ST2132 Mathematical Statistics Confidence intervals

Lecture 5

Lecturer: Michael Choi *Scribe: Michael Choi*

1 Goal of this lecture

In this lecture we will begin our discussion on confidence intervals, one popular approach in interval estimation.
Suggested reading: Chapter 7.1, 7.2 in the book.

2 Confidence interval

For $\alpha \in [0,1]$, a $100(1-\alpha)\%$ confidence interval for parameter θ is an interval (that depends on a random sample $X_1,\ldots,X_n$) such that the probability that θ is in the interval is $1-\alpha$.

3 (Two-sided) Confidence intervals for estimating μ

Suppose that $X_1,\ldots,X_n$ are i.i.d. random samples of size n. We are interested in estimating the mean $\mu = E(X_1)$. We will discuss 4 cases:

- Case 1: $X_1,\ldots,X_n \overset{i.i.d.}{\sim} N(\mu,\sigma^2)$ and σ^2 is known
- Case 2: $X_1,\ldots,X_n$ are i.i.d. with mean $\mu = E(X_1)$ and variance $\sigma^2 = \mathrm{Var}(X_1)$, and σ^2 is known
- Case 3: $X_1,\ldots,X_n \overset{i.i.d.}{\sim} N(\mu,\sigma^2)$ and σ^2 is unknown
- Case 4: $X_1,\ldots,X_n$ are i.i.d. with mean $\mu = E(X_1)$ and variance $\sigma^2 = \mathrm{Var}(X_1)$, and σ^2 is unknown

3.1 Case 1: $X_1,\ldots,X_n \overset{i.i.d.}{\sim} N(\mu,\sigma^2)$ and σ^2 is known

Let

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

Then

$$P\left(-z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1-\alpha$$
$$P\left(-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le \overline{X}-\mu \le z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1-\alpha$$
$$P\left(-\overline{X}-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le -\mu \le -\overline{X}+z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1-\alpha$$
$$P\left(\overline{X}+z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \ge \mu \ge \overline{X}-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1-\alpha.$$

As a result, the probability that the random interval

$$\left[\overline{X}-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right), \overline{X}+z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

includes μ is $1-\alpha$.

Once the sample is observed, we can compute the sample mean $\overline{x}$. The computed interval $\overline{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$ is called a $100(1-\alpha)\%$ **confidence interval (C.I.)** of μ.

Note that the C.I. of μ is

- The C.I. is centered at $\overline{x}$, a point estimate of μ, and the C.I. has a width of $2z_{\alpha/2}(\sigma/\sqrt{n})$.
- The bigger the sample size n, the smaller the width of the C.I., resulting in a shorter C.I..
- The bigger the α, the smaller the z_α, resulting in a shorter C.I..

Example 1 *This is Example 7.1-1 in book. Let X be the length of life of a light bulb, and assume that $X \sim N(\mu, 1296)$. We have a random sample of $n = 27$ with $\overline{x} = 1478$. A 95% confidence interval for μ is*

$$\left[\overline{x}-z_{0.025}\left(\frac{\sigma}{\sqrt{n}}\right), \overline{x}+z_{0.025}\left(\frac{\sigma}{\sqrt{n}}\right)\right] = \left[1478-1.96\left(\frac{36}{\sqrt{27}}\right), 1478+1.96\left(\frac{36}{\sqrt{27}}\right)\right]$$
$$= [1464.42, 1491.58].$$

3.2 Case 2: $X_1, \ldots, X_n$ are i.i.d. with mean $\mu = E(X_1)$ and variance $\sigma^2 = \mathrm{Var}(X_1)$, and σ^2 is known

Note that $X_1, \ldots, X_n$ do not necessarily follow a normal distribution.

When n is large enough, by central limit theorem,

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \overset{\text{approx}}{\sim} N(0,1).$$

As a result, approximately we have

$$P\left(-z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) \approx 1-\alpha$$
$$P\left(\overline{X}-z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le \mu \le \overline{X}+z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right) \approx 1-\alpha.$$

$\overline{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$ is called an **approximate** $100(1-\alpha)\%$ **confidence interval (C.I.)** of μ.

