Sampling and statistical modeling

Olga Klopp

1 Statistical problem

1) Starting point : data (e.g. real numbers)

$$\mathtt{x_1}, \ldots, \mathtt{x_n}$$

2) Statistical modeling :

   – the data are realizations of random variables

$$X_1(\omega), \ldots, X_n(\omega) \text{ of r.v. } X_1, \ldots, X_n.$$

   (that is, for some ω, $X_1(\omega) = \mathtt{x_1}, \ldots, X_n(\omega) = \mathtt{x_n}$)

   – The law $\mathbb{P}^{(X_1,\ldots,X_n)}$ of $(X_1, \ldots, X_n)$ is unknown, but belongs to a given family of distributions

$$\boxed{\left\{\mathbb{P}_\theta^n, \theta \in \Theta\right\}} : \text{the model}$$

   We think that there exists $\theta \in \Theta$ such that $\mathbb{P}^{(X_1,\ldots,X_n)} = \mathbb{P}_\theta^n$.

3) Problem: from the observations $X_1, \ldots, X_n$, can we estimate θ? test its properties?

- θ is the parameter and Θ is the set of parameters.
- Estimation: using $X_1, \ldots, X_n$, construct $\varphi_n(X_1, \ldots, X_n)$ which "approaches at best" θ.
- Test: using data $X_1, \ldots, X_n$, take a decision $\varphi_n(X_1, \ldots, X_n) \in \{\text{set of decisions}\}$ about a hypothesis on θ.

Definition 1. *A statistic (singular) or sample statistic is a measurable function of a sample.*

!ATTENTION! A statistic can not depend on unknown parameter: a statistic is constructed only from the data!

Example:

- We throw a coin 18 times and we observe ($H = 0$, $T = 1$)

$$0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0$$

- Statistical model : we observe $n = 18$ independent Bernoulli random variables X_i of **unknown** parameter $\theta \in \Theta = [0, 1]$.

  – Estimation. Estimator $\bar{X}_{18} = \frac{1}{18}\sum_{i=1}^{18} X_i \overset{\text{here}}{=} 8/18 = 0.44$ (sample mean) What is the precision?

  – Testing. Decision to make: "Is the coin balanced?". One possible solution: we compare $\bar{X}_{18}$ to 0.5. If $|\bar{X}_{18} - 0.5|$ is "small", we accept the hypothesis "the coin is balanced ". If not, we reject. How to choose the threshold, and with what consequences (e.g. probability of being wrong)?

- The basic statistical trial : we observe realization of $X_1, \ldots, X_n$, where the X_i are , independent identically distributed (i.i.d.) random variables, of the same common law $\mathbb{P}^X \in \{\mathbb{P}_\theta : \theta \in \Theta\}$.
- **Problem** : using data $X_1, \ldots, X_n$ what can be said about $\mathbb{P}^X$ the commune law of X_i? (expectation, moments, symmetry, density, etc.)

2 Statistical model

- $\{\mathbb{P}_\theta,\ \theta \in \Theta\}$ is called model
- when for some k, $\Theta \subset \mathbb{R}^k$, we say that the model is parametric
- when θ is an infinite dimensional parameter, we say that the model is non-parametric (e.g., the set of all distributions with mean 0)
- when the parameter $\theta = (f, \theta_0)$ has both, f infinite-dimensional component (e.g., a real-valued function defined on the real line; often nuisance parameter) and a finite-dimensional component $\theta_0 \in \mathbb{R}^k$ (parameter of interest), we say that the model is semi-parametric
- when $\theta \in \Theta \mapsto \mathbb{P}_\theta$ is injective, we say that the model is **identifiable** (an injective function or one-to-one function is a function that preserves distinctness: it never maps distinct elements of its domain to the same element of its codomain).

Example: A well-known example of a semi-parametric model is the Cox proportional hazards model. If we are interested in studying the time T to an event such as death due to cancer or failure of a light bulb, the Cox model specifies the following distribution function for T:

$$F(t) = 1 - \exp\left(-\int_0^t \lambda_0(u) e^{\beta' x} du\right)$$

where x is the covariate vector, and β and $\lambda_0(u)$ are unknown parameters. $\theta = (\beta, \lambda_0(u))$. Here β is finite-dimensional and is of interest; $\lambda_0(u)$ is an unknown non-negative function of time (known as the baseline hazard function) and is often a nuisance parameter. The collection of possible candidates for $\lambda_0(u)$ is infinite-dimensional.

