Hypothesis test of sample mean of random intervals and comparing between methods based on Hausdorff distance and the maximum likelihood ratio

Ahssaine Bourakadi*
Department of Mathematics
Faculty of Sciences
Ibn Tofail University Kenitra City
Morocco
ahssaine91@gmail.com

Driss Mentagui
Department of Mathematics,
Faculty of Sciences
Ibn Tofail University Kenitra City
Morocco
dri_mentagui@yahoo.fr

Baraka Achraf Chakir
Department of Mathematics
Faculty of Sciences
Ibn Tofail University Kenitra City
Morocco
Baraka.achraf.chakir@gmail.com

Abstract. In this article, we will introduce some important elements of random sets theory, and we will use the tools of topology, probability theory, and statistical inference aims to develop a coherent mathematical framework to analyze random elements whose realization are intervals.

Hypothesis testing is one of the most widely used methods in statistical inference that involves asking a question, collecting data, and then examining what the data tells us about how to proceed.

The hypothesis tests that we will examine in this paper is the test of sample mean of random sets in a Euclidean space, and in particular in the case where these sets are intervals in a Euclidean space. We will use two test statistics, one is based on Hausdorff distance with using central limit theorem and the second is based on the maximum likelihood ratio.

Keywords: random set, random interval, Hausdorff distance, Minkowski average, maximum likelihood ratio and R software.

*. Corresponding author

1. Introduction

Statistical inference is a very important field in the theory of random sets, including the search for estimator of the expectation of random sets which are considered independent and identically distributed and follow the same capacity functional, in this article, we will demonstrate that the maximum likelihood estimator of this expectation is the Minkowski average of these random sets. This naturally leads to testing hypotheses and building confidence sets for expectation of random set and its subsets in a manner that is analogous to classical methods for scalar-valued random variables. Among the test statistics used for this unilateral hypothesis test which treats the equality of sample mean of independent and identically distributed random sets to a deterministic set, we find the test statistic based on the Hausdorff distance, under the assumption that the size of this sample is very large and using the central limit theorem. In that sense, we propose another method based on the maximum likelihood ratio for testing the equality of sample mean of the convex compact random sets to a deterministic compact and convex set in euclidean space R^d. Finally, we will apply those two methods of test for hypotheses testing of random intervals which are a particular case of random sets in IR , and we will study an empirical example with using the " R" Project for Statistical Computing and discuss the different results obtained from " R" code .

2. Definition of random closed set in Euclidean space IR^d and the Capacity functional

In the theory of random set (see, [1], [2] and [3]), X is called a random closed set in euclidean space IR^d, if X is a map from a probalility space (Ω, T, P) to the family of closed set F in IR^d and $X^-(K)= \{\omega : X(\omega) \cap K \neq \emptyset\}$ belongs to the σ-algebra T on Ω for each compact set K in IR^d. Each random element must be modeled by a probability distribution, in our case this random element is a set, then we use the capacity functional, which have an important role in the theory of random set.

2.1 Definition of the capacity functional

The capacity functional (see, [1], [4], [5] and [6]) of a random set X, is a map from $\mathcal{K}$ a family of compacts sets in IR^d to $[0, 1]$ defining by: $T_X(K) = P(X \cap K \neq \emptyset)$ with $K \in \mathcal{K}$. If $X = \{x\}$, then $T_X(K) = P(\{x\} \cap K \neq \emptyset) = P_x(K)$ and in this case T_X becomes the probability distribution of random variable x.

2.2 Proprieties of the capacity functional

In general case, T_X is a sub-additive capacity functional (see, [7] and [8]), and for each two compacts sets K_1 and K_2 from $\mathcal{K}$ we have:

$$T_X(K_1 \cup K_2) \leq T_X(K_1) + T_X(K_2).$$

And we have the equality, if and only if K_1 and K_2 are disjoints, it means $K_1 \cap K_2 = \emptyset$.

3. Selections of random set

3.1 Definition of selection of random set

We consider a random set X in IR^d, if x is a random element with values in IR^d, such that $x(\omega) \in X(\omega)$ almost surely, then we called x a selection of X. We denote by $Sel(X)(orL^0(X))$ a set of all selections ([2] and [3]) from X, then: $Sel(X) = \{x(\omega) : x(\omega) \in X(\omega), \omega \in \Omega\}$. In particular, if $X = \{x\}$, then $Sel(X) = \{x\}$.

3.2 Fundamental selection theorem

If $X : \Omega \to F$ is an almost surely non-empty random closed set in IR^d, then X has a measurable selection (see, [1] and [2]).

4. Expectation of random set in IR^d

It is difficult to define the expectation of a random set in euclidean space, because the space of closed sets is not linear (see, [2] and [10]), one approach relies on representing a random set using the family of its selections, and taking their expectation. For to have an integrable random closed set X in IR^d, it suffices that X possesses at least one an integrable selection, it means $\exists x(\omega) \in X(\omega)$ almost surly and $\|E(x)\| < +\infty$. Then, an integrable random closed set is non-empty, with probability one. We denote by $L^p(X)$ the family of all p-integrable selections of X, with $p \in [0, +\infty[$. Usually, we are interested in $L^1(X)$ and $L^2(X)$ which are respectively the families of all integrable and square-integrable selections of set X. And we called X respectively an integrable and square-integrable random closed set.

