ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

PHAN VĂN TÂN

PHƯƠNG PHÁP THỐNG KÊ TRONG KHÍ HẬU

HÀ NỘI - 1999

LỜI NÓI ĐẦU

Khí hậu luôn là bộ phận quan trọng của điều kiện tự nhiên và môi trường. Khí hậu có ý nghĩa quyết định đến nhiều mặt hoạt động sản xuất và đời sống. Điều kiện khí hậu là một trong những nhân tố tạo nên sự hình thành, tồn tại và phát triển của thế giới sinh vật, ảnh hưởng quan trọng đến nhiều lĩnh vực kinh tế và xã hội nhân văn của loài người. Bởi vậy, khi nói đến một miền đất nào đó người ta không thể không nhắc tới điều kiện khí hậu của nó.

Trong quá trình tồn tại và phát triển, con người luôn phải tìm hiểu, nghiên cứu điều kiện tự nhiên và môi trường để nắm bắt được các qui luật biến đổi của nó với mục đích cải tạo, chinh phục và khai thác nó. Vì vậy khí hậu cũng luôn là một đối tượng cần được tìm hiểu và nghiên cứu.

Một trong những phương pháp được ứng dụng phổ biến trong nghiên cứu khí hậu là phương pháp xác suất thống kê. Đây là một công cụ toán học được áp dụng rất rộng rãi và có hiệu quả trong nhiều lĩnh vực. "Phương pháp thống kê trong khí hậu" vận dụng một số nguyên lý của lý thuyết xác suất thống kê toán học, tính toán thông kê các đặc trưng khí tượng, khí hậu, giải quyết một số bài toán trong nghiên cứu qui luật, bản chất, đặc tính cũng như các vấn đề liên quan đến cấu trúc các trường khí quyển. Nó là cầu nối giữa lý thuyết xác suất thống kê toán học và khoa học khí quyển, là một môn học mang tính phương pháp.

Hiện nay có rất nhiều tài liệu viết về lý thuyết xác suất thống kê đang được lưu hành. Tuy vậy, một cách tương đối có thể phân chia các tài liệu này ra làm hai loại. Loại thứ nhất thiên về toán học, trong đó trình bày chặt chẽ lý thuyết xác suất dựa trên nền toán học ở trình độ cao. Những tài liệu này thường dùng cho các chuyên gia về toán nên rất khó đối với sinh viên cũng như một số ít chuyên gia ngành khí tượng thuỷ văn. Loại thứ hai bao gồm các tài liệu thống kê trong chuyên ngành, do các chuyên gia thuộc nhiều lĩnh vực chuyên môn khác nhau viết. Đối với loại tài liệu này, tuỳ thuộc vào từng chuyên ngành mà nội dung khai thác những kiến thức về lý thuyết xác suất thống kê cũng không nhất quán. Nói chung những tài liệu này thường chỉ đi sâu về một số khía cạnh và coi nhẹ những phần khác, đặc biệt trong đó chú trọng trình bày những ví dụ mang tính đặc thù chuyên ngành hẹp. Điều này cũng gây không ít khó khăn cho việc ứng dụng chúng trong chuyên ngành khí tượng khí hậu.

Trước tình hình đó, quyển sách này được biên soạn như là việc giải quyết một yêu cầu thúc bách của thực tế. Đúng với tên gọi của nó – "Phương pháp thống kê

trong khí hậu" – nội dung quyển sách chú trọng trình bày khía cạnh ứng dụng công cụ thống kê toán học vào chuyên ngành khí hậu. Quyển sách được viết trên cơ sở tập bài giảng mà tác giả đã dùng để giảng dạy cho sinh viên ngành khí tượng khí hậu trường Đại học Tổng hợp Hà Nội, nay là Đại học Quốc gia Hà Nội, trong nhiều năm gần đây. Mục đích viết cuốn sách này nhằm tạo cho sinh viên có được một tài liệu chính thống trong quá trình tiếp thu môn học "Phương pháp thống kê trong khí hậu" ở trường. Quyển sách cũng có thể dùng làm tài liệu tham khảo bổ ích cho các cán bộ, kỹ sư thuộc ngành khí tượng khí hậu và các độc giả thuộc những chuyên ngành gần gũi như thuỷ văn, hải dương trong quá trình làm công tác nghiên cứu và ứng dụng nghiệp vụ. Ngoài ra, những độc giả khác có quan tâm đến lĩnh vực ứng dụng của lý thuyết xác suất thống kê cũng có thể đọc và khai thác nó.

Quyển sách được viết cho những đối tượng đã được trang bị kiến thức toán cao cấp và lý thuyết xác suất thống kê toán học dành cho sinh viên ngành khí tượng thuỷ văn. Bởi vậy, trong quá trình trình bày, một số khái niệm, định nghĩa được xem là đã biết, do đó chúng chỉ được nêu ra một cách ngắn gọn mà không đi sâu chi tiết. Mặt khác, bám sát mục tiêu của chương trình đào tạo đại học chuyên ngành khí tượng khí hậu, quyển sách được viết dưới hình thức là một giáo trình môn học.

Trừ phần mở đầu và phụ lục, quyển sách được bố cục trong 7 chương:

Chương 1. Một số kiến thức cơ bản của lý thuyết xác suất và úng dụng trong khí tượng khí hậu. Chương này trình bày những khái niệm cơ bản nhất của lý thuyết xác suất và phương thức vận dụng chúng để giải quyết một số bài toán thường gặp trong thực tế.

Chương 2. Các đặc trưng số của phân bố và vấn đề phân tích khảo sát số liệu. Ở đây, trình bày những đặc trưng số quan trọng thường được ứng dụng trong phân tích khảo sát và nghiên cứu các tập số liệu khí tượng khí hậu cũng như các phương pháp ước lượng chúng.

Chương 3. Một số phân bố lý thuyết. Trình bày những phân bố xác suất lý thuyết thường được ứng dụng trong nghiên cứu các hiện tượng khí quyển và các bài toán kiểm nghiệm giả thiết thống kê trong khí hậu.

Chương 4. Kiểm nghiệm giả thiết thống kê trong khí hậu. Chương này đề cập đến một loạt bài toán liên quan đến vấn đề kiểm nghiệm giả thiết thống kê thường gặp trong khí hậu, cách thức nêu bài toán và các bước tiến hành kiểm nghiệm.

Chương 5. Phân tích tương quan và hồi qui. Ở đây trình bày các phương pháp xác định mức độ và dạng thức liên hệ giữa các chuỗi số liệu khí tượng, khí hậu trên cơ sở các phương pháp phân tích tương quan và hồi qui của thống kê toán học, trong đó chú trọng các phương pháp nghiên cứu quan hệ tuyến tính và biến đổi các mối quan hệ phi tuyến về dạng tuyến tính.

Chương 6. Chỉnh lý số liệu khí hậu. Trên cơ sở những kiến thức về phân tích

tương quan và hồi qui, chương này trình bày phương pháp xử lý ban đầu các chuỗi số liệu khí hậu, phương pháp giải quyết một trong những vấn đề cơ bản luôn tồn tại trong các chuỗi số liệu khí hậu là chuỗi ngắn và gián đoạn. Ngoài ra ở đây còn nêu một số phương pháp xác định các đặc trưng của chuỗi ngắn thông qua việc bổ khuyết và kéo dài chuỗi.

Chương 7. Phân tích chuỗi thời gian. Chương này trình bày một số phương pháp thông dụng nghiên cứu hai đặc tính cơ bản nhất của các chuỗi số liệu khí hậu là tính xu thế và tính chu kỳ, qua đó nhằm trang bị những công cụ hữu hiệu cho việc giải quyết một trong những nhiệm vụ thời sự của khí hậu hiện đại là nghiên cứu biến đổi khí hậu.

Nhằm giúp cho người đọc có thể tiếp cận vấn đề một cách nhanh chóng, tác giả đã cố gắng tuân thủ nguyên tắc trình bày là sau mỗi một phần lý thuyết sẽ có các ví dụ minh hoạ gần sát với những bài toán thực tế. Tuy vậy, do khuôn khổ quyển sách có hạn, hệ thống các bài tập không được đưa vào đây mà sẽ dành cho một cuốn sách khác. Một số ví dụ cũng không được trình bày chi tiết. Mặt khác quyển sách cũng chưa chú trọng đến những nội dung liên quan với việc phân tích không gian, phân vùng và lập bản đồ khí hậu.

