Bayes Interval Estimation for binomial proportion and difference of two binomial proportions with Simulation Study

Masoud Ganji[1] , Solmaz Aghlmandi [2]

[1]Department of Statistics, Faculty Mathematical Science, University of Mohaghegh Ardabili

Ardabil, Iran

[2]Department of Statistics, Faculty Mathematical Science, University of Mohaghegh Ardabili

Ardabil, Iran

Abstract

This study is considered a number of popular confidence interval for binomial proportion and the difference of two binomial proportions. A new approach is proposed base on Bayesian view for binomial proportion and also for difference of two binomial proportions. The Bayes confidence intervals compared with other confidence intervals of coverage probability and expected length. Based on this analysisand t ,he simulation study recommend the Bayes confidence interval of binomial proportion and difference of two binomial proportions for small sample, and show their superior performance from both criteria.

Keywords: *Bayes interval, coverage probability, confidence interval, expected lengths.*

1. Introduction

In recent years the interval estimation of binomial proportion and difference of two binomial proportions has been reviewed. This is due to poor and irregular behavior of Wald confidence interval of coverage probability, it is mentioned by Agresti and Coull[1], Agresti and Caffo[2], Brown et al.[5] and Newcomb[7]. Standard confidence interval is used globally, since it's structure is simple, and it can be accounted as the representative of confidence interval for binomial proportion and difference of two binomial proportions. This confidence interval, is acquired of inverting Wald test, and is known to Wald confidence interval. Nevertheless, the erratic behavior of the coverage probability of confidence interval is mentioned in several papers. For example, Agresti and Caffo[2] shows the poor performance of Wald confidence interval, by using of drawing and numerical methods and by adding 2 successes and 2 failures to the sample size, they improved confidence interval for the binomial proportions which is show the performance of new confidence interval of coverage probability, which improved confidence interval is known as Agresti-Coull confidence interval. Agresti and Coull[1] by adding 1 success and 1 failure to the sample size, they improved confidence interval for the difference of two binomial proportions, which is show the superior performance of new confidence interval of coverage probability. Newcomb[7] by using lower and upper limits of score confidence interval, made the new confidence interval by combination of these limits for difference of two binomial proportions, and by comparison of this confidence interval with ten others famous confidence interval, then he showed it's good performance of coverage probability. There are much confidence interval (CI) such as Wilson CI, arcsine CI, logit CI, exact CI, likelihood ratio CI and Jefferys CI for binomial proportion. It should explained that, the Bayes estimator $\tilde{p} = \frac{x+0.5}{m+1}$ is obtained by using the Jeffery Beta form, with parameters (0.5,0.5) as a prior distribution, and the Jeffery approximate CI obtained by using the normal approximation. For such estimator, this CI, is recommended for small sample size (less than 40) (Brown et al.[5], [6]).

There for, we apply the $Beta(\alpha,\beta)$ prior distribution and normal approximation to make Bayes approximate CI. We know that, the prior distribution $Beta(\alpha,\beta)$ for binomial distribution is conjugate. If we assume that $X \sim binom(n,p)$, and p has the prior distribution $Beta(\alpha,\beta)$, the posterior distribution of p is $Beta(x+\alpha, m-x+\beta)$.

By using of the normal approximate, we get:

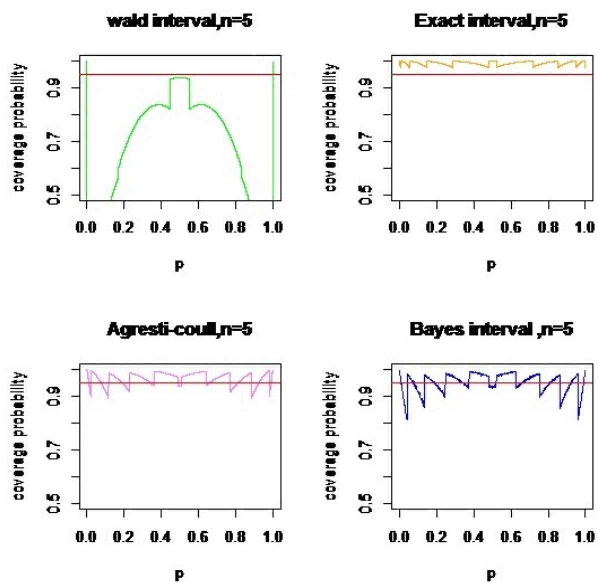
$$P \mid x \sim N\left(\tilde{p}, \frac{\tilde{p}(1-\tilde{p})}{(n+\alpha+\beta+1)}\right) \qquad (1)$$

There for by notice to the above approximation, we can proposed the Bayes approximate of CI and in section 2 and 3 we show that, for suitable amount of α and β in comparison with the other CIs, the Bayes CI involve the best performance in the coverage probability and expected length.

