STA355H1
Excerpt From Personal Notes

Omerzahid Ali

November 20, 2024

Contents

1 Inference

1.1 Maximum Likelihood Estimators

1.1.1 Definition

Suppose $(X_1, \ldots, X_n)$ are random variables with joint pdf/pmf $f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_k)$ where $\theta_1, \ldots, \theta_k$ are unknown. Then, given data $x_1, \ldots, x_n$, we can define the likelihood function.

Definition 1.1 (Likelihood Function)**.**

$$\mathcal{L}(\theta_1, \cdots, \theta_k) = f(\underbrace{x_1, \cdots, x_n}_{data}; \theta_1, \cdots, \theta_k)$$

We commonly refer to the natural log of the likelihood, $\ln \mathcal{L}(\theta)$, as the log-likelihood.

Example 1.1. **Model**: $X_1, \cdots, X_n$ independent random variables with pdf

$$f(x;\theta) = \frac{|x-\theta|^{-1/2}}{2\sqrt{\pi}} \exp\left(-|x-\theta|\right).$$

The likelihood function is

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \left\{ \frac{|x_i-\theta|^{-1/2}}{2\sqrt{\pi}} \exp\left(-|x_i-\theta|\right) \right\}.$$

□

Definition 1.2 (Maximum Likelihood Estimator)**.** Suppose that for each $\mathbf{x} = (x_1, \ldots, x_n)$, $(T_1(x), \ldots, T_k(x))$ maximize $\mathcal{L}(\theta_1, \ldots, \theta_k)$. Then the maximum likelihood estimators (MLEs) of $\theta_1, \ldots, \theta_k$ are

$$\hat{\theta}_j = T_j(X_1, \ldots, X_n) \text{ for } j = 1, \ldots, k.$$

To compute such an estimator, it depends on our parameter space Θ. If $\mathcal{L}(\theta)$ is differentiable, Θ is an open set, and an MLE exists, it satisfies the **likelihood equation**:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \ln \mathcal{L}(\hat{\theta}) = 0$$

This may not always be the case, as Θ could be closed, potentially leaving $\hat{\theta}$ to be a boundary point. Another possibility is that $\hat{\theta}$ may be an extremum of the data (Like $\hat{\theta} = X_{(n)}$). If this happens, we must directly maximize $\mathcal{L}(\theta)$.

Lemma 1 (MLE Invariance)**.** *If $\hat{\theta}$ is an MLE of θ and $u(\theta)$ is a function of θ, then $u(\hat{\theta})$ is an MLE for $u(\theta)$.*

Example 1.2. Let $X_1, \ldots, X_n$ be a random sample from an Exponential distribution with scale parameter β. Since the MLE, $\hat{\beta}$, is $\bar{X}_n$, then the MLE for $p(\beta) = P(X \geq 1) = e^{-1/\beta}$ is $\widehat{p(\beta)} = p(\hat{\beta}) = e^{-1/\bar{X}_n}$.

□

1.1.2 Consistency

Definition 1.3 (Consistent Estimator)**.** An estimator $\hat{\theta}$ is consistent if it approaches the true value θ_0 as more data is observed. That is, for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[|\hat{\theta}_n - \theta_0| > \epsilon\right] = 0.$$

The MLE is consistent when particular conditions are met:

1. The model must be identifiable.
2. The parameter space Θ must be compact.
3. The density function must be continuous.
4. The log-likelihood must converge uniformly:

$$\sup_{\theta\in\Theta} ||\ell_n(\theta) - \ell(\theta)|| \xrightarrow{p} 0$$

   where $\ell(\theta)$ is the expected log-likelihood, $\ell(\theta) = \mathbb{E}[\log f(X_i, \theta)]$.

These conditions are somewhat mild, and hold for most i.i.d. samples with most common distributions, but still must not be assumed.

