Đặng Kim Phương (2020)
(18): 22-28

NHẬN DẠNG VÀ ỨNG DỤNG PHÂN PHỐI NHỊ THỨC TRONG THỐNG KÊ

Đặng Kim Phương
Trường Đại học Tây Bắc

Tóm tắt: *Trong khuôn khổ của bài viết này, chúng tôi sẽ trình bày về nhận dạng và ứng dụng qui luật phân phối nhị thức cho sự đo lường được thực hiện trong các điều kiện quan sát hay thí nghiệm, để giải một số bài toán xác suất thống kê, trong đó có những bài toán thống kê có ý nghĩa trong nghiên cứu khoa học thực nghiệm. Đồng thời chúng tôi cũng đưa ra một hệ thống ví dụ minh họa nhằm cung cấp một số kĩ năng giải quyết bài toán trong thực tiễn khi nghiên cứu khoa học thực nghiệm.*

Từ khóa: *Đại lượng ngẫu nhiên, Trung bình, Phương sai, Độ lệch chuẩn, Kiểm định giả thiết thống kê.*

1. Đặt vấn đề

Nghiên cứu xã hội học cho thấy, tình yêu của người Mỹ dành cho xe hơi là rất lớn. Số ngày mà một người Mỹ có sở hữu xe hơi không ngồi sau tay lái để lái xe đi làm, đi mua sắm, hay lái xe chỉ vì yêu thích,… chẳng còn là bao. Tuy nhiên theo *Fank Newport* và *Leslie McAneny* (1993) khi điều tra 1.003 người lớn vào tháng sáu và 803 thiếu niên vào tháng chín năm 1993 thì cả người lớn và thiếu niên Mỹ đều cho rằng bằng lái xe không phải là một quyền lợi mà là một đặc quyền. Theo kết quả điều tra họ thấy rằng: 70% số người lớn được hỏi ủng hộ một kỳ thi mang tính bắt buộc 3 năm 1 lần đối với những người lái xe trên 65 tuổi và 56% số thiều niên được hỏi đã ủng hộ điều luật từ chối cấp bằng lái xe cho những ai dưới 21 tuổi mà đã bỏ học trung học. Báo cáo của hai tác giả này khẳng định rằng: Kết quả điều tra tỷ lệ % người lớn ủng hộ một kỳ thi mang tính bắt buộc 3 năm 1 lần chỉ khác với tỷ lệ % thực tế với toàn bộ số người lớn ở Mỹ không lớn hơn 3% và kết quả điều tra tỷ lệ % thiếu niên ủng hộ điều luật từ chối cấp bằng lái xe cho những ai dưới 21 tuổi mà đã bỏ học trung học chỉ khác với tỷ lệ % thực tế với toàn bộ số thiếu niên ở Mỹ không lớn hơn 4%. Vấn đề được đặt ra là:

- Bằng cách nào mà có thể khẳng định chắc chắn rằng các tỷ lệ % được báo cáo là chính xác khi cuộc điều tra được thực hiện bằng cách sử dụng câu hỏi trả lời là "có " và "không".

- Mô hình thống kê nào là thích hợp trong những tình huống như thế này.

- Việc sử dụng mô hình này để đánh giá độ tin cậy của kết luận dựa trên các câu hỏi trả lời là "có " và "không", xác định giá trị trung bình, độ lệch chuẩn,… được thực hiện như thế nào?

Trong bài báo này, chúng tôi sẽ trình bày phương pháp nhận dạng qui luật phân phối nhị thức và ứng dụng của qui luật phân phối này thông qua nội dung của những bài toán thống kê có ý nghĩa trong nghiên cứu khoa học thực nghiệm.

2. Phương pháp nghiên cứu

Trước hết, chúng tôi nhắc lại một số khái niệm và kết quả cần thiết sau trong [2] và [4].

2.1 Định nghĩa. Đại lượng ngẫu nhiên X được gọi là có phân phối nhị thức với tham số (n, p) nếu phân phối xác suất của nó có dạng

$$P(X = k) = C_n^k p^k q^{n-k}$$

trong đó:

n là số lần thực hiện phép thử.

