Athens Institute for Education and Research
ATINER

ATINER's Conference Paper Series
MAT2013-0753

Finding the Ratio of Two Percentiles by Way of Non Parametric Methods

Li-Fei Huang
Assistant Professor
Department of Applied Statistics and Information Science
Ming Chuan University
Taiwan

Athens Institute for Education and Research
8 Valaoritou Street, Kolonaki, 10671 Athens, Greece
Tel: + 30 210 3634210 Fax: + 30 210 3634209
Email: info@atiner.gr URL: www.atiner.gr
URL Conference Papers Series: www.atiner.gr/papers.htm

Printed in Athens, Greece by the Athens Institute for Education and Research.

ISSN **2241-2891**
9/12/2013

An Introduction to ATINER's Conference Paper Series

ATINER started to publish this conference papers series in 2012. It includes only the papers submitted for publication after they were presented at one of the conferences organized by our Institute every year. The papers published in the series have not been refereed and are published as they were submitted by the author. The series serves two purposes. First, we want to disseminate the information as fast as possible. Second, by doing so, the authors can receive comments useful to revise their papers before they are considered for publication in one of ATINER's books, following our standard procedures of a blind review.

Dr. Gregory T. Papanikos
President
Athens Institute for Education and Research

This paper should be cited as follows:

Huang, L-F. (2013) **"Finding the Ratio of Two Percentiles by Way of Non Parametric Methods"** Athens: ATINER'S Conference Paper Series, No: **MAT2013-0753**.

Finding the Ratio of Two Percentiles by Way of Non Parametric Methods

Li-Fei Huang
Assistant Professor
Department of Applied Statistics and Information Science
Ming Chuan University
Taiwan

Abstract

In the wood industry, it is common practice to compare in terms of the ratio of two different strength properties for lumber of the same dimention, grade and species or the same strength property for lumber of two different dimensions, grades or species. Because United States lumber standards are given in terms of population fifth percentile, and strength problems arised from the weaker fifth percentile rather than the stronger mean, the ratio should be expressed in terms of the fifth percentiles of two strength distributions rather than the mean.

Exact confidence regions for the ratio of percentiles for two independent normal distributions when the ratio of variances is known are obtained. The confidence region can be a bounded interval, the complement of an interval, or the whole real line. When large samples are available, confidence intervals for the ratio of percentiles are also obtained even when the ratio of variances is unknown. The confidence region is always a bounded interval, but it shows poor coverage rates when the percentile in the denominator is near zero.

When sample sizes are large, non parametric approaches are possible. If percentiles are estimated by order statistics, the resulting confidence region is always a bounded interval with poor coverage rates. This paper will assume small samples to derive new non parametric method which is similar to the Wilcoxon rank-sum test, find ratio of percentiles in original measurements and in ranks, and compute confidence regions which should be intervals and hopefully can show good coverage rates.

Keywords: Strength of lumber, Ratio of percentiles, Non parametric methods

Acknowledgements: The author wishes to thank James Evans, United States Forest Products Laboratory, for providing the data for the example.

Corresponding Author:

1. Introduction

Alougth the purpose is to compare two different strength properties, at first only one sample is considered and made inferences. The 100 p th percentile of the population is

$$\xi_p = \inf\{x : F(x) \geq p\}.$$

2. Sign Test for the Percentile

Define a new random variable D to be the result of original random variable X subtracting the conjectural 100 p th percentile. Then D is negative with probability p and positive with probability 1- p. Let P- be the prabiblity that D is negative, then test hyphotheses become

$$H_0: P- = p,$$

$$H_1: P- \neq p.$$

Let T_- be the number of observations of negative D. If the null hypothesis is true, T_- follows a binomial distribution with parameters n and p. The null hypothesis is rejected when

$$\text{p-value}= 2\sum_{k=0}^{T_-}\binom{n}{k}p^k(1-p)^{n-k} < \alpha .$$

To construct an exact confidence interval for the 100 p th percentile, find two orders L and U such that

$$L = \sup\left\{L : \sum_{k=0}^{L-1}\binom{n}{k}p^k(1-p)^{n-k} \leq \frac{\alpha}{2}\right\}, \quad U = \inf\left\{U : \sum_{k=U+1}^{n}\binom{n}{k}p^k(1-p)^{n-k} \leq \frac{\alpha}{2}\right\}.$$

Finally the exact 100(1-α)% confidence interval calculated by Thompson-Savur method for the 100 p th percentile ξ_p is $\left(X_{(L)},\quad X_{(U)}\right)$.

