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 từ thực tiễn, dựa trên những thành tựu của lý thuyết XS.

*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 5: LÝ THUYẾT MẪU

Chương 6: LÝ THUYẾT ƯỚC LƯỢNG

Chương 7: * KIỂM ĐỊNH GIẢ THIẾT THỐNG KÊ

* PHÂN TÍCH PHƯƠNG SAI

Chương 8: LÝ THUYẾT HỒI QUY ĐƠN

Chương 5: LÝ THUYẾT MẪU

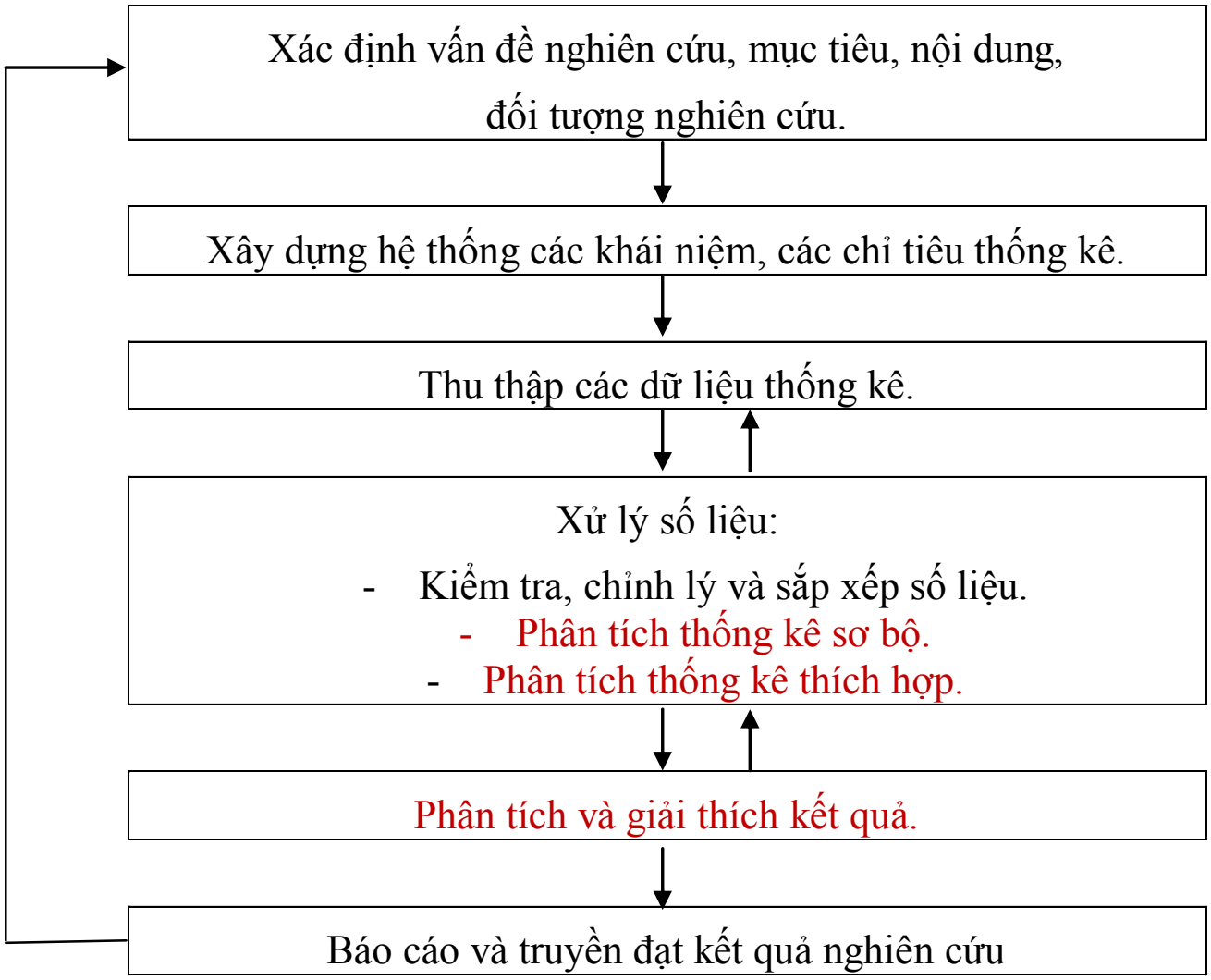
I.1. Một số khái niệm:

- **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 (Ch3-giáo trình TKƯD)

- 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 Bernoulli (hay là pp 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: $f=\frac{m}{n}$

Ví dụ 1: Phòng CTSV khảo sát về chi phí trong 1 học kỳ của SV K18 dành cho các hoạt động ngoại khóa. Có 50 sinh viên được lựa chọn ngẫu nhiên để trả lời bảng khảo sát.

Gọi X là chi phí của 1 sinh viên K18 cho hoạt động ngoại khóa.

Giả sử những SV có chi phí cho HĐNK trên 1 triệu/ 1 hk được gọi là có chi phí HĐNK cao.

Các đặc trưng của ...	Tổng thể	Mẫu
Kích thước	(N)=	$n = 50$
Trung bình	$E(X) = a$	$\overline{x}$
Phương sai	$D(X) = \sigma^2$	$\widehat{s}^2$ $(xσn \mid σx)^2$ (PS mẫu)
		s^2 $(xσn-1 \mid sx)^2$ (PS mẫu hiệu chỉnh)
Tỉ lệ	p	f

⇩

Hằng số

Nói chung chưa biết

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: $x_1 \leq x_2 \leq .... \leq x_{n-1} \leq x_n$.

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 Outlier: 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 ($Q1 - 1{,}5 \times IQR$; $Q3 + 1{,}5 \times IQR$) .

II.2.7 Vẽ biểu đồ hộp và râu:

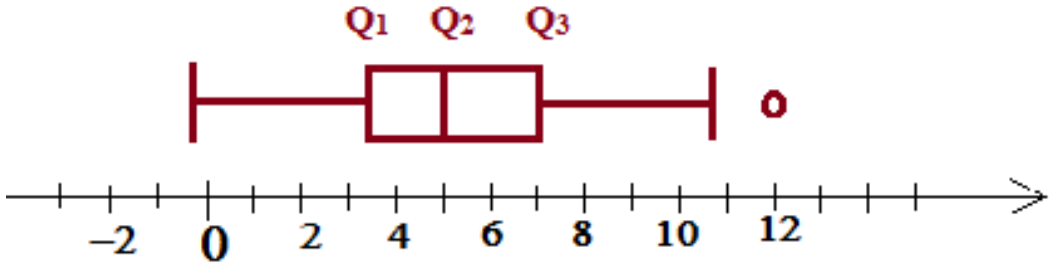
Xét mẫu có kích thước n = 9 đã được sắp theo thứ tự tăng dần:

1 3 4 4 5 5 6 8 12

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

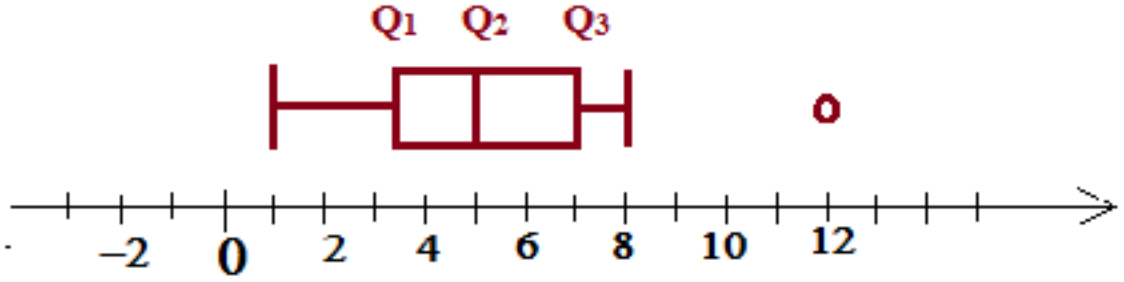
Khoảng trải giữa IQR = $Q_3 - Q_1 = 7 - 3,5 = 2,5$