2. Confidence interval of binomial proportion and difference of two binomial proportions

2.1. Confidence interval of binomial proportion

Fig. 1 coverage probability of 95% confidence intervals for n=5



Wald CI (CI_w): If $x_1, x_2, ., x_n$ is a random sample from binomial distribution with unknown parameter P and known parameter n, the MLE of p is $\hat{p} = \frac{x}{n}$. As it is mentioned in most of text books CI_w of p is equal with:

$$CI_w = \hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p})/n} \qquad (2)$$

Where $z_{\alpha/2}$ denote the upper $\alpha/2$ quintile of the standard normal distribution.

Agresti-Coull CI (CI_{AC}): Agresti and Coull[1] improved CI_w, by adding two successes and two failures to the sample size, that is, they replaced n by $\tilde{n} = n+4$ and $\hat{p}$ with $\tilde{p} = \frac{x+2}{n+4}$, which is performance improved the CI_w. The CI_{AC} is as:

$$CI_{AC} = \tilde{p} \pm z_{\frac{\alpha}{2}} \sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}} \qquad (3)$$

Exact CI (CI_E): In contrast to CI_w, most advanced statistical text books recommend the

CI_E which presented by Clopper-Pears on (1934).

CI_E of p, obtained by inverting the binomial test of $H_0: p = p_0$ with equal-taild and by solving the following equations with respect to p_0, i.e:

$$\sum_{k=0}^{x} \binom{n}{k} p_0^{\,k} (1-p_0)^{n-k} = \frac{\alpha}{2}$$

and

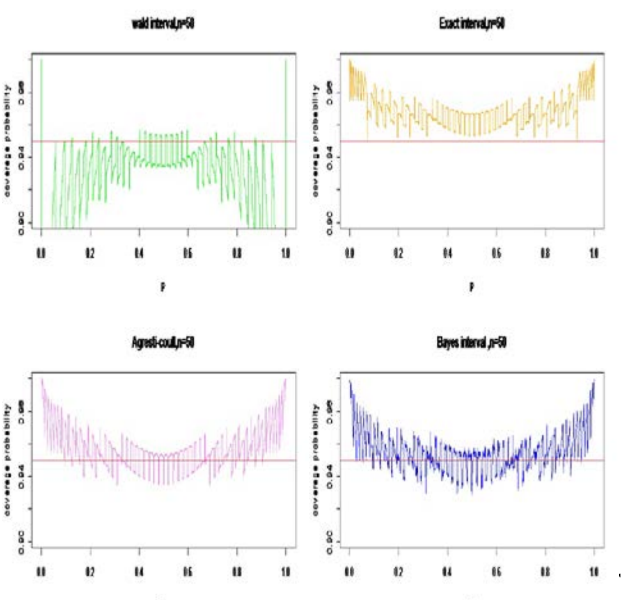
$$\sum_{k=x}^{n} \binom{n}{k} p_0^{\,k} (1-p_0)^{n-k} = \frac{\alpha}{2}$$

This estimated CI_E ensures that, coverage probability for all of the amount of p is at least $1-\alpha$, for x=1,2,...n-1. The CI_E is as follows:

$$CI_E = [1+\frac{n-x+1}{xF_{2x,2(n-x+1),1-\alpha/2}}]^{-1} < p < [1+\frac{n-x}{(x+1)F_{2(x+1),2(n-x),\alpha/2}}]^{-1} \quad (4)$$

Where $F_{a,b,c}$ is measure of F distribution with "a" and "b" degree of freedom and "1-c" quintile.

