CONFIDENCE INTERVALS

1 Sampling distribution of sample mean ($\bar{X}$)

Let $X_1, X_2, \dots, X_n$ be independent and identically distributed (i.i.d) random variables with $E(X) = \mu$ and $Var(X) = \sigma^2$. Consider the expectation and standard deviation of the sample mean ($\bar{X}$) :

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + X_3 + \cdots + X_n)$$

$$E(\bar{X}) = \frac{1}{n}\big(E(X_1) + E(X_2) + \cdots + E(X_n)\big) = \mu$$

$$Var(\bar{X}) = \frac{1}{n^2}\big(Var(X_1) + Var(X_2) + \cdots + Var(X_n)\big) = \frac{\sigma^2}{n}$$

$$SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

We can standardize (compute the z-score of) the sample mean:

$$Standardized\ sample\ mean:\ \ Z_{\bar{X}} = \left(\frac{\bar{X} - E(\bar{X})}{SD(\bar{X})}\right) = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

The Central Limit Theorem (CLT) tells us that the distribution of $Z_{\bar{X}}$ approaches standard normal as n approaches infinity:

$$The\ Central\ Limit\ Theorem:\ \ \lim_{n\to\infty} \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right) \sim N(0, 1)$$

If the population is normally distributed, then $Z_{\bar{X}} \sim N(0, 1)$ regardless of the sample size.

2 Confidence Intervals

A confidence interval (CI) is an interval that is likely to contain a parameter of interest. A two-sided confidence interval has the following form:

$$(point\ estimate) \pm (margin\ of\ error)$$

$$width\ of\ the\ CI:\ \ 2 \times (margin\ of\ error)$$

The confidence level indicates the proportion of intervals that contain the parameter, if the experiment is repeated a large number of times. For example, a 95% confidence level means that if we repeat the experiment many times and construct the 95% CI for each experiment, we expect that 95% of these intervals contain the parameter. Once the CI is constructed, the probability of the parameter lying in the interval is either 0 (if the parameter is outside the interval) or 1 (if the parameter is inside the interval).

3 Confidence intervals for population mean ($\boldsymbol{\mu}$)

Let $X_1, X_2, \dots, X_n$ be independent and identically distributed (i.i.d) random variables with $E(X) = \mu$ and $Var(X) = \sigma^2$. We know from the CLT that

$$Z_{\bar{X}} = \left(\frac{\bar{X} - E(\bar{X})}{SD(\bar{X})}\right) \sim N(0,1)$$

provided that n is large. If X is normally distributed, then $Z_{\bar{X}} \sim N(0,1)$ regardless of the sample size n.

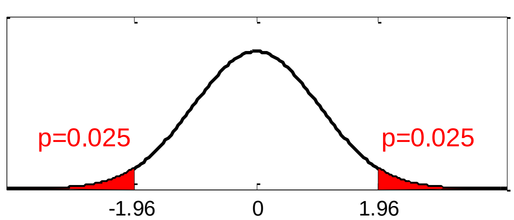
3.1 Known population standard deviation

Two-sided CI for population mean μ has the following form

$$Two\text{-}sided\ CI\ for\ \ \mu: \quad \bar{X} \pm z_{cr}\, SD(\bar{X}) \qquad \text{where } SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

$$margin\ of\ error: \ z_{cr} SD(\bar{X}) = z_{cr} \frac{\sigma}{\sqrt{n}}$$

where z_{cr} is the critical value of standard normal distribution that depends on the confidence level: For $(1-\alpha) \times 100\%$ confidence: $z_{cr} = \Phi^{-1}(1 - \alpha/2)$ where Φ is the standard normal CDF.



$$P(-z_{cr} \le Z_{\bar{X}} \le z_{cr}) = 1 - \alpha$$

For 95% confidence ($\alpha = 0.05$), $z_{cr} = 1.96$.

Example: Generate a single sample of 4 observations from a normal distribution with mean 12 and standard deviation 10. Construct a two-sided, 95% confidence interval for the population mean, and check whether the confidence interval captures the population mean.

