PHẦN II: THỐNG KÊ

Thống kê toán là bộ môn toán học nghiên cứu quy luật của các hiện tượng ngẫu nhiên có tính chất số lớn trên cơ sở thu nhập và xử lý các số liệu thống kê (các kết quả quan sát). Nội dung chủ yếu của thống kê toán là xây dựng các phương pháp thu nhập và xử lý các số liệu thống kê nhằm rút ra các kết luận khoa học và thực tiễn, dựa trên những thành tựu của lý thuyết xác suất.

Việc thu thập, sắp xếp, trình bày các số liệu của tổng thể hay của một mẫu được gọi là **thống kê mô tả**. Còn việc sử dụng các thông tin của mẫu để tiến hành các suy đoán, kết luận về tổng thể gọi là **thống kê suy diễn**.

Thống kê được ứng dụng vào mọi lĩnh vực. Một số ngành đã phát triển thống kê ứng dụng chuyên sâu trong ngành như thống kê trong xã hội học, trong y khoa, trong giáo dục học, trong tâm lý học, trong kỹ thuật, trong sinh học, trong phân tích hóa học, trong thể thao, trong hệ thống thông tin địa lý, trong xử lý hình ảnh...

Chương I: LÝ THUYẾT MẪU

Chương II: LÝ THUYẾT ƯỚC LƯỢNG

Chương III: KIỂM ĐỊNH GIẢ THIẾT THỐNG KÊ

Chương IV: PHÂN TÍCH PHƯƠNG SAI (chỉ trong BTL)

Chương V: LÝ THUYẾT HỒI QUY ĐƠN (sơ lược)

Chương I: LÝ THUYẾT MẪU

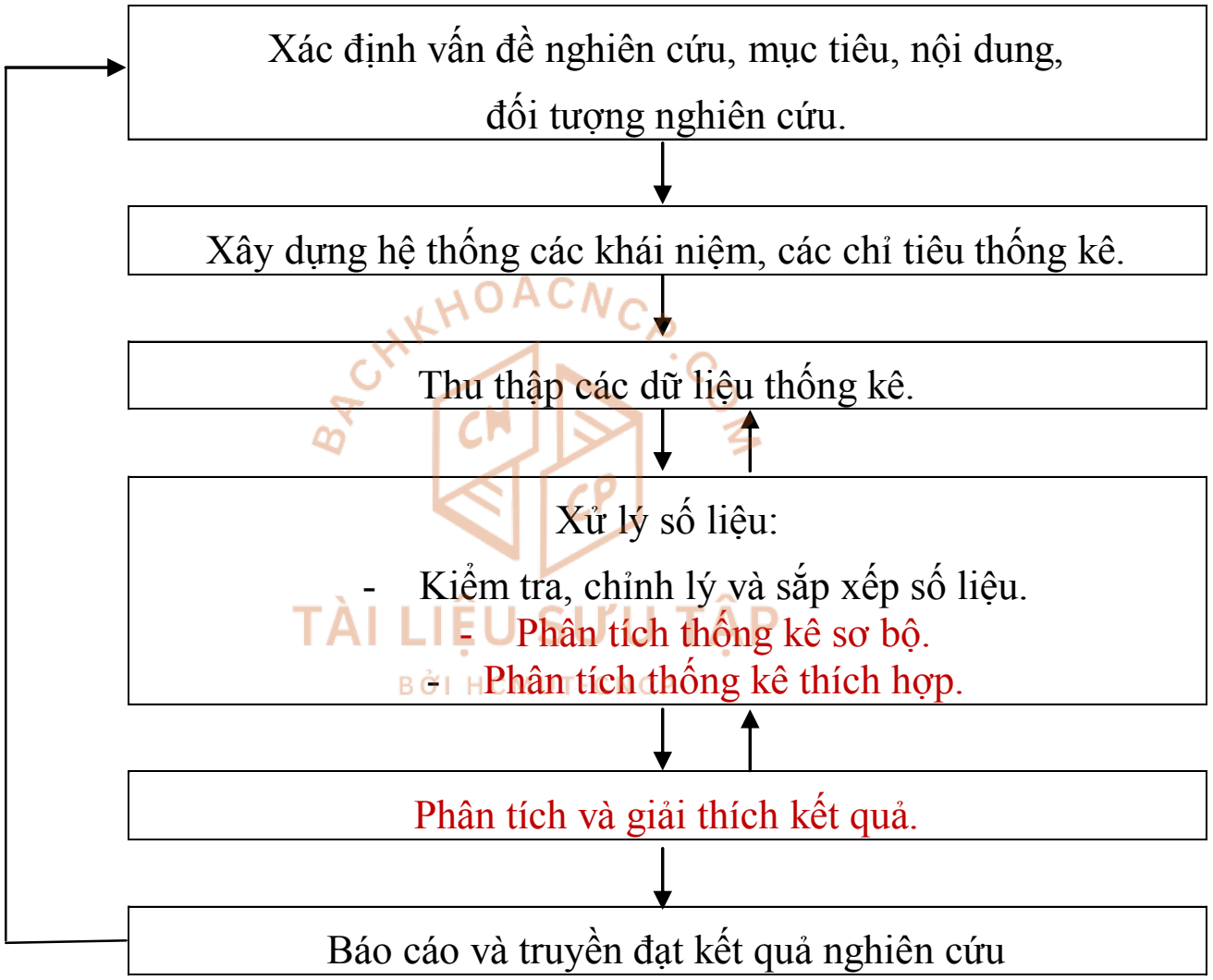
I.1. Một số khái niệm:

- **Tổng thể thống kê** là tập hợp các phần tử thuộc đối tượng nghiên cứu, cần được quan sát, thu thập và phân tích theo một hoặc một số đặc trưng nào đó. Các phần tử tạo thành tổng thể thống kê được gọi là đơn vị tổng thể.
- **Mẫu** là một số đơn vị được chọn ra từ tổng thể theo một phương pháp lấy mẫu nào đó. Các đặc trưng mẫu được sử dụng để suy rộng ra các đặc trưng của tổng thể nói chung.
- **Đặc điểm thống kê** (dấu hiệu nghiên cứu) là các tính chất quan trọng liên quan trực tiếp đến nội dung nghiên cứu và khảo sát cần thu thập dữ liệu trên các đơn vị tổng thể; Người ta chia làm 2 loại: *đặc điểm thuộc tính* và *đặc điểm số lượng*.

- Trong thực tế, phương pháp nghiên cứu toàn bộ tổng thể chỉ áp dụng được với các tập hợp có qui mô nhỏ, còn chủ yếu người ta áp dụng phương pháp nghiên cứu không toàn bộ, đặc biệt là phương pháp chọn mẫu.
- Nếu mẫu được chọn ra một cách ngẫu nhiên và xử lý bằng các phương pháp xác suất thì thu được kết luận một cách nhanh chóng, đỡ tốn kém mà vẫn đảm bảo độ chính xác cần thiết.
- Có 2 phương pháp để lấy một mẫu có n phần tử : lấy có hoàn lại và lấy không hoàn lại. Nếu kích thước mẫu rất bé so với kích thước tổng thể thì hai phương pháp này được coi là cho kết quả như nhau.
- Về mặt lý thuyết, ta giả định rằng các phần tử được lấy vào mẫu theo phương thức có hoàn lại và mỗi phần tử của tổng thể đều được lấy vào mẫu với khả năng như nhau.

- Việc sử dụng bất kỳ phương pháp thống kê nào cũng chỉ đúng đắn khi tổng thể nghiên cứu thỏa mãn những giả thiết toán học cần thiết của phương pháp. Việc sử dụng sai dữ liệu thống kê có thể tạo ra những sai lầm nghiêm trọng trong việc mô tả và diễn giải. Bằng việc chọn (hoặc bác bỏ, hay thay đổi) một giá trị nào đó, hay việc bỏ đi các giá trị quan sát quá lớn hoặc quá nhỏ cũng là một cách làm thay đổi kết quả; và đôi khi những kết quả thú vị khi nghiên cứu với mẫu nhỏ lại không còn đúng với mẫu lớn.
- Dữ liệu sơ cấp là dữ liệu người làm nghiên cứu thu thập trực tiếp từ đối tượng nghiên cứu hoặc thuê các công ty, các tổ chức khác thu thập theo yêu cầu của mình.
- Dữ liệu thứ cấp là dữ liệu thu thập từ những nguồn có sẵn, thường đã qua tổng hợp, xử lý. Dữ liệu thứ cấp thường có ưu điểm là thu nhập nhanh, ít tốn kém công sức và chi phí so với việc thu thập dữ liệu sơ cấp; tuy nhiên dữ liệu này thường ít chi tiết và đôi khi không đáp ứng được yêu cầu nghiên cứu.

Khái quát quá trình nghiên cứu thống kê



Có 2 nhóm kỹ thuật lấy mẫu là kỹ thuật lấy mẫu xác suất (probability sampling) , trên nguyên tắc mọi phần tử trong tổng thể đều có cơ hội được lấy vào mẫu như nhau) và lấy mẫu phi xác suất (non- probability sampling) .

I.2 CÁC KỸ THUẬT LẤY MẪU XÁC SUẤT:

I.2.1 Lấy mẫu ngẫu nhiên đơn giản (simple random sampling):

Cách tiến hành:

- Lập danh sách tổng thể theo số thứ tự, gọi là khung lấy mẫu.
- Xác định số phần tử n cần lấy vào mẫu (sample size).
- Chọn 1 mẫu gồm các đối tượng có số thứ tự được lựa chọn ra 1 cách ngẫu nhiên bằng cách bốc thăm, lấy từ 1 bảng số ngẫu nhiên; bằng MTBT hay 1 phần mềm thống kê nào đó.

- *Ưu điểm:* Tính đại diện cao.
- *Hạn chế:* Mẫu phải không có kích thước quá lớn; Người nghiên cứu phải lập được danh sách tổng thể cần khảo sát.

I.2.2 Lấy mẫu hệ thống (systematic sampling):

Cách tiến hành:

- Lập danh sách *N* phần tử của tổng thể, có mã là số thứ tự.
- Xác định số phần tử *n* cần lấy vào mẫu (sample size).
- Xác định số nguyên k gọi là khoảng cách, k lấy giá trị làm tròn của N/n. Chọn phần tử đầu tiên vào mẫu 1 cách ngẫu nhiên (có số thứ tự trong khoảng 1 đến k hay 1 đến N). Các phần tử tiếp theo là các phần tử có STT = STT phần tử đầu tiên + k/2k/3k/…

Có thể quay vòng lại để tiếp tục nếu lấy mẫu chưa đủ n phần tử; khi đó coi phần tử số 1 có STT là N+1,…

- *Ưu điểm:* Tiết kiệm thời gian khi cần mẫu có kích thước lớn.
- *Hạn chế:* Người nghiên cứu phải lập được danh sách tổng thể cần khảo sát. Thứ tự trong danh sách tổng thể chỉ để mã hóa, không được sắp xếp theo các đặc điểm khảo sát.

I.2.3 Lấy mẫu phân tầng (stratified sampling):

Cách tiến hành:

- Chia tổng thể thành nhiều tầng khác nhau dựa vào các tính chất liên quan đến đặc điểm cần khảo sát. Trên mỗi tầng thực hiện lấy mẫu ngẫu nhiên đơn giản với số lượng phần tử cần lấy vào mẫu là n_i được phân bổ theo tỉ lệ các phần tử ở mỗi tầng.
- Trong thực tế, với mẫu được chọn, người ta có thể kết hợp khảo sát thêm các đặc điểm riêng lẻ đối với những phần tử trong cùng 1 tầng. Khi đó nếu nhận thấy 1 vài giá trị m_i quá nhỏ làm các khảo sát riêng lẻ đó không đủ độ tin cậy thì chúng ta cần lấy mẫu không cân đối (*disproportionately)* và phải quan tâm đến việc hiệu chỉnh kết quả theo trọng số. (xem thêm tài liệu).
- *Ưu điểm:* Kỹ thuật này làm tăng khả năng đại diện của mẫu theo đặc điểm cần khảo sát. Ở các nghiên cứu có quy mô lớn, người ta thường kết hợp với cách lấy mẫu cả cụm.

I.2.4 Lấy mẫu cả cụm(cluster sampling) và lấy mẫu nhiều giai đoạn (multi- stage sampling):

Cách tiến hành:

- Chia tổng thể thành nhiều cụm theo các tính chất nào đó ít liên quan đến đặc tính cần khảo sát, chọn ra *m* cụm ngẫu nhiên. Khảo sát hết các phần tử trong các cụm đã lấy ra. Theo cách này số phần tử lấy vào mẫu có thể nhiều hơn số cần thiết *n* và các phần tử trong cùng cụm có thể có khuynh hướng giống nhau.
- Để khắc phục, ta chọn *m* cụm gọi là mẫu bậc 1 nhưng không khảo sát hết mà trong từng cụm bậc 1 lại chọn ngẫu nhiên k_i cụm nhỏ gọi là mẫu bậc 2;...làm như vậy cho đến khi đủ số lượng cần. Khảo sát tất cả các phần tử đã được chọn ở bậc cuối cùng.
- *Ưu điểm:* Kỹ thuật này xử lý tốt các khó khăn gặp phải khi tổng thể có phân bố rộng về mặt địa lý (thời gian, tiền bạc, nhân lực, bảo quản dữ liệu...), hay khi lập 1 danh sách tổng thể đầy đủ.

I.3 MỘT SỐ KỸ THUẬT LẤY MẪU PHI XÁC SUẤT:

I.3.1 Lấy mẫu thuận tiện (convenient sampling):

Người lấy mẫu lấy thông tin cần khảo sát ở những nơi mà người đó nghĩ là thuận tiện.

I.3.1 Lấy mẫu định mức (quota sampling):

Người lấy mẫu chia tổng thể thành các tổng thể con (tương tự như phân tầng trong lấy mẫu phi xác suất) rồi dựa vào kinh nghiệm tự định mức số phần từ cần lấy vào mẫu theo 1 tỷ lệ nào đó.

I.3.1 Lấy mẫu phán đoán (judgement sampling):

Người lấy mẫu dựa vào năng lực và kinh nghiệm của mình để tự phán đoán cần khảo sát trong phạm vi nào, những phần tử nào cần chọn vào mẫu.

Mẫu phi xác suất không đại diện cho toàn bộ tổng thể nhưng được chấp nhận trong nghiên cứu khám phá; trong việc ước lượng sơ bộ do việc nghiên cứu bị hạn chế thời gian, kinh phí, hay đôi khi chỉ để hoàn thiện một bộ câu hỏi khảo sát.

I.4 MỘT SỐ VẤN ĐỀ LIÊN QUAN:

1.4.1 Cỡ mẫu được tính như thế nào?

Mặc dù có thể đưa số công thức cho 1 số trường hợp nhưng đáp án duy nhất là không có. Về nguyên tắc, mẫu càng lớn thì càng chính xác vì sai số lấy mẫu có thể giảm khi tăng kích thước mẫu. Tuy nhiên thời gian và nguồn lực của nhà nghiên cứu có hạn nên người ta phải cân nhắc chúng với yêu cầu về độ chính xác, độ tin cậy của khảo sát, loại phân tích sẽ dùng để xử lý dữ liệu.

I.4.2 Sai lệch hệ thống (Bias) trong chọn mẫu:

- Sai lệch (hay thiên lệch) trong lấy mẫu thể hiện việc lấy mẫu có xu hướng không đại diện cho tổng thể, sai lệch này nằm trong cách thức lấy mẫu và cách thức thu thập thông tin từ mẫu. Có các loại sai lệch thường gặp sau:

- **Sai lệch lựa chọn mẫu** (Selection Bias): sai lệch này xuất hiện khi cách thức lấy mẫu đã làm loại trừ hay hạn chế cơ hội được lấy vào mẫu của bộ phận trong tổng thể.
- **Sai lệch đo lường hay sai lệch phản hồi** (Measurement or Response Bias): sai lệch này làm cho thông tin chúng ta nhận được từ mẫu đã chọn không đúng với giá trị thực của nó. Sai lệch này xảy ra có thể do cách đo lường không chuẩn (cách thiết kế bảng câu hỏi, cách đặt vấn đề, cách dùng từ ngữ, cách thức tiếp cận mẫu,…)
- **Sai lệch do không phản hồi** (Non-Response Bias): do không có thông tin phản hồi từ 1 bộ phận trong mẫu đã thiết kế nên có thể ảnh hưởng đến tính đại diện của mẫu. Các cuộc điều tra qua email thường ít tốn kém nhưng tỷ lệ phản hồi thấp; các cuộc phỏng vấn cá nhân có tỷ lệ phản hồi cao hơn.

