A note on the asymptotic law of the histogram without continuity assumptions

Thomas Laloë, Rémi Servien

To cite this version:

Thomas Laloë, Rémi Servien. A note on the asymptotic law of the histogram without continuity assumptions. Brazilian Journal of Probability and Statistics, São Paulo, SP : Associação Brasileira de Estatística, 2016, 30 (4), pp.1. 10.1214/15-BJPS294 . hal-01153062v2

HAL Id: hal-01153062

https://hal.archives-ouvertes.fr/hal-01153062v2

Submitted on 6 Oct 2015

Submitted to the Brazilian Journal of Probability and Statistics
arXiv: arXiv:0000.0000

A note on the asymptotic law of the histogram without continuity assumptions

Thomas Laloë[a] **and Rémi Servien**[b]

[a] *Université de Nice Sophia-Antipolis, Laboratoire J-A Dieudonné, Parc Valrose, 06108 Nice Cedex 02, France.*
[b] *INRA-ENVT, Université de Toulouse, UMR1331 Toxalim, Research Centre in Food Toxicology, F-31027 Toulouse, France.*

Abstract. Asymptotic normality of density estimates often requires the continuity of the underlying density and assumptions on its derivatives. Recently, these assumptions have been weakened for some estimates using the less restrictive notion of regularity index. However, the particular definition of this index makes it unusable for many estimates. In this paper, we define a more general regularity concept : the r-regularity. This concept is used to obtain asymptotic law of the histogram without hypothesis on the continuity of the underlying density. As expected, when it does exist, the limit distribution is a standard Gaussian. Then, to illustrate the new definition of r-regularity, examples are studied.

1 Introduction

The subject of this paper is related to the general problem of derivation of measures (Rudin, 1987; Dudley, 2002) and is motivated by a paper by Berlinet and Levallois (2000). In their paper, Berlinet and Levallois address the problem of the asymptotic normality of the nearest neighbor density estimator when the density has bad local behavior (e.g. it is not continuous or has infinite derivative). If you denote λ the Lebesgue measure and $B_\delta(x)$ the open ball of radius δ and center $x \in \mathbb{R}$, and if, for fixed $x \in \mathbb{R}$, the following limit

$$\ell(x) = \lim_{\delta \to 0} \frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} \tag{1.1}$$

exists, then x is called a Lebesgue point of the measure μ. As shown in Berlinet and Levallois (2000), the rate of convergence of $\frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))}$ towards $\ell(x)$ plays a key role in the density estimation. In particular, they showed

MSC 2010 subject classifications: 62G07.

Keywords and phrases. Lebesgue point; asymptotic normality; histogram; probability measure; regularity index.

that it is possible to weaken the classical assumption of continuity for the density, replacing it by the less restrictive notion of Lebesgue point. In this context, they define a ρ-regularity point of the measure μ as any Lebesgue point x of μ satisfying

$$\left|\frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} - \ell(x)\right| \leq \rho(\delta).$$

where ρ is a measurable function such that $\lim_{\delta\downarrow 0}\rho(\delta) = 0$ and they use it to obtain nice properties (such as asymptotic normality) for the nearest neighbor density estimate. Besides, Beirlant, Berlinet and Biau (2008) assume that

$$\frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} = \ell(x) + C_x\delta^{\alpha_x} + o(\delta^{\alpha_x}) \text{ when } \delta\downarrow 0, \tag{1.2}$$

in order to obtain improvements on nearest-neighbor estimation (such as removing bias or selection of the number of neighbors). Moreover, Berlinet and Servien (2011) state a necessary and sufficient condition for the existence of a limit distribution of the nearest neighbor density estimate using the regularity index α_x.

Nevertheless, this definition suffers some flaws. First, some measures with ρ-regularity have no regularity index α_x. Second, many density estimates require a development for a ratio of set measures which are not centered around the estimation point x and which are not balls. Definition (1.2) of the regularity index is useless in these cases. These flaws represent a major restriction in practice, since we can not obtain similar results for an estimate such as the histogram, even for measures that could have a regularity index α_x. To circumvent these problems, we propose in Section 2 a new definition, called r-regularity which is then used to obtain a sufficient condition for the existence of a limit distribution for the histogram without any continuity assumptions concerning the underlying density. Examples are given in Section 3 using this definition which is shown to be useful in some cases when the regularity index α_x does not exist. Proofs are presented in Section 4.