X là số lần xuất hiện biến cố A trong n lần thực hiện phép thử.

p là xác suất xuất hiện biến cố A trong mỗi lần thực hiện phép thử $(0 < p < 1)$.

$k = 0, 1, 2, ..., n$; $q = 1 - p$.
$C_n^k = \frac{n!}{k!(n-k)!}$ với $n! = 1.2...n$ và $0! = 1$.

Ký hiệu đại lượng ngẫu nhiên X phân phối theo quy luật nhị thức với tham số n và p là $X \sim B(n, p)$.

2.2 Các số đặc trưng của phân phối nhị thức

Nếu đại lượng ngẫu nhiên X có phân phối nhị thức với tham số (n, p) thì

i) Kỳ vọng $EX = np$.

ii) Phương sai $DX = npq$.

iii) Độ lệch chuẩn $\sigma = \sqrt{DX}$.

iiii) $Mod(X) = [(n+1)p]$; ($[a]$ chỉ phần nguyên của a).

3. Kết quả nghiên cứu

Trong xác suất thống kê, mỗi dấu hiệu nghiên cứu đều có một qui luật phân phối nhất định, trong đó qui luật phân phối nhị thức có tần suất gặp khá phổ biến. Để nhận dạng qui luật phân phối nhị thức có thể dùng tiêu chuẩn Kolmogorov, tiêu chuẩn Palowski,... Trong bài viết này sẽ trình bày cách nhận dạng phân phối nhị thức bằng phương pháp: sử dụng tiêu chuẩn kiểm định khi bình phương và thông qua các đặc trưng của phép thử nhị thức. Kết quả chính của chúng tôi là cung cấp hệ thống ví dụ minh họa, trong đó chúng tôi sử dụng hệ thống kiến thức liên quan vào phân tích dữ liệu thực nghiệm để giải một số bài toán thống kê cụ thể.

3.1 Sử dụng tiêu chuẩn kiểm định khi bình phương nhận dạng phân phối nhị thức

Các bước sử dụng tiêu chuẩn kiểm định khi bình phương để kiểm định giả thiết về qui luật phân phối nhị thức được thực hiện như sau:

Giả sử $(X_1, X_2, ..., X_n)$ là mẫu quan sát của dấu hiệu nghiên cứu X. Kiểm định giả thiết: X là đại lượng ngẫu nhiên có phân phối nhị thức $B(n, p)$ ở mức ý nghĩa α.

Xét khoảng (a, b) trên trục số sao cho mọi quan sát của mẫu $(X_1, X_2, ..., X_n)$ đều nằm trong khoảng này. Chia (a, b) thành k khoảng (hay còn gọi là tổ): $C_1, C_2, ..., C_k$. Gọi n_i là tần số của các quan sát X_i trong mẫu $(X_1, X_2, ..., X_n)$ thuộc khoảng $C_i, i = \overline{1,k}$; $\sum_{i=1}^{k} n_i = n$.

Thay p bởi ước lượng điểm của p là $\hat{p}$, tính xác suất $\hat{p}_i = P[X \in C_i]; i = 1, 2, ..., k$.

Tính tiêu chuẩn kiểm định

$$Z = \sum_{i=1}^{k} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

và so sánh Z với C_α (C_α là giá trị tra trong bảng phân phối khi bình phương với $k - r - 1$ bậc tự do, mức ý nghĩa α). Nếu $Z > C_\alpha$ thì bác bỏ giả thiết cho rằng dấu hiệu nghiên cứu X có phân phối nhị thức $B(n, p)$.

Lưu ý, tiêu chuẩn kiểm định khi bình phương được sử dụng tốt khi kích thước mẫu n đủ lớn và tần số n_i trong mỗi khoảng lớn hơn hoặc bằng 5, do đó nếu trong số liệu của mẫu đã cho có khoảng nào có tần số nhỏ hơn 5 thì phải gộp khoảng đó vào khoảng trước hoặc sau nó.

Ví dụ 1. Để đánh giá chất lượng sản phẩm do doanh nghiệp A sản xuất, người ta tiến hành chọn ngẫu nhiên từ mỗi kiện hàng ra 3 sản phẩm để kiểm tra. Kết quả thu được như sau:

Số sản phẩm loại I	0	1	2	3
Số kiện hàng	13	107	376	504

Với mức ý nghĩa $\alpha = 0,05$ có thể khẳng định tỷ lệ sản phẩm loại I trong mỗi kiện hàng do doanh nghiệp A sản xuất là 80% không?

