The Bayesian approach to parameter estimation

From Lec 3

Three interesting examples -- 3. Bayesian Inference

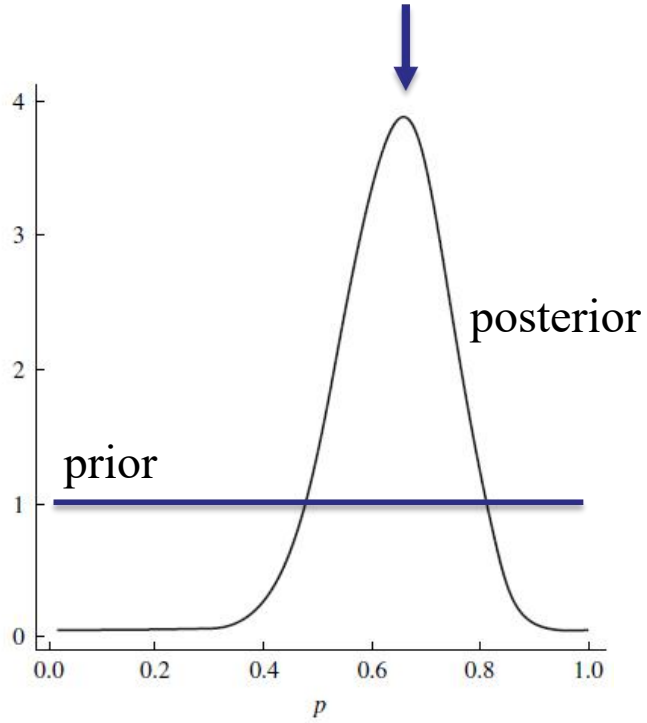
A freshly minted coin has a certain probability of coming up heads if it is spun on its edge (may not be ½). Say, you spin it n times and see X heads. What has been learned about the chance Θ it comes up heads?

Totally ignorant about it, we might represent our knowledge by a uniform density on [0, 1], the prior density

$$f_\Theta(\theta) = 1, \quad 0 \le \theta \le 1.$$

$$f_{\Theta|X}(\theta|x) = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\theta^x(1-\theta)^{n-x}$$

Posterior is ***Beta density****, a=x+1, b=n−x+1.*



Basic thoughts

Unknown parameter θ treated as a random variable

- Assumed to be continuous without loss of generality
- No longer an unknown constant as before!

Prior distribution $f_{\Theta}(\theta)$ represents what we know about it before observing data, X.

For a given value, $\Theta=\theta$, data have the probability distribution $f_{X|\Theta}(x|\theta)$.

Joint distribution of X, Θ: $f_{X,\Theta}(x,\theta) = f_{X|\Theta}(x|\theta) f_{\Theta}(\theta)$

Basic thoughts

Unknown parameter θ treated as a random variable

- Assumed to be continuous without loss of generality
- No longer an unknown constant as before!

Prior distribution $f_\Theta(\theta)$ represents what we know about it before observing data, X.

For a given value, $\Theta=\theta$, data have the probability distribution $f_{X|\Theta}(x|\theta)$.

Joint distribution of X, Θ: $$f_{X,\Theta}(x,\theta) = f_{X|\Theta}(x|\theta)\, f_\Theta(\theta)$$

Marginal distribution of X: $$f_X(x) = \int f_{X,\Theta}(x,\theta)d\theta = \int f_{X|\Theta}(x|\theta)\, f_\Theta(\theta)d\theta$$

likelihood is $f_{X|\Theta}(x|\theta)$

Distribution of Θ given data X: $$f_{\Theta|X}(\theta|x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} = \frac{f_{X|\Theta}(x|\theta)\, f_\Theta(\theta)}{\int f_{X|\Theta}(x|\theta)\, f_\Theta(\theta)d\theta}$$

Posterior distribution
= what's known about
Θ having observed data

$$f_{\Theta|X}(\theta|x) \propto f_{X|\Theta}(x|\theta) \times f_\Theta(\theta)$$

Posterior density $\propto$ Likelihood $\times$ Prior density

Example: Poisson. $P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$

λ is an unknown parameter, with a prior distribution $f_\Lambda(\lambda)$. Data are n i.i.d. observations $X_1, \ldots, X_n$ are i.i.d. and Poisson for a given λ:

$$f_{X_i|\Lambda}(x_i|\lambda) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}, \qquad x_i = 0, 1, 2, \ldots$$

Their joint distribution given λ is the product of marginals:

$$f_{X|\Lambda}(x|\lambda) = \frac{\lambda^{\Sigma_{i=1}^n x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!} \qquad X \text{ denotes } (X_1, X_2, \ldots, X_n)$$

Posterior distribution of Λ given X:

$$f_{\Lambda|X}(\lambda|x) = \frac{\lambda^{\Sigma_{i=1}^n x_i} e^{-n\lambda} f_\Lambda(\lambda)}{\int \lambda^{\Sigma_{i=1}^n x_i} e^{-n\lambda} f_\Lambda(\lambda)\, d\lambda}$$

To evaluate posterior distribution, we have to do two things:

1. Specify prior distribution
2. Carry out the integration in denominator

Two Bayesians: “he” and “she”…

He is an orthodox Bayesian, takeing very seriously the model that the prior distribution specifies his prior opinion -- a meticulous approach.

