Modern Statistics	**Lecture 8 - 03/17/2025**

Lecture 8 Statistical Inference

Lecturer:Xiangyu Chang *Scribe:Zhiwen Li,Le Chang*

Edited by: Zhihong Liu

1 Introduction

Statistical inference, or "learning" as it is called in computer science, is the process of using data $\{Z_i\}_{i=1}^n$ to infer the distribution that generated the data.

1.1 Examples

Example 1 (Inferring the Population Mean from the Sample Mean) *Suppose there is a population X with an expected value μ. A random sample of size n, $X_1, X_2, \ldots, X_n$, is drawn from this population, and the sample mean is represented as:*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

According to WLLN, we can use the sample mean $\overline{X}$ to infer the population mean μ.

Example 2 (Estimating Parameters of a Normal Distribution) *Let $X_1, X_2, \ldots, X_n$ be independent observations from a normal distribution $N(\mu, \sigma^2)$. The problem is to estimate the parameters μ and σ^2. The maximum likelihood estimates are given by*

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Example 3 (Estimating Linear Coefficient Vector) *Consider independent data pairs $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n)$, where $\mathbf{X}_i$ is a p-dimensional vector and Y_i is a scalar observation. Suppose the relationship between Y_i and $\mathbf{X}_i$ is given by an unknown function $Y_i = r(\mathbf{X}_i) + \epsilon_i$ with ϵ_i being independent errors. Assume further that $r(\mathbf{X})$ is linear, i.e., $r(\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is an unknown p-dimensional parameter vector. The problem then becomes estimating the parameter vector $\boldsymbol{\beta}$. This assumption transforms the inference problem from estimating the function $r(\mathbf{X})$ itself to estimating its parameter $\boldsymbol{\beta}$.*

Example 4 (k-Nearest Neighbors Function Estimation) *Continuing with the setup of independent data pairs $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n)$, where $\mathbf{X}_i$ is a p-dimensional vector and Y_i is a scalar* *observation, we now consider a non-parametric approach to estimating the unknown function $r(\mathbf{X})$. Using the k-nearest neighbors (kNN) method, for a new input point $\mathbf{X}^*$, we identify the k closest points $\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, \ldots, \mathbf{X}_{(k)}$ in the training data, along with their corresponding $Y_{(1)}, Y_{(2)}, \ldots, Y_{(k)}$. The function value $r(\mathbf{X}^*)$ is then estimated as the average of these k $Y_{(i)}$ values:*

$$\hat{r}(\mathbf{X}^*) = \frac{1}{k}\sum_{i=1}^{k} Y_{(i)}.$$

This approach does not assume a specific parametric form for $r(\mathbf{X})$ but instead relies on local averaging of neighboring data points for estimation.

2 Fundamental Concepts in Inference

2.1 Estimating

Let $X_1, \ldots, X_n$ be n IID data points from some distribution F. A point estimator $\widehat{\theta}_n$ of a parameter θ is some function of $X_1, \ldots, X_n$:

$$\widehat{\theta}_n = g(X_1, \ldots, X_n).$$

2.1.1 Criteria Performance Metrics

1. MSE(Mean Square Error)

$$\text{MSE} = \mathbb{E}[(\hat{\theta_n} - \theta)^2].$$

2. Bias

$$\text{Bias} = \mathbb{E}(\hat{\theta_n} - \theta) = \mathbb{E}(\hat{\theta_n}) - \theta.$$

