Pengantar Statistika Bayes - STA1312

Frequentist vs Bayesian

Dr. Kusman Sadik, S.Si, M.Si
Program Studi Statistika dan Sains Data IPB
Tahun 2024

Pendekatan Utama dalam Statistika

Ada dua pendekatan filosofis utama dalam statistika:

- Pendekatan frequentist (*frequentist approach*) atau sering pula disebut pendekatan klasik (*classical approach*).
- Pendekatan Bayesian (*Bayesian approach*). Pendekatan ini berpotensi memberikan banyak keuntungan mendasar dibandingkan pendekatan frequentist.

KS - STK - IPB

What is statistics?

- The study about uncertainty

- Statistics is the exploration of a parameter (theta) in light of data X.

KS - STK - IPB

Perbedaan utama antara pendekatan classical dan Bayesian:

	Classical	Bayesian
(1). What's probability?		
(2). Data X		
(3). Parameter θ		
(4). Confidence Interval		

KS - STK - IPB

What is probability?

- **Classical**
  - frequency of a certain event when repeating a random experiment infinite times
- **Bayesian**
  - the quantification of uncertainty –de Finetti
  - this uncertainty can be equated with personal belief of a quantity

KS - STK - IPB

- **Examples:**
  - probability of heads in a coin toss experiment: defined
  - probability that it will rain Jan 27, 2024: not defined.
  - probability of me getting cancer by 70 if smoking 1 pack cig a day: not defined
- **Examples**
  - belief (can be updated by observation)
  - belief (can be refined by modeling)
  - belief (as above)

Data and Parameter

❖ Classical:

- Parameter is fixed
- Data is random

❖ Bayesian

- Parameter has uncertainty ("random")
- Data is given ("fixed" in operational sense)

KS - STK - IPB

- Goal of both: draw conclusion/inference for θ

Data and Parameter (cont.)

- **Classical:**
  - Parameter θ is fixed
  - Data X is random
- **Bayesian**
  - Inference should be the same if data are the same: likelihood principle
  - We can talk about distribution of θ, to express uncertainty/belief about θ, whereas classical stats cannot.

KS - STK - IPB

Confidence/Credible Interval for θ

- **Classical**
  - Confidence Interval
  - If samples of the same size are drawn repeatedly from a population specified by θ, and a confidence interval is calculated from each sample for θ, then 95% of these intervals should contain θ.

- **Bayesian**
  - Credible Interval
  - Probability of 95% that the credible interval contains θ.

KS - STK - IPB

Pendekatan Bayesians Menggunakan Prior

- Prior: non-experimental knowledge.
- Examples from Berger's book:
  - A man tasting tea: tea poured into milk or vice versa.
  - A man predicts coin toss.
  - Each of the 10 times, they got it right.
  - Inference may differ due to different prior beliefs.

KS - STK - IPB

Criticisms of Bayesianism

- **Use of prior: not objective; different people come to different conclusions; this is not scientific.**
- **Answers:**
  - Absolute "objectivity" does not exist.
  - This provides a formal way of incorporating subjective belief.
  - To be objective: use noninformative priors.
  - Evidence-based: when data accumulate, all rational individuals will come to the same conclusion.

KS - STK - IPB

Bayesian Machinery

data

prior belief ------------------------------→ posterior belief

likelihood

prior distribution ----------------------→ posterior

$L(\theta)=P(X|\theta)$

$p(\theta)$ -------------------------------------→ $p(\theta|X)$

Ringkasan: perbedaan utama antara pendekatan *classical* (*frequentist*) dan Bayesian

Approach	What does it assume?	What does it ask?	What do you need?	What does it give?	Main advantages	Main drawbacks	When to use it
Frequentist	Parameters that you estimate are fixed and are a single point	Is the hypothesis true or false?	A stopping criterion A predetermined experimental design	A point estimate (*p*-value)	Simple and easy to use Widely accepted	Significance depends on the sample size Only gives you a yes/no answer	When you have a large amount of data
Bayesian	There's a probability distribution around the parameters	What is the probability of the hypothesis given the data?	A prior Any data set	A probability for or against the hypothesis	You get strength of evidence Can update it with new information	Priors can be subjective or arbitrary **Requires advanced statistics**	When have limited data When you have priors **When you have computing power**

Ilustrasi Statistika Bayesian:

- Ini merupakan ilustrasi sederhana untuk menunjukkan dasar-dasar cara kerja statistika Bayesian.
- Dimulai dengan beberapa nilai peluang di awal (*prior*) dan bagaimana persisnya nilai ini diperbarui saat telah mendapatkan lebih banyak informasi.
- Nilai peluang yang diperbarui ini disebut sebagai peluang *posterior*.
- Untuk membantu memperjelas step-by-step, akan digunakan tabel yang biasanya disebut *Bayes' Box* untuk menghitung peluang posterior.

KS - STK - IPB

Penggunaan Bayes' Box:

a posterior probability.

