Chapter 4: Interval estimation

STAT2602A Probability and statistics II
(2024-2025 1st Semester)

Contents

4.1 Basic concepts

- **Motivation:** A point estimator for θ does not provide much information about the accuracy of the estimator. It is desirable to generate a narrow interval that will cover the unknown parameter θ with a large probability (confidence).
- **Definition 4.1.** (Interval estimator)An interval estimator of θ is a random interval $[L(\mathbf{X}), U(\mathbf{X})]$, where $L(\mathbf{X}) := L(X_1, \cdots, X_n)$ and $U(\mathbf{X}) := U(X_1, \cdots, X_n)$ are two statistics such that $L(\mathbf{X}) \leq U(\mathbf{X})$ with probability one.
- **Definition 4.2.** (Interval estimate) If $\mathbf{X} = \mathbf{x}$ is observed, $[L(\mathbf{x}), U(\mathbf{x})]$ is the interval estimate of θ.
- Remark: *it will sometimes be more natural to use an open interval* $(L(\mathbf{X}), U(\mathbf{X}))$*, a half-open and half-closed interval* $(L(\mathbf{X}), U(\mathbf{X})]$ *(or* $[L(\mathbf{X}), U(\mathbf{X}))$*), or an one-sided interval* $(\infty, U(\mathbf{X})]$ *(or* $[L(\mathbf{X}), \infty)$*).*

4.1 Basic concepts

Example 4.1. For an independent random sample X_1, X_2, X_3, X_4 from $\mathrm{N}(\mu, 1)$, consider an interval estimator of μ by $[\overline{X} - 1, \overline{X} + 1]$. Then, the probability that μ is covered by the interval $[\overline{X} - 1, \overline{X} + 1]$ can be calculated by

$$\begin{aligned}
\mathrm{P}\left(\mu \in [\overline{X} - 1, \overline{X} + 1]\right) &= \mathrm{P}\left(\overline{X} - 1 \leq \mu \leq \overline{X} + 1\right) \\
&= \mathrm{P}\left(-2 \leq (\overline{X} - \mu)/\sqrt{1/4} \leq 2\right) \\
&= \mathrm{P}\left(-2 \leq Z \leq 2\right) \\
&\approx 0.9544,
\end{aligned}$$

where $Z \sim \mathrm{N}(0, 1)$ and with the fact that $\overline{X} \sim \mathrm{N}(\mu, 1/4)$. Thus, we have over a 95% change of covering the unknown parameter with our interval estimator. Note that for any point estimator $\hat{\mu}$ of μ, we have $\mathrm{P}(\hat{\mu} = \mu) = 0$. Sacrificing some precision in the interval estimator, in moving from a point to an interval, has resulted in increased confidence that our assertion about μ is correct. □

4.1 Basic concepts

- **Definition 4.3.** (*Confidence coefficient*) For an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of θ, the confidence coefficient of $[L(\mathbf{X}), U(\mathbf{X})]$, denoted by $(1-\alpha)$, is

$$1-\alpha = \mathrm{P}(\theta \in [L(\mathbf{X}), U(\mathbf{X})]),$$

  where $\mathrm{P}(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$ is the coverage probability of $[L(\mathbf{X}), U(\mathbf{X})]$.

- Remark: In some situations, the coverage probability $\mathrm{P}(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$ may depend on θ, and then the confidence coefficient is defined as

$$1-\alpha = \inf_{\theta} \mathrm{P}(\theta \in [L(\mathbf{X}), U(\mathbf{X})]).$$

- Interval estimator, together with a measure of confidence (say, the confidence coefficient), is sometimes known as **confidence interval**. A confidence interval with confidence coefficient equal to $1-\alpha$, is called a $1-\alpha$ confidence interval.

