KIỂM TRA NGẪU NHIÊN HÀNG HÓA XUẤT NHẬP KHẨU TRƯỜNG HỢP MẪU NHỎ

Võ Minh Trí [1], Ung Ngọc Quang [2]
Dương Tôn Đảm [2], Tô Anh Dũng [2], Nguyễn Minh Hải [2]
[1] Cục Hải Quan TP.Hồ Chí Minh
[2] Trường Đại học Khoa học Tự nhiên TP.HCM

TÓM TẮT : Bài này nối tiếp bài [1]. Trong bài, các tác giả khảo sát vấn đề kiểm tra ngẫu nhiên hàng hóa xuất nhập khẩu trong Hải quan cho trường hợp mẫu nhỏ.
Từ khóa : Đại lượng nhị thức, kiểm định giả thuyết, mẫu nhỏ.

1. ĐẶT VẤN ĐỀ

Bài báo [1] đã giải quyết bài toán kiểm tra ngẫu nhiên hàng hoá xuất nhập khẩu trong trường hợp hàng hoá được đóng theo kiện và các kiện hàng là đồng nhất.

Nhắc lại rằng, trong [1] vấn đề kiểm tra ngẫu nhiên hàng hóa xuất nhập khẩu đã đưa về bài toán kiểm định giả thuyết thống kê theo tham số tỷ lệ có dạng như sau :

$$\begin{cases} H : p = p_0 \\ K : p > p_0 \end{cases}, \text{ với } H \text{ là giả thuyết}, K \text{ là đối thuyết}.$$

Trong đó, p là tỷ lệ kiện hàng "Khai báo đúng" trong toàn bộ lô hàng và p_0 là tỷ lệ giả thuyết về kiện hàng "Khai báo đúng" trong toàn bộ lô hàng.

Lời giải bài toán này dựa trên việc kiểm tra ngẫu nhiên n kiện hàng từ lô hàng. Trong [1] nếu n và p_0 thỏa điều kiện :

$$\begin{cases} np_0 \geq 5 \\ n(1-p_0) \geq 5 \end{cases} \quad (*)$$

thì n được gọi là cỡ mẫu lớn và bài toán là trường hợp mẫu lớn. Bài báo [1] đã giải quyết xong bài toán kiểm tra ngẫu nhiên hàng hóa xuất nhập khẩu đối với trường hợp mẫu lớn.

Khi cỡ mẫu n không thỏa điều kiện (*) thì bài toán được gọi là trường hợp mẫu nhỏ. Bài báo này giải quyết vấn đề kiểm tra hàng hóa xuất nhập khẩu cho trường hợp mẫu nhỏ.

Cần nhấn mạnh rằng, hiện nay trong việc kiểm tra hàng hóa xuất nhập khẩu của Hải quan, người ta thường gặp bài toán với cỡ mẫu nhỏ. Vì vậy việc giải quyết bài toán này là vấn đề thời sự.

Để giải quyết bài toán mẫu nhỏ nêu trên, trước hết ta xét một số vấn đề có liên quan tới đại lượng ngẫu nhiên nhị thức $B(n,p)$ (xem [2], [3]).

2. ĐẠI LƯỢNG NHỊ THỨC VÀ BÀI TOÁN KIỂM ĐỊNH GIẢ THUYẾT VỀ TỶ LỆ CHO TRƯỜNG HỢP MẪU NHỎ

2.1. Định nghĩa

Định nghĩa 2.1.1 Xét dạng thí nghiệm ngẫu nhiên chỉ xảy ra 2 loại biến cố là A và A^C. Đặt $p = P(A)$ và $q = P(A^C) = 1 - p$. Lặp lại thí nghiệm ngẫu nhiên nói trên n lần độc lập và đặt S_n= "là số lần xuất hiện biến cố A trong n lần thí nghiệm".Khi ấy S_n là một đại lượng ngẫu nhiên rời rạc có tập giá trị $\{0, 1, 2, \cdots, n\}$ và được gọi là đại lượng nhị thức. Ký hiệu $S_n \sim B(n,p)$. Luật phân phối này có dạng :

$$P\{S_n = x\} = C_n^x p^x q^{n-x}, \ x = 0, 1, \cdots, n \text{ (xem [2], [3])}.$$

Trong thực tế, đôi khi người ta quan tâm tới tần suất xuất hiện biến cố A hơn là quan tâm tới số lần xuất hiện biến cố A. Do đó người ta đưa ra một đại lượng ngẫu nhiên mới, liên quan với $B(n,p)$, như sau.

