Chapter 1 Introduction

Data doesn't come with a model on its back — John Tukey

1. Example: Let $F(\cdot)$ be the cumulative density function of the monthly household of income over Hong Kong.

$X_1, X_2, \cdots, X_n$ is randomly and independently generated from $F(\cdot)$.

• Two ways to modeling the $F(\cdot)$:

Model 1: $F(\cdot)$ is $N(\mu, 1) \Leftrightarrow F(\cdot) \in \{N(\mu, 1) : \mu \in (-\infty, +\infty)\}$, where "$F(\cdot)$" is known except for one dimensional number μ. This is a parametric model.

Model 2: $F(\cdot)$ is a symmetric distribution $\Leftrightarrow$ the density function is symmetric about (median) μ

$$\Leftrightarrow F(\cdot) \in \{ \text{ All symmetric distribution}\},$$

where "$F(\cdot)$" can't be described by a few finite-dimensional parameters.

• Estimation and Accuracy:

Model 1: μ is estimated by

$$\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_i \ \ and \ \ \bar{X} - \mu \sim N(0, \frac{1}{n}).$$

Model 2: μ is estimated by the sample median

$$\hat{X} = \begin{cases} \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} & if \ n \text{ is even} \\ X_{\left(\frac{n+1}{2}\right)} & if \ n \text{ is odd} \end{cases}$$

where $X_{(1)} \leq X_{(2)}, \cdots \leq X_{(n)}$ are ordered sample of $X_1, X_2. \cdots, X_n$.

$$\hat{X} - \mu \sim N\left(0, \frac{1}{4f^2(\mu)n}\right) \quad \text{approximately}$$

—-f is the density function of X_1.

• Comparison of Model 1 and Model 2 in terms of statistical analysis.

(1) If Model 1 is correct (then model 2 is also correct), we have

$$\bar{X} - \mu \sim N\left(0, \frac{1}{n}\right),$$

$$\begin{aligned} \hat{X} - \mu \sim N\left(0, \frac{1}{4f^2(\mu)n}\right) &= N\left(0, \frac{\pi}{2n}\right) \\ &= N\left(0, \frac{1.57}{n}\right). \end{aligned}$$

Since $\frac{1.57}{n} > \frac{1}{n}$, $\bar{X}$ is a more accurate estimator than $\hat{X}$. (Parametric model beats nonparametric model in the sense of estimation accuracy when it's correct).

(2) If Model 2 is correct but Model 1 is wrong, still we have

$$\hat{X} - \mu \sim N\left(0, \frac{1}{4f^2(\mu)n}\right)$$

Here $\hat{X}$ is better than $\bar{X}$.

2. Para/nopara -metric models

A. Parametric model: Making strong assumption: the population of distribution is known except for finite dimensional parameter.

• Example: One sample

$$\begin{cases} X_1, X_2, \cdots, X_n \text{ i.i.d.} \sim N(\mu, \sigma^2) \\ X_1, X_2, \cdots, X_n \text{ i.i.d.} \sim P(\lambda) \\ X_1, X_2, \cdots, X_n \text{ i.i.d.} \sim Bin(1, p) \end{cases}$$

Two samples (independent)

$$\begin{cases} X_1, X_2, \cdots, X_n \text{ i.i.d.} \sim N(\mu_1, \sigma_1^2) \\ Y_1, Y_2, \cdots, Y_n \text{ i.i.d.} \sim N(\mu_2, \sigma_2^2) \end{cases}$$

B. Nonparametric model: Making little or minimal assumption about population.

• Example: One sample

$$X_1, X_2, \cdots, X_n \text{ i.i.d.} \sim F,$$

where F is continuous or F is symmetric about median Two samples

$$\begin{cases} X_1, X_2, \cdots, X_n \text{ i.i.d.} \sim F \\ Y_1, Y_2, \cdots, Y_n \text{ i.i.d.} \sim G \end{cases}$$

where $F(x) = G(x - \mu)$, μ is called location shift.

Relative advantage/disadvantage:

Parametric model

•**Adv**: More accurate estimation inference and statistic analysis. Sharper conclusions and better interpretation, if the model is correct.

•**Disadv**: Strong assumption about population, more likely to be wrong. Interpretation and conclusions may be seriously misleading, if the model is wrong.

Nonparametric model

•**Adv**: Little or minimal assumption. Less likely to be wrong in terms of modeling. More reliable conclusions and interpretation.

•**Disadv**: Loss of accuracy. Less powerful conclusion.