Foundations of Probability

Marco Ramoni
Children's Hospital Informatics Program
Division of Health Sciences and Technology
Harvard Medical School and
Massachusetts Institute of Technology

statistical recipes

bayesians

Profile: Hates the cookbook style of classical statistics.
Tools: Bayes Theorem and Total Probability Theorem.
Main Belief: Everything has a probability.

- Unknown events have a probability: inference is updating prior knowledge (even when it is knowledge of being ignorant).
- Parameters have a probability: the parameters of interest are random variables with a probability distribution.
- Hypothesis have a probability: we can compute the probability of an hypothesis rather than the occurrence in 1000 experiments.
- Statistical models have a probability: fitting a model means to find the most probable model given the data.

definitions

Data: n cases \( y = (y_1, \ldots, y_n) \).
Parameters: the object of inference \( \theta \).
Inference: use the data \( y \) to refine (or infer) \( \theta \).

Frequentists: likelihood function \( L(\theta) = p(y|\theta) \) is the function of interest and the maximum likelihood estimator is the ratio between the positive events and all the events.
Prior: It makes no sense, for a frequentist, to talk before seeing any data - all the information is in the data.

Rev. Thomas Bayes

Name: Thomas Bayes.
Death: 1761, Tunbridge, Kent.
Job title: Presbyterian minister.
Publications: 3 (1 posthumous).
Trivia: Unrelated to Bayesians.
Bayes proved a basic result in probability theory:

\[ p(h|e) \propto p(e|h)p(h) \]

The probability of \( h \) in the context of \( e \), i.e., rescaled by \( e \):

\[ p(e|h)p(h) = p(h) p(e) \]

To “go back” we need to expand \( p(e|h) \) by \( h \) and then rescale the result by \( e \).

We can include a prior on \( h \).

\[ p(h|e) = \frac{p(e|h)p(h)}{p(e)} \]

### The Single Case Problem

**Problem**: A dataset comprises one (1) coin flip (head).

**Question**: What is the probability of coin=H?

**Frequentist solution**: \( 1/1 = 1 \). The probability is 100%.

**Bayesian solution**: Flips follow a \( \beta \) distribution: \( \beta(H,T) \) and \( \mu(H) = H/H+T \).

We assume ignorance as \( \beta(1,1) \).

We observe coin H and hence \( \beta(2,1) \).

\( \mu(H) = 2/(2+1) = 2/3 \) and hence the probability is 66%.

**Prior**: I can include formally information about a bias.

**Note**: Frequentist and Bayesian agree, asymptotically.

**Inference**

- The probability distribution of your parameter: \( p(\theta|I_0) \), a random variable with a probability distribution.
- The available information changes the conditional distribution of \( \theta \) through Bayesian Theorem:
  \[ p(\theta|y, I_0) \propto p(y|\theta, I_0)p(\theta|I_0) \]
- Distribution \( p(\theta|y, I_0) \) is compared via Total Probability Theorem:
  \[ p(y|I_0) = \int p(y|\theta, I_0)p(\theta|I_0)d\theta \]
- and it is also called the marginal density of data, to stress the fact that it is no longer conditional on \( \theta \).

**Prediction**

- The posterior distribution is the result of Bayesian inference. This distribution is then used to find a point estimate of the parameters, to test a hypothesis or, in general, to find credibility intervals, or to predict future data \( y \), conditional on the posterior information \( I_1 \):
  \[ p(y|I_1) \equiv \int p(y|\theta, I_1)p(\theta|I_1)d\theta \]
- This is called the predictive distribution.
- The process of updating/prediction is iterative and incremental.

**Prior Distributions**

- The use of previous posterior as prior allows the principled integration of results from multiple studies.
- Priors can be elicited from humans or can encode background information in a principled way (not by tricking the design of the experiment).
- Priors can be set to be non informative by maximizing the entropy of the prior distribution.
- Priors should be conjugate to the posterior (i.e. have the same distributional form).
- Non informative priors can present problems of conjugacy because the posterior does not necessarily have the same form as the prior.
Bayesian inference returns the posterior density \( p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \).

