情報学基礎 B

統計 3
Confidence Interval and its Determination, Central Limit Theorem

Confidence Interval (信頼区間)for μ of the Normal Distribution with Known σ^2 :
Determination of a Confidence Interval for the Mean μ of a Normal Distribution with Known Variance σ^2
Here, Sample mean

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}(x_1 \ + \ x_2 \ + \cdots + \ x_n)$$

Here, n the sample size. The variance σ^2 is either known or calculated as follows:,

$$\sigma^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 \ + \ \cdots \ + \ (x_n - \bar{x})^2]$$

Confidence interval for mean of a normal distribution
In sampling from a normal distribution whose mean takes some unknown value μ, it is still true that $\bar{X}$ will follow $N(\mu, \ \sigma^2/n)$ and therefore that $Z = \sqrt{n}(\bar{X} - \mu)/\sigma$ will be Standardized normal distribution $\Phi(z)$. Using tha table of the standard normal distribution, we can then say that

$$Pr(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq +1.96) = 0.95$$

If we are given the values of σ^2 and n, we can find the set of values of μ that satisfy the inequality within the brackets. Now

$$-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq +1.96$$

if and only if

$$-1.96\sigma/\sqrt{n} \leq \bar{X} - \mu \leq 1.96\sigma/\sqrt{n}$$

The above statement is equivalent to

$$Pr(-1.96\sigma/\sqrt{n} \leq \bar{X} - \mu \leq 1.96\sigma/\sqrt{n}) = 0.95$$

The inequality in brackets is true with probability 0.95, and hence the inequality for μ inside the brackets is also true with probability 0.95. Any μ within the limits given is acceptable, at this level of probability, as the true mean in the normal distribution from with the sample of n has been drawn. We call the interval in the brackets a 95% confidence interval for μ and the two ends of the interval the 95% confidence limits. Note that μ is fixed (though unknown), and it is the interval which is the random variable: for if we take a new sample of n from the same distribution, μ does not change, but $\bar{x}$ does and therefore so does the interval. From any one sample, with mean $\bar{x}$, we claim that the interval contains the true (but unknown) μ. From many samples we would obtain many intervals; if we claimed in each case that the interval contained μ then in the long run 95% of such claims would be true. We denote range as $\pm\sigma\, c/\sqrt{n}$. In the following table we list the confidence level γ and corresponding range c.

confidence level γ	0.90	0.95	0.98	0.99	0.999
corresponding c	1.645	1.960	2.326	2.576	3.291

Example:
A machine which packs sugar has a normal distribution of weights of filled packets, and the standard deviation of weight has been 2.5 grams. 20 of the new packets are weighed. Their mean is 1002 grams. Set up 95% confidence limits for the true mean weight after adjustment.
(ある砂糖を詰める機械ではいっぱいに詰めた重量は標準分布で、その重さの標準偏差は 2.5g である。20 個の袋の重さを量ったところ、平均は 1002g であった。調整後に正しい平均重量になるように 95%信頼度上限を決めよ)

Example:
Determine a 95% confidence interval for the mean of a normal distribution with variance $\sigma^2 = 9$, using a sample of $n = 100$ with mean $\bar{x} = 5$?
(標準分布で サンプルの個数が 100、平均が 5 で $\sigma^2 = 9$ のとき、95% 信頼区間を求めよ)

Example:
How large must n be in Example 1 if we want to obtain a 95% confidence interval of length $L = 0.4$?
(Example 1 おいて $L = 0.4$ の長さの 95% 信頼区間を得るように n を決めよ)

Example:
If we wish to set up 99% confidence limits for the true μ, we replace 1.96 by 2.576.
(正しい μ で 99%の信頼区間を得るために、1.96 ではなく 2.576 を使う)

Example:
It is required to estimate μ the mean length of a mas-produced screws, to within limits of $\pm\frac{1}{2}mm$. What is the minimum sample size needed to achieve this, if the standard deviation of length is known to be1.2mm, and the

probability attached to the limits is to be 95% ?
(限界値 $\pm\frac{1}{2}mm$ でねじの長さの平均から μ を計算する。長さの標準偏差が 1.2mm で、95% がその制限値に含まれるときの最小のサンプルのサイズを求めよ)

