PHẦN THỐNG KÊ

CHƯƠNG 4 LÝ THUYẾT MẪU

4.1. CÁC PHƯƠNG PHÁP CHỌN MẪU

4.1.1. Các vấn đề thống kê

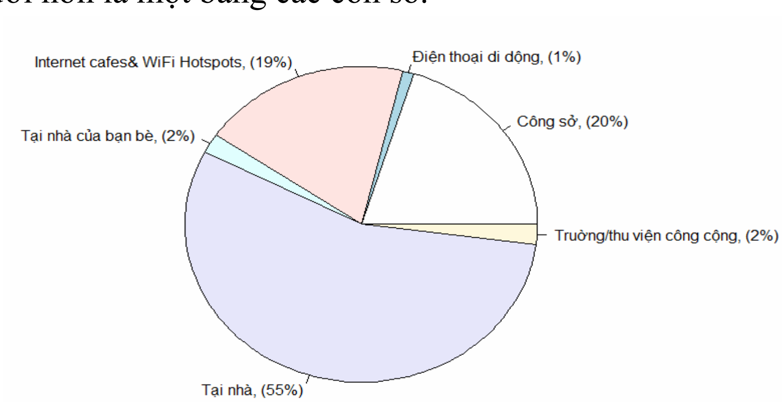
Thống kế toán học có thể coi là tổng thể các phương pháp toán học, dựa trên lý thuyết xác suất và các công cụ khác, nhằm đưa ra được những thông tin mới, kết luận mới, có giá trị, từ những bảng số liệu thô ban đầu, và nhằm giải quyết những vấn đề nào đó nảy sinh từ thực tế. Có thể kể tên một số mục đích chính của thống kê như sau:

- Mô tả số liệu.
- Ước lượng và dự đoán các đại lượng.
- Tìm ra các mối quan hệ giữa các đại lượng .
- Kiểm định các giả thuyết.

Thống kê học là một ngành lớn, với nhiều phương pháp khác nhau để dùng cho các tình huống khác nhau (có người ví các phương pháp thông kê như là các cách nấu ăn, rất đa dạng phong phú), và có nhiều điểm cần chú ý để khỏi dẫn đến các kết luận thống kê sai lệch (hoặc là bị mắc lừa bởi những người cố tình làm thống kê theo các phương pháp sai lệch). Trong chương này chúng ta sẽ chỉ bàn tới một số vấn đề và phương pháp thống kê toán học cơ bản nhất. Trước khi đi vào lý thuyết, ở phần này chúng ta sẽ điểm qua các mục đích chính trên của thống kê, qua một số ví dụ.

Ví dụ 4.1. (Biểu đồ thống kê).

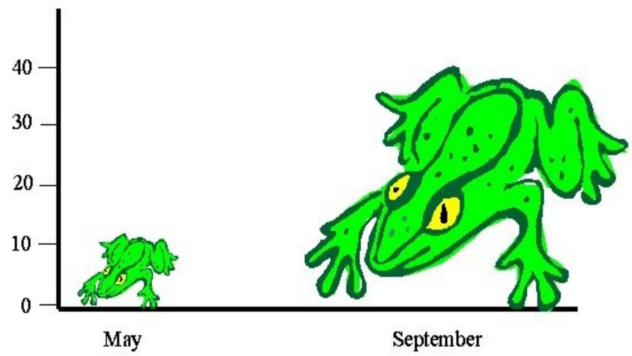
Trong thống kê mô tả, ngoài các bảng số liệu (cùng với một số đại lượng đặc trưng tiêu biểu nhất như trung vị, kỳ vọng, phương sai), các biểu đồ cũng hay được dùng, để giúp người đọc nắm bắt thông tin về số liệu một cách nhanh chóng. Một số loại biểu đồ hay gặp là: biểu đồ tần số, đồ thị phát tán, biểu đồ hình quạt (pie chart), v.v. Hình 4.1, là một ví dụ về biểu đồ hình quạt, phản ảnh tỷ lệ thời gian dùng internet ở Việt Nam vào năm 2009 (theo báo Lao Động). So với các bảng số liệu, các biểu đồ có thể có nhược điểm là cho thông tin không được chính xác bằng (độ sai số cao hơn), nhưng có ưu điểm là cho được cùng một lúc nhiều thông tin trên một hình ảnh, dễ tiếp thu đối với não người hơn là một bảng các con số.



Hình 4.1: Tỷ lệ thời gian dùng internet ở Việt Nam năm 2009

Các lệnh trong R

```
> noisudung<-c("Tại nhà, (55%)", "Công sở, (20%)","Internet cafes& WiFi Hotspots, (19%)", "Tại nhà của bạn bè, (2%)","Trường/thư viện công cộng, (2%)","Điện thoại di động, (1%)")
> tylekhachhang<-c(55,20,19,2,2,1)
> data<- rep(noisudung, tylekhachhang)
> data.freq<-table(data)
> data.freq
data
 Công sở, (20%)   Điện thoại di động, (1%)   Internet cafes& WiFi Hotspots, (19%)
       20                    1                                  19
Tại nhà của bạn bè, (2%)   Tại nhà, (55%)   Truờng/thu viện công cộng, (2%)
        2                        55                      2
> pie(data.freq)
```



Hình 4.2: Số ếch trong hồ vào tháng 5 và tháng 9

Tất nhiên, có những biểu đồ có thể phản ánh rất sai lệch về các đại lượng. Hình 4.2 là một ví dụ đơn giản về đề tài nói dối bằng thống kê. Đồ thị đó xuất phát từ số liệu thống kê số ếch trong 1 cái hồ: hơn 10 con vào tháng 5, và nhiều gấp 3 lần như vậy vào tháng 9. Nhưng nhìn vào đồ thị người ta sẽ có cảm giác là số ếch vào tháng 9 gấp $3 \times 3 = 9$ lần tháng 5.

Ví dụ 4.2. (Phát xít Đức sản xuất bao nhiêu máy bay và xe cơ giới?)

Trong chiến tranh, việc ước lượng được đúng sức mạnh của quân địch là một việc nhiều khi có tính chất sống còn. Trong chiến tranh thế giới lần thứ II, các cơ quan tình báo quân đồng minh Anh-Mỹ đã cung cấp nhiều thông tin rất sai lệch về lực lượng quân Đức. Thế nhưng, bằng phương pháp thống kê (thu nhặt các mã số trên các xác máy bay, lốp xe, v.v. của quân Đức bị bắn cháy, bỏ rơi, rồi từ đó giải mã và dùng các hàm ước lượng để ước lượng), nhà thống kê học Richard Ruggles cùng với các cộng sự của mình, lúc đó làm tại Cục tình báo kinh tế của Anh, đã ước lượng được rất chính xác số máy bay và xe cơ giới mà Đức sản xuất được hàng tháng.

Công suất hàng tháng của Phát xít Đức	Máy bay	Xe cơ giới
Ước lượng của Ruggles	28500	147000
Số liệu thực theo tài liệu của Đức	26400	159000

Trong khi đó, ước lượng của các tình báo viên Anh-Mỹ là công suất của Đức khoảng 1 triệu xe cơ giới một tháng!

Ví dụ 4.3. (Thần dược chống béo phì?)

Nguồn: http://ir.vivus.com/releasedetail.cfm?ReleaseID=407933

Tỷ lệ số người bị béo phì (obesity) tăng rất nhanh trên thế giới (kể cả ở Việt Nam, châu Âu, và Mỹ) trong những thập kỷ cuối thế kỷ 20 - đầu thế kỷ 21, và trở thành một vấn đề xã hội lớn, vì béo phì hay dẫn đến nhiều căn bệnh khác (tim mạch, tiểu đường, đột quị, vô sinh, v.v.), và có thể làm giảm đáng kể tuổi thọ của người. Chống béo phì là một vấn đề nóng hổi, nhưng cho đến năm 2009 chưa có thuốc nào thật hiệu quả được bán trên thị trường. Điều này có thể thay đổi trong những năm sau đó, vì trong năm 2009 có 3 hãng dược phẩm công bố các kết quả thử nghiệm lâm sàng giai đoạn III (phase III clinical trial) cho các loại thuốc chống béo phì mới có nhiều triển vọng. Trong đó đáng chú ý nhất có lẽ là thuốc Qnexa của hãng Vivus. Công bố kết quả về Qnexa của Vivus vào ngày 09/09/2009[2] có một bảng thống kê sau (trong số nhiều bảng thống kê):

	ITT-LOCF			Completers		
EQUIP (OB-302) 56 Weeks	Placebo (n=498)	Qnexa Low Dose (n=234)	Qnexa Full Dose (n=498)	Placebo (n=241)	Qnexa Low Dose (n=138)	Qnexa Full Dose (n=301)
Mean Weight Loss (%)	1.6%	5.1%*	11.0%*	2.5%	7.0%*	14.7%*
Greater than or equal to 5% weight loss rate	17%	45%*	67%*	26%	59%*	84%*

ITT-LOCF: Intent-to-treat with last observation carried forward
*p<0.0001 vs. placebo

Theo bảng trên, tổng số người tham gia thử nghiệm lâm sàng (trong thử nghiệm đó) là 498+234+498 = 1300 người. Đợt thử nghiệm kéo dài 56 tuần, nhưng có những người bỏ dở giữa chừng: trong số 498 người được nhận placebo (trông giống như viên thuốc thật, nhưng không có thuốc trong đó) thì chỉ có 241 người theo đến cùng cuộc thử nghiệm, còn trong số 498 người được nhận liều đầy đủ của thuốc, có 301 người (61%) theo đến cùng. Trong số những người được nhận đủ liều và theo đến cùng, thì có 84% số người giảm được ít nhất 5% trọng lượng, và trung bình mỗi người giảm được 14,7% trọng lượng.

Trong bảng trên có viết $p < 0{,}0001$ vs. placebo. Điều đó có nghĩa là, với độ tin cậy bằng $1 - p > 99{,}99\%$ (hay nói cách khác, với khả năng kết luận sai lầm nhỏ hơn

0,01%), các con số thống kê cho thấy kết quả đạt được (ở đây là giảm cân) tốt hơn khi có thuốc so với khi không có thuốc.

Thông thường, khi $p < 0,01$ thì người ta chấp nhận giả thuyết là thuốc có hiệu ứng thực sự, còn nếu $p \geq 0,05$ thì hiệu ứng đó không rõ ràng, có thể là do ngẫu nhiên.

Các hãng dược phẩm trên thế giới, trước khi được quyền bán một loại thuốc mới nào đó, thông thường đều phải qua thử nghiệm lâm sàng diện rộng (trên ít nhất mấy trăm bệnh nhân), và các kết quả thống kê phải chứng tỏ rõ ràng công dụng và sự an toàn của thuốc, tức là phải qua được kiểm định thống kê cho giả thuyết "thuốc có công dụng và an toàn", với độ tin cậy cao.

Ví dụ 4.4. (London nguy hiểm hay an toàn?)

Ngày 10/07/2008, có 4 vụ giết người bằng dao ở 4 nơi khác nhau ở London. Sự kiện này làm náo loạn dư luận đến mức thủ tướng Anh là Gordon Brown phải tuyên bố hứa sẽ tìm cách làm giảm các vụ đâm dao. London có trở nên nguy hiểm cho tính mạng hơn những năm trước không? Để trả lời câu hỏi đó, chúng ta có thể dựa trên một số số liệu thống kê sau:

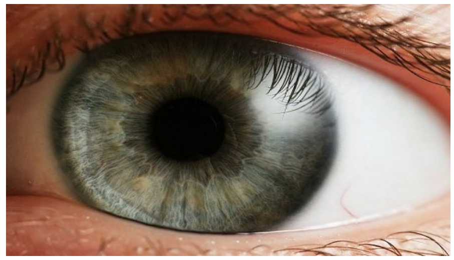
- Trong 5 năm trước đó, mỗi năm ở London có khoảng 170 người bị giết, và con số này khá ổn định hàng năm.
- Khoảng 41% các vụ giết người là dùng dao, 17% là dùng súng, 9% là đánh đập (không vũ khí), 5% là đánh bằng vật không phải dao, 3% là bóp cổ, 3% là dùng thuốc độc, v.v., và 17% là không xác định được phương pháp.

Trong thời gian 3 năm 04/2004 – 03/2007, có 713 ngày không có vụ án mạng nào, 299 ngày có 1 vụ, 66 ngày có 2 vụ, 16 ngày có 3 vụ, 1 ngày có 4 vụ, và không có ngày nào có từ 5 vụ trở lên.

Từ các số liệu thống kê, người ta tính được một số ước lượng sau về số vụ án mạng ở London:

- Số vụ án mạng xảy ra trong ngày tuân theo phân bố Poisson với kỳ vọng là 0,44 (tức là trung bình mỗi ngày có 0,44 vụ).
- Kỳ vọng là mỗi năm có khoảng 3-4 ngày với 3 vụ án mạng, cứ khoảng gần 3 năm thì có 1 ngày với 4 vụ án mạng, và khoảng 30 năm thì mới có một ngày với 5 vụ án mạng.

Việc xảy ra 1 ngày vào năm 2008 với 4 vụ án mạng không nằm ngoài các con số ước lượng trên. Đâm bằng dao là phương pháp gây án mạng phổ biến nhất (41% tổng số các vụ). Khi có 4 vụ án mạng, thì xác suất để cả 4 vụ đều do đâm dao là $(0,41)^4 =$ 2,8%, một con số khá nhỏ, nhưng cũng không nhỏ đến mức "không thể xảy ra". Khi có 4 vụ án mạng xảy ra cùng ngày, thì có rất nhiều tổ hợp các khả năng xảy ra về phương pháp gây án mạng trong 4 vụ đó (ví dụ 2 vụ dùng dao, 1 vụ dùng súng, 1 vụ thắt cổ), và tất cả các tổ hợp đó đều có xác suất nhỏ, tổ hợp với xác suất lớn nhất cũng không vượt quá 6%. Từ đó, có thể kết luận là, việc hôm 10/07/2008 xảy ra 4 án mạng ở London, và cả 4 đều bằng đâm dao, hoàn toàn nằm trong các ước lượng về án mạng xảy ra ở London, và không hề chứng tỏ xu thế gì mới. Tổng kết năm 2008, ở London có 152 án mạng xảy ra năm đó. Phương tiện truyền thông được dịp vui mừng vì "đã lâu rồi chưa năm nào London được an toàn như vậy". Nhưng con số đó có chứng tỏ xu thế gì không, hay chẳng qua cũng chỉ là một sự ngẫu nhiên không nằm ngoài qui luật chung?



Hình 4.3: Các đường vân trong màng mắt

Ví dụ 4.5. (Con mắt trở thành chìa khóa).

Đầu thế kỷ 21, đã có những khách sạn mà khách không cần chìa khóa phòng, chỉ cần nhìn vào camera ở cửa phòng, là phòng tự động mở cửa. Sự tiện lợi này dựa trên công nghệ nhận biết danh tính của người qua màng mắt (iris). Một điều thú vị là, kể cả khi hai người sinh đôi và trông giống hệt nhau, thì các đường nét trong màng mắt của họ vẫn rất khác nhau, do quá trình phát triển các đường nét trong màng mắt ở thai nhi phụ thuộc vào nhiều yếu tố ngẫu nhiên (không do di truyền). Từ những năm 1930, các bác sĩ mắt đã nói rằng có thể dùng màng mắt để nhận biết danh tính người. John Daugman là một trong những người làm ra công nghệ nhận biết danh tính bằng màng mắt, từ cuối thế kỷ 20. Thuật toán của ông ta tách ra được từ ảnh màng mắt 1 mã với 266 đơn vị thông tin có thể coi là ngẫu nhiên và độc lập với nhau (mỗi đơn đơn vị ở đây là một biến ngẫu nhiên nhận 2 giá trị 0 và 1, với xác suất 50% − 50%, và các biến này gần như độc lập với nhau). Để tìm ra 266 đơn vị thông tin độc lập đó (xuất phát từ 2048 đơn vị thông tin không độc lập với nhau) và kiểm định sự độc lập của chúng, Daugman đã làm thống kê so sánh hơn 222 nghìn lần cặp ảnh màng mắt khác chủ (2 mắt trong 1 cặp là của hai người khác nhau), và hơn 500 cặp ảnh màng mắt cùng chủ. Một trong các kết quả là, tỷ lệ đơn vị thông tin chệch nhau giữa mã của 2 mắt khác chủ tuân theo phân bố normal với kỳ vọng là 45.6% (tức là trung bình hai mắt khác chủ thì có 45.6% đơn vị thông tin chệch nhau) với độ lệch chuẩn là 0.18%, và không có cặp mắt khác chủ nào (trong các thử nghiệm) có dưới 37% đơn vị thông tin lệch nhau. Mặt khác, hai ảnh màng mắt khác nhau của cùng một chủ thì trung bình chỉ có 9% các đơn vị thông tin bị lệch nhau trong số 266 đơn vị, và không có cặp ảnh mắt cùng chủ nào bị lệch nhau quá 31% đơn vị thông tin. Từ đó dẫn đến thuật toán phân biệt: coi rằng nếu hai mã bị lệch nhau không quá 34% số đơn vị thông tin, thì vẫn là của cùng một người, còn nếu trên 34% thì coi là của hai người khác nhau.

Một điều cần chú ý là, thống kê hay bị các tổ chức hay cá nhân lạm dụng để bóp méo sự thật theo hướng có lợi cho mình, hoặc có khi tự dối mình, nếu như làm không đúng cách. Có rất nhiều cách nói dối khác nhau bằng thống kê, chẳng hạn như: bịa đặt các con số không có thật, lựa chọn các con số có lợi, giấu đi các con số bất lợi, lệch (bias) trong việc chọn mẫu thí nghiệm, v.v. Ví dụ về nói dối trắng trợn: Bộ quốc phòng Mỹ có tuyên bố rằng, trong cuộc chiến với Irak năm 1991, các tên lửa Patriot của Mỹ đã bắn rụng 41 tên lửa Scud của Irak, nhưng khi Quốc hội Mỹ điều tra lại thấy chỉ có 4 tên lửa Scud bị bắn rụng. Ví dụ về lệch (bias) làm hỏng kết quả thống kê: Báo Literacy Digest thăm dò ý kiến cử tri về bầu cử tổng thống ở Mỹ năm 1936, qua điện thoại và qua các độc giả đặt báo. Kết quả thăm dò trên phạm vi rất rộng cho dự đoán là Landon sẽ được 370 phiếu (đại cử tri) còn Roosevelt sẽ chỉ được 161 phiếu. Thế nhưng lúc bầu thật thì Roosevelt thắng. Hoá ra, đối tượng mà Literacy Digest thăm dò năm đó, những người có tiền đặt điện thoại hay đặt báo, là những người thuộc tầng lớp khá giả, có lệch (bias) theo phía Landon (Đảng Cộng hòa), không đặc trưng cho toàn dân chúng Mỹ.

Nói chung, để thống kê toán học cho ra được các kết quả đáng tin cậy, ngoài các công thức toán học đúng đắn, còn cần đảm bảo sự trung thực của các số liệu, có mẫu thực nghiệm (lượng số liệu) đủ lớn, và loại đi được ảnh hưởng của các lệch (bias) để đảm bảo tính ngẫu nhiên của số liệu. Nhiều khi việc loại đi các kết quả có lệch (bias) cao từ mẫu thực nghiệm là công việc hiệu quả, cho ra kết luận thống kê chính xác và đỡ tốn kém hơn là tăng cỡ của mẫu thực nghiệm lên thêm nhiều. Ở chương này, chúng ta sẽ chỉ bàn đến một số phương pháp thống kê cơ bản, dựa trên giả sử là số liệu mà chúng ta nhận được là đúng thực và không bị lệch.

4.1.2. Tổng thể và mẫu

Có hai loại dữ liệu thường được đề cập đến trong các phương pháp thống kê. Đó là các dữ liệu của mẫu và dữ liệu của tổng thể.

4.1.2.1. Tổng thể

Một tập hợp dữ liệu đầy đủ liên quan đến lĩnh vực nào đó mà chúng ta quan tâm tìm hiểu một hoặc một số dấu hiệu thì được gọi là tổng thể.

Ví dụ 4.6

Một ông hiệu trưởng cần tìm hiểu phân bố tuổi của tất cả các học sinh của trường mình. Như vậy, tuổi của tất cả các học sinh trong trường của ông tạo thành một tổng thể (tổng thể tuổi). Nếu có 3240 học sinh trong trường ông thì đây là một tổng thể 3240 tuổi. Ta có thể gọi đây là một tổng thể 3240 học sinh, nhưng mối quan tâm của hiệu trưởng trong trường hợp này là dấu hiệu tuổi của học sinh.

* Tổng thể vô hạn: là tập hợp số các đơn vị (phần tử) không giới hạn, trong thực tế, tổng thể hữu hạn lấy mẫu có hoàn lại xem như tổng thể vô hạn.

Vì phần nhiều các tổng thể đều rất lớn (có thể vô hạn) cho nên đa số các nghiên cứu không thể đo lường, xác định được hết từng phần tử trong tổng thể lớn (vô hạn). Có những trường hợp điều tra sẽ làm phá hủy các phần tử được điều tra, chẳng hạn kiểm tra chất lượng đồ hộp. Vì vậy chúng ta sẽ lựa chọn mẫu lấy từ tổng thể.

4.1.2.2. Mẫu

Là một bộ phận hay một tập hợp con (subset) của tổng thể và có cỡ nhỏ hơn tổng thể ấy. Chúng ta lựa chọn một mẫu từ tổng thể để thu thập dữ kiện, từ đó suy diễn, khái quát về các đặc trưng của dấu hiệu tổng thể.

Nếu tổng thể là 10.000 người thì mẫu của nó có thể rất nhỏ (là 1) hoặc rất lớn (là 9.999). Thông thường cỡ của một mẫu nằm giữa hai con số này. Nó có thể là 30 hay 100 người, đó là cỡ mẫu thông thường trong các nghiên cứu thực nghiệm. Nhưng dù nhỏ hay lớn điều quan trọng là mẫu phải có tính cách đại diện cho tổng thể mà người ta muốn tìm hiểu. Vậy thế nào là mẫu đại diện?

4.1.3. Phân loại dữ liệu

4.1.3.1. Dữ liệu định lượng: là dữ liệu mà giá trị của nó có thể xác định bằng việc đo lường được gồm:

a. Dữ liệu ngẫu nhiên rời rạc: là dữ liệu nhận giá trị hữu hạn hoặc vô hạn đếm được.

Ví dụ 4.7: số học sinh vắng mặt, số sản phẩm tốt,…

b. Dữ liệu ngẫu nhiên liên tục: là dữ liệu nhận giá trị trên một khoảng số thực (a,b).

Ví dụ 4.8: Chiều cao của học sinh, tuổi thọ của bóng đèn, năng suất của một loại cây trồng,…

4.1.3.2. Dữ liệu định tính: là dữ liệu mà giá trị của nó được gán để phân loại hay phân biệt, gồm:

a. Dữ liệu danh mục:chỉ dùng để gán tên, không xếp thứ tự, không quan trọng về mức độ

Ví dụ 4.9: Phân loại nhóm máu: 1-A, 2-B, 3-AB, 4-O

Trường hợp riêng là biến nhị phân (chỉ có 2 giá trị):1-có, 2-không ; 1-nam, 2-nữ

Giữa các số 1 và 2 không có quan hệ hơn kém, mà chỉ dùng để đếm số lần xuất hiện của các dấu hiệu.

b. Dữ liệu thứ hạng: quan trọng về thứ tự, không quan trọng về mức độ.

Ví dụ 4.10:

Mức độ suy dinh dưỡng của trẻ:1-nhẹ, 2-vừa, 3-nặng

Trình độ học vấn: 1-không biết chữ, 2-cấp I, 3-cấp II, 4-cấp III, 5-Đại học – cao đẳng.

4.1.3.3 Các loại thang đo

a. Thang đo định danh (Nominal Scale) là loại thang đo dùng cho các dữ liệu thuộc tính. Người ta sử dụng các mã số để phân loại các đối tượng và chúng không mang ý nghĩa nào khác. Ví dụ giới tính, nam kí hiệu là số 1, nữ kí hiệu là số 0.

Thước đo đo tập trung duy nhất là Yếu vị (ModX); độ phân tán thống kê có thể đo bằng các tỷ lệ, không tính được độ lệch chuẩn. Chúng ta hay gặp thang đo định danh trong các câu hỏi về thông tin cá nhân của từng người hay của thông tin về danh nghiệp.

Ví dụ 4.11

Trong một chủ đề nghiên cứu, người ta đưa ra hai câu hỏi sau:

Câu 1. Tình trạng hôn nhân của Anh/Chị là:

A. Độc thân B. Đã lập gia đình C. Khác

Câu 2. Công ty của Anh/Chị đang hoạt động chính trong lĩnh vực nào:

A. Sản xuất B. Xây dựng C. Dịch vụ D. Thương mại

Mỗi người sẽ chọn một trong các mã số A, B, C cho câu 1 hoặc A, B, C, D cho câu 2. Các mã số này là thang đo định danh.

b. Thang đo thứ bậc (Ordinal scale) là thang đo mà giữa các biểu hiện của dữ liệu có quan hệ thứ bậc hơn kém. Sự chênh lệch giữa các biểu hiện không nhất thiết phải bằng nhau.

Ví dụ 4.12

Với các câu hỏi được cho như sau:

Câu 1. Anh/Chị vui lòng cho biết kết quả học tập Anh/Chị đạt được khi tốt nghiệp Đại học:

□ Trung bình □ Trung bình-Khá □ Khá □ Giỏi

Câu 2. Anh/Chị hãy xếp hạng các chủ đề sau trên báo Phụ nữ theo mức độ quan tâm (chủ đề nào quan tâm nhất thì ghi số 1, quan tâm thứ hai thì ghi số 2 và quan tâm thứ ba thì ghi số 3):

□ Hôn nhân gia đình □ Thời trang □ Nuôi dạy con cái

Câu 3. Thu nhập của Anh/Chị hằng tháng là:

□ Dưới 6 triệu đồng □ Từ 6 đến 10 triệu đồng

□ Từ 10 đến 14 triệu đồng □ Từ 14 triệu đồng

Thì việc chọn lựa các đáp án sẽ cho chúng ta dữ liệu thu thập được có sự hơn kém nhưng không thể hiện rõ rệt sự hơn kém này, vì thế các câu hỏi trên có dữ liệu thu được thuộc thang đo thứ bậc.

c. Thang đo khoảng (Interval scale) là thang đo dùng cho dữ liệu định lượng và các dữ liệu định tính có phản ánh sự hơn kém được thể hiện bằng những con số nhưng không mang một ý nghĩa tuyệt đối.

Ví dụ 4.13

Rõ nhất cho thang đo này là thang đo nhiệt độ, chẳng hạn $32^0C>30^0C$ và $80^0C>78^0C$, sự chênh lệch giữa 32^0C và 30^0C cũng giống như sự chênh lệch giữa 80^0C và 78^0C cách đều 2^0C.

Như vậy, thang đo khoảng cho phép chúng ta đo lường một cách tương đối sự khác nhau giữa hai giá trị bất kỳ, còn trong thang đo thứ bậc thì không thể, vì nó chỉ cho biết giá trị này lớn hơn giá trị khác mà thôi.

Chúng ta cũng gặp loại thang đo này trong câu hỏi phỏng vấn dạng:

Đề nghị quý thầy/cô cho biết ý kiến của mình về tầm quan trọng của các mục tiêu đào tạo sinh viên đại học sau đây bằng cách khoanh tròn các con số tương ứng trên thang đo đánh giá chỉ mức độ từ 1 đến 5: (1:không quan trọng→5: rất quan trọng)

Năng lực giải quyết vấn đề	1	2	3	4	5
Tư duy logic	1	2	3	4	5
Khả năng làm việc độc lập	1	2	3	4	5
Năng lực nghiên cứu khoa học	1	2	3	4	5

d. Thang đo tỷ lệ (Ratio scale) là loại thang đo dùng cho dữ liệu số lượng. Thang đo tỷ lệ có đầy đủ các đặc tính của thang đo khoảng, tức là có thể áp dụng các phép tính cộng trừ. Ngoài ra, thang đo này có một giá trị số 0 thật, cho phép lấy tỷ lệ so sánh giữa hai giá trị thu thập cho nên gọi là thang đo tỷ lệ. Đây là thang đo có ý nghĩa nhất trong các loại thang đo.

Sự khác nhau giữa thang đo khoảng và thang đo tỷ lệ thường bị lẫn lộn vì hai điểm sau:

1. Điểm 0 trong thang đo tỷ lệ là trị số thật.
2. Trong thang đo khoảng, sự so sánh về mặt tỷ lệ giữa các giá trị không có ý nghĩa.

Ví dụ 4.14

Bạn có 5 triệu đồng và anh của bạn có 10 triệu đồng. Như vậy số tiền của anh bạn gấp đôi số tiền của bạn. Nếu chúng ta đổi sang dollar, pound, yên,… thì số tiền của anh bạn vẫn gấp đôi số tiền của bạn. Nếu số tiền của bạn bị mất hay bị đánh cắp thì bạn có 0 đồng. Số 0 ở đây là một trị số thật vì thật sự bạn không có đồng nào cả. Như vậy tiền tệ có trị số 0 thật và là thang đo tỷ lệ. Các loại thang đo tỷ lệ khác là: mét, kg, tấn, tạ,… Trái lại, nhiệt độ là thang đo khoảng, ví dụ nhiệt độ hôm nay là 12^0C ($53,6^0$F) và hôm qua là 6^0C ($42,8^0$F), chúng ta không thể nói rằng hôm nay ấm áp gấp hai lần hôm qua (vì đổi từ ^{0}C sang ^{0}F thì không còn gấp đôi nữa). Hơn nữa, nếu nhiệt độ là 0^0C, không có nghĩa là không có nhiệt độ, 0^0C dĩ nhiên lạnh hơn 6^0C. Như vậy nhiệt độ không có giá trị 0 thật.

Tóm lại, hai thang đo đầu tiên sử dụng cho dữ liệu định tính cho nên còn có tên gọi là thang đo định tính. Hai thang đo còn lại sử dụng dữ liệu định lượng nên còn có tên gọi là thang đo định lượng. Trong thực tế, việc sử dụng loại thang đo nào phụ thuộc rất nhiều vào chủ đề nghiên cứu cũng như mục đích của chủ đề nghiên cứu, theo đó các nhà thống kê sẽ chọn loại thang đo nào phù hợp nhất để giải quyết vấn đề nghiên cứu. Dữ liệu thuộc loại thang đo bậc cao có thể chuyển thành các thang đo bậc thấp hơn nhưng không có chiều ngược lại, cụ thể như sau:

Thang đo tỷ lệ → thang đo khoảng→thang đo thứ bậc→thang đo định danh.

4.1.4. Các phương pháp chọn mẫu

Từ tổng thể lấy ra n phần tử để nghiên cứu suy ra các đặc trưng của dấu hiệu tổng thể ta gọi là một mẫu có kích thước n.

Để có thể làm công việc này, ta cần có những mẫu càng có tính cách đại diện cho tổng thể ta muốn tìm hiểu càng tốt. Có nhiều yếu tố khiến ta chọn phải những mẫu không đại diện cho tổng thể. Chẳng hạn, ta muốn tìm hiểu về khả năng học tập của học sinh lớp 10 trong thành phố Thủ Dầu Một, nhưng mẫu của ta chỉ bao gồm những học sinh lớp 10 tại các trường điểm hay các trường nổi tiếng, hoặc ngược lại, những trường đầu vào thấp nhất trong thành phố, như vậy mẫu của ta sẽ không đại diện cho tất cả các học sinh lớp 10 trong thành phố Thủ Dầu Một. Mẫu không đại diện như vậy bị xem là mẫu lệch (biased). Ngoài ra, mẫu của ta có thể không đại diện cho tổng thể vì sai số mẫu là do may rủi.

Một trong các phương pháp để làm tăng xác suất lựa chọn một mẫu đại diện là lựa chọn một mẫu ngẫu nhiên từ tổng thể mà ta quan tâm.
Một mẫu ngẫu nhiên là mẫu trong đó mỗi phần tử (đơn vị) trong tổng thể đều có cơ hội đồng đều được lựa chọn làm phần tử trong mẫu. Ý tưởng mẫu ngẫu nhiên trong xác suất liên quan đến tính khách quan trong thu thập dữ kiện. Do đó trong nghiên cứu, cần xác định rõ các nguyên tắc, định nghĩa về đơn vị đo đếm, về cách bố trí trước khi thu thập. Khi thực hiện phải tôn trọng các nguyên tắc đã xác định. Chúng ta xem qua các phương pháp lấy mẫu sẽ rõ hơn.

4.1.4.1. Các yêu cầu của chọn mẫu:

- Mang tính ngẫu nhiên: Đảm bảo tính khách quan, không phụ thuộc vào ý muốn chủ quan.
- Đảm bảo tính đại diện: Mẫu được chọn phải mang nhiều đặc tính với tổng thể.
- Mang tính đồng nhất: Mẫu được chọn phải cùng chủng loại, hoặc có đặc tính gần chủng loại.

4.1.4.2. Các phương pháp chọn mẫu:

*** Có nhiều cách lấy mẫu như sau:**

a. Lấy mẫu đơn giản ngẫu nhiên

- Mọi đối tượng có khả năng được chọn như nhau.
- Việc chọn có thể hoàn lại hoặc không hoàn lại.
- Để lấy mẫu n phần tử, ta đánh số từ 1 đến N phần tử tổng thể, sau đó dùng bảng số ngẫu nhiên hoặc dùng cách bốc thăm lấy đủ n phần tử.

Ví dụ 4.15

Khảo sát môi trường tại một khu vực, chúng ta chia ô cho khu vực trên bản đồ, đánh số, làm thăm và bốc thăm trùng số nào thì ra thực địa lấy mẫu đó.

Ví dụ 4.16

Khảo sát thực trạng hiểu biết về luật giao thông đường bộ trong danh sách 1000 học sinh phổ thông trung học.
Học sinh được sắp xếp từ 1 đến 1000.
Cỡ mẫu là 100 học sinh.
Chọn ngẫu nhiên ra 100 học sinh từ học sinh thứ 1 đến thứ 1000.

• Ưu điểm 4.1

- Đơn giản
- Các mẫu chọn trước không tác động đến sự lựa chọn các mẫu khác.
- Các ước lượng trung bình và phương sai không bị chệch.

• Nhược điểm 4.1

- Cần danh sách đầy đủ tất cả các đơn vị của tổng thể (phải nắm hết tất cả phần tử tổng thể).
- Không phải luôn luôn có tính đại diện tốt nhất. (không có tính đại diện cao).

- Các phần tử có thể bị phân tán và khó tiếp cận. (khó khăn trong việc đi đến địa điểm đã chọn).

b. Lấy mẫu hệ thống (Systematic)

- Mẫu được chọn một cách hệ thống theo bước nhảy k=N/n.

Ví dụ 4.17

N= 1200 , n= 60, bước nhảy k= N/n=1200/60=20

Lấy danh sách của 1200 đơn vị.

Lựa chọn ngẫu nhiên bất kỳ một số trong 20 số đầu.(ví dụ là số 5)

Cách 20 người nữa lại chọn người tiếp theo.