You can skip the comments after % sign. The line numbers (e.g. `*Line1*`) will be needed when modifying the code.

```
n=4; % *Line 1*
pm=12;  ps=10;  % population mean and std    *Line 2*
figure;  hold on;  % start an empty figure *Line 3*
%*Line 4*
i=1;    % *Line 5*
x=pm+ps*randn(n,1);  % *Line 6*
mx=mean(x); % *Line 7*
zcr=1.96;  % *Line 8*
sem=ps/sqrt(n);  % standard error of the mean  *Line 9*
me=zcr*sem; % margin of error,  *Line 10*
CI1= mx-me;    % lower CI bound *Line 11*
CI2= mx+me;   % upper CI bound *Line 12*
```

Plot the observation points as black dots scattered vertically. On the same plot, show the confidence limits as green pluses. Check whether the confidence interval was able to capture the true mean. If the confidence interval has missed the true mean, show a red asterisk where the true mean is:

```
plot(i*ones(1,n), x, 'k.','markersize',5); % show points *Line 13*
plot(i*ones(1,2),[CI1, CI2],'g+') % show CI * Line 14*
if pm<CI1 || pm>CI2  % missed    *Line 15*
 plot(i,pm,'r*','markersize',10)  %*Line 16*
% *Line 17*
end  % *Line 18*
%*Line 19*
xlabel('experiment number')  % *Line 20*
ylabel('observations ')  % *Line 21*
title(['95% CI:',num2str(mx,3), '\pm',num2str(me,3)]), shg  %*Line 22 *
```

Run this code several times, until you get at least one unsuccessful interval (shown as a red asterisk). <u>As a very rough average</u>, you need 20 trials (5% of the time) to get an unsuccessful interval. For each trial, check the sample mean and the margin of error on the title. The sample mean should change every time you run the code, while the margin of error should not.

The margin of error ($z_{cr}\frac{\sigma}{\sqrt{n}}$) is inversely proportional to the square root of the sample size. To halve the margin of error, we need to increase the sample size by 4. Modify the sample size and confirm that margin of error is halved:

```
n=16          % Line 1
```

To see what happens when we repeat the experiment many times, we will use a `for` loop (unless willing to run the code by hand many times). We will start with a relatively small number of repetitions (20). Modify the two lines (Line 5 and Line 19) as follows:

```
for i=1:20 ;   % Line 5
```

```
end            % Line 19
```

Count the number of misses (shown as red *). The number of experiments (20) is still not very big, and the number of misses will depend on our luck (or, what the current "seed" is). On average, we expect to see one unsuccessful interval every time we run the code. Run the code several times to see how the number of misses changes.

If we perform the experiment a large number of times, we expect that 95% of our intervals contain the population mean. With 1000 repetitions, approximately 950 should capture the population mean.

To test this, we will count the number of misses and show it on the title:

```
miss_count=0;   % Line 4
```

```
for i=1:1000;   % Line 5
```

```
miss_count=miss_count+1;   % Line 17
```

```
title([num2str(miss_count),' misses out of 1000']) % Line 22
```

As we increase the number of experiments (and thus the number of confidence intervals), the proportion of misses should get closer to 5%.

The previous example was just an academic exercise explaining the meaning of confidence intervals. In real life, we face two challenges:

- We do not know the population standard deviation. If we knew it, we probably would know the population mean as well, and there wouldn't be any need to construct a confidence interval.
- We do not have the opportunity to repeat the experiment many times, and we have to rely on a single confidence interval, based on a single sample. We will not know if we are lucky or not with the particular confidence interval we computed, we are just confident that if we repeated the experiment many times, a certain percentage of such intervals will contain the population mean.

3.2 Unknown population standard deviation

If we do not know the population standard deviation (σ), we can substitute sample standard deviation (S)

$$\text{Sample standard deviation:} \quad S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

However, when we standardize the sample mean using $\left(\frac{S}{\sqrt{n}}\right)$, the resulting statistic has student-t distribution with (n-1) degrees of freedom.

<u>Gosset's Theorem</u>: Suppose $X_1, X_2, \ldots, X_n$, be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. Let the sample mean and sample standard deviation be $\bar{X}$ and S, respectively. Then the random variable

$$T_{\bar{X}} = \frac{\bar{X} - \mu}{\left(\frac{S}{\sqrt{n}}\right)}$$

has student-t distribution with $(n-1)$ degrees of freedom.