Problem

1. A random sample of one observation x is chosen from a normal distribution with unknown mean μ and variance 4. Give a 95% confidence interval for μ
(標準分布で平均は未知、分散が 4 のである x の 95% 信頼区間の μ を求めよ)

2. A random sample of 16 observations, with mean $\bar{x}$, is chosen from a normal distribution with unknown mean μ and variance 25. Give a:(i) 95%; (ii) 99%; (iii) 99.9% confidence interval for μ.
(平均が $\bar{x}$ である 16 個のサンプルから、標準分布で未知の平均で分散が 25 の時の (i) 95%; (ii) 99%; (iii) 99.9% である μ を求めよ)

3 . A random sample of 100 packs of apples with a professed weight of 1kg were found to have mean 1020g; the estimated variance from the sample was 144g^2. Give an approximate 95% confidence interval for the mean of the apple packs.
(1kg と重量の決められたリンゴが 100 パックの平均が 1020g であった。サンプルの分散は 144g^2 であるとき、このリンゴのパックの平均が 95% 信頼区間になるように求めよ)

Central Limit Theorem :
The central idea of **Central Limit Theorem** is the concept of sampling distribution. How the mean and standard deviation of sample distribution changes with sample size. Suppose we have 1000 random numbers from 0 to 10. The mean will be very near to 5 (if the samples are true uniform random). Now out of these 1000 samples, if we take just 5 samples, we can not expect the mean to be 5. We can create 200 samples (data), all of them are mean to 5 randomly selected samples out of the total of 1000. What will be the distribution of these 200 data? What will be their standard deviation? **Central limit theorem** gives an answer to them.
Suppose we have the following 500 data of weight (in pounds) of 3 years old children. This data has a normal distribution with $\mu = 38.72$ pounds and $\sigma = 3.17$. We randomly took 5 data out of the whole set and put them in one row. Finally we have 100 rows of data, and the mean of each row of 5 data is written on the right. Thus, we get 100 samples, which are mean of randomly selected 5 samples of the original normally distributed data set. What is the distribution of these 100 samples. If we draw histogram, it shows that the distribution is still normal. Now what is the mean of these 100 samples? What is their s.d.? The mean is still 38.72 pounds, as was that of the original sample. But the standard deviation is now only 1.374. If we take 10 samples in a row, we would get 50 samples which are mean of 10 samples each from the original data. The mean of thes 50 samples is still the same, 38.72 pounds, but the s.d. is reduced further to 1.0.
Theorem *The law of large number:* As the sample size increases, the sample mean gets closer to the population mean. Thus, when an additional observation is added to the sample, the difference between the sample mean, $\overline{x}$, and the population mean μ approaches to zero.
Theorem *The mean and s.d. of the sampling distribution of $\overline{x}$:* Suppose n samples are drawn from a population with mean μ and s.d. σ. The sampling distribution of $\overline{x}$ will have a mean $\mu_{\overline{x}} = \mu$ and s.d. $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$. The term $\sigma_{\overline{x}}$ is often called the **standard error of the mean**.