I.5 THIẾT KẾ THÍ NGHIỆM

- Xem giáo trình XSTK và PTSL (Nguyễn Tiến Dũng, Ng. Đình Huy).
- Xem file tài liệu tham khảo kèm theo (Nguyễn Văn Tuấn).

I.6 MÔ TẢ DỮ LIỆU BẰNG BIỂU ĐỒ VÀ ĐỒ THỊ (tự đọc tài liệu)

- Dữ liệu định tính (Biểu đồ cột; biểu đồ Pie).
- Dữ liệu định lượng: Biểu đồ cành lá; biểu đồ phân bố tần số hoặc tần suất (Histograms); biểu đồ mật độ tần suất trong cả trường hợp các khoảng chia bằng nhau và các khoảng chia không bằng nhau.

I.7 TÓM TẮT DỮ LIỆU BẰNG CÁC ĐẠI LƯỢNG SỐ (Ch4-giáo trình TKƯD)

- TrB nhân; TrB điều hòa. Ý nghĩa của hệ số biến thiên CV.
- Hình dáng phân phối của dữ liệu, liên hệ với biểu đồ hộp và râu.
- Quy tắc phân phối dữ liệu thực nghiệm.
- Chuẩn hóa dữ liệu.

I.8 Tìm hiểu 1 số phần mềm máy tính có chức năng thống kê được dùng để mô tả dữ liệu mẫu: EXCEL; SPSS; STATA; R, MFIT…

II.1 CÁC ĐẶC TRƯNG TỔNG THỂ VÀ MẪU:

- Số lượng N các phần tử của tổng thể được gọi là kích thước tổng thể. Trong nhiều trường hợp, ta không biết được N.
- Khi khảo sát tổng thể theo một dấu hiệu nghiên cứu nào đó, người ta mô hình hóa nó bởi một biến ngẫu nhiên X, gọi là **biến ngẫu nhiên gốc**. Các đặc trưng thường gặp khi dấu hiệu nc là định lượng:

- Trung bình tt (Kỳ vọng) E(X)	*Kí hiệu :*	*a* hoặc μ
- Phương sai tổng thể D(X)	$\rightarrow$	σ^2
- Độ lệch chuẩn tổng thể $\sqrt{D(X)}$	$\rightarrow$	σ

- Trường hợp dấu hiệu nghiên cứu mang tính chất định tính thì ta coi *X có phân phối không – một*. **Tỉ lệ tổng thể** là **xác suất** lấy được phần tử mang dấu hiệu nghiên cứu từ tổng thể.

- Tỉ lệ tổng thể:	*Kí hiệu :*	**p**

- **Mẫu ngẫu nhiên 1 chiều kích thước n** là tập hợp của n biến ngẫu nhiên độc lập $X_1, X_2, .., X_n$ được thành lập từ biến ngẫu nhiên X của tổng thể nghiên cứu và có cùng quy luật phân phối xác suất với X.

- K/h của mẫu nn tổng quát kích thước n là: $\mathbf{W = (X_1, X_2, .., X_n)}$

  với $\mathbf{E(X_i) = E(X) = a;}$ $\mathbf{D(X_i) = D(X) = \sigma^2,}$ $\forall i$.

- Việc thực hiện một phép thử đối với mẫu ngẫu nhiên W chính là thực hiện một phép thử đối với mỗi thành phần X_i. Ta gọi kết quả $w_n = (x_1, x_2, .., x_n)$ tạo thành là mẫu cụ thể.

- Bảng phân phối tần số thực nghiệm của mẫu cụ thể:

x_i	x_1	x_2	....	x_k
n_i	n_1	n_2	....	n_k

với $\sum_{i=1}^{k} n_i = n$

CÁC ĐẶC TRƯNG CỦA MẪU TỔNG QUÁT	CÁC ĐẶC TRƯNG CỦA MẪU CỤ THỂ
TRUNG BÌNH MẪU $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$	Trung bình mẫu (Mean): $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ hay $\bar{x}=\frac{1}{n}\sum_{i=1}^{k}n_i x_i$
PHƯƠNG SAI MẪU $\hat{S}^2 \overset{ct1}{=} \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$	Phương sai mẫu: $\hat{s}^2$ Độ lệch mẫu: $\hat{s}$ $\hat{s}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ hay $\hat{s}^2=\frac{1}{n}\sum_{i=1}^{k}n_i x_i^2-\left(\bar{x}\right)^2=\overline{x^2}-\left(\bar{x}\right)^2$
PHƯƠNG SAI MẪU HIỆU CHỈNH $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2=\frac{n}{n-1}\hat{S}^2$	Phương sai mẫu hiệu chỉnh (Sample variance): s^2 Độ lệch mẫu hiệu chỉnh (SD- Standard Deviation): s $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ hay $s^2=\frac{n}{n-1}\hat{s}^2$
TỈ LỆ MẪU $F=\frac{M}{N}$	Tỉ lệ mẫu: $f=\frac{m}{n}$

II.2 Một số đặc trưng khác của mẫu dữ liệu định lượng:

II.2.1 Yếu vị (Mode)

II.2.2 Hệ số biến thiên (Coefficient of variation - CV)

Hệ số biến thiên đo lường mức độ biến động tương đối của mẫu dữ liệu, được dùng khi người ta muốn so sánh mức độ biến động của các mẫu không cùng đơn vị đo.

CV (của tổng thể) = $\frac{\sigma}{a} \times 100\%$ CV (của mẫu) = $\frac{s}{\overline{x}} \times 100\%$

II.2.3 Sai số chuẩn của trung bình mẫu (Standard error): SE = $\frac{s}{\sqrt{n}}$

II.2.4 Trung vị (Median) *(Trường hợp mẫu không được phân tổ dữ liệu)*

Giả sử mẫu có kích thước n được sắp xếp tăng dần theo giá trị được khảo sát: $x_1 \le x_2 \le .... \le x_{n-1} \le x_n$.

Nếu n = 2k+1 thì trung vị mẫu là giá trị x_{k+1} .

Nếu n = 2k thì trung vị mẫu là giá trị ($x_k + x_{k+1}$): 2.

II.2.5 Tứ phân vị (Quartiles)

Giá trị trung vị chia mẫu dữ liệu đã sắp thứ tự thành 2 tập có số phần tử bằng nhau. Trung vị của tập dữ liệu nhỏ hơn là Q_1 (gọi là tứ phân vị dưới) và trung vị của tập dữ liệu lớn hơn là Q_3 (gọi là tứ phân vị trên). Q_2 được lấy bằng giá trị trung vị.

Độ trải giữa IQR $\equiv R_Q = Q_3 - Q_1$.

II.2.6 Điểm Outlies: còn gọi là điểm dị biệt, điểm ngoại lệ, điểm ngoại lai…. Đó là các phần tử của mẫu có giá trị nằm ngoài khoảng (Q1 – 1,5 × IQR; Q3 + 1,5 × IQR) .

II.2.7 Vẽ biểu đồ hộp và râu:

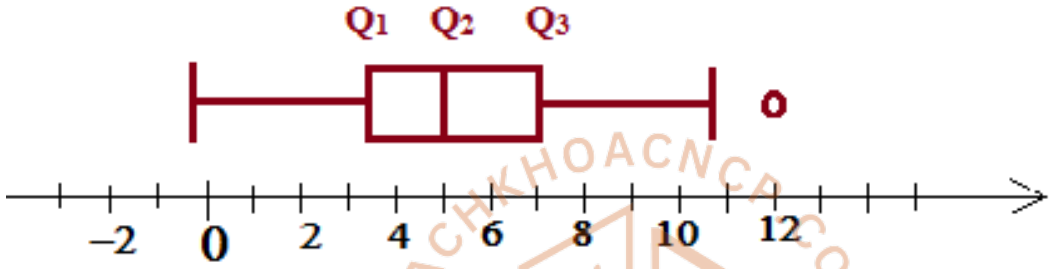
Xét mẫu có kích thước n = 9 đã được sắp theo thứ tự tăng dần:

1 3 4 4 5 5 6 8 12

$Q_1 = 3,5$ Q_2 $Q_3 = 7$

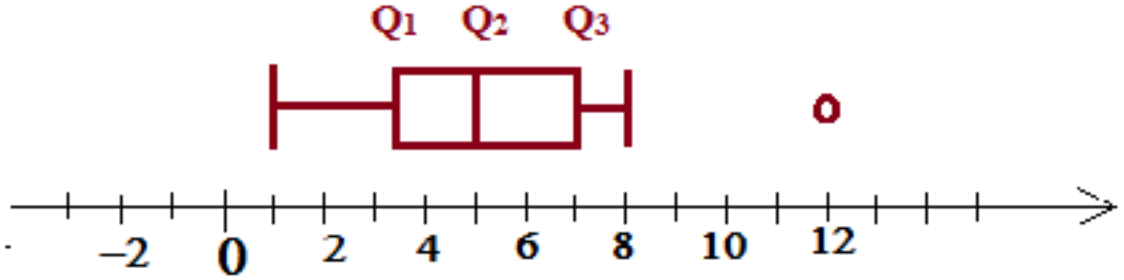
Khoảng trải giữa IQR = $Q_3 - Q_1 = 7 - 3,5 = 2,5$

$Q_1 - 1,5 \times IQR = -0,25$ $Q_3 + 1,5 \times IQR = 10,75$

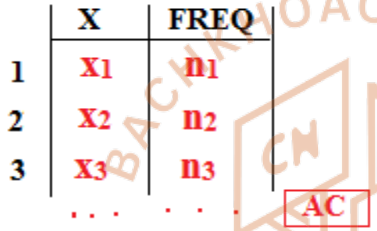


Có 1 giá trị outlier là 12

Điều chỉnh lại 2 râu của hình hộp đến 2 giá trị nhỏ nhất và lớn nhất của dữ liệu, không tính các giá trị outlier .



HD Sử dụng MTBT tìm 1 số đặc trưng của BNN rời rạc:

Các bước	Máy CASIO fx 570 ES PLUS…	Máy CASIO fx 500 MS….
Vào TK 1 biến.	MODE -- 3 (STAT) -- 1 (1-VAR)	MODE -- MODE --…-- 1 (SD)
Mở cột tần số (nếu chưa có)	SHIFT -- MODE (SETUP) – -- ▼ -- ---4 (STAT) -- 1 (ON)	
Nhập dữ liệu		x_1 ; n_1 M+ x_2 ; n_2 M+ ……
Đọc kết quả n	SHIFT – 1 (STAT)- 4 (VAR) -- 1 (n) -- =	SHIFT – 1 (SSUM) - 3 (n)
Đọc kq $\bar{x}$	SHIFT – 1 (STAT)- 4 (VAR) -- 2 ($\bar{x}$) -- =	SHIFT – 2 (SVAR) -1 ($\bar{x}$)-- =
Đọc kq $\hat{s}$	SHIFT – 1 (STAT)- 4 (VAR) - 3 (σx) -=	SHIFT – 2 (SVAR)- 2 ($x\sigma_n$)-- =
Đọc kq s	SHIFT – 1 (STAT)- 4 (VAR) - 4 (sx) -=	SHIFT – 2 (SVAR)- 3($x\sigma_{n-1}$)-- =
Kq trung gian $\sum_{i=1}^{k} x_i n_i \equiv \sum_{i=1}^{n} x_i$	SHIFT – 1 (STAT)- 3 (SUM) –2 ($\sum x$) =	SHIFT – 1 (SSUM)- ($\sum x$)--- =
$\sum_{i=1}^{k} x_i^2 n_i \equiv \sum_{i=1}^{n} x_i^2$	SHIFT – 1 (STAT)- 3 (SUM) –1 ($\sum x^2$)=	SHIFT – 1 (SSUM)- ($\sum x^2$)-- =

Ví dụ 1: Người ta lấy 16 mẫu nước trên 1 dòng sông để phân tích hàm lượng BOD (đơn vị mg/l), kết quả thu được:

125 205 134 137 168 174 158 172

98 113 174 185 197 163 168 141

Hãy tìm các tham số mẫu:

a) Trung bình mẫu (TB cộng), trung vị mẫu, mode và CV.

b) Độ lệch mẫu và độ lệch mẫu hiệu chỉnh.

Ví dụ 2:

Khảo sát thời gian gia công của 1 số chi tiết máy được chọn ngẫu nhiên, người ta ghi nhận số liệu:

Thời gian gia công (phút)	15-17	17-19	19-21	21-23	23-25	25-28
Số chi tiết máy tương ứng	11	32	54	32	23	22

a) Tính các đặc trưng mẫu sau: $n; \bar{x}; \hat{s}; s.$

b) Tìm tỷ lệ các chi tiết được gia công dưới 19 phút.

II.3. Quy luật phân phối xác suất của các đặc trưng mẫu:

1- Phân phối xác suất của tỷ lệ mẫu

Vì $E(F) = p$ và $D(F) = \frac{pq}{n}$ nên theo định lý 4.5 chương 4 (xem giáo trình XS) thì với n≥30 ta có thể coi $F \sim N(p, \frac{pq}{n})$.

Với một mẫu cụ thể kích thước n, tỷ lệ mẫu f, ta có $p \approx f$, nên:

$$F \sim N(p, \frac{f(1-f)}{n}) \text{ hay } \boxed{\frac{(F-p)}{\sqrt{\frac{f(1-f)}{n}}} \sim N(0,1)}$$

2- Phân phối xác suất của trung bình mẫu

- Vì $E(\overline{X}) = a,\ D(\overline{X}) = \frac{\sigma^2}{n}$ nên nếu tổng thể có phân phối chuẩn thì

$$\overline{X} \sim N(a, \frac{\sigma^2}{n}) \text{ hay } \boxed{\frac{\overline{X}-a}{\sigma}\sqrt{n} \sim N(0,1)}$$

- **Nếu $n \geq 30$** thì với một mẫu cụ thể kích thước n ta có $\sigma^2 \approx s^2$

Do đó $\overline{X} \sim N(a, \frac{s^2}{n})$ hay $\boxed{\frac{\overline{X}-a}{s}\sqrt{n} \sim N(0,1)}$

trong đó s^2 là phương sai mẫu hiệu chỉnh của một mẫu kích thước n bất kỳ.

- **Trường hợp $n < 30$, tổng thể có phân phối chuẩn**, ta có

$$\boxed{\frac{\overline{X}-a}{s}\sqrt{n} \sim T(n-1)}$$

3- Phân phối xác suất của phương sai mẫu

Nếu **tổng thể có phân phối chuẩn** thì ta có

$$\boxed{\frac{n\hat{S}^2}{\sigma^2} = \frac{n-1}{\sigma^2}S^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2(n-1)}$$

Chương II: LÝ THUYẾT ƯỚC LƯỢNG

Giả thiết một dấu hiệu nghiên cứu trong tổng thể được xem như một biến ngẫu nhiên X mà ta chưa biết một tham số θ nào đó của X. Ta cần phải ước lượng (xác định một cách gần đúng) giá trị tham số θ. Trong chương này, giá trị cần ước lượng θ được đề cập đến là trung bình tổng thể, phương sai tổng thể hoặc tỉ lệ tổng thể.

Phương pháp mẫu cho phép giải bài toán trên như sau: Từ tổng thể nghiên cứu, người ta rút ra 1 mẫu ngẫu nhiên kích thước n (gọi là mẫu thực nghiệm _ *empirical*) và dựa vào đó xây dựng một hàm thống kê $\hat{\theta} = f(X_1, X_2, .., X_n)$ dùng để ước lượng θ bằng cách này hay cách khác, gọi là hàm ước lượng *(estimator)*.

Có 2 phương pháp ước lượng: ƯL điểm và ƯL khoảng.