2.2. Định nghĩa: Cho đại lượng ngẫu nhiên nhị thức $S_n \sim B(n,p)$. Đặt $f_n = \dfrac{S_n}{n}$. Người ta gọi f_n là đại lượng ngẫu nhiên nhị thức theo tần suất. Hiển nhiên f_n là một đại lượng ngẫu nhiên rời rạc có tập giá trị $\left\{0, \dfrac{1}{n}, \dfrac{2}{n}, \cdots, 1\right\}$.

Chú ý rằng, việc chia đại lượng ngẫu nhiên S_n cho hằng số n không làm thay đổi phân phối xác suất của S_n. Vậy nên phân phối xác suất của f_n có dạng như sau :

$$P\left\{f_n = \frac{x}{n}\right\} = C_n^x p^x q^{n-x} \quad , \quad x = 0, 1, \cdots, n .$$

Ta sẽ dùng đại lượng nhị thức tần suất f_n để giải quyết bài toán kiểm định giả thuyết cho tham số tỷ lệ trong trường hợp mẫu nhỏ.

2.2. Bài toán kiểm định giả thuyết trong trường hợp mẫu nhỏ

Xét các đại lượng nhị thức $S_n \sim B(n,p)$ và với p là tham số tỷ lệ chưa biết và n là cỡ mẫu nhỏ.

Đặt p_0 là tỷ lệ giả thuyết cho trước, $p_0 \in (0,1)$. Khi ấy bài toán kiểm định giả thuyết cho tham số tỷ lệ p có dạng :

$$\begin{cases} H : p = p_0 \\ K : p > p_0 \end{cases}$$, với H là giả thuyết, K là đối thuyết.

Cho $\alpha > 0$ là xác suất sai lầm loại I, tức là xác suất bác bỏ giả thuyết H khi H đúng ($\alpha > 0$ còn gọi là mức ý nghĩa của bài toán kiểm định).

Cho $G = G(X_1, X_2, \ldots, X_n)$ là một thống kê được xác định từ mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$. Ký hiệu W_α là miền bác bỏ giả thuyết H với mức ý nghĩa α. Trong bài toán này thì

$$G = G(X_1, X_2, \ldots, X_n) \text{ chính là } f_n,\ W_\alpha = \left\{f_n \ge \frac{x}{n}\right\} \text{ và } H \text{ là } p = p_0 .$$

Khi ấy, người ta đã chứng minh rằng bài toán kiểm định giả thuyết có lời giải tối ưu được viết dưới dạng sau :

Bác bỏ H khi $x \ge x_C$ và chấp nhận H khi $x < x_C$ (xem [3], chương VII).

Trong đó, hằng số x_C được xác định bởi các đại lượng f_n, W_α, H như sau :

$$\begin{aligned} P\{G \in W_\alpha \mid H\} &= P\left\{f_n \ge \frac{x_C}{n} \mid p = p_0\right\} \\ &= \sum_{\frac{x}{n} = \frac{x_C}{n}}^{1} C_n^x p_0^x (1-p_0)^{n-x} \\ &= P\{S_n \ge x_C \mid p = p_0\} \end{aligned}$$

Với n và p_0 cho trước có thể tìm được số nguyên nhỏ nhất x_C (bằng cách tra bảng phân phối nhị thức) đảm bảo cho xác suất mắc sai lầm loại một nhỏ hơn hoặc bằng α.

Để thấy rõ ý nghĩa của lời giải trên, ta xét thí dụ sau.

2.3. Thí dụ

Một máy gia công một loại chi tiết kỹ thuật có tỷ lệ phế phẩm 10 %. Kiểm tra ngẫu nhiên 15 chi tiết thấy có 6 phế phẩm. Với xác suất sai lầm loại I là α =0.05, có thể cho rằng tỷ lệ phế phẩm của máy đó tăng lên hay không ?