Ngoài những tài liệu đã được liệt kê trong danh mục tài liệu tham khảo, khi biên soạn quyển sách tác giả còn tham khảo thêm tập bài giảng mà GS–PTS Nguyễn Trọng Hiệu đã dùng để giảng dạy cho sinh viên ngành khí tượng khí hậu trong những năm của thập kỷ bảy mươi. Đó là một nguồn tư liệu quí giá giúp cho tác giả định hướng lựa chọn phương pháp trình bày nội dung cũng như bố cục của cuốn sách.

Trong quá trình biên soạn quyển sách, tác giả đã nhận được những ý kiến đóng góp quí báu của các đồng nghiệp thuộc Đại học Quốc gia Hà Nội; nhận được sự giúp đỡ tận tình, những lời động viên chân thành và những ý kiến bổ sung về mặt học thuật của các thành viên Hội đồng Khoa học khoa Khí tượng Thuỷ văn & Hải dương học, trường Đại học Khoa học Tự nhiên. Nhân đây tác giả xin bày tỏ lòng biết ơn sâu sắc. Đặc biệt tác giả xin chân thành cám ơn PGS–PTS Nguyễn Văn Tuyên và PGS–PTS Nguyễn Văn Hữu, những người đã đọc kỹ bản thảo của cuốn sách và cho những nhận xét quí báu.

Do trình độ và kinh nghiệm còn hạn chế, chắc chắn quyển sách còn những khiếm khuyết nhất định. Tác giả hy vọng nhận được sự góp ý của các đồng nghiệp và các độc giả.

Hà Nội, tháng 01 năm 1999

Tác giả

MỞ ĐẦU

Khi nghiên cứu một hiện tượng nào đó xảy ra trong khí quyển ta cần phải quan sát nó, trắc lượng nó. Hiện tượng được nghiên cứu nói chung luôn luôn liên hệ với các hiện tượng khác bởi những mối phụ thuộc có tính nguyên nhân, và vì vậy tiến trình của nó phụ thuộc vào vô số các nhân tố bên ngoài. Về nguyên tắc ta không thể theo dõi được tất cả các nguyên nhân xác định tiến trình của hiện tượng nghiên cứu và cũng không thể thiết lập được tất cả các mối liên hệ giữa hiện tượng đang xét với toàn bộ những yếu tố bên ngoài. Ta chỉ có thể thiết lập và theo dõi được một số nhất định các mối liên hệ giữa hiện tượng nghiên cứu với những nhân tố khác, và đương nhiên còn vô số những nhân tố nữa chưa được tính đến, chúng có tác dụng nào đó đến tiến trình của hiện tượng khảo sát. Chính vì vậy mà khi quan sát hiện tượng nhiều lần, bên cạnh những đặc điểm chung nhất, ta thấy mỗi lần hiện tượng xuất hiện với một dáng vẻ khác nhau, mang những đặc điểm riêng đặc trưng cho từng lần quan sát. Kết quả là các lần quan sát khác nhau không hoàn toàn giống nhau. Chẳng hạn, trong trường hợp lý tưởng, nếu chúng ta đồng thời đo nhiệt độ không khí tại một địa điểm nào đó vào một thời điểm nhất định bằng nhiều nhiệt kế giống nhau, có thể nhận được những trị số khác nhau dao động xung quanh một giá trị nền nào đó. Sự khác nhau này phụ thuộc vào rất nhiều nhân tố khách quan, như mức độ đồng nhất của các nhiệt kế về độ nhạy, độ chính xác, tác dụng bức xạ của mặt trời, mặt đệm đến các bầu nhiệt kế,...

Vì lẽ đó, khi nghiên cứu mỗi hiện tượng cho trước, người ta tách tất cả những mối liên hệ thành hai loại: các *mối liên hệ cơ bản* xác định những nét chung tiến trình của hiện tượng, mà khi quan sát chúng được lặp đi lặp lại nhiều lần, và các *mối liên hệ thứ yếu* có ảnh hưởng khác nhau đến tiến trình tại mỗi lần quan sát. Các mối liên hệ cơ bản xác định cái gọi là tính qui luật của hiện tượng. Các mối liên hệ thứ yếu làm cho kết quả quan sát hiện tượng sai lệch khác nhau so với qui luật tại mỗi lần quan sát. Những sai lệch đó được gọi là những hiện tượng ngẫu nhiên.

Mỗi một mối liên hệ thứ yếu riêng biệt nói chung chỉ có thể ảnh hưởng rất ít đến tiến trình của hiện tượng. Tuy nhiên, vì có vô số các mối liên hệ thứ yếu cùng tác động nên ảnh hưởng tổng cộng của chúng có khi lại rất đáng kể, thậm chí chúng xác định tất cả tiến trình của hiện tượng, làm cho hiện tượng không còn một tính qui luật rõ rệt nào cả.

Do tác dụng đồng thời của các mối liên hệ cơ bản và các mối liên hệ thứ yếu

nên tính qui luật và tính ngẫu nhiên trong mọi hiện tượng luôn luôn liên hệ mật thiết với nhau, gắn chặt với nhau.

Vì hiện tượng ngẫu nhiên được sinh ra bởi vô số mối liên hệ thứ yếu trong hiện tượng cần khảo sát nên, về nguyên tắc, việc nghiên cứu chúng bằng cách theo dõi tất cả các mối liên hệ này là không thể được. Chúng ta chỉ có thể nghiên cứu hiện tượng ngẫu nhiên bằng cách phát hiện tính qui luật trong bản thân chúng.

Lý thuyết xác xuất là một ngành toán học nghiên cứu tính quy luật của những hiện tượng ngẫu nhiên. Để xác định được tính quy luật cần phải biết được các đặc trưng xác suất của hiện tượng ngẫu nhiên. Muốn vậy, không còn cách nào khác là phải trở về với thực nghiệm. Việc xây dựng được các phương pháp hợp lý để xử lý các kết quả quan sát thực nghiệm là nội dung cơ bản của lý thuyết thống kê.

Theo nghĩa đó, "Phương pháp thống kê trong khí hậu" là môn học vận dụng một số nguyên lý của lý thuyết xác suất thống kê toán học, tính toán thống kê các đặc trưng khí hậu, giải quyết một số bài toán trong nghiên cứu các hiện tượng khí hậu. Nó là một môn học mang tính phương pháp, là cầu nối giữa lý thuyết xác suất thống kê toán học và khí hậu học.

Khí hậu là trạng thái trung bình của thời tiết. Thời tiết là trạng thái tức thời của khí quyển, được qui định bởi các quá trình, các đặc trưng vật lý của khí quyển. Nghiên cứu khí hậu là xác định được những qui luật diễn biến của khí hậu theo không gian và thời gian, thiết lập được những mối liên hệ bên trong và bên ngoài của các đặc trưng yếu tố khí hậu, từ đó tiến hành đánh giá tài nguyên khí hậu, phán đoán về sự biển đổi khí hậu và giải bài toán dự báo khí hậu.

Trên cơ sở các chuỗi số liệu khí hậu "Phương pháp thống kê trong khí hậu" căn cứ vào tính hai mặt của các quá trình và hiện tượng khí hậu là tính quy luật và tính ngẫu nhiên để:

1) Thống kê, tính toán và ước lượng các trị số khí hậu;
2) Phán đoán và kiểm nghiệm luật phân bố của một số đặc trưng yếu tố khí hậu;
3) Phân tích mối liên hệ tương quan và hồi qui giữa các đặc trưng yếu tố khí hậu;
4) Phân tích qui luật biến đổi của các chuỗi số liệu khí hậu;
5) Chỉnh lý, bổ sung các chuỗi số liệu khí hậu.

