Foundations of Statistical Inference

Statistical Decision Theory and Bayesian Analysis

Author: Massimiliano Pronesti

Instructor: **Motonobu Kanagawa**

November 30, 2021

1 Introduction to Decision Theory

We refer to **statistical decision theory** as the branch of statistics concerned with the problem of making decisions in the presence of statistical knowledge. Differently from classical statistics - which is mainly directed towards the use of sample information -, decision statistics takes **also**[1] into account other relevant aspects of the problem, which usually prove crucial in the decision process:

- **knowledge of possible consequences** of the decision, typically quantified from the loss (or the gain) arising from each possible decision;
- **prior information**, typically arising from past experience of similar situations

This report introduces the main concepts and tools of statistical decision theory as described in the introduction of *James O Berger. Statistical Decision Theory and Bayesian Analysis. Springer 1985.*

2 Key elements of decision statistics

2.1 Loss function

A key element of the decision theory is the loss function, which defines the result of taking an action a given the unknown quantity θ affecting the decision process, commonly referred as **the state of nature** .

For technical convenience, the loss function satisfies the following chain of inequalities

$$L(\theta, a) \geq -K > -\infty$$

assuming, of course, that $L(\theta, a)$ is defined for all $(\theta, a) \in \Theta \times \mathscr{A}$, being Θ the parameter space and $\mathscr{A}$ the set of all possible actions . Nevertheless, a critical **caveat** is that, in many problems, loss function and prior information might not be well defined or even nonunique (at the time of decision making).

2.2 Bayesian expected loss

In light of the above caveat, the natural way of proceeding consists in considering the **expected loss** of making a decision and then picking the "optimal" one according to it.

The **Bayesian expected loss** refers to the uncertainty in θ, treating it as a random quantity associated to a probability distribution $\pi^*(\theta)$ [2]

$$\rho(\pi^*, a) = E^{\pi^*} L(\theta, a) = \int_\Theta L(\theta, a) dF^{\pi^*}(\theta)$$

2.3 Frequentist risk and its implications

The so-called "classical statistics" employs a rather different approach towards loss, i.e. the **risk function** given a decision rule $\delta(X)$, trying to minimize that:

$$R(\theta, \delta) = E_\theta^X [L(\theta, \delta(x))] = \int_{\mathscr{X}} L(\theta, \delta(x)) dF^X(x|\theta)$$

Notice that, in a no data problem, $R(\theta, \delta) \equiv L(\theta, a)$. Moreover, differently from Bayesian expected loss - which is a number -, the risk is a function on Θ which drives us to the following problem: **being θ unknown, it's not clear what "small" risk means**. Hence, we need to define a (partial) ordering for decision rules to introduce a choice criteria: given two decision rules δ_1, δ_2, we say δ_1 is R-better than δ_2 if

$$R(\theta, \delta_1) \leq R(\theta, \delta_2), \ \forall \theta \in \Theta$$

with strict inequality holding for some θ. If no R-better decision rule exists, a decision rule δ is called *admissible* (the definition of *inadmissible* is dual). Nevertheless, the choice of a decision rule is quite more complicated than what we just introduced: there is usually a large class of admissible decision rules for a large problem, which, typically, have risk functions being better than the others only locally.

[1] Sample information is still considered

[2] We use π^* rather than π as the latter one typically refers to the initial prior distribution, whereas the former one is usually the final posterior distribution of θ (after seeing the data)

3 Decision principles

This far, we introduced how to perform a statistical analysis. We now describe the major methodologies to actually make a decision:

- **conditional Bayes decision principle**: choose an action $a \in \mathscr{A}$ such that it minimizes the expected loss $\rho(\pi^*, a)$. Such an action will be called a **Bayes action** and will be denoted a^{π^*}.
- **Bayes risk principle:** as remarked in **section 2.3**, Bayes risk is a number, hence we define a decision rule minimizing it. Notice that, in a no-data problem, the risk is nothing but the loss which implies that this principle gives the same answer of the conditional Bayes decision one;
- **invariance principle:** if two problems have same formal structure, then same decision must be taken;
- **minimax principle:** this principle is often called for consideration of randomized decision rules. According to it, a rule δ^{*M} is a *minmax decision rule* if it minimizes $\sup_\theta \ R(\theta, \delta^*)$ among all randomized rules in $\mathscr{D}^*$ i.e.

$$\sup_{\theta \in \Theta} \ R(\theta, \delta^*) = \inf_{\delta^* \in \mathscr{D}^*} \sup_{\theta \in \Theta} \ R(\theta, \delta^*)$$

4 Foundations

In this last section, we're gonna expand even further the conditional versus frequentist controversy - crucial in statistics - as well as discuss some incorrect usages of classical inference in decision problems.

4.1 Misused inference procedures

A common mistake while analyzing the null hypothesis is to run tests on it without getting deep into its purposes, with the consequence of often discarding a "useful" hypothesis. An interesting example involves **Kepler's law** of planetary motion, which is in fact an approximation of the reality (like every physics model). Nonetheless, if blindly statistically tested with too accurate data, such a rather essentially correct model would be rejected. In other words, a **statistically significant difference** between the true model and the null hypothesis might be **unimportant practically**.

4.2 Frequentist Perspective

The frequentist approach aims at producing measures which don't depend upon θ, or any prior knowledge about it. It considers a procedure $\delta(x)$ and a criterion function $L(\theta, \delta(x))$ to determine a quantity $\bar{R}$ s.t. repeated use of δ in different problems would yield average the long run performance of at least $\bar{R}$.

4.3 Conditional Perspective

The conditional approach mainly focuses on the performance of a procedure $\delta(x)$ on *actual* data while the overall performance is considered almost secondary. This might definitely lead to different results with respect to the frequentist approach introduced above. In order to distinguish between those two categories of results, Savage (1962) introduced the terms *initial precision* and *final precision* to, respectively, refer to frequentist and conditional measures. In fact, before seeing the data we can only use frequentist analysis, while after observing it we can make a more precise one.

4.4 Likelihood Principle

This principle makes explicit the natural conditional idea that only the actual observed x should be relevant to conclusions or evidence about θ. The key concept in the Likelihood Principle is that of the likelihood function $l(\theta) = f(x|\theta)$ which, intuitively, takes its name from the fact that an input θ for which $l(\theta)$ is large is more "likely" to be the true θ than a θ for which $l(\theta)$ is small. Notice that two likelihood values $l_1(\theta), l_2(\theta)$ contain the same information about θ if they are proportional[3].

It is important to highlight, however, that the likelihood function contains **all experimental information** about θ, **but not all the information in general**: there could be some other important statistical information, as the prior information or considerations of loss.

Limitations: this principle is not enough to decide, at a given stage of our analysis, whether or not to take another observation. Besides, it is not enough to predict a future value of X and does not indicate how the likelihood function should be used to make decisions or inferences about θ.

[3] It's implicit that l_1, l_2 must refer to the same θ