4.1 Basic concepts

- **Definition 4.4.** (*Pivotal Quantity*) A random variable $Q(\mathbf{X}, \theta) = Q(X_1, \cdots, X_n, \theta)$ is a pivotal quantity if the distribution of $Q(\mathbf{X}, \theta)$ is free of θ. That is, regardless of the distribution of $\mathbf{X}$, $Q(\mathbf{X}, \theta)$ has the same distribution for all values of θ.
- Remark: Logically, when $Q(\mathbf{X}, \theta)$ is a pivotal quantity, we can easily construct a $1 - \alpha$ confidence interval for $Q(\mathbf{X}, \theta)$ by

$$1 - \alpha = \mathrm{P}\left(\widetilde{L}_\alpha \leq Q(\mathbf{X}, \theta) \leq \widetilde{U}_\alpha\right), \tag{1.1}$$

where $\widetilde{L}_\alpha$ and $\widetilde{U}_\alpha$ do not depend on θ. Suppose that the inequalities $\widetilde{L}_\alpha \leq Q(\mathbf{X}, \theta) \leq \widetilde{U}_\alpha$ in (1.1) are equivalent to the inequalities $L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})$, Then, from (1.1), a $1 - \alpha$ confidence interval of θ is $[L(\mathbf{X}), U(\mathbf{X})]$.

4.2 Confidence intervals for means - One-sample case

Formula derivation Let $\mathbf{X} = \{X_1, \cdots, X_n\}$ be an independent random sample from the population $\mathrm{N}(\mu, \sigma^2)$. We first consider the interval estimator of μ when σ^2 is known. Note that

$$\overline{X} \sim \mathrm{N}(\mu, \sigma^2/n). \tag{2.1}$$

Hence, when σ^2 is known, $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$ is a pivotal quantity involving μ. Let

$$\begin{aligned} 1-\alpha &= \mathrm{P}\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) \\ &= \mathrm{P}\left(-z_{\alpha/2} \le \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) \\ &= \mathrm{P}\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right), \end{aligned}$$

where z_α satisfies

$$\mathrm{P}(Z \ge z_\alpha) = \alpha$$

for $Z \sim \mathrm{N}(0,1)$

4.2 Confidence intervals for means - One-sample case

Formula derivation (con't) Usually, we call z_α the upper percentile of N(0, 1) at the level α. When σ^2 is known, a $1-\alpha$ confidence interval of μ is

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]. \tag{2.2}$$

Given the observed value of $\overline{X} = \overline{x}$ and the value of $z_{\alpha/2}$, we can calculate the interval estimate of μ by

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

As the point estimator, the $1-\alpha$ confidence interval is also not unique. Ideally, we should choose it as narrow as possible in some sense, but in practice, we usually choose the equal-tail confidence interval as in (2.2) for convenience, since tables for selecting equal probabilities in the two tails are readily available.

4.2 Confidence intervals for means - One-sample case

Example 4.2. A publishing company has just published a new college textbook. Before the company decides the price of the book, it wants to know the average price of all such textbooks in the market. The research department at the company took a sample of 36 such textbooks and collected information on their prices. This information produced a mean price of \$48.40 for this sample. It is known that the standard deviation of the prices of all such textbooks is \$4.50. Construct a 90% confidence interval for the mean price of all such college textbooks assuming that the underlying population is normal.

Solution. From the given information, $n = 36$, $\overline{x} = 48.40$ and $\sigma = 4.50$. Now, $1 - \alpha = 0.9$, i.e., $\alpha = 0.1$, and by (2.2), the 90% confidence interval for the mean price of all such college textbooks is given by

$$[\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}] = [48.40 - z_{0.05}\frac{4.50}{\sqrt{36}}, 48.40 + z_{0.05}\frac{4.50}{\sqrt{36}}]$$

$$\approx [47.1662, 49.6338].$$

4.2 Confidence intervals for means - One-sample case

Example 4.3. Suppose the bureau of the census and statistics of a city wants to estimate the mean family annual income μ for all families in the city. It is known that the standard deviation σ for the family annual income is 60 thousand dollars. How large a sample should the bureau select so that it can assert with probability 0.99 that the sample mean will differ from μ by no more than 5 thousand dollars?

