Kiểm định phân bố chuẩn (test for normal distribution)

Đặng Khải Hoàn

2/25/2020

Khi chúng ta muốn kiểm định sự khác nhau của 2 phân nhóm hay trong việc phân tích các mô hình hồi quy, thì luôn có một giả thiét rất quan trọng đó là các biến phân tích phải có phân phối chuẩn. Do đó bài này sẽ giúp chúng ta việc đó.

Load Dataset.

rexam = read.delim("D://Data files/Data files/RExam.dat",header = TRUE)

head(rexam)

## exam computer lectures numeracy uni

## 1 18 54 75.0 7 0

## 2 30 47 8.5 1 0

## 3 40 58 69.5 6 0

## 4 30 37 67.0 6 0

## 5 40 53 44.5 2 0

## 6 15 48 76.5 8 0

1.Kiểm định phân phối chuẩn bằng biểu đồ.

**library**(ggplot2)

hist.exam = ggplot(data = rexam,aes(exam))+

geom\_histogram(aes(y=..density..),color="black",fill="white")+

stat\_function(fun = dnorm,

args = list(mean = mean(rexam$exam,na.rm = TRUE),

sd = sd(rexam$exam,na.rm = TRUE)),

color ='blue',size = 1)

hist.exam

Dựa vào biểu đồ trên ta thấy biểu đồ histogram với đường cong biểu diễn cho phân phối chuẩn không khớp với nhau nên ta có thể nói biến exam không tuân theo phân phối chuẩn.

Và một cách cũng dùng biểu đồ để kiểm định phân bố chuẩn đó là QQ-plot.

**library**(ggpubr)

qqplot1 = ggqqplot(rexam$exam)

qqplot1

Biểu đồ QQ-plot cho ta thấy những giá trị quan sát đa phần không nằm trên đường thẳng kì vọng của phân phối chuẩn do đó biến exam không có tuân theo phân bố chuẩn.

Một ví dụ khác :

hist.computer = ggplot(data = rexam,aes(lectures))+

geom\_histogram(aes(y=..density..),color="black",fill="white")+

stat\_function(fun = dnorm,

args = list(mean = mean(rexam$lectures,na.rm = TRUE),

sd = sd(rexam$lectures,na.rm = TRUE)),

color ='blue',size = 1)

hist.computer

Biểu đồ trên cho ta thấy sự trùng khớp đa phần của biểu đồ histgram và đường cong của phân bố chuẩn nên ta có thể kết luận biến lectures tuân theo luật phân bố chuẩn.

Và để yên tâm hơn ta sẽ kiểm tra lại bằng QQ-plot.

ggqqplot(rexam$lectures)

Và biểu đồ QQ-plot cũng cho kết luận biến lectures cũng tuân theo luật phân bố chuẩn.

2 Sử dụng kiểm định Shapiro-Wilk:

Phát biểu giả thuyết thống kê:

H0H0: Biến cần kiểm định tuân theo phân phối chuẩn.

H1H1: Biến cần kiểm định không tuân theo phân phối chuẩn

Giá trị của thống kê Shapiro-Wilk:

W=(∑ni=1aixi)2∑ni=1(xi−x¯¯¯)2W=(∑i=1naixi)2∑i=1n(xi−x¯)2.

Với: xi:xi: giá trị thứ i nhỏ nhất của x.

ai:ai: Hằng số Shapiro - Wilk.

Xem cụ thể ở đây: “<http://www.real-statistics.com/statistics-tables/shapiro-wilk-table/>”.

Code trong R:

shapiro.test(rexam$exam)

##

## Shapiro-Wilk normality test

##

## data: rexam$exam

## W = 0.96131, p-value = 0.004991

Nhìn vào p-value = 0.004991 < 0.05.Ta có thể bác bỏ giả thuyết H0H0 hay nói cách khác biến exam không tuân theo luật phân bố chuẩn.

shapiro.test(rexam$lectures)

##

## Shapiro-Wilk normality test

##

## data: rexam$lectures

## W = 0.97698, p-value = 0.07712

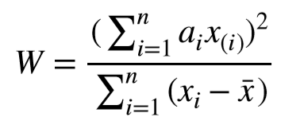
Nhìn vào p-value = 0.07712 > 0.05.Nên ta không đủ cơ sở bác bỏ giả thuyết H0H0 hay biến lectures tuân theo luật phân bố chuẩn.