- **ƯL điểm** là dùng một tham số thống kê mẫu đơn lẻ để ước lượng giá trị tham số của tổng thể. Ví dụ dùng một giá trị cụ thể của trung bình mẫu $\overline{X}$ để ước lượng trung bình tổng thể *a*.

Có nhiều cách chọn hàm ước lượng $\hat{\theta}$ khác nhau, vì vậy người ta đưa ra một số tiêu chuẩn để đánh giá chất lượng của các hàm này, để từ đó lựa chọn được hàm "xấp xỉ một cách tốt nhất" tham số cần ước lượng.

- *Ước lượng không chệch:* $\hat{\theta}$ là ước lượng không chệch của θ nếu $E(\hat{\theta}) = \theta$.
- *Ước lượng hiệu quả:* $\hat{\theta}$ là ước lượng hiệu quả của θ nếu nó là ước lượng không chệch của θ và có phương sai nhỏ nhất so với các ước lượng không chệch khác được xây dựng trên cùng mẫu đó.
- *Ước lượng vững:* $\hat{\theta}$ là ước lượng vững (hay ước lượng nhất quán) của θ nếu $\hat{\theta}$ hội tụ theo xác suất đến θ khi $n \to \infty$.
- *Ước lượng đủ:* $\hat{\theta}$ được gọi là ước lượng đủ nếu nó chứa toàn bộ các thông tin trong mẫu về tham số θ của ước lượng.

Phương pháp ước lượng hợp lý cực đại:

Có nhiều phương pháp ước lượng tổng quát như phương pháp moment, phương pháp Bayes, phương pháp minimax,.., nhưng thông dụng nhất là phương pháp ước lượng hợp lý cực đại (maximal likelyhood). Phương pháp này do Ronald Fisher đề ra, nó là một trong những phương pháp quan trọng và hay dùng nhất để tìm hàm ước lượng.

Giả sử ta đã biết phân phối xác suất tổng quát của biến ngẫu nhiên gốc X dưới dạng hàm mật độ $f(x, \theta)$. Đó cũng có thể là biểu thức xác suất nếu X là biến ngẫu nhiên rời rạc. Để ước lượng θ, ta lấy mẫu ngẫu nhiên $(X_1,X_2,\ldots,X_n)$ và lập hàm số:

$$L(\theta)= f(X_1, \theta).\ f(X_2, \theta)\ldots.f(X_n, \theta).$$

Hàm L được gọi là *hàm hợp lý* của mẫu, nó phụ thuộc vào $X_1, X_2,\ldots.,X_n$ và θ nhưng ta coi $X_1, X_2,\ldots.,X_n$ là các hằng số, còn θ được coi là biến số. Từ đó tìm hàm ước lượng $\hat{\theta}$ phụ thuộc $X_1, X_2,\ldots.,X_n$ sao cho $L(\theta)$ đạt GTLN tại $\hat{\theta}$.

Bảng 1- Tóm tắt một số hàm ước lượng tham số thông dụng:

Tham số θ cần ước lượng	Chọn thống kê $\hat{\theta}$ để ước lượng	$E[\hat{\theta}]$	$D[\hat{\theta}]$	Tính chất của ước lượng
Tỉ lệ p (xác suất)	$F = \frac{m}{n}$	$E(F) = p$	$D(F) = \frac{p(1-p)}{n}$	Không chệch, vững, hiệu quả, đủ; hợp lý cực đại.
Kỳ vọng $a = E(X)$	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$E(\overline{X}) = a$	$D(\overline{X}) = \frac{\sigma^2}{n}$	Không chệch, vững, hiệu quả, đủ; hợp lý cực đại.
Phương sai $\sigma^2 = D(X)$	$\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$	$E(\hat{S}^2) = \frac{n-1}{n}\sigma^2$	…	Chệch, vững, đủ; hợp lý cực đại.
	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$	$E(S^2) = \sigma^2$	…	Không chệch, vững, đủ.

Ví dụ: Khảo sát thu nhập hàng tháng của 50 công nhân được lựa chọn ngẫu nhiên từ các xí nghiệp may trong khu vực, người ta tính được thu nhập bình quân của 50 người này là 4,2 triệu đồng. Phương pháp ước lượng điểm cho phép ta đánh giá thu nhập trung bình của công nhân ở các nhà máy này là 4,2 triệu.

Một nhược điểm cơ bản của phương pháp ước lượng điểm là khi kích thước mẫu chưa thực sự lớn thì ước lượng điểm tìm được có thể sai lệch rất nhiều so với giá trị của tham số cần ước lượng. Mặt khác, dùng các phương pháp ước lượng đều có thể có sai lầm nhưng phương pháp ƯL điểm không đánh giá được khả năng mắc sai lầm là bao nhiêu.

-**Ước lượng bằng khoảng tin cậy** chính là tìm ra khoảng ước lượng $(G_1;G_2)$ cho tham số θ trong tổng thể sao cho ứng với độ tin cậy *(confidence)* bằng $(1-\alpha)$ cho trước, $P(G_1 < \theta < G_2) = 1-\alpha$.

Phương pháp ƯL bằng khoảng tin cậy có ưu thế hơn phương pháp ƯL điểm vì nó làm tăng độ chính xác của ước lượng và còn đánh giá được mức độ tin cậy của ước lượng. Nó chứa đựng khả năng mắc sai lầm là α .

Phương pháp tìm khoảng tin cậy cho tham số θ với độ tin cậy *1-*α cho trước:

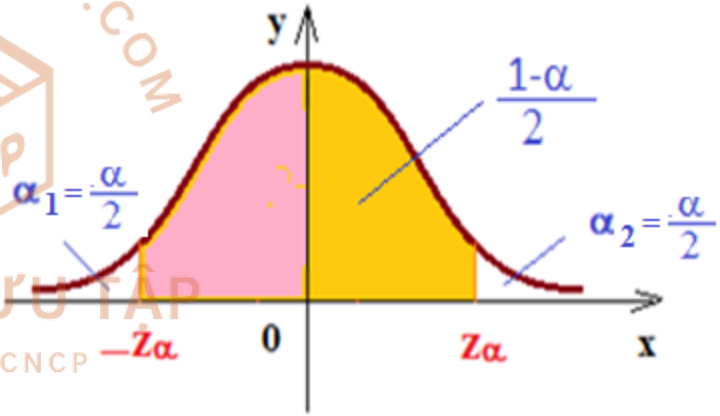
- Trước tiên ta tìm hàm ước lượng G = f(X1 , X2 , .., Xn , θ) sao cho quy luật phân phối xác suất của G hoàn toàn xác định, không phụ thuộc vào các đối số. Chọn cặp giá trị $\alpha 1, \alpha 2 \geq 0$ sao cho $\alpha 1 + \alpha 2 = \alpha$ và tìm $G\alpha 1$, $G\alpha 2$ mà ($G < G\alpha 1$) $= \alpha 1$ và P ($G > G\alpha 2$) = $\alpha 2$, suy ra P($G\alpha 1 < G < G\alpha 2$) = 1 - α. Biến đổi để tìm được các giá trị G1, G2 sao cho P($G1 < \theta < G2$) = 1-α. Khi đó khoảng (G1, G2) chính là một trong các khoảng tin cậy *(confidence interval)* cần tìm.
- Theo nguyên lý xác suất lớn thì với độ tin cậy (1 -α) đủ lớn, hầu như chắc chắn biến cố ($G1 < \theta < G2$) sẽ xảy ra trong một phép thử. Vì vậy trong thực tế chỉ cần thực hiện phép thử để có được một mẫu cụ thể w = (x1, x2 , .., xn) rồi tính giá trị của G1 và G2 ứng với mẫu đã cho sẽ cho ta một khoảng ước lượng thỏa yêu cầu.

Bài toán minh họa 1: *Xét mẫu tổng quát có kích thước n (đủ lớn) và tỉ lệ mẫu F. Ký hiệu f là tỉ lệ của một mẫu cụ thể. Tìm khoảng tin cậy đối xứng cho tỉ lệ tổng thể p với độ tin cậy 1-α.*

Từ kết quả đã nêu ở mục II.2, khi n ≥ 30, nếu ta đặt: $Z = \dfrac{F-p}{\sqrt{f(1-f)}}\sqrt{n}$

thì $Z \sim N(0,1)$. Chọn $\alpha_1 = \alpha_2 = \alpha/2$;

Chọn z_α thỏa $\Phi(z_\alpha) = \dfrac{1-\alpha}{2}$



thì $P(-z_\alpha < Z < z_\alpha) = 1-\alpha$

$$\Leftrightarrow P\left(-z_\alpha < \frac{F-p}{\sqrt{f(1-f)}}\sqrt{n} < z_\alpha\right) = 1-\alpha$$

$$\Leftrightarrow P\left(-\frac{z_\alpha \cdot \sqrt{f(1-f)}}{\sqrt{n}} < F-p < \frac{z_\alpha \cdot \sqrt{f(1-f)}}{\sqrt{n}}\right) = 1-\alpha$$

$$\Leftrightarrow P(F-\varepsilon < p < F+\varepsilon) = 1-\alpha \text{ ;ở đây } \varepsilon = \frac{z_\alpha\sqrt{f(1-f)}}{\sqrt{n}} .$$

Người ta gọi ε là *sai số của ƯL* hay *độ chính xác của ƯL.*

Vậy khoảng ước lượng cho p là (F-ε; F+ε); có độ dài là 2ε.

Tham khảo cách trình bày khác:

Ta chọn F là để ước lượng cho tỉ lệ tổng thể p chưa biết (Bảng 1), và chọn khoảng ước lượng có dạng (F- ε, F +ε), còn gọi là khoảng tin cậy đối xứng. Vì thế ta sẽ tìm ε sao cho: P (F- ε < p < F +ε) = 1 - α (1)

Từ (1) suy ra P (- ε < F- p < ε) = 1 - α hay

$$P\left(-\frac{\varepsilon}{\sqrt{f(1-f)}}\sqrt{n} < \frac{F-p}{\sqrt{f(1-f)}}\sqrt{n} < \frac{\varepsilon}{\sqrt{f(1-f)}}\sqrt{n}\right) = 1-\alpha \quad . \quad (2)$$

Do hàm $Z = \frac{F-p}{\sqrt{f(1-f)}}\sqrt{n} \sim N(0,1)$

nên (2) $\Leftrightarrow P(-z_\alpha < Z < z_\alpha) = 1-\alpha \Leftrightarrow 2.\Phi(z_\alpha) = 1-\alpha$

dẫn đến $\Phi(z_\alpha) = \frac{1-\alpha}{2}$. Tìm z_α bằng cách tra (ngược) bảng giá trị hàm tp Laplace (PLII), từ đó sẽ tìm được công thức $\varepsilon = \frac{z_\alpha \cdot \sqrt{f(1-f)}}{\sqrt{n}}$.

Lưu ý:

* Đối với mẫu đã xác định, khoảng tin cậy đối xứng có độ dài càng hẹp thì độ tin cậy càng thấp. Nếu chúng ta muốn có được sai số nhỏ (khoảng tin cậy hẹp) và độ tin cậy như mong muốn thì chúng ta phải tăng kích thước mẫu hợp lý.

* Có vô số khoảng ước lượng cho giá trị p của tổng thể tùy theo cách chọn α_1, α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$. Đối với bài toán ƯL tỉ lệ hay ƯL trung bình thì khoảng ƯL được trình bày ở trên chính là khoảng ƯL đối xứng và nó có độ dài ngắn nhất.

* Ở bài toán trên, nếu ta chọn trước $\alpha_1 = 0$ và $\alpha_2 = \alpha$ thì ta có khoảng ƯL bên trái $\left(0, F + \frac{z_{2\alpha}\sqrt{f(1-f)}}{\sqrt{n}}\right)$ với $\Phi(z_{2\alpha}) = \frac{1-2\alpha}{2}$; Người ta nói $F + \frac{z_{2\alpha}\sqrt{f(1-f)}}{\sqrt{n}}$ là ***ước lượng giá trị tối đa*** của p .

* Nếu chọn $\alpha_1 = \alpha$; $\alpha_2 = 0$ thì ta được khoảng ƯL bên phải $\left(F - \frac{z_{2\alpha}\sqrt{f(1-f)}}{\sqrt{n}}, 1\right)$ và ***ƯL giá trị tối thiểu*** của p là $F - \frac{z_{2\alpha}\sqrt{f(1-f)}}{\sqrt{n}}$.

Bài toán minh họa 2: *Giả sử tổng thể X có phân phối chuẩn, chưa biết trung bình tổng thể a và phương sai tổng thể σ^2. Từ tổng thể, người ta lấy được mẫu tổng quát với kích thước n, trung bình mẫu* $\overline{X}$ *và phương sai mẫu hiệu chỉnh* S^2.

Tìm khoảng tin cậy cho trung bình tổng thể a với độ tin cậy 1-α*; trong trường hợp mẫu có kích thước nhỏ.*

Theo kết quả ở II.2, khi n <30 thì hàm: $Q = \dfrac{\overline{X} - a}{\dfrac{S}{}}\sqrt{n} \sim T(n-1)$

Chọn khoảng ước lượng đối xứng có dạng $(\overline{X} - \varepsilon; \overline{X} + \varepsilon)$

Dẫn đến bài toán tìm ε để $P(\overline{X} - \varepsilon < a < \overline{X} + \varepsilon) = 1 - \alpha$

$\Rightarrow P(-\dfrac{\varepsilon}{s}\sqrt{n} < Q = \dfrac{\overline{X} - a}{s}\sqrt{n} < \dfrac{\varepsilon}{s}\sqrt{n}) = 1 - \alpha$. Đặt : $T_\alpha = \dfrac{\varepsilon}{s}\sqrt{n}$

$\Rightarrow$ Dựa vào bảng tra 1 phía trong Phụ lục VII cho hàm Student, ta tìm được giá trị $T_\alpha = t_{\alpha/2}^{(n-1)}$ bằng cách tìm số nằm ở cột $\alpha/2$, dòng thứ (n-1). Từ đó suy ra ε cần tìm.

(Nhắc lại: Khi n ≥ 30, phân phối Student xấp xỉ phân phối Chuẩn tắc.)

Bảng 2: ***Một số bài toán ước lượng khoảng thông dụng***

Tham số cần ước lượng	Phân bố của tổng thể	Thông tin bổ sung	Khoảng tin cậy khi chọn $\alpha_1 = \alpha_2 = \alpha/2$
Tỉ lệ p (xác suất)	Nhị thức B(1, p)	Mẫu lớn ($n \geq 30$)	$(F \pm \varepsilon)$; $\varepsilon = z_\alpha . \frac{\sqrt{f(1-f)}}{\sqrt{n}}$
Trung bình a	Bất kỳ	Mẫu lớn ($n \geq 30$)	$(\overline{X} \pm \varepsilon)$; $\varepsilon = z_\alpha . \frac{s}{\sqrt{n}}$
	Chuẩn $N(a,\sigma^2)$	σ^2 đã biết	$(\overline{X} \pm \varepsilon)$; $\varepsilon = z_\alpha . \frac{\sigma}{\sqrt{n}}$
	Chuẩn $N(a,\sigma^2)$	σ^2 chưa biết Mẫu nhỏ ($n < 30$)	$(\overline{X} \pm \varepsilon)$ $\varepsilon = T_\alpha . \frac{s}{\sqrt{n}}$
Phương sai σ^2	Chuẩn $N(a,\sigma^2)$	a chưa biết	$\left(\frac{(n-1).s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}, \frac{(n-1).s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \right)$

* Tìm giá trị z_α *thỏa* $\Phi(z_\alpha) = (1-\alpha)/2$ *: tra ngược bảng tích phân Laplace.*

* Tìm giá trị $T_\alpha = t_{\alpha/2}(n-1)$ *: tra bảng Student (PL VII), cột* $\alpha/2$*; dòng n-1.*

* Tìm giá trị $\chi^2_{\alpha/2}(n-1)$ *: tra bảng Chi bình phương (PL VI), cột* $\alpha/2$*; dòng n-1.*

Ví dụ 1:

Tìm khoảng ƯL cho tỉ lệ hạt lúa nảy mầm với độ tin cậy 98% trên cơ sở gieo 1000 hạt thì có 140 hạt không nảy mầm.