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

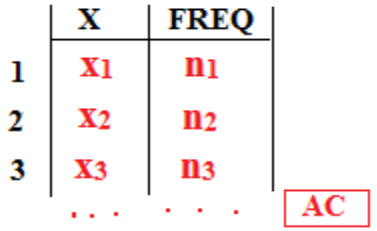


Có 1 giá trị outlier là 12

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 580 vnx….
Vào TK 1 biến.	MODE -- 3 (STAT) -- 1 (1-VAR)	MENU – 6 - 1
Mở cột tần số (nếu chưa có)	SHIFT -- MODE (SETUP) – -- ▼ -- ---4 (STAT) -- 1 (ON)	SHIFT -- MODE– -- ▼ ---3 -- 1
Nhập dữ liệu		
Đọc kết quả n	SHIFT – 1 (STAT)- 4 (VAR) -- 1 (n) -- =	OPTIONS - ▼ - 2
Đọc kq $\bar{x}$	SHIFT – 1 (STAT)- 4 (VAR) -- 2 ($\bar{x}$) -- =	
Đọc kq $\hat{s}$	SHIFT – 1 (STAT)- 4 (VAR) - 3 (σx) -=	
Đọc kq s	SHIFT – 1 (STAT)- 4 (VAR) - 4 (sx) -=	
Kq trung gian $\sum_{i=1}^{k} x_i n_i \equiv \sum_{i=1}^{n} x_i$	SHIFT – 1 (STAT)- 3 (SUM) –2 ($\sum x$) =	
$\sum_{i=1}^{k} x_i^2 n_i \equiv \sum_{i=1}^{n} x_i^2$	SHIFT – 1 (STAT)- 3 (SUM) –1 ($\sum x^2$)=	

Ví dụ 2: 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:

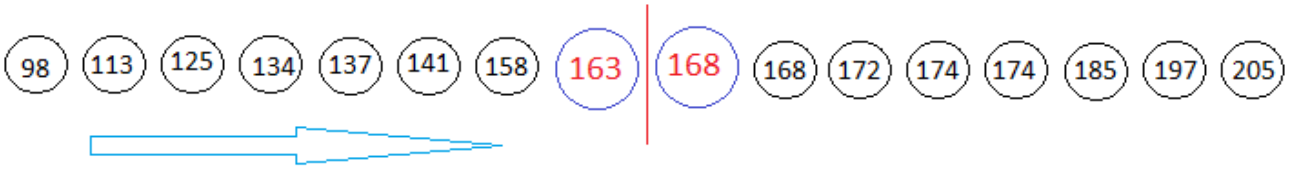
a) Trung bình mẫu (TB cộng), trung vị mẫu, mode và CV.

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

X	Freq
125	1
205	1
134	1
---	---
141	1
	AC

$a)\ n=16; \qquad \bar{x}=157 \qquad cv=19,03\%$

$\text{mod}=\{168;174\} \qquad med=\dfrac{163+168}{2}=165,5;$



$b)\ \hat{s}=28,9353\ ; s=29,8842$

Ví dụ 3:

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.

X	Freq
16	11
18	32
---	---
24	23
26.5	22
	AC

$a)\ n = 174; \qquad \bar{x} = 21,0977$

$\hat{s} = 2,9555; \qquad s = 2,9640$

$b)\ f = \dfrac{11+32}{174} = \dfrac{43}{174}$

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:

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

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

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

- **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

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

Chương 6: 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:

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

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

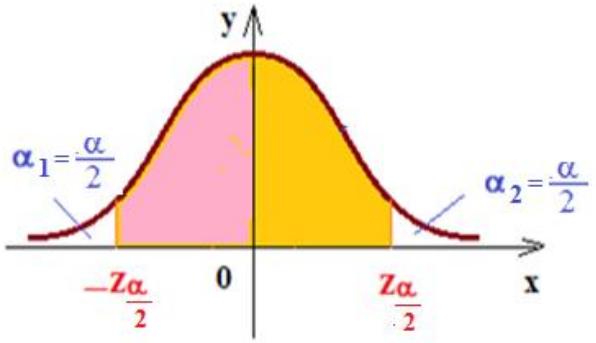
Phương pháp tìm khoảng tin cậy cho tham số θ với độ tin cậy *1-*α cho trước:

- 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.*

Nếu ta đặt $Z = \dfrac{F-p}{\sqrt{f(1-f)}}\sqrt{n}$ thì Z ~ N(0;1) (Ch 5, mục II.3, slide 25)

Chọn $\alpha_1 = \alpha_2 = \alpha/2$.