$\hat{\theta_n}$ is unbiased if $\mathbb{E}(\hat{\theta_n}) = 0$.
Connection:MSE $=$ **Bias**$^2(\hat{\theta_n}) +$ **Var**$(\hat{\theta_n})$
Prove:

$$\begin{aligned}
\text{MSE} &= \mathbb{E}[(\hat{\theta_n} - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta_n} - \mathbb{E}(\hat{\theta_n}) + \mathbb{E}(\hat{\theta_n}) - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta_n} - \mathbb{E}(\hat{\theta_n}))^2 + (\mathbb{E}(\hat{\theta_n}) - \theta)^2 + 2(\hat{\theta_n} - \mathbb{E}(\hat{\theta_n}))(\mathbb{E}(\hat{\theta_n}) - \theta)] \\
&= \text{Var}(\hat{\theta_n}) + \text{Bias}^2(\hat{\theta_n}) + 2\mathbb{E}[(\hat{\theta_n} - \mathbb{E}(\hat{\theta_n}))(\mathbb{E}(\hat{\theta_n}) - \theta)] \\
&= \text{Var}(\hat{\theta_n}) + \text{Bias}^2(\hat{\theta_n}) + 2(\mathbb{E}(\hat{\theta_n}) - \theta)\mathbb{E}[\hat{\theta_n} - \mathbb{E}(\hat{\theta_n})] \\
&= \text{Bias}^2(\hat{\theta_n}) + \text{Var}(\hat{\theta_n}).
\end{aligned}$$

3. Consistency
A point estimator $\hat{\theta}_n$ of a parameter θ is **consistent** if $\hat{\theta}_n \xrightarrow{\text{P}} \theta$.

Theorem 1 *If bias* $\to 0$ *and* $Var(\hat{\theta_n}) \to 0$ *as* $n \to \infty$ *then* $\hat{\theta_n}$ *is consistent, that is,* $\hat{\theta}_n \xrightarrow{P} \theta$.

Proof 1

$$\mathbb{P}(\left|\hat{\theta_n} - \theta\right| > \epsilon) \leq \frac{\mathbb{E}[(\hat{\theta_n} - \theta)^2]}{\epsilon^2} = \frac{Bias^2(\hat{\theta_n}) + Var(\theta_n)}{\epsilon^2},$$

if $n \to \infty$,*bias* $\to 0$ *and* $Var(\hat{\theta_n}) \to 0$,*then* $\hat{\theta}_n \xrightarrow{P} \theta$.

4. Standard Error

$$\text{se}(\hat{\theta_n}) = \sqrt{\text{Var}(\hat{\theta_n})}.$$

2.2 Confidence Sets

A $1-\alpha$ **confidence interval** for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ are functions of the data such that

$$\mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha, \quad \text{for all } \theta \in \Theta.$$

In words, (a, b) traps θ with probability $1-\alpha$. We call $1-\alpha$ the **coverage** of the confidence interval.

Warning! C_n is random and θ is fixed.

2.2.1 Eg. Expectation and Variance of an Estimator

Let $\hat{\mu}_n$ be an estimator of μ. Then, its expectation and variance are given by:

$$\mathbb{E}(\hat{\mu}_n) = \mu, \quad \text{Var}(\hat{\mu}_n) = \frac{\sigma^2}{n}.$$

Note: The estimator $\hat{\mu}_n$ is unbiased since its expectation equals the true parameter μ. The variance decreases as the sample size n increases, indicating greater precision.

For any $\varepsilon > 0$, Chebyshev's inequality states:

$$P(|\hat{\mu}_n - \mu| > \varepsilon) \leq \frac{\text{Var}(\hat{\mu}_n)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2 n}.$$

Consequently,

$$P(|\hat{\mu}_n - \mu| \leq \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2 n}.$$

Let $\alpha = \frac{\sigma^2}{\varepsilon^2 n}$. Solving for ε, we obtain:

$$\varepsilon = \sqrt{\frac{\sigma^2}{n\alpha}}.$$

Thus, an approximate $1-\alpha$ confidence interval for μ is:

$$\mu \in \left[\hat{\mu}_n - \sqrt{\frac{\sigma^2}{n\alpha}}, \hat{\mu}_n + \sqrt{\frac{\sigma^2}{n\alpha}}\right].$$

This interval has a confidence level of $1-\alpha$.