作为一个具体的例子，我们考虑美国国家科学与技术研究所的一项研究 (Steel 等 1980). 观测滤光片上石棉样纤维的个数，以制定石棉浓度的测量标准. 将石棉溶解在水中，然后散布在滤光片上，同时在滤光片上取直径 3 毫米的小孔，最后将其安放在透射电子显微镜下观察. 操作员计数下 23 个网格中的纤维数，得到的数据如下：

31	29	19	18	31	28	34	27
34	30	16	18	26	27	27	18
24	22	28	24	21	17	24	

在这种情况下，泊松分布适合描述不同网格之间纤维数的变异性，并用来刻画未来观测的内在变异性. λ 的矩估计方法是上述观测值的算术平均，或 $\hat{\lambda} = 24.9$.

$$n=23, \quad \Sigma\, x_i=573$$

Two Bayesians: "he" and "she"…

He is an orthodox Bayesian, takeing very seriously the model that the prior distribution specifies his prior opinion -- a meticulous approach.

- Note: this specification should be done before seeing the data, X.
- He decides to quantify his opinion by specifying a prior mean μ_1=15, σ=5, Gamma distribution (math will work out nicely!)

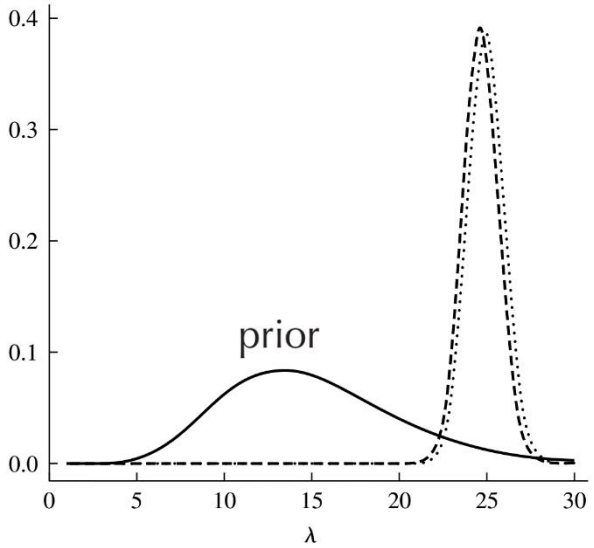
2nd moment is $$\mu_2 = \mu_1^2 + \sigma^2 = 250$$

Parameters ($\lambda \rightarrow \nu$): $$\nu = \frac{\mu_1}{\mu_2 - \mu_1^2} = 0.6$$

$$\alpha = \nu\mu_1 = 9$$

Prior distribution

$$f_\Lambda(\lambda) = \frac{\nu^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\nu\lambda}$$



Posterior density

$$f_{\Lambda|X}(\lambda|x) = \frac{\lambda^{\Sigma x_i+\alpha-1}e^{-(n+\nu)\lambda}}{\int_0^\infty \lambda^{\Sigma x_i+\alpha-1}e^{-(n+\nu)\lambda}d\lambda}$$

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$f_{\Lambda|X}(\lambda|x) = \frac{\lambda^{\Sigma x_i+\alpha-1} e^{-(n+\nu)\lambda}}{\int_0^\infty \lambda^{\Sigma x_i+\alpha-1} e^{-(n+\nu)\lambda} d\lambda}$$

$$\alpha' = \sum x_i + \alpha = 582$$
$$\nu' = n + \nu = 23.6$$

It must be a gamma density!

$$\mu_1 = \frac{\alpha}{\lambda}$$
$$\mu_2 = \frac{\alpha(\alpha+1)}{\lambda^2}$$

Posterior mean $\mu_{\text{post}} = \frac{\alpha'}{\nu'} = 24.7$

Posterior mode $(\alpha - 1)/\nu = 24.6$

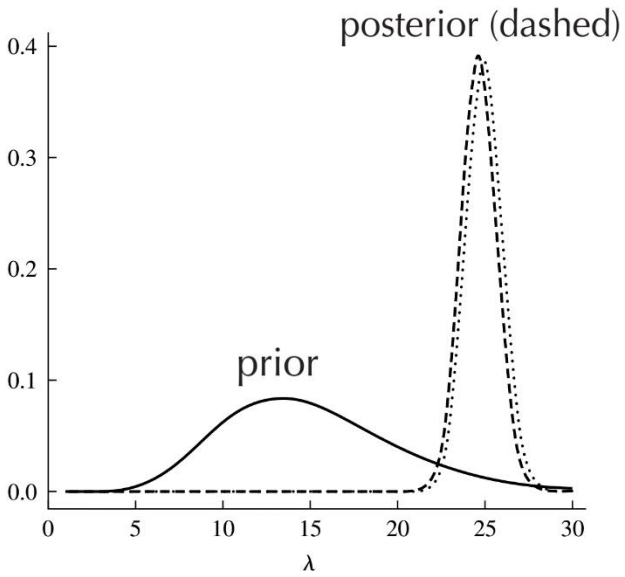
Variance of posterior distribution

$$\sigma^2_{\text{post}} = \frac{\alpha'}{\nu'^2} = 1.04 \qquad \sigma_{\text{post}} = 1.02$$

Bayesian analogue of 90% confidence interval:

Interval from the 5th percentile to the 95th percentile of the posterior, [23.02, 26.34].

A common alternative: use **high posterior density (HPD) interval**, i.e. a horizontal line cuts the density corresponding to 90% probability.