Nếu lấy $z_{\alpha/2}$ thỏa $F(z_{\frac{\alpha}{2}}) = 1 - \dfrac{\alpha}{2}$

thì $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$

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

$$\Leftrightarrow P\left(-\frac{z_{\frac{\alpha}{2}}\cdot\sqrt{f(1-f)}}{\sqrt{n}} < F-p < \frac{z_{\frac{\alpha}{2}}\cdot\sqrt{f(1-f)}}{\sqrt{n}}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(F-\varepsilon < p < F+\varepsilon\right) = 1-\alpha \quad \text{ở đây} \quad \varepsilon = \frac{z_{\alpha/2}\sqrt{f(1-f)}}{\sqrt{n}}$$

Ta gọi ε là *ngưỡng 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) $\Leftrightarrow P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1-\alpha \Leftrightarrow P(Z < z_{\frac{\alpha}{2}}) = 1-\frac{\alpha}{2}$

dẫn đến $F(z_{\frac{\alpha}{2}}) = 1-\frac{\alpha}{2}$ .Tìm $z_{\alpha/2}$ bằng cách tra ngược bảng Hàm phân phối chuẩn tắc, sau đó sẽ tìm được công thức $\varepsilon = \frac{z_{\alpha/2}\cdot\sqrt{f(1-f)}}{\sqrt{n}}$.

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 trong bài 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 $\left(0; F + \frac{z_\alpha \sqrt{f(1-f)}}{\sqrt{n}}\right)$ với $\Phi(z_\alpha) = P(Z \le z_\alpha) = 1-\alpha$;

Người ta nói $F + \frac{z_\alpha \sqrt{f(1-f)}}{\sqrt{n}}$ là ***ước lượng giá trị tối đa*** của p.

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

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.*

Từ Ch5, mục II.3, slide 26, khi n <30 thì: $Q = \dfrac{\overline{X} - a}{s}\sqrt{n} \sim T(n-1)$

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.

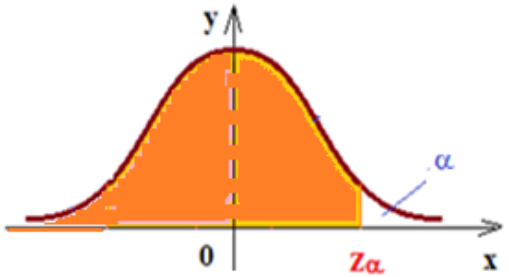
(Nhắc lại: Khi n ≥ 30*, phân phối Student xấp xỉ phân phối Chuẩn tắc.)*

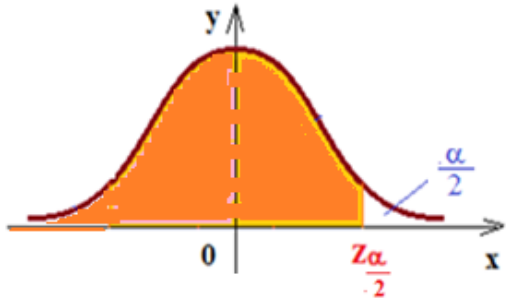
Bảng 2: ***Một số bài toán ước lượng khoảng thông dụng***

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

* Mức phân vị α (trên) là giá trị z_α mà $P(Z \le z_\alpha) = 1-\alpha$

* Mức phân vị $\alpha/2$ là giá trị $z_{\alpha/2}$ mà $P(Z \le z_{\alpha/2}) = 1-\alpha/2$





1) Tìm giá trị $z_{\alpha/2}$ *thỏa* $F(z_{\alpha/2}) = 1-\alpha/2$:

Cách 1: Tra ngược bảng Hàm phân phối chuẩn tắc.

