Statistical Inference

Lecture 10b

ANU - RSFAS

Last Updated: Tue May 8 16:45:05 2018

Bayesian Testing

- Consider testing:

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1$$

- The classical statistician considers θ a fixed unknown, thus a hypothesis test is true or false. Either θ is in Θ_0 or Θ_1!
- Bayesians however consider θ to be random and it is quite natural to consider:

$$P(\theta \in \Theta_0 | \boldsymbol{x}) \quad \text{vs.} \quad P(\theta \in \Theta_1 | \boldsymbol{x})$$

Bayesian Testing

- One approach to Bayesian testing is to reject H_0 if:

$$P(\theta \in \Theta_1|\boldsymbol{x}) > P(\theta \in \Theta_0|\boldsymbol{x})$$

Bayesian Testing

Example 8.2.7: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{normal}(\theta, \sigma^2)$. Let $\theta \sim \text{normal}(\mu, \tau^2)$, where σ^2, μ, τ^2 are known. Consider testing:

$$H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0$$

normal posterior

$$[\theta | \boldsymbol{x}] \sim \text{normal}\left(\frac{\sigma^2\mu + n\tau^2\bar{x}}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)$$

- To compare we simply examine:

$$\int_{-\infty}^{\theta_0} p(\theta|\boldsymbol{x})d\theta \quad \text{vs.} \quad \int_{\theta_0}^{\infty} p(\theta|\boldsymbol{x})d\theta$$

Bayesian Testing

- Based on this Bayesian approach to hypothesis testing, what if we wanted to test:

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0$$

$$[\theta|\boldsymbol{x}] \sim \text{normal}\left(\frac{\sigma^2\mu + n\tau^2\bar{x}}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)$$

- $[\theta|\boldsymbol{x}]$ is a continuous distribution, so the probability of any single point (such as θ_0) is zero. This approach does not seem to work.

Bayesian Testing

1

- Notice that when we test through a Bayesian approach, we model the parameter of interest (i.e. we put a prior on it).
- If our scientific question concerns whether a parameter can be exactly θ_0 or not, then we should model that:

"prior is a mixture"

$$\theta \sim p\mathbf{1}_{\theta=\theta_0} + (1-p)\text{normal}(\mu, \tau^2)$$

where $0 \leq p \leq 1$.

- In a regression or GLM setting many times we are interested in testing whether $\beta = 0$ vs. $\beta \neq 0$ and these priors or variants on them can be quite useful.

Bayesian Testing

2

- Another approach that does allow for consideration of $\beta = 0$ is Bayes factors. Let's rephrase hypothesis testing as choosing between competing models:

$$\begin{cases} \text{Model 1 } (M_1)\text{: } y_i = \alpha + \epsilon_i & \epsilon_i \stackrel{\text{iid}}{\sim} \text{normal}(0, \sigma^2); \quad \theta_1 = \{\alpha, \sigma^2\} \\ \text{Model 2 } (M_2)\text{: } y_i = \alpha + \beta x_i + \epsilon_i & \epsilon_i \stackrel{\text{iid}}{\sim} \text{normal}(0, \sigma^2); \quad \theta_2 = \{\alpha, \beta, \sigma^2\} \end{cases}$$

Bayesian Testing

based on p.p. choose a model.

- Let's figure out the posterior probability for a given model i:

$$\pi(M_i|\boldsymbol{x}) = \frac{\overbrace{f(\boldsymbol{x}|M_i)}^{\text{marginal prob. of the data given the model.}}\pi(M_i)}{m(\boldsymbol{x})}$$

→ marginal prob of the data.

$$f(\boldsymbol{x}|M_i) = \int_\theta f(\boldsymbol{x}|\theta, M_i)\pi(\theta|M_i)d\theta_i$$

$$m(\boldsymbol{x}) = \sum_{i=1}^{2} f(\boldsymbol{x}|M_i)\pi(M_i)$$

Bayesian Testing

- Now consider the following ratio of the posterior model probabilities:

interest!

$$\frac{\pi(M_2|\boldsymbol{x})}{\pi(M_1|\boldsymbol{x})} = \frac{f(\boldsymbol{x}|M_2)}{f(\boldsymbol{x}|M_1)} \times \frac{\pi(M_2)}{\pi(M_1)} = BF(M_2; M_1) \times \frac{\pi(M_2)}{\pi(M_1)}$$

Where $BF(M_2; M_1)$ is called the Bayes factor.