Two Bayesians: "he" and "she"…

She is a more utilitarian Bayesian, believing that it is implausible that the mean count λ could be larger than 100. She uses a simple prior uniform on [0, 100], without trying to quantify her opinion more precisely.

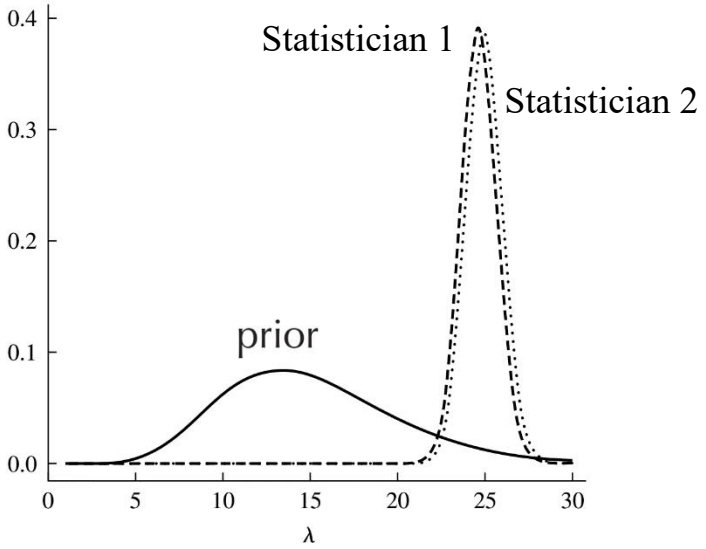
Posterior density is

$$f_{\Lambda|X}(\lambda|x) = \frac{\lambda^{\Sigma_{i=1}^n x_i} e^{-n\lambda} \frac{1}{100}}{\frac{1}{100}\int_0^{100} \lambda^{\Sigma_{i=1}^n x_i} e^{-n\lambda} d\lambda}, \qquad 0 \le \lambda \le 100$$

Gaussian-ish…
Why?

Numerical evaluations:

- posterior mode = 24.9
- posterior mean = 25.0
- posterior standard deviation = 1.04.
- Interval from 5th to 95th percentile = [23.3, 26.7]



	"he"	"she"	
Estimate	Bayes 1	Bayes 2	Maximum Likelihood
mode	24.6	24.9	24.9
mean	24.7	25.0	—
standard deviation	1.02	1.04	1.04
upper limit	26.3	26.7	26.6
lower limit	23.0	23.3	23.2

- Important: Her posterior density is directly proportional to the likelihood for $0 \leq \lambda \leq 100$, because prior is flat over this range.
- His prior opinion was inconsistent with data, but data strongly modified the prior to produce a posterior close to hers.

	"he"	"she"	
Estimate	Bayes 1	Bayes 2	Maximum Likelihood
mode	24.6	24.9	24.9
mean	24.7	25.0	—
standard deviation	1.02	1.04	1.04
upper limit	26.3	26.7	26.6
lower limit	23.0	23.3	23.2

- Important: Her posterior density is directly proportional to the likelihood for $0 \leq \lambda \leq 100$, because prior is flat over this range.
- His prior opinion was inconsistent with data, but data strongly modified the prior to produce a posterior close to hers.

Bayesian interpretation of the confidence intervals:

Λ is a random variable, "Given the observations, the probability that it is in the interval [23.3, 26.7] is 90%." The interval refers to the state of knowledge about λ and not to λ itself.

	"he"	"she"	
Estimate	Bayes 1	Bayes 2	Maximum Likelihood
mode	24.6	24.9	24.9
mean	24.7	25.0	—
standard deviation	1.02	1.04	1.04
upper limit	26.3	26.7	26.6
lower limit	23.0	23.3	23.2

- Important: Her posterior density is directly proportional to the likelihood for $0 \leq \lambda \leq 100$, because prior is flat over this range.
- His prior opinion was inconsistent with data, but data strongly modified the prior to produce a posterior close to hers.

Bayesian interpretation of the confidence intervals:

Λ is a random variable, "Given the observations, the probability that it is in the interval [23.3, 26.7] is 90%." The interval refers to the state of knowledge about λ and not to λ itself.

Frequentist framework:

Such a statement makes no sense, λ is a constant, it either lies in [23.3, 26.7] or doesn't – no probability is involved. Before the data are observed, the interval is random, one can state that the probability that the interval contains the true value is 90%, but after the data are observed, nothing is random anymore.

Example: Gaussian.

Replacing σ^2 by $\xi=1/\sigma^2$; ξ is called the **precision**; also using θ in place of μ,

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

Consider three cases:

1. Unknown mean, known variance
2. Unknown variance, known mean
3. Unknown mean, unknown variance

Case 1: precision is known, $\xi=\xi_0$, mean θ is unknown (random variable Θ)
Prior distribution is $N(\theta_0,\xi_{\text{prior}}^{-1})$, flat and uninformative when ξ_{prior} is small.