Solution. From the construction of a confidence interval, we have

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

where $1-\alpha = 0.99$ and $\sigma = 60$ thousand dollars. It suffices to have $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le 5$ or

$$n \ge \left(60 z_{\alpha/2}/5\right)^2 = (60 \times 2.576/5)^2 \approx 955.5517.$$

Thus, the sample size should be at least 956. (Note that we have to round 955.5517 up to the next higher integer. This is always the case when determining the sample size.)

4.2 Confidence intervals for means - One-sample case

Case: consider the interval estimator of μ when σ^2 is unknown.

- **Property 4.1.** (i) $\overline{X}$ and S^2 are independent;

  (ii) $\dfrac{nS^2}{\sigma^2} = \dfrac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sigma^2}$ is χ^2_{n-1}, where χ^2_k is a chi-square distribution with k degrees of freedom;

  (iii) $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n-1}}$ is t_{n-1}, where t_k is a t distribution with k degrees of freedom.

- **Property 4.2.** (i) If $Z_1, \cdots, Z_k$ are k independent $\mathrm{N}(0,1)$ random variables, then $\sum_{i=1}^{k} Z_i^2$ is χ^2_k;

  (ii) If Z is $\mathrm{N}(0,1)$, U is χ^2_k, and Z and U are independent, then $T = \dfrac{Z}{\sqrt{U/k}}$ is t_k.

The proof of Property 4.1. could be found in the lecture notes.

4.2 Confidence intervals for means - One-sample case

Derivation of the interval estimator of μ when σ^2 is unknown.
From Property 4.1(iii), we know that T is a pivotal quantity of μ.
Let

$$\begin{aligned}
1-\alpha &= \mathrm{P}\left(-t_{\alpha/2,df=n-1} \leq T \leq t_{\alpha/2,df=n-1}\right) \\
&= \mathrm{P}\left(-t_{\alpha/2,df=n-1} \leq \frac{\overline{X}-\mu}{S/\sqrt{n-1}} \leq t_{\alpha/2,df=n-1}\right) \\
&= \mathrm{P}\left(\overline{X} - t_{\alpha/2,df=n-1}\frac{S}{\sqrt{n-1}} \leq \mu \leq \overline{X} + t_{\alpha/2,df=n-1}\frac{S}{\sqrt{n-1}}\right)
\end{aligned}$$

where $t_{\alpha,df=k}$ satisfies

$$\mathrm{P}(T \geq t_{\alpha,df=k}) = \alpha$$

for a random variable $T \sim t_k$.

4.2 Confidence intervals for means - One-sample case

Derivation con't

Therefore, when σ^2 is unknown, a $1-\alpha$ confidence interval of μ is

$$\left[\overline{X} - t_{\alpha/2,df=n-1}\frac{S}{\sqrt{n-1}}, \overline{X} + t_{\alpha/2,df=n-1}\frac{S}{\sqrt{n-1}}\right]. \qquad (2.3)$$

Given the observed value of $\overline{X} = \overline{x}$, $S = s$, and the value of $t_{\alpha/2,df=n-1}$, we can calculate the interval estimate of μ by

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

- Remark: Usually there is a row with ∞ degrees of freedom in a t-distribution table, which actually shows values of z_α. In fact, when $n \to \infty$, the distribution function of t_n tends to that of $\mathrm{N}(0,1)$. That is, in tests or exams, if n is so large that the value of $t_{\alpha,df=n}$ cannot be found, you may use z_α instead.

4.2 Confidence intervals for means - One-sample case

Example 4.4 A paint manufacturer wants to determine the average drying time of a new brand of interior wall paint. If for 12 test areas of equal size he obtained a mean drying time of 66.3 minutes and a standard deviation of 8.4 minutes, construct a 95% confidence interval for the true population mean assuming normality.

