Statistics for Applications

Chapter 8: Bayesian Statistics

what we have been doing so far is frequentist approach.

The Bayesian approach (1)

- So far, we have studied the frequentist approach of statistics.
- The frequentist approach:
  - Observe data
  - These data were generated randomly (by Nature, by measurements, by designing a survey, etc...)
  - We made assumptions on the generating process (e.g., i.i.d., Gaussian data, smooth density, linear regression function, etc...)
  - The generating process was associated to some object of interest (e.g., a parameter, a density, etc...)
  - This object was unknown but fixed and we wanted to find it: we either estimated it or tested a hypothesis about this object, etc...

* The First thing we had data, we observed some data and we assumed that this data is generated randomly, The reason we did that is because this will allow to leverage tools from probability.

* Then we made some assumption's on the data generating Process, for Eg, we assumed they are iid, sometimes we assume gaussian, etc...

* So this data, we are interested in some Parameter θ, or β in regression model, so we have this unknown problem, unknown Parameter we want to find it. we want it either Estimate it, or test it, or may be find confidence interval on this object.

"The Object was unknown but fixed" is different for Bayesian approach

The Bayesian approach (2)

- Now, we still observe data, assumed to be randomly generated by some process. Under some assumptions (e.g., parametric distribution), this process is associated with some fixed object.

- We have a **prior belief** about it.

- Using the data, we want to update that belief and transform it into a **posterior belief**.

in the Bayesian approach we still assume we observe some random data, But the generating process is slightly different, its sort of two layer process,

① That generate the Parameter
② Given this parameter generate the data

SO what the 1st layer does, nobody really believes that there is some random process that's happening, about generating what is gonna to be the true expected number of people when they kiss.

* But often we actually have Prior Belief

Ex: when we did least squares, we looked over for all of the vectors in all of the $\mathbb{R}^p$ including the ones that are 50 million. SO Those are things we might be able to rule out, may be we might be rule out that on a much smaller scale.

* This prior belief is gonna play hopefully less and less important role as we collect more and more data, but if we have smaller amount of data, we might want to be able to use this information, rather than just shooting in the dark.

* So idea is to have this prior belief & then we want to update this prior belief what's called posterior belief, after we have seen some data.

* So belief encompasses basically what you think and how strongly you think about it.

* Ex: if we have some belief about parameter θ, may be my belief is telling me, where the θ should be and how strongle I believe mean's I have a very narrow region where θ could be.

The Bayesian approach (3)

Example

- Let p be the proportion of woman in the population.
- Sample n people randomly with replacement in the population and denote by $X_1, \ldots, X_n$ their gender (1 for woman, 0 otherwise).
- In the frequentist approach, we estimated p (using the MLE), we constructed some confidence interval for p, we did hypothesis testing (e.g., $H_0 : p = .5$ v.s. $H_1 : p \neq .5$).
- Before analyzing the data, we may believe that p is likely to be close to $1/2$.
- The Bayesian approach is a tool to:
  1. include mathematically our prior belief in statistical procedures.
  2. update our prior belief using the data.

P: Proportion of women in population

Collect data: $X_1, X_2, \ldots, X_n \sim Ber(p)$ $p \in (0,1)$

frequentist approach: Let's just estimate the $\hat{p} = \frac{1}{n}\sum X_i = \bar{X}_n$ & do some tests whether $p = 0.5$ or not etc..

But here this is the case where we definitely have prior belief of what p should be. We are pretty confident that p is not going to be 0.7. We actually believe p is extremely close to 0.5, may be not exactly.

* So, we are going to want to integrate that knowledge, so we could integrate it in blunt manner by saying, like dis card the data and say that $p = 0.5$.
* But may be that's just too much, so how do we do this trade off b/w adding the data & combining with this prior knowledge?

in many ways, in many instances, essentially what's gonna happen in this O.S is going to act like one new observation. So if we have five observations, this will be sixth observation, which will play a role.

