STAT 390 A
Statistical Methods in Engineering and Science

Week 7 Lectures – Part 2 – Spring 2023

Statistical Models and Point Estimation

Alexander Giessing

Department of Statistics

University of Washington

May 10, 2023

Outline

How to systematically find estimators?

- So far we have justified the use of certain estimators via the WLLN:
  - sample average $\bar{X}_n$ as estimator for mean μ.
  - empirical cdf $F_n(a)$ as estimator for the population cdf $F(a)$.
  - ...
- This approach may seem limited to statistics that can be written as averages. But the WLLN implies the following more general result: for any continuous function $g : \mathbb{R}^m \to \mathbb{R}$ and for all $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|g(\widehat{\mu}_1, \widehat{\mu}_2, \ldots, \widehat{\mu}_m) - g(\mu, \mu_2, \ldots, \mu_m)| > \varepsilon) = 0,$$

  where $\widehat{\mu}_k = n^{-1} \sum_{i=1}^{n} X_i^k$ and $\mu_k = \mathrm{E}[X^k]$.
- How can we use this?

  Note that $\sigma^2 = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = g(\mu, \mu_2)$ for $g(x, y) = y - x^2$.
  $\implies \widehat{\sigma}^2 = \widehat{\mu}_2 - \widehat{\mu}_1^2$ is a "sensible" estimator for the variance.

Method of Moments Estimator

- Suppose that the random sample $X_1, \dots X_n$ has joint pdf or pmf $f(x_1, \dots, x_n; \theta_1, \dots, \theta_m) \equiv f(x; \theta)$.
- Let $\mu_k(\theta) = \mathrm{E}[X^k]$ be the kth **moment** of X and $\widehat{\mu}_k = n^{-1} \sum_{i=1}^n x_i^k$ be the kth **empirical moment**, where x_i is the realization of X_i.
- Note: Typically, kth moment of X is a function of θ, i.e. $\mu_k(\theta)$.

Method of Moments Estimate

Let $x_1, \dots x_n$ be a realization of a random sample with joint pdf or pmf $f(\cdot; \theta)$. The method of moment estimate for $\theta = (\theta_1, \dots, \theta_m)$ is the vector $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_m)$ which solves the m equations

$$\widehat{\mu}_k = \mu_k(\hat{\theta}) \qquad \text{for} \qquad 1 \le k \le m.$$

Example: Menstrual cycles (Revisited)

Suppose that we want to investigate the number of menstrual cylces it takes women to become pregnant, measured from the moment they have decided to become pregnant. We collect a random sample $X_1, \ldots, X_n$*, where* X_i *denotes the number of cyles up to pregnancy of the* i*th woman.*

- X is a waiting time, we decide to model it as $X \sim Geo(p)$ with one unknown parameter $0 < p < 1$.

- Recall $\mathrm{E}[X] = 1/p \equiv \mu_1(p)$. Thus, the method of moment estimator $\hat{p}$ for p solves

$$\frac{1}{n}\sum_{i=1}^{n} X_i \equiv \widehat{\mu}_1 = \mu_1(\hat{p}) \equiv \frac{1}{\hat{p}} \Leftrightarrow \hat{p} = \frac{n}{\sum_{i=1}^{n} X_i}.$$

- **Note:** $\hat{p}$ is the method of moment estimator. if we had a specific data set we could now use the realizations $x_1, \ldots, x_n$ to compute the method of moment estimate.

Example: Defective items in a shipment (Revisited)

Suppose that there is a shipment of N flashlights. An unknown number θN for $0 < \theta < 1$ of the flashlights are defective. It is too expensive to examine all of the flashlights. How can we get information about θ, the percentage of defective flashlights?

- **Idea:** Draw a random sample of n flashlights with out replacement and let X denote the number of defective lights in this random sample. We know that $X \sim HyperGeo(N, \theta N, n)$, i.e.

$$P(X = k) = \frac{\binom{\theta N}{k}\binom{N-\theta N}{n-k}}{\binom{N}{n}},$$

  for $\max\{0, n - N + \theta N\} \leq k \leq \min\{\theta N, n\}$.

- The unknown parameter is $0 < \theta < 1$.

- Let x be the number of defective flashlights in the subsample of size n. Since $\mathrm{E}[X] = n(\theta N)/N = n\theta \equiv \mu_1(\theta)$, the method of moment estimator for θ is

$$x \equiv \widehat{\mu}_1 = \mu_1(\hat{\theta}) \equiv \hat{\theta} n \Leftrightarrow \hat{\theta} = \frac{x}{n}.$$

Example: Gaussian Error Model (Revisited)

Suppose certain measurements can be modeled as $X = \mu + \varepsilon$, *where* $\varepsilon \sim N(0, \sigma^2)$. *Find the MME estimators of* μ *and* σ.

- We have

$$\mathrm{E}[X] = \mu + \mathrm{E}[\varepsilon] = \mu \equiv \mu_1(\mu, \sigma^2)$$
$$\mathrm{E}[X^2] = \ldots = \mu^2 + \sigma^2 \equiv \mu_2(\mu, \sigma^2).$$

- Thus, we have to solve the following system of equations:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \equiv \hat{\mu}_1 = \mu_1(\hat{\mu}, \hat{\sigma}^2) \equiv \hat{\mu}$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \equiv \hat{\mu}_2 = \mu_2(\hat{\mu}, \hat{\sigma}^2) \equiv \hat{\mu}^2 + \hat{\sigma}^2.$$

- We find $\hat{\mu} = \bar{X}_n$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.