$$\begin{aligned}
f_{\Theta|X}(\theta|x) &\propto f_{X|\Theta}(x|\theta)\times f_{\Theta}(\theta)\\
&=\left(\frac{\xi_0}{2\pi}\right)^{n/2}\prod_{i=1}^{n}\exp\left(\frac{-\xi_0}{2}(x_i-\theta)^2\right)\times\left(\frac{\xi_{\text{prior}}}{2\pi}\right)^{1/2}\\
&\quad\times\exp\left(\frac{-\xi_{\text{prior}}}{2}(\theta-\theta_0)^2\right)\\
&\propto\exp\left(-\frac{1}{2}\left[\xi_0\sum_{i=1}^{n}(x_i-\theta)^2+\xi_{\text{prior}}(\theta-\theta_0)^2\right]\right)
\end{aligned}$$

θ-dependent terms only

Case 1: Unknown mean, known precision

$$f_{\Theta|X}(\theta|x) \propto \exp\left(-\frac{1}{2}\left[\xi_0 \sum_{i=1}^{n}(x_i-\theta)^2 + \xi_{\text{prior}}(\theta-\theta_0)^2\right]\right)$$

二项式，肯定可以配方

$$\sum(x_i-\theta)^2 = \sum(x_i-\bar{x})^2 + n(\theta-\bar{x})^2$$

$$f_{\Theta|X}(\theta|x) \propto \exp\left(-\frac{1}{2}[n\xi_0(\theta-\bar{x})^2 + \xi_{\text{prior}}(\theta-\theta_0)^2]\right)$$

Let's find $\xi_{\text{post}}(\theta-\theta_{\text{post}})^2 + \text{terms that do not depend on } \theta$

Expand the terms and identify the coefficients of θ^2, θ,

$$\xi_{\text{post}} = n\xi_0 + \xi_{\text{prior}}$$

$$\theta_{\text{post}} = \bar{x}\frac{n\xi_0}{n\xi_0+\xi_{\text{prior}}} + \theta_0\frac{\xi_{\text{prior}}}{n\xi_0+\xi_{\text{prior}}}$$

$$\xi_{\text{post}} = n\xi_0 + \xi_{\text{prior}}$$

$$\theta_{\text{post}} = \bar{x}\frac{n\xi_0}{n\xi_0 + \xi_{\text{prior}}} + \theta_0\frac{\xi_{\text{prior}}}{n\xi_0 + \xi_{\text{prior}}}$$

Posterior density of θ is normal:
precision has increased (surely it should!)
posterior mean is a weighted combination of sample mean and prior mean.

Try to better understand it --

Consider what happens when $\xi_{\text{prior}} << n\xi_0$, which would be the case if n were sufficiently large, or if ξ_{prior} were small (as for a very flat prior)

Posterior mean (same as MLE) $\theta_{\text{post}} \approx \bar{x}$

Posterior precision $\xi_{\text{post}} \approx n\xi_0 \Longrightarrow \sigma^2_{\text{post}} = \sigma_0^2/n$

X-bar in non-Bayesian setting

When a flat prior with very small ξ_{prior} is used,
posterior density of θ is very close to normal with mean $\bar{x}$ and variance σ_0^2/n.

Case 2: Known mean, unknown precision

Precision is treated as a random variable Ξ, prior distribution $f_\Xi(\xi)$.

Given ξ, the X_i are independent $N(\theta_0, \xi^{-1})$.

$$\begin{aligned} f_{\Xi|X}(\xi|x) &\propto f_{X|\Xi}(x|\xi) f_\Xi(\xi) \\ &\propto \xi^{n/2} \exp\left(-\frac{1}{2}\xi \sum (x_i - \theta_0)^2\right) f_\Xi(\xi) \end{aligned}$$

Dependence on ξ indicates analytical convenience to specify the prior to be a gamma density.

$$f_{\Xi|X}(\xi|x) \propto \xi^{n/2} \exp\left(-\frac{1}{2}\xi \sum (x_i - \theta_0)^2\right) \xi^{\alpha-1} e^{-\lambda\xi}$$

Posterior mean	$\mu_{\text{post}} = \frac{\alpha'}{\nu'}$
Posterior mode	$(\alpha - 1)/\nu$

$$\alpha_{\text{post}} = \alpha + \frac{n}{2}$$

$$\lambda_{\text{post}} = \lambda + \frac{1}{2}\sum (x_i - \theta_0)^2$$

$$\text{Posterior mean}^{-1} \approx \frac{1}{n}\sum (x_i - \theta_0)^2$$

$$\text{Posterior mode}^{-1} \approx \frac{1}{n-2}\sum (x_i - \theta_0)^2$$

$$\lambda \to 0, \alpha \to 0, \qquad f_{\Xi|X}(\xi|x) \propto \xi^{n/2-1} \exp\left(-\frac{1}{2}\xi \sum (x_i - \theta_0)^2\right)$$

Case 3: Unknown mean, unknown precision

Bayesian approach requires specification of a joint 2-d prior distribution. For convenience we take the priors to be independent:

$$\Theta \sim N\left(\theta_0, \xi_{\text{prior}}^{-1}\right)$$

$$\Xi \sim \Gamma(\alpha, \lambda)$$

$$f_{\Theta,\Xi|X}(\theta, \xi|x) \propto f_{X|\Theta,\Xi}(x|\theta, \xi) f_\Theta(\theta) f_\Xi(\xi)$$

$$\propto \xi^{n/2} \exp\left(-\frac{\xi}{2}\sum(x_i - \theta)^2\right) \quad \text{Gamma form}$$

$$\times \exp\left(-\frac{\xi_{\text{prior}}}{2}(\theta - \theta_0)^2\right) \xi^{\alpha-1} \exp(-\lambda\xi)$$

