TRƯỜNG ĐẠI HỌC ĐÀ LẠT
KHOA SAU ĐẠI HỌC

GIÁO TRÌNH

XỬ LÝ SỐ LIỆU THỰC NGHIỆM

TS. MAI XUÂN TRUNG

2013

MỤC LỤC

CHƯƠNG I : 5
ĐÁNH GIÁ CÁC SAI SỐ TRONG THỰC NGHIỆM 5
A. Các biểu diễn số liệu 5
1. Sai số tuyệt đối: 5
2. Sai số tương đối 5
3. Các chữ số có nghĩa 5
B. Phân loại sai số 7
1. Sai số hệ thống: 8
2. Sai số ngẫu nhiên: 8
3. Sai số thô: 9
C. Các khái niệm cơ sở và mối liên hệ với thống kê toán học 9
1. Trung bình số học: 9
2. Trung bình trọng số 10
3. Phương sai. 12
4. Cộng phương sai và tương quan 16
5. Công thức truyền sai số 18
6. Ước lượng khoảng 26
CHƯƠNG II: 43
CÁC PHÂN BỐ THƯỜNG DÙNG TRONG XỬ LÝ SỐ LIỆU 43
1. Tính qui luật xác suất 43
2. Phương sai nội và phương sai ngoại 47
3. Hàm phân bố khi bình phương χ^2 48
4. Hàm phân bố nhị thức 56
5. Hàm phân bố Poisson 60
6. Hàm phân bố Gauss 66
7. Hàm phân bố Log-Normal 72
CHƯƠNG III: 76
PHÂN TÍCH CÁC QUY LUẬT THỰC NGHIỆM DỰA TRÊN CÁC CƠ SỞ HỒI QUY TUYẾN TÍNH 76
I. Mô hình đường thẳng 77
I.1 Mô hình đường thẳng tổng quát........
80
a. Xác định các tham số b_0 *và* b_1 78
b. Các tính chất của đường thẳng hồi qui 80
c. Xác định sai số của các tham số 81
d. Xác định phương sai σ^2 82
e. Cộng phương sai của hệ số b_0, b_1 84
f. Xác định phương sai của một giá trị nội suy 85
g. Khoảng tin cậy cho các giá trị 86
h. Hệ số tương quan 88
i. Phân tích các đại lượng phương sai 89
I.2 Mô hình đường thẳng xuyên qua gốc
94
II. Mô hình tổng quát 92
1. Bài toán bình phương tối thiểu 92
2. Mô hình hồi qui tổng quát ở các số hạng ma trận 94

3. Mô hình hồi qui chuẩn hóa
116
3. Mô hình các đa thức trực giao 117
4. Bình phương tối thiểu có trọng số 121
CHƯƠNG IV: 128
PHƯƠNG PHÁP BÌNH PHƯƠNG TỐI THIỂU PHI TUYẾN 128
1. Vectơ Gradient và ma trận Hessian 129
2. Khai triển chuỗi Taylor 131
3. Các điều kiện tối ưu không ràng buộc 132
4. Thuật toán tối ưu tổng quát 133
5. Đảm bảo sự giảm dốc 134
6. Đảm bảo sự hội tụ : Các phương pháp tìm kiếm đường 139
7. Khớp dữ liệu bình phương tối thiểu phi tuyến. 141
8. Phương pháp Gauss-Newton 143
9. Phương pháp Levenberg-Marquardt 147

Vật lý học là một khoa học quan sát, tìm hiểu những cái gì xảy ra, định tính về chúng, đo lường chúng để tìm ra các đại lượng quan trọng. Đầu tiên các mô tả vật lý cần phải định lượng. Thí dụ một vật chuyển động nhanh hay chậm cần xác định được tốc độ của nó hay một vật nặng hay nhẹ cần xác định được khối lượng của nó. Thứ đến các đại lượng xác định cần được biết với ý nghĩa chính xác nhất định và cuối cùng là phân tích các giá trị thu được liên hệ đến các qui luật logic mà nó tuân theo để luận ra các tham số của mô hình mà nó tuân theo thông qua các phương pháp toán học hồi qui. Nội dung môn học đươc chia thành 4 chương:

