BIOSTAT 701

Introduction to Statistical Theory and Methods I

Lynn Lin

Preparation

- **Likelihood function**: https://duke.zoom.us/rec/share/ebl8UloqK0Jbs_NJ2Zbrhs7FjO8iIRNpkfXEuSUTUue3wjWd5erds6oKyBC2SYb9.jvbiPIMr8A3JPP3M?startTime=1649272987000

Introduction to inference

- Up until now, we have considered RVs, their distributions, and some properties of those distributions.
- Importantly, in actual practice distributions (or distributional forms) are selected based on the characteristics of the population under study.
- E.g., if the distribution of LDL cholesterol values appears to be approximately bell-shaped then the normal distribution is a natural choice to model these data.
- Distributions can be assigned based on empirical considerations (e.g., the data appear normal), theoretical considerations (e.g., LDL values are determined by a large number of small, independent perturbations), or both.
- From now on, we assume that the distributions in question have been well chosen.

Introduction to inference

- The linkage between actual data and statistical inference is through the parameters of a well-chosen distribution that is used as a model for the data.
  - E.g., for LDL, assumed to be normally distributed.
  - But μ and σ are typically unknown. They are called parameters. A parameter is a number describing a whole population.
  - The goal is then to identify the true but unknown μ and σ values.
  - This is call point estimation
  - The idea is to use data (statistic) to “guess” the value of the (unknown) parameters which is hopefully close to the true values. A statistic is a descriptive measure of a sample.

Introduction to inference

- Populations have parameters; Samples have statistics.
- Statistical inference is about how and what can we infer about the population's parameters by using the sample's statistics.

Likelihood

- The likelihood is the PMF or PDF thought of as a function of parameters (rather then as a function of data)
  - $L_x(\theta) = f_\theta(x)$, where θ denote the (unknown) parameter(s) of the distribution
  - Since it is a function of θ (not x), for an observed sample, it gives the “likelihood” or “plausibility” of various parameter values.

Likelihood

- E.g., for the binomial PMF $f_p(x) = \binom{n}{x} p^x (1-p)^{(n-x)}$, where $\theta = p$ and $\theta \in \Theta \equiv [0,1]$, Θ is called the parameter space.
  - Sample space S: the set of possible data values
  - $L_x(p) = \binom{n}{x} p^x (1-p)^{(n-x)}$
  - Or, equivalently by dropping the multiplicative term that does not contain the parameter $L_x(p) = p^x (1-p)^{(n-x)}$

Likelihood

- What is the likelihood for the Poisson PMF $f_\lambda(x) = \dfrac{\lambda^x}{x!} e^{-\lambda}$?

Likelihood

- If X is discrete, $L_x(\theta) = P_\theta(X = x)$. Consider the likelihood at 2 parameter points, θ_1 and θ_2. If $L_x(\theta_1) > L_x(\theta_2)$, then $P_{\theta_1}(X = x) > P_{\theta_2}(X = x)$, implying that θ_1 is a more plausible value for the true value of θ than θ_2 for the observed data.

- If X is continuous, then for small ϵ, $P_\theta(x - \epsilon < X < x + \epsilon) \approx 2\epsilon f_\theta(x) = 2\epsilon L_x(\theta)$

- Then $\dfrac{P_{\theta_1}(x - \epsilon < X < x + \epsilon)}{P_{\theta_2}(x - \epsilon < X < x + \epsilon)} \approx \dfrac{L_x(\theta_1)}{L_x(\theta_2)}$ provides an approximate comparison of the probability of the observed sample under 2 parameter values.

Likelihood principle

- Let x and y are 2 sample points such that $L_x(\theta) = h(x, y)L_y(\theta)$.
- Then the same inference for θ should be drawn from x and y.

Parametric point estimation

- Observed data $x_1, \dots, x_n$ regarded as outcomes/realizations of RVs $X_1, \dots, X_n$.
  - E.g., tossing a coin n times
  - $S = \{0,1\}^n$
  - $X_i = \begin{cases} 1 & x \text{ if the ith toss is H} \\ 0 & x \text{ if the ith toss is T} \end{cases}$

Parametric point estimation

- Statistical model: joint distribution of $X_1, \ldots, X_n$.
    - Suppose $X_1, \ldots, X_n \sim F_\theta$, where θ is unknown.
    - We denote F_θ the joint distribution with (unknown) parameter θ.
    - Thus, the joint distribution of $X_1, \ldots, X_n$ belongs to some parametric model and $\theta \in \Theta$ represents the unspecified part of model.
    - E.g., $F_\theta = f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i)$ by assuming independence among $X_1, \ldots, X_n$.

Parametric point estimation

- If $X_1, \ldots, X_n \sim f_\theta$ iid (independently and identically distributed), then the likelihood function is $L_x(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$, treated as a function of θ.

Parametric point estimation

- We can interpret the likelihood as ranking all the possible θ values in terms of how well the corresponding model fits the observed data.
- The larger the likelihood the better the model fits the data.
- Maximum likelihood estimator (MLE)
- Definition: for a given observed data x, let $\hat{\theta}(x)$ be a value of the parameter space Θ at which the likelihood function $L_x(\theta)$ attains its maximum. The statistic $\hat{\theta}(x)$ is called a MLE of θ.

Estimator

- Any statistic used to estimate the value of some known function of unknown parameter θ, say $\tau(\theta)$, is called an estimator of $\tau(\theta)$.
- An observed value of the statistic is called an estimate of $\tau(\theta)$.
- An estimator is a function of RVs $X_1, \ldots, X_n$, while an estimate is a function of observed values $x_1, \ldots, x_n$.