4.1 Definition of expectation of random set in IR^d

The selection expectation EX (see, [1], [2] and [3]) of an integrable random closed set X is the closure of the family of expectations of all its integrable selections $x \in L^1(X)$. Then, we denote $EX = cl\{Ex, x \in L^1(X)\}$.

4.2 Minkowski sum of expectation of random set in IR^d

The Minkowski sum of expectation of two random sets X and Y in IR^d is defining by the formula:

$$E(X+Y) = EX + EY = \{a + b : a \in EX, b \in EY\}.$$

5. Law of large numbers for integrably bouneded random sets

If $X_1, \ldots, X_n$ is a n-sample of independent and identically distributed integrably bounded random compact sets in IR^d, then $d_H(\frac{S_n}{n}, EX_i) \to 0$, almost surely as $n \to \infty$ (see, [2] and [3]). With $S_n = X_1 + \ldots + X_n$, $EX_i = EX_j$, $T_{X_i} = T_{X_j}$ for each$(i,j) \in \{1, \ldots, n\}$, and d_H is the Hausdorff distance of two sets, and we denote:

$$d_H(X,Y) = inf\{r : X \subset Y^r, Y \subset X^r\}$$

with

$$X^r = \{a \in IR^d, d(a,X) \leq r\}$$

and

$$Y^r = \{a \in IR^d, d(a,Y) \leq r\}.$$

According to the law of large numbers, we can sense that estimation of sample mean of random closed sets can be the Minkowski average.

6. Estimation of sample mean of random sets in Euclidean space IR^d

Theorem 6.1. *Let $X_1, \ldots, X_n$ a n-sample of compact random sets, integrable and non-empty in IR^d, Suppose that $(X_i)_{i\in\{1,\ldots,n\}}$ are independent and identically distributed. It means $T_{X_i} = T_{X_j}$ for all $i \neq j$, and $EX_i = EX_j = K$. Then the estimation of sample mean is the Minkowski average and we denote:*

$$\hat{EX} = \hat{K} = \frac{S_n}{n} = \bar{X}_n$$

with S_n is the Minkowski sum of X_i, and K a compact set in IR^d.

Proof. We use the Maximum likelihood function denoted L, we have:

$$L(X_1, \ldots, X_n, K) = P(X_1 \cap K \neq \emptyset, ..., X_n \cap K \neq \emptyset).$$

The events $\{\omega, X_i(\omega) \cap K \neq \emptyset\}$ are independents for each $i \in \{1, \ldots, n\}$, then:

$$L(X_1, \ldots, X_n, K) = \prod_{i=1}^{n} P(X_i \cap K \neq \emptyset).$$

These random sets are identically distributed, then we have:

$$L(X_1, \ldots, X_n, K) = \prod_{i=1}^{n} T_{X_i}(K) = T_{X_i}(K)^n$$

The estimator of K is denoted by $\hat{K}$, and it takes the value that maximizes L. For maximizing L it suffices to maximize $T_{X_i}(K)^n$ for each $i \in \{1, ..., n\}$.

We have $T_{X_i}(K) \in [0,1]$, therefore : $\max_K T_{X_i}(K) = T_{X_i}(\hat{K})$, Then $P(X_i \cap K \neq \emptyset) = 1$, it means the event $\{\omega, X_i(\omega) \cap K \neq \emptyset\}$ is certain, and it suffices to take $\hat{K} = X_i$ for each $i \in \{1, ..., n\}$. For having an exhaustive estimator, we take $n\hat{K} = S_n = \sum_{i=1}^{n} X_i$, then $\hat{K} = \frac{Sn}{n} = \bar{X}_n$. □

6.1 Expectation of K

We have the expectation of Minkowski sum is the Minkowski sum of expectation of random sets $X_1, \ldots, X_n$ then:

$$E(X_1 + \ldots + X_n) = EX_1 + \ldots + EX_n.$$

Then:

$$E\hat{K} = E(\frac{X_1 + \ldots + X_n}{n}) = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n}\sum_{i=1}^{n} K = K.$$

And we conclude that $\hat{K}$ is an unbiased estimator.

7. Central limit theorem of random sets

Let $X_1, \ldots, X_n$ n-sample of random closed sets in IR^d, such that $E\|X_i\|^2 < \infty$, then:

$$\sqrt{n} d_H(\frac{S_n}{n}, EX_i) ==> \|\xi(u)\|_\infty = \sup\{|\xi(u)|; u \in S^{d-1}\}, \text{ as } n \to \infty.$$

With $\|X_i\| = d_H(X_i, \{0\})$ and $\xi(u)$ is a centered Gaussian random field on the unit sphere S^{d-1} in IR^d, with the covariance function $E(\xi(u)\xi(v))$ (see, [1], [2] and [3]).

8. Hypothesis testing

8.1 Hypothesis testing with using central limit theorem

Let $X_1, \ldots, X_n$ n-sample of random compact sets in IR^d, independent and identically distributed such that $EX_i = K$ and $E\|X_i\|^2 < \infty$, for each $i \in 1, \ldots, n$ with $\bar{X_n}$ the Minkowski average of this sample.