Central question in statistics:

What model is the most suitable for our data?

There are two equivalent ways to define a model:

1. by giving a family of laws $\{\mathbb{P}_\theta,\ \theta \in \Theta\}$
2. by giving an equation: everything in statistics essentially boils down to one equation:

$$\text{outcome}_i = (\text{model}) + \text{error}_i$$

This means that the data we observe can be predicted from the model we choose plus some amount of error.

Example of model/modeling (1): We observe a n-tuple of real random variables :

$$Z = (X_1, \ldots, X_n)$$

These observations can be modeled in two equivalent ways:

- Family of laws : $\{\mathbb{P}_\theta : \theta \in \mathbb{R}\}$, for example,

$$\boxed{\mathbb{P}_\theta = \big(\mathcal{N}(\theta, 1)\big)^{\otimes n}}$$

- By using an equation: for any $i \in 1, \ldots, n$,

$$\boxed{X_i = \theta + g_i}$$

where $g_1, \ldots, g_n$ are n independent standard Gaussian random variables.

Example of model/modeling (2): We observe a n-tuple of real random variables :

$$Z = (X_1, \dots, X_n)$$

These observations can be modeled in two equivalent ways:

- By using an equation: $X_1 = g_1$ and for any $i \in 1, \dots, n-1$,

$$\boxed{X_{i+1} = \theta X_i + g_i}$$

  where $g_1, \dots, g_n$ are iid $\mathcal{N}(0,1)$.

- Family of laws: $\{\mathbb{P}_\theta : \theta \in \mathbb{R}\}$ where $\mathbb{P}_\theta$ has the density f_θ given by

$$\boxed{f_\theta(x_1, \dots, x_n) = f(x_1) f(x_2 - \theta x_1) \cdots f(x_n - \theta x_{n-1})}$$

  where $f(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}}$.

If we're looking at how well a model fits the data then we generally look at deviation from the model: we look at the sum of squared error

$$\text{deviation} = \sum (\text{obseved} - \text{model})^2.$$

That is, we assess models by comparing the data we observe to the model we have fitted to the data.

2.1 3 classic (non-parametric) models

1. **Density model**: we observe a n -sample

$$X_1, \dots, X_n \text{ of r.v. having distribution } f \text{ such that} f \in \mathcal{C}$$

   where $\mathcal{C}$ is a set of densities on $\mathbb{R}$.

2. **Regression**: we observe a n -sample of couples $(X_i, Y_i)_{i=1}^n$ such that $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^d$ and

$$Y_i = g(X_i) + \xi_i$$

   where ξ_i are i.i.d. random variables independent of X_i and g.

   - when $g(X_i) = \langle \theta, X_i \rangle$: linear regression model,
   - and when $\xi_i \sim \mathcal{N}(0, \sigma^2)$: linear Gaussien model

3. **Classification**: we observe a n-sample $(Y_i, X_i)_{i=1}^n$ such that $Y_i \in \{0, 1\}$ and $X_i \in \mathcal{X}$. For example:

$$\mathbb{P}[Y_i = 1 | X_i = x] = \sigma(\langle x, \theta \rangle) \text{ where } \sigma(x) = (1 + e^{-x})$$

3 Fundamental theorems

Let $X_1, \dots, X_n$ be a random sample of size n (that is, a sequence of independent and identically distributed random variables drawn from a distribution of a random variable X with expected value given by μ and finite variance given by σ^2). Suppose we are interested in the sample average

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

Remember that $\bar{X}_n$ **is a random variable**. Each time we draw a sample of size n from a population, we can get a different value for $\bar{X}_n$.

Theorem 1 (**Strong law of large numbers**). *Let (X_n) be a sequence of i.i.d. random variables such that $\mathbb{E}|X_1| < \infty$. Then,*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mathbb{E}\, X_1$$

That is, the average of the results obtained from a large number of trials should be close to the expected value.

The Law of Large Numbers (LLN) is important because it "guarantees" stable long-term results for the averages of some random events. The law of large numbers is the reason such businesses as health insurance, automobile insurance, and gambling casinos can exist and make a profit: while a casino may lose money in a single spin of the roulette wheel, its earnings will tend towards a predictable percentage over a large number of spins.