Lời giải: Gọi p là tỷ lệ phế phẩm của máy đó, p_0 là tỷ lệ giả thuyết.

Cho trước $p_0 = 0{,}1$; $\alpha = 0{,}05$; $n = 15$; $x = 6$.

Với các số liệu trên, ta có bài toán kiểm định giả thuyết : $\begin{cases} H : p = 0{,}1 \\ K : p > 0{,}1 \end{cases}$.

Trước hết, kiểm tra điều kiện $^{(*)}$, ta thấy :

$$\begin{cases} np_0 = 15.0{,}1 = 1{,}5 < 5 \\ n(1 - p_0) = 15.0{,}9 = 4{,}5 < 5 \end{cases}$$

Vậy điều kiện $^{(*)}$ không thoả nên bài toán này thuộc trường hợp mẫu nhỏ. Để giải quyết nó, ta sẽ dùng lời giải tối ưu ở mục 2.2 . Bằng cách sử dụng bảng phân phối nhị thức trong [4], với $n = 15$ và $p_0 = 0{,}1$, ta sẽ tìm được miền bác bỏ W_α theo bảng dưới đây :

x	$P\{S_n = x\}$	$P\{S_n \geq x\}$	x	$P\{S_n = x\}$	$P\{S_n \geq x\}$
0	0,2059	1	6	0,0019	0,0022
1	0,3432	0,7941	7	0,0003	0,0003
2	0,2669	0,4509	8	0,0000	0,0000
3	0,1258	0,1840	⋮	⋮	⋮
4	0,0428	0,0555	⋮	⋮	⋮
5	0,0105	0,0127	15	0,0000	0,0000

Để bảo đảm $\alpha \leq 0{,}05$, khi nhìn vào bảng trên, ta thấy phải lấy $\alpha = 0{,}0127$. Nhưng cũng theo bảng trên thì ứng với $\alpha = 0{,}0127$ ta có $P\{S_n \geq 5\}$. Vậy $x_C = 5$. Mặt khác, trong mẫu ta đã có $x = 6$, nên $x > x_C$. Do đó ta phải bác bỏ H, tức là ta nói tỷ lệ phế phẩm p có tăng lên.

3. BÀI TOÁN KIỂM TRA NGẪU NHIÊN HÀNG HOÁ XUẤT NHẬP KHẨU CHO TRƯỜNG HỢP MẪU NHỎ

Mục này nhằm ứng dụng lời giải tối ưu của bài toán kiểm định giả thuyết về tham số tỷ lệ p đã xét trong mục 2.2 vào việc giải quyết vấn đề kiểm tra ngẫu nhiên hàng hoá xuất nhập khẩu của Hải quan cho trường hợp mẫu nhỏ. Trước hết ta có nhận xét sau.

3.1 Nhận xét

Ta thấy lời giải tối ưu trong mục 2.2 chưa thể áp dụng ngay cho vấn đề kiểm tra ngẫu nhiên hàng hoá xuất nhập khẩu. Muốn giải quyết nó, cần chú ý rằng: Tất cả n kiện hàng được lấy ra từ lô hàng để kiểm tra đều phải " Khai báo đúng ". (Vì chỉ cần một kiện hàng " Khai báo sai" thì Hải quan phải kiểm tra lại toàn bộ lô hàng và khi ấy tính ngẫu nhiên không còn nữa, do đó bài toán của ta trở nên vô nghĩa).

Như vậy, ở đây ta có $S_n = n$. Với nhận xét này lời giải cho bài toán kiểm tra ngẫu nhiên hàng hoá xuất nhập khẩu sẽ được viết lại như sau.

3.2 Lời giải tối ưu của bài toán kiểm tra hàng hoá xuất nhập khẩu

Trước hết, bài toán được viết dưới dạng :

$$\begin{cases} H: p = p_0 \\ K: p > p_0 \end{cases}$$, với H là giả thuyết , K là đối thuyết .

Lời giải tối ưu có dạng : Bác bỏ H khi $x \geq x_C$ và chấp nhận H khi $x < x_C$. Trong đó, hằng số x_C được xác định theo công thức :

$$P\{S_n \geq x_C \mid p = p_0\} = \alpha$$

và $$P\{S_n \geq x_C \mid p = p_0\} = \sum_{x=x_C}^{n} C_n^x p_0^x (1-p_0)^{n-x}$$

với α là mức ý nghĩa và n là cỡ mẫu cho trước.