Fig. 2 coverage probability of 95% confidence intervals for n=50



distribution Beta(2, 2), the posterior of p is :

$$P \mid x \sim Beta(x+2, n-x+2)$$

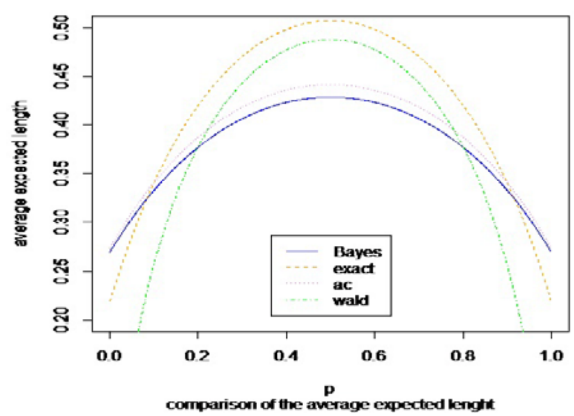
By using the normal approximate we get:

$$P \mid x \sim N(\tilde{p}, \frac{\tilde{p}(1-\tilde{p})}{(n+5)})$$

and by using $\tilde{p} = \frac{x+2}{n+4}$ as a estimation of p, CI_B can be obtained as follows:

$$CI_B = \tilde{p} \pm z_{\frac{\alpha}{2}} \sqrt{\tilde{p}(1-\tilde{p})/(n+5)} \quad (5)$$

Fig. 3 average expected lengths of 95% confidence intervals, n=15



2.2. Confidence interval of difference of two binomial proportions

The following notation used frequently in this article.

$X \sim binom(n, p_1)$, $y \sim binom(n, p_2)$, $\Delta = p_1 - p_2$, $q_i = 1 - p_i$, i=1, 2 and the MLE of p_1 and p_2 are $\hat{p}_1 = \frac{X}{n}$, $\hat{p}_2 = \frac{y}{m}$ respectively.

Wald CI (CI_W^*): The distribution of statistics $T = \frac{(\hat{\Delta}-\Delta)}{\sigma_{\hat{\Delta}}}$ approximately tends to the standard normal distribution, since $\sigma_{\hat{\Delta}}$ is the consistent estimator of $Var(\Delta) = p_1 q_1 / n + p_2 q_2 / m$, we can replace MLE of the parameters in above formula i.e. $Var(\hat{\Delta}) = \hat{p}_1\hat{q}_1 / n + \hat{p}_2\hat{q}_2 / m$, where $\hat{\Delta} = \hat{p}_1 - \hat{p}_2$ and get CI_W^* as;

$$CI_w^* = \hat{\Delta} \pm z_{\frac{\alpha}{2}} \sqrt{\hat{p}_1\hat{q}_1 / n + \hat{p}_2\hat{q}_2 / m} \quad (6)$$

Agresti CI (CI_{Ag}^*) : Agresti and Caffo[2] estimated both of populations parameters by adding one success and one failure for each sample and obtained $p_1' = \frac{X+1}{n+2}$ and $p_2' = \frac{Y+1}{m+2}$ as a estimation of p_1 and p_2 respectively. They gat CI_{Ag}^* for difference of binomial proportions as follows:

$$CI_{Ag}^* = p_1' - p_2' \pm z_{\frac{\alpha}{2}} \sqrt{p_1'q_1' / (n+2) + p_2'q_2' / (m+2)} \quad (7)$$

Newcomb CI (CI_{new}^*)**:** Newcomb[7] by using lower and upper limits of Wilson CI for p_1 and p_2 separately, made the new CI with combining of this limitation. (see Newcomb[7]). (l_i, u_i) are becomes from the solution of the quadratic equation $z_{\frac{\alpha}{2}} = \frac{(\hat{p}_i - p_i)}{\sqrt{p_i(1-p_i)/n_i}}$, with respective to p_i , i=1, 2, with $n_1 = n$ and $n_2 = m$, as follow:

$$\left(\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{l_1(1-l_1)}{n} + \frac{u_2(1-u_2)}{m}}, \right. \quad (8)$$
$$\left. \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{l_2(1-l_2)}{m} + \frac{u_1(1-u_1)}{n}}\right)$$

