Week 5: Distribution-Free Estimation

STA238: Probability, Statistics, and Data Analysis II

Luis Sierra Muntané

2026-02-03

Recap Quiz

What does a sampling distribution of the mean represent?

a. The distribution of all possible sample means of a specific size taken from a population.

b. The distribution of the entire population from which samples are drawn.

c. The distribution of individual values within a single sample.

d. The spread of data points around the median of a specific sample.

Recap Quiz

According to the Central Limit Theorem (CLT), what happens to the shape of the sampling distribution of the mean as the sample size n increases?

a. It becomes more spread out with higher variance.

b. It becomes approximately normal, regardless of the population's shape.

c. It becomes a uniform distribution.

d. It becomes more skewed to match the population distribution.

Recap Quiz

What is the primary objective of the Maximum Likelihood Estimation (MLE) method?

a. To minimize the variance of the observed data points.

b. To calculate the probability that a specific parameter is correct.

c. To ensure the sample mean is always equal to the population mean.

d. To find the parameter values that make the observed data most probable.

Recap WLLN and CLT

In the context of the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ calculated from iid. samples drawn from a distribution with finite mean:

- The WLLN states that $\bar{X} \xrightarrow{p} \mu$.
- The CLT states that $\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \xrightarrow{d} \mathcal{N}(0,1)$.

Stronger convergence

Need $\mathrm{Var}(X) < \infty$

More information

Also applies to $S_n = \sum_{i=1}^{n} X_i$

Example: MLE

An environmental scientist is measuring the total amount of rainfall (in mm) during storm events in a specific region. Based on historical data, the scientist models the rainfall amount X of a single storm using a Gamma distribution with a known shape parameter $\alpha = 2$ and an unknown rate parameter $\lambda > 0$.

The pdf of a $\Gamma(\alpha, \lambda)$ is

$$f(x_i; \alpha = 2, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i} = \lambda^2 x_i e^{-\lambda x_i}, \quad x_i > 0.$$

The log-likelihood is then given by:

$$\ell(x_i; \lambda) = 2n \ln(\lambda) + \log\left(\prod_{i=1}^{n} x_i\right) - \lambda \sum_{i=1}^{n} x_i.$$

$$f(x_i;\lambda) = \lambda^2 x_i e^{-\lambda x_i}, \quad \log f(x_i;\lambda) = 2\log(\lambda) + \log(x_i) - \lambda x_i$$

$$\ell(x_i;\lambda) = \sum_{i=1}^{n} \log f(x_i;\lambda) = 2n\log(\lambda) + \sum_{i=1}^{n} \log(x_i) + (-1)\sum_{i=1}^{n} x_i$$

$$\frac{\partial}{\partial\lambda}\ell = \frac{2n}{\lambda} + -\sum_{i=1}^{n} x_i = 0 \Longrightarrow \lambda = \frac{2}{\bar{x}}.$$

Example: Cramer Rao Lower Bound

Suppose you are monitoring the wait times between customers at a local coffee shop. You assume these wait times, $X_1, \ldots, X_n$, are iid. random variables following an Exponential distribution with rate parameter θ:

$$f(x;\theta) = \theta e^{-\theta x}, \quad x \geq 0, \theta > 0.$$

The log-likelihood is given by:

$$\ell(x_i;\theta) = \sum_{i=1}^{n} \log f(x_i;\theta) = \sum_{i=1}^{n} \log\theta - \theta \sum_{i=1}^{n} x_i$$

$$\frac{\partial}{\partial\theta}\ell(X_i;\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} X_i, \quad \frac{\partial^2}{\partial\theta^2}\ell(X_i;\theta) = \frac{-n}{\theta^2} \longrightarrow \mathbb{E}\frac{\partial^2}{\partial\theta^2}\ell(X_i;\theta) = \frac{-n}{\theta^2} = -I(\theta)$$

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{n\,I(\theta)} = \frac{\theta^2}{n}, \quad I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right]$$

$\longrightarrow \hat{\theta}_{MLE}$ is the most efficient estimator possible in this case.

Model Mispecification

A class takes a test with $n = 10$ questions. The score of each student X_i follows a Binomial distribution $\mathbf{Bin}(10, p)$.

We collect data from the class grades and find the average score is $\bar{x} = 8$.

Assume we use a $\mathbf{Pois}(\lambda)$ model. We can show the Poisson MLE for the rate parameter λ is $\hat{\lambda}_{\mathbf{MLE}} = 8$.

Our model for the grades would then be $P(X = k) = \frac{e^{-8}8^k}{k!}$

Outcome	Real World (Binomial)	Misspecified Model (Poisson)
Score of 8	≈30%	≈14% (Underestimates the peak)
Score of 10	≈10%	≈10% (Accidentally close)
Score ≥ 11	0%	≈18%

Model Mispecification and the MLE

Suppose we have some data $x = \{10, 11, 25\}$ and assuming it comes from a normal distribution $\mathcal{N}(\theta, \sigma^2)$, we want to estimate the mean parameter θ.

- The MLE of a Gaussian is given simply by $\bar{x} = \frac{1}{3}(10 + 11 + 25) = \underline{15.33}$.