Số liệu khí hậu, kết quả thực nghiệm của việc quan sát các hiện tượng khí quyển, là yếu tố quan trọng, cần thiết và không thể thiếu được đối với việc sử dụng phương pháp thống kê trong nghiên cứu khí hậu. Thông thường số liệu khí hậu được thành lập từ các số liệu khí tượng. Số liệu khi tượng là số liệu thu thập được

từ những quan trắc khí tượng. Nghĩa là:

Quan trắc khí tượng ⟶ Số liệu khí tượng ⟶ Chuỗi số liệu khí hậu.

Quan trắc khí tượng được tiến hành để theo dõi sự xuất hiện của các hiện tượng vật lý xảy ra trong khí quyển, đo đạc một số tính chất vật lý của khí quyển cấu thành thời tiết.

Khi nghiên cứu một hiện tượng nào đó người ta thường tiến hành khảo sát nhiều lần trong cùng những điều kiện như nhau nhằm mục đích giảm bớt sự tác động của các mối liên hệ thứ yếu, làm nổi bật những mối liên hệ cơ bản để xác định qui luật của hiện tượng. Chính vì vậy việc quan trắc khí tượng nói chung được tiến hành tại những địa điểm được chọn sẵn (là vị trí trạm khí tượng), vào những thời điểm qui định (là kỳ quan trắc) và theo một thể thức bắt buộc (qui trình, qui phạm quan trắc). Các yếu tố được quan trắc phải mô tả đầy đủ trạng thái thời tiết. Vị trí các trạm quan trắc được lựa chọn sao cho có thể bao quát được một vùng không gian nhất định. Các kỳ quan trắc phải được ấn dịnh vào những thời điểm điển hình, đủ để mô tả được biến trình thời gian của yếu tố. Việc tuân thủ qui trình, qui phạm quan trắc bảo đảm tính nhất quán trong số liệu thu nhập được.

Kết quả của quan trắc khí tượng cho ta tập số liệu đo đạc thực nghiệm các hiện tượng khí tượng, các tính chất vật lý của khí quyển mô tả điều kiện thời tiết. Từ tập số liệu này, bằng các phương pháp chọn mẫu khác nhau người ta mới thành lập các chuỗi số liệu khí hậu.

Chuỗi số liệu khí hậu là một bộ phận của tổng thể khí hậu. Nó là bộ phận duy nhất mà ta có thể có để từ đó tiến hành thống kê tính toán và nhận định phán đoán. Tổng thể khí hậu là tập hợp mọi thành phần có thể của đặc trưng yếu tố khí hậu. Tổng thể khí hậu bao gồm 3 nhóm: 1) Nhóm các trị số đã xảy ra nhưng không được quan trắc; 2) Nhóm các trị số đã xảy ra và đã được quan trắc; 3) Nhóm các trị số chưa xảy ra. Số thành phần của tổng thể là vô hạn. Tổng thể luôn luôn bao quát đầy đủ mọi sắc thái hình thù của đặc trưng yếu tố khí hậu.

Trên cơ sở các chuỗi số liệu khí hậu ta có thể tiến hành xử lý, tính toán các đặc trưng tham số khí hậu, phân tích, phán đoán và mô tả đặc điểm, tính chất, cấu trúc bên trong, tiến đến dự báo khí hậu. Chất lượng tính toán phụ thuộc vào khả năng của chuỗi (dung lượng mẫu – độ dài chuỗi). Thông thường các thành phần của chuỗi cách nhau một năm, nên số lượng các năm quan trắc càng nhiều thì dung lượng mẫu càng lớn, kết quả tính toán sẽ càng đảm bảo độ ổn định thống kê và do đó những phân tích, phán đoán càng chính xác.

CHƯƠNG 1
MỘT SỐ KIẾN THỨC CƠ BẢN CỦA LÝ THUYẾT XÁC SUẤT VÀ ỨNG DỤNG TRONG KHÍ TƯỢNG KHÍ HẬU

1.1 Sự kiện, không gian sự kiện và tần suất sự kiện

1.1.1 Phép thử và sự kiện

Các khái niệm đầu tiên của lý thuyết xác suất là "*phép thử*" và "*sự kiện*". "*Phép thử*" được hiểu là việc thực hiện một bộ điều kiện xác định nào đó khi nghiên cứu một hiện tượng. "*Phép thử*" cũng có thể hiểu là "*thí nghiệm*" hoặc "*quan sát*" hay "*quan trắc*", "*trắc lượng*",... về sự xuất hiện một hiện tượng nào đó. Quan trắc khí tượng là một kiểu mô phỏng "*phép thử*" như vậy. Kế quả của "*phép thử*" là kết cục. Một phép thử có thể có nhiều kết cục. Các kết cục này được gọi là các "*sự kiện*". Người ta chia các sự kiện thành sự kiện cơ sở và sự kiện phức hợp.

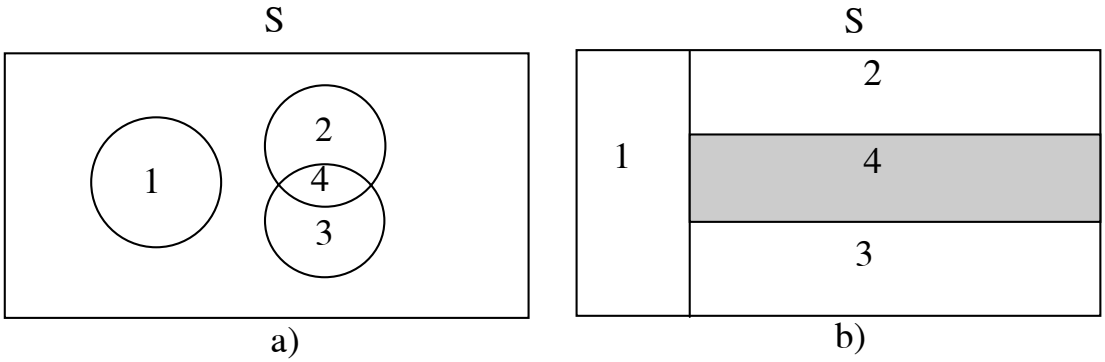
Trong những trường hợp đơn giản có thể phân biệt được rõ ràng sự kiện cơ sở và sự kiện phức hợp. Chẳng hạn sự kiện con xúc xắc nhận mặt nào khi ta gieo là sự kiện cơ sở. Nhưng trong khí tượng khí hậu, việc phân chia sự kiện cơ sở và sự kiện phức hợp nhiều khi cần phải căn cứ vào cách nhìn nhận vấn đề. Chẳng hạn, nếu chỉ quan tâm đến việc có giáng thuỷ hay không thì các sự kiện "ngày mai có giáng thuỷ" và "ngày mai không có giáng thuỷ" có thể được xem là những sự kiện cơ sở. Song, nếu xét thêm giáng thuỷ dạng nào – "lỏng" hay "rắn", thì sự kiện "ngày mai có giáng thuỷ" là sự kiện phức hợp, nó có thể được chia thành các sự kiện cơ sở: "ngày mai có giáng thuỷ lỏng" – mưa, "ngày mai có giáng thuỷ rắn" – tuyết rơi chẳng hạn và "ngày mai có giáng thuỷ hỗn hợp cả lỏng và rắn" – mưa và tuyết rơi. Nếu còn xét đến lượng giáng thuỷ thì các sự kiện này sẽ trở thành những sự kiện phức hợp, ta có thể chia chúng thành những sự kiện nhỏ hơn, chẳng hạn giáng thuỷ trên 10mm và dưới 10mm, v.v.

1.1.2 Không gian sự kiện

Không gian sự kiện, hay không gian mẫu, là tập hợp tất cả những sự kiện cơ sở có thể có. Như vậy không gian mẫu biểu diễn mọi kết cục hay sự kiện có thể có. Nó tương đương với sự kiện phức hợp lớn nhất.

Mối quan hệ giữa các sự kiện có thể được mô tả bằng hình học. Thông thường người ta biểu diễn không gian mẫu bởi một hình chữ nhật mà bên trong nó là các hình tròn biểu thị những sự kiện. Ví dụ trên hình 1.1a, không gian mẫu là hình chữ nhật S biểu thị những kết cục giáng thuỷ trong ngày mai. Bốn sự kiện cơ sở được mô tả bởi phần bên trong của ba hình tròn (được đánh số 1, 2, 3, 4). Hình tròn đứng độc lập tương ứng với sự kiện "không có giáng thuỷ". Phần giao nhau của hai hình tròn còn lại biểu thị có giáng thuỷ hỗn hợp cả hai dạng (lỏng và rắn), còn phần của hình chữ nhật nằm ngoài các hình tròn tương ứng với sự kiện trống rỗng, nó không thể xuất hiện.