If a large sample is available, T_- can be approximated by a normal distribution with $E(T_-) = np$ and $Var(T_-) = np(1-p)$. Now the null hypothesis is rejected when

$$Z = \frac{|T_- - np| - 0.5}{\sqrt{np(1-p)}} > Z_{\frac{\alpha}{2}} .$$

To construct an approximated confidence interval for the 100 p th percentile, find two orders L and U such that

$$L = \sup\left\{ L : L \le np - 0.5 - Z_{\frac{\alpha}{2}} \sqrt{np(1-p)}, \quad L \in \mathrm{Z} \right\},$$

$$U = \inf\left\{ U : U \ge np + 0.5 + Z_{\frac{\alpha}{2}} \sqrt{np(1-p)}, \quad U \in \mathrm{Z} \right\}.$$

The approximated 100(1-α)% confidence interval for the 100 p th percentile ξ_p is $\left(X_{(L)}, \quad X_{(U)} \right)$.

3. Wilcoxon signed Rank Test for the Percentile

Only the sign of the new random variable D is considered in the sign test. Hence Wilcoxon signed rank test is created to take care of the size of the new random variable D.

Ranks 1, 2, ... n are given to the values of |D| in ascending order. Define

$$R_i = \begin{cases} i & w.p. \quad 1-p \\ 0 & w.p. \quad p-0 \end{cases}, \text{ then}$$

$E(R_i) = i(1-p)$,

$E(R_i^2) = i^2(1-p)$,

$Var(R_i) = E(R_i^2) - [E(R_i)]^2 = i^2(1-p) - i^2(1-p)^2 = i^2(1-p)(1-1+p) = i^2 p(1-p)$
,

$T_+ = \sum_{i=1}^{n} R_i = \frac{n(n+1)}{2} - T_-$,

and the null hypothesis is rejected if the corresponding p-value of T_- in Table 1 < α.

Table 1 is the new Wilcoxon signed rank test table adjusted for percentiles.

To construct an confidence interval for the 100 p th percentile, the empirical 100 p th percentiles of $\binom{n}{2} + n = \frac{n(n+1)}{2}$ pairs of observations need to be found. If X_S is the smaller value and X_L is the larger value of a pair of observations, then the empirical 100 p th percentile is $E = (1-p)X_S + pX_L$.

Two orders L and U are created in the following way to construct an exact confidence interval:

$$L = \sup\left\{ L : P(T_- \le L) \le \frac{\alpha}{2} \right\}, \quad U = \inf\left\{ U : P(T_- \ge U) \le \frac{\alpha}{2} \right\}.$$

Table 1. *Wilcoxon signed rank test table adjusted for percentiles when n=12*

10%		50%					
T-	p-value	T-	p-value	T-	p-value	T-	p-value
0	0.28243	0	0.00024	23	0.11670	46	0.71533
1	0.31381	1	0.00049	24	0.13306	47	0.74072
2	0.34519	2	0.00073	25	0.15063	48	0.76514
3	0.38006	3	0.00122	26	0.16968	49	0.78809
4	0.41493	4	0.00171	27	0.19019	50	0.80981
5	0.45328	5	0.00244	28	0.21191	51	0.83032
6	0.49202	6	0.00342	29	0.23486	52	0.84937
7	0.53425	7	0.00464	30	0.25928	53	0.86694
8	0.57687	8	0.00610	31	0.28467	54	0.88330
9	0.62336	9	0.00806	32	0.31104	55	0.89819
10	0.67028	10	0.01050	33	0.33862	56	0.91187
11	0.72108	11	0.01343	34	0.36670	57	0.92432
12	0.77269	12	0.01709	35	0.39551	58	0.93530
13	0.79684	13	0.02124	36	0.42505	59	0.94507
14	0.81836	14	0.02612	37	0.45483	60	0.95386
15	0.84071	15	0.03198	38	0.48486	61	0.96143
16	0.86008	16	0.03857	39	0.51514	62	0.96802
17	0.87994	17	0.04614	40	0.54517	63	0.97388
18	0.89687	18	0.05493	41	0.57495	64	0.97876
19	0.91390	19	0.06470	42	0.60449		
20	0.92762	20	0.07568	43	0.63330		
21	0.94144	21	0.08813	44	0.66138		
22	0.95158	22	0.10181	45	0.68896		
23	0.96139						
24	0.96747						
25	0.97281						
26	0.97748						

Finally the exact 100(1-α)% confidence interval calculated by Hodges-Lehmann method for the 100 p th percentile ξ_p is $\left(E_{(L)},\quad E_{(U)}\right)$.