Chương 1: Đánh giá các sai số trong thực nghiệm

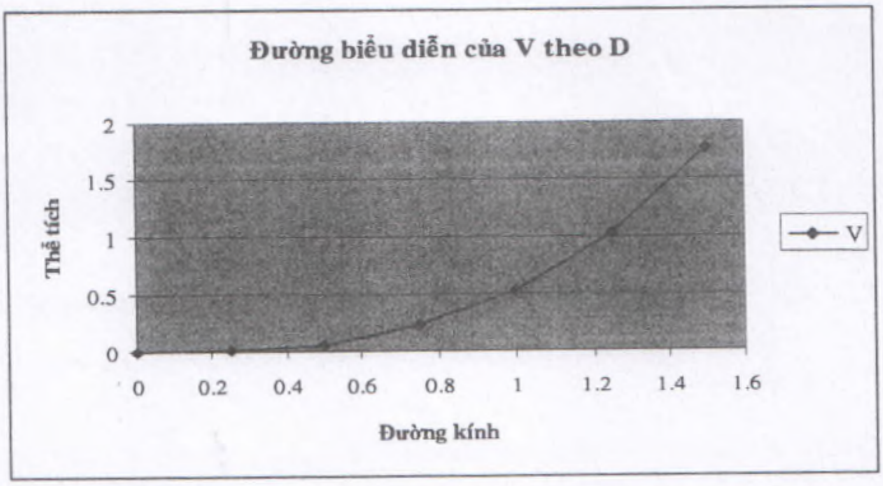
Chương 2: Các phân bố thương dùng trong xử lý sô liệu

Chương 3: Phân tích các qui luật thực nghiệm dựa trên cơ sở hồi qui tuyến tính

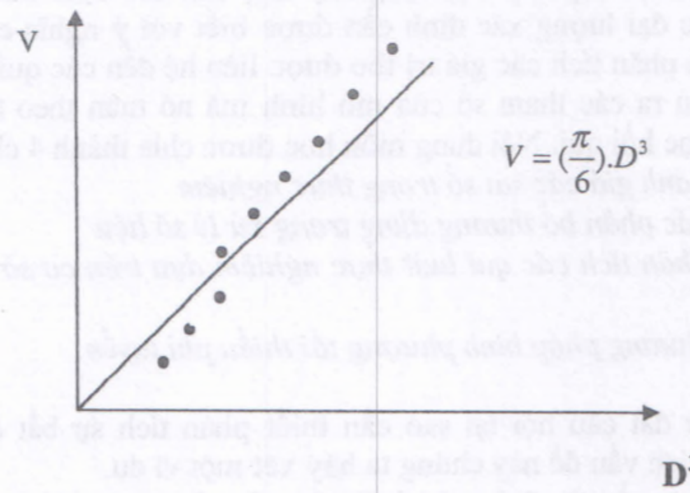
Chương 4: Phương pháp bình phương tối thiểu phi tuyến

Chúng ta thử đặt câu hỏi tại sao cần thiết phân tích sự bất định trong thí nghiệm và để giải thích vấn đề này chúng ta hãy xét một ví dụ.

Giả sử chúng ta cần đo đường kính D của các giọt nước rơi và cân chúng để tìm các thể tích V tương ứng. Chúng ta thấy sai số của thể tích giọt nước theo sai số đường kính của nó là $\sigma_V = \frac{\pi D^2}{2}\sigma_D$ và $\sigma_V = \sigma_D$ chỉ khi D = 0,79788 đơn vị đo. Sai số thể tích của giọt nước sẽ tăng dần mặc dù nếu sai số đường kính của nó giử không đổi. Nếu mô hình các giọt nước là hình cầu thì các điểm thực nghiệm đo được không nằm trên đường cong lý thuyết bởi ví các đo lường tùy thuộc vào các sai số rút ra từ sự xét đoán



và các giới hạn của dung cụ đo. Nếu chúng ta vẽ thể tích V theo D^3, các điểm đo với D lớn hơn nó tuân theo qui luật D^3 nhưng với độ dốc khác ,các điểm thấp hơn cũng rời khỏi đường thẳng nầy phản ánh sự đo lường khó khăn hơn và khó chính xác hơn. Qua đó chúng ta có thể rút ra một số kết luận từ phép đo là " Các dữ liệu biểu thị một đường thẳng trên một vùng nào đó của số liệu đo hay trên toàn bộ tầm của nó ở trong một khoảng bất định của các điểm dữ liệu riêng lẻ. " . Chúng ta thấy rằng sự phân tích sai số (sự bất định) là cần thiết để thiết lập các biên giá trị của các kết luận mà chúng ta rút ra từ các thí nghiệm. Rõ ràng sự phân tích sai số là rất quan trọng.