- Typically $\pi(M_2) = \pi(M_1)$, so the ratio of the posterior probabilities is the Bayes factor.
- The Bayes factor looks "like" a likelihood ratio. However, the difference is that θ has been integrated out in both the numerator and denominator, so we have the marginal distribution of the data given the model.

Bayesian Testing

- If $f(\boldsymbol{x}|M_2) > f(\boldsymbol{x}|M_1)$ or $\frac{f(\boldsymbol{x}|M_2)}{f(\boldsymbol{x}|M_1)} > 1$ then we have support for M_2 against M_1.
- Jeffreys, H. (1961 appendix B) suggested the following:

$BF(M_2; M_1) = B_{21}$	Evidence against model 1 (H_0)
1 to 3.2	Not worth more than a bare mention
3.2 to 10	Substantial
10 to 100	Strong
> 100	Decisive

Bayesian Interval Estimation

- Suppose we have data $X_1, \dots, X_n$ from density $f_X(x|\theta)$ along with a prior distribution $\pi(\theta)$. As we saw we use Bayes' rule to update our 'beliefs' about θ once we observe the data:

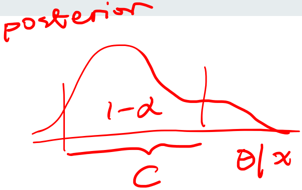
$$\pi(\theta|\boldsymbol{x}) = \frac{L(\theta|\boldsymbol{x})\pi(\theta)}{\int_{\theta\in\Theta} L(\theta|\boldsymbol{x})\pi(\theta)d\theta} = \frac{L(\theta|\boldsymbol{x})\pi(\theta)}{m(\boldsymbol{x})}$$

- So we have the whole distribution for

$$\pi(\theta|\boldsymbol{x})$$

- This is different than the frequentist approach where find an estimator for θ, say $\hat{\theta}$ and then try to determine the distribution of $\hat{\theta}$.

Bayesian Interval Estimation



- To obtain an interval we simply consider:

$$P\pi(\theta|x)(C) = \int_C \pi(\theta|\boldsymbol{x})d\theta = 1 - \alpha$$

- Be careful. We are using α quite generically. Recall that α does have a formal definition: The probability of a Type-I error. This is based on repeated sampling. For the Bayesian case we only think about one data set an infinite number of possible data sets.
- There are quite a lot of choices for C. We will consider the 3 most common.

Bayesian Interval Estimation

①
- Equal tailed: both sides with $\frac{\alpha}{2}$ tails

$$\int_{-\infty}^{\theta_L} \pi(\theta|\boldsymbol{x})d\theta = \alpha/2, \quad \int_{\theta_U}^{\infty} \pi(\theta|\boldsymbol{x})d\theta = \alpha/2$$

②
- Smallest length: We can choose C to minimize $\theta_U - \theta_L$.

Bayesian Interval Estimation

3

- Highest posterior density region (HPD): We define C to be that set with posterior probability $1-\alpha$ which satisfies the criterion:

"prob"

$$\theta_1 \in C \quad \text{and} \quad \pi(\theta_2|\boldsymbol{x}) > \pi(\theta_1|\boldsymbol{x}) \Rightarrow \theta_2 \in C$$

C contains the values of θ which have the highest posterior density values, so that we can determine HPD regions as the set:

$$C = \{\theta \in \Theta : \pi(\theta|\boldsymbol{x}) > c_\alpha\}$$

- If the posterior is unimodal then this will be the smallest length interval!

"centred around the peak"

Bayesian Interval Estimation

$E(X) = \frac{1}{\theta}$

Example: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{exponential}(1/\theta)$ and $\pi(\theta) = \theta exp(-\theta)$.

$$
\begin{aligned}
\pi(\theta|\boldsymbol{x}) \quad \propto \quad & \underbrace{\left\{\prod_{i=1}^{n} \theta exp(-x_i\theta)\right\}}_{\text{likelihood}} \underbrace{\theta exp(-\theta)}_{\text{prior}} \\
& = \theta^n exp(-\sum x_i\theta)\theta exp(-\theta) \\
& = \theta^{n+1} exp(-\theta(n\bar{x}+1)) \\
& = \theta^{n+2-1} exp(-\theta(n\bar{x}+1))
\end{aligned}
$$

