Statistics and Scipy Tools

Daniel de Jesus

April 2022

1 Introduction

In this document, I'm going to resume the main tools of Scipy library. This material is indicated to those who the python documentations do not seems strange. For beginners, I recommend the python basics courses, some are listed below.

Neste documento, eu irei resumir algumas das principais ferramentas da biblioteca Scipy. Este material é indicado àqueles para os quais a documentação do python não pareça estranha. Para iniciantes, eu recomendo os cursos básicos, alguns são listasdos abaixo.

Curso Python Essencial para Data Science no canal xavecoding - prof. samuka no youtube [1].

Curso Pandas Essencial no canal xavecoding - prof. samuka no youtube [2].

2 Scipy probability

Scipy is organized into subpackages covering different scientific computing domains.

- cluster,
- constants,
- fftpack(fast Fourier transform routines),
- integrate(and ODE),
- interpolate(splines),
- io(in-output),
- linalg(Linear algebra),
- ndimage(n-dimensional image processing),
- odr(orthogonal distance regression),
- optimize(and rout-finding routines),
- signal(Signal processing),
- sparse(matrix),
- spatial(spatial data structure and algorithms),
- special,
- stats(Statistical distributions and functions!!)

From the scipy oficial cite[3].

3 First contact

The usage of the Scipy library has the common head command

from scipy.stats import binom.

Where scipy.stats is the subpackage to binomial statistical class. We can use the objects from this class, the scipy.stats objects has a standard usage:

subpackage: stats

object: binom

features: cdf, pdf, pmf, ppf,(...)

Hence, for usage of the binomial pmf we perform the command line:

MyDistribution= binom.pmf(x,n,p)

The PMF, Probability Mass Function, is the distribution when we have discrete data values.

The PDF, Probability density function,is the distribution when we have continuous data values. The CDF, Cumulative Density Function, is the sum of a specific PMF or the integral of PDF. After calling to a specific distribution, we can use the features described earlier, as shown in the examples above.

MyDiscreteDistribution= binom.pmf(x,n,p)
MyContinuousDistribution= norm.pdf(x,n,p)
MyCumulativeDistribution= binom.cdf(x,n,p)

4 Fundamental Statistical background

4.1 Conditional probability

4.1.1 Classical definition

In any event of equal probability the traditional probabibily is given by:

$$P(A) = \frac{Results_{Desired}}{Results_{Total}} \tag{1}$$

This definition can be used when the probabilities are the same for each outcome. As application example , we have the toss-a-coin and roll the dice experiment [4, 5].

4.1.2 Frequency definition

In a event where the probabilities are different with a known frequency, the use of the previous definition consist of an conceptual error and should be avoided. In such case, we have a distribution of probability better suitable for each specific situation. In next section, I'm going to discuss the statistical frequency and probability distribution function definition by its usage [4, 5].

5 PROBABILITY DISTRIBUTIONS

5.1 Bernoulli Distribution

This distribution is indicated when ... [6]

This is the simplest distribution used as a benchmark for further distributions.

5.1.1 The shape:

$$\text{P(X= x)} = \begin{cases} 1-p & x=0 \\ p & x=1 \end{cases}$$

$$E(X) = p$$

$$Var(X) = pq = p(1-p)$$

where x is the number of successes, p the probability of success, $q = (1-p)$ the probability of failure.

5.1.2 Conditions under application

1. Unic tryal
2. Outcome is success or failure

5.2 Binomial Distribution

This distribution is indicated when ... [6]

5.2.1 The shape:

$$P(X = x) = \binom{n}{x} p^x q^{n-x}$$

$$\binom{n}{x} = \frac{n!}{(n-x)!x!}$$

$$E(X) = np$$

$$Var(X) = \sqrt{npq} = \sqrt{np(1-p)}$$

where n is the number of trials, x number of successes, p the probability of success, $q = (1-p)$ the probability of failure.