Solution. As $n = 12$, $\overline{x} = 66.3$, $s = 8.4$, $\alpha = 1 - 0.95 = 0.05$ and $t_{\alpha/2, df=n-1} = t_{0.025,11} \approx 2.201$, the 95% confidence interval for μ is

$$\left[66.3 - 2.201 \times \frac{8.4}{\sqrt{12-1}},\ 66.3 + 2.201 \times \frac{8.4}{\sqrt{12-1}}\right],$$

that is, $[61.1722, 71.4278]$. □

4.2 Confidence intervals for means - One-sample case

Example 4.5 Construct a 95% confidence interval for the mean hourly wage of apprentice geologists employed by the top 5 oil companies. For a sample of 50 apprentice geologists, $\overline{x} = 14.75$ and $s = 3.0$ (in dollars).

Solution. As $n = 50$, $\alpha = 1 - 0.95 = 0.05$, $t_{0.025, df=49} \approx 2.010$, we have

$$t_{\alpha/2, df=n-1} \frac{s}{\sqrt{n-1}} \approx 2.010 \times \frac{3.0}{\sqrt{50-1}} = 0.8614.$$

Thus, the 95% confidence interval is $[14.75 - 0.86,\ 14.75 + 0.86]$, or $[13.89, 15.59]$. □

4.2 Confidence intervals for means - Two-sample case

In the next part, we shall consider the problem of constructing confidence intervals for the difference of the means of two normal distributions when the variances are unknown.

Formula derivation

Let $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ and $\mathbf{Y} = \{Y_1, Y_2, \cdots, Y_m\}$ be random samples from independent distributions $\mathrm{N}(\mu_X, \sigma_X^2)$ and $\mathrm{N}(\mu_Y, \sigma_Y^2)$, respectively. We are of interest to construct the confidence interval for $\mu_X - \mu_Y$ when $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

First, we can show that

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2/n + \sigma^2/m}}$$

is $\mathrm{N}(0, 1)$. Also, by the independence of $\mathbf{X}$ and $\mathbf{Y}$, from Property 4.1(ii), we know that

$$U = \frac{nS_X^2}{\sigma^2} + \frac{mS_Y^2}{\sigma^2}$$

is χ^2_{n+m-2}.

4.2 Confidence intervals for means - Two-sample case

Formula derivation con't

Moreover, by Property 4.1(i), Z and U are independent. Hence,

$$\begin{aligned} T &= \frac{Z}{\sqrt{U/(n+m-2)}} \\ &= \frac{[(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)]/\sqrt{\sigma^2/n+\sigma^2/m}}{\sqrt{(nS_X^2+mS_Y^2)/[\sigma^2(n+m-2)]}} \\ &= \frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{R} \end{aligned}$$

is t_{n+m-2}, where

$$R = \sqrt{\frac{nS_X^2+mS_Y^2}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)}.$$

That is, T is a pivotal quantity of $\mu_X-\mu_Y$.

4.2 Confidence intervals for means - Two-sample case

Formula derivation con't

Let

$$\begin{aligned} 1-\alpha &= \mathrm{P}\left(-t_{\alpha/2,df=n+m-2} \leq T \leq t_{\alpha/2,df=n+m-2}\right) \\ &= \mathrm{P}\left(-t \leq \frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{R} \leq t\right) \\ &= \mathrm{P}\left((\overline{X}-\overline{Y})-tR \leq \mu_X-\mu_Y \leq (\overline{X}-\overline{Y})+tR\right), \end{aligned}$$

So, when $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ is unknown, a $1-\alpha$ confidence interval of $\mu_X - \mu_Y$ is

$$\left[(\overline{X}-\overline{Y}) - t_{\alpha/2,df=n+m-2}R, (\overline{X}-\overline{Y}) + t_{\alpha/2,df=n+m-2}R\right]. \quad (2.4)$$