* if we have million observations, we are going to have one million and one observations, and its not going to play so much of a role.

* we have 5 observation $\Rightarrow$ 5 + 1 (Prior belief)
= 6 observations

if we have 1 milion observations $\Rightarrow$ 1 million + 1 (Prior belief)
(Not gonna play big role)

(1) The Bayesian approach is a tool to one to include mathematically our Prior Belief into Statistical Procedures,

(2) update this Prior Belief into a Posterior Belief by using the data.

How do we do this?

there is two layer's.

① one is where we draw the parameter at random

② once we have the parameter, condition this parameter and draw our data.

No Body Believes this actually happening that Nature is Just rolling dice, for us and Choosing parameter's at random. This idea that Parameter coming from some random distribution actually Capture very well that this idea that how you would encompass your prior.

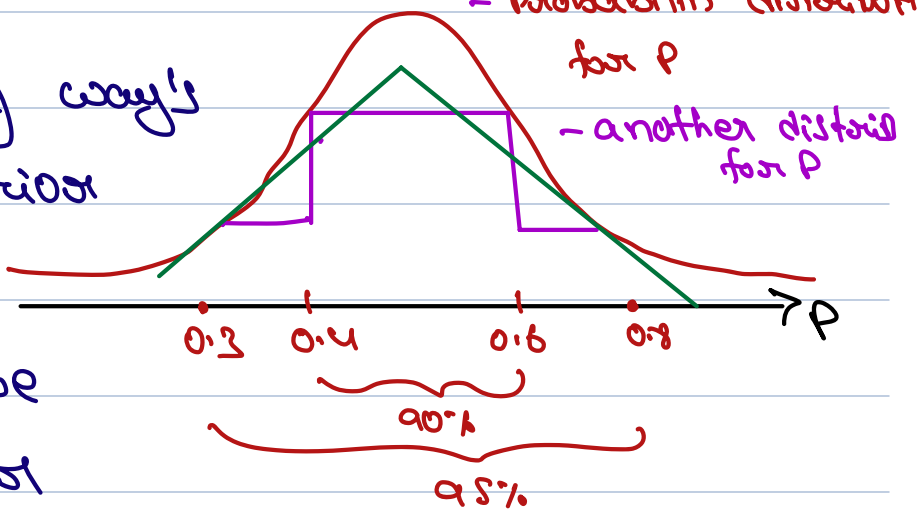
The Bayesian approach (4)

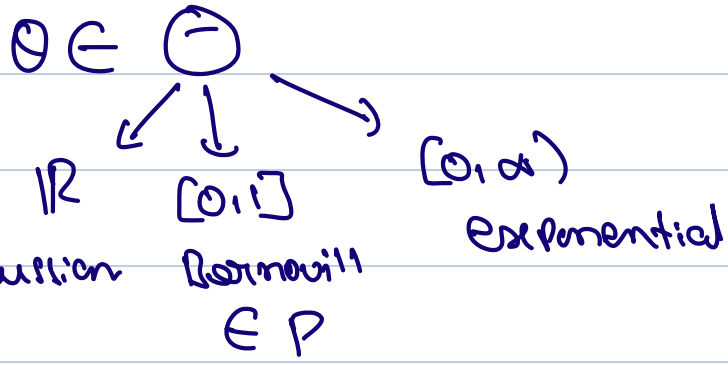
Example (continued)

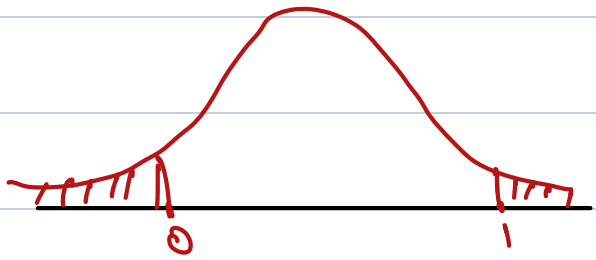
- Our prior belief about p can be quantified:
- E.g., we are 90% sure that p is between $.4$ and $.6$, 95% that it is between $.3$ and $.8$, etc...
- Hence, we can model our prior belief using a distribution for p, *as if* p was random.
- In reality, the true parameter is not random ! However, the Bayesian approach is a way of modeling our belief about the parameter by doing **as if** it was random.
- E.g., $p \sim \mathcal{B}(a, a)$ (*Beta distribution*) for some $a > 0$.
- This distribution is called the *prior distribution*.