Do không biết tổng số sản phẩm trong 1000 kiện hàng do doanh nghiệp A sản suất, nên không thể dùng tiêu chuẩn kiểm định về tỷ lệ để kiểm định giả thiết cho rằng "tỷ lệ sản phẩm loại I trong mỗi kiện hàng do doanh nghiệp A sản xuất là 80% ". Để kiểm định được giả thiết này phải sử dụng tiêu chuẩn khi bình phương:

Gọi X là số sản phẩm loại I có thể được lấy ra trong mỗi kiện hàng.

Thiết lập bài toán kiểm định giả thiết:

$H: X$ có phân phối nhị thức $B(3;0,8)$.

$K: X$ không có phân phối nhị thức $B(3;0,8)$ ở mức ý nghĩa $\alpha = 0,05$.

Gọi $\hat{p}_i$ là xác suất trong kiện hàng có i sản phẩm loại I thì

$\hat{p}_i = C_3^i \hat{p}^i (1-\hat{p})^{3-i}$; $i = 0;1;2;3$. Ta có

$$\hat{p}_0 = C_3^0 0,8^0.0,2^3 = 0,008$$
$$\hat{p}_1 = C_3^1 0,8^1.0,2^2 = 0,096$$
$$\hat{p}_2 = C_3^2 0,8^2.0,2^1 = 0,384$$
$$\hat{p}_3 = C_3^3 0,8^3.0,2^0 = 0,512.$$

Tính tiêu chuẩn kiểm định

$$Z = \frac{(13-8)^2}{8} + \frac{(107-96)^2}{96} + \frac{(376-384)^2}{384} + \frac{(504-512)^2}{512} = 4,676.$$

Tra bảng giá trị hàm phân phối khi bình phương: $C_\alpha = \chi^2(3;0,05) = 7,8$. Do $Z < C_\alpha$ nên giả thiết H được chấp nhận ở mức ý nghĩa $\alpha = 0,05$ tức là X là đại lượng ngẫu nhiên tuân theo qui luật phân phối nhị thức $B(3;0,8)$. Vậy tỷ lệ sản phẩm loại I trong mỗi kiện hàng do doanh nghiệp A sản xuất là 80%. Với số liệu thống kê và kết quả kiểm định X là đại lượng ngẫu nhiên tuân theo qui luật phân phối nhị thức $B(3;0,8)$ có thể giải quyết được một số bài toán đặt ra như:

Tính các xác suất:

$$P(X=0) = C_3^0 0,8^0.0,2^3 = 0,008$$
$$P(X=1) = C_3^1 0,8^1.0,2^2 = 0,096$$
$$P(X=2) = C_3^2 0,8^2.0,2^1 = 0,384$$
$$P(X=3) = C_3^3 0,8^3.0,2^0 = 0,512.$$

Tính giá trị trung bình của X:

$$EX = np = 3.0,8 = 2,4.$$

Tính phương sai và độ lệch chuẩn của X:

$$DX = npq = 3.0,8.0,2 = 0,48$$
$$\sigma = \sqrt{DX} = \sqrt{0,48} = 0,69.$$

Để nhận biết một dấu hiệu cần nghiên cứu nào đó có tuân theo qui luật phân phối nhị thức hay không, ngoài cách sử dụng tiêu chuẩn kiểm định ở trên còn có thể nhận dạng được qui luật phân phối nhị thức thông qua phép thử tạo nên qui luật phân phối này, đó là phép thử nhị thức. Phép thử nhị thức là một mô hình tuyệt vời cho nhiều tình huống chọn mẫu trong thống kê, đặc biệt là các cuộc điều tra tạo ra loại hình dữ liệu "có" hoặc "không". Sau đây chúng tôi sẽ trình bày các đặc trưng của phép thử nhị thức và thông qua các ví dụ giúp cho bạn đọc nắm được qui trình phân tích số liệu thống kê để nhận dạng phân phối nhị thức và ứng dụng phân phối này vào giải những bài toán trong thực tiễn khi nghiên cứu khoa học thực nghiệm [1], [2], [3].

