Statistical Estimation

EES 4891/5891

Probability & Statistics for Geosciences

Jonathan Gilligan

Class #12: Thursday, February 13 2025

Learning Goals

Learning Goals

- Learn about the method of moments and its limitations
- Learn how to use the method of maximum likelihood estimation
- Learn the properties of estimation:
  - Accuracy and bias
  - Precision
- Learn about the tradeoffs between bias and variance
- Today is very mathematical
  - Don’t worry about the proofs and derivations
  - Focus on why the results are important

Statistical Estimation

Introduction

- Context:
  - We've learned about statistical distributions
  - Parametric distributions:
    - Probability mass or density can be written as a function with some *parameters*:
      - Normal: $\mathcal{N}(\mu, \sigma)$
      - Binomial: $\mathcal{B}(n, p)$
      - Poisson: $\text{Poisson}(\lambda)$
      - Gamma: $\text{Gamma}(k, \theta)$,
        - $k = \text{shape}, \theta = \text{scale}$
      - Weibull: $\mathcal{W}(k, \theta)$
      - ...
  - Given the parameters, we know how to generate a random sample from the distribution
    - `rnorm(N, mu, sigma)`, `rbinom(N, n, p)`, `rpois(N, lambda)`, `rgamma(N, shape = k, scale = theta)`, ...

- The problem:
  - Given N points $\mathbf{X} = x_1, x_2, \ldots, x_N$ sampled from a distribution $\mathbb{P}(x, \theta_1, \theta_2, \ldots)$, with parameters $\theta_1, \theta_2, \ldots$, estimate the parameter values $\theta_1, \theta_2, \ldots$
  - Point vs. Interval Estimation
    - **Point estimate:** The most likely value for θ_i
    - **Interval estimate:** A range of values for θ_i, where we are confident there's a certain probability (e.g., 95%) that the true value of θ lies within the interval.
  - Today we focus on point estimation.

Method of Moments

- Not very reliable, but easy to work
- Definitions:
  - k^{th} moment

$$\mu_k = E(x^k) \approx \hat{\mu}_k = \frac{1}{N}\sum_{i=1}^{N} x_i^k$$

  - 1st moment:

$$\mu_1 = E(x) \approx \hat{\mu}_1 = \frac{1}{N}\sum_{i=1}^{N} x_i$$

  - 2nd moment:

$$\mu_2 = E(x^2) \approx \hat{\mu}_2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2$$

  - μ_k is the true value, $\hat{\mu}_k$ is an approximation based on N observations

- Method:
  1. Write the parameter as a function of the moments μ_k
  2. Substitute the estimates $\hat{\mu}_k$ to estimate the parameter

Example

- There are θ balls in a jar, and you draw n balls and try to estimate θ

$$\mu_1 = E(x) = \sum_{i=1}^{\theta} i \times P(x = i) = \sum_{i=1}^{N} i \times \frac{1}{\theta}$$

$$= \frac{1}{\theta} \sum_{i=1}^{\theta} i = \frac{1}{\theta} \frac{\theta(\theta + 1)}{2} = \frac{\theta + 1}{2}$$

$$\theta = 2\mu_1 - 1 \approx 2\hat{\mu}_1 - 1$$

- Try this in R

```
theta <- 14
N <- 5
x <- sample(1:theta, N)
print(x)
```

```
## [1] 13  4  9  1  3
```

```
mu_1 <- mean(x)
print(2 * mu_1 - 1)
```

```
## [1] 11
```

  - There's a problem: We estimate that $\theta = 11$, but we drew a ball with 13.

Maximum Likelihood Estimation

Overview

- Likelihood $L(x|\theta)$ is the conditional probability of observing x if the parameter θ has a certain value.
  - We often say it's the probability of x, given θ.
- The big idea is that if we have observations $\mathbf{X} = x_1, x_2, \ldots, x_N$, the best estimate for θ is the value that has the largest likelihood $L(\mathbf{X}|\theta)$
- If $x_1, x_2, \ldots, x_N$ are **iid** observations (*independent, identically distributed*), then

$$L(x_1, x_2, \ldots x_N|\theta) = \prod_{i=1}^{N} L(x_i|\theta)$$

and

$$\ell(x_1, x_2, \ldots x_N|\theta) = \sum_{i=1}^{N} \ell(x_i|\theta),$$

where

$$\ell(x|\theta) = \log(L(x|\theta))$$

  - It's much easier to add numbers than to multiply them, so we often work with the log-likelihood ℓ instead of L

Example of Maximum Likelihood Estimation

- Suppose $x_1, x_2, \ldots, x_N$ are drawn from a normal distribution $\mathcal{N}(\mu, \sigma)$, with unknown parameters
- The log-likelihood is

$$\ell(x_1, x_2, \ldots, x_N | \mu, \sigma) = \sum_{i=1}^{N} \ell(x_i | \mu, \sigma)$$

where

$$\begin{aligned}\ell(x_i | \mu, \sigma) &= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right) \\ &= -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{(x_i-\mu)^2}{2\sigma^2},\end{aligned}$$

so

$$\ell(x_1, x_2, \ldots, x_N | \mu, \sigma) = -N\left(\frac{\log(2\pi)}{2} + \log(\sigma)\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2$$

Finding the Maximum Likelihood

- When a function reaches its maximum or minimum values, the derivative is zero, so find the value of θ the makes the derivative of ℓ zero:

$$\frac{\partial \ell(\mathbf{X}|\mu,\sigma)}{\partial \theta} = -\frac{1}{\sigma^2}\sum_{i=1}^{N} x_i - \mu,$$

which is zero if

$$0 = \sum_{i=1}^{N}(x_i - \mu) = \left(\sum_{i=1}^{N} x_i\right) - N\mu$$

$$\hat{\mu}_{\text{MLE}} = \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x}$$

  - So the maximum-likelihood estimate of μ is just $\bar{x}$, the mean of $x_1, x_2, \ldots, x_N$.