Example 2 *This is Example 7.1-3 in book. Let X be the amount of orange juice consumed by an American per day. We know that $\sigma = 96$. To estimate μ, we have a sample of $n = 576$ and $\overline{x} = 133$. An approximate 90% confidence interval for μ is*

$$133 \pm 1.645 \left(\frac{96}{\sqrt{576}} \right) = [126.42, 139.58].$$

3.3 Case 3: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ and σ^2 is unknown

As σ is unknown, we cannot use the C.I. in Case 1 since the interval

$$\overline{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$$

depends on σ.

Instead, recalling the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2,$$

and

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

Essentially, we replace σ by S and we "lose" one degree of freedom. Consider

$$P\left(-t_{\alpha/2}(n-1) \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2}(n-1) \right) = 1 - \alpha$$
$$P\left(\overline{X} - t_{\alpha/2}(n-1) \left(\frac{S}{\sqrt{n}} \right) \leq \mu \leq \overline{X} + t_{\alpha/2}(n-1) \left(\frac{S}{\sqrt{n}} \right) \right) = 1 - \alpha.$$

Once we have the sample, we can compute $\overline{x}$ and s, and

$$\left[\overline{x} - t_{\alpha/2}(n-1) \left(\frac{s}{\sqrt{n}} \right), \overline{x} + t_{\alpha/2}(n-1) \left(\frac{s}{\sqrt{n}} \right) \right]$$

is called a $100(1-\alpha)\%$ **confidence interval (C.I.)** of μ.

Example 3 *This is Example 7.1-5 in book. Let $X \sim N(\mu, \sigma^2)$, and both μ and σ^2 are unknown. We compute that $n = 20$, $\overline{x} = 507.50$ and $s = 89.75$. Since $t_{0.05}(19) = 1.729$, a 90% confidence interval for μ is*

$$507.50 \pm 1.729 \left(\frac{89.75}{\sqrt{20}} \right) = [472.80, 542.20].$$

3.4 Case 4: $X_1, \ldots, X_n$ are i.i.d. with mean $\mu = E(X_1)$ and variance $\sigma^2 = \mathrm{Var}(X_1)$, and σ^2 is unknown

If n is large enough (say $n \geq 30$), or if each X_i is approximately normal, then

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \overset{\text{approx}}{\sim} t(n-1).$$

As a result,

$$\left[\overline{x} - t_{\alpha/2}(n-1)\left(\frac{s}{\sqrt{n}}\right), \overline{x} + t_{\alpha/2}(n-1)\left(\frac{s}{\sqrt{n}}\right)\right]$$

is called an **approximate** $100(1-\alpha)\%$ **confidence interval (C.I.)** of μ. Note that

- If n is large enough, then $t_{\alpha/2}(n-1) \approx z_{\alpha/2}$, and so we can also use the interval

  $$\left[\overline{x} - z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right), \overline{x} + z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)\right]$$

  as an approximate $100(1-\alpha)\%$ C.I. for μ.

- If n is small or when each X_i does not look like a normal distribution (say when X_i are very skewed), this method has to be used with caution.

4 One-sided confidence intervals for estimating μ

Up until now we have only been discussing **two-sided** confidence intervals, since the C.I. are often of the form $\overline{x} \pm a$ for some constants a that depends on α, n and the samples. Suppose that $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ and σ^2 is known. Consider

$$P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_\alpha\right) = 1 - \alpha,$$

$$P\left(\overline{X} - z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu\right) = 1 - \alpha.$$

As a result, we have

$$\left[\overline{x} - z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right), \infty\right)$$

an **one-sided** $100(1-\alpha)\%$ **confidence interval** for μ. Similarly, we can show that

$$\left(-\infty, \overline{x} + z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

is another one-sided $100(1-\alpha)\%$ confidence interval for μ.

5 Confidence intervals for the difference of two means

Suppose that we have two populations from which we draw i.i.d. random samples, say $X_1, \ldots, X_n \sim N(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_m \sim N(\mu_Y, \sigma_Y^2)$. We would like to develop confidence intervals for $\mu_X - \mu_Y$, the difference between the two population means.