2 Definitions and results

2.1 The r-regularity

To mitigate the two problems of the definition of Beirlant et al., we propose the following definition. Given $x \in \mathbb{R}$ we set $\mathcal{I}_x$ the set of all the intervals which contain x and we define E_x by

$$\begin{aligned} E_x = \Big\{ & r > 0 \text{ such that } \exists C > 0, \exists \lambda_0 > 0, \text{ such that } \forall I \in \mathcal{I}_x \\ & \text{verifying } \lambda(I) < \lambda_0 \text{ we have } \left| \frac{\mu(I)}{\lambda(I)} - l(x) \right| \leq C \lambda(I)^r \Big\}. \end{aligned}$$

If there exists a real r_x such that

$$r_x = \sup E_x, \tag{2.1}$$

r_x is the r-regularity index of the measure μ at x. If $\sup E_x = +\infty$, we set $r_x = +\infty$. Examples of r-regularity are provided in the next section.

With this definition, the r-regularity can be viewed as an intermediate stage between the ρ-regularity and the regularity index: it gives us a bound for the rate of convergence of the measures. Furthermore, the r-regularity does not involve a ball centered on x and, consequently, can be used with a larger class of density estimates. Note that, as for the regularity index, the larger the value of r_x, the more regular the derivative of μ with respect to λ. Using the known estimates of the regularity index obtained by Beirlant, Berlinet and Biau (2008) or Berlinet and Servien (2012), a bound could be trivially obtained for the r-regularity index.

2.2 Main results

The histograms are probably the oldest and simplest method to estimate an unknown density. The simplest histogram methods partition the space into congruent intervals or cubes whose size and position depends on the number of available data points, but not on the data itself. They are meant to approximate the data distribution in the best manner possible within a bounded amount of space. These methods provide estimates that are consistent, regardless of the underlying distribution of the data. For a more detailed literature on the subject, we refer the reader to Ioannidis (2003) and the references therein.

Asymptotic results have been derived with a continuity assumption on the density to estimate (Stadtmüller, 1983; Devroye and Györfi, 1985). In Theorem 2.1 below we state the asymptotic normality of the histogram estimate of the density function, removing this continuity assumption by using the r-regularity index. This index could also be used to obtain similar results for other density estimates.

Assume that $X_1, \dots, X_n$ are i.i.d. observations drawn from an unknown probability measure μ with density function f. A histogram f_h consists of a partition of the space $\mathbb{R}$ of Borel-measurable subsets of $\mathbb{R}$, referred to as cells. We consider here partitions with the same size h_n such that

$$B_{nq} = [(q-1)h_n, qh_n[, q \in \mathbb{Z}$$

with the property that (i) $\cup_{q\in\mathbb{Z}} B_{nq} = \mathbb{R}$ and (ii) $B_{nq} \cap B_{nq'} = \emptyset$ if $q \neq q'$. Using these notations, the histogram estimate is

$$f_h(x) = \frac{\nu_{nq}}{nh_n}$$

with $x \in B_{nq}$ and ν_{nq} the number of X_i in the B_{nq} cell. By consequence, the function f_h is constant in a cell. So, to obtain the consistency of f_h towards f, the cells need to become smaller and smaller with n. This estimate is proven (Bosq and Lecoutre, 1987) to be convergent in quadratic mean if

$$\lim_{n\to\infty} h_n = 0 \quad \text{and} \quad \lim_{n\to\infty} nh_n = +\infty. \tag{2.2}$$

Here we are interested in the limit distribution of the histogram. Contrary to the case of the k_n-nearest neighbor estimate (Berlinet and Servien, 2011), the α-regularity is useless. Indeed, the cells of the histogram are not balls centered on the point of estimation. By consequence, we rather use the r-regularity defined in Section 2.

Theorem 2.1. *Suppose that the convergence conditions (2.2) hold and that x is a Lebesgue point in $\mathbb{R}$ where (2.1) is satisfied with $l(x) > 0$. Then the condition*

$$\lim_{n\to\infty} nh_n^{2r+1} = 0 \tag{2.3}$$

for some $r \in]0; r_x[$ implies that

$$H_n(x) = \sqrt{nh_n}\frac{f_h(x) - l(x)}{\sqrt{l(x)}}$$

converges in distribution towards a $\mathcal{N}(0,1)$.

Remark 1: Remind that the definition of E_x excludes the case $r_x = 0$. Indeed in this case the condition of Theorem 2.1 would be $\lim_{n\to\infty} nh_n = 0$, which is in contradiction with the second condition of (2.2).

Remark 2: Note that the case $r_x = +\infty$ is covered by Theorem 2.1 (and an example is provided on next section). In this case, the conditions (2.2) implies condition (2.3) and, by consequence, is sufficient to ensure Theorem 2.1.

