STA 532: Theory of Statistical Inference

Robert L. Wolpert
Department of Statistical Science
Duke University, Durham, NC, USA

3 Parametric Inference I

We now turn attention to statistical models in which the family $\mathfrak{F}$ of possible pdfs for the observable $X \in \mathcal{X}$ are a k-dimensional parametric family $\mathfrak{F} = \{f(x \mid \theta) : \theta \in \Theta\}$ for some parameter space $\Theta \subseteq \mathbb{R}^k$ and function $f : \mathcal{X} \times \Theta \to \mathbb{R}_+$. Examples include the Poisson distribution $\mathsf{Po}(\theta)$ with $\Theta = \mathbb{R}_+$ and the $\mathsf{Be}(\alpha, \beta)$ distribution with $\theta = (\alpha, \beta) \in \Theta \subset \mathbb{R}_+^2$. Other examples include the univariate normal distribution $\mathsf{No}(\mu, \sigma^2)$, with $k = 2$ and $\Theta = \mathbb{R} \times \mathbb{R}_+$ with $\theta = (\mu, \sigma^2)$, and the p-dimensional multivariate normal distribution $\mathsf{No}(\mu, \Sigma)$ with $k = p(p+3)/2$-dimensional parameter $\theta = (\mu, \Sigma)$, with mean vector $\mu \in \mathbb{R}^p$ and $p \times p$ positive-definite covariance matrix $\Sigma \in \mathcal{S}_+^p$.

3.1 Change of Parameters

Any k-dimensional parametric family $\mathfrak{F} = \{F(\cdot \mid \theta) : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}^k$ can also be written as $\mathfrak{F} = \{G(\cdot \mid \phi) : \phi \in \Phi\}$ for a different parameter ϕ, also k-dimensional, with $\phi = H(\theta)$, $\theta = H^{-1}(\phi)$ for any invertible 1:1 transformation $H : \Theta \to \Phi$. The statistician is free to use whichever parameterization is most convenient. For example, we will use the "shape/rate" parameterization for the Gamma distribution $\mathsf{Ga}(\alpha, \lambda)$, with mean α/λ and pdf on the left in (1), while some authors use the "shape/scale" parameterization in which the $\mathsf{Ga}(\alpha, \beta)$ distribution has mean $\alpha\beta$ and the pdf given on the right. The two are related by the invertible transformation $H(\alpha, \lambda) = (\alpha, \beta = 1/\lambda)$, while their pdfs at $x > 0$ are

$$\frac{\lambda^\alpha \, x^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x} = \frac{x^{\alpha-1}}{\beta^\alpha \, \Gamma(\alpha)} e^{-x/\beta}. \tag{1}$$

3.2 Likelihood Functions

When evaluated at the observed value $X = x$ of the data, and viewed as a function only of $\theta \in \Theta$, the function

$$\mathcal{L}(\theta \mid x) \propto f(x \mid \theta) \tag{2}$$

(or any positive multiple of it) is called the "likelihood function" (LH). Some have argued that *all* inference about $\theta \in \Theta$ should depend on the sampling model and the data $X = x$ *only* through this function (Birnbaum, 1962; Berger and Wolpert, 1988), a proposition known as the Likelihood Principle (LP). As we consider various ways of estimating or making other inference about θ below, try to see which ones are consistent with LP and which aren't.

3.3 Parameters of Interest & Nuisance Parameters

Often some feature of θ is "interesting", *i.e.*, of use in a particular analysis, while the rest of θ is uncertain but not interesting. For example, in $\mathsf{No}(\mu, \sigma^2)$ problems, often it is only the mean that is interesting and not the variance. We can always write the parameter vector in the form $\theta = (\eta, \lambda)$ with "parameter of interest" η and "nuisance parameter" λ.