2.3 Theorem: Asymptotic Normality and Confidence Interval

If we know that :

$$\frac{\hat{\theta}_n - \theta}{\text{Se}(\hat{\theta}_n)} \xrightarrow{d} N(0,1).$$

where $\hat{\theta}_n$ is an estimator of θ, and $\text{Se}(\hat{\theta}_n)$ is the standard error of $\hat{\theta}_n$. Then, the probability that θ lies in the confidence interval C_n converges to $1 - \alpha$:

$$P(\theta \in C_n) \to 1 - \alpha.$$

where the confidence interval C_n is given by:

$$C_n = \left[\hat{\theta}_n - z_{\alpha/2} \cdot \text{Se}(\hat{\theta}_n), \hat{\theta}_n + z_{\alpha/2} \cdot \text{Se}(\hat{\theta}_n)\right].$$

Here, $z_{\alpha/2}$ is the critical value of the standard normal distribution such that $P(Z > z_{\alpha/2}) = \alpha/2$.

Under the same assumptions, the scaled estimator satisfies:

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma^2} \xrightarrow{d} N(0,1).$$

Note: The term σ^2 is replaced by the standard error $\text{Se}(\hat{\theta}_n)$ in practice. This substitution is illustrated below:

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma^2} \xrightarrow{\text{Se}(\hat{\theta}_n) = \frac{\sigma}{\sqrt{n}}} N(0,1).$$

Proof 2 *The probability that θ lies in the confidence interval C_n is:*

$$P(\theta \in C_n) = P\left(-z_{\alpha/2} \leq \frac{\hat{\theta}_n - \theta}{Se(\hat{\theta}_n)} \leq z_{\alpha/2}\right).$$

By the asymptotic normality assumption, this probability converges to:

$$P(\theta \in C_n) \to 1 - 2\Phi(-z_{\alpha/2}) = 1 - \alpha.$$

2.4 Hypothesis testing

2.4.1 Key Components

- **Null Hypothesis (H_0)**: A statement that there is no effect or no difference. It represents the default or status quo assumption.
- **Alternative Hypothesis (H_1 or H_a)**: A statement that contradicts the null hypothesis. It represents the research question or the effect we are testing for.

- **Test Statistic**: A numerical value calculated from the sample data, used to assess the strength of evidence against the null hypothesis.
- **Significance Level (α)**: The probability of rejecting the null hypothesis when it is true (Type I error). Common choices are $\alpha = 0.05$ or $\alpha = 0.01$.
- **p-value**: The probability of observing the test statistic or something more extreme under the null hypothesis. If the p-value is less than α, we reject the null hypothesis.

2.4.2 Eg: Testing a Bernoulli Parameter

Consider a dataset $\{x_i\}$ for $i = 1, \ldots, n$, where each x_i is independently drawn from a Bernoulli distribution with parameter p:

$$x_i \sim \text{Ber}(p).$$

We want to test whether the parameter p is equal to $1/2$.

2.4.3 Hypotheses

- Null Hypothesis (H_0): $p = \frac{1}{2}$.
- Alternative Hypothesis (H_1): $p \neq \frac{1}{2}$.

2.4.4 Test Statistic

Under the null hypothesis, the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an estimator of p. The test statistic for this problem is:

$$Z = \frac{\bar{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

where $p_0 = \frac{1}{2}$ is the value of p under the null hypothesis. For large n, Z approximately follows a standard normal distribution:

$$Z \sim N(0, 1).$$

2.4.5 Decision Rule

- If $|Z| > z_{\alpha/2}$, reject the null hypothesis.
- If $|Z| \leq z_{\alpha/2}$, fail to reject the null hypothesis.

Here, $z_{\alpha/2}$ is the critical value from the standard normal distribution corresponding to the significance level α.

2.4.6 Interpretation

- Rejecting H_0 suggests that there is sufficient evidence to conclude that $p \neq \frac{1}{2}$.
- Failing to reject H_0 suggests that there is not enough evidence to conclude that $p \neq \frac{1}{2}$.