§5.3 Interval Estimation

For a parameter θ in the **pdf** of a random variable, we have an estimate θ_e based on a sample. We consider θ_e as a **point** estimation.

We want to find an **interval** $[\theta_e - d, \theta_e + d]$ with **confidence** $(1-\alpha)100\%$.

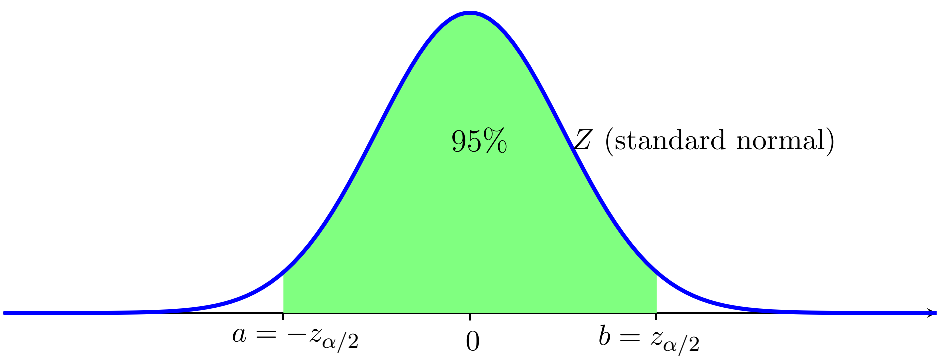
1. (Normal) Suppose $X \sim \text{Normal}(\mu, \sigma^2)$ with known σ. From §5.4, Example 4, we know that the maximum likelihood estimator for μ based on a sample $X_1 = x_1$, $X_2 = x_2$, ... , $X_n = x_n$ is

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

By CLT, we know that $\overline{X} \sim \text{Normal}(\mu, \sigma^2/n)$. Then,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0,1)$$

For example, if we want 95% confidence interval, $\alpha = 0.05$.



By calculator, $-z_{\alpha/2} = \textbf{invNorm}(0.025, 0, 1) \approx -1.96$, or
$z_{\alpha/2} = \textbf{invNorm}(0.975, 0, 1) \approx 1.96$. (We only need one of them)

Then,

$$P(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha = 95\%$$

We solve μ, $\mu = \overline{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$ Hence,

$$P(\overline{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \overline{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)) = 1 - \alpha = 95\%$$

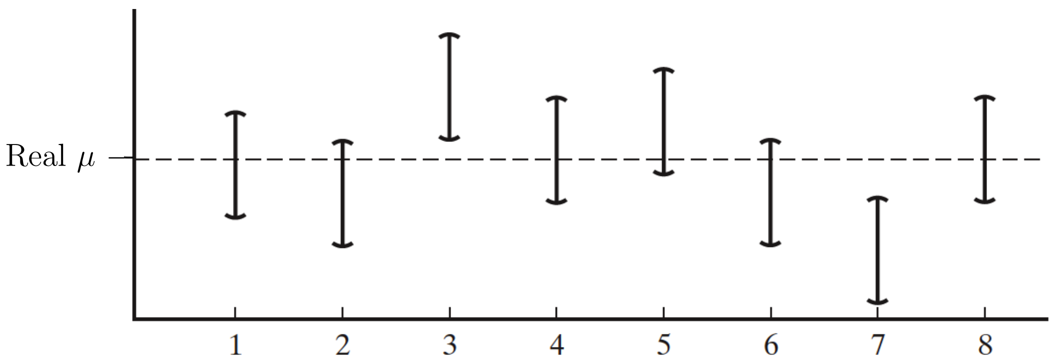
The $100(1-\alpha)\%$ **confidence interval** for μ

$$\overline{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \overline{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Here, $\bar{x}$ is called **basis estimate** and

$$d = \frac{(z_{\alpha/2})\sigma}{\sqrt{n}}$$

is called **margin of error**.



The probability that μ is in the confidence interval is $100(1-\alpha)\%$.

