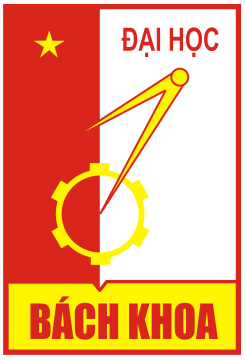
ĐẠI HỌC BÁCH KHOA HÀ NỘI

KHOA TOÁN - TIN



ĐỒ ÁN I

KIỂM ĐỊNH PHÂN PHỐI CHUẨN (TEST OF NORMALITY)

Chuyên ngành: Toán cơ bản

Giảng viên hướng dẫn: TS. Nguyễn Văn Hạnh

Sinh viên thực hiện: Phạm Thị Thanh Hà

MSSV: 20216823

Lớp: Toán Tin 01 - K66

Hà Nội, Ngày 21 tháng 6 năm 2024

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

1. Mục tiêu và nội dung của đồ án

(a) Mục tiêu: Hiểu được các phương pháp phổ biến để kiểm định phân phối chuẩn và áp dụng trong bộ dữ liệu số ngẫu nhiên.

(b) Nội dung: Cơ sở lý thuyết kiểm định phân phối chuẩn với phương pháp đồ thị và phương pháp suy luận thống kê. Xác định tính chuẩn của bộ dữ liệu số ngẫu nhiên bằng cách áp dụng các phương pháp đã nêu .

2. Kết quả đạt được

(a)

(b)

(c)

3. Ý thức làm việc của sinh viên

(a)

(b)

(c)

Hà Nội, ngày ... tháng ... năm 2023

Giảng viên hướng dẫn

TS. Nguyễn Văn Hạnh

Lời cảm ơn

Để có thể hoàn thành đồ án này, em xin được gửi lời cám ơn chân thành và sâu sắc nhất đến giảng viên hướng dẫn - TS. Nguyễn Văn Hạnh. Thầy đã tận tình hướng dẫn, góp ý, và hỗ trợ em trong suốt quá trình thực hiện đồ án. Nhờ đó, em học hỏi được nhiều điều cũng như phát triển kỹ năng một cách đáng kể. Đồ án này sẽ không thể hoàn thiện được nếu thiếu đi những sự giúp đỡ vô cùng to lớn và kịp thời của thầy.

Em cũng xin gửi lời cám ơn đến Khoa Toán - Tin đã thiết kế các học phần cần thiết trong chương trình giảng dạy của ngành Toán Tin để em được trang bị những kiến thức nền tảng sử dụng trong nội dung đồ án. Những kiến thức, kĩ năng và cách tư duy ấy sẽ luôn là hành trang quan trọng nhất của em trên con đường học tập hiện tại và sau này.

Em rất hy vọng đồ án sẽ cung cấp được những thông tin hữu ích, góp phần vào việc nâng cao hiểu biết về việc điểm định phân phối chuẩn của bộ dữ liệu. Song, bởi kiến thức và kinh nghiệm của bản thân còn nhiều hạn chế, em hiểu rằng mình không thể tránh khỏi những sai lầm và thiếu sót. Vì vậy, em rất mong nhận được những lời nhận xét, những ý kiến đóng góp, phê bình từ phía Thầy/Cô để đồ án được hoàn thiện hơn.

Cuối cùng, em xin kính chúc quý thầy cô luôn dồi dào sức khỏe để tiếp tục cống hiến cho sự nghiệp giáo dục và luôn là tấm gương sáng cho sinh viên Đại học Bách khoa Hà Nội chúng em noi theo.

Em xin chân thành cảm ơn!

Hà Nội, ngày 08 tháng 06 năm 2023

Tác giả đồ án

Phạm Thị Thanh Hà

Mục lục

Lời mở đầu

Trong thống kê, việc giả định rằng các quan sát tuân theo phân phối chuẩn là điều rất thường xuyên và không thể thiếu. Toàn bộ cấu trúc thống kê dựa trên giả định này và nếu giả định này bị vi phạm, suy luận sẽ bị phá vỡ. Chính vì lý do này, việc kiểm định hoặc thử nghiệm giả định này trước khi phân tích bất kì thống kê nào của dữ liệu là rất quan trọng.

Trong suốt 100 năm qua, quan điểm về việc sử dụng phân phối chuẩn trong các mô hình thống kê đã thay đổi một cách rõ rệt. Pearson (1905) [1] từng nói: "Ngay cả vào cuối thế kỷ XIX, không phải ai cũng nghĩ là cần có những đường cong khác ngoài đường cong chuẩn." Đến giữa thế kỷ XX, Geary (1947) [2] lại nhận xét: "Chuẩn chỉ là chuyện hoang đường; chưa từng và sẽ không bao giờ có một phân phối chuẩn thực sự." Tuy ý kiến này có vẻ hơi cực đoan, nhưng quả thật các phân phối không chuẩn xuất hiện trong thực tế nhiều hơn người ta tưởng trước đây.

Gnanadesikan (1977) [3] chỉ ra rằng: "Tác động của việc lệch khỏi phân phối chuẩn lên các phương pháp thống kê cổ điển vẫn chưa được hiểu rõ". Tuy nhiên, nhiều bằng chứng cho thấy những sai lệch này có thể gây ra nhiều hệ quả không mong muốn trong nhiều tình huống khác nhau. Trong lĩnh vực hồi quy, Huber (1973) [4] đã nghiên cứu ảnh hưởng của sự lệch chuẩn trong ước lượng. Ông nhận thấy rằng khi dữ liệu không tuân theo phân phối chuẩn, việc tìm ra các điều kiện cần và đủ để mọi ước lượng tham số đều tiệm cận chuẩn là rất khó khăn. Về kiểm định giả thuyết, nhiều nhà thống kê đã tìm hiểu tác động của sự lệch chuẩn. Judge (1985) [5] đã tổng hợp một cách khá toàn diện các nghiên cứu này. Khi các quan sát không tuân theo phân phối chuẩn, các kiểm định chuẩn và kiểm định khi bình phương trở nên thiếu chính xác, dẫn đến việc các kiểm định t và F mất hiệu lực với các mẫu có kích thước hữu hạn. Tuy nhiên, chúng vẫn có cơ sở hợp lý khi kích thước mẫu tiến tới vô cùng. Theo Pearson và Please (1975) [6], mức ý nghĩa của các kiểm định t và F có vẻ khá bền vững trước sự lệch chuẩn. Dù đây là một đặc tính hấp dẫn, nhưng điều quan trọng là phải nghiên cứu cả độ mạnh của kiểm định và mức ý nghĩa dưới tác động của sự lệch chuẩn. Koenker (1982) [7] chỉ ra rằng độ mạnh của các kiểm định t và F rất nhạy

cảm với phân phối giả định và có thể giảm nhanh chóng khi phân phối có đuôi dày hơn. Hơn nữa, Bera và Jarque (1982) [8] phát hiện rằng các kiểm định phương sai đồng nhất và độc lập chuỗi được đề xuất cho quan sát chuẩn có thể dẫn đến kết luận sai lầm trong điều kiện phi chuẩn. Việc hiểu rõ bản chất của các quan sát cũng rất cần thiết trong dự báo và xác định khoảng tin cậy của dự báo. Hầu hết các kết quả chuẩn trong lĩnh vực này đều dựa trên giả định về tính chuẩn, và toàn bộ quy trình suy luận có thể bị sai lệch nếu giả định này không đúng.

Nhìn chung, vi phạm giả định về tính chuẩn có thể dẫn đến việc sử dụng các ước lượng không tối ưu, đưa ra các kết luận suy luận không hợp lý và các dự báo thiếu chính xác. Do đó, để đảm bảo tính đúng đắn của các kết luận, chúng ta cần kiểm định kỹ lưỡng giả định về tính chuẩn. Mục tiêu chính của đồ án này là tổng hợp các phương pháp có thể sử dụng để kiểm định phân phối chuẩn.

Cấu trúc đồ án

1. Các phương pháp kiểm định phân phối chuẩn.
2. Ứng dụng của kiểm định phân phối chuẩn vào phân tích thống kê.
3. Kết luận