If a large sample is available, then

$$E(T_+) = E\left(\sum_{i=1}^{n} R_i\right) = \sum_{i=1}^{n} i(1-p) = \frac{n(n+1)(1-p)}{2},$$

$$E(T_-) = \frac{n(n+1)}{2} - E(T_+) = \frac{n(n+1)p}{2},$$

$$Var(T_-) = p^2 \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)p^2}{6}.$$

Now the null hypothesis is rejected when

$$Z = \frac{\left| T_- - \frac{n(n+1)p}{2} \right| - 0.5}{\sqrt{\frac{n(n+1)(2n+1)p^2}{6} - \frac{\sum (t_i^3 - t_i)}{48}}} > Z_{\frac{\alpha}{2}} .$$

To construct an approximated confidence interval for the $100\,p$ th percentile, find two orders L and U such that

$$L = \sup\left\{ L : L \le \frac{n(n+1)p}{2} - 0.5 - Z_{\frac{\alpha}{2}} \sqrt{\frac{n(n+1)(2n+1)p^2}{6} - \frac{\sum (t_i^3 - t_i)}{48}}, \quad L \in Z \right\},$$

$$U = \inf\left\{ U : U \ge \frac{n(n+1)p}{2} + 0.5 + Z_{\frac{\alpha}{2}} \sqrt{\frac{n(n+1)(2n+1)p^2}{6} - \frac{\sum (t_i^3 - t_i)}{48}}, \quad U \in Z \right\}$$

.

The approximated 100(1-α)% confidence interval for the $100\,p$ th percentile ξ_p is $\left(E_{(L)}, \quad E_{(U)}\right)$.

4. Wilcoxon Rank-sum Test for Ratio of Percentiles

Suppose that 100(1-α)% confidence intervals for the $100\,p$ th percentile ξ_{1p} of the numerator population and the $100\,p$ th percentile ξ_{2p} of the denominator population are $\left(X_{(L1)}, \quad X_{(U1)}\right)$ and $\left(Y_{(L2)}, \quad Y_{(U2)}\right)$, respectively. Then the 100(1-α)% confidence interval for ξ_{1p} / ξ_{2p} can be easily calculated by

$$\left(\frac{X_{(L1)}}{Y_{(U2)}}, \frac{X_{(U1)}}{Y_{(L2)}} \right).$$

Wilcoxon rank-sum test becomes very complicate for ratio of percentiles. Let us focus only on the steps of constructing the exact confidence interval.

At first the logrithms of $N_1 = n_1(n_1+1)/2$ empirical $100\,p$ th percentiles of the numerator sample and $N_2 = n_2(n_2+1)/2$ empirical $100\,p$ th percentiles of the denominator sample are taken. Then $N_1 - N_2$ differences are computed.

Given N_1, N_2 and α, an order r will be computed. If the (r+1)th small difference is L and the (r+1)th large difference is U, then the 100(1-α)% confidence interval for ξ_{1p} / ξ_{2p} is $\left(e^L, \quad e^U\right)$.

5. Results

Two small examples from Lehmann (1975) and their 95% confidence intervals for percentiles are given in Table 2. Since the sample size is only 12, $P(T_- \leq 0) = 0.28$ becomes too large to create confidence intervals for the 10th percentile. When confidence intervals for the 50th percentile are compared, (approximated) Hodges-Lehmann method can create narrower confidence intervals because $n(n+1)/2$ rather than just n values are utilized.