CHƯƠNG I :
ĐÁNH GIÁ CÁC SAI SỐ TRONG THỰC NGHIỆM

A. Các biểu diễn số liệu

Để đo chiều dài với một thước mét chúng ta định vị cẩn thận điểm đầu và cuối ví dụ chúng ta liếc nhìn tại điểm cuối vạch cm gần nhất là 54 cm rồi thì đếm số milimet qua điểm ghi nhận centimet và nhận được 7 mm . Chúng ta ghi nhận rằng điểm cuối ở một vị trí nào đó giửa 2 vạch milimet có khả năng giửa 0,2 và 0,4 mm hay sự bất định là $\pm 0,2mm$ và làm thẳng hàng giửa điểm đầu và điểm cuối thì sự bất định của việc đọc giai đo trên thước mét là $\pm 0,5mm$. Ở toàn bộ có một sự bất định tối thiểu là $\pm 0,7mm$. Nếu chúng ta thêm vào những nguồn bất định khác mà cho sự bất định tối thiểu là $\pm 1mm$ hay $\pm 0,1cm$. Chúng ta có thể viết

(54,7 ± 0,1) cm

Khi đối với các số lớn chúng ta tránh viết các zero mà không có nghĩa, ví dụ 24000 ± 400 calo. Trong trường hợp nầy là tốt hơn viết $(24,0 \pm 0,4)\times 10^3$ calo. Trong khoa học từ sai số có nghĩa là sự bất định trong một đo lường và các sai số trong nghĩa nầy là sự đo lường không thể có độ chính xác hoàn hảo.

1. Sai số tuyệt đối:

Giả sử đại lượng x đo được , thông thường là không khó khăn để có sự ước đoán hợp lý Δx tầm của sự bất định phép đo. Kết quả được biểu diễn như

$$x \pm \Delta x$$

và chúng ta thường gọi Δx là sai số tuyệt đối

2. Sai số tương đối

Người ta cũng dùng sai số tương đối $\delta x = \dfrac{\Delta x}{x}$ và sai số phần trăm $\delta x\% = 100.(\dfrac{\Delta x}{x})\%$

Ví dụ: $R = 6\Omega \pm 3\%$ có nghĩa là giá trị R đo được có sai số là ± 0,18 Ω

3. Các chữ số có nghĩa

Chúng ta xét các đại lượng sau. Một vận động viên chạy được một quảng đường 91 m trong 37 s, tốc độ trung bình là

$$v = \frac{91}{37} = 2,459459459 m/s$$

Vận-tốc v trong trường hợp nầy không cần thiết chính xác như tính được ở trên mà có thể có các giá trị sau

$$v_{trên} = \frac{91}{36} = 2,52777777 \text{ m/s}$$

$$v_{duoi} = \frac{91}{38} = 2,3947368 \text{ m/s}$$

Trong các đo lường ở phòng thí nghiệm qui tắc tổng quát là giữ lại một và chỉ một chử số nghi ngờ. Do đó sự trả lời là v = 2,5 m/s. Chữ số đầu tiên của sự trả lời (2) là chắc chắn (vì cả hai $v_{trên}$ và v_{duoi} bắt đầu với số 2). Chữ số thứ hai (ở vị trí thập phân đầu tiên) là không chắc chắn. Do đó sự trả lời cho v được làm tròn đến chữ số nầy và phần còn lại được khử.

Các qui tắc thuộc về các chữ số có nghĩa là như sau:

1/ Khi ghi nhận dữ liệu hay tính một kết quả, chữ số thập phân cuối cùng (và duy nhất chỉ chữ số cuối cùng) là bất định.

