1 Introduction

The Glivenko-Cantelli theorem is discussed in details in [1] section 36. Nice on-line presentations of the χ^2 and Kolmogorov-Smirnov statistics can be found from OpenCourseWare [2].

Glivenko-Cantelli theorem

The theorem of Glivenko-Cantelli is often referred to in the literature as the basis for mathematical statistics. It allows us to estimate whether data collected from empirical observations pertain to the realizations of a certain random variable. Let the collection of numbers $\{\tilde{x}_i\}_{i=1}^n$ be the result of re-iterated observations of an indicator $\mathcal{X}$. Let us also suppose that they are in ascending order

$$\tilde{x}_1 \leq \tilde{x}_2 \leq \ldots \leq \tilde{x}_n$$

Such data arrangement is often referred to as *order statistics* (*variational series* in the Russian literature).

Definition 1.1 (*Empirical cumulative distribution*)**.** *We call the empirical distribution of an order statistics the step function*

$$F_n[\mathcal{X}](x) = \frac{1}{n}\sum_{i=1}^{n} \chi_{(-\infty\,,\tilde{x}_i]}(x)$$

with χ_A the characteristic function of the set A. By construction the empirical cumulative distribution is monotonic, continuous from the left with jumps corresponding to *differing* entries of the order statistics. If $x_\star$ is the location of a jump, the size of this latter is a multiple of $1/n$ equal to the number of entries in the order statistics with value $x_\star$ If the entries in the order statistics are indeed the realization of n samples of a random variable

$$\xi : \Omega \to \mathbb{R}$$

upon setting

$$F_\xi(x) = P(\xi \leq x) \tag{1.1}$$

we find that for $\{\xi_i\}_{i=1}^n$ a sequence of i.i.d. random variables $\xi_i \overset{d}{=} \xi$ we have

$$P\left(F_n[\xi](x) = \frac{k}{n}\right) = \frac{\Gamma(n+1)\,\left[F_\xi(x)\right]^k\,\left[1-F_\xi(x)\right]^{n-k}}{\Gamma\left(k+1\right)\Gamma\left(n-k+1\right)}$$

where now

$$F_n[\xi](x) = \frac{1}{n}\sum_{i=1}^{n} \chi(\xi_i \leq x)$$

is the *empirical distribution* associated to the sequence $\{x_i\}_{i=1}^n$. In the simplest case when

$$\xi : \Omega \to x_1\,, \ldots\,, x_s$$

the set of frequencies

$$\tilde{\boldsymbol{p}} = \left(\frac{m_1}{n}\,, \ldots\,, \frac{m_s}{n}\right)$$

with

$$m_i = \text{number of observations of } x_i \text{ out of a total of } n \text{ observations}$$

the strong law of large numbers ensures us that

$$\lim_{n\uparrow\infty} \frac{m_i}{n} = P_\xi(\xi = x_i) \qquad a.s.$$

In general the following result holds true

Theorem 1.1 (*Glivenko-Cantelli*)**.** *Let $F_\xi(x)$ the cumulative distribution function of a a random variable ξ and $F_n[\xi](x)$ the cumulative empirical distribution associated to a sequence of i.i.d. copies of ξ. Then we have*

$$\lim_{n\uparrow\infty} \sup_x |F_n[\xi](x) - F_\xi(x)| = 0 \qquad a.s.$$

Proof. Let us distinguish two cases

- $F_\xi(x)$ is continuous (and by definition bounded). Define then for any fixed x

  $$\eta = \chi_{(-\infty\,,x)}(\xi)$$

  so that

  $$\prec \eta \succ = \prec \eta^4 \succ = F_\xi(x)$$

  The strong law of large numbers then implies for i.i.d. copies of η

  $$\lim_{n\uparrow\infty} \frac{1}{n} \sum_{i=1}^{\infty} \eta_i = \prec \eta \succ \qquad a.s.$$

  which yields the proof of the claim by the assumed continuity of F_ξ.

- Suppose now that F_ξ is only left continuous with a countable number of finite jumps. In such a case

  $$\prec \eta \succ = F_\xi(x-)$$

  and the strong la of large numbers yields

  $$\lim_{n\uparrow\infty} \frac{1}{n} \sum_{i=1}^{\infty} F_n[\xi](x) = F_\xi(x-) \qquad a.s.$$

Let now choose an arbitrary $j \in \mathbb{N}$ and define

$$x_{i\,,j} = \inf \left\{ x \,:\, F_\xi(x) \geq \frac{i}{j} \right\} \qquad 1 \leq i \leq j-1$$

The point-wise convergence of $F_n[\xi](x)$ and $F_n[\xi](x-)$ imply that there is an $N_k(\omega)$ such that for any $n \geq N_k(\omega)$

$$|F_n[\xi](x_{ij}) - F_\xi(x_{ij})| < \frac{1}{j} \qquad \& \qquad |F_n[\xi](x_{ij}-) - F_\xi(x_{ij}-)| < \frac{1}{j} \tag{1.2}$$

