Analysis of the probabilistic information of random sets

One of the problems that can be encountered during the statistical analysis of data is the existence of some imprecision in the observation process. This can appear in both the recollection and in the transmission of the data, and usually passes on the conclusions that are derived from them. Because of this, there have appeared several models in the literature that permit working with imprecise information and quantifying the influence that this imprecision has in the inferences we make. We can think for instance of the models of Robust Statistics ([3, 4]), which study when the conclusions are not affected in a drastic way by small deviations in the observations, or by the presence of outliers; we can also consider random sets ([5, 8]) and fuzzy random variables ([9]), where we establish a classification in the sets of possible values for each data, showing which of them are more plausible, taking into account the available information. And we can also use the models of sets of probabilities ([7, 10]), non-additive measures ([2]), etc, which summarize the available information when the imprecision does not allow to determine a unique probability measure to represent that information.

In this work, we consider one of the models listed above: random sets. These are multi-valued mappings satisfying some measurability condition, and constitute a generalization of the notion of random variable where the image of any element of the initial space is a subset of the final space. We are going to interpret them following the ideas considered by Kruse and Meyer in [6]: we will assume the existence of a random variable whose values cannot be observed with precision and will give, for any element of the initial space, a subset of the final space which includes with certainty its image by the random variable.

Obviously, if our observations are not precise we cannot expect precision from the conclusions we derive from them; but it is nevertheless worth trying to quantify which is the available information about the characteristic of interest. In the case of random sets, and under our interpretation, there are basically two ways of modelling the information about the probability distribution induced by the "original" random variable: on the one hand, we can consider the probability distributions induced by the so-called *measurable selections* of the random set. These are the random variables whose values are included in the images of the random set. Hence, the "true" random variable, which models the behaviour of the characteristic of interest, is one of the measurable selections of the random set, and its distribution belongs to the class of distributions of the measurable selections. This is the most precise model at our disposal. The second model is based in the *upper* and *lower* probabilities of the random set, which were defined by Dempster in [1]. It can be checked that the distribution of any measurable selection (and in particular that of the imprecisely observed random variable) belongs to the class of the probabilities bounded by these two functions. This is a convex set, and possesses certain features which lacks the class of distributions of the measurable selections. Moreover, it is uniquely determined by the upper and lower probabilities, which are sometimes characterized by their values on certain sets. All these reasons make it more desirable to work with this second set of probabilities. However, in many cases it includes strictly the class of the distributions of the measurable selections, and consequently its use could add imprecision to the one we already have in our experiment.

The goal of this thesis is to determine under which conditions it is advisable to use the model based in the upper and lower probabilities to summarize the available information. We will investigate its relationship with the class of the distributions of the measurable selections, studying when they coincide and the magnitude of their difference.

Summary

In Chapter 1 we introduce the concepts and notations that will be used throughout the work, and present the problem that will be studied.

First, we compare the information provided by the upper and lower probabilities of the random set and the class of distributions of the measurable selections about the probability that the original random variable takes values in some set of the final σ-field. That is, we investigate whether the upper and lower probabilities of a set are the most precise bounds one can give, taking into account the available information, of its probability by the original random variable. We relate this problem to the existence of measurable selections and to the inner approximations of the upper probability. This is done in Chapter 2.

Next, we study if the upper and lower probabilities keep all the information about the original random variable. In order to do this, we focus previously in a number of problems: first, we investigate the properties, in the topology of the weak convergence, of the class of distributions of the measurable selections and of those bounded by the upper probability. These properties are interesting for us because they allow us later to relate the closures of these two sets. In Chapter 3 we study under which conditions these sets are closed or compact under the aforementioned topology, and we also study their convexity. We prove that these properties are sometimes related to the outer continuity of the upper probability and to its approximation by its values in the open sets, and investigate also these two properties.

Chapter 4 is devoted to the study of the extreme points of the class of the probabilities dominated by the upper probability. This, together with the relationship between the extreme points and the distributions of the measurable selections, allows us later to establish relationships between the the class of the distributions of the measurable selections and those bounded by the upper probability. In order to characterize the extreme points, we embed first the probabilities in a locally convex and Hausdorff linear topological space, and such that the subspace topology coincides with the weak convergence topology. Then, we study the characteristics of the extreme points, connecting our results with the ones established for the finite case.

The results from Chapters 2, 3 and 4 allow us to study in Chapter 5 the relationships between the class of the distributions of the measurable selections and those bounded by the upper probability. We obtain sufficient conditions for the equality between the closures of these two sets of probabilities, and also study if under those conditions the two sets coincide.

Chapter 6 contains a more detailed study on three types of random sets which are very common in the literature: random sets on finite spaces, random intervals and consonant random sets. In the first case, we see that the study benefits from a number of simplifications, which are the cause of some additional properties. In the case of random intervals, we study first random closed intervals and later random intervals with one of their extremes open. We study the relationships existing between them. Concerning consonant random sets, we study the different degrees of consonance and prove that they are related to possibility measures.

Finally, Chapter 7 contains our conclusions about the problem at hand and points several open lines of research.

References

[1] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[2] D. Denneberg. *Non-additive measure and integral.* Kluwer Academic Publishers, 1994.

[3] F. Hampel. A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 42:1887–1896, 1971.

[4] P. J. Huber. *Robust Statistics.* Wiley, New York, 1981.

[5] D. G. Kendall. Foundations of a theory of random sets. In E. F. Harding and D. G. Kendall, editors, *Stochastic Geometry*, pages 322–376. Wiley, New York, 1974.

[6] R. Kruse and K. D. Meyer. *Statistics with vague data.* D. Reidel Publishing Company, Dordretch, 1987.

[7] I. Levi. *The enterprise of knowledge.* MIT Press, 1980.

[8] G. Matheron. *Random sets and integral geometry.* Wiley, New York, 1975.

[9] M. Puri and D. Ralescu. Fuzzy random variables. *Journal of Mathematical Analysis and Applications*, 114(2):409–422, 1986.

[10] P. Walley. *Statistical reasoning with imprecise probabilities.* Chapman and Hall, London, 1991.