Choosing Sample Size:

In order for $\bar{x}$ to have $100(1-\alpha)\%$ confidence interval of width at most $2d$, the **sample size** n should be no smaller than

$$n = \frac{z_{\alpha/2}^2\sigma^2}{d^2}$$

Example 1. Suppose $X \sim \text{Normal}(\mu, \sigma^2)$ with known $\sigma = 2$. Suppose a sample of size 6 is $\{10.1, 15, 11.7, 14.2, 10, 11\}$ with a sample mean (average) of $\overline{x} = 12$.

(1) Find the 95% confidence interval for μ.

$$\overline{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \overline{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right).$$

Here, $z_{\alpha/2} = \textbf{invNorm}(0.975, 0, 1) \approx 1.96$. So, the 95% confidence interval is

$$[10.4, 13.6]$$

(2) In order for $\bar{x}$ to have 95% confidence interval of width at most 3, how large is the sample size have to be?

The sample size n should be no smaller than

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{d^2}$$

Here, $d = 3/2$. So, $\dfrac{1.96^2(2^2)}{1.5^2} \approx 6.8$. So $n = 7$.

(3) Find the 99% confidence interval for μ.

(Solution: $\alpha = 0.1$, $z_{\alpha/2} = 2.576$, the 99% confidence interval for μ is $[9.9, 14.1]$)

Example 2. An institute wants to estimate the household income in a country. The incomes are normally distributed with standard deviation \$26,000. The institute take a survey of 2416 households randomly. The average household income in the survey is \$56,000.

(1) Find a 95% confidence interval for the average household income in the country.

The 95% confidence interval is

$$\overline{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \overline{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right).$$

Here, $\overline{x} = 56000$, $z_{\alpha/2} = \textbf{invNorm}(0.975, 0, 1) \approx 1.96$, $\sigma = 26000$ and $n = 2416$
So, the 95% confidence interval is

$$[54963, 57037]$$

(2) How large does the sample size have to be to guarantee that the length of the 95% confidence interval for μ will be less than \$1000.

The sample size n should be no smaller than

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{d^2}$$

Here, $d = 1000/2$. So, $\dfrac{1.96^2(26000^2)}{500^2} \approx 10387.7$. So $n = 10388$.

Do the same practice for 99% confidence interval.

To summarize, we need to know three concepts in this section:

1. Confidence Interval; 2. Margin of error; 3. Sample size.

2. (Bernoulli/Proportion/Binomial) Binomial random variable is a sum of IID Bernoulli random variables. Suppose $X \sim Bernoulli(p)$ with unknown p. From §5.2, Example 3, based on a sample of size n: $X_1 = k_1$, $X_2 = k_2$, ... , $X_n = k_n$, ($k_i \in \{0, 1\}$) we calculated the maximum likelihood estimate(MLE)

$$p_e = \frac{k_1 + \cdots + k_n}{n} = \frac{\#\text{ of success}}{n} = \bar{k}$$

and the estimator for p

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

Recall that each $X_i \sim Bernoulli(p)$ with $\mu = E(X_i) = p$ and variance $\text{Var}(X_i) = \sigma^2 = p(1-p)$.

By CLT, we know that $\overline{X} \sim \text{Normal}\left(p, \frac{\bar{k}(1-\bar{k})}{n}\right)$. Then,

$$Z = \frac{\overline{X} - p}{\sqrt{\frac{\bar{k}(1-\bar{k})}{n}}} \sim \text{Normal}(0, 1)$$

Then,

$$P(-z_{\alpha/2} \leq \frac{\overline{X} - p}{\sqrt{\frac{\bar{k}(1-\bar{k})}{n}}} \leq z_{\alpha/2}) = 1 - \alpha$$

Solve p we have

$$p = \bar{k} \pm z_{\alpha/2}\sqrt{\frac{\bar{k}(1-\bar{k})}{n}}$$

Hence,

$$P(\bar{k} - z_{\alpha/2}\sqrt{\frac{\bar{k}(1-\bar{k})}{n}} \leq p \leq \bar{k} + z_{\alpha/2}\sqrt{\frac{\bar{k}(1-\bar{k})}{n}}) = 1 - \alpha$$

The $100(1-\alpha)\%$ **confidence interval** can be computed by

$$\bar{k} - z_{\alpha/2}\sqrt{\frac{\bar{k}(1-\bar{k})}{n}} \leq p \leq \bar{k} + z_{\alpha/2}\sqrt{\frac{\bar{k}(1-\bar{k})}{n}}$$

Theorem.