Ex:

there are many ways I can assume my prior belief. some of them are definitely to be mathematically more convenient than other's. Hopefully we are gonna have things that we can parameterize very well.







cannot have Gaussian as it has prob outside $[0,1]$

Beta distribution:

$$x \sim \text{Beta}(a,b) \qquad f(x) = \begin{cases} \frac{1}{c} x^{a-1} \cdot (1-x)^{b-1} & \text{if } x \in (0,1) \\ 0 & \text{o.w} \end{cases}$$

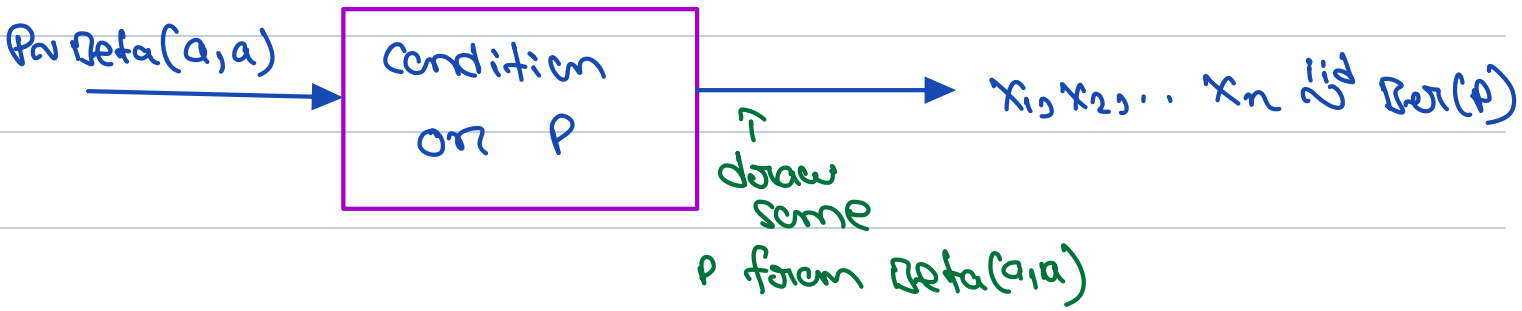
The Bayesian approach (5)

Example (continued)

- In our statistical experiment, $X_1, \ldots, X_n$ are assumed to be i.i.d. Bernoulli r.v. with parameter p **conditionally on** p.
- After observing the available sample $X_1, \ldots, X_n$, we can update our belief about p by taking its distribution conditionally on the data.
- The distribution of p conditionally on the data is called the *posterior distribution*.
- Here, the posterior distribution is