3.2 Nhận dạng phân phối nhị thức thông qua các đặc trưng của phép thử nhị thức

Phép thử nhị thức có các đặc trưng sau:

1. Phép thử đó được thực hiện n lần giống nhau.

2. Mỗi lần thử chỉ có một trong hai kết quả: "thành công" hoặc "thất bại".

3. Xác suất thành công trong mỗi lần thử luôn bằng $p\ (0 < p < 1)$, xác suất thất bại trong mỗi lần thử luôn bằng $1 - p = q$.

4. Các lần thử độc lập với nhau.

5. Ta quan tâm đến là số lần thành công trong n lần thử.

Gọi X là số lần thành công trong n lần thử thì X là đại lượng ngẫu nhiên có phân phối nhị thức với tham số (n, p).

Ví dụ 2. Một chủ doanh nghiệp nhận ra rằng, một số nhân viên trong doanh nghiệp đã làm giả mạo thông tin trong hồ sơ xin việc và xác suất một nhân viên làm giả mạo thông tin trong hồ sơ xin việc là 0,35 . Doanh nghiệp tiến hành kiểm tra hồ sơ xin việc của 5 nhân viên mới được nhận vào làm việc. Việc chọn mẫu này có phải là phép thử nhị thức không?

Ta thấy:

1. Việc kiểm tra hồ sơ xin việc của 5 nhân viên là thực hiện 5 lần thử giống nhau.

2. Mỗi lần thử chỉ có một trong hai kết quả: Hồ sơ đó "có" hoặc "không" làm giả mạo thông tin. Hai kết quả này có thể liên tưởng đến sự "thành công" hay "thất bại" của một phép thử.

3. Xác suất "thành công" của một lần thử luôn bằng 0,35.

4. Các lần thử là độc lập với nhau, vì xác suất "thành công" của lần thử này không bị tác động bởi kết quả của các lần thử khác.

5. Ta quan tâm tới số hồ sơ xin việc làm giả mạo thông tin.

Vậy, việc kiểm tra hồ sơ xin việc của 5 nhân viên mới thỏa mãn các đặc trưng của phép thử nhị thức.

Gọi X là số hồ sơ xin việc làm giả mạo thông tin thì X là đại lượng ngẫu nhiên có phân phối nhị thức với tham số $(5;0,35)$.

Ví dụ 3. Trở lại với nghiên cứu điển hình đã được trình bày trong phần mở đầu.

Sự ước tính tỷ lệ người lớn ở Mỹ ủng hộ một kỳ thi mang tính bắt buộc 3 năm 1 lần đối với những người lái xe trên 65 tuổi, phụ thuộc vào số người trong cuộc điều tra ủng hộ bài kiểm tra mang tính bắt buộc đối với những người lái xe trên 65 tuổi.

Việc thực hiện cuộc điều tra thỏa mãn các đặc trưng của phép thử nhị thức:

1. Việc chọn mẫu này bao gồm $n = 1.003$ lần thử giống nhau. Mỗi lần thử là sự lựa chọn 1 người duy nhất từ một số lớn người dân Mỹ.

2. Mỗi lần thử chỉ có một trong hai kết quả: Người được hỏi trả lời "có" hoặc "không" ủng hộ một kỳ thi bắt buộc. Hai kết quả này có thể liên tưởng đến sự "thành công" hay "thất bại" của một phép thử.

3. Xác suất của sự "thành công" của mỗi lần thử luôn bằng 0,7 và xác suất này giữ nguyên từ lần thử này đến lần thử khác.

4. Các lần thử là độc lập vì xác suất "thành công" trong bất cứ lần thử nào sẽ không bị tác động bởi kết quả của bất kỳ lần thử khác.

5. Ta quan tâm tới số người trong mẫu $n = 1.003$ ủng hộ bài kiểm tra mang tính bắt buộc đối với những người lái xe trên 65 tuổi.