Fisherian and Bayesian practitioners differ in how they dispose of the λ in order to make inference about η (Berger et al., 1999). Typically the Fisherian approach is to replace the LH of (2) with the "profile likelihood"

$$\mathcal{L}_P(\eta \mid x) = \sup_{\lambda} f\big(x \mid (\eta, \lambda)\big) \tag{3a}$$

while the Bayesian approach is to determine a "conditional prior distribution" $\pi(\lambda \mid \eta)$ quantifying the plausibility of various possible values of the nuisance parameter λ, possibly depending on the parameter of interest η, and then base inference about η on the "marginal likelihood"

$$\mathcal{L}_M(\eta \mid x) = \int f(x \mid \eta, \lambda)\ \pi(\lambda \mid \eta)\, d\eta \tag{3b}$$

We'll see examples below, once we've introduced and compared estimation for these two approaches.

3.4 Methods of Estimation

Here we present four possible ways of estimating an uncertain parameter $\theta \in \Theta$ on the basis of a sequence $T_n(X)$ of statistics based on a random sample of size n from an uncertain distribution $F(\cdot \mid \theta)$. As an example, consider estimating the rate parameter λ for a sequence of n iid draws $\{X_i\} \overset{\text{iid}}{\sim} \mathsf{Ga}(\alpha, \lambda)$ from a gamma distribution with known shape parameter α (for example, perhaps $\alpha = 1$ in which case the data are iid $\mathsf{Ex}(\lambda)$). The joint pdf for $X_1 \cdots X_n$ at a point $x \in (0, \infty)^n$ is:

$$f_n(x) = \prod_{i=1}^{n} \left\{ \frac{\lambda^\alpha\, x_i^{\alpha-1}}{\Gamma(\alpha)}\, e^{-\lambda x_i} \right\} = \frac{\lambda^{n\alpha} \big(\prod x_i\big)^{\alpha-1}}{\Gamma(\alpha)^n} \exp\left\{ -\lambda \sum x_i \right\} \tag{4}$$

3.4.1 Method of Moments

The Method of Moments (MoM) procedure is to estimate $\theta \in \Theta \subset \mathbb{R}^k$ by that value (if any exists) $\hat\theta_n$ such that the first k moments of the parametric and empirical distributions agree, *i.e.*,

$$\mathsf{E}_{\hat\theta_n} X^m = \frac{1}{n} \sum_{i=1}^{n} (X_i)^m, \qquad 1 \le m \le k.$$

For the $\mathsf{Ga}(\alpha, \lambda)$ example with α known, $k = 1$ and the mean is $\mathsf{E}_\lambda[X] = \alpha/\lambda$ so the MoM estimate is

$$\tilde\lambda_n = \alpha n / \sum_{i=1}^{n} X_i = \alpha / \bar X_n.$$

For any one-dimensional inference problem the MoM estimator of any parameter $\theta \in \Theta \subset \mathbb{R}$ is just the solution $\tilde\theta$ to the equation

$$\mathsf{E}_\theta[X] = \bar X_n := \frac{1}{n}\sum_{i=1}^n X_i \tag{5a}$$

equating the population and sample means. For example, the MoM estimator of θ for $\{X_i\} \overset{\text{iid}}{\sim} \mathsf{Un}\big([0,\theta]\big)$ is $2\bar X_n$, since $\mathsf{E}_\theta X = \theta/2$. For $k=2$ the MoM is the value $\tilde\theta \in \Theta \subset \mathbb{R}^2$ satisfying both (5a) and $\mathsf{E}_\theta X^2 = (1/n)\sum X_i^2$ or, equivalently, (5a) and

$$\mathsf{V}_\theta[X] = S_n^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \bar X_n)^2. \tag{5b}$$

