CALIFORNIA STATE UNIVERSITY, LONG BEACH EE 381 - Probability, Statistics, and Stochastic Modeling

Projects

Confidence Intervals

0. Introduction and Background Material

0.1. Sample size and confidence intervals

Assume that you are measuring a statistic in a large population of size *N*. The statistic has mean μ and standard deviation σ. Drawing a sample of size n from the population, produces a distribution for the sample mean ($\overline{X}$) with:

$$E[\overline{X}] = \mu_{\overline{X}} = \mu \qquad \text{and} \qquad E[(\overline{X} - \mu_{\overline{X}})^2] = \sigma_{\overline{X}}^2 = \frac{\sigma^2}{n} \Rightarrow \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

In this project we will explore the relation of $\overline{X}$ to the population mean μ.

- As a first example consider a barrel of a million ball bearings (i.e. population size $N = 1,000,000$) where someone has actually weighed all one million of them and found the exact mean to be $\mu = 100$ grams and the exact standard deviation to be $\sigma = 12$ grams. This is obviously an unrealistic assumption, but assume for the time being that these parameters have been measured exactly.
- Now pick a sample of size n (for example $n = 5$) of bearings from the barrel, weigh them and find the mean of the sample, $\overline{X}_5 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$.
- Next take a larger sample (for example $n = 10$) and find the new mean $\overline{X}_{10} = \frac{X_1 + X_2 + \cdots + X_{10}}{10}$.
- Continue this process for larger and larger n, until $n = 100$.
- Plot the points $(n, \overline{X}_n)$ using a point marker (for example a blue 'x') as shown in Figure 1.

- Next for each value of n, calculate the standard deviation of the sample from $\sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}}$ and plot:

  (i) The values of $\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$ with respect to n, shown as the solid red curve in the figure. *These curves define the 95% confidence interval,* which means that approximately 95% of the sample means will fall within the two solid red curves. This can also be visually confirmed by looking at how many of the sample means fall outside of the solid red curves (approx. 5%).

  (ii) The values of $\mu \pm 2.58\frac{\sigma}{\sqrt{n}}$ with respect to n, shown as the dashed red curve in the figure. *These curves define the 99% confidence interval,* which means that approximately 99% of the sample means fall within the dashed red curves.