2/Số của các chữ số có nghĩa là số của các chữ số thập phân đếm từ chữ số cao nhất đến chữ số nghi ngờ đầu tiên

0.002477 có 4 chữ số có nghĩa (các zero dẫn đầu không tính)

$3,45\times10^6$ có 3 chữ số có nghĩa

3/ Số các chữ số có nghĩa trong một kết quả tính thường là bằng số nhỏ nhất của các chữ số có nghĩa trong các dữ liệu nhập vào.

Ví dụ : Biểu thị gia tốc trọng trường tại một nơi nào đó với 3 chử số có nghĩa thì chúng ta chỉ ra sự tính toán giá trị gia tốc trọng trường ở phép đo

$$g = \frac{4\pi^2 l}{T^2}$$ l và T là các giá trị đo

Cũng có 3 chữ số có nghĩa.

4/ Phép tính cộng trừ ở đây chữ số nghi ngờ cao nhất của các dữ liệu nhập vào trở thành chữ số nghi ngờ đầu tiên của lời giải

Ví dụ $4,3\underline{7} + 0,014\underline{8} - 0,16\underline{5} = 4,2\underline{1}98 = 4,2\underline{2}$

Ở đây chữ số nghi ngờ đầu tiên của sự trả lời được xác định bởi vị trí của chữ số cuối cùng của 4,37.

Khi trừ các đại lượng gần bằng nhau, kết quả có thể có các chữ số có nghĩa ít hơn các số ban đầu.

Ví dụ $21,437\underline{5} - 21,3\underline{8} = 0,0\underline{5}75$

$= 0,06$

5/ Đối với các tính toán trung gian nó là thích hợp giữ lại các chữ số vượt quá tránh sự tích lủy sai số làm tròn

6/ Làm tròn số

Khi các số làm tròn để khử các chữ số không có nghĩa, luôn luôn làm tròn lên các số trên 5 đến 10, và làm tròn xuống các số bên dưới 5 đến 0.

Ví dụ : Một số có 3 chử số có nghĩa

10,67 làm tròn thành 10,7 (vì 10,67 gần 10,7 hơn 10,6)

10,63 làm tròn thành 10,6 (vì 10,63 gần 10,6 hơn 10,7)

Để làm tròn các số cuối cùng ở số 5 một qui tắc thường dùng cho các số 5 là làm tròn đến số chẵn gần nhất. Chúng ta xét các số trên đó số cuối là số 5

Bảng 1

Thực tế	Cắt khỏi	làm tròn lên	làm tròn đến số chẵn
10,65	10,6	10,7	10,6
10,55	10,5	10,6	10,6
10,15	10,1	10,2	10,2
10,45	10,4	10,5	10,4
10,85	10,8	10,9	10,8
10,55	10,5	10,6	10,6
10,35	10,3	10,4	10,4
10,75	10,7	10,8	10,8
10,45	10,4	10,5	10,4
10,65	10,6	10,7	10,6
10,55	10,5	10,6	10,6
10,25	10,2	10,3	10,2
---------	--------	----------	---------
136,35	135,9	137,2	136,6

Rõ ràng rằng các sai số từ sự làm tròn số 5 là nhỏ nhất nếu người ta làm tròn đến số chẵn gần nhất, nghĩa là nếu trước số 5 là số chẵn thì giữ nguyên chữ số đó và nếu trước số 5 là số lẽ thì tăng chữ số đó lên 1.

7/ *Nhân, chia:*

Muốn nhân các số thập phân ta làm tròn các chữ số chính xác nhiều hơn, đến một chữ số có nghĩa hơn số chính xác thấp nhất trước khi tính kết quả của phép tính sẽ cho có cùng chữ số có nghĩa như số chính xác thấp hơn.

Ví dụ: $\dfrac{(1,2)(6,335)(0,0072)}{3,14159}$

Ta làm tròn như sau: $\dfrac{(1,2)(6,34)(0,0072)}{3,14} = 0,017$

Kết quả cuối cùng là 0,017 ít nhất có hai chữ số có nghĩa

B. Phân loại sai số

Trong các quá trình thực nghiệm, người ta thu được các giá trị nào đó của phép thí nghiệm và so sánh với một mô hình nào đó. Tùy theo mức độ phù hợp của chúng người thí nghiệm có thể đánh giá sai số của phép đo như thế nào và trong trường hợp cần thiết có thể tiến hành xử lý thống kê các dữ kiện thực nghiệm thô.