- người thứ 1: đứng thứ 5
- người thứ 2: đứng thứ 25
- người thứ 3: đứng thứ 45

…

- người thứ 60: đứng thứ 1185

• Ưu điểm 4.2

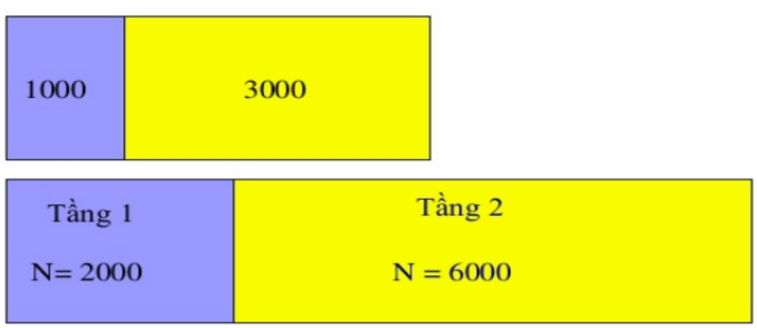
- Đơn giản
- Không có chiều hướng biến thiên theo không gian (dốc) hay thời gian (mưa nắng…)

• Nhược điểm 4.2

- Cần danh sách đầy đủ tất cả các đơn vị của tổng thể.
- Theo chu trình nên không phân tầng tự nhiên.
- Không có liên hệ giữa các mẫu
- Không phải luôn luôn có tính đại diện tốt nhất.

c. Lấy mẫu phân tầng (phân tổ)

- Tổng thể được chia thành các tầng theo các đặc trưng thông dụng.
- Mẫu ngẫu nhiên đơn giản được chọn từ các tầng vừa phân chia.
- Áp dụng cho tổng thể gồm các nhóm khác biệt hay tầng (strata), khác nhau về các đặc tính nghiên cứu cần quan tâm. Chẳng hạn các đoạn sông: thượng lưu, trung lưu, hạ lưu.
- Chiến lược sử dụng mẫu phân tầng là chọn các đơn vị trong mỗi tầng với tỷ lệ như nhau (lược đồ epsem) nghĩa là sử dụng cùng một phân số lấy mẫu (sampling fraction) cho mỗi tầng như hình minh hoạ là 1/2=1000/2000=3000/6000.



Hình 4.4: Lấy mẫu cho các tầng có cùng phân số

Ví dụ 4.18

Chọn mẫu các tỉnh để điều tra trong nghiên cứu kỳ gốc của dự án TAMP- GDT (Tax Administration Modernization Project - General Department of Taxation) với 2 biến vùng (Bắc/trung/nam) và mức đóng góp thuế vào ngân sách nhà nước (cao/trung bình/thấp) ---> bao nhiêu phân tầng?

Phân thành 9 tầng như sau:

Tầng	Vùng	Mức đóng thuế
1	Bắc	Cao
2	Bắc	Trung bình
3	Bắc	Thấp
4	Trung	Cao
5	Trung	Trung bình
6	Trung	Thấp
7	Nam	Cao
8	Nam	Trung Bình
9	Nam	Thấp

• **Ưu điểm 4.3**

- Có thể có được thông tin về toàn bộ tổng thể và thông tin về mỗi tầng. (tính đại diện cao).
- Hai hoặc nhiều mẫu được kết nối với nhau.
- Độ chính xác được tăng lên nếu như sự biến đổi trong mỗi tầng là ít hơn so với giữa các tầng với nhau.

• **Nhược điểm 4.3**

- Có thể sẽ khó xác định các tầng.
- Sẽ giảm độ chính xác nếu các đơn vị trong mỗi tầng nhỏ.

Ví dụ 4.19: Phân các điểm số từ 0 đến 10 thành 5 tầng: 0 – 2; 3 – 4; 5 – 6; 7 – 8; 9 – 10.

Các tầng mà có giới hạn trên hoặc dưới không xác định gọi là tầng mở.

Ví dụ 4.20: Trên 30 tuổi; dưới 50kg, …

d. Chọn mẫu cụm (cluster sampling) hay mẫu chùm:

- Tổng thể được chia thành các cụm.
- Chọn ngẫu nhiên ra một số cụm.
- Từ các cụm này chọn tất cả các đơn vị (phần tử) không bỏ sót các đơn vị nào để lấy mẫu.

Ví dụ 4.21

Giả sử cần nghiên cứu ô nhiễm chì trong sơn tại các trường học trong tỉnh Bình Dương có 7 huyện/thị. Ta có thể lấy ngẫu nhiên các trường học trong tỉnh nhưng khó khăn trong việc đi lại và chi phí khảo sát sẽ cao. Thay vào đó ta xem huyện/thị là cụm, lấy ngẫu nhiên ra 2 huyện/thị. Từ đó chúng ta chọn tất cả các trường trong 2 huyện/thị này. Như vậy đỡ chi phí di chuyển từ huyện này sang huyện khác.

• Ưu điểm 4.4

- Đơn giản vì nó không yêu cầu danh sách đầy đủ các đơn vị trong tổng thể.
- Ít phải đi lại, ít tốn kém.
- Ưu tiên dùng khi tổng thể quá lớn, ví dụ như cả một quốc gia.

• Nhược điểm 4.4

- Vấn đề tiềm ẩn ở đây là các đơn vị trong cụm thường giống nhau hơn là so với các đơn vị ở cụm khác (tính đồng nhất không cao giữa các cụm dẫn đến mẫu không đồng nhất).

Chú ý:

Việc lấy mẫu tiến hành chủ yếu theo hai phương thức:

- Lấy mẫu có hoàn lại
- Lấy mẫu không hoàn lại.

Theo định lý giới hạn của xác suất, người ta chứng minh được rằng: khi số phần tử của tổng thể đủ lớn thì coi hai cách lấy mẫu theo phương thức trên là như nhau.

4.1.4.3. Ưu và nhược điểm của điều tra chọn mẫu:

• Ưu điểm 4.5

- Tiết kiệm thời gian, chi phí, nhân lực và phương tiện.
- Áp dụng cho trường hợp không xác định được hết toàn bộ phần tử của tổng thể.

• Nhược điểm 4.5

- Số liệu thống kê có sai lệch so với thực tế.
- Đôi khi bị nhiễu thông tin.

4.1.5. Sai số chọn mẫu

Khái niệm 4.1

Sai số chọn mẫu là sự chênh lệch về trị số giữa các chỉ tiêu tính ra được trong điều tra chọn mẫu và các chỉ tiêu tương ứng của tổng thể.

Ví dụ 4.22

Độ chênh lệch về chiều cao trung bình tính được đối với mẫu gồm 500 sinh viên so với chiều cao trung bình của sinh viên toàn trường.

4.1.5.1 Nguyên nhân sai số:

- Sai số do tính chất đại biểu, tức là:
  + Mẫu được chọn có kích thước quá nhỏ so với tổng thể nhưng kết quả lại suy rộng thành
  kết quả của tổng thể.
  + Tổng thể có độ chênh lệch nhiều về dấu hiệu cần nghiên cứu.
  + Cách chọn mẫu không khách quan.
- Sai số do ghi chép gồm: hiểu chưa đúng, ghi nhầm, dụng cụ đo không phù hợp, …

* Sai số là điều không thể tránh khỏi dù mẫu được chọn có đúng quy cách đến bao nhiêu.

4.1.5.2. Một số biện pháp làm giảm sai số:

- Khâu chuẩn bị phải tốt về: quy trình, thiết bị hỗ trợ phù hợp, lựa chọn người điều tra, tuyên truyền mục đích, ý nghĩa của việc điều tra.
- Thiết kế các câu hỏi ngắn gọn, khoa học, dễ hiểu.

4.1.5.3. Một số yêu cầu, nguyên tắc chọn mẫu thống kê:

- Xác định kích cỡ mẫu đủ khả năng đại diện cho tổng thể.
- Quy định sai số cho phép, độ tin cậy.
- Ước tính độ lệch tiêu chuẩn.
- Suy rộng các kết quả điều tra.
- Chọn thời điểm điều tra phù hợp.
- Kiểm định mẫu, kiểm định số liệu nghi ngờ.

CHƯƠNG 6 KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

6.1. CÁC KHÁI NIỆM

6.1.1. Kiểm định giả thuyết thống kê

Khái niệm 6.1

Trong lĩnh vực sản xuất, nghiên cứu, chúng ta luôn luôn phải đưa ra những quyết định: quyết định dùng phương pháp này hay một phương pháp kia, quyết định thay thế một biện pháp này bằng một biện pháp khác tốt hơn hay giữ nguyên như cũ. Các quyết định của chúng ta hầu như luôn luôn phải dựa vào những hiểu biết có giới hạn của chúng ta nên không thể hoàn toàn chắc chắn. Luôn luôn có yếu tố rủi ro trong các quyết định của chúng ta. Ta có lúc đưa ra quyết định sai lầm. Các phương pháp suy diễn có thể giúp ta trong quá trình quyết định ấy. Đó là một lĩnh vực thống kê suy diễn được gọi là kiểm định giả thuyết hay kiểm định ý nghĩa (test of significance).

Kiểm định thống kê (statistical test) hay kiểm định giả thuyết thống kê (statistical hypothezis test) là một họ các phương pháp thống kê giúp cho người xử lý số liệu kết luận, quyết định về giả thuyết thống kê, giúp khẳng định các kết luận khoa học từ kết quả xử lý số liệu. Trong kiểm định, chúng ta thường đặt ra hai giả thuyết thống kê:

- Giả thuyết H_0: gọi là giả thuyết "null" (null hypothesis), giả định rằng không có sự khác biệt, không có liên hệ, hay là một kết luận mà chúng ta khả dĩ tin nó đúng từ dữ liệu thực nghiệm.
- Giả thuyết H_1: gọi là đối thuyết, là giả thuyết sẽ được chấp nhận khi chúng ta bác bỏ H_0.

Các kiểm định giả thuyết đòi hỏi phải đưa ra một đối thuyết H_1 nhất định, để khi giả thuyết H_0 bị bác bỏ thì ta chấp nhận đối thuyết H_1. Đối thuyết này chỉ rõ chiều hướng thay đổi (nhỏ hơn hoặc lớn hơn), đó là kiểm nghiệm giả thuyết có chiều hướng (directional hypothesis test). Hoặc chỉ nêu lên rằng có sự khác biệt mà thôi, đó là kiểm nghiệm giả thuyết không có chiều hướng (nondirectional hypothesis test).

Ví dụ 6.1

Giám đốc xí nghiệp cho rằng dường như chất lượng sản phẩm năm nay kém hơn năm trước. Nếu thật sự là vậy, giám đốc xí nghiệp sẽ phải sửa đổi chương trình và phương pháp sản xuất. Nhưng trước khi quyết định về vấn đề này ông giám đốc xí nghiệp phải làm một cuộc khảo sát về chất lượng sản phẩm năm nay so với chất lượng sản phẩm năm trước. Căn cứ vào kết quả khảo sát chất lượng sản phẩm các năm trước bằng một phương pháp tiêu chuẩn hóa mà giá trị đã được xác nhận thì điểm chất lượng sản phẩm các năm trước là 100. Năm nay ông giám đốc xí nghiệp quyết định chọn một mẫu ngẫu nhiên gồm các sản phẩm năm nay và cũng áp dụng phương pháp tiêu chuẩn hóa này cho nhóm sản phẩm mới này.

Nếu ông muốn biết rằng điểm trung bình chất lượng sản phẩm của tổng thể sản phẩm năm nay có phải là 100 như các năm trước hay không, các giả thuyết ông đưa ra để kiểm định là:

Giả thuyết H_0: $\mu = 100$

Đối thuyết H_1: $\mu \neq 100$

Đó là một kiểm định giả thuyết không có chiều hướng. Nó cho biết μ không bằng 100, nhưng không tiên đoán rằng μ nhỏ hơn hay lớn hơn 100. Trong trường hợp này, ông giám đốc chỉ muốn biết rằng chất lượng sản phẩm năm nay có khác với các chất lượng sản phẩm các năm trước hay không mà thôi. Nếu mẫu sản phẩm mới này cho kết quả điểm trung bình chất lượng sản phẩm là xấp xỉ 100 cho tổng thể sản phẩm năm nay thì hẳn ông giám đốc không phải băn khoăn gì nữa.

Nhưng nếu ông giám đốc tiên đoán rằng chất lượng sản phẩm năm nay kém thua chất lượng sản phẩm các năm trước. Trong trường hợp này, các giả thuyết của ông sẽ là:

Giả thuyết H_0: $\mu = 100$

Đối thuyết H_1: $\mu < 100$

Đây là một kiểm định giả thuyết có chiều hướng. Nếu H_0 bị bác bỏ, ông có thể chấp nhận H_1, và như vậy ông sẽ phải sửa đổi chương trình và phương pháp sản xuất.

Tương tự nếu ông giám đốc tiên đoán rằng chất lượng sản phẩm năm nay vượt xa hơn chất lượng sản phẩm các năm trước. Trong trường hợp này, các giả thuyết của ông sẽ là:

Giả thuyết H_0: $\mu = 100$

Đối thuyết H_1: $\mu > 100$

Đây là một kiểm định giả thuyết có chiều hướng. Nếu H_0 bị bác bỏ, ông có thể chấp nhận H_1, và như vậy ông sẽ giữ chương trình và phương pháp sản xuất hiện hành.

Kiểm nghiệm giả thuyết không có chiều hướng nói trên thường được gọi là kiểm nghiệm hai đuôi (two-tailed test), và kiểm nghiệm giả thuyết có chiều hướng được gọi là kiểm nghiệm một đuôi (one-tailed test).

Tóm lại

- Giả thuyết thống kê là một giả sử hay một phát biểu có thể đúng hoặc sai liên quan đến các tham số đặc trưng của một hay nhiều tổng thể dựa trên dữ liệu quan trắc thực nghiệm.
- Kiểm định giả thuyết thống kê là tìm ra kết luận chấp nhận hay bác bỏ một giả thuyết thống kê.
- Giả sử tham số θ là các đặc trưng của tổng thể chưa biết và dựa vào dữ liệu quan trắc thực nghiệm để nêu ra giả thiết $H_0 : \theta = \theta_0$ là giả thuyết cần kiểm định.
- Giả thuyết đối của H_0 kí hiệu là H_1. Có 3 trường hợp

$$\begin{cases} \mathrm{H}_0 : \theta = \theta_0; & \mathrm{H}_1 : \theta \neq \theta_0 \\ \mathrm{H}_0 : \theta = \theta_0; & \mathrm{H}_1 : \theta > \theta_0 \\ \mathrm{H}_0 : \theta = \theta_0; & \mathrm{H}_1 : \theta < \theta_0 \end{cases}$$

Trên đây ta chỉ mới đưa ví dụ về kiểm định thống kê với một mẫu. Ngoài ra ta còn có trường hợp kiểm định thống kê với nhiều mẫu. Ví dụ ta muốn biết công nhân nam và công nhân nữ có khác biệt hay không về năng suất lao động. Ta ra một khảo sát phương pháp tiêu chuẩn hóa cho một mẫu công nhân nam và một mẫu công nhân nữ. So sánh điểm trung bình của hai mẫu, ta có thể đánh giá được sự khác biệt giữa hai trung bình tổng thể nam và nữ. Trong ví dụ này, số thống kê hiệu số giữa hai trung bình mẫu được dùng để đánh giá tham số hiệu số giữa hai trung bình tổng thể.

6.1.1.1. Tiêu chuẩn kiểm định

Để kiểm định cặp giả thuyết thống kê H_0 và H_1, từ tổng thể, ta chọn mẫu ngẫu nhiên ($X_1, X_2, ..., X_n$). Dựa trên mẫu này ta xây dựng thống kê: $G = f(X_1, X_2, ..., X_n, \theta_0)$.

Trong đó θ_0 là một tham số liên quan đến H_0 sao cho H_0 đúng thì quy luật phân phối xác suất của G hoàn toàn xác định. Khi đó thống kê G được gọi là tiêu chuẩn kiểm định.

6.1.1.2. Các kiểm định thống kê thường gặp

Trong kiểm định thống kê thường sử dụng một trong các trị thống kê đã được nghiên cứu như trị Z (kiểm định chuẩn z), Student (kiểm định t), trị Fisher (kiểm định F), Chi bình phương (kiểm định χ^2) dùng cho việc đánh giá kiểm định giả thuyết. Ta có bảng tóm tắt các trị thống kê:

Kiểm định	Ý nghĩa – Áp dụng	Các phương pháp thống kê có sử dụng
Z	Dùng kiểm định trung bình, áp dụng cho phân phối chuẩn.	Khoảng tin cậy, so sánh
t	Dùng kiểm định trung bình, áp dụng cho phân phối chuẩn, đặc biệt khi cỡ mẫu nhỏ.	So sánh, hồi qui, tương quan
F	Phân tích phương sai ANOVA	ANOVA, hồi qui
χ^2	Kiểm định phân phối, các tỷ lệ trong các bảng thống kê	Xử lý bảng và các phân phối dữ liệu

6.1.2. Miền bác bỏ, miền chấp nhận

Để xây dựng miền bác bỏ ta sử dụng nguyên lý xác suất nhỏ: Nếu một biến cố có xác suất nhỏ chúng ta có thể coi nó không xảy ra trong một lần thực hiện phép thử. Vì đã biết quy luật phân phối xác suất của G, nên với một số α khá bé cho trước ta có thể tìm được miền W_α gọi là miền bác bỏ, sao cho nếu giả thuyết H_0 đúng thì xác suất để G nhận giá trị thuộc miền W_α bằng α: $P(G \in W_\alpha \mid H_0) = \alpha$.

Nếu trong một lần lấy mẫu, G nhận giá trị cụ thể g_{qs} sao cho:

- Nếu $g_{qs} \in W_\alpha$, bác bỏ H_0 và chấp nhận H_1;
- Nếu $g_{qs} \notin W_\alpha$, chưa đủ cơ sở bác bỏ H_0.

Tóm lại

- Miền bác bỏ là miền chứa giá trị của các đại lượng thống kê làm cho H_0 bị bác bỏ. Kí hiệu: W_α
- Miền chấp nhận là miền chứa giá trị của các đại lượng thống kê làm cho H_0 được chấp nhận. Kí hiệu: $\overline{W_\alpha}$.

6.1.2.1. Các bước kiểm định

i. Xác định bài toán kiểm định H_0 và H_1.

ii. Xây dựng tiêu chuẩn kiểm định G.

iii. Tìm miền bác bỏ W_α.

iv. Dựa vào mẫu quan sát, tính giá trị g_{qs} và kết luận.

6.1.3. Mức ý nghĩa

6.1.3.1. Các loại sai lầm

Việc kiểm định giả thuyết có thể dẫn đến một trong ba loại quyết định như sau:

1. Nếu giả thuyết H_0 bị bác bỏ, chấp nhận đối thuyết H_1.
2. Nếu giả thuyết H_0 không bị bác bỏ, chấp nhận giả thuyết H_0.
3. Nếu giả thuyết H_0 không bị bác bỏ, không quyết định gì cả. Chờ một cuộc thử nghiệm khác về sau.

Nhưng dù lựa chọn quyết định nào trên đây, ta cũng có thể phạm phải sai lầm. Có hai loại sai lầm, được gọi là: sai lầm loại 1 và sai lầm loại 2.

Nếu H_0 thực sự đúng nhưng kiểm định dẫn đến việc bác bỏ nó nên chúng ta chấp nhận đối thuyết H_1, đó là sai lầm loại 1.

Mặt khác, nếu H_0 thực sự sai, nhưng kiểm định không dẫn đến việc bác bỏ nó nên chúng ta chấp nhận H_0, đó là sai lầm loại 2.

Do đó sai lầm loại 1 là sai lầm mà ta phạm phải khi ta bác bỏ một giả thuyết kiểm định nếu giả thuyết ấy đúng. Sai lầm loại 2 là sai lầm chấp nhận giả thuyết kiểm định nếu giả thuyết ấy là sai.

6.1.3.2. Các mức ý nghĩa

Mỗi khi kiểm định giả thuyết, người nghiên cứu cần phải xác định xác suất sai lầm loại 1 mà mình có thể phạm phải. Xác suất lớn nhất mắc sai lầm loại 1 này được gọi là mức alpha (ký hiệu α) hay cũng gọi là mức ý nghĩa thống kê (level of significance).

$$\alpha = \max P(\text{bác bỏ } H_0 \mid H_0 \text{ đúng}) = \max P(H_1 \mid H_0)$$

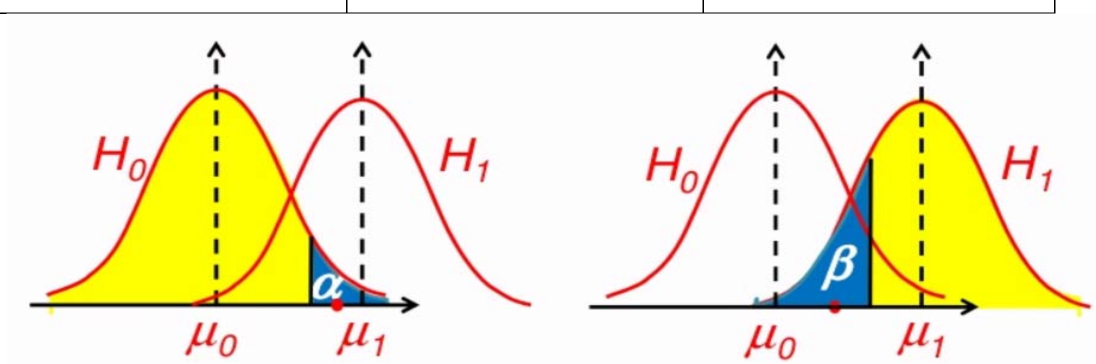
Mức alpha này có thể là bất cứ trị số nào ta muốn, nhưng thông thường, người ta đặt mức alpha này là 0,05. Nếu $\alpha = 0,05$, như vậy có nghĩa là có 5% may rủi H_0 là đúng và quyết định bác bỏ H_0 là một sai lầm. Quyết định này dựa vào dữ liệu của mẫu, chứ không phải căn cứ trên tổng thể, cho nên người ta không thể nào hoàn toàn chắc chắn về kết quả kiểm nghiệm giả thuyết được. Vì vậy ta cần phải đưa ra quyết định của ta với một xác suất sai lầm nào đó mà ta có thể phạm phải. Nếu ta bác bỏ giả thiết H_0 với mức $\alpha = 0,05$, như vậy có nghĩa là, nếu lặp lại cuộc thử nghiệm 100 lần, với 100 mẫu khác nhau, và cứ mỗi lần như vậy lại bác bỏ giả thuyết H_0 thì sẽ có khoảng 5 lần ta có thể phạm phải sai lầm loại 1 (nói cách khác, ta có may rủi 5% phạm phải sai lầm loại 1). Xác suất phạm phải sai lầm loại 2 được biểu thị bằng ký hiệu bê-ta β.

$$\beta = \max P(\text{chấp nhận } H_0 \mid H_0 \text{ sai}) = \max P(H_0 \mid H_1)$$

Không giống như α, ta không thể kiểm soát trực tiếp β được. Thế nhưng có điều quan trọng cần phải biết là có mối liên hệ nghịch giữa α và β. Nói cách khác, khi α tăng lên thì β giảm đi. Do đó người ta thường chọn α trong khoảng từ 1% đến 10%. Thế nhưng số lượng thay đổi ở β không trực tiếp tỷ lệ với số lượng thay đổi ở α. Một yếu tố quan trọng nhằm giảm thiểu cả hai loại sai lầm nói trên là cỡ mẫu. Khi cỡ mẫu tăng lên, các xác suất sai lầm loại 1 và 2 sẽ giảm đi.

● Các trường hợp xảy ra khi tiến hành kiểm định giả thuyết thống kê có thể tóm tắt dưới dạng bảng sau:

Thực tế / Kết luận	H_0 đúng	H_0 sai
Bác bỏ H_0	Sai lầm loại 1 (xác suất $=\alpha$)	Kết luận đúng (xác suất$=1-\beta$)
Chấp nhận H_0	Kết luận đúng (xác suất $=1-\alpha$)	Sai lầm loại 2 (xác suất$=\beta$)



Hình 6.1: Minh hoạ sai lầm α, β

Tóm lại

- Sai lầm loại 1: là sai lầm mắc phải khi ta bác bỏ H_0 trong khi H_0 đúng. Xác suất bác bỏ H_0 trong khi H_0 đúng là α.

Kí hiệu: $P(Z \in W_\alpha) = \alpha$ (đây là xác suất để tiêu chuẩn Z thuộc miền bác bỏ W_α khi giả thuyết H_0 đúng)

- Sai lầm loại 2: là sai lầm mắc phải khi ta chấp nhận H_0 trong khi H_0 sai. Xác suất chấp nhận H_0 trong khi H_0 sai là β.

Kí hiệu: $P(Z \notin W_\alpha | H_1) = \beta$ (đây là xác suất để tiêu chuẩn Z không thuộc miền bác bỏ W_α khi giả thuyết H_0 sai).

Khi đó biến cố không mắc sai lầm loại 2 là biến cố để Z thuộc miền bác bỏ W_α khi thực tế H_1 đúng. Kí hiệu: biến cố này là $(Z \in W_\alpha | H_1)$

Biến cố $(Z \in W_\alpha | H_1)$ đối lập với biến cố $(Z \notin W_\alpha | H_1)$ nên xác suất của nó là

$$P(Z \in W_\alpha | H_1) = 1 - \beta$$

$1 - \beta$ được gọi là lực kiểm định giả thuyết H_0(power of the hypothesis test) hay độ mạnh của kiểm định. Nó chính là xác suất không mắc sai lầm loại 2. β càng nhỏ lực kiểm định càng lớn.

-Trong thống kê qui ước sai lầm loại 1 tác hại hơn sai lầm loại 2 nên cần tránh hơn. Số α là mức ý nghĩa, thường $\leq 5\%$.

6.1.3.3. Giá trị xác suất của kiểm định:**giá trị P (P_value, p_value)**

Thay vì kiểm định giả thuyết với một mức ý nghĩa α định trước thì người ta cho rằng sau khi định rõ các giả thuyết kiểm định H_0 và giả thuyết đối H_1, chúng ta thu thập các số liệu mẫu và xác định mức độ khẳng định việc bác bỏ giả thiết H_0. Mức độ khẳng định này thường được gọi là giá trị xác suất P hay P_value.

Vậy phương pháp giá trị P dựa trên lý luận phản chứng: xuất phát trên số liệu mẫu quan sát được, giả sử giả thuyết H_0 đúng chúng ta tìm xác suất để đối thuyết H_1 xảy ra.

Định nghĩa 6.1

Giả định rằng giả thuyết H_0 là đúng.

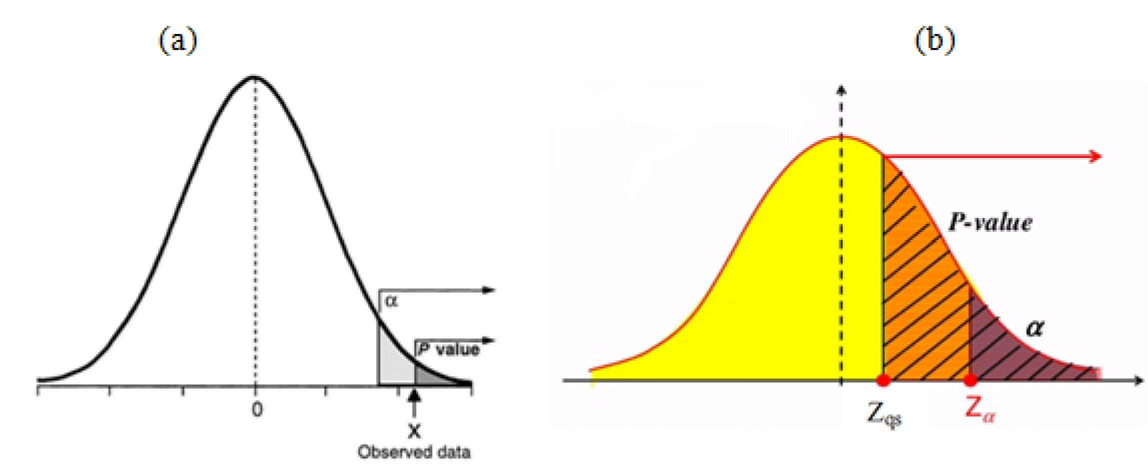
• Giá trị P (P_value) là mức ý nghĩa thống kê của đối thuyết H_1.

• Hay Giá trị P (P_value) là mức ý nghĩa quan sát , nó cho biết xác suất mắc sai lầm loại 1 tối đa khi bác bỏ giả thuyết H_0 với tập dữ liệu mẫu đang quan sát.

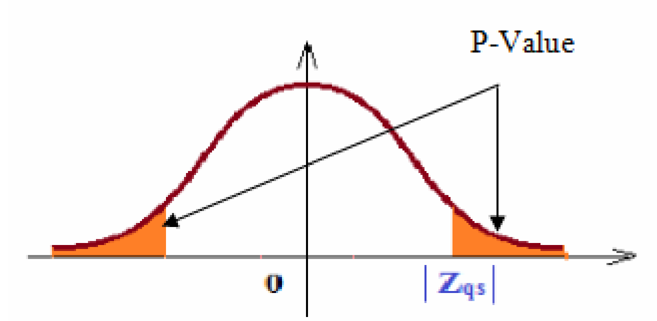
• Hay Giá trị P (P_value) là mức ý nghĩa nhỏ nhất (thấp hơn α) tại đó giả thuyết H_0 được bác bỏ và nghiêng về H_1.

+Nếu P_value < mức ý nghĩa α thì đối thuyết H_1 có ý nghĩa thống kê (bác bỏ H_0).

+Nếu P_value > mức ý nghĩa α thì đối thuyết H_1 không có ý nghĩa thống kê, chưa có cơ sở để chấp nhận đối thuyết H_1(chấp nhận H_0).



Hình 6.2: (a) Mức ý nghĩa quan sát P_value < mức ý nghĩa α,
(b) P_value > mức ý nghĩa α



Hình 6.3: Mức ý nghĩa quan sát P_value 2 phía

Ghi chú 6.1

Trong thực tế, việc kiểm định giả thuyết theo P_value thường theo nguyên tắc sau:

P_value	Kết luận H_0	Mức ý nghĩa
P_value < 0,001	Hoàn toàn yên tâm bác bỏ	Rất có ý nghĩa, rất thuyết phục về thống kê(very strong evidence), nhưng cũng đáng để nghi ngờ kết quả này
0,001 $\le$ P_value < 0,01	Ít băn khoăn khi bác bỏ hay có một cơ sở mạnh bác bỏ	Rất có ý nghĩa, thuyết phục về thống kê (strong evidence)
0,01 $\le$ P_value $\le$ 0,05	Nghiêng về hướng bác bỏ nhiều hơn hay có đủ cơ sở bác bỏ	Có ý nghĩa thống kê (substantial evidence)
0,05< P_value $\le$ 0,1	Cân nhắc cẩn thận trước khi bác bỏ (tuỳ chọn α)	Vùng xám (Gray area)
P_value > 0,1	Chấp nhận	Không có ý nghĩa thống kê

Như vậy P_value càng nhỏ thì mức độ khẳng định mẫu về việc bác bỏ H_0 càng rõ rệt hơn. Theo cách kiểm định giả thuyết này thì việc sử dụng P_value lại chính là kiểm định theo cách tiếp cận truyền thống.

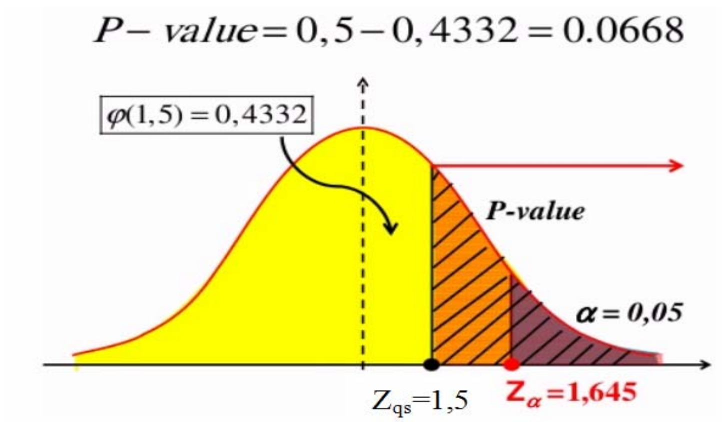
● **Cách tính giá trị p-value**

a. Trường hợp: σ^2 **đã biết hoặc** σ^2 **chưa biết,** $n \geq 30$

Nếu miền bác bỏ bên phải: P_value= $P(Z > z_{qs}) = 0,5 - \varphi(z_{qs})$

Ví dụ 6.1

Nếu giá trị kiểm định z_{qs} =1,5, ta tính P_value= $P(Z > 1,5) = 0,5 - \varphi(1,5)$ như hình



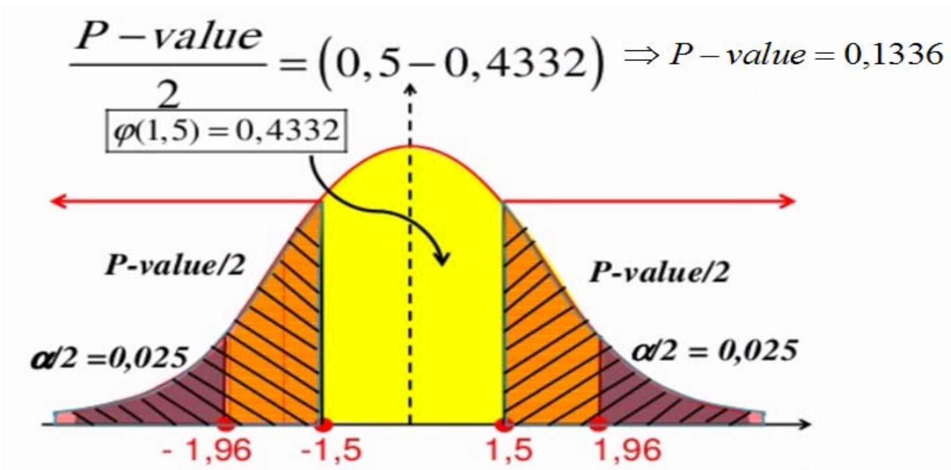
Hình 6.3 : P_value= $P(Z > 1,5)$

Nếu miền bác bỏ bên trái: P_value= $P(Z < z_{qs}) = 0,5 + \varphi(z_{qs})$

Nếu miền bác bỏ hai phía: P_value= $2P(Z > |z_{qs}|) = 1 - 2\varphi(z_{qs})$

Ví dụ 6.2 (tiếp theo Ví dụ 6.1)

Nếu giá trị kiểm định z_{qs} =1,5, ta tính P_value= $2P(Z > |z_{qs}|) = 1 - 2\varphi(z_{qs})$ như hình



Hình 6.4: P_value= $2P(Z > 1,5)$

b. Trường hợp : σ^2 **chưa biết,** $n < 30$ **và X có pp chuẩn**

Nếu miền bác bỏ bên phải: P_value= $P(T > t_{qs})$ tra bảng phụ lục 9.

Nếu miền bác bỏ bên trái: P_value= $P(T < t_{qs})$ tra bảng phụ lục 9.

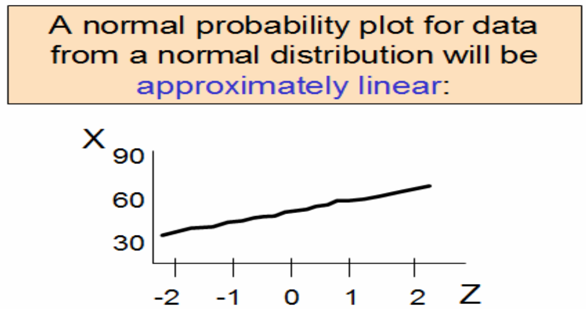
Nếu miền bác bỏ hai phía: P_value= $2P(T > |t_{qs}|)$ tra bảng phụ lục 9.