Để thấy rõ ý nghĩa của bài toán, ta xét các thí dụ ứng dụng sau đây.

3.3 Thí dụ

Thí dụ 3.3.1 : Lấy ngẫu nhiên 2 kiện hàng từ một lô hàng ra kiểm tra thì thấy cả hai kiện hàng đều khai báo đúng (tức là ở đây $S_2 = 2$). Có người cho rằng, tỷ lệ kiện hàng khai báo đúng của toàn bộ lô hàng là 50%. Hãy cho kết luận về tỷ lệ kiện hàng khải báo đúng trong toàn bộ lô hàng với xác suất sai lầm loại I là $\alpha = 0.25$.

Lời giải : Đặt p làtỷ lệ kiện hàng " Khai báo đúng " trong toàn bộ lô hàng. Với các số liệu cho trước như trên, ta có bài toán kiểm định như sau:

$$\begin{cases} H: p = 0{,}5 \\ K: p > 0{,}5 \end{cases}$$

Theo mục 2.2, lời giải tối ưu của bài toán có dạng :

Bác bỏ H khi $x \geq x_C$ và chấp nhận H khi $x < x_C$.

Nhờ vào bảng phân phối nhị thức trong [4], ta thiết lập được bảng sau đây để tìm hằng số x_C (chú ý $P\{S_n \geq x \mid p = p_0\} = \alpha$) :

x	$P\{S_n = x \mid p = 0{,}5\}$	$P\{S_n \geq x \mid p = 0{,}5\}$
0	0,25	1
1	0,50	0,75
2	0,25	0,25

Ta thấy rằng, để bảo đảm $\alpha \leq 0{,}25$ thì theo bảng trên ta phải lấy $\alpha = 0.25$.

Nhưng cũng theo bảng trên, ứng với $\alpha = 0{,}25$, ta có $x_C = 2$. Mặt khác, trong mẫu ta có $x = 2$. Vậy $x = x_C$. Nên ta bác bỏ H. Do đó ta chấp nhận K, tức là nói rằng $p > 0{,}5$. Điều này có nghĩa tỷ lệ kiện hàng khai báo đúng của toàn bộ lô hàng cao hơn 50%.

Thí dụ 3.3.2 : Lấy 5 kiện hàng từ một lô hàng ra kiểm tra thì thấy cả 5 kiện hàng đều khai báo đúng (tức là ta có $S_5 = 5$). Có người đưa ra giả thuyết : Tỷ lệ kiện hàng khai báo đúng trong toàn bộ lô hàng là 80%. Với xác suất sai lầm loại I là $\alpha = 0{,}33$, hãy cho biết kết luận về tỷ lệ kiện hàng khai báo đúng trong toàn bộ lô hàng.

Lời giải : Gọi p là tỷ lệ kiện hàng " Khai báo đúng " trong toàn bộ lô hàng. Với các số liệu cho trước như trên, ta có bài toán kiểm định như sau :

$$\begin{cases} H: p = 0{,}8 \\ K: p > 0{,}8 \end{cases}$$

Theo mục 2.2 , lời giải tối ưu của bài toán có dạng :

Bác bỏ H khi $x \geq x_C$ và chấp nhận H khi $x < x_C$.

Nhờ vào bảng phân phối nhị thức trong [4], ta thiết lập được bảng sau đây để tìm hằng số x_C (chú ý $P\{S_n \geq x_C \mid p = p_0\} = \alpha$):

x	$P\{S_n = x \mid p = 0{,}8\}$	$P\{S_n \geq x \mid p = 0{,}8\}$
0	0,003	1
1	0,0064	0,997
2	0,0512	0,9906
3	0,2048	0,9394
4	0,4096	0,7346
5	0,3277	0,325

Chú ý rằng để bảo đảm $\alpha \leq 0{,}33$ thì trong bảng trên ta chỉ có thể lấy $\alpha = 0.325$. Nhưng cũng theo bảng trên, ứng với $\alpha = 0.325$, ta có $x_C = 5$. Mặt khác, trong mẫu ta có $x = 5$. Vậy $x = x_C$. Nên ta bác bỏ *H*. Do đó ta chấp nhận *K*, tức là nói rằng $p > 0{,}8$. Điều này có nghĩa tỷ lệ kiện hàng khai báo đúng của toàn bộ lô hàng cao hơn 80% , với xác suất sai lầm loại I $\alpha = 0{,}33$.