Example 1.3. **Neyman-Scott Problem**

Model: $(X_1, Y_1), \ldots, (X_n, Y_n)$ independent pairs of independent Normal random variables, where for all fixed i, $X_i, Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$.
Note that X_i, Y_i are independent measurements of the same quantity, and $\mu_1, \ldots, \mu_n$ are unknown. Then, the likelihood function would be,

$$\mathcal{L}(\mu_1, \ldots, \mu_n, \sigma) = \prod_{i=1}^{n} \left[\frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x_i - \mu_i)^2 + (y_i - \mu_i)^2}{2\sigma^2}\right)\right]$$

Maximizing this over σ^2 and $\{\hat{\mu}_i\}$, you get MLEs:

$$\hat{\mu}_i = \frac{X_i + Y_i}{2},\ i = 1, \ldots, n.$$

$$\hat{\sigma}^2 = \frac{1}{4n} \sum_{i=1}^{n} (X_i - Y_i)^2.$$

By WLLN, we know

$$\frac{1}{n} \sum_{i=0}^{n} (X_i - Y_i) \xrightarrow{p} 2\sigma^2$$

Thus, for our MLE of σ^2, our limiting behaviour ends up being,

$$\hat{\sigma}^2 = \frac{1}{4n} \sum_{i=1}^{n} (X_i - Y_i)^2 \xrightarrow{p} \sigma^2/2$$

Thus, $\hat{\sigma}^2$ is not consistent.

□

This happens because we're using $2n$ observations to estimate n means, so we overfit.

1.1.3 Fisher Information

Definition 1.4 (Fisher Information)**.** Suppose $\ln \mathcal{L}(\theta) = \log f(X;\theta)$ is the log-likelihood function. Then, we say that

$$\mathcal{I}(\theta) = E\left[\frac{\partial}{\partial\theta}\ln \mathcal{L}(\theta)\right]^2$$

is the fisher information. Under appropriate conditions on $f(x;\theta)$ and $\mathcal{I}(\theta)$, it may also (and more commonly) be expressed as:

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln \mathcal{L}(\theta)\right]$$

Note that the first equation may also be expressed as the variance of the **score**. However, oftentimes this is not readily available as we work with samples, suggesting we need an analogous sample version.

Definition 1.5 (Observed Fisher Information)**.** Suppose $\ell(\hat\theta) = \ln \mathcal{L}(\hat\theta)$ is the observed log-likelihood function. Then, we say that

$$I(\theta) = -\frac{\mathrm{d}}{\mathrm{d}\theta^2}\ln \mathcal{L}(\hat\theta)$$

is the observed fisher information for a particular observation. Note that for an i.i.d. sample, the fisher information produced is identical, and thus the observed fisher information of the entire sample is $nI(\theta)$.

From basic calculus techniques, it is easy to observe that the observed Fisher information is simply the (absolute) curvature of the log-likelihood function at its maximum. The greater the curvature, the more well-defined the maximizer $\hat\theta$ is.
This suggests that as the observed Fisher information increases, the more prominent our $\hat\theta$, resulting with a better estimator and less uncertainty. This is the motivation behind our standard error estimate,

$$\widehat{\mathrm{se}}(\hat\theta) = \left\{-\frac{\mathrm{d}}{\mathrm{d}\theta^2}\ln \mathcal{L}(\hat\theta)\right\}$$

1.1.4 Exponential MLE Convergence

Some common distributions that can be parameterized "nicely" are distributions from the exponential class or family. These include the normal distribution, bernoulli, poisson, and more. Because of their special parameterizations, it is often easier to prove their convergence properties in a more general fashion.

Definition 1.6 (One-parameter exponential family)**.** Assume $f(x;\theta)$ is the pmf of a random variable X. If f has the form,

$$f(x;\theta) = \exp[c(\theta)T(x) - d(\theta) + h(x)] \text{ for } x \in A$$

then X is considered to be from the one-parameter exponential family of distributions.