4.2 Confidence intervals for means - Two-sample case

Formula derivation con't

Given the observed value of $\overline{X} = \overline{x}$, $\overline{Y} = \overline{y}$, $S_X = s_X$, $S_Y = s_Y$, and the value of $t_{\alpha/2, df=n+m-2}$, we can calculate the interval estimate of $\mu_X - \mu_Y$ by

$$\left[(\overline{x} - \overline{y}) - t_{\alpha/2, df=n+m-2} r, (\overline{x} - \overline{y}) + t_{\alpha/2, df=n+m-2} r\right].$$

where

$$r = \sqrt{\frac{n s_X^2 + m s_Y^2}{n + m - 2} \left(\frac{1}{n} + \frac{1}{m}\right)}.$$

4.2 Confidence intervals for means - Two-sample case

Example 4.6 Suppose that scores on a standardized test in mathematics taken by students from large and small high schools are $\mathrm{N}(\mu_X, \sigma^2)$ and $\mathrm{N}(\mu_Y, \sigma^2)$, respectively, where σ^2 is unknown. If a random sample of $n = 9$ students from large high schools yielded $\bar{x} = 81.31$, $s_X^2 = 60.76$ and a random sample of $m = 15$ students from small high schools yielded $\bar{y} = 78.61$, $s_Y^2 = 48.24$, the endpoints for a 95% confidence interval for $\mu_X - \mu_Y$ are given by

$$81.31 - 78.61 \pm 2.074 \sqrt{\frac{9 \times 60.76 + 15 \times 48.24}{22} \left(\frac{1}{9} + \frac{1}{15} \right)},$$

since $\mathrm{P}(T \leq 2.074) = 0.975$. So, the 95% confidence interval is $[-3.95, 9.35]$. □

4.3 Confidence intervals for variances - One-sample case

Formula derivation First, we consider the one-sample case. By Property 4.1(ii),

$$\frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$$

is a pivotal quantity involving σ^2. Let

$$\begin{aligned} 1-\alpha &= \mathrm{P}\left(\chi^2_{1-\alpha/2,df=n-1} \leq \frac{nS^2}{\sigma^2} \leq \chi^2_{\alpha/2,df=n-1}\right) \\ &= \mathrm{P}\left(\frac{nS^2}{\chi^2_{\alpha/2,df=n-1}} \leq \sigma^2 \leq \frac{nS^2}{\chi^2_{1-\alpha/2,df=n-1}}\right), \end{aligned}$$

where $\chi^2_{\alpha,df=n}$ satisfies

$$\mathrm{P}(T \geq \chi^2_{\alpha,df=n}) = \alpha$$

for a random variable $T \sim \chi^2_n$.

4.3 Confidence intervals for variances - One-sample case

Formula derivation con't

Therefore, a $1-\alpha$ confidence interval of σ^2 is

$$\left[\frac{nS^2}{\chi^2_{\alpha/2,df=n-1}},\ \frac{nS^2}{\chi^2_{1-\alpha/2,df=n-1}}\right]. \tag{3.1}$$

Given the observed value of $S=s$ and the values of $\chi^2_{\alpha/2,df=n-1}$ and $\chi^2_{1-\alpha/2,df=n-1}$, we can calculate the interval estimate of σ^2 by

$$\left[\frac{ns^2}{\chi^2_{\alpha/2,df=n-1}},\ \frac{ns^2}{\chi^2_{1-\alpha/2,df=n-1}}\right].$$

4.3 Confidence intervals for variances - One-sample case

Example 4.7. A machine is set up to fill packages of cookies. A recently taken random sample of the weights of 25 packages from the production line gave a variance of 2.9 g^2. Construct a 95% confidence interval for the standard deviation of the weight of a randomly selected package from the production line.