Hướng dẫn:

Gọi p là tỉ lệ hạt nảy mầm của tổng thể (đề bài không nhắc đến tổng thể cụ thể).

Khoảng UL (đối xứng) cho p có dạng (f- ε; f + ε)

Tính các đặc trưng mẫu: n = 1000; f= 860/1000 = 0,86.

Độ tin cậy 1 -α = 0,98 $\Rightarrow \Phi(z_\alpha) = (1-\alpha)/2 = 0,49 \Rightarrow z_\alpha = 2,33$.

Tìm độ chính xác của ƯL:

$$\varepsilon = \frac{z_\alpha \sqrt{f(1-f)}}{\sqrt{n}} = \frac{2,33 \times \sqrt{0,86 \times 0,14}}{\sqrt{1000}} \approx 0,0256$$

$\Rightarrow$ KƯL cho p: (f-ε; f+ε) = (0,8344; 0,8856) = (83,44%;88,56%)

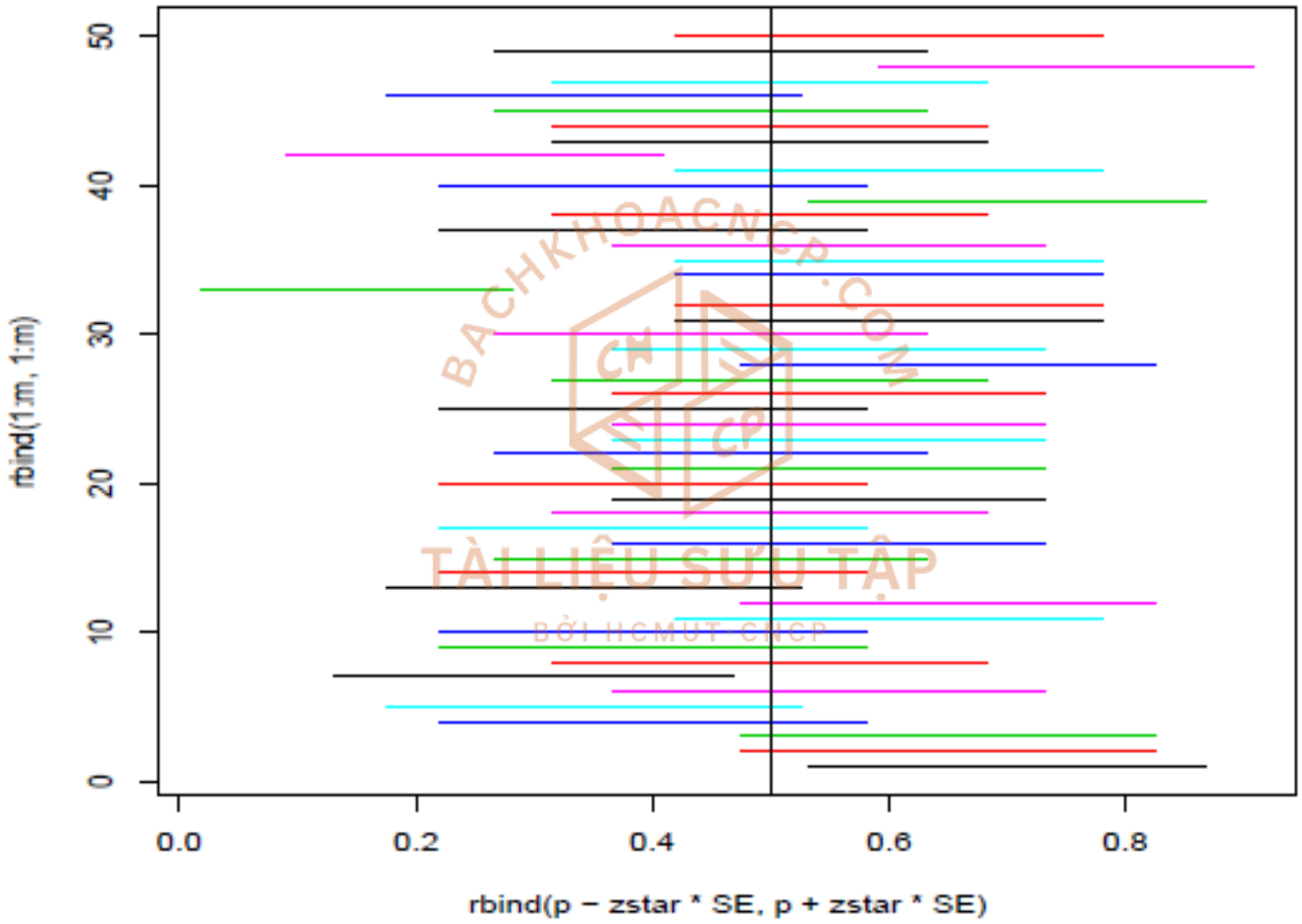
Lưu ý: *Vì p là 1 số , không phải BNN nên chỉ xảy ra 1 trong 2 khả năng:*

- *Nếu $p \in (0{,}8344; 0{,}8856)$ _ tức là kết quả đưa ra đúng.*
- *Nếu $p \notin (0{,}8344; 0{,}8856)$_ kết quả sai. KƯL trên không chứa p.*

Do đó người ta không viết $P(0{,}8344 < p < 0{,}8856) = 98\%$.

Độ tin cậy 98% được hiểu là trong tất cả các khoảng ƯL được xây dựng theo cách trên, (các khoảng ƯL này khác nhau do các mẫu cụ thể khác nhau), thì có 98% KƯL chứa giá trị p. Theo nguyên lý xác suất lớn, nếu ta lấy 1 mẫu cụ thể thì KƯL ta tìm được sẽ chứa p.

Hình ảnh sau đây minh họa cho kết quả của việc người ta dùng mô hình để tạo ngẫu nhiên 50 khoảng ước lượng có cùng độ tin cậy 90% cho giá trị p là xác suất tung đồng xu được mặt sấp. Với mỗi lần thực nghiệm, ta tung ngẫu nhiên 20 đồng xu.
(Giả thiết $n*F \geq 5$ và $n*(1-F) \geq 5$).



Verzani-SimpleR.pdf

Ví dụ 2:

Trong đợt vận động bầu cử ở một bang có khoảng 4 triệu cử tri, người ta phỏng vấn 1600 cử tri thì có 960 cử tri ủng hộ ứng cử viên A. Với độ tin cậy 97% , hãy dự đoán xem ứng cử viên A có khoảng bao nhiêu phiếu ủng hộ ở bang này?

Ví dụ 3:

Người ta muốn ước lượng tỉ lệ phế phẩm trong một lô hàng mới nhập về với độ tin cậy 99% và sai số không vượt quá 3%. Hãy cho biết để thỏa yêu cầu đó người ta phải kiểm tra ít nhất bao nhiêu sản phẩm với mỗi giả thiết sau:

a) Người ta đã lấy một mẫu sơ bộ thì thấy tỉ lệ phế phẩm trong mẫu này là 20%.

b) Chưa có thông tin gì liên quan đến tỉ lệ phế phẩm của lô hàng.

Ví dụ 4:

Để điều tra số cá trong một hồ, cơ quan quản lý đánh bắt 300 con, làm dấu rồi thả xuống hồ. Lần sau người ta bắt ngẫu nhiên 400 con thì thấy có 60 con đã được đánh dấu. Hãy xác định số cá trong hồ với độ tin cậy 96%.

Ví dụ 5:

Để nghiên cứu độ ổn định của 1 loại máy tiện người ta đo ngẫu nhiên đường kính (có phân phối chuẩn và đơn vị là mm) 24 trục máy do loại máy tiện này làm ra thì có kết quả dưới đây. Với độ tin cậy 98 %, hãy ước lượng đường kính trung bình và độ phân tán của đường kính trục máy.

24,1;	27,2;	26,7;	23,6;	24,6;	24,5;	26,4;	26,1;
25,8;	27,3;	23,2;	26,9;	27,1;	25,4;	23,3;	25,9;
22,7;	26,9;	24,8;	24,0;	23,4;	23,0;	24,3;	25,4.

Ví dụ 6:

Để xác định giá trung bình của mặt hàng B trên thị trường, người ta khảo sát ngẫu nhiên 100 cửa hàng và thu được số liệu:

Giá (nghìn đồng)	83	84	85	86	87	88	89	90
Số cửa hàng	6	7	12	15	30	10	10	10

a) Hãy tìm khoảng tin cậy cho giá trung bình của loại hàng hóa trên tại thời điểm đang xét với độ tin cậy 97% .

b) Nếu muốn độ dài của khoảng ước lượng không vượt quá 600 đồng và độ tin cậy của ước lượng là 99% thì cần phải điều tra thêm ít nhất bao nhiêu cửa hàng?

c) Với độ tin cậy 0,98, hãy ƯL số cửa hàng trong 8000 cửa hàng ở vùng đó bán thấp hơn giá bán lẻ 88 ngàn mà công ty đề nghị.

Hướng dẫn: a) $$\varepsilon = \frac{2,17 \times 1,8969}{\sqrt{100}} = 0,4116$$

KƯL cần tìm: (86,76-0,4116; 86,76+0,4116)= (86,3484; 87,1716)

$$Do\ \varepsilon' \le 0,3 \quad \Rightarrow n' \ge \left(\frac{2,58 \times 1,8969}{0,3} \right)^2 = 266,1251 \Rightarrow n' = 267.$$

b) Từ công thức: $$\varepsilon = \frac{z_\alpha \times s}{\sqrt{n}} \Rightarrow n = \left(\frac{z_\alpha \times s}{\varepsilon} \right)^2$$ *n ∈N, (làm tròn lên)*

KQ: Cần khảo sát thêm 267 – 100 = 167 cửa hàng nữa.

Lưu ý: Trong công thức trên, ε'; z_α' và n' là các kí hiệu trong mẫu cần tìm. Nhưng giá trị s' được lấy bằng giá trị s từ mẫu ban đầu đã có, mẫu này gọi là mẫu sơ bộ.

Ví dụ 7: Biết rằng thời gian thi công một chi tiết máy tuân theo quy luật phân phối chuẩn. Để định mức thời gian gia công một chi tiết máy, người ta theo dõi ngẫu nhiên quá trình thi công của 25 chi tiết và có được số liệu ở bảng sau:

Thời gian gia công (phút)	15-17	17-19	19-21	21-23	23-25	25-27
Số chi tiết máy tương ứng	1	3	4	12	3	2

a) Hãy tìm khoảng ước lượng cho thời gian gia công trung bình một chi tiết máy với độ tin cậy 0,95.

b) Hãy tìm khoảng ƯL cho phương sai với độ tin cậy 0,95.

Ví dụ 8: Để ước lượng doanh thu của 1 công ty có 380 cửa hàng trên toàn quốc trong 1 tháng, người ta chọn ngẫu nhiên 10% số cửa hàng và có bảng thống kê doanh thu trong 1 tháng như sau:

Doanh thu (triệu đồng / tháng)	20	40	60	80
Số cửa hàng	8	16	12	2

a) Với độ tin cậy 97%, hãy ƯL doanh thu trung bình của mỗi cửa hàng và doanh thu trung bình của công ty trong 1 tháng.

b) Nếu lấy độ dài của KƯL doanh thu trung bình mỗi cửa hàng trong 1 tháng là 6 triệu đồng thì độ tin cậy của khoảng ƯL khi đó là bao nhiêu?

Ví dụ 9:

Trọng lượng sản phẩm do một máy đóng gói là biến ngẫu nhiên tuân theo quy luật chuẩn với độ lệch chuẩn là 2,5 gram. Để ước lượng trọng lượng trung bình, người ta cân ngẫu nhiên 36 sản phẩm thì có được số liệu: $\bar{x} = 124,5\,gram;\ \ s = 2,35\,gram$

a) Hãy ước lượng trọng lượng trung bình của sản phẩm với độ tin cậy 95%.

b) Nếu muốn độ dài khoảng tin cậy trong câu a) không vượt quá 0,4 gram thì cần phải cân bao nhiêu sản phẩm?

c) Nếu người ta sử dụng mẫu đã có và quy ước lấy độ dài khoảng ước lượng đối xứng là 1 gram thì độ tin cậy tương ứng của khoảng ước lượng là bao nhiêu?

(Lưu ý: Vừa cho σ, vừa có s $\Rightarrow$ Sử dụng σ)

Ví dụ 10: Khảo sát chiều cao và cân nặng của một số bé trai 10 tuổi được lựa chọn ngẫu nhiên trong vùng, người ta có được số liệu mẫu dưới đây:

Y=Cân nặng (kg)	20-30	30-40	40-50	50-60
X=Chiều cao (cm)				
110-120	2	5		
120-130	4	9	6	
130-140	3	15	25	1
140-150		12	20	2
150-160		2	10	4

Với độ tin cậy 95%, hãy tìm các khoảng ước lượng cho:

a) Chiều cao trung bình và cân nặng trung bình của trẻ em trong vùng ở độ tuổi này.

b) Cân nặng trung bình của những trẻ có chiều cao từ 150cm trở lên.

c) Nếu muốn 2 khoảng ƯL trong câu a) có sai số tương ứng không vượt quá lần lượt là 1,5 cm và 1 kg thì ta cần lấy mẫu có kích thước tối thiểu là bao nhiêu?

d) Tỉ lệ trẻ có chiều cao từ 150 cm trở lên ở độ tuổi 10.

Giả thiết chiều cao và cân nặng của các bé trai ở độ tuổi này tuân theo quy luật phân phối chuẩn.

HD: *a) n=120; b) n=16, tính lại các đặc trưng và dùng công thức với mẫu nhỏ; c) $n' = \max\{n_1, n_2\}$; d) n = 120*

Chương III: KIỂM ĐỊNH GIẢ THIẾT THỐNG KÊ

III.1 Một số khái niệm:

- **Giả thiết kiểm định H_0** (*Null Hypothesis) gồm:*
  - Giả thiết về *tham số của tổng thể*
  - GT về dạng *phân phối của tổng thể*.
  - GT về *tính độc lập* của các BNN.

Giả thiết H_0 là giả thiết về yếu tố cần kiểm định của tổng thể ở trạng thái bình thường, không chịu tác động của các hiện tượng liên quan. Yếu tố trong H_0 phải được xác định cụ thể, ví dụ:

+ H_0 : Tỉ lệ nảy mầm của 1 loại hạt giống là 70%.
+ H_0: Thời gian công nhân hoàn thành 1 sản phẩm là BNN có phân phối chuẩn với kỳ vọng là 20 phút và phương sai là 9 phút^2.
+ H_0: Mức độ yêu thích của khán giả với chương trình truyền hình *"Tìm kiếm tài năng "* không phụ thuộc vào lứa tuổi.

- **Giả thiết đối H_1** *(Alternative Hypothesis)* là một mệnh đề mâu thuẫn với H_0, H_1 thể hiện xu hướng cần kiểm định.

Vì ta sẽ dựa vào thông tin thực nghiệm của mẫu để kết luận xem có thừa nhận các giả thiết nêu trên hay không nên công việc này gọi là *kiểm định thống kê*.

- **Tiêu chuẩn kiểm định** là hàm thống kê $G = G(X_1, X_2, .., X_n, \theta_0)$, xây dựng trên mẫu ngẫu nhiên $W = (X_1, X_2, .., X_n)$ và tham số θ_0 liên quan đến H_0; Điều kiện đặt ra với thống kê G là nếu H_0 đúng thì quy luật phân phối xác suất của G phải hoàn toàn xác định.

- **Miền bác bỏ giả thiết W_α** là miền thỏa $P(G \in W_\alpha / H_0 \text{ đúng}) = \alpha$. α là một số khá bé, thường không quá 0,05 và gọi là ***mức ý nghĩa*** của kiểm định. Có vô số miền W_α như vậy.

- Quy tắc kiểm định: Từ mẫu thực nghiệm, ta tính được một giá trị cụ thể của tiêu chuẩn kiểm định là thống kê $g_{qs} = G(x_1, x_2, .., x_n, \theta_0)$. Theo nguyên lý xác suất bé, biến cố $G \in W_\alpha$ có xác suất nhỏ nên với 1 mẫu thực nghiệm, nó không thể xảy ra. Do đó:

+ Nếu $g_{qs} \in W_\alpha$ thì bác bỏ H_0 , thừa nhận giả thiết H_1 .