Using Maximum Likelihood for Other Distributions

- Maximum Likelihood isn’t always as easy as it is for the normal distribution.
- Computers can find maximum likelihood estimates for most distributions
  - `fitdistr()` function from the `MASS` package
  - `mle()` function from the `stat4` package
  - More about this on Tuesday

Properties of Maximum Likelihood

- The maximum-likelihood estimate of $\hat{\mu}$ is $\overline{x}$
- The maximum-likelihood estimate of $\hat{\sigma}^2$ is $\frac{1}{N}\sum_i (x_i - \overline{x})^2$, the variance of the sample (details in the textbook)
- Accuracy of estimates:
  - Define $\text{bias}(\hat{\theta}, \theta) = E(\hat{\theta}) - \theta$.
  - For accuracy, we want bias to be as close to zero as possible
- Precision:
  - We want the variance of $\hat{\theta}$ to be as small as possible

- Mean-Squared Error (MSE)

$$\text{MSE}(\hat{\theta}, \theta) = E\left[(\hat{\theta} - \theta)^2\right]$$

- We want our estimator to have the smallest possible MSE.

Bias-Variance Decomposition

- Examine the MSE:

$$
\begin{aligned}
\mathrm{MSE}(\hat{\theta}, \theta) &= E\left[(\hat{\theta} - \theta)^2\right] \\
&= E\left[(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2\right] \\
&= E\left[(\hat{\theta} - E(\hat{\theta}))^2 + (E(\hat{\theta}) - \theta)^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)\right] \\
&= \underbrace{E\left[(\hat{\theta} - E(\hat{\theta}))^2\right]}_{\text{variance}} + \underbrace{E\left[(E(\hat{\theta}) - \theta)^2\right]}_{\text{bias}^2} + \underbrace{2E\left[\hat{\theta} - E(\hat{\theta})\right](E(\hat{\theta}) - \theta)}_{=0} \\
&= V(\hat{\theta}) + \text{bias}^2
\end{aligned}
$$

- There is a trade-off: Making bias smaller generally makes variance larger and vice-versa.

Example: Normal Distribution

- Maximum-Likelihood Estimators:
  - $\hat{\mu} = \bar{x}$
  - $\hat{\sigma}^2 = \frac{1}{N} \sum_i (x_i - \bar{x})^2$
- $\hat{\mu}$ is unbiased:

$$\begin{aligned}
E(\hat{\mu}) &= E(\bar{x}) = \mu \\
V(\hat{\mu}) &= V(\bar{x}) = \frac{\sigma^2}{N} \\
\text{bias}(\hat{\mu}, \mu) &= E(\hat{\mu}) - \mu \\
&= \mu - \mu \\
&= 0
\end{aligned}$$

- $\hat{\sigma}^2$ is biased:

$$\begin{aligned}
E(\hat{\sigma}^2) &= \frac{N-1}{N}\sigma^2 \quad \text{(see textbook)} \\
V(\hat{\sigma}^2) &= \frac{2(N-1)}{N^2}\sigma^4 \quad \text{(see textbook)} \\
\text{bias}(\hat{\sigma}^2, \sigma^2) &= E(\hat{\sigma}^2) - \sigma^2 \\
&= \frac{N-1}{N}\sigma^2 - \sigma^2 \\
&= \frac{-1}{N}\sigma^2
\end{aligned}$$

Bias-Variance Tradeoff

- Suppose we choose an *unbiased estimator* for σ^2:
  - MLE estimate:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{N} \sum_i (x_i - \bar{x})^2$$

  - Unbiased estimate:

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_i (x_i - \bar{x})^2 = \frac{N}{N-1} \hat{\sigma}^2_{\text{MLE}}$$

- Then:

$$\begin{aligned} E(\hat{\sigma}^2) &= \sigma^2 \\ \text{bias}(\hat{\sigma}^2, \sigma^2) &= E(\hat{\sigma}^2) = \sigma^2 \\ &= \sigma^2 - \sigma^2 \\ &= 0 \end{aligned}$$

- Now look at MSE:

$$\text{MSE}\left(\hat{\sigma}^2, \sigma^2\right) = V\left(\sigma^2\right) = \frac{2}{N-1}\sigma$$

But

$$\text{MSE}\left(\hat{\sigma}^2_{\text{MLE}}\right) = \frac{2N-1}{N^2}\sigma^2$$

And

$$\frac{2}{N-1} > \frac{2N-1}{N^2},$$

So the unbiased estimate gives a greater MSE because the variance increases by more than the original bias.

- Bias-variance tradeoff applies to all estimates, no just MLE

Cramér-Rao Bound

- From information theory, every unbiased estimator has a minimum possible variance.

$$\mathsf{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)},$$

  Where $I(\theta)$ is the Fisher information

$$I(\theta) = \mathsf{E}\left[\left(\frac{\partial \ell(x|\theta)}{\partial \theta}\right)^2\right]$$
$$= -\mathsf{E}\left[\frac{\partial^2 \ell(x|\theta)}{\partial \theta^2}\right]$$

- We can then define the *efficiency* of an estimator: How close is the variance to this lower bound:

$$e(\hat{\theta}) = \frac{1/I(\theta)}{\mathsf{Var}(\hat{\theta})} \leq 1$$

  - The Cramér-Rao bound sets a limit on how small the variance can be, and therefore how good the efficiency can be, so $e(\hat{\theta}) \leq 1$.
  - The MLE estimator has efficiency of 1.
    - It's the best possible point-estimator