6.2. KIỂM ĐỊNH SỰ PHÙ HỢP VỀ QUY LUẬT PHÂN PHỐI XÁC SUẤT

6.2.1. Sự cần thiết kiểm định dạng phân phối của dữ liệu

+ Phương pháp xử lý, kiểm định thống kê phần lớn dựa trên giả thiết “tập hợp tổng thể có phân phối chuẩn (bình thường).

=Phương pháp dựa trên tham số: Trung bình, phương sai, độ lệch chuẩn…

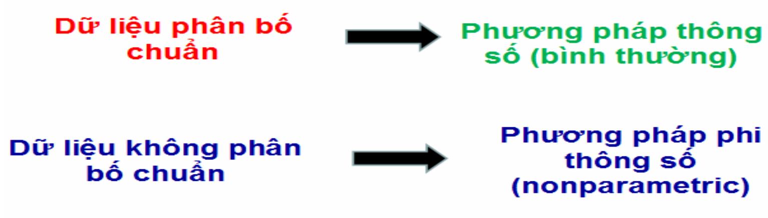


Hình 6.5: dữ liệu có phân phối chuẩn xấp xỉ đường thẳng tuyến tính

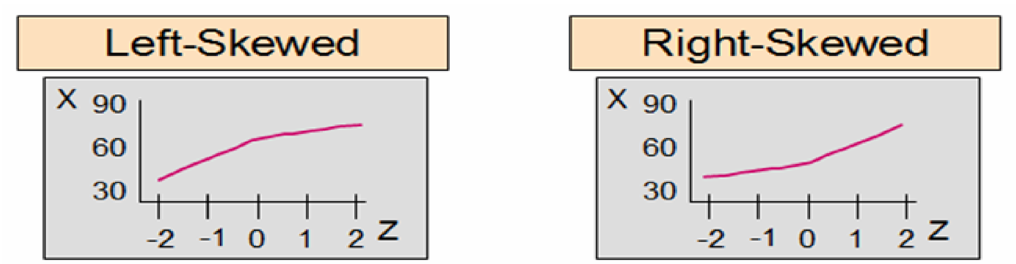
Nếu phân phối của dữ liệu là phân phối không chuẩn (lệch phải, lệch trái..)

=> Nên xử lý bằng phương pháp phi tham số (nonparametric method)

- ⇒ Phải kiểm định xem dạng phân bố trước khi xử lý
- ⇒ Đối với các kết luận có tính hệ trọng đối với thực tế => xử lý đúng phương pháp.



Hình 6.6: Kiểm định tham số và phi tham số
cho dữ liệu phân phối chuẩn và không chuẩn



(a) (b)

Hình 6.7: Phân phối không chuẩn lệch trái (a) và lệch phải (b)

6.2.2. Xem dạng phân phối chuẩn bằng đồ thị xác suất chuẩn (Normal Probability Plot)

H_0: X tuân theo quy luật phân phối chuẩn

H_1: X không tuân theo quy luật phân phối chuẩn

• Giả thuyết H_0: phân phối tuân theo luật chuẩn (p >0.05), nếu kết quả thu được p <0.05 thì không tuân theo luật chuẩn (chấp nhận H_1).

• Cần chú ý: các test này rất nhạy nên cần phải xem xét các yếu tố khác: độ méo lệch (skewness) và độ nhọn (kurtosis) của đường cong phân phối chuẩn.

Skewness (độ méo lệch) và Kurtosis (độ nhọn) là hai chỉ số chính chúng ta cần xem xét để quyết định biến định lượng có phân phối chuẩn hay không. Một biến ngẫu nhiên có phân phối chuẩn khi giá trị Skewness và Kurtosis tiến gần đến giá trị 0 và 0.

Minh hoạ skewness và kurtosis của phân phối chuẩn.

Các lệnh trong R

```
#Simulation
> n.sample <- rnorm(n = 10000, mean = 55, sd = 4.5)
#Skewness and Kurtosis
> library(e1071)
> skewness(n.sample)
[1] 0.01862268
> kurtosis(n.sample)
[1] 0.01611428
```

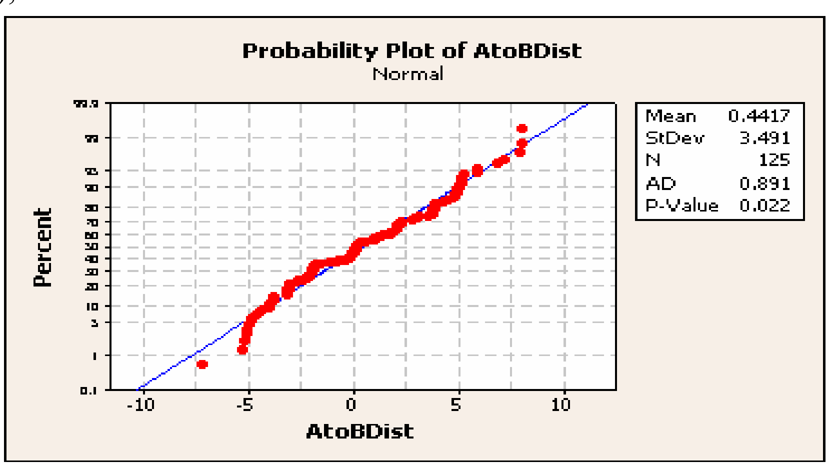
6.2.2.1. Trong MINITAB:

1. Stat/Basic Statistics/Normality Test
2. Khai biến số

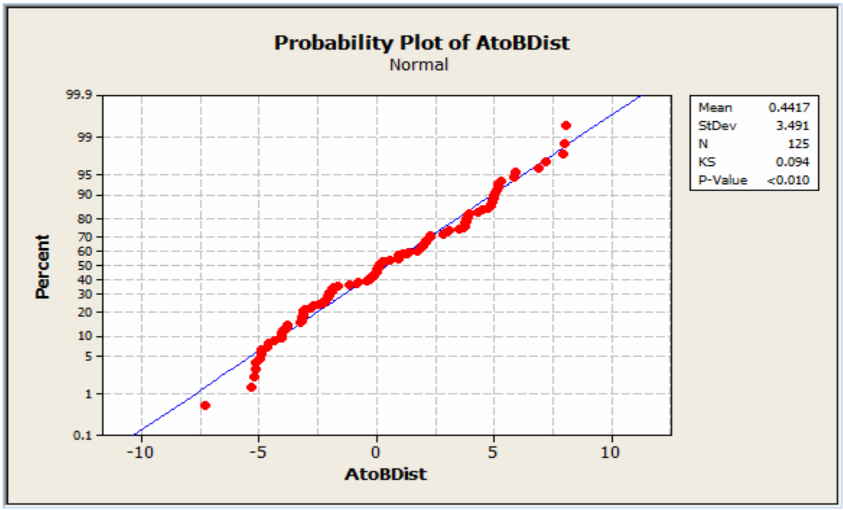
Ví dụ 6.3

Tập dữ liệu CRANKSH.MTW xem [3. Mô tả số liệu trong Minitab]

1. Open the worksheet CRANKSH.MTW.
2. Choose Stat > Basic Statistics > Normality Test.
3. Chuyển AtoBDist từ ô bên trái vào ô Variable bằng cách nhấn nút Select. Trong Tests for Nomality, đánh dấu tròn chọn Anderson-Darling (hay Kolmogorov-Smirnov), Click OK.



Hình 6.8: Biểu đồ kiểm định Anderson-Darling cho AtoBDist



Hình 6.9: Biểu đồ kiểm định Kolmogorov-Smirnov cho AtoBDist

* Tính Skewness (độ méo lệch) và Kurtosis (độ nhọn)

1. Choose Stat > Basic Statistics >Display Descriptive Statistics.
2. Chuyển AtoBDist từ ô bên trái vào ô Variable bằng cách nhấn nút Select. Trong Statistics, đánh dấu Skewness và Kurtosis, Click OK.

Descriptive Statistics: AtoBDist

```
Variable    N  N*   Mean  SE Mean  StDev  Minimum      Q1  Median     Q3
AtoBDist  125   0  0.442    0.312  3.491   -7.303  -2.243   0.130  3.607
Variable  Maximum  Skewness  Kurtosis
AtoBDist    8.023      0.12     -0.82
```

Giải thích kết quả

Đây là biểu đồ so sánh xác suất chuẩn với dữ liệu. Sự sai lệch của dữ liệu so với đường thẳng chuẩn, dễ thấy nhất là ở các điểm cực trị, hoặc ở hai đuôi. Biểu đồ trên cho thấy hai đuôi chệch so với đường chéo chuẩn. Và kiểm định Anderson-Darling cho giá trị p=0,022 < 0,05 hay kiểm định Kolmogorov-Smirnov cho giá trị p< 0,01 chỉ ra rằng có bằng chứng dữ liệu không tuân theo sự phân phối chuẩn. Thật vậy, các giá trị Skewness và Kurtosis lần lượt là 0,12 và -0,82 chưa tiến về 0.

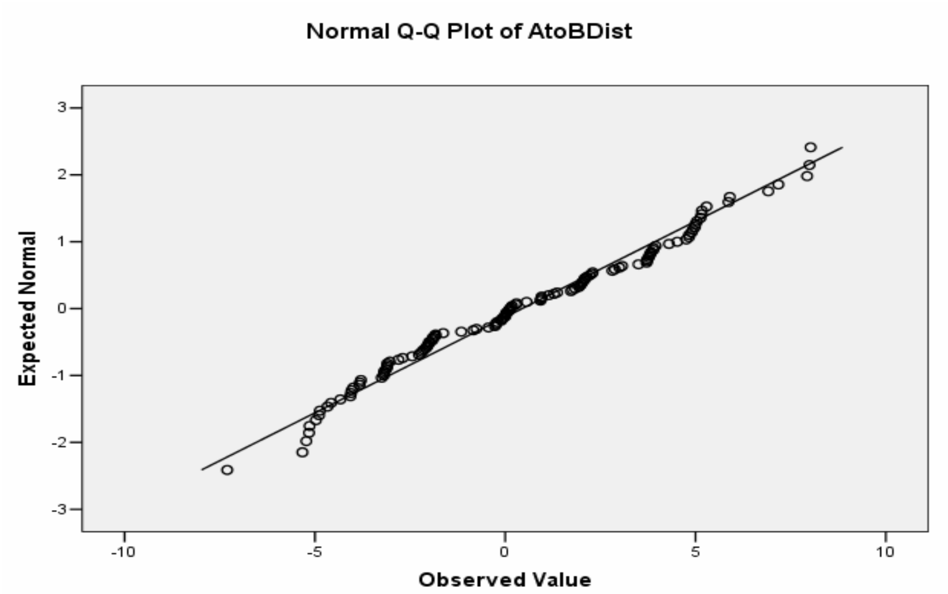
Ghi chú 6.3

Dùng phép kiểm định Kolmogorov-Smirnov khi cỡ mẫu lớn hơn 50 hoặc phép kiểm Anderson-Darling khi cỡ mẫu nhỏ hơn 50. Được coi là có phân phối chuẩn khi giá trị p <0,05.

6.2.2.2. Tương tự trong SPSS

Vẽ biểu đồ xác suất chuẩn Q-Q (Normal QQ plot) vào thực đơn: Analyze> Descriptive Statistics> Explore

Khi xuất hiện màn hình Explore, chuyển AtoBDist từ ô bên trái vào ô Dependent List Nhấn vào nút Plots. Đánh dấu stick vào ô Histogram và ô Normality plots with tests. Nhắp Continue và nhắp OK.



Hình 6.10: Biểu đồ Normal Q-Q plot

Descriptives

			Statistic	Std. Error
AtoBDist	Mean		.4417036	.31227646
	95% Confidence Interval for Mean	Lower Bound	-.1763790	
		Upper Bound	1.0597862	
	5% Trimmed Mean		.3925091	
	Median		.1300100	
	Variance		12.190	
	Std. Deviation		3.491357	
	Minimum		-7.303	
	Maximum		8.0232	
	Range		15.32608	
	Interquartile Range		5.84981	
	Skewness		.125	.217
	Kurtosis		-.823	.430

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
AtoBDist	.094	125	.008	.976	125	.028

Hình 6.11: Kết quả các giá trị Skewness và Kurtosis

Vì cỡ mẫu N=125 (lớn hơn 50), dùng kiểm định Kolmogorov-Smirnov với Sig.=0,008 < 0,05. Chứng tỏ phân phối này không là phân phối chuẩn. Thật vậy, các giá trị Skewness và Kurtosis lần lượt là 0,125 và -0,823 chưa tiến về 0.

Ghi chú 6.4

Dùng phép kiểm định Kolmogorov-Smirnov khi cỡ mẫu lớn hơn 50 hoặc
phép kiểm Shapiro-Wilk khi cỡ mẫu nhỏ hơn 50. Được coi là có phân phối chuẩn khi mức ý nghĩa (Sig.) lớn hơn 0,05.

6.2.2.3. Tương tự trong R

Vẽ biểu đồ xác suất chuẩn Q-Q (Normal quantile-quantile plot)

> data<-c(-0.44025, 5.90038, 2.08965, 0.09998, 2.01594, 4.83012, 3.78732, 4.99821, 6.91169,1.93847, -3.09907, -3.18827, 5.28978, 0.56182, -3.18960, 7.93177, 3.72692, 3.83152,-2.17454, 2.81598, 4.52023, 3.95372, 7.99326, 4.98677, -2.03427, 3.89134, 1.99825,0.01028, -0.24542, 2.08175, -4.86937, -2.69206, -3.02947, 2.99932, 3.50123, -1.99506,-1.62939, 2.14395, -1.90688,8.02322,4.75466,1.14240,0.93790,-7.30286, -5.22516,-4.06527,-1.91314,2.04590,4.93029,0.03095,-2.80363,-3.12681,-4.57793,-3.17924,-2.44537,1.36225,0.92825,-0.24151,-0.83762,-1.99674,4.90024,1.28079,2.87917,1.83867,-0.75614,3.72977,3.77141,-4.04994,3.89824,1.76868,2.27310,-3.82297,-2.26821,-2.07973,0.01739,3.71309,1.72573,3.07264,0.15676,-0.05666,3.81341,-3.78952,-3.81635,-4.88820,-3.24534,-0.27272,-4.33095,-1.83547,-3.98876,-4.97431,-5.14050,-0.10379,2.21033,5.13041,-1.89455,0.95119,-5.15414,4.82794,0.13001,-0.09911,-1.15453,2.29868,5.15847,0.08558,-3.09574,5.16744,0.29748,-4.66858,-2.13787,-0.00450,0.18096,4.30247,-2.21708,7.17603,5.86525,0.95699,-4.03441,-2.05086,-3.10319,-1.83001,5.03945,1.96583,-0.21026,0.27517,-5.32797)# 125 số liệu của dữ liệu CRANKSH.MTW

> mean(data)

[1] 0.4417036

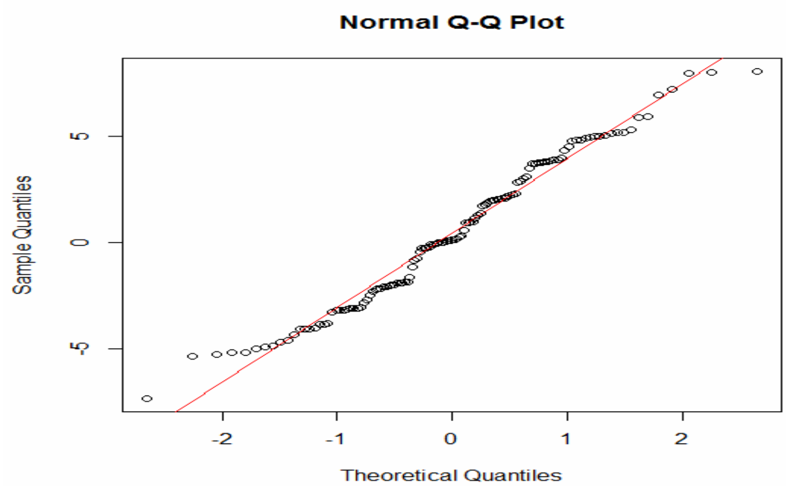
> sd(data)

[1] 3.491357

> qqnorm(data)

> abline(a=0.4417036, b=3.491357,col=2)

đường thẳng tuyến tính của phân phối chuẩn $N(\mu = 0.4417036, \sigma^2 = 3.491357^2)$



Hình 6.12: Biểu đồ Normal Q-Q plot trong R đối trục so với Biểu đồ Q-Q trong SPSS.

Kiểm định t dựa vào giả thuyết là phân phối của một biến phải tuân theo luật phân phối chuẩn (sẽ trình bày rõ trong phần III tiếp theo). Nếu giả định này không đúng, kết quả của kiểm định t có thể không hợp lí (valid). Để kiểm định phân phối chuẩn của dữ liệu trong R, chúng ta có thể dùng hàm shapiro.test như sau:

> shapiro.test(data)

Shapiro-Wilk normality test

data: data

W = 0.97647, p-value = 0.0279

>install.packages (c(' e1071'))

> library(e1071)

> skewness(data)

[1] 0.1215466

> kurtosis(data)

[1] -0.8725303

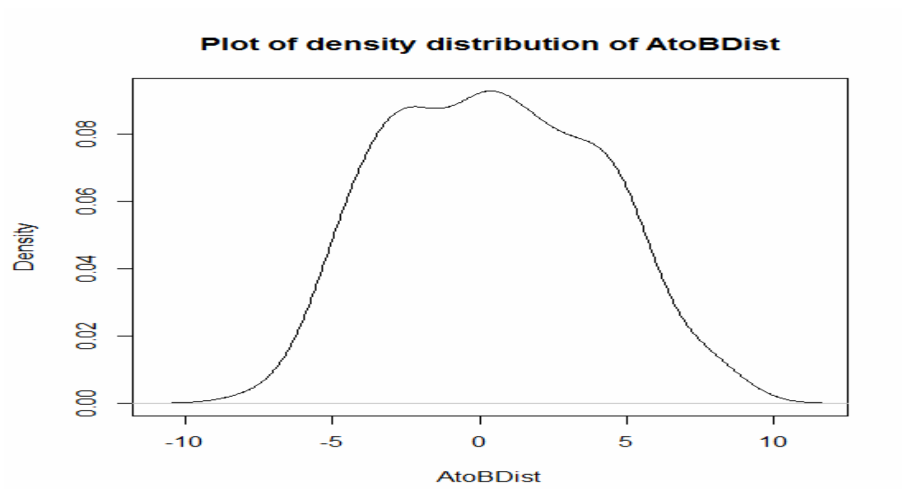
Giải thích:(Interpretation)

Giá trị p-value = 0.0279 của kiểm định Shapiro-Wilk nhỏ hơn 0,05 cho nên chúng ta có thể nói rằng dữ liệu không tuân theo sự phân phối chuẩn. Thật vậy, các giá trị Skewness và excess Kurtosis lần lượt là 0,121 và -0,872 chưa tiến về 0.

Giải thích

Độ méo lệch ở đây là 0,1215466. Giá trị này hàm ý rằng sự phân phối dữ liệu có độ lệch trái (có "đuôi" phía phải) hay lệch dương, vì giá trị tính toán lớn hơn 0. Đối với kurtosis, chúng ta có -0,8725303 hàm ý rằng sự phân phối dữ liệu là platykurtic, vì giá trị tính toán nhỏ hơn 0. Hình minh họa của dữ liệu được trong Hình 1.

> plot(density(data), main="Plot of density distribution of AtoBDist", xlab="AtoBDist", add=TRUE)



Hình 6.13: Hàm mật độ của AtoBDist

6.2.2.4. Tóm lại:

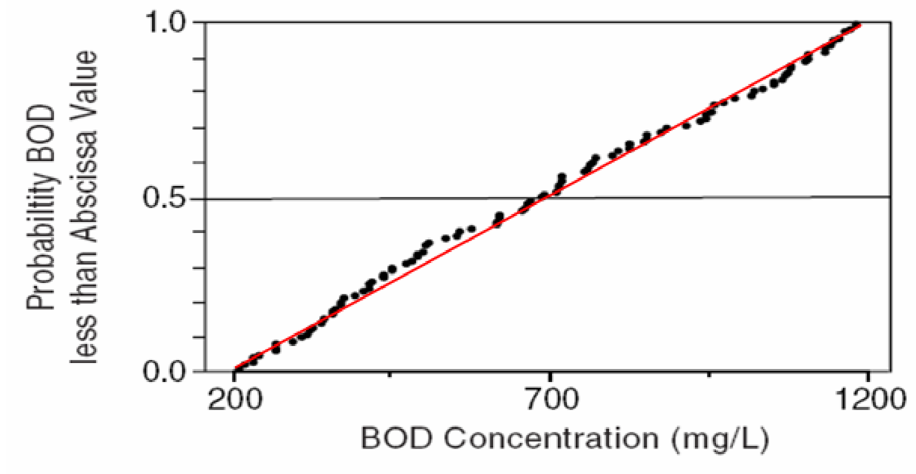
* Dấu hiệu nhận biết phân phối chuẩn trong biểu đồ xác suất chuẩn (Normal Probability plot hay Normal Q-Q plot): phân phối chuẩn khi biểu đồ xác suất này có quan hệ tuyến tính (đường thẳng)

Biểu đồ xác suất chuẩn (Normal Probability plot) gồm:

+ Trục tung biểu thị tần suất tích lũy (cumulative percent) , trục hoành biểu thị trị của các đơn vị trong mẫu từ thấp đến cao.

+ Đường thẳng có độ dốc biểu thị cho phân phối chuẩn có thông số (μ,σ) diễn đạt dưới dạng tuyến tính.

+ Nếu các điểm dữ liệu quan sát có phân phối "chập" với đường thẳng phân phối chuẩn, tức là các điểm nằm đè và sát với đường chéo chuẩn, ta kết luận dữ liệu có phân phối chuẩn . Xem hình minh hoạ 6.14



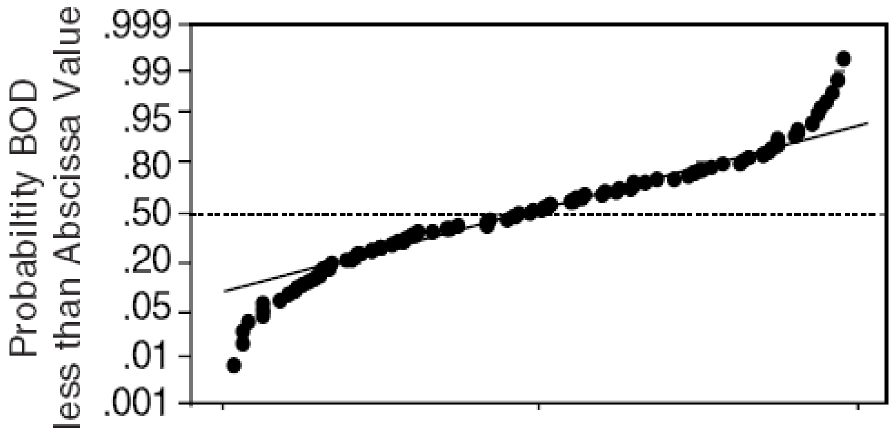
Hình 6.14: Dữ liệu BOD tuân theo phân phối chuẩn

Ghi chú 6.5

BOD**:** chỉ số nhu cầu ôxy sinh hóa (Biochemical (hay Biological) Oxygen Demand**)** Trong môi trường nước, khi quá trình oxy hoá sinh học xảy ra thì các vi sinh vật sử dụng oxy hoà tan, vì vậy xác định tổng lượng oxy hoà tan cần thiết cho quá trình phân huỷ sinh học là phép đo quan trọng đánh giá ảnh hưởng của một dòng thải đối với nguồn nước. BOD có ý nghĩa biểu thị lượng các chất thải hữu cơ trong nước có thể bị phân huỷ bằng các vi sinh vật.

Phần lớn các con sông còn nguyên sơ sẽ có BOD 5 ngày (ký hiệu là BOD_5) là nhỏ hơn 1 mg/L. Các con sông bị ô nhiễm ở mức độ nhẹ sẽ có giá trị BOD_5 trong khoảng 2–8 mg/L. Nước thải đô thị được xử lý có hiệu quả bằng công nghệ ba giai đoạn có thể có giá trị của BOD_5 vào khoảng 20 mg/L. Nước thải chưa xử lý thì có giá trị BOD_5 không cố định, nhưng trung bình vào khoảng 600 mg/L tại châu Âu và khoảng 200 mg/L tại Hoa Kỳ hay tại các khu vực mà nó bị thấm lọc qua nước ngầm hay nước bề mặt. Các giá trị nói chung của Hoa Kỳ thấp chủ yếu là do tại đây lượng nước tiêu thụ trên đầu người là cao hơn rất nhiều so với các khu vực khác của thế giới. Bùn sệt từ các trang trại chăn nuôi bò sữa có giá trị BOD_5 vào khoảng 8.000 mg/L còn thức ăn ủ thành xi lô có giá trị BOD_5 vào khoảng 60.000 mg/L.

+ Nếu các điểm dữ liệu có phân phối chệch (trên, dưới) ở hai đuôi so với đường chéo chuẩn thì ta kết luận phân phối không chuẩn, chuyển sang kiểm định phi tham số. Xem hình minh hoạ 6.15.



Hình 6.15: Dữ liệu BOD không tuân theo phân phối chuẩn

6.2.2.5. Cơ sở toán thống kê cho các điểm dữ liệu quan sát có phân phối "chập" với đường thẳng phân phối chuẩn như sau:

Chúng ta kiểm nghiệm dựa trên điểm chuẩn (NSCORES) của các giá trị mẫu. Các điểm chuẩn tương ứng với một mẫu $x_1, x_2, ..., x_n$ thu được bằng cách sau đây:

• Trước tiên chúng ta gọi

$$r_i = \text{thứ hạng của } x_i, \quad i = 1, ..., n$$

Thứ hạng của x_i là vị trí của x_i trong danh sách mẫu được sếp tăng dần. Vì vậy, hạng của giá trị nhỏ nhất là 1, của giá trị nhỏ nhất thứ hai là 2,v.v.

• Tiếp theo ta đặt

$$p_i = \frac{r_i - 3/8}{n + 1/4}, \; i = 1, ..., n$$

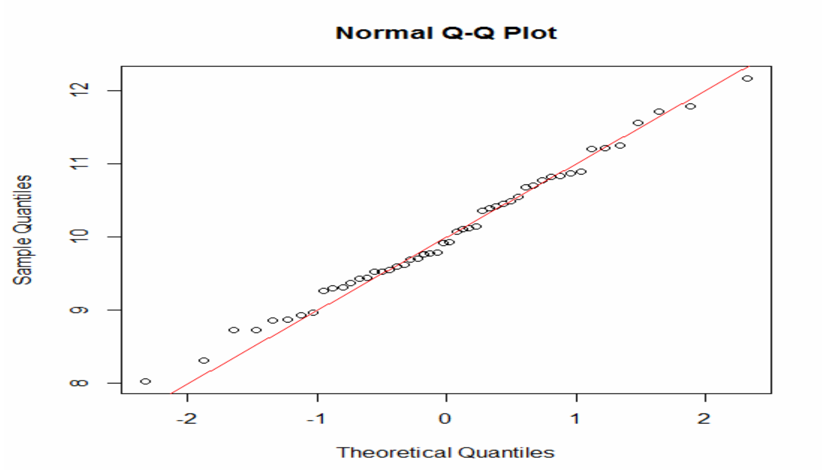
• Khi đó điểm chuẩn của x_i là $z_i = \varphi^{-1}(p_i)$, là phân vị thứ p_i của phân phối chuẩn tắc.

• Quy tắc kiểm định phân phối chuẩn của dữ liệu mẫu

Nếu mẫu được lấy ngẫu nhiên từ một phân phối chuẩn $N(\mu, \sigma^2)$ thì mối quan hệ giữa các điểm chuẩn (NSCORES) và x_i sẽ là xấp xỉ tuyến tính. Theo đó, hệ số tương quan giữa các giá trị mẫu $x_1, x_2, ..., x_n$ và NSCORES của chúng sẽ gần 1 trong các mẫu lớn.

Biểu đồ của các giá trị mẫu $x_1, x_2, ..., x_n$ và NSCORES của chúng được gọi là sơ đồ Q-Q chuẩn (normal Q-Q plot) cho trong hình 6.16. Ta vẽ được hình 6.16 khi áp dụng các lệnh trong R sau:

```
> set.seed(123)
> X<-rnorm(50,10,1)
> qqnorm(y=X)# biểu đồ xác suất chuẩn của dữ liệu
> abline(a=10,b=1,col=2)# đường thẳng tuyến tính của phân phối chuẩn chuẩn tắc
```



Hình 6.16: Q-Q chuẩn của các giá trị mô phỏng từ phân phối N(10,1)

Ta có sơ đồ xác suất chuẩn của n=50 giá trị mô phỏng từ N(10,1). Nếu mô phỏng tốt và mẫu thực sự được tạo ra từ N(10,1) thì các giá trị của X so với NSCORES sẽ được phân phối ngẫu nhiên xung quanh đường thẳng X=10 + NSCORES. Ta thấy trong hình 6.16 rằng đây thực sự là trường hợp như vậy.

Hệ số tương quan giữa giá trị x và NSCORES của nó là 0,976. Hồi quy tuyến tính của giá trị x theo NSCORES là $X = 10,043 + 0,976 * \text{NSCORES}$.

Ta thấy rằng cả hai tung độ gốc và độ dốc của phương trình hồi quy là gần với giá trị danh nghĩa của $\mu = 10$ và $\sigma = 1$.

6.2.3. Xem dạng phân phối bằng kiểm định giả thuyết (dùng cho dữ liệu có dạng bảng tần số)

6.2.3.1.Kiểm định sự phù hợp về quy luật phân phối xác suất

H_0: X tuân theo quy luật đã cho và H_1: X không tuân theo quy luật đã cho

Tính toán	- $\chi^2_{qs} = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i}$ với $n'_i = n.p_i$ - $\chi^2_\alpha(k-1-m)$
So sánh	- $\chi^2_{qs} \le \chi^2_\alpha$: chấp nhận giả thiết H_0. - $\chi^2_\alpha < \chi^2_{qs}$:chấp nhận giả thiết H_1.

Trong đó m là số tham số cần ước lượng (để tính các p_i),

k là số giá trị của x_i để tính các p_i.

Từ mẫu đã cho ta tính được các ước lượng hợp lí cực đại các tham số của F(x). Từ các tham số đó ta sẽ tính được các xác suất:

- Nếu X là biến ngẫu nhiên rời rạc có dãy phân phối xác suất là $\{p_i\}$ thì $n'_i = n.p_i$

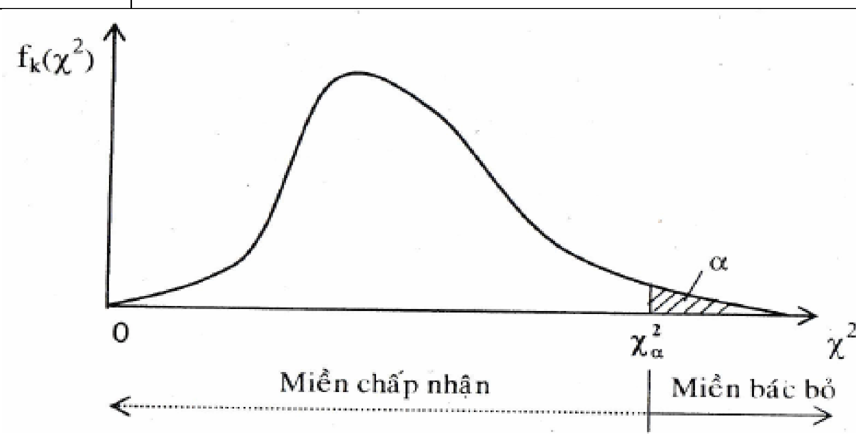
- Nếu X là biến ngẫu nhiên liên tục có hàm phân phối xác suất F(x) hoặc hàm mật độ f(x) thì $n'_i = n.P(x_{i-1} < X < x_i) = n.[F(x_i) - F(x_{i-1})] = n.\left(\int_{x_{i-1}}^{x_i} f(x)dx\right)$

6.6.2. Kiểm định sự phù hợp về quy luật phân phối chuẩn

H_0: X tuân theo quy luật phân phối chuẩn

H_1: X không tuân theo quy luật phân phối chuẩn

Tính toán	- $\chi^2_{qs} = n\left[\frac{a_3^2}{6} + \frac{(a_4-3)^2}{24}\right]$ với + Hệ số bất đối xứng $a_3 = \frac{\frac{1}{n}\sum_1^n n_i(x_i - \overline{x})^3}{s^3}$ + Hệ số nhọn $a_4 = \frac{\frac{1}{n}\sum_1^n n_i(x_i - \overline{x})^4}{s^4}$ - $\chi^2_\alpha(k-1-m)$
So sánh	- $\chi^2_{qs} \le \chi^2_\alpha$: chấp nhận giả thuyết H_0. - $\chi^2_\alpha < \chi^2_{qs}$:chấp nhận giả thuyết H_1.



Hình 6.17: Kiểm định Chi-bình phương một phía-bên phải

Ví dụ 6.4

Đo chiều cao của một loại cây mọc tự nhiên trong một khu rừng:

Chiều cao	3-8	8-13	13-18	18-23	23-28	28-33	33-38
Số cây	6	8	15	40	16	8	7

Có thể cho rằng chiều cao của cây đó là biến ngẫu nhiên tuân theo phân phối chuẩn. Cho kết luận với mức ý nghĩa 5%.

Giải

H_0: X tuân theo quy luật phân phối chuẩn $N(\mu,\sigma^2)$, μ,σ^2 chưa biết.

H_1: X không tuân theo quy luật phân phối chuẩn

Trước hết ta ước lượng μ,σ. Ta thay thế 2 tham số bằng ước lượng hợp lý cực đại là $\overline{x}, \hat{s}$. Dựa vào bảng số liệu ta có $\overline{x} = 20,7;\ \hat{s} = 7,28$.

i	x_{i-1}	x_i	n_i	$\varphi\left(\frac{x_{i-1}-\overline{x}}{\hat{s}}\right)$	$\varphi\left(\frac{x_i-\overline{x}}{\hat{s}}\right)$	$n_i' = n.p_i$	$(n_i - n_i')^2 / n_i'$
1	3	8	6	-0,4924	-0,4591	3,33	2,1408
2	8	13	8	-0,4591	-0,3554	10,37	0,542
3	13	18	15	-0,3554	-0,1443	21,11	1,768
4	18	23	40	-0,1443	0,1255	26,98	6,283
5	23	28	16	0,1255	0,3413	21,58	1,443
6	28	33	8	0,3413	0,4545	11,32	0,974
7	33	38	7	0,4545	0,4911	3,66	3,047
			100				16,19

$\chi^2_{qs} = \sum_{i=1}^{k} \frac{(n_i - n_i')^2}{n_i'}$ =16,19

Tra bảng $\chi^2_\alpha(k-1-m) = \chi^2_\alpha(7-1-2) = \chi^2_{0,05}(4) = 9,487$

$\chi^2_{qs} = 16,19 > \chi^2_{0,05}(4) = 9,487$ suy ra bác bỏ H_0.

Kết luận: X không tuân theo quy luật phân phối chuẩn.

Ví dụ 6.5

Giá của một loại cổ phiếu bán trên thị trường chứng khoán trong 100 phiên giao dịch được cho ở bảng sau:

Giá cổ phiếu (1000đ)	13-15	15-17	17-19	19-21	21-23
Số phiên giao dịch	5	18	42	27	8

Với mức ý nghĩa 5% có thể cho rằng giá của một loại cổ phiếu nói trên là biến ngẫu nhiên tuân theo phân phối chuẩn.