Tuy nhiên cũng không nhất thiết phải biểu diễn mối quan hệ giữa các sự kiện theo sơ đồ trên đây. Thông thường người ta xem không gian sự kiện lấp đầy toàn bộ hình chữ nhật S mà trong đó các sự kiện cơ sở phủ vừa kín nó (hình 1.1b). Với cách biểu diễn này hình chhữ nhật S được xem như là sự kiện phức hợp lớn nhất, trong đó có thể chia thành các miền không giao nhau biểu thị các sự kiện xung khắc với nhau. Chẳng hạn trên hình 1.1b, bốn miền không giao nhau tương ứng với bốn sự kiện cơ sở đã nói trên đây. Trong trường hợp này, nhất thiết một trong bốn sự kiện phải xảy ra. Mặt khác cũng cần lưu ý rằng mỗi một trong các sự kiện cơ sở biểu thị có giáng thuỷ ta có thể thêm vào các đường phân chia để biểu diễn những sự kiện nhỏ hơn, chẳng hạn lượng giáng thuỷ trên 10mm và dưới 10mm.



Hình 1.1 Sơ đồ biểu diễn không gian mẫu.

1) Không có giáng thuỷ; 2) Giáng thuỷ lỏng; 3) Giáng thuỷ rắn; 4) Giáng thuỷ hỗn hợp

1.1.3 Tần suất sự kiện

Khi tiến hành phép thử, hiện tượng có thể xuất hiện cũng có thể không xuất hiện. Để đo độ chắc chắn của sự kiện "hiện tượng xuất hiện" hay "hiện tượng không xuất hiện" trong lần thử người ta sử dụng khái niệm "*xác suất sự kiện*". Xác suất của sự kiện A nào đó nằm trong khoảng từ 0 đến 1:

$$0 \le P(A) \le 1 \qquad (1.1.1)$$

Sự kiện có xác suất xuất hiện bằng 0 ứng với sự kiện bất khả V còn sự kiện có xác suất xuất hiện bằng 1 ứng với sự kiện chắc chắn U, tức P(V) = 0, P(U) = 1.

Theo định nghĩa cổ điển, xác suất của sự kiện A là tỷ số giữa số kết cục thuận lợi cho A so với tổng số kết cục đồng khả năng. Tuy nhiên, định nghĩa này chỉ áp dụng được khi số kết cục đồng khả năng là hữu hạn. Để tính được xác suất của sự kiện cho một lớp phép thử rộng lớn hơn, người ta đưa đưa vào định nghĩa xác suất theo quan điểm thống kê. Khái niệm cơ bản đưa tới định nghĩa này là khái niệm tần suất.

Giả sử tiến hành (trên thực tế) n phép thử cùng loại khi nghiên cứu một hiện tượng nào đó. Gọi A là sự kiện "hiện tượng xuất hiện" và gọi m là số các phép thử quan sát thấy A. Khi đó tỷ số $\frac{m}{n}$ được gọi là tần suất xuất hiện sự kiện A trong loạt phép thử đã được tiến hành:

$$p = \frac{m}{n} \tag{1.1.2}$$

Trị số của tần suất nói chung phụ thuộc vào số lượng n phép thử được tiến hành. Khi n bé, tần suất thay đổi rõ rệt nếu ta chuyển từ loạt n phép thử này sang loạt n phép thử khác. Tuy nhiên thực nghiệm chứng tỏ rằng đối với phạm vi khá rộng, tần suất có tính ổn định, nghĩa là khi số phép thử n khá lớn thì trị số của tần suất biến thiên rất ít xung quanh một hằng số xác định nào đó. Ký hiệu xác suất của sự kiện A là P(A), theo định luật số lớn ta có:

$$P\left(\left|\frac{m}{n} - P(A)\right| \le \varepsilon\right) \to 0 \qquad \text{khi} n \to \infty \tag{1.1.3}$$

trong đó ε là một số dương bé tuỳ ý.

Khái niệm tần suất là một khái niệm mang tính trực giác, kinh nghiệm nhưng có cơ sở lý thuyết vững chắc. Nó được ứng dụng rất có hiệu quả để ước lượng xác suất khí hậu. Nếu gọi A là sự kiện *hiện tượng khí hậu xuất hiện*, n là số lần quan sát hiện tượng, m là số lần xuất hiện hiện tượng trong n lần quan sát thì p là *tần suất xuất hiện hiện tượng*. Đại lượng p được dùng để ước lượng giá trị xác suất xuất hiện hiện tượng.

Ví dụ, từ số liệu mưa ngày lịch sử 50 năm của tháng 5 ở một trạm người ta quan sát thấy có có 487 ngày có mưa. Vậy xác suất xuất hiện mưa trong những ngày tháng 5 ở trạm này được xác định bởi trị số tần suất 487/(31 x 50) = 487/1550 = 0.314.

1.2 Một số phép tính và quan hệ về sự kiện và xác suất sự kiện

1) Hai sự kiện A và B được gọi là xung khắc với nhau nếu A xuất hiện thì B không xuất hiện và ngược lại. Các sự kiện A_1, A_2,..., A_n được gọi là lập thành nhóm đầy đủ các sự kiện nếu chúng xung khắc với nhau từng đôi một và nhất thiết một trong chúng phải xuất hiện.

2) Sự kiện B được gọi là sự kiện đối lập với sự kiện A nếu chúng không đồng thời xuất hiện và chúng lập thành nhóm đầy đủ các sự kiện. Ví dụ, các sự kiện "có giáng thuỷ" và "không có giáng thuỷ" là hai sự kiện đối lập. Trong trường hợp này ta có hệ thức:

$$P(B) = 1 - P(A) \tag{1.2.1}$$

3) Sự kiện B được gọi là tổng của hai sự kiện A_1 và A_2 nếu B xuất hiện kéo theo A_1 hoặc A_2 hoặc đồng thời cả A_1 và A_2 xuất hiện. Xác suất của sự kiện B trong trường hợp này bằng xác suất của tổng các sự kiện A_1 và A_2:

$$P(B) = P(A_1+A_2) = P(A_1) + P(A_2) - P(A_1.A_2) \tag{1.2.2}$$

Công thức này còn được gọi là qui tắc cộng xác suất.

Trong công thức (1.2.2) sự kiện $(A_1.A_2)$ được gọi là tích của các sự kiện A_1 và A_2, xuất hiện khi đồng thời cả A_1 và A_2 cùng xuất hiện.

$$P(A_1.A_2) = \text{Xác suất để } A_1 \text{ và } A_2 \text{ đồng thời xuất hiện} \tag{1.2.3}$$

Nếu A_1 và A_2 xung khắc với nhau thì $P(A_1.A_2) = 0$.

