Extension of the Glivenko-Cantelli theorem to the resampling.

Renato del Canto Jarpa

Instituto de Estadística Aplicada y Computación
Universidad de Los Andes
Mérida, Venezuela.

Abstract: In this article, the extension of the Glivenko-Cantelli theorem to the resampling is proved in its version of convergence in quadratic mean. The consequences for the development of the statistical inference based on resampling is analyzed. The results are extended to the case of the mixture of information of the sample with the sample of the sample.

Key Words: Probability space, random sample, empirical sample distribution function of the random sample, resampling, sample of the sample, empirical sample distribution function of the sample of the sample, convergence in quadratic mean.

1. Introduction

The Glivenko-Cantelli theorem or central theorem of statistics is the one which enables best to lay the foundation of statistics. The inferential base composed of a probability space and a random sample has served to justify the foundation of classical statistical inference based on the fact that the empirical sample distribution function converges to the population distribution function (Glivenko-Cantelli theorem).

In the context of classical statistical inference, methodologies have been developed which consider samples of the sample or randomization of the same with different purposes. "Jackknife" was developed by Quenouille-Tukey in this way in order to reduce the bias of an estimator.

Later, a modification of jackknife was developed by Efron who called it "bootstrap", with the purpose to estimate the distribution of a statistic or its standard error. See Efron (1979).

The jackknife was borne with its basis already justified, but in the case of bootstrap it was not the same. We will attempt here to justify the basis ...

1.1. Inferential base of the classical statistics and Glivenko-Cantelli theorem

Let X be the random variable with probability distribution function $F_x(x;\theta)$, induced by X in the real line, according to a probability space (Ω,A,P), that is,

(1.1) $X \sim F_x(x;\theta) = P[X \leq x ; \theta] = F_X(x) = F(x)$, $\theta \in \Theta \subseteq |R$.

Let $\mathbf{X}_n$ be a random sample of size n of X, that is,

(1.2) $\mathbf{X}_n = (X_1 , ... , X_i, ... , X_n)'$ in which,

(1.3) $X_i \sim_{ind} F(x_i)$ for i = 1,2, ... , n

The inferential base of the classical statistics is constituted by a probability space and a random sample, given by

(1.4) $(\Omega,A,P),X_n$

Empirical sample distribution function. Let the n-th empirical sample distribution function of $\mathbf{X}_n$ be denoted by $F_n(x;\mathbf{X}_n)$ and be defined for each real x by the random variable

(1.5) $F_n(x;\mathbf{X}_n) = (1/n) \Sigma_i I_{(-\infty,x]}(x_i)$

The random variable $n F_n(x;\mathbf{X}_n)$ has a binomial distribution with parameters n and F(x), for each real x, that is,

(1.6) $n F_n(x;\mathbf{X}_n) \sim bin(. ; n, F(x))$

The foundation of classical statistics is based on the Glivenko-Cantelli theorem and its consequences.

Glivenko-Cantelli theorem. For each real x the random variable $F_n(x;\mathbf{X}_n)$ converges almost sure to F(x), which we denote by

(1.7) $F_n(x;\mathbf{X}_n) \rightarrow^{a.s.} F(x)$, as $n \rightarrow \infty$

1.2. **Consequences of the Glivenko-Cantelli theorem.** Given that the almost sure convergence implies the convergence in probability and this in turn the convergence in distribution, from this it follows that $F_n(x;\mathbf{X}_n)$ converges to F(x) in probability and in distribution. But, since the former types of convergence do not imply convergence in quadratic mean, it becomes necessary to prove separately this type of convergence.

The Glivenko-Cantelli theorem assures us for large samples $F_n(x;\mathbf{X}_n)$ is close to $F(x)$ and therefore an unknown functional $\theta(F)$ may be estimated by an appropriate $\theta(F_n)$. The meaning of the proximity is given by the significance and consequences of each type of convergence.

Finally, we may state that the knowledge of the distribution of probability of the estimator $\theta(F_n)$ makes possible to do the statistical inference. There exists an extensive theoretical support to derive these distributions. The work in this direction has been tedious and not always easy.