Note it is $\frac{1}{n}$ and not $\frac{1}{n-1}$ that appears in (5b) for the variance of the empirical distribution $\hat F_n$. For example, if $\{X_i\} \sim \mathsf{Ga}(\alpha,\lambda)$ with both α and λ unknown, the population mean and variance are $\mu = \mathsf{E}[X_i] = (\alpha/\lambda)$ and $\sigma^2 = \mathsf{V}[X_i] = \alpha/\lambda^2$. We can solve to find α and λ as functions of the population mean and variance, to find $\lambda = \mu/\sigma^2$ and $\alpha = \mu^2/\sigma^2$, so the MoM estimators for the shape α and rate λ are

$$\tilde\alpha_n = (\bar X_n)^2/S_n^2 \qquad \tilde\lambda_n = \bar X_n/S_n^2.$$

Under a change of parameters $\phi = H(\theta)$ as in Section (3.1), the MoM estimators $\tilde\phi = H(\tilde\theta)$ change with the same transformation. For example, with $H(\alpha,\lambda) = (\alpha, 1/\lambda)$, the MoM estimators for the shape α and *scale* $\beta = 1/\lambda$ are $(\tilde\alpha_n, \tilde\beta_n) = H(\tilde\alpha_n, \tilde\lambda_n)$, or

$$\tilde\alpha_n = (\bar X_n)^2/S_n^2 \qquad \tilde\beta_n = S_n^2/\bar X_n.$$

3.4.2 Maximum Likelihood

The Maximum Likelihood Estimator (MLE) is whatever function value $\hat\theta_n(X) \in \Theta$ maximizes the likelihood function (2). In our example, the LH is any function proportional to the joint pdf (4). Such a function will attain its maximum at the same place as its logarithm $\ell(\theta \mid X) := \log \mathcal{L}(\theta \mid X)$ does. This is a very general and useful observation, because typically the log LH is easier to maximize.

For the example of $X_i \overset{\text{iid}}{\sim} \mathsf{Ga}(\alpha,\lambda)$ with known α, the log LH and its derivative are:

$$\begin{aligned} \ell_n(\lambda \mid x) &= n\alpha\log\lambda + \alpha\sum \log(x_i) - n\log\Gamma(\alpha) - \lambda\sum x_i \\ \frac{\partial}{\partial\lambda}\ell_n(\lambda \mid x) &= n\alpha/\lambda - \sum x_i \end{aligned} \tag{6}$$

so the MLE is again $\hat\lambda_n = \alpha/\bar X_n$, the same as the MoM.

Finding the MLE for $\mathsf{Ga}(\alpha,\lambda)$ with both parameters unknown is a bit more involved. For each fixed $\alpha > 0$ the optimal rate is $\hat\lambda_n(\alpha) = \alpha/\bar X_n$, but we must also maximize

$$\ell_n\big(\alpha, \hat\lambda(\alpha)\big) = n\alpha\log(\alpha/\bar X_n) + \alpha\sum\log(x_i) - n\log\Gamma(\alpha) - (\alpha/\bar X_n)n\bar X_n$$

by finding the unique solution $\hat\alpha_n$ to the equation $0 = (\partial/\partial\alpha)\ell_n\big(\alpha, \hat\lambda(\alpha)\big)$,

$$0 = n\log(\alpha/\bar X_n) + n + \sum \log(X_i) - n\psi(\alpha) - n,$$

i.e., by setting $\hat\alpha$ to the unique solution of

$$\psi(\alpha) - \log\alpha = \frac{1}{n}\sum \log(X_i/\bar X_n). \tag{7}$$

Here $\psi(z) := (d/dz)\log\Gamma(z) = \Gamma'(z)/\Gamma(z)$ is the "digamma function", given for integer arguments by

$$\psi(n) = -\gamma_e + \sum_{k=1}^{n-1}\frac{1}{k}$$

where $\gamma_e \approx 0.577216$ is Euler's gamma constant. Equation (7) always has a unique solution, because $[\psi(\alpha) - \log\alpha]$ increases monotonically from $-\infty$ to 0 as α increases from 0 to ∞, and $\sum\log(X_i/\bar X_n) < 0$ by the arithmetic-geometric mean inequality.

Under a change of parameters $\phi = H(\theta)$ as in Section (3.1), the MLEs $\hat\phi = H(\hat\theta)$ change with the same transformation. For example, with α known the MLE for $\beta = 1/\lambda$ is $\hat\beta_n = 1/\hat\lambda_n = \bar X_n/\alpha$, while when both α and β are unknown the MLEs are $\hat\beta_n = 1/\hat\lambda_n$ and $\hat\beta_n = \bar X_n/\hat\alpha_n$ for the solution $\hat\alpha_n$ to (7).

