CS 215
Data Analysis and Interpretation

Bayesian Statistical Analysis

Suyash P. Awate

Bayesian Statistical Analysis

- Thomas Bayes
  - 18th-century mathematician, statistician

Portrait purportedly of Bayes used in a 1936 book,[1] but it is doubtful whether the portrait is actually of him.[2] No earlier portrait or claimed portrait survives.

Bayesian Statistical Analysis

- List of things named after Thomas Bayes

Bayes [edit]

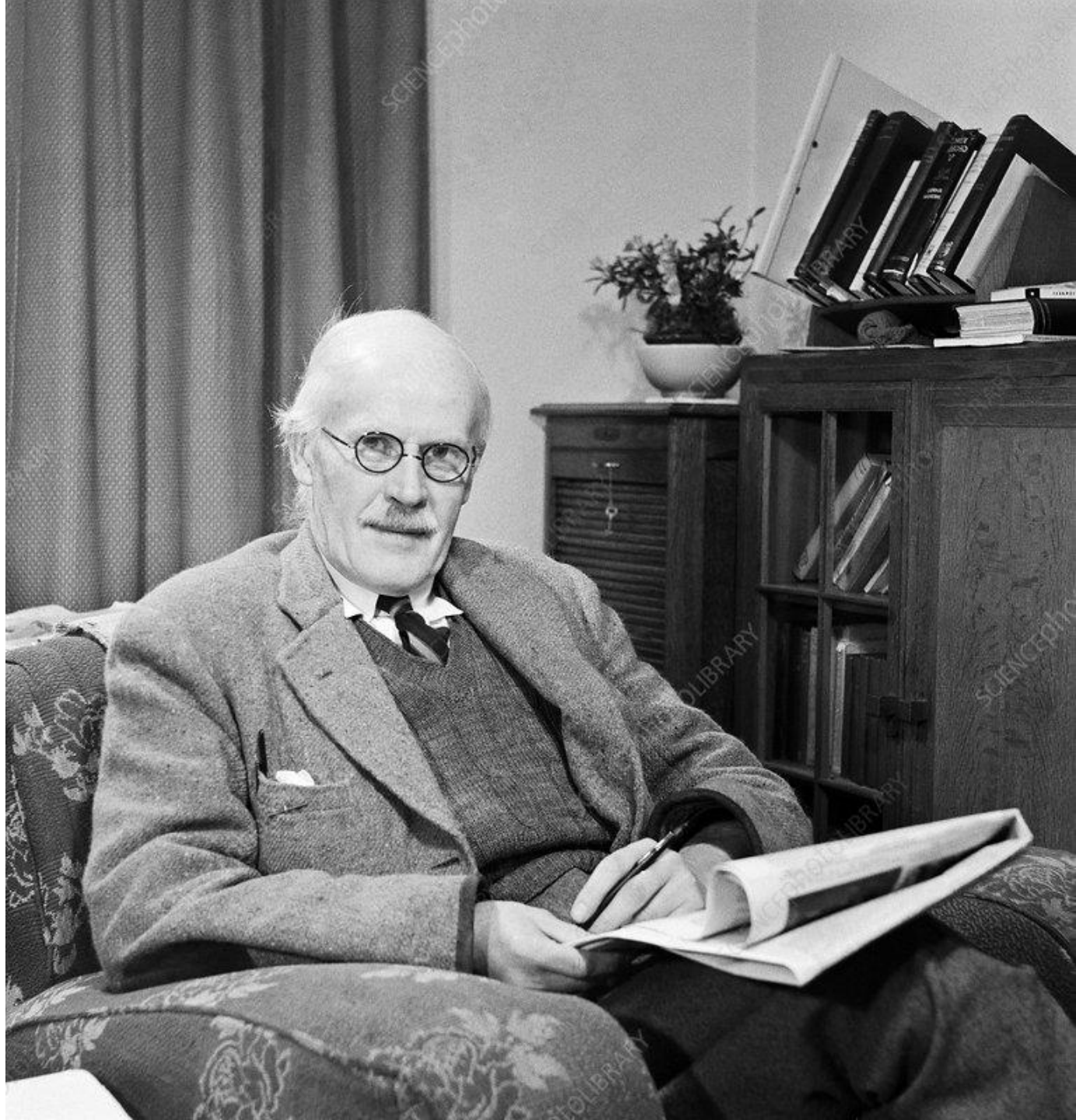
- Bayes error rate
- Bayes estimator
- Bayes factor
- Bayes linear statistics
- Bayes prior
- Bayes' theorem (or rule)
- Empirical Bayes method
- Evidence under Bayes theorem
- Hierarchical Bayes model
- Laplace–Bayes estimator
- Naive Bayes classifier
- Random naive Bayes

Bayesian [edit]

- Approximate Bayesian computation
- Bayesian average
- Bayesian approaches to brain function
- Bayesian econometrics
- Bayesian efficiency
- Bayesian experimental design
- Bayesian game
- Bayesian inference
- Bayesian inference in phylogeny
- Bayesian information criterion
- Bayesian linear regression
- Bayesian model selection
- Bayesian multivariate linear regression
- Bayesian network
- Bayesian poisoning
- Bayesian probability
- Bayesian programming
- Bayesian search theory
- Bayesian spam filtering
- Bayesian statistics
- Bayesian tool for methylation analysis
- Bayesian vector autoregression
- Dynamic Bayesian network
- International Society for Bayesian Analysis
- Quantum Bayesianism
- Recursive Bayesian estimation
- Robust Bayesian analysis
- Variable-order Bayesian network

Bayesian Statistical Analysis

- Sir Harold Jeffreys
  - 20th-century mathematician, statistician
  - “Bayes’ theorem is to the theory of probability what Pythagoras’s theorem is to geometry”



Bayesian Statistical Analysis

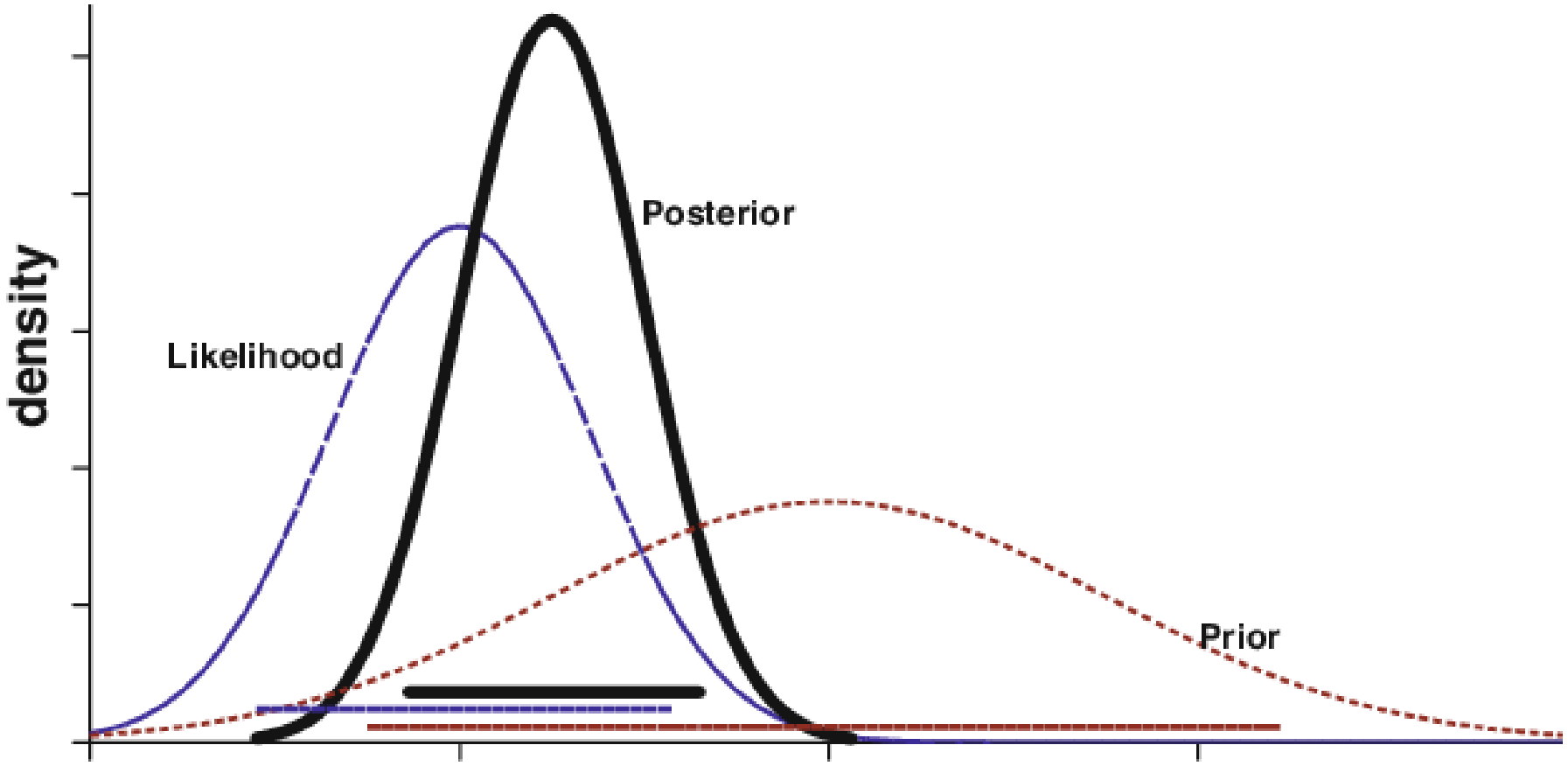
- Bayes theorem
  - X = **discrete** random variable (e.g., modeling some unknown parameter)
  - Y = continuous/discrete random variable (e.g., modeling observed data)
  - Consider joint distribution: P (X,Y)
  - Likelihood : P (Y=y | X=x)
    - Conditional
  - Evidence : $P (Y=y) = \Sigma_x P (X=x, Y=y)$
    - Marginal. Accounts for various ways in which data could have been generated.
  - Prior : P (X)
    - Before data is observed. Models beliefs (e.g., evolved over past experiences / learning)
  - Posterior : P (X=x | Y=y)
    - After data is observed. Data updates the beliefs.
  - Posterior P (X=x | Y=y)
    = Likelihood P (Y=y | X=x) * Prior P (X=x)
    / P (Y=y)

Bayesian Statistical Analysis

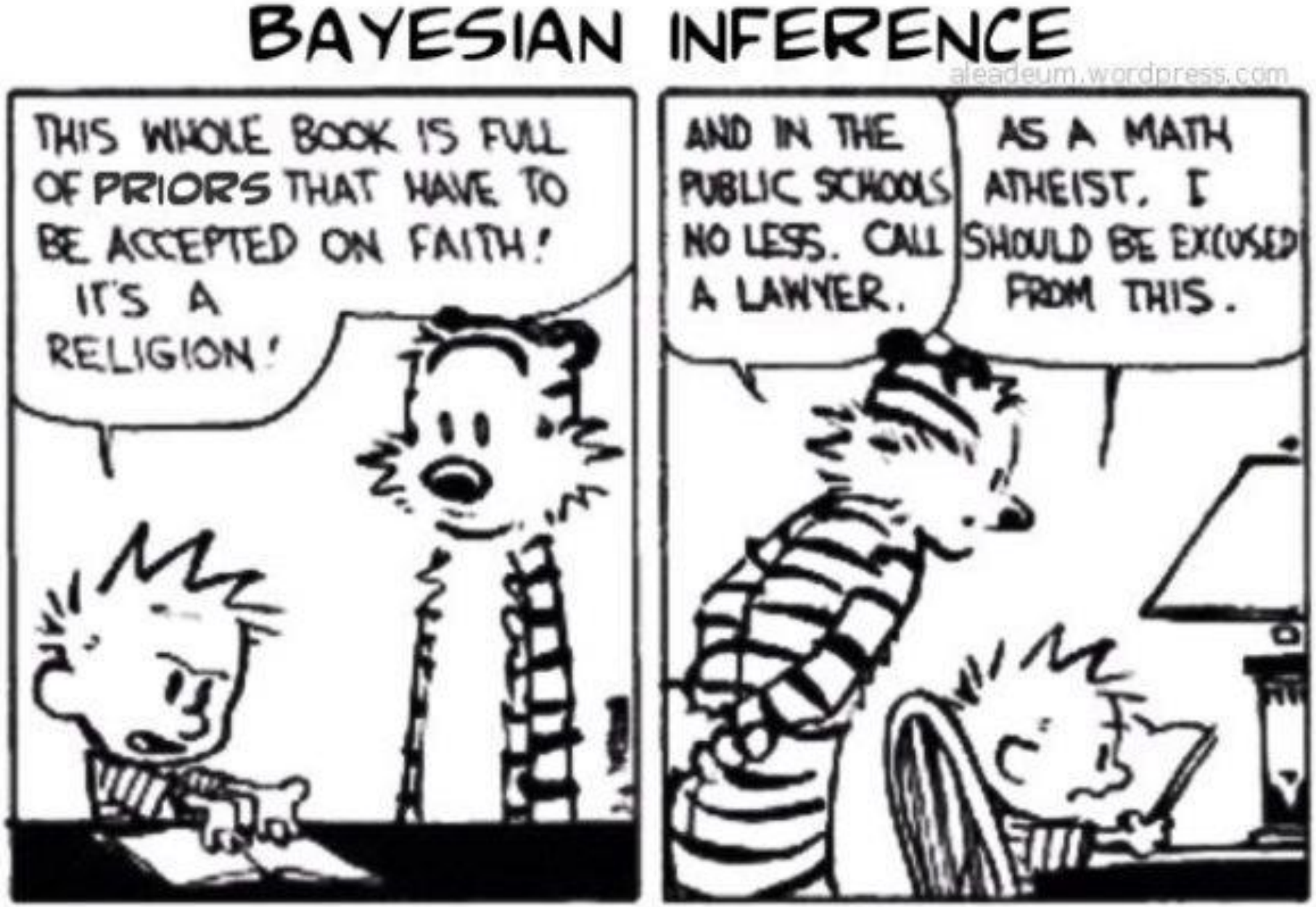
- Bayes theorem
  - X = **continuous** random variable (e.g., modeling some unknown parameter)
  - Y = continuous/discrete random variable (e.g., modeling observed data)
  - Consider joint distribution: P (X,Y)
  - Likelihood : P (Y=y | X=x)
    - Conditional
  - Evidence : $P (Y=y) = \int_x P (X=x, Y=y)$ **dx**
    - Marginal. Accounts for various ways in which data could have been generated.
  - Prior : P (X)
    - Before data is observed. Models beliefs (e.g., evolved over past experiences / learning)
  - Posterior : P (X=x | Y=y) **dx**
    - After data is observed. Data updates the beliefs.
  - Posterior P (X=x | Y=y) **dx**
    = Likelihood P (Y=y | X=x) * Prior P (X=x) **dx**
    / P (Y=y)

Bayesian Statistical Analysis

- Instead of making inferences about X based on **likelihood** function, Bayesian analysis makes inferences about X based on the **posterior**
- When can/should/do prior models make a difference ?
  - When the data sample size is **finite**



Bayesian Statistical Analysis



Bayesian Estimation

- Example: Coin Flip

Consider that we don't know if a coin is fair / unfair

We have 2 possibilities in our mind:
(1) Coin fair, i.e., $P(\text{head}) = p = 0.5$
(2) Coin biased towards heads with $P(\text{head}) = q = 0.7$

We have a belief (**prior** to observing data) that $P(\text{CoinFair}) = 0.8$

Now we experiment with the coin, collect data, and
recompute the probability that the coin is fair

$$P(\text{CoinFair}|\text{Data}) = P(\text{Data}|\text{CoinFair})P(\text{CoinFair})/P(\text{Data})$$

Bayesian Estimation

- Example: Coin Flip

Given: We have data as n observations with r heads and $(n-r)$ tails.
What does the data do to our belief ?

$$P(\text{Data}|\text{CoinFair}) = C_r^n 0.5^r 0.5^{n-r}$$

$$P(\text{Data}|\text{CoinUnfair}) = C_r^n 0.7^r 0.3^{n-r}$$

$$P(\text{Data}) = P(\text{Data}|\text{CoinFair})P(\text{CoinFair}) + P(\text{Data}|\text{CoinUnfair})P(\text{CoinUnfair})$$

$$P(\text{CoinFair}|\text{Data}) = \frac{0.5^r 0.5^{n-r} \times 0.8}{0.5^r 0.5^{n-r} \times 0.8 + 0.7^r 0.3^{n-r} \times 0.2}$$

Bayesian Estimation

- Example: Coin Flip

We have 2 possibilities in our mind:
(1) Coin fair, i.e., $P(\text{head}) = p = 0.5$
(2) Coin biased towards heads with $P(\text{head}) = q = 0.7$

Case 1: We have a belief (**prior** to observing data) that $P(\text{CoinFair}) = 0.8$
If $n = 20, r = 11$,
then $P(\text{CoinFair}|\text{Data}) = 0.9074$ that is more than 0.8.
So the data has strengthened our belief !