We will discuss 3 cases:

1. Case 1: $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are independent
   (a) $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, and σ^2 is unknown (Two populations have the same variance σ^2)
   (b) $\sigma_X^2 \neq \sigma_Y^2$, and they are unknown
2. Case 2: $n = m$ and X_i, Y_i are dependent for $i = 1, \ldots, n$. However, the n pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent.

5.1 Case 1(a): Two-sample pooled t-interval

Theorem 1 *Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_X, \sigma^2)$ and $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} N(\mu_Y, \sigma^2)$ be independent random variables. Then a $100(1-\alpha)\%$ C.I. for $\mu_X - \mu_Y$ is*

$$\overline{X} - \overline{Y} \pm t_{\alpha/2}(n+m-2) S_p \sqrt{\frac{1}{n} + \frac{1}{m}},$$

where S_p^2 is the pooled estimator of σ^2:

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2 + \sum_{i=1}^m (Y_i - \overline{Y})^2}{n+m-2},$$

which is an unbiased estimator of σ^2.

Proof: Since X_i and Y_i are independent, then their sample means

$$\overline{X} \sim N\left(\mu_X, \frac{\sigma^2}{n}\right)$$
$$\overline{Y} \sim N\left(\mu_Y, \frac{\sigma^2}{m}\right)$$

are independent, and so

$$\overline{X} - \overline{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$$
$$Z = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma^2/n + \sigma^2/m}} \sim N(0,1).$$

Recall that in Lecture 1 and 2 we have recalled that

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2(n-1),$$
$$\frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2(m-1),$$

and since they are independent,

$$U = \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2(n+m-2).$$

Since $E(U) = n+m-2$,

$$E(S_p^2) = \frac{\sigma^2}{n+m-2} E(U) = \sigma^2,$$

S_p^2 is an unbiased estimator of σ^2. Since Z and U are independent, recall the definition of t distribution that

$$T = \frac{Z}{\sqrt{U/(n+m-2)}} \sim t(n+m-2).$$

Note that

$$\begin{aligned} T &= \frac{\frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{\sqrt{\sigma^2/n+\sigma^2/m}}}{\sqrt{\left[\frac{(n-1)S_X^2}{\sigma^2}+\frac{(m-1)S_Y^2}{\sigma^2}\right]/(n+m-2)}} \\ &= \frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{S_p \cdot \sqrt{\frac{1}{n}+\frac{1}{m}}}. \end{aligned}$$

As a result,

$$P\left(-t_{\alpha/2}(n+m-2) \le \frac{\overline{X}-\overline{Y}-(\mu_X-\mu_Y)}{S_p \cdot \sqrt{\frac{1}{n}+\frac{1}{m}}} \le t_{\alpha/2}(n+m-2)\right) = 1-\alpha$$

Rearranging we have

$$P\left(\overline{X}-\overline{Y}-t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}} \le \mu_X-\mu_Y \le \overline{X}-\overline{Y}+t_{\alpha/2}(n+m-2)S_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right) = 1-\alpha.$$

□

Example 4 *This is Example 7.2-2 in book. Let $X \sim N(\mu_X, \sigma^2)$ be the score on a standardized test in a large high school, and $Y \sim N(\mu_Y, \sigma^2)$ be the score on a standardized test in a small high school. We have*

$$\begin{aligned} n &= 9 \\ \overline{x} &= 81.31 \\ s_x^2 &= 60.76 \\ m &= 15 \\ \overline{y} &= 78.61 \\ s_y^2 &= 48.24. \end{aligned}$$

To construct a 95% confidence interval for $\mu_X - \mu_Y$, we calculate

$$t_{0.025}(22) = 2.074$$
$$s_p = \sqrt{\frac{8(60.76) + 14(48.24)}{22}}$$

The required 95% confidence interval for $\mu_X - \mu_Y$ is

$$81.31 - 78.61 \pm 2.074\sqrt{\frac{8(60.76) + 14(48.24)}{22}}\sqrt{\frac{1}{9} + \frac{1}{15}} = [-3.65, 9.05].$$