Other Examples:

- The MLE for the mean of $\{X_i\} \overset{\text{iid}}{\sim} \mathsf{Po}(\lambda)$ is $\hat\lambda_n = \bar X_n$;
- The MLE for the success probability of $\{X_i\} \overset{\text{iid}}{\sim} \mathsf{Ge}(p)$ is $\hat p_n = 1/(1+\bar X_n)$;
- The MLE for the mean and variance of $\{X_i\} \overset{\text{iid}}{\sim} \mathsf{No}(\mu,\sigma^2)$ are $\hat\mu_n = \bar X_n := (1/n)\sum X_i$ and $\hat\sigma_n^2 = S_n^2 := (1/n)\sum(X_i - \bar X_n)^2$. The MLE for the *precision* $\tau = 1/\sigma^2$ is $\hat\tau_n = 1/S_n^2$.

3.4.3 Bayesian Inference

In the Bayesian approach the data pdf $f(x \mid \theta)$ is treated as a *conditional* pdf for X, given the parameter vector θ. If the investigator can identify in some way a marginal pdf $\pi(\theta)$ for θ unrelated to the data X, called the *prior pdf*, then the joint pdf for X and θ will be $f(x \mid \theta)\pi(\theta)$ and the conditional (given $X = x$) or *posterior pdf* for θ, given the observed data, will be given by the elementary probability calculation

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\ \pi(\theta)}{\int_\Theta f(x \mid \theta)\ \pi(\theta)\,d\theta} = \frac{\mathcal{L}(\theta \mid x)\ \pi(\theta)}{\int_\Theta \mathcal{L}(\theta \mid x)\ \pi(\theta)\,d\theta} \propto \mathcal{L}(\theta \mid x)\ \pi(\theta) \tag{8a}$$

called *Bayes' Formula*. The denominator in (8a) is just a normalizing constant to ensure that $\pi(\theta \mid x)$ integrates to one; usually it needn't be computed explicitly (see examples below).

The estimator $\delta(x)$ that minimizes the expected squared error

$$\mathsf{E}\big[|\delta(x) - \theta|^2 \mid x\big] = \int_\Theta |\delta(x) - \theta|^2\ \pi(\theta \mid x)\,d\theta$$

can easily be shown to be the mean of the posterior distribution, $\delta(x) = \bar\theta$ given by

$$\bar\theta := \int \theta\, \pi(\theta \mid x)\, d\theta. \tag{8b}$$

Under a change of parameters $\phi = H(\theta)$ as in Section (3.1), Bayesian posterior means $\bar\phi$ are *not* simply the transformed values $H(\bar\theta)$; rather, a Jacobian enters with the change of variables in the integral of (8b).

For example, if $\{X_i\} \overset{\text{iid}}{\sim} \mathsf{Ga}(\alpha, \lambda)$ with known α as in (4), and we represent uncertainty about λ before observing $X = \{X_i : 1 \le i \le n\}$ using a $\mathsf{Ga}(a, b)$ distribution, then

$$\begin{aligned}
\pi(\lambda \mid x) &\propto \frac{\lambda^{n\alpha} \prod (x_i)^\alpha}{\Gamma(\alpha)^n} \exp\Big\{-\lambda \sum x_i\Big\} \times \frac{b^a \lambda^{a-1}}{\Gamma(a)} \ \exp\{-b\lambda\} \\
&\propto \lambda^{a+n\alpha-1} \exp\Big\{-\lambda\big(b + \sum x_i\big)\Big\} \\
&\sim \mathsf{Ga}\Big(a + n\alpha, b + \sum x_i\Big)
\end{aligned}$$

with mean

$$\bar\lambda_n = \frac{a + n\alpha}{b + \sum x_i} = \frac{a/n + \alpha}{b/n + \bar X_n}, \tag{9}$$

the same asymptotically as $n \to \infty$ but a bit different for small n from the MoM and MLE estimators $\alpha/\bar X_n$. Note we didn't need to calculate the normalizing constant, because we recognized the form of $\pi(\lambda \mid x)$ as that of a gamma pdf. The posterior mean of $\mu := \mathsf{E}[X]$, or $\bar\mu_n := \mathsf{E}[\alpha/\lambda \mid \{X_i\}]$, is *not* just $\alpha/\bar\lambda_n$; can you calculate what $\bar\mu_n$ is? Suggestion: First find $\mathsf{E}[X^p]$ for $X \sim \mathsf{Ga}(\alpha, \lambda)$, for all $-\infty < p < \infty$ (you'll need it for $p < 0$).

