Project 3 -- Confidence Intervals

Lu Dong   Gina Girgis

Introduction:

In this project, we generate datasets from three different distributions: Normal distribution (N), Bernoulli distribution(B), and Uniform distribution (U). We mainly discuss and reason three things for the given ten functions: whether the confidence intervals are valid? Valid at what level? Valid for all dataset sizes or only asymptotically?

1. Experiment conclusions

Function_ No.	Valid?	What Level? Distribution:N/B/U	For all datasizes or asymptotically?
1	Yes	0/0/0	all datasizes
2	Yes	0/0/0	all datasizes
3	Yes	0.05/0.05/0.05	asymptotically
4	No		
5	No		
6	No		
7	Yes	0.4/0.1/0.005	all dataset sizes
8	Yes	0/0/0	all datasizes
9	Yes	0.01/0.01/0.01	asymptotically
10	Yes	0.3/0.05/0.001	all dataset sizes

2. How did you design the data to input to the test functions?

Since it's a black-box problem, we don't know the sample distribution, so we separately test data from three distributions. For normal distribution (N), we set X= sample_normal (N,1,0), where the true mean is 0; for bernoulli distribution (B), we set X= sample_bernoulli (N,0.5), where the true mean is 0.5; for uniform distribution (U), we set X= sample_uniform(N,0,1), where the true mean is 0.5. As for the data size-N, we change values from [10 100 1000 10000].

3. How did you design your testing procedure?

- After we get datasets X, we send it to Ci function like this, [a,b]=ci(X,f), where f is the function number (from 1 to 10). For each X, it returns a pair of [a,b].
- Supposing the given functions are using different ways to compute confidence intervals, they may have various ways to calculate the distance (ep) between the sample mean (Xbar) and the true mean (mu). However, as we already have the claimed [a, b] pairs, we could compute ep by equation ep=1/2*(b-a).
- If the sample mean captures the true mean (which can be expressed in Matlab as if a<=mu && b>=mu), we let count value plus one, and we repeat this process 10000 times (reps=1:10000).
- Finally, we print the following results: (function_no, N, ep, frac missed), where ep is the mean_ep, and frac missed computed by 1-count/10000.

The reasoning for valid:
First, whether the CI is valid can be tested by the CI definition if it doesn't satisfy the Pr[An<=mu<=Bn]>=1-alpha, then it is invalid. For this, we can observe the frac missed value. If frac missed close to 1, then it is invalid. Second, we also purposely set a false mu to see the frac missed performance. If frac missed close to 0, reflecting they capture a wrong mean, which can also help to double-check the validity. Third, based on CLT, the sample mean follows a normal distribution, so the ep value is supposed drop as N increases. Under normal distribution, we find function 4,5,6 frac missed value is 1. Under uniform distribution U(0,0.4), we set wrong mu=0.5, but frac missed value for function 4,5,6 are 0. Also, function 4 ,5, 6's ep value doesn't drop. So the confidence intervals (CI) of function 4, 5, 6 are invalid.

The reasoning for all dataset sizes or asymptotically:
If it's valid for all dataset sizes, the frac missed value will converge no matter how large the N value is. However, if it's valid asymptotically, the frac missed value will only converge when N is larger. Our experiments indicate the CI from function 1,2,7,8,10 are valid for all dataset sizes, and Ci from function 3 and 9 are valid only asymptotically.

The reasoning for CI level:
The converge value of frac missed will reflect the CI level. For function 1,2,8, their CI level is 0. For function 3 and 9, the CI level is respectively 0.05 and 0.01. For function 7 and 10, the CI level is affected by the sample distribution. In terms of normal/bernoulli/uniform distribution, the CI level of function 7 is 0.4/0.1/0.005, and the CI level of function 10 is 0.3/0.05/0.001.