If the underlying model was actually that of a Laplace distribution, with pdf

$$f(x; \theta) = \frac{1}{2b} \exp\left(-\frac{|x - \theta|}{b}\right),$$

then the MLE for θ would actually be given by:

$\log f(x;\theta) = \log \frac{1}{2b} - \frac{|x-\theta|}{b}$, $\max \ell(x;\theta) = \max\left\{\log \frac{1}{2b} - \frac{1}{b}\sum_{i=1}^{n} |x_i - \theta|\right\}$

$\max\left\{-\frac{1}{b}\sum_{i=1}^{n} |x_i - \theta|\right\} = \min_{\theta}\left\{\sum_{i=1}^{n} |x_i - \theta|\right\} \longrightarrow \hat{\theta} = \text{med}\{x_i\} = 11.$

Main limitation of MLE

Besides not meeting the assumptions required for the MLE, its biggest limitation in real-world applications is that of requiring **distributional assumptions**.

If we assume our true model is in

$$\{P(\theta) \,:\, \theta \in \Theta\}\,,$$

when in reality it isn't, the maximum of the likelihood need not be close to the true parameter value. Furthermore, the predictions we will make from it will be far away from the true values.

Method of Moments

A distribution-free method of estimating a vector of parameters $(\theta_1, \ldots, \theta_k)$.

Moments of a distribution

Let F be a distribution whose first k moments are well defined and given by

$$\begin{aligned}\mu_1 &= \mathbb{E}[X] \\ \mu_2 &= \mathbb{E}[X^2] \\ &\vdots \\ \mu_k &= \mathbb{E}[X^k]\end{aligned}$$

> The first moment $\mu_1 = \mathbb{E}[X]$ is the mean, while the variance would be $\mu_2 - \mu_1^2$.

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Sample Moments

Let $X_1, \ldots, X_n \sim F$ be a random sample, then the first k *sample moments* are given by

$$m_1 = \mathbb{E}_{F_n}[X] = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$m_2 = \mathbb{E}_{F_n}[X^2] = \frac{1}{n}\sum_{i=1}^{n} X_i^2,$$

$$\vdots$$

$$m_k = \mathbb{E}_{F_n}[X^k] = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Method of Moments

The *Method of Moments* (MoM) produces estimators $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ by solving the system of equations $\mu = f^{-1}(\theta)$ and plugging in the sample moments:

$$\begin{aligned}
\hat{\theta}_1 &= f_1(m_1, \ldots, m_k), \\
\hat{\theta}_2 &= f_2(m_1, \ldots, m_k), \\
&\vdots \\
\hat{\theta}_k &= f_k(m_1, \ldots, m_k).
\end{aligned}$$

These equations come from the relationship between the parameters $\theta_1, \ldots, \theta_k$ and the moments $\mu_1, \ldots, \mu_k$. Eg. for a normal distribution $\begin{cases} \mu = \mu_1 \\ \sigma^2 = \mu_2 - \mu_1^2 \end{cases}$.

Method of Moments

All of these statistics m_i define unbiased estimators for the population moments μ_i.

- By the WLLN, $m_i \xrightarrow{p} \mu_i$.
- By the CLT, $\frac{m_i - \mu_i}{\sqrt{\text{Var}(m_i)}} \xrightarrow{d} \mathcal{N}(0,1)$, provided $\text{Var}(m_i) < \infty$. $\Longrightarrow$ need μ_{2i} to be well-defined

Then, for the map $\theta_i = f_i(\mu_1, \ldots, \mu_k)$ we estimate

$$\hat{\theta}_i = f_i(m_1, \ldots, m_k), \quad i = 1, 2, \ldots, k.$$

- These approximate inferences are appropriate for every distribution in the model.

> Sometimes, the different methods MLE and MoM can result in the same final expression/rule for the estimator!

Example: MoM I

A city transit authority wants to estimate the maximum possible delay for a specific bus route. They assume the delay time (in minutes), X, follows a Uniform distribution on the interval $[0, \theta]$. How can we obtain an estimator for θ using the Method of Moments?

$\mu_1 = \mathbb{E}[X] = \frac{\theta}{2}$, $m_1 = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, replacing μ_1 for m_1:

$\frac{\hat{\theta}}{2} = m_1$, $\hat{\theta}_{MM} = 2\bar{X}$. (unbiased)

$\hat{\theta}_{MLE} = \max X_i = X_{(n)}$, which was biased.

Example: MoM II

A factory produces "10-gram" weights. Because of slight variations in the machinery, the actual weight of each piece, X, follows a Normal distribution $\mathcal{N}(\mu, \sigma^2)$. Estimate both the average weight μ and the variance σ^2 from a random sample $X_1, \ldots, X_n$.

$\mathbb{E}[X] = \mu_1$, $\longrightarrow$ $\bar{X} = \mu_1 = m_1$

$\mathbb{E}[X^2] = \sigma^2 + \mu_1^2$, $\longrightarrow$ $\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \sigma^2 + \mu_1^2 = \mu_2$

$\sigma^2 = \mu_2 - \mu_1^2$, $\longrightarrow$ $\hat{\sigma}^2_{MM} = m_2 - m_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$

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

Method of Moments: Properties

- It produces unbiased estimators for the population moments μ_i that are also (under the right conditions) consistent.
- There is no guarantee that the functions of these moments $\theta = f(\mu)$ will be themselves unbiased (f can be strictly convex/concave).
- Still, $\hat{\theta}_i = f_i(\mu_1, \ldots, \mu_k)$ will converge in distribution to a normal with mean θ_i whenever the first $2k$ moments exist.
- Beyond our toy examples, it can be hard to solve the system of equations algebraically.

Summary

- It is not necessary to make strong distributional assumptions to estimate all parameters.
- The Method of Moments provides a way of constructing estimators for the moments of a distribution without assuming a particular model.
- We still need to make *some* assumptions.