Theorem 2 (**Central Limit Theorem**). *If a random variable X possesses any distribution with mean μ and standard deviation σ, then, if the sample size is large, $\bar{X}_n$ has a distribution that is approximately normal with mean μ and and standard deviation $\sigma/\sqrt{n}$:*

$$\frac{\sqrt{n}}{\sigma}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\, X\Big) \overset{d}{\to} \mathcal{N}(0,1)$$

The central limit theorem is surprising. It says that X can have any distribution whatsoever, but as the sample size gets larger and larger, the distribution of $\bar{X}_n$ will approach a normal distribution. The central limit theorem (CLT) implies that statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions. In the CLT, the degree to which the distribution of $\bar{X}_n$ fits a normal distribution depends on both the selected value of n and the original distribution of X.

A natural question: How large should the sample size be if we want to apply the CLT? For the CLT to hold $n \geq 30$ is adequate in almost all practical applications. However, if the X distribution is definitely not symmetrical about its mean, then the $\bar{X}_n$ distribution also will display a lack of symmetry. In such a case, a sample size larger than 30 may be required to get a reasonable approximation to the normal. In practice, it is a good idea to make a histogram of sample values.

- CLT : "rate of convergence" in the LLN
- Interpretation of the CLT:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mu + \frac{\sigma}{\sqrt{n}}\,\xi^{(n)}, \quad \xi^{(n)} \overset{d}{\approx} \mathcal{N}(0,1).$$

- The convergence is convergence in distribution. Can not take place in probability.

3.1 Confidence Interval

We can use $\bar{X}_n$ (the sample mean) to estimate the population mean μ. An estimate of a population parameter given by a single number is called a **point estimate** for that parameter. Even with a large random sample, the value of $\bar{X}_n$ usually is not exactly equal to the population mean μ. An estimate is not very valuable unless we have some kind of measure of how "good" it is. Of cause, we cannot say exactly how close $\bar{X}_n$ is to μ when μ is unknown.

The reliability of an estimate is usually measured by the **confidence level**. Suppose we want a confidence level of c. Theoretically, we can choose c to be any value between 0 and 1, but usually c is equal to a number such as 0.90, 0.95, or 0.99. In each case, the value z_c is the number such that the area under the standard normal curve falling between $-z_c$ and z_c is equal to c:

$$\mathbb{P}\,(-z_c < Z < z_c) = c.$$

The value z_c is called the critical value for a confidence level of c.

According to the CTL, if the sample size is large, then $\bar{X}_n$ has a distribution that is approximately normal with mean μ, the population mean we are trying to estimate and the standard deviation $\sigma/\sqrt{n}$ (if X has a normal distribution, these results are true for any sample size). This information leads to the following probability statement:

$$\mathbb{P}\left(-z_c\frac{\sigma}{\sqrt{n}} < \bar{X}_n - \mu < z_c\frac{\sigma}{\sqrt{n}}\right) = c.$$

This is a probabilistic statement: we have a chance c of obtaining a sample such that the interval will contain the parameter μ. That is, we may get a different confidence interval for each different sample that is taken. Some intervals will contain the population mean μ and others will not. However, in the long run, the proportion of confidence intervals that contain μ is c.

3.2 Slutsky's theorem

- We say that a sequence of random vectors $(X_n, Y_n) \xrightarrow{d} (X, Y)$ if

  $$\mathbb{E}\left[\varphi(X_n, Y_n)\right] \to \mathbb{E}\left[\varphi(X, Y)\right],$$

  for any bounded continuous function φ.

- Attention ! If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$, in general we do not have $(X_n, Y_n) \xrightarrow{d} (X, Y)$.

- But (Slutsky's theorem) if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{\mathbb{P}} c$ (constant), then $(X_n, Y_n) \xrightarrow{d} (X, Y)$.

- Under the assumptions of the theorem, we also have that $g(X_n, Y_n) \xrightarrow{d} g(X, Y)$ for any continuous function g.

3.3 Continuous map theorem

Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a continuous function and (X_n) a sequence of r.v.

1. if (X_n) converges in distribution to X then $f(X_n)$ converges in distribution to $f(X)$
2. if (X_n) converges in probability to X then $f(X_n)$ converges in probability to $f(X)$
3. if (X_n) converges a.s. to X then $f(X_n)$ converges a.s. to $f(X)$