$$\mathcal{B}\left(a + \sum_{i=1}^{n} X_i, a + n - \sum_{i=1}^{n} X_i\right).$$

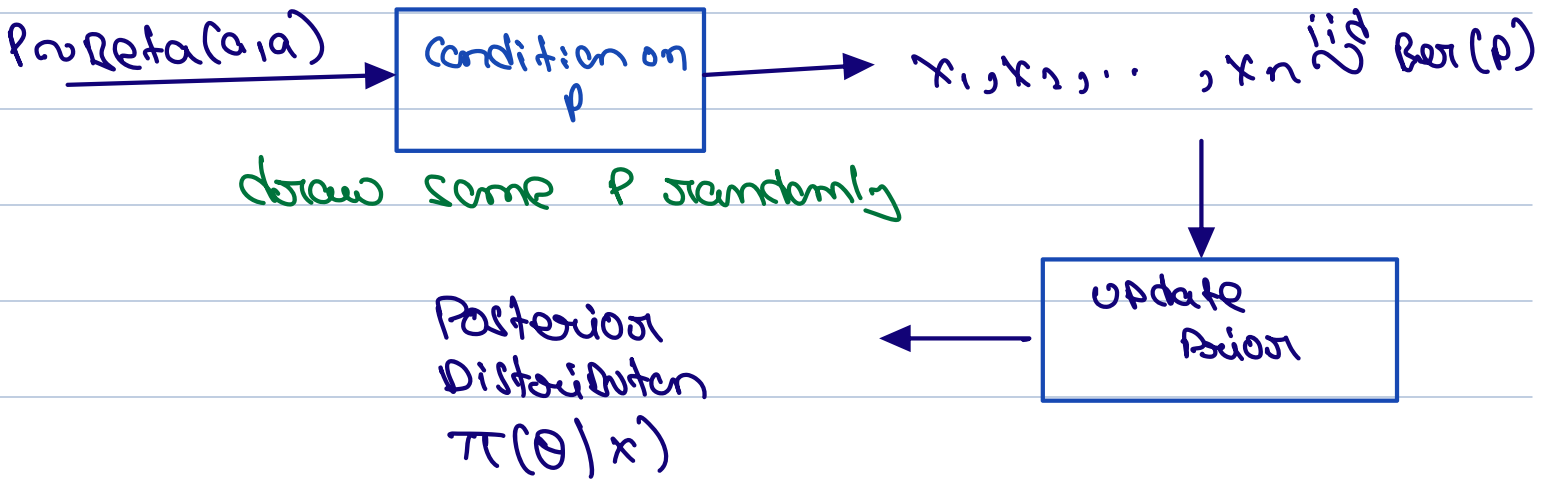
Flow Chat:



p is randomly drawn from Beta(a,a)

* So we first draw parameter $p \sim Beta(a,a)$ (assume $p=0.52$) and then we just flip those independent biased coin's with this particular $p=0.52$ (2 layer process)

* This is just a through process. its not anything that actually happen's in general. This is our way of thinking about how the data was generated.



in this case we will see the posterior distribution is still Beta

Beta is Conjugate: meaning I put Beta as Prior and I get Beta as Posterior.

$$\pi(p(x) \sim \text{Beta}\left(a + \sum x_i,\ a + n - \sum x_i\right)$$

So how do we get this Posterior distribution given the Prior? How do we update this? This is called Bayesian Statistics.

Bayes Formula:

B = data $(x_1, x_2, \ldots x_n)$

A = Parameter (θ)

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \mathbb{P}(B|A) \cdot \frac{P(A)}{\mathbb{P}(B)}$$

($P(A)$: $\mathbb{P}_\theta(\theta)$; $\mathbb{P}(B)$: $P_x(x)$)

$\mathbb{P}(A|B)$: dist of Parameter θ, given I have seen the data $\mathbb{P}_{\theta|x}(\theta | x_1, x_2, \ldots x_n)$

$\mathbb{P}(B|A)$: distribution of data given that I know what my parameter θ is $\mathbb{P}_{x|\theta}(x_1, x_2, \ldots x_n | \theta)$

$\mathbb{P}(A) = \mathbb{P}(\theta)$ = Prior distribution

$\mathbb{P}(B) = \mathbb{P}_x(x_1, x_2, \ldots x_n)$ = distribution of data itself.

The Bayes rule and the posterior distribution (1)

- Consider a probability distribution on a parameter space Θ with some pdf $\pi(\cdot)$: the *prior distribution*.
- Let $X_1, \ldots, X_n$ be a sample of n random variables.
- Denote by $p_n(\cdot|\theta)$ the joint pdf of $X_1, \ldots, X_n$ conditionally on θ, where $\theta \sim \pi$.
- Usually, one assumes that $X_1, \ldots, X_n$ are i.i.d. conditionally on θ.
- The conditional distribution of θ given $X_1, \ldots, X_n$ is called the *posterior distribution*. Denote by $\pi(\cdot|X_1, \ldots, X_n)$ its pdf.