For a parameter \( \theta \), we can compute the marginal posterior distribution:
\[
p(\theta|y) = \int p(\theta|j_i,y) \, d\theta_i,
\]

A standard point estimate of \( \theta \) is the posterior expectation:
\[
E(\theta|y) = \int \theta \, p(\theta|y) \, d\theta.
\]

The posterior mode is based on a principle similar to Maximum Likelihood:
\[
\hat{\theta} = \arg \max \, p(\theta|y).
\]

Since \( p(\theta|y) \) is proportional to \( p(y|\theta)p(\theta) \), the vector of posterior modes \( \hat{\theta} \) maximizes the augmented likelihood \( p(y|\theta)p(\theta) \), and it is therefore called the Generalized Maximum Likelihood Estimate (GML estimate) of \( \theta \).

Model averaging:
1. Compute the probability of each assumption \( D_j \) and \( D_i \), given the data \( \mathcal{D}_j(y) \) and \( \mathcal{D}_i(y) \).
2. Predict the value of the variable under each assumption:
\[
p(y|D_j\mathcal{D}_j) \quad \text{and} \quad p(y|D_i\mathcal{D}_j).
\]
3. Weight each prediction by the probability of its assumption:
\[
p(y|\mathcal{D}_j) = \frac{p(y|D_j\mathcal{D}_j) \times p(D_j\mathcal{D}_j)}{p(D_j\mathcal{D}_j) \times p(D_j\mathcal{D}_j) + p(D_i\mathcal{D}_j) \times p(D_i\mathcal{D}_j)}
\]

Credibility Region: Posterior expectation and mode provide simple summaries, but we can use them to compute that \( \theta \) is in some given region \( R \) or to find a region \( R \) that contains \( \theta \) with probability \( 1 - \alpha \).

Posterior Highest Density Region: When \( R \) is the region of smallest volume, it is also called the Posterior Highest Density (PHD) region.

Credibility vs Confidence: In some cases, credibility intervals are identical to the classical \((1-\alpha)\% \) confidence interval. For instance, for the mean of a normal population when the variance is known.

Confidence Intervals: The frequentist interpretation of the \((1-\alpha)\% \) confidence interval is based on the repeatability of the sampling process: if we could take \( n \) samples we would expect that in \((1-\alpha)\% \) of cases the interval contains the true value of \( \theta \).

Credibility Intervals: The \((1-\alpha)\% \) PHD interval returned by the Bayesian method is a credibility statement, conditional on the information \( I_i \). With probability \((1-\alpha)\% \), \( \theta \) belongs to the interval.

Hypothesis testing:

Scoring: The Bayesian approach to hypothesis testing is based on the computation of the conditional probability of a hypothesis \( H \) given the available information and the data \( p(H|y, I_i) \).

Testing: hypothesis testing is formulated by setting the null hypothesis \( H_0: \theta < \theta_0 \) and the alternative hypothesis \( H_1: \theta > \theta_0 \), with prior probabilities \( p(H_0|I_0) + p(H_1|I_0) = 1 \).

Prior Density: By Total Probability Theorem, the prior density of \( \theta \) is:
\[
p(\theta|I_0) = p(\theta|H_0)p(H_0|I_0) + p(\theta|H_1)p(H_1|I_0)
\]

where \( p(\theta|H_0) \) and \( p(\theta|H_1) \) are the prior densities of \( \theta \), conditional on hypothesis.

Odds: The sample information is used to compute from prior odds
\[
p(p(H_0|I_0)/p(H_1|I_0)),
\]

the posterior odds in favor of \( H_0 \) as
\[
p(p(H_0|y|I_0)/p(H_1|y|I_0)) = p(p(y|H_0)p(H_0|I_0))/p(p(y|H_1)p(H_1|I_0))
\]

Bayes factors: The ratio \( p(y|H_0)p(H_0|I_0)/p(y|H_1)p(H_1|I_0) \) is called the Bayes factor, and when the prior probabilities of \( H_0 \) and \( H_1 \) are equal, the Bayes factor determines the decision rule.

Likelihood ratio test: The classical likelihood ratio test is a special case of this that occurs when the hypothesis \( H_1 \) and \( H_2 \) are simple, that is, they specify completely the parameters values.

Model averaging:

The Bayesian approach is more general and allows us to get a straight answer for both hypothesis, just computing the posterior probability.

We can apply the same principle to fitting a model.

We can consider each model (or decomposable part of it) as an hypothesis and compute its posterior probability given the data.

The advantage is that we can compute how much more probable is a model given the data compared to an alternative model.

As for prediction, I can do even better, and use model averaging.