Has to be done numerically… but often the primary interest is θ, good thing about Bayesian: θ marginalized

$$f_{\Theta|X}(\theta|x) = \int_0^\infty f_{\Theta,\Xi|X}(\theta, \xi|x) d\xi$$

Recognizing Gamma density,

$$\tilde{\alpha} = \alpha + n/2 \qquad \tilde{\lambda} = \lambda + (1/2)\sum(x_i - \theta)^2$$

$$f_{\Theta|X}(\theta|x) \propto \exp\left(-\frac{\xi_{\text{prior}}}{2}(\theta - \theta_0)^2\right) \frac{\Gamma(\alpha + n/2)}{[\lambda + \frac{1}{2}\sum(x_i - \theta)^2]^{\alpha+n/2}}$$

$$f_{\Theta|X}(\theta|x) \propto \exp\left(-\frac{\xi_{\text{prior}}}{2}(\theta-\theta_0)^2\right)\frac{\Gamma(\alpha+n/2)}{[\lambda+\frac{1}{2}\sum(x_i-\theta)^2]^{\alpha+n/2}}$$

Still consider the case n is large or the prior is quite flat (α, λ, ξ_{prior} are small)

$$f_{\Theta|X}(\theta|x) \propto \left(\sum(x_i-\theta)^2\right)^{-n/2}$$

$\theta = \bar{x}$ maximize the posterior.

How is this related to MLE results?

$$\begin{aligned}\sum(x_i-\theta)^2 &= \sum(x_i-\bar{x})^2 + n(\theta-\bar{x})^2 \\ &= (n-1)s^2 + n(\theta-\bar{x})^2 \\ &= (n-1)s^2\left(1+\frac{n(\theta-\bar{x})^2}{(n-1)s^2}\right)\end{aligned}$$

Rearrange it,

$$f_{\Theta|X}(\theta|x) \propto \left(1+\frac{1}{n-1}\frac{n(\theta-\bar{x})^2}{s^2}\right)^{-n/2}$$

$$\frac{\sqrt{n}(\Theta-\bar{x})}{s} \sim t_{n-1}$$

自由度为 n 的 t 分布的密度函数是

$$f(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\Gamma(n/2)}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}$$

We saw this in confidence interval for MLE (exact method)

Prior, prior…

We saw: if prior for a Poisson parameter is chosen to gamma, posterior is gamma.
If prior for a normal mean with known variance is normal, then posterior is normal.

"**Conjugate priors**": if the prior distribution belongs to a family G and, conditional on the parameters of G, the data have a distribution H, then G is conjugate to H if the posterior is in the family G.

In scientific applications, it is usually desirable to use a flat, or "uninformative", prior so that the data can speak for themselves.

Even if a scientific investigator actually had a strong prior opinion, he or she might want to present an "objective" analysis so that the conclusions, as summarized in the posterior density, are those of one who is initially unopinionated or unprejudiced.

The objective prior thus has a hypothetical status: if one was initially indifferent to parameter values in the range in which the likelihood is large, then one's opinion after observing the data would be expressed as a posterior proportional to the likelihood.

Prior, prior… Improper?!

If α and v are very small, the gamma prior is quite flat and the posterior is proportional to the likelihood function. Formally, if α and v are set equal to zero, then the prior is

$$f_{\Lambda|\alpha,v}(\lambda) = \lambda^{-1}, \quad 0 \leq \lambda < \infty$$

But this is not probability density, it does not integrate to 1!!!
Similarly, for a Gaussian prior, the precision is set to 0 to make the prior flat:

$$f_{\Theta}(\theta) \propto 1, \quad -\infty < \theta < \infty$$

These are **improper priors**.

Using an improper prior may still lead to well-defined posterior density.

For the Poisson example, if $f_{\Lambda}(\lambda) \propto \lambda^{-1}$, then the denominator is

$$\int_0^{\infty} \lambda^{\sum x_i - 1} e^{-n\lambda} d\lambda < \infty$$

Resulting in a proper (Gamma) posterior

$$f_{\Lambda|X}(\lambda|x) \propto \lambda^{\sum x_i - 1} e^{-n\lambda}$$

Prior, prior… Improper?!

In the normal example with unknown mean and variance, we can take θ and ξ to be independent with improper priors $f_\Theta(\theta) = 1$ and $f_\Xi(\xi) = \xi^{-1}$. The joint posterior of θ and ξ is then (we used Gaussian, gamma)

$$f_{\Theta,\Xi|X}(\theta, \xi|x) \propto \xi^{n/2-1} \exp\left(-\frac{\xi}{2}\sum(x_i - \theta)^2\right)$$

Expressing $\sum_{i=1}^{n}(x_i - \theta)^2 = (n-1)s^2 + n(\theta - \bar{x})^2$, we have

$$f_{\Theta,\Xi|X}(\theta, \xi|x) \propto \xi^{n/2-1} \exp\left(-\frac{\xi}{2}(n-1)s^2\right) \exp\left(-\frac{n\xi}{2}(\theta - \bar{x})^2\right)$$

Conditional on ξ, θ is normal with mean *x-bar* and precision $n\xi$.
By integrating out ξ, we can find the marginal distribution of θ and relate it to the t distribution as was done earlier.