posterior is a Gamma

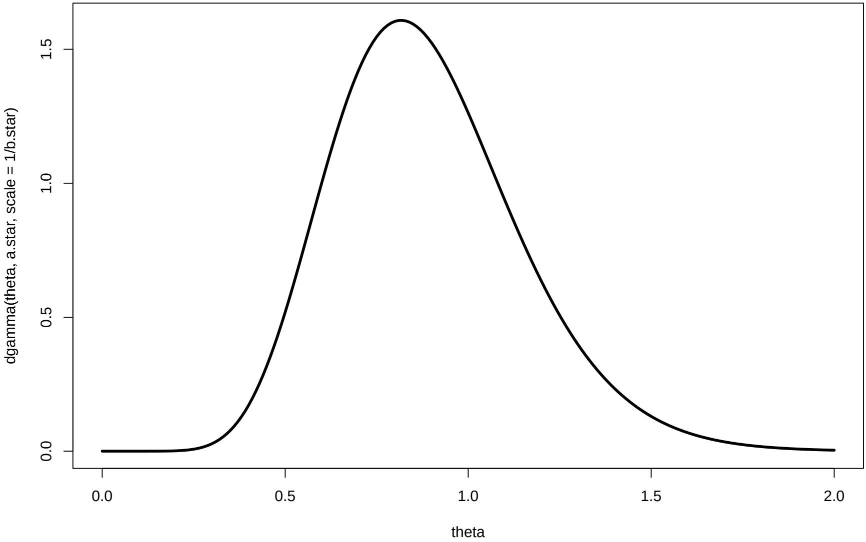
$$[\theta|\boldsymbol{x}] \sim gamma\left(n+2, \frac{1}{n\bar{x}+1}\right)$$

Bayesian Interval Estimation

- Let's plot the density for $n = 10$ and $\bar{x} = 1.247$.

```
n=10; x.bar <- 1.247
a.star <- n+2; b.star <- n*x.bar+1
theta <- seq(0, 2, by=0.01)
plot(theta, dgamma(theta, a.star,
                   scale=1/b.star), type="l", lwd=3)
```

Bayesian Interval Estimation



Bayesian Interval Estimation

- An equal-tailed 95% interval is given by $[\theta_l, \theta_u]$:

$$\begin{aligned}\int_0^{\theta_l} \pi(\theta|\boldsymbol{x}) &= 0.025 \\ F_{[\theta|\boldsymbol{x}]}(\theta_l) &= 0.025\end{aligned}$$

$$\begin{aligned}\int_0^{\theta_u} \pi(\theta|\boldsymbol{x}) &= 1 - 0.025 = 0.975 \\ F_{[\theta|\boldsymbol{x}]}(\theta_u) &= 0.975\end{aligned}$$

Bayesian Interval Estimation

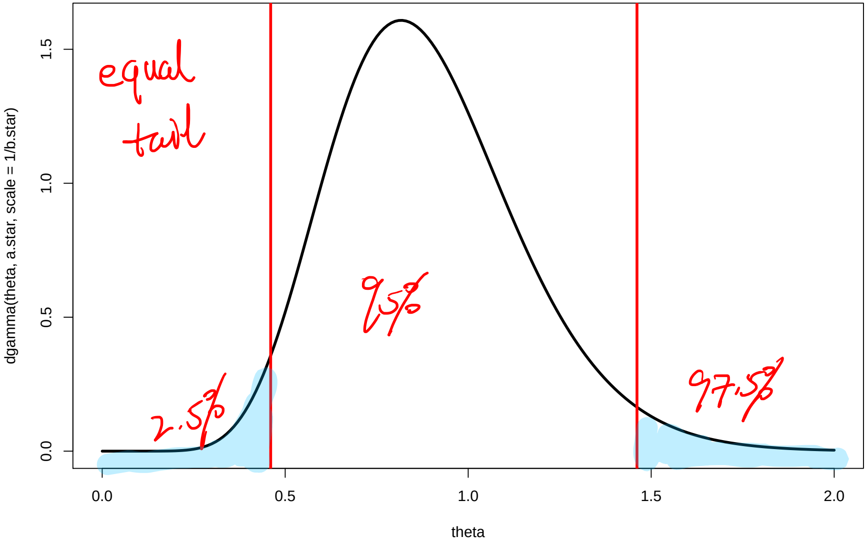
```r
theta.L <- qgamma(0.025,  a.star, scale=1/b.star)
theta.U <- qgamma(0.975,  a.star, scale=1/b.star)

c(theta.L, theta.U)
```

```
## [1] 0.4603248 1.4611758
```