• The **margin of error** associated to $\bar{k}$ is $100d\%$ with $d = \dfrac{z_{\alpha/2}}{2\sqrt{n}}$.

• In order for $\bar{k}$ to have $100(1-\alpha)\%$ confidence interval of width at most $2d$, the **sample size** should be no smaller than

$$n = \frac{z_{\alpha/2}^2}{4d^2}$$

Proof of the theorem: The margin of error associated to $\bar{k}$ is $100d\%$ with

$$d = \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = \frac{z_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}$$

In order for $\bar{k}$ to have $100(1-\alpha)\%$ confidence interval of width at most d, the sample size should be no smaller than

$$n = \frac{z_{\alpha/2}^2\sigma^2}{d^2} = \frac{z_{\alpha/2}^2 p(1-p)}{d^2}$$

We are not satisfied with the above two formulas, because we don't know p in the formula. But we know that $0 \le p \le 1$. By optimization in Calculus 1, we have $p(1-p) \le (1/4)$.

Example 3. In the example of flipping a coin 10 times with 6 faces, $n = 10$ and $\bar{k} = 6/10$.

(1) Find the 90% confidence interval for p.

The 90% confidence interval for p is calculated by

$$\bar{k} - z_{\alpha/2}\sqrt{\frac{\bar{k}(1-\bar{k})}{n}} \le p \le \bar{k} + z_{\alpha/2}\sqrt{\frac{\bar{k}(1-\bar{k})}{n}}$$

Here, $z_{\alpha/2} = \textbf{invNorm}(0.95, 0, 1) \approx 1.64$,

$$0.6 - 1.64\sqrt{0.6(0.4)/10} \le p \le 0.6 + 1.64\sqrt{0.6(0.4)/10}$$

which is $[0.346, 0.854]$.

(2) The margin of error is associated to $\bar{k}$ is

The margin of error is associated to $\bar{k}$ is $d = \dfrac{z_{\alpha/2}}{2\sqrt{n}} = \dfrac{1.64}{2\sqrt{10}} = 0.26$

(3) In order for $\bar{k}$ to have 90% confidence interval of width at most 0.2, how large does the sample size have to be?

$$n = \frac{z_{\alpha/2}^2}{4d^2} = \frac{1.64^2}{4(0.1)^2} \approx 67.2$$

The sample size should be no smaller than 68.

Example 4. A poll was conducted to find out the percentage of people who will vote A or B for mayor of a city. Out of 500 people polled, 263 said A and the rest said B.

(1) The MLE for p.

The MLE for p is

$$\overline{k} = \frac{263}{500} = 0.526 = 52.6\%$$

" Can we conclude that A is 5 percent head? " No, because this is only for the sample. We want to find the information for the population.

(2) The 95% confidence interval for p.

The 95% confidence interval for p is calculated by

$$\overline{k} - z_{\alpha/2}\sqrt{\frac{\overline{k}(1-\overline{k})}{n}} \leq p \leq \overline{k} + z_{\alpha/2}\sqrt{\frac{\overline{k}(1-\overline{k})}{n}}$$

Here, $z_{\alpha/2} = \textbf{invNorm}(0.975, 0, 1) \approx 1.96$,

$$0.526 - 1.96\sqrt{0.526(0.474)/500} \leq p \leq 0.526 + 1.96\sqrt{0.526(0.474)/500}$$

which is $[0.482, 0.57]$.

(3) The margin of error at the 95% confidence interval for p.