Cách 2: Bấm mò qua phím chức năng P(?) = $1-\alpha/2$

Cách 3: Sử dụng chức năng hàm ngược của hàm chuẩn

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

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

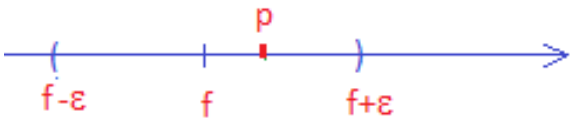
$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

TABLE • III Cumulative Standard Normal Distribution (*Continued*)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.532922	0.527903	0.531881	0.535856
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989

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 .

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 $\Rightarrow$ F($z_{\alpha/2}$) = 1-α/2 = 0,99

$\Rightarrow z_{\alpha/2} = 2,33.$

Tìm ngưỡng sai số *(hay là độ chính xác)* của ƯL:

$$\varepsilon = \frac{z_{\alpha/2}\sqrt{f(1-f)}}{\sqrt{n}} = \frac{2,33\times\sqrt{0,86\times0,14}}{\sqrt{1000}} \approx 0,0256$$

$\Rightarrow$ KƯL cho p: (f-ε; f+ε) = (0,8344; 0,8856)

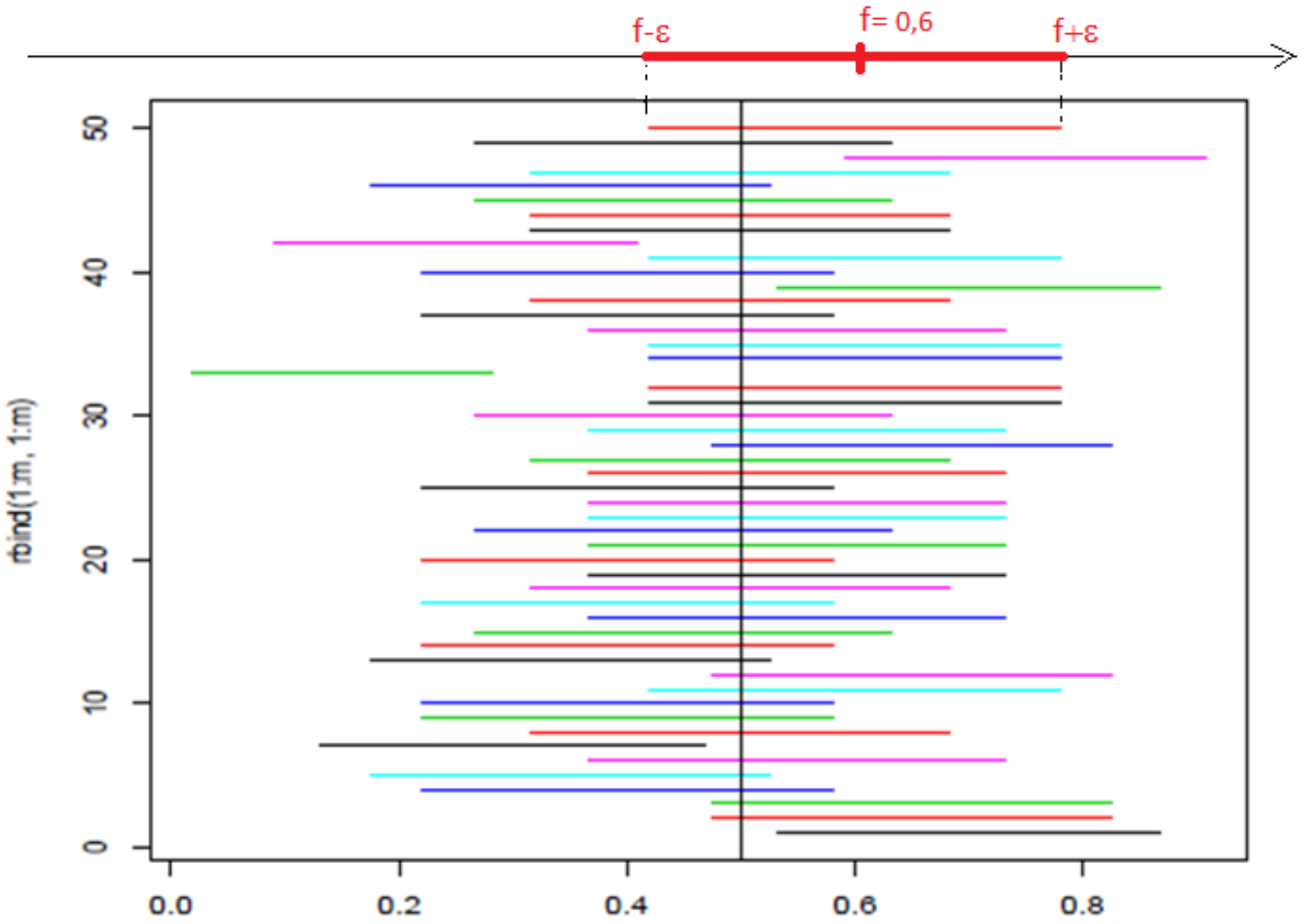
Lưu ý: *Vì p là 1 số , không phải BNN nên chỉ xảy ra 1 trong 2 khả năng:*

