ASSIGNMENT 3 – PROBLEM 4

P4. In this problem, you will explore confidence intervals.

(1) Generate 100 datasets, each with 50 instances. Each data point is drawn from the Gaussian distribution $\mathcal{N}(67, 3.8)$. That is, the true mean is $\mu = 67$ and the standard deviation is $\sigma = 3.8$.
Hint: You can generate a random two-dimensional array with the shape 100×50 in one go.

(2) Compute the sample mean $\hat{\mu}$ and sample standard deviation $\hat{\sigma}$ for each dataset. You will get an array of 100 for each.
Hint: For the standard deviation, there is a Numpy function `np.std()`. Here, you need to be careful about the normalization. By default, the function first computes the sample variance with a normalization factor $1/N$, where N is the number of data points in a set, in this case $N = 50$. However, it is known that this is a biased estimator for the true variance σ^2. Instead, you should use the unbiased estimator, which has the normalization factor $1/(N-1)$. You can do this by setting the parameter `ddof` to 1 when using `np.std()`. For more details, please check the user guide for `np.std()` online: https://numpy.org/doc/stable/reference/generated/numpy.std.html.

(3) For each dataset, compute the 95%-confidence interval for the true mean using the Gaussian formula for confidence interval. In other words, here you assume the true variance is known. You can check the lecture notes in `endtoend-pt2.pdf` about this. The 95%-confidence interval for μ is:

$$\{\bar{Z} - 1.96\frac{\sigma}{\sqrt{N}} < \mu < \bar{Z} + 1.96\frac{\sigma}{\sqrt{N}}\}, \tag{1}$$

where $\bar{Z}$ is the sample mean for each dataset.

(4) Plot the above results for all 100 datasets. For each dataset, plot the confidence interval and the sample mean. Also plot the true mean on the same figure. In other words, plot something like Fig. 1. You should observe that the sample mean is a random quantity, which is different for different datasets. Also note that the confidence intervals are also random. Count the number of datasets for which μ is outside the confidence interval.

(5) We next explore whether we truly get 95%-confidence levels and what we even mean by that. Generate $10,000$ datasets, each with 50 data points. Generate the 95%-confidence interval for μ using the

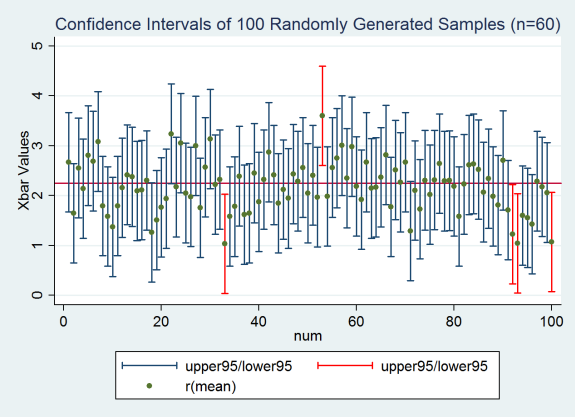


FIGURE 1

Gaussian formula. You will get 10,000 confidence intervals. Compute the percent of intervals that do not contain the true mean μ.

Repeat the above 10 times. You should get 10 values for the percentage. Show the 10 values. You will see that you get something close to 5%, but not close enough in some of the 10 cases.

(6) Repeat the previous step for $100,000$ datasets. In other words, generate $100,000$ datasets, each with 50 data points. Generate the 95%-confidence interval for μ using the Gaussian formula. You will get 100,000 confidence intervals. Compute the percent of intervals that do not contain the true mean μ.

Repeat the above 10 times. You should get 10 values for the percentage. Show the 10 values. You will see that you get much closer to 5% in all the 10 cases.

(7) We now compare confidence intervals based on Gaussian with those based on Stduent's t-distribution. This is most relevant when a dataset has a small number of samples, e.g., 20 or less data points.