Why has this happened ? Because 11 heads out of 20 is more like the fair coin.

Case 2:
If $n = 20, r = 13$,
then $P(\text{CoinFair}|\text{Data}) = 0.6429$ that is less than 0.8.
So the data has weakened our belief !

Why has this happened ? Because 13 heads out of 20 is more like the unfair coin.

Case 3:
If $n = 20, r = 12$,
then $P(\text{CoinFair}|\text{Data}) = 0.8077$ that is close to 0.8.

Bayesian Estimation

- Example: Gaussian (**Unknown mean**, Known variance)

Given: Data $\{x_i\}_{i=1}^N$ drawn from a Gaussian PDF with known variance σ^2, but unknown mean μ

Bayesian strategy: Model mean as a random variable M along with an associated PDF $P(M)$

Bayesian strategy: Prior belief on M is that it is drawn from a Gaussian with mean μ_0 and variance σ_0^2

Associated Generative Model here: first draw μ from prior $P(M)$, then draw data from $P(X|M)$ given $M = \mu$

Goal: Estimate μ, given:
(i) prior PDF $P(M)$,
(ii) likelihood function using the model $P(X|M)$, and
(iii) observed data $\{x_i\}_{i=1}^N$

[What if we ignore the likelihood / data ? ($\mu = \mu_0$)]

[What if we ignore the prior ? (ML estimation seen before)]

Bayesian Estimation

- Example: Gaussian (Unknown mean, Known variance)

Given: Data $\{x_i\}_{i=1}^N$ drawn from a Gaussian PDF with known variance σ^2, but unknown mean μ

Goal: Estimate μ, given:
(i) prior PDF $P(M)$,
(ii) likelihood function using the model $P(X|M)$, and
(iii) observed data $\{x_i\}_{i=1}^N$

Assume sample mean = $\overline{x}$

- Prior: $P(M=\mu) = G(\mu; \text{mean} = \mu_0, \text{variance} = \sigma_0^2)$
- Likelihood: $P(\text{data}|M=\mu) = \Pi_i\, G(x_i ; \mu, \sigma^2) = \Pi_i\, G(\mu ; x_i, \sigma^2)$
  - Negative exponent here can be written as:
    $0.5\,[\, N\,\mu^2 - 2\,(\Sigma_i\, x_i)\;\; \mu\,]\,/\;\sigma^2\;\;$ + terms independent of μ
  - Negative exponent here can be written as:
    $= 0.5\,[\;\; \mu^2 - 2\,(\Sigma_i\, x_i/N)\,\mu\,]\,/\,(\sigma^2/N)$ + terms independent of μ
  - Thus, likelihood is proportional to $G(\mu ; \text{mean} = (\Sigma_i\, x_i/N), \text{variance} = \sigma^2/N)$
    - Gaussian mean = sample mean = $(\Sigma_i\, x_i\,/\,N)$
    - Gaussian variance = σ^2/N

Bayesian Estimation

- Example: Gaussian (Unknown mean, Known variance)

Given: Data $\{x_i\}_{i=1}^N$ drawn from a Gaussian PDF with known variance σ^2, but unknown mean μ

Goal: Estimate μ, given:
(i) prior PDF $P(M)$,
(ii) likelihood function using the model $P(X|M)$, and
(iii) observed data $\{x_i\}_{i=1}^N$

Assume sample mean = $\overline{x}$

Then MAP estimate for the mean $\widehat{\mu}$ is:

$$\widehat{\mu} = \frac{\overline{x}\sigma_0^2 + \mu_0\sigma^2/N}{\sigma_0^2 + \sigma^2/N}$$

What if $N = 1$?

What if $N \to \infty$? (data dominates the prior)

What if $\sigma_0 \to \infty$? (very weak prior: ignore the prior)

What if $\sigma_0 \to 0$? (very strong prior: ignore the data)

Bayesian Estimation

- Example: Gaussian (Unknown mean, Known variance)
  - Instead of finding posterior mode, we could have done something different
  - "Posterior mean" estimate to minimize mean/expected squared error

Given data: $\{x_i\}_{i=1}^n$ drawn from $P(X|\theta^*)$

We have a prior PDF $P(\Theta)$ on RV Θ

Posterior PDF := conditional density :=

$$P(\Theta|x_1, \cdots, x_n) = \frac{P(x_1, \cdots, x_n|\Theta)P(\Theta)}{\int_\theta P(x_1, \cdots, x_n, \theta)d\theta}$$

Question: Given a PDF $P(\Theta|x_1, \cdots, x_n)$ on the variable Θ,
what is the best estimate $\widehat{\theta}^*$ to
minimize mean squared error $E_{P(\Theta|x_1,\cdots,x_n)}[(\widehat{\theta} - \Theta)^2]$?

Answer: The PDF mean $E_{P(\Theta|x_1,\cdots,x_n)}[\Theta]$. This is also a Bayes estimate.

Bayesian Estimation

- Question (Bayes interval estimate):
  - Suppose signal (scalar) of value “s” is sent from location A to B.
  - Because of the noisy communication channel,
    signal received at B has a Gaussian PDF with mean “s” and variance 60.
  - Value received at B is 40.
  - Find interval (a,b) s.t. the probability of signal “s” being within (a,b) is 0.9
  - A priori, it’s known that “s” was drawn from Gaussian (mean 50, variance 100)

Bayesian Estimation

- Question (Bayes interval estimate):
  - Product of two Gaussians: $G(z;\mu_1,\sigma_1^2)G(z;\mu_2,\sigma_2^2) \propto G(z;\mu_3,\sigma_3^2)$

Numerator exponent $= \frac{(z-\mu_1)^2}{2\sigma_1^2} + \frac{(z-\mu_2)^2}{2\sigma_2^2}$

$$= \frac{1}{2\sigma_1^2\sigma_2^2}\left(z^2(\sigma_2^2+\sigma_1^2) - (2\mu_1\sigma_2^2 + 2\mu_2\sigma_1^2)z + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_2^2\right)$$

$$= \frac{1}{2\sigma_1^2\sigma_2^2}\left(z^2(\sigma_2^2+\sigma_1^2) - (2\mu_1\sigma_2^2 + 2\mu_2\sigma_1^2)z\right) + c, \quad \text{where } c = \text{constant independent of } z$$

$$= \frac{\sigma_2^2+\sigma_1^2}{2\sigma_1^2\sigma_2^2}\left(z^2 - \frac{2\mu_1\sigma_2^2 + 2\mu_2\sigma_1^2}{\sigma_2^2+\sigma_1^2}z\right) + c$$

$$= \frac{\sigma_2^2+\sigma_1^2}{2\sigma_1^2\sigma_2^2}\left(z^2 - 2\mu_3 z + \mu_3^2\right) + c' = \frac{1}{2\sigma_3^2}(z-\mu_3)^2 + c'$$

$$\boxed{\mu_3 = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}, \sigma_3^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}$$

where $c' =$ constant independent of z and

Bayesian Estimation

- Question (Bayes interval estimate):
  - Suppose signal (scalar) of value "s" is sent from location A to B.
  - Because of the noisy communication channel, signal received at B has a Gaussian PDF with mean "s" and variance 60.
  - Value received at B is 40.
  - Find interval (a,b) s.t. the probability of signal "s" being within (a,b) is 0.9
  - A priori, it's known that "s" was drawn from Gaussian (mean 50, variance 100)

Using formulas derived before for the posterior $P(s|x_1 = 40)$ of parameter s given data x_1,

Posterior mean = $\frac{50*60+40*100}{60+100} = 43.75$

Posterior variance = $\frac{60*100}{60+100} = 37.5$

$$\mu_3 = \frac{\mu_1\sigma_2^2 + \mu_2\sigma_1^2}{\sigma_1^2 + \sigma_2^2}, \sigma_3^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

We know that the posterior PDF is Gaussian

Thus, $Z := \frac{S-43.75}{\sqrt{37.5}}$ has a standard normal PDF

For a standard normal Z, we know that probability mass within $Z \in (-1.645, +1.645)$ is 0.9

Bayesian Estimation

- Question (Bayes interval estimate):
  - Suppose signal (scalar) of value “s” is sent from location A to B.
  - Because of the noisy communication channel,
    signal received at B has a Gaussian PDF with mean “s” and variance 60.
  - Value received at B is 40.
  - Find interval (a,b) s.t. the probability of signal “s” being within (a,b) is 0.9
  - A priori, it’s known that “s” was drawn from Gaussian (mean 50, variance 100)

Thus, we want to find S s.t. $P(-1.645 < Z < 1.645|\text{data}) = 0.9$
i.e., $P(-1.645 < \frac{S-43.75}{\sqrt{37.5}} < 1.645|\text{data}) = 0.9$
i.e., $P(33.68 < S < 53.83|\text{data}) = 0.9$

Thus, the desired interval is $(a = 33.68, b = 53.83)$

Bayesian Estimation

- **Loss** functions and **Risk** functions

If the true value was θ, and our estimator produced an estimate as $\widehat{\theta}$,
lets say we incur a "loss" based on the "loss function" $L(\widehat{\theta}|\theta)$

We know that, given the data, the true value θ is distributed as per the posterior PDF $P(\Theta|x_1, \cdots, x_n)$

We define a "risk function" $R(\widehat{\theta})$:= expected loss
:= expectation of loss function $L(\widehat{\theta}|\Theta)$ under posterior PDF $P(\Theta|x_1, \cdots, x_n)$

Goal: Choose optimal estimate as $\widehat{\theta}^*$ that minimizes the risk function $R(\widehat{\theta})$

Bayesian Estimation

- Loss functions and Risk functions

Example: Squared-error loss function: $L(\widehat{\theta}|\theta) = (\widehat{\theta} - \theta)^2$

Risk function $R(\widehat{\theta}) = E_{P(\Theta|x_1,\cdots,x_n)}[(\widehat{\theta} - \Theta)^2]$ = mean squared error

Let risk-function minimizer be $\widehat{\theta}^*$

Then,

$$\left(\frac{\partial}{\partial\widehat{\theta}} E_{P(\Theta|x_1,\cdots,x_n)}[(\widehat{\theta} - \Theta)^2]\right)\Big|_{\widehat{\theta}=\widehat{\theta}^*} = 0$$

Thus, $\widehat{\theta}^* = E_{P(\Theta|x_1,\cdots,x_n)}[\Theta]$ = Posterior mean

Bayesian Estimation

- Loss functions and Risk functions

Example: Zero-one loss function (case of discrete RV Θ): $L(\widehat{\theta}|\theta) = I(\widehat{\theta} \neq \theta)$

Risk function $R(\widehat{\theta}) = E_{P(\Theta|x_1,\cdots,x_n)}[I(\widehat{\theta} \neq \Theta)]$

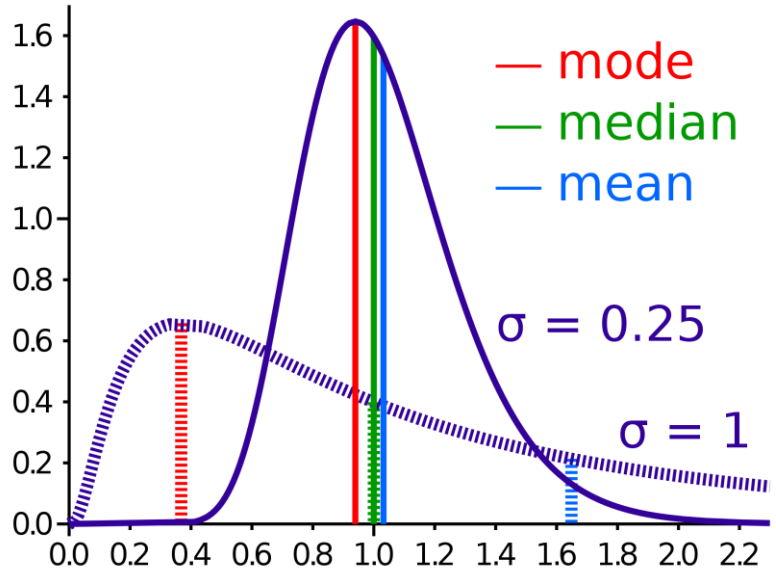
$= \sum_{\theta \neq \widehat{\theta}} P(\theta|x_1, \cdots, x_n)$

$= 1 - P(\theta = \widehat{\theta}|x_1, \cdots, x_n)$

Thus, the risk function is minimized at $\widehat{\theta}^* := \arg\max_\theta P(\theta|x_1, \cdots, x_n)$ = MAP estimate

Bayesian Estimation

- Loss functions and Risk functions

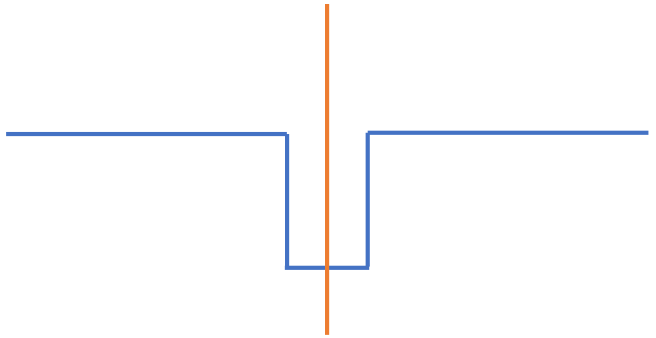


Example: Zero-one loss function (case of continuous RV Θ)