Only in the range where likelihood is large, the prior makes practical difference
— truncate the improper prior well outside this range to produce a proper prior

Large sample normal approximation to the posterior

We often see that posterior is nearly normal, mean=MLE, standard deviation close to asymptotic deviation of MLE. Why??

Posterior is

$$\begin{aligned} f_{\Theta|X}(\theta|x) &\propto f_{\Theta}(\theta) f_{X|\Theta}(x|\theta) \\ &= \exp[\log f_{\Theta}(\theta)] \exp[\log f_{X|\Theta}(x|\theta)] \\ &= \exp[\log f_{\Theta}(\theta)] \exp[l(\theta)] \end{aligned}$$

If sample is large, posterior dominated by likelihood, prior is nearly flat where likelihood is large,

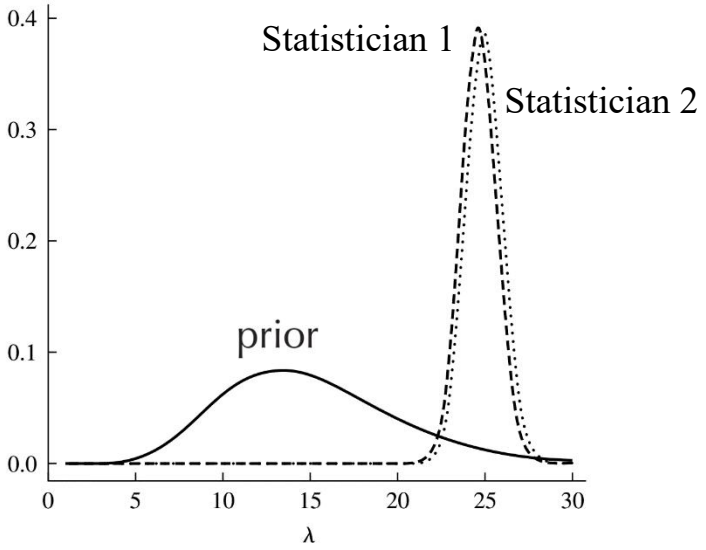
MLE

$l'(\hat{\theta}) = 0$

$$\begin{aligned} f_{\Theta|X}(\theta|x) &\propto \exp\left[l(\hat{\theta}) + (\theta - \hat{\theta})l'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 l''(\hat{\theta})\right] \\ &\propto \exp\left[\frac{1}{2}(\theta - \hat{\theta})^2 l''(\hat{\theta})\right] \end{aligned}$$

Posterior is approximately normal with the mean = the MLE, $\hat{\theta}$, and variance approximately equal to $-[l''(\hat{\theta})]^{-1}$.

Gaussian-ish…
Why?



Efficiency and the Cramér-Rao lower bound

Efficiency and the Cramér-Rao lower bound

There are a variety of possible parameter estimates: sample mean estimate, method of moments, MLE… How would we choose which to use? Choose the one with a sampling distribution most highly concentrated about the true parameter value.

To define this aim operationally, we need to specify a quantitative measure of such concentration. Mean squared error is the most commonly used, largely because of its analytic simplicity.

$$\begin{aligned} MSE(\hat{\theta}) &= E(\hat{\theta} - \theta_0)^2 \\ &= \mathrm{Var}(\hat{\theta}) + (E(\hat{\theta}) - \theta_0)^2 \end{aligned}$$

If the estimate $\hat{\theta}$ is unbiased $[E(\hat{\theta}) = \theta_0]$, $\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta})$.

Given two estimates, $\hat{\theta}$ and $\tilde{\theta}$, of a parameter θ, the **efficiency** of $\hat{\theta}$ relative to $\tilde{\theta}$ is defined to be

$$\mathrm{eff}(\hat{\theta}, \tilde{\theta}) = \frac{\mathrm{Var}(\tilde{\theta})}{\mathrm{Var}(\hat{\theta})}$$

The variances are often of the form $\mathrm{Var}(\hat{\theta}) = \frac{c_1}{n}$ $\mathrm{Var}(\tilde{\theta}) = \frac{c_2}{n}$

In this case, interpreted as ratio of sample sizes necessary to obtain same variance for both estimates.

Example: Muon decay

在介子衰变中，电子散射角度 θ 具有密度

$$f(x|\alpha)=\frac{1+\alpha x}{2},\quad -1\leqslant x\leqslant 1,\quad -1\leqslant \alpha\leqslant 1 \qquad x=\cos\theta.$$

Mean: $$\mu=\int_{-1}^{1} x\frac{1+\alpha x}{2}\mathrm{d}x=\frac{\alpha}{3}$$

So the methods of moments estimate is $\tilde{\alpha}=3\overline{X}$

While MLE is the solution of this equation $$\sum_{i=1}^{n}\frac{X_i}{1+\hat{\alpha}X_i}=0$$

To find efficiency, need variances.

$$\begin{aligned}\sigma^2 &= E(X^2)-[E(X)]^2\\ &=\int_{-1}^{1} x^2\frac{1+\alpha x}{2}\,dx-\frac{\alpha^2}{9}=\frac{1}{3}-\frac{\alpha^2}{9}\end{aligned}$$

the variance of the method of moments estimate is

$$\mathrm{Var}(\tilde{\alpha})=9\ \mathrm{Var}(\overline{X})=\frac{3-\alpha^2}{n}$$