5.2.2 Conditions under application

1. The experiment performed n fixed times.
2. The tryals are independent in identical conditions.
3. Outcome is success or failure.
4. Success probability in one tryal is known(relative frequency) p.
5. Random variable X is the success count.

5.2.3 Python example of Binomial distribution

importing the binomial object:

$$from\ scipy.stats\ import\ binom$$

Now, lets perform the binomial probability distribution. The probability of score 5 in 10 question test, with 3 alternatives each question betting.

$x = 5$ — target value
$p = 1/3$ — frequency
$n = 10$ — tryals

$$prob1 = binom.pmf(x, n, p).$$

Desired probability of pass using CDF, $score >= 5$

$$probpass = 1 - binom.cdf(4, n, p).$$

Desired probability of pass using SURVIVOR FUNCTION (1-CDF).

$$probpass = binom.sf(4, n, p).$$

5.3 Poisson Distribution

This distribution is indicated when ... [6]

5.3.1 The shape:

$$P(X = x) = \frac{e^{\lambda}\lambda^{x}}{x!}$$

$$E(X) = \lambda$$

$$Var(X) = \lambda$$

5.3.2 Conditions under application

1. Success probability in one tryal is known(relative frequency) p.
2. The success frequency are the same for fixed length intervals..
3. The success probability approach to zero with interval tends to zero.
4. Outcome is success or failure.
5. The events are independents..

5.3.3 Python Poisson Distribution Example

$$from\ scipy.stats\ import\ poisson$$

importing the Poisson object

Now, lets perform the Poisson probability distribution

A restaurant receives 20 orders per hour. What is the chance that, at a given hour chosen at random, the restaurant will receive 15 orders?

x= 15
λ= 20

$$prob = poisson.pmf(x, \lambda)$$

5.4 Normal Distribution

This distribution is indicated when ... [6]

5.4.1 The shape:

$$P(X = x) = ggggg$$

$$E(X) = dddd$$

$$Var(X) = dddd$$

5.4.2 Conditions under application

1. The experiment performed .
2. The tryals are .
3. Outcome is .
4. Success probability in .
5. Random variable X is .

5.4.3 Python example of normal distribution

importing the normal object:

from scipy.stats import norm

Desired probability of pass using CDF, *score* >= 5
Desired probability of pass using SURVIVOR FUNCTION (1-CDF).

6 Statistical Inference

The sampling process does not guarantee the replication of a population, in fact, different samples may have a different distributions and the relation between the sample and the populations is no so simple to understand. Steady of use the distribution and the statistical basic numbers to describe some characteristic of the population another parameter is needed. This parameter has to carry the information of this uncertainty. That's the meaning of interval of confidence and significance level.

6.1 The central limit theorem

Consider a random experiment and the probability distribution over a variable of interest. If we perform the independent experiments, and get a set of probability distribution, one for each aleatory and independent experiment, we may want to understand the statistical behavior of the distributions. The mean of the sample is closely to a normal, or Gaussian, distribution. The approximation to a normal distribution well-fits with the increase of the sample size.

> Theorem: *For a random experiment with a well defined distribution, the distribution of sample mean is close to a normal distribution and gets closer to a normal distribution with the increase of sample size [4]*

6.2 Confidence and significance level

1. 95 % confidence
2. higher confidence
3. Proportion confidence
4. Significance level
5. Confidence interval comparison
6. Sample Dimensioning

References

[1] Samuel Botter Martins. Python essencial para data science.

[2] Samuel Botter Martins. Pandas essencial.

[3] Scipy Org. Numpy and scipy documentation.

[4] Silvia Shimakura. Notas de aula do curso de bioestatística.

[5] Charles Wheelan. *Estatística: o que é, para que serve, como funciona.* Editora Schwarcz-Companhia das Letras, 2016.

[6] Marcos Nascimento Magalhães and Antônio Carlos Pedroso De Lima. *Noções de probabilidade e estatística*, volume 5. Editora da Universidade de São Paulo, 2002.