Bayes CI (CI_B^*)**:** To find CI_B for binomial proportion, we get prior distribution $Beta(\alpha, \beta)$ with $\alpha = \beta = 2$, which is acquires Bayes estimate of p as $\tilde{p} = \frac{x+2}{n+4}$. By regarding this fact and considering the independent prior distribution Beta(1,1), for each populations, the posterior distribution of p_1 and p_2 are as follows:

$$P_1 | x \sim Beta(x+1, n-x+1)$$

and

$$P_2 | y \sim Beta(y+1, m-y+1)$$

Now by using the normal approximation, we get:

$$P_1 | x \sim N\left(\tilde{p}_1, \frac{\tilde{p}_1(1-\tilde{p}_1)}{(n+3)}\right)$$

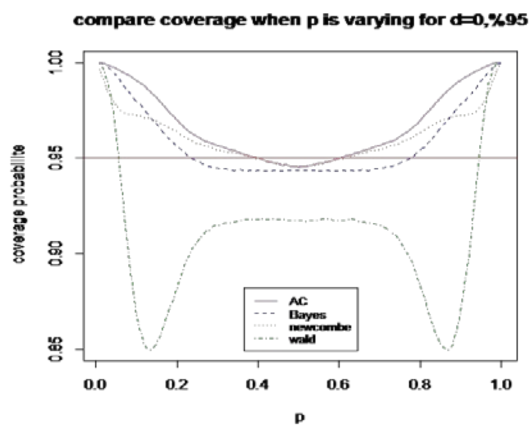
and

$$P_2 | y \sim N\left(\tilde{p}_2, \frac{\tilde{p}_2(1-\tilde{p}_2)}{(m+3)}\right)$$

Which implies that, $\tilde{p}_1 = \frac{x+1}{n+2}$ and $\tilde{p}_2 = \frac{y+1}{m+2}$. So CI_B^* of two binomial proportions is:

$$CI_B^* = \tilde{p}_1 - \tilde{p}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\tilde{p}_1\tilde{q}_1 / (n+3) + \tilde{p}_2\tilde{q}_2 / (m+3)} \quad (9)$$

Fig. 4 Compare coverage probability of 95% confidence intervals for fixed Δ=0 , variable p and n=m=20



3. comparison of confidence interval of binomial proportion

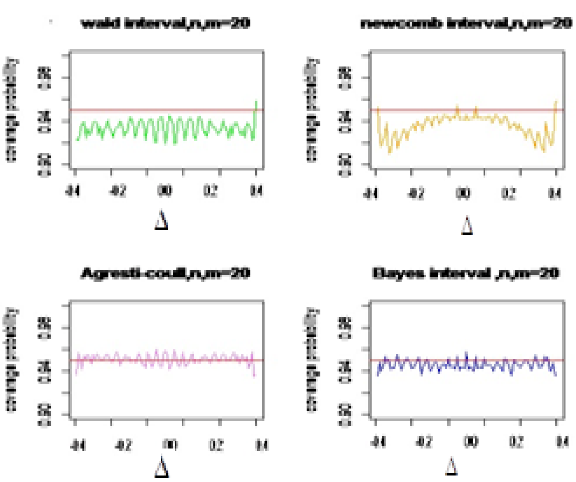
3.1. Coverage probability of confidence interval of binomial proportion

We will compare four CIs in term of coverage probability with this definition of performance, the CI which is closer to nominal level has preference than to other CIs, in other words, the lowest oscillation is noticeable. In this view, an approximate of CI of $(1-\alpha)$ of level should cover the right number of parameter approximately $100 \times (1-\alpha)$ times, so we can consider the coverage probability of CI as follows:

$$C(p)=\sum_{k=0}^{n}\binom{n}{k}I(k,p)p^{k}(1-p)^{n-k}$$

Where I(k, p) is equal to one, if CI includes p, otherwise is equal to zero. Figs. 1 and 2 indicate the coverage probability of four CIs for a sample size n=5 and n=50, respectively. It seems that, except CI_w ,the other CIs have acceptable performance in term of coverage probability. CI_{AC} and CI_B are closer in oscillation to nominal level, and CI_E is upper than nominal level. By notice to Table 1, we can see that, CI_B has the less distance of nominal level with respect to other CIs, such as, CI_E has the upper nominal level with respect to other CIs.

Fig. 5 Compare coverage probability of 95% confidence intervals for variable Δ , fixed p=0.5 and n=m=20



3.2. Expected length of confidence interval of binomial proportion

The length of CI is one of the most important performance criteria. In Fig. 3, four CIs from expected length are in comparison for n=15, and as it is seen, CI_B is shorter than CI_E and CI_{AC} and this is true even for CI_w for distance of 0.2 until 0.8. We see that, the CI_{AC} can never be shorter than CI_B .