Assume that the loss function is an *inverted* rectangular pulse with height 1 and an infinitesimally small width $\epsilon > 0$ (we do NOT set $\epsilon = 0$), with center of the pulse at the true value θ. i.e.,

$L(\widehat{\theta}|\theta) = 0$; if $\widehat{\theta} \in (\theta - \epsilon/2, \theta + \epsilon/2)$

$L(\widehat{\theta}|\theta) = 1$; otherwise



The risk function

$$1 - \int_{\widehat{\theta}-\epsilon/2}^{\widehat{\theta}+\epsilon/2} P(\theta|x_1, \cdots, x_n) d\theta$$

Now take the limit

$$\lim_{\epsilon \to 0} \arg\max_{\widehat{\theta}} \int_{\widehat{\theta}-\epsilon/2}^{\widehat{\theta}+\epsilon/2} P(\theta|x_1, \cdots, x_n) d\theta$$

The optimum $\widehat{\theta}^*$ is when the pulse center is placed at the mode of the PDF, i.e.,

$$\widehat{\theta}^* = \arg\max_{\widehat{\theta}} P(\theta|x_1, \cdots, x_n)$$

Bayesian Estimation

- Loss functions and Risk functions

Example: Absolute-error loss function $L(\widehat{\theta}|\theta) = |\widehat{\theta} - \theta|$

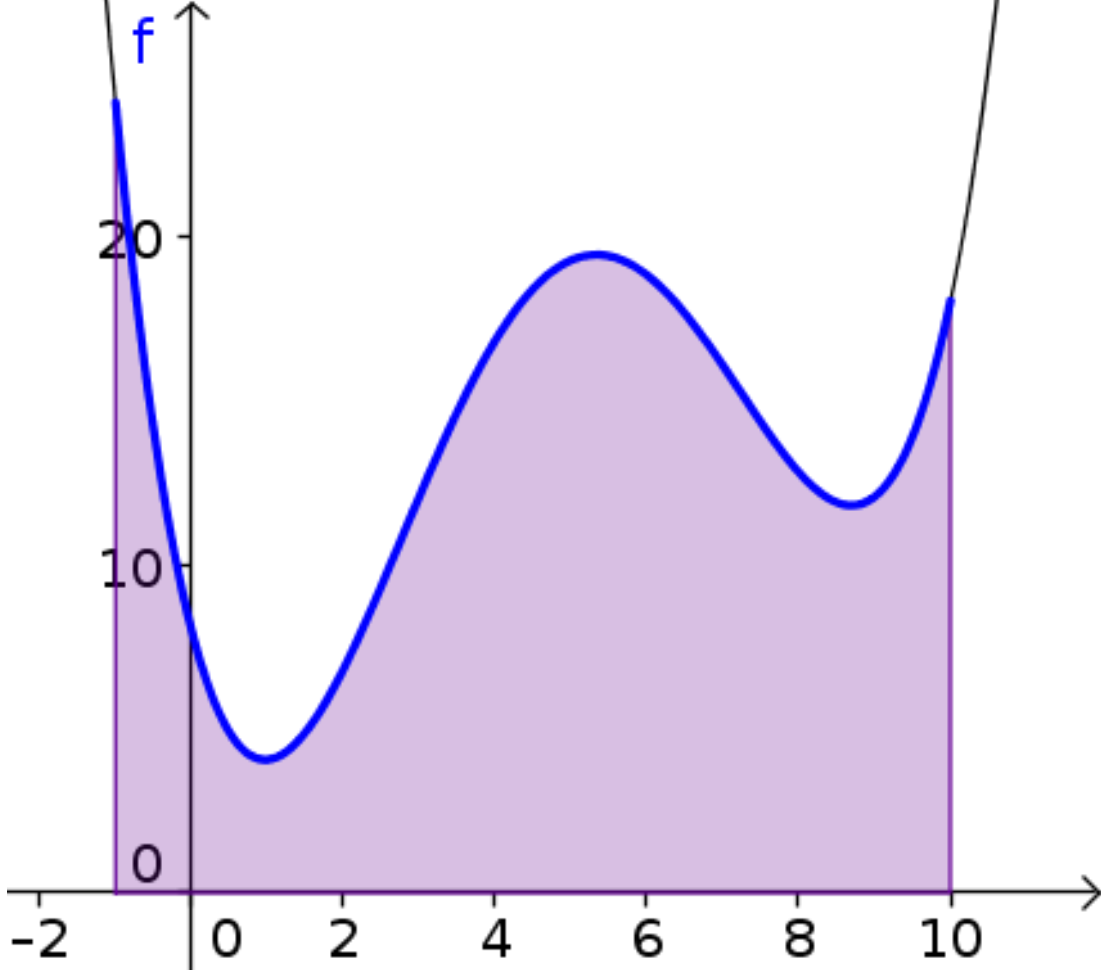
Let $x = \{x_i\}_{i=1}^{N}$

Risk function $= E_{P(\Theta|x)}[|\widehat{\theta} - \Theta|]$

$= \int_{-\infty}^{\infty} |\widehat{\theta} - \theta| P(\theta|x) d\theta$

$= \int_{-\infty}^{\widehat{\theta}} (\widehat{\theta} - \theta) P(\theta|x) d\theta + \int_{\widehat{\theta}}^{\infty} (\theta - \widehat{\theta}) P(\theta|x) d\theta$

The risk function is minimized when its derivative is zero



How to take the derivative of an integral where the limits of the integral are also a function of the variable of interest ?

Leibniz's Integral Rule:

$$\frac{\partial}{\partial a} \int_{l(a)}^{u(a)} f(z,a) dz = \int_{l(a)}^{u(a)} \frac{\partial f}{\partial a} dz + f(z = u(a), a) \frac{\partial u}{\partial a} - f(z = l(a), a) \frac{\partial l}{\partial a}$$

Bayesian Estimation

- Loss functions and Risk functions

In our case, $f(z \equiv \theta, a \equiv \widehat{\theta}) \propto (\widehat{\theta} - \theta) P(\theta|x)$

Leibniz's Integral Rule:

$$\frac{\partial}{\partial a} \int_{l(a)}^{u(a)} f(z,a)dz = \int_{l(a)}^{u(a)} \frac{\partial f}{\partial a} dz + f(z = u(a), a)\frac{\partial u}{\partial a} - f(z = l(a), a)\frac{\partial l}{\partial a}$$

Example: Absolute-error loss function $L(\widehat{\theta}|\theta) = |\widehat{\theta} - \theta|$

Let $x = \{x_i\}_{i=1}^{N}$

Risk function $= E_{P(\Theta|x)}[|\widehat{\theta} - \Theta|]$

$= \int_{-\infty}^{\infty} |\widehat{\theta} - \theta| P(\theta|x) d\theta$

$= \int_{-\infty}^{\widehat{\theta}} (\widehat{\theta} - \theta) P(\theta|x) d\theta + \int_{\widehat{\theta}}^{\infty} (\theta - \widehat{\theta}) P(\theta|x) d\theta$

The risk function is minimized when its derivative is zero

In our case, for the 1st integral: $f(z = u(a), a) = 0$ and the lower-limit term doesn't arise

In our case, for the 2nd integral: $f(z = l(a), a) = 0$ and the upper-limit term doesn't arise

Bayesian Estimation

- Loss functions and Risk functions

Example: Absolute-error loss function $L(\widehat{\theta}|\theta) = |\widehat{\theta} - \theta|$

Let $x = \{x_i\}_{i=1}^{N}$

Risk function $= E_{P(\Theta|x)}[|\widehat{\theta} - \Theta|]$

$= \int_{-\infty}^{\infty} |\widehat{\theta} - \theta| P(\theta|x) d\theta$

$= \int_{-\infty}^{\widehat{\theta}} (\widehat{\theta} - \theta) P(\theta|x) d\theta + \int_{\widehat{\theta}}^{\infty} (\theta - \widehat{\theta}) P(\theta|x) d\theta$

The risk function is minimized when its derivative is zero

In our case, $f(z \equiv \theta, a \equiv \widehat{\theta}) \propto (\widehat{\theta} - \theta) P(\theta|x)$

Leibniz's Integral Rule:

$$\frac{\partial}{\partial a} \int_{l(a)}^{u(a)} f(z,a) dz = \int_{l(a)}^{u(a)} \frac{\partial f}{\partial a} dz + f(z = u(a), a) \frac{\partial u}{\partial a} - f(z = l(a), a) \frac{\partial l}{\partial a}$$

Thus, the derivative of our risk function w.r.t. $\widehat{\theta}$ is:

$$= \int_{-\infty}^{\widehat{\theta}} (+1) P(\theta|x) d\theta + \int_{\widehat{\theta}}^{\infty} (-1) P(\theta|x) d\theta$$

$$= \int_{-\infty}^{\widehat{\theta}} P(\theta|x) d\theta - \int_{\widehat{\theta}}^{\infty} P(\theta|x) d\theta$$

This is zero when $\widehat{\theta}$ = median of $P(\theta|x)$

Bayesian Estimation

- Loss functions and Risk functions

Thus, the derivative of our risk function w.r.t. $\widehat{\theta}$ is:

$$\int_{-\infty}^{\widehat{\theta}} P(\theta|x)d\theta - \int_{\widehat{\theta}}^{\infty} P(\theta|x)d\theta$$

This is zero when $\widehat{\theta}$ = median of $P(\theta|x)$

Example: Absolute-error loss function $L(\widehat{\theta}|\theta) = |\widehat{\theta} - \theta|$

Let $x = \{x_i\}_{i=1}^{N}$

Risk function $= E_{P(\Theta|x)}[|\widehat{\theta} - \Theta|]$

$= \int_{-\infty}^{\infty} |\widehat{\theta} - \theta| P(\theta|x)d\theta$

$= \int_{-\infty}^{\widehat{\theta}} (\widehat{\theta} - \theta) P(\theta|x)d\theta + \int_{\widehat{\theta}}^{\infty} (\theta - \widehat{\theta}) P(\theta|x)d\theta$

The risk function is minimized when its derivative is zero

The median will be a minimizer if the 2nd derivative is positive. Is that so ?

In this case, for both integrals, $\frac{\partial f}{\partial a} = 0$
In this case, for 1st integral, the lower-limit term doesn't arise
In this case, for 2nd integral, the upper-limit term doesn't arise
Thus, the 2nd derivative of our risk function w.r.t. $\widehat{\theta}$, evaluated at $\widehat{\theta}$ = median of $P(\theta|x)$, is:
$= P(\widehat{\theta}|x) + P(\widehat{\theta}|x) \geq 0$

Note: for the median $\widehat{\theta}$ to be unique, the CDF must be continuous and strictly increasing at $\widehat{\theta}$, which means that the PDF must be positive at $\widehat{\theta}$.

Bayesian Estimation

- Example: i.i.d. Bernoulli

Given: $X_1, \cdots, X_n$ are i.i.d. Bernoulli with parameter θ and PDF $P(x = 1|\theta) = \theta, P(x = 0|\theta) = 1 - \theta$

Data: $x_1, \cdots, x_n$

Estimate $\theta \in (0, 1)$

Prior $P(\theta) = 1, \forall \theta \in (0, 1)$

Rewrite PDF as $P(x|\theta) = \theta^x(1-\theta)^{1-x}$, where $x \in \{0, 1\}$

$$P(\theta|x_1, \cdots, x_n) = P(x_1, \cdots, x_n|\theta)/P(x_1, \cdots, x_n)$$

where Numerator $= \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}$

If we want the posterior mean, then we need to care about the denominator as well

Denominator

$$= \int_0^1 \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i} d\theta$$

To handle the integral in the denominator, we use the result / trick:

$$\int_0^1 \theta^m (1-\theta)^r d\theta = m!r!/(m+r+1)!$$

Let $x = \sum_i x_i$

Then, $P(\theta|x_1, \cdots, x_n) = \frac{(n+1)!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}$

Bayesian Estimation

- Example: i.i.d. Bernoulli

ML estimator $\equiv$ MAP estimator; because $P(\theta) = 1$

Note: ML estimator $= \max_\theta \log\left(\theta^{\sum_i X_i}(1-\theta)^{n-\sum_i X_i}\right)$
$= \max_\theta X \log\theta + (n-X)\log(1-\theta)$, where $X := \sum_i X_i$
$= X/n$
$= \sum_i X_i/n$

Thus, $E_{P(\theta|x_1,\cdots,x_n)}[\theta] = \int_0^1 \theta \frac{(n+1)!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}d\theta = \frac{x+1}{n+2}$

Thus, Bayes posterior-mean estimator $= \frac{\sum_i X_i+1}{n+2}$

Check that the 2nd derivative is negative

(Use the facts: $X \geq 0$ and $n - X \geq 0$ and $0 < \theta < 1$)

Given: $X_1, \cdots, X_n$ are i.i.d. Bernoulli with parameter θ and PDF $P(x = 1|\theta) = \theta, P(x = 0|\theta) = 1 - \theta$

Data: $x_1, \cdots, x_n$

Estimate $\theta \in (0, 1)$

Prior $P(\theta) = 1, \forall\theta \in (0, 1)$

Rewrite PDF as $P(x|\theta) = \theta^x(1-\theta)^{1-x}$, where $x \in \{0,$

$P(\theta|x_1, \cdots, x_n) = P(x_1, \cdots, x_n|\theta)/P(x_1, \cdots, x_n)$

where Numerator $= \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}$

$$\int_0^1 \theta^m(1-\theta)^r d\theta = m!r!/(m+r+1)!$$

Let $x = \sum_i x_i$

Then, $P(\theta|x_1, \cdots, x_n) = \frac{(n+1)!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}$

Bayesian Estimation

- Example: i.i.d. Bernoulli

Thus, $E_{P(\theta|x_1,\cdots,x_n)}[\theta] = \int_0^1 \theta \frac{(n+1)!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}d\theta = \frac{x+1}{n+2}$

Thus, Bayes posterior-mean estimator $= \frac{\sum_i X_i+1}{n+2}$

Note: ML estimator $= \max_\theta \log\left(\theta^{\sum_i X_i}(1-\theta)^{n-\sum_i X_i}\right)$
$= \max_\theta X\log\theta + (n-X)\log(1-\theta)$, where $X := \sum_i X_i$
$= X/n$
$= \sum_i X_i/n$

Given: $X_1, \cdots, X_n$ are i.i.d. Bernoulli with parameter θ and PDF $P(x=1|\theta) = \theta, P(x=0|\theta) = 1-\theta$

Data: $x_1, \cdots, x_n$

Estimate $\theta \in (0,1)$

Prior $P(\theta) = 1, \forall\theta \in (0,1)$

Note: When $n = 0$, Bayes estimate $= 0.5$, the mid-point of the interval $(0,1)$. This is what we get when we solely rely on the prior

Note: Asymptotically, i.e., as $n \to \infty$, the Bayes (mean) estimator tends to the ML estimator

What happens to the Bayes estimate and ML estimate when true $\theta = 0$ or true $\theta = 1$? Assume n is large.