Qui tắc cộng xác suất có thể được mở rộng cho trường hợp nhiều sự kiện:

$$P(A_1+A_2+A_3) = P(A_1)+P(A_2)+P(A_3) - P(A_1.A_2)-P(A_2.A_3)-$$

$$-P(A_3.A_1)-P(A_1.A_2.A_3) \tag{1.2.4}$$

4) Xác suất có điều kiện

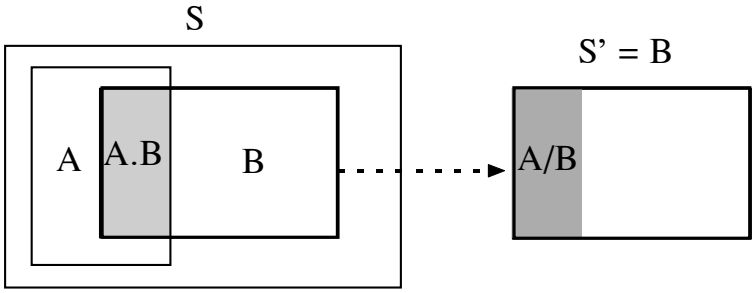
Trong thực tế người ta thường quan tâm đến xác suất của một sự kiện nào đó khi cho trước một vài sự kiện khác đã hoặc sẽ xảy ra. Chẳng hạn, tính xác suất của sự kiện xuất hiện mưa đá khi biết rằng có giáng thuỷ xảy ra; hoặc tính xác suất các cấp tốc độ gió ở một số vị trí nào đó ven bờ biển khi biết rằng bão đang đi đến gần và sẽ đổ bộ vào đất liền. Ở đây sự kiện được quan tâm là "mưa đá" và "tốc độ gió", còn sự kiện cho trước là "có giáng thuỷ" và "bão sẽ đổ bộ vào đất liền". Người ta gọi các sự kiện cho trước là những điều kiện hay sự kiện điều kiện, còn xác suất của sự kiện được quan tâm khi cho trước các điều kiện được gọi là xác suất có điều kiện. Nếu A là sự kiện đang xét, B là điều kiện cho trước thì xác suất có điều kiện của A là *xác suất của sự kiện A khi cho trước điều kiện B đã hoặc sẽ xuất hiện.* Ký hiệu

xác suất này là P(A/B). Nếu sự kiện B đã xuất hiện hoặc sẽ xuất hiện thì xác suất của sự kiện A là xác suất có điều kiện P(A/B). Nếu B không xuất hiện thì tự nó không cho thông tin gì đối với xác suất của sự kiện A.

Xác suất có điều kiện P(A/B) có thể được xác định bởi:

$$P(A/B) = \frac{P(A.B)}{P(B)} \qquad (1.2.5)$$

Có thể minh hoạ cách tính xác suất này trên hình 1.2.



Hình 1.2 Minh hoạ cách tính xác suất có điều kiện

Xác suất (không điều kiện) của A là tỷ số giữa diện tích miền A và S (hình bên trái). Xác suất có điều kiện của A với điều kiện B được xác định khi xét miền B như một không gian mẫu mới trên đó sự kiện A được biểu diễn bởi miền giao nhau A.B (hình bên trái)

5) Các sự kiện độc lập

Có thể viết lại công thức (1.2.5) dưới dạng qui tắc nhân xác suất:

$$P(A.B) = P(A/B).P(B) = P(B/A).P(A) \qquad (1.2.6)$$

Từ đó, hai sự kiện được gọi là độc lập với nhau nếu sự xuất hiện hoặc không xuất hiện của sự kiện này không làm ảnh hưởng đến xác suất xuất hiện của sự kiện kia và ngược lại. Chẳng hạn, kết cục của việc gieo đồng thời hai con xúc xắc là độc lập nhau. Sự độc lập giữa các sự kiện A và B cũng có nghĩa là:

$$P(A/B) = P(A) \text{ và } P(B/A) = P(B)$$

Từ tính chất độc lập của các sự kiện A và B suy ra:

$$P(A.B) = P(A).P(B) \qquad (1.2.7)$$

Ví dụ 1.2.1. Xét ước lượng xác suất khí hậu (tần suất) từ tập số liệu cho trong bảng 1.1. Giả sử ta quan tâm đến việc ước lượng xác suất để lượng mưa ở điểm A vào tháng 1 không dưới 0.3mm trong điều kiện nhiệt độ tối thấp không dưới 0^oC. Về mặt vật lý có thể nhận thấy rằng, nhiệt độ thường hạ xuống rất thấp vào những đêm trời quang, còn để xuất hiện mưa thì bầu trời phải có mây. Điều đó gợi cho ta ý tưởng rằng hai sự kiện *lượng mưa không dưới 0.3mm* và *nhiệt độ tối thấp không dưới* 0^oC có liên hệ thống kê với nhau (tức chúng không độc lập) và xác suất có điều

kiện của mưa được cho bởi những điều kiện nhiệt độ khác nhau sẽ khác nhau và khác với xác suất không điều kiện. Từ những kiến thức về bản chất vật lý của quá trình, có thể suy ra rằng xác suất có điều kiện của mưa với điều kiện nhiệt độ tối thấp ≥0°C sẽ lớn hơn xác suất có điều kiện này trong trường hợp ngược lại (nhiệt độ tối thấp nhỏ hơn 0°C).

Để tính tần suất có điều kiện này ta chỉ cần xem xét đến những trường hợp số liệu có *nhiệt độ tối thấp* $T_m \geq 0^oC$. Từ bảng 1.1 ta thấy có tất cả 24 ngày như vậy, trong đó có 14 ngày mưa với lượng mưa đo được R≥0.3mm. Do đó ta có ước lượng:

$$P(R\geq 0.3/\ T_m\geq 0) = 14/24 = 0.58$$

Trong số 7 ngày còn lại có nhiệt độ tối thấp dưới 0°C chỉ có 1 ngày có lượng mưa đo được R≥0.3mm. Do đó xác suất mưa trong trường hợp ngược lại (nhiệt độ tối thấp nhỏ hơn 0°C) sẽ là:

$$P(R\geq 0.3/\ T_m<0) = 1/7 = 0.14$$

Bảng 1.1 Số liệu nhiệt độ tối thấp và lượng mưa ngày điểm A tháng 1–1973

Ngày	R	T_m	Ngày	R	T_m	Ngày	R	T_m	Ngày	R	T_m
1	0.0	14.3	9	0.5	17.3	17	0.0	0.0	25	0.0	–9.8
2	1.8	18.8	10	1.3	20.3	18	0.0	1.5	26	0.0	–9.8
3	28.2	16.5	11	8.6	21.8	19	0.0	19.5	27	0.0	–8.3
4	0.0	–0.8	12	1.5	18.8	20	11.4	12.8	28	0.0	–3.0
5	0.0	3.0	13	4.6	21.8	21	0.0	14.3	29	0.3	–3.0
6	0.0	10.5	14	0.5	11.3	22	0.0	6.8	30	0.8	8.3
7	0.0	15.8	15	0.5	21.8	23	17.8	15.0	31	1.3	17.3
8	1.0	16.5	16	0.0	18.0	24	0.0	–4.5			

Tương tự như vậy, xác suất không điều kiện của lượng mưa trên 0.3mm bằng:

$$P(R\geq 0.3) = 15/31 = 0.48$$

Sự khác nhau của các xác suất có điều kiện nhận được trong ví dụ trên đây phản ánh sự phụ thuộc thống kê giữa hai đại lượng nhiệt độ tối thấp và lượng mưa. Tuy nhiên, khi đã hiểu biết tốt bản chất vật lý của quá trình ta sẽ không đi sâu vào việc nghiên cứu mối liên hệ tại sao nhiệt độ tối thấp càng cao sẽ là nguyên nhân gây mưa. Đúng hơn là giữa các sự kiện nhiệt độ và mưa tồn tại mối liên hệ thống kê vì chúng đều có mối quan hệ vật lý khác nhau với lượng mây. Vì sự phụ thuộc thống kê không nhất thiết bao hàm cả mối quan hệ nhân quả vật lý, nên khi đề cập đến sự phụ thuộc thống kê giữa các biến có thể không nhất thiết phải gắn nó với mối quan hệ vật lý của chúng.

Ví dụ 1.2.2. Tính xác suất có điều kiện theo chuỗi thời gian. Các biến khí quyển thường biểu lộ sự phụ thuộc thống kê giữa những trị số của chúng với những giá trị trong quá khứ hoặc tương lai. Mối phụ thuộc này xuyên suốt thời gian và được gọi là tính ổn định. Tính ổn định có thể được định nghĩa như là sự tồn tại mối phụ thuộc thống kê (dương) giữa những giá trị liên tiếp của cùng một biến, hoặc giữa sự xuất hiện liên tiếp các sự kiện cho trước nào đó. Sự phụ thuộc dương ở đây có nghĩa là những trị số lớn của biến có xu hướng sẽ kéo theo những trị số lớn tương ứng và ngược lại. Thông thường mối phụ thuộc thống kê của các biến khí tượng theo thời gian là dương. Ví dụ, xác suất để *nhiệt độ ngày mai vượt quá trung bình* sẽ lớn nếu *nhiệt độ ngày hôm nay đã trên trung bình*. Như vậy, cách gọi khác của tính ổn định là sự phụ thuộc dương của chuỗi.