Giải

H_0: X tuân theo quy luật phân phối chuẩn $N(\mu,\sigma^2)$, μ,σ^2 chưa biết.

H_1: X không tuân theo quy luật phân phối chuẩn

i	x_i	n_i	$n_i.x_i$	$n_i.x_i^2$	$n_i.(x_i-\overline{x})^3$	$n_i.(x_i-\overline{x})^4$
1	14	5	70	980	-397,54	1709,40
2	16	18	288	4608	-219,01	503,71
3	18	42	756	13608	-1,13	0,34

4	20	27	540	10800	132,65	225,51
5	22	8	176	3872	405,22	1499,33
		100	1830	33868	-79,80	3938,29

Tính được $\overline{x} = 18,3;\ s = 1,96$. Hệ số bất đối xứng $a_3 = \dfrac{\frac{1}{n}\sum_1^n n_i(x_i - \overline{x})^3}{s^3} = -1,0815$

+ Hệ số nhọn $a_4 = \dfrac{\frac{1}{n}\sum_1^n n_i(x_i - \overline{x})^4}{s^4} = 2,74176$

+ $\chi_{qs}^2 = n\left[\dfrac{a_3^2}{6} + \dfrac{(a_4 - 3)^2}{24}\right] = 0,4728$

+ Tra bảng $\chi_\alpha^2(k-1-m) = \chi_\alpha^2(5-1-2) = \chi_{0,05}^2(2) = 5,99$

+ $\chi_{qs}^2 = 0,4728 < \chi_{0,05}^2(2) = 5,99$ suy ra chấp nhận H_0.

Kết luận: Giá của một loại cổ phiếu tuân theo quy luật phân phối chuẩn.

Ví dụ 6.6

Ở một cửa hàng bán xăng dầu, theo dõi nhu cầu của mặt hàng xăng trong một số ngày có luật phân phối chuẩn, chúng ta có kết quả ở bảng sau dạng [a,b):

Số bán ra (lít)	Số ngày	Số bán ra (lít)	Số ngày
20-30	3	70-80	25
30-40	8	80-90	17
40-50	30	90-100	9
50-60	45	100-110	4
60-70	20		

Ta sẽ kiểm định sự phù hợp về quy luật phân phối chuẩn

Gọi X là lượng xăng bán ra trong ngày, μ là lượng xăng trung bình bán ra trong ngày.

H_0: X tuân theo quy luật phân phối chuẩn $N(\mu, \sigma^2)$, μ, σ^2 chưa biết.

H_1: X không tuân theo quy luật phân phối chuẩn

x_{i0}	n_i	$x_{i0} * n_i$	$x_{i0}^2 * n_i$	$n_i.(x_i - \overline{x})^3$	$n_i.(x_i - \overline{x})^4$
25	3	75	1875	-159090.717	5977038.248
35	8	280	9800	-167648.737	4622075.672
45	30	1350	60750	-162718.353	2858961.459
55	45	2475	136125	-19520.9142	147773.3204
65	20	1300	84500	286.97814	697.3568802
75	25	1875	140625	48012.3977	596794.1031
85	17	1445	122825	191838.929	4302947.187
95	9	855	81225	306961.109	9954748.77

105	4	420	44100	305547.748	12964390.93
Tổng	n=161	$\sum_{i=1}^{9} n_i x_{i0} = 10075$	$\sum_{i=1}^{9} n_i x_{i0}^2 = 681825$	343668,441	41425427,05

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{9} n_i x_{i0} = \frac{10075}{161} = 62,57$$

$$\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{9} n_i x_{i0}^2 - \left(\overline{x}\right)^2 = \frac{681825}{161} - (62,57)^2 = 319,93$$

$$s^2 = \frac{161}{160}\cdot 319,93 = 321,92 \Rightarrow \sqrt{s^2} = s = 17,94$$

Ta tính được: $\overline{x} = 62,57$ và $s = 17,94$.

- Hệ số bất đối xứng $a_3 = \dfrac{\frac{1}{n}\sum_{1}^{n} n_i (x_i - \overline{x})^3}{s^3} = \dfrac{343668,441/161}{17,94^3} = 0,369$

- Hệ số nhọn $a_4 = \dfrac{\frac{1}{n}\sum_{1}^{n} n_i (x_i - \overline{x})^4}{s^4} = \dfrac{41425427.05/161}{17,94^4} = 2,483$

$$\chi_{qs}^2 = n\left[\frac{a_3^2}{6} + \frac{(a_4-3)^2}{24}\right] = 161\left[\frac{0,369^2}{6} + \frac{(2,483-3)^2}{24}\right] = 5,446$$

Tra bảng phụ lục 5, $\chi_\alpha^2(k-1-m) = \chi_\alpha^2(9-1-2) = \chi_{0,01}^2(6) = 16,81$

$\chi_{qs}^2 = 5,446 < \chi_{0,01}^2(6) = 16,81$ suy ra chấp nhận H_0.

Kết luận: X tuân theo quy luật phân phối chuẩn.

Ví dụ 6.7

Sản phẩm sản xuất ra trên dây chuyền tự động được đóng gói ngẫu nhiên theo quy cách 3 sản phẩm /1 hộp. Có thể xem số chính phẩm của một hộp là biến ngẫu nhiên tuân theo quy luật nhị thức không, biết rằng kiểm tra 100 hộp người ta 75 hộp không có phế phẩm, 20 hộp có 1 phế phẩm, 5 hộp có 2 phế phẩm, không hộp nào có 3 phế phẩm ? cho kết luận với mức ý nghĩa 5%.

Giải

H_0: X tuân theo quy luật nhị thức B(n=3; p)

H_1: X không tuân theo quy luật nhị thức

Ta ước lượng p. Ta thay thế p bằng ước lượng không chệch là f. Theo số liệu ta có

n=100 x 3=300 sản phẩm

m=5 x 1+20 x 2+75 x 3=270 chính phẩm

f=m/n=270/300=0,9

Do đó p_i được tính như sau

$$p_i = C_3^i 0,9^i (1-0,9)^{3-i}, i = \overline{0,3}$$

Tham số cần ước lượng là p.

x_i	n_i	p_i	$n'_i = n.p_i$	$(n_i - n'_i)^2 / n'_i$
0	0	0,001	0,1	0,1
1	5	0,027	2,7	1,959
2	20	0,243	24,3	0,761
3	75	0,729	72,9	0,06
N=	100			2,881

$\chi^2_{qs} = \sum_{i=1}^{k} \frac{(n_i - n'_i)^2}{n'_i}$ =2,881

Tra bảng $\chi^2_\alpha(k-1-m) = \chi^2_\alpha(4-1-1) = \chi^2_{0,05}(2) = 5,991$

$\chi^2_{qs} < \chi^2_{0,05}(2)$ suy ra chấp nhận H_0.

Kết luận: X tuân theo quy luật nhị thức B(n=3; p=0,9)

Ví dụ 6.8

Kiểm tra 200 thùng một loại đồ hộp người ta được số liệu sau

Số hộp bị hỏng/thùng	0	1	2	3	4
Số thùng	116	56	22	4	2

Chứng tỏ rằng số hộp bị hỏng của một thùng là biến ngẫu nhiên tuân theo quy luật Poisson với mức ý nghĩa 5%.

Giải

H_0: X tuân theo quy luật Poisson $P(\lambda)$, λ chưa biết.

H_1: X không tuân theo quy luật Poisson

Ta thay thế λ bằng $\overline{x}$. Tính $\overline{x}$ ta được

$\overline{x} = (116 \times 0 + 56 \times 1 + 22 \times 2 + 4 \times 3 + 2 \times 4)/200 = 0,6$

Vậy $p_i = P(X = i) = e^{-\lambda} \cdot \frac{\lambda^i}{i!} = e^{-0,6} \cdot \frac{0,6^i}{i!}$, $i = \overline{0,4}$ ta tính được

$p_0 = 0,588;\ p_1 = 0,3293;\ p_2 = 0,0988;\ p_3 = 0,0198;\ p_4 = 0,003$

Chú ý: Do tần số thực nghiệm 2 giá trị cuối nhỏ hơn 5 nên tiến hành ghép lớp (4+2=6) và tần số lý thuyết là (3,96+0,6=4,56).

x_i	n_i	p_i	$n'_i = n.p_i$	$(n_i - n'_i)^2 / n'_i$
0	116	0,588	109,76	0,3548
1	56	0,3293	65,86	1,4762
2	22	0,0988	19,76	0,2539
3	4 } 6	0,0198	3,96 } 4,56	0,4547
4	2	0,003	0,6	
N=	200			2,5396

$\chi_{qs}^2 = \sum_{i=1}^{k} \frac{(n_i - n_i')^2}{n_i'}$ =2,5396

Tra bảng phụ lục 5, $\chi_{\alpha}^2(k-1-m) = \chi_{\alpha}^2(4-1-1) = \chi_{0,05}^2(2) = 5,991$

$\chi_{qs}^2 < \chi_{0,05}^2(2)$ suy ra chấp nhận H_0.

Kết luận: X tuân theo quy luật Poisson $P(\lambda = 0,6)$.

6.3. KIỂM ĐỊNH GIẢ THUYẾT CÓ THAM SỐ

6.3.1. Kiểm định giả thuyết về trung bình tổng thể

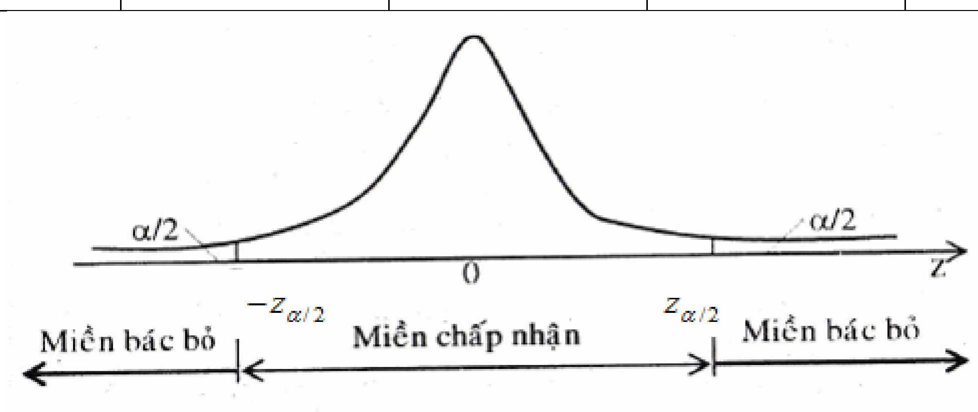
Bài toán 6.1

Giả sử tổng thể X có phân phối chuẩn có trung bình μ chưa biết. Ta cần kiểm định giả thuyết $H_0 : \mu = \mu_0$. Hãy đưa ra quy tắc căn cứ vào mẫu gồm n quan sát độc lập $(x_1,...,x_n)$ mà chấp nhận hay bác bỏ giả thiết trên với mức ý nghĩa α. Biết biến định lượng trong tổng thể có phân phối chuẩn.

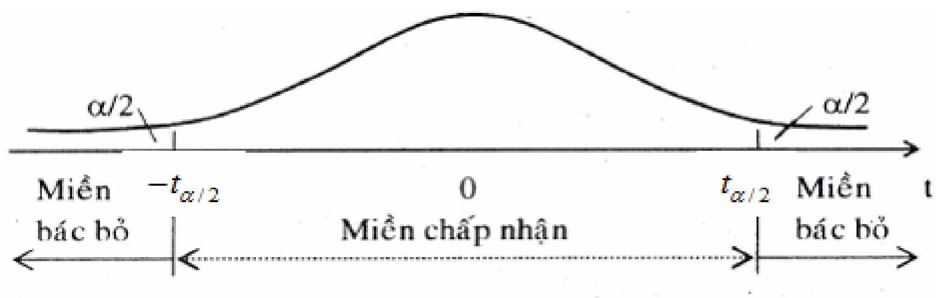
Quy tắc thực hành 6.1

● **Bảng kiểm định 2 phía (two tail):** Giả thuyết $H_0 : \mu = \mu_0$ và Đối thuyết $H_1 : \mu \neq \mu_0$

Bước	TH1: $n \geq 30$	TH2: $n \geq 30$	TH3: $n < 30$	TH4: $n < 30$
	σ^2 đã biết	σ^2 chưa biết	σ^2 đã biết	σ^2 chưa biết
1.Tính z hoặc t	$z = \frac{\overline{x} - \mu_0}{\sigma} \cdot \sqrt{n}$	$z = \frac{\overline{x} - \mu_0}{s} \cdot \sqrt{n}$	$z = \frac{\overline{x} - \mu_0}{\sigma} \cdot \sqrt{n}$	$t = \frac{\overline{x} - \mu_0}{s} \cdot \sqrt{n}$
2.Tra bảng	$z_{\alpha/2}$ phụ lục 6	$z_{\alpha/2}$ phụ lục 6	$z_{\alpha/2}$ phụ lục 6	$t_{\alpha/2}^{n-1}$ phụ lục 8
3a. Chấp nhận H_0	$\vert z \vert \leq z_{\alpha/2}$	$\vert z \vert \leq z_{\alpha/2}$	$\vert z \vert \leq z_{\alpha/2}$	$\vert t \vert \leq t_{\alpha/2}^{n-1}$
3a. Bác bỏ H_0	$\vert z \vert > z_{\alpha/2}$	$\vert z \vert > z_{\alpha/2}$	$\vert z \vert > z_{\alpha/2}$	$\vert t \vert > t_{\alpha/2}^{n-1}$



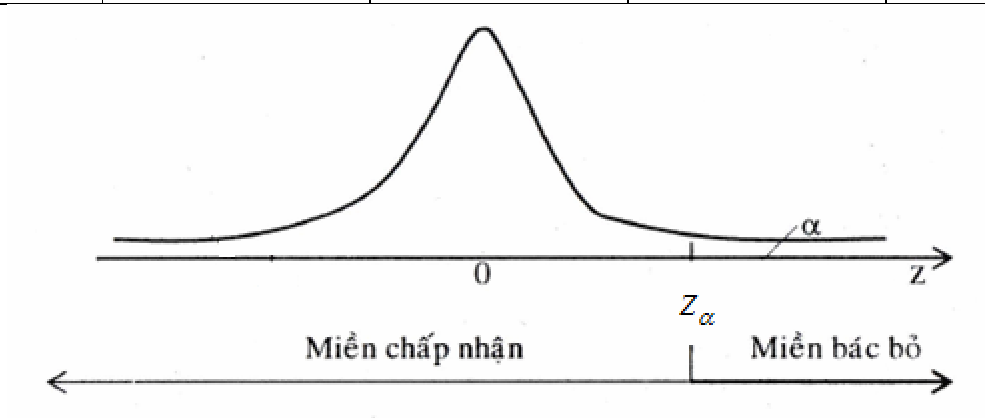
Hình 6.18: Kiểm định Z hai phía mức $\alpha/2$



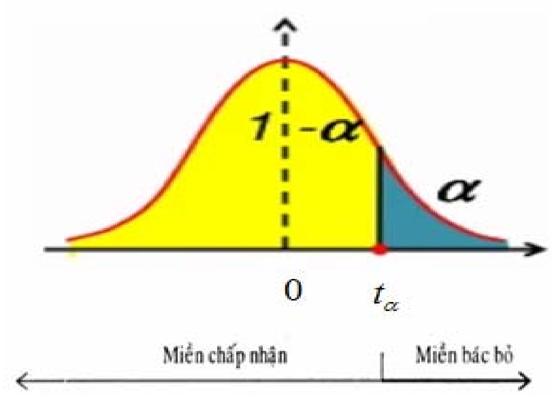
Hình 6.19: Kiểm định t hai phía mức $\alpha/2$

● **Bảng kiểm định 1 phía (one tail):** Giả thuyết $\mathrm{H}_0 : \mu = \mu_0$ và Đối thuyết $H_1 : \mu > \mu_0$

Bước	TH1: $n \geq 30$	TH2: $n \geq 30$	TH3: $n < 30$	TH4: $n < 30$
	σ^2 đã biết	σ^2 chưa biết	σ^2 đã biết	σ^2 chưa biết
1.Tính z hoặc t	$z = \dfrac{\overline{x} - \mu_0}{\sigma} \cdot \sqrt{n}$	$z = \dfrac{\overline{x} - \mu_0}{s} \cdot \sqrt{n}$	$z = \dfrac{\overline{x} - \mu_0}{\sigma} \cdot \sqrt{n}$	$t = \dfrac{\overline{x} - \mu_0}{s} \cdot \sqrt{n}$
2.Tra bảng	z_α phụ lục 7	z_α phụ lục 7	z_α phụ lục 7	t_α^{n-1} phụ lục 9
3a. Chấp nhận H_0	$z \leq z_\alpha$	$z \leq z_\alpha$	$z \leq z_\alpha$	$t \leq t_\alpha^{n-1}$
3a. Bác bỏ H_0	$z > z_\alpha$	$z > z_\alpha$	$z > z_\alpha$	$t > t_\alpha^{n-1}$



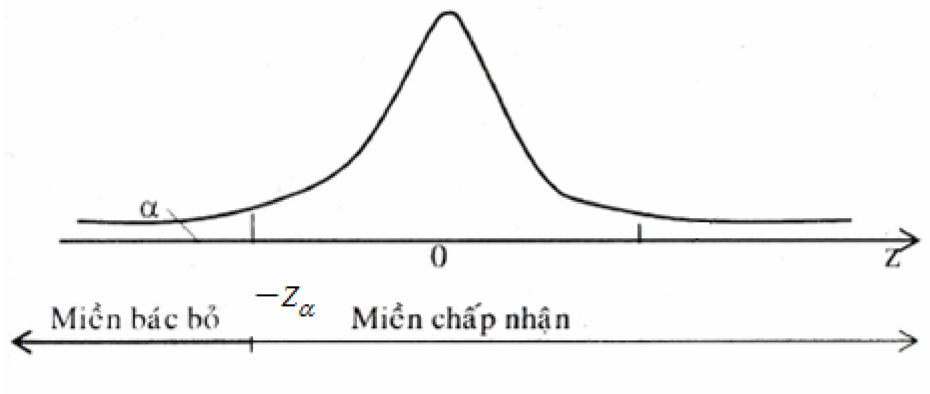
Hình 6.20: Kiểm định Z một phía-bên phải mức α



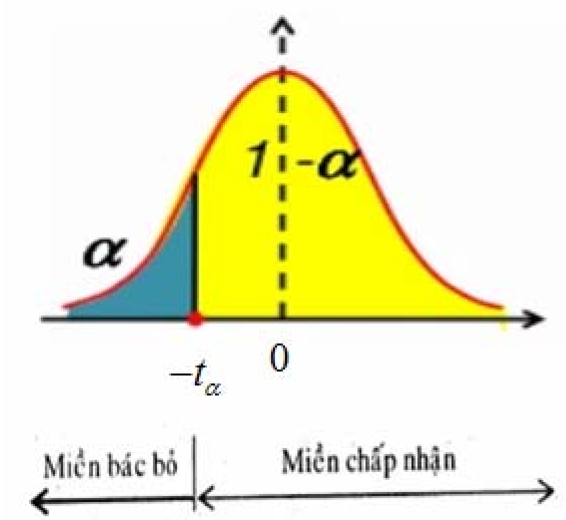
Hình 6.21: Kiểm định t một phía-bên phải mức α

● **Bảng kiểm định 1 phía (one tail):** Giả thuyết $\mathrm{H}_0 : \mu = \mu_0$ và Đối thuyết $H_1 : \mu < \mu_0$

Bước	TH1: $n \geq 30$	TH2: $n \geq 30$	TH3: $n < 30$	TH4: $n < 30$
	σ^2 đã biết	σ^2 chưa biết	σ^2 đã biết	σ^2 chưa biết
1.Tính z hoặc t	$z = \frac{\overline{x} - \mu_0}{\sigma} \cdot \sqrt{n}$	$z = \frac{\overline{x} - \mu_0}{s} \cdot \sqrt{n}$	$z = \frac{\overline{x} - \mu_0}{\sigma} \cdot \sqrt{n}$	$t = \frac{\overline{x} - \mu_0}{s} \cdot \sqrt{n}$
2.Tra bảng	z_α phụ lục 7	z_α phụ lục 7	z_α phụ lục 7	t_α^{n-1} phụ lục 9
3a. Chấp nhận H_0	$z \geq -z_\alpha$	$z \geq -z_\alpha$	$z \geq -z_\alpha$	$t \geq -t_\alpha^{n-1}$
3a. Bác bỏ H_0	$z < -z_\alpha$	$z < -z_\alpha$	$z < -z_\alpha$	$t < -t_\alpha^{n-1}$



Hình 6.22: Kiểm định Z một phía-bên trái mức α



Hình 6.23: Kiểm định t một phía-bên trái mức α

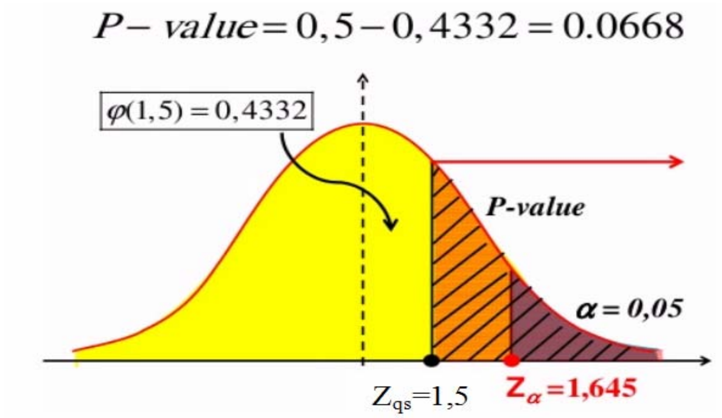
● **Dựa vào giá trị p-value**

a. Trường hợp: σ^2 đã biết hoặc σ^2 chưa biết, $n \geq 30$

Nếu miền bác bỏ bên phải: P_value= $P(Z > z_{qs}) = 0,5 - \varphi(z_{qs})$

Ví dụ 6.8

Nếu giá trị kiểm định z_{qs} =1,5, ta tính P_value= $P(Z > 1,5) = 0,5 - \varphi(1,5)$ như hình



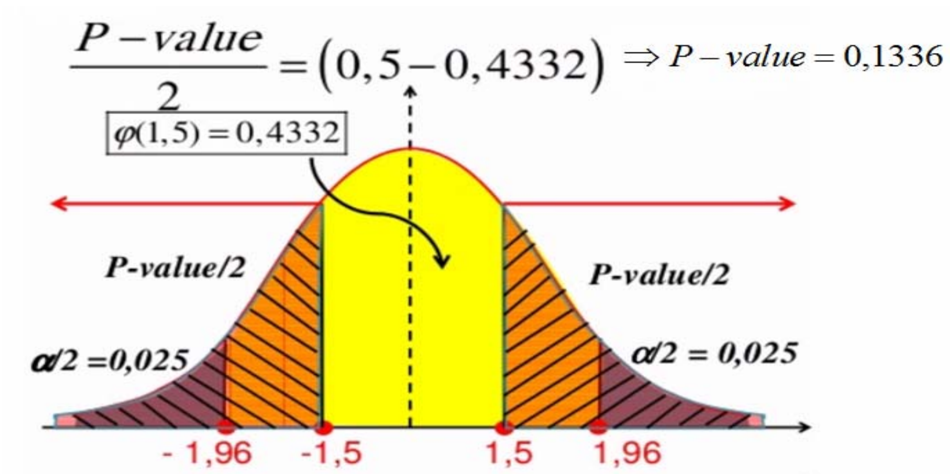
Hình 6.24 : P_value= $P(Z > 1,5)$

Nếu miền bác bỏ bên trái: P_value= $P(Z < z_{qs}) = 0,5 + \varphi(z_{qs})$

Nếu miền bác bỏ hai phía: P_value= $2P(Z > |z_{qs}|) = 1 - 2\varphi(z_{qs})$

Ví dụ 6.9 (tiếp theo Ví dụ 6.8)

Nếu giá trị kiểm định z_{qs} =1,5, ta tính P_value= $2P(Z > |z_{qs}|) = 1 - 2\varphi(z_{qs})$ như hình



Hình 6.25: P_value= $2P(Z > 1,5)$

b. Trường hợp : σ^2 chưa biết, $n < 30$ và X có pp chuẩn

Nếu miền bác bỏ bên phải: P_value= $P(T > t_{qs})$ tra bảng phụ lục 9.

Nếu miền bác bỏ bên trái: P_value= $P(T < t_{qs})$ tra bảng phụ lục 9.

Nếu miền bác bỏ hai phía: P_value= $2P(T > |t_{qs}|)$ tra bảng phụ lục 9.

Ví dụ 6.10

Một nhà máy thép sản xuất sợi thép có khả năng chịu lực X tuân theo phân phối chuẩn với $\mu_0 = 1000$ (đơn vị N=Newton) và $\sigma = 21$. Một quy trình sản xuất mới được đưa vào thử nghiệm, người ta sản xuất thử nghiệm ra mẫu gồm 50 sợi thép thì thấy khả năng chịu lực trung bình của sợi thép là 1010N, dùng nó kiểm định xem liệu khả năng chịu tải của sợi thép có thay đổi không với mức ý nghĩa $\alpha = 5\%$.

Giải

Gọi X là khả năng chịu lực (N) của sợi thép do nhà máy sản xuất.

Trung bình quy định $\mu_0 = 1000$

Trung bình thực tế khả năng chịu lực (N) của sợi thép do nhà máy sản xuất là μ chưa biết.

Đưa giả thuyết $H_0 : \mu = 1000$

Đối thuyết $H_1 : \mu \neq 1000$

Ta thấy n=50>30, X có phân phối chuẩn, σ^2 đã biết nên thuộc trường hợp 1.

+ Tính kiểm định: $z = \dfrac{\overline{x} - \mu_0}{\sigma} \cdot \sqrt{n} = \dfrac{1010 - 1000}{21} \cdot \sqrt{50} = 3,36$

+Tìm $z_{\alpha/2}$: $\alpha = 5\% \Rightarrow z_{\alpha/2} = 1,96$ tra bảng phụ lục 6.

So sánh: |z|=3,36 > 1,96. Vậy ta chấp nhận đối thiết H_1, xem $\mu \neq 1000$.

Ta kết luận khả năng chịu tải của sợi thép do qui trình mới sản xuất có thay đổi.

Lệnh trong R

> library(BSDA)

> zsum.test(mean.x=1010, n.x=50, sigma.x=21, mu=1000, alt="t", conf.level=0.95)

One-sample z-Test

data: Summarized x

z = 3.3672, p-value = 0.0007594 # kiểm định z và giá trị p_value

alternative hypothesis: true mean is not equal to 1000

95 percent confidence interval:

1004.179 1015.821

sample estimates: mean of x 1010

*** Tính P_value bằng lệnh pnorm(q, mean, sd)**

> (1-pnorm(3.3672, 0, 1))*2

[1] 0.000759356

*** Tính P_value theo công thức kiểm định hai phía**

P_value= $2.P(Z > |z_{qs}|) = 2.P(Z > 3,3672) = 1 - 2.\varphi(3,3672)$

$= 1 - 2 \times 0,49961 = 0,00078$

Nhận xét: p-value = 0.0007594 $< \alpha = 5\%$ nên đối thiết H_1 có ý nghĩa thống kê.

Chú ý: alt="t": two.sided, alt="g": greater, alt="l": less

Ví dụ 6.11 (tiếp VD 6.10)

Kiểm định xem liệu khả năng chịu tải của sợi thép do quy trình sản xuất mới có hiệu quả không với mức ý nghĩa $\alpha = 5\%$.

Giải

Đưa giả thuyết $H_0 : \mu = 1000$

Đối thuyết $H_1 : \mu > 1000$

+ Tính kiểm định: $z = \dfrac{\overline{x} - \mu_0}{\sigma} \cdot \sqrt{n} = \dfrac{1010 - 1000}{21} \cdot \sqrt{50} = 3,36$

+Tìm z_α : $\alpha = 5\% \Rightarrow z_\alpha = 1,645$ tra bảng phụ lục 7.

So sánh: z=3,36 > 1,645. Vậy ta chấp nhận đối thiết H_1, xem $\mu > 1000$.

Ta kết luận khả năng chịu tải của sợi thép do qui trình mới sản xuất có hiệu quả.

Lệnh trong R

```
> library(BSDA)
> zsum.test(mean.x=1010, n.x=50, sigma.x=21, mu=1000, alt="g", conf.level=0.95)
        One-sample z-Test
data:  Summarized x
z = 3.3672, p-value = 0.0003797 # kiểm định z và giá trị p_value
alternative hypothesis: true mean is greater than 1000
95 percent confidence interval:
 1005.115       NA
sample estimates: mean of x      1010
```

*** Tính P_value bằng lệnh pnorm(q, mean, sd)**

>1-pnorm(3.3672, 0, 1)

[1] 0.000379678

*** Tính P_value theo công thức kiểm định một phía**

P_value=$P(Z > |z_{qs}|) = P(Z > 3,3672) = 0,5 - \varphi(3,3672) = 0,5 - 0,49961 = 0,00039$

Nhận xét: p-value = 0.0003797 <$\alpha = 5\%$ nên đối thiết H_1 có ý nghĩa thống kê.

Ví dụ 6.12

Giám đốc xí nghiệp cho biết lương trung bình của 1 công nhân trong xí nghiệp là 3,8 triệu đồng/tháng. Chọn ngẫu nhiên 36 công nhân thấy lương trung bình là 3,5 triệu đồng/tháng và độ lệch chuẩn $\sigma = 400$ ngàn đồng. Lời báo cáo của giám đốc có tin cậy không với mức ý nghĩa $\alpha = 5\%$. Biết lương của các công nhân trong xí nghiệp tuân theo phân phối chuẩn.

Giải

Tiền lương trung bình của 1 công nhân theo lời giám đốc: $\mu_0 = 3,8$.

Tiền lương trung bình tực tế của 1 công nhân là μ chưa biết.

Đưa ra giả thuyết : $H_0 : \mu = 3,8$

Đối thuyết $H_1 : \mu \neq 3,8$

Kiểm tra giả thuyết:

$\sigma = 0,4$ đã biết , vậy kiểm tra giả thuyết là trường hợp 1.

+ Tính kiểm định: $z = \dfrac{\overline{x} - \mu_0}{\sigma} \cdot \sqrt{n} = \dfrac{3,5 - 3,8}{0,4} \cdot \sqrt{36} = -4,5$

+ Tìm $z_{\alpha/2}$: $\alpha = 5\% \Rightarrow z_{\alpha/2} = 1,96$ tra bảng phụ lục 6.

So sánh: |z|=4,5 > 1,96. Vậy ta chấp nhận đối thuyết H_1, xem $\mu \neq 3,8$.

Ta kết luận lời báo cáo của giám đốc không đáng tin.

Lệnh trong R

> library(BSDA)

> zsum.test(mean.x=3.5, n.x=36, sigma.x=0.4, mu=3.8, alt="t", conf.level=0.95)

One-sample z-Test

data: Summarized x

z = -4.5, p-value = 6.795e-06 # kiểm định z và giá trị p_value

alternative hypothesis: true mean is not equal to 3.8

95 percent confidence interval:

3.369336 3.630664

sample estimates: mean of x 3.5

*** Tính P_value bằng lệnh pnorm(q, mean, sd)**

> (1-pnorm(4.5, 0, 1))*2

[1] 6.795346e-06

*** Tính P_value theo công thức kiểm định hai phía**

P_value= $2.P(Z >| z_{qs} |) = 2.P(Z > 4,5) = 1 - 2.\varphi(4,5) = 1 - 2 \times 0,499999 = 2.10^{-6}$

Nhận xét: p-value = 6.795e-06 < $\alpha = 5\%$ nên đối thuyết H_1 có ý nghĩa thống kê.

Ví dụ 6.13

Trọng lượng trung bình của một loại sản phẩm theo quy định là 6 kg. Sau một thời gian sản xuất, kiểm tra 121 sản phẩm ta có $\overline{x} = 5,9\,kg$ và $s^2 = 5,76$. Giả thiết trọng lượng của các sản phẩm có phân phối chuẩn. Cho kết luận về tình hình sản xuất có bình thường, đúng quy định không với mức ý nghĩa $\alpha = 5\%$. Biết trọng lượng của các sản phẩm tuân theo phân phối chuẩn.

Giải

Trung bình quy định $\mu_0 = 6\,kg$.

Trung bình thực tế sản xuất là μ chưa biết.

Đưa ra giả thuyết: $H_0 : \mu = 6$ Đối thuyết $H_1 : \mu \neq 6$

Kiểm tra giả thuyết:

σ^2 chưa biết, n=121>30, vậy kiểm tra giả thiết là trường hợp 2.

+ Tính kiểm định: $z = \dfrac{\overline{x} - \mu_0}{s} \cdot \sqrt{n} = \dfrac{5,9 - 6}{\sqrt{5,76}} \cdot \sqrt{121} = -0,45$

+Tìm $z_{\alpha/2}$: $\alpha = 5\% \Rightarrow z_{\alpha/2} = 1,96$ tra bảng phụ lục 6.

So sánh: $| z |= 0,45 < z_{\alpha/2} = 1,96$. Vậy ta chấp nhận giả thuyết $H_0 : \mu = 6$.

Ta kết luận tình hình sản xuất bình thường, đúng quy định.

Lệnh trong R

> library(BSDA)

> zsum.test(mean.x=5.9, n.x=121, sigma.x=2.4, mu=6, alt="t", conf.level=0.95)

One-sample z-Test

data: Summarized x

z = -0.45833, p-value = 0.6467 # kiểm định z và giá trị p_value

alternative hypothesis: true mean is not equal to 6

95 percent confidence interval:

5.472371 6.327629

sample estimates: mean of x 5.9

*** Tính P_value bằng lệnh pnorm(q, mean, sd)**

>(1-pnorm(0.45833, 0, 1))*2

[1] 0.6467154

*** Tính P_value theo công thức kiểm định hai phía**

P_value= $2.P(Z > |z_{qs}|) = 2.P(Z > 0,45) = 1 - 2.\varphi(0,45) = 1 - 2 \times 0,17364 = 0,65272$

Nhận xét: p-value = 0.6467 $> \alpha = 5\%$ nên chấp nhận giả thuyết H_0.

Ví dụ 6.14 (tiếp Ví dụ 6.13)

Có báo cáo cho rằng tình hình sản xuất không bình thường, trọng lượng trung bình của sản phẩm thấp hơn 6 kg theo quy định. Hãy kiểm định lời báo cáo đó với mức ý nghĩa $\alpha = 5\%$.

Giải

Đưa ra giả thuyết: $H_0 : \mu = 6$

Đối thuyết $H_1 : \mu < 6$

+ Tính kiểm định: $z = \frac{\overline{x} - \mu_0}{s} \cdot \sqrt{n} = \frac{5,9 - 6}{\sqrt{5,76}} \cdot \sqrt{121} = -0,45$

+Tìm z_α : $\alpha = 5\% \Rightarrow z_\alpha = 1,645$ tra bảng phụ lục 7.

So sánh: $z = -0,45 > -z_\alpha = -1,645$. Vậy ta chấp nhận giả thuyết $H_0 : \mu = 6$.

Ta kết luận lời báo cáo đó chưa đủ cơ sở để kết luận trọng lượng trung bình của sản phẩm thấp hơn 6 kg theo quy định.