Fisher Information

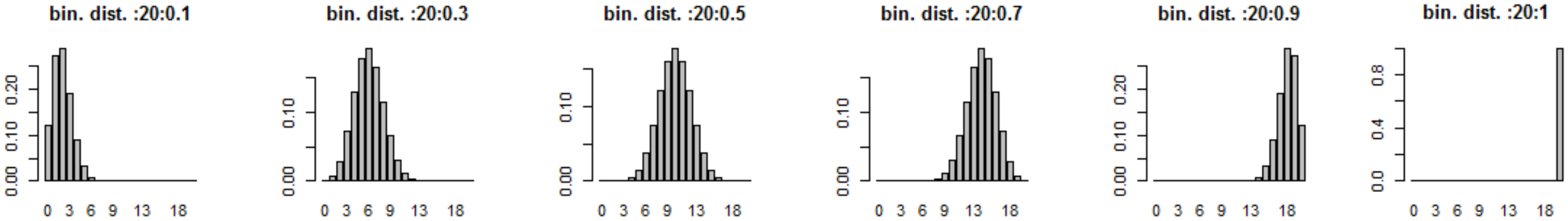
- Key questions:
  - How much information can a sample of data provide about the unknown parameter ?
  - Comparing two scenarios of parameter estimation, when is parameter estimation easier ?

Fisher Information

- Observations:

(1) If likelihood function $P(\text{data}|\theta)$ changes (more) sharply w.r.t. Δ changes in θ around $\theta = \theta_{\text{true}}$,
then it is easier to estimate θ_{true} from the given data sample of size N.

  - Bernoulli distribution with success probability (p) close to 0 or 1
    - Variance = p(1-p)
  - Binomial distribution (n tries) with success probability (p) close to 0 or 1
    - Variance = np(1-p)
    - Bottom: n=20; p = 0.1, 0.3, 0.5, 0.7, 0.9, 1

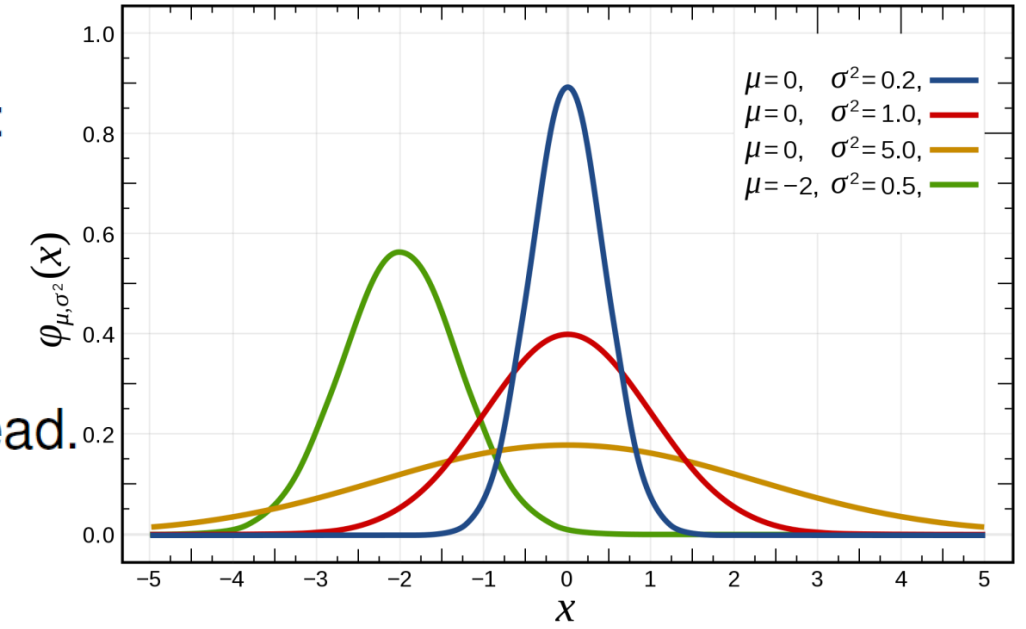


Fisher Information

- Observations:

(1) If likelihood function $P(\text{data}|\theta)$ changes (more) sharply w.r.t. Δ changes in θ around $\theta = \theta_{\text{true}}$,
then it is easier to estimate θ_{true} from the given data sample of size N.

Example 2: Estimating Gaussian mean $\theta := \mu$ in two cases:
(i) when variance σ^2 (known) is huge, and
(ii) when σ^2 (known) is tiny.



Data drawn from $G(x; \mu, \sigma^2)$ in 2nd case has a smaller spread.

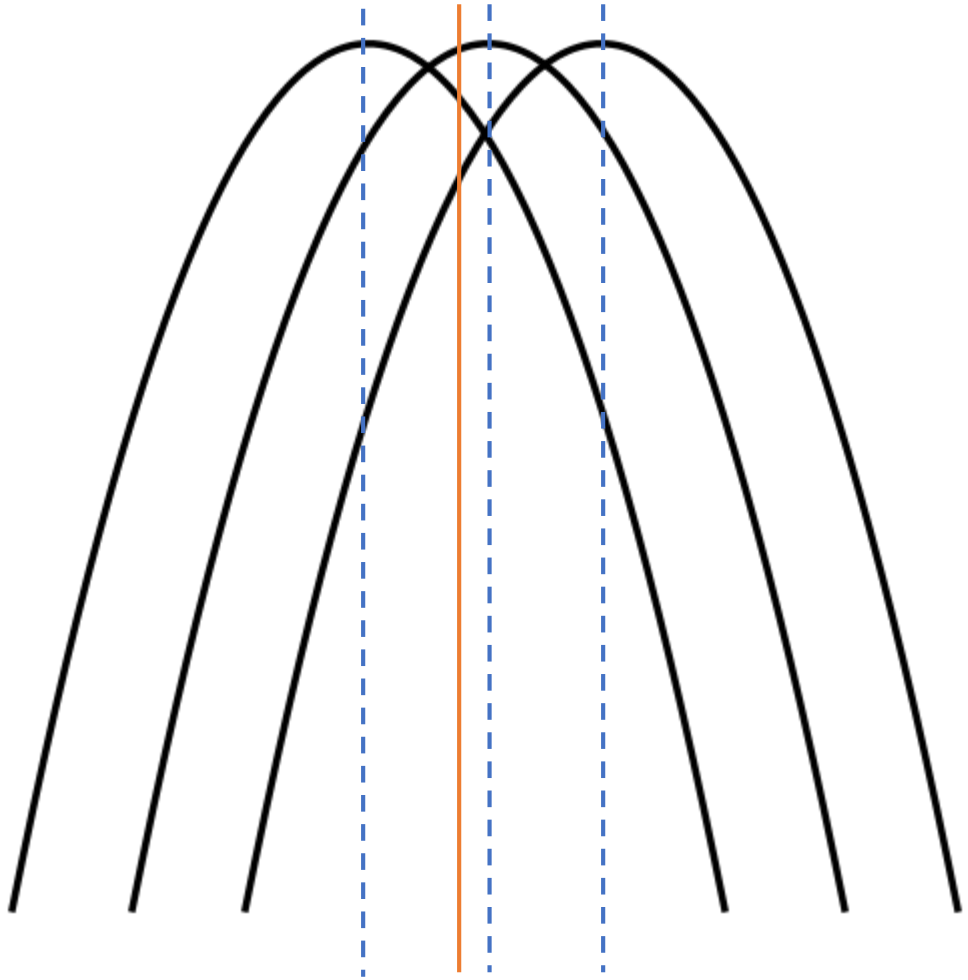
Likelihood function in 2nd case more peaked.

For a small sample of size N (say, N = 5),
mean estimate (sample mean; always unbiased = always high accuracy)
is much more precise (= much lower variance) in 2nd case

Fisher Information

- Observations:

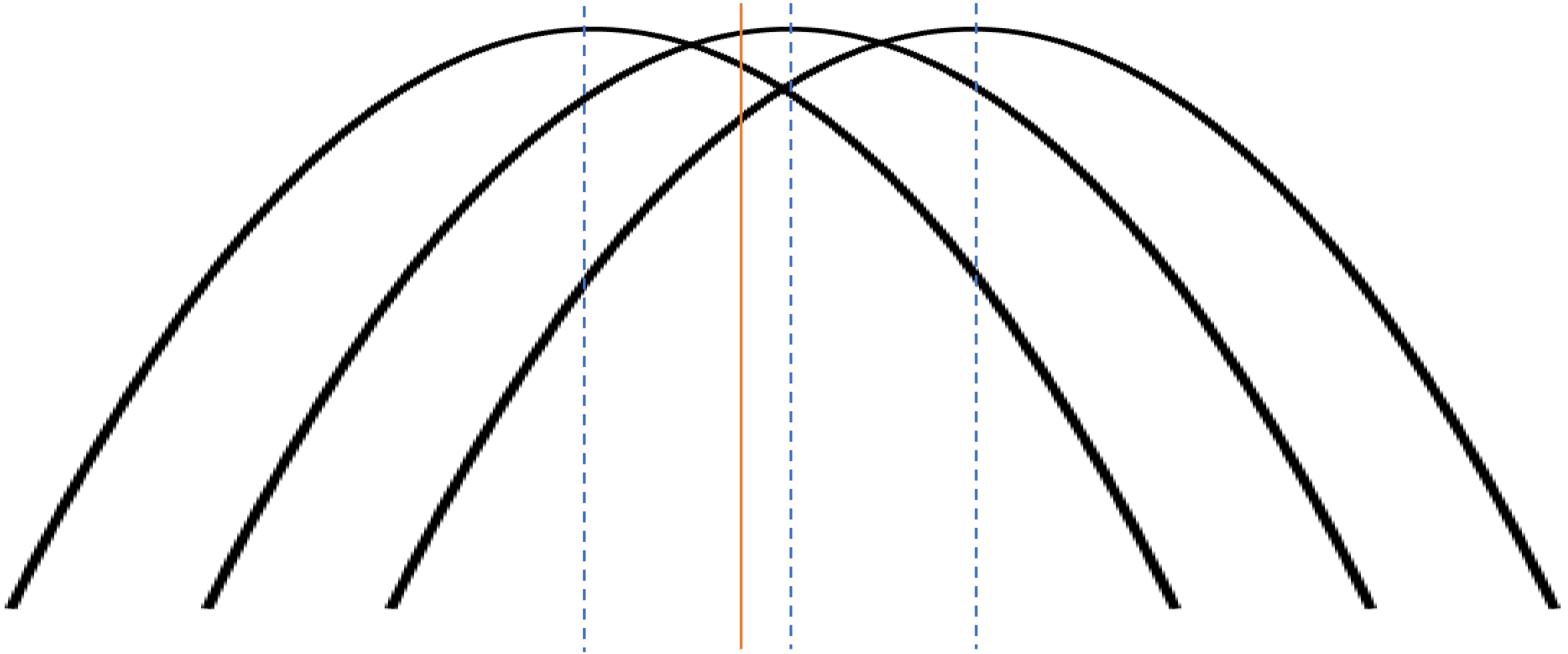
(2) If likelihood function $P(\text{data}|\theta)$ has a large spread w.r.t. changes in θ around θ_{true}, it will take very many N-sized data samples to get the ML estimate of θ to be at / close to θ_{true}



Fisher Information

- Observations:

(2) If likelihood function $P(\text{data}|\theta)$ has a large spread w.r.t. changes in θ around θ_{true}, it will take very many N-sized data samples to get the ML estimate of θ to be at / close to θ_{true}

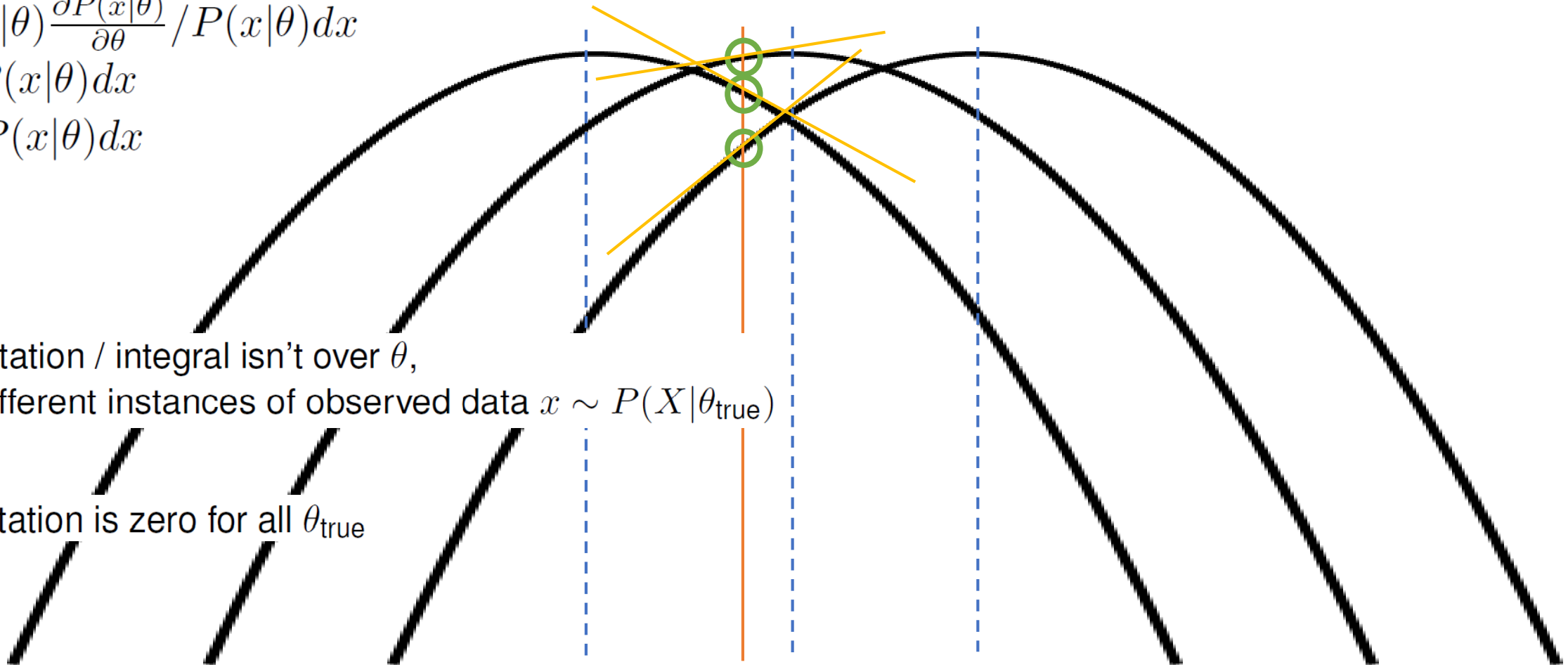


Fisher Information

- Towards formulating Fisher information

First, consider the average (expected) derivative of the log-likelihood function:

$$E_{P(X|\theta_{\text{true}})}\left[\frac{\partial}{\partial\theta}\log P(X|\theta)\Big|_{\theta=\theta_{\text{true}}}\right]$$
$$= \int_x P(x|\theta)\frac{\partial P(x|\theta)}{\partial\theta}/P(x|\theta)dx$$
$$= \int_x \frac{\partial}{\partial\theta}P(x|\theta)dx$$
$$= \frac{\partial}{\partial\theta}\int_x P(x|\theta)dx$$
$$= \frac{\partial}{\partial\theta}1$$
$$= 0$$