Example: Muon decay

The exact variance of the mle, $\hat{\theta}$, cannot be computed in closed form, so we approximate it by the asymptotic variance,

$$\text{Var}(\hat{\alpha}) \approx \frac{1}{nI(\alpha)}$$

$$I(\alpha) = E\left[\frac{\partial}{\partial\alpha}\log f(x|\alpha)\right]^2$$

$$= \int_{-1}^{1} \frac{x^2}{(1+\alpha x)^2}\left(\frac{1+\alpha x}{2}\right) dx \quad = \frac{\log\left(\frac{1+\alpha}{1-\alpha}\right) - 2\alpha}{2\alpha^3}, \qquad -1 < \alpha < 1, \alpha \neq 0$$

$$= \frac{1}{3}, \qquad \alpha = 0$$

The **asymptotic relative efficiency**:

$$\frac{\text{Var}(\hat{\alpha})}{\text{Var}(\tilde{\alpha})} = \frac{2\alpha^3}{3-\alpha^2}\left[\frac{1}{\log\left(\frac{1+\alpha}{1-\alpha}\right) - 2\alpha}\right]$$

Example: Muon decay

- When α~0, MLE is not much better than moments
- As α tends to 1, MLE is increasingly better
- Note: we used the asymptotic variance of MLE, calculated an asymptotic relative efficiency
- To gain more precise information for a given sample size, conduct a simulation of the sampling distribution of MLE
- Simulation studies allow for analyzing the behavior of the bias as *n* and α vary (MLE is only *asymptotically unbiased*, there may be bias for a finite sample size)

α	Efficiency
0.0	1.0
.1	.997
.2	.989
.3	.975
.4	.953
.5	.931
.6	.878
.7	.817
.8	.727
.9	.582
.95	.464

Cramér-Rao lower bound

In searching for an optimal estimate, we want to ask:
Is there a lower bound for the MSE of *any* estimate?

If it exists, it would function as a benchmark against which estimates could be compared. If our estimate achieves this lower bound, we know that it could not be improved upon.

Answer: For unbiased estimates, Yes!

THEOREM **A** *Cramér-Rao Inequality*

Let $X_1, \ldots, X_n$ be i.i.d. with density function $f(x|\theta)$. Let $T = t(X_1, \ldots, X_n)$ be an unbiased estimate of θ. Then, under smoothness assumptions on $f(x|\theta)$,

$$\text{Var}(T) \geq \frac{1}{nI(\theta)}$$

Does it ring a bell to you?

I is the **Fisher information metric**

Cramér-Rao lower bound

- A lower bound on the variance of *any* unbiased estimate is given
- An unbiased estimate whose variance achieves this bound is **efficient**
- Asymptotic variance of MLEs = the lower bound, MLEs are **asymptotically efficient**
- If $n<\infty$, MLEs may be not efficient. Not the only asymptotically efficient estimates

Cramér-Rao lower bound

- A lower bound on the variance of *any* unbiased estimate is given
- An unbiased estimate whose variance achieves this bound is **efficient**
- Asymptotic variance of MLEs = the lower bound, MLEs are **asymptotically efficient**
- If $n<\infty$, MLEs may be not efficient. Not the only asymptotically efficient estimates

Example. Poisson distribution

Now we know for any unbiased estimate T of λ, based on an i.i.d. Poisson sample,

$$I(\lambda) = -E\left[\frac{\partial^2}{\partial\lambda^2}\log f(X|\lambda)\right] = -\frac{X}{\lambda^2} \qquad I(\lambda) = \frac{1}{\lambda} \qquad \mathrm{Var}(T) \geq \frac{\lambda}{n}$$

MLE of λ was found to be $\hat{\lambda} = \overline{X} = S/n$, where $S = X_1 + \cdots + X_n$.

S follows a Poisson distribution with parameter $n\lambda$, $\mathrm{Var}(S) = n\lambda$, $\mathrm{Var}(\overline{X}) = \lambda/n$.

$\overline{X}$ attains the Cramér-Rao lower bound, and we know that no unbiased estimator of λ can have a smaller variance. In this sense, $\overline{X}$ is optimal for the Poisson distribution.

*But there could be a **biased** estimator of λ with a smaller mean squared error!*

Sufficiency and theoretical support of using MLEs

Sufficiency

Suppose $X_1, X_2, \ldots, X_N$ is a sample from a probability distribution $f(x|\theta)$.

Question: Is there a statistic, a function $T(X_1, X_2, \ldots, X_N)$, that contains all the information in the sample about θ?

If so, reduction of the original data to a statistics without loss of information is possible.

Sufficiency

Suppose $X_1, X_2, \ldots, X_N$ is a sample from a probability distribution $f(x|\theta)$.

Question: Is there a statistic, a function $T(X_1, X_2, \ldots, X_N)$, that contains all the information in the sample about θ?

If so, reduction of the original data to a statistics without loss of information is possible.

Example.
A sequence of Bernoulli trials with unknown probability of success, θ.
The total # of successes is sufficient, by intuition (e.g. order of successes adds nothing to it.)

A statistic $T(X_1, \ldots, X_n)$ is said to be **sufficient** for θ if the conditional distribution of $X_1, \ldots, X_n$, given $T = t$, does not depend on θ for any value of t. ■

Given T, the **sufficient statistic**, we can gain no more knowledge about θ from knowing more about the probability distribution of $X_1, X_2, \ldots, X_N$.

Sufficiency

Example.