Gọi X là số người trong mẫu $n = 1.003$ ủng hộ bài kiểm tra mang tính bắt buộc 3 năm 1 lần đối với những người lái xe trên 65 tuổi thì X là đại lượng ngẫu nhiên có phân phối nhị thức

$B(1003;0,7)$ với trung bình và độ lệch chuẩn:

$$EX = np = 1003.0,7 = 702,1.$$
$$\sigma = \sqrt{npq} = \sqrt{1003.0,7.0,3} = 14,51.$$

Với kết quả điều tra thực tế, tỷ lệ người lớn ở Mỹ ủng hộ một kỳ thi mang tính bắt buộc 3 năm 1 lần đối với những người lái xe trên 65 tuổi là $p = 0,7$ thì theo qui tắc thực chứng ta biết được rằng, có khoảng 95% số người trong mẫu ủng hộ một kỳ thi mang tính bắt buộc 3 năm 1 lần đối với những người lái xe trên 65 tuổi nằm trong khoảng 2 lần độ lệch chuẩn so với giá trị trung bình:

$$P\big[EX - 2\sigma \le X \le EX + 2\sigma\big] = 0,95$$
$$P\big[673,08 \le X \le 731,12\big] = 0,95.$$

Tức là, với xác suất 0,95 có khoảng 673 đến 731 người ủng hộ kỳ thi mang tính bắt buộc đối với người lớn và ta có

$$P\left[\frac{673}{1003} \le \frac{X}{n} \le \frac{731}{1003}\right] = 0,95$$
$$P\big[0,67 \le p \le 0,729\big] = 0,95.$$

Với độ tin cậy 0,95 có thể khẳng định tỷ lệ người lớn ở Mỹ ủng hộ một kỳ thi mang tính bắt buộc 3 năm 1 lần đối với những người lái xe trên 65 tuổi nằm trong khoảng 67% đến 72,9%.

Vậy, báo cáo của hai tác giả khẳng định rằng: Kết quả điều tra tỷ lệ % người lớn ủng hộ một kỳ thi mang tính bắt buộc 3 năm 1 lần chỉ khác với tỷ lệ % thực tế với toàn bộ số người lớn ở Mỹ không lớn hơn 3% là đúng.

Tương tự, có thể kiểm tra được kết quả báo cáo về tỷ lệ % thiếu niên ủng hộ điều luật từ chối cấp bằng lái xe cho những ai dưới 21 tuổi mà đã bỏ học trung học.

Ví dụ 4. Giả sử có khoảng 1 triệu người trong một khu vực bán hàng nào đó là người mua tiềm năng của một sản phẩm mới. Để ước lượng tỷ lệ người sẽ mua sản phẩm này nếu như nó được đưa ra chào bán. Người ta đã chọn một mẫu gồm 1.000 người theo cách thức, mỗi người trong số 1 triệu người trong khu vực bán hàng này sẽ có cơ hội ngang nhau của việc lựa chọn. Mỗi người trong mẫu sẽ được hỏi rằng: Ông/bà có mua sản phẩm mới này không nếu như nó được chào bán?

Ta sẽ kiểm tra việc chọn mẫu trong ví dụ này có thỏa mãn các đặc trưng của phép thử nhị thức được mô tả ở trên hay không?

1. Việc chọn mẫu này bao gồm $n = 1.000$ lần thử giống nhau. Mỗi lần thử là sự lựa chọn 1 người duy nhất từ 1 triệu người trong khu vực bán hàng.

2. Mỗi lần thử chỉ có một trong hai kết quả: Người được hỏi trả lời "có" hoặc "không" mua sản phẩm. Hai kết quả này có thể liên tưởng đến sự "thành công" hay "thất bại" của một phép thử.

3. Xác suất của sự "thành công" sẽ bằng với tỷ lệ của 1 triệu người sẽ mua sản phẩm mới. Theo luật số lớn, xác suất này giữ nguyên từ lần thử này đến lần thử khác.

4. Các lần thử là độc lập vì xác suất "thành công" trong bất cứ lần thử nào sẽ không bị tác động bởi kết quả của bất kỳ lần thử khác.