A major point is that we obtain the asymptotic normality of the histogram without a continuity assumption on the density function f at the point of estimation x. Nevertheless, this result provides a necessary condition for having a limit distribution, but not a sufficient one. This comes from the fact that, unlike the regularity index, the r-regularity does not provide us with an exact rate, but only an upper bound of the rate.

3 Examples of r-regularity

In a sake of compactness, the proof of the following examples are not presented in the present note. It can be found in Servien (2010).

3.1 When the regularity index does not exist

Let f_1 be the probability density defined by

$$f_1(x) = \frac{2 - \cos(1/x) + 2x\sin(1/x)}{c}$$

in $\mathbb{R}$ for $x \in [-1,1]\backslash 0$ with $c = 4 + 2\sin 1$ and μ_1 its probability measure.

The densityf_1 is differentiable at any point of $[-1,1]$ except at the point 0 where it has no left and no right limits. We have

$$\lim_{h\to 0} \frac{F_1(h) - F_1(-h)}{2h} = \frac{2}{c}.$$

Thus 0 is a Lebesgue point of μ_1 but, by setting, $f_1(0) = 2/c$,we still have discontinuity of the second kind at the point 0. Now,

$$\frac{\mu_1([-h,h])}{2h} - f_1(0) = \frac{1}{c} h \sin\left(\frac{1}{h}\right)$$

so, at any point of $[-1,1]$ we have ρ-regularity with $\rho(\delta) = \delta/c$ but no regularity index at the point 0.

If we choose h_1 and g_1 two positive integers with $0 \in I = [-h_1, g_1]$ and $\lambda(I) < \lambda_0$. We obtain

$$\frac{\mu_1(I)}{\lambda(I)} = \frac{2}{c} + \frac{g_1^2 \sin(1/g_1) + h_1^2 \sin(1/h_1)}{c\lambda(I)}$$

and, with $f_1(0) = 2/c$, we have

$$\left|\frac{\mu_1(I)}{\lambda(I)} - f_1(0)\right| \leq \frac{g_1^2 + h_1^2}{c\lambda(I)} \leq \frac{1}{c}\lambda(I)$$

which gives us a r-regularity at the point 0 with $r_0 = 1$.

3.2 Lipschitz case

Lemma 3.1. *Assume that h is a Lipschitz density with order β at the point $x \in \mathbb{R}$ such that, for all t,*

$$|h(x) - h(t)| \leq C_x |x - t|^{\beta},$$

and μ_h its associated measure. For all intervals I with $\lambda(I) \neq 0$ and $x \in I$, we have

$$\left|\frac{\mu_h(I)}{\lambda(I)} - h(x)\right| \leq C_x \lambda(I)^{\beta}.$$

Example :
Let f_2 be the probability density function on $[-1/2, 1/2]$ defined by

$$f_2(x) = 1 - \frac{\sqrt{2}}{3} + \sqrt{|x|}$$

and μ_2 its associated measure. At the point $x = 0$, f_2 is continuous, not differentiable and $\frac{1}{2}$-lipschitzian. If $0 \in I$ and $\lambda(I) < \lambda_0$, a straightforward development gives us

$$\left|\frac{\mu_0(I)}{\lambda(I)} - f_0(0)\right| \leq \frac{2}{3}\lambda(I)^{1/2}.$$

3.3 Constant density

Assume that d is a constant density on an interval $I \neq \emptyset$ and that μ_d is its associated measure. Thus, for all $x \in I$, we have

$$\mu_d(I) = \int_I d(t)dt = d(x)\lambda(I)$$

which leads us to

$$\left|\frac{\mu_d(I)}{\lambda(I)} - d(x)\right| = 0$$

and, by consequence, $r_x = +\infty$. As r_x increases with the regularity of the derivative of the measure, it is logical to find that it is maximum when the density is constant.

3.4 When r_x does not exist

Consider the probability density function f_3 defined on $[-0.5; 0.5]$ by

$$\begin{aligned} f_3(x) &= \frac{1}{\log|x|} + 1 - a \text{ if } x \neq 0 \\ &= 1 - a \qquad\qquad\quad \text{if } x = 0 \end{aligned}$$

where a is a normalization constant and μ_3 its associated probability measure. At the Lebesgue point 0, Lemma 3.2 in Berlinet and Levallois (2000) gives us ρ-regularity with $\rho(\delta) = -1/\log(\delta) \quad (\delta < 1)$. By consequence, we have $E_0 = \emptyset$ and μ_3 does not admit r-regularity at the point 0.