Let K_0 be a deterministic non-empty convex compact set in IR^d, in order to test at a prespecified significance level $\alpha \in [0, 1]$, then we consider the following bilateral hypothesis testing $H_0 : K = K_0$, vs $H_1 : K \neq K_0$.

We use the test statistic based on Hausdorff distance (see, [2]), as $n \to \infty$

$$T_n = \sqrt{n} d_H(\bar{X_n}, EX_i) => \|\xi(u)\|_\infty = \sup\{|\xi(u)|; u \in S^{d-1}\}$$

Under H_0, we denote $T_n(K_0) = \sqrt{n} d_H(\bar{X_n}, K_0)$, and we use the following criterion:

$$\begin{cases} \text{Reject } H_0, \text{ if } T_n(K_0) > C_\alpha \\ \text{Fail reject } H_0, \text{ if } T_n(K_0) < C_\alpha \end{cases},$$

where C_α is chosen so that $P(\|\xi(u)\|_\infty > C_\alpha) = \alpha$.

8.2 Hypothesis testing with using maximum liklihood ratio

Now, we started from the above hypothesis, and using the Maximum likelihood function$L$, then we have :

$$\Lambda_n = \frac{L(X_1, \ldots, X_n, K)_{H_1}}{L(X_1, \ldots, X_n, K)_{H_0}} = \frac{L(X_1, \ldots, X_n, \hat{K})}{L(X_1, \ldots, X_n, K_0)}$$

with $\hat{K}$ is the maximum likelihood estimator of K and verifying $\sqrt{n} d_H(\hat{K}, K) => \|\xi(u)\|_\infty$with $\xi(u)$ is a centered Gaussian random field on the unit sphere S^{d-1} in IR^d. Then, the test statistic becomes:

$$T_n = 2\ln(\Lambda_n) \sim \chi^2_m.$$

Such that $m = df(H_1) - df(H_0)$, with $df(H_1)$ is the degree of freedom of parameters of K under H_1, and $df(H_0)$ is the degree of freedom of parameters of K under H_0. Then, we calculate $2\ln(\Lambda_n)$ and we will compare it with $\chi^2_{m,1-\alpha}$, such that α is the type I error.

We have $L(X_1, \ldots, X_n, K) = T_{X_i}(K)^n$, and $\bar{X}_n = \hat{K}$, then:

$$2\ln(\Lambda_n) = 2n[\ln T_{X_i}(\bar{X}_n) - \ln T_{X_i}(K_0)].$$

Then, we define the region of acceptance of hypothesis H_0 :

$$R^{acc} = \{\Lambda_n^{obs}, 2\ln(\Lambda_n) \leq \chi^2_{m,1-\alpha}\}.$$

9. Case of random intervals in IR

Random intervals are a special case of random sets on IR, let $Y = [Y^l, Y^u]$ be a random compact interval on the line IR, with square integrable end points (Y^l and Y^u are two dependent random variables). The unit sphere on the line IR is $\{-1, 1\}$ (see, [2]).

The expectation of Y denoted $EY = [EY^l, EY^u]$ is a compact set in IR, while the covariance function of this random interval can be identified as a 2x2 matrix:

$$C_Y = \begin{pmatrix} var(Y^l) & -cov(Y^l, Y^u) \\ -cov(Y^l, Y^u) & var(Y^u) \end{pmatrix}.$$

9.1 Hypothesis testing with using central limit theorem: case of random interval

Let $Y_1, \ldots, Y_n$ be independent and identically distributed copies of a random compact interval, $Y = [Y^l, Y^u]$ in IR which are closed and bounded, such that $E\|Y\|^2 < \infty$. Using above result in section 8.1, we consider the following bilateral hypothesis testing $H_0 : [EY^l, EY^u] = [a_0, b_0]$ vs $H_1 : [EY^l, EY^u] \neq [a_0, b_0]$.

We use the test statistic based on Hausdorff distance, as $n \to \infty$

$$T_n = \sqrt{n} d_H(\bar{Y}_n, [EY^l, EY^u]) => \|\xi(u)\|_\infty = \sup\{|\xi(u)|; u \in \{-1, 1\}\}$$

with $\bar{Y}_n$ is the Minkowski average of $Y_1, \ldots, Y_n$ denoted $\bar{Y}_n = [\bar{Y}_n^l, \bar{Y}_n^u]$, and $\bar{Y}_n^k = \frac{1}{n}\sum_{i=1}^n Y_i^k, k \in l, u$. Then:

$$T_n = \sqrt{n} d_H([\bar{Y}_n^l, \bar{Y}_n^u], [EY^l, EY^u]) => \sup\{|\xi(-1)|, |\xi(1)|\}.$$

The Hausdorff distance between two intervals equals the maximum of the absolute difference of their corresponding end-points. Then:

$$T_n = \sqrt{n} \max(|\bar{Y}_n^l - EY^l|, |\bar{Y}_n^u - EY^u|) => \sup\{|\xi(-1)|, |\xi(1)|\}$$

$\bar{Y}_n^l$ and $\bar{Y}_n^u$ are the averages of the left and right end points of independent and identically distributed random intervals $Y_1, \ldots, Y_n$. And $(\xi(-1), \xi(1))$ is the bivariate centred Gaussian vector with covariance matrix C_Y (see, [2] and [9]). We denote $\bar{y}_n^l$ and $\bar{y}_n^u$ the realizations of $\bar{Y}_n^l$ and $\bar{Y}_n^u$, respectively.