## Shapiro-Wilk Test

When the distribution of a real valued continuous random variable is unknown, it is convenient to assume that it is normally distributed. However, this may not always be true leading to incorrect results. To avert this problem, there is a statistical test by the name of Shapiro-Wilk Test that gives us an idea whether a given sample is normally distributed or not. The test works as follows:

Specify the null hypothesis and the alternative hypothesis as:  
H0 : the sample is normally distributed  
HA : the sample is not normally distributed

A test statistic is computed as follows:



This W is also referred to as the Shapiro-Wilk statistic W (W for Wilk) and its range is 0<W 1. The lower bound on W is actually determined by the size of the sample. Normally distributed samples will result in a high value of W and samples deviating away from a normal distribution will have a lower value of W. Based on the value of W, we accept or reject the null hypothesis. Accepting the null hypothesis implies that we have suﬃcient evidence to claim that our data is normally distributed. Likewise, rejecting the null hypothesis in favor of the alternate hypothesis means that our data sample does not provide us suﬃcient evidence to claim that the sample is normally distributed.

In the expression, is the sample mean, x(i) is the ith smallest value in the given sample x (also called order statistic). ai are coeﬃcients computed from the order statistics of the standard normal distribution. Let’s now apply this test in R.

## Shapiro-Wilk Test in R

In R, the Shapiro-Wilk test can be applied to a vector whose length is in the range [3,5000]. At the R console, type:

> shapiro.test(x)

You will see the following output:

Shapiro-Wilk normality test

data: x

W = 0.99969, p-value = 0.671

The function shapiro.test(x) returns the name of data, W and p-value. Let us now talk about how to interpret this result.

## **P-values**

When looking at the p-values, there are diﬀerent guidelines on when to accept or reject the null hypothesis, (recall from our earlier.discussion that the null hypothesis states that the sample values are normally distributed). Generally we compare the p-value with a user deﬁned level of signiﬁcance denoted by alpha or a and make a decision as:

If p > a then accept H0  
If p </= a then reject H0 in favor of HA.

The question remains on what should be the value of a . Depending upon your application you can choose a diﬀerent signiﬁcance level, e.g., 0.1, 0.05, 0.01 etc.. Michael Baron in his book: “[Probability and Statistics for Computer Scientists](https://www.amazon.com/Probability-Statistics-Computer-Scientists-Michael/dp/%201584886412)” recommends choosing an alpha in the range [0.01, 0.1]. So for most applications you can safely accept H0 if p > 0.1 and safely reject H0 if p<0.01. For values of p in this range [0.01,0.1], it may be a good idea to collect more data if your application is a critical one. In the example above x is randomly sampled from a normal distribution and hence we get a p-value of 0.671 and we are sure to accept the null hypothesis that x is normally distributed.

3.Sử dụng kiểm định Anderson-Darling:

Giá trị của thống kê AD:

AD=−n−1n∑ni=1(2i−1)[lnF(xi)+ln(1−F(xn−i+1))]AD=−n−1n∑i=1n(2i−1)[lnF(xi)+ln(1−F(xn−i+1))].

Với F(x) :hàm phân phối tích lũy của biến cần phân tích.

ithith Giá trị nhỏ nhất thứ i.

Code R:

**library**(nortest)

ad.test(rexam$exam)

##

## Anderson-Darling normality test

##

## data: rexam$exam

## A = 1.4031, p-value = 0.001181

Dựa vào p-value = 0.001181 < 0.05.Nên ta có thể bác bỏ giả thuyết H0H0 hay biến exam không tuân theo luật phân bố chuẩn.

ad.test(rexam$lectures)

##

## Anderson-Darling normality test

##

## data: rexam$lectures

## A = 0.49945, p-value = 0.2048

Nhìn vào p-value = 0.2048 > 0.05.Nên ta có không đủ cơ sở bác bỏ H0H0 hay biến lectures tuân theo luật phân bố chuẩn.