- *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?

Hướng dẫn:

Gọi p là tỉ lệ cử tri ủng hộ ƯCV A trong toàn bang .

Tính các đặc trưng mẫu: n = 1600; f= 960/1600 = 0,6.

Đtc 0,97$\Rightarrow \alpha$= 0,03 $\Rightarrow$ F($z_{\alpha/2}$) = 1-α/2 = 0,985 $\Rightarrow$ $z_{\alpha/2}$ = 2,17.

Tìm ngưỡng sai số của ƯL:

$$\varepsilon = \frac{z_{\alpha/2}\sqrt{f(1-f)}}{\sqrt{n}} = \frac{2,17\times\sqrt{0,6\times 0,4}}{\sqrt{1600}} \approx 0,0266$$

$\Rightarrow$ KƯL cho p: (f-ε; f+ε) = (0,5734; 0,6266)

KƯL cho số phiếu cần tìm: (0,5734$\times$ 4.10^6 ; 0,6266$\times$ 4.10^6)

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.

Hướng dẫn:

1- α	z_α
99%	2,58
98%	2,33
---	---
95%	1,96

$1-\alpha$

$$\varepsilon = \frac{z_\alpha \sqrt{f(1-f)}}{\sqrt{n}}$$

$Z_\alpha =$ ⇨ 1- α

n = (làm tròn lên)

Hướng dẫn:

Đtc 0,99 $\Rightarrow$ $F(z_{\alpha/2}) = 1-\alpha/2 = 0,995 \Rightarrow z_{\alpha/2} = F^{-1}(0,995) = 2,58.$

$$\varepsilon = \frac{z_{\frac{\alpha}{2}}\sqrt{f(1-f)}}{\sqrt{n}} \quad \Rightarrow n = \left(\frac{z_{\frac{\alpha}{2}}\sqrt{f(1-f)}}{\varepsilon}\right)^2$$ + làm tròn lên

a) $$\varepsilon < 0,03 \Rightarrow n \geq \left(\frac{2,58\times\sqrt{0,2\times 0,8}}{0,03}\right)^2 = 1183,36 \quad \Rightarrow n = 1184$$

b) Cần tìm n để $$n \geq \left(\frac{2,58\times\sqrt{f\times(1-f)}}{0,03}\right)^2, \forall f \in (0;1)$$

$$\Rightarrow n \geq \underset{f\in(0;1)}{GTLN}\left(\frac{2,58\times\sqrt{f\times(1-f)}}{0,03}\right)^2 = \left(\frac{2,58\times\sqrt{\frac{1}{2}\times(1-\frac{1}{2})}}{0,03}\right)^2 = 1849$$

Hãy nêu nhận xét !

- Nhận xét về ưu điểm và nhược điểm của 2 cách lấy mẫu trên?

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%.

Hướng dẫn:

n = 400; f = 60/400 = 0,15 là tỉ lệ cá được đánh dấu trong mẫu.

Gọi p là tỉ lệ cá được đánh dấu trong hồ và N là số cá trong hồ.

Từ giả thiết suy ra p = 300/N.

Trước tiên tìm KƯL cho p ở dạng: $(f - \varepsilon; f + \varepsilon)$

Suy ra $f - \varepsilon < 300/N < f + \varepsilon \Rightarrow$ Khoảng ƯL cho N.