Lemma 2 (Exponential MLE Consistency)**.** *If $\hat\theta_{ML}$ is the MLE of a one-parameter exponential family, then:*

$$\hat\theta_{ML} \xrightarrow{p} \theta$$

Proof. For this model, the log-likelihood becomes

$$\ln \mathcal{L}(\theta) = \sum_{i=1}^{n}[c(\theta)T(x_i) - d(\theta) + h(x_i)]$$

and its derivative is

$$\frac{\partial}{\partial\theta}\ln \mathcal{L}(\theta) = \sum_{i=1}^{n}[c'(\theta)T(x_i) - d'(\theta)]$$

Thus, the MLE $\hat\theta_n$ satisfies the equation

$$\frac{1}{n}\sum_{i=1}^{n}T(x_i) = \frac{d'(\hat\theta_n)}{c'(\hat\theta_n)}$$

We now want to show that

$$\mathbb{E}_\theta\left[T(X_i)\right] = \frac{d'(\theta)}{c'(\theta)}$$

and

$$\mathrm{Var}_\theta[T(X_i)] = \frac{d''(\theta)c'(\theta) - c''(\theta)d'(\theta)}{[c'(\theta)]^3}$$

Subproof: Expectation / Variance of $T(X_i)$.

Let $m_\theta(s)$ be the mgf of $T(X_i)$.

$$\begin{aligned} m_\theta(s) &= \mathbb{E}_\theta\left[\exp(sT(X_i))\right] \\ &= \int_A \exp[sT(x)] \cdot \exp[c(\theta)T(x) - d(\theta) + h(x)]\, dx \\ &= \int_A \exp[(s + c(\theta))T(x) - d(\theta) + h(x)]\, dx \end{aligned}$$

Let $c(\theta') := s + c(\theta)$. If $\theta' \in \Theta$, then, by integrating $f(x;\theta')$

$$\implies \int_A \exp[c(\theta')T(x) - d(\theta') + h(x)]\, dx = 1$$

$$\implies \int_A \exp[c(\theta')T(x) + h(x)] = \exp[d(\theta')]$$

$$\implies \int_A \exp[c(\theta')T(x) - d(\theta) + h(x)] = \exp[d(\theta') - d(\theta)]$$

Thus, substituting back $c(\theta')$, then differentiating with respect to s,

$$m_\theta(s) = \exp[d(\theta') - d(\theta)]$$

$$\implies m'_\theta(s) = \exp[d(\theta') - d(\theta)] \cdot \frac{\partial d(\theta')}{\partial s}$$

By the chain rule,

$$\frac{\partial d(\theta')}{\partial s} = \frac{\partial d(\theta')}{\partial \theta'} \cdot \frac{\partial \theta'}{\partial s}$$

Recall that $c(\theta') = s + c(\theta)$, meaning that $c'(\theta')$ with respect to s is 1. Thus, by chain rule,

$$1 = \frac{\partial c(\theta')}{\partial s} = \frac{\partial c(\theta')}{\partial \theta'} \cdot \frac{\partial \theta'}{\partial s}$$

$$\implies \frac{\partial \theta'}{\partial s} = \frac{1}{\partial c(\theta')/\partial \theta'} = \frac{1}{c'(\theta')}$$

$$\implies \frac{\partial d(\theta')}{\partial s} = d'(\theta') \cdot \frac{1}{c'(\theta')}$$

Thus,

$$\implies m'_\theta(s) = \exp[d(\theta') - d(\theta)] \cdot \frac{d'(\theta')}{c'(\theta')}$$

Note that at $s = 0$, $\theta = \theta'$. Using the property of moment generating functions,

$$\mathbb{E}\left[T(X_i)\right] = m'_\theta(0) = \exp(0) \cdot \frac{d'(\theta)}{c'(\theta)} = \frac{d'(\theta)}{c'(\theta)}$$

By the quotient and product rule,

$$m''_\theta(s) = m_\theta(s) \cdot \frac{d''(\theta') \cdot c'(\theta') - d'(\theta') \cdot c''(\theta')}{(c'(\theta'))^2} \cdot \frac{\partial \theta'}{\partial s} + m'_\theta(s) \cdot \frac{d'(\theta')}{c'(\theta')}$$

Similar to before, $s = 0 \implies \theta = \theta'$,

$$m''_\theta(0) = \exp(0) \cdot \frac{d''(\theta) \cdot c'(\theta) - d'(\theta) \cdot c''(\theta)}{(c'(\theta))^3} + \left(\frac{d'(\theta)}{c'(\theta)}\right)^2$$