+ Nếu $g_{qs} \notin W_\alpha$: ta chưa đủ dữ liệu khẳng định H_0 sai. Ta nói "có thể chấp nhận H_0 " hay " không bác bỏ H_0".

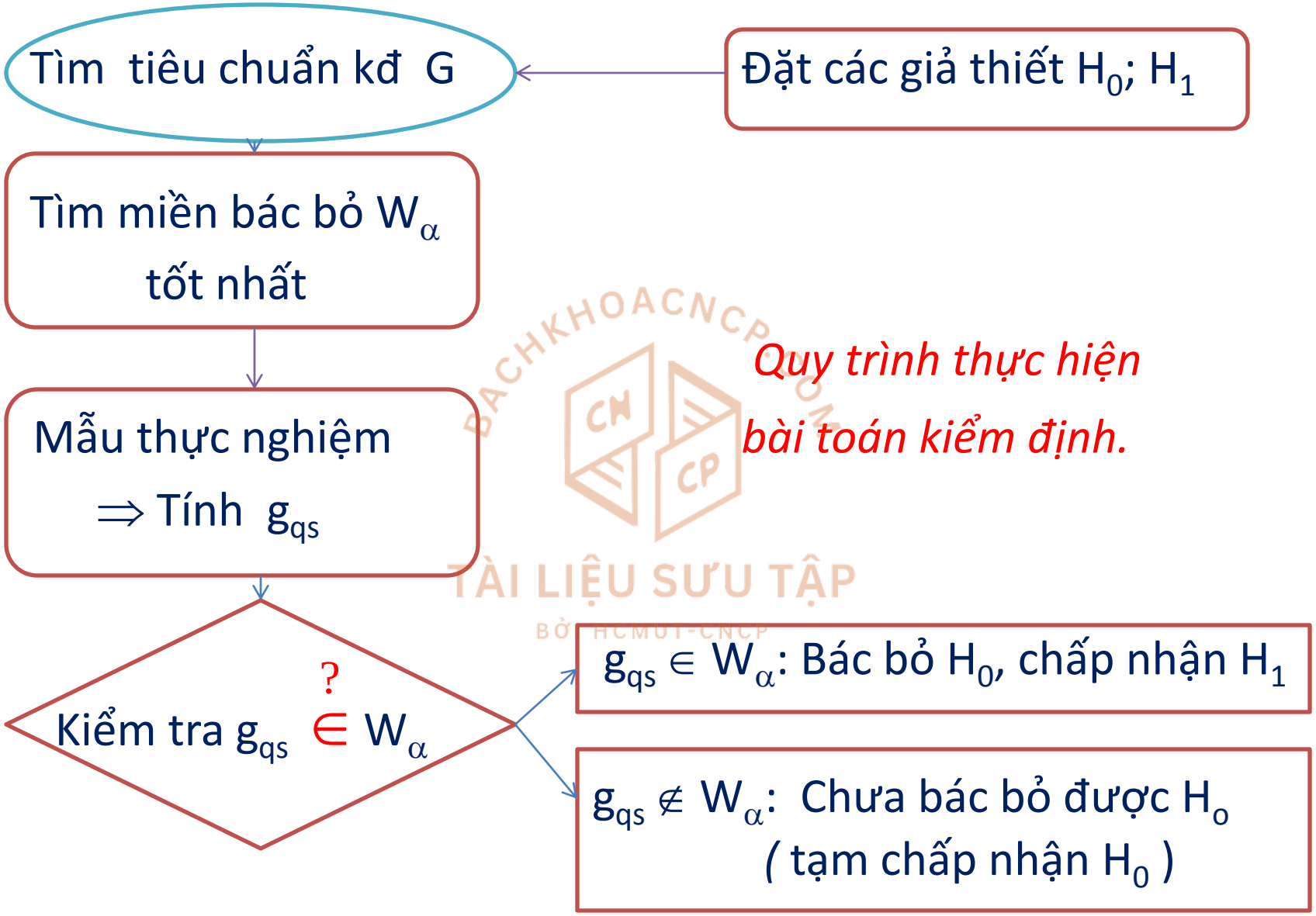
Kết luận của một bài toán kiểm định có thể mắc các sai lầm sau:

- **Sai lầm loại I**: Bác bỏ giả thiết H_0 trong khi H_0 đúng. Xác suất mắc phải sai lầm này nếu H_0 đúng chính bằng mức ý nghĩa α. Nguyên nhân mắc phải sai lầm loại I thường có thể do kích thước mẫu quá nhỏ, có thể do phương pháp lấy mẫu ...
- **Sai lầm loại II**: Thừa nhận H_0 trong khi H_0 sai, tức là mặc dù thực tế H_1 đúng nhưng giá trị thực nghiệm g_{qs} không thuộc W_α.

Tình huống / Quyết định	**H_0 đúng**	**H_0 sai**
Bác bỏ H_0	Sai lầm loại I. Xác suất = α	Quyết định đúng.
Không bác bỏ H_0	Quyết định đúng.	Sai lầm loại II. Xác suất =β

Ví dụ: Người bán hàng nói rằng tỉ lệ phế phẩm trong mỗi lô hàng không quá 5%. Người mua quyết đinh kiểm ngẫu nhiên 10 sản phẩm, nếu được cả 10 sản phẩm tốt thì mới mua lô hàng. Sai lầm loại I xảy ra khi người mua từ chối mua hàng trong khi thực sự lô hàng có không quá 5% phế phẩm; α là mức rủi ro cho bên bán. Sai lầm loại II xảy ra khi người mua nhận hàng nhưng tỉ lệ phế phẩm thực ra trên 5%; β chính là mức rủi ro cho bên mua.

Với một mẫu xác định, khi ta giảm α đi thì đồng thời sẽ làm tăng β và ngược lại. Chỉ có thể cùng giảm α, β nếu tăng kích thước mẫu. Người ta thường có xu hướng coi trọng xác suất mắc sai lầm loại I nên sẽ hạn chế trước giá trị α tùy thực tế, và sau đó phải tìm miền W_α sao cho xác suất mắc sai lầm loại II là nhỏ nhất. Miền W_α thỏa yêu cầu này được gọi là miền bác bỏ ***tốt nhất*** dựa trên các cơ sở toán học chặt chẽ.



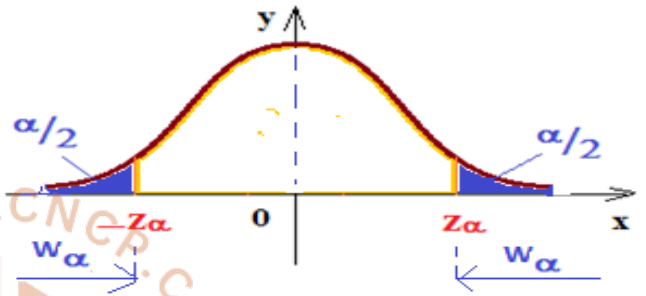
Quy trình thực hiện bài toán kiểm định.

Ví dụ minh họa cho các miền bác bỏ khi tiêu chuẩn kiểm định Z có phân phối chuẩn N(0,1).

1. *Miền bác bỏ 2 phía:*

$W_\alpha = (-\infty, -Z_\alpha) \cup (Z_\alpha, +\infty)$

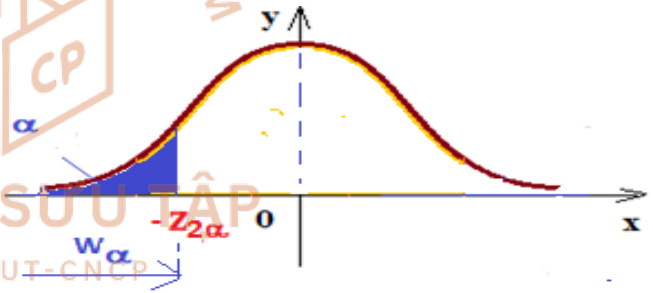
ở đây $\Phi(Z_\alpha) = \dfrac{(1-\alpha)}{2}$



2. *Miền bác bỏ bên trái:*

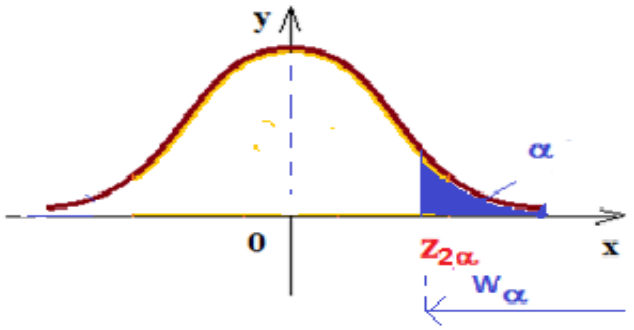
$W_\alpha = (-\infty, -Z_{2\alpha})$

ở đây $\Phi(Z_{2\alpha}) = \dfrac{(1-2\alpha)}{2}$



3. *Miền bác bỏ bên phải:*

$W_\alpha = (Z_{2\alpha}, +\infty)$



III.2 Bài toán kiểm định tham số:

III.2.1 Bài toán kiểm định tỉ lệ:

	Giả thiết KĐ H_0	Giả thiết đối H_1	Tiêu chuẩn kiểm định	Miền bác bỏ H_0 với mức ý nghĩa α
BT 1 mẫu $n \geq 30$	$p = p_0$	$p \neq p_0$	$Z_{qs} = \dfrac{F - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$	$W_\alpha = (-\infty, -z_\alpha) \cup (z_\alpha, +\infty)$
		$p < p_0$		$W_\alpha = (-\infty, -z_{2\alpha})$
		$p > p_0$		$W_\alpha = (z_{2\alpha}, +\infty)$
BT 2 mẫu $n_1 \geq 30$ $n_2 \geq 30$	$p_1 = p_2$	$p_1 \neq p_2$	$Z_{qs} = \dfrac{F_1 - F_2}{\sqrt{\overline{f}(1-\overline{f})\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$; mẫu gộp: $\overline{f} = \dfrac{n_1F_1 + n_2F_2}{n_1 + n_2}$	$W_\alpha = (-\infty, -z_\alpha) \cup (z_\alpha, +\infty)$
		$p_1 < p_2$		$W_\alpha = (-\infty, -z_{2\alpha})$
		$p_1 > p_2$		$W_\alpha = (z_{2\alpha}, +\infty)$

Bảng 3: Tóm tắt một số công thức của bài toán kiểm định tỉ lệ

Ở BT 2 mẫu, khi dùng mẫu cụ thể $f_1 = \dfrac{m_1}{n_1}; f_2 = \dfrac{m_2}{n_2} \Rightarrow \overline{f} = \dfrac{m_1 + m_2}{n_1 + n_2}$

Ví dụ 11: *Theo số liệu công bố của một công ty dịch vụ tin học, tỷ lệ khách hàng hài lòng với dịch vụ của công ty là 85%. Một khảo sát độc lập cho thấy trong mẫu gồm 145 khách hàng của công ty có 120 khách hàng hài lòng. Với mức ý nghĩa 3%, có thể coi số liệu của công ty là đáng tin cậy không?*

Hướng dẫn: Gọi p là tỉ lệ khách hàng hài lòng với dịch vụ của CT.

GtKđ H_0: $p = 85\%$

Giả thiết đối H_1: $p \neq 85\%$

+ Mức ý nghĩa $\alpha = 3\% \Rightarrow \Phi(z_\alpha) = (1- 0,03)/2 = 0,485 \Rightarrow z_\alpha = 2,17$

Miền b/bỏ $W_\alpha = (-\infty, - z_\alpha) \cup (z_\alpha ,+\infty) = (-\infty; - 2,17) \cup (2,17 ; +\infty)$

Kích thước mẫu: n = 145; Tỉ lệ mẫu: f = 120/145= 0,8276

+ Tiêu chuẩn kđ:

$$Z_{qs} = \frac{f - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} = \frac{0,8276 - 0,85}{\sqrt{0,85(1-0,85)}}\sqrt{145} = -0,7559$$

Do $Z_{qs} \notin W_\alpha$ nên ta chưa đủ dữ kiện bác bỏ H_0.

Có thể tạm xem như số liệu của công ty là đáng tin.

Ví dụ 12: *Theo tiêu chuẩn của công ty thì một lô hàng nguyên liệu được chấp nhận nếu không có quá 3% phế phẩm. Kiểm tra ngẫu nhiên 400 sản phẩm từ lô hàng này thì thấy 16 phế phẩm. Với mức ý nghĩa 5%, hãy xem xét lô hàng này có thể được chấp nhận không?*

Hướng dẫn: + Gọi p là tỉ lệ phế phẩm thực sự của lô hàng.

GtKđ H_0: $p = 3\%$ (hay $p \le 3\%$)

Giả thiết đối H_1: $p > 3\%$

+ Myn $\alpha = 5\% \Rightarrow \Phi(z_{2\alpha}) = (1- 2*0,05)/2 = 0,45 \Rightarrow z_{2\alpha} = 1,645$

Miền bác bỏ $W_\alpha = (z_{2\alpha} ; +\infty) = (1,645 ; +\infty)$

Kích thước mẫu: n = 400; Tỉ lệ mẫu: f = 16/400 = 0,04.

+ TC kiểm định: $$Z_{qs} = \frac{f - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} = \frac{0,04-0,03}{\sqrt{0,03(1-0,03)}}\sqrt{400} = 1,172$$

Do $Z_{qs} \notin W_\alpha$ nên ta chưa bác bỏ H_0, tức là chưa thể kết luận tỉ lệ phế phẩm của lô hàng vượt ngưỡng cho phép.

Ví dụ 13: *Tỉ lệ bệnh nhân bị bệnh T được chữa khỏi bệnh bằng thuốc A là 85%. Khi dùng thuốc B điều trị thì trong 1100 bệnh nhân bị bệnh T người ta thấy có 903 người khỏi bệnh. Có thể nói rằng thuốc B điều trị ít hiệu quả hơn thuốc A được không, kết luận với mức ý nghĩa 4%?*

Hướng dẫn: + Gọi p là tỉ lệ BN khỏi bệnh khi dùng thuốc B.

GtKđ H_0: $p = 85\%$

Giả thiết đối H_1: $p < 85\%$

+ Myn $\alpha = 4\% \Rightarrow \Phi(z_{2\alpha}) = (1- 2*0,04)/2 = 0,46 \Rightarrow z_{2\alpha} = 1,75$

Miền bác bỏ $W_\alpha = (-\infty; - z_{2\alpha}) = (-\infty; - 1,75)$

Kích thước mẫu: n = 1100; Tỉ lệ mẫu: f = 903/1100 .

+ Tiêu chuẩn kđ:

$$Z_{qs} = \frac{f - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} = \frac{\frac{903}{1100} - 0,85}{\sqrt{0,85(1-0,85)}}\sqrt{1100} = -2,7021$$

+ Do $Z_{qs} \in W_\alpha$ nên bác bỏ H_0, chấp nhận H_1. Xem như tỉ lệ BN khỏi bệnh khi dùng thuốc B là thấp hơn so với dùng thuốc A.

Ví dụ 14: *Khảo sát ngẫu nhiên 80 sinh viên nam thấy có 56 bạn thường xuyên đi xe buýt; trong 60 SV nữ thì con số này là 48. Có thể coi như tỷ lệ SV nam đi xe buýt thường xuyên là thấp hơn so với SV nữ hay không? Hãy kết luận với mức ý nghĩa 5%?*

Hướng dẫn:+ Gọi p_1, p_2 lần lượt là tỉ lệ SV nam & nữ đi xe buýt tx.