Let $X_1, \ldots, X_n$ be a sequence of independent Bernoulli random variables with $P(X_i = 1) = \theta$. We will verify that $T = \sum_{i=1}^{n} X_i$ is sufficient for θ.

$$P(X_1 = x_1, \ldots, X_n = x_n | T = t) = \frac{P(X_1 = x_1, \ldots, X_n = x_n, T = t)}{P(T = t)}$$

Numerator: t 1s and $(n-t)$ 0s, probability $\theta^t(1-\theta)^{n-t}$

Denominator: total # of 1s is binomial with n trials. RHS is

$$\frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}$$

Given the total # of 1s, the probability that they occur on any particular set of t trials is the same for any value of θ, so that set of trials contains no additional information about θ.

Sufficiency: A factorization theorem

Motivation. The definition of sufficiency is hard to work with:

- it does not indicate how to go about finding a sufficient statistic
- Given a candidate statistic, T, it is typically very hard to conclude whether it was sufficient due to the difficulty in evaluating the conditional distribution.

Sufficiency: A factorization theorem

Motivation. The definition of sufficiency is hard to work with:

- it does not indicate how to go about finding a sufficient statistic
- Given a candidate statistic, T, it is typically very hard to conclude whether it was sufficient due to the difficulty in evaluating the conditional distribution.

Fisher–Neyman factorization theorem

A necessary and sufficient condition for $T(X_1, \ldots, X_n)$ to be sufficient for a parameter θ is that the joint probability function (density function or frequency function) factors in the form

$$f(x_1, \ldots, x_n|\theta) = g[T(x_1, \ldots, x_n), \theta]h(x_1, \ldots, x_n)$$

The factorization theorem provides a convenient means of identifying sufficient statistics.

Sufficiency: A factorization theorem

Example.
Consider a sequence of independent Bernoulli random variables, $X_1, X_2, \ldots, X_n$.

$$P(X_i = x) = \theta^x (1-\theta)^{1-x}, \qquad x = 0 \text{ or } x = 1$$

$$\begin{aligned} f(\mathbf{x}|\theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\Sigma_{i=1}^n x_i}(1-\theta)^{n-\Sigma_{i=1}^n x_i} \\ &= \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^n x_i}(1-\theta)^n \end{aligned}$$

RHS depends only on $x_1, x_2, \ldots, x_n$ through the sufficient statistic $t = \sum_{i=1}^n x_i$

$$g(t,\theta) = \left(\frac{\theta}{1-\theta}\right)^t (1-\theta)^n \qquad h(\mathbf{x}) = 1$$

$$f(x_1, \ldots, x_n|\theta) = g[T(x_1, \ldots, x_n), \theta]h(x_1, \ldots, x_n)$$

Sufficiency: A factorization theorem

Example.
Consider a random sample from a normal distribution, mean and variance unknown.

$$
\begin{aligned}
f(\mathbf{x}|\mu,\sigma) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2\sigma^2}(x_i-\mu)^2\right] \\
&= \frac{1}{\sigma^n(2\pi)^{n/2}} \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\
&= \frac{1}{\sigma^n(2\pi)^{n/2}} \exp\left[\frac{-1}{2\sigma^2}\left(\sum_{i=1}^{n}x_i^2 - 2\mu\sum_{i=1}^{n}x_i + n\mu^2\right)\right]
\end{aligned}
$$

Just a function of $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n} x_i^2$, sufficient statistics (2-dimensional).

$$f(x_1,\ldots,x_n|\theta) = g[T(x_1,\ldots,x_n),\theta]h(x_1,\ldots,x_n)$$

Sufficiency: A factorization theorem

The likelihood now

$$f(x_1, \ldots, x_n; \theta) = g[T(x_1, \ldots, x_n), \theta]h(x_1, \ldots, x_n)$$

depends only on data through T, so MLE is maximizing $g[T, \theta]$.
In the Bernoulli example, likelihood is function of $t = \sum_{i=1}^{n} x_i$, MLE is $\hat{\theta} = t/n$.

Bayesian framework: posterior distribution of θ is proportional to prior • likelihood.

So, posterior depends only on the data through $g[T(x_1, x_2, \ldots, x_n), \theta]$

— the posterior probability of θ is the same for all $\{x_1, x_2, \ldots, x_n\}$ which have a common value of $T(x_1, x_2, \ldots, x_n)$. The sufficient statistic carries all the information about θ that is contained in the data $x_1, x_2, \ldots, x_n$.

Therefore, we have the corollary:

If T is sufficient for θ, the maximum likelihood estimate is a function of T.

Theoretical support to use MLEs

If T is sufficient for θ, the maximum likelihood estimate is a function of T.

The Rao-Blackwell theorem

Let $\hat{\theta}$ be an estimator of θ with $E(\hat{\theta}^2) < \infty$ for all θ. Suppose that T is sufficient for θ, and let $\tilde{\theta} = E(\hat{\theta}|T)$. Then, for all θ,

$$E(\tilde{\theta} - \theta)^2 \leq E(\hat{\theta} - \theta)^2$$

The inequality is strict unless $\hat{\theta} = \tilde{\theta}$.

Since $E(\hat{\theta}|T)$ is a function of the sufficient statistic T, the Rao-Blackwell theorem gives a strong rationale for basing estimators on sufficient statistics if they exist. If an estimator is not a function of a sufficient statistic, it can be improved.