Sample	Sample of Size n = 5					Sample Mean
1	36.48	39.94	42.57	39.53	33.81	38.47
2	43.13	37.97	42.41	39.61	43.30	41.28
3	41.64	39.01	37.77	38.94	41.10	39.69
4	40.37	43.49	37.60	40.14	38.88	40.10
5	38.62	33.43	45.17	42.66	39.98	39.97
6	38.98	41.35	36.80	43.56	39.92	40.12
7	42.48	37.00	35.87	39.62	38.74	38.74
8	39.38	37.02	41.60	40.34	37.62	39.19
9	42.82	45.77	35.16	42.56	39.75	41.21
10	36.19	35.20	37.74	40.46	37.47	37.41
11	36.59	41.62	42.18	39.23	39.26	39.77
12	38.57	42.13	45.39	38.22	46.18	42.10
13	38.40	39.06	43.60	31.46	37.03	37.91
14	34.29	47.73	37.27	41.82	33.33	38.89
15	42.28	43.29	37.69	37.32	40.06	40.13
16	34.31	43.58	40.02	41.13	42.99	40.41
17	38.71	39.03	39.39	42.62	38.41	39.63
18	38.63	39.66	39.47	41.13	38.01	39.38
19	39.09	33.86	37.57	41.65	35.22	37.48
20	40.94	37.50	38.72	41.64	35.48	38.86
21	38.72	35.89	37.82	35.04	37.06	36.91
22	39.64	36.30	35.54	40.40	38.74	38.12
23	38.22	38.49	33.60	40.18	39.07	37.91
24	40.93	40.53	37.55	37.30	37.16	38.70
25	33.27	38.92	37.14	39.90	33.83	36.61
26	39.44	37.28	35.70	41.97	36.80	38.24

27	38.83	41.41	38.87	39.40	37.20	39.14
28	40.10	36.96	35.73	43.00	38.11	38.78
29	41.93	36.57	37.55	35.14	38.75	37.99
30	31.25	38.85	39.25	35.07	39.77	36.84
31	38.47	34.45	30.43	41.76	41.61	37.34
32	37.98	35.56	43.97	44.96	37.81	40.06
33	43.34	40.94	35.17	41.74	37.59	39.76
34	39.80	44.44	37.53	40.52	41.95	40.85
35	41.98	42.02	40.73	40.47	36.81	40.40
36	40.98	35.08	34.61	40.78	37.26	37.74
37	35.75	40.81	40.13	35.99	36.52	37.84
38	36.39	45.97	40.59	37.64	42.42	40.60
39	36.20	35.63	37.43	38.35	34.81	36.48
40	33.58	33.87	41.60	45.10	38.68	38.57
41	31.77	38.34	41.79	37.93	40.83	38.13
42	43.03	33.12	34.98	36.58	37.78	37.10
43	35.76	35.17	42.58	39.10	41.08	38.74
44	38.44	38.45	35.93	35.32	44.60	38.55
45	44.54	41.88	35.84	42.64	42.38	41.46
46	41.89	36.81	41.83	40.24	39.28	40.01
47	38.00	40.08	35.57	34.44	39.51	37.52
48	39.92	38.05	39.96	38.04	32.11	37.61
49	36.37	38.62	32.25	41.35	40.91	37.90
50	34.38	36.65	32.97	39.93	41.34	37.05
51	40.32	39.80	41.00	38.62	38.24	39.59
52	37.95	45.26	38.67	34.96	41.13	39.60
53	36.82	42.63	41.62	39.43	37.48	39.59
54	41.63	37.65	38.58	39.03	37.53	38.88
55	37.91	37.20	38.72	36.87	45.40	39.22
56	41.05	34.01	39.11	38.23	35.74	37.63
57	42.09	45.44	35.52	39.87	37.28	40.04
58	39.31	35.79	37.82	39.15	35.57	37.53
59	41.16	39.98	41.11	39.21	39.98	40.29
60	35.68	45.60	39.34	36.65	43.30	40.12
61	36.07	39.63	42.55	41.72	36.81	39.36
62	38.97	36.83	41.01	38.12	35.27	38.04
63	33.70	39.15	34.81	34.13	39.00	36.16
64	37.19	34.69	36.21	34.34	39.07	36.30
65	33.99	44.87	42.52	40.22	39.26	40.17
66	41.40	27.62	34.57	40.08	34.65	35.66
67	40.14	34.45	38.26	38.09	39.72	38.13
68	33.64	42.62	32.08	34.30	37.34	35.99
69	35.36	39.02	43.98	41.19	32.47	38.40
70	43.26	37.85	35.82	37.11	36.22	38.05
71	36.24	38.07	33.38	38.43	39.88	37.20
72	38.55	43.06	41.07	36.58	37.02	39.25
73	41.26	36.99	36.17	38.98	36.03	37.89
74	37.31	38.41	41.18	39.76	39.64	39.26
75	32.26	41.84	42.50	37.70	41.21	39.90
76	39.27	38.61	44.53	38.08	35.01	39.10
77	39.14	40.83	39.83	37.78	36.51	38.82
78	42.53	43.41	41.01	33.71	39.47	40.03
79	45.34	32.61	33.81	39.03	40.32	38.22
80	36.31	35.55	37.12	38.74	40.80	37.70
81	31.40	41.80	40.15	42.53	37.62	38.70
82	41.01	39.02	39.68	36.61	38.44	38.95
83	34.15	36.19	35.98	36.02	36.32	35.73
84	31.50	37.61	43.29	39.82	38.78	38.20