"credible interval"

```r
plot(theta, dgamma(theta, a.star, scale=1/b.star), type="l", l
abline(v=c(theta.L, theta.U), lwd=3, col="red")
```

Bayesian Interval Estimation



Bayesian Interval Estimation

- If we only have tables in front of us, we can relate the gamma distribution to a χ^2 distribution as was discussed in tutorial the other week.
- If $[\theta|\boldsymbol{x}] \sim \mathrm{gamma}(a^*, b^*)$ then

$$\left[\frac{2\theta}{b^*}\middle|\boldsymbol{x}\right] \sim \mathrm{gamma}(a^*, 2)$$
$$\sim \chi^2_{p=2a^*}$$

Bayesian Interval Estimation

- Using probabilities to the left. $p = 2a^* = 2n + 4$.

pivoting

$$\left[\chi^2_{0.025,p} \le \frac{2\theta}{b^*}\Big|\boldsymbol{x} \le \chi^2_{0.975,p}\right]$$

$$\left[\chi^2_{0.025,p} \le 2\theta(n\bar{x}+1)|\boldsymbol{x} \le \chi^2_{0.975,p}\right]$$

$$\left[\frac{\chi^2_{0.025,p}}{2(n\bar{x}+1)} \le \theta|\boldsymbol{x} \le \frac{\chi^2_{0.975,p}}{2(n\bar{x}+1)}\right]$$

```
p <- 2*n + 4
theta.L <- qchisq(0.025, p)/(2*(n*x.bar+1))
theta.U <- qchisq(0.975, p)/(2*(n*x.bar+1))

c(theta.L, theta.U)
```

```
## [1] 0.4603248 1.4611758
```

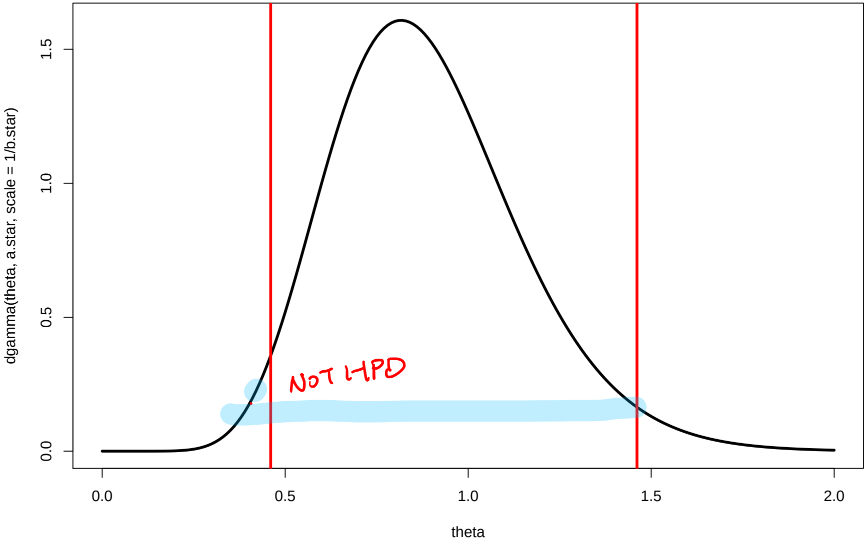
Bayesian Interval Estimation

- Is the interval $[0.4603, 1.4612]$ a HPD (highest posterior density) interval (the posterior is unimodal)?
- Recall:

$$\theta_1 \in C \quad \text{and} \quad \pi(\theta_2|\boldsymbol{x}) > \pi(\theta_1|\boldsymbol{x}) \Rightarrow \theta_2 \in C$$

- Let's see the density with the equal-tailed interval again.

Bayesian Interval Estimation



- Note that the density seems to be higher for $\theta = 0.40$ than $\theta = 1.4612$:

```
dgamma(0.4, a.star, scale=1/b.star)
```

```
## [1] 0.1713707
```

```
dgamma(1.4612, a.star, scale=1/b.star)
```

```
## [1] 0.1641042
```

- So the equal-tailed interval is not a HPD interval!

density should be the same!

