5)
			(4,6)	(5,6)	(6,6)	(7,6)	(7,6)	(8,6)	(9,6)	(9,6)	(10,6)	(11,6)	(11,6)	(12,6)
				(5,7)	(6,7)	(7,7)	(8,7)	(9,7)	(9,7)	(10,7)	(11,7)	(12,7)	(12,7)	(13,7)
						(7,8)	(8,8)	(9,8)	(10,8)	(11,8)	(12,8)	(12,8)	(13,8)	(14,8)
										(11,9)	(12,9)	(13,9)	(14,9)	(15,9)

TAB. 1 – Couples (a, b) pour lesquels un taux de succès $\geq a/n$ est significativement plus élevé qu'un taux de b/m, pour un niveau de confiance de 95%.

Remerciements Le présent travail a été en partie financé par la réserve stratégique de la Haute École Spécialisée de Suisse Occidentale (HES-SO), projet LQF01-03.

RÉFÉRENCES

[1] W. J. Conover, *Practical Nonparametric Statistics*, Wiley, troisième édition, 1999.

[2] É. D. Taillard, «A Statistical Test for Comparing Success Rates», actes de : *Metaheuristic international conference MIC'03*, Kyoto, Japon, août 2003.