**6.S951 Modern Mathematical Statistics** Fall 2024

# Lecture 11 — November 21, 2024

*Prof. Stephen Bates*

*Scribe: Kai Chang*

## 1 Outline

**Agenda:**

1. Overview of asymptotic normality (optimality and confidence intervals)
2. Probability recap
3. Delta Method

**Last time:**

1. Prediction confidence intervals in linear model
2. Conformal prediction
3. Risk-controllability
4. Multiple testing
5. FWER (Bonferroni)
6. FDR (Benjamini-Hochberg)

## 2 Asymptotics

### Problem Setting

Let $X_i \in \mathcal{X} \sim P, \quad i = 1, 2, \ldots, n$ be IID data points. We want to estimate $\theta \in \mathbb{R}^d$ $(\theta \mapsto P)$ with an estimator: $\hat{\theta}_n : \mathcal{X}^n \to \mathbb{R}^d$ (the form of the estimator is predetermined, but data points are random).

### Goal

Wish to understand the behavior of $\hat{\theta}_n$ (limiting behavior):

- A good estimator (theoretical).
- Confidence intervals (CIs) based on $\hat{\theta}_n$ (fixed on a prediction).

### Convergence

It turns out for most $\hat{\theta}_n$,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma), \quad \hat{\theta}_n \sim \mathcal{N}(\theta, \Sigma_n).$$

### Optimality

The smaller $\Sigma$ is, the better the estimator (in the sense of SPD matrices).

### Inference

Also want to characterize approximate confidence intervals:

$$\hat{\theta}_n^{(j)} \pm 1.96\sqrt{\Sigma_{jj}/n}.$$

## 3 Probability Reminder

**Theorem 1** (CLT)**.** *Let $X_i \in \mathbb{R}^k$, $i = 1, 2, \ldots,$ be i.i.d. $\sim P$, with finite $\mu = \mathbb{E}[X_i]$ and $\Sigma = \mathbb{E}[(X_i - \mu)(X_i - \mu)^\top]$ . Then:*

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

**Theorem 2** (Continuous Mapping)**.** *Let $Y_n \xrightarrow{d} Y^*$ and $g$ be a continuous function. Then the followings hold.*

1. *$g(Y_n) \xrightarrow{d} g(Y^*)$, if $g$ is continuous.*
2. *If $Y_n \xrightarrow{p} c$, $g$ is continuous at $c$, then $g(Y_n) \xrightarrow{p} g(c)$.*

**Theorem 3** (Slutsky)**.** *Let $Y_n \xrightarrow{d} Y^*$ and $Z_n \xrightarrow{p} c$ (where $c$ is a constant). Then the followings hold.*

1. *$Y_n + Z_n \xrightarrow{d} Y^* + c$.*
2. *$Y_n Z_n \xrightarrow{d} Y^* c$.*
3. *$Y_n / Z_n \xrightarrow{d} Y^*/c$, if $c \neq 0$.*

**Definition 4** (Uniform Tightness)**.** *Let $Y_n$ be random vectors in $\mathbb{R}^k$. The sequence $\{Y_n\}$ is uniformly tight if $\forall \epsilon > 0$, $\exists M > 0$ such that:*

$$\sup_n \mathbb{P}(\|Y_n\| > M) < \epsilon.$$

Uniform Tightness is an analogy of a bounded deterministic sequence (*sequential compactness*). With such a property, no probability mass is escaping to infinity.

**Definition 5.** *Some other convenient definitions/notations are summarized here.*

1. $o_p(1)$ *is a sequence* $Y_n \xrightarrow{p} 0 \quad$ (*where* $Y_n$ *is a random vector*).
2. $O_p(1)$ *is a sequence that is uniformly tight.*
3. *For random variables* $R_n$, *we have:*
   - $X_n = o_p(R_n)$ *if* $X_n = Y_n R_n$, *where* $Y_n = o_p(1)$.
   - $X_n = O_p(R_n)$ *if* $X_n = Y_n R_n$, *where* $Y_n = O_p(1)$.
4. *Analogy in scalar sequences:*
   - $o(1)$: *Sequence converges to* 0.
   - $O(1)$: *Bounded sequence.*

**Proposition 6.** *With the definitions above, the following statements hold:*

- $o_p(1) + o_p(1) = o_p(1)$.
- $o_p(1) + O_p(1) = O_p(1)$.
- $o_p(R_n) = R_n o_p(1) \quad$ *by definition.*
- $O_p(R_n) = R_n O_p(1)$.