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)}$$

- $P(H|D)$ is the **posterior probability**. It describes how certain or confident we are that hypothesis H is true, given that we have observed data D. Calculating posterior probabilities is the main goal of Bayesian statistics!
- $P(H)$ is the **prior probability**, which describes how sure we were that H was true, before we observed the data D.
- $P(D|H)$ is the **likelihood**. If you were to assume that H is true, this is the probability that you would have observed data D.
- $P(D)$ is the **marginal likelihood**. This is the probability that you would have observed data D, *whether H is true or not.*

Bayes' Box:

Hypotheses	prior	likelihood	prior × likelihood	posterior
H_1	$P(H_1)$	$P(D\|H_1)$	$P(H_1) \times P(D\|H_1)$	$P(H_1\|D)$
H_2	$P(H_2)$	$P(D\|H_2)$	$P(H_2) \times P(D\|H_2)$	$P(H_2\|D)$
...	...	...	...	...
Totals:	1		$P(D)$	1

Ilustrasi Statistika Bayesian:

Suppose there are two balls in a bag. We know in advance that at least one of them is black, but we're not sure whether they're both black, or whether one is black and one is white. These are the only two possibilities we will consider. To keep things concise, we can label our two competing hypotheses. We could call them whatever we want, but I will call them `BB` and `BW`. So, at the beginning of the problem, we know that *one and only one* of the following statements/hypotheses is true:

`BB`: Both balls are black
`BW`: One ball is black and the other is white.

Ilustrasi Statistika Bayesian:

BB: Both balls are black
BW: One ball is black and the other is white.

Suppose an experiment is performed to help us determine which of these two hypotheses is true. The experimenter reaches into the bag, pulls out one of the balls, and observes its colour. The result of this experiment is (drumroll please!):

D: The ball that was removed from the bag was black.

We will now do a Bayesian analysis of this result.

Ilustrasi Statistika Bayesian:

Bayes' Box

A Bayesian analysis starts by choosing some values for the prior probabilities. We have our two competing hypotheses BB and BW, and we need to choose some probability values to describe how sure we are that each of these is true. Since we are talking about two hypotheses, there will be two prior probabilities, one for BB and one for BW. For simplicity, we will assume that we don't have much of an idea which is true, and so we will use the following prior probabilities:

$$
\begin{aligned}
P(\texttt{BB}) &= 0.5 \\
P(\texttt{BW}) &= 0.5.
\end{aligned}
$$

The first column of a Bayes' Box is just the list of hypotheses we are considering.

Hypotheses	prior	likelihood	prior × likelihood	posterior
BB	0.5			
BW	0.5			
Totals:	1			

Ilustrasi Statistika Bayesian:

D: The ball that was removed from the bag was black.

The likelihood for a hypothesis is the probability that you would have observed the data, if that hypothesis were true. The values can be found by going through each hypothesis in turn, imagining it is true, and asking, "What is the probability of getting the data that I observed?".

Likelihood = Probability(Data | Hypothesis)

Hypotheses	Possible Data	Probability
BB	Black Ball	1
	White Ball	0
BW	Black Ball	0.5
	White Ball	0.5

Likelihood

This table demonstrates a method for calculating the likelihood values, by considering not just the data that actually occurred, but all data that might have occurred. Ultimately, it is only the probability of the data which actually occurred that matters, so this is highlighted in blue.

Ilustrasi Statistika Bayesian:

To find the posterior probabilities, we take the `prior` $\times$ `likelihood` column and divide it by its sum, producing numbers that do sum to 1. This gives us the final posterior probabilities, which were the goal all along. The completed Bayes' Box is shown below:

Hypotheses	prior	likelihood	h = prior × likelihood	posterior
BB	0.5	1	0.5	0.667
BW	0.5	0.5	0.25	0.333
Totals:	1		0.75	1

We can see that the posterior probabilities <u>are not the same as</u> the prior probabilities, because we have more information now! The experimental result made BB a little bit more plausible than it was before. Its probability has increased from 1/2 to 2/3.

Likelihood = Probability(Data = Black | BB) = 1

Likelihood = Probability(Data = Black | BW) = 0.5

$$\boxed{0.667 = \frac{0.5}{0.75}}$$

Materi Praktikum

Selesaikan problem ini melalui Bayes' Box untuk memperoleh posterior: (a).Secara manual; (b).Menggunakan Program R.

You move into a new house which has a phone installed. You can't remember the phone number, but you suspect it might be 555-3226 (some of you may recognise this as being the phone number for Homer Simpson's "Mr Plow" business). To test this hypothesis, you carry out an experiment by picking up the phone and dialing 555-3226.

If you are correct about the phone number, you will definitely hear a busy signal because you are calling yourself. If you are incorrect, the probability of hearing a busy signal is 1/100. However, all of that is only true if you assume the phone is working, and it might be broken! If the phone is broken, it will always give a busy signal.

When you do the experiment, the outcome (the data) is that you do actually get the busy signal. The question asked us to consider the following four hypotheses, and to calculate their posterior probabilities:

Hypothesis	Description	Prior Probability
H_1	Phone is working and 555-3226 is correct	0.4
H_2	Phone is working and 555-3226 is incorrect	0.4
H_3	Phone is broken and 555-3226 is correct	0.1
H_4	Phone is broken and 555-3226 is incorrect	0.1

Pustaka

1. Reich BJ dan Ghosh SK. (2019). *Bayesian Statistical Methods*. Taylor and Francis Group.

2. Bolstad WM dan Curran JM. (2017). *Introduction to Bayesian Statistics 3th Edition*. John Wiley and Sons, Inc.

3. Lee, Peter M. (2012). *Bayesian Statistics, An Introduction 4th Edition*. John Wiley and Sons, Inc.

4. Albert, Jim. (2009). *Bayesian Computation with R*. Springer Science and Business Media.

5. Ghosh JK, Delampady M, dan Samanta T. (2006). *An Introduction to Bayesian Analysis, Theory and Methods*. Springer Science and Business Media.

KS STK IPB

Terima Kasih