GtKđ H_0: $p_1 = p_2$; Giả thiết đối H_1: $p_1 < p_2$

+ Myn $\alpha = 5\% \Rightarrow \Phi(z_{2\alpha}) = (1- 2*0,05)/2 = 0,45 \Rightarrow z_{2\alpha} = 1,645$

Miền bác bỏ $W_\alpha = (-\infty; -1,645)$

$n_1 = 80$; $f_1 = 56/80$; $n_2 = 60$; $f_2 = 48/60$; $\overline{f}$ =(56+48)/(60+80)

+ Tiêu chuẩn kđ:

$$Z_{qs} = \frac{f_1 - f_2}{\sqrt{\overline{f}(1-\overline{f})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{\frac{56}{80} - \frac{48}{60}}{\sqrt{\frac{104}{140}\left(1-\frac{104}{140}\right)\left(\frac{1}{80}+\frac{1}{60}\right)}} = -1,3397$$

+ Do $Z_{qs} \notin W_\alpha$ nên chưa bác bỏ được H_0. Xem như tỉ lệ sinh viên nam thường xuyên đi xe buýt không thấp hơn so với SV nữ.

Ví dụ 15: *Tỉ lệ phế phẩm của 1 nhà máy là 10%. Sau khi cải tiến quy trình sản xuất, người ta kiểm tra thử 250 sản phẩm thì thấy có 17 phế phẩm.*

Bài toán a) Với mức ý nghĩa 5%, hãy cho biết có thể coi như việc cải tiến quy trình sản xuất đã làm thay đổi tỷ lệ phế phẩm của nhà máy không?

Bài toán b) Với mức ý nghĩa 5%, có thể cho rằng việc cải tiến quy trình sản xuất đã có hiệu quả hay không?

Ví dụ 16:

Người ta bảo quản cùng 1 loại hạt giống theo 2 phương pháp khác nhau trong thời gian như nhau. Gieo thử ngẫu nhiên 500 hạt giống đã được bảo quản theo phương pháp I thì thấy có 450 hạt nảy mầm; gieo thử 700 hạt giống đã được bảo quản theo phương pháp II thì có 600 hạt nảy mầm. Với myn 2%, có thể xem như phương pháp I có hiệu quả hơn phương pháp II hay không?

III.2.2 Bài toán kiểm định trung bình:

	GT KĐ H_0	GT đối H_1	Tiêu chuẩn kiểm định	Miền bác bỏ H_0 với mức ý nghĩa α	
BT 1 mẫu			- Nếu đã cho σ^2 : $Z_{qs} = \frac{\overline{X} - a_0}{\sigma}\sqrt{n}$ - Nếu chưa cho σ^2 : $Z_{qs} = \frac{\overline{X} - a_0}{s}\sqrt{n}$	*-Tổng thể phân phối chuẩn, đã biết σ^2. - Hoặc tt tùy ý, $n \geq 30$*	*- Tổng thể phân phối chuẩn; - Chưa biết σ^2. - $n < 30$.*
	$a = a_0$	$a \neq a_0$		$W_\alpha = (-\infty, -z_\alpha) \cup (z_\alpha, +\infty)$	$W_\alpha = (-\infty, -t_{\frac{\alpha}{2}}(n-1)) \cup (t_{\frac{\alpha}{2}}(n-1), +\infty)$
		$a < a_0$		$W_\alpha = (-\infty, -z_{2\alpha})$	$W_\alpha = (-\infty, -t_\alpha^{(n-1)})$
		$a > a_0$		$W_\alpha = (z_{2\alpha}, +\infty)$	$W_\alpha = (t_\alpha^{(n-1)}, +\infty)$

SV cần sử dụng bảng công thức kiểm định so sánh 2 trung bình (BT 2 mẫu) đầy đủ hơn trong file kèm theo: "Tóm tắt công thức bài toán kiểm định trung bình 2 tổng thể"

Ví dụ 17: *Một công ty sản xuất phomát nghi ngờ một nhà cung cấp sữa cho công ty đã pha thêm nước vào sữa để làm tăng lượng sữa cung cấp. Nếu sữa có pha nhiều nước quá mức bình thường thì nhiệt độ đông của nó sẽ thấp hơn so với sữa tự nhiên. Biết rằng điểm đông của sữa tự nhiên tuân theo quy luật phân phối chuẩn với trung bình khoảng -0,545^{0}C, độ lệch chuẩn 0,008^{0}C. Người ta kiểm định chất lượng sữa trong các container hàng mới nhập bằng cách lấy ra 25 mẫu ngẫu nhiên thì thấy nhiệt độ đông trung bình của sữa trong mẫu là -0,55^{0}C. Hãy kết luận về chất lượng sữa mà công ty mua với mức ý nghĩa 1%.*

Hướng dẫn:

+ Gọi a là nhiệt độ đông trung bình của lượng sữa mới nhập.

GtKđ H_0: $a = -0{,}545^0C$; Giả thiết đối H_1: $a < -0{,}545^0C$

$n_1 = 25 < 30$; $\sigma = 0{,}008$ (đã biết); $a_0 = -0{,}545^0C$. $\bar{x} = -0{,}55^0C$

+ Myn $\alpha = 1\% \Rightarrow \Phi(z_{2\alpha}) = (1 - 2*0{,}01)/2 = 0{,}49 \Rightarrow z_{2\alpha} = 2{,}33$

Miền bác bỏ $W_\alpha = (-\infty; -2{,}33)$

+ Tiêu chuẩn kđ:

$$Z_{qs} = \frac{\bar{x} - a_0}{\sigma}\sqrt{n} = \frac{(-0,55) - (-0,545)}{0,008}\sqrt{25} = -3,125$$

+ Do $Z_{qs} \in W_\alpha$ nên bác bỏ H_0, chấp nhận H_1. Ta kết luận lượng sữa công ty mới mua đã bị pha nước.

Ví dụ 18 *Người ta đã thực hiện một cải tiến kỹ thuật trong bộ hòa khí của xe ôtô với hy vọng sẽ tiết kiệm được xăng hơn. Cho xe chạy thử 12 lần thì họ có số km chạy được cho 1 lít xăng:*

20,6 20,5 20,8 20,8 20,7 20,6

21 20,6 20,5 20,4 20,3 20,7

Nếu trước khi cải tiến, 1 lít xăng trung bình chạy được 20,4 km thì với số liệu này người ta đã có thể kết luận việc cải tiến mang lại hiệu quả đáng kể hay không, với mức ý nghĩa 5% ?

Hướng dẫn: *Giả thiết bổ sung: Quãng đường xe chạy được khi tiêu thụ 1 lít xăng là biến ngẫu nhiên có phân phối chuẩn.*

+ Gọi a là quãng đường trung bình ôtô chạy được với 1 lít xăng sau khi cải tiến kỹ thuật.

$n_1 = 12 < 30$; $\overline{x} = 20{,}625$ $a_0 = 20{,}4$ $s = 0{,}1913$

Giả thiết kiểm định H_0: $a = 20{,}4$

Giả thiết đối H_1: $a > 20{,}4$

+ Myn $\alpha = 5\% \Rightarrow$ Tra bảng Student 1 phía: $t_\alpha(n-1) = t_{0,05}(11) = 1{,}796$

Miền bác bỏ $W_\alpha = (1{,}796; +\infty)$

+ Tiêu chuẩn kiểm định:

$$Z_{qs} = \frac{\overline{x} - a_0}{s}\sqrt{n} = \frac{20{,}625 - 20{,}4}{0{,}1913}\sqrt{12} = 4{,}0743$$

+ Do $Z_{qs} \in W_\alpha$ nên bác bỏ H_0, chấp nhận H_1. Việc cải tiến kỹ thuật đã có hiệu quả.

Ví dụ 19 *Ở một phân xưởng, người ta định mức thời gian gia công 1 chi tiết cho mỗi công nhân là 12 phút. Sau khi thay đổi nguyên liệu, người ta khảo sát ngẫu nhiên quá trình gia công của 1 số chi tiết và thu được số liệu dưới đây. Với myn 5%, hãy quyết định xem có cần thay đổi định mức gia công hay không?*

Thời gian gia công 1 chi tiết (phút)	10-10,5	10,5-11	11-11,5	11,5-12	12-12,5	12,5-13	13-13,5
Số chi tiết t/ư	4	12	26	37	43	28	10

Hướng dẫn:

+ Gọi a là thời gian gia công TB 1 chi tiết ở thời điểm hiện tại.

+ GTKĐ H_0: a = 12 phút. GTĐ H_1: a ≠ 12 phút

+ Miền bác bỏ $W_\alpha = (-\infty; -1,96) \cup (1,96; +\infty)$

+ Tckđ: $$Z_{qs} = \frac{\overline{x} - a_0}{s}\sqrt{n} = \frac{11,9594 - 12}{0,7170}\sqrt{160} = -0,7163$$

+ Do $Z_{qs} \notin W_\alpha$ nên chưa bác bỏ được H_0. Vì vậy chưa cần thay đổi định mức.

Ví dụ 20 *Người ta trồng cùng 1 giống lúa trên 2 thửa rộng như nhau và bón 2 loại phân khác nhau, đến ngày thu hoạch họ lấy mẫu trên 2 thửa ruộng và có kết quả khảo sát như sau:*

	Số bông k/s	Số hạt trung bình/1 bông	Độ lệch mẫu HC
Thửa ruộng 1	1000	70	10
Thửa ruộng 2	500	72	20

Với mức ý nghĩa 5%, hãy kết luận xem sự khác nhau giữa 2 trung bình mẫu là ngẫu nhiên hay bản chất.

Hướng dẫn: + Gọi a_1; a_2 là số hạt lúa TB trên 1 bông ở mỗi thửa.

+ GTKĐ H_0: $a_1 = a_2$. GTĐ H_1: $a_1 \neq a_2$

+ Miền bác bỏ $W_\alpha = (-\infty; -1,96) \cup (1,96; +\infty)$

+ Tckđ: $$Z_{qs} = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{70 - 72}{\sqrt{\frac{10^2}{1000} + \frac{20^2}{500}}} = 2,1082$$

+ Do $Z_{qs} \in W_\alpha$ nên bác bỏ H_0. Số hạt TB trên mỗi bông ở 2 thửa ruộng là khác nhau, nên sự khác nhau giữa 2 TB mẫu là có nghĩa.

Ví dụ 21

Khảo sát thu nhập (đơn vị: triệu đồng) trong 3 tháng đầu năm của các công nhân trong 2 nhà máy có điều kiện làm việc như nhau, người ta có được kết quả:

Nhà máy 1	**18.5**	**19**	**19.3**	**20**	**20.2**	**21**	**21.5**	**19**	**19.7**	**20**
Nhà máy 2	**17.3**	**18**	**19**	**20**	**20.6**	**20.9**	**18.2**	**19.6**	**20.8**	

Với mức ý nghĩa 5%, có thể cho rằng thu nhập trung bình của công nhân 2 nhà máy đó trong 3 tháng đầu năm là như nhau hay không, biết thu nhập của công nhân ở 2 nhà máy có phân phối chuẩn và có phương sai bằng nhau.

Hướng dẫn:

Đây là bài toán t-test với giả thiết 2 phương sai tổng thể như nhau.

Gọi a_1; a_2 là thu nhập trung bình 3 tháng đầu năm của công nhân 2 nhà máy.

Giả thiết kiểm định H_o: $a_1 = a_2$; H_1: $a_1 \neq a_2$

$n_1 = 10$ $\overline{x_1} = 19,82$ $s_1^2 = 0,8662$

$n_2 = 9$ $\overline{x_2} = 19,3777$ $s_2^2 = 1,7519$

+ Miền bác bỏ $W_\alpha = (-\infty; - t_{\alpha/2} (n_1+n_2-2) \cup (t_{\alpha/2} (n_1+n_2-2); +\infty)$

$= (-\infty; - 2,1098) \cup (2,1098; +\infty)$

+ Do giả thiết phương sai 2 tổng thể chưa biết và $\sigma_1^2 = \sigma_2^2$, nên ta cần tính thêm phương sai gộp:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = 1,2830$$

suy ra tiêu chuẩn kiểm định:

$$T_{qs} = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = 0,8497 \notin W\alpha$$

+ Có thể chấp nhận H_0. Thu nhập của CN 2 nhà máy là như nhau.

III.2.3 Bài toán kiểm định phương sai:

	Giả thiết KĐ H_0	Giả thiết đối H_1	ĐK của PP tổng thể	Tiêu chuẩn kiểm định	Miền bác bỏ H_0 với mức ý nghĩa α
BT 1 mẫu	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	-Bất kỳ khi mẫu lớn. -PP chuẩn, khi n nhỏ.	$\chi^2_{qs} = \frac{(n-1)S^2}{\sigma_0^2}$	$W_\alpha = [0, \chi^2_{1-\frac{\alpha}{2}}(n-1)) \cup (\chi^2_{\frac{\alpha}{2}}(n-1), +\infty)$
		$\sigma^2 < \sigma_0^2$			$W_\alpha = [0, \chi^2_{1-\alpha}(n-1))$
		$\sigma^2 > \sigma_0^2$			$W_\alpha = (\chi^2_{\alpha}(n-1), +\infty)$
BT 2 mẫu	$\sigma^2_1 = \sigma^2_2$	$\sigma^2_1 \neq \sigma^2_2$	-Bất kỳ khi mẫu lớn. -PP chuẩn, khi n nhỏ. - Chưa biết a_1, a_2.	$F_{qs} = \frac{S_1^2}{S_2^2}$ **Luôn lưu ý đặt $S_1 > S_2$**	$W_\alpha = (f_{\frac{\alpha}{2}}(n_1-1; n_2-1), +\infty)$
		$\sigma^2_1 > \sigma^2_2$			$W_\alpha = (f_{\alpha}(n_1-1; n_2-1), +\infty)$ *Tra bảng Fisher*

Trong bài toán so sánh phương sai 2 tổng thể, để xác định miền bác bỏ 1 cách đơn giản ta có thể chọn mẫu 1 là mẫu có phương sai mẫu hiệu chỉnh lớn hơn. Công thức đầy đủ cho BTL xem trong file " Tóm tắt..."

Ví dụ 22:

Chọn ngẫu nhiên đường kính 41 vòng bi do một máy sản xuất thì thấy độ lệch chuẩn trong mẫu là 0,003 cm. Theo quy định thì độ lệch chuẩn của vòng bi không được vượt quá 0,0025 cm. Với mức ý nghĩa 5%, hãy kết luận về độ ổn định của máy.

Hướng dẫn:

Gọi σ^2 là phương sai của đường kính các vòng bi do máy sản xuất

Giả thiết kiểm định H_0: $\sigma^2 = (0{,}0025\ \text{cm})^2$

Giả thiết đối: H_1: $\sigma^2 > (0{,}0025\ \text{cm})^2$

Miền bác bỏ $W_\alpha = (\chi^2_{0,05}(40); +\infty) = (55{,}76; +\infty)$

Tiêu chuẩn kiểm định:

$$\chi^2_{qs} = \frac{(n-1)s^2}{\sigma_0^2} = \frac{40 \times 0{,}003^2}{0{,}0025^2} = 57{,}6 \quad \in W_\alpha$$

Từ đó bác bỏ giả thiết H_0. Ta coi như máy hoạt động không ổn định do đường kính các vòng bi phân tán quá mức cho phép.

Ví dụ 23: *(tham khảo cho BTL)*

Một nhà máy đang thử nghiệm 2 quy trình khác nhau cùng sản xuất một loại sản phẩm. Để kiểm tra sự ổn định của hàm lượng chất A trong các sản phẩm ở 2 quy trình có như nhau không, người ta khảo sát 2 mẫu và có được kết quả:

Quy trình 1: $n_1 = 41$ Độ lệch mẫu HC: $s_1 = 2,889$

Quy trình 2: $n_2 = 30$ Độ lệch mẫu HC: $s_1 = 2,113$

Với mức ý nghĩa 5%, hãy nêu kết luận về sự đồng đều của hàm lượng chất A trong các sản phẩm ở 2 quy trình trên.

Hướng dẫn:

Gọi σ_1^2 ; σ_2^2 là phương sai của hàm lượng chất A trong sp ở Qt1; Qt2.

Giả thiết kiểm định H_0: $\sigma_1^2 = \sigma_2^2$ H_1: $\sigma_1^2 > \sigma_2^2$

Miền bác bỏ $W_\alpha = (f_\alpha(n_1-1; n_2-1) + \infty) = (1,8; +\infty)$

Tiêu chuẩn kiểm định:

$$F_{qs} = \frac{S_1^2}{S_2^2} = \frac{2,889^2}{2,113^2} = 1,8694 \in W_\alpha$$

Từ đó bác bỏ giả thiết H_0, chấp nhận H_1. Hàm lượng chất A trong các sản phẩm ở quy trình 1 kém đồng đều hơn so với quy trình 2.

