Bayesian Statistics W. Bolstad - Book Notes

Max Lang

3/20/2022

Logic, Probabiltiy and Uncertanty

Deductive Logic

- A logical process for determining the truth of a statement from knowing the truth or falsehood of other statements that the first statement is a consequence of.
- Deduction works from **the general to the particular**
- Deductions do not have the possibility for error

Example Socrates was a man (premise)

All men are mortal (premise).

Socrates was mortal (conclusion)

Inductive Logic

- A process based on *plausible* reasoning, for inferring the truth of the statement from knowing the truth or falsehood of other statements which are consequences of the first statement
- Statistical Inference is a inductive process, hence there is always the prossibility of error when making an inference

Example Socrates was Greek (premise).

Most Greeks eat fish (premise).

Socrates ate fish (conclusion).

Probability

- Plausible reasoning should be based on the rules of probability to be consistent. They are:

  Probability of an event is a nonnegative number

  Probability of the sample space (universe) equal 1

  The probabiltiy is additive over disjoint events

- Bayes' theorem is the key to Bayesian Statistics

$$P(B_i \mid A) = \frac{P(B_i) \times P(A \mid B_i)}{\sum_j P(B_j) \times P(A \mid B_j)}$$

This follows from the definition of conditional probability. $P(B_i)$ is called the **prior probability** of event B_i and $P(B_i \mid A)$ is called the **posterior probability** of event $P(B_i \mid A)$

- Bayes' Theorem can be summarized by saying that the posterior probability is the prior times the likelihood divided by the sum of the prior times likelihood (basically scalled back to 1)

- The Bayesian universe has two dimensions

1. The sample space forms the observable (horizontal) dimension
2. The parameter space is the unobservable (vertical) dimension

In Bayesian statistics the probabilities are defined on both dimension of the Bayesian universe.

Bayesian Inference for Discrete Random Variables

Like stated before the Bayesian Universe has two dimensions. The vertical dimension is the parameter space and is unobservable. The horizontal dimesnion the sample space and we observe which value occurs.

- The reduced universe is the column for the observed value.

The simplified table for finding posterior distributuion given Y=2

μ	prior	likelihood	prior $\times$ likelihood	posterior
1.0	$\frac{1}{6}$	$\frac{1.0^2 e^{-1.0}}{2!} = .1839$	.0307	$\frac{.0307}{2473} = .124$
1.5	$\frac{1}{3}$	$\frac{1.5^2 e^{-1.5}}{2!} = .2510$	.0837	$\frac{.0837}{.2473} = .338$
2.0	$\frac{1}{3}$	$\frac{2.0^2 e^{-2.0}}{2!} = .2707$	.0902	$\frac{.0902}{.2473} = .365$
2.5	$\frac{1}{6}$	$\frac{2.5^2 e^{-2.5}}{2!} = .2565$	.0428	$\frac{.0428}{.2473} = .173$
marginal $P(Y=2)$	.2473		1.000	

One can see that for discrete observations the posterior probabilities are found by multiplying the prior x liklihood and then dividing by their sum (norming back to 1, because it should be a probability / normed measure)

- Neither multiplying the prior nor the likelihood will change the result, because only the relative wheights are important

Bayesian Inference for Binomial Proportion

- The key relationship is posterior $\propto$ prior $\times$ likelihood. This gives us the shape of the posterior density. We must find the constant to divide this by to make it a density, eg., integrate to 1 over its whole range.
- The constant we need is $k = \int_0^1 g(\pi) \times f(y \mid \pi) d\pi$. In general, this integral does not have a closed form, so we have to evaluate it numerically.
- If you have some prior knowledge, and you can't find a member of the conjugate family that matches it, you can construct a discrete prior at several values over the range and interpolate between them to make the prior continuous. Of course, you may ignore the constant needed to make this a density, since any constant gets cancelled out by when you divide by $\int$ prior $\times$ likelihood to find the exact posterior.
- The main thing is that your prior must have reasonable probability over all values that realistically are possible. If that is the case, the actual shape doesn't matter very much. If there is a reasonable amount of data, different people will get similar posteriors, despite starting from quite different shaped priors.
- The posterior mean is the estimate that has the smallest posterior mean square. This means that, on average (with respect to posterior), it is closer to the parameter than any other estimate. In other words, given our prior belief and the observed data, the posterior mean will be, on average, closer to the parameter than any other estimate. It is the most widely used Bayesian estimate because it is optimal post-data.
- A $(1-\alpha) \times 100\%$ Bayesian credible interval is an interval that has a posterior probability of $1-\alpha$ of containing the parameter.

Comparing Bayesian and Frequentist Inferences for Proportion

Frequentist Interpretation of probability and parameters

- The parameter is a **fixed but unknown** constant. This doesn't allow any probability distribution associated with it.

- The only probability considered is the **probability of the random sample of size n, given the parameter**. This explains how the randoms sample varies over all possible random samples given the fixed but unknown parameter value.

Sampling Distribution of Statistic

Let $Y_1, \ldots, Y_n$ be a random sample from a distribution that depends on a parameter θ. Suppose a statistic S is calculated from the random sample. This statistic can be interpreted as a random variable, since the random sample can vary over all possible samples. Calculate the statistic for each possible random sample of size n. The **distribution of these values is called the sampling distribution of the statistic**. Of course, the *sampling distribution also depends on the unknown value of the parameter* θ. We will write this sampling distribution as

$$f(s \mid \theta)$$

However, we must remember that in frequentist statistics, the **parameter θ is a fixed but unknown constant**, not a random variable. The **sampling distribution measures how the statistic varies over all possible samples**, given the unknown fixed parameter value. This **distribution does not have anything to do with the actual data that occurred**. It is the distribution of values of the statistic that could have occurred, given that specific parameter value. Frequentist statistics uses the sampling distribution of the statistic to perform inference on the parameter.