So far we have mentioned only the classical statistical inference. Now in the sequel we consider **a new way to carry out the statistical inference.**

2.- Statistical inference based on the random sampling of a random sample.

Let $\mathbf{x}_n$ be an observed random sample of the random variable X given by (1.1). That is, let $\mathbf{X}_n = \mathbf{x_n}$, in which:

(2.1) $\mathbf{x_n} = (x_1, \dots , x_i , \dots , x_n)'$

Let $\mathbf{X}^*_{n,m}$ be a random sample of size m of the observed random sample $\mathbf{x}_n$, obtained by assigning the probability (1/n) to each element of $\mathbf{x_n}$. That is,

(2.2) $\mathbf{X}^*_{n,m} = (X^*_1, \dots , X^*_j , \dots , X^*_m)'$ in which

(2.3) $\mathbf{X}^*_j \sim_{ind} F_n(x;\mathbf{x}_n)$, $j = 1,2, \dots , m$

The inferential base of the statistics based in the random sampling of a random sample is composed of a probability space, a random sample of size n, $\mathbf{X}_n$, which is the population random sample, and a random sample of size m denoted by $\mathbf{X}^*_{n,m}$ which is obtained from the observed random sample $\mathbf{x}_n$

(2.4) $(\Omega,A,P),\mathbf{X}_n,\mathbf{X}^*_{n,m}$

We denote the (n,m)-th empirical sample distribution function of $\mathbf{X}^*_{n,m}$ by $F_{n,m}(x;\mathbf{X}^*_{n,m})$, which is defined for all real x by the following random variable

(2.5) $F_{n,m}(x;\mathbf{X}^*_{n,m}) = (1/m)\ \Sigma_j\ I_{(-\infty,x]}(X^*_j)$

The random variable $m\ F_{n,m}(x;\mathbf{X}^*_{n,m})$ has a binomial distribution with parameters m and $F_n(x;\mathbf{x}_n)$ for each real x. That is,

(2.6) $m\ F_{n,m}(x;\mathbf{X}^*_{n,m}) \sim bin(. ; m , F_n(x;\mathbf{x}_n))$

The justification of the statistical inference based on random sampling of a random sample, and in consequence the bootstrap is based on an extension of the Glivenko-Cantelli theorem to the sampling of samples, is given in the following theorem:

3.- Theorem (extension of the Glivenko-Cantelli to the resampling): For each real x, the random variable $F_{n,m}(x;\mathbf{X}^*_{n,m})$ converges almost sure to $F(x)$, for n and m which tend to infinity in both processes of sampling and in the order indicated. We denote this by

(3.1) $F_{n,m}(x;\mathbf{X}^*_{n,m}) \rightarrow^{a.s.} F(x)$ as $n \rightarrow\infty, m \rightarrow\infty$

Proof: By Glivenko-Cantelli theorem, as n tends to infinity, F_n converges almost sure to $F(x)$; besides when m tends to infinity, Fn,m converges to Fn which in turn converges to $F(x)$. Hence, $F_{n,m}(x;\mathbf{X}^*_{n,m})$ converges almost sure to $F(x)$.See Bickel and Freedman (1981)

3.1.- Consequences of the extension of the Glivenko-Cantelli theorem: These are similar to the Glivenko-Cantelli theorem: $F_{n,m}(x;\mathbf{X}^*_{n,m})$ converges in probability and in distribution to $F(x)$.

Since the convergence in quadratic mean is not implied by the almost sure convergence, we will now state and prove **the extension of the Glivenko-Cantelli theorem to the resampling** in its version of convergence in quadratic mean:

3.2.- Theorema (extension of the theorem of Glivenko-Cantelli to the resampling). For each real x the (n,m)-th empirical distribution function $F_{n,m}(x;\mathbf{X}^*_{n,m})$ converges in quadratic mean to F(x), when n , m tend to infinity in the order indicated. We denote this by:

(3.2) $F_{n,m}(x;\mathbf{X}^*_{n,m}) \rightarrow^{q.m.} F(x)$ as $n \rightarrow \infty$ and $m \rightarrow \infty$