III.3 Bài toán kiểm định phi tham số: (xét KĐ Chi Bình Phương)

III.3.1 Bài toán kiểm định tính độc lập: (so sánh các ti lệ)

Xét một mẫu kích thước n của BNN **định tính** 2 chiều (X,Y).
X nhận các giá trị A_1; A_2;….; A_k. Y nhận các giá trị B_1; B_2; …; B_h.

Y / X	B_1	B_2	…	B_h	Tổng hàng
A_1	n_{11}	n_{12}	…	n_{1h}	n_1
A_2	n_{21}	n_{22}	…	n_{2h}	n_2
…	…	…	…	…	…
A_k	n_{k1}	n_{k2}	…	n_{kh}	n_k
Tổng cột	m_1	m_2	…	m_h	$\sum ni = n$

Hãy kiểm định xem X,Y có độc lập hay không với mức ý nghĩa α.

* Giả thiết kiểm định H_0: X,Y độc lập.
  Giả thiết đối H_1: X, Y không độc lập.
* Miền bác bỏ $W_\alpha = (\chi^2_\alpha$ (số hàng-1)*(số cột-1) ; $+\infty)$
* Do giả thiết X,Y độc lập nên xác suất tính theo lý thuyết là

$$p_{ij} = P(X= A_i ; Y= B_j) = P(X= A_i)*P(Y= B_j) = \frac{n_i}{n} * \frac{m_j}{n}$$

suy ra tần số lý thuyết là:

$$E_{ij} = n*p_{ij} = \frac{n_i * m_j}{n} = \frac{tong\ hang\ i * tong\ cot\ j}{kich\ thuoc\ mau}$$

Tính tiêu chuẩn kiểm định:

$$\chi^2_{qs} = \sum_{i;j} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = \sum_{i;j} \frac{\left(Observed_{ij} - Expected_{ij}\right)^2}{Expected_{ij}}$$

hoặc:

$$\chi^2_{qs} = n.\left[\sum_{i;j} \frac{n_{ij}^{\ 2}}{n_i \times m_j}\right] - n$$

* B4: Kết luận.

Ví dụ 24: *Ở 1 trường đại học, để nghiên cứu xem khả năng học toán của sinh viên có tương quan gì với sự yêu thích môn thống kê hay không, người ta khảo sát ngẫu nhiên 200 SV và có kết quả:*

Mức độ yêu thích môn thống kê	Khả năng học toán		
	Thấp	Trung bình	Cao
Ít thích	60	15	15
Thích vừa	15	45	10
Rất thích	5	10	25

Với myn α = 0,05, kiểm định xem sự yêu thích môn thống kê có phụ thuộc vào khả năng học toán của sinh viên trường này không?

Hướng dẫn: Gọi X : mức độ yêu thích đối với môn TK của SV.

Gọi Y : mức độ thể hiện khả năng học toán của sinh viên.

+ Giả thiết kiểm định H_0: X, Y độc lập

H_1: X,Y không độc lập.

+ Miền bác bỏ: $W_\alpha = (\chi^2_\alpha (3-1)*(3-1) ; +\infty) = (9,49; +\infty)$.

+ Tính tiêu chuẩn kiểm định:

Cách 1: ***Lập bảng tần số thực nghiệm*** O_{ij} ***Bảng tần số lý thuyết*** E_{ij}

60	15	15	90
15	45	10	70
5	10	25	40
80	70	50	n =200

→*tổng hàng1*

↓ *Tổng cột 1*

$\frac{90\times80}{200}=36$	$\frac{90\times70}{200}=31,5$	$\frac{90\times50}{200}=22,5$
...	...	...
...	$\frac{40\times70}{200}=14$	$\frac{40\times50}{200}=10$

$$\chi^2_{qs}=\sum_{i;j}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}$$

$$=\left[\frac{(60-36)^2}{36}+\frac{(15-31,5)^2}{31,5}+\ldots+\frac{(25-10)^2}{10}\right]=84,7513\in W_\alpha$$

Cách 2: (dùng công thức tắt, không lập bảng E_{ij})

$$\chi^2_{qs}=200\times\left[\frac{60^2}{90\times80}+\frac{15^2}{90\times70}+\ldots+\frac{25^2}{40\times50}\right]-200=84,7513\in W_\alpha$$

Bác bỏ H_0, chấp nhận H_1. Mức độ yêu thích của SV đối với môn học thống kê có liên quan đến khả năng học toán.

Ví dụ 25: *Khi khảo sát tình trạng việc làm của sinh viên ngành QTKD sau khi ra trường 1 năm, ta có được số liệu trong mẫu sau:*

Trường mà SV tốt nghiệp	SV chưa có việc làm	SV có việc chưa phù hợp	SV có việc làm phù hợp	SV chưa lựa chọn việc làm
Trường A	3	17	58	2
Trường B	5	31	48	6
Trường C	2	40	80	8

Với myn 1%, có thể coi tình trạng việc làm của sinh viên phụ thuộc vào trường mà sinh viên đó đã tốt nghiệp hay không?
(Với myn 1%, có thể coi tình trạng việc làm của sinh viên các trường A,B,C có phân bố tỉ lệ như nhau hay không?)

Ví dụ 26:

Để so sánh trình độ tiếng Anh đầu vào ở 2 trường THPT A và B, người ta đã khảo sát điểm kiểm tra của 290 học sinh lớp 10. Kết quả xếp loại thu được từ mẫu như sau:

Xếp loại	Kém	Trung bình	Khá	Giỏi
Trường A	15	57	34	24
Trường B	13	46	68	33

a) Với mức ý nghĩa 1%, có thể xem như tỉ lệ học sinh giỏi của trường A bằng với trường B hay không?

b) Hãy kiểm định xem trình độ Anh văn đầu vào của học sinh 2 trường có phân bố tỉ lệ như nhau hay không, kết luận với mức ý nghĩa 1%.

<u>Lưu ý:</u> Kiểm định Chi-Bình-Phương được coi là chính xác hơn khi tất cả giá trị trong bảng tần số lý thuyết lớn hơn hay bằng 5. Vì vậy khi trong bảng tần số lý thuyết xuất hiện số nhỏ hơn 5 thì ta nên sắp xếp lại dữ liệu ban đầu cho thích hợp rồi thực hiện lại bài toán.

Ví dụ 14 (dạng bài so sánh tỷ lệ, bài toán 2 mẫu) có thể giải cách giải thứ 2 là áp dụng dạng bài kiểm định tính độc lập.

III.3.2 Bài toán kiểm định dạng phân phối XS của tổng thể :

Chúng ta chỉ xét các bài toán kiểm định sau:

- Kiểm định phân phối Poisson
- Kiểm định phân phối chuẩn
- Kiểm định sự phù hợp (tham khảo)

(trường hợp riêng: kiểm định phân phối đều rời rạc)

Các bước tiến hành chung:

+ B1: Đặt giả thiết kiểm định: H_0: Tổng thể có phân phối F(x)

H_1: Tổng thể không có phân phối F(x).

Tính các đặc trưng mẫu cần thiết ở dạng ước lượng hợp lý cực đại.

+ B2: Tìm miền bác bỏ. $W_\alpha = (\chi^2_\alpha (k-r-1); +\infty)$

k: số hàng (cột) được chia trong bảng dữ liệu mẫu.

r: số tham số chưa biết của phân phối F(x).(chính là số tham số cần ước lượng từ mẫu để sử dụng trong công thức tính các p_i).

+ B3: Tính tiêu chuẩn kiểm định: $\chi^2_{qs} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

ở đây $O_i = n_i$ là tần số từ mẫu thực nghiệm;

E_i là tần số theo lý thuyết nếu giả thiết H_0 đúng.

+ B4: Kết luận. Nếu $\chi^2_{qs} \in W_\alpha$ thì ta bác bỏ giả thiết H_0.

Chấp nhận H_0 trong trường hợp ngược lại

Ví dụ 27: ***Kiểm định phân phối Poisson.***

Một hãng bảo hiểm nghiên cứu về số tai nạn xảy ra trong các gia đình có từ 2 con nhỏ trở lên trong một năm. Dưới đây là một bảng số liệu thống kê mẫu:

Số tai nạn	0	1	2	3	4	≥5
Số gia đình	135	344	257	165	78	21

Với mức ý nghĩa 5%, có thể xem như số vụ tai nạn loại này tuân theo quy luật phân bố Poisson hay không?

Hướng dẫn: Gọi X là số vụ tai nạn trong một năm của các gia đình có từ 2 con nhỏ trở lên.

+ Giả thiết kiểm định H_0: X có phân phối Poisson với $\lambda \overset{ul}{=} \overline{x} = 1,77$

H_1: X không có phân phối Poisson.

Tra bảng Chi-Bình-Phương với k= 6 ; r = 1 tìm được :

$\chi^2_\alpha(k-r-1) = \chi^2_{0,05}(6-1-1) = 9,49$ Miền bb W_α= (9,49; +∞)

x_i	$n_i \equiv O_i$	$p_i = P(X= x_i) = \frac{e^{-\lambda}\lambda^{x_i}}{(x_i)!}$	$n*p_i \equiv E_i$	$\frac{(O_i - E_i)^2}{E_i} \equiv \frac{(n_i - np_i)^2}{np_i}$
0	135	0.1703	170.3	7.3293
1	344	0.3015	301.49	5.9941
2	257	0.2668	266.82	0.3613
3	165	0.1574	157.42	0.3647
4	78	0.0697	69.660	0.9986
5	21	0.0247	24.659	0.5430
	n=1000	Tổng: χ^2_{qs} =		15.59106

Do $\chi^2_{qs} = \sum_i \frac{(O_i - E_i)^2}{E_i} = 15,59106 \in W_\alpha$ nên bác bỏ H_0.

Số tai nạn trong các gia đình không tuân theo phân phối Poisson.

Ví dụ 28: ***(Kiểm định phân phối Chuẩn)***

Khảo sát chiều cao các cây con được chọn ngẫu nhiên từ vườn ươm, người ta có được kết quả sau:

x_i (cm)	5 – 15	15 – 25	25 - 35	35 – 45	45 - 55	55 - 65	65 - 75
n_i	25	67	191	273	202	54	18

Với mức ý nghĩa 1%, có thể coi mẫu trên phù hợp phân phối chuẩn hay không?

Hướng dẫn: Gtkđ H_0: Mẫu phù hợp với phân phối Chuẩn $N(a, \sigma^2)$

Gt đối H_1: Mẫu không phù hợp với phân phối Chuẩn.

Tính các đặc trưng mẫu: n= 830 ; $\overline{x}$ = 39.5663; $\hat{s}$ = 12.3329

$\overline{x}$ là ước lượng hợp lý cực đại cho a $\Rightarrow$ a = 39,5663

còn $\hat{s}^2$ là ước lượng hợp lý cực đại cho $\sigma^2 \Rightarrow \sigma$ = 12,3329

Tra bảng Chi-BP với k= 7 ; r = 2 $\Rightarrow$ $\chi^2_\alpha(k-r-1) = 13,28$

Miền bác bỏ $W_\alpha = (13{,}28; +\infty)$.

Tính tiêu chuẩn kiểm định:

Khoảng $(\alpha; \beta)$	$n_i \equiv O_i$	$p_i = P(\alpha<X<\beta) = \Phi\left(\frac{\beta-a}{\sigma}\right) - \Phi\left(\frac{\alpha-a}{\sigma}\right)$	$n^*p_i \equiv E_i$	$\frac{(O_i-E_i)^2}{E_i} \equiv \frac{(n_i-np_i)^2}{np_i}$
$(-\infty; 15)$	25	$\Phi\left(\frac{15-a}{\sigma}\right) - (-0,5) =$ 0.0231	19.247	1.719326
(15; 25)	67	$\Phi\left(\frac{25-a}{\sigma}\right) - \Phi\left(\frac{15-a}{\sigma}\right) =$ 0.0956	79.343	.....
(25; 35)	191	0.2368	196.556	.....
(35; 45)	273	0.3146	261.156	.....
(45; 55)	202	0.2244	186.224	.....
(55; 65)	54	0.0858	71.213	.....
$(65; +\infty)$	18	$0,5 - \Phi\left(\frac{65-a}{\sigma}\right) =$ 0.0196	16.261	.....
	n= 830	=1	Tổng: $\chi^2_{qs}=$	10.0166

Cách 1: tính $\chi^2_{qs} = \sum_i \frac{(O_i - E_i)^2}{E_i} = 10,0166 \notin W_\alpha$ Chấp nhận H_0.

Cách 2: Do cách điều chỉnh 2 cận về +/- ∞ nên tổng các p_i = 1. Vì vậy ta có 1 công thức tắt để tính nhanh tiêu chuẩn kiểm định:

$$\chi^2_{qs} \equiv \frac{1}{n} \times \left[\sum_i \frac{n_i^2}{p_i} \right] - n = \frac{1}{830} \left[\frac{25^2}{0,0231} + \frac{67^2}{0,0956} + ... + \frac{18^2}{0,0196} \right] - 830 = 10,0166$$

Lưu ý:

** Nếu thay đổi yêu cầu của đề bài là kiểm định mẫu trên có phù hợp phân phối chuẩn đã xác định như N(a=40; σ^2= 150) chẳng hạn, thì ta không phải sử dụng các đặc trưng mẫu để ước lượng cho a và σ nữa, do đó số tham số cần ước lượng r = 0.*

** Phân phối chuẩn là 1 phân phối liên tục, nên cần lưu ý xử lý số liệu thích hợp khi đề cho các gía trị mẫu ở dạng rút gọn rời rạc.*

Ví dụ 29: Khi khảo sát công việc của 200 công nhân, người ta kiểm tra ngẫu nhiên 1000 sản phẩm của mỗi người thì thu được kết quả dưới đây. Với mức ý nghĩa 1%, có thể coi mẫu này phù hợp với phân phối Poisson hay không?

Số phế phẩm trên 1000 sản phẩm	0	1	2	3	4
Số công nhân	109	65	22	3	1

Ví dụ 30:

Người ta đo bán kính của 1 số chi tiết được chọn ngẫu nhiên từ các sản phẩm của 1 máy tiện. Có thể xem như kết quả nhận được dưới đây là phù hợp với phân phối chuẩn hay không, với mức ý nghĩa 5%?

Bán kính chi tiết (cm)	3,2	3,22	3,24	3,26	3,28	3,3	3,32
Số chi tiết tương ứng	20	41	69	78	54	36	12

Ví dụ 31: ***(Kiểm định sự phù hợp*** *– tham khảo* ***)***

Một công ty dược phẩm cho biết lượng thuốc cảm họ bán ra hàng năm thay đổi theo mùa. Lượng thuốc cảm bán ra vào mùa đông chiếm 40%; 30% lượng thuốc bán được vào mùa xuân, còn lại chia đều vào 2 mùa thu và mùa hè.

Để đánh giá xem lượng thuốc năm nay có phân bố theo mùa như mọi năm hay không, người ta khảo sát ngẫu nhiên hồ sơ của 1000 lô thuốc được tiêu thụ trong năm và có số liệu:

Được bán vào mùa Xuân : ..... 282 hộp
..... Hè : ..... 185 hộp
..... Thu : .....159 hộp
..... Đông: ..... 374 hộp

Với mức ý nghĩa 1%, hãy nêu kết luận cho yêu cầu bài toán?

Hướng dẫn:

Giả thiết kđ H_0: lượng thuốc bán trong năm nay phân bố phù hợp với các năm trước.

Giả thiết đối H_1: lượng thuốc bán trong năm nay có phân bố thay đổi so với các năm trước.