Bảng ký hiệu và chữ viết tắt

Phần I

Các phương pháp kiểm định phân phối chuẩn

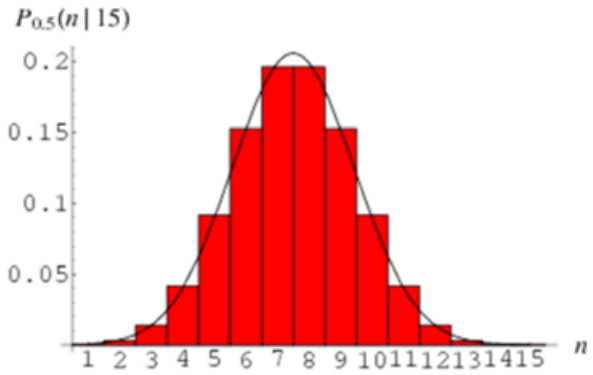
Chương 1

Phương pháp đồ thị để kiểm định phân phối chuẩn

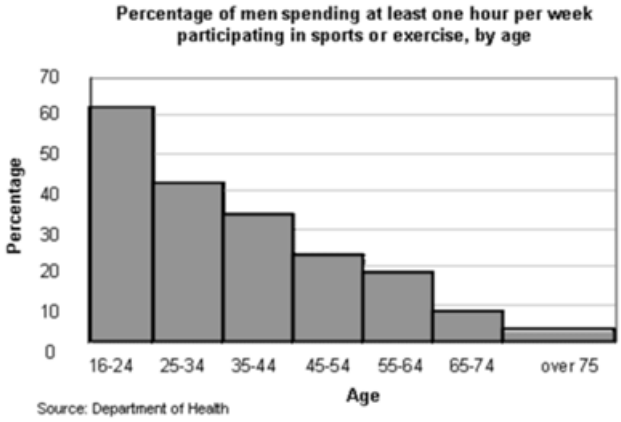
Có thể ta đã biết, bất kỳ phân tích thống kê nào cũng trở nên phong phú hơn khi kết hợp với việc xem xét các biểu đồ quan sát. Như Chambers (1983) đã nói: "Các phương pháp đồ thị cung cấp những công cụ chẩn đoán mạnh mẽ để kiểm chứng các giả định hoặc khi giả định không được đáp ứng thì dễ đề xuất các biện pháp khắc phục. Thiếu những công cụ này, việc kiểm chứng giả định chỉ còn là hy vọng suông". Một số loại biểu đồ thống kê như biểu đồ phân tán, biểu đồ phần dư được khuyến khích sử dụng để kiểm tra hoặc chẩn đoán các phương pháp thống kê. Đối với việc kiểm tra độ phù hợp và vẽ đường cong phân phối, các biểu đồ thống kê rất cần thiết để có cái nhìn tổng quan về mẫu. Các phương pháp kiểm định hiện có giúp đưa ra quyết định khách quan về tính chuẩn, nhưng chúng không cung cấp gợi ý tổng quát về nguyên nhân bác bỏ giả thuyết không. Vì vậy, chúng ta quan tâm đến việc trình bày các loại đồ thị khác nhau để kiểm tra tính chuẩn cũng như các quy trình kiểm định khác nhau. Thông thường, các biểu đồ được sử dụng nhiều nhất để kiểm tra giả định về tính chuẩn bao gồm: biểu đồ tần suất, biểu đồ thân-lá, biểu đồ hộp, đồ thị phần trăm-phần trăm (P-P), biểu đồ xác suất chuẩn (Q-Q), biểu đồ hàm phân phối tích lũy thực nghiệm và các biến thể khác của đồ thị xác suất.

1.1 Biểu đồ tần suất

Biểu đồ dễ dàng và đơn giản nhất là biểu đồ tần suất. Biểu tần suất trong đó các giá trị quan sát được vẽ bằng tần suất của chúng, cho phép ước lượng trực quan xem phân phối có hình chuông hay không. Đồng thời, nó cũng cung cấp chỉ báo về các khoảng trống trong dữ liệu và các điểm ngoại lai. Nó cũng đưa ra ý tưởng về độ nghiêng hay tính đối xứng. Dữ liệu có thể được biểu diễn bằng loại đường cong chuông lý tưởng như trong biểu đồ đầu tiên được gọi là dữ liệu có phân phối chuẩn. Tất nhiên đối với biểu đồ thứ hai, dữ liệu không được phân phối chuẩn.



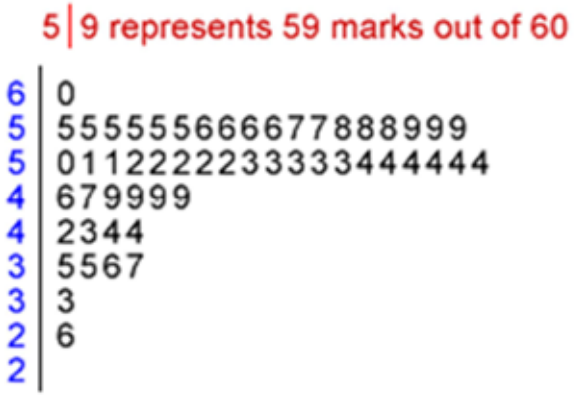
Hình 1.1: Biểu đồ tần suất thể hiện dữ liệu tuân theo phân phối chuẩn



Hình 1.2: Biểu đồ tần suất thể hiện dữ liệu không tuân theo phân phối chuẩn

1.2 Biểu đồ thân và lá

Biểu đồ thân và lá thể hiện cùng một loại nội dung như biểu đồ tần suất, nhưng khi các quan sát xuất hiện cùng với giá trị thực của chúng thì dường như chúng không bị mất bất kỳ thông tin nào về dữ liệu gốc. Giống như biểu đồ tần suất, chúng hiển thị tần số của các quan sát cùng với giá trị trung vị, giá trị cao nhất và thấp nhất của phân phối, cũng như các tỷ lệ phần trăm mẫu khác từ cách hiển thị dữ liệu. Có một "thân" và một "lá" cho mỗi giá trị, trong đó thân biểu diễn một tập hợp các ngăn mà các lá được nhóm lại, và các lá này phản ánh các thanh giống như trong biểu đồ tần suất.



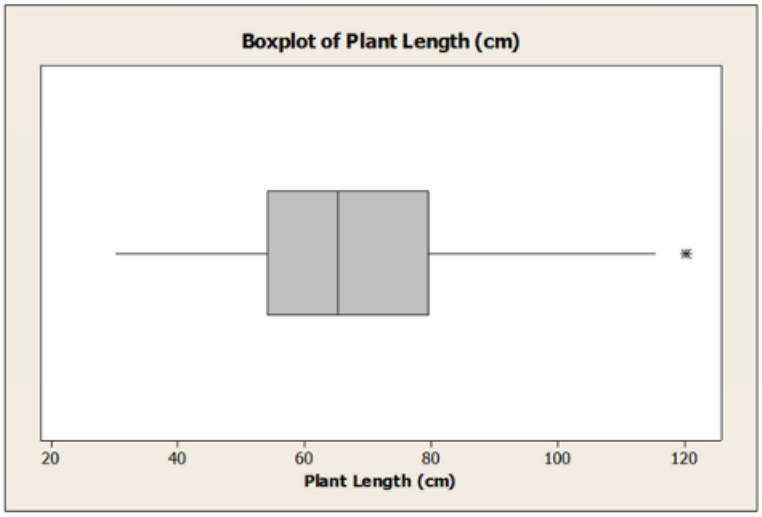
Hình 1.3: Biểu đồ thân và lá thể hiện dữ liệu không tuân theo phân phối chuẩn

Với biểu đồ thân và lá về điểm số của sinh viên trên đây, ta có thể nhận thấy rõ ràng dữ liệu không tuân theo phân phối chuẩn.

1.3 Biểu đồ hộp

Nó có một tên gọi khác là tóm tắt năm số, trong đó cần phân vị thứ nhất, phân vị thứ hai (trung vị), phân vị thứ ba, giá trị nhỏ nhất và giá trị lớn nhất để hiển thị. Ở đây, chúng ta cố gắng biểu diễn dữ liệu của mình trong một hộp có điểm giữa là trung vị của mẫu, phần trên cùng của hộp là phân vị thứ ba (Q3) và phần đáy của hộp là phân vị thứ nhất (Q1). Đường thước trên kéo dài đến giá trị liền kề này - giá trị dữ liệu cao nhất trong giới hạn trên = Q3 + 1,5 IQR, trong đó khoảng tứ phân vị IQR được định nghĩa là IQR = Q3 - Q1. Tương tự, đường thước dưới kéo dài đến

giá trị liền kề này - giá trị thấp nhất trong giới hạn dưới = Q1 - 1,5 IQR.



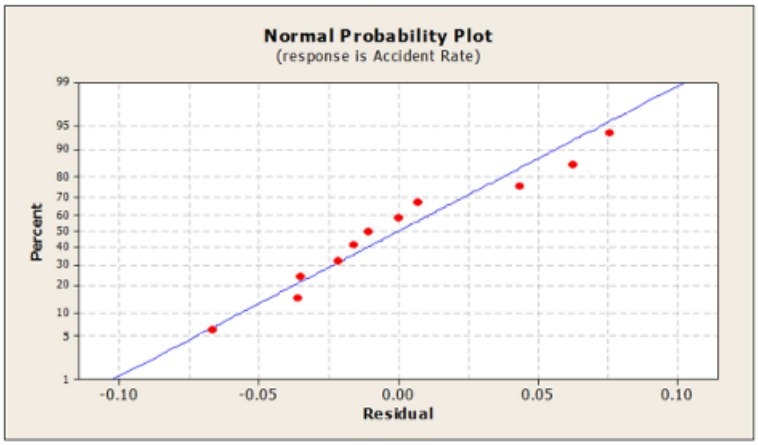
Hình 1.4: Biểu đồ hộp và râu thể hiện dữ liệu không được phân phối chuẩn

Chúng ta coi một quan sát là bất thường lớn hay nhỏ khi nó được biểu diễn ngoài các đường thước và chúng được xem là những giá trị ngoại lệ. Từ đồ thị này, chúng ta có thể nhận được sự chỉ dẫn rõ ràng về tính đối xứng của tập dữ liệu. Đồng thời nó cũng cho ta ý tưởng về sự phân tán của các quan sát. Do đó, mô hình phân phối chuẩn của dữ liệu cũng được hiểu từ đồ thị này. Đồ thị hộp được trình bày trong Hình 1.4 được lấy từ Imon và Das (2015) [9]. Đồ thị này rõ ràng cho thấy mô hình phi chuẩn của dữ liệu. Nó chứa giá trị ngoại lệ và dữ liệu thậm chí không đối xứng mà thực tế là bị nghiêng về bên phải.