$X_1, X_2, \ldots, X_n$ iid

↓ Give Parameter θ

Conditioned on random Parameter, as if it was a fixed Number

$X_1, X_2, \ldots, X_n \sim P_n(x_1, x_2, \ldots, x_n | \theta)$

Pdf or Pmf

Ex: $X_1, X_2, \ldots X_n$ iid $N(\theta, 1)$

$$\Rightarrow P_n(x_1, x_2, \ldots x_n | \theta) = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\sum(x_i-\theta)^2\right\}$$

$\Rightarrow$ θ has a Prior Distribution $\pi(\cdot)$ pdf or pmf

Ex: $\pi(\theta) = C \cdot \theta^{a-1}(1-\theta)^{a-1}$, $\theta \sim Beta(a, a)$

Posterior Distribution:

$$\pi(\theta | x_1, x_2, \ldots, x_n) = \frac{P_n(x_1, x_2, \ldots x_n | \theta)\, \pi(\theta)}{\int_\theta P_n(x_1, x_2, \ldots x_n | \theta)\, \pi(\theta)\, d\theta}$$

↓ marginal distribution

weighted version of Likelihood

Posterior dist

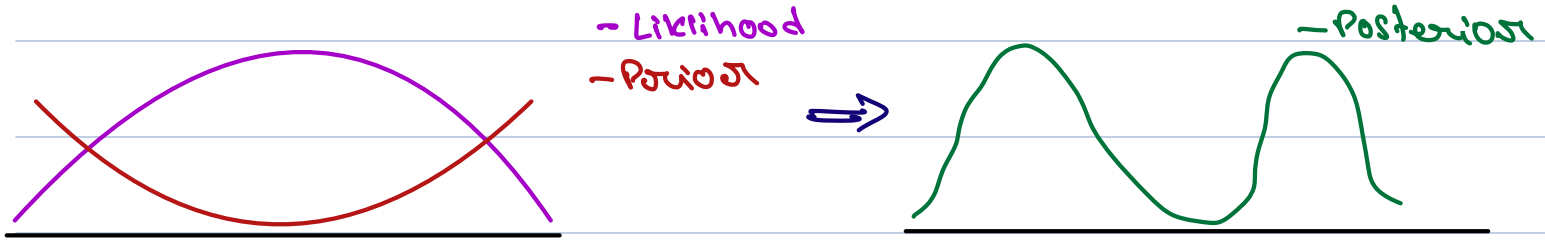
$$\pi(\theta|x_1,x_2,\ldots,x_n) = \frac{P_n(x_1,x_2,\ldots x_n|\theta)\,\pi(\theta)}{\int_\theta P_n(x_1,x_2,\ldots,x_n|\theta)\,\pi(\theta)\,d\theta}$$

Likelihood

Prior dist

Constant that does not depend on θ

* Important to remember: MLE & Bayesian are really not that different. because our posterior is really just like your likelihood time's something that's just poking some weight's on the θ's. dependending on where the θ should be.



Likelihood Prior dist

$P_n(x_1,x_2,\ldots x_n|\theta)\;\pi(\theta)$ ⇒ weighted version of Likelihood.

* weighting the Likelihood using my prior belief on θ.
* if we follow MLE principle, if we maximize the weighted Likelihood we would get Maximum a Posterior Estimator $\hat{\theta}_{MAP}$

MAP → Maximum a posterior

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \pi(\theta | x_1, x_2, \ldots x_n)$$

The Beauty of Bayesian Statistics is we Don't have to take any number in particular (like maximum, mean, median etc). We have an entire posterior distribution.