Các lệnh trong R

> library(BSDA)

> zsum.test(mean.x=5.9, n.x=121, sigma.x=2.4, mu=6, alt="l", conf.level=0.95)

One-sample z-Test

data: Summarized x

z = -0.45833, p-value = 0.3234 # kiểm định z và giá trị p_value

alternative hypothesis: true mean is less than 6

95 percent confidence interval:

NA 6.258877

sample estimates: mean of x 5.9

*** Tính P_value bằng lệnh pnorm(q, mean, sd)**

> (1-pnorm(0.45833, 0, 1))

[1] 0.3233577

*** Tính P_value theo công thức kiểm định một phía**

P_value= $P(Z < z_{qs}) = P(Z < -0,45) = 0,5 + \varphi(-0,45) = 0,5 - 0,17361 = 0,32639$

Nhận xét: p-value = 0.3234 > $\alpha = 5\%$ nên chấp nhận giả thuyết H_0.

Ví dụ 6.15

Trước đây máy móc của một công ty sản xuất ra các vòng đệm (thiết bị cơ khí nhỏ như camera số, tablet,…) có bề dày trung bình là 0,05 cm. Để xác định liệu các máy này vẫn còn hoạt động bình thường hay không, người ta chọn một mẫu gồm 10 vòng đệm và tính được bề dày trung bình là 0.053 cm và độ lệch chuẩn mẫu 0,003 cm. Với mức ý nghĩa $\alpha = 5\%$, hãy kiểm định giả thuyết máy này có hoạt động bình thường không? Biết bề dày của các vòng đệm tuân theo phân phối chuẩn.

Giải

Gọi X là bề dày vòng đệm (bearing) do máy móc công ty sản xuất.

Trung bình quy định $\mu_0 = 0.05$

Trung bình thực tế sản xuất là μ chưa biết.

Đưa giả thuyết $H_0 : \mu = 0,05$

Đối thuyết $H_1 : \mu \neq 0,05$

Ta thấy n=10<30, X có phân phối chuẩn, σ^2 chưa biết nên thuộc trường hợp 3.

+Tính kiểm định $t = \dfrac{\overline{x} - \mu_0}{s} \cdot \sqrt{n} = \dfrac{0,053 - 0,05}{0,003} \cdot \sqrt{10} = 3,1623$

+Tìm $t_{\alpha/2}^{n-1}$: $\alpha = 5\% \Rightarrow t_{\alpha/2}^{9} = 2,2622$ tra bảng phụ lục 8.

So sánh: |t| = 3,1623 > $t_{\alpha/2}^{9} = 2,2622$ nên ta bác bỏ giả thuyết $H_0 : \mu = 0,05$.

Kết luận tình hình hoạt động của máy không bình thường.

Các lệnh trong R

> library(BSDA)

> tsum.test(mean.x=0.053, s.x = 0.003, n.x = 10,mu=0.05,alternative = "two.sided", conf.level=0.95)

One-sample t-Test

data: Summarized x

t = 3.1623, df = 9, p-value = 0.01151 # kiểm định t, bậc tự do là 9 và giá trị p_value

alternative hypothesis: true mean is not equal to 0.05

95 percent confidence interval: 0.05085393 0.05514607

sample estimates: mean of x 0.053

*** Tính P_value bằng lệnh pt(q, df)**

> 1-(1-((1 - pt(3.1623, 9))*2))

[1] 0.01150757

***Tính P_value theo công thức kiểm định 2 phía bên phải**

P_value= $2.P(T > |t_{qs}|) = 2.P(T > 3{,}1623)$

Tra bảng phụ lục 9, ta có $P(T(9) > 2{,}821) = 0{,}01$ và $P(T(9) > 3{,}25) = 0{,}005$

Suy ra 0,005 x 2=0,01< P_value <0,01 x 2=0,02

Nhận xét: p-value =0.01151< $\alpha = 5\%$ nên đối thuyết H_1 có ý nghĩa thống kê.

Ví dụ 6.16

Năm trước tiền lương trung bình của các cử nhân quản trị kinh doanh làm việc tại các công ty liên doanh nước ngoài là 210 USD/tháng. Năm nay điều tra ngẫu nhiên lương tháng của 25 cử nhân đang làm việc cho các công ty đó tìm được $\overline{x} = 218$ USD/tháng, độ lệch chuẩn $s = 10$.

a. Nếu giả thuyết tiền lương là biến ngẫu nhiên phân phối chuẩn thì với mức ý nghĩa $\alpha = 5\%$ có thể cho rằng năm nay các nhân viên đó có hưởng mức lương cao hơn hay không?

b. Hãy tìm P_value để kết luận vấn đề trên.

Giải

a. Tiền lương trung bình của các cử nhân quản trị kinh doanh năm trước là $\mu_0 = 210$.

Tiền lương trung bình của các cử nhân quản trị kinh doanh năm nay là μ chưa biết.

Đưa giả thuyết $H_0 : \mu = 210$

Đối thuyết $H_1 : \mu > 210$

Ta thấy n=25 < 30, X có phân phối chuẩn, σ^2 chưa biết nên thuộc trường hợp 4.

+ Tính kiểm định $t = \dfrac{\overline{x} - \mu_0}{s} \cdot \sqrt{n} = \dfrac{218 - 210}{10} \cdot \sqrt{25} = 4$

+ Tìm t_α^{n-1}: $\alpha = 5\% \Rightarrow t_\alpha^{24} = 1{,}7109$ tra bảng phụ lục 9.

So sánh: t =4 > $t_\alpha^{24} = 1{,}7109$ nên ta chấp nhận đối thuyết $H_1 : \mu > 210$.

Kết luận: năm nay các nhân viên đó có hưởng mức lương cao hơn năm trước.

b. Tính P_value theo công thức kiểm định 1 phía bên phải

P_value= $P(T > t_{qs}) = P(T > 4)$

Tra bảng phụ lục 9, ta có $P(T(24) > 2{,}797) = 0{,}005$.

Nên P_value = $P(T > 4)$ < 0,01 suy ra P_value < $\alpha = 5\%$. Vậy đối thuyết H_1 có ý nghĩa thống kê.

*** Tính P_value bằng lệnh pt(q, df)**

> 1-(1-(1 - pt(4, 24)))

[1] 0.000263454

Các lệnh trong R

> tsum.test(mean.x=218, s.x = 10, n.x = 25,mu=210,alternative = "greater", conf.level=0.95)

One-sample t-Test

data: Summarized x

t = 4, df = 24, p-value = 0.0002635 # kiểm định t, bậc tự do là 24 và giá trị p_value

alternative hypothesis: true mean is greater than 210

95 percent confidence interval: 214.5782 NA

sample estimates: mean of x 218

Nhận xét: p-value = 0.0002635 $<\alpha = 5\%$ nên đối thuyết H_1 có ý nghĩa thống kê.

6.3.2. Kiểm định giả thuyết về tỷ lệ tổng thể ($n \geq 30$)

Bài Toán 6.2

Giả sử tổng thể chia 2 loại phần tử. Tỷ lệ phần tử có tính chất A là p chưa biết. Ta cần kiểm tra giả thuyết $H_0 : p = p_0$.

Hãy đưa ra quy tắc căn cứ vào mẫu gồm n quan sát độc lập ($x_1,...,x_n$) mà chấp nhận hay bác bỏ giả thuyết trên với mức ý nghĩa α.

Quy tắc thực hành 6.2

Bước	$H_0 : p = p_0$ $H_1 : p \neq p_0$	$H_0 : p = p_0$ $H_1 : p < p_0$	$H_0 : p = p_0$ $H_1 : p > p_0$
1. Tính z	$z = \frac{f - p_0}{\sqrt{p_0(1-p_0)}} \cdot \sqrt{n}$	$z = \frac{f - p_0}{\sqrt{p_0(1-p_0)}} \cdot \sqrt{n}$	$z = \frac{f - p_0}{\sqrt{p_0(1-p_0)}} \cdot \sqrt{n}$
2. Tra bảng	$z_{\alpha/2} : \alpha \Rightarrow z_{\alpha/2}$ tra bảng phụ lục 6.	$z_\alpha : \alpha \Rightarrow z_\alpha$ tra bảng phụ lục 7.	$z_\alpha : \alpha \Rightarrow z_\alpha$ tra bảng phụ lục 7.
3a. Chấp nhận H_0	$\lvert z \rvert \leq z_{\alpha/2}$	$z \geq -z_\alpha$	$z \leq z_\alpha$
3b. Bác bỏ H_0	$\lvert z \rvert > z_{\alpha/2}$	$z < -z_\alpha$	$z > z_\alpha$

Ví dụ 6.17

Một vùng biển bị tràn dầu với tỷ lệ sinh vật biển nhiễm dầu sống sót lúc đầu là $p_0 = 0{,}155$. Sau khi áp dụng biện pháp hút dầu, kiểm tra 1000 sinh vật biển ta thấy tỷ lệ sinh vật biển nhiễm dầu sống sót trong mẫu là f=0,2. Cho kết luận về biện pháp hút dầu này với $\alpha = 5\%$.

Giải

Tỷ lệ sinh vật biển nhiễm dầu sống sót lúc đầu là $p_0 = 0{,}155$;

Tỷ lệ sinh vật biển nhiễm dầu sống sót sau khi áp dụng biện pháp hút dầu là p chưa biết.

Ta đặt giả thuyết $H_0 : p = 0{,}155$

Đối thuyết $H_1 : p \neq 0{,}155$

- Kiểm tra giả thuyết:

+ Tính kiểm định $z = \dfrac{f - p_0}{\sqrt{p_0(1-p_0)}} \cdot \sqrt{n} = \dfrac{0,2 - 0,155}{\sqrt{0,155(1-0,155)}}\sqrt{1000} = 3,932$

+Tìm $z_{\alpha/2}$: $\alpha = 5\% \Rightarrow z_{\alpha/2} = 1,96$ tra bảng phụ lục 6.

So sánh: |z| = 3,932 > $z_{\alpha/2} = 1,96$ nên bác bỏ giả thuyết H_0, xem $p \neq p_0$.

Vậy biện pháp hút dầu làm thay đổi tỷ lệ sinh vật biển nhiễm dầu sống sót.

Chú ý: Kiểm định tỷ lệ 2 phía chỉ có thể kết luận đơn giản: bằng hay khác, thay đổi hay không thay đổi, muốn kết luận mạnh hơn: hiệu quả hay không hiệu quả, chúng ta phải chuyển qua kiểm định tỷ lệ 1 phía sẽ trình bày tiếp theo.

Các lệnh trong R

> library(BSDA)

> prop.test(200, 1000, 0.155)

1-sample proportions test with continuity correction

data: 200 out of 1000, null probability 0.155

X-squared = 15.119, df = 1, p-value = 0.0001009 # kiểm định Chi-bình phương, bậc tự do là 1 và giá trị p_value

alternative hypothesis: true p is not equal to 0.155

95 percent confidence interval: 0.1759021 0.2264401

sample estimates: p 0.2

Nhận xét: p-value = 0,0001009 < $\alpha = 5\%$ nên đối thiết H_1 có ý nghĩa thống kê.

Ở đây hàm prop.test (x, n, p_0) sử dụng Z^2 với $Z = \dfrac{(F-p)\sqrt{n}}{\sqrt{F(1-F)}} \sim N(0,1)$; ta có thể nói Z^2 tuân theo phân phối Chi-bình phương với bậc tự do bằng 1.

*** Tính P_value theo công thức kiểm định hai phía**

P_value $= 2P(Z > |z_{qs}|) = 2P(Z > 3,932) = 1 - 2\varphi(3,932)$

$= 1 - 2 \times 0,4999575 = 0,000085$

Ví dụ 6.18

Một trường trung học có tỷ lệ học sinh giỏi lúc đầu là $p_0 = 0,19$. Sau khi áp dụng phương pháp dạy học mới, kiểm tra 1000 học sinh thấy tỷ lệ học sinh giỏi trong mẫu là f=0,2. Có thể xem phương pháp dạy học mới này có hiệu quả không với $\alpha = 5\%$.

Giải

Tỷ lệ học sinh giỏi lúc đầu là $p_0 = 0,19$.

Tỷ lệ học sinh giỏi sau khi áp dụng phương pháp dạy học mới là p chưa biết.

Ta đặt giả thuyết $H_0 : p = 0,19$

Đối thuyết $H_1 : p > 0,19$

- Kiểm tra giả thuyết:

+ Tính kiểm định $z = \frac{f - p_0}{\sqrt{p_0(1-p_0)}} \cdot \sqrt{n} = \frac{0,2 - 0,19}{\sqrt{0,19(1-0,19)}}\sqrt{1000} = 0,8$

+ Tìm z_α : $\alpha = 5\% \Rightarrow z_\alpha = 1,645$ tra bảng phụ lục 7.

So sánh: z = 0,8 < $z_\alpha = 1,645$ nên ta chấp nhận giả thuyết $H_0 : p = 0,19$.

Vậy phương pháp dạy học mới này chưa có hiệu quả.

Lệnh trong R

> library(BSDA)

> prop.test(200, 1000, p = 0.19, alternative = c("greater"), conf.level = 0.95, correct = TRUE)

1-sample proportions test with continuity correction

data: 200 out of 1000, null probability 0.19

X-squared = 0.58642, df = 1, p-value = 0.2219 # kiểm định Chi-bình phương, bậc tự do là 1 và giá trị p_value

alternative hypothesis: true p is greater than 0.19

95 percent confidence interval: 0.1795367 1.0000000

sample estimates: p 0.2

Nhận xét: p-value = 0,2219 > $\alpha = 5\%$ nên đối thiết H_1 không có ý nghĩa thống kê.

Ví dụ 6.19

Một luật sư khu vực hạt muốn chuyển đến văn phòng luật sư cấp tiểu bang. Cô đã quyết định rằng cô sẽ từ bỏ văn phòng của mình tại hạt và điều hành văn phòng cấp tiểu bang nếu hơn 65% cử tri tiểu bang ủng hộ cô. Bạn cần phải kiểm tra H_0: p = 0,65 so với H_1: p> 0,65. Là người quản lý chiến dịch của mình, bạn đã thu thập dữ liệu trên 950 cử tri tiểu bang được chọn ngẫu nhiên và thấy rằng 560 cử tri ủng hộ ứng cử viên. Một kiểm định tỷ lệ sẽ được thực hiện để xác định xem tỷ lệ của người ủng hộ có lớn hơn tỷ lệ yêu cầu là 0,65 với 95% sự tự tin bị ràng buộc.

Giải

Lệnh trong R

> library(BSDA)

> prop.test(560, 950, p = 0.65, alternative = c("greater"), conf.level = 0.95, correct = TRUE)

1-sample proportions test with continuity correction

data: 560 out of 950, null probability 0.65

X-squared = 15.033, df = 1, p-value = 0.9999 # kiểm định Chi-bình phương, bậc tự do là 1 và giá trị p_value

alternative hypothesis: true p is greater than 0.65

95 percent confidence interval:

0.5624734 1.0000000

sample estimates: p 0.5894737

Nhận xét: p-value = 0,9999 > $\alpha = 5\%$ nên đối thiết H_1 không có ý nghĩa thống kê.

Giải thích kết quả

Giá trị p là 0,9999 cho thấy rằng dữ liệu phù hợp với giả thuyết không (H_0: p = 0.65), nghĩa là tỷ lệ cử tri tiểu bang ủng hộ ứng cử viên không lớn hơn tỷ lệ yêu cầu là 0,65. Với tư cách là người quản lý chiến dịch của mình, bạn sẽ khuyên cô đừng chạy vào văn phòng của luật sư tiểu bang.

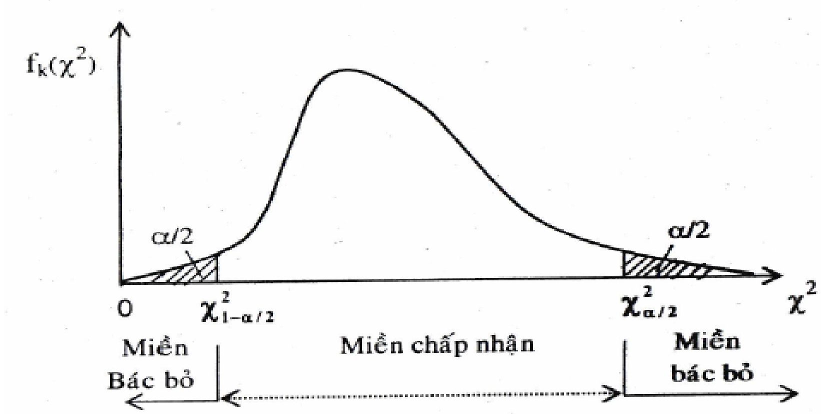
6.3.3. Kiểm định giả thiết về phương sai ($n < 30$)

Bài toán 6.3

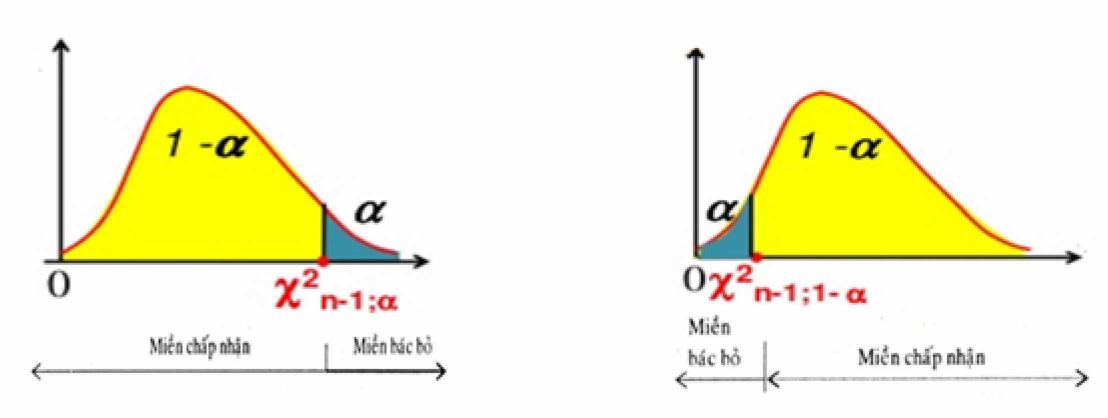
Giả sử tổng thể X có phân phối chuẩn có phương sai σ^2 chưa biết. Ta cần kiểm định giả thuyết $H_0 : \sigma^2 = \sigma_0^2$. Hãy đưa quy tắc căn cứ vào mẫu gồm n quan sát độc lập ($x_1,...,x_n$) mà chấp nhận hay bác bỏ giả thuyết trên với mức ý nghĩa α.

Quy tắc thực hành 6.3

Bước	$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 \neq \sigma_0^2$	$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 < \sigma_0^2$	$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 > \sigma_0^2$
1. Tính z	$\chi_{qs}^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$	$\chi_{qs}^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$	$\chi_{qs}^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$
2. Tra bảng	$\chi_{\alpha/2}^2(n-1)$ và $\chi_{1-\alpha/2}^2(n-1)$ phụ lục 5	$\chi_{1-\alpha}^2(n-1)$ phụ lục 5	$\chi_{\alpha}^2(n-1)$ phụ lục 5
3a. Chấp nhận H_0	$\chi_{1-\alpha/2}^2(n-1) \le \chi_{qs}^2 \le \chi_{\alpha/2}^2(n-1)$	$\chi_{qs}^2 \ge \chi_{1-\alpha}^2(n-1)$	$\chi_{qs}^2 \le \chi_{\alpha}^2(n-1)$
3b. Bác bỏ H_0	$\chi_{qs}^2 < \chi_{1-\alpha/2}^2(n-1)$ hoặc $\chi_{\alpha/2}^2(n-1) < \chi_{qs}^2$	$\chi_{qs}^2 < \chi_{1-\alpha}^2(n-1)$	$\chi_{qs}^2 > \chi_{\alpha}^2(n-1)$



Hình 6.26: Kiểm định Chi-bình phương hai phía mức $\alpha/2$



Hình 6.27: Kiểm định Chi-bình phương một phía mức α

Ví dụ 6.20

Tiến hành 25 quan sát về chỉ tiêu X của 1 loại sản phẩm có phân phối chuẩn, tính được $s^2 = 416,667$. Có tài liệu nói rằng phương sai của chỉ tiêu X là 400. Với mức ý nghĩa 2%, hãy cho nhận xét về tài liệu này.

Giải

Phương sai của chỉ tiêu X theo tài liệu là $\sigma_0^2 = 400$

Phương sai của chỉ tiêu X thực tế là σ^2 chưa biết.

Ta đặt giả thuyết $H_0 : \sigma^2 = \sigma_0^2 = 400$

Đối thuyết $H_1 : \sigma^2 \neq 400$

- Kiểm tra giả thiết:

+ Tính $\chi_{qs}^2 = \dfrac{(n-1)s^2}{\sigma_0^2} = \dfrac{24 \times 416,667}{400} = 25$

+ Tìm $\alpha = 2\% \Rightarrow \alpha/2 = 0,01 \Rightarrow \chi_1^2 = \chi_{0,99}^2(24) = 10,856$; $\chi_2^2 = \chi_{0,01}^2(24) = 42,98$ tra bảng phụ lục 5.

So sánh: $\chi_1^2 \leq \chi_{qs}^2 \leq \chi_2^2$: chấp nhận giả thuyết H_0.

Vậy tài liệu đáng tin cậy.

Ví dụ 6.21 (tiếp Ví dụ 6.20)

Có thể xem phương sai của chỉ tiêu X lớn hơn phương sai của chỉ tiêu X trong tài liệu không, với mức ý nghĩa 5%.

Giải

Ta đặt giả thuyết $H_0 : \sigma^2 = \sigma_0^2 = 400$

Đối thuyết $H_1 : \sigma^2 > 400$

- Kiểm tra giả thuyết:

+ Tính $\chi_{qs}^2 = \dfrac{(n-1)s^2}{\sigma_0^2} = \dfrac{24 \times 416,667}{400} = 25$

+ Tìm $\alpha = 5\% \Rightarrow \chi^2_{0,05}(24) = 36,415$ tra bảng phụ lục 5.

So sánh: $\chi^2_{qs} = 25 < \chi^2_{0,05}(24) = 36,415$, ta chấp nhận giả thuyết H_0.

Kết luận chưa có cơ sở kết luận phương sai của chỉ tiêu X lớn hơn phương sai của chỉ tiêu X trong tài liệu

Ví dụ 6.22

Độ rủi ro của một cổ phiếu thường được đo bằng phương sai của lãi suất của cổ phiếu đó. Theo dõi lãi suất cổ phiếu của một công ty trong vòng 5 năm qua thu được kết quả là 15%, 10%, 20%, 7%, 14%. Với mức ý nghĩa 5%, hãy kiểm định phương sai của lãi suất cổ phiếu công ty có bằng 25% không. Biết lãi suất cổ phiếu công ty tuân theo phân phối chuẩn.

Giải

Trong R có lệnh **var.test()** để kiểm định phương sai của hai mẫu có bằng nhau không. Trong khi bài toán của chúng ta chỉ kiểm định phương sai của một mẫu so với giá trị phương sai cho trước có bằng nhau không. Cho nên chúng ta tính thủ công các bước:

Phương sai của lãi suất cổ phiếu công ty thực tế là σ^2 chưa biết.

Phương sai của lãi suất cổ phiếu công ty theo báo cáo là $\sigma_0^2 = 25\%$

Ta đặt giả thuyết $H_0 : \sigma^2 = 25\%$

Đối thuyết $H_1 : \sigma^2 \neq 25\%$

- Kiểm tra giả thiết:

+ Tính $\chi^2_{qs} = \dfrac{(n-1)s^2}{\sigma_0^2} = \dfrac{4 \times 24,7}{25} = 3,952$

Ta có $n = 5;\ \alpha = 0,05 \Rightarrow \dfrac{\alpha}{2} = 0,025;\ 1 - \dfrac{\alpha}{2} = 0,975$

$\Rightarrow \chi^2_{0,975}(4) = 0,4844;\quad \chi^2_{0,025}(4) = 11,1433$ tra bảng phụ lục 5.

So sánh: $\chi^2_{0,975} \leq \chi^2_{qs} \leq \chi^2_{0,025}$: chấp nhận giả thuyết H_0.

Kết luận: phương sai của lãi suất cổ phiếu công ty bằng 25%.

Ví dụ 6.23 (tiếp Ví dụ 6.22)

Với mức ý nghĩa 5%, hãy kiểm định phương sai của lãi suất cổ phiếu công ty có nhỏ hơn 25% không. Biết lãi suất cổ phiếu công ty tuân theo phân phối chuẩn.

Giải

Phương sai của lãi suất cổ phiếu công ty thực tế là σ^2 chưa biết.

Phương sai của lãi suất cổ phiếu công ty theo báo cáo là $\sigma^2 < 25\%$

Ta đặt giả thuyết $H_0 : \sigma^2 \geq 25\%$

Đối thuyết $H_1 : \sigma^2 < 25\%$

- Kiểm tra giả thuyết:

+ Tính $\chi^2_{qs} = \dfrac{(n-1)s^2}{\sigma_0^2} = \dfrac{4\times 24,7}{25} = 3,952$

+ Tìm $\alpha = 5\% \Rightarrow \chi^2_{0,95}(4) = 0,711$ tra bảng phụ lục 5.

So sánh: $\chi^2_{qs} = 3,952 > \chi^2_{0,95}(4) = 0,711$, ta chấp nhận giả thuyết $H_0 : \sigma^2 \geq \sigma_0^2 = 25\%$.

Kết luận: phương sai của lãi suất cổ phiếu công ty nhỏ hơn 25% là chưa đúng.

6.3.4. Kiểm định giả thuyết về sự bằng nhau của hai trung bình

Bài toán 6.4

Giả sử X và Y là hai đại lượng ngẫu nhiên độc lập có cùng phân phối chuẩn với μ_X, μ_Y chưa biết. Ta cần kiểm định giả thuyết $H_0 : \mu_X = \mu_Y$

Lấy mẫu ngẫu nhiên kích thước n_1 đối với X và mẫu ngẫu nhiên kích thước n_2 đối với Y ta có 2 trường hợp sau:

Quy tắc thực hành 6.4

□ **TH 1: σ_X^2, σ_Y^2 đã biết**

Bước	$H_0 : \mu_X = \mu_Y$ $H_1 : \mu_X \neq \mu_Y$	$H_0 : \mu_X = \mu_Y$ $H_1 : \mu_X < \mu_Y$	$H_0 : \mu_X = \mu_Y$ $H_1 : \mu_X > \mu_Y$
1. Tính z	$z = \dfrac{\overline{x}-\overline{y}}{\sqrt{\dfrac{\sigma_X^2}{n_1}+\dfrac{\sigma_Y^2}{n_2}}}$	$z = \dfrac{\overline{x}-\overline{y}}{\sqrt{\dfrac{\sigma_X^2}{n_1}+\dfrac{\sigma_Y^2}{n_2}}}$	$z = \dfrac{\overline{x}-\overline{y}}{\sqrt{\dfrac{\sigma_X^2}{n_1}+\dfrac{\sigma_Y^2}{n_2}}}$
2. Tra bảng	$z_{\alpha/2} : \alpha \Rightarrow z_{\alpha/2}$ tra bảng phụ lục 6.	$z_\alpha : \alpha \Rightarrow z_\alpha$ tra bảng phụ lục 7.	$z_\alpha : \alpha \Rightarrow z_\alpha$ tra bảng phụ lục 7.
3a. Chấp nhận H_0	$\lvert z\rvert \leq z_{\alpha/2}$	$z \geq -z_\alpha$	$z \leq z_\alpha$
3b. Bác bỏ H_0	$\lvert z\rvert > z_{\alpha/2}$	$z < -z_\alpha$	$z > z_\alpha$

□ **TH 2: σ_X^2, σ_Y^2 chưa biết**

● **Bảng kiểm định 2 phía (two tail):** Giả thuyết $H_0 : \mu_X = \mu_Y$ và Đối thuyết $H_1 : \mu_X \neq \mu_Y$

Bước	Trường hợp		
	$n_1, n_2 \geq 30$	$n_1, n_2 < 30$ Biết $\sigma_X^2 = \sigma_Y^2$	$n_1, n_2 < 30$ Chưa biết $\sigma_X^2 = \sigma_Y^2$
1. Tính z	$z = \dfrac{\overline{x}-\overline{y}}{\sqrt{\dfrac{s_x^2}{n_1}+\dfrac{s_y^2}{n_2}}}$	$t = \dfrac{\overline{x}-\overline{y}}{s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ $s_p = \sqrt{\dfrac{(n_1-1)s_x^2+(n_2-1)s_y^2}{n_1+n_2-2}}$	$t = \dfrac{\overline{x}-\overline{y}}{\sqrt{\dfrac{s_x^2}{n_1}+\dfrac{s_y^2}{n_2}}}$

2. Tra bảng	$z_{\alpha/2}:\alpha \Rightarrow z_{\alpha/2}$ phụ lục 6.	$t_{\alpha/2}^{n_1+n_2-2}:\alpha \Rightarrow t_{\alpha/2}^{n_1+n_2-2}$ phụ lục 8	$df=\dfrac{\left(\dfrac{s_x^2}{n_1}+\dfrac{s_y^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_x^2}{n_1}\right)^2}{n_1-1}+\dfrac{\left(\dfrac{s_y^2}{n_2}\right)^2}{n_2-1}}$ $t_{\alpha/2}^{df}:\alpha \Rightarrow t_{\alpha/2}^{df}$ phụ lục 8
3a. Chấp nhận H_0	$\lvert z\rvert \le z_{\alpha/2}$	$\lvert t\rvert \le t_{\alpha/2}^{n_1+n_2-2}$	$\lvert t\rvert \le t_{\alpha/2}^{df}$
3b. Bác bỏ H_0	$\lvert z\rvert > z_{\alpha/2}$	$\lvert t\rvert > t_{\alpha/2}^{n_1+n_2-2}$	$\lvert t\rvert > t_{\alpha/2}^{df}$

● **Bảng kiểm định 1 phía (one tail):** Giả thuyết $H_0:\mu_X=\mu_Y$ và Đối thuyết $H_1:\mu_X<\mu_Y$

Bước	Trường hợp		
	$n_1,n_2\ge 30$	$n_1,n_2<30$ Biết $\sigma_X^2=\sigma_Y^2$	$n_1,n_2<30$ Chưa biết $\sigma_X^2=\sigma_Y^2$
1. Tính z	$z=\dfrac{\overline{x}-\overline{y}}{\sqrt{\dfrac{s_x^2}{n_1}+\dfrac{s_y^2}{n_2}}}$	$t=\dfrac{\overline{x}-\overline{y}}{s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$ $s_p=\sqrt{\dfrac{(n_1-1)s_x^2+(n_2-1)s_y^2}{n_1+n_2-2}}$	$t=\dfrac{\overline{x}-\overline{y}}{\sqrt{\dfrac{s_x^2}{n_1}+\dfrac{s_y^2}{n_2}}}$
2. Tra bảng	$z_\alpha:\alpha \Rightarrow z_\alpha$ phụ lục 7.	$t_\alpha^{n_1+n_2-2}:\alpha \Rightarrow t_\alpha^{n_1+n_2-2}$ phụ lục 9	$df=\dfrac{\left(\dfrac{s_x^2}{n_1}+\dfrac{s_y^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_x^2}{n_1}\right)^2}{n_1-1}+\dfrac{\left(\dfrac{s_y^2}{n_2}\right)^2}{n_2-1}}$ $t_\alpha^{df}:\alpha \Rightarrow t_\alpha^{df}$ phụ lục 9
3a. Chấp nhận H_0	$z\ge -z_\alpha$	$t\ge -t_\alpha^{n_1+n_2-2}$	$t\ge -t_\alpha^{df}$
3b. Bác bỏ H_0	$z<-z_\alpha$	$t<-t_\alpha^{n_1+n_2-2}$	$t<-t_\alpha^{df}$

● **Bảng kiểm định 1 phía (one tail):** Giả thuyết $H_0 : \mu_X = \mu_Y$ và Đối thuyết $H_1 : \mu_X > \mu_Y$

Bước	Trường hợp		
	$n_1, n_2 \geq 30$	$n_1, n_2 < 30$ Biết $\sigma_X^2 = \sigma_Y^2$	$n_1, n_2 < 30$ Chưa biết $\sigma_X^2 = \sigma_Y^2$
1. Tính z	$z = \dfrac{\overline{x} - \overline{y}}{\sqrt{\dfrac{s_x^2}{n_1} + \dfrac{s_y^2}{n_2}}}$	$t = \dfrac{\overline{x} - \overline{y}}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ $s_p = \sqrt{\dfrac{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}{n_1 + n_2 - 2}}$	$t = \dfrac{\overline{x} - \overline{y}}{\sqrt{\dfrac{s_x^2}{n_1} + \dfrac{s_y^2}{n_2}}}$
2. Tra bảng	$z_\alpha : \alpha \Rightarrow z_\alpha$ phụ lục 7.	$t_\alpha^{n_1+n_2-2} : \alpha \Rightarrow t_\alpha^{n_1+n_2-2}$ phụ lục 9	$df = \dfrac{\left(\dfrac{s_x^2}{n_1} + \dfrac{s_y^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_x^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_y^2}{n_2}\right)^2}{n_2 - 1}}$ $t_\alpha^{df} : \alpha \Rightarrow t_\alpha^{df}$ phụ lục 9
3a. Chấp nhận H_0	$z \leq z_\alpha$	$t \leq t_\alpha^{n_1+n_2-2}$	$t \leq t_\alpha^{df}$
3b. Bác bỏ H_0	$z > z_\alpha$	$t > t_\alpha^{n_1+n_2-2}$	$t > t_\alpha^{df}$

Ví dụ 6.24

Trọng lượng sản phẩm do hai nhà máy sản xuất là các đại lượng ngẫu nhiên có phân phối chuẩn và có cùng độ lệch chuẩn là $\sigma = 1$kg. Với mức ý nghĩa $\alpha = 5\%$, có thể xem trọng lượng trung bình của sản phẩm do hai nhà máy sản xuất là như nhau không? Nếu cân thử 25 sản phẩm của nhà máy A ta tính được $\overline{x} = 50$ kg, cân 20 sản phẩm của nhà máy B thì tính được $\overline{y} = 50,6$ kg.

Giải

Gọi trọng lượng sản phẩm của nhà máy A là X; trọng lượng sản phẩm của nhà máy B là Y thì X,Y là các đại lượng ngẫu nhiên có phân phối chuẩn với $\sigma_X^2 = \sigma_Y^2 = 1$.

Ta kiểm định giả thuyết $H_0 : \mu_X = \mu_Y$

Đối thuyết $H_1 : \mu_X \neq \mu_Y$

+ Tính $z = \dfrac{50-50,6}{\sqrt{\dfrac{1}{25}+\dfrac{1}{20}}} = -2$

+Tìm $z_{\alpha/2}$: $\alpha = 5\% \Rightarrow z_{\alpha/2} = 1,96$ tra bảng phụ lục 6.

+ Ta thấy $|z| > z_{\alpha/2}$, ta chấp nhận đối thuyết H_1, tức là trọng lượng trung bình của sản phẩm ở hai nhà máy là khác nhau.

Lệnh trong R

Giả thuyết $H_0 : \mu_X - \mu_Y = 0$

Đối thuyết $H_1 : \mu_X - \mu_Y \neq 0$

> library(BSDA)

> zsum.test(mean.x=50, n.x=25, sigma.x=1, mean.y=50.6, n.y=20, sigma.y=1, mu=0, alt="t", conf.level=0.95)

Two-sample z-Test

data: Summarized x and y

z = -2, p-value = 0.0455 # kiểm định z và giá trị p_value

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.1879892 -0.0120108

sample estimates:

mean of x mean of y 50.0 50.6

Nhận xét: p-value = 0.0455 < $\alpha = 5\%$ nên đối thiết H_1 có ý nghĩa thống kê.