According the Table 2, we see that, CI_w has the shortest length in comparison with other intervals in small sample size and it can be argued that, CI_B acts as short as CI_w and has much better performance than the other two CIs. But whenever, when the sample size is greater, the difference between CIs becomes less, in term of length criteria. This was expected for us, according to CIs structure.

4. The compare of confidence interval for difference of two binomial proportions

4.1. Coverage probability of confidence interval for difference of two binomial proportions

With respect to structure of CIs of two binomial proportion, which have four parameters as (m, n, p, Δ), we compared here, four CIs for different values of their parameters.

Fig. 4 is drown with assumption fixed Δ and variable p and n=m=20. Simply it is recognizable that CI^*_{new} and CI^*_B have the better performance than CI^*_{Ag} and especially performance than CI^*_W .

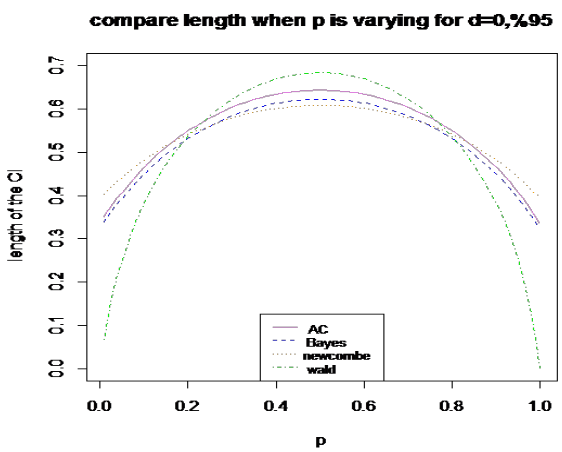
Fig. 5 is drown for p=0.5 and n=m=20 and for Δ as a variable, here it is recognizable that, CI^*_{Ag} has the least oscillation from nominal level, and CI^*_W never reach to nominal level and CI^*_{new} below nominal level is oscillation.

Table 3 indicates the average coverage probability of four CIs in different situation. According on this table, we can remark that CI^*_B has better performance in average coverage probability than the CI^*_{new} , CI^*_{Ag} and especially CI^*_W .

4.2. Expected length of confidence interval for difference of two binomial proportion

The length of CI is another important attribute in evaluation of a CI. As shown in Fig. 6 and Table 4, expected length of the CI^*_W is much shorter than CI^*_{Ag} and CI^*_{new} for smaller sample size. While CI^*_B performs as short as CI^*_W in term of expected length and also, for larger sample their performance is very similar from this viewpoint.

Fig. 6 Compare expected lengths of 95% confidence intervals for Δ=0, n=30 and m=10



5. Conclusion

In this study, we develop by simulation study a CI for binomial proportion and difference of two binomial proportions in the hope that, new CI is simple and useful for nearly all sample size. We reviewed the popular CI of binomial proportion and difference of two binomial proportions. By doing similar process we obtained other Bayes CI for binomial proportion and difference of two binomial proportions. Applying of numerical tables and drawing diagrams, we showed that, CI^*_B has the good performance in both of the coverage probability and expected length for small sample size. For a larger sample size (>40) except CI^*_W, the others CI have acceptable coverage probability such that, the CI^*_{Ag} is preferred for its simplicity in structure. In this paper we compared CIs of binomial proportion, and indicate the poor performance of CI^*_W in coverage probability. Also for fixed value of p and variable Δ, we recommended the CI^*_{Ag} and CI^*_B with respect to others. And when the value of Δ is fixed and parameter p is variable, we proved that, CI^*_{new} and CI^*_B have the better performance than the two other CIs, especially CI^*_W. At last, it is possible that we find out that the CI_B and CI^*_B for all sample sizes in different situation, has better performance from both comparative criteria.

References

[1]A. Agresti, B.A. Coull "Approximation is better than "exact" for interval estimation of binomial proportions", Amer. Statist. 52, 1998, 119–126.

[2] A. Agresti , B. Caffo "Simple effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures." Amer. Statist, 54, 2000, 280–288.