Solution. As $n = 25$, $s^2 = 2.9$, $\alpha = 0.05$,

$$\frac{ns^2}{\chi^2_{\alpha/2,df=n-1}} = \frac{25(2.9)}{\chi^2_{0.025,df=24}} \approx \frac{25(2.9)}{39.36} \approx 1.8420,$$

$$\frac{ns^2}{\chi^2_{1-\alpha/2,df=n-1}} = \frac{25(2.9)}{\chi^2_{0.975,df=24}} \approx \frac{25(2.9)}{12.40} \approx 5.8468,$$

the 95% confidence interval for the population variance is $(1.8420,\ 5.8468)$. Taking positive square roots, we obtain the 95% confidence interval for the population standard deviation to be $(1.3572,\ 2.4180)$. □

4.3 Confidence intervals for variances - Two-sample case

- **Basic settings:** Consider the two-sample case. Let

$$\mathbf{X} = \{X_1, X_2, \cdots, X_n\} \text{ and } \mathbf{Y} = \{Y_1, Y_2, \cdots, Y_m\}$$

be random samples from independent distributions $\mathrm{N}(\mu_X, \sigma_X^2)$ and $\mathrm{N}(\mu_Y, \sigma_Y^2)$, respectively. We are of interest to construct the confidence interval for σ_X^2/σ_Y^2.

- **Property 4.3.** Suppose that $U \sim \chi^2_{r_1}$ and $V \sim \chi^2_{r_2}$ are independent. Then,

$$F_{r_1,r_2} = \frac{U/r_1}{V/r_2}$$

has an F_{r_1,r_2} distribution with r_1 and r_2 degrees of freedom.

4.3 Confidence intervals for variances - Two-sample case

Formula derivation

By Property 4.1(ii),

$$\frac{nS_X^2}{\sigma_X^2} \sim \chi_{n-1}^2 \quad \text{and} \quad \frac{mS_Y^2}{\sigma_Y^2} \sim \chi_{m-1}^2.$$

Then, by Property 4.3, it follows that

$$\left[\frac{mS_Y^2}{\sigma_Y^2(m-1)}\right] \Big/ \left[\frac{nS_X^2}{\sigma_X^2(n-1)}\right] \sim F_{m-1,n-1},$$

which is a pivotal quantity involving σ_X^2/σ_Y^2.

4.3 Confidence intervals for variances - Two-sample case

Formula derivation con't

Let

$$1-\alpha = \mathrm{P}\left(F_{1-\alpha/2,df=(m-1,n-1)} \leq \frac{\left[\frac{mS_Y^2}{\sigma_Y^2(m-1)}\right]}{\left[\frac{nS_X^2}{\sigma_X^2(n-1)}\right]} \leq F_{\alpha/2,df=(m-1,n-1)} \right)$$

$$= \mathrm{P}\left(\frac{n(m-1)S_X^2}{m(n-1)S_Y^2} F \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq \frac{n(m-1)S_X^2}{m(n-1)S_Y^2} F \right),$$

where $F_{\alpha,df=(m,n)}$ satisfies

$$\mathrm{P}(T \geq F_{\alpha,df=(m,n)}) = \alpha$$

for a random variable $T \sim F_{m,n}$.

4.3 Confidence intervals for variances - Two-sample case

Formula derivation con't

Therefore, a $1-\alpha$ confidence interval of σ_X^2/σ_Y^2 is

$$\left[\frac{n(m-1)S_X^2}{m(n-1)S_Y^2}F_{1-\alpha/2,df=(m-1,n-1)},\ \frac{n(m-1)S_X^2}{m(n-1)S_Y^2}F_{\alpha/2,df=(m-1,n-1)}\right]. \tag{3.2}$$

Given the observed value of $S_X = s_X$, $S_Y = s_Y$, and the values of $F_{\alpha/2,df=(m-1,n-1)}$ and $F_{1-\alpha/2,df=(m-1,n-1)}$, we can calculate the interval estimate of σ_X^2/σ_Y^2 by

$$\left[\frac{n(m-1)s_X^2}{m(n-1)s_Y^2}F_{1-\alpha/2,df=(m-1,n-1)},\ \frac{n(m-1)s_X^2}{m(n-1)s_Y^2}F_{\alpha/2,df=(m-1,n-1)}\right].$$