Ví dụ 6.25

Trọng lượng học sinh của hai trường là các đại lượng ngẫu nhiên có phân phối chuẩn và có cùng độ lệch chuẩn là $\sigma = 1$kg. Cân thử 40 học sinh của trường A ta tính được $\overline{x} = 35$kg, cân 30 học sinh của trường B thì tính được $\overline{y} = 36$kg. Với mức ý nghĩa $\alpha = 5\%$, có thể xem trọng lượng trung bình của học sinh ở trường A thấp hơn trọng lượng trung bình của học sinh ở trường B không?

Giải

Gọi trọng lượng học sinh của trường A là X; trọng lượng học sinh của trường B là Y thì X,Y là các đại lượng ngẫu nhiên có phân phối chuẩn với $\sigma_X^2 = \sigma_Y^2 = 2,5$.

Ta kiểm định giả thuyết $H_0 : \mu_X = \mu_Y$

Đối thuyết $H_1 : \mu_X < \mu_Y$

+ Tính $z = \dfrac{35-36}{\sqrt{\dfrac{1}{40}+\dfrac{1}{30}}} = -4,149$

+ Tìm z_α: $\alpha = 5\% \Rightarrow z_\alpha = 1,645$ tra bảng phụ lục 7.

+ Ta thấy $z = -4,149 < -z_{\alpha} = -1,645$, ta chấp nhận đối thuyết H_1, tức là trọng lượng trung bình của học sinh ở trường A thấp hơn trọng lượng trung bình của học sinh ở trường B.

Lệnh trong R

Giả thuyết $H_0 : \mu_X - \mu_Y = 0$

Đối thuyết $H_1 : \mu_X - \mu_Y < 0$

> library(BSDA)

>zsum.test(mean.x=35, n.x=40, sigma.x=1, mean.y=36, n.y=30, sigma.y=1, mu=0, alt="l", conf.level=0.95)

Two-sample z-Test

data: Summarized x and y

z = -4.1404, p-value = 1.734e-05 # kiểm định z và giá trị p_value

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

NA -0.6027301

sample estimates:

mean of x mean of y 35 36

Nhận xét: p-value = 1.734e-05 < $\alpha = 5\%$ nên đối thuyết H_1 có ý nghĩa thống kê.

Ví dụ 6.26

Theo dõi giá cổ phiếu của hai công ty A và B trong một số ngày người ta tính được các giá trị sau đây:

	Trung bình $\overline{x}$	Độ lệch mẫu hiệu chỉnh s
Công ty A	38,24	2,20
Công ty B	37,10	1,50

Giả thiết rằng giá cổ phiếu của hai công ty A và B là hai biến ngẫu nhiên phân phối chuẩn.

a. Cho biết số liệu trên có được từ 31 ngày theo dõi giá trị cổ phiếu (mỗi ngày một giá trị cho mỗi công ty). Vậy với mức ý nghĩa 1%, có thể nói rằng có sự khác biệt thực sự về giá cổ phiếu trung bình của hai công ty A và B hay không?

b. Cho biết số liệu trên có được từ 20 ngày theo dõi giá trị cổ phiếu (mỗi ngày một giá trị cho mỗi công ty). Vậy với mức ý nghĩa 2%, có thể nói rằng giá cổ phiếu trung bình của công ty A thực sự cao hơn công ty B hay không?

Giải

a. Gọi giá cổ phiếu của công ty A là X; giá cổ phiếu của công ty B là Y thì X,Y là các đại lượng ngẫu nhiên có phân phối chuẩn.

Ta có $n_1 = n_2 = 31 > 30$

Ta kiểm định giả thuyết $H_0 : \mu_X = \mu_Y$

Đối thuyết $H_1 : \mu_X \neq \mu_Y$

+ Tính $z = \dfrac{38,24 - 37,10}{\sqrt{\dfrac{2,2^2}{31} + \dfrac{1,5^2}{31}}} = 2,3838$

+Tìm $z_{\alpha/2}$: $\alpha = 1\% \Rightarrow z_{\alpha/2} = 2,576$ tra bảng phụ lục 6.

+ Ta thấy $|z| = 2,3838 < z_{\alpha/2} = 2,576$, ta chấp nhận giả thiết H_0, tức là giá cổ phiếu trung bình của hai công ty A và B là như nhau.

> library(BSDA)

> zsum.test(mean.x=38.24, n.x=31, sigma.x=2.20, mean.y=37.1, n.y=31, sigma.y=1.5, mu=0, alt="t", conf.level=0.99)

Two-sample z-Test

data: Summarized x and y

z = 2.3838, p-value = 0.01714 # kiểm định z và giá trị p_value

alternative hypothesis: true difference in means is not equal to 0

99 percent confidence interval:

-0.09185432 2.37185432

sample estimates:

mean of x mean of y 38.24 37.10

Nhận xét: p-value = 0.01714 > $\alpha = 1\%$ nên chấp nhận giả thuyết H_0.

b. Ta có $n_1 = n_2 = 20 < 30$ chia 2 trường hợp

***Trường hợp $\sigma_X^2 = \sigma_Y^2$**

Ta kiểm định giả thuyết $H_0 : \mu_X = \mu_Y$

Đối thuyết $H_1 : \mu_X > \mu_Y$

+ Tính $t = \dfrac{38,24 - 37,10}{1,8828\sqrt{\dfrac{1}{20} + \dfrac{1}{20}}} = 1,9147$

với $s_p = \sqrt{\dfrac{19 \times 2,2^2 + 19 \times 1,5^2}{20 + 20 - 2}} = 1,8828$

+Tìm t_α^k: $k = n_1 + n_2 - 2 = 38$, $t_{0,02}^{38} = 2,1267$ tra bảng phụ lục 9.

+ Ta thấy $t = 1,9147 < t_{0,02}^{38} = 2,1267$, ta chấp nhận giả thiết H_0, tức là giá cổ phiếu trung bình của hai công ty A và B là như nhau.

Các lệnh trong R

> library(BSDA)

> tsum.test(mean.x=38.24, s.x = 2.20, n.x = 20, mean.y = 37.1, s.y = 1.5, n.y = 20, alternative = "greater", mu = 0, var.equal = TRUE, conf.level = 0.98)

Standard Two-Sample t-Test

data: Summarized x and y

t = 1.9147, df = 38, p-value = 0.03154 # kiểm định t, bậc tự do 38 và giá trị p_value
alternative hypothesis: true difference in means is greater than 0
98 percent confidence interval:
-0.1262196 NA
sample estimates:
mean of x mean of y 38.24 37.10

Nhận xét: p-value = 0.03154 > $\alpha = 2\%$ nên chấp nhận giả thuyết H_0.

***Trường hợp chưa biết $\sigma_X^2 = \sigma_Y^2$**

Ta kiểm định giả thuyết $H_0 : \mu_X = \mu_Y$

Đối thuyết $H_1 : \mu_X > \mu_Y$

\+ Tính $t = \dfrac{38,24 - 37,10}{\sqrt{\dfrac{2,2^2}{20} + \dfrac{1,5^2}{20}}} = 1,9147$

+Tìm t_α^{df} :

$$df = \frac{\left(\dfrac{s_x^2}{n_1} + \dfrac{s_y^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_x^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_y^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\dfrac{2,2^2}{20} + \dfrac{1,5^2}{20}\right)^2}{\dfrac{\left(\dfrac{2,2^2}{20}\right)^2}{19} + \dfrac{\left(\dfrac{1,5^2}{20}\right)^2}{19}} = 33,529$$

Tra bảng phụ lục 9, ta có $t_{0,02}^{34} = 2.1356 < t_{0,02}^{33,529} < t_{0,02}^{32} = 2.1409$.

Sử dụng lệnh trong R
> qt(0.98, 33.529)
[1] 2.136775

Ta có $t_{0,02}^{33,529} = 2,136775$

\+ Ta thấy $t = 1,9147 < t_{0,02}^{33,529} = 2.136775$, ta chấp nhận giả thiết H_0, tức là giá cổ phiếu trung bình của hai công ty A và B là như nhau.

Các lệnh trong R

> tsum.test(mean.x=38.24, s.x = 2.20, n.x = 20, mean.y = 37.1, s.y = 1.5, n.y = 20, alternative = "greater", mu = 0, var.equal = FALSE, conf.level = 0.98)

Welch Modified Two-Sample t-Test
data: Summarized x and y
t = 1.9147, df = 33.526, p-value = 0.03205 # kiểm định t, bậc tự do 33,526 và giá trị p_value
alternative hypothesis: true difference in means is greater than 0
98 percent confidence interval:

-0.1322385 NA

sample estimates:

mean of x mean of y 38.24 37.10

Nhận xét: p-value = 0.03205 > $\alpha = 2\%$ nên chấp nhận giả thuyết H_0.

6.3.5. Kiểm định giả thuyết về sự bằng nhau của hai tỷ lệ

Bài toán 6.5

Giả sử p_1, p_2 tương ứng là tỷ lệ các phần tử có tính chất A của tổng thể thứ nhất và tổng thể thứ hai. Ta cần kiểm định giả thuyết

$$\mathrm{H}_0 : p_1 = p_2$$

Quy tắc thực hành 6.5

●**Bảng kiểm định 2 phía (two tail):** Giả thuyết $\mathrm{H}_0 : p_1 = p_2$ và Đối thuyết $\mathrm{H}_1 : p_1 \neq p_2$

Bước	Trường hợp	
	p_0 đã biết	p_0 chưa biết
1. Tính z	$z = \dfrac{f_1 - f_2}{\sqrt{p_0(1-p_0)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$	$z = \dfrac{f_1 - f_2}{\sqrt{p^*(1-p^*)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ với $p^* = \dfrac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$
2. Tra bảng	$z_{\alpha/2} : \alpha \Rightarrow z_{\alpha/2}$ phụ lục 6.	$z_{\alpha/2} : \alpha \Rightarrow z_{\alpha/2}$ phụ lục 6.
3a. Chấp nhận H_0	$\lvert z \rvert \leq z_{\alpha/2}$	$\lvert z \rvert \leq z_{\alpha/2}$
3b. Bác bỏ H_0	$\lvert z \rvert > z_{\alpha/2}$	$\lvert z \rvert > z_{\alpha/2}$

● **Bảng kiểm định 1 phía (one tail):** Giả thuyết $\mathrm{H}_0 : p_1 = p_2$ và Đối thuyết $\mathrm{H}_1 : p_1 < p_2$

Bước	Trường hợp	
	p_0 đã biết	p_0 chưa biết
1. Tính z	$z = \dfrac{f_1 - f_2}{\sqrt{p_0(1-p_0)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$	$z = \dfrac{f_1 - f_2}{\sqrt{p^*(1-p^*)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ với $p^* = \dfrac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$
2. Tra bảng	$z_{\alpha} : \alpha \Rightarrow z_{\alpha}$ phụ lục 7.	$z_{\alpha} : \alpha \Rightarrow z_{\alpha}$ phụ lục 7.
3a. Chấp nhận H_0	$z \geq -z_{\alpha}$	$z \geq -z_{\alpha}$
3b. Bác bỏ H_0	$z < -z_{\alpha}$	$z < -z_{\alpha}$

● **Bảng kiểm định 1 phía (one tail):** Giả thuyết $H_0 : p_1 = p_2$ và Đối thuyết $H_1 : p_1 > p_2$

Bước	Trường hợp	
	p_0 đã biết	p_0 chưa biết
1. Tính z	$z = \dfrac{f_1 - f_2}{\sqrt{p_0(1-p_0)\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$	$z = \dfrac{f_1 - f_2}{\sqrt{p^*(1-p^*)\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$ với $p^* = \dfrac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$
2. Tra bảng	$z_\alpha : \alpha \Rightarrow z_\alpha$ phụ lục 7.	$z_\alpha : \alpha \Rightarrow z_\alpha$ phụ lục 7.
3a. Chấp nhận H_0	$z \le z_\alpha$	$z \le z_\alpha$
3b. Bác bỏ H_0	$z > z_\alpha$	$z > z_\alpha$

Ví dụ 6.27

Kiểm tra các sản phẩm được chọn ở hai nhà máy sản xuất ta được số liệu sau

Nhà máy	Số sản phẩm được kiểm tra	Số phế phẩm
I	$n_1 = 100$	20
II	$n_2 = 120$	36

Với mức ý nghĩa $\alpha = 1\%$, có thể coi tỷ lệ phế phẩm của hai nhà máy là như nhau không ?

Giải

Gọi p_1, p_2 là tỷ lệ phế phẩm của nhà máy I, II.

Ta cần kiểm định giả thuyết $H_0 : p_1 = p_2$

Đối thuyết $H_1 : p_1 \ne p_2$

Từ các số liệu đã cho ta có

$$+\ f_1 = \frac{20}{100} = 0,2;\ f_2 = \frac{36}{120} = 0,3$$

$$+\ p^* = \frac{100\times 0,2 + 120\times 0,3}{100+120} = 0,254 \Rightarrow 1 - p^* = 0,746$$

$$+\ z = \frac{0,2 - 0,3}{\sqrt{0,254\times 0,746\times\left(\dfrac{1}{100}+\dfrac{1}{120}\right)}} \approx -1,69$$

+Tìm $z_{\alpha/2}$: $\alpha = 1\% \Rightarrow z_{\alpha/2} = 2,576$ tra bảng phụ lục 6.

Ta thấy $|z| < z_{\alpha/2}$ nên chấp nhận giả thuyết H_0, tức là tỷ lệ phế phẩm của hai nhà máy là như nhau.

Lệnh trong R

Giả thuyết $H_0 : p_1 - p_2 = 0$

Đối thuyết $H_1 : p_1 - p_2 \neq 0$

> fracture <- c(20, 36)

> total <- c(100, 120)

> prop.test(fracture, total, conf.level=0.99)

2-sample test for equality of proportions with continuity correction

data: fracture out of total

X-squared = 2.3717, df = 1, p-value = 0.1236 # kiểm định Chi-bình phương, bậc tự do 1 và giá trị p_value

alternative hypothesis: two.sided

99 percent confidence interval: -0.25825357 0.05825357

sample estimates: prop 1 prop 2 0.2 0.3

Nhận xét: p-value = 0,1236 > $\alpha = 1\%$ nên chấp nhận H_0.

Ví dụ 6.28

Kiểm tra các học sinh được chọn ở hai trường ta được số liệu sau

Trường	Số học sinh được kiểm tra	Số học sinh giỏi
I	$n_1 = 100$	32
II	$n_2 = 120$	48

Với mức ý nghĩa $\alpha = 1\%$, có thể coi tỷ lệ học sinh giỏi của trường II cao hơn trường I không ?

Giải

Gọi p_1, p_2 là tỷ lệ học sinh giỏi của hai trường I,II.

Ta cần kiểm định giả thuyết $H_0 : p_1 = p_2$

Đối thuyết $H_1 : p_1 < p_2$

Từ các số liệu đã cho ta có

$$+ f_1 = \frac{32}{100} = 0,32;\ f_2 = \frac{48}{120} = 0,4$$

$$+ p^* = \frac{100 \times 0,32 + 120 \times 0,4}{100 + 120} = 0,363 \Rightarrow 1 - p^* = 0,637$$

$$+ z = \frac{0,32 - 0,4}{\sqrt{0,363 \times 0,637 \times \left(\frac{1}{100} + \frac{1}{120}\right)}} \approx -1,22$$

+ Tìm z_α : $\alpha = 1\% \Rightarrow z_\alpha = 2,326$ tra bảng phụ lục 7.

Ta thấy $z = -1,22 > z_\alpha = -2.326$ nên chấp nhận giả thuyết H_0,

Kết luận chưa có cơ sở để kết luận tỷ lệ học sinh giỏi của trường II cao hơn trường I.

Lệnh trong R

> fracture <- c(32,48)

> total <- c(100, 120)

> prop.test(fracture, total,alternative = "less", conf.level=0.99)

2-sample test for equality of proportions with continuity correction

data: fracture out of total

X-squared = 1.1827, df = 1, p-value = 0.1384 # kiểm định Chi-bình phương, bậc tự do 1 và giá trị p_value

alternative hypothesis: less

99 percent confidence interval: -1.00000000 0.07949987

sample estimates:prop 1 prop 2 0.32 0.40

Nhận xét: p-value = 0.1384 > $\alpha = 1\%$ nên chấp nhận H_0.

6.3.6. Kiểm định giả thuyết về sự bằng nhau của hai phương sai

Bài toán 6.7

Giả sử X và Y là hai đại lượng ngẫu nhiên độc lập có cùng phân phối chuẩn với σ_X^2, σ_Y^2 chưa biết. Ta cần kiểm định giả thuyết $H_0 : \sigma_X^2 = \sigma_Y^2$

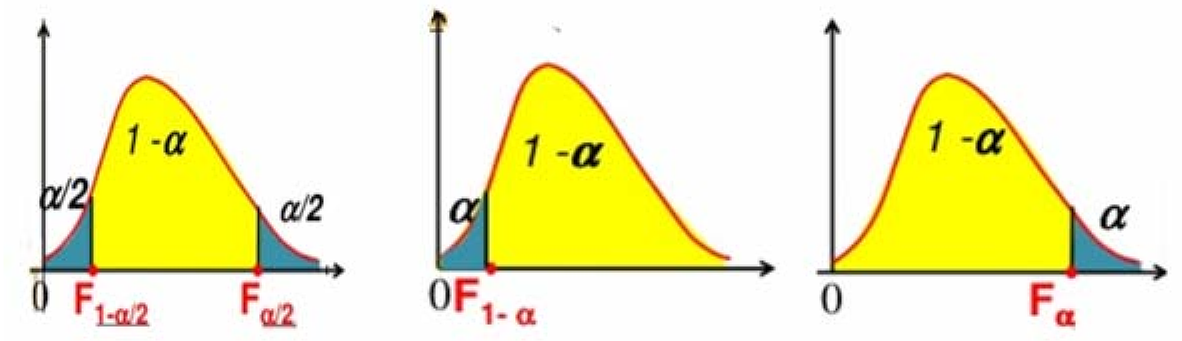
Để so sánh phương sai của hai mẫu có bằng nhau không, phương pháp đơn giản nhất là dùng phương pháp kiểm định F. Kiểm định F tính tỉ số phương sai của hai mẫu và so sánh với giá trị tham chiếu của phân phối F.

Quy tắc thực hành 6.4

Lấy mẫu ngẫu nhiên kích thước n_1 đối với X và mẫu ngẫu nhiên kích thước n_2 đối với Y ta có:

Bước	$H_0 : \sigma_X^2 = \sigma_Y^2$ $H_1 : \sigma_X^2 \neq \sigma_Y^2$	$H_0 : \sigma_X^2 = \sigma_Y^2$ $H_1 : \sigma_X^2 < \sigma_Y^2$	$H_0 : \sigma_X^2 = \sigma_Y^2$ $H_1 : \sigma_X^2 > \sigma_Y^2$
1. Tính F	$F_{qs} = \dfrac{S_X^2}{S_Y^2}(\text{khi } S_X^2 > S_Y^2)$ hoặc $F_{qs} = \dfrac{S_Y^2}{S_X^2}(\text{khi } S_Y^2 > S_X^2)$		
2. Tra bảng	$F_{\alpha/2}(n_1-1, n_2-1)$ và $F_{1-\alpha/2}(n_1-1, n_2-1)$ phụ lục 10	$F_{1-\alpha}(n_1-1, n_2-1)$ phụ lục 10	$F_{\alpha}(n_1-1, n_2-1)$ phụ lục 10
3a. Chấp nhận H_0	$F_{1-\alpha/2}(n_1-1, n_2-1) \le F$ $\le F_{\alpha/2}(n_1-1, n_2-1)$	$F_{qs} \ge F_{1-\alpha}(n_1-1, n_2-1)$	$F_{qs} \le F_{\alpha}(n_1-1, n_2-1)$
3b. Bác bỏ H_0	$F_{qs} < F_{1-\alpha/2}(n_1-1, n_2-1)$ hoặc $F_{\alpha/2}(n_1-1, n_2-1) < F_{qs}$	$F_{qs} < F_{1-\alpha}(n_1-1, n_2-1)$	$F_{qs} > F_{\alpha}(n_1-1, n_2-1)$

Chú ý: $F_{1-\alpha}(n_1-1, n_2-1) = \dfrac{1}{F_{\alpha}(n_1-1, n_2-1)}$



Hình 6.21: Kiểm Fisher hai phía và một phía

Ví dụ 6.29

Một giáo sư cho rằng các bạn sinh viên nữ học đều hơn các bạn sinh viên nam trong lớp. Để kiểm tra tính đúng đắn của lời nhận xét này, ông chủ nhiệm khoa đã chọn một số bài thi của sinh viên nam và nữ cùng học lớp do vị giáo sư trên so sánh. Kết quả thu được như sau (điểm tối đa mỗi bài là 100):

Sinh viên nam: 75, 86, 77, 72, 89, 94, 81, 83, 77, 73, 86, 90, 90.

Sinh viên nữ: 77, 83, 72, 67, 84, 91, 82, 73, 65, 72, 70, 72, 65.

Với mức ý nghĩa $\alpha = 5\%$, có thể cho rằng phương sai điểm của sinh viên nam lớn hơn phương sai điểm của sinh viên nữ hay không ? Giả sử điểm của sinh viên nam và nữ là hai đại lượng ngẫu nhiên độc lập có cùng phân phối chuẩn

Giải

Giả sử X và Y là điểm của sinh viên nam và nữ của lớp do vị giáo sư trên so sánh.

Ta cần kiểm định giả thuyết $H_0: \sigma_X^2 = \sigma_Y^2$ và đối thuyết $H_1: \sigma_X^2 > \sigma_Y^2$

Ta tính được: $s_X^2 = 52{,}60256$; $s_Y^2 = 64{,}47436$

+Tính $F_{qs} = \dfrac{s_Y^2}{s_X^2} = \dfrac{64{,}47436}{52{,}60256} = 1.225689$

+Tìm $F_\alpha(n_1, n_2) = F_{0,05}(12,12) = 2{,}69$

+ So sánh: $F_{qs} = 1{,}225689 < F_{0,05}(12,12) = 2{,}69$, chấp nhận H_0. Chưa có cơ sở để cho rằng các bạn sinh viên nữ học đều hơn các bạn sinh viên nam.

Các lệnh trong R

> data1<-c(75, 86, 77, 72, 89, 94, 81, 83, 77, 73, 86, 90, 90)

> data2<-c(77, 83, 72, 67, 84, 91, 82, 73, 65, 72, 70, 72, 65)

> var.test(data1, data2, alternative = "greater", conf.level = 0.95)

F test to compare two variances

data: data1 and data2

F = 0.81587, num df = 12, denom df = 12, p-value = 0.6349 # kiểm định Fisher, bậc tự do (12,12) và giá trị p_value

alternative hypothesis: true ratio of variances is greater than 1

95 percent confidence interval: 0.3036763 Inf

sample estimates: ratio of variances 0.815868

Nhận xét: p-value = 0.6349 > $\alpha = 5\%$ nên chấp nhận H_0.

Chú ý: Phương sai của nhóm sinh viên nam và nhóm sinh viên nữ dùng lệnh var() là:

```
> var(data1)
[1] 52.60256
> var(data2)
[1] 64.47436
```

Do đó, tỉ số của hai phương sai (ratio of variances) là 52,60256/ 64,47436 = 0,81587 nên F = 0,81587.

6.3.7. Phân tích Phương sai

Chúng ta đã kiểm định sự khác biệt về giá trị trung bình của hai tổng thể có dựa trên hai mẫu khảo sát ngẫu nhiên. Tuy nhiên, trên thực tế đôi khi chúng ta cần so sánh sự khác biệt về giá trị trung bình, không chỉ trên hai tổng thể mà có thể là trên nhiều tổng thể. Trong nghiên cứu về sự phụ thuộc giữa kết quả học tập và mức lương ở cùng lĩnh vực, người ta quan tâm đến việc có hay không sự phụ thuộc mức lương vào kết quả học tập với giả thuyết rằng kết quả học tập được chia làm 4 nhóm: trung bình, trung bình-khá, khá và giỏi. Mức lương được khảo sát dựa trên các nhóm kết quả học tập này. Lúc này việc so sánh mức lương trung bình trên các nhóm kết quả học tập sẽ là phương án phù hợp để trả lời cho câu hỏi trên. Việc phân tích sự phụ thuộc giữa một biến nguyên nhân (thường là biến định tính hoặc là biến định lượng nhưng được chuyển về biến định tính) với ít nhất là ba nhóm tính chất khác nhau với một biến định lượng dựa trên giá trị trung bình của mỗi nhóm tính chất thường được gọi là phân tích phương sai một yếu tố (ANOVA). Còn việc phân tích sự phụ thuộc của hai biến định tính, mỗi biến có từ ba nhóm tính chất trở lên với một biến định lượng được gọi là phân tích phương sai hai yếu tố (2-ANOVA).

6.3.7.1. Phân tích phương sai một yếu tố (ANOVA)

Giả sử chúng ta cần phân tích xem có hay không sự phụ thuộc giữa một biến định tính (biến nguyên nhân) và một biến định lượng (biến kết quả) dựa trên k nhóm giá trị của biến định tính. Nếu biến định tính chỉ có hai giá trị khác nhau thì sự phân tích này có thể đơn giản là so sánh giá trị trung bình của biến kết quả dựa trên hai nhóm giá trị của biến nguyên nhân xem có bằng nhau hay khác nhau để từ đó rút ra kết luận là có hay không sự phụ thuộc. Tuy nhiên, nếu số nhóm tính chất của biến định tính (biến nguyên nhân) từ 3 trở lên thì việc so sánh các nhóm giá trị trung bình sẽ trở nên phức tạp hơn rất nhiều. Chẳng hạn như khi muốn xem kết quả học tập của sinh viên có chịu ảnh hưởng bởi giới tính hay không, chúng ta hoàn toàn có thể chỉ cần so sánh điểm trung bình học tập của nhóm giới tính nam và nữ xem có sự khác biệt không? Phương pháp so sánh hai giá trị trung bình đã được trình bày. Song nếu chúng ta muốn tìm hiểu xem kết quả học tập của sinh viên có chịu ảnh hưởng của khu vực sinh sống không, với giả sử khu vực sinh sống được chia làm ba nhóm: sống với gia đình; sống ở ký túc xá; ở trọ bên ngoài thì việc so sánh kết quả học tập trung bình giữa ba nhóm này sẽ không thể giải quyết dễ dàng. Để giải quyết vấn đề này, phân tích phương sai một yếu tố (ANOVA) là sự lựa chọn phù hợp khi cần so sánh k (k>2) nhóm tính chất của biến nguyên nhân mà vẫn dựa trên những giả định cơ bản như so sánh giá trị trung bình của hai tổng thể. Các bước cụ thể của phân tích phương sai một yếu tố (ANOVA) được thực hiện như sau:

Giả sử chúng ta cần so sánh giá trung bình của k tổng thể ($k \geq 3$), cặp giả thuyết cần kiểm định có dạng sau:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_k \\ H_1 : \text{Không phải tất cả các } \mu_i \text{ đều bằng nhau} (i = 1,...,k) \end{cases}$$

Ở đây, μ_i là giá trị trung bình của tổng thể thứ i ($i = 1,...,k$).

Việc kiểm định cặp giả thuyết trên đòi hỏi một số giả thiết cơ bản sau:

- Các tổng thể có phân phối chuẩn;
- Phương sai của các tổng thể bằng nhau;
- Các mẫu được chọn ngẫu nhiên và độc lập với nhau.

Lưu ý: Nếu bác bỏ H_0, chúng ta không thể kết luận tất cả μ_i khác nhau mà chỉ có thể kết luận ít nhất hai giá trị trung bình của tổng thể là μ_i và μ_j khác nhau. Để có thể đưa ra kết luận chính xác là các μ_i nào khác nhau, chúng ta cần tìm hiểu ở mục tiếp theo 6.12.1. phân tích sâu phương sai một yếu tố.

Do việc khảo sát để tìm được các giá trị $\mu_1, \mu_2, ..., \mu_k$ thường khó xảy ra trong thực tế nên người ta thường dựa vào các mẫu ngẫu nhiên $n_1, n_2, ..., n_k$ ứng với các tổng thể này để tìm các giá trị đại diện tương ứng $\overline{X_1}, \overline{X_2}, ..., \overline{X_k}$. Điều này được tóm tắt trong bảng sau:

Biến 1 Biến 2	Mẫu 1(n_1)	Mẫu 2(n_2)	…	Mẫu k(n_k)
1	X_{11}	X_{12}	…	X_{1k}
2	X_{21}	X_{22}	…	X_{2k}
…	…	…	…	
			…	
Giá trị trung bình	$\overline{X_1}$	$\overline{X_2}$	…	$\overline{X_k}$

Từ mẫu này, chúng ta xây dựng các bước để kiểm định cặp giả thuyết như sau:

Bước 1: Đặt giả thuyết

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_k \\ H_1 : \text{Không phải tất cả các } \mu_i \text{ đều bằng nhau} (i = 1,...,k) \end{cases}$$

Bước 2:

- Tính trung bình mẫu của từng nhóm theo công thức $\overline{X_j} = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$

- Tính trung bình chung của k nhóm theo công thức: $\overline{X} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{\sum_{j=1}^{7} n_j} = \frac{\sum_{j=1}^{k} n_j \overline{x_j}}{\sum_{j=1}^{7} n_j}$

Bước 3:

- Tính tổng các chênh lệch bình phương trong nội bộ nhóm: $SSW = \sum_{j=1}^{k} SS_j$

với $SS_j = \sum_{i=1}^{n_j}\left(x_{ij} - \overline{X_j}\right)^2$

- Tính tổng các chệnh lệch bình phương giữa các nhóm:

$$SSG = \sum_{j=1}^{k} n_j\left(\overline{X_j} - \overline{X}\right)^2$$

- Tính tổng các chệnh lệch bình phương toàn bộ của toàn bộ các phần tử theo công thức:

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \overline{X}\right)^2 = SSW + SSG$$

với

- k là số mẫu khảo sát;
- n_j là cỡ mẫu của nhóm thứ j;
- x_{ij} là giá trị của phần tử thứ i trong mẫu thứ j;
- $\overline{X_j}$ là trung bình mẫu của mẫu thứ j;
- $\overline{X}$ là trung bình mẫu chung của toàn bộ các phần tử trong k mẫu.

Ý tưởng chính của bước 3 là đo lường sự biến động ở dạng bình phương của tất cả các giá trị quan sát (SST) trên các mẫu riêng biệt ứng với từng tổng thể so với giá trị trung bình chung $\overline{X}$. Sự biến động (SST) này được chia tách làm hai thành phần chính: một phần là sự biến động do các yếu tố nghiên cứu tạo ra (SSG) và phần còn lại là do các yếu tố khác không nghiên cứu tạo ra (SSW). Nếu (SSG) càng nhỏ (sự biến động do các yếu tố nghiên cứu quá ít) thì càng có cơ sở để chấp nhận H_0 và ngược lại.

Bước 4:

- Tính phương sai trong nội bộ nhóm theo công thức:

$$MSW = \frac{SSW}{n-k}$$

- Tính phương sai giữa các nhóm theo công thức:

$$MSG = \frac{SSG}{k-1}$$

Bước 5: Tính giá trị kiểm định F theo quy luật Fisher-Snedecor qua công thức:

$$F_0 = \frac{MSG}{MSW}$$

- Tra bảng phụ lục 10, ta tìm được giá trị $F_{(df_1=k-1, df_2=n-k);\alpha}$.

- Nếu $F_0 > F_{(df_1=k-1, df_2=n-k);\alpha}$ thì bác bỏ H_0 và ngược lại.

Nhận xét: Ngoài cách so sánh giá trị kiểm định F với giá trị tra bảng Fisher-Snecdecor, chúng ta có thể sử dụng giá trị p_value=P(F>F_0). Nếu p_value < mức ý nghĩa α thì bác bỏ H_0 và ngược lại.

Chúng ta có bảng tóm tắt các bước kiểm định trên trong bảng sau:

Phân tích phương sai một yếu tố (one-way ANOVA)				
	Tổng các chênh lệch bình phương	Bậc tự do	Phương sai	Tỉ số (F)
Giữa các nhóm	SSG	k-1	MSG	MSG/MSW
Nội bộ nhóm	SSW	n-k	MSW	
Tổng	SST	n		

Ví dụ 6.30

Theo số liệu quan trắc tại tỉnh Ninh Thuận, Sở Tài nguyên và Môi trường tỉnh muốn so sánh xem có hay không sự khác biệt về mức độ ô nhiễm thông qua hàm lượng NO_2 (mg/lít) có trong nước tại 7 cây cầu của tỉnh. Số liệu được thu thập trong vòng 12 tháng tại 7 cây cầu như sau:

	Tháng	Cầu sông Cái	Cầu Ninh Bình	Cầu Tân Mỹ	Thôn Phú Thạnh	Đập Lâm Cấm	Cầu Móng (Bảo An)	Cầu Đạo Long 1
2015	1	0,009	0,009	0,005	0,004	0,012	0,008	0,01
	2	0,012	0,022	0,011	0,009	0,031	0,051	0,036
	3	0,005	0,008	0,004	0,005	0,056	0,117	0,094
	4	0,013	0,006	0,006	0,003	0,012	0,009	0,12
	5	0,01	0,024	0,014	0,013	0,015	0,019	0,019
	6	0,011	0,014	0,008	0,009	0,017	0,029	0,373
	7	0,009	0,013	0,014	0,013	0,02	0,016	0,045
	8	0,005	0,017	0,011	0,012	0,013	0,01	0,013
	9	0,014	0,008	0,007	0,008	0,009	0,008	0,006
	10	0,018	0,032	0,01	0,015	0,009	0,012	0,014
	11	0,008	0,029	0,024	0,019	0,015	0,017	0,025
	12		0,007	0,006	0,006	0,01	0,009	0,01

Với mức ý nghĩa 5%, bạn hãy kiểm định xem có hay không sự khác biệt về mức độ ô nhiễm nguồn nước thông qua hàm lượng NO_2 tại 7 cây cầu của tỉnh (giả định là phương sai các tổng thể về hàm lượng NO_2 tại 7 cây cầu là bằng nhau, tuân theo phân phối chuẩn, các mẫu được chọn là ngẫu nhiên và độc lập nhau).

Giải

Bước 1: Đặt giả thuyết

Gọi μ_i là hàm lượng NO_2 trung bình đo được tại cây cầu thứ i ($i = \overline{1,7}$).