And thus,

$$\mathrm{Var}[T(X_i)] = \mathbb{E}\left[T(X_i)^2\right] - \left(\mathbb{E}\left[T(X_i)\right]\right)^2 = m''_\theta(0) - m'_\theta(0)^2$$

$$= \frac{d''(\theta) \cdot c'(\theta) - d'(\theta) \cdot c''(\theta)}{(c'(\theta))^3} + \left(\frac{d'(\theta)}{c'(\theta)}\right)^2 - \left(\frac{d'(\theta)}{c'(\theta)}\right)^2$$

$$\implies \mathrm{Var}[T(X_i)] = \frac{d''(\theta) \cdot c'(\theta) - d'(\theta) \cdot c''(\theta)}{(c'(\theta))^3}$$

Which concludes the subproof. ■

Now that we know the expectation and variance, let $h'(\theta) = \frac{d'(\theta)}{c'(\theta)}$. By CLT,

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} T(X_i) - h'(\theta)\right] \xrightarrow{p} \mathcal{N}(0, h''(\theta)/c'(\theta)).$$

Define g to be the inverse of $h'(\theta)$, such that $g(h'(\theta)) = \theta$. Thus,

$$g'(h'(\theta))h''(\theta) = 1$$

$$\implies g'(h'(\theta)) = \frac{1}{h''(\theta)}$$

We can now use the Delta Method, and result with

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, [g'(h'(\theta))]^2 \cdot h''(\theta)/c'(\theta)\right)$$

$$\implies \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{c'(\theta) \cdot h''(\theta)}\right)$$

□

Now, we can use this to estimate the standard error. Using the previous result,

$$\hat{\theta} \approx \mathcal{N}\left(\theta, \frac{1}{n \cdot c'\theta \cdot h''(\theta)}\right)$$

This suggests that we can estimate the standard error by

$$\widehat{\mathrm{se}}(\hat{\theta}) = [n \cdot c'(\theta) \cdot h''(\theta)]^{-1/2}$$

Define $\ell(x;\theta) = \ln f(x;\theta)$ and $l'(x;\theta)$, $l''(x;\theta)$ to be its first two partial derivatives with respect to θ. Then,

$$\mathrm{Var}_\theta[\ell'(X_i;\theta)] = [c'(\theta)]^2 \cdot \mathrm{Var}_\theta[T(X_i)] = c'(\theta) \cdot h''(\theta)$$

We then have that,

$$I(\theta) = c'(\theta) \cdot h''(\theta) = \left(\frac{\partial^2}{\partial\theta^2} \ln \mathcal{L}(\hat{\theta})\right)$$

Recall that $nI(\theta)$ is the **observed Fisher Information**, and the following holds,

$$\hat{\theta} \approx \mathcal{N}\left(\theta, \frac{1}{nI(\theta)}\right)$$

1.2 Sufficiency

Suppose $(X_1, \dots, X_n)$ is a random sample with joint pdf/pmf $f(\mathbf{x}; \boldsymbol{\theta})$, where $\mathbf{x} = (x_1, \dots, x_n)$, and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$.

Often times, $\mathbf{X} = (X_1, \dots, X_n)$ contains information not relevant to $\boldsymbol{\theta}$, which we can reduce in complexity. We would like to form a statistic, $T(\mathbf{X})$, in which it is a data reduction or data summary of $\mathbf{X}$, and ideally, $T(\mathbf{X})$ would map to a smaller dimension than $\mathbf{X}$.

1.2.1 Sufficient Statistics

Definition 1.7 (Sufficient Statistic)**.** A statistic $T(\mathbf{X})$ is a *sufficient statistic* for $\boldsymbol{\theta}$ if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on $\boldsymbol{\theta}$.

We notationally define $T = (T_1(\mathbf{X}), \dots, T_m(\mathbf{X}))$.

