Week 13: Estimation

Armenak Petrosyan

Tuesday class

Data

- Data in general: just a collection of numbers

$$\{x_1, \dots, x_n\}$$

- Sample data:

$$\begin{cases} x_1 = X_1(s) \\ \quad \vdots \\ x_n = X_n(s) \end{cases} \qquad s \in S$$

where $X_1, \dots, X_n$ are i.i.d. random variables.

Discrete-type data:

- Data can only take values from a given set $R = \{r_1, r_2, \ldots\}$: $x_i \in R$.
- **Relative frequency** of r_j:
$$f_j = \frac{\text{number of data } = r_j}{n}.$$
- If data is a sample from a distribution, we use it as an estimate for the pmf (**empirical pmf**).
- Can be checked that
$$\bar{\mu} = \sum_j r_j f_j \qquad \text{(mean)}$$
$$\bar{\sigma}^2 = \sum_j (r_j - \bar{\mu})^2 f_j \qquad \text{(variance)}$$

Continuous-type data:

- Data can take any values from an interval in $\mathbb{R}$.
- **Relative frequency** of being in interval $[a, b]$:

$$\frac{\text{number of data in } [a, b]}{n}.$$

- If data is a sample from a distribution, relative frequency is used to estimate $P(X \in [a, b])$.
- **Density:**

$$\frac{\text{number of data in } [a, b]}{n(b-a)}.$$

  Compare to

$$f(x) = F'(x) = \lim_{\delta \to 0} \frac{P(X \in (x, x+\delta])}{\delta}$$

To plot the density histogram:

1. Compute the minimum and the maximum of data: y_1, y_n.
2. Divide the interval $[y_1, y_n]$ into m (often we take equally spaced) subintervals

$$(c_0, c_1], (c_1, c_2], \ldots, (c_{m-1}, c_m].$$

3. Compute densities d_i on each subinterval $(c_{j-1}, c_j]$.
4. Plot histogram with bases $(c_{j-1}, c_j]$ and height d_j.

- The intervals $(c_{j-1}, c_j]$ are called **class intervals** or **bins**.
- The midpoint $u_j = \frac{c_{j-1}+c_j}{2}$ of the class interval is called class mark $(c_{j-1}, c_j]$.
- Note that

$$\bar{\mu} \approx \sum_j u_j d_j \, (c_j - c_{j-1}) \qquad \text{(empirical approximation)}$$

$$\bar{\sigma}^2 \approx \sum_j (u_j - \bar{\mu})^2 d_j \, (c_j - c_{j-1}) \qquad \text{(variance approximation)}$$

Definition

$X_1, \ldots, X_n$ be a random sample of size n. For any function $u : \mathbb{R}^n \to \mathbb{R}$, the random variable

$$Y = u(X_1, \ldots, X_n)$$

is called a **sample statistic**.

- If a statistic is designed to estimate a quantity (parameter) associated with the underlying distribution (mean,variance, moments, etc) then it is called the **estimator** of that quantity.
- The sample mean is an estimator for the population mean.
- Let $\theta \in \mathbb{R}$ be a quantity associated with the population distribution. If $u : \mathbb{R}^n \to \mathbb{R}$ provides an estimator for θ then, for random sample $X_1, \ldots, X_n$, we denote

$$\bar{\theta} = u(X_1, \ldots, X_n).$$

Definition

Let $\theta \in \mathbb{R}$ be a quantity associated with the population distribution. An estimator $\bar{\theta}$ is called **unbiased** if

$$E[\bar{\theta}] = \theta.$$

- The sample mean (empirical mean) is an unbiased estimator:

$$E[\bar{X}] = \mu.$$

Theorem

$$\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is **not** *an unbiased estimator of the population variance.*