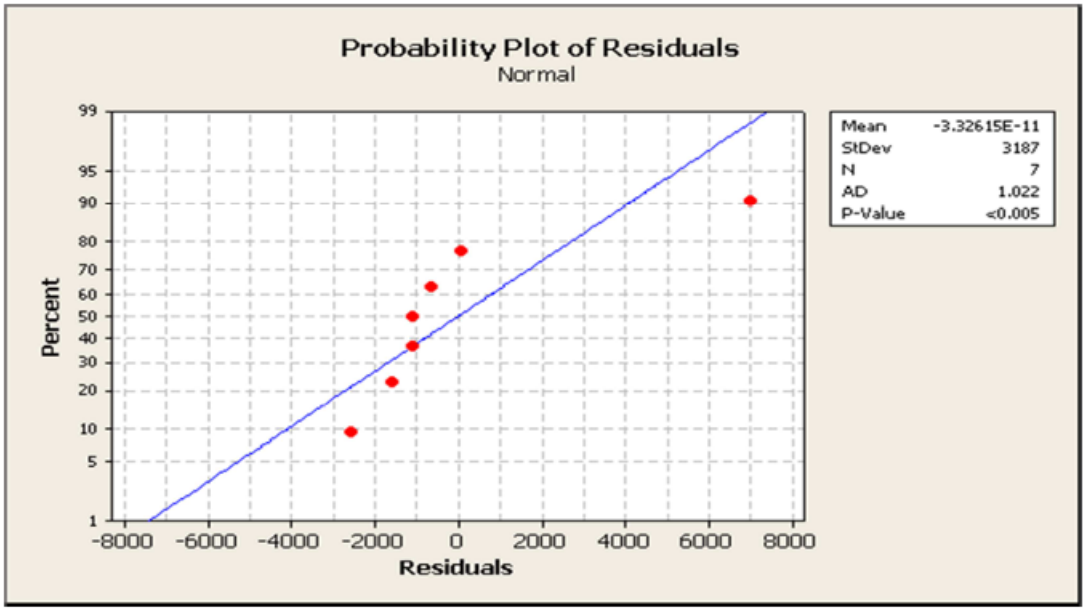
1.4 Biểu đồ phần trăm - phần trăm chuẩn

Trong thống kê, biểu đồ P-P (biểu đồ xác suất–xác suất hoặc biểu đồ phần trăm–phần trăm) là một biểu đồ xác suất để đánh giá mức độ tương đồng giữa hai tập dữ liệu, nó vẽ hai hàm phân phối tích lũy so sánh với nhau. Từ biểu đồ này, chúng ta có ý tưởng về các điểm ngoại lai, sự lệch và độ nhọn, và vì lý do này, nó đã trở thành một công cụ rất phổ biến để kiểm tra giả định về tính chuẩn. Biểu đồ P-P so sánh hàm phân phối tích lũy thực nghiệm của một tập dữ liệu với một hàm phân phối tích lũy lý thuyết đã được chỉ định F(·). Nếu nó trông giống như một đường thẳng hoặc không có đường cong, thì nó không chứa điểm ngoại lai và giả định được

cho là được thực hiện, và nếu nó hiển thị một cái nhìn khác ngoài đường thẳng (ví dụ: đường cong), thì giả định được cho là thất bại. Biểu đồ P-P chuẩn được trình bày trong Hình 1.5 và Hình 1.6 được lấy từ Imon (2015) [10]. Biểu đồ đầu tiên cho thấy một mẫu phân phối chuẩn và biểu đồ thứ hai cho thấy sự không chuẩn và sự tồn tại của một điểm ngoại lai.



Hình 1.5: Biểu đồ P-P thể hiện dữ liệu được phân phối chuẩn



Hình 1.6: Biểu đồ P-P thể hiện dữ liệu là phi chuẩn

1.5 Biểu đồ xác suất chuẩn

Đồ thị xác suất chuẩn (Q-Q) so sánh các phân vị của một phân phối dữ liệu với các phân vị của một phân phối lý thuyết chuẩn hóa từ một họ phân phối xác định.

Một đồ thị xác suất chuẩn là đồ thị mà chúng ta có thể định hình bằng cách vẽ đồ thị các phân vị của một phân phối so với các phân vị của phân phối chuẩn. Khi các phân vị của hai phân phối trùng nhau, các điểm được vẽ sẽ nằm trên đường thẳng $y = x$. Nếu đồ thị cho thấy một đường cong với độ dốc tăng dần từ trái sang phải, điều đó chỉ ra rằng phân phối dữ liệu bị nghiêng về phía bên phải, và nếu đường cong có độ dốc giảm dần từ trái sang phải, điều đó cho thấy sự nghiêng về bên trái của phân phối. Bằng cách sử dụng giấy xác suất chuẩn, một đồ thị Q-Q có thể được vẽ bằng tay dễ dàng. Trục hoành trên giấy xác suất được tỷ lệ cho các phân vị dự kiến của một phân phối chuẩn tiêu chuẩn, sao cho đồ thị của $(p, \phi^{-1}(p))$ là đường thẳng. Giới hạn trục hoành thường chạy từ 0,0001 đến 0,9999. Trục tung là tuyến tính và không đòi hỏi dữ liệu phải được chuẩn hóa theo bất kỳ cách nào; cũng có giấy xác suất được tỷ lệ logarit trên trục tung để kiểm tra xem dữ liệu có tuân theo phân phối chuẩn logarit hay không. Trên giấy xác suất, các cặp điểm $(p_i, x_{(i)})$ được vẽ. Đối với các đồ thị được vẽ bằng tay, ưu điểm của đồ thị Q-Q trên giấy xác suất chuẩn là các phân vị và xác suất tích lũy có thể được ước tính trực tiếp, và không cần phải tính toán $\phi^{-1}(p_i)$ để tạo đồ thị.

Có một nhầm lẫn lớn giữa đồ thị P-P và đồ thị Q-Q, đôi khi người ta nghĩ rằng chúng là đồng nghĩa. Nhưng có ba khác biệt quan trọng trong cách xây dựng và diễn giải đồ thị P-P và Q-Q:

- Việc xây dựng đồ thị Q-Q không đòi hỏi phải chỉ định tham số vị trí hoặc tham số tỷ lệ của $F(\cdot)$. Các phân vị lý thuyết được tính toán từ một phân phối chuẩn hóa trong họ phân phối đã chỉ định. Một mô hình điểm tuyến tính cho thấy họ phân phối đã chỉ định mô tả hợp lý phân phối dữ liệu, và các tham số vị trí và tỷ lệ có thể được ước lượng bằng mắt thông qua giao điểm và hệ số góc của mô hình tuyến tính.Ngược lại, việc xây dựng đồ thị P-P đòi hỏi phải có tham số vị trí và tỷ lệ của $F(\cdot)$ để tính toán hàm phân phối tích lũy tại các giá trị dữ liệu đã được sắp xếp.

- Tính tuyến tính của mô hình điểm trên đồ thị Q-Q không bị ảnh hưởng bởi thay đổi vị trí hoặc tỷ lệ. Trên đồ thị P-P, sự thay đổi vị trí hoặc tỷ lệ không nhất thiết bảo toàn tính tuyến tính.

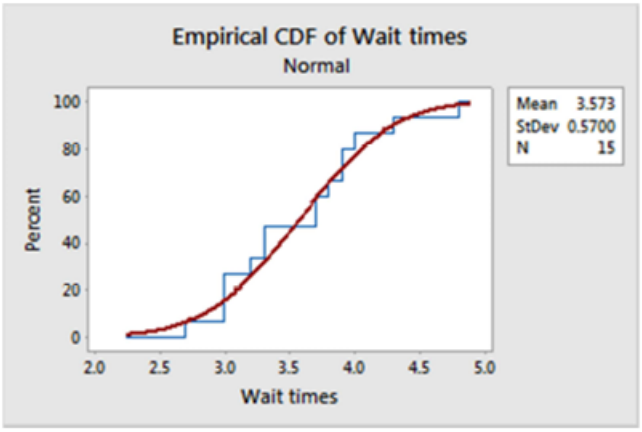
- Trên đồ thị Q-Q, đường tham chiếu đại diện cho một phân phối lý thuyết cụ thể phụ thuộc vào tham số vị trí và tỷ lệ của phân phối đó, có giao điểm và hệ số góc tương ứng với tham số vị trí và tỷ lệ. Trên đồ thị P-P, đường tham chiếu cho bất kỳ phân phối nào luôn là đường chéo $y = x$.

Do đó, bạn nên sử dụng đồ thị Q-Q nếu mục tiêu của bạn là so sánh phân phối dữ liệu với một họ phân phối chỉ khác nhau về vị trí và tỷ lệ, đặc biệt nếu bạn muốn ước lượng các tham số vị trí và tỷ lệ từ đồ thị.

Một lợi thế của đồ thị P-P là chúng có khả năng phân biệt ở các vùng có mật độ xác suất cao, bởi vì ở những vùng này, các phân phối tích lũy thực nghiệm và lý thuyết thay đổi nhanh hơn so với các vùng có mật độ xác suất thấp. Ví dụ, nếu bạn so sánh phân phối dữ liệu với một phân phối chuẩn cụ thể, thì sự khác biệt ở giữa hai phân phối sẽ rõ ràng hơn trên đồ thị P-P so với đồ thị Q-Q.

1.6 Biểu đồ hàm phân phối tích lũy thực nghiệm

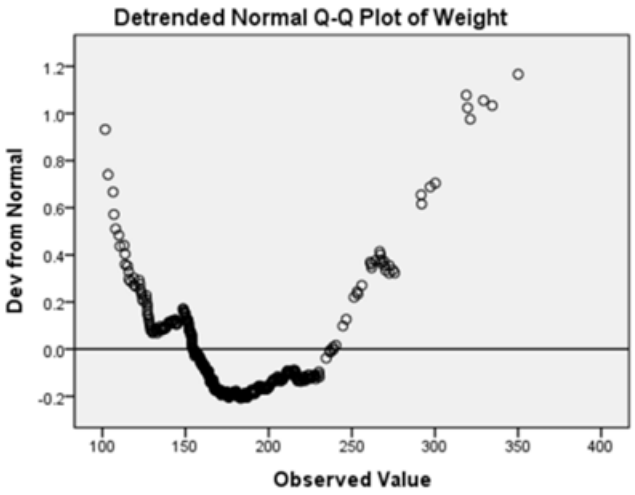
Một đồ thị hàm phân phối tích lũy thực nghiệm thực hiện chức năng tương tự như một đồ thị xác suất. Tuy nhiên, khác với đồ thị xác suất, đồ thị hàm phân phối tích lũy thực nghiệm có các thang đo không được biến đổi và phân phối phù hợp không tạo thành một đường thẳng, mà nó cho ra một đường cong hình chữ S dưới giả định phân phối chuẩn. Các xác suất tích lũy thực nghiệm gần với đường cong chữ S này thỏa mãn giả định phân phối chuẩn.