This contrasts with **Bayesian statistics** where the complete inference is the posterior distribution of the parameter given the actual data that occurred:

$$g(\theta \mid \text{ data })$$

Any **subsequent Bayesian inference** such as a Bayesian estimate or a Bayesian credible interval **is calculated from the posterior distribution**. Thus the estimate or the credible interval depends on the data that actuallv occurred.

Frequentist and Bayesian Inference approaches

- The posterior distribution of the parameter given the data is the **entire** inference from a Bayesian perspective
- Under frequentist perspective there are specific inference about the parameter: *point estimation, confidence intervals and hypothesis tests*
- Frequentists consider the parameter fixed but unknown constant, the only kind of probability allowed is long-rong relative frequency
- The sampling distribution of a statistic is its distribution over all possible random samples given the fixed parameter value. Frequentist statistics is based on the sampling distribution
- All Probabilities calculated using the sampling distribution are *pre-data* because they are based on all possible random samples not the specific random sample obtained
- Bayesian estimators are often better than frequentist estimators even when judged by frequentist criteria such as mean squared error, this can be seen by performing a *pre-posterior* analysis

Confidence intervals (Frequentist) A $(1-\alpha) \times 100\%$ confidence interval for a parameter θ is an interval (l, u) such that

$$P(l \leq \theta \leq u) = 1 - \alpha,$$

where the probability is found using the **sampling distribution of an estimator for θ*.

Correct interpretation:

$(1-\alpha) \times 100\%$ of the random intervals calculated this way do contain the true value. When the actual data are put in and the endpoints calculated, there is nothing left to be random. The endpoints are numbers; the parameter is fixed but unknown.

We say that we are $(1-\alpha)\times 100\%$ confident that the calculated interval covers the true parameter.

The confidence comes from our belief in the method used to calculate the interval. **It does not say anything about the actual interval we got for that particular data set.**

Credible intervals (Bayesian) A $(1-\alpha)\times 100\%$ Bayesian credible interval for θ is a range of parameter values that has posterior probability $(1-\alpha)$. **Probability statement about the parameter is allowed!** This is again obtained from the posteriori distribution.

Hypothesis testing

- **Frequentist approach** divides the sample space in a rejection region and an acceptance region such that the probability the test statistic lies in the rejection region if the null hypothesis is true is less than the level of significance α.
- Or we could calculate the p-value. If the p-value< a, we reject the null hypothesis at level α.
- The p-value is not the probability the null hypothesis is true. Rather, it is the probability of observing what we observed, or even something more extreme, given that the null hypothesis is true.
- With the **Bayesian approach** we can test one-sided hypothesis by computing the posterior probability of the null hypothesis. The probability is found by integrating the posterior density over the null region. If this probability is less thant the level of significance then we reject the null hypothesis
- We cannot test **two-sided hypothesis** by intergrating the posterior probability over the null region, because with a **continous prior** the prior probability of a point null hypothesis is zero, so posterior will also be zero. Instead we **test the credibility of the null value by observing wheter or not it lies within the Bayesian credible interval**. If it does the null value remains credible and we can't reject it.

Comparing Bayesian and Frequentist Inferences for Mean

- When we have prior information on the values of the parameter that are realistic, we can find a prior distribution so that the mean of the posterior distribution of μ (the Bayesian estimator) has a smaller mean squared error than the sample mean (the frequentist estimator) over the range of realistic values. This means that on the average, it will be closer to the true value of the parameter.
- A confidence interval for μ is found by inverting a probability statement for $\bar{y}$, and then plugging in the sample value to compute the endpoints. *It is called a confidence interval because there is nothing left to be random*, so no probability statement can be made after the sample value is plugged in.
- *The interpretation of a* $(1-\alpha)\times 100\%$ *frequentist confidence interval for* μ *is that* $(1-\alpha)\times 100\%$ *of the random intervals calculated this way would cover the true parameter*, so we are $(1-\alpha)\times 100\%$ confident that the interval we calculated does.
- A $(1-\alpha)\times 100\%$ *Bayesian credible interval is an interval such that the posterior probability it contains the random parameter* is $(1-\alpha)\times 100\%$.
- This is more useful to the scientist because he/she is only interested in his/her particular interval.
- The $(1-\alpha)\times 100\%$ *frequentist confidence interval for* μ *corresponds to the* $(1-\alpha)\times 100\%$ *Bayesian credible interval for* μ *when we used the "flat prior." So, in this case, frequentist statisticians can get away with misinterpreting their confidence interval for* μ *as a probability interval.*
- In the general, misinterpreting a frequentist confidence interval as a probability interval for the parameter will be wrong.
- Hypothesis testing is how we protect our credibility, by not attributing an effect to a cause if that effect could be due to chance alone.
- Frequentist hyptohesis tests are based on the sample space

Robust Bayesian Methods

- If the prior places high probability on values that have low likelihood, and low probability on values that have high likelihood, the posterior will place high probability on values that are not supported either by the prior or by the likelihood. This is not satisfactory.
- This could be caused by a miss-specified prior
- Using mixture priors protects against this possible misspecification of the prior. We use mixtures of conjugate priors. We do this by introducing a mixture index random variable that takes on the values 0 or 1. If the likelihood has most of its value far from the original prior, the mixture posterior will be close to the likelihood. This is a much more satisfactory result. When the prior and likelihood are conflicting, we should base our posterior belief mostly on the likelihood, because it is based on the data. Our prior was based on faulty reasoning from past data that failed to note some important change in the process we are drawing the data from.