STAT24510 (W25): Statistical Theory and Methods IIa

Lecturer: Mei Wang
Notes by: Aden Chen

Tuesday 7th January, 2025

Contents

1 Introduction

The goal of statistics is often to estimate a (population) parameter θ. From data, we may obtain point estimates $\hat{\theta}$ that depends on data, and construct confidence intervals to quantify the uncertainty of the estimate, with which we can conduct hypothesis testing.

In this course we will start from confidence intervals and hypothesis testing, and then move on to linear models.

2 Confidence Intervals

2.1 Constructing CI

2.1.1 Pivotal method

Example 2.1 (Pivotal method). Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 5^2)$, $i = 1, \ldots, n$. Our goal is to construct a $1 - \alpha$ CI for μ. We have the MLE estimator $\hat{\mu} = \overline{X} = n^{-1} \sum_{i=1}^{n} X_i$.

Note that

$$\frac{\overline{X} - \mathrm{E}[X]}{\sqrt{\mathrm{Var}(X)}} = \frac{\overline{X} - \mu}{\sqrt{5^2/n}} \sim \mathcal{N}(0, 1).$$

In particular, note that the left side is a function of data and parameters, while the right side is free of parameters.

Last updated: Tuesday 7th January, 2025.

We thus have

$$\mathbb{P}\left(z_{\alpha/2} \le \frac{\overline{X} - \mu}{\sqrt{5^2/n}} \le z_{1-\alpha/2}\right) = 0.95,$$

using which we can construct the CI of μ: With $I := \left[\overline{X} - z_{1-\alpha/2}\sqrt{5^2/n}, \overline{X} + z_{1-\alpha/2}\sqrt{5^2/n}\right]$, we have $\mathbb{P}(\mu \in I) = 1 - \alpha$.

Notice that we obtain a probability statement of random interval containing a fixed quantity from a probability statement of a fixed interval containing a random quantity.

Example 2.2 (Pivotal method II). Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(4, \sigma^2)$, $i = 1, \dots, n$. The goal: CI for σ^2, i.e., to find random variables L and U such that $\mathbb{P}\left(L \le \sigma^2 \le U\right) = 1 - \alpha$.

Note that

$$Y_i := \frac{X_i - 4}{\sigma} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$$

and thus

$$T_n := \sum Y_i^2 = \sum \left(\frac{X_i - 4}{\sigma}\right)^2 \sim \chi_n^2.$$

Again, we obtained a function of data and parameters that follows a known distribution. From

$$\mathbb{P}\left(\chi^2_{n,\alpha/2} \le T_n \le \chi^2_{n,1-\alpha/2}\right) = 1 - \alpha$$

we may again obtain the CI for σ^2,

$$\left[\frac{\sum(X_i - 4)^2}{\chi^2_{n,1-\alpha/2}}, \frac{\sum(X_i - 4)^2}{\chi^2_{n,\alpha/2}}\right].$$

Example 2.3 (Pivot failing). Let $X_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. The goal: CI for p. The MLE for p is $\hat{p} = \overline{X}$, thus we may be temped to try

$$T_n := \frac{\hat{p} - p}{\sqrt{\text{Var}(\hat{p})}},$$

but the distribution of T_n depends on p. The method of pivots fail.

2.1.2 Asymptotic CI

I.e., when we have large sample size n.

Example 2.4 (Wald CI). Let X_i be iid with mean μ and variance σ^2. From the CLT we have

$$\sqrt{n}\left(\overline{X} - \mu\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

that is,

$$\lim_{n\to\infty} \mathbb{P}(T_n \le x) = P(Z \le x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, \mathrm{d}t.$$

Thus we have

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

When σ^2 is known, we may derive an approximate CI for μ:

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

When σ^2 is unknown: If there exists random variables $U_n \to_{\mathrm{p}} \sigma^2$ (that is, $\lim \mathbb{P}\left(U_n = \sigma^2\right) = 1$), then

$$T_n \coloneqq \frac{\overline{X} - \mu}{\sqrt{U_n/n}} = \frac{\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{U_n/\sigma^2}}$$

where $(\overline{X} - \mu)/\sqrt{\sigma^2/n} \xrightarrow{\mathscr{D}} \mathcal{N}(0, 1)$ and $\sqrt{U_n/\sigma^2} \to_{\mathrm{p}} 1$, and thus by Slutsky's theorem we have

$$T_n \xrightarrow{\mathscr{D}} \mathcal{N}(0, 1),$$

using which we can again construct an approximate CI. Note that we used asymptotic approximation multiple times. This is called the Wald confidence interval.

Example 2.5 (Wald CI). Let $X_i \overset{\text{iid}}{\sim}$ Poisson(λ). The goal: asymptotic CI for λ. Note that we have the MLE of λ, $\hat{\lambda} = \overline{X}$, with $\mathrm{E}[\hat{\lambda}] = \lambda$ and $\mathrm{Var}[\hat{\lambda}] = \lambda/n$. We then have

$$\frac{\overline{X} - \lambda}{\sqrt{\lambda/n}} \approx Z \sim \mathcal{N}(0, 1).$$

We approximate a second time: $\frac{\overline{X}-\lambda}{\sqrt{\hat{\lambda}/n}} \approx Z$, from which we obtain the Wald CI for λ:

$$\left[\hat{\lambda} - z_{1-\alpha/2}\sqrt{\hat{\lambda}/n}, \hat{\lambda} + z_{1-\alpha/2}\sqrt{\hat{\lambda}/n}\right].$$

Example 2.6 (Wilson's method). Assume the same setup as above. Again, we use

$$\frac{\overline{X} - \lambda}{\sqrt{\lambda/n}} \approx Z \sim \mathcal{N}(0, 1),$$

which gives

$$\mathbb{P}\left(z_{\alpha/2} \leq \frac{\overline{X} - \lambda}{\sqrt{\lambda/n}} \leq z_{1-\alpha/2}\right) = \mathbb{P}\left(\left(\frac{\overline{X} - \lambda}{\sqrt{\lambda/n}}\right)^2 \leq z_{1-\alpha/2}^2\right) \approx 1 - \alpha.$$

Solving for λ in the middle expression gives the Wilson CI. We used one fewer approximation.