Theorem 1 (Factorization Theorem)**.** *Let $f(\boldsymbol{x}|\boldsymbol{\theta})$ denote the joint pdf/pmf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for $\boldsymbol{\theta}$ if and only if there exist functions $g(t|\boldsymbol{\theta})$ and $h(\boldsymbol{x})$ such that for all $(\boldsymbol{x}, \boldsymbol{\theta})$,*

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = g(T(\boldsymbol{x})|\boldsymbol{\theta})h(\boldsymbol{x})$$

Note that the likelihood function, $\mathcal{L}(\boldsymbol{\theta})$, is defined as $f(\mathbf{x}; \boldsymbol{\theta})$. Thus, for the factorization, we end up with,

$$\mathcal{L}(\boldsymbol{\theta}) = f(\mathbf{x}; \boldsymbol{\theta}) = g(T(\mathbf{x}); \boldsymbol{\theta})h(\mathbf{x})$$

This tells us that $h(\mathbf{x})$ is only a multiplicative constant in $\mathcal{L}(\boldsymbol{\theta})$. Thus, maximizing $\mathcal{L}(\boldsymbol{\theta})$ is equivalent to maximizing $g(T(\mathbf{x}); \boldsymbol{\theta})$.

1.3 Bayesian Inference

Suppose we have a model $(X_1, \ldots, X_n)$ that have a joint pdf/pmf $f(x_1, \ldots, x_n; \theta)$ where $\theta \in \Theta$ is real. Suppose we want to incorporate previous information about θ into our estimate of our model. In bayesian statistics, this is done by incorporating *a priori* information about θ using probability distributions. We thus treat θ as a random variable, representing what we know about the parameter before observing data $\mathbf{X}$.

1.3.1 Prior and Posterior Distribution

We characterize the prior distribution with a density function.

Suppose $\pi(\theta)$, is the prior density function for θ. We see it as the distribution of the parameter based on personal opinion, experts' knowledge, and other subjective information about the parameter.

One could think of our observed data having the joint pdf of $(X_1 \ldots, X_n, \theta)$. Thus, for a given value θ, a random sample comes from a distribution $\pi(\theta) f(x_1, \ldots, x_n; \theta)$. This leads us to the following definition.

Definition 1.8 (Posterior Density Function)**.**

Given a continuous parameter space Θ, we define our posterior density function as:

$$\begin{aligned}\pi(\theta | x_1, \ldots, x_n) &= \frac{\pi(\theta) f(x_1, \ldots, x_n; \theta)}{\int_\Theta \pi(s) f(x_1, \ldots, x_n; s)\, ds} \\ &= c(x_1, \ldots, x_n) \pi(\theta) \mathcal{L}(\theta_j)\end{aligned}$$

where the function c is our normalizing constant,

$$c(x_1, \ldots, x_n) = \int_\Theta \pi(s) f(x_1, \ldots, x_n; s)\, ds$$

For the analogous discrete parameter space,

$$c(x_1, \ldots, x_n) = \left\{ \sum_k \pi(\theta_k) f(x_1, \ldots, x_n l \theta_k) \right\}^{-1}$$

Thus, for continuous parameter spaces,

$$\pi(\theta | x_1, \ldots, x_n) \propto \pi(\theta) \mathcal{L}(\theta)$$

which is similar to the discrete case, with θ_j instead of θ.

Computing Posterior Densities
As noted earlier, for a single parameter, the posterior density takes the form

$$\pi(\theta | x_1, \ldots, x_n) = c(x_1, \ldots, x_n) \pi(\theta) \mathcal{L}(\theta)$$

Typically, $\pi(\theta)$ and $\mathcal{L}(\theta)$ are easy to evaluate, so you only need to compute

$$c(x_1, \ldots, x_n) = \left\{ \int_\theta \pi(\theta) \mathcal{L}(\theta)\, d\theta \right\}^{-1}$$

Sometimes this integral may be simple, such as if we have a conjugate prior, however oftentimes we may need to use numerical integration:

$$\int_\Theta \pi(\theta) \mathcal{L}(\theta)\, d\theta \approx \sum_{k=1}^N \mathbf{w}_k \pi(\theta_k) \mathcal{L}(\theta_k)$$

1.3.2 Choosing Priors

Ideally, we would want our prior density $\pi(\theta)$ to reflect *a priori* information about θ. This information could come from previous studies, expert opinions, or other informative processes.