Finding MLEs

- Direct maximization: Examine the likelihood directly to determine which value of θ maximizes $L_x(\theta)$.
- E.g., let $X_1, \ldots, X_n$ be independent uniform RVs on the interval $[0,\theta]$, where $\theta > 0$.
- $L_x(\theta) = \dfrac{1}{\theta^n} \prod_{i=1}^{n} 1_{[0,\theta]}(x_i) = \dfrac{1}{\theta^n} 1_{[x_{(n)},\infty)}(\theta)$.
- If $\theta < x_{(n)}$, $L_x(\theta) = 0$. If $\theta \geq x_{(n)}$, $L_x(\theta)$ is a decreasing function of θ. $L_x(\theta)$ is maximized at $\theta = x_{(n)}$. Thus, $X_{(n)}$ is the MLE of θ.

Finding MLEs

- Likelihood equations: If the support of $f(x \mid \theta_1, \ldots, \theta_p)$ does not depend on $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, and $L_x(\boldsymbol{\theta})$ is differentiable w.r.t. $\boldsymbol{\theta}$, then an MLE will be a solution of the likelihood equations
  - $\dfrac{\partial}{\partial \theta_j} L_x(\theta_1, \ldots, \theta_p) = 0$, for j = 1,…,p
- If is often easier to differentiate $\log L_x(\boldsymbol{\theta})$, known as log-likelihood function.

Finding MLEs

- Solutions to likelihood equations are only positive candidates for an MLE.
  - Points are with the 1st order partial derivative are zero may be local/global minima, local/global maxima, or saddle points.
  - 1st order partial derivatives may not be zero if extrema occur on boundary. Therefore, boundary much be checked separately.
- In maximum likelihood estimation, our job is to find a global maximum.

Example

- Y has a Poisson distribution with unknown parameter $\lambda \geq 0$. What is the MLE for λ?

Example

- Y has a Poisson distribution with unknown parameter $\lambda \geq 0$. What is the MLE for λ?
- First, collect data from independent trials:
  - $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$

Example

- Y has a Poisson distribution with unknown parameter $\lambda \geq 0$. What is the MLE for λ?
- First, collect data from independent trials:
  - $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$
- **Likelihood**: $L_{x_{1:n}}(\lambda) = \prod_{i=1}^{n} f_\lambda(x_i) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{x_1+\cdots x_n}}{x_1!\cdots x_n!}$

Example

- Y has a Poisson distribution with unknown parameter $\lambda \geq 0$. What is the MLE for λ?
- First, collect data from independent trials:
  - $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$
- **Likelihood**: $L_{x_{1:n}}(\lambda) = \prod_{i=1}^{n} f_\lambda(x_i) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda} \lambda^{x_1 + \cdots x_n}}{x_1! \cdots x_n!}$
- **Log likelihood** (easier to be maximized):
  $\log L = -n\lambda + (x_1 + \cdots x_n) \log \lambda - \log(x_1! \cdots x_n!)$

Example

- **Critical point**: solve $d(\log L)/d\lambda = 0$ ==> $-n + (x_1 + \cdots + x_n)/\lambda = 0$ ==> $\lambda = \bar{x}$
- **Check 2nd derivative is negative**: $-(x_1 + \cdots + x_n)/\lambda^2 < 0$
  - So it is a max unless $x_1 + \cdots + x_n = 0$
  - $X_1 = x_1, X_2 = x_2, \ \cdots, X_n = x_n$
- **Boundary for range $\lambda \geq 0$**: Check $\lambda \to 0^+$ and $\lambda \to \infty$. Both let $\log L \to -\infty$. So $\lambda = \bar{x}$ gives the max.

Example

- The exceptional case is when $x_1 + \cdots + x_n = 0$
    - Giving $x_1 = x_2 = \cdots x_n = 0$
    - In this case, $\log L = -n\lambda + 0 \log \lambda - \log(0! \cdots 0!) = -n\lambda$
- On the range $\lambda \geq 0$, this is maximized at $\hat{\lambda} = 0$, which agrees with the main formula $\hat{\lambda} = \bar{x}$.

Example

- Let $X_1, \ldots, X_n$ be RVs from $N(\mu, 1)$, where $-\infty < \mu < \infty$
- What is the MLE for μ?

Example

- Let $X_1, \ldots, X_n$ be RVs from $N(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma > 0$.
- What is the MLE for μ and σ?

Example

- Let $X_1, \ldots, X_n$ be RVs from $N(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma > 0$.
- Log Likelihood: $\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$
- Likelihood equations:
  - $\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$
  - $\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$

Example

- Solving, we obtain
    - $\mu = \bar{x}$
    - $\sigma^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$

Theorem

- Let $g(\cdot, \cdot)$ be a function of 2 variables, for which 1st and 2nd-order partial derivatives are continuous in a neighborhood of (x_0, y_0). Then $g(\cdot, \cdot)$ has a maximum at (x_0, y_0) if following conditions are satisfied:
  - The 1st-order partial derivatives are zero: $\frac{\partial g(x, y)}{\partial x}|_{x=x_0, y=y_0} = 0$ and $\frac{\partial g(x, y)}{\partial y}|_{x=x_0, y=y_0} = 0$
  - At lease one 2nd-order partial derivative is negative: $\frac{\partial^2 g(x, y)}{\partial x^2}|_{x=x_0, y=y_0} < 0$ or $\frac{\partial^2 g(x, y)}{\partial y^2}|_{x=x_0, y=y_0} < 0$
  - The determinant of the matrix of 2nd-order partial derivatives (Jacobian) is positive at (x_0, y_0).