The margin of error is associated to $\bar{k}$ is

$$d = \frac{z_{\alpha/2}}{2\sqrt{n}} = \frac{1.96}{2\sqrt{500}} = 0.04$$

(4) Find the minimal number of people to be polled for error $\leq 2.6\%$.

$$n = \frac{z_{\alpha/2}^2}{4d^2} = \frac{1.96^2}{4(0.026)^2} = 1420.7$$

The sample size should be no smaller than 1421.

Any question about **polling** background are the same.

More interesting examples

Example 5. People use electronic devices every day. Some people even have smartphone addiction.

(1) A health research institute claims that in a survey, 57% college students use smartphone more than 4 hours per day. How many college students must be surveyed in order to be 95% confident that the sample percentage is in error $\leq 2\%$?

To satisfy the conditions, the smallest number of surveyed students is

$$n = \frac{z_{\alpha/2}^2}{4d^2} = 1.96^2/(4(0.02^2)) \approx 2401$$

(2) A smart phone company want to know the percentage of people who will update their phones to a new version. In a survey of 350 users, 189 said that they plan to update their phones.

(i) Find a 95% confidence interval for the population proportion.

The 95% confidence interval for p is calculated by

$$\overline{k} - z_{\alpha/2}\sqrt{\frac{\overline{k}(1-\overline{k})}{n}} \leq p \leq \overline{k} + z_{\alpha/2}\sqrt{\frac{\overline{k}(1-\overline{k})}{n}}$$

Here, $z_{\alpha/2} = \textbf{invNorm}(0.975, 0, 1) \approx 1.96$, $n = 350$, $\overline{k} = 189/350$.
So, the 95% confidence interval for p is

$$[0.4878, 0.5922]$$

(ii) Find the margin of error corresponding to a 95% confidence interval.

The margin of error is associated to $\bar{k}$ is

$$d = \frac{z_{\alpha/2}}{2\sqrt{n}} = \frac{1.96}{2\sqrt{350}} \approx 0.0524$$

Example 6. According to stats.nba.com, in NBA 2019-2020 season, Boston Celtics win 42 games in the first 61 games. Find a 95% confidence interval for Boston Celtics's winning population proportion.

The 95% confidence interval for p is calculated by

$$\overline{k} - z_{\alpha/2}\sqrt{\frac{\overline{k}(1-\overline{k})}{n}} \leq p \leq \overline{k} + z_{\alpha/2}\sqrt{\frac{\overline{k}(1-\overline{k})}{n}}$$

Here, $z_{\alpha/2} = \textbf{invNorm}(0.975, 0, 1) \approx 1.96$, $\overline{k} = 42/61$, and $n = 61$.
So, the 95% confidence interval for p is

$$[0.5723, 0.805]$$

Example 7. According to stats.nba.com, in NBA 2019-2020 season, Boston Celtics got an average 113.4 points in the first 61 games. Suppose the number of points is normally distributed with population standard deviation $\sigma = 10.5$. Find a 95% confidence interval for the average number of points for Boston Celtics.

The 95% confidence interval is

$$\overline{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \overline{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right).$$

Here, $\overline{x} = 113.4$, $z_{\alpha/2} = \textbf{invNorm}(0.975, 0, 1) \approx 1.96$, $\sigma = 10.5$ and $n = 61$.
The 95% confidence interval is

$$[110.8077, 116.0776]$$

The R computation from the raw data is in the Celtics-R lab.

Example 8. According to www.bostonglobe.com, Massachusetts average commute time increases to fifth longest in US in 2019. US Census make a survey of 1220 people to estimate the commute time in MA. They found a mean of 29.5 minutes with a standard deviation of σ =16.3 minutes. Find a 95% confidence interval for the average commuting time in MA.

The 95% confidence interval is

$$\overline{x} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le \mu \le \overline{x} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right).$$

Here, $\overline{x} = 29.5$, $z_{\alpha/2} = \textbf{invNorm}(0.975, 0, 1) \approx 1.96$, $\sigma = 16.3$ and $n = 1220$.
The 95% confidence interval is

$$[28.58535, 30.41465]$$