The expectation / integral isn't over θ,
but over different instances of observed data $x \sim P(X|\theta_{\text{true}})$

The expectation is zero for all θ_{true}

Fisher Information

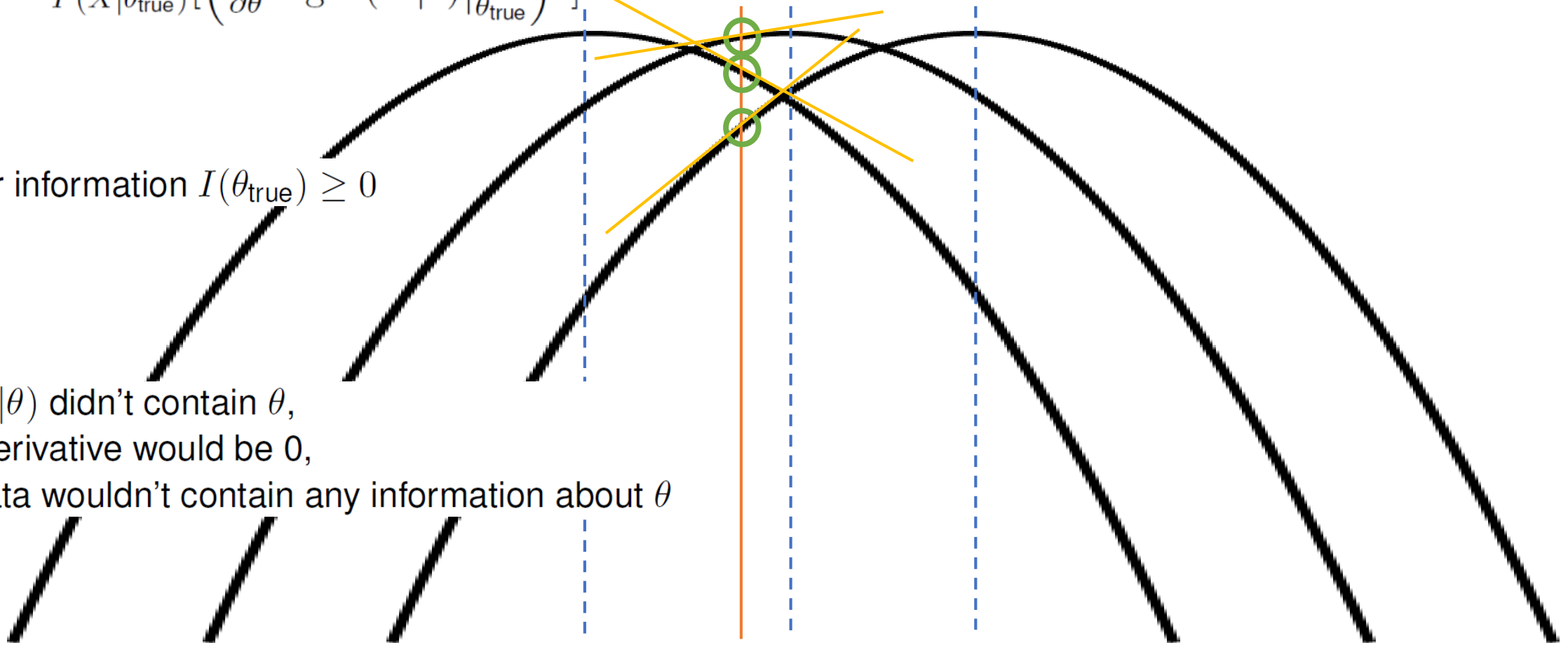
- Towards formulating Fisher information I(.)

Now, consider the expected squared slope (slope variance) of the log-likelihood function $\log P(X|\theta)$, evaluated at $\theta = \theta_{\text{true}}$, i.e.,

$$I(\theta_{\text{true}}) := E_{P(X|\theta_{\text{true}})}[\left(\frac{\partial}{\partial\theta}\log P(X|\theta)\Big|_{\theta_{\text{true}}}\right)^2]$$

The Fisher information $I(\theta_{\text{true}}) \geq 0$

If $\log P(X|\theta)$ didn't contain θ,
then the derivative would be 0,
and the data wouldn't contain any information about θ



Fisher Information

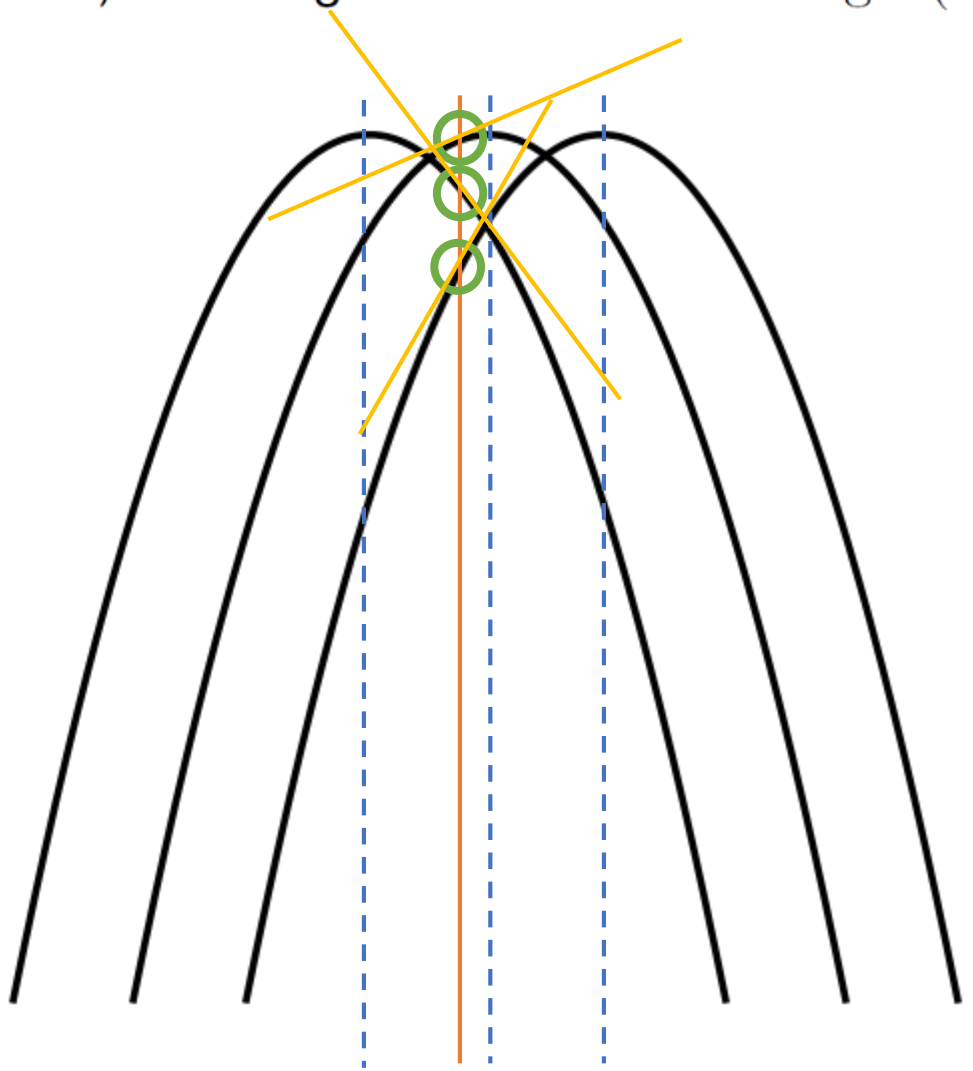
- Towards formulating Fisher information I(.)

Now, consider the expected squared slope (slope variance) of the log-likelihood function $\log P(X|\theta)$, evaluated at $\theta = \theta_{\text{true}}$, i.e.,

$$I(\theta_{\text{true}}) := E_{P(X|\theta_{\text{true}})}[\left(\frac{\partial}{\partial\theta}\log P(X|\theta)\Big|_{\theta_{\text{true}}}\right)^2]$$

The Fisher information $I(\theta_{\text{true}}) \geq 0$

If $\log P(X|\theta)$ didn't contain θ,
then the derivative would be 0,
and the data wouldn't contain any information about θ



Fisher Information

- Alternate (equivalent) formulation for Fisher information
  - Consider

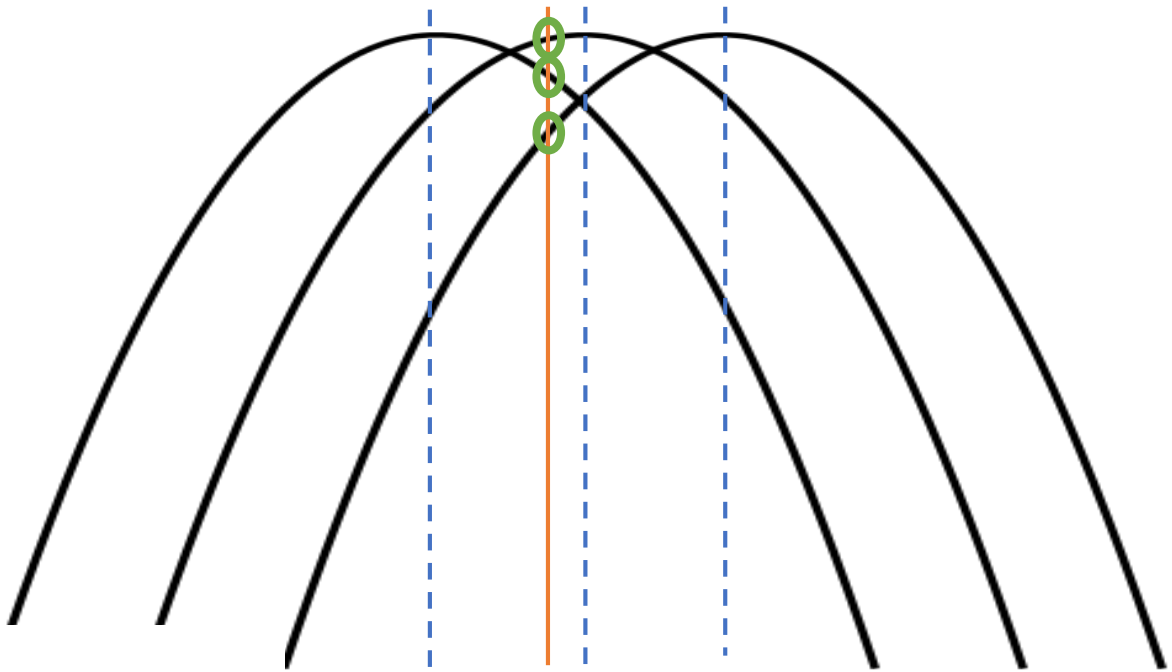
$$\frac{\partial^2}{\partial\theta^2}\log P(X|\theta) = \frac{\frac{\partial^2 P(X|\theta)}{\partial\theta^2}}{P(X|\theta)} - \left(\frac{\frac{\partial P(X|\theta)}{\partial\theta}}{P(X|\theta)}\right)^2 = \frac{\frac{\partial^2 P(X|\theta)}{\partial\theta^2}}{P(X|\theta)} - \left(\frac{\partial \log P(X|\theta)}{\partial\theta}\right)^2$$

Now, (i) evaluate LHS and RHS at $\theta := \theta_{\text{true}}$ and (ii) take expectation w.r.t. $P(X|\theta_{\text{true}})$:

$$E_{P(X|\theta_{\text{true}})}\left[\frac{\partial^2}{\partial\theta^2}\log P(X|\theta)\Big|_{\theta=\theta_{\text{true}}}\right]$$

$$= E_{P(X|\theta_{\text{true}})}\left[\frac{\frac{\partial^2 P(X|\theta)}{\partial\theta^2}}{P(X|\theta)}\Big|_{\theta=\theta_{\text{true}}}\right] - I(\theta_{\text{true}})$$

$= -I(\theta_{\text{true}})$, because



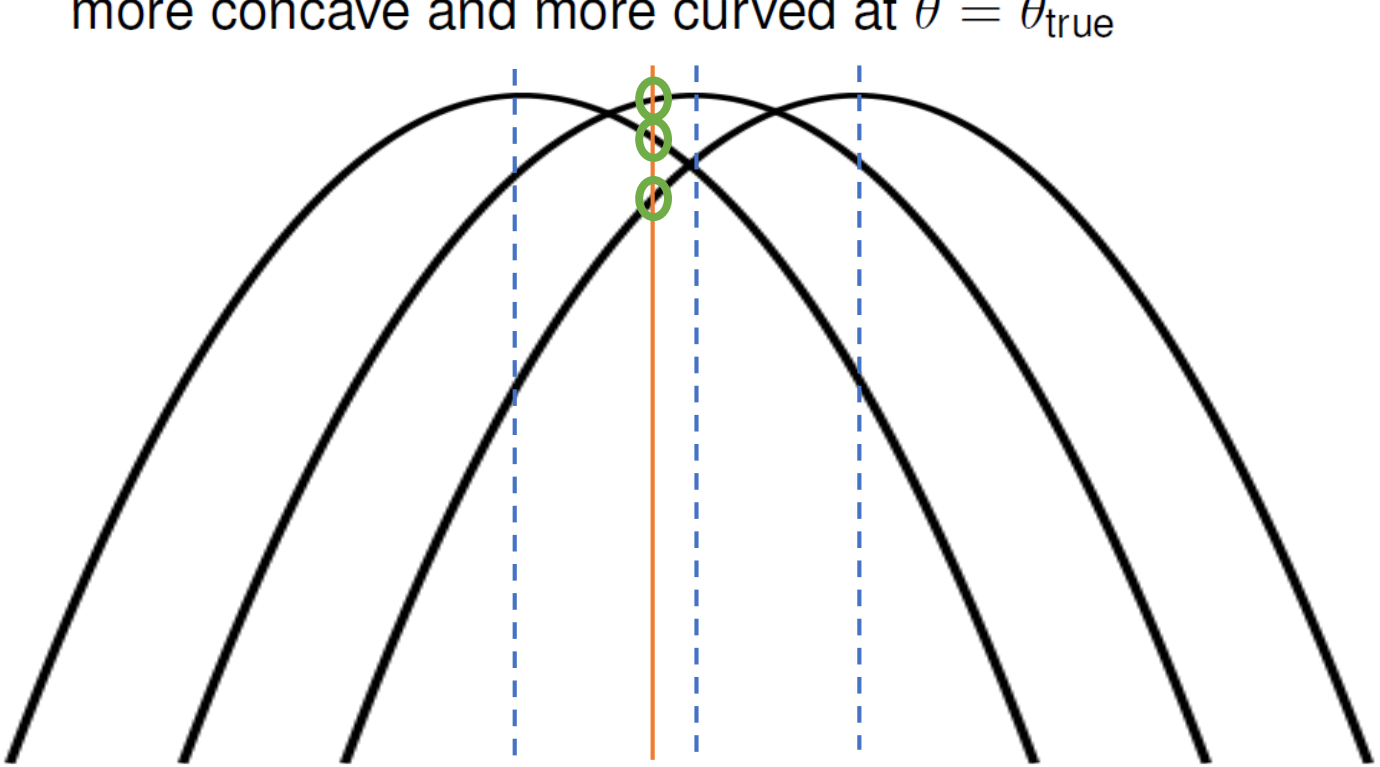
$$E_{P(X|\theta_{\text{true}})}\left[\frac{\frac{\partial^2 P(X|\theta)}{\partial\theta^2}}{P(X|\theta)}\Big|_{\theta=\theta_{\text{true}}}\right] = \int_x \frac{\partial^2 P(x|\theta)}{\partial\theta^2}dx = \frac{\partial^2}{\partial\theta^2}\int_x P(X|\theta)dx = 0$$