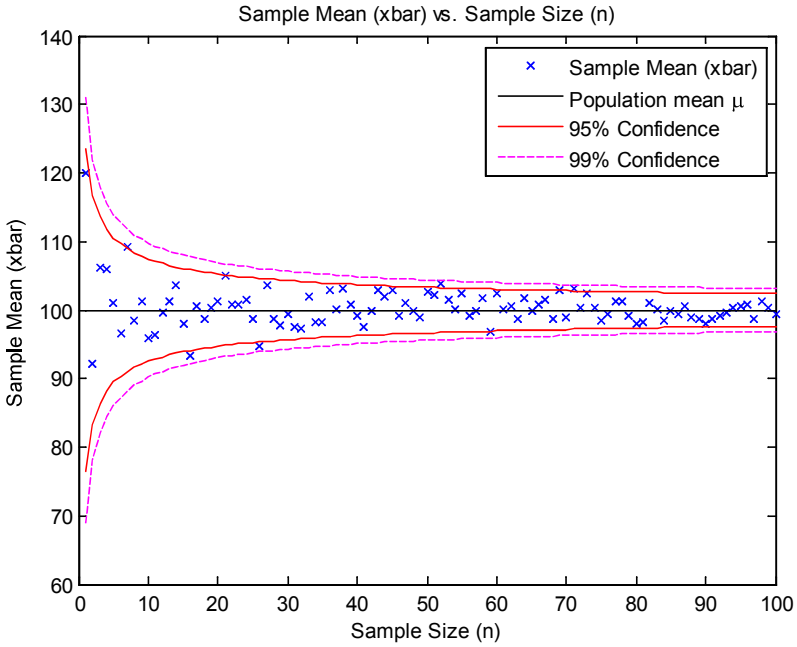

Figure 1. Sample mean as a function of the sample size

0.2. Example code

The following example code will plot the values of the mean $\overline{X}$ for six different sample sizes n: $n = \{5,6,7,8,9,10\}$. The code uses the Python function "mean".

```
clear
close all
mu_p=100;  % Population mean
sig_p=12;  % Population SD
%
Nb=1e6;          % Number of bearings in production
n=5:10; n=n';    % sample sizes
kmax=max(size(n));
M=zeros(length(n),2);
%
Bearings=sig_p*randn(Nb,1)+mu_p;     % Create the population of bearings

for k=1:kmax
        x_index=ceil(Nb*rand(k,1));
        x=Bearings(x_index);
        x_bar=mean(x);
        M(k,:)=[k x_bar ];
end
figure(1); plot(n,M(:,2),'o', n, mu_p*ones(size(n)),'k')
figure(1); hold on
plot(n, mu_p+1.96*sig_p./sqrt(n),'r', n, mu_p+2.58*sig_p./sqrt(n),'m--')
plot(n, mu_p-1.96*sig_p./sqrt(n),'r', n, mu_p-2.58*sig_p./sqrt(n),'m--')
figure(1); hold off
```

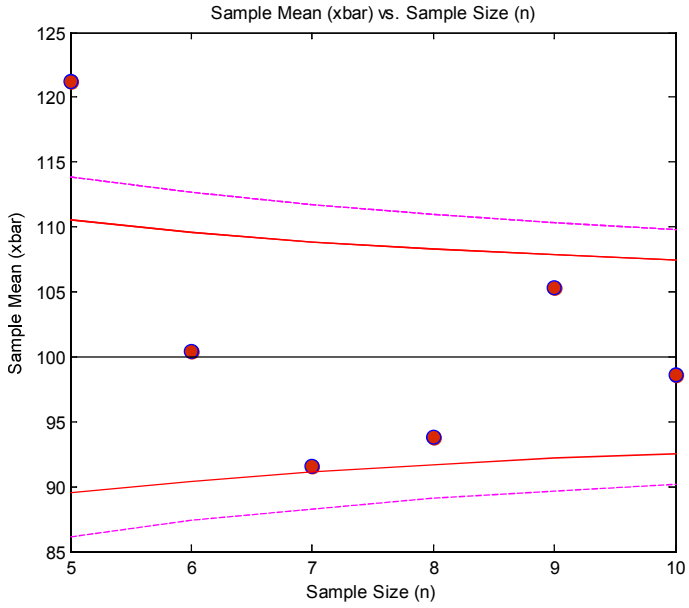

Figure 2. Example plot for $n = \{5,6,7,8,9,10\}$

0.3. Using the sample mean to estimate the population mean

In reference to the previous section, it is obviously unrealistic to think that anyone actually measured the exact mean and standard deviation of all one million ball bearings. More realistically, you would not have any idea what the mean or standard deviation was, and you would need to weigh random samples of different sizes (for example $n = 5, 35,$ or 100 bearings) and then draw reasonable conclusions about the weight distribution of all one million bearings.

To simulate this problem, generate a barrel of a million ball bearings with weights normally distributed, with a mean μ and a standard deviation σ.

As an example, take a sample of n bearings from the population of $N = 1,000,000$.

Then calculate the mean of the sample:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

The standard deviation of the sample mean can be calculated by:

$$\sigma_n = \frac{\hat{S}}{\sqrt{n}} \text{ , where } \hat{S} = \left\{ \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} \right\}^{1/2}$$

The question is: *Can the value of $\overline{X}$, which is calculated for a sample of size n, be used to estimate the mean μ of the population of $N = 1,000,000$ bearings?*

The answer is given in terms of confidence intervals, typically the 95% and 99% confidence intervals.

<u>Large samples</u> ($n \geq 30$)**.**

Consider the standardized variable $z = \dfrac{\overline{X} - \mu}{\sigma_n}$. Based on the Central Limit Theorem, it is known that for large samples the standardized variable z will approach the normal distribution.

The 95% confidence interval for large samples. *The 95% confidence interval* is determined by the critical values $[-z_c, z_c]$ such that

$$P\{-z_c < z < z_c\} = 0.95$$

From the tables of the normal distribution it is seen that these critical values correspond to $z_c = 1.96$ and $-z_c = -1.96$. Hence:

$$P\{-z_c < z < z_c\} = 0.95 \Rightarrow P\left\{-1.96 < \frac{\overline{X} - \mu}{\sigma_n} < 1.96\right\} = 0.95 \Rightarrow P\left\{\overline{X} - 1.96\sigma_n < \mu < \overline{X} + 1.96\sigma_n\right\} = 0.95$$

which is written as

$$P\left\{\bar{X}-1.96\frac{\hat{S}}{\sqrt{n}}<\mu<\bar{X}+1.96\frac{\hat{S}}{\sqrt{n}}\right\}=0.95$$