We calculate the value of $\max(|\bar{y}_n^l - EY^l|, |\bar{y}_n^u - EY^u|)$ and we conclude that:

$$\begin{cases} \text{If } \max(|\bar{y}_n^l - EY^l|, |\bar{y}_n^u - EY^u|) = |\bar{y}_n^l - EY^l|, \text{ then } T_n \sim N(0, V(Y_i^l)) \\ \text{Else if } \max(|\bar{y}_n^l - EY^l|, |\bar{y}_n^u - EY^u|) = |\bar{y}_n^u - EY^u|, \text{ then } T_n \sim N(0, V(Y_i^u)) \end{cases}.$$

The variances of Y_i^l, and Y_i^u are unknown, but as $n \to \infty$, we have:

$$V(Y_i^l) = S_{Y_i^l}^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i^l - \bar{Y}^l)^2$$

and

$$V(Y_i^u) = S_{Y_i^u}^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i^u - \bar{Y}^u)^2.$$

Under H_0, $T_n = \sqrt{n} \max(|\bar{Y}_n^l - a_0|, |\bar{Y}_n^u - b_0|)$ and we conclude the following result:

- if $\max(|\bar{y}_n^l - a_0|, |\bar{y}_n^u - b_0|) = |\bar{y}_n^l - a_0|$, then, $\frac{|\bar{Y_n}^l - a_0|}{S_{Y_i^l}} \sim N(0,1)$ and $R^{accp} = \{{Y_i^l}^{obs}, \frac{|\bar{Y}_n^l - a_0|}{S_{Y_i^l}} \leq z_{\frac{\alpha}{2}}\}$, else, if $\max(|\bar{y}_n^l - a_0|, |\bar{y}_n^u - b_0|) = |\bar{y}_n^u - b_0|$, then, $\frac{|\bar{Y}_n^u - b_0|}{S_{Y_i^u}} \sim N(0,1)$ and $R^{accp} = \{{Y_i^l}^{obs}, \frac{|\bar{Y}_n^u - a_0|}{S_{Y_i^u}} \leq z_{\frac{\alpha}{2}}\}$.

9.2 Hypothesis testing with using maximum likelihood ratio: case of random interval

We started from the above hypothesis, and using the Maximum likelihood function L, then we have :

$$\Lambda_n = \frac{L(Y_1, \ldots, Y_n, [EY^l, EY^u])_{H_1}}{L(Y_1, \ldots, Y_n, [EY^l, EY^u])_{H_0}} = \frac{L(Y_1, \ldots, Y_n, \bar{Y}_n)}{L(Y_1, \ldots, Y_n, [a_0, b_0])}$$

with $\bar{Y_n} = [\bar{Y}_n^l, \bar{Y}_n^u]$ is the maximum likelihood estimator of $[EY^l, EY^u]$. The test statistic becomes:

$$T_n = 2\ln(\Lambda_n) \sim \chi^2_m.$$

Such that $m = df(H_1) - df(H_0)$, with $df(H_1)$ is the degree of freedom of parameters of K under H_1, and $df(H_0)$ is the degree of freedom of parameters of K under H_0.

In this case $df(H_1) = 2$ and $df(H_0) = 0$, then $m = 2$ and under H_0 we have:

$$T_n = 2\ln(\Lambda_n) \sim \chi^2_2.$$

Then, we calculate $2\ln(\Lambda_n)$ and we will compare it with $\chi^2_{2,1-\alpha}$, such that $1-\alpha$ is the confidence level of test.

On the other hand we have $L(Y_1, \ldots, Y_n, [EY^l, EY^u]) = T_{Y_i}([EY^l, EY^u])^n$ that implies:

$$T_n = 2\ln(\Lambda_n) = 2n[\ln T_{Y_i}([\bar{Y}_n^l, \bar{Y}_n^u]) - \ln T_{Y_i}([a_0, b_0])].$$

We define the region of acceptance of hypothesis H_0 :

$$R^{acc} = \left\{\Lambda_n^{obs}, T_n \leq \chi^2_{2,1-\alpha}\right\}.$$

9.3 Simplification of expression of T_n

The problem now is how to calculate the test statistic T_n and to compare it with the value of $\chi^2_{2,1-\alpha}$.

On the one hand, we should to simplify the expression of T_n.

We have:

$$T_{Y_i}([\bar{Y}_n^l, \bar{Y}_n^u]) = P([\bar{Y}_n^l, \bar{Y}_n^u] \cap Y_i \neq \emptyset] = P([\bar{Y}_n^l, \bar{Y}_n^u] \cap [Y_i^l, Y_i^u] \neq \emptyset)$$
$$= P((Y_i^l < \bar{y}_n^l \leq Y_i^u) \cup (\bar{Y}_n^l \leq Y_i^l \leq \bar{Y}_n^u)).$$

The events $\{\omega \in \Omega, Y_i^l(\omega) < \bar{Y}_n^l(\omega) \leq Y_i^u(\omega)\}$ and $\{\omega \in \Omega, \bar{Y}_n^l(\omega) \leq Y_i^l(\omega) \leq \bar{Y}_n^u(\omega)\}$ are independents then:

$$T_{Y_i}([\bar{Y}_n^l, \bar{Y}_n^u]) = P(Y_i^l < \bar{Y}_n^l \leq Y_i^u) + P(\bar{Y}_n^l \leq Y_i^l \leq \bar{Y}_n^u).$$