4 Proofs

Proof of Theorem 2.1 :
We have

$$H_n(x) = R_n(x)\sqrt{\frac{\mu(B_{nq})}{h_n}}\frac{1}{\sqrt{l(x)}}$$

with

$$R_n(x) = \sqrt{nh_n}\,(f_h(x) - l(x))\sqrt{\frac{h_n}{\mu(B_{nq})}}.$$

As x is a Lebesgue point with $l(x) > 0$, Lemma 4.1 gives us the result. □

Lemma 4.1. *Under assumptions of Theorem 2.1, we have*

$$R_n(x) \xrightarrow{L} \mathcal{N}(0,1).$$

Proof of Lemma 4.1 :
We have

$$R_n(x) = S_n(x) + P_n(x)$$

where

$$S_n(x) = \frac{\nu_{nq} - n\mu(B_{nq})}{\sqrt{n\mu(B_{nq})}}$$

and

$$P_n(x) = \sqrt{nh_n}\sqrt{\frac{h_n}{\mu(B_{nq})}}\left(\frac{\mu(B_{nq})}{h_n} - l(x)\right)$$

and Lemmas 4.2 and 4.3 achieve the proof. □

Lemma 4.2. *Under assumptions of Theorem 2.1, we have*

$$S_n(x) \xrightarrow{L} \mathcal{N}(0,1).$$

Proof of Lemma 4.2 :
The random variable ν_{nq} follows a binomial distribution with parameters n and $\mu(B_{nq})$. So, central limit theorem (Papoulis and Pillai, 2002) concludes the proof. □

Lemma 4.3. *Under assumptions of Theorem 2.1, we have*

$$P_n(x) \to 0.$$

Proof of Lemma 4.3 :
Using the definition of $P_n(x)$ and relation (2.1), we have that, for all $r \in]0; r_x[$ and as soon as $h_n < \lambda_0$,

$$|P_n(x)| \leq \frac{\sqrt{n} h_n^{r+1} C_r}{\sqrt{\mu(B_{nq})}}.$$

As x is a Lebesgue point we get, for all $r \in]0; r_x[$,

$$|P_n(x)| \leq \frac{\sqrt{n} h_n^{r+1/2} C_r}{\sqrt{f(x)}}.$$

Then, as $f(x) > 0$ and according to condition (2.3), we obtain the lemma.□

Acknowledgements : The authors thank the associate editor and the anonymous referee for relevant remarks and constructive comments on a previous version of the paper.

References

Beirlant, J., Berlinet, A. and Biau, G. (2008). Higher order estimation at Lebesgue points. *Annals of the Institute of Statistical Mathematics* **60** 651-677. MR2434416

Berlinet, A. and Levallois, S. (2000). Higher order analysis at Lebesgue points. In *Asymptotics in Statistics and Probability in G.G. Roussas Festschrift* (M. L. Puri, ed.), VSP Leiden Netherlands, 1-16.

Berlinet, A. and Servien, R. (2011) Necessary and sufficient condition for the existence of a limit distribution of the nearest neighbor density estimator. *Journal of Nonparametric Statistics* **23** 633-643. MR2836281

Berlinet, A. and Servien, R. (2012) Empirical estimator of the regularity index of a probability measure. *Kybernetika* **48** 589-599. MR3013392

Bosq, D. and Lecoutre, J.-P. (1987) *Théorie de l'Estimation Fonctionnelle.* Economica, Paris.

Devroye, L. and Györfi, L., (1985) *Nonparametric density estimation: the L^1 view*. Wiley series in probability and mathematical statistics, Wiley, New-York. MR0780746

Dudley, R. M. (2002) *Real Analysis and Probability*. Cambridge University Press, Cambridge. MR1932358

Ioannidis, Y. E. (2003) The History of Histograms (abridged). *Proceedings of VLDB Conference* 19-30.

Papoulis, A. and Pillai, S.U. (2002) *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New-York.

Rudin, W. (1987) *Real and Complex Analysis*. McGraw-Hill, New-York. MR0924157

Servien, R. (2010) *Estimation de régularité locale*. Université Montpellier II, Montpellier.

Stadtmüller, U. (1983) Asymptotic distributions of smoothed histograms. *Metrika* **30** 145-158. MR0726014

T.Laloë
Université de Nice Sophia-Antipolis
France
E-mail: tlaloe@unice.fr

R.Servien
INRA Toulouse-ENVT
France.
E-mail: remi.servien@toulouse.inra.fr