Thí dụ 3.3.3 : Lấy 10 kiện hàng từ một lô hàng ra kiểm tra thì thấy cả 10 kiện đều khai báo đúng (tức là ta có $S_{10} = 10$). Có ý kiến cho rằng tỷ lệ kiện hàng khai báo đúng của toàn bộ lô hàng là 99%. Hãy cho biết kết luận về tỷ lệ kiện hàng khai báo đúng trong toàn bộ lô hàng với xác suất sai lầm loại I là $\alpha = 0.50$.

Lời giải : Gọi *p* là tỷ lệ kiện hàng " Khai báo đúng " trong toàn bộ lô hàng. Với các số liệu như trên, ta có bài toán kiểm định như sau:

$$\begin{cases} H : p = 0{,}99 \\ K : p > 0{,}99 \end{cases}$$

Theo mục 2.2 , lời giải tối ưu của bài toán có dạng :

Bác bỏ *H* khi $x \geq x_C$ và chấp nhận *H* khi $x < x_C$.

Từ bảng phân phối nhị thức trong [4], ta xây dựng được bảng sau đây để tìm hằng số x_C (chú ý $P\{S_n \geq x_C \mid p = p_0\} = \alpha$):

x	$P\{S_n = x \mid p = 0{,}99\}$	$P\{S_n \geq x \mid p = 0{,}99\}$	x	$P\{S_n = x \mid p = 0{,}99\}$	$P\{S_n \geq x \mid p = 0{,}99\}$
0	0	1	6	0	1
1	0	1	7	0,0001	1
2	0	1	8	0,0042	0,9999
3	0	1	9	0,0914	0,9957
4	0	1	10	0,9043	0,9043
5	0	1			

Theo bảng trên thì tất cả $\alpha > 0{,}9043$ nên không thể có $\alpha \leq 0{,}50$. Vậy nên không thể tìm được hằng số x_C thoả yêu cầu đầu bài. Tức là bài toán này không có lời giải tối ưu.

4. THẢO LUẬN

4.1 Thí dụ 3.3.1 cho thấy lời giải tối ưu tồn tại nhưng không tốt lắm, vì tỷ lệ kiện hàng khai báo đúng $p > 50\%$, tức là không cao. Hơn nữa xác suất mắc sai lầm loại một $\alpha = 0.25$ là quá lớn, nghĩa là khả năng mắc sai lầm quá cao.

Tình hình tương tự cho Thí dụ 3.3.2 .

Thí dụ 3.3.3 cho thấy rằng với yêu cầu về tỷ lệ khai báo đúng rất cao ($p_0 = 0.99$) thì mặc dù xác suất sai lầm loại một rất lớn ($\alpha = 0.5$), bài toán vẫn không có lời giải.

Điều đó chứng tỏ, thống kê toán học cho ta thấy rằng với mẫu nhỏ thì tỷ lệ kiện hàng khai báo đúng thường là một tỷ lệ không cao, trong khi khả năng mắc sai lầm là lớn. Như vậy, khi kiểm tra hàng hoá với cở mẫu quá nhỏ thì việc kiểm tra đó không còn ý nghĩa thực tế.

4.2 Để tiện cho công chức Hải quan làm việc, tác giả Nguyễn Minh Hải đã tính toán vàxây dựng được các bảng số liệu thống kê cho các cỡ mẫu n khác với $n < 50$. Khi nhìn vào bảng đó, công chức Hải quan sẽ thấy tỷ lệ khai báo đúng của kiện hàng là bao nhiêu với cỡ mẫu n cho trước và với mức ý nghĩa α cho trước.

Vì khuôn khổ bài báo có hạn nên phần phụ lục ở cuối bài chỉ đưa ra một bảng số liệu với cỡ mẫu $n = 5$. Độc giả quan tâm tới các bảng số liệu với các cỡ mẫu khác nhau có thể liên hệ địa chỉ email : minhhaikhtn@yahoo.com .