Hình 1.7: Biểu đồ hàm phân phối tích lũy thực nghiệm thể hiện dữ liệu được phân phối chuẩn.

1.7 Biểu đồ xác suất tách xu hướng

Trong thống kê, một đồ thị biểu diễn sự khác biệt giữa các giá trị quan sát và giá trị dự kiến, trong đó giá trị dự kiến dựa trên giả định phân phối chuẩn. Nếu các điểm số quan sát tuân theo phân phối chuẩn, thì các điểm sẽ tụ lại trong một dải nằm ngang gần với không mà không có bất kỳ mô hình nào đáng kể. Đây cũng được gọi là đồ thị Q-Q tách xu hướng vì ở đây $(x_{(i)} - \widehat{\sigma}\Phi^{-1}(p_i))$ được vẽ đối với vị trí đồ thị p_i hoặc phân vị dự kiến $\Phi^{-1}(p_i)$ cho một ước lượng nào đó của độ lệch chuẩn $\widehat{\sigma}$. Nếu các quan sát đến từ phân phối chuẩn, kết quả sẽ là một đường thẳng với hệ số góc bằng không.



Hình 1.8: Biểu đồ xác suất chuẩn tách xu hướng.

Chương 2

Các phương pháp suy luận thống kê để kiểm định phân phối chuẩn

Các loại đo lường mô tả khác nhau như mô-men, cumulant[1], hệ số nghiêng và nhọn, trung bình tuyệt đối, phạm vi của mẫu, v.v. và hàm phân phối thực nghiệm đã được đề xuất để sử dụng trong các kiểm tra phân phối chuẩn, nhưng chỉ có một số ít trong số đó được sử dụng thường xuyên trên thực tế. Ở đây, chúng ta phân loại các kiểm định thành hai nhóm: kiểm định dựa trên hàm phân phối thực nghiệm (EDF) và kiểm định dựa trên các đo lường mô tả.

2.1 Kiểm định hàm phân phối thực nghiệm (EDF)

Dựa trên đo lường sai lệch giữa phân phối thực nghiệm và phân phối giả định, thường được đề cập đến là hàm phân phối thực nghiệm, chúng ta có thể xác định các kiểm định sau.

2.1.1 Kiểm định Kolmogorov-Smirnov

Kiểm định Kolmogorov-Smirnov lần đầu tiên được xây dựng bởi Kolmogorov (1933) [11] và sau đó được sửa đổi và đề xuất bởi Smirnov (1948) [12]. Đại lượng kiểm định là :

$$D = sup_x|F_n(X) - F(X, \mu, \sigma)| \tag{1}$$

[1]cumulant : bất kỳ hệ số thống kê nào phát sinh trong khai triển chuỗi theo lũy thừa của x của logarit hàm tạo mô-men (xem `https://www.merriam-webster.com/dictionary/cumulant`)

Trong đó, $F(X, \mu, \sigma)$ là hàm phân phối tích lũy lý thuyết của phân phối chuẩn và $F_n(X)$ là hàm phân phối thực nghiệm của dữ liệu. Nếu D cho ra giá trị lớn thì thể hiện rằng dữ liệu không tuân theo phân phối chuẩn. Khi các tham số quần thể (μ và σ) không được biết thì ước lượng mẫu sẽ được sử dụng thay cho giá trị tham số.

Giá trị tới hạn của kiểm định thống kê $D_{n,\alpha}$ được cho trong bảng 2.5 tại phụ lục. Nếu $D \geq D_{n,\alpha}$ thì bác bỏ giả thuyết không, nếu không thì không thì chấp nhận giả thuyết không.

2.1.2 Kiểm định Shapiro-Wilk

Kiểm định Shapiro-Wilk là một trong những kiểm định phổ biến nhất đối với giả định phân phối chuẩn, có tính năng về sức mạnh tốt và dựa trên tương quan bên trong các quan sát cho trước và điểm số chuẩn hóa tương ứng. Đại lượng kiểm định Shapiro-Wilk được xây dựng bởi Shapiro và Wilk (1965) [13]. Dạng của đại lượng kiểm định là :

$$W = \frac{\left(\sum a_i y_{(i)}\right)^2}{\sum (y - \bar{y})^2} \tag{2}$$

Trong đó $y_(i)$ là thống kê hạng thứ i và a_i là giá trị kỳ vọng của thống kê chuẩn hóa hạng thứ i. Đối với các quan sát độc lập và tuân theo cùng một phân phối, các giá trị của a_i có thể được lấy từ bảng do Shapiro và Wilk (1965) [13] trình bày cho kích thước mẫu lên đến 50. W có thể biểu diễn dưới dạng bình phương của hệ số tương quan giữa a_i và $y_{(i)}$. Do đó, W không phụ thuộc vào vị trí và tỷ lệ và luôn nhỏ hơn hoặc bằng 1. Trên đồ thị $y_{(i)}$ đối với a_i, một đường thẳng chính xác sẽ dẫn đến W rất gần 1. Vì vậy, nếu W nhỏ hơn 1 rất nhiều, giả thuyết phân phối chuẩn sẽ bị bác bỏ.

Công thức tính giá trị p:

$$\mu = 0.0038915 \cdot \ln(N)^3 - 0.083751 \cdot \ln(N)^2 - 0.31082 \cdot \ln(N) - 1.5861$$

$$\sigma = e^{0.0030302 \cdot \ln(N)^2 - 0.082676 \cdot \ln(N) - 0.4803}$$

$$z = \frac{\ln(1 - W) - \mu}{\sigma}$$

$$p = 1 - P(Z < z)$$

Mặc dù kiểm định Shapiro-Wilk W rất phổ biến, nhưng nó phụ thuộc vào khả năng cung cấp các giá trị của a_i, và đối với trường hợp mẫu lớn, việc tính toán chúng

có thể phức tạp hơn nhiều. Một số điều chỉnh nhỏ đối với kiểm định W đã được đề xuất bởi Shapiro và Francia (1972) [14], Weisberg và Bingham (1975) [15] và Royston (1982) [16]. Một kiểm định thay thế cùng bản chất đối với mẫu lớn hơn 50 đã được thiết kế bởi D'Agostino (1971) [17].

2.1.3 Kiểm định Anderson-Darling

Stephens (1974) [18] đã đề xuất một kiểm định dựa trên phân phối thực nghiệm bằng cách mở rộng công trình của Anderson và Darling (1952) [19]. Kiểm định này thường được gọi là kiểm định phân phối chuẩn Anderson-Darling. Đối với các quan sát ngẫu nhiên $y_(i)$ tuân theo phân phối chuẩn với trung bình μ và phương sai σ^2, đại lượng kiểm định Anderson-Darling được cho bởi:

$$AD = -\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right)\left[\frac{\sum_{i=1}^{n}\left[(2i-1)\log\{\hat{z}_i(1-\hat{z}_{n+1-i})\}\right]}{n} + n\right] \tag{3}$$

Trong đó $\hat{z}_i = \Phi\left[(y_{(i)} - \hat{\mu})/\hat{\sigma}\right]$ and $\Phi(\bullet)$ là hàm phân phối của biến ngẫu nhiên N (0,1). Stephen (1974) [18] đã cung cấp các điểm phần trăm cho kiểm định này.

Công thức tính giá trị p:

AD	Giá trị p
$AD \leq .2$	$1 - \exp\left(-13.436 + 101.14AD - 223.73AD^2\right)$
$.2 < AD \leq .34$	$1 - \exp\left(-8.318 + 42.796AD - 59.938AD^2\right)$
$.34 < AD < .6$	$\exp\left(0.9177 - 4.279AD - 1.38AD^2\right)$
$AD \geq .6$	$\exp\left(1.2937 - 5.709AD + 0.0186AD^2\right)$

2.2 Kiểm định dựa trên đo lường thống kê mô tả

Fisher (1930) [20] đã đề xuất sử dụng cumulant. Dựa trên kết quả của ông, Pearson (1930) [21] đã thu được bốn mô-men đầu tiên của phân phối mẫu của hệ số nghiêng và nhọn, dưới giả thuyết không có phân phối chuẩn. Ông đã sử dụng những kết quả đó để phát triển các tiêu chí kiểm tra phân phối chuẩn bằng cách sử dụng các giá trị mẫu của hệ số nghiêng và nhọn riêng rẽ. Tỷ lệ trung bình tuyệt đối so với độ lệch chuẩn [xem Geary (1935) [22]] và tỷ lệ phạm vi mẫu so với độ lệch chuẩn [xem David, Hartley và Pearson (1954) [23]] cũng được đề xuất cho cùng mục đích. Dựa trên các

mô-men, các kiểm tra phổ biến nhất là kiểm định toàn diện D'Agostino-Pearson và kiểm định Jarqua-Bera.

2.2.1 Kiểm định D'Agostino-Pearson

Để đánh giá sự đối xứng hoặc thiên lệch, thường đo lường độ nghiêng và để đánh giá hình dạng của phân phối, độ nhọn được bỏ qua. Kiểm tra D'Agostino-Pearson (1973) [24] dựa trên kiểm tra độ nghiêng và nhọn, và những đặc tính này cũng được đánh giá thông qua các mô-men. Đại lượng kiểm tra DAP là:

$$K^2 = Z^2(\sqrt{b_1}) + Z^2(b_2) \tag{4}$$

Trong đó $Z(\sqrt{b_1})$ and $Z(b_2)$ là xấp xỉ chuẩn tương ứng với $\sqrt{b_1}$ và b_2, lần lượt là độ nghiêng và nhọn của mẫu. Đại lượng thống kê này tuân theo phân phối χ^2 với hai bậc tự do nếu quần thể đến từ phân phối chuẩn. Một giá trị lớn của K^2 dẫn đến việc bác bỏ giả thuyết phân phối chuẩn.