On the other hand, we have:

$$T_{Y_i}([a_0, b_0]) = P([a_0, b_0] \cap Y_i \neq \emptyset] = P([a_0, b_0] \cap [Y_i^l, Y_i^u] \neq \emptyset)$$
$$= P[(Y_i^l < a_0 \leq Y_i^u) \cup (a_0 \leq Y_i^l \leq b_0)].$$

The events $\{\omega \in \Omega, Y_i^l(\omega) < a_0 \leq Y_i^u(\omega)\}$ and $\{\omega \in \Omega, a_0 \leq Y_i^l(\omega) \leq b_0\}$ are independents then:

$$T_{Y_i}([a_0, b_0]) = P(Y_i^l < a_0 \leq Y_i^u) + P(a_0 \leq Y_i^l \leq b_0).$$

We conclude that:

$$2\ln(\Lambda_n) = 2n(\ln T_{Y_i}([\bar{y}_n^l, \bar{y}_n^u]) - lnT_{Y_i}([a_0, b_0]))$$

$$= 2n \ln \frac{P(Y_i^l < \bar{y}_n^l \leq Y_i^u) + P(\bar{y}_n^l \leq Y_i^l \leq \bar{y}_n^u)}{P(Y_i^l < a_0 \leq Y_i^u) + P(a_0 \leq Y_i^l \leq b_0)}.$$

Y_i^l and Y_i^u are dependent variables, with $P(\ Y_i^l < Y_i^u) = 1$ then, we simplify the expression of $P(Y_i^l < \bar{Y}_n^l \leq Y_i^u)$ and $P(Y_i^l < a_0 \leq Y_i^u)$.

On the one hand, we have:

$$\{\omega \in \Omega, Y_i^l(\omega) < \bar{Y}_n^l(\omega) \leq Y_i^u(\omega)\}$$
$$= \{\omega \in \Omega, Y_i^l(\omega) < \bar{Y}_n^l(\omega)\} \cap \{\omega \in \Omega, \bar{Y}_n^l(\omega) \leq Y_i^u(\omega)\}$$

and $(\{\omega \in \Omega, Y_i^l(\omega) < \bar{Y}_n^l(\omega)\} \cap \{\omega \in \Omega, , \bar{Y}_n^l(\omega) \leq Y_i^u(\omega)\}) \cap \{\omega \in \Omega, Y_i^u(\omega) < \bar{Y}_n^l(\omega)\} = \emptyset$. Then, the events $\{\omega \in \Omega, Y_i^l(\omega) < \bar{Y}_n^l(\omega)\} \cap \{\omega \in \Omega, \bar{Y}_n^l(\omega) \leq Y_i^u(\omega)\}$ and $\{\omega \in \Omega, Y_i^u(\omega) < \bar{Y}_n^l(\omega)\}$ are disjoints, and we have:

$$P\{\omega \in \Omega, Y_i^l(\omega) < \bar{Y}_n^l(\omega) \leq Y_i^u(\omega)\} + P\{\omega \in \Omega, Y_i^u(\omega) < \bar{Y}_n^l(\omega)\}$$
$$= P[(\{\omega \in \Omega, Y_i^l(\omega) < \bar{Y}_n^l(\omega)\} \cap \{\omega \in \Omega, \bar{Y}_n^l(\omega)$$
$$\leq Y_i^u(\omega)\}) \cup \{\omega \in \Omega, Y_i^u(\omega) < \bar{Y}_n^l(\omega)\}]$$
$$= P[(\{\omega \in \Omega, Y_i^l(\omega) < \bar{Y}_n^l(\omega)\} \cup \{\omega \in \Omega, Y_i^u(\omega) < \bar{Y}_n^l(\omega)\}) \cap \Omega]$$
$$= P(\{\omega \in \Omega, Y_i^l(\omega) < \bar{Y}_n^l(\omega)\} \cup \{\omega \in \Omega, Y_i^u(\omega) < \bar{Y}_n^l(\omega)\}).$$

On the other hand, we have $Y_i^l < Y_i^u$ almost surely, then $Y_i^u < \bar{Y}_n^l$ implies $Y_i^l < \bar{Y}_n^l$, then:

$$\{\omega \in \Omega, Y_i^u(\omega) < \bar{Y}_n^l(\omega)\} \subset \{\omega \in \Omega, Y_i^l(\omega) < \bar{Y}_n^l(\omega)\}.$$

And we have:

$$P(\{\omega \in \Omega, Y_i^l(\omega) < \bar{Y}_n^l(\omega) \leq Y_i^u(\omega)\}) + P(\{\omega \in \Omega, Y_i^u(\omega) < \bar{Y}_n^l(\omega)\})$$
$$= P(\{\omega \in \Omega, Y_i^l(\omega) < \bar{Y}_n^l(\omega)\}).$$