5. Ta quan tâm tới số người trong mẫu $n = 1.000$ sẽ mua sản phẩm này.

Cuộc điều tra này thỏa mãn cả năm đặc trưng của phép thử nhị thức nên đây là một phép thử nhị thức. Giả sử kết quả khảo sát trong mẫu có 650 người trả lời "có mua sản phẩm mới nếu như nó được chào bán" thì để ước lượng tỷ lệ người sẽ mua sản phẩm mới nếu như nó được đưa ra chào bán sẽ được thực hiện như sau:

Gọi p là tỷ lệ người sẽ mua sản phẩm mới nếu như nó được đưa ra chào bán. Với độ tin cậy 0,95 ta có

$$0,65 - 1,96.\sqrt{\frac{0,65.0,35}{1000}} < p < 0,65 + 1,96.\sqrt{\frac{0,65.0,35}{1000}}$$
$$0,621 < p < 0,679.$$

Như vậy, với độ tin cậy 0,95 tỷ lệ người sẽ mua sản phẩm mới nếu như nó được đưa ra chào bán nằm trong khoảng 62,1% đến 67,9%.

Kiểm định giả thiết

$$\begin{cases} \text{H} : \text{p} = 0,67 \\ \text{K} : \text{p} \neq 0,67 \end{cases}$$

ở mức ý nghĩa $\alpha = 0,05$. Tính giá trị kiểm định

$$|Z| = \frac{|650 - 1000.0,67|}{\sqrt{1000.0,67.0,33}} = 1,34 < 1,96.$$

Ta chấp nhận giả thiết: tỷ lệ người sẽ mua sản phẩm mới nếu như nó được đưa ra chào bán là 67%. Gọi X là số người trong mẫu sẽ mua sản phẩm mới nếu như nó được đưa ra chào bán thì X là đại lượng ngẫu nhiên có qui luật phân phối nhị thức $B(1000;0,67)$ và ta có thể tính được:

Số người trung bình trong mẫu sẽ mua sản phẩm mới nếu như nó được đưa ra chào bán:

$$EX = np = 1000.0,67 = 670 \text{ (người)}$$

Độ lệch chuẩn:

$$\sigma = \sqrt{npq} = \sqrt{1000.0,67.0,33} = 14,86$$

4. Kết luận

Trong xác suất thống kê, phân phối nhị thức là một trong những phân phối quan trọng và thông dụng, những tính chất của qui luật phân phối này đã được ứng dụng để giải quyết rất nhiều bài toán trong nghiên cứu Khoa học kỹ thuật, Kinh tế, Giáo dục, Xã hội, … Việc quen thuộc với phân phối nhị thức và nhận biết được những đặc tính của phép thử tạo ra qui luật phân phối này là hết sức hữu ích. Nó giúp cho các nhà nghiên cứu, không những tính được xác suất của số lần "thành công" trong n lần thử độc lập giống nhau, trong đó xác suất của một "thành công" trong mỗi lần thử luôn

bằng p, mà còn xác định được các thông tin về giá trị trung bình, độ lệch chuẩn, mod,… của dấu hiệu cần nghiên cứu một cách dễ dàng mà không cần phải qua các qui trình tính toán phức tạp.

TÀI LIỆU THAM KHẢO

1 Đặng Hùng Thắng (2011). *Mở đầu về lý thuyết xác suất và các ứng dụng*. Nxb Giáo dục,47-48.

2 Đào Hữu Hồ (2000). *Thống kê xã hội học*. Nxb ĐHQG Hà Nội,57-70.

3 Đinh Văn Gắng (2003). *Lý thuyết xác suất và thống kê*. Nxb Giáo dục,42-50.

4 Phạm Văn Kiều (1998). *Xác suất thống kê*. Nxb Giáo dục, 62-68.

IDENTIFICATION AND APPLICATION OF BINOMIAL DISTRIBUTION IN STATISTICS

Dang Kim Phuong
Tay Bac University

Abstract: *In this article, we shall present the identification and application of binomial distribution for measurement conducted under the observational or experimental conditions to solve some statistical probability problems including those of significance in experimental scientific research. We also offer a series of illustrative examples to provide some practical problem-solving skills when carrying out empirical scientific research.*

Keywords: *Random variables, Average, Expected Value, Standard deviation, Statistical hypothesis testing.*

Ngày nhận bài: 14/8/2019. Ngày nhận đăng: 29/09/2019
Liên lạc: Đặng Kim Phương; Email: dangkimphuongtbu@gmail.com