2.2.2 Kiểm định Jarqua-Bera

Kiểm định Jarqua-Bera ban đầu được đề xuất bởi Bowman và Shenton (1975) [25]. Họ kết hợp bình phương của độ nghiêng và nhọn chuẩn hóa trong một đại lượng thống kê duy nhất như sau:

$$JB = \frac{n}{6}\left[S^2 + \frac{(K-3)^2}{4}\right] \tag{5}$$

Việc chuẩn hóa này dựa trên tính phân phối chuẩn vì $S = 0$ và $K = 3$ đối với phân phối chuẩn và phương sai giới hạn tương ứng là $6/n$ và $24/n$. Do đó, dưới giả định phân phối chuẩn, đại lượng kiểm định JB cũng tuân theo phân phối χ^2 với hai bậc tự do. Một giá trị JB lớn đáng kể dẫn đến việc bác bỏ giả định phân phối chuẩn.

Phần II

Ứng dụng của kiểm định phân phối chuẩn vào phân tích thống kê

Bài toán: Cho bộ dữ liệu kích thước n = 42, nghiên cứu xem số ngẫu nhiên được tạo ra có tuân theo phân phối chuẩn hay không. Với mức ý nghĩa $\alpha = 0.05$, sử dụng các phương pháp kiểm định đơn biến khác nhau.

Giả thuyết không (H_0): Số ngẫu nhiên tuân theo phân phối chuẩn.

Giả thuyết khác (H_a): Số ngẫu nhiên không tuân theo phân phối chuẩn.

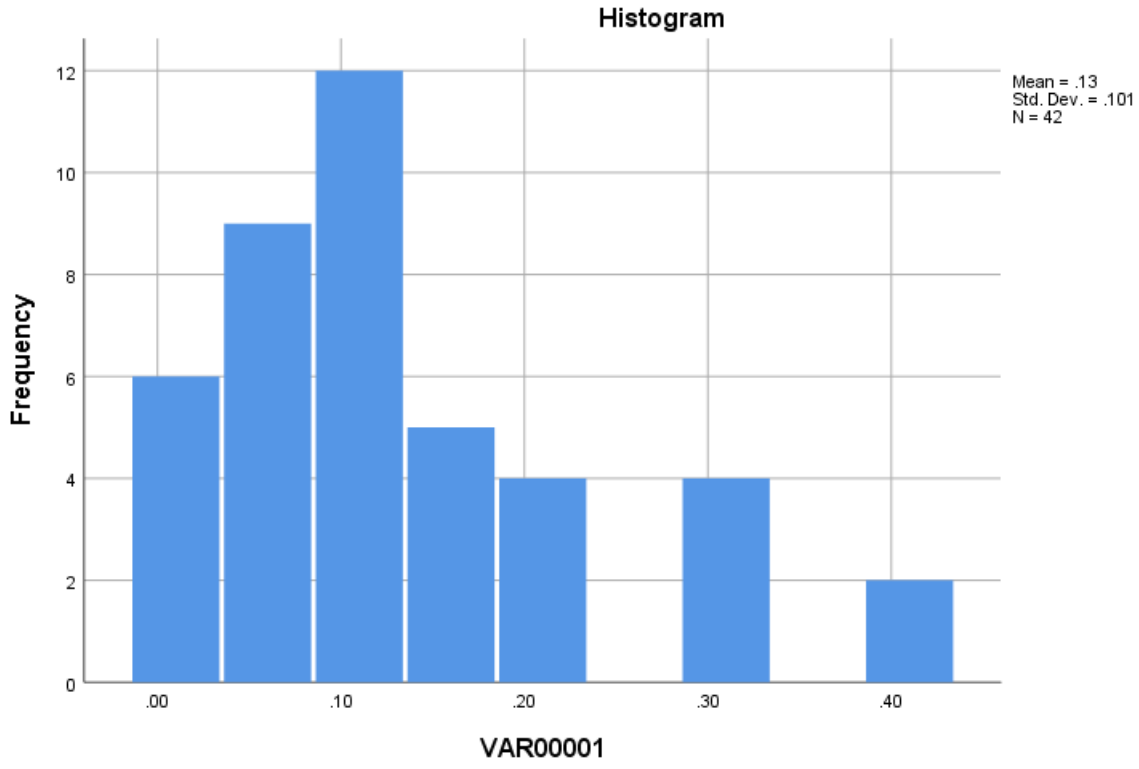
STT	**Giá trị**	**STT**	**Giá trị**	**STT**	**Giá trị**	**STT**	**Giá trị**
1	0.15	12	0.02	23	0.03	34	0.30
2	0.09	13	0.01	24	0.05	35	0.02
3	0.18	14	0.10	25	0.15	36	0.20
4	0.10	15	0.10	26	0.10	37	0.20
5	0.05	16	0.20	27	0.15	38	0.30
6	0.12	17	0.02	28	0.09	39	0.30
7	0.08	18	0.10	29	0.08	40	0.40
8	0.05	19	0.01	30	0.18	41	0.30
9	0.09	20	0.40	31	0.10	42	0.05
10	0.10	21	0.10	32	0.20		
11	0.07	22	0.05	33	0.11		

Bảng 2.1: Số ngẫu nhiên được tạo.

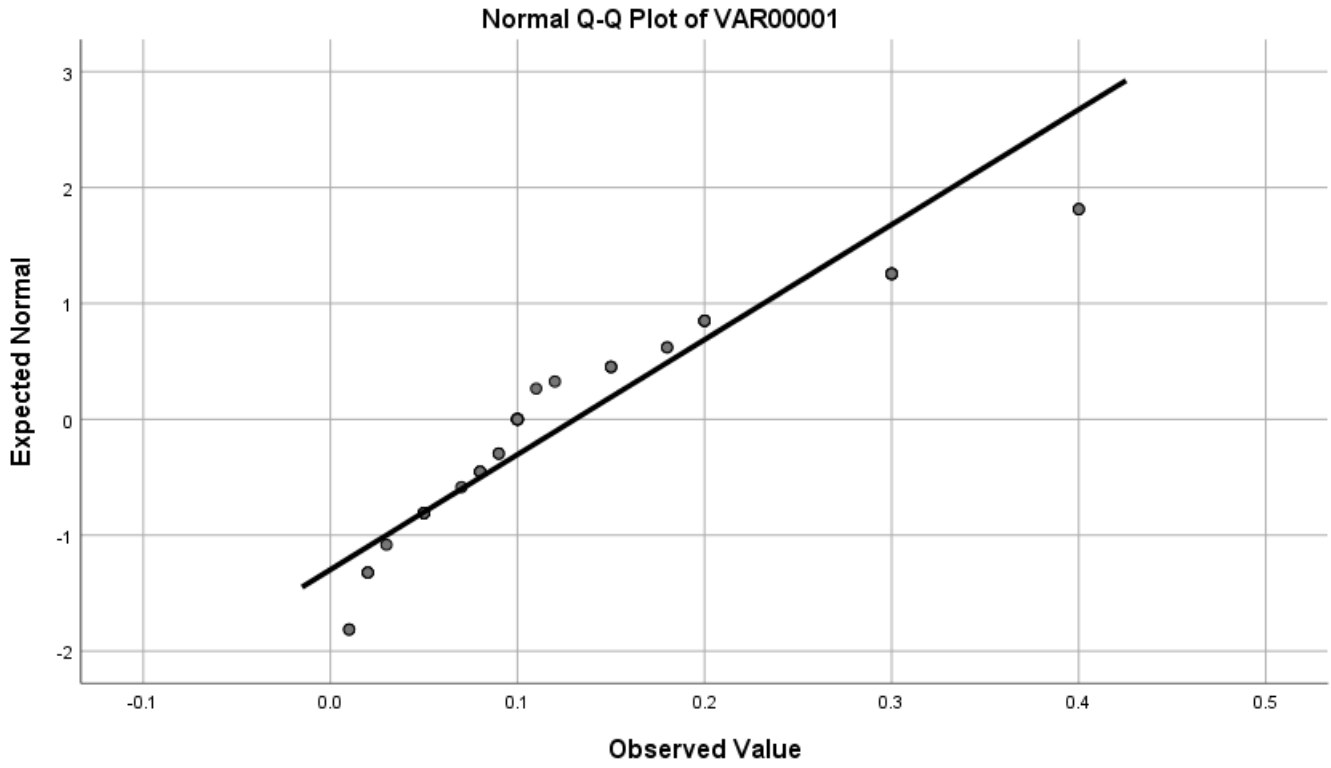
Trung bình mẫu $\mu = 0.1307$, độ lệch chuẩn $\sigma = 0.10076$