Then $P(Y_i^l < \bar{Y}_n^l \leq Y_i^u) + P(Y_i^u < \bar{Y}_n^l) = P(Y_i^l < \bar{Y}_n^l)$ and we conclude that:

$$\begin{cases} P(Y_i^l < \bar{Y}_n^l \leq Y_i^u) = P(Y_i^l < \bar{Y}_n^l) - P(Y_i^u < \bar{Y}_n^l), \\ P(Y_i^l < a_0 \leq Y_i^u) = P(Y_i^l < a_0) - P(Y_i^u < a_0). \end{cases}$$

On the other hand, we have:

$$\begin{cases} P(\bar{Y}_n^l \leq Y_i^l \leq \bar{Y}_n^u) = P(Y_i^l \leq \bar{Y}_n^u) - P(Y_i^l \leq \bar{Y}_n^l), \\ P(a_0 \leq Y_i^l \leq b_0) = P(Y_i^l \leq b_0) - P(Y_i^l \leq a_0). \end{cases}$$

Then:

$$T_n = 2n \ln \frac{P(Y_i^l < \bar{Y}_n^l) - P(Y_i^u < \bar{Y}_n^l) + P(Y_i^l \leq \bar{Y}_n^u) - P(Y_i^l \leq \bar{Y}_n^l)}{P(Y_i^l < a_0) - P(Y_i^u < a_0) + P(Y_i^l \leq b_0) - P(Y_i^l \leq a_0)}.$$

And finally we conclude that:

$$T_n = 2n[\ln(P(Y_i^l \leq \bar{Y}_n^u) - P(Y_i^u < \bar{Y}_n^l)) - \ln(P(Y_i^l \leq b_0) - P(Y_i^u < a_0))].$$

We denote by $Z_i^l = Y_i^l - \bar{Y}_n^u$ and $Z_i^u = Y_i^u - \bar{Y}_n^l$ then:

$$\begin{cases} P(Y_i^l \leq \bar{Y}_n^u) = P(Z_i^l \leq 0) = F_{Z_i^l}(0), \\ P(Y_i^u < \bar{Y}_n^l) =)P(Z_i^u \leq 0) = F_{Z_i^u}(0) \end{cases}$$

and

$$\begin{cases} P(Y_i^l \leq b_0) = F_{Y_i^l}(b_0), \\ P(Y_i^u < a_0) = F_{Y_i^u}(a_0), \end{cases}$$

we replace these equations in the formula of T_n, and we have:

$$T_n = 2n[\ln(F_{Z_i^l}(0) - F_{Z_i^u}(0)) - \ln(F_{Y_i^l}(b_0) - F_{Y_i^u}(a_0))].$$

9.4 Estimation of test statistic T_n

Now, we will estimate the expression of T_n, and for do it, we must to estimate the values of the distribution functions $F_{Z_i^l}(0)$, $F_{Z_i^u}(0)$, $F_{Y_i^l}(b_0)$ and $F_{Y_i^u}(a_0)$.

For that, we use the empirical distribution function F_n denoted by: $\forall x \in IR, F_n(x) = \frac{1}{n}\sum_{i=1}^n 1_{X_i \leq x}$ for a n-sample $X_1, \ldots, X_n$ of random variables, then:

$$\begin{cases} \hat{F}_{Z_i^l}(0) = F_n^{Z_i^l}(0) = \dfrac{1}{n}\displaystyle\sum_{i=1}^n 1_{Z_i^l \leq 0}, \\ \hat{F}_{Z_i^u}(0) = F_n^{Z_i^u}(0) = \dfrac{1}{n}\displaystyle\sum_{i=1}^n 1_{Z_i^u \leq 0} \end{cases}$$

and

$$\begin{cases} \hat{F}_{Y_i^l}(b_0) = F_n^{Y_i^l}(b_0) = \dfrac{1}{n}\displaystyle\sum_{i=1}^n 1_{Y_i^l \leq b_0}, \\ \hat{F}_{Y_i^u}(a_0) = F_n^{Y_i^u}(a_0) = \dfrac{1}{n}\displaystyle\sum_{i=1}^n 1_{Y_i^u \leq a_0}. \end{cases}$$

Then:

$$\begin{aligned} \hat{T}_n &= 2n[\ln(\hat{F}_{Z_i^l}(0) - \hat{F}_{Z_i^u}(0)) - \ln(\hat{F}_{Y_i^l}(b_0) - \hat{F}_{Y_i^u}(a_0))] \\ &= 2n[\ln(F_n^{Z_i^l}(0) - F_n^{Z_i^u}(0)) - \ln(F_n^{Y^l}(b_0) - F_n^{Y^u}(a_0))]. \end{aligned}$$

And, finally, we have:

$$\hat{T}_n = 2n[\ln(\frac{1}{n}\sum_{i=1}^{n} 1_{Z_i^l \leq 0} - \frac{1}{n}\sum_{i=1}^{n} 1_{Z_i^u \leq 0}) - \ln(\frac{1}{n}\sum_{i=1}^{n} 1_{Y_i^l \leq b_0} - \frac{1}{n}\sum_{i=1}^{n} 1_{Y_i^u \leq a_0})]$$
$$= 2n[\ln(\sum_{i=1}^{n} 1_{Z_i^l \leq 0} - \sum_{i=1}^{n} 1_{Z_i^u \leq 0}) - \ln(\sum_{i=1}^{n} 1_{Y_i^l \leq b_0} - \sum_{i=1}^{n} 1_{Y_i^u \leq a_0})].$$

9.5 The Glivenko-Cantelli theorem

Let $X_1, ..., X_n$ be *i.i.d* real-valued random variables with distribution function $F(x) = P(X_i \leq x)$. We denote the standard empirical distribution function by $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i < x}$ then $\sup_{x \in R} |F(x) - F_n(x)| \to 0$ almost surely as $n \to \infty$.