Ta hãy xét tính ổn định của sự kiện xuất hiện mưa tại điểm A với tập số liệu nhỏ trong bảng 1.1 trên đây. Để đánh giá sự phụ thuộc của hiện tượng mưa trong chuỗi cần phải ước lượng xác suất có điều kiện dạng:

$$P(R_{hn}/R_{hq}),$$

trong đó: R_{hn} là có mưa ngày "hôm nay", R_{hq}– có mưa ngày "hôm qua".

Vì trong bảng 1.1 không chứa số liệu của ngày 31/12/72 và ngày 1/2/73 nên ta chỉ có 30 cặp "*hôm qua*/*hôm nay*" tham gia tính toán. Để tính $P(R_{hn}/R_{hq})$ ta chỉ cần đếm số ngày có mưa (như là điều kiện hoặc sự kiện "*hôm qua*") mà ngày tiếp sau cũng có mưa (như là sự kiện cần quan tâm hay sự kiện "*hôm nay*"). Khi ước lượng xác suất có điều kiện này người ta không quan tâm đến điều gì xảy ra ở những ngày tiếp theo không mưa. Trừ ngày 31/1, có tất cả 14 ngày có mưa, trong đó có 10 ngày mưa mà hôm sau cũng xảy ra mưa và 4 ngày có mưa mà hôm sau không mưa. Vì vậy tần suất có điều kiện sẽ được tính bởi:

$$P(R_{hn}/R_{hq}) = 10/14 = 0.71.$$

(10 ngày "*hôm nay*" có mưa trên tổng số 14 ngày có mưa được xét).

Bằng cách tương tự, xác xuất để "*hôm nay*" có mưa với điều kiện "*hôm qua*" không mưa được tính bởi:

$$P(R_{hn}/\overline{R_{hq}}) = 5/16= 0.31$$

(5 ngày "*hôm nay*" có mưa, 16 ngày "*hôm qua*" không mưa).

Sự khác nhau giữa các ước lượng xác suất có điều kiện này khẳng định sự phụ thuộc của các thành phần trong chuỗi số liệu. Xác suất $P(R_{hn}/R_{hq})$ chính là xác suất để hai ngày mưa liên tiếp. Bằng cách tương tự ta có thể tính được xác suất để 3

ngày, 4 ngày,... có mưa liên tiếp. Còn xác suất $P(R_{hn}/\overline{R_{hq}})$ là xác suất để ngày hôm sau có mưa nếu ngày hôm trước không mưa.

6) Qui tắc cộng xác suất

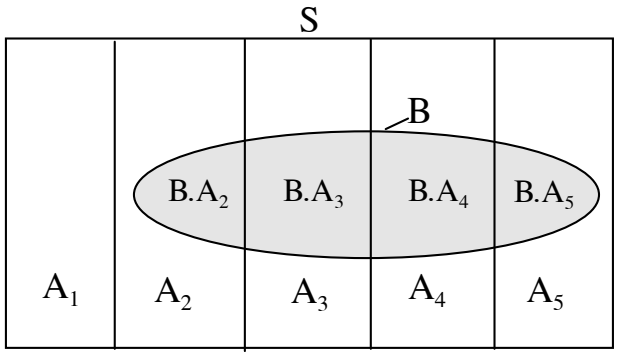
Xét nhóm đầy đủ các sự kiện xung khắc (MECE) A_i, i=1..L trên không gian mẫu được quan tâm và B cũng là một sự kiện được xác định trên không gian mẫu này (hình 1.3). Khi đó xác suất của sự kiện B có thể được tính bởi:

$$P(B) = \sum_{i=1}^{L} P(B.A_i) \tag{1.2.8}$$

Theo qui tắc nhân xác suất ta có:

$$P(B) = \sum_{i=1}^{L} P(B/A_i)P(A_i) \tag{1.2.9}$$

Như vậy, có thể tính được xác suất không điều kiện của B khi biết các xác suất có điều kiện của B và xác suất không điều kiện của các A_i. Cần chú ý rằng phương trình (1.2.9) chỉ đúng khi các sự kiện A_i tạo thành nhóm đầy đủ các sự kiện xung khắc của không gian mẫu.



Hình 1.3 Minh hoạ qui tắc cộng xác suất

Không gian mẫu S chứa sự kiện B (hình ellip) và 5 sự kiện xung khắc $A_1,...,A_5$

Ví dụ 1.2.3. Có thể xem xét ví dụ 1.2.2 trên đây dưới góc độ qui tắc cộng xác suất. Giả sử chỉ có L=2 sự kiện xung khắc lập thành nhóm đầy đủ trên không gian mẫu: A_1 là sự kiện *hôm qua có mưa* và $A_2 = \overline{A_1}$ là sự kiện *hôm qua không mưa*. Ký hiệu sự kiện B là *hôm nay có mưa*. Khi đó xác suất của B có thể được xác định bởi:

$$P(B) = P(B/A_1).P(A_1) + P(B/A_2).P(A_2)$$

Từ số liệu trong bảng, trừ ngày 31/1, số trường hợp được xét đến là 30 (ngày), trong đó 14 ngày có mưa (tức: $P(A_1) = 14/30$ và $P(A_2) = 16/30$). Trong số những ngày có mưa thì có 10 trường hợp thoả mãn hai ngày mưa liên tiếp (tức $P(B/A_1)=10/14$), với 16 ngày không mưa còn lại có 5 trường hợp ngày tiếp theo xảy ra mưa (nên $P(B/A_2)=5/16$). Vậy ta có:

$$P(B)=(10/14)(14/30)+(5/16)(16/30)=0.5$$

7) Định lý Bayes

Định lý Bayes là sự kết hợp lý thú của qui tắc cộng và nhân xác suất. Trong tính toán thông thường, định lý Bayes được dùng để tính ngược xác suất có điều kiện.

Ta hãy xét lại tình huống như đã chỉ ra trên hình 1.3, trong đó nhóm đầy đủ các sự kiện xung khắc A_i đã được xác định, còn B là một sự kiện khác xảy ra trên nền các sự kiện A_i. Từ qui tắc nhân xác suất và công thức (1.2.9) ta suy ra:

$$P(A_i/B) = \frac{P(B/A_i)P(A_i)}{P(B)} = \frac{P(B/A_i)P(A_i)}{\sum_{j=1}^{L} P(B/A_j)P(A_j)} \qquad (1.2.10)$$

Phương trình (1.2.10) là biểu thức của định lý Bayes. Nó được ứng dụng để tính xác suất có điều kiện của các sự kiện thành phần trong nhóm đầy đủ các sự kiện xung khắc A_i.

Ví dụ 1.2.4 Định lý Bayes từ quan điểm tần suất. Trong ví dụ 1.2.1 đã trình bày cách ước lượng xác suất có điều kiện đối với sự xuất hiện mưa với các điều kiện nhiệt độ tối thấp $T_m \geq 0^oC$ và $T_m < 0^oC$. Ta có thể sử dụng định lý Bayes để tính xác suất có điều kiện của T_m khi cho trước sự kiện mưa *có* hoặc ***không*** xuất hiện. Ký hiệu A_1 là sự kiện nhiệt độ tối thấp $T_m \geq 0^oC$, $A_2 = \overline{A_1}$ là sự kiện đối lập, tức nhiệt độ tối thấp $T_m < 0^oC$ và B là sự kiện xảy ra mưa. Rõ ràng hai sự kiện A_1 và A_2 lập thành nhóm đầy đủ các sự kiện trên không gian mẫu.

Từ số liệu ta có 24 trường hợp nhiệt độ tối thấp $T_m \geq 0^oC$ trên tổng số 31 ngày, vì vậy ước lượng xác suất không điều kiện đối với nhiệt độ tối thấp sẽ là:

$$P(A_1) = 24/31 \text{ và } P(A_2) = 7/31$$

Từ ví dụ 1.2.1 ta đã tính được $P(B/A_1) = 14/24$ và $P(B/A_2) = 1/7$.