3.4.4 Objective Bayes Inference

How can an investigator find the prior or marginal density $\pi(\theta)$ needed to compute the posterior density $\pi(\theta \mid x)$ in (8a)? We'll see more about this in Week 5 of the course, but for now here are three suggestions:

- Historical Records: If similar analyses have been performed in the past, the values $\{\theta_j\}$ (true or estimated) may be available to offer a guide for what is the distribution of possible values of θ in the current study;
- Personal Opinion: An experienced investigator may have informed opinions about which values of θ are plausible and which are not. This subjective approach is particularly well-suited to problems of personal finance, sports bets, and other situations where long-term historical evidence isn't available even in principle. It's not well-suited to problems in scientific exploration, litigation, or other areas where objectivity is paramount.
- Objective Bayesian Analysis: In the earliest days of Bayesian analysis, Laplace (1774) and Bayes (1763) himself used uniform (or "flat") prior densities to represent ignorance about a parameter. In our example this would be $\pi(\lambda) \equiv 1$ on $\mathbb{R}_+$, a prior that is "improper" in the

sense that $\int \pi(\lambda)\,d\lambda = \infty$ but which nevertheless leads to the proper posterior distribution with density $\pi(\lambda \mid x) \sim \mathsf{Ga}\big(1 + n\alpha, \sum x_i\big)$ distribution with well-defined posterior mean $\bar\lambda_n = (\alpha + 1/n)/\bar X_n$. This approach is not invariant under changes of parameters.

A modern approach that *is* invariant would recommend the "Jeffreys" or "Reference" prior $\pi_R(\lambda) \propto 1/\lambda$ leading to posterior distribution $\pi_R(\lambda \mid x) \sim \mathsf{Ga}\big(n\alpha, \sum x_i\big)$, with mean $\bar\lambda_n = \alpha/\bar X_n$ identical in this example (but not always) to the MLE and MoM (Jeffreys, 1961; Bernardo, 1979; Berger et al., 2009). This is a Bayesian approach that *can* be used in problems where objectivity is key.

We'll see more about this in Week 5.

References

Bayes, T. (1763), "An essay toward solving a problem in the doctrine of chances," *Philosophical Transactions of the Royal Society*, pp. 370–418.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009), "The Formal Definition of Reference Priors," *Annals of Statistics*, 37, 905–938.

Berger, J. O., Liseo, B., and Wolpert, R. L. (1999), "Integrated likelihood methods for eliminating nuisance parameters," *Statistical Science*, 14, 1–28.

Berger, J. O. and Wolpert, R. L. (1988), *The Likelihood Principle: A Review, Generalizations, and Statistical Implications (with discussion)*, *IMS Lecture Notes-Monograph Series*, volume 6, Hayward, CA: Institute of Mathematical Statistics, second edition.

Bernardo, J. M. (1979), "Reference posterior distributions for Bayesian inference (with discussion)," *Journal of the Royal Statistical Society, Ser.* B*: Statistical Methodology*, 41, 113–147.

Birnbaum, A. (1962), "On the Foundations of Statistical Inference (with discussion)," *Journal of the American Statistical Association*, 57, 269–326.

Jeffreys, H. (1961), *Theory of Probability*, Oxford University Press.

Laplace, P. S. (1774), *A philocophical essay on probabilities*, New York, NY: Dover, translated from 6th French edn (1774) by Frederick Wilson Truscott and Frederick Lincoln Emory; Dover edition, 1952.