We conclude that $F_n(x)$ uniformly converges to its theoretical counterpart $F(x)$ (see, [11], [12], [13] and [14]). Then, $F_n^{Z_i^l}(0)$, $F_n^{Z_i^u}(0)$, $F_n^{Y^l}(b_0)$ and $F_n^{Y^u}(a_0)$ uniformly converges to $F_{Z_i^l}(0)$, $F_{Z_i^u}(0)$,$F_{Y_i^l}(b_0)$ and $F_{Y_i^u}(a_0)$ respectively as $n \to \infty$. And, we have:

$$\lim_{n\infty} \frac{\hat{T}_n}{T_n} = \lim_{n\infty} \frac{2n[\ln(F_n^{Z_i^l}(0) - F_n^{Z_i^u}(0)) - \ln(F_n^{Y_i^l}(b_0) - F_n^{Y_i^u}(a_0))]}{2n[\ln(F_{Z_i^l}(0) - F_{Z_i^u}(0)) - \ln(F_{Y_i^l}(b_0) - F_{Y_i^u}(a_0))]}$$
$$= \lim_{n\infty} \frac{\ln(F_n^{Z_i^l}(0) - F_n^{Z_i^u}(0)) - \ln(F_n^{Y_i^l}(b_0) - F_n^{Y_i^u}(a_0))}{\ln(F_{Z_i^l}(0) - F_{Z_i^u}(0)) - \ln(F_{Y_i^l}(b_0) - F_{Y_i^u}(a_0))}.$$

Such that:

$$\begin{cases} F_n^{Z_i^l}(0) - F_n^{Z_i^u}(0) \to F_{Z_i^l}(0) - F_{Z_i^u}(0) \text{ as } n \to \infty \\ F_n^{Y_i^l}(b_0) - F_n^{Y_i^u}(a_0) \to F_{Y_i^l}(b_0) - F_{Y_i^u}(a_0) \text{ as } n \to \infty \end{cases}$$

Then $\lim_{n\infty} \frac{\hat{T}_n}{T_n} = 1$, and we conclude that $\hat{T}_n \sim T_n$ as $n \to \infty$. Then, the test statistic using for hypothesis testing H_0 is $\hat{T}_n$, and finally we calculate the value of $\hat{T}_n$, and we will compare it with $\chi^2_{2,1-\alpha}$, and we conclude the region of acceptance of this hypothesis:

$$R^{acc} = \{Y_i^{u^{obs}}, Y_i^{l^{obs}}, \hat{T}_n \leq \chi^2_{2,1-\alpha}\}.$$

10. Simulation study

In this paragraph we will to compare between decisions taking for hypothesis testing of random interval with using the test statistics based on Hausdorff distance and the maximum likelihood ratio. Let's be 4000 observations of random interval in IR, our objective is to test the equality between the expectation of this sample and a deterministic interval. Using the " R" Project for Statistical Computing, we denote the first 20 observations and the last 20 observations from our simulation data denoted " simulation.csv"

Table1: The first 20 and last 20 observations of random interval from R Software

Row	Y^l	Y^u	*Row*	Y^l	Y^u
1	20	35	3981	99520	99535
2	45	60	3982	99545	99560
3	70	85	3983	99570	99585
4	95	110	3984	99595	99610
5	120	135	3985	99620	99635
6	145	160	3986	99645	99660
7	170	185	3987	99670	99685
8	195	210	3988	99695	99710
9	220	235	3989	99720	99735
10	245	260	3990	99745	99760
11	270	285	3991	99770	99785
12	295	310	3992	99795	99810
13	320	335	3993	99820	99835
14	345	360	3994	99845	99860
15	370	385	3995	99870	99885
16	395	410	3996	99895	99910
17	420	435	3997	99920	99935
18	445	460	3998	99945	99960
19	470	485	3999	99970	99985
20	495	510	4000	99995	100010

Source: Author

We consider the hypothesis testing:

$H_0 : [EY_l, EY_u] = [50000, 50020]$ vs $H_1 : [EY_l, EY_u] \neq [50000, 50020]$

10.1 Test statistic based on Hausdorff distance using the central limit theorem

We have the test statistic is denoted by:

$$T_n = \sqrt{n} \max(|\bar{Y}_n^l - EY^l|, |\bar{Y}_n^u - EY^u|) => \sup\{|\xi(-1)|, |\xi(1)|\}.$$

Under H_0, we have:

$$T_{4000} = \sqrt{4000} \max(|\bar{Y}^l - 50000|, |\bar{Y}^u - 50020|).$$

We have:

$$\max(|\bar{Y}^l - 50000|, |\bar{Y}^u - 50020|) = |\bar{Y}^l - 50000|,$$

and

$$\frac{\bar{Y}^l - 50000}{S_{Y_l}} \sim N(0, 1).$$

Taken the type 1 error $\alpha = 0.05$, we have:

$$\frac{\bar{Y}^l - 50000}{S_{Y_l}} = 0,01642962 \leq Z_{0.025} = 1,96.$$

Then, we accept the hypothesisH_0 (see, appendices).