[3] E.B. Wilson, "Probable inference the law of succession and statistical inference, " J. Amer. Statist. Assoc., 22, 1927, 209–212.

[4] L. D. Brown, "Confidence interval for two sample binomial distribution", J.Statist Plan. Infer.130, 2004, 359-375.

[5] L. D. Brown, T.T. Caio, A. DasGupta, "Interval estimation for a binomial proportion". Statist. Sci, 16, 2001, 101–133.

[6] L. D. Brown, T.T. Caio, A. DasGupta, "Confidence intervals for a binomial proportion and asymptotic expansions", Ann. Statist, 30, 2002, 160–201.

[7] R. Newcombe, "Interval estimation for the difference between independent propo-rtionns: comparison of eleven methods", Statist. Med, 17, 1998, 873–890.

[8] S. Lee, "Interval estimation of binomial proportions based on weighted Polya post-erior",Com.Statist.51, 2005, 1012- 10 21.

First Author: Masoud Ganji is Associate Professor at Department of Statistics, Faculty of Mathematical Science, University of Mohaghegh Ardabili, Ardabil, Iran. He obtained his Ph. D. from the National University of Delhi, India. His research interests are in the areas of Applied Mathematics, Statistical inference. He has published research articles in reputable international journals of Statistics and sciences. He is referee and editor of mathematical journals in the frame of Applied Mathematics and Statistics. He is currently Director of the journal of Hyperstructures.

Second Author: Solmaz Aghlmandi is M.Sc. student of Statistics in the university of Mohaghegh Ardabili, Ardabil, Iran. Her main research interests are Distribution theory, Parameter estimation, Bayes estimation.

Table Titles and Legends

Table 1. Compare coverage probability of 95% confidence intervals for variable n

Method	n=5	n=15	n=30	n=40	n=50
Wald	0.651	0.829	0.884	0.901	0.909
AC	0.967	0.963	0.961	0.959	0.957
Bayes	0.950	0.954	0.956	0.955	0.955
Exact	0.990	0.980	0.974	0.971	0.968

Table 2. Compare average expected lengths of 95% confidence intervals for variable

Method	n=5	n=15	n=30	n=40	n=50
Wald	0.519	0.365	0.270	0.236	0.212
AC	0.606	0.389	0.280	0.243	0.218
Bayes	0.571	0.379	0.276	0.241	0.216
Exact	0.676	0.418	0.297	0.256	0.228

Table. 3 Compare coverage probability of 95% confidence intervals

Method	Wald	AC	Bayes	newcomb
Δ=0,m=n=5	0.941	0.990	0.959	0.960
Δ=0.2,m=n=5	0.840	0.980	0.963	0.962
Δ=0,m=n=50	0.947	0.956	0.956	0.957
Δ=0.2,m=n=50	0.943	0.952	0.950	0.951
Δ=0,n=40,m=30	0.947	0.960	0.957	0.956
Δ=0.2,n=40,m=30	0.939	0.952	0.951	0.954
Δ=0,n=10,m=15	0.946	0.963	0.957	0.957
Δ=0.2,n=10,m=15	0.914	0.959	0.957	0.961
P=0.5,m=40,n=15	0.929	0.955	0.950	0.953
P=0.2,m=40,n=15	0.921	0.956	0.952	0.955

Table. 4 Compare average expected lengths of 95% confidence intervals

Method	Wald	AC	Bayes	newcomb
Δ=0,m=n=5	0.8044002	0.9084248	0.8499346	0.9740487
Δ=0.2,m=n=5	0.9079542	0.929898	0.8703352	0.9746431
Δ=0,m=n=50	0.302792	0.3076849	0.3047229	0.3102968
Δ=0.2,m=n=50	0.331739	0.3304354	0.3273357	0.3293394
Δ=0,n=40,m=30	0.3625507	0.3707945	0.3655725	0.3748306
Δ=0.2,n=40,m=30	0.3981733	0.3962452	0.3907539	0.3945575
Δ=0,n=10,m=15	0.5826204	0.6155786	0.594109	0.6320222
Δ=0.2,n=10,m=15	0.6473352	0.643662	0.6215511	0.6460252
P=0.5,m=40,n=15	0.547842	0.5250228	0.5122558	0.5250228
P=0.2,m=40,n=15	0.4569661	0.4633073	0.4521536	0.4629565