If you define: $\mu_{\text{Lower}}=\bar{X}-1.96\frac{\hat{S}}{\sqrt{n}}$ and $\mu_{\text{Upper}}=\bar{X}+1.96\frac{\hat{S}}{\sqrt{n}}$, then the above equation is written as:

$$P\left\{\mu_{\text{Lower}}<\mu<\mu_{\text{Upper}}\right\}=0.95$$

This equation can be interpreted as follows:

Based on a sample of size n, we are 95% confident that the mean μ of the population lies in the interval $[\mu_{\text{Lower}},\mu_{\text{Upper}}]$.

Another way to interpret this statement is:

(i) Obtain a large number of different samples, of size n each

(ii) For each of these samples calculate $\bar{X}$ and $\sigma_n=\frac{\hat{S}}{\sqrt{n}}$

(iii) For each of these samples generate the corresponding intervals $[\mu_{\text{Lower}},\mu_{\text{Upper}}]$

(iv) Then *the mean μ of the population will be contained in 95% of these intervals*

For example if you obtain 500 samples of size n each, and you create the corresponding 500 intervals $[\mu_{\text{Lower}},\mu_{\text{Upper}}]$, then 475 of these intervals (95% of 500) will contain the population mean μ.

The 99% confidence interval for large samples. *Similarly, the 99% confidence interval* is determined by the critical values $[-z_c,z_c]$ such that

$$P\{-z_c<z<z_c\}=0.99$$

From the tables of the normal distribution it is seen that these critical values correspond to $z_c=2.58$ and $-z_c=-2.58$.

Hence, if you define: $\mu_{\text{Lower}}=\bar{X}-2.58\frac{\hat{S}}{\sqrt{n}}$ and $\mu_{\text{Upper}}=\bar{X}+2.58\frac{\hat{S}}{\sqrt{n}}$, then you can write:

$$P\left\{\mu_{\text{Lower}}<\mu<\mu_{\text{Upper}}\right\}=0.99$$

which can be interpreted as follows:

Based on a sample of size n, we are 99% confident that the mean μ of the population lies in the interval $[\mu_{\text{Lower}},\mu_{\text{Upper}}]$.

Small samples ($n < 30$).

When the sample size is small, the standardized variable $\frac{\overline{X} - \mu}{\sigma_n}$ does not approach the normal distribution, since the Central Limit Theorem does not hold for small n.

In this case the standardized variable $T = \frac{\overline{X} - \mu}{\sigma_n}$ follows the *Student's t distribution* with $\nu = n - 1$ degrees of freedom.

The 95% confidence interval for small samples. *The 95% confidence interval* is determined by the critical values $[-t_c, t_c]$ such that

$$P\{-t_c < z < t_c\} = 0.95$$

Note that the critical values $[-t_c, t_c]$ depend on two values: (i) the probability value (0.95) and (ii) the degrees of freedom (ν).

Example: sample size $n = 5$. In that case $\nu = n - 1 = 4$ degrees of freedom. For the 95% confidence interval, the critical value is: $t_c = t_{0.975} = 2.78$, as seen from the *Student's t distribution* tables.

Hence, if you define: $\mu_{\text{Lower}} = \overline{X} - 2.78\frac{\hat{S}}{\sqrt{5}}$ and $\mu_{\text{Upper}} = \overline{X} + 2.78\frac{\hat{S}}{\sqrt{5}}$, then the above equation is written as:

$$P\left\{\mu_{\text{Lower}} < \mu < \mu_{\text{Upper}}\right\} = 0.95$$

which can be interpreted as follows:

Based on a sample of size $n = 5$, we are 95% confident that the mean μ of the population lies in the interval $[\mu_{\text{Lower}}, \mu_{\text{Upper}}]$.

The 99% confidence interval small samples. *Similarly the 99% confidence interval* is determined by the critical values $[-t_c, t_c]$ such that

$$P\{-t_c < z < t_c\} = 0.99$$

Example: sample size $n = 5$. In that case $\nu = n - 1 = 4$ degrees of freedom. For the 99% confidence interval, the critical value is: $t_c = t_{0.995} = 4.60$, as seen from the *Student's t distribution* tables.