10.2 Test statistic based on maximum likelihood ratio

Test statistic based on maximum likelihood ratio is denoted by:

$$\hat{T}_n = 2n[\ln(\sum_{i=1}^{n} 1_{Z_i^l \leq 0} - \sum_{i=1}^{n} 1_{Z_i^u \leq 0}) - \ln(\sum_{i=1}^{n} 1_{Y_i^l \leq b_0} - \sum_{i=1}^{n} 1_{Y_i^u \leq a_0})].$$

In our simulation data, we have n=4000, and under H_0 :

$$\hat{T}_{4000} = 2 * 4000[\ln(\sum_{i=1}^{4000} 1_{Z_i^l \leq 0} - \sum_{i=1}^{4000} 1_{Z_i^u \leq 0}) - \ln(\sum_{i=1}^{4000} 1_{Y_i^l \leq 50020} - \sum_{i=1}^{4000} 1_{Y_i^u \leq 50000})].$$

With $Z_i^l = Y_i^l - \bar{Y}^u$ and $Z_i^u = Y_i^u - \bar{Y}^l$.

We have $\hat{T}_{4000} \leq \chi^2_{2,0.05} = 5.991$, then we accept the hypothesis H_0(see appendices).

11. Discussion

In this part we made the comparison between two statistical test methods in the theory of random sets especially the random intervals, the first is that based on the central limit theorem, using the Hausdorff distance which is most known when the size of the sample is very large.

In this perspective, we proposed another test method based on the maximum likelihood ratio, the test statistic of which depends on unknown probability distributions, which forced us to estimate them, in order to be able to calculate this test statistic and compare it with the value of the chi-square distribution.

The decision taken by the two tests are equivalent, so we can replace the test method based on the central limit theorem by the maximum likelihood ratio test which is a very powerful test.

Appendices

Code from R software of test statistic based on Hausdorff distance

data = read.csv("C:/Users/abourakadi/Desktop/simulation.csv",sep=";")
summary (data)

Yl		Yu	
Min.	20	Min.	35
1st Qu.	25014	1st Qu.	25029
Median	50008	Median	50023
Mean	50008	Mean	50023
3rd Qu.	75001	3rd Qu.	75016
Max.	99995	Max.	100010

```
a0=50000
b0=50020
N=sqrt(nrow(data))*max(mean(data[,1])-a0,mean(data[,2]-b0))
N
[1]474.3416
T=N/sd(data[,1])
T
[1] 0.01642962
```

Source: Author

Code from R software of test statistic based on maximum likelihood ratio

$table(data[,1] \leq mean(data[,2]))$

FALSE	TRUE
1999	2001

$table(data[,2] \leq mean(data[,1]))$

FALSE	TRUE
2001	1999

$table(data[,1] \leq 50020)$

FALSE	TRUE
1999	2001

$table(data[,2] \leq 50000)$

FALSE	TRUE
2001	1999

```
T=2*4000*(log(2001-1999)-log(2001-1999))
T
[1] 0
```

Source: Author

References

[1] I. Molchanov, I.S. Molchanov, *Theory of random sets*, 19 (2005), London: Springer.

[2] I. Molchanov, F. Molinari, *Random sets in econometrics* (Vol. 60), Cambridge University Press, 2018.

[3] I. Molchanov, F. Molinari, *Applications of random set theory in econometrics*, Annu. Rev. Econ., 6 (2014), 229-251.

[4] G. Matheron, *Random sets theory and its applications to stereology*, Journal of Microscopy, 95 (1972), 15-23.

[5] G. Matheron, *Ensembles fermés aléatoires, ensembles semi-markoviens et polyedres poissoniens,* Advances in Applied Probability, 508-541, 1972.

[6] J. Serra, *The Boolean model and random sets*, In Image Modeling (pp. 343-370), Academic Press, 1981.

[7] H.T. Nguyen, *On random sets and belief functions*, In Classic Works of the Dempster-Shafer Theory of Belief Functions (pp. 105-116). Springer, Berlin, Heidelberg, 2008.

[8] R. Kruse, *On the variance of random sets,* Journal of Mathematical Analysis and Applications, 122 (1987), 469-473.

[9] A.J. Baddeley, I.S. Molchanov, *On the expected measure of a random set*, In Proceedings of the International Symposium on Advances in Theory and Applications of Random Sets (Fontainebleau, 1996), (pp. 3-20).

[10] H.G. Tucker, *A generalization of the Glivenko-Cantelli theorem*, The Annals of Mathematical Statistics, 30 (1959), 828-830.

[11] J. DeHardt, *Generalizations of the Glivenko-Cantelli theorem*, The Annals of Mathematical Statistics, 42 (1971), 2050-2055.

[12] J. Wolfowitz, *Generalization of the theorem of Glivenko-Cantelli,* The Annals of Mathematical Statistics, 25 (1954), 131-138.

[13] A. Van Der Vaart, J.A. Wellner, *Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes*, In High dimensional probability II (pp. 115-133), Birkhäuser, Boston, MA, 2000.

Accepted: June 19, 2021