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_8 \\ H_1 : \text{Tồn tại ít nhất một cặp } (\mu_i \neq \mu_j) \end{cases}$$

Bước 2:

- Tính trung bình mẫu của từng nhóm

Nhóm 1 (Cầu sông Cái): $\overline{X_1} = 0,01$

Nhóm 2 (Cầu Ninh Bình): $\overline{X_2} = 0,016$

Nhóm 3 (Cầu Tân Mỹ): $\overline{X_3} = 0,01$

Nhóm 4 (Cầu thôn Phú Thạnh): $\overline{X_4} = 0,01$

Nhóm 5 (Đập Lâm Cấm): $\overline{X_5} = 0,018$

Nhóm 6 (Cầu Móng): $\overline{X_6} = 0,025$

Nhóm 7 (Cầu Đạo Long 1): $\overline{X_7} = 0,064$

- Tính trung bình chung của 7 nhóm: $\overline{X} = \dfrac{\sum_{j=1}^{7}\sum_{i=1}^{n_j} x_{ij}}{\sum_{j=1}^{7} n_j} = 0,022$

Bước 3:

- Tính tổng các chênh lệch bình phương trong nội bộ nhóm: $SSW = \sum_{j=1}^{7} SS_j = 0,13292$

với $SS_j = \sum_{i=1}^{n_j}\left(x_{ij} - \overline{X_j}\right)^2$, cụ thể là:

$$SS_1 = \sum_{i=1}^{11}\left(x_{i1} - \overline{X_1}\right)^2 = (0,009 - \overline{X_1})^2 + (0,012 - \overline{X_1})^2 + ... + (0,008 - \overline{X_1})^2 = 0,00015$$

$SS_2 = 0,000897$

$SS_3 = 0,000336$

$SS_4 = 0,00026$

$SS_5 = 0,001959$

$SS_6 = 0,010821$

$SS_7 = 0,118505$

- Tính tổng các chệnh lệch bình phương giữa các nhóm (SSG):

$$SSG = \sum_{j=1}^{7} n_j \left(\overline{X_j} - \overline{X} \right)^2 = 11 \times (0,01 - 0,022)^2 + ... + 12 \times (0,064 - 0,022)^2 = 0,02673$$

- Tính tổng các chệnh lệch bình phương toàn bộ của toàn bộ các phần tử theo công thức:

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_j} \left(x_{ij} - \overline{X} \right)^2 = SSW + SSG = 0,15965$$

Bước 4:

- Tính phương sai trong nội bộ nhóm theo công thức:

$$MSW = \frac{SSW}{n-k} = \frac{0,13292}{76} = 0,001749$$

- Tính phương sai giữa các nhóm theo công thức:

$$MSG = \frac{SSG}{k-1} = \frac{0,02673}{6} = 0,00445$$

Bước 5: Tính giá trị kiểm định F theo quy luật Fisher-Snedecor qua công thức:

$$F = \frac{MSG}{MSW} = \frac{0,00445}{0,001749} = 2,544$$

Tra bảng phụ lục 10, ta tìm được giá trị $F_{(6,76);0,05} = 2,22$.

Kết luận: Do giá trị kiểm định $F = 2,544 > F_{(6,76);0,05} = 2,22$ nên bác bỏ H_0.

Vậy mức độ ô nhiễm nguồn nước thông qua hàm lượng NO_2 tại 7 cây cầu của tỉnh Ninh thuận là có sự khác biệt nhau; hay chúng ta có thể kết luận rằng mức độ ô nhiễm thể hiện qua hàm lượng NO_2 phụ thuộc vào nguồn nước tại 7 cây cầu này.

6.3.7.2. Phân tích sâu phương sai một yếu tố

Khi phân tích ANOVA một yếu tố, nếu giả thuyết H_0 xảy ra thì giá trị trung bình của các tổng thể trên các nhóm tính chất của biến nguyên nhân là bằng nhau với một mức ý nghĩa cho sẵn, khi đó quá trình kiểm định kết thúc. Ngược lại, nếu giả thuyết H_1 xảy ra, nghĩa là tồn tại ít nhất một cặp giá trị trung bình khác nhau nhưng chúng ta không biết được có bao nhiêu cặp giá trị trung bình khác nhau và bao nhiêu cặp không khác nhau, khi đó người ta sẽ tiếp tục việc kiểm định để xác định xem cặp giá trị trung bình nào khác nhau và không khác nhau. Công việc này được thực hiện qua kiểm định Tukey-Kramer với các bước thực hiện như sau:

Bước 1: Xác định số cặp trung bình cần kiểm định và đặt giả thuyết

- Số cặp giả thuyết được xác định qua công thức:

Số cặp=$C_k^2 = \dfrac{k!}{2! \times (k-2)!}$ trong đó k là số nhóm.

- Xây dựng các cặp giả thuyết dựa trên số cặp cần kiểm định:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}; \begin{cases} H_0 : \mu_1 = \mu_3 \\ H_1 : \mu_1 \neq \mu_3 \end{cases}$$

Bước 2: Tính giá trị kiểm định Tukey-Kramer cho cặp giả thuyết cần kiểm định theo

công thức $T = q_\alpha^{n-k} \times \sqrt{\frac{MSW}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$

trong đó:

- q_α^{n-k} là giá trị tra bảng trong phân phối Tukey;
- MSW là phương sai trong nội bộ nhóm;
- n_i và n_j lần lượt là cỡ mẫu trong nhóm thứ i và j của cặp giả thuyết cần kiểm định;
- n là tổng số phần tử của k nhóm.

Bước 3: So sánh và kết luận

Nếu độ lệch tuyệt đối cặp trung bình mẫu là $\left|\overline{X_i} - \overline{X_j}\right| > T$ thì bác bỏ giả thuyết H_0 của cặp giả thuyết đó và ngược lại.

Lưu ý:

- Bước 3 sẽ được thực hiện trên mọi cặp giả thuyết đã được xây dựng ở bước 1 để xác định chính xác cặp giá trị trung bình nào sẽ khác biệt và ngược lại.
- Ngoài cách tính giá trị kiểm định và so sánh như ở bước 2 và bước 3 ở trên, chúng ta cũng có thể sử dụng phương pháp kiểm định so sánh 2 giá trị trung bình trên từng cặp giả thuyết như sau:

Bước 1: Đặt giả thuyết $\begin{cases} H_0 : \mu_i = \mu_j \\ H_1 : \mu_i \neq \mu_j \end{cases}$

Bước 2: Tính giá trị kiểm định: $T = \dfrac{\left(\overline{X_i} - \overline{X_j}\right)}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$

Bước 3: So sánh và kết luận

Nếu $|T| \geq q_\alpha^{n-k}$ thì bác bỏ H_0 và ngược lại.

Ví dụ 6.31 (tiếp theo Ví dụ 6.30)

Khi bác bỏ H_0 ở mức ý nghĩa 5%, chúng ta sẽ tiếp tục kiểm định xem mức độ ô nhiễm nguồn nước thông qua hàm lượng NO_2 tại các cây cầu nào sẽ có sự khác biệt nhau và ngược lại. Do đó, chúng ta sẽ áp dụng phân tích sâu ANOVA trong trường hợp này.

Giải

Bước 1:

- Xác định số cặp giả thuyết cần kiểm định là: $C_7^2 = \dfrac{7!}{2!\times(7-2)!} = 21$ cặp

- Xây dựng các cặp giả thuyết dựa trên số cặp cần kiểm định là:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}; \begin{cases} H_0 : \mu_1 = \mu_3 \\ H_1 : \mu_1 \neq \mu_3 \end{cases}; \ldots; \begin{cases} H_0 : \mu_6 = \mu_7 \\ H_1 : \mu_6 \neq \mu_7 \end{cases}$$

Bước 2: Tính giá trị kiểm định Tukey-Kramer cho từng cặp giả thuyết cần kiểm định theo công thức $T = q_\alpha^{n-k} \times \sqrt{\dfrac{MSW}{2}\left(\dfrac{1}{n_i} + \dfrac{1}{n_j}\right)}$

Nhận xét: Trong 21 cặp giả thuyết cần kiểm định thì chỉ có cỡ mẫu của nhóm thứ nhất là 11 còn lại các cỡ mẫu của các nhóm khác đều là 12 nên giá trị cần kiểm định T này chỉ khác nhau ở những cặp giả thuyết có nhóm 1 và ngược lại. Do đó chúng ta sẽ chỉ có 2 giá trị kiểm định T: một là giá trị T cho các cặp giả thuyết có nhóm 1 và hai là các cặp giả thuyết không có nhóm 1.

Trường hợp 1: các cặp cần kiểm định có nhóm 1

$$T = q_\alpha^{n-k} \times \sqrt{\frac{MSW}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = q_{0,05}^{76}\sqrt{\frac{0,001749}{2}\left(\frac{1}{11} + \frac{1}{12}\right)} = 0,0532$$

Biết $q_{0,05}^{76} = 4,307$

Trường hợp 2: các cặp cần kiểm định không có nhóm 1

$$T = q_\alpha^{n-k} \times \sqrt{\frac{MSW}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = q_{0,05}^{76}\sqrt{\frac{0,001749}{2}\left(\frac{1}{12} + \frac{1}{12}\right)} = 0,052$$

Bước 3: So sánh và kết luận

Cặp 1: $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \Rightarrow \left|\overline{X_1} - \overline{X_2}\right| = |\,0,01 - 0,016\,| = 0,006 < T.$

Kết luận: Không có sự khác biệt về mức độ ô nhiễm thông qua hàm lượng NO_2 giữa nhóm 1 (Cầu sông Cái) và nhóm 2 (Cầu Ninh Bình).

Cặp 2: $\begin{cases} H_0 : \mu_1 = \mu_3 \\ H_1 : \mu_1 \neq \mu_3 \end{cases} \Rightarrow \left|\overline{X_1} - \overline{X_3}\right| = |\,0,01 - 0,01\,| = 0 < T.$

Kết luận: Không có sự khác biệt về mức độ ô nhiễm thông qua hàm lượng NO_2 giữa nhóm 1 (Cầu sông Cái) và nhóm 3 (Cầu Tân Mỹ).

Nhận xét: Chúng ta sẽ thực hiện bước 3 này thêm 19 lần nữa để đưa ra kết luận là sự khác biệt về mức độ ô nhiễm thông qua hàm lượng NO_2 ở các cây cầu nào là khác nhau và các cây cầu nào là không. Kết quả thu được như sau:

1. Giữa các cây cầu sau có sự khác biệt về mức độ ô nhiễm thông qua hàm lượng NO_2, cụ thể là:

- Cầu đập Lâm Cấm và sông Cái; - Cầu Ninh Bình và Đạo Long 1;
- Cầu thôn Phú Thạnh và sông Cái.

2. Các cây cầu khác không có sự khác biệt về mức độ ô nhiễm thông qua hàm lượng NO_2.

6.4. KIỂM ĐỊNH PHI THAM SỐ

6.4.1. Kiểm định phi tham số (non-parametric tests)

Các phương pháp thống kê vẫn dùng như ước lượng, kiểm định giá trị trung bình, so sánh các trung bình, phân tích phương sai một nhân tố, hai nhân tố, tương quan, hồi quy đơn, hồi quy bội tuyến tính liên quan đến biến định lượng, đều dựa trên giả thuyết biến kết quả (biến định lượng) tuân theo phân phối chuẩn (Gauss), và dữ liệu thu thập dưới dạng các thang đo có độ tin cậy cao như thang đo khoảng (interval scale) và thang đo tỷ lệ (ratio scale).

Trong những năm gần đây do nhu cầu xử lý các biến định tính, và xử lý các biến định lượng không phân phối chuẩn (dùng các phép biến đổi thông thường như χ^2, Log(X), Arcsin hay biến đổi tổng quát Box - Cox cũng không đưa được về chuẩn) đã ra đời nhiều phương pháp thống kê gọi chung là phương pháp phi tham số (non parametric method).

Thống kê phi tham số dựa trên việc xếp hạng các số liệu do đó nhiều giả thuyết và kết luận liên quan đến trung vị (median) chứ không liên quan đến trung bình.

Tổng quát, chúng ta có thể nói rằng kiểm định phi tham số sẽ phù hợp với dữ liệu trong các trường hợp sau:

- Dữ liệu thu thập thuộc vào các thang đo có độ tin cậy thấp như: thang đo định danh (nominal) hay thang đo thứ bậc (ordinal).
- Dữ liệu thu thập dù thuộc vào thang đo khoảng hay tỷ lệ nhưng hình dáng phân phối dữ liệu không tuân theo phân phối chuẩn hay phương sai không bằng nhau.

• Thống kê phi tham số có ưu điểm:

+ Không dùng các tham số của tổng thể, không cần giả thiết phân phối chuẩn, nói chung không đòi hỏi nhiều điều kiện đối với dữ liệu.

+ Dùng được cho nhiều loại biến chứ không riêng gì biến định lượng.

+ Việc tính các thống kê phi tham số thường đơn giản dễ tính.

• Thống kê phi tham số có nhược điểm:

+ Chuyển sang thứ hạng nên mất nhiều thông tin về bản chất dữ liệu do đó nếu gặp biến định lượng có phân phối chuẩn thì thống kê phi tham số không mạnh bằng thống kê tham số, thành thử nếu điều kiện cho phép dùng kiểm định tham số được thỏa mãn, thì ta nên dùng kiểm định có tham số. Điều này tương tự như khi tổ chức chạy thi 100 m, nếu có dụng cụ đo chính xác thì biết rõ thành tích người thứ nhất chạy hết 9,70 giây, người thứ hai 9,80 giây,... Còn nếu không có dụng cụ đo thì chỉ ghi lại: về đầu,về thứ hai,...

+ Không có nhiều kết luận đi sâu về mặt định lượng hoặc nếu có thì việc tính toán tương đối khó.

• Trong các mẫu nhỏ, nếu ta nghi ngờ không biết sự phân phối của tổng thể là ra sao, ta nên thực hiện một kiểm định phi phân phối hay phi tham số (không cần giả định nào về phân phối tổng thể) , hoặc tính toán giá trị P của thống kê kiểm định theo phương pháp bootstrapping.

• Bảng 1: So sánh kiểm định phi tham số và kiểm định tham số

Kiểm định	Kiểm định phi tham số	Kiểm định tham số
Kiểm định tương quan	Spearman	Pearson
Mẫu từng cặp	Kiểm định dấu (Sign test) hoặc kiểm định dấu và hạng Wilcoxon (Wilcoxon signed rank test)	Kiểm định t từng cặp (Paired sample test)
Hai mẫu độc lập	Kiểm định tổng hạng Wilcoxon (Wilcoxon rank-sum test) hay kiểm định Man-Whitney	Kiểm định t độc lập (Independent sample t test)
Nhiều hơn 2 mẫu độc lập	Kiểm định Kruskal-Wallis	ANOVA một chiều (Kiểm định F)

• Bảng 2: So sánh kiểm định phi tham số và kiểm định tham số

Thống kê tham số	Thống kê phi tham số	Mục đích
t- test		Kiểm định trung bình trên một mẫu
	Wilcoxon signed rank test	Kiểm định trung vị trên một mẫu
Paired t-test		Kiểm định trung bình trên hai mẫu từng cặp
	Signed test hoặc Wilcoxon signed rank test	Kiểm định trung vị trên hai mẫu từng cặp
Independent t- test		Kiểm định trung bình trên hai mẫu độc lập
	Wilcoxon rank-sum test (Man-Whitney)	Kiểm định trung vị trên hai mẫu độc lập
One way anova	Kruskal Wallis	Kiểm định giả thuyết các trung bình của các mức bằng nhau
Pearson R	Spearman Rs χ^2	Tính mối quan hệ giữa hai biến

• Bảng 3: Phương pháp phân tích thống kê thích hợp với các thang đo

Loại thang đo	Đo lường độ tập trung	Kiểm định phi tham số	Kiểm định tham số
Thang đo định danh	Mod	Chi-bình phương χ^2	
Thang đo thứ bậc	Trung vị	- Kiểm định dấu - Kiểm định dấu và hạng Wilcoxon - Kiểm định tổng hạng Wilcoxon	
Thang đo khoảng	Trung bình	- Kiểm định dấu và hạng Wilcoxon - Kiểm định tổng hạng Wilcoxon (Man-Whitney) - Kiểm định Kruskal-Wallis	- Kiểm định t - ANOVA một chiều (Kiểm định F)
Thang đo tỷ lệ	Trung bình tỷ lệ	- Kiểm định dấu và hạng Wilcoxon - Kiểm định tổng hạng Wilcoxon (Man-Whitney) - Kiểm định Kruskal-Wallis	- Kiểm định t - Kiểm định χ^2 - ANOVA một chiều (Kiểm định F)

• Trong phần này chúng ta thảo luận hai phương pháp kiểm định phi tham số thông dụng nhất, được gọi là kiểm định dấu và hạng có dấu Wilcoxon, tổng hạng Wilcoxon (Mann-Whitney), bảng tóm tắt như sau:

Kiểm định phi tham số	Kiểm định tham số
Kiểm định trung vị (Med) $\begin{cases} H_0 : Med = Med_0 \\ H_1 : Med \neq Med_0 \end{cases}$	Kiểm định trung bình trên 1 mẫu $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$
Kiểm định sự bằng nhau của 2 trung vị trong trường hợp mẫu bắt cặp $\begin{cases} H_0 : Med_d = 0 \\ H_1 : Med_d \neq 0 \end{cases}$ Với $Med_d = Med_1 - Med_2$	Kiểm định sự bằng nhau của 2 trung bình trong trường hợp mẫu bắt cặp $\begin{cases} H_0 : \mu_d = 0 \\ H_1 : \mu_d \neq 0 \end{cases}$ Với $\mu_d = \mu_1 - \mu_2$

6.4.2. Kiểm định dấu (Sign Test)

Giả sử (X,Y) là một cặp gồm hai ĐLNN. Ta có thể coi thành phần thứ nhất X là hiệu quả của phương pháp thứ nhất, còn Y hiệu quả của phương pháp thứ hai tác động lên cùng một cá thể (đối tượng). Ta kiểm định giả thuyết

H_0: "Hiệu quả của phương pháp thứ nhất và thứ hai là như nhau".

Giả sử $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ là n quan sát độc lập về (X,Y). Đặt $d_i = x_i - y_i$. Ta loại bỏ các d_i có giá trị bằng 0 vì chúng không đem lại thông tin gì. Gọi $\tilde{n}$ là số các d_i có giá trị khác 0 và n^+ là số các số hạng d_i mang dấu +. Nếu giả thiết H_0 là đúng thì số các số hạng mang dấu + có xu hướng bằng số các số hạng mang dấu -. Thành thử khi H_0 đúng thì n^+ sẽ có phân phối nhị thức với tham số p=0,5 và $\tilde{n}$. Ta biết rằng nếu $\tilde{n} \times 0,5 > 5 \Leftrightarrow \tilde{n} > 10$ thì tần suất $f = \dfrac{n^+}{\tilde{n}}$ sẽ có phân phối xấp xỉ chuẩn với kỳ vọng 0,5 và độ lệch chuẩn là $\sigma = \sqrt{\dfrac{1/2 \times 1/2}{\tilde{n}}} = \dfrac{1}{2\sqrt{\tilde{n}}}$. Thành thử test thống kê sau đây

$$Z = (f - 0,5).2\sqrt{\tilde{n}} = \frac{2n^+ - \tilde{n}}{\sqrt{\tilde{n}}}$$

sẽ có phân phối chuẩn tắc. Do đó với mức ý nghĩa α đã cho và đối thiết hai phía:

H_1: "Có sự khác nhau"

Ta sẽ bác bỏ H_0 khi $|Z| > z_{\alpha/2}$.

Còn đối thuyết một phía

H_1: "Phương pháp thứ nhất có hiệu quả hơn phương pháp thứ hai"

Thì ta sẽ bác bỏ H_0 khi $Z > z_\alpha$ (ở đây z_α kí hiệu phân vị mức α của phân phối chuẩn tắc).

Vi dụ 6.32

Một thầy giáo dạy Toán cho rằng việc cho học sinh ôn tập 1 tiết cuối kỳ kiểm tra có tác dụng tốt đến kết quả học tập của các em. Một mẫu gồm 21 học sinh được chọn để theo dõi điểm thi của các em trước và sau khi ôn tập. Kết quả ghi lại bảng sau:

Học sinh	Điểm thi trước	Điểm thi sau
1	22	21
2	26	29
3	17	15
4	20	20
5	28	26
6	31	32

7	23	25
8	13	14
9	19	19
10	25	27
11	28	27
12	24	25
13	27	27
14	18	20
15	20	23
16	14	16
17	24	26
18	15	20
19	19	20
20	18	17
21	27	29

Trên cơ sở khảo sát này có thể kết luận rằng sau khi ôn tập, kết quả thi của các em có tốt hơn không ? Với mức ý nghĩa $\alpha = 5\%$.

Giải

Kí hiệu p là tỷ lệ học sinh có điểm thi sau cao hơn điểm thi trước. Ta có bài toán kiểm định giả thuyết

$$H_0: p = 0,5$$

Và đối thuyết một phía

$$H_1: p > 0,5$$

Kí hiệu d là hiệu số giữa điểm thi sau và điểm thi trước. Ta có bảng sau đây:

Học sinh	Hiệu số d	Dấu của d
1	-1	-
2	3	+
3	-2	-
4	0	0
5	-2	-
6	1	+
7	2	+
8	1	+
9	0	0
10	2	+
11	-1	-
12	1	+
13	0	0
14	2	+

15	3	+
16	2	+
17	2	+
18	5	+
19	1	+
20	-1	-
21	2	+

Ta có $\tilde{n} = 18;\ n^+ = 13;\ f = \dfrac{13}{18} = 0,722;\ Z = \dfrac{2n^+ - \tilde{n}}{\sqrt{\tilde{n}}} = \dfrac{8}{4,242} = 1,886$.

Với mức ý nghĩa $\alpha = 0,05$ ta có $z_{0,05} = 1,64$ tra bảng phu lục 7

Ta có $Z > z_\alpha$, vậy bác bỏ H_0. Nghĩa là việc cho học sinh ôn tập có tác dụng cải tiến kết quả học tập của các bạn.

Chú ý: Nếu chúng ta có thêm thông tin để cho rằng điểm thi môn Toán của các em học sinh có phân phối chuẩn hay xấp xỉ chuẩn với phương sai bằng nhau (hoặc khác nhau), thì sử dụng kiểm định t cho mẫu bắt cặp (paired t-test) như sau:

Lệnh trong R

> t.test(data1, data2, paired=TRUE))# hai nhóm có phương sai khác nhau

Paired t-test

data: data1 and data2

t = -2.46, df = 20, p-value = 0.02312 # kiểm định t, bậc tự do 20 và giá trị p_value

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -1.7599657 -0.1447962

sample estimates:

mean of the differences -0.952381

+ Nhận xét: p-value = 0.02312< $\alpha = 5\%$ nên đối thiết H_1 có ý nghĩa thống kê.

6.4.3. Kiểm định hạng có dấu Wilcoxon (Wilcoxon signed rank test)

Trong khi tiêu chuẩn dấu chỉ quan tâm tới dấu của các hiệu số d_i, thì trong tiêu chuẩn hạng có dấu Wilcoxon ta còn tính đến độ lớn của $|d_i|$. Như vậy tiêu chuẩn này sẽ hiệu quả hơn tiêu chuẩn dấu.

6.4.3.1. Kiểm định dấu và hạng Wilcoxon về trung vị của 1 tổng thể

• Trường hợp 1: Mẫu nhỏ ($n \le 20$)

B1: Đặt giả thuyết

(1)	(2)	(3)
$\begin{cases} H_0: Med = Med_0 \\ H_1: Med \neq Med_0 \end{cases}$	$\begin{cases} H_0: Med = Med_0 \\ H_1: Med > Med_0 \end{cases}$	$\begin{cases} H_0: Med = Med_0 \\ H_1: Med < Med_0 \end{cases}$

B2: Tính chệnh lệch D_i giữa giá trị quan sát và giá trị trung vị $D_i = X_i - Med_0$

B3: Lấy trị tuyệt đối $|D_i|$

B4: Xếp hạng cho $|D_i|$,

+ Nếu $|D_i| = 0$ thì không xếp hạng

+ Nếu $|D_i| \neq 0$ thì nguyên tắc xếp hạng như sau:

- Giá trị $|D_i|$ nhỏ nhất xếp hạng 1, lớn nhất xếp hạng n,
- Nếu tồn tại các $|D_i|$ bằng nhau thì tính hạng trung bình cho tất cả các $|D_i|$ này.

B5: Thêm 2 cột R+ và R-:

+ R+: gồm những hạng của $D_i > 0$

+ R-: gồm những hạng của $D_i < 0$

B6: Tính giá trị kiểm định W

Kiểm định 2 bên (1): $W = \min\{\sum R+; \sum R-\}$

Kiểm định bên phải (2): $W = \sum R+$

Kiểm định bên trái (3): $W = \sum R-$

B7: Tra bảng phụ lục 11, tìm giá trị cận dưới và cận trên (W_L, W_M) của $W_{\alpha/2}(n')$ (KĐ 2 bên) hoặc $W_\alpha(n')$ (KĐ 1 bên) với n'=số lượng $D_i \neq 0$.

(1)	$\begin{cases} H_0: Med = Med_0 \Leftrightarrow W \in (W_L, W_M) \\ H_1: Med \neq Med_0 \Leftrightarrow W \notin (W_L, W_M) \end{cases}$
(2)	$\begin{cases} H_0: Med = Med_0 \Leftrightarrow W \leq W_M \\ H_1: Med > Med_0 \Leftrightarrow W > W_M \end{cases}$
(3)	$\begin{cases} H_0: Med = Med_0 \Leftrightarrow W \geq W_L \\ H_1: Med < Med_0 \Leftrightarrow W < W_L \end{cases}$

Kiểm định 2 bên (1): Bác bỏ H_0 nếu $W \notin (W_L, W_M)$,

Kiểm định bên phải (2): Bác bỏ H_0 nếu $W > W_M$,

Kiểm định bên trái (3): Bác bỏ H_0 nếu $W < W_L$.

Ví dụ 6.33

Khảo sát mẫu ngẫu nhiên gồm thu nhập của 11 sinh viên đã tốt nghiệp đi làm việc được 2 năm được ghi nhận ở bảng sau:

Đơn vị tính: đô la

Sinh viên	1	2	3	4	5	6	7	8	9	10	11
Mức lương	300	320	340	380	420	400	300	340	360	400	410

Với mức ý nghĩa 5%, hãy cho kết luận về giả thuyết cho rằng thu nhập sinh viên đã tốt nghiệp đi làm việc được 2 năm vượt quá con số 350 đô la?

Giải

Đặt giả thuyết:

$$\begin{cases} H_0 : Med = 350 \\ H_1 : Med > 350 \end{cases}$$

• Tính D_i và $|D_i|$

X_i	D_i	$\|D_i\|$
300	-50	50
320	-30	30
340	-10	10
380	30	30
420	70	70
400	50	50
300	-50	50
340	-10	10
360	10	10
400	50	50
410	60	60

• Xếp hạng $|D_i|$

Xếp thứ tự	$\|D_i\|$	Hạng
1	10	2
2	10	2
3	10	2
4	30	4,5
5	30	4,5
6	50	7,5
7	50	7,5
8	50	7,5
9	50	7,5
10	60	10
11	70	11

• Thêm R+ và R-

X_i	D_i	$\|D_i\|$	Hạng $\|D_i\|$	R+	R-
300	-50	50	7,5		7,5
320	-30	30	4,5		4,5
340	-10	10	2		2
380	30	30	4,5	4,5	

420	70	70	11	11	
400	50	50	7,5	7,5	
300	-50	50	7,5		7,5
340	-10	10	2		2
360	10	10	2	2	
400	50	50	7,5	7,5	
410	60	60	10	10	

Tính W : do KĐ 1 bên phải nên $W = 42,5$

Tra bảng phụ lục 11, ta có $W_{\alpha(n)} = W_{0,05(11)} = (13,53)$

So sánh: $W = 42,5 \le W_M = 53$ nên chấp nhận H_0.

Kết luận: Chưa thể nói rằng thu nhập sinh viên đã tốt nghiệp đi làm việc được 2 năm vượt quá con số 350 đô la.

• Trường hợp 2: Mẫu lớn ($n > 20$)

Giá trị kiểm định W sẽ xấp xỉ phân phối chuẩn với giá trị kiểm định Z tính theo công thức sau:

B6: Tính giá trị kiểm định W

Kiểm định 2 bên (1): $W = \min\{\sum R+; \sum R-\}$

Kiểm định bên phải (2): $W = \sum R+$

Kiểm định bên trái (3): $W = \sum R-$

Tính $Z = \dfrac{W - \mu_W}{\sigma_W}$

Với $\mu_W = \dfrac{n'(n'+1)}{4}$, $\sigma_W = \sqrt{\dfrac{n'(n'+1)(2n'+1)}{24}}$, n' =số lượng $D_i \ne 0$

B7: Quy tắc bác bỏ

(1)	$\begin{cases} H_0 : Med = Med_0 \Leftrightarrow \lvert Z \rvert \le z_{\alpha/2} \\ H_1 : Med \ne Med_0 \Leftrightarrow \lvert Z \rvert > z_{\alpha/2} \end{cases}$
(2)	$\begin{cases} H_0 : Med = Med_0 \Leftrightarrow Z \le z_{\alpha} \\ H_1 : Med > Med_0 \Leftrightarrow Z > z_{\alpha} \end{cases}$
(3)	$\begin{cases} H_0 : Med = Med_0 \Leftrightarrow Z \ge -z_{\alpha} \\ H_1 : Med < Med_0 \Leftrightarrow Z < -z_{\alpha} \end{cases}$

KĐ (1): Bác bỏ H_0 khi $\lvert Z \rvert > z_{\alpha/2}$

KĐ (2): Bác bỏ H_0 khi $Z > z_{\alpha}$

KĐ (3): Bác bỏ H_0 khi $Z < -z_{\alpha}$

Ví dụ 6.34

Kiểm tra trọng lượng của một số quả, người ta có bảng kết quả sau:

$X_i(g)$	300	320	340	380	420	400	300	340	360	400	410
$X_i(g)$	310	330	350	370	390	400	320	360	410	390	340

Với mức ý nghĩa 5%, có thể cho rằng Med=350(g)?

Giải

Đặt giả thuyết:

$$\begin{cases} H_0 : Med = 350 \\ H_1 : Med \neq 350 \end{cases}$$

• Tính D_i và $|D_i|$

$X_i(g)$	D_i	$\|D_i\|$
300	-50	50
320	-30	30
340	-10	10
380	30	30
420	70	70
400	50	50
300	-50	50
340	-10	10
360	10	10
400	50	50
410	60	60
310	-40	40
330	-20	20
350	0	0
370	20	20
390	40	40
400	50	50
320	-30	30
360	10	10
410	60	60
390	40	40
340	-10	10

• Xếp hạng $|D_i|$

Xếp thứ tự	$\| D_i \|$	Hạng
1	10	3
2	10	3
3	10	3
4	10	3
5	10	3
6	20	6,5
7	20	6,5
8	30	9
9	30	9
10	30	9
11	40	12
12	40	12
13	40	12
14	50	16
15	50	16
16	50	16
17	50	16
18	50	16
19	60	19,5
20	60	19,5
21	70	21

• Thêm R+ và R-

$X_i(g)$	D_i	$\| D_i \|$	Hạng $\| D_i \|$	R+	R-
300	-50	50	16		16
320	-30	30	9		9
340	-10	10	3		3
380	30	30	9	9	
420	70	70	21	21	
400	50	50	16	16	
300	-50	50	16		16
340	-10	10	3		3
360	10	10	3	3	
400	50	50	16	16	
410	60	60	19,5	19,5	
310	-40	40	12		12
330	-20	20	6,5		6,5

350	0	0			
370	20	20	6,5	6,5	
390	40	40	12	12	
400	50	50	16	16	
320	-30	30	9		9
360	10	10	3	3	
410	60	60	19,5	19,5	
390	40	40	12	12	
340	-10	10	3		3
Tổng				153,5	77,5

Tính W (do KĐ 2 bên nên)

Ta có W=min(153,5; 77,5)=77,5

$n' = 21$

Test thống kê Z là

Tính $Z = \dfrac{W - \mu_W}{\sigma_W} = \dfrac{77,5 - 115}{28,77} = -1,3$

$$\mu_W = \frac{n'(n'+1)}{4} = \frac{21 \times 22}{4} = 115,5,\ \sigma_W = \sqrt{\frac{n'(n'+1)(2n'+1)}{24}} = \sqrt{\frac{21 \times 22 \times 43}{24}} = 28,77$$

Với mức ý nghĩa $\alpha = 0,05$ ta có $z_{\alpha/2} = 1,96$

Vì $|Z| = 1,3 < z_{\alpha/2}$ nên chấp nhận H_0.

Vậy trung vị trọng lượng của một số quả là 350 g.

6.4.3.2. Kiểm định dấu và hạng Wilcoxon về sự bằng nhau của 2 trung vị trong trường hợp mẫu từng cặp

• Trường hợp 1: Mẫu nhỏ ($n' \le 20$)

Các bước tiến hành thực hiện như sau:

B1: Đặt giả thuyết

(1)	(2)	(3)
$\begin{cases} H_0 : Med_d = 0 \\ H_1 : Med_d \ne 0 \end{cases}$	$\begin{cases} H_0 : Med_d = 0 \\ H_1 : Med_d > 0 \end{cases}$	$\begin{cases} H_0 : Med_d = 0 \\ H_1 : Med_d < 0 \end{cases}$

Với $Med_d = Med_1 - Med_2$

B2: Tính chệnh lệch D_i giữa giá trị quan sát của 2 mẫu $D_i = x_i - y_i$

B3: Lấy trị tuyệt đối $|D_i|$

B4: Xếp hạng cho $|D_i|$,

+ Nếu $|D_i| = 0$ thì không xếp hạng

+ Nếu $|D_i| \ne 0$ thì nguyên tắc xếp hạng như sau:

+ Giá trị $|D_i|$ nhỏ nhất xếp hạng 1, lớn nhất xếp hạng n,

+ Nếu tồn tại các $|D_i|$ bằng nhau thì tính hạng trung bình cho tất cả các $|D_i|$ này.

B5: Thêm 2 cột R+ và R-:

+ R+: gồm những hạng của $D_i > 0$

+ R-: gồm những hạng của $D_i < 0$

B6: Tính giá trị kiểm định W

Kiểm định 2 bên (1): $W = \min(\sum R+; \sum R-)$

Kiểm định bên phải (2): $W = \sum R+$

Kiểm định bên trái (3): $W = \sum R-$

B7: Tra bảng phụ lục 11, tìm giá trị cận dưới và cận trên (W_L, W_M) của $W_{\alpha/2}(n')$ (KĐ 2 bên) hoặc $W_\alpha(n')$ (KĐ 1 bên) với n'=số lượng $D_i \neq 0$.

(1)	$\begin{cases} H_0: Med_d = 0 \Leftrightarrow W \in (W_L, W_M) \\ H_1: Med_d \neq 0 \Leftrightarrow W \notin (W_L, W_M) \end{cases}$
(2)	$\begin{cases} H_0: Med_d = 0 \Leftrightarrow W \leq W_M \\ H_1: Med_d > 0 \Leftrightarrow W > W_M \end{cases}$
(3)	$\begin{cases} H_0: Med_d = 0 \Leftrightarrow W \geq W_L \\ H_1: Med_d < 0 \Leftrightarrow W < W_L \end{cases}$

Kiểm định 2 bên (1): Bác bỏ H_0 nếu $W \notin (W_L, W_M)$,

Kiểm định bên phải (2): Bác bỏ H_0 nếu $W > W_M$,

Kiểm định bên trái (3): Bác bỏ H_0 nếu $W < W_L$.

Ví dụ 6.35

Để kiểm tra hiệu quả của 1 khoá học, người ta theo dõi kĩ khả năng đọc của trẻ em trước và sau khi học. Kết quả như sau:

Trước	60	40	78	53	67	88	77	60	64	75
Sau	63	38	77	50	74	96	80	70	65	75

Với mức ý nghĩa 5%, khoá học này có hiệu quả hay không?

Giải

Đặt giả thuyết

$$\begin{cases} H_0: Med_d = 0 \\ H_1: Med_d < 0 \end{cases}$$

Với Med_d = trước - sau

B1: Đặt X_1: trước X_2: sau

$D_i = X_{1i} - X_{2i}$

B2: $|D_i|$

B3: Xếp hạng cho $|D_i|$, $|D_i| = 0$ thì không xếp hạng.