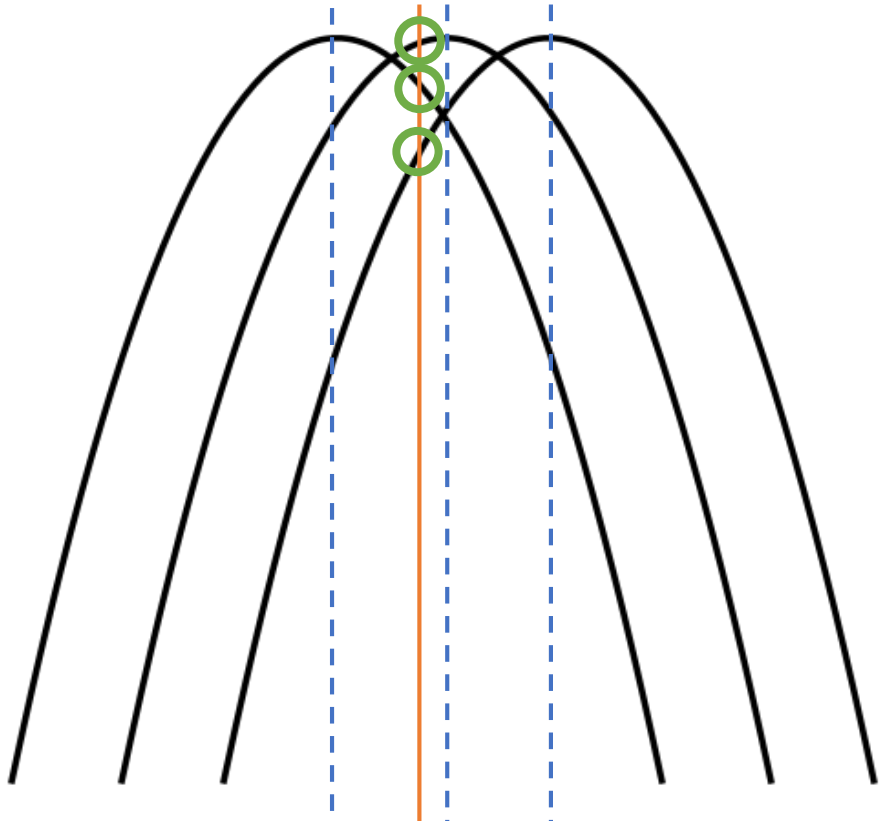
Fisher Information

- Alternate (equivalent) formulation for Fisher information

So, Fisher information is the expectation (over $x \sim P(X|\theta_{\text{true}})$) of the negative 2nd-derivative (curvature) of the log-likelihood function $\log P(x|\theta)$ evaluated at $\theta = \theta_{\text{true}}$

So, larger Fisher information means the log-likelihood function $\log P(x|\theta)$ is expected to be more concave and more curved at $\theta = \theta_{\text{true}}$



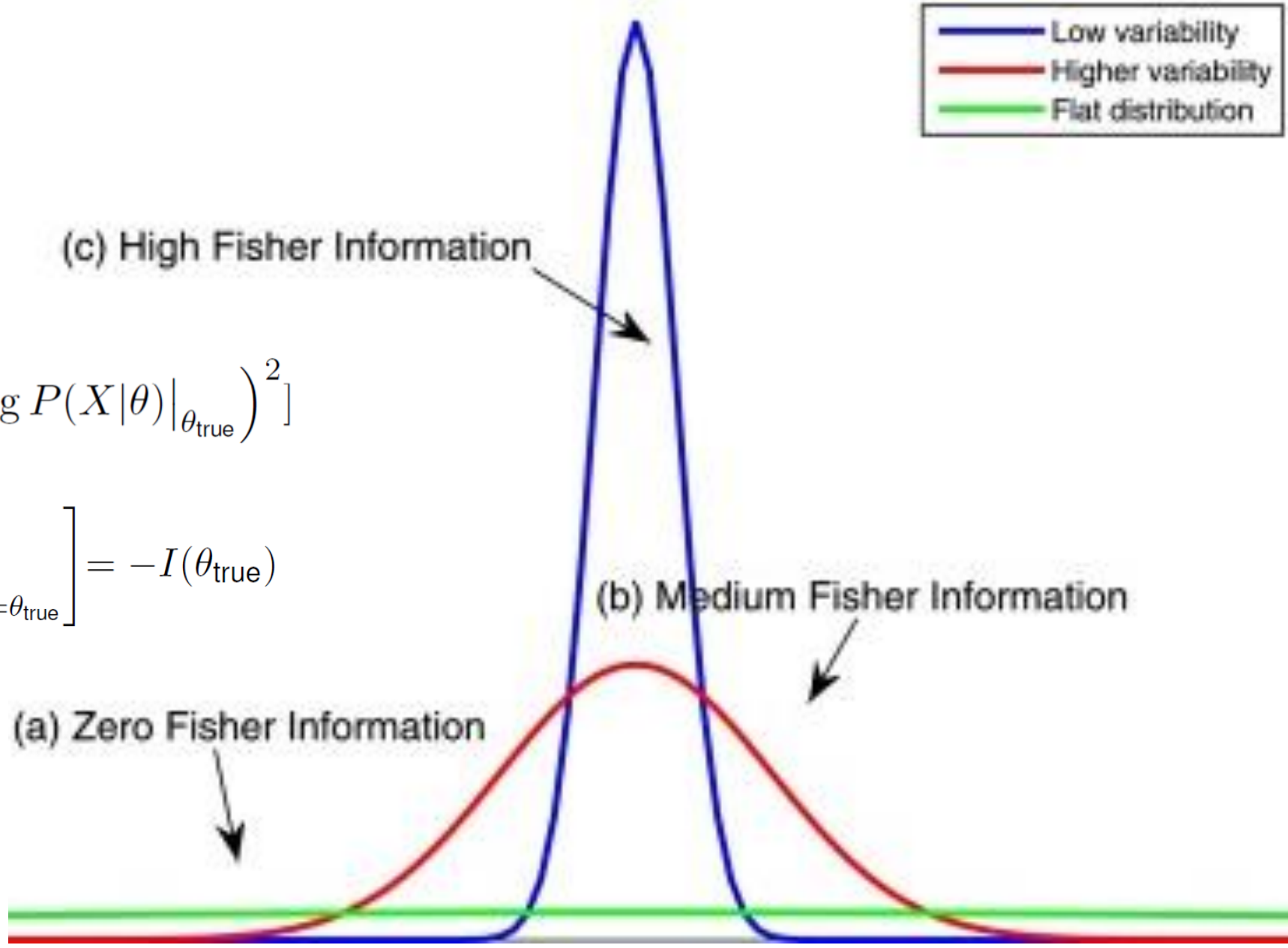


Fisher Information

- Two formulations

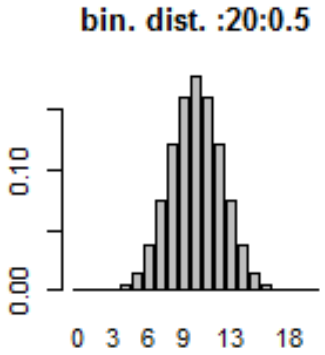
$$I(\theta_{\text{true}}) := E_{P(X|\theta_{\text{true}})}[\left(\frac{\partial}{\partial\theta}\log P(X|\theta)\Big|_{\theta_{\text{true}}}\right)^2]$$

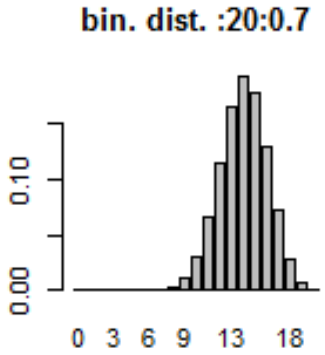
$$E_{P(X|\theta_{\text{true}})}\left[\frac{\partial^2}{\partial\theta^2}\log P(X|\theta)\Big|_{\theta=\theta_{\text{true}}}\right] = -I(\theta_{\text{true}})$$

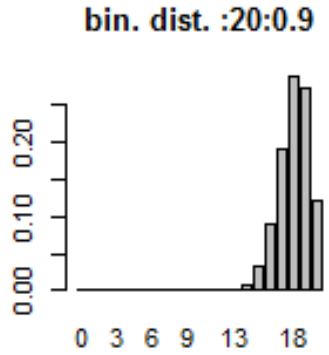


Fisher Information

- Example: Bernoulli random variable



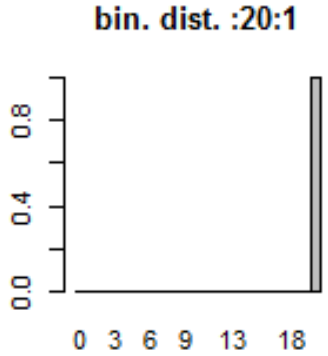




$$\log P(x|\theta) = x \log \theta + (1-x) \log(1-\theta)$$

$$\frac{\partial}{\partial \theta} \log P(x|\theta) = x/\theta - (1-x)/(1-\theta)$$

$$\frac{\partial^2}{\partial \theta^2} \log P(x|\theta) = -x/\theta^2 - (1-x)/(1-\theta)^2$$



$$I(\theta) = -E[\frac{\partial^2}{\partial \theta^2} \log P(x|\theta)] = \theta/\theta^2 + (1-\theta)/(1-\theta)^2 = 1/(\theta(1-\theta))$$

So, $I(\theta)$ is large when θ close to 0 or 1

For a dataset of size N, $I_N(\theta) = N/(\theta(1-\theta))$

So, $I_N(\theta)$ increases with N (as we should expect)

Fisher Information

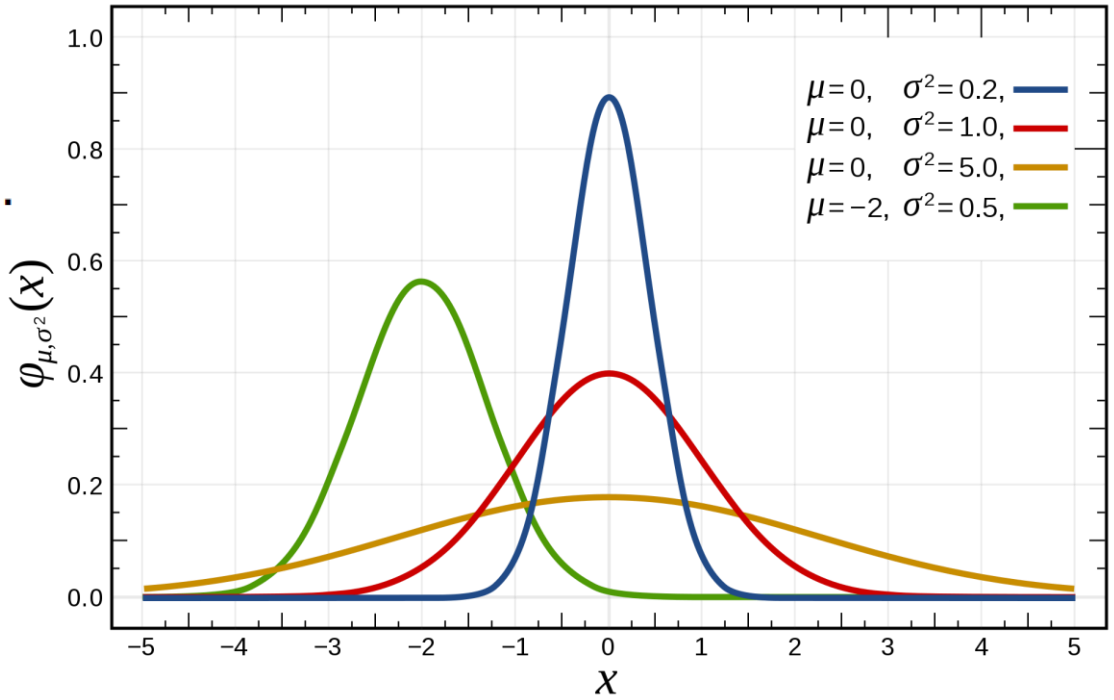
- Example: Gaussian random variable

Unknown mean parameter $\theta = \mu$. Known variance σ^2.

$$\frac{\partial}{\partial \mu} \log P(x|\mu) = (x - \mu)/\sigma^2$$

$$\frac{\partial^2}{\partial \mu^2} \log P(x|\mu) = -1/\sigma^2$$

$$I(\mu) = 1/\sigma^2$$



Here, $I(\mu)$ is independent of μ, but rather depends on the other parameter σ^2

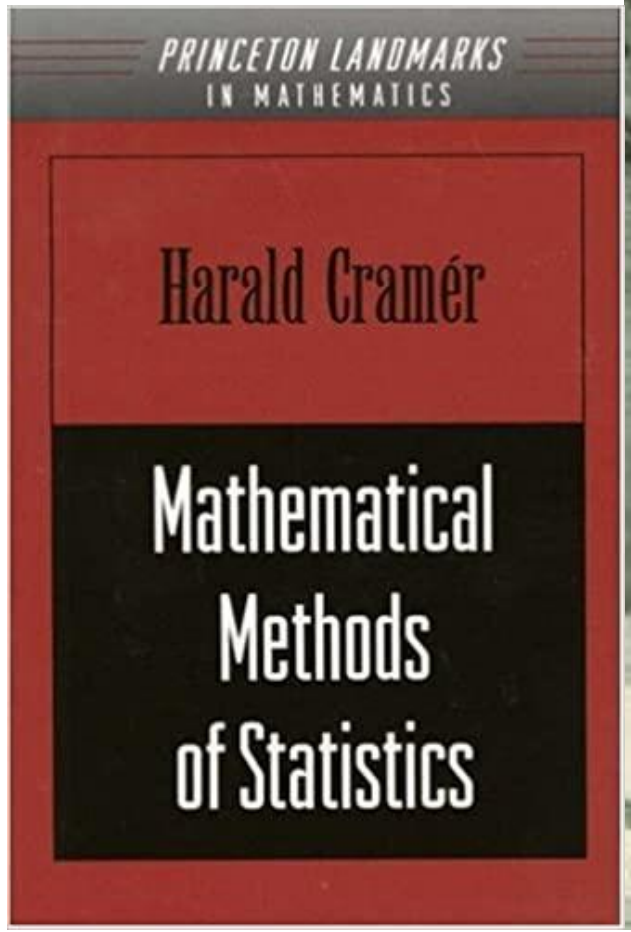
For a dataset of size N, $I_N(\mu) = N/\sigma^2$

Cramer-Rao Lower Bound

- How good can (some classes of) estimators ever be ?