2.3 Phương pháp đồ thị

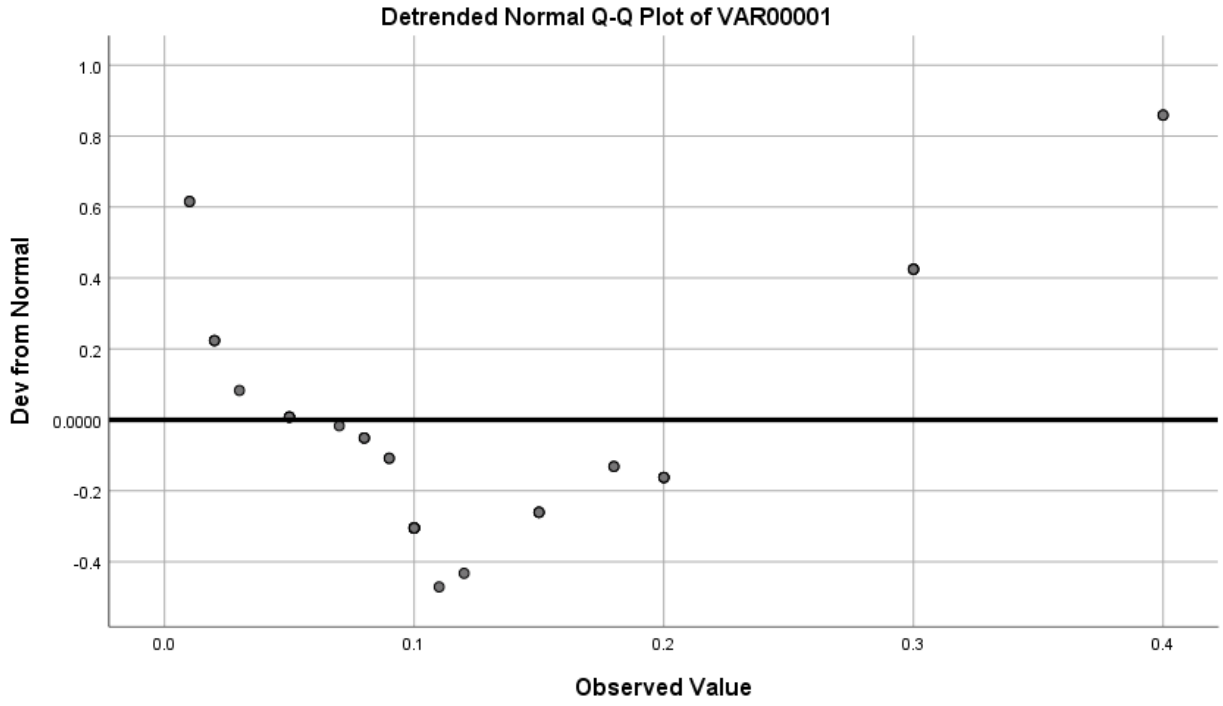
Qua một số biểu đồ được thể hiện ở trên(2.3, 2.3, 2.3, 2.4), ta có một cái nhìn tổng quan về bộ dữ liệu. Dữ liệu được phân bố không đều ở trung tâm, cách xa đường chuẩn và phân tán. Chính vì vậy, ta dự đoán bộ dữ liệu số ngẫu nhiên này không tuân theo phân phối chuẩn. Ta sẽ thực hiện các thủ tục kiểm định để kiểm chứng dự đoán này.



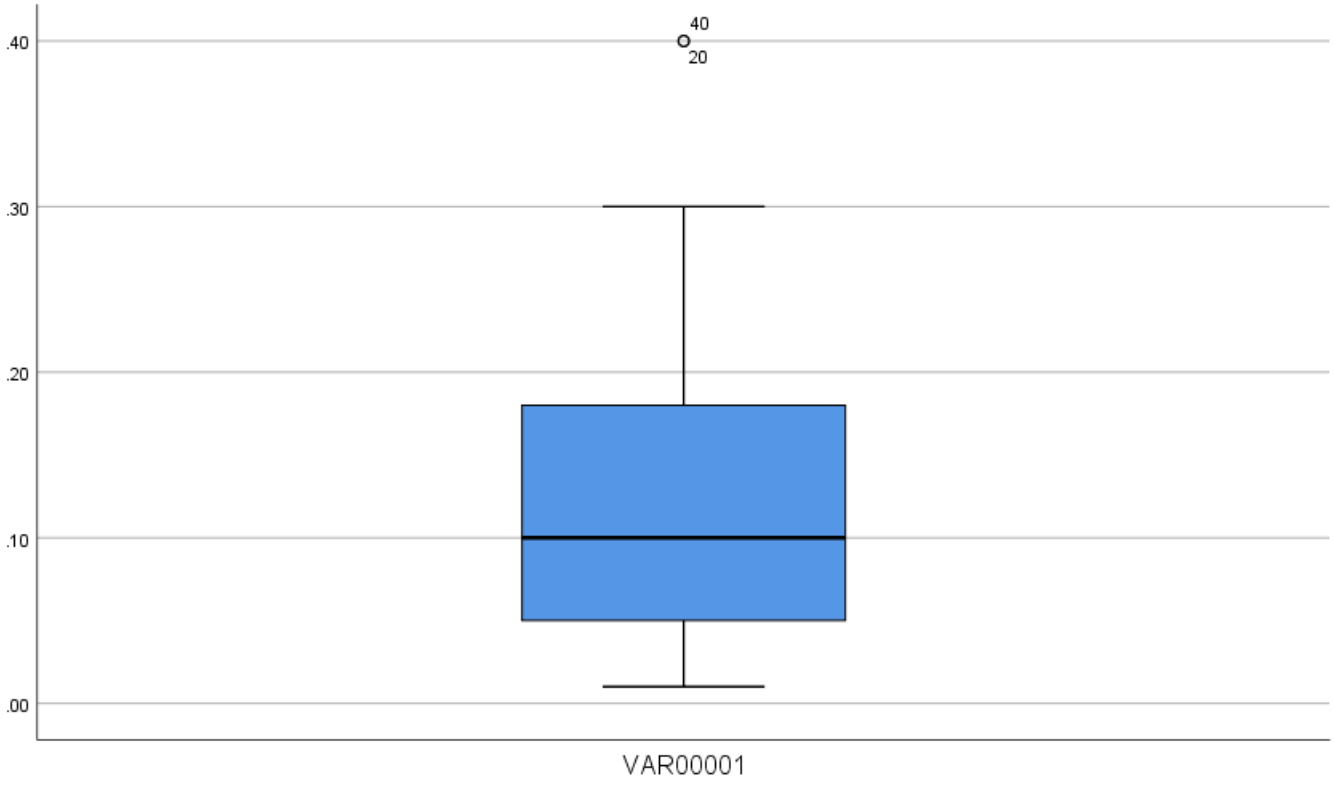
Hình 2.1: Biểu đồ tần suất.



Hình 2.2: Biểu đồ xác suất chuẩn.



Hình 2.3: Biểu đồ xác suất tách xu hướng.



Hình 2.4: Biểu đồ hộp và râu.

i	x_i	**Tần suất**	f_0	**Phần trăm tích lũy** (F_n)	z_i	$F_0 = P(Z < x_i)$	$\lvert F_n - F_0 \rvert$
1	0.01	2	0.048	0.048	-1.198	0.115458	0.0678
2	0.02	3	0.071	0.119	-1.0976	0.135936	0.0169
3	0.03	1	0.024	0.143	-0.99952	0.158772	0.0158
4	0.05	5	0.119	0.262	-0.80103	0.211556	0.0503
5	0.06	1	0.024	0.286	-0.72645	0.232005	0.0123
6	0.07	3	0.071	0.357	-0.5033	0.307375	0.0496
7	0.09	2	0.048	0.405	-0.40406	0.343084	0.0619
8	0.1	8	0.190	0.595	-0.30482	0.380252	0.2148
9	0.11	1	0.024	0.619	-0.20557	0.418562	0.2004
10	0.12	1	0.024	0.643	-0.10632	0.45766	0.1853
11	0.15	3	0.071	0.714	0.191397	0.575893	0.1381
12	0.2	2	0.048	0.762	0.489126	0.687624	0.0744
13	0.3	4	0.095	0.857	0.878612	0.754152	0.1028
14	0.35	4	0.095	0.952	1.680042	0.953525	0.0011
15	0.4	2	0.048	1.000	2.672472	0.996235	0.0038
Tổng		42	1.000				

Bảng 2.2: Tính toán kiểm định KS.

2.4 Kiểm định Komgorov-Smirnov

Kết quả từ bảng trên, ta có đại lượng kiểm định $D = sup_x|F_n(X) - F(X, \mu, \sigma)| = 0.215$. Với $\alpha = 0.05$ và $n = 42$, giá trị tới hạn được lấy từ bảng 2.5 $D_{n,\alpha} = 0.2099$. Do $D = 0.215 > D_{n,\alpha} = 0.2099$ nên ta bác bỏ giả thuyết H_0 tức là các số ngãu nhiên này không tuân theo phân phối chuẩn.

2.5 Kiểm định Shapiro-Wilk

Từ bảng 2.3 ta tính được đại lượng kiểm định $W = \frac{\left(\sum a_i y_{(i)}\right)^2}{\sum (y-\bar{y})^2} = 0.8663$. Áp dụng công thức tính p ở phần lí thuyết, ta có $p = 0.000163 < 0.05 = \alpha$. Do đó, ta bác bỏ H_0 tức là số ngẫu nhiên không tuân theo phân phối chuẩn.

Bảng 2.3: Tính toán Shapiro-Wilk

i	y_i	$y_{(n+1-i)}$	$(y_i - \bar{y})^2$	a_i	$a_i(y_{(n+1-i)} - y_i)$
1	0.01	0.4	0.01456849	0.3917	0.152763
2	0.01	0.4	0.01456849	0.2701	0.105339
3	0.02	0.3	0.01225449	0.2345	0.06566
4	0.02	0.3	0.01225449	0.2085	0.05838
5	0.02	0.3	0.01225449	0.1874	0.052472
6	0.03	0.3	0.01014049	0.1694	0.045738
7	0.05	0.2	0.00651249	0.1535	0.023025
8	0.05	0.2	0.00651249	0.1392	0.02088
9	0.05	0.2	0.00651249	0.1259	0.018885
10	0.05	0.2	0.00651249	0.1136	0.01704
11	0.05	0.18	0.00651249	0.1020	0.01326
12	0.07	0.18	0.00368449	0.0909	0.009999
13	0.08	0.15	0.00257049	0.0804	0.005628
14	0.08	0.15	0.00257049	0.0701	0.004907
15	0.08	0.15	0.00257049	0.0602	0.004214
16	0.09	0.12	0.00165649	0.0506	0.001518
17	0.09	0.11	0.00165649	0.0411	0.000822
18	0.1	0.1	0.00094249	0.0318	0
19	0.1	0.1	0.00094249	0.0227	0
20	0.1	0.1	0.00094249	0.0136	0
21	0.1	0.1	0.00094249	0.0045	0
22	0.1	0.1	0.00094249	0	0
23	0.1	0.1	0.00094249	0	0
24	0.1	0.1	0.00094249	0	0
25	0.1	0.1	0.00094249	0	0
26	0.11	0.09	0.00042849	0	0
27	0.12	0.09	0.00011449	0	0
28	0.15	0.08	0.00037249	0	0
29	0.15	0.08	0.00037249	0	0
30	0.15	0.08	0.00037249	0	0
31	0.18	0.07	0.00243049	0	0
32	0.18	0.05	0.00243049	0	0
33	0.2	0.05	0.00480249	0	0
34	0.2	0.05	0.00480249	0	0
35	0.2	0.05	0.00480249	0	0
36	0.2	0.05	0.00480249	0	0
37	0.3	0.03	0.02866249	0	0
38	0.3	0.02	0.02866249	0	0
39	0.3	0.02	0.02866249	0	0
40	0.3	0.02	0.02866249	0	0
41	0.4	0.01	0.07252249	0	0
42	0.4	0.01	0.07252249	0	0

2.6 Kiểm định Anderson-Darling

Trong đó, $S_i = [(2i-1)\log\{\hat{z}_i(1-\hat{z}_{n+1-i})\}]$.