Note that the Gaussian confidence interval has the true standard deviation σ in it. If we don't know the true σ, can we use the computed sample standard deviation $\hat{\sigma}$ to replace the true σ in the formula (1)? In other words, let us try the 95%-confidence interval for μ is:

$$\{\bar{Z} - 1.96\frac{\hat{\sigma}}{\sqrt{N}} < \mu < \bar{Z} + 1.96\frac{\hat{\sigma}}{\sqrt{N}}\}. \tag{2}$$

Generate $100,000$ datasets, each with 10 data points. Generate the 95%-confidence interval for μ using the formula in (2). You will get 100,000 confidence intervals. Compute the percent of intervals that do not contain the true mean μ.

Repeat the above 10 times. You should get 10 values for the percentage. Show the 10 values. You will see that you don't get near 5% in these ten runs.

(8) We now try the Student's t-distribution. The confidence interval has the same form:

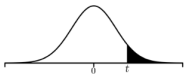
$$\{\bar{Z} - t\frac{\hat{\sigma}}{\sqrt{N}} < \mu < \bar{Z} + t\frac{\hat{\sigma}}{\sqrt{N}}\}. \tag{3}$$

However, the t value not only depends on the confidence level, but also the degree of freedom. For a dataset with N samples (data points), the degree of freedom is $N - 1$. In our case, the degree of freedom is 9. You can then look up in a table about the t value base on the confidence level and the degree of freedom. For instance, Figure 2 shows part of such a table. For 95%-confidence level, we should choose t such that the upper-tail probability is $P(T > t) = 0.025$, and by the symmetry of the Student's t-distribution, we also have $P(T < t) = 0.025$. The degree of freedom is 9 in the case of $N = 10$. Therefore, we find in the table that the correct t value is 2.262.

Now, generate 100,000 datasets, each with 10 data points. Generate the 95%-confidence interval for μ using the formula in (3) with $t = 2.262$. You will get 100,000 confidence intervals. Compute the percent of intervals that do not contain the true mean μ.

Repeat the above 10 times. You should get 10 values for the percentage. Show the 10 values. You will see that indeed you get near 5% in these ten runs.

Critical Values for Student's t-Distribution.



	Upper Tail Probability: $\Pr(T > t)$									
df	0.2	0.1	0.05	0.04	0.03	0.025	0.02	0.01	0.005	0.0005
1	1.376	3.078	6.314	7.916	10.579	12.706	15.895	31.821	63.657	636.619
2	1.061	1.886	2.920	3.320	3.896	4.303	4.849	6.965	9.925	31.599
3	0.978	1.638	2.353	2.605	2.951	3.182	3.482	4.541	5.841	12.924
4	0.941	1.533	2.132	2.333	2.601	2.776	2.999	3.747	4.604	8.610
5	0.920	1.476	2.015	2.191	2.422	2.571	2.757	3.365	4.032	6.869
6	0.906	1.440	1.943	2.104	2.313	2.447	2.612	3.143	3.707	5.959
7	0.896	1.415	1.895	2.046	2.241	2.365	2.517	2.998	3.499	5.408
8	0.889	1.397	1.860	2.004	2.189	2.306	2.449	2.896	3.355	5.041
9	0.883	1.383	1.833	1.973	2.150	2.262	2.398	2.821	3.250	4.781
10	0.879	1.372	1.812	1.948	2.120	2.228	2.359	2.764	3.169	4.587
11	0.876	1.363	1.796	1.928	2.096	2.201	2.328	2.718	3.106	4.437
12	0.873	1.356	1.782	1.912	2.076	2.179	2.303	2.681	3.055	4.318
13	0.870	1.350	1.771	1.899	2.060	2.160	2.282	2.650	3.012	4.221
14	0.868	1.345	1.761	1.887	2.046	2.145	2.264	2.624	2.977	4.140
15	0.866	1.341	1.753	1.878	2.034	2.131	2.249	2.602	2.947	4.073
16	0.865	1.337	1.746	1.869	2.024	2.120	2.235	2.583	2.921	4.015
17	0.863	1.333	1.740	1.862	2.015	2.110	2.224	2.567	2.898	3.965
18	0.862	1.330	1.734	1.855	2.007	2.101	2.214	2.552	2.878	3.922
19	0.861	1.328	1.729	1.850	2.000	2.093	2.205	2.539	2.861	3.883
20	0.860	1.325	1.725	1.844	1.994	2.086	2.197	2.528	2.845	3.850

FIGURE 2. t values