By setting

$$x_{0j} = -\infty \qquad \& \qquad x_{jj} = \infty$$

the bounds can be extended to any $0 \leq i \leq j$. Consider now any x in between x_{i-1j} and x_{ij}. The monotonicity of the empirical and of the probability distribution functions allows us to write

$$0 \leq F_\xi(x_{ij}-) - F_\xi(x_{i-1j}) \leq \frac{1}{j} \tag{1.3}$$

which in turn implies

$$F_n[\xi](x) \leq F_n[\xi](x_{ij}-) \leq F_n[\xi](x_{i-1j}) + \frac{1}{j} \leq F_\xi(x_{i-1j}) + \frac{2}{j} \leq F_\xi(x) + \frac{2}{j}$$
$$F_n[\xi](x) \geq F_n[\xi](x_{i-1j}) \geq F_n[\xi](x_{ij}) - \frac{1}{j} \geq F_\xi(x_{ij}) - \frac{2}{j} \geq F_\xi(x) - \frac{2}{j}$$

Gleaning together the above inequalities, we have proved that for *any* x

$$|F_n[\xi](x) - F_\xi(x)| \leq \frac{2}{j} \qquad \Rightarrow \qquad \sup_x |F_n[\xi](x) - F_\xi(x)| \leq \frac{2}{j}$$

The arbitrariness of j yields the proof of the claim. □

The Glivenko-Cantelli theorem is a strong law of large number type result. Bulk fluctuations around this result are taken into account by a *central limit type* result:

$$P\left(\sqrt{n}[F_n[\xi](x) - F_\xi(x)] \leq t\right) \to \int_{-\infty}^{\frac{t}{\sigma(x)}} du\, g_{0,1}(u)$$

with

$$\sigma(x) = F_\xi(x)\,[1 - F_\xi(x)]$$

the variance of $\chi_{(-\infty,x)}(\xi)$.

1.1 Kolmogorov-Smirnov test

The practical importance of the Glivenko-Cantelli theorem for statistics comes from the observation in that the distribution of the supremum of the distance between the empirical and the theoretical cumulative distributions does not depend on the distribution F_ξ of the sample, if F_ξ is a continuous distribution. Theoretical background on the tests is also provided by [3] (chapter 1) and [4] (chapter 14).

Proposition 1.1. *If $F_\xi(x)$ is continuous then the distribution of*

$$\eta_n(x) := \sup_x |F_n[\xi](x) - F_\xi(x)|$$

does not depend on F_ξ .

Proof. Let us define the inverse of F_ξ by

$$F_\xi^{-1}(y) = \min\{x \,:\, F_\xi(x) \geq y\}$$

Using the definition we can write

$$P\left(\sup_x |F_n[\xi](x) - F_\xi(x)| \leq y\right) =$$
$$P\left(\sup_{0\leq y\leq 1} |F_n[\xi](F_\xi^{-1}(y)) - y| \leq t\right) = P\left(\sup_{0\leq y\leq 1} |F_n[F_\xi(\xi)](y) - y| \leq t\right)$$

The random variable $\zeta := F_\xi(\xi)$ is uniform in the unit interval:

$$P\left(F_\xi(\xi) \le y\right) \equiv P\left(\xi \le F_\xi^{-1}(y)\right) = F_\xi\left(F_\xi^{-1}(y)\right) = y$$

hence

$$P\left(\sup_x |F_n[\xi](x) - F_\xi(x)| \le y\right) = P\left(\sup_x |F_n[\zeta](y) - y| \le t\right)$$

where the right hand side is independent of F_ξ. □

The explicit form of the distribution of $\eta(x)$ is given by a theorem proved by Kolmogorov:

Theorem 1.2 (*Kolmogorov*)**.**

$$P\left(\sqrt{n}\sup_x |F_n[\xi](x) - F_\xi(x)| \le y\right) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2\,i^2\,y}$$

We can now use the above result to set up a statistical test. We are given a set of data organized in an order statistic and we wan to test two hypotheses

- null hypothesis (H_0): the unknown cumulative distribution F_ζ to which the data pertain *is equal* to F_ξ
- alternative hypothesis (H_1): the unknown cumulative distribution F_ζ *is not equal* to F_ξ

If the null hypothesis holds true in the large n limit $F_n[\mathcal{X}](x)$ converges to $F_\xi(x)$ and

$$\eta_n(x) \to 0 \qquad a.s.$$

On the other hand if H_0 fails even for large n

$$\eta_n(x) > \delta > 0 \qquad \Rightarrow \qquad \sqrt{n}\eta_n(x) > \sqrt{n}\delta \uparrow \infty$$