Rarely, we would either have no information at all about θ, or that there are regions in the parameter space that are implausible. In these cases, we want to use an uninformative or weakly informative prior.

Definition 1.9 (Uniform Prior)**.** If the parameter space is bounded, a uniform prior is a prior distribution where for all $\theta \in \Theta$ and fixed $c \in \mathbb{R}$, θ has the prior density function

$$\pi(\theta) = c$$

Note that if we transform θ, such that we have $\phi = g(\theta)$, if g is a non-linear transformation, then the prior for ϕ is no longer uniform on $g(\Theta)$.

Some controversy arises with the subjectivity of choosing prior distributions. Since the choice of prior is largely subjective, an unscrupulous investigator could ”massage” results by adjusting prior distribution.

However, a common counter-argument is that non-Bayesian methods are victim to this as well. Choosing the model (likelihood), choosing thresholds for p-values, and model selection methods are largely subjective too.

The most important thing in bayesian methods is transparency. Informing and justifying why you chose the prior distribution, and if possible, making the data available increases the legitimacy of bayesian methods.

1.3.3 Conjugate Priors

Definition 1.10 (Conjugate Prior)**.** Given a model, i.e. a joint pdf/pmf $f(x_1, \ldots, x_n; \theta)$, choose a prior density (indexed by some hyperparameter) such that the posterior density has the same form as the prior:

$$\pi_\alpha(\theta) \xrightarrow{\text{data}} \pi_{\alpha'}(\theta | x_1, \ldots, x_n)$$

where the hyperparameter in the posterior will depend on the data $x_1, \ldots, x_n$. Then, $\pi_\alpha(\theta)$ is a **conjugate prior**.

Some of the reasons behind using conjugate priors include the ease of computation. Since the data only changes the values of the hyperparameters, we don't need advanced model selection. In simple models, conjugate priors can provide a very rich choice of prior densities.

However, we shouldn't restrict ourselves to conjugate priors.

Application to one parameter exponential families

Suppose we have a model $(X_1, \ldots, X_n)$ that is from the one parameter exponential family, which we defined here: 1.6.

We try a prior density of the form:

$$\pi(\theta) = K(\alpha, \beta) \exp[\alpha c(\theta) - \beta d(\theta)]$$

for α, β in some set such that $\pi(\theta)$ is a density function.

Note that

$$\pi(\theta)\mathcal{L}(\theta) \propto \exp[(\alpha + T(\mathbf{x}))c(\theta) - (\beta + 1)d(\theta)]$$

Thus, the posterior density is:

$$\pi(\theta|\mathbf{x}) = K(\alpha + T(\mathbf{x}), \beta + 1) \exp[(\alpha + T(\mathbf{x}))c(\theta) - (\beta + 1)d(\theta)]$$

Example 1.4. [Binary Data]
Suppose we have a model with $X_1, \dots, X_n$ independent binary random variables that have the pmf:

$$f(x;\theta) = \theta^x(1-\theta)^{1-x} \text{ for } x = 0, 1$$

where $0 < \theta < 1$.

Then, the likelihood, or joint pmf of $(X_1, \dots, X_n)$ is

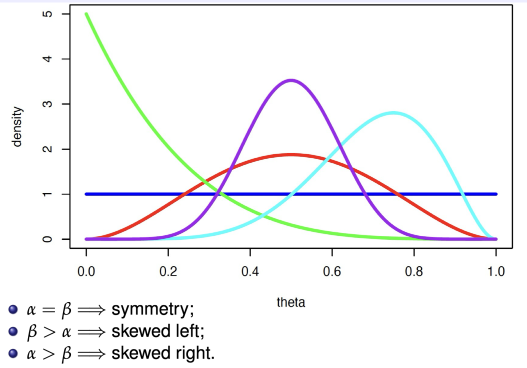
$$f(x_1, \dots, x_n; \theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} = \exp\left[\ln\left(\frac{\theta}{1-\theta}\right)\sum_{i=1}^{n} x_i + n\ln(1-\theta)\right]$$

Which is an exponential family distribution with
$c(\theta) = \ln\left(\frac{\theta}{1-\theta}\right)$, $T(\mathbf{x}) = \sum_{i=1}^{n} x_i$, $d(\theta) = -n\ln(1-\theta)$, $h(x) = 0$.