Outline

Example: Defective items in a shipment (black market)

Suppose that a dealer of computer chips is offered on the black market two batches of 10,000 chips each. According to the seller, in one batch about 50% of the chips are defective, while in the other batch only 10%. Obviously, the dealer is interested only in the latter batch. The seller offers the dealer to sample and test 10 chips from one of the two batches.

- Dealer samples 10 chips from the second batch and finds that only the fourth chip is defective.
- The dealer decides to buy this batch. Is this the right decision?
- **Yes!** With a batch in which 50% of the chips are defective it is more likely that defective chips will appear in the sub-sample of 10, whereas with a batch in which 10% are defective one would expect fewer defective chips in the sub-sample of 10. The dealer has therefore chosen the batch for which it is most likely that a sub-sample of 10 contains only 1 defective chip.

Maximum Likelihood Principle

Maximum likelihood principle

Given a data set, choose the parameter(s) of interest in such a way that the data are most likely.

- Preceding example:
  - the data set is the sub-sample of 10 chips;
  - the parameter is the percentage θ of defective chips;
  - candidate values for θ are 0.1 and 0.5;
  - the data set "suggests" that the true value of θ is indeed 0.1.
- How can we formalize this principle/ decision making process in a rigorous, mathematical way?

Maximum Likelihood Function

- Suppose that the random sample $X_1, \dots X_n$ has joint pdf or pmf $f(x_1, \dots, x_n; \theta_1, \dots, \theta_m) \equiv f(x; \theta)$.

- Define the **likelihood function** as

$$L(\theta) := L(\theta_1, \dots, \theta_m) = f(x_1, \dots, x_n; \theta_1, \dots, \theta_m).$$

  - This function shows how likely it is to observe the data set $x_1, \dots, x_n$.
  - The likelihood function of a random sample from pdf (or pmf) $f(x; \theta)$ is

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta).$$

- **Note:** Density and likelihood are two sides of the same coin.
  - **density** regards $f(x; \theta)$ as a function of x with known θ (probability).
  - **likelihood** regards $f(x; \theta)$ as a function of θ while data x is given (statistics).

Maximum Likelihood Estimator

Maximum likelihood estimate (MLE)

Let $x_1, \ldots x_n$ be a realization of a random sample with joint pdf or pmf $f(\cdot;\theta)$. The maximum likelihood estimate (MLE) of $\theta \in \Theta$ is the maximizer $\hat{\theta}$ of the likelihood function $L(\theta)$, i.e.

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta).$$

- We will use the same notation to denote the likelihood function evaluated at a realized data set $x_1, \ldots, x_n$ and evaluated at a random sample $X_1, \ldots X_n$.
- The maximum likelihood estimator is the maximizer of the $L(\theta)$ evaluated at a random sample $X_1, \ldots, X_n$.
- **Note:** The MLE $\hat{\theta}$ is also the maximizer of the **log-likelihood**,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \log L(\theta).$$

Example: MLE for the rate parameter of $Exp(\lambda)$

- $X \sim Exp(\lambda)$ with one unknown parameter $\lambda > 0$.
- Given a random sample $X_1, \dots X_n$ the likelihood function is

$$L(\lambda) = \prod_{i=1}^{n} f(X_i; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} X_i} = \lambda^n e^{-\lambda n \bar{X}_n}.$$

- Maximizing the log-likelihood is simpler, since

$$\log L(\lambda) = n \log(\lambda) - \lambda n \bar{X}_n.$$

- We find the MLE $\hat{\lambda}$ by solving the first order condition

$$\frac{d}{d\lambda} \log L(\lambda) = \frac{n}{\lambda} - n\bar{X}_n \overset{!}{=} 0 \implies \hat{\lambda} = \frac{1}{\bar{X}_n} = \frac{n}{\sum_{i=1}^{n} X_i}.$$

Example: Menstrual cycles (Revisited)

- X is a waiting time, we decide to model it as $X \sim Geo(p)$ with one unknown parameter $0 < p < 1$.

- Given a random sample $X_1, \dots X_n$ the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(X_i; p) = \prod_{i=1}^{n} (1-p)^{X_i - 1} p = \left(\frac{p}{1-p}\right)^n (1-p)^{n\bar{X}_n}.$$

- Maximizing the log-likelihood is simpler, since

$$\log L(p) = n \log\left(\frac{p}{1-p}\right) - n\bar{X}_n \log(1-p).$$

- We find the MLE $\hat{p}$ by solving the first order condition (FOC)

$$\frac{d}{dp} \log L(p) = \frac{n}{p(1-p)} + \frac{n\bar{X}_n}{(1-p)} \stackrel{!}{=} 0 \implies \hat{p} = \frac{1}{\bar{X}_n} = \frac{n}{\sum_{i=1}^{n} X_i}.$$

- **Note:** The MLE coincides with the MME estimator; however, this is rather an exception than the rule.

Example: Gaussian Error Model (Revisited)

- Note: $X = \mu + \varepsilon$ has $N(\mu, \sigma^2)$ distribution!
- Given a random sample $X_1, \ldots X_n$ the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \mu, \sigma) = \prod_{i=1}^{n} \frac{e^{-\frac{1}{2}\left(\frac{X_i - \mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2}.$$

- Maximizing the log-likelihood is simpler, since

$$\log L(\theta) = -n\log(\sigma) - n\log(\sqrt{2n}) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2.$$

- We find the MLE $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$ by simultaneously solving the FOC

$$\frac{\partial}{\partial\mu}\log L(\mu, \sigma) = -\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma^2} \stackrel{!}{=} 0$$

$$\frac{\partial}{\partial\sigma}\log L(\mu, \sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i - \mu) \stackrel{!}{=} 0.$$

- We find $\hat{\mu} = \bar{X}_n$ and $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$.