Hence, if you define: $\mu_{\text{Lower}} = \overline{X} - 4.60\frac{\hat{S}}{\sqrt{5}}$ and $\mu_{\text{Upper}} = \overline{X} + 4.60\frac{\hat{S}}{\sqrt{5}}$, then the above equation is written as:

$$P\left\{\mu_{\text{Lower}} < \mu < \mu_{\text{Upper}}\right\} = 0.99$$

which can be interpreted as follows:

Based on a sample of size $n = 5$, we are 99% confident that the mean μ of the population lies in the interval $[\mu_{\text{Lower}}, \mu_{\text{Upper}}]$.

1. Effect of sample size on confidence intervals

Create the plot of Figure 1, showing the effect on sample size on the confidence intervals.
The values of the following parameters have been provided to you:

- Total number of bearings: N ;
- Population mean: μ (grams) ;
- Population standard deviation: σ (grams) ;
- Sample sizes: $n = 1, 2, \cdots 200$

SUBMIT a report with the plot and the MATLAB code in a Word file. You must follow the guidelines given in the syllabus regarding the structure of the report. Points will be taken off if you do not follow the guidelines.

2. Using the sample mean to estimate the population mean

(A) Perform the following simulation experiment. Use Table 1 to tabulate the results.

1. Choose a random sample of $n = 5$ bearings from the N bearings you created in the previous problem. Calculate the sample mean and the sample standard deviation:

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j \quad \text{and} \quad \hat{S} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

2. Create the 95% confidence interval using the normal distribution to fill in the first two entries in the top row. You realize, however, that this is not an appropriate distribution to use because you have a small sample $n = 5 < 30$

$$[\mu_{\text{Lower}}, \mu_{\text{Upper}}] = [\bar{X} - 1.96\frac{\hat{S}}{\sqrt{n}}, \bar{X} + 1.96\frac{\hat{S}}{\sqrt{n}}]$$

3. Check if the confidence interval includes the actual mean μ of the population of N bearings. If it does, then Step 2 is considered a success.
4. The appropriate distribution for small samples ($n \le 30$) is the t-distribution. Create the 95% confidence interval using the t-distribution with $\nu = n - 1 = 4$

$$[\mu_{\text{Lower}}, \mu_{\text{Upper}}] = [\bar{X} - t_{0.975}\frac{\hat{S}}{\sqrt{n}}, \bar{X} + t_{0.975}\frac{\hat{S}}{\sqrt{n}}]$$

At the 95% confidence level with $\nu = 4$ degrees of freedom, the value of $t_{0.975}$ can be found from the tables, and it seen to be: $t_{0.975} = 2.78$. This is the value that will be used to determine the 95% confidence interval: $[\mu_{\text{Lower}}, \mu_{\text{Upper}}] = [\bar{X} - 2.78\frac{\hat{S}}{\sqrt{n}}, \bar{X} + 2.78\frac{\hat{S}}{\sqrt{n}}]$

5. Check if the confidence interval includes the actual mean μ of the population. If it does, then Step 4 is considered a success.
6. Repeat the experiment for $M = 10,000$ times and count the number of successes.
7. Enter the percentage of successful outcomes in Table 1.
8. Repeat steps 1-7 above with $n = 5$ and 99% confidence interval.

9. After completing all of the above steps you will have filled out the first row of the table.

(B) Repeat part (A) with $n = 40$ using the normal distribution and the t-distribution to complete the second row of the table.

(C) Repeat part (A) with $n = 120$ using the normal distribution and the t-distribution to complete the third row of the table. You realize, however, that for a large sample ($n > 30$) the t-distribution will be very close to normal, so the differences between Student's -t and Normal will be minimal.

SUBMIT a report with the results and the MATLAB code in a Word file. You must use Table 1 to report the results, and you must follow the guidelines given in the syllabus regarding the structure of the report. Points will be taken off, if you do not follow the guidelines and if you do not use Table 1 to report the results.

Sample size (n)	**95% Confidence** (Using Normal distribution)	**99% Confidence** (Using Normal distribution)	**95% Confidence** (Using Student's t distribution)	**99% Confidence** (Using Student's t distribution)
5	88.3	93.83	94.93	97.87
40	94.18	98.52	94.85	99.06
120	94.99	99.01	94.7	98.94
200	94.59	98.9	94.88	99.0

Table 1. Success rate (percentage) for different sample sizes