Từ bảng 2.4, tính

$$AD = -\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right)\left[\frac{\sum_{i=1}^{n} S_i}{n} + n\right] = 1.8441$$

Sử dụng công thức tính giá trị p trong phần lí thuyết,

$$p = \exp\left(1.2937 - 5.709AD + 0.0186AD^2\right) = 1.04 \cdot 10^{-4}$$

Do $p = 1.04 \cdot 10^{-4} < \alpha = 0.05$ nên ta bác bỏ giả thuyết H_0 tức là số ngẫu nhiên không tuân theo phân phối chuẩn.

i	y_i	$\hat{z}_i$	$1-\hat{z}_i$	$1-\hat{z}_{n+1-i}$	S_i
1	0.01	0.1155	0.8845	0.0038	-7.741
2	0.01	0.1155	0.8845	0.0038	-23.224
3	0.02	0.1360	0.8640	0.0465	-25.323
4	0.02	0.1360	0.8640	0.0465	-35.452
5	0.02	0.1360	0.8640	0.0465	-45.582
6	0.03	0.1588	0.8412	0.0465	-54.003
7	0.05	0.2116	0.7884	0.2458	-38.432
8	0.05	0.2116	0.7884	0.2458	-44.345
9	0.05	0.2116	0.7884	0.2458	-50.258
10	0.05	0.2116	0.7884	0.2458	-56.171
11	0.05	0.2116	0.7884	0.3123	-57.053
12	0.07	0.2734	0.7266	0.3123	-56.589
13	0.08	0.3074	0.6926	0.4240	-50.936
14	0.08	0.3074	0.6926	0.4240	-55.011
15	0.08	0.3074	0.6926	0.4240	-59.086
16	0.09	0.3431	0.6569	0.5423	-52.130
17	0.09	0.3431	0.6569	0.5814	-53.195
18	0.10	0.3803	0.6197	0.6197	-50.586
19	0.10	0.3803	0.6197	0.6197	-53.477
20	0.10	0.3803	0.6197	0.6197	-56.367
21	0.10	0.3803	0.6197	0.6197	-59.258
22	0.10	0.3803	0.6197	0.6197	-62.148
23	0.10	0.3803	0.6197	0.6197	-65.039
24	0.10	0.3803	0.6197	0.6197	-67.930
25	0.10	0.3803	0.6197	0.6197	-70.820
26	0.11	0.4186	0.5814	0.6569	-65.845
27	0.12	0.4577	0.5423	0.6569	-63.694
28	0.15	0.5760	0.4240	0.6926	-50.549
29	0.15	0.5760	0.4240	0.6926	-52.387
30	0.15	0.5760	0.4240	0.6926	-54.225
31	0.18	0.6877	0.3123	0.7266	-42.326
32	0.18	0.6877	0.3123	0.7884	-38.567
33	0.20	0.7542	0.2458	0.7884	-33.789
34	0.20	0.7542	0.2458	0.7884	-34.829
35	0.20	0.7542	0.2458	0.7884	-35.868
36	0.20	0.7542	0.2458	0.7884	-36.908
37	0.30	0.9535	0.0465	0.8412	-16.096
38	0.30	0.9535	0.0465	0.8640	-14.528
39	0.30	0.9535	0.0465	0.8640	-14.915
40	0.30	0.9535	0.0465	0.8640	-15.303
41	0.40	0.9962	0.0038	0.8845	-10.245
42	0.40	0.9962	0.0038	0.8845	-10.498

Bảng 2.4: Tính toán AD

2.7 Kiểm định D'Agostino-Pearson

a) Độ nghiêng

Tính toán độ nghiêng mẫu:

$$\sqrt{b_1} = \frac{m_3}{m_2^{3/2}} = 1.1544$$

trong đó $m_k = \sum_i (X_i - \overline{X})^k / n$, $\overline{X} = \sum_i X_i / n$.

Tính:

$$Y = \sqrt{b_1} \left(\frac{(n+1)(n+3)}{6(n+2)} \right)^{1/2} = 3.1253$$

$$\beta_2(\sqrt{b_1}) = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)} = 3.4943$$

$$W^2 = -1 + \left(2(\beta_2(\sqrt{b_1}) - 1) \right)^{1/2} = 1.2335$$

$$\delta = \frac{1}{\sqrt{\ln W}} = 3.0870$$

$$\alpha = \left(\frac{2}{W^2 - 1} \right)^{1/2} = 2.9267$$

$$Z(\sqrt{b_1}) = \delta \ln \left(\frac{Y}{\alpha} + \left(\left(\frac{Y}{\alpha} \right)^2 + 1 \right)^{1/2} \right) = 2.8664$$

b) Độ nhọn

Tính độ nhọn mẫu:

$$b_2 = \frac{m_4}{m_2^2} = 3.6751$$

trong đó $m_k = \sum_i (X_i - \overline{X})^k / n$, $\overline{X} = \sum_i X_i / n$.

Tính:

$$E\{b_2\} = \frac{3(n-1)}{n+1} = 2.8605$$

$$\text{var}\{b_2\} = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)} = 0.4021$$

$$x = \frac{(b_2 - E\{b_2\})}{\sqrt{\text{var}\{b_2\}}} = 1.2846$$

$$\sqrt{\beta_1(b_2)} = \frac{6(n^2-5n+2)}{(n+7)(n+9)} \sqrt{\frac{6(n+3)(n+5)}{n(n-2)(n-3)}} = 1.6441$$

$$A = 6 + \frac{8}{\sqrt{\beta_1(b_2)}} \left(\frac{2}{\sqrt{\beta_1(b_2)}} + \sqrt{1 + \frac{4}{\beta_1(b_2)}} \right) = 19.5817$$

$$Z(b_2) = \left(1 - \frac{2}{9A} - \left(\frac{1-\frac{2}{A}}{1+x\sqrt{\frac{2}{A-4}}}\right)^{1/3}\right) \Big/ \sqrt{\frac{2}{9A}} = 1.2983$$

c) Kiểm định D'Agostino-Pearson

Đại lượng kiểm định tính toán dựa trên xấp xỉ chuẩn của độ nghiêng và độ nhọn của mẫu:

$$K^2 = Z^2(\sqrt{b_1}) + Z^2(b_2) = 9.9018$$

Do $Z^2(\sqrt{b_1}) + Z^2(b_2) \sim \chi^2(2)$ nên ta tính giá trị p từ K^2 sử dụng phân phối khi bình phương với 2 bậc tự do: $p = 0.0071 < \alpha = 0.05$. Vậy ta bác bỏ giả thuyết H_0 .

2.8 Kiểm định Jarqua-Bera

Trước tiên, ta tính các giá trị độ nghiêng và độ nhọn lấy từ phần trước:

$$S = \sqrt{b_1} = 1.1544$$

$$K = b_2 = 3.6751$$

$$JB = \frac{n}{6}\left[S^2 + \frac{(K-3)^2}{4}\right] = 10.1261$$

Do $JB \sim \chi^2(2)$ nên ta tính giá trị p từ thống kê JB sử dụng phân phối khi bình phương với 2 bậc tự do: $p = 0.00633 < \alpha = 0.05$ nên ta bác bỏ giả thuyết H_0.

2.9 Nhận xét

Với các phương pháp đã được trình bày ở trên để kiểm định phân phối chuẩn của các số ngẫu nhiên, ta có nhận xét khái quát:

Phương pháp đồ thị

Các phương pháp đồ thị như biểu đồ tần suất, biểu đồ thân và lá, biểu đồ hộp và râu, biểu đồ P-P, biểu đồ Q-Q và biểu đồ hàm phân phối tích lũy thực nghiệm đều đã được sử dụng để đánh giá tính chuẩn của dữ liệu. Các đồ thị này giúp chúng ta có cái nhìn trực quan về sự phân bố của dữ liệu, phát hiện các điểm ngoại lệ, độ nghiêng và tính đối xứng của dữ liệu. Kết quả từ các đồ thị cho thấy:

- Biểu đồ tần suất và biểu đồ thân và lá cho thấy dữ liệu không có dạng hình chuông điển hình của phân phối chuẩn.

- Biểu đồ hộp và râu cùng với các biểu đồ P-P và Q-Q chỉ ra sự tồn tại của các điểm ngoại lệ và sự lệch của dữ liệu so với phân phối chuẩn lý thuyết.

Phương pháp suy luận thống kê

Chúng ta đã áp dụng các kiểm định thống kê khác nhau bao gồm kiểm định Kolmogorov-Smirnov, kiểm định Shapiro-Wilk, kiểm định Anderson-Darling, kiểm định D'Agostino-Pearson và kiểm định Jarqua-Bera. Các kiểm định này đánh giá sự khác biệt giữa phân phối thực nghiệm và phân phối chuẩn lý thuyết dựa trên các mô-men và hàm phân phối tích lũy. Kết quả từ các kiểm định này cho thấy:

- Kiểm định Kolmogorov-Smirnov và kiểm định Shapiro-Wilk cho kết quả bác bỏ giả thuyết không với mức ý nghĩa $\alpha = 0.05$, cho thấy dữ liệu không tuân theo phân phối chuẩn.

- Kiểm định Anderson-Darling cũng bác bỏ giả thuyết không, cho thấy sự khác biệt rõ rệt giữa phân phối dữ liệu và phân phối chuẩn lý thuyết.