Proof

$$\begin{aligned} E[\sum_{i=1}^{n}\frac{1}{n}(X_i - \bar{X})^2] &= E[\sum_{i=1}^{n}\frac{1}{n}(X_i - \mu - (\bar{X} - \mu))^2] \\ &= E[\sum_{i=1}^{n}\frac{1}{n}\left((X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2\right)] \\ &= E[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - 2\sum_{i=1}^{n}\frac{1}{n}(X_i - \mu)(\bar{X} - \mu) + \frac{1}{n}\sum_{i=1}^{n}(\bar{X} - \mu)^2] \end{aligned}$$

Proof.

$$
\begin{aligned}
&= \sum_{i=1}^{n} \frac{1}{n} E[(X_i - \mu)^2] - 2E[(\bar{X} - \mu)^2] + \frac{1}{n} n E[(\bar{X} - \mu)^2] \\
&= \sum_{i=1}^{n} \frac{1}{n} E[(X_i - \mu)^2] - E[(\bar{X} - \mu)^2] \\
&= \sum_{i=1}^{n} \frac{1}{n} \sigma^2 - \frac{\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n} = \boxed{\frac{n-1}{n} \sigma^2}
\end{aligned}
$$

$\square$

To remove the bias, we define

Definition

Let $X_1, \ldots, X_n$ be a random sample (i.i.d. random variables) with mean μ and variance σ. The random variable

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

is called the **sample variance**.

- Sample variance is an unbiased estimator due to previous theorem.
- The data variance $\bar{s}igma^2$ measures the dispersion of a standalone data set.
- The sample variance S^2 estimates the variance of the population using the sample data.

Thursday class

Parametric estimation

- In practice, we usually assume or model the underlying distribution from which data is sampled to have a certain function that depends on a parameter $\theta = (\theta_1, \ldots, \theta_m) \in \Omega$.
- $\Omega \subset \mathbb{R}^m$ is called the **parameter space**.
- The parameter is unknown but we have sampled data.
- Any estimator of θ is called **point estimator**.

Example

- If the model distribution is normal, then $(\mu, \sigma) \in (-\infty, \infty) \times (0, \infty)$ are parameters.
- For exponential distribution $\lambda \in (0, \infty)$ is a parameter.

Example

- Assume the random variable X has Bernoulli distribution ($X = 1$ or 0).
- The $P(X = 1) = p$ is the unknown parameter.
- We can also write
$$P(X = x) = p^x(1-p)^{1-x}, \quad x = 0, 1.$$

We have sample data
$$\{x_1, \ldots, x_n\}.$$
We want to estimate the $p \in [0, 1]$.

- Note that

$$\begin{aligned} P(X_1 = x_1, \ldots, X_n = x_n) & P(X_1 = x_1) \cdots P(X_n = x_n) \\ &= p^{x_1}(1-p)^{1-x_1} \cdots p^{x_n}(1-p)^{1-x_n} \\ &= p^{\sum\limits_{i=1}^{n} x_i}(1-p)^{n-\sum\limits_{i=1}^{n} x_i}. \end{aligned}$$

Find p as the value for which the probability is the largest. It is called maximum likelihood estimator.

Example (cont.)

Let

$$L(p) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}.$$

- We want to maximize $L(p)$.
- It is equivalent to maximizing

$$\ln L(p) = \sum_{i=1}^{n} x_i \cdot \ln p + (n - \sum_{i=1}^{n} x_i) \cdot \ln(1-p).$$

- To do that:

$$(\ln L(p))' = \sum_{i=1}^{n} x_i \cdot \frac{1}{p} - (n - \sum_{i=1}^{n} x_i) \cdot \frac{1}{1-p} = 0.$$

- Solving from here:

$$p = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

- The maximum likelihood estimator for p in Bernoulli distribution is given by

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

Maximum likelihood estimator

- Let $X_1, \ldots, X_n$ be a random sample from a distribution with pmf or pdf $f(x, \theta)$ where $\theta \in \Omega \subseteq \mathbb{R}^m$.
- We have sample data
$$\{x_1, \ldots, x_n\}.$$

Definition

The function
$$L(\theta) = f(x_1, \theta) \cdots f(x_n, \theta)$$
is called **likelihood function**.