B4: Thêm cột R- gồm những hạng của $D_i < 0$

B5: $W = \sum R- = 35,5$

Trước	Sau	D_i	$\|D_i\|$	Hạng	R-
60	63	-3	3	5	5
40	38	2	2	3	
78	77	1	1	1,5	
53	50	3	3	5	
67	74	-7	7	7	7
88	96	-8	8	8	8
77	80	-3	3	5	5
60	70	-10	10	9	9
64	65	-1	1	1,5	1,5
75	75	0	0		
Tổng					35,5

B6: Quy tắc bác bỏ

Tra bảng phụ lục 11, ta có $W_{\alpha(n')} = W_{0,05(9)} = (8,37)$

Kiểm định (3): (KĐ bên trái)

Ta có $W = 35,5 > W_L = 8$, chấp nhận H_0.

Vậy khoá học này chưa có hiệu quả.

Các lệnh trong R

> data3<-c(60,40,78,53,67,88,77,60,64,75)

> data4<-c(63,38,77,50,74,96,80,70,65,75)

> wilcox.test(data3, data4, paired=TRUE, alt="l") # Kiểm định Wilcoxon cho hai mẫu từng cặp

Wilcoxon signed rank test with continuity correction

data: data3 and data4

V = 9.5, p-value = 0.06845 # kiểm định Wilcoxon hạng có dấu và giá trị p_value

alternative hypothesis: true location shift is less than 0

+ Nhận xét: p-value = 0.06845> $\alpha = 5\%$ nên đối thiết H_1 không có ý nghĩa thống kê, tứ là chấp nhận H_0.

Kết luận: Khoá học này chưa có hiệu quả.

• Trường hợp 2: Mẫu lớn ($n' > 20$)

B6: Tính giá trị kiểm định W

Kiểm định 2 bên (1): $W = \min\{\sum R+; \sum R-\}$

Kiểm định bên phải (2): $W = \sum R+$

Kiểm định bên trái (3): $W = \sum R-$

Tính $Z = \dfrac{W - \mu_W}{\sigma_W}$

Với $\mu_W = \dfrac{n'(n'+1)}{4}$, $\sigma_W = \sqrt{\dfrac{n'(n'+1)(2n'+1)}{24}}$, n'=số lượng $D_i \neq 0$

B7: Quy tắc bác bỏ

(1)	$\begin{cases} H_0 : Med_d = 0 \Leftrightarrow \lvert Z \rvert \le z_{\alpha/2} \\ H_1 : Med_d \neq 0 \Leftrightarrow \lvert Z \rvert > z_{\alpha/2} \end{cases}$
(2)	$\begin{cases} H_0 : Med_d = 0 \Leftrightarrow Z \le z_{\alpha} \\ H_1 : Med_d > 0 \Leftrightarrow Z > z_{\alpha} \end{cases}$
(3)	$\begin{cases} H_0 : Med_d = 0 \Leftrightarrow Z \ge -z_{\alpha} \\ H_1 : Med_d < 0 \Leftrightarrow Z < -z_{\alpha} \end{cases}$

KĐ (1): Bác bỏ H_0 khi $\lvert Z \rvert > Z_{\alpha/2}$

KĐ (2): Bác bỏ H_0 khi $Z > Z_{\alpha}$

KĐ (3): Bác bỏ H_0 khi $Z < -Z_{\alpha}$

Ví dụ 6.36

Có ý kiến cho rằng trong hai anh em trai người em luôn cao hơn người anh. Một mẫu ngẫu nhiên gồm 25 cặp anh em trai được chọn; chiều cao của người anh (X) và của người em (Y) được ghi nhận như sau (đơn vị là cm):

X	170 169 167 168 166
Y	175 172 167 166 163

X	165 165 164 164 165
Y	166 164 167 163 167

X	166 166 169 168 168
Y	168 164 170 172 171

X	166 168 168 169 169
Y	170 167 165 166 171

X	164 170 169 169 166
Y	163 168 166 167 163

Với mức ý nghĩa $\alpha = 5\%$, sử dụng tiêu chuẩn hạng có dấu Wilcoxon kiểm định xem liệu trung vị chiều cao của hai anh em trai là khác nhau hay giống nhau.

Giải

Đặt giả thuyết

$$\begin{cases} H_0 : Med_d = 0 \\ H_1 : Med_d \neq 0 \end{cases}$$

Từ số liệu trên ta tính $d_i = x_i - y_i$ và hạng của $|d_i| (\mathrm{d}_i \neq 0)$.

x_i	y_i	d_i	hạng của $\lvert d_i \rvert$	hạng với $d_i > 0$	hạng với $d_i < 0$
170	175	-5	23		23
169	172	-3	17		17
167	167	0	0		
168	166	+2	10	10	
166	163	+3	17	17	
165	166	-1	3,5		3,5
165	164	+1	3,5	3,5	
164	167	-3	17		17
164	163	+1	3,5	3,5	
165	167	-2	10		10
166	168	-2	10		10
166	164	+2	10	10	
169	170	-1	3,5		3,5
168	172	-4	21,5		21,5
168	171	-3	17		17
166	170	-4	21,5		21,5
168	167	+1	3,5	3,5	
168	165	+3	17	17	
169	166	+3	17	17	
169	171	-2	10		10
164	163	+1	3,5	3,5	
170	168	+2	10	10	
169	166	+3	17	17	
169	167	+2	10	10	
166	166	0	0		
				R^+=122	R^-=154

Chú thích (về cách tính hạng). Ta sắp xếp các $|d_i|$ theo thứ tự tăng dần:

1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 5.

Số 1 có hạng là $\dfrac{1+2+3+4+5+6}{6} = 3,5$

Số 2 có hạng là $\frac{7+8+9+10+11+12+13}{7}=10$

Số 3 có hạng là $\frac{14+15+16+17+18+19+20}{7}=17$

Số 4 có hạng là $\frac{21+22}{2}=21,5$

Số 5 có hạng là 23

Ta có W=min(122, 154)=122

$n'=23$

Test thống kê Z là

$$Z=\frac{122-23\times(23+1)/4}{\sqrt{\frac{23\times(23+1)\times(40+1)}{24}}}=\frac{122-138}{32,88}=-0,4866$$

Với mức ý nghĩa $\alpha=0,05$ ta có $z_{\alpha/2}=1,96$

Vì $|Z|=0,4866<z_{\alpha/2}$ nên ta không có cơ sở bác bỏ H_0. Vậy chiều cao trung bình của hai anh em trai không khác nhau.

Các lệnh trong R

> data5<-c(170, 169, 167, 168, 166, 165, 165, 164,164, 165, 166, 166,169,168,168, 166, 168,168,169,169, 164,170,169,169,166)

> data6<-c(175,172,167,166,163,166,164,167,163,167, 168, 164,170,172,171, 170,167, 165,166,171, 163,168,166,167,163)

> wilcox.test(data5, data6, paired=TRUE)

Wilcoxon signed rank test with continuity correction

data: data5 and data6

V = 141.5, p-value = 0.8176 # kiểm định Wilcoxon hạng có dấu và giá trị p_value

alternative hypothesis: true location shift is not equal to 0

+ Nhận xét: p-value = 0.8176 > $\alpha=5\%$ nên đối thiết H_1 không có ý nghĩa thống kê, tứ là chấp nhận H_0.

Kết luận: Chiều cao trung bình của hai anh em trai không khác nhau.

6.4.4. Tổng hạng Wilcoxon (Wilcoxon Rank-Sum Test - còn gọi là Mann-Whitney)

Giả sử ta có hai mẫu ngẫu nhiên độc lập nhau:

Mẫu thứ nhất $\{x_1,x_2,...,x_n\}$ là n quan sát độc lập về ĐLNN X, còn mẫu thứ hai $\{y_1,y_2,...,y_n\}$ là n quan sát độc lập về ĐLNN Y. Phân phối của X và Y chưa biết và không nhất thiết là phân phối chuẩn.

• Các bước kiểm định

Bước 1: Đặt giả thuyết

(1)	(2)	(3)
$\begin{cases} H_0 : Med_1 = Med_2 \\ H_1 : Med_1 \neq Med_2 \end{cases}$	$\begin{cases} H_0 : Med_1 = Med_2 \\ H_1 : Med_1 > Med_2 \end{cases}$	$\begin{cases} H_0 : Med_1 = Med_2 \\ H_1 : Med_1 < Med_2 \end{cases}$

Bước 2: Xếp hạng tất cả các giá trị của 2 mẫu theo thứ tự tăng dần. Những giá trị bằng nhau sẽ nhận giá trị trung bình.

Bước 3: có 2 trường hợp

Trường hợp mẫu nhỏ ($n_1, n_2 \leq 10$)	Trường hợp mẫu lớn ($n_1 + n_2 > 20$)
Lấy tổng hạng T_1 của mẫu nhỏ. Nếu hai mẫu bằng nhau thì lấy tổng hạng của mẫu nào cũng được B3: Quy tắc bác bỏ Tra bảng phụ lục 12 để tìm giới hạn trên và dưới của $W_{\alpha/2(m,n)}$ (KĐ 2 bên) và $W_{\alpha(m,n)}$ (KĐ 1 bên) KĐ (1): Bác bỏ H_0 khi $T_1 \notin (W_L, W_M)$ KĐ (2): Bác bỏ H_0 khi $T_1 > W_M$ KĐ (3): Bác bỏ H_0 khi $T_1 < W_L$	Lấy tổng hạng T_1 của mẫu nhỏ. Phân phối của T_1 được xem như chuẩn với $\mu_{T_1} = \frac{n_1(n+1)}{2}; \sigma_{T_1} = \sqrt{\frac{n_1 n_2 (n+1)}{12}}$ Tính $Z = \frac{T_1 - \mu_{T_1}}{\sigma_{T_1}}$ KĐ (1): Bác bỏ H_0 khi $\lvert Z \rvert > z_{\alpha/2}$ KĐ (2): Bác bỏ H_0 khi $Z > z_\alpha$ KĐ (3): Bác bỏ H_0 khi $Z < -z_\alpha$

• Trường hợp 1: Mẫu nhỏ ($n_1, n_2 \leq 10$)

Ví dụ 6.37

Theo dõi doanh thu bán hàng của 2 nhân viên. Với mức ý nghĩa 5%, có thể cho rằng doanh thu bán hàng là như nhau?

NV1	NV2
60	63
61	64
63	67
72	40
68	50
70	90
80	80
90	70
85	85
	93

Giải

B1: Đặt giả thuyết

$$\begin{cases} H_0: Med_1 = Med_2 \\ H_1: Med_1 \neq Med_2 \end{cases}$$

B2: Xếp hạng theo thứ tự tăng dần.

Lấy tổng hạng T_1 của mẫu nhỏ. Nếu 2 mẫu bằng nhau thì lấy tổng hạng của mẫu nào cũng được.

NV1	Hạng	NV2	Hạng
60	3	63	5,5
61	4	64	7
63	5,5	67	8
72	12	40	1
68	9	50	2
70	10,5	90	17,5
80	13,5	80	13,5
90	17,5	70	10,5
85	15,5	85	15,5
		93	19
Tổng	90,5		99,5

B3: Quy tắc bác bỏ

Kiểm định (1): bác bỏ H_0 khi $T_1 \notin (W_L; W_M)$

Theo bảng tính , ta có T_1=90,5

Dùng bảng tra phụ lục 12, ta có $W_{\alpha/2(m,n)} = W_{0,025(9,10)} = (65;115)$.

So sánh: $T_1 = 90,5 \in (65;\, 115)$. Nên chấp nhận H_0.

Kết luận: doanh thu bán hàng của 2 nhân viên là như nhau.

Các lệnh trong R

```
> data1<-c(60,61,63,72,68,70,80,90,85)
> data2<-c(63,64,67,40,50,90,80,70,85,93)
> wilcox.test(data1,data2) # Kiểm định Wilcoxon cho hai mẫu độc lập
        Wilcoxon rank sum test with continuity correction
data:  data1 and data2
W = 45.5, p-value = 1 # kiểm định tổng hạng Wilcoxon và giá trị p_value
alternative hypothesis: true location shift is not equal to 0
```

+ Nhận xét: p-value = 1 > $\alpha = 5\%$ nên đối thiết H_1 không có ý nghĩa thống kê (chấp nhận H_0).

Kết luận: doanh thu bán hàng của 2 nhân viên là như nhau.

• **Trường hợp 2: Mẫu lớn (**$n_1 + n_2 > 20$ **hay m+n >20)**

Tiêu chuẩn Mann-Whitney được xây dựng như sau:

i) Gộp hai mẫu trên thành một mẫu với cỡ mẫu là n + m .

ii) Sắp xếp n + m giá trị $\{x_1, x_2, ..., x_n, y_1, y_2, ..., y_n\}$ theo thứ tự tăng dần. Giả sử sau khi sắp xếp thu được dãy sau đây.

$$c_1 < c_2 < .... < c_k < .... < c_{n+m}$$

Nếu $x_i = c_k$ thì ta nói hạng của x_i là k. Tương tự $y_j = c_k$ thì ta nói hạng của y_j là k.

iii) Giả sử x_i có hạng là r_i (i=1,2…,n). Ta tính tổng các hạng của x_i

$$T_1 = r_1 + r_2 + ... + r_n$$

Giả sử y_j có hạng là s_j (j=1,2…,n). Ta tính tổng các hạng của y_j

$$T_2 = s_1 + s_2 + ... + s_m$$

Hiển nhiên

$$T_1 + T_2 = r_1 + r_2 + ... + r_n + s_1 + s_2 + ... + s_m = 1 + 2 + ... + (\mathrm{n+m}) = \frac{(\mathrm{n+m+1})(\mathrm{n+m})}{2}$$

Người ta chứng minh được rằng nếu H_0 đúng và $n, m \geq 8$ thì T_1 có phân bố xấp xỉ chuẩn với giá trị trung bình là:

$$\mu_{T_1} = \frac{n(\mathrm{n+m+1})}{2}$$

Và phương sai là $$\sigma^2_{T_1} = \frac{nm(\mathrm{n+m+1})}{12}$$

Tương tự T_2 có phân bố xấp xỉ chuẩn với giá trị trung bình là:

$$\mu_{T_2} = \frac{\mathrm{m(n+m+1)}}{2}$$

Và phương sai là

$$\sigma^2_{T_2} = \frac{nm(\mathrm{n+m+1})}{12}$$

Thông thường chúng ta chọn số nhỏ nhất giữa R_1 và R_2. Giả sử $T_1 < T_2$, khi đó test thống kê ta sử dụng là

$$Z = \frac{T_1 - \mu_{T_1}}{\sigma_{T_1}}$$

sẽ có phân phối chuẩn tắc. Quy tắc bác bỏ:

KĐ (1): Bác bỏ H_0 khi $| Z |> z_{\alpha/2}$

KĐ (2): Bác bỏ H_0 khi $Z > z_\alpha$

KĐ (3): Bác bỏ H_0 khi $Z < -z_\alpha$

Ví dụ 6.38

Một người lái xe thường xuyên đi lại giữa hai địa điểm A và B. Có hai con đường nối A và B: đường X và đường Y. Anh ta muốn chọn con đường đi nào mất ít thời gian nhất. Chọn ngẫu nhiên 10 ngày đi trên đường X và 10 ngày đi trên đường Y, anh ta có số liệu sau đây (thời gian tính bằng phút):

Đường X: 34, 28, 46, 42, 56, 85, 48, 25, 37, 49.

Đường Y: 43, 49, 41, 55, 39, 45, 65, 50, 47, 51.

Với mức ý nghĩa $\alpha = 5\%$, hãy nhận định xem có sự khác nhau về thời gian đi lại khi sử dụng đường X và đường Y hay không?

Giải

Đầu tiên ta nhận xét rằng thời gian trung bình đi trên đường X là 45 phút, trong khi thời gian trung bình đi trên đường Y là 48,5 phút. Tuy nhiên ta không có cơ sở để cho rằng thời gian đi trên đường X và thời gian đi trên đường Y có phân phối chuẩn hay xấp xỉ chuẩn với phương sai bằng nhau. Do đó việc áp dụng test thống kê Student đã trình bày ở phần trước là không phù hợp. Thành thử ta hãy áp dụng tiêu chuẩn hạng Mann – Whitney.

Ta kiểm định

Giả thuyết H_0: Thời gian đi lại khi sử dụng đường X và đường Y là như nhau

Đối thuyết H_1: Thời gian đi lại khi sử dụng đường X và đường Y là khác nhau

Đầu tiên ta lập bảng xếp hạng các số liệu:

Hạng	Thời gian	Đường
1	25	X
2	28	X
3	34	X
4	37	X
5	39	Y
6	41	Y
7	42	X
8	43	Y
9	45	Y
10	46	X
11	47	Y
12	48	X
13	49	X
14	49	Y
15	50	Y
16	51	Y
17	55	Y
18	56	X
19	65	Y
20	85	X

Ta thấy có hai số liệu trùng nhau đều bằng 49. chúng ở vị trí 13 và 14 do đó ta gán cho chúng cùng hạng 13,5.

Tổng các hạng của đường X là

T_1=1+2+3+4+7+10+12+13,5+18+20=90,5

Vì n=10 và m=10 lớn hơn 8 nên T_1 có phân phối xấp xỉ chuẩn với kỳ vọng

$\mu_{T_1} = \frac{10(10+10+1)}{2} = 105$ và phương sai: $\sigma^2_{T_1} = \frac{10 \times 10(10+10+1)}{12} = 175$

Giá trị của test thống kê là

$$Z = \frac{T_1 - \mu_{T_1}}{\sigma_{T_1}} = \frac{90,5-105}{\sqrt{175}} = -1,1$$

Với mức ý nghĩa $\alpha = 0,05$ ta có $z_{\alpha/2} = 1,96$ tra bảng phụ lục 6

Ta có $|Z| = 1,1 < z_{\alpha/2} = 1,96$, do đó ta không có cơ sở bác bỏ H_0. Chúng ta tạm thời kết luận rằng thời gian đi đường giữa hai con đường X và Y không khác nhau.

Các lệnh trong R

> data1<-c(34, 28, 46, 42, 56, 85, 48, 25, 37, 49)

> data2<-c(43, 49, 41, 55, 39, 45, 65, 50, 47, 51)

> wilcox.test(data1,data2) # Kiểm định Wilcoxon cho hai mẫu độc lập

Wilcoxon rank sum test with continuity correction

data: data1 and data2

W = 35.5, p-value = 0.2897 # kiểm định tổng hạng Wilcoxon và giá trị p_value

alternative hypothesis: true location shift is not equal to 0

+ Nhận xét: p-value = 0.2897 > $\alpha = 5\%$ nên đối thiết H_1 không có ý nghĩa thống kê (chấp nhận H_0).

Kết luận: rằng thời gian đi đường giữa hai con đường X và Y không khác nhau.

Chú ý: Nếu chúng ta có thêm thông tin để cho rằng thời gian đi trên đường X và thời gian đi trên đường Y có phân phối chuẩn hay xấp xỉ chuẩn với phương sai bằng nhau (hoặc khác nhau), thì sử dụng kiểm định t.test hai mẫu như sau:

Lệnh trong R

> t.test(data1,data2)# hai nhóm có phương sai khác nhau

Welch Two Sample t-test

data: data1 and data2

t = -0.59213, df = 12.391, p-value = 0.5644

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-16.333788 9.333788

sample estimates:

mean of x mean of y 45.0 48.5

+ Nhận xét: p-value = 0.5644 > $\alpha = 5\%$ nên đối thiết H_1 không có ý nghĩa thống kê (chấp nhận H_0).

> t.test(data1,data2,var.equal=TRUE)# hai nhóm có phương sai bằng nhau

Two Sample t-test

data: data1 and data2

t = -0.59213, df = 18, p-value = 0.5611

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-15.918367 8.918367

sample estimates:

mean of x mean of y 45.0 48.5

+ Nhận xét: p-value = 0.5644 > $\alpha = 5\%$ nên đối thiết H_1 không có ý nghĩa thống kê (chấp nhận H_0).

6.4.5. Kiểm định Kruskal-Wallis

Trong phân tích phương sai một yếu tố để có thể thực hiện được các bước kiểm định đòi hỏi phải xảy ra ba giả thuyết quan trọng là tổng thể có phân phối chuẩn, phương sai giữa các tổng thể bằng nhau và mẫu được chọn ngẫu nhiên, độc lập. Các giả thuyết này trong thực tế khó đạt được, nhất là khi dữ liệu thuộc thang đo thứ bậc (ordinal) hoặc khoảng (interval) là những thang đo rất phổ biến trong một số lĩnh vực như quản trị kinh doanh; vì vậy yêu cầu đặt ra là cần có một kiểm định mà sự yêu cầu các giả thuyết trên hoặc không có hoặc có thể giảm bớt đi. Vì thế kiểm định phi tham số Kruskal-Wallis là một sự lựa chọn phù hợp khi phân tích sự phụ thuộc giữa một biến định tính (biến nguyên nhân) và biến định lượng (biến kết quả). Các bước của kiểm định phi tham số Kruskal-Wallis được thực hiện như sau:

Bước 1: Xây dựng cặp giả thuyết

$$\begin{cases} H_0 : \mu_1 = \mu_2 = ... = \mu_k \\ H_1 : \exists!\left(\mu_i \neq \mu_j\right) \end{cases}$$

Với giả thuyết H_0 hàm ý rằng không có sự khác biệt về mặt trung bình giữa các nhóm tính chất của biến nguyên nhân dựa trên giá trị thu nhận của biến kết quả và giả thuyết H_1 hàm ý rằng tồn tại ít nhất hai nhóm tính chất của biến nguyên nhân có sự khác biệt về mặt trung bình.

Bước 2: Tính giá trị kiểm định $\chi_0^2 = \dfrac{12}{n(n+1)}\sum_{i=1}^{k}\dfrac{R_i^2}{n_i} - 3(n+1)$ với

- R_i là tổng hạng của mẫu có nhóm tính chất i của biến nguyên nhân dựa trên hảng của mẫu kết hợp.
- k là số nhóm tính chất.
- n là tổng số phần tử khảo sát trên toàn bộ các nhóm tính chất.

Bước 3: So sánh và kết luận

- Nếu $\chi_0^2 < \chi_{\alpha;k-1}^2$(tra bảng phụ lục 5) thì chấp nhận H_0 và ngược lại.

Trong trường hợp giả thuyết H_0 bị bác bỏ thì có sự phụ thuộc giữa biến nguyên nhân và biến kết quả. Lúc này chúng ta sẽ tiếp tục tìm hiểu là sự phụ thuộc nếu xảy ra thì sẽ xảy ra trên các nhóm tính chất nào của biến nguyên nhân. Quá trình thực hiện này hoàn toàn tương tự như phân tích sâu phương sai một yếu tố với các bước thực hiện cụ thể sau:

Bước 4:

- Xác định số cặp giả thuyết cần kiểm định qua công thức:

$$C_k^2 = \frac{k!}{2!(k-2)!}$$

- Xây dựng các cặp giả thuyết cần kiểm định

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases};...;\begin{cases} H_0 : \mu_{k-1} = \mu_k \\ H_1 : \mu_{k-1} \neq \mu_k \end{cases}$$

Bước 5:

- Tính hạng trung bình cho từng nhóm tính chất: $\overline{R_i} = \dfrac{R_i}{n_i}$

- Tính chênh lệch về hạng trung bình giữa hai nhóm tính chất cần kiểm định:

$D_{ij} = \left|\overline{R_i} - \overline{R_j}\right|$.

Bước 6: Tính giá trị kiểm định theo công thức:

$$C_{ij} = \sqrt{\chi_{\alpha;k-1}^2\left(\frac{n(n+1)}{12}\right)\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

C_{ij} thay đổi giá trị tuỳ theo cặp giả thuyết cần kiểm định.

Bước 7: So sánh và kết luận

Ứng với cặp giả thuyết $\begin{cases} H_0 : \mu_i = \mu_j \\ H_1 : \mu_i \neq \mu_j \end{cases}$, nếu $D_{ij} > C_{ij}$ thì bác bỏ H_0 và ngược lại.

Ví dụ 6.39

Để kiểm định xem việc làm thêm có ảnh hưởng đến kết quả học tập không, người ta chọn ngẫu nhiên một số sinh viên và hỏi họ về kết quả học tập với thời gian làm thêm trong các khoảng:<8 giờ/tuần; 8-16 giờ/tuần; >16 giờ/tuần thì thu được bảng số liệu sau:

Thời gian làm thêm		
<8 giờ/tuần	8-16 giờ/tuần	>16 giờ/tuần
6,3	7,0	6,3
7,2	6,6	5,8
6,5	6,1	6,0
6,6	5,8	5,5
7,3	6,8	5,3
-	7,1	6,5

Nếu các giả định về những tổng thể có phân phối chuẩn và phương sai bằng nhau không được thoả mãn thì việc kiểm định sẽ thực hiện như thế nào, với mức ý nghĩa 5%?

Giải

Bước 1: Xây dựng cặp giả thuyết

H_0: Kết quả học tập trung bình trên các nhóm thời gian làm thêm là giống nhau.

H_1: Tồn tại ít nhất hai nhóm thời gian làm thêm có kết quả học tập trung bình khác nhau.

Bước 2: Tính giá trị kiểm định

$$\chi_0^2 = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(n+1)$$

$$= \frac{12}{17(17+1)}\left(\frac{61,5^2}{5} + \frac{62,5^2}{6} + \frac{29^2}{6}\right) - 3(17+1) = 6,6925$$

với R_i được tính trong bảng sau:

Thời gian làm thêm	Hạng kết hợp
5,3	1
5,5	2
5,8	3,5
5,8	3,5
6,0	5
6,1	6
6,3	7,5
6,3	7,5
6,5	9,5
6,5	9,5
6,6	11,5
6,6	11,5
6,8	13
7,0	14

7,1	15
7,2	16
7,3	17

Thời gian làm thêm					
<8 giờ/tuần	R_1	8-16 giờ/tuần	R_2	>16 giờ/tuần	R_3
6,3	7,5	7,0	14	6,3	7,5
7,2	16	6,6	11,5	5,8	3,5
6,5	9,5	6,1	6	6,0	5
6,6	11,5	5,8	3,5	5,5	2
7,3	17	6,8	13	5,3	1
-	-	7,1	15	6,5	9,5
Tổng hạng	61,5		63		28,5

Bước 3: So sánh và kết luận

- Vì $\chi_0^2 > \chi^2_{0,05;2} = 5,99$ (tra bảng phụ lục 5) nên bác bỏ H_0.

Kết luận: Thời gian làm thêm có ảnh hưởng đến kết quả học tập.

Nhận xét: Khi bác bỏ H_0, chúng ta chỉ biết được thời gian làm thêm có ảnh hưởng đến kết quả học tập chứ hoàn toàn không thể đưa ra kết luận là thời gian làm thêm trên nhóm nào có sự khác biệt với nhóm còn lại. Để trả lời cho vấn đề còn vướng mắc này, chúng ta sẽ tiến hành kiểm định sâu phi tham số Kruskal-Wallis tiếp như sau:

Bước 4:

- Xác định số cặp giả thuyết cần kiểm định qua công thức:

$$C_3^2 = \frac{3!}{2!(3-2)!} = 3$$

- Xây dựng các cặp giả thuyết cần kiểm định

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}; \begin{cases} H_0 : \mu_1 = \mu_3 \\ H_1 : \mu_1 \neq \mu_3 \end{cases}; \begin{cases} H_0 : \mu_2 = \mu_3 \\ H_1 : \mu_2 \neq \mu_3 \end{cases}$$

với μ_1 là kết quả học tập trung bình trên nhóm có thời gian làm thêm <8 giờ/tuần,

μ_2 là kết quả học tập trung bình trên nhóm có thời gian làm thêm 8-16 giờ/tuần,

μ_3 là kết quả học tập trung bình trên nhóm có thời gian làm thêm >16 giờ/tuần.

Bước 5:

- Tính hạng trung bình cho từng nhóm tính chất:

$\overline{R_1} = \frac{R_1}{n_1} = \frac{61,5}{5} = 12,3$; $\overline{R_2} = 10,4167$; $\overline{R_3} = 4,833$.

- Tính chênh lệch về hạng trung bình giữa hai nhóm tính chất cần kiểm định:

$D_{ij} = \left|\overline{R_i} - \overline{R_j}\right| \Rightarrow D_{12} = \left|\overline{R_1} - \overline{R_2}\right| = 1{,}8833;\ D_{23} = 5{,}5837;\ D_{13} = 7{,}467$.

Bước 6: Tính giá trị kiểm định theo công thức:

$$C_{12} = \sqrt{\chi^2_{\alpha;k-1}\left(\frac{n(n+1)}{12}\right)\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = \sqrt{5{,}99\left(\frac{17\times 18}{12}\right)\left(\frac{1}{5} + \frac{1}{6}\right)} = 7{,}4837$$

$C_{13} = C_{23} = 7{,}1355$

Bước 7: So sánh và kết luận

- $D_{12} < C_{12}$, chấp nhận H_0. Kết luận: Kết quả học tập trung bình không có sự khác biệt giữa nhóm có thời gian làm thêm <8 giờ/tuần và nhóm có thời gian làm thêm 8-16 giờ/tuần.
- $D_{23} < C_{23}$, chấp nhận H_0. Kết luận: Kết quả học tập trung bình không có sự khác biệt giữa nhóm có thời gian làm thêm 8-16 giờ/tuần và nhóm có thời gian làm thêm >16 giờ/tuần.
- $D_{13} > C_{13}$, bác bỏ H_0. Kết luận: Kết quả học tập trung bình có sự khác biệt giữa nhóm có thời gian làm thêm <8 giờ/tuần và nhóm có thời gian làm thêm >16 giờ/tuần.

6.4.6. Kiểm định Chi-bình phương về tính phụ thuộc

Chúng ta đã sử dụng kiểm định Kruskal-Wallis để kiểm định sự phụ thuộc giữa một biến định tính (biến nguyên nhân) và một biến định lượng (biến kết quả) trong điều kiện dữ liệu thu thập không thoả mãn các giả thuyết của phân tích phương sai một yếu tố. Song đôi khi trong thực tế chúng ta vẫn gặp phải các vấn đề khác như: mức độ hài lòng và việc lựa chọn khách sạn có phụ thuộc vào nhau không; hoặc giới tính và sự ưa thích các loại nước giải khát có phụ thuộc không; đặc biệt khi dữ liệu thu thập thuộc vào một thang đo có độ tin cậy thấp nhất là thang đo định danh.

Tổng quát hơn, có lúc chúng ta sẽ cần phân tích sự phụ thuộc giữa một biến nguyên nhân và một biến kết quả mà cả hai biến này đều là biến định tính. Lúc này, rõ ràng số liệu thu thập ở biến kết quả không phải là các giá trị cụ thể mà là số lượng phần tử; và điều này sẽ là hoàn toàn không phù hợp nếu sử dụng kiểm định phi tham số Kruskal-Wallis, vốn là kiểm định đòi hỏi biến kết quả phải là biến định lượng. Khi gặp phải những tình huống này thì kiểm định Chi-bình phương là một trong những giải pháp khả thi; nó sẽ giúp chúng ta kiểm định sự phụ thuộc giữa các biến định tính mà thang đo có độ tin cậy thấp với các bước thực hiện như sau:

Bước 1: Xây dựng cặp giả thuyết

H_0: Hai biến định tính không có sự phụ thuộc

H_1: Hai biến định tính có sự phụ thuộc

Bước 2: Tính giá trị kiểm định Chi-bình phương theo công thức

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

với

- $E_{ij} = \dfrac{(\text{Tổng hàng i})\times(\text{Tổng cột j})}{\text{n}}$
- O_{ij} là tần số quan sát thực tế của ô ở địa chỉ ij
- E_{ij} là tần số lý thuyết của ô ở địa chỉ ij
- r là số hàng của bảng; c là số cột của bảng.

Bước 3: So sánh và kết luận

- Nếu $\chi_0^2 < \chi^2_{\alpha;(r-1)(c-1)}$ (tra bảng phụ lục 5) chấp nhận H_0 và ngược lại.

Ví dụ 6.40

Để đánh giá xem mức độ hài lòng và việc lựa chọn khách sạn có phụ thuộc vào nhau không ở hai khách sạn A và B, người ta đã khảo sát ngẫu nhiên một số khách, bảng số liệu thu được như sau:

Khách sạn / Khách hàng	A	B	Tổng
Quay trở lại	163	154	317
Không quay trở lại	64	108	172
Tổng	227	262	489

Với mức ý nghĩa 5%, hãy cho biết có hay không mối liên hệ giữa mức độ hài lòng và việc chọn lựa khách sạn?

Giải

Bước 1: Xây dựng cặp giả thuyết

H_0: Mức độ hài lòng và khách sạn độc lập

H_1: Mức độ hài lòng và khách sạn phụ thuộc

Bước 2: Tính giá trị kiểm định Chi-bình phương

Mẫu được viết lại như sau:

Khách sạn / Khách hàng	A	B	Tổng
Quay trở lại	163 $E_{11} = 147,1$	154 $E_{12} = 169,8$	317
Không quay trở lại	64 $E_{21} = 79,8$	108 $E_{22} = 92,1$	172
Tổng	227	262	489

$$\chi_0^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} = \frac{(163-147,1)^2}{147,1} + \frac{(154-169,8)^2}{169,8} + \frac{(64-79,8)^2}{79,8} + \frac{(108-92,1)^2}{92,1} = 9,06$$

Bước 3: So sánh và kết luận

- Vì $\chi_0^2 > \chi^2_{\alpha;(r-1)(c-1)} = \chi^2_{0,05;1} = 3,8415$ (tra bảng phụ lục 5) nên bác bỏ H_0.

Kết luận: Mức độ hài lòng và việc lựa chọn khách sạn có sự phụ thuộc.

Ví dụ 6.41

Ba mẫu vật liệu được thử sức bền dưới ảnh hưởng của việc thay đổi nhiệt độ vô cùng lớn đưa ra các kết quả sau:

Loại vật liệu Kết cục	VL1	VL2	VL3
Vỡ vụn	25	45	41
Bị phá huỷ một phần	40	35	33
Còn toàn vẹn	35	20	26

Kiểm định xem có sự phụ thuộc giữa loại vật liệu với tác động thay đổi nhiệt độ thể hiện qua kết cục không, với mức ý nghĩa 5%.

Giải

Bước 1: Xây dựng cặp giả thuyết

H_0: Loại vật liệu không có sự phụ thuộc với tác động thay đổi nhiệt độ.

H_1: Loại vật liệu có sự phụ thuộc với tác động thay đổi nhiệt độ.

Bước 2: Tính giá trị kiểm định Chi-bình phương

Mẫu được viết lại như sau:

Loại vật liệu Kết cục	VL1	VL2	VL3	Tổng
Vỡ vụn	25 $E_{11}=37$	45 $E_{12}=37$	41 $E_{13}=37$	111
Bị phá huỷ một phần	40 $E_{21}=36$	35 $E_{22}=36$	33 $E_{23}=36$	108
Còn toàn vẹn	35 $E_{31}=27$	20 $E_{32}=27$	26 $E_{33}=27$	81
Tổng	100	100	100	300

$$\chi_0^2 = \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} = 10,998$$

Bước 3: So sánh và kết luận

- Vì $\chi_0^2 > \chi^2_{\alpha;(r-1)(c-1)} = \chi^2_{0,05;4} = 9,4877$ (tra bảng phụ lục 5) nên bác bỏ H_0.

Kết luận: Các loại vật liệu khác nhau sẽ chịu ảnh hưởng của sự thay đổi nhiệt độ.