The Bayes rule and the posterior distribution (2)

- Bayes' formula states that:

$$\pi(\theta|X_1,\ldots,X_n) \propto \pi(\theta)p_n(X_1,\ldots,X_n|\theta), \quad \forall\theta\in\Theta.$$

- The constant does not depend on θ:

$$\pi(\theta|X_1,\ldots,X_n) = \frac{\pi(\theta)p_n(X_1,\ldots,X_n|\theta)}{\int_\Theta p_n(X_1,\ldots,X_n|t)\,\mathrm{d}\pi(t)}, \quad \forall\theta\in\Theta.$$

The Bayes rule and the posterior distribution (3)

In the previous example:

- $\pi(p) \propto p^{a-1}(1-p)^{a-1}, p \in (0,1)$.
- Given p, $X_1, \ldots, X_n \overset{i.i.d.}{\sim} Ber(p)$, so

$$p_n(X_1, \ldots, X_n|\theta) = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}.$$

- Hence,

$$\pi(\theta|X_1, \ldots, X_n) \propto p^{a-1+\sum_{i=1}^n X_i}(1-p)^{a-1+n-\sum_{i=1}^n X_i}.$$

- The posterior distribution is

$$\mathcal{B}\left(a + \sum_{i=1}^n X_i, a + n - \sum_{i=1}^n X_i\right).$$

Prior $\pi(\theta) \sim \text{Beta}(a, a)$

$\Rightarrow \quad \pi(p) \propto p^{a-1}(1-p)^{a-1}$ (Density of p)

Likelihood: $P_n(x_1, x_2, \ldots, x_n \mid p) = p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}$

Posterior distribution: $\propto p^{a-1}(1-p)^{a-1} p^{\sum x_i} (1-p)^{n-\sum x_i}$

$$\propto p^{a+\sum x_i - 1} (1-p)^{a+n-\sum x_i - 1}$$

$$\pi(\theta \mid x_1, x_2, \ldots x_n) \sim \text{Beta}(a + \sum x_i, a + n - \sum x_i)$$

* The Prior $\pi(\theta)$ in the Posterior formula is to weigh some theta's more than other's, depending on their Prior Belief
* if our Prior Belief does not want to put any Preference towards some theta's than to other's, what do we do? $\pi(\theta) = 1$

* if $\pi(\theta) = 1$, constant and does not depend on θ, that would mean that we are not preferring anything, and we are looking at the Likelihood.

⇒ we are looking at the likelihood, but not as a function that we are trying to maximize, but it is a function that we normalize in such a way its actually a distribution.

Non informative priors (1)

- Idea: In case of ignorance, or of lack of prior information, one may want to use a prior that is as little informative as possible.
- Good candidate: $\pi(\theta) \propto 1$, i.e., constant pdf on Θ.
- If Θ is bounded, this is the uniform prior on Θ.
- If Θ is unbounded, this does not define a proper pdf on Θ !
- An *improper prior* on Θ is a measurable, nonnegative function $\pi(\cdot)$ defined on Θ that is not integrable.
- In general, one can still define a posterior distribution using an improper prior, using Bayes' formula.

Non informative priors (2)

Examples:

- If $p \sim \mathcal{U}(0,1)$ and given p, $X_1, \ldots, X_n \overset{i.i.d.}{\sim} Ber(p)$:

$$\pi(p|X_1, \ldots, X_n) \propto p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i},$$

i.e., the posterior distribution is

$$\mathcal{B}\left(1 + \sum_{i=1}^n X_i, 1 + n - \sum_{i=1}^n X_i\right).$$

- If $\pi(\theta) = 1, \forall \theta \in \mathbb{R}$ and given θ, $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$:

$$\pi(\theta|X_1, \ldots, X_n) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^n (X_i - \theta)^2\right),$$

i.e., the posterior distribution is

$$\mathcal{N}\left(\bar{X}_n, \frac{1}{n}\right).$$

Prior: $p \sim Unif(0,1)$

Likelihood: $x_1, x_2, \ldots x_n \overset{iid}{\sim} Ber(p)$

$\Rightarrow p_n(x_1, x_2, \ldots, x_n | p) =$

$= p^{\sum x_i} (1-p)^{n - \sum x_i}$

Posterior:

$$\pi(p | x_1, x_2, \ldots x_n) = \frac{p^{\sum x_i} (1-p)^{n-\sum x_i} \cdot 1}{\int_0^1 p^{\sum x_i} (1-p)^{n-\sum x_i} dp}$$

$$\pi(\theta) = Beta\left(1 + \sum_{i=1}^{n} x_i, 1 + n - \sum_{i=1}^{n} x_i\right)$$