Definition

The value $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m) \in \Omega$ at which the likelihood function takes its maximum value in Ω is called the **maximum likelihood estimator** or **MLE** of θ.

- For many distributions, the maximum likelihood estimator exists and is unique.
- Often it cannot be computed exactly but only approximately.
- In practice, we often instead minimize
$$-\ln L(\theta) = \sum_{i=1}^{n} f(x,\theta)$$

Example

- Suppose the parametric family of distribution is the family of normal distributions

$$f(x, \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{(x-\theta_1)^2}{2\theta_2}}.$$

- $\theta_1 = \mu \in (-\infty, \infty)$ and $\theta_1 = \sigma^2 \in (0, \infty)$.
- $$L(\theta_1, \theta_2) = \left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\theta_1)^2}{2\theta_2}}.$$
- It is equivalent to maximizing

$$\ln L(\theta_1, \theta_2) = -\frac{n}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2.$$

- $$\frac{\partial L(\theta_1, \theta_2)}{\partial \theta_1} = \frac{1}{\theta_2}\sum_{i=1}^{n}(x_i - \theta_1) = \frac{1}{\theta_2}[\sum_{i=1}^{n} x_i - n\theta_1]$$
$$\frac{\partial L(\theta_1, \theta_2)}{\partial \theta_2} = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^{n}(x_i - \theta_1)^2$$

- The maximum likelihood estimators are

$$\theta_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{\mu}$$

$$\theta_2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{\mu})^2 = \bar{\sigma}^2.$$

- **The maximum likelihood estimator does not have to be unbiased.**

Distributions with a random parameter

- Sometimes, the parameters in the distribution can be random variables themselves.

Example

From year to year the height X of Georgia Tech students may be distributed differently: have different mean, variance, etc.

Example

The salaries of employees at various companies are distributed differently.

- Let Θ be a random variable representing the unknown parameter. It has range $\Omega \subseteq \mathbb{R}$ and pdf $h(\theta)$ called **prior**.
- Let $g(x|\theta)$ be the conditional pdf of X given $\Theta = \theta$.
- Let $X_1 = x_1, \ldots, X_n = x_n$ be conditionally sampled data from $g(x|\theta)$ (i.e. this can be salaries of n employees at Georgia Tech).
- We want to estimate the value of θ from these samples.

Bayes estimator for a single data sample

- Assume we have a single data point for now: $X = x$.
- Let $k(\theta|x)$ be the conditional distribution of Θ given $X = x$ ($k(\theta|x)$ provides the likelihood of θ being our unknown parameter if we know the sampled data had the value $X = x$).
- Given $X = x$, the "best" guess for the value of θ is the conditional mean $\theta_B = E[\Theta|x]$.
- It is the value that minimizes

$$g(z) = E[(z - \Theta)^2 \,|X = x]$$

and is the center of the conditional distribution (we have discussed this before).

Definition

$\theta_B = E[\Theta|x]$ is called the **Bayes estimator** of θ.

- If, to find the center, we minimized

$$g(z) = E(|z - \Theta| \,|x)$$

instead then the median of $g(x|\theta)$ would have been the best guess.
- If we took the best guess value to be the value where the likelihood is the largest, i.e. the maximum of $k(\theta|x)$, this would correspond to the maximum likelihood estimator.