Thông thường, thí nghiệm càng chính xác khi các dụng cụ thí nghiệm càng tinh vi và phức tạp hơn. Tuy nhiên sự xử lý toán học tốt các kết quả thu được cho phép một phần nào loại trừ sai số phép đo và điều nầy chỉ ra rằng không kém hữu hiệu hơn so với khi sử dụng các dụng cụ chính xác đắt tiền. Trong phần nầy chúng ta dành cho sự khảo sát xử lý thống kê các kết quả thực nghiệm , nó cho phép ước lượng và đánh giá sai số của phép đo. Các sai số thực nghiệm có thể chia làm ba loại. Sai số hệ thống, sai số ngẫu nhiên và sai số thô. Chúng ta khảo sát các loại sai số này.

1. Sai số hệ thống:

Các hiệu ứng hệ thống là một dạng tổng quát của thiết bị khi tiến hành thí nghiệm. Đây là các sai số nó ảnh hưởng tất cả số đọc trong cùng cách hay trong một số các cư xử thông thường, và là các bất định thuộc về dụng cụ đo.

Thí dụ nó bao hàm các hiệu ứng như phông , thế hiệu dịch lựa chọn, độ phân giải năng lượng, biến thiên hiệu suất ông đếm đối với vị trí nguồn và năng lượng, thời gian chết của hệ thiết bị..vv..trong một thí nghiệm ở vật lý hạt nhân. Các dạng nầy không thay đổi khi lặp lại nhiều lần thí nghiệm đo.

Người ta phân biệt ra 3 dạng sai số hệ thống

a. Sai số hệ thống tự nhiên

Như độ lớn của phông phóng xạ trong phép đo bức xạ luôn tồn tại trong phổ bức xạ đo được . Trong quá trình phân tích sử lý phổ người ta phải loại bỏ sự đóng góp các bức xạ phông nầy.

b. Sai số hệ thống nguồn gốc được biết

- Sai số giới hạn của dụng cụ: Nó liên quan đến các loại dụng cụ đo và xác định chúng bằng cấp chính xác tương ứng. Các sai số như thế có thể biết được giới hạn trên và không thể tính hiệu chính chúng

Ví dụ Độ phân giải của hệ thống phổ kế gamma

với detectơ nhấp nháy $\Delta E_{662kev} = 40keV$, với detectơ bán dẫn $\Delta E_{662keV} = 1keV$

- Sai số zero: Hầu hết các phép đo là hiệu số giữa số liệu đọc dụng cụ và số liệu đọc dụng cụ có lối vào zero. Thường sai số zero do sự bỏ qua việc kiểm tra zero của dụng cụ. Sai số zero là một sai số hệ thống bởi vì tất cả các điểm bị ảnh hưởng bởi cùng một lượng . Tuy nhiên nếu mỗi điểm đo lường bao hàm một điểm zero mới (giống các đo lường định thời gian) thì sai số zero được bao hàm như phần của sai số ngẫu nhiên.

c. Sai số hệ thống nguồn gốc không được biết

- Sai số chuẩn: Thí dụ như sử dụng các thiết bị đã củ và nó đã có các sai hỏng ngầm, độ chính xác thực tế xấu hơn so với tiêu chuẩn kỹ thuật đã cho của dụng cụ. Để tìm ra sai số hệ thống thông thường các dụng cụ cần được lấy chuẩn với các tiêu chuẩn được định trước bởi nhà sản xuất

2. Sai số ngẫu nhiên:

Là hệ quả của một số lớn các nhân tố ngẫu nhiên mà tác dụng của chúng nhỏ không thể tách riêng cho từng nhân tố. Các nhân tố nầy rất nhiều và chúng phân biệt bằng bản chất vật lý khác nhau và ảnh hưởng ở quá trình đo.