- Kiểm định D'Agostino-Pearson và kiểm định Jarqua-Bera chỉ ra rằng dữ liệu có độ nghiêng và nhọn không phù hợp với phân phối chuẩn.

Kết hợp các kết quả từ phương pháp đồ thị và các kiểm định thống kê, chúng ta có thể kết luận rằng bộ dữ liệu ngẫu nhiên với kích thước n = 42 không tuân theo phân phối chuẩn. Các phương pháp đồ thị cung cấp cái nhìn trực quan về sự phân bố dữ liệu, trong khi các kiểm định thống kê cung cấp các đánh giá khách quan dựa trên các số liệu thống kê cụ thể. Sự kết hợp này cho phép chúng ta có một đánh giá toàn diện và chính xác về tính chuẩn của dữ liệu.

Các kết quả này có ý nghĩa quan trọng trong việc áp dụng các phương pháp thống kê phù hợp trong phân tích dữ liệu. Việc xác định liệu dữ liệu có tuân theo phân phối chuẩn hay không sẽ giúp lựa chọn các phương pháp phân tích và mô hình hóa dữ liệu thích hợp, đảm bảo tính chính xác và độ tin cậy của các kết luận thống kê.

Kết luận

Trong đồ án này, em đã nghiên cứu và áp dụng các phương pháp kiểm định phân phối chuẩn nhằm đánh giá tính chuẩn của bộ dữ liệu ngẫu nhiên. Các phương pháp được sử dụng bao gồm phương pháp đồ thị và các phương pháp suy luận thống kê phổ biến.

Ta nhận thấy rằng, phương pháp đồ thị cho ta một cái nhìn trực quan nhất về tính chuẩn. Đồng thời, việc đánh giá tính chuẩn bằng phương pháp đồ thị là thiếu khách quan, không giống như khi sử dụng các kiểm định thống kê. Tuy nhiên, đánh giá tính chuẩn bằng các kiểm định thống kê lại nhạy cảm với cỡ mẫu. Trong trường hợp mẫu nhỏ, giả thuyết không có phân phối chuẩn thường không bị bác bỏ hoặc là giả thuyết phân phối chuẩn bị bác bỏ ngay cả khi vi phạm nhỏ. Do đó, các phương pháp đồ thị nên được sử dụng để phân tích vi phạm tính chuẩn dựa trên cỡ mẫu lớn. Tóm lại, kết hợp các phương pháp đồ thị và các phương pháp suy luận thống kê chắc chắn sẽ cải thiện đánh giá của chúng ta về tính chuẩn của dữ liệu. Khi n tăng lên, sức mạnh tổng thể tăng lên nhưng đồng thời ta thấy kiểm định Shapiro-Wilk (SW), Anderson-Darling (AD) là những kiểm định mạnh nhất trong số các kiểm định khác.

Để tiếp tục nghiên cứu, có thể xem xét áp dụng thêm các phương pháp kiểm định khác hoặc thử nghiệm với các bộ dữ liệu khác nhau để kiểm tra tính ổn định của các kết quả. Ngoài ra, việc áp dụng các kỹ thuật biến đổi dữ liệu hoặc sử dụng các mô hình phi chuẩn có thể giúp cải thiện kết quả phân tích thống kê đối với các bộ dữ liệu không tuân theo phân phối chuẩn. Bên cạnh đó, chúng ta cũng có thể mở rộng các phương pháp này đối với phân phối chuẩn đa biến.

Tài liệu tham khảo

[1] Pearson, K. 1905. "On the general theory of skew correlation and non-linear regression." Biometrika 4: 171-212.

[2] Geary, R. C. 1947. "Testing for normality." Biometrika 34: 209-242. http://webspace.ship.ec

[3] Gnanadesikan, R. 1977. Methods for Statistical Analysis of Multivariate Data. New York. Wiley.

[4] Huber, P. J. 1973. "Robust regression: Asymptotics, conjectures, and Monte Carlo." The Annals of Statistics 1(5): 799-821. DOI: 10.1214/aos/1176342503.

[5] Judge, G. G., Griffith, W. E., Hill, R. C., Lutkepohl, H., and Lee, T. 1985. Theory and Practice of Econometrics. 2nd. Ed. New York. Wiley.

[6] Pearson, E. S., and Please, N. W. 1975. "Relation between the shape of population distribution and the robustness of four simple statistical tests." Biometrika 62: 223-241.

[7] Koenker, R. W. 1982. "Robust methods in econometrics." Econometric Reviews 1: 213-290.

[8] Bera, A. K., and Jarque, C. M. 1982. "Model specification tests: A simultaneous approach." Journal of Econometrics 20: 59-82.

[9] Imon, A. H. M. R., and Das, K. 2015. "Analyzing length or size based data: A study on the lengths of peas plants." Malaysian Journal of Mathematical Sciences 9(1): 1-20. http://einspem.upm.edu.my/journal/fullpaper/vol9/1.

[10] Imon, A. H. M. R. 2015. "An Introduction to Regression, Time Series, and Forecasting."

[11] Kolmogorov, A. 1933. "Sulla determinazione empirica di una legge di distribuzione." G. Ist. Ital. Attuari 4, 83–91.

[12] Smirnov, N. 1948. "Table for estimating the goodness of fit of empirical distributions." Annals of Mathematical Statistics 19(2): 279–281. doi: 10.1214/aoms/1177730256.

[13] Shapiro, S. S., and Wilk, M. B. 1965. "An analysis of variance test for normality (complete samples)." Biometrika 52(3/4): 591-611. http://sci2s.ugr.es/keel/pdf/algorithm/articu

[14] Shapiro, S. S., and Francia, R. S. 1972. "An approximate analysis of variance test for normality." Journal of the American Statistical Association 67(337): 215-216.

DOI: 10.1080/01621459.1972.10481232.

[15] Weisberg, S., and Bingham, C. 1975. "An approximate analysis of variance test for non-normality suitable for machine calculation." Technometrics 17(1): 133-134.

[16] Royston, J. P. 1982. "An extension of Shapiro-Wilk's W test for non-normality to large samples." Applied Statistics 31: 115-124.

[17] D'Agostino, R. B. 1971. "An omnibus test of normality for moderate and large sample sizes." Biometrika 58(August): 341-348.

[18] Stephens, M. A. 1974. "EDF statistics for goodness of fit and some comparisons." Journal of the American Statistical Association 69(347): 730-737.

[19] Anderson, T. W., and Darling, D. A. 1952. "Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes." The Annals of Mathematical Statistics 23(2): 193-212. http://www.cithep.caltech.edu/ fcp/statistics/hypothesisTest/P oissonConsistency/AndersonDarling1952.pdf.

[20] Fisher, R. A. 1930. "The moments of the distribution for normal samples of measures of departure from normality." Proceedings of the Royal Society of London 130(December): 16-28.

[21] Pearson, E. S. 1930. "A further development of tests for normality." Biometrika 10.1093/biomet/22.1-2.239.

[22] Geary, R. C. 1935. "The ratio of mean deviation to the standard deviation as a test of normality." Biometrika 27: 310-332.

[23] David, H. A., Hartley, H. O., and Pearson, E. S. 1954. "The distribution of the ratio, in a single normal sample of range to standard deviation." Biometrika 41: 482-93.

[24] DAgostino R, and Pearson E. S. 1973. "Tests for departure from normality. Empirical results for the distributions of b_2 and $\sqrt{b_1}$." Biometrika. 60(3), 613-622.

[25] Bowman, K. O., and Shenton, B. R. 1975. "Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2 Biometrika 64: 243-50.

Phụ lục

n	$\alpha = 0.20$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.02$	$\alpha = 0.01$
1	0.900	0.950	0.975	0.990	0.995
2	0.684	0.776	0.842	0.899	0.929
3	0.565	0.636	0.708	0.785	0.829
4	0.494	0.565	0.624	0.703	0.751
5	0.447	0.509	0.563	0.627	0.669
6	0.410	0.468	0.519	0.577	0.617
7	0.381	0.438	0.483	0.538	0.575
8	0.358	0.410	0.454	0.507	0.542
9	0.339	0.387	0.430	0.479	0.513
10	0.323	0.369	0.409	0.457	0.489
11	0.308	0.352	0.391	0.439	0.468
12	0.296	0.338	0.375	0.423	0.450
13	0.285	0.325	0.361	0.404	0.432
14	0.275	0.314	0.349	0.393	0.418
15	0.266	0.304	0.338	0.377	0.404
16	0.258	0.295	0.327	0.366	0.392
17	0.250	0.286	0.318	0.355	0.381
18	0.244	0.279	0.309	0.346	0.371
19	0.237	0.271	0.301	0.338	0.361
20	0.232	0.265	0.294	0.328	0.352
21	0.226	0.259	0.287	0.321	0.345
22	0.221	0.253	0.281	0.314	0.337
23	0.216	0.247	0.275	0.307	0.331
24	0.212	0.242	0.269	0.301	0.323
25	0.208	0.238	0.264	0.295	0.317
26	0.204	0.233	0.259	0.290	0.311
27	0.200	0.229	0.254	0.284	0.305
28	0.197	0.225	0.250	0.279	0.300
29	0.193	0.221	0.246	0.275	0.295
30	0.190	0.218	0.242	0.270	0.290
31	0.187	0.214	0.238	0.266	0.285
32	0.184	0.211	0.234	0.262	0.281
33	0.182	0.208	0.231	0.258	0.277
34	0.179	0.205	0.227	0.255	0.273
35	0.177	0.202	0.224	0.251	0.269
36	0.174	0.199	0.221	0.247	0.265
37	0.172	0.197	0.218	0.244	0.262
38	0.170	0.194	0.215	0.241	0.258
39	0.168	0.192	0.212	0.238	0.255
40	0.165	0.189	0.210	0.235	0.252
> 40	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Bảng 2.5: Giá trị tới hạn $D_{n,\alpha}$ với các mức ý nghĩa khác nhau α