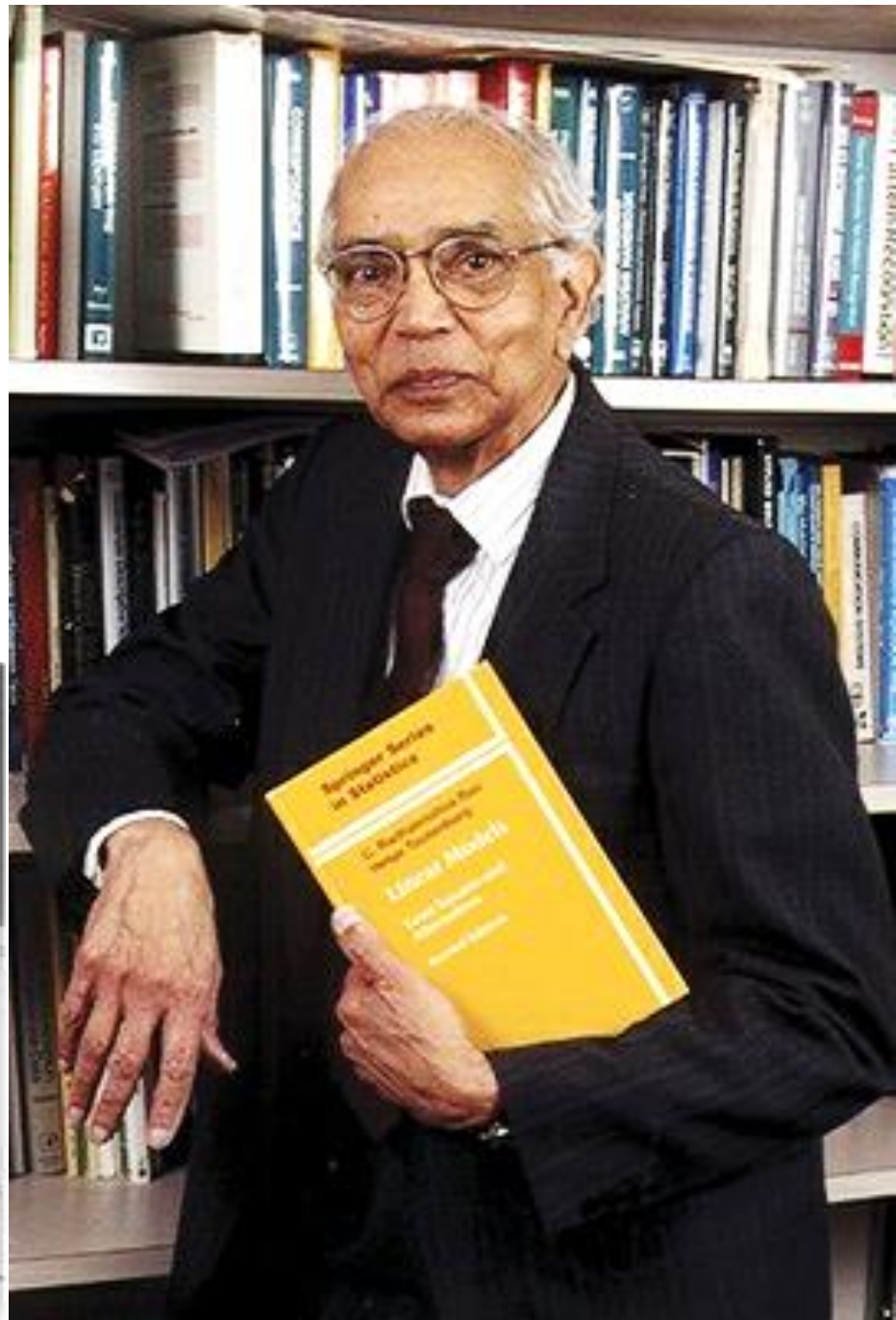
Cramer-Rao Lower Bound

- Harald Cramer
  - Swedish mathematician, actuary, statistician, specializing in mathematical statistics
  - “One of the giants of statistical theory“ – Sir John Kingman



Cramer-Rao Lower Bound

- Calyampudi Radhakrishna Rao (Born 1920)
  - Doctoral advisor: Fisher
  - American Statistical Assoc.: “a living legend”
  - bhavana.org.in/c-r-rao-a-life-in-statistics/
  - magazine.amstat.org/blog/2020/09/01/crrao/



Cramer-Rao Lower Bound

- Analysis

Let RV X model a dataset.

Assumption: Consider an **unbiased** estimator $\widehat{\theta}(X)$

Then, $E_{P(X|\theta_{\text{true}})}[\widehat{\theta}(X) - \theta_{\text{true}}] = 0 = \left(\int_x P(x|\theta)[\widehat{\theta}(x) - \theta]dx\right)\Big|_{\theta=\theta_{\text{true}}}$

This holds for all θ_{true}.

That is, $\int_x P(x|\theta')[\widehat{\theta}(x) - \theta']dx$ is a function of θ' that is identically zero. So, its derivative is also identically zero.

Thus, $0 = \frac{\partial}{\partial\theta}\left(\int_x P(x|\theta)[\widehat{\theta}(x) - \theta]dx\right)\Big|_{\theta=\theta_{\text{true}}}$

For convenience, lets call θ_{true} as θ

Thus, $\int_x [\widehat{\theta}(x) - \theta]\frac{\partial}{\partial\theta}P(x|\theta)dx = \int_x P(x|\theta)dx = 1$

Cramer-Rao Lower Bound

- Analysis

For convenience, lets call θ_{true} as θ

Thus, $\int_x [\widehat{\theta}(x) - \theta] \frac{\partial}{\partial \theta} P(x|\theta) dx = \int_x P(x|\theta) dx = 1$

Thus, $1 = \int_x [\widehat{\theta}(x) - \theta] P(x|\theta) \frac{\partial}{\partial \theta} \log P(x|\theta) dx$

$$= \int_x \left([\widehat{\theta}(x) - \theta] \sqrt{P(x|\theta)} \right) \left(\sqrt{P(x|\theta)} \frac{\partial}{\partial \theta} \log P(x|\theta) \right) dx$$

$$= \left[\int_x \left([\widehat{\theta}(x) - \theta] \sqrt{P(x|\theta)} \right) \left(\sqrt{P(x|\theta)} \frac{\partial}{\partial \theta} \log P(x|\theta) \right) dx \right]^2$$

Using Cauchy-Schwarz inequality, $1 \leq \int_x [\widehat{\theta}(x) - \theta]^2 P(x|\theta) dx \cdot \int_x P(x|\theta) \left(\frac{\partial}{\partial \theta} \log P(x|\theta) \right)^2 dx$

Thus, $\text{Var}(\widehat{\theta}(X)) \geq I(\theta)^{-1}$

Cramer-Rao Lower Bound

$$\mathrm{Var}(\widehat{\theta}(X)) \geq I(\theta)^{-1}$$

- Application example
  - For any Gaussian random variable,
    when mean is unknown, when variance is known (σ^2), when sample size = n:
    - We knew that ML estimator (sample mean) is unbiased
    - We knew that variance of ML estimator = σ^2/n
    - We know that Fisher information = n/σ^2
    - So, CRLB implies that any unbiased estimator of the mean will have variance $\geq \sigma^2/n$
    - Thus, ML estimator is an efficient estimator: minimum-variance unbiased estimator
- When ML estimator is unbiased and minimum-variance (as per CRLB), then why do we need the Bayesian estimator ?
  - For a finite data sample,
    Bayesian can reduce mean/expected squared error,
    at the cost of introducing a bias in the estimator

Bayesian Cramer-Rao Lower Bound

Bayesian Cramer-Rao Lower Bound

- A research paper
  - Applications of the van Trees Inequality: A Bayesian Cramer-Rao Bound. Bernoulli 1995, https://www.jstor.org/stable/3318681
  - Unlike CRLB, the Bayesian-CRLB gives us a lower bound for all (biased and unbiased both) estimators
- Statement

Let X model a dataset.
Consider likelihood $P(X|\theta)$ with "parameter" / RV θ
Consider a prior PDF $Q(\theta|\alpha)$ on "parameter" / RV θ with hyper-parameter α

$$E_{Q(\theta|\alpha)}[E_{P(X|\theta)}[\widehat{\theta}(X) - \theta]^2] \geq (E_{Q(\theta|\alpha)}[I_P(\theta)] + J_Q(\theta))^{-1}$$

where
$I_P(\theta)$ is the Fisher information of the likelihood associated with PDF / model $P(X|\theta)$, and
$J(Q;\alpha)$ is the "prior information" of the prior PDF / model $Q(\theta|\alpha)$

Bayesian Cramer-Rao Lower Bound

- Analysis

Assumption: Consider the prior θ defined on (compact) interval (a, b) such that:
$Q(\theta|\alpha) \to 0$ as $\theta \to a$ and as $\theta \to b$

$$\int_x [\widehat{\theta}(x) - \theta] \frac{\partial}{\partial \theta} P(x|\theta) dx = \int_x P(x|\theta) dx = 1$$

Then, similar to our strategy in proving CRLB, lets consider

$$\int_{\theta=a}^{b} \int_x \left(\widehat{\theta}(x) - \theta\right) \frac{\partial}{\partial \theta} \left(P(x|\theta) Q(\theta|\alpha)\right) dx d\theta$$

$$= \int_x \int_\theta \widehat{\theta}(x) \frac{\partial}{\partial \theta} \left(P(x|\theta) Q(\theta|\alpha)\right) d\theta dx - \int_x \int_\theta \theta \frac{\partial}{\partial \theta} \left(P(x|\theta) Q(\theta|\alpha)\right) d\theta dx$$

1st term includes the inner integral:

$$\int_\theta \widehat{\theta}(x) \frac{\partial}{\partial \theta} [P(x|\theta) Q(\theta|\alpha)] d\theta \quad = \widehat{\theta}(x) \int_\theta \frac{\partial}{\partial \theta} [P(x|\theta) Q(\theta|\alpha)] d\theta \quad = \widehat{\theta}(x) [P(x|\theta) Q(\theta|\alpha)]_a^b$$

$= 0$, because the prior $Q(\theta|\alpha)$ goes to zero at the boundary points a and b
So, the 1st term reduces to zero

Bayesian Cramer-Rao Lower Bound

- Analysis

Then, similar to our strategy in proving CRLB, lets consider

$$\int_{\theta=a}^{b}\int_{x}\left(\widehat{\theta}(x)-\theta\right)\frac{\partial}{\partial\theta}\left(P(x|\theta)Q(\theta|\alpha)\right)dxd\theta$$

$$=\int_{x}\int_{\theta}\widehat{\theta}(x)\frac{\partial}{\partial\theta}\left(P(x|\theta)Q(\theta|\alpha)\right)d\theta dx-\int_{x}\int_{\theta}\theta\frac{\partial}{\partial\theta}\left(P(x|\theta)Q(\theta|\alpha)\right)d\theta dx$$

2nd term (without the negative sign) includes an inner integral:

$$\int_{\theta}\theta\frac{\partial}{\partial\theta}\left[P(x|\theta)Q(\theta|\alpha)\right]d\theta=\left[\theta P(x|\theta)Q(\theta|\alpha)\right]_{a}^{b}-\int_{\theta}P(x|\theta)Q(\theta|\alpha)d\theta$$

$$=0-\int_{\theta}P(x|\theta)Q(\theta|\alpha)d\theta$$

So, 2nd term (with the negative sign) equals:

$$\int_{x}\int_{\theta}P(x|\theta)Q(\theta|\alpha)d\theta dx=\int_{\theta}Q(\theta|\alpha)\left(\int_{x}P(x|\theta)dx\right)d\theta=1$$

So, our original term equals 1:

$$1=\int_{\theta=a}^{b}\int_{x}\left(\widehat{\theta}(x)-\theta\right)\frac{\partial}{\partial\theta}\left(P(x|\theta)Q(\theta|\alpha)\right)dxd\theta$$

Bayesian Cramer-Rao Lower Bound

- Analysis

So, our original term equals 1:

$$1 = \int_{\theta=a}^{b} \int_{x} \left(\widehat{\theta}(x) - \theta\right) \frac{\partial}{\partial \theta} \left(P(x|\theta)Q(\theta|\alpha)\right) dx d\theta$$

$$= \int_{\theta=a}^{b} \int_{x} \left(\widehat{\theta}(x) - \theta\right) P(x|\theta)Q(\theta|\alpha) \frac{1}{P(x|\theta)Q(\theta|\alpha)} \frac{\partial}{\partial \theta} \left(P(x|\theta)Q(\theta|\alpha)\right) dx d\theta$$

$$= \left[\int_{\theta=a}^{b} \int_{x} \left(\widehat{\theta}(x) - \theta\right) \sqrt{P(x|\theta)Q(\theta|\alpha)} \sqrt{P(x|\theta)Q(\theta|\alpha)} \frac{\partial}{\partial \theta} \log \left(P(x|\theta)Q(\theta|\alpha)\right) dx d\theta\right]^2$$

Now, we apply the Cauchy-Schwarz inequality:

$$1 \leq \int_{\theta=a}^{b} \int_{x} \left(\widehat{\theta}(x) - \theta\right)^2 P(x|\theta)Q(\theta|\alpha) dx d\theta \cdot \int_{\theta=a}^{b} \int_{x} P(x|\theta)Q(\theta|\alpha) \left[\frac{\partial}{\partial \theta} \log P(x|\theta)Q(\theta|\alpha)\right]^2 dx d\theta$$

where

1st integral = expected squared error (NOT variance; because bias of estimator $\widehat{\theta}(x)$ may be non-zero)

2nd integral:

Bayesian Cramer-Rao Lower Bound

- Analysis

$$1 \leq \int_{\theta=a}^{b} \int_{x} \left(\widehat{\theta}(x) - \theta\right)^2 P(x|\theta)Q(\theta|\alpha)dxd\theta \cdot \int_{\theta=a}^{b} \int_{x} P(x|\theta)Q(\theta|\alpha) \left[\frac{\partial}{\partial\theta} \log P(x|\theta)Q(\theta|\alpha)\right]^2 dxd\theta$$

