A Non-Parametric Statistical Test for Comparing Success Rates

Éric Taillard, Ph. Wälti, J. Zuber

EIVD

University of Applied Sciences of Western Switzerland,

Yverdon-les-bains, Switzerland

TYPICAL NUMERICAL RESULTS :

From Kim[3], JoH 9 (3), June 2003

Problem instances	Number or runs	TCC	RSC	FSC	SSC	SHC	TMC
Sorting network design, n = 7	10	7	5	**8**	**8**	**8**	5
Sorting network design, n = 7	10	3	2	3	4	3	2
Sorting network design, n = 13	10	0	0	0	0	0	0
2DTTTgame	10	6	8	4	**9**	6	6
Nim(3,4,5,4)	10	3	2	**6**	**6**	4	3
Nim(5,7,11,6)	10	0	0	**1**	**1**	0	0

Conditions of the experiments :

A number of runs of non-deterministic methods on the same instance, or

Deterministic (or non-deterministic) methods, randomly generated instances (same difficulty)

Typical question :

Is SSC significantly better than FSC for 2DTTT game ?

i. e. is a 9/10 rate of success significantly better than a 4/10 rate ?

STANDARD TEST FOR BINARY COMPARISON

Conditions of the experiments :

n runs of method A, a successes, (unknown) probability of success of A : p_a (= $1 - q_a$)

m runs of method B, b successes, (unknown) probability of success of B : p_b (= $1 - q_b$)

X_A (resp. X_B) be the random variable associated with successful run of A (resp. B)

If *n* and *m* > 14 then :

$U = X_A/n - X_B/m$ is approximately normally distributed

Average : $p_a - p_b$

Variance $p_a q_a/n + p_b q_b/m$.

STANDARD PARAMETRIC TEST

Null hypothesis :

$$p_a = p_b = p$$

Alternate hypothesis :

$$p_a > p_b$$

Parameter p can be approximated by the pooled estimate :

$$\hat{p} = 1 - \hat{q} = \frac{a+b}{n+m}$$

Compute :

$$\hat{u} = \frac{a/n - b/m}{\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{\hat{p}\hat{q}}{m}}}$$

Reject the null hypothesis with confidence level α if :

$\Phi(\hat{u}) > 1 - \alpha$ (Φ : cumulative normal distribution)

($\hat{u} > 1.65$ for $\alpha = 95\%$ and $\hat{u} > 2.33$ for $\alpha = 99\%$)

NON PARAMETRIC TEST

Problem : Compare proportions p_a and p_b of Yes answers in two samples

Samples : (n runs of method A, a Yes), (m runs of method B, b Yes), $a/n > b/m$

Null hypothesis : $p_a = p_b = p$

Alternate hypothesis : $p_a > p_b$ (unilateral test)

Method :

Suppose that the Yes answer has the same unknown probability p to appear for both methods A and B

The probability $S(p, a, n, b, m)$ to observe :

a Yes answers or more for method A

b Yes answers or less for method B

is given by $S(p, a, n, b, m) = \sum_{i=a}^{n} \sum_{j=0}^{b} \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i} \cdot \binom{m}{j} \cdot p^j \cdot (1-p)^{n_b - j}$

Reject null hypothesis at confidence level α if : $\max_{0<p<1} S(p, a, n, b, m) < 1-\alpha$

Note : Since $\max_{0<p<1}$ is taken, the statistical test does not depend on a parameter.

EXAMPLES

1) All *n* runs of *A* were successes, none of the *m* runs of *B* were successes.

$S(p, n, n, 0, m) = 1 \cdot p^n \cdot (1-p)^0 \cdot 1 \cdot p^0 \cdot (1-p)^m = p^n(1-p)^m$

Maximum of S over p : $dS/dp = 0 = n\,p^{n-1}(1-p)^m - mp^n(1-p)^{m-1} \Rightarrow p = \frac{n}{n+m} = \frac{a+b}{n+m}$

For $n = 3$, $m = 2$, max $S(p, n, n, 0, m) = 108/3125 < 5\%$

It can be assumed, at 95% confidence level, that a success rate of 3/3 is significantly higher than 0/2.