The form of the joint pmf suggests a Beta prior:

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta) \text{ for } 0 < \theta < 1$$

where $\alpha, \beta > 0$ are hyperparameters. We can vary α, β to approximate a wide variety of prior densities on $(0, 1)$.



To obtain the posterior density, recall that

$$\pi(\theta)\mathcal{L}(\theta) \propto \theta^{\alpha-1+\sum x_i}(1-\theta)^{\beta-1+n-\sum x_i}$$

Thus, the posterior distribution is Beta with hyperparameters:

$$\alpha' = \alpha + \sum_{i=1}^{n} x_i, \quad \beta' = \beta + n - \sum_{i=1}^{n} x_i$$

Note that as n increases, the hyperparameters (and posterior density) are dominated by the data.

Thus, for large n, the posterior is approximately Normal with mean and variance

$$\hat{\mu} = \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{\sigma}^2 = \hat{\theta}(1-\hat{\theta})/n$$

□

1.3.4 Credible and Confidence Intervals

Credible and confidence intervals attempt to do the same thing, albeit in different ways. Firstly, credible intervals are defined in terms of the posterior distribution of θ (which depends on the prior distribution and data), as defined below.

Definition 1.11 (Credible Interval)**.** Given a posterior density $\pi(\theta|x_1,\ldots,x_n)$, an interval (or set) $\mathcal{I}=\mathcal{I}(\mathbf{x})$ is a $100p\%$ **credible interval** (or region) for θ if

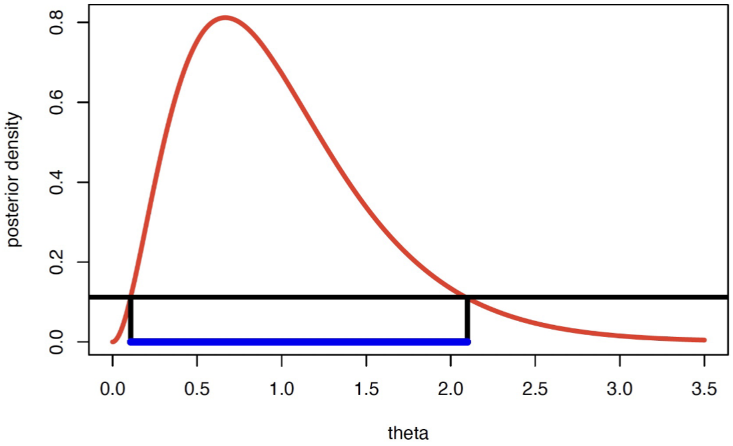
$$\int_{\mathcal{I}(\mathbf{X})} \pi(\theta|x_1,\ldots,x_n)\,d\theta = p$$

If for all $\theta\in\mathcal{I}$ and all $\theta'\in\mathcal{I}$,

$$\pi(\theta|x_1,\ldots,x_n) > \pi(\theta'|x_1,\ldots,x_n)$$

then $\mathcal{I}$ is called a $100p\%$ **highest posterior density (HPD) interval (region)** for θ.

Note that an HPD interval (region) will be the smallest credible interval (or region). Computation of HPD intervals are not always trivial.



At the end points, $\ell(\mathbf{x})$ and $u(\mathbf{x})$, we have

$$\pi(\ell(\mathbf{x})|\mathbf{x}) = \pi(u(\mathbf{x})|\mathbf{x})$$

On the other hand, confidence intervals are defined in terms of coverage over (conceptually) repeated experiments, as defined below.

Definition 1.12 (Confidence Interval)**.** Suppose P_θ is the probability distribution of $\mathbf{X}=(X_1,\ldots,X_n)$. Then, a confidence interval, $\mathcal{I}(\mathbf{X})$, is any interval that satisfies the following:

$$P_\theta[\theta\in\mathcal{I}(\mathbf{X})] = P_\theta[\ell(\mathbf{X})\le\theta\le u(\mathbf{X})] = p \text{ for all } \theta\in\Theta$$