Ví dụ: Đo vận tốc âm trong không khí từ một ống cộng hưởng kết quả đo được phụ thuộc vào nhiệt độ, áp suất không khí , độ ẩm và thời tiết môi trường

Các sai số ngẫu nhiên có thể làm sai lệch các kết quả phép đo ở cả hai phía khi lặp lại cùng một thí nghiệm. Các thăng giáng là ngẫu nhiên chỉ nếu giá trị trung bình không thay đổi ý nghĩa với thời gian. Trong các hiện tượng vi mô như các bức xạ phát ra từ hạt nhân nguyên tử, chúng tuân theo qui luật thống kê các giá trị đo được mang đặc trưng thống kê và kết quả có thể nhận với độ chính xác đến thăng giáng thông kê cho phép như chúng ta gặp ở hiện tượng phân rã, sự tán xạ, sự làm chậm (sự mất năng lượng trung bình) và bằng phương pháp của lý thuyết xác suất thống kê có thể tính được ảnh hưởng của chúng nghĩa là biết được qui luật phân bố thống kê của các đại lượng ngẫu nhiên.

3. Sai số thô:

Là kết quả của sự không cẩn thận của người quan sát có thể viết một số khác thay cho một số nào đó. Ở một phép đo duy nhất sai số thô không thể nhận ra nhưng nếu phép đo lặp lại một số lần nào đó thì sự xử lý thống kê các giới hạn của sai số ngẫu nhiên sẽ chỉ ra được sai số thô.

C. Các khái niệm cơ sở và mối liên hệ với thống kê toán học

Mục đích của hầu hết các đo lường là tìm giá trị của đại lượng vật lý. Giá trị nầy có thể được dẫn xuất trực tiếp từ các đo lường hoặc được dẫn xuất từ các giá trị đo bằng một hệ thức toán học. Trong các trường hợp như vậy, chúng ta cần có sự phân tích thống kê nhằm thu giá trị cuối cùng và sự bất định của nó. Nếu một đại lượng được đo lặp lại với cùng hay khác phương pháp trong cùng hoặc ở phòng thí nghiệm khác nhau. Các giá trị quan sát sai lệch từ một giá trị đúng không biết được gọi là sai số của phép đo và chúng có thể được sử lý một cách định lượng bởi các phương pháp thông kê.

1. Trung bình số học:

Nếu n đo lường độc lập các giá x_i của một đại lượng nào đó ở đây i=1,2, …n. Giá trị trung bình của chúng là

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1.1)$$

+ Nó là giá trị biểu diễn kết quả đo của phép thí nghiệm

+ Nó tham gia vào việc xác định sai số của phép đo.

Ví dụ: Trong phép đo chu kì T của một dao động người ta thông thường đo một bội số nhiều lần chu kỳ T bằng một giá trị x như 10T = x thì giá trị $T = \frac{x}{10}$ chính là giá trị trung bình của phép thí nghiệm hay ở thí nghiệm đo khoảng cách vân i giao thoa để tìm bước sóng ánh sáng người ta đo khoảng cách giữa nhiều vân tiếp trong trường giao thoa bằng một giá trị x như 5.i = x thì giá trị $i = \frac{x}{5}$ là giá trị của phép thí nghiệm.

Người ta cũng có thể xem x_i như là giá trị của đại lượng ngẫu nhiên X, phân bố theo qui luật nào đó chưa biết . Rỏ ràng kỳ vọng EX của nó (có thể viết $<x>$) bằng giá trị chính xác của phép đo. Sự xử lý các giá trị đo dựa trên định lý giới hạn trung tâm của lý thuyết xác suất.

Nếu X là đại lượng ngẫu nhiên phân bố theo qui luật bất kỳ thì $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ cũng là đại lượng ngẫu nhiên , hơn nữa

$$E\bar{x}_n = <\bar{x}_n> = <\frac{1}{n}\sum_{i=1}^{n} x_i> = \frac{1}{n}\sum_{i=1}^{n} <x_i> = \frac{n<x_i>}{n} = <x_i>$$

Điều nầy nói lên rằng giá trị trung bình của trung bình một mẫu n quan sát rút ra từ cùng phân bố là cùng giá trị trung bình của mỗi quan sát. Điều nầy nói lên rằng trung bình mẫu như một ước lượng tốt nhất của trung bình phân bố.