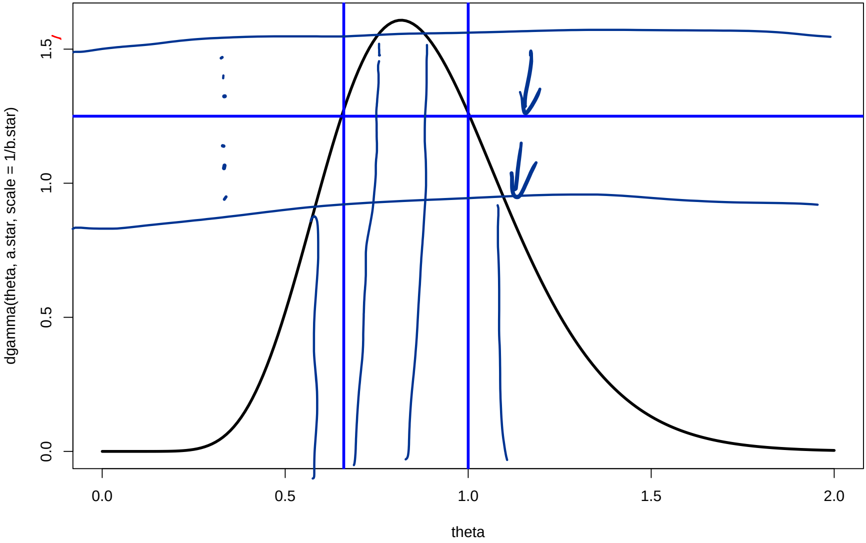
leftmost & rightmost

Bayesian Interval Estimation

- To get the HPD interval we take horizontal slices across the density till we get the appropriate probability.

```
plot(theta, dgamma(theta, a.star, scale=1/b.star), type="l", lwd=3)
abline(h=1.25, lwd=3, col="blue")

##
theta <- seq(0, 2, by=0.01)
dens <- dgamma(theta, a.star, scale=1/b.star)

##
hpd.cut <- 1.25
theta.L <- min(theta[dens>=hpd.cut])
theta.U <- max(theta[dens>=hpd.cut])
abline(v=c(theta.L, theta.U), lwd=3, col="blue")

## interval probability
pgamma(theta.U, a.star, scale=1/b.star) -
  pgamma(theta.L, a.star, scale=1/b.star)
```

Bayesian Interval Estimation



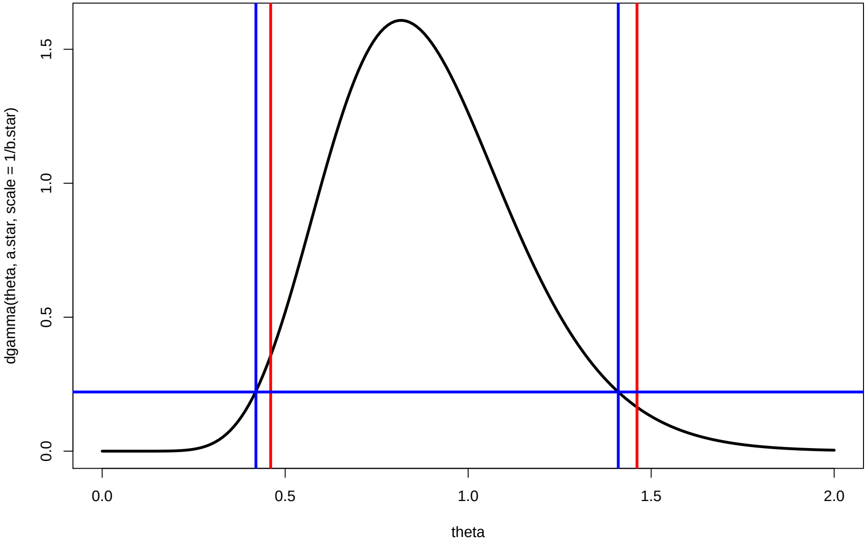

```
## [1] 0.5062717
```

Bayesian Interval Estimation

```r
hpd.cut <- sort(seq(0.1, 1.25, by=0.0001), decreasing =TRUE)
c <- 1
cred.int <- 0.5063

while(cred.int<0.95){
theta.L <- min(theta[dens>=hpd.cut[c]])
theta.U <- max(theta[dens>=hpd.cut[c]])

## interval probability
cred.int <- pgamma(theta.U, a.star, scale=1/b.star) -
  pgamma(theta.L, a.star, scale=1/b.star)
c <- c+1
}

HPD <- c(theta.L,theta.U)
HPD
```

```
## [1] 0.42 1.41
```

Bayesian Interval Estimation



Bayesian Interval Estimation

- Let's check the length of each interval:
  - equal-tailed: $1.46 - 0.460 = 1.00$
  - HPD: $1.41 - 0.42 = 0.99$
- HPD is the shorter interval, but not by much.