4.4 Confidence intervals: Large samples

- **Motivation:** When the population is not normal, we can make use of the CLT to propose confidence intervals, which has an approximated confidence coefficient $1-\alpha$ for large n.
- **Theorem 4.1.** (Slutsky's theorem) If $X_n \to_d X$ and $Y_n \to_p C$ (a constant), then
  (i) $X_n + Y_n \to_d X + C$;
  (ii) $X_n \cdot Y_n \to_d X \cdot C$;
  (iii) $X_n / Y_n \to_d X / C$ provided that $C \neq 0$.

Remark: Note that Theorem 4.1 fails if C is not a constant.

4.4 Confidence intervals: Large samples

Formula derivation

Let **X** be an independent random sample from a population, which has the mean μ and the variance $\sigma^2 < \infty$. According to Theorem 4.1, CLT and the fact that $S \to_p \sigma$, we have

$$\frac{\sqrt{n}(\overline{X} - \mu)}{S} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \cdot \frac{S}{\sigma} \to_d \mathrm{N}(0, 1), \tag{4.1}$$

for large n. Hence, by (4.1), it follows that for large n,

$$\begin{aligned} 1 - \alpha &\approx \mathrm{P}\left(-z_{\alpha/2} \leq \frac{\sqrt{n}(\overline{X} - \mu)}{S} \leq z_{\alpha/2}\right) \\ &= \mathrm{P}\left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right). \end{aligned}$$

4.4 Confidence intervals: Large samples

Formula derivation con't

So, an approximated $1-\alpha$ confidence interval of μ is

$$\left[\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right]. \tag{4.2}$$

Given the observed value of $\overline{x}$ and s, we can calculate the interval estimate of μ by

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

Note that the confidence interval in (4.2) only requires a large n but not the normal population assumption. Clearly, the similar idea can be applied to the two-sample case.

4.4 Confidence intervals: Large samples

End of the Chapter

To end this chapter, we consider the interval estimator for percentage p, where

$$p = \mathrm{P}\left(X \in (a, b)\right).$$

Define $\xi = \mathrm{I}(a < X < b)$. Then, $\mathrm{E}(\xi) = p$. This indicates that p is the theoretical mean of ξ. Hence, by (4.2), an approximated $1 - \alpha$ confidence interval of p is

$$\left[\overline{\xi} - z_{\alpha/2}\frac{S_\xi}{\sqrt{n}},\ \overline{\xi} + z_{\alpha/2}\frac{S_\xi}{\sqrt{n}}\right], \tag{4.3}$$

where $\overline{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$ and $S_\xi^2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \overline{\xi})^2 = \overline{\xi}(1 - \overline{\xi})$ with $\xi_i = \mathrm{I}(a < X_i < b)$.

In general, we can treat the interval (a, b) as "success", $p = \mathrm{P}(\text{"success"})$, and $\overline{\xi}$ = relative frequence of "success".

4.4 Confidence intervals: Large samples

Example 4.8. In a certain political campaign, one candidate has a poll taken at random among the voting population. The results are $n = 112$ and $y = 59$ (for "Yes"). Should the candidate feel very confident of winning?

Solution. Let $p = \mathrm{P}(\text{"the condidate wins the campaign"})$. Then, $\overline{\xi} = 59/112 \approx 0.527$. According to (4.3), since $z_{0.025} \approx 1.96$, an approximated 95% confident interval estimate for p is

$$\left[0.527 - z\sqrt{\frac{0.527 * (1 - 0.527)}{112}},\ 0.527 + z\sqrt{\frac{0.527 * (1 - 0.527)}{112}}\right]$$
$$\approx [0.435, 0.619].$$

There has certain possibility that p is less than 50%, and the candidate should take this into account in campaigning. □