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.

Hướng dẫn:

a) Kí hiệu a là đường kính trung bình trục máy

KƯL đường kính trung bình trục máy có dạng: $(\overline{x}-\varepsilon;\ \overline{x}+\varepsilon)$

$n=24;\ \ \overline{x}=25,1083;\ \ s=1,5036$

Do n < 30 + Đk trục máy có phân phối chuẩn + chưa biết σ^2 nên ta sử dụng bảng tra Student.

$t_{\alpha/2}(n-1) = t_{0,02/2}(24-1) = 2,5$

Tính ngưỡng sai số : $\varepsilon = t_{\frac{\alpha}{2}(n-1)} \times \frac{s}{\sqrt{n}} = 2,5 \times \frac{1,5036}{\sqrt{24}} = 0,7673$

KƯL đường kính trung bình trục máy : (25,1083 $\pm$ 0,7673)

b) Kí hiệu σ^2 là phương sai của đường kính trục máy.

KƯL σ^2 có dạng:

$$\left(\frac{(n-1)\times s^2}{\chi^2_{\frac{\alpha}{2};(n-1)}}; \frac{(n-1)\times s^2}{\chi^2_{1-\frac{\alpha}{2};(n-1)}} \right) = \left(\frac{23\times 1,5036^2}{41,64}; \frac{23\times 1,5036^2}{10,20} \right)$$

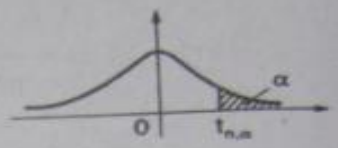
Tra bảng Chi Bình phương:

$$\chi^2_{\frac{\alpha}{2};(n-1)} = \chi^2_{0,01;(23)} = 41,64; \qquad \chi^2_{1-\frac{\alpha}{2};(n-1)} = \chi^2_{0,99;(23)} = 10,20;$$

Bảng VII

Các trị số $t_{n\alpha}$ được xác định từ đại lượng ngẫu nhiên t_n có phân phối Student với n bậc tự do theo công thức

$$P(t_n > t_{n\alpha}) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}} \int_{t_{n,\alpha}}^{+\infty} \left(1 + \frac{t^2}{2}\right)^{-\frac{n+1}{2}} dt = \alpha$$



n \ α	,25	,1	,05	,025	,01	,005
1	4,000	3,078	6,314	12,706	31,821	63,657
2	,816	1,886	2,920	4,303	6,965	9,925
3	,765	1,638	2,353	3,182	4,541	5,841
4	,741	1,533	2,132	2,776	3,747	4,604
5	,727	1,476	2,015	2,571	3,365	4,032
6	,718	1,440	1,943	2,447	3,143	3,707
7	,711	1,415	1,895	2,365	2,998	3,499
8	,706	1,397	1,860	2,306	2,896	3,355
9	,703	1,383	1,833	2,262	2,821	3,250
10	,700	1,372	1,812	2,228	2,764	3,169
11	,697	1,363	1,796	2,201	2,718	3,106
12	,695	1,356	1,782	2,179	2,681	3,055
13	,694	1,350	1,771	2,160	2,650	3,012
14	,692	1,345	1,761	2,145	2,624	2,977
15	,691	1,341	1,753	2,131	2,602	2,947
16	,690	1,337	1,746	2,120	2,583	2,921
17	,689	1,333	1,740	2,110	2,567	2,898
18	,688	1,330	1,734	2,101	2,552	2,878
19	,688	1,328	1,729	2,093	2,539	2,861
20	,687	1,325	1,725	2,086	2,528	2,845
21	,686	1,323	1,721	2,080	2,518	2,831
22	,686	1,321	1,717	2,074	2,508	2,819
23	,685	1,319	1,714	2,069	2,500	2,807
24	,685	1,318	1,711	2,064	2,492	2,797
25	,684	1,316	1,708	2,060	2,485	2,787