Để tính các xác suất $P(A_i/B)$ theo công thức (1.2.10) cần phải tính giá trị P(B) ở mẫu số cho tất cả các trường hợp:

$$P(B) = P(B/A_1).P(A_1) + P(B/A_2).P(A_2)$$
$$= (14/24)(24/31) + (1/7)(7/31) = 15/31$$

(Kết quả này khác chút ít so với ước lượng xác suất mưa nhận được trong ví dụ 1.2.2, vì ở đó số liệu ngày 31/12 không được đưa vào tính).

Vậy, xác suất có điều kiện của nhiệt độ tối thấp $T_m \geq 0^oC$ với điều kiện có mưa là:

$$P(A_1/B) = (14/24)(24/31)(15/31) = 14/15$$

Tương tự, ta có xác suất có điều kiện đối với nhiệt độ tối thấp $T_m<0^oC$ với điều kiện có mưa là:

$$P(A_2/B) = (1/7)(7/31)(15/31) = 1/15$$

Những kết quả nhận được trong ví dụ trên đây đã khẳng định vai trò đóng góp thông tin của những sự kiện phụ thuộc. Giả sử dự báo viên đã đưa ra kết luận "nhiệt độ tối thấp $T_m\geq0^oC$". Nếu không có thông tin gì thêm ta có thể sử dụng xác suất không điều kiện $P(A_1) = 24/31$ để đánh giá *mức độ tin tưởng* vào kết luận dự báo. Người ta gọi xác suất $P(A_1)$ là xác suất tiên nghiệm (prior probability). Bây giờ giả sử rằng, bằng cách nào đó có thể biết được mưa sẽ xuất hiện (hay không xuất hiện), *mức độ tin tưởng* vào kết luận dự báo lúc này phụ thuộc vào mối quan hệ thống kê giữa nhiệt độ tối thấp và mưa, và sẽ được đánh giá thông qua xác suất có điều kiện $P(A_1/B)$ và $P(A_1/\overline{B})$ tương ứng với hai trường hợp có mưa (sự kiện B) và không mưa (sự kiện $\overline{B}$). Vì $P(A_1/B)=14/15 > P(A_1) = 24/31$ nên nếu mưa xuất hiện, kết luận dự báo "nhiệt độ tối thấp $T_m\geq0^oC$" có độ tin cây cao hơn. Hay nói cách khác, khi có thêm thông tin *mưa xuất hiện* xác suất dự báo đã bị thay đổi (tăng lên). Người ta gọi xác suất này là xác suất hậu nghiệm. Ở đây, xác suất hậu nghiệm lớn hơn xác suất tiên nghiệm.

1.3 Công thức Bernoulli và xác suất các sự kiện thông thường

Bài toán: Giả sử tiến hành n phép thử độc lập cùng loại và trong cùng một điều kiện như nhau. Mỗi một phép thử chỉ có 2 kết cục là A và $\overline{A}$. Xác suất xuất hiện sự kiện A ở mỗi phép thử không đổi, bằng p và không phụ thuộc vào chỉ số phép thử. Hãy tính xác suất để trong n lần trắc nghiệm, sự kiện A xuất hiện k lần.

Gọi B là sự kiện "trong n lần trắc nghiệm sự kiện A xuất hiện k lần". Sự kiện B có thể được thực hiện theo nhiều cách khác nhau: *Sự kiện A xuất hiện trong tổ hợp k phép thử bất kỳ của n phép thử*. Như vậy có tất cả C_n^k cách.

Ta có:

Xác suất xuất hiện sự kiện A là P(A) = p.

Xác suất xuất hiện sự kiện $\overline{A}$ là $P(\overline{A}) = 1–p = q$.

Vì các phép thử là độc lập nên xác suất hiện sự kiện B sẽ là:

$$P(B) = C_n^k p^k q^{n-k} \qquad (1.3.1)$$

Biểu thức (1.3.1) được gọi là công thức Bernoulli. Trong khí hậu công thức này thường được ứng dụng để tính xác suất các sự kiện thông thường.

Sự kiện thông thường là sự kiện có xác suất xuất hiện và không xuất hiện gần tương đương nhau. Bài toán được đặt ra ở đây là *hãy tính xác suất để trong n lần trắc nghiệm hiện tượng khí hậu xuất hiện k lần*. Ký hiệu xác suất này là $P_n(k)$, ta có:

$$P_n(k) = C_n^k p^k q^{n-k}. \quad (1.3.2)$$

Cần lưu ý rằng, công thức Bernoulli chỉ được áp dụng khi xác suất xuất hiện sự kiện không đổi và không phụ thuộc vào số thứ tự lần trắc nghiệm.

Ví dụ 1.3. Giả sử khảo sát chuỗi số liệu 100 năm tổng lượng mưa năm ở trạm A người ta thấy có 46 năm có lượng mưa vượt quá chuẩn khí hậu. Hãy tính xác suất để trong 10 năm quan trắc có 1, 2, 3, 5, 7 năm có lượng mưa vượt chuẩn khí hậu.

Gọi A là sự kiện "tổng lượng mưa năm vượt quá chuẩn khí hậu". Sự kiện A có thể được xem là sự kiện thông thường bởi, về ý nghĩa khí hậu, mưa là một yếu tố biến đổi thất thường, giá trị tổng lượng mưa năm nói chung thường dao động lên xuống xung quanh chuẩn khí hậu từ năm này sang năm khác. Xác suất sự kiện A có thể được ước lượng bởi tần suất $P(A) \approx p = 46/100 = 0.46$.

Từ đó, với n = 10 (10 năm quan trắc), p = 0.46, q = 1–p=0.54, k = 1, 2, 3, 5, 7 ta có:

$P_{12}(2)= C_{10}^2 (0.46)^2(0.54)^8$, $P_{10}(3)= C_{12}^3 (0.46)^3(0.54)^7$,

$P_{10}(5)= C_{10}^5 (0.46)^5(0.54)^5$, $P_{10}(7)= C_{10}^7 (0.46)^7(0.46)^3$.

1.4. Định lý Poisson và xác suất các sự kiện hiếm

Công thức Bernoulli trên đây chỉ cho kết quả chính xác khi số lượng phép thử n bé và p càng gần 0.5; khi p quá bé hoặc quá lớn thì sai số mắc phải sẽ khá lớn, hơn nữa khi *n* rất lớn việc tính toán càng trở nên phức tạp. Trong trường hợp này ta có thể áp dụng định lý Poisson sau đây:

Giả sử tiến hành n phép thử độc lập, mỗi phép thử sự kiện A xuất hiện với xác suất P(A) = p. Nếu khi $n \to \infty$ mà $p \to 0$ sao cho $np = \lambda = const$ thì:

$$\lim_{n\to\infty} P_n(k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad (1.4.1)$$

Từ đó ta có công thức xấp xỉ để tính xác suất "trong *n* lần trắc nghiệm sự kiện A xuất hiện *k* lần":

$$P_n(k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad (1.4.2)$$

Ở đây n là số lần quan sát, k là số lần xuất hiện hiện tượng, p là xác suất hiện hiện tượng, λ là trung bình số lần xuất hiện hiện tượng. Điều kiện ràng buộc là các lần trắc nghiệm đều phải thoả mãn tiêu chuẩn Bernoulli và xác suất xuất hiện hiện tượng phải khá nhỏ ($p << 1$). Trong trường hợp p khá gần với 1 ($p \approx 1$) thì thay cho việc xét sự kiện A là "sự kiện xuất hiện hiện tượng" ta xét sự kiện B là "sự kiện không xuất hiện hiện tượng" ($B=\overline{A}$).