Bayesian Inference: Properties

Definition 7.1: A statistic $T(\boldsymbol{X})$ is **sufficient** for $\boldsymbol{\theta}$ *if and only if* the posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{X}$ is the same as the posterior distribution of $\boldsymbol{\theta}$ given $T(\boldsymbol{X})$.

Proof: Note that Definitions 2.5 and 7.1 are the same!

Suppose that $T(\boldsymbol{X})$ satisfies Definition 2.5. Then:

no longer have theta there

$$f(\boldsymbol{x};\boldsymbol{\theta}) = g(\boldsymbol{x}|t,\boldsymbol{\theta})h(t|\boldsymbol{\theta}) = g(\boldsymbol{x}|t)h(t|\boldsymbol{\theta})$$

- The posterior is

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}) &\propto f(\boldsymbol{x};\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &\propto h(t|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &\propto p(\boldsymbol{\theta}|t) \end{aligned}$$

- Now assume that $T(\boldsymbol{X})$ satisfies Definition 7.1.

likelihood

$$
\begin{aligned}
f(\boldsymbol{x}|\boldsymbol{\theta}) &= \frac{p(\boldsymbol{\theta}|\boldsymbol{x})h(\boldsymbol{x})}{p(\boldsymbol{\theta})} \\
&= \frac{p(\boldsymbol{\theta}|t)h(\boldsymbol{x})}{p(\boldsymbol{\theta})} \\
&= K_1[t|\boldsymbol{\theta}]\ K_2[\boldsymbol{x}]
\end{aligned}
$$

- From the **factorization theorem**, it follows that $T(\boldsymbol{X})$ is a sufficient statistic.

Bayesian Inference: Asymptotics - Rough Idea

- Suppose we have $y_1, \ldots, y_n \sim p(y|\theta)$.
- Let's consider the posterior distribution:

$$
\begin{aligned}
p(\theta|\boldsymbol{y}) \quad &\propto \quad p(\boldsymbol{y}|\theta)\ p(\theta) \\
&= exp\left[log\ \ p(\boldsymbol{y}|\theta)\right] exp\left[log\ \ p(\theta)\right]
\end{aligned}
$$

- As $n \rightarrow \infty$ the posterior is dominated by the likelihood.

$$
p(\theta|\boldsymbol{y}) \propto exp\left[log\ \ p(\boldsymbol{y}|\theta)\right]
$$

- Thus to an approximation we have the following:

$$
\begin{aligned}
p(\theta|\mathbf{y}) &\propto exp[log \quad p(\mathbf{y}|\theta)] \\
&\propto exp\,[\ell(\theta)] \\
&\propto exp\left[\ell(\hat{\theta}) + \overset{0}{(\theta-\hat{\theta})\ell'(\hat{\theta})} + \frac{1}{2}(\theta-\hat{\theta})^2\ell''(\hat{\theta})\right] \\
&\propto exp\left[\frac{1}{2}(\theta-\hat{\theta})^2\ell''(\hat{\theta})\right]
\end{aligned}
$$

- Where: $\hat{\theta}$ is the MLE.
- Note: $\ell(\hat{\theta})$ is a constant.
- Note: $\ell'(\hat{\theta}) = 0$

$X \sim P(\theta)$

$E(\bar{X}) \to \theta$

$\bar{X} \xrightarrow{LLN} E(\theta) = \theta$

sample mean

dist mean

$$
\begin{aligned}
p(\theta|\mathbf{y}) &\propto exp\left[\frac{1}{2}(\theta - \hat{\theta})^2 \ell''(\hat{\theta})\right] \\
&\propto exp\left[\frac{1}{2}(\theta - \hat{\theta})^2 \left[-I(\hat{\theta})\right]\right] \\
&\propto \underbrace{exp\left[-\frac{1}{2\left[I(\hat{\theta})\right]^{-1}}(\theta - \hat{\theta})^2\right]}_{\text{the kernel for normal distribution}}
\end{aligned}
$$

converges to that

$\ell''(\hat{\theta}) = H(\hat{\theta})$

- We see that this expression is proportional to a normal distribution. So we have:

$$
p(\theta|\mathbf{y}) \approx \text{normal}\left(\hat{\theta}, \left[I(\hat{\theta})\right]^{-1}\right)
$$