- Let $h(\theta)$ be the marginal pdf of Θ.
- Let $g(x|\theta)$ be the conditional pdf given $\Theta = \theta$.
- The joint pdf:

$$f(x,\theta) = g(x|\theta)h(\theta).$$

- The marginal pdf:

$$f_X(x) = \int\limits_\Omega f(x,\theta)\,\mathrm{d}\theta = \int\limits_\Omega g(x|\theta)h(\theta)\,\mathrm{d}\theta$$

- The conditional pdf is also called **posterior**:

$$k(\theta|x) = \frac{f(\theta,x)}{f_X(x)} = \frac{g(x|\theta)h(\theta)}{\int\limits_\Omega g(x|\theta)h(\theta)\,\mathrm{d}\theta}.$$

- The formula for the Bayes estimator:

$$\boxed{\theta_B = E[\Theta|x] = \int\limits_\Omega \theta\, k(\theta|x)\,\mathrm{d}\theta = \frac{\int\limits_\Omega \theta\, g(x|\theta)h(\theta)\,\mathrm{d}\theta}{\int\limits_\Omega g(x|\theta)h(\theta)\,\mathrm{d}\theta}}$$

Bayes estimator for multiple samples

- Let $X_1, \ldots, X_n$ be random variables which, given $\Theta = \theta$, are i.i.d. with pdf $g(x|\theta)$ for any $\theta \in \Omega$.
- We have n sampled values $X = x_1, \ldots, X_n = x_n$, given $\Theta = \theta$.
- The Bayes estimator is given by
$$\theta_B = E[\Theta | X_1 = x_1, \ldots, X_n = x_n].$$
- The joint pdf:
$$f(x_1, \ldots, x_n, \theta) = g(x_1, \ldots, x_n|\theta) h(\theta) = g(x_1|\theta) \cdots g(x_n|\theta)\, h(\theta) = L(\theta)\, h(\theta)$$
where $L(\theta)$ is the **likelihood function**. We used the fact that $X_1, \ldots, X_n$ are independent, given $\Theta = \theta$, to write the $g(x_1, \ldots, x_n|\theta)$ as a product of $g(x_i|\theta)$.
- $$f_X(x_1, \ldots, x_n) = \int_\Omega f(x_1, \ldots, x_n, \theta)\, \mathrm{d}\theta = \int_\Omega L(\theta)\, h(\theta)\, \mathrm{d}\theta$$
- Posterior
$$k(\theta | x_1, \ldots, x_n) = \frac{L(\theta)\, h(\theta)}{\int_\Omega L(\theta)\, h(\theta)\, \mathrm{d}\theta}$$
- $$\boxed{\theta_B = E[\Theta | x_1, \ldots, x_n] = \int_\Omega \theta\, k(\theta | x_1, \ldots, x_n)\, \mathrm{d}\theta = \frac{\int_\Omega \theta\, L(\theta)\, h(\theta)\, \mathrm{d}\theta}{\int_\Omega L(\theta)\, h(\theta)\, \mathrm{d}\theta}}$$

- If we know the prior and the conditional probabilities, we can compute the posterior and find the Bayes estimator.
- The conditional pdf-s are modeled (assumed to have a certain parametric form).
- The prior is typically very hard to know exactly. In Bayesian statistics, the prior is typically guessed or estimated based on prior information about the distribution of Θ.
- If no prior information is known, it is taken to be the uniform distribution (**noninformative prior**).
- The advantage of the Bayesian method is that it uses past knowledge to make a more accurate guess.
- As new data arrives, the Bayesian estimator can be updated to get a better estimator (called r**ecursive Bayesian estimation** or **Bayesian filter**).
- Drawbacks are
  1. The performance depends on how well the prior is chosen.
  2. The integrals in the formula are expensive to compute numerically (see Markov chain Monet Carlo methods).

Example

- Assume salaries are distributed uniformly in $[0, \theta]$ in companies.
- An investigative reporter is trying to guess θ in a given company (CEO's salary).
- He asks n-employees their salaries $x_1, \dots, x_n$.
- If he used MLE, his guess for the salary would have been

$$\bar{\theta} = \max\{x_1, \dots, x_n\}.$$

- But this maybe a wrong estimate with high probability if he sampled low-payed employees.
- But other investigating journalists before him found that salaries of CEO-s have a distribution

$$P(\Theta \leq x) = \begin{cases} 1 - \left(\frac{\theta_0}{\theta}\right)^{\alpha} & \theta \geq \theta_0 \\ 0 & \theta < \theta_0 \end{cases} \implies \boxed{h(\theta) = \frac{\alpha(\theta_0)^{\alpha}}{\theta_0^{\alpha+1}}, \quad \theta_0 > 0, \alpha > 0}.$$