Proof: We should show that as $n \rightarrow \infty$ and $m \rightarrow \infty$:

(3.3) $\lim E[(F_{n,m}(x;\mathbf{X}^*_{n,m}) - F(x))^2] = 0.$ as $n \rightarrow \infty$ and $m \rightarrow \infty$

Using the exact distribution of the random vector $(F_n , F_{n,m})'$, see del Canto (1985), we get the result that the expected value of $F_{n,m}(x)$ is F(x) and in consequence the limit that should be evaluated in (3.3) is the limit of the variance of $F_{n,m}(x)$. Consequently we should evaluate

(3.4) $\lim(n+m-1)F(x)(1-F(x))/(nm)$ as $n \rightarrow \infty$ and $m \rightarrow \infty$

Hence this limit is 0 (zero). This in turn proves the extension of the Glivenko-Cantelli theorem to the random resampling of random samples.

4.- Discussion.

4.1 Consequences of the extension of the Glivenko-Cantelli theorem to resampling.

The most important consequence of the implications of this extension of the Glivenko-Cantelli theorem is that it enables us to assert that the statistical inference can be done using a random sample of the population random sample. The estimators of $\theta(F)$ may be constructed using appropriate $\theta(F_{n,m})$. Next, the probability distribution of such estimators may be determined in order to construct confidence intervals and test of hypotheses. Besides there also exist the alternative to estimate the distribution of the estimator by using bootstrap.

Remark: In the context of statistical inference based on resampling, the bootstrap can be considered as the Monte Carlo method with the restriction that only the sample elements of the observed population sample can be extracted. See del Canto (1992).

4.2 Criticism about statistical inference based on resampling.

In the resampling process, the sample size m can be less than, equal to or greater than n, which is the size of the population random sample. Whatever be the case, some elements of the random sample may not appear in the sample of the sample and this “is considered as a loss of information”. The answer to this criticism is that we may put together all the data of the random sample in the sample of the sample, constituting in this way a new joint sample, whose empirical sample distribution function also converges to F(x) and therefore a new extension of the Glivenko-Cantelli can be proved. This new sample distribution function results to be a convex linear combination of the sample distribution function considered before and therefore also converges to F(x).

Another natural criticism is that “it is necessary to revise completely all the classical concepts of the statistical inference, so that no contradiction may arise or appear as improper manipulation of the data”. Even if it is like this, the advancement of the statistical inference requires it and besides it is a very motivated work.

Finally, another criticism comes from the area of sampling of the finite populations. ... "Why use random sampling and not simple random sample ?." Some of the reasons are obvious but there is an interesting approach with respect to this. It is possible to apply random sampling to simple random samples and we would come across important situations which should be considered in such inferences.

5. Conclusion

The extension of the Glivenko-Cantelli theorem to the resampling enables us to provide a complete justification of statistical inference based on the random sample of an observed population random sample. The estimator $\theta(F_n)$ of the population parameter $\theta(F)$, using a random sampling has an expression which is similar to the estimator $\theta(F_{m,n})$ of the same parameter using a random sample of the observed population random sample. Finally, since both F_n and $F_{n,m}$ converges to F, both $\theta(F_n)$ and $\theta(F_{n,m})$ are very close to each other in values for large samples.

References

[1] Bickel and Freedman. Some asymtotic theory for the bootstrap. The annals of Statistics, Vol. 9 N_o 6, 1196-1217 (1981)

[2] del Canto J.,R. Distribución exacta del vector aleatorio $(F_r(x),F_{r,n}(x))'$ en el muestreo aleatorio de una muestra aleatoria. IEAC – Instituto de Estadística Aplicada y Computación. Universidad de Los Andes, Venezuela. (1985)

[3] del Canto J.,R. Fundamentación del bootstrap mediante una extensión del teorema de Glivenko-Cantelli, Jornadas Matemáticas del Departamento de Matemáticas de la Universidad de Concepción, Concepción, Chile (1992)

[4] Efron, B. Bootstrap Methods: another look at the Jackknife, Annals of Statistics. 7p. 1-26 (1979)

E-mail of the author: delcanto@cantv.net and delcanto@faces.ula.ve