Miền bác bỏ $W_\alpha = (\chi^2_{0,01}(4-0-1); +\infty) = (11,34; +\infty)$

x_i	$n_i \equiv O_i$	$p_i = P(X = x_i)$	$n^* p_i \equiv E_i$	$\frac{(O_i - E_i)^2}{E_i} \equiv \frac{(n_i - np_i)^2}{np_i}$
Xuân	**282**	**30%**	**300**	**1,08**
Hạ	**185**	**15%**	**150**	**8,1667**
Thu	**159**	**15%**	**150**	**0,54**
Đông	**374**	**40%**	**400**	**1,69**
	n = 1000		Tổng: $\chi^2_{qs} =$	**11,4767**

Do $\chi^2_{qs} = \sum_i \frac{(O_i - E_i)^2}{E_i} = 11,4767 \in W_\alpha$ nên bác bỏ H_0, chấp nhận H_1.

Ví dụ 32: ***Kiểm định phân phối đều rời rạc*** *(tham khảo).*

Để kiểm tra sự cân đối giữa các mặt của 1 con xúc xắc, người ta tung ngẫu nhiên con xúc xắc đó 120 lần và thống kê được kết quả sau:

Mặt xuất hiện	1	2	3	4	5	6
Số lần xuất hiện	23	19	24	21	18	15

Với mức ý nghĩa 5%, có thể xem con xúc xắc này là cân đối hay không?

Hướng dẫn:

Gọi X là số chấm xuất hiện khi tung con xúc xắc.

Giả thiết kiểm định H_0: X có phân phối đều rời rạc,
hay là con xúc xắc cân đối.

Giả thiết đối H_1: Con xúc xắc không cân đối.

Miền bác bỏ $W_\alpha = (\chi^2_{0,05}(6-0-1); +\infty) = (11,07; +\infty)$

x_i	$n_i \equiv O_i$	$p_i = P(X = x_i)$	$n^*p_i \equiv E_i$	$\frac{(O_i - E_i)^2}{E_i} \equiv \frac{(n_i - np_i)^2}{np_i}$
1	23	1/6	20	0,45
2	19	1/6	20	0,05
3	24	1/6	20	0,8
4	21	1/6	20	0,05
5	18	1/6	20	...0,2
6	15	1/6	20	1,25
	n = 120	Tổng: χ^2_{qs} =		2,8

Do $\chi^2_{qs} = \sum_i \frac{(O_i - E_i)^2}{E_i} = 2,8 \notin W_\alpha$

nên chưa bác bỏ được H_0.

Chưa có cơ sở để nói rằng con xúc xắc không cân đối.

Chương V: HỒI QUY TUYẾN TÍNH ĐƠN

(Giới thiệu sơ lược)

V.1 Các đặc trưng mẫu và hệ số tương quan mẫu:

Một bảng tương quan mẫu 2 chiều (X,Y) kích thước *n* có dạng như sau:

X \ Y	y_1	y_2	….	y_h	n_i
x_1	n_{11}	n_{12}	….	n_{1h}	n_1
x_2	n_{21}	n_{22}	….	n_{2h}	n_2
….	….	….	….	….	….
x_k	n_{k1}	n_{k2}	….	n_{kh}	n_k
m_j	m_1	m_2	….	m_h	$\Sigma = n$

Các đặc trưng mẫu:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i; \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i; \quad \hat{s}_X^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i \equiv \overline{x^2} - (\bar{x})^2$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i \equiv \frac{n}{n-1}\hat{s}_X^2$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{h} y_j m_j; \quad \overline{y^2} = \frac{1}{n}\sum_{j=1}^{h} y_j^2 m_j; \quad \hat{s}_Y^2 = \frac{1}{n}\sum_{i=1}^{k}(y_j - \bar{y})^2 m_i \equiv \overline{y^2} - (\bar{y})^2$$

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 m_j \equiv \frac{n}{n-1}\hat{s}_Y^2$$

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij} x_i y_j$$

Hệ số tương quan mẫu:

$$r_{XY} = \frac{\overline{xy} - \bar{x}.\bar{y}}{\hat{s}_X.\hat{s}_Y}$$

Các bước thực hiện	Máy CASIO fx 570 ES (PLUS)...	Máy CASIO fx 500 MS....
Vào chế độ thống kê hai biến.	MODE -- 3 (STAT) -- 2 (A+BX)	(MODE) – 2 (REG) --- 1 (Lin)
Mở cột tần số (nếu máy chưa mở)	SHIFT -- MODE (SETUP) -- ▼ -- -- 4 (STAT) -- 1 (ON)	
Nhập dữ liệu	X Y FREQ 1 x_1 y_1 n_{11} 2 x_1 y_2 n_{12} ... ... x_k y_h n_{kh} AC	Nhập lần lượt theo từng dòng, thứ tự nhập như sau: X_i , Y_j ; n_{ij} M+
Đọc kết quả $\overline{x}; \overline{y}$	SHIFT – 1 (STAT)- 4 (VAR) – --- 2 ($\overline{x}$) -- = *Tương tự ta chọn* $\overline{y}$	SHIFT – 2 (SVAR) -1 ($\overline{x}$)-- = SHIFT – 2(SVAR) - ▶ -1 ($\overline{y}$)-- =
Đọc kết quả $\hat{s}_x; \hat{s}_y$	SHIFT – 1 (STAT)- 4 (VAR) – --- 3 (σ_X) -- = *Tương tự ta chọn* σ_Y	SHIFT – 2 (SVAR)- 2 ($x\sigma n$)-- = SHIFT – 2(SVAR) - ▶ -1 ($y\sigma n$) -- =
Đọc kết quả $\overline{xy}$	SHIFT – 1 (STAT)- 3 (SUM) – -- 5 ($\sum xy$) -- ÷ -- n -- =	SHIFT – 1 (SSUM) – ÷ --- n --- = $\sum xy$
Đọc kết quả R_{XY}	SHIFT – 1 (STAT)-6(REG)–3 (r) --=	SHIFT – 2 (SVAR) - ▶ - ▶ 3(r)-=

Ví dụ 1: Xét bảng tương quan mẫu 2 chiều (X,Y) thu được khi người ta sơ chế một loại nông sản, ở đây X (đơn vị: phút) biểu diễn thời gian chế biến, và Y (đơn vị: %) thể hiện mức suy giảm lượng đường trong sản phẩm. Hãy tính các đặc trưng mẫu và hệ số tương quan mẫu.

X	Y				
	30	35	40	45	50
2	4				
4		7	3		
6		1	16	4	
8			2	10	3
10				4	6

Ví dụ 2: Theo dõi kết quả thực hành của sinh viên, người ta có được số liệu mẫu sau đây. Tìm các đặc trưng mẫu và tìm hệ số tương quan của X,Y.

Thời gian thí nghiệm *(phút)*	3	4	3	5	6	4	6	5	7	8
Khối lượng sản phẩm tạo thành *(gram)*	7	8	7.3	9	10.5	8.2	10.8	9.5	11	12

$VD1:\quad n=60 \quad \bar{x}=6,5667 \quad \hat{s}_x=2,2536 \quad s_X=2,2727$

$\overline{xy}=284,5 \quad \bar{y}=41,6667 \quad \hat{s}_y=5,4518 \quad s_Y=5,4978 \quad r_{XY}=0,8863$

$VD2:\quad n=10 \quad \bar{x}=5,1 \quad \hat{s}_x=1,5780 \quad s_X=1,6633$

$\overline{xy}=50,1 \quad \bar{y}=9,33 \quad \hat{s}_y=1,6181 \quad s_Y=1,7056 \quad r_{XY}=0,9858$

Hướng dẫn nhập dữ liệu:

X	Y	Freq
2	30	4
4	35	7
4	40	3
6	35	1
6	40	16
6	45	4
8	40	2
8	45	10
8	50	3
10	45	4
10	50	6

Ví dụ 1

X	Y	Freq
3	7	1
4	8	1
3	7.3	1
5	9	1
6	10.5	1
4	8.2	1
6	10.8	1
5	9.5	1
7	11	1
8	12	1

Ví dụ 2

V.2 LÝ THUYẾT HỒI QUY:

Lý thuyết hồi quy (đơn biến) nghiên cứu bài toán dự báo biến ngẫu nhiên Y trên cơ sở đã biết về biến ngẫu nhiên X. Biến X được gọi là biến độc lập, hay gọi là biến giải thích. Y gọi là biến phụ thuộc, hay biến được giải thích. Người ta tìm cách thay Y bởi hàm f(X) sao cho "chính xác nhất".

Trong mối liên hệ hàm số, với mỗi một giá trị X ta tìm được duy nhất một giá trị Y. Tuy nhiên trong thống kê, một giá trị X có thể cho tương ứng nhiều giá trị Y khác nhau, bởi vì ngoài biến chính là X, biến Y có thể còn chịu tác động bởi một số yếu tố khác.

Định nghĩa hồi quy: *Hàm hồi quy của Y theo X chính là kỳ vọng có điều kiện của Y đối với X, tức là E(Y|X).* Nếu đồ thị hàm hồi quy là đường thẳng thì ta nói *hồi quy tuyến tính*. Phần sau đây chỉ đề cập đến đường hồi quy mẫu và đường hồi quy tuyến tính mẫu.

V.2.1 Đường hồi quy mẫu:

Xét bảng tương quan mẫu 2 chiều. Với mỗi *i =1,2,..,k*; ta đặt:

$$\overline{y}_{|X=x_i} = E\left(Y \mid X = x_i\right) = \frac{1}{n_i}\sum_{j=1}^{h} n_{ij} y_j$$

Nối các điểm có tọa độ $(X_i; \overline{y}_{|X=x_i})$ theo thứ tự của i bởi các đoạn thẳng, ta được một đường gấp khúc, gọi là *đường hồi quy mẫu của Y theo X.*

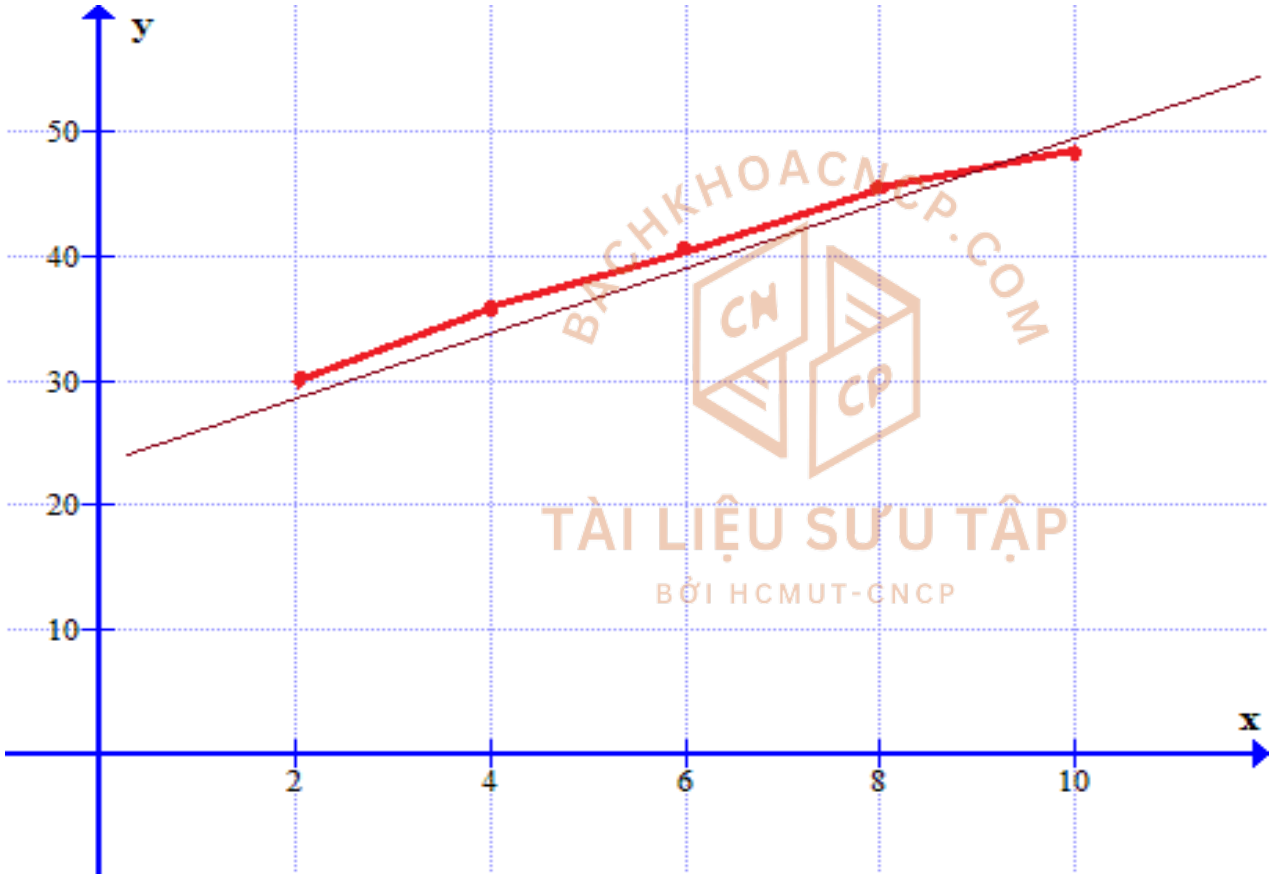
Ví dụ 3: *Sử dụng bảng tương quan trong ví dụ 1 để vẽ đường hồi quy mẫu Y theo X:*

$$\overline{y}_{|X=2} = 30$$

$$\overline{y}_{|X=4} = \frac{35*7+40*3}{10} = 36,5 \qquad \overline{y}_{|X=8} = \frac{40*2+45*10+50*3}{15} = 45,3333$$

$$\overline{y}_{|X=6} = \frac{35*1+40*16+45*4}{21} = 40,7143 \qquad \overline{y}_{|X=10} = \frac{45*4+50*6}{10} = 48$$

Đường hồi quy mẫu của Y theo X (đường gấp khúc):



V.2.2 Đường hồi quy tuyến tính mẫu:

Đường hồi quy tuyến tính mẫu của Y theo X là đường thẳng có dạng y = A + Bx, xấp xỉ *"gần nhất"* với đường hồi quy mẫu.

Người ta dùng *phương pháp bình phương bé nhất* để tìm các hệ số A, B này, tức là (A,B) làm cho hàm Q đạt giá trị nhỏ nhất.

$$Q(\mathrm{A},\ B) = \sum_{i=1}^{k} n_i \left[(A + Bx_i) - \overline{y}_{|X=x_i} \right]^2$$

Giải bài toán cực trị tự do, ta tìm được:

$$\begin{cases} B = \dfrac{\overline{xy} - \overline{x}.\overline{y}}{\hat{s}_x^{\,2}} \\ A = \overline{y} - B.\overline{x} \end{cases}$$

Giá trị Y được dự báo qua đường hồi quy tuyến tính mẫu thường được viết ở dạng $\mathrm{Y} = A + B.\mathrm{X}$. (Có thể sử dụng MTBT tìm A,B, Y)

Ví dụ 4: *Sử dụng bảng tương quan của (X,Y) trong ví dụ 1.*

a) Tìm phương trình đường hồi quy tuyến tính của Y theo X. Từ đó dự đoán mức suy giảm lượng đường trong sản phẩm khi thời gian sơ chế là 9 phút; 11 phút?

b) Tìm phương trình đường hồi quy tuyến tính của X theo Y.

a)
$$\begin{cases} B = \dfrac{\overline{xy} - \overline{x}.\overline{y}}{\hat{s}_x^{\,2}} = \dfrac{284,5 - 6,5667 \times 41,6667}{2,2536^2} = 2,1440 \\ A = \overline{y} - B.\overline{x} = 41,6667 - 2,1440 \times 6,6567 = 27,5881 \end{cases}$$

Đường hồi quy tuyến tính mẫu Y theo X: y = 27,5881 + 2,1440x.

Dự đoán: $Y(X = 9) = 46,8836(\%)$ $Y(X = 11) = 51,1715(\%)$

b) Phương trình hồi quy tuyến tính X theo Y có dạng x = C+Dy, với:
$$\begin{cases} D = \dfrac{\overline{xy} - \overline{x}.\overline{y}}{\hat{s}_y^{\,2}} = \dfrac{284,5 - 6,5667 \times 41,6667}{5,4518^2} = 0,3664 \\ C = \overline{x} - D.\overline{y} = 6,6567 - 0,3664 \times 41,6667 = -8,6981 \end{cases}$$