Table 2. *Two Small Examples and their 95% Confidence Intervals for Percentiles*

Ex1	20.3	23.5	4.7	21.9	15.6	20.3	26.6	21.9	-9.4	4.7	-1.6	25.0
(i)	(6)	(10)	(3)	(8)	(5)	(7)	(12)	(9)	(1)	(4)	(2)	(11)
Method			10th percentile				1-α	50th percentile				1-α
Binomial test			N/A		N/A		N/A	$(X_{(3)},X_{(9)})$		(4.7, 21.9)		0.96
Approx. Bin. test			$(X_{(1)},X_{(3)})$		(-9.4,4.7)		0.97	$(X_{(3)},X_{(9)})$		(4.7, 21.9)		0.96
Hodges-Lehmann			N/A		N/A		N/A	$(E_{(14)},E_{(63)})$		(6.25,22.65)		0.95
Approx. H-L			N/A		N/A		N/A	$(E_{(15)},E_{(63)})$		(6.25,22.65)		0.95
Ex2	6.2	15.6	25	4.7	28.1	17.2	14.1	31.1	12.6	9.4	17.2	23.4
(i)	(2)	(6)	(10)	(1)	(11)	(7)	(5)	(12)	(4)	(3)	(8)	(9)
Method			10th percentile				1-α	50th percentile				1-α
Binomial test			N/A		N/A		N/A	$(X_{(3)},X_{(9)})$		(9.4, 23.4)		0.96
Approx. Bin. test			$(X_{(1)},X_{(3)})$		(4.7, 9.4)		0.97	$(X_{(3)},X_{(9)})$		(9.4, 23.4)		0.96
Hodges-Lehmann			N/A		N/A		N/A	$(E_{(14)},E_{(63)})$		(10.95,22.6)		0.95
Approx. H-L			N/A		N/A		N/A	$(E_{(15)},E_{(63)})$		(11, 22.6)		0.95

95% confidence intervals for ratios of percentiles of two small examples are given in Table 3. Hodges-Lehmann method can create narrower confidence intervals because $n(n+1)/2$ rather than just n values are utilized in both samples. The problem that the upper bound of the confidence interval of the ratio of 10th percentiles goes to infinity arised when the 10th percentile of the denominator population is close to zero. It is suggested that the population with a small percentile which is close to zero should always be the numerator population.

One large example from United States Forest Products Laboratory and its 95% confidence intervals for percentiles are given in Table 4. It’s hard to create the new Wilcoxon signed rank test table adjusted for percentiles for such a large sample, so only the results of approximated Hodges-Lehmann method are available.

Table 3. *95% confidence intervals for ratios of percentiles of two small examples*

Numerator	Method	Ratio of 10%ile	Ratio of 50%ile
Ex1	Binomial test	N/A	(0.201, 2.330)
	Approx. Bin. test	(-1, 1)	(0.201, 2.330)
	Hodges-Lehmann	N/A	(0.277, 2.068)
	Approx. H-L	N/A	(0.277, 2.059)
Ex2	Binomial test	N/A	(0.429,4.979)
	Approx. Bin. test	$(-\infty,-1)\cup(1,\infty)$	(0.429,4.979)
	Hodges-Lehmann	N/A	(0.483, 3.616)
	Approx. H-L	N/A	(0.486, 3.616)

Table 4. *One large example and its 95% confidence intervals for percentiles*

5418.6	4795.9	7061.8	6307.9	6964.0	6674.1	8153.4	6843.5
7011.3	5817.3	6617.3	6136.7	7529.2	6357.9	7643.6	7311.8
6997.6	4533.1	5691.8	6245.9	6455.7	6082.7	5511.1	5976.8
4607.6	4414.5	5268.4	8145.4	4616.1	3508.1	5231.6	5851.4
4281.8	9213.0	6051.4	4050.5	5677.2	5531.8	4872.4	3677.3
4230.9	2524.8	4896.3	4161.0	4818.4	5325.4	5818.8	4787.2
5988.6	5530.9	6351.2	2764.8	4432.9	5325.4	5651.6	3917.7
4429.6	3938.0	5143.3	4044.5	5128.6	5681.0	4690.8	5894.0
2352.0	2822.7	5465.8	3770.9	3168.1	4994.6	2327.6	2642.7
4760.3	3022.1	4187.7	4420.0	3095.2	5289.3	3440.7	4533.1
3340.9	3207.7	5270.6	4274.4	3854.5	2748.2	4461.2	4892.6
5078.2	5278.7	1420.3	3015.2	4451.7	1931.2	3556.0	3396.4
6767.2	2566.1	1228.4	5883.6	2444.5	3747.0	3879.0	4392.0
2105.5	5161.1	4658.4					

Method	5%	10%	50%
Binomial test	(1228.4, 2566.1)	(2327.6, 3095.2)	(4451.7, 5231.6)
Approx. Bin. test	(1228.4, 2642.7)	(2327.6, 3168.1)	(4451.7, 5268.4)
Approx. H-L	(1989.40, 2220.95)	(2499.89, 2713.94)	(4536.2, 5140.5)

6. Future Studies

Simulation will be made to show better coverage rates than those non parametric approaches of Huang & Johnson (2006).

Bibliography

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, Inc.

Huang, Li-Fei & Johnson, R. A. (2006). Confidence regions for the ratio of percentiles. *Statistics & Probability Letters* 76: 384-392.