Lời cảm ơn : Các tác cảm ơn Giáo sư Nguyễn Bác Văn về những ý kiến đóng góp bổ ích cho bài toán kiểm tra ngẫu nhiên hàng hoá xuất nhập khẩu trong trường hợp mẫu nhỏ.

ON THE PROBLEM OF RANDOMLY CHECKING IMPORT – EXPORT GOODS : SMALL SAMPLE CASE

Vo Minh Tri (1), **Ung Ngoc Quang** (2)
Duong Ton Dam (2), **To Anh Dung** (2), **Nguyen Minh Hai** (2)
(1) Customs of Ho Chi Minh City
(2) University of Natural Sciences – Ho Chi Minh City

ABSTRAST : *In this article, we apply the Hypothesis testing method to the problem of randomly checking import – export for the small sample case in the customs.*

Keywords : *Binomial distribution, hypothesis testing, small sample case.*

TÀI LIỆU THAM KHẢO

[1] Võ Minh Trí, Ung Ngọc Quang, Dương Tôn Đảm, Tô Anh Dũng, Nguyễn Minh Hải. *Kiểm tra ngẫu nhiên hàng hoá xuất nhập khẩu*. Tạp chí Phát triển khoa học và công nghệ, Đại học quốc gia TP. HCM, Tập 7, số 11 / 2004.

[2] Nguyễn Bác Văn. *Xác suất và Xử lý số liệu thống kê*. NXB Giáo dục, 1996.

[3] Nguyễn Cao Văn, Trần Thái Ninh. *Lý thuyết Xác suất vàThống kê toán học*. NXB Khoa học và Kỹ thuật, 1999.

[4] Terry Sincich. *Statistics by Example*. Macmillan, London, 1987.

PHỤ LỤC

BẢNG SỐ LIỆU VỀ TỶ LỆ KHAI BÁO ĐÚNG TRONG LÔ HÀNG

n : cỡ mẫu . Bảng này ta lấy $n = 5$

p_0 : Tỷ lệ giả thuyết

α : Xác suất sai lầm loại I (mức ý nghĩa)

n	5	5	5	5	5	5	5
p_0	0,50	0,60	0,70	0,80	0,90	0,95	0,99
α	0,03	0,078	0,17	0,33	0,59	0,77	0,95

Chú giải :

- Nhắc lại : $P\{S_n \geq x \mid p = p_0\} = \alpha$.
- Bảng này chỉ ra rằng khi cố định $n = 5$, nếu cho $p_0 = 50\%$ thì ta nói rằng tỷ lệ khai báo đúng của lô hàng cao hơn 50% với xác suất sai lầm loại một là α =0,03.
- Tương tự, khi cố định $n = 5$, nếu cho $p_0 = 60\%$ thì ta nói rằng tỷ lệ khai báo đúng của lô hàng cao hơn 60% với xác suất sai lầm loại một là $\alpha = 0{,}078$.
- Khi cố định $n = 5$, nếu cho $p_0 = 70\%$ thì ta nói rằng tỷ lệ khai báo đúng của lô hàng cao hơn 70% với xác suất sai lầm loại một là $\alpha = 0{,}17$. . . .
- Khi cố định $n = 5$, nếu cho $p_0 = 99\%$ thì ta nói rằng tỷ lệ khai báo đúng của lô hàng cao hơn 99% với xác suất sai lầm loại một là $\alpha = 0{,}95$.
- Như vậy với $n = 5$ thì tỷ lệ khai báo đúng chỉ có ý nghĩa khi ta lấy $p_0 = 50\%$, $p_0 = 60\%$, vì khi đó xác suất sai lầm loại I tương ứng là khá nhỏ. Còn đối với các p_0 khác thì xác suất sai lầm loại I quá lớn. Nhất là khi $p_0 = 99\%$ thì $\alpha = 0{,}95$. Điều này làm cho giả thuyết $p_0 = 99\%$ trở nên vô nghĩa.
- Tóm lại với cỡ mẫu nhỏ thì tỷ lệ khai báo đúng không thể là một tỷ lệ cao như ta đã nêu ở phần thảo luận 4.1 .