1st integral = expected squared error

2nd integral:

$$= \int_{\theta=a}^{b} \int_{x} P(x|\theta)Q(\theta|\alpha) \left[\frac{\partial}{\partial\theta} \log P(x|\theta)\right]^2 dxd\theta + \int_{\theta=a}^{b} \int_{x} P(x|\theta)Q(\theta|\alpha) \left[\frac{\partial}{\partial\theta} \log Q(\theta|\alpha)\right]^2 dxd\theta$$

$$+2 \int_{\theta=a}^{b} \int_{x} P(x|\theta)Q(\theta|\alpha) \frac{\partial}{\partial\theta} \log P(x|\theta) \frac{\partial}{\partial\theta} \log Q(\theta|\alpha)dxd\theta$$

where

1st term $= \int_{\theta=a}^{b} Q(\theta|\alpha) \left(\int_{x} P(x|\theta) \left[\frac{\partial}{\partial\theta} \log P(x|\theta)\right]^2 dx\right) d\theta = E_{Q(\theta|\alpha)}[I_P(\theta)]$

2nd term $= \int_{\theta=a}^{b} \left(\int_{x} P(x|\theta)dx\right) Q(\theta|\alpha) \left[\frac{\partial}{\partial\theta} \log Q(\theta|\alpha)\right]^2 d\theta = \int_{\theta=a}^{b} Q(\theta|\alpha) \left[\frac{\partial}{\partial\theta} \log Q(\theta|\alpha)\right]^2 d\theta$
$= J(Q;\alpha)$

3rd term $= 2 \int_{\theta=a}^{b} \frac{\partial}{\partial\theta} Q(\theta|\alpha) \cdot \int_{x} \frac{\partial}{\partial\theta} P(x|\theta)dx \cdot d\theta = 0$, because the inner integral is zero

Jeffreys Prior

Jeffreys Prior

- What happens to the functional form of the prior PDF on “parameter” if that “parameter” gets transformed ?
  - In general, the prior PDF’s functional form will change
    - e.g., uniform PDF on original parameter may become a non-uniform PDF on transformed parameter

Jeffreys Prior

- Effect of re-parametrization on functional form of prior PDF

Consider likelihood $P_\theta(X|\theta)$

Define prior $Q(\theta)$

Let transformed / reparametrized random variable be $\beta := f(\theta)$, where $f(\cdot)$ is strictly monotonic

For example, reparametrizing the univariate zero-mean Gaussian PDF by the variance $\beta := \sigma^2$, instead of the standard deviation $\theta := \sigma$.

Thus, instead of modeling likelihood as $P_\sigma(x|\sigma) := \exp(-0.5x^2/\sigma^2)/(\sigma\sqrt{2\pi})$ with prior as $Q(\sigma)$,

we model likelihood as $P_\beta(x|\beta) := \exp(-0.5x^2/\beta)/(\sqrt{\beta 2\pi})$ with prior as $R(\beta)$.

Note that $P_\beta(x|\beta) = P_{\sigma=\sqrt{\beta}}(x|\sigma = \sqrt{\beta})$, or

in general, $P_\beta(x|\beta) = P_{\theta=f^{-1}(\beta)}(x|\theta = f^{-1}(\beta))$ (because the transformation isn't on X).

Jeffreys Prior

- How can prior be invariant to transformation of parameter ?

Consider likelihood $P_\theta(X|\theta)$

Let Fisher information $I(\theta) := E_{P_\theta(X|\theta)}\left[\left(\frac{\partial}{\partial\theta}\log P_\theta(X|\theta)\right)^2\right]$

Define prior $Q(\theta)$ as $\propto \sqrt{I(\theta)}$

Let transformed / reparametrized random variable be $\beta := f(\theta)$, where $f(\cdot)$ is strictly monotonic

Then, what is the prior PDF $R(\beta)$ on the transformed random variable $\beta := f(\theta)$?

The transformation on the random variable gives $R(\beta) \propto Q(f^{-1}(\beta))\left|\frac{\partial f^{-1}(\beta)}{\partial\beta}\right|$

$$\propto \sqrt{E_{P_{f^{-1}(\beta)}(X|f^{-1}(\beta))}\left[\left(\frac{\partial \log P_{f^{-1}(\beta)}(X|f^{-1}(\beta))}{\partial f^{-1}(\beta)}\frac{\partial f^{-1}(\beta)}{\partial\beta}\right)^2\right]}$$

$$= \sqrt{E_{P_{f^{-1}(\beta)}(X|f^{-1}(\beta))}\left[\left(\frac{\partial \log P_{f^{-1}(\beta)}(X|f^{-1}(\beta))}{\partial\beta}\right)^2\right]} = \sqrt{E_{P_\beta(X|\beta)}\left[\left(\frac{\partial \log P_\beta(X|\beta)}{\partial\beta}\right)^2\right]} = \sqrt{I(\beta)}$$

(chain rule)

Jeffreys Prior

- Designing a prior to be invariant to transformations on parameters is a reasonable way of making the prior "non-informative"
  - Otherwise, e.g.,
    a flat prior on a parameter (indicating no preference towards specific values)
    becomes non-flat after some nonlinear transformation of that parameter
- Useful for scale parameters underlying a distribution
  - Scale-parameter example motivating Jeffreys prior
    - For Gaussian(0,σ^2): consider flat prior on standard deviation σ as Uniform(0,infinity)
      - Then, what is the prior on variance $v=\sigma^2$?
      - P(v) proportional to 1/sqrt(v), for positive v: prefers small variance (inconsistent with uniform)
    - For Gaussian(0,σ^2): consider prior on variance $v=\sigma^2$ as Uniform(0,infinity)
      - Then, what is the prior on standard deviation σ ?
      - P(σ) proportional to σ, for positive σ: prefers large standard deviation (inconsistent with uniform)
    - Jeffreys prior for Gaussian scale-parameter standard deviation σ is: P(σ) prop. to (1/σ)
    - Jeffreys prior for Gaussian scale-parameter variance $v=\sigma^2$ is: P(v) prop. to (1/v)

Conjugate Prior

- Motivation

If the posterior PDFs $P(\theta|x)$ are in the same family as the prior PDF $P(\theta)$, then:
(i) the prior and posterior are called *conjugate* PDFs, and
(ii) the prior is called the conjugate prior for the likelihood function

Advantage of conjugate priors: When the prior gives closed-form analytical expressions, then the posterior also gives closed-form analytical expressions, and its denominator / normalizing constant has a closed-form expression

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{\int P(x|\theta)P(\theta)d\theta}$$

Otherwise, a difficult numerical integration may be required to approximate the normalizing factor

Conjugate Prior

- Example

Example: Binomial Likelihood and Beta prior

1) Likelihood of s successes in n tries: $P(s, n|\theta) = {}^{n}C_{s}\theta^{s}(1-\theta)^{n-s}$, where $n \in \mathbb{N}$, $s \in \mathbb{I}_{\geq 0}$
2) Prior: $P(\theta) = \text{beta}(\theta; a \in \mathbb{R}^{+}, b \in \mathbb{R}^{+}) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$, Note: $a > 0, b > 0$
3) Posterior $\propto \theta^{s+a-1}(1-\theta)^{n-s+b-1} \equiv \text{beta}(\theta; a+s, b+n-s)$

- We know that the **mean** of the beta PDF $\text{beta}(\theta; a, b)$ is $a/(a+b)$

Thus, Bayes (mean) estimate = posterior mean = $(a+s)/(a+b+n)$
$= w(a/(a+b)) + (1-w)(s/n)$, where weight $w = (a+b)/(a+b+n)$

Note: When the sample size $n = 0$, the posterior mean $= a/(a+b) =$ prior mean

Note: As the sample size $n \to \infty$, the weight $w \to 0$ and the posterior mean $\to$ ML estimate

If prior $P(\theta) = 1$ is uniform over $\theta \in (0, 1)$, i.e., $\text{beta}(\theta, 1, 1)$

In that case, the likelihood determines the posterior

Conjugate Prior

- Example

Example: Binomial Likelihood and Beta prior

1) Likelihood of s successes in n tries: $P(s, n|\theta) = {}^{n}C_{s}\theta^{s}(1-\theta)^{n-s}$, where $n \in \mathbb{N}$, $s \in \mathbb{I}_{\geq 0}$
2) Prior: $P(\theta) = \text{beta}(\theta; a \in \mathbb{R}^{+}, b \in \mathbb{R}^{+}) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$, Note: $a > 0, b > 0$
3) Posterior $\propto \theta^{s+a-1}(1-\theta)^{n-s+b-1} \equiv \text{beta}(\theta; a+s, b+n-s)$

- We know that the **mode** of the beta PDF $\text{beta}(\theta; a, b)$ is $(a-1)/(a+b-2)$ for $a, b > 1$

So, posterior mode $= (a+s-1)/(a+b+n-2)$
$= w((a-1)/(a+b-2)) + (1-w)(s/n)$, where weight $w = (a+b-2)/(a+b+n-2)$

Note: When the sample size $n = 0$, the posterior mode $= (a-1)/(a+b-2) =$ prior mode

Note: As the sample size $n \to \infty$, the weight $w \to 0$ and the posterior mode $\to$ ML estimate

Conjugate Prior

- Example

Example: Gaussian (known mean μ, unknown variance θ) and Inverse Gamma

1) Likelihood: $P(x_1, \cdots, x_n|\mu, \theta) \propto \prod_{i=1}^{n} \theta^{-0.5} \exp(-0.5(x_i - \mu)^2/\theta)$
2) Prior = Inverse Gamma PDF: $P(\theta; a, b) \propto \theta^{-a-1} \exp(-b/\theta)$, where $a > 0, b > 0$
3) Posterior = Inverse Gamma PDF: $P(\theta; a + n/2, b + \sum_i (x_i - \mu)^2/2)$

- **Mean** of the inverse Gamma $P(\theta; a, b) = b/(a-1)$, for $a > 1$

Thus, Bayes estimate = posterior mean = $(b + \sum_i (x_i - \mu)^2/2)/(a + n/2 - 1)$
$= (2b + \sum_i (x_i - \mu)^2)/(2a + n - 2)$
$= w(b/(a-1)) + (1-w)\sum_i (x_i - \mu)^2/n$, where weight $w = (2a-2)/(2a+n-2)$

Note: When the sample size $n = 0$, the weight $w = 1$ and the posterior mean $= b/(a-1) =$ prior mean

Note: As the sample size $n \to \infty$, the weight $w \to 0$ and the posterior mean $\to$ ML estimate

Conjugate Prior

- Example

Example: Gaussian (known mean μ, unknown variance θ) and Inverse Gamma

1) Likelihood: $P(x_1, \cdots, x_n|\mu, \theta) \propto \prod_{i=1}^{n} \theta^{-0.5} \exp(-0.5(x_i - \mu)^2/\theta)$
2) Prior = Inverse Gamma PDF: $P(\theta; a, b) \propto \theta^{-a-1} \exp(-b/\theta)$, where $a > 0, b > 0$
3) Posterior = Inverse Gamma PDF: $P(\theta; a + n/2, b + \sum_i (x_i - \mu)^2/2)$

- **Mode** of the inverse Gamma $P(\theta; a, b) = b/(a + 1)$

So, posterior mode = $(b + \sum_i (x_i - \mu)^2/2)/(a + n/2 + 1)$
$= (2b + \sum_i (x_i - \mu)^2)/(2a + n + 2)$
$= w(b/(a + 1)) + (1 - w) \sum_i (x_i - \mu)^2/n$, where weight $w = (2a + 2)/(2a + n + 2)$

Note: When the sample size $n = 0$, the weight $w = 1$ and the posterior mode $= b/(a + 1) =$ prior mode

Note: As the sample size $n \to \infty$, the weight $w \to 0$ and the posterior mode $\to$ ML estimate

Conjugate Prior

- Example

Example: Poisson PDF and Gamma prior

Use this example to motivate the general result for exponential families later
1) Likelihood: $P(k_1, \cdots, k_n|\lambda) = \prod_i \lambda^{k_i} \exp(-\lambda)/k_i!$, where $\lambda \in \mathbb{R}^+, k_i \in \mathbb{I}^+$
2) Prior: $P(\theta) = \text{Gamma}(\lambda|\alpha, \beta) \propto \lambda^{\alpha-1} \exp(-\beta\lambda)$, where $\alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+, \lambda \in \mathbb{R}^+$
3) Posterior: $\propto \lambda^{\sum_i k_i + \alpha - 1} \exp(-n\lambda - \beta\lambda) \equiv \text{Gamma}(\lambda; \sum_i k_i + \alpha, n + \beta)$

- For a Gamma distribution $\text{Gamma}(\lambda|\alpha, \beta)$, we know that the **mean** is α/β

Thus, the Bayes estimate = posterior mean = $(\sum_i k_i + \alpha)/(n + \beta)$
$= w(\alpha/\beta) + (1 - w)\sum_i k_i/n$, where weight $w = \beta/(\beta + n)$
$= w(\alpha/\beta) + (1 - w)\widehat{\lambda}_{\text{MLE}}$

Note: When the sample size $n = 0$, the weight $w = 1$ and the posterior mean $= \alpha/\beta =$ prior mean

Note: As the sample size $n \to \infty$, the weight $w \to 0$ and the posterior mean $\to$ ML estimate

Conjugate Prior

- Example

Example: Poisson PDF and Gamma prior

Use this example to motivate the general result for exponential families later
1) Likelihood: $P(k_1, \cdots, k_n|\lambda) = \prod_i \lambda^{k_i} \exp(-\lambda)/k_i!$, where $\lambda \in \mathbb{R}^+, k_i \in \mathbb{I}^+$
2) Prior: $P(\theta) = \text{Gamma}(\lambda|\alpha, \beta) \propto \lambda^{\alpha-1} \exp(-\beta\lambda)$, where $\alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+, \lambda \in \mathbb{R}^+$
3) Posterior: $\propto \lambda^{\sum_i k_i + \alpha - 1} \exp(-n\lambda - \beta\lambda) \equiv \text{Gamma}(\lambda; \sum_i k_i + \alpha, n + \beta)$

• For a Gamma distribution $\text{Gamma}(\lambda|\alpha, \beta)$, we know that the **mode** is $(\alpha - 1)/\beta$ when $\alpha \geq 1$. When $\alpha < 1$, the case is tricky.

Then, posterior mode = $(\sum_i k_i + \alpha - 1)/(n + \beta)$
$= w((\alpha - 1)/\beta) + (1 - w)\sum_i k_i/n$, where weight $w = \beta/(\beta + n)$

Note: When the sample size $n = 0$, the weight $w = 1$ and the posterior mode $= (\alpha - 1)/\beta =$ prior mode

Note: As the sample size $n \to \infty$, the weight $w \to 0$ and the posterior mode $\to$ ML estimate