Gosset's theorem holds approximately for samples from non-normal populations, for $n \geq 30$.

A two-sided CI for population mean μ, with unknown σ has the following form:

$$Two\text{-}sided\ CI\ for\ \mu\ when\ \sigma\ is\ unknown: \quad \bar{X} \pm t_{cr}\frac{s}{\sqrt{n}}$$

$$margin\ of\ error: \ t_{cr}\frac{s}{\sqrt{n}}$$

where t_{cr} is the critical value of student-t distribution that depends on the confidence level and the degrees of freedom $(n-1)$. For $(1-\alpha) \times 100\%$ confidence, $t_{cr} = F^{-1}(1-\alpha/2)$ where F is the CDF of the student t-distribution with $(n-1)$ degrees of freedom, and F^{-1} is the inverse of F. The critical t value satisfies $P(-t_{cr} \leq T_{\bar{X}} \leq t_{cr}) = 1-\alpha$. Because student t- distribution is symmetric, the shortest interval is obtained when the upper and lower tail probabilities are the same. That is,

$P(T_{\bar{X}} < -\ t_{cr}) = \alpha/2$ and $P(T_{\bar{X}} > t_{cr}) = \alpha/2$.

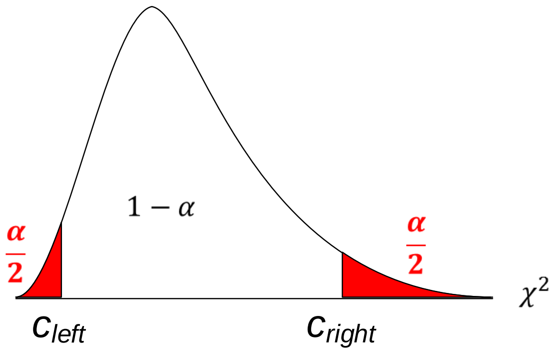
4 Confidence intervals for variance (σ^2) of a normally distributed population

To construct a confidence interval for population variance, we use the following theorem:

Theorem: Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2, and let S^2 be the sample variance. Then the random variable $(n-1)S^2/\sigma^2$ follows a chi-squared $(\varkappa^2)$ distribution with $(n-1)$ degrees of freedom. That is,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Given a sample of size n with s^2, a two-sided, $(1-\alpha) \times 100\%$ **c**onfidence interval for the population variance (σ^2) can be constructed by assigning equal probability to the two tails:



$$\left[\frac{(n-1)s^2}{c_{right}}, \frac{(n-1)s^2}{c_{left}}\right]$$

The critical values $z_{cr},\ t_{cr},\ c_{right}\ and\ c_{left}$ values can be determined using inverse cumulative distribution function, as shown below.

5 Example:

The cross sectional area measurements (in^2) from ten steel specimens are as follows:

$x = [105, 118, 77, 108, 103, 87, 96, 103, 136, 128]$. Assuming that the cross sectional areas of the specimens are normally distributed as $X \sim N(\mu, \sigma^2)$ construct a two-sided 95% confidence interval for the following cases:

(a) Construct an interval for μ, assuming $\sigma = 10$.

(b) Construct an interval for μ, assuming σ is unknown.

(c) Construct an interval for σ^2.

```
x=[105, 118, 77, 108, 103, 87, 96, 103, 136, 128] % measurements
n=length(x);   % sample size
m=mean(x);     s=std(x);
alpha=0.05;  % for 95% confidence level
% ------  part a --------
sigma=10; % population standard deviation
zcr=norminv(1-alpha/2)   % zcr for two-sided CI
me=zcr*sigma/sqrt(n);   % margin of error
lower = m-me;   upper = m+me;
CIa=[lower,upper]
%------  part b --------
tcr=tinv(1-alpha/2, n-1)  % tcr for two-sided CI
me = tcr*s/sqrt(n);
lower = m-me;   upper = m+me;
CIb=[lower,upper]
%------- part c ----------
cright=chi2inv(1-alpha/2,n-1);
cleft=chi2inv(alpha/2,n-1);
lower = (n-1)*s^2/cright;   upper = (n-1)*s^2/cleft;
CIc=[lower,upper]
```