2) 3 successes of *A* over 4 runs, 0 success of *B* over 3 runs

a) Parametric test : pooled estimate $\hat{p} = \frac{3}{7}$, S(3/7, 3, 4, 0, 3) < 4%

At 96% confidence level it can be assumed that *A* is better than *B*

b) Non parametric test : $S\left(\frac{6-2\sqrt{2}}{7}, 3, 4, 0, 3\right) > 4\%$

At 96% confidence level it cannot be assumed that *A* is better than *B*

NON-PARAMETRIC BILATERAL TEST

Problem : Compare proportions p_a and p_b of Yes answers in two samples

Samples : (n runs of method A, a Yes), (m runs of method B, b Yes)

Null hypothesis : $p_a = p_b = p$

Alternate hypothesis : $p_a \neq p_b$ (bilateral test)

Direct computation :

Probability P to observe a/n and b/n under null hypothesis :

$$P = \frac{\binom{n}{a} \cdot p^a \cdot (1-p)^{n-a} \cdot \binom{m}{b} \cdot p^b \cdot (1-p)^{m-b}}{\binom{n+m}{a+b} \cdot p^{a+b} \cdot (1-p)^{n+m-a-b}} = \frac{\binom{n}{a} \cdot \binom{m}{b}}{\binom{n+m}{a+b}}$$

Reject the null hypothesis with confidence level α if $P > 1 - \alpha$

McNemar non-parametric Test

Another name of the sign test in case of nominal data.

Comparison of bivariate (paired) data :

Methods A and B are run on the **same** set of instances.

Ties are discarded.

a successes for method A, b (complementary) successes for method B on $a + b$ instances

Null Hypothesis : $P(X_a = 0, X_b = 1) = \mathrm{P}(X_a = 1, X_b = 0) = 0.5$

Alternate hypothesis : $P(X_a = 0, X_b = 1) \neq \mathrm{P}(X_a = 1, X_b = 0)$ (Bilateral test)

Direct computation :

Probability P to observe b successes over $a + b$ instances under null hypothesis :

$$P = \binom{a+b}{b} \cdot 0.5^b \cdot (1 - 0.5)^{a+b-b} = \binom{a+b}{b} \cdot 0.5^{a+b}$$

Reject the null hypothesis with confidence level α if $2P > 1 - \alpha$

TABLES FOR $N = M = 6$

Confidence level = 95%

		McNemar (uni or bilateral)							New unilateral							Bilateral						
		b							*b*							*b*						
		0	1	2	3	4	5	6	0	1	2	3	4	5	6	0	1	2	3	4	5	6
	0							B				B	B	B	B					B	B	B
	1												B	B	B						B	B
	2													B	B							B
a	3								A						B							
	4								A	A						A						
	5								A	A	A					A	A					
	6	A							A	A	A	A				A	A	A				

Confidence level = 99%

		McNemar							New unilateral							Bilateral						
		b							*b*							*b*						
		0	1	2	3	4	5	6	0	1	2	3	4	5	6	0	1	2	3	4	5	6
	0												B	B	B						B	B
	1														B							B
	2														B							
a	3																					
	4								A													
	5								A							A						
	6								A	A	A					A	A					

WEB SITE

HTTP://INA.EIVD.CH/PROJECTS/STAMP

On-line computation of the p-value

C code available

CONCLUSIONS

A new powerful, non-parametric statistical test has been developed

Allows to compare success rates of two methods when

The methods are non-deterministic (one problem instance or more with same difficulty)

The methods are deterministic, randomly generated problem instances (with same difficulty)

Can be applied to very small samples, not necessarily of the same size

Allows to reduce computational effort for parameter tuning

Available on-line.

Compares favourably with McNemar non-parametric test