Definition 1.2 (*Kolmogorov-Smirnov statistic*)**.** *If F_ξ is the conjectured cumulative distribution the Kolmogorov-Smirnov statistic is*

$$\mathsf{KS}_n := \sqrt{n}\sup_x |F_n[\mathcal{X}](x) - F_\xi(x)|$$

A *decision rule* $\mathcal{D}$ can be defined by choosing a *threshold value* $\mathtt{t}$ so that

$$\mathcal{D} = \begin{cases} H_0 & \text{if} \quad \sqrt{n}\,\eta_n(x) \le \mathtt{t} \\ H_1 & \text{if} \quad \sqrt{n}\,\eta_n(x) > \mathtt{t} \end{cases}$$

The threshold value is determined by the required *level of significance* α of the decision rule

$$\alpha = P(\mathcal{D} = H_0) = P\left(\sqrt{n}\,\eta_n(x) \le \mathtt{t}|H_0\right)$$

For finite n we can tabulate the probability in (1.4) using its very definition, in the large n limit we can resort to Kolmogorov's theorem. In summary:

- Formulate a conjecture for the data cumulative distribution, say F_ξ.
- Compute the Kolmogorov-Smirnov statistical indicator KS_n.
- Fix a significance level, i.e. the probability that it is consistent to accept the null hypothesis.
- Use the significance level to determine the acceptance threshold $\mathtt{t}$.
- Compare KS_n with $\mathtt{t}$.

2 Empirical description of a probability density

The histogram is a graphical tool to display the content of an order statistics. Let us suppose that we gathered n observations of an indicator $\mathcal{X}$ which we conjecture to be modeled by a continuous random variable ξ

$$\xi : \Omega \to [x_{min}\,, x_{max}]$$

we can use these observation to approximate the probability density p_ξ by means of simple functions over $I = [x_{min}\,, x_{max}]$. This can be done e.g. by defining a uniform partition of I in n_ℓ bins of width ℓ so that

$$|I| = n_\ell\, \ell$$

The i-th bin of the partition the interval $x \in (x_{i-1}\,, x_i]$ with $i = 1\,, n_{ell}$ and $x_0 = x_{min}$, $x_{n_\ell} = x_{max}$. We can then count the number m_i of events coorresponding to values of x in the range of the $i - th$ bin. We have

$$n = \sum_{i=1}^{n_\ell} m_i \qquad \Rightarrow \qquad 1 = \sum_{i=1}^{n_\ell} \frac{m_i}{n}$$

The graph of $(i\,, m_i)$ is called the (uniform) histogram of $\mathcal{X}$. In order to define a discrete approximation to a continuous function we further observe that

$$\sum_{i=1}^{n_\ell} \frac{m_i}{n} = \sum_{i=1}^{n_\ell} \ell \left(\frac{m_i}{n\,\ell}\right) \tag{2.1}$$

We can interpret the right hand side of (2.1) as a Riemann sum where

$$dx \sim \ell = \frac{|I|}{n} \qquad \& \qquad p_\xi\,(\bar{x}_i) = \frac{m_i}{n\,\ell} \quad \bar{x}_i \in (x_{i-1}\,, x_i]$$

From the algorithmic point of view, an histogram is efficiently constructed as follows Two observation are in order:

Algorithm 1 histogram algorithm

READ ℓ	(set bin width)
$n_\ell = \lceil (x_{max} - x_{min})/\ell \rceil$	(compute the integer part e.g. by taking the ceiling)
for $i = 1, n_\ell$ **do**	
$histogram[i] = 0$	(initialize histogram bins)
end for	
for $i = 1, n$ **do**	
READ $data[i]$	(store the i-th empirical data into data[i])
$j = \lceil (data[i] - x_{min})/\ell \rceil$	
$histogram[j] = histogram[j] + 1/(n * \ell)$	
end for	
stop	

- The accuracy of the approximation depends upon the choice of the bin width. This latter is in general chosen taking into account the size of the observation sample n and the known statistical properties of the data. In some cases non uniform partitions may be more attuned to encode the data.
- If the distribution is expected to have infinite support, tail events can be included in dedicated bin of infinite width e.g. $x \leq x_{min}$ or $x \geq x_{max}$.

References

[1] B.V. Gnedenko, *Theory of probability*, Publisher: CRC Press. 1

[2] MIT OpenCourseWare: *18.443 Statistics for Applications: lectures L11 and L14*, http://ocw.mit.edu/OcwWeb/Mathematics/18-443Fall-2006/LectureNotes/index.htm 1

[3] P. E. Kloeden, E. Platen, H. Schurz, *"Numerical Solution Of Sde Through Computer Experiments"*, Springer (Universitext) (2003) preview from http://books.google.com/. 3

[4] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery,
Numerical recipes in C: the art of scientific computing
Cambridge University Press (1992) and
http://www.fizyka.umk.pl/nrbook/bookcpdf.html,
see also http://www.nr.com. 3