$\hat{p}_{MLE} = \bar{x}$ (MLE)

Non informative priors (3)

- *Jeffreys prior:*

$$\pi_J(\theta) \propto \sqrt{\det I(\theta)},$$

where $I(\theta)$ is the Fisher information matrix of the statistical model associated with $X_1, \ldots, X_n$ in the frequentist approach (provided it exists).

- In the previous examples:
  - Ex. 1: $\pi_J(p) \propto \frac{1}{\sqrt{p(1-p)}}$, $p \in (0,1)$: the prior is $\mathcal{B}(1/2, 1/2)$.
  - Ex. 2: $\pi_J(\theta) \propto 1$, $\theta \in \mathbb{R}$ is an improper prior.

Non informative priors (4)

- Jeffreys prior satisfies a reparametrization invariance principle: If η is a reparametrization of θ (i.e., $\eta = \phi(\theta)$ for some one-to-one map ϕ), then the pdf $\tilde{\pi}(\cdot)$ of η satisfies:

$$\tilde{\pi}(\eta) \propto \sqrt{\det \tilde{I}(\eta)},$$

  where $\tilde{I}(\eta)$ is the Fisher information of the statistical model parametrized by η instead of θ.

Bayesian confidence regions

- For $\alpha \in (0,1)$, a Bayesian confidence region with level α is a random subset $\mathcal{R}$ of the parameter space Θ, which depends on the sample $X_1, \ldots, X_n$, such that:

$$\mathbb{P}[\theta \in \mathcal{R} | X_1, \ldots, X_n] = 1 - \alpha.$$

- Note that $\mathcal{R}$ depends on the prior $\pi(\cdot)$.

- "Bayesian confidence region" and "confidence interval" are two **distinct** notions.

Bayesian estimation (1)

- The Bayesian framework can also be used to estimate the true underlying parameter (hence, in a frequentist approach).
- In this case, the prior distribution does not reflect a prior belief: It is just an artificial tool used in order to define a new class of estimators.
- **Back to the frequentist approach:** The sample $X_1, \ldots, X_n$ is associated with a statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$.
- Define a distribution (that can be improper) with pdf π on the parameter space Θ.
- Compute the posterior pdf $\pi(\cdot|X_1, \ldots, X_n)$ associated with π, seen as a prior distribution.

Bayesian estimation (2)

- *Bayes estimator:*

$$\hat{\theta}^{(\pi)} = \int_{\Theta} \theta \, \mathrm{d}\pi(\theta | X_1, \ldots, X_n) :$$

  This is the *posterior mean*.

- The Bayesian estimator depends on the choice of the prior distribution π (hence the superscript π).

Bayesian estimation (3)

- In the previous examples:
  - Ex. 1 with prior $\mathcal{B}(a, a)$ $(a > 0)$:

$$\hat{p}^{(\pi)} = \frac{a + \sum_{i=1}^{n} X_i}{2a + n} = \frac{a/n + \bar{X}_n}{2a/n + 1}.$$

    In particular, for $a = 1/2$ (Jeffreys prior),

$$\hat{p}^{(\pi_J)} = \frac{1/(2n) + \bar{X}_n}{1/n + 1}.$$

  - Ex. 2: $\hat{\theta}^{(\pi_J)} = \bar{X}_n$.
- In each of these examples, the Bayes estimator is consistent and asymptotically normal.
- In general, the asymptotic properties of the Bayes estimator do not depend on the choice of the prior.

18.650 / 18.6501 Statistics for Applications

Fall 2016