- This is called **Pareto distribution** based on the Pareto principle, which states that a large portion of wealth of CEOs is held by a small fraction of them. θ_0 is the minimum salary.
- Using Bayesian distribution he will find a more accurate estimate of the CEO salary.
- It can be checked that

$$\theta_B = \frac{\alpha + n}{\alpha + n - 1} \max\{\theta_0, x_1, \dots, x_n\}.$$

Example

- The data is sampled from some distribution with known variance and unknown mean.
- The mean $Y = \bar{X}$ is approximately $N(\theta, \frac{\sigma^2}{n})$ due to CLT so we will use it as the conditional probability $g(y|\theta)$.
- Let
$$\bar{\mu} = \frac{x_1 + \cdots + x_n}{n}.$$
- We want to find the Bayes estimator of the mean θ given $Y = \bar{\mu}$ (i.e. finding a Bayes estimator with a single sample).
- Take the prior on θ to be $N(\mu_0, \sigma_0^2)$ where μ_0 and θ_0 are some values we have picked.
-
$$h(\theta) = \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}}$$
$$g(y|\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{-n(y-\theta)^2}{2\sigma^2}}.$$

- $\propto$ means two functions are proportional to each other (equal up to a constant multiple independent of θ).
- Notice that, ignoring the constants,

$$
\begin{aligned}
k(\theta|Y=\bar{\mu}) &\propto g(y|\theta)\,h(\theta) \\
&\propto e^{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}}\,\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{-n(\bar{\mu}-\theta)^2}{2\sigma^2}} \\
&= e^{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}-\frac{n(\bar{\mu}-\theta)^2}{2\sigma^2}} \\
&\propto e^{-\frac{\theta^2-2\theta\mu_0}{2\sigma_0^2}-\frac{n\theta^2-2n\theta\bar{\mu}}{2\sigma^2}} \\
&= e^{-\theta^2\left[\frac{1}{2\sigma_0^2}+\frac{n}{2\sigma^2}\right]+2\theta\left[\frac{\mu_0}{2\sigma_0^2}+\frac{n}{2\sigma^2}\bar{\mu}\right]} \\
&= e^{-\theta^2\frac{1}{2\tau_n^2}+2\theta\frac{\mu_n}{2\tau_n^2}} \\
&\propto \frac{1}{\tau_n\sqrt{2\pi}}e^{-\frac{(\theta-\mu_n)^2}{2\tau_n^2}}
\end{aligned}
$$

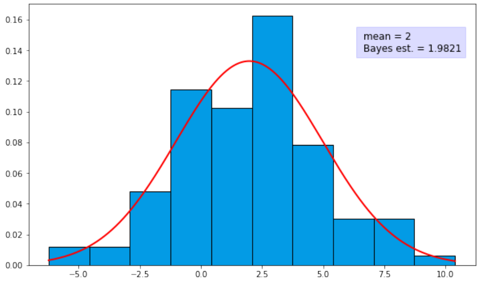
where

$$
\tau_n^2=\left[\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right]^{-1},\quad \mu_n=\tau_n^2\left[\frac{1}{\sigma_0^2}\cdot\mu_0+\frac{n}{\sigma^2}\cdot\bar{\mu}\right].
$$

- $k(\theta|Y = \bar{\mu})$ is the $N(\mu_n, \tau_n^2)$.
- Therefore, the Bayes estimator of the mean is

$$E[\Theta|Y = \bar{\mu}] = \mu_n.$$

- When n is large, it is close to the MLE.



- Data with 300 points sampled from a normal distribution.
- $\sigma = 3$ and $\mu = 2$ in the original distribution.
- Prior is taken to be $N(0, 1)$.
- The density histogram plotted with 10 class intervals.