85	43.26	34.01	41.18	40.23	39.28	39.59
86	41.76	41.40	39.02	38.20	39.42	39.96
87	37.06	35.95	39.98	40.00	43.36	39.27
88	41.01	37.56	36.95	39.71	37.97	38.64
89	34.97	38.36	36.30	38.48	34.24	36.47
90	38.38	38.94	40.96	36.13	35.98	38.08
91	39.41	30.78	37.66	37.31	42.04	37.44
92	39.83	35.88	30.20	45.07	40.06	38.21
93	36.25	39.56	34.53	40.69	37.03	37.61
94	45.46	40.66	44.51	40.50	39.43	42.15
95	37.63	44.77	38.31	36.53	38.41	39.13
96	39.78	33.34	43.42	43.63	38.77	39.79
97	41.48	37.39	38.62	43.83	34.26	39.12
98	37.68	40.66	38.93	40.94	37.54	39.15
99	39.72	32.61	32.62	40.35	38.65	36.79
100	39.25	41.06	41.17	38.30	38.24	39.60

Theorem *The central limit theorem:* Suppose a random variable X has population mean μ and s.d. σ. Sample size of n is randomly taken from it. Then the value of $\overline{x}$ becomes approximately normal as n increases, irrespective the distribution of the original X. The mean of the distribution is $\mu_{\overline{x}} = \mu$ and the s.d. $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$.
To illustrate *The central limit theorem* we will give an example here. The data is the waiting time in queue as described below. First you need to draw the histogram and see that it fits to exponential distribution. For exponential distribution, let me repeat that the distribution function is:

$$f(x) = \begin{cases} \lambda\, e^{-\lambda\, x} & \text{if } x \geq 0 \\ 0 & x < 0 \end{cases}$$

and both the average (μ) and s.d. (σ) are $\frac{1}{\lambda}$ (verify analytically using the above equation0. From the 90 data generated randomly using the above equation is listed below. Draw the histogram of the 90 data. Show that the distribution is exponential. Calculate μ and σ.
Now get the average of the five entries per row, and draw the histogram of these 18 average values. Show that they resemble normal distribution. Find the average and s.d. of that distribution. Next time take the average of 10 data (two rows of original data at a time) and find the average and s.d.. See how it is changed. Does all the results conform to *the central limit theorem*?

```
Roller coaster Wait Times

7.00   33.0   30.0   4.00   35.0
94.0   3.00   76.0   6.00   3.00
39.0   18.0   5.00   14.0   21.0
2.00   9.00   4.00   107.   8.00
0.00   7.00   15.0   61.0   8.00
9.00   86.0   18.0   53.0   38.0
37.0   45.0   6.00   0.00   9.00
38.0   16.0   15.0   11.0   93.0
47.0   21.0   41.0   81.0   1.00
68.0   5.00   61.0   7.00   94.0
1.00   6.00   55.0   9.00   25.0
10.0   3.00   94.0   64.0   80.0
22.0   115.   9.00   16.0   51.0
18.0   19.0   18.0   79.0   13.0
80.0   21.0   1.00   0.00   41.0
24.0   26.0   8.00   40.0   18.0
2.00   6.00   24.0   14.0   1.00
11.0   9.00   12.0   12.0   47.0
```