# 4 Delta Method (Taylor Theorem + Probability)

Delta method is a method to figure out the behavior of functions of sequences. Suppose

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

and we want to know the behavior of $\phi(\hat{\theta}_n)$, where $\phi : \mathbb{R}^d \to \mathbb{R}^m$.

### Idea: Taylor Expansion

Since $\hat{\theta}_n \xrightarrow{p} \theta$, $\hat{\theta}_n$ is close to $\theta$. By Taylor Expansion, we have

$$\phi(\hat{\theta}_n) \approx \phi(\theta) + \phi'(\theta)(\hat{\theta}_n - \theta),$$

where $\phi'$ is the Jacobian (gradient transpose). Since this is an affine transform, we have:

$$\phi(\hat{\theta}_n) - \phi(\theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\phi'(\theta)\Sigma\phi'(\theta)^\top}{n}\right).$$

**Theorem 7** (Delta Method)**.** *Let* $\phi : \mathbb{R}^d \to \mathbb{R}^m$ *be differentiable at* $\theta$. *If*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma),$$

*then*

$$\sqrt{n}\left(\phi(\hat{\theta}_n) - \phi(\theta)\right) \xrightarrow{d} \mathcal{N}(0, \phi'(\theta)\Sigma\phi'(\theta)^\top).$$

*In particular,*

$$\sqrt{n}\left(\phi(\hat{\theta}_n) - \phi(\theta)\right) \xrightarrow{d} \mathcal{N}(0, \phi'(\theta)\Sigma\phi'(\theta)^\top).$$

**Example: Sample Variance**

Let $X_i \in \mathbb{R}$, i.i.d., with finite 4th moment. Let $\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$. What is $\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d}$?

**Ans:** Note $\hat{\sigma}^2 = \phi\left(\bar{X}, \bar{X^2}\right)$, where $\phi(x, y) = y - x^2$. By CLT,

$$\sqrt{n}\left(\begin{pmatrix}\bar{X} \\ \bar{X}^2\end{pmatrix} - \begin{pmatrix}\alpha_1 \\ \alpha_2\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\alpha_2 & \alpha_3 - \alpha_1\alpha_2 \\ \alpha_3 - \alpha_1\alpha_2 & \alpha_4 - \alpha_2\end{pmatrix}\right)$$

where $\alpha_k = \mathbb{E}[X^k]$. Using $\phi'(\theta) = (-2\alpha_1, 1)$ and applying the Delta Method yield

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{d} \mathcal{N}\left(0, \phi'(\theta)\Sigma\phi'(\theta)^\top\right).$$

What about $\frac{1}{n-1}\sum(X_i - \bar{X})^2$? Simply rewrite

$$\frac{1}{n-1}\sum(X_i - \bar{X})^2 = \frac{n}{n-1}\hat{\sigma}^2 \approx \hat{\sigma}^2,$$

which asymptotically stays the same.

**Procedure**

1. Write statistics as a function of simple statistics where we can apply the CLT.
2. Apply the Delta Method.

**Confidence Intervals for $\sigma$:**

Let

$$\hat{\alpha}_k = \frac{1}{n}\sum X_i^k, \quad k = 1, \cdots, 4.$$

Then

$$\hat{\gamma}^2 = (-2\hat{\alpha}_1, 1)\,\hat{\Sigma}\,(-2\hat{\alpha}_1, 1)^\top,$$

and

$$\hat{\sigma}^2 \pm 1.96\hat{\gamma}/\sqrt{n}$$

is the asymptotic 95% confidence interval.

**Example:**

$$T_n = \left(\hat{\sigma}^2, \frac{\bar{x}}{\hat{\sigma}}\right)$$
$$\sqrt{n}\left(T_n - \left(\sigma^2, \frac{\mu}{\sigma}\right)\right) \xrightarrow{d} ?$$