Bảng VI

Các trị số $\chi^2_{n,\alpha}$ được xác định từ phân phối χ^2 với n bậc tự do

$$P(\chi^2 > \chi^2_{n,\alpha}) = \frac{1}{\Gamma\left(\frac{n}{2}\right)2^{n/2}} \int_{\chi^2_{n,\alpha}}^{+\infty} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} dx = \alpha$$

n	α									
	0,995	0,990	0,975	0,950	0,900	0,10	0,05	0,025	0,010	0,005
1	0,000039	0,00016	0,00098	0,0039	0,0158	2,71	3,84	5,02	6,63	7,88
2	0,0100	0,0201	0,0506	0,103	0,211	4,61	5,99	7,38	9,21	1,60
3	0,0717	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,34	12,84
4	0,207	0,297	0,484	0,711	1,06	7,78	9,49	11,14	13,28	14,86
5	0,412	0,554	0,831	1,15	1,61	9,24	11,07	12,83	15,09	16,75
6	0,676	0,872	1,24	1,64	2,20	10,64	12,59	14,45	16,81	18,55
7	0,989	1,24	1,69	2,17	2,83	12,02	14,07	16,04	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	21,96
9	1,73	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	25,19
11	2,60	3,05	3,82	4,57	5,58	1728	19,68	21,92	24,73	26,76
12	3,07	3,57	4,40	5,23	6,30	18,55	21,03	23,34	2622	28,30
13	3,57	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	29,82
14	4,07	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14	31,32
15	4,60	5,23	6,26	7,27	8,55	22,31	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	9,31	23,54	2630	28,85	32,00	34,27
17	5,70	6,41	7,56	8,67	10,08	24,77	27,59	30,19	33,41	35,72
18	6,26	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,81	37,16
19	6,84	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19	38,58
20	7,43	8,25	9,59	10,85	12,44	28,41	31,41	34,17	37,57	40,00
21	8,03	8,90	10,28	11,59	13,24	29,62	32,67	35,48	38,93	41,40
22	8,64	9,54	10,98	12,34	14,04	30,81	33,92	36,78	40,29	42,80
23	9,26	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64	44,18
24	9,89	10,86	12,40	13,85	15,66	33,20	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31	46,93

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) KƯL cho gía bán trung bình (a) của mặt hàng này có dạng

$(\bar{x}-\varepsilon;\ \bar{x}+\varepsilon).$ $\quad n=100;\ \bar{x}=86,76;\ s=1,8969$

Ngưỡng sai số của ước lượng: $\varepsilon=\dfrac{z_{\alpha/2}\times s}{\sqrt{n}}=\dfrac{2,17\times 1,8969}{\sqrt{100}}=0,4116$

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

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

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

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.

c) Tỉ lệ cửa hàng bán thấp hơn giá của công ti trong mẫu: $f = 0,7$

Gọi p là tỉ lệ cửa hàng bán thấp hơn giá của công ti trong vùng.

KUL cho p là (f-ε; f+ε); với $\varepsilon = \frac{z_{\alpha/2} \times \sqrt{f(1-f)}}{\sqrt{n}} = \frac{2,33 \times \sqrt{0,7 \times 0,3}}{\sqrt{100}} = 0,1068$

Nhân 2 vế của KUL p với 8000, ta suy ra KUL cần tìm cho số cửa hàng.

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

BT : 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	25-27
Số chi tiết máy tương ứng	11	18	30	32	23	16

Những chi tiết được gia công dưới 19 phút gọi là chi tiết được gia công nhanh. GT thời gian gia công các chi tiết tuân theo pp Chuẩn.

a) Với độ tin cậy 99%, hãy tìm các khoảng ước lượng đối xứng cho:

a1) Thời gian gia công trung bình 1 chi tiết.

a2) Tỉ lệ chi tiết được gia công nhanh.

a3) Thời gian gia công trung bình của các chi tiết được g.công nhanh.

b) Nếu muốn khoảng tin cậy 99% cho thời gian gia công trung bình 1 chi tiết có chiều dài là 1 phút thì cần phải khảo sát bao nhiêu chi tiết?

c) Tìm khoảng tin cậy 95% cho số các chi tiết được gia công nhanh trong 1500 chi tiết của phân xưởng.

d) Tìm KUL 95% cho tổng thời gian cần có để gia công 1500 chi tiết.