Trong khí hậu, công thức này thường được ứng dụng để tính xác suất hiện sự kiện hiếm. Cũng cần nói rằng, thật khó có thể đưa ra được một định nghĩa chính xác khái niệm "sự kiện hiếm". Tuy nhiên, để có một khái niệm chung nhất ta có thể chấp nhận định nghĩa sau đây: "Sự kiện hiếm là sự kiện có xác suất xuất hiện rất nhỏ so với đơn vị". Tính mập mờ trong định nghĩa này là ở chỗ khái niệm "xác suất xuất hiện rất nhỏ" không được định lượng hoá một cách cụ thể; có thể xem đó là một khiếm khuyết buộc người sử dụng phải cân nhắc một cách kỹ lưỡng trên cơ sở những kiến thức chuyên môn của mình. Như vậy, khi nghiên cứu một hiện tượng nào đó trên các vùng địa lý khác nhau, có thể xảy ra trường hợp ở nơi này thì hiện tượng đang xét là hiện tượng hiếm nhưng ở nơi khác nó lại không còn là hiện tượng hiếm nữa.

Ví dụ 1.4 Giả sử ở điểm B trung bình hàng năm có 2 ngày sương muối. Tính xác suất hàng năm ở B có 0, 1, 2,..., 6 ngày có sương muối.

Ta thấy hiện tượng sương muối ở địa điểm B là một hiện tượng hiếm khi xuất hiện (bình quân một năm chỉ có 2 ngày, $\lambda=2$). Ta lập bảng tính sau đây:

Bảng 1.2. Xác suất xuất hiện sương muối

Số ngày (k)	0	1	2	3	4	5	6
$P_n(k) = e^{-2}\frac{2^k}{k!}$	0.14	0.27	0.27	0.18	0.09	0.04	0.01

Như vậy với các giá trị k lân cận $\lambda=2$ thì xác suất $P_n(k)$ lớn đáng kể, k càng nhỏ hoặc càng lớn hơn λ thì xác suất $P_n(k)$ càng giảm dần.

Có thể nhận thấy ở đây tính tương đối của khái niệm "sự kiện hiếm". Nếu quan niệm rằng tất cả các ngày trong năm đều quan trắc sương muối thì rõ ràng xác suất xuất hiện "hiện tượng sương muối" rất nhỏ ($2/365 \approx 0.0055$). Tuy nhiên, nếu tại địa điểm xét sương muối chỉ có thể xuất hiện vào những ngày chính đông (từ tháng 12 đến tháng 2 năm sau) thì việc quan trắc sương muối không phải được thực hiện ở tất cả các ngày trong năm mà chỉ trong 3 tháng chính đông (90 ngày). Trong trường hợp này xác suất xuất hiện hiện tượng lớn hơn đáng kể so với trường hợp trên ($2/90 \approx 0.02222$).

1.5 Đại lượng ngẫu nhiên và hàm phân bố xác suất

Khi nghiên cứu một hiện tượng nào đó ta cần tiến hành các phép thử, trong mỗi phép thử có thể nhận được các kết cục khác nhau. Chẳng hạn, kết quả của một lần quan trắc lượng mây có thể nhận một trong các tình huống "trời quang", "ít mây", "mây rải rác" hoặc "nhiều mây". Những tình huống như vậy đặc trưng về chất lượng cho phép thử, chúng chỉ mang tính chất định tính. Để đặc trưng định lượng cho phép thử người ta đưa vào khái hiệm đại lượng ngẫu nhiên.

Đại lượng ngẫu nhiên là đại lượng mà trong kết quả của phép thử, hay một lần thí nghiệm, nó nhận một và chỉ một giá trị từ tập những giá trị có thể, giá trị này hoàn toàn không thể đoán trước được.

Ví dụ, trong trường hợp quan trắc lượng mây trên đây, bầu trời có thể được chia làm 10 phần. Kết quả mỗi lần quan trắc giá trị của lượng mây chỉ có thể nhận một trong các trị số 0,1,...,10 (phần mười bầu trời) và ta chỉ có thể biết được giá trị này sau khi tiến hành quan trắc.

Người ta thường ký hiệu đại lượng ngẫu nhiên bởi các chữ cái in hoa X, Y, Z,..., còn các chữ cái in thường tương ứng x, y, z,... được dùng để chỉ các giá trị có thể của chúng. Đặc trưng có thể mô tả một cách đầy đủ đại lượng ngẫu nhiên là luật phân bố xác suất. Dạng tổng quát của luật phân bố của đại lượng ngẫu nhiên là hàm phân bố. Theo định nghĩa, hàm phân bố của đại lượng ngẫu nhiên X là hàm một biến F(x) được xác định bởi:

$$F(x) = P(X < x) \tag{1.5.1}$$

Trong đó $P(X < x)$ là xác suất để đại lượng ngẫu nhiên X nhận giá trị nhỏ hơn x. Người ta còn gọi F(x) là xác suất tích luỹ của X tại giá trị X=x. Hàm phân bố có các tính chất sau:

1) $0 \le F(x) \le 1$

2) $P(\alpha \le X < \beta) = F(\beta) - F(\alpha)$

3) Nếu $\alpha < \beta$ thì $F(\alpha) \le F(\beta)$

4) $\lim\limits_{x \to +\infty} F(x) = 1$ và $\lim\limits_{x \to -\infty} F(x) = 0$

Đồ thị hàm phân bố xác suất có dạng như trên hình 1.4a. Trong khí hậu tính chất 2) được ứng dụng để tính xác suất mà đại lượng khí hậu X nhận giá trị trong một khoảng (a_j,b_j) nào đó khi đã biết hàm phân bố F(x):

$$P(a_j \le X < b_j) = F(b_j) - F(a_j) \tag{1.5.2}$$

Người ta còn gọi $F(a_J)$ và $F(b_j)$ là xác suất tích luỹ của X tại a_j và b_j.

Từ (1.5.1) và tính chất 1) suy ra rằng:

$$P(X \geq x) = 1 - F(x) = \Phi(x) \tag{1.5.3}$$

Trong khí hậu $\Phi(x)$ được gọi là suất bảo đảm, tức là xác suất để X nhận giá trị vượt quá x. Đồ thị hàm suất bảo đảm có dạng như trên hình 1.4b. Nếu cho x nhận một giá trị a_j nào đó thì:

$$\Phi(a_j) = P(X \geq a_j) \tag{1.5.4}$$

Khi đã biết được F(x) ta dễ dàng suy ra được $\Phi(x)$, và như vậy, nếu cho trước suất bảo đảm $\Phi(x) = \alpha$ nào đó ta hoàn toàn có thể tính được x_α sao cho:

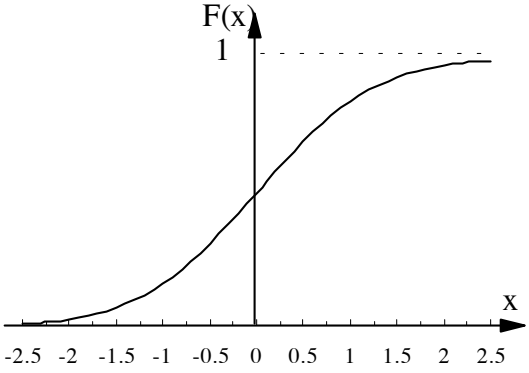
$$\Phi(x_\alpha) = P(X \geq x_\alpha) = \alpha \tag{1.5.5}$$

Kết hợp (1.5.3) và (1.5.5) ta cũng có thể tính được x_α, từ F(x) và α:

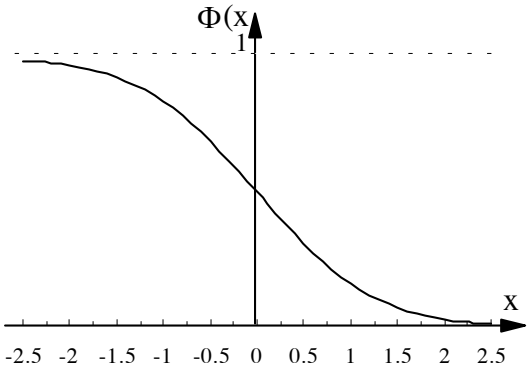
$$F(x_\alpha) = P(X < x_\alpha) = 1 - \alpha \tag{1.5.6}$$

Từ các tính chất 3) và 4) suy ra:

$$\lim_{x \to +\infty} \Phi(x) = 0 \text{ và } \lim_{x \to -\infty} \Phi(x) = 1 \tag{1.5.7}$$



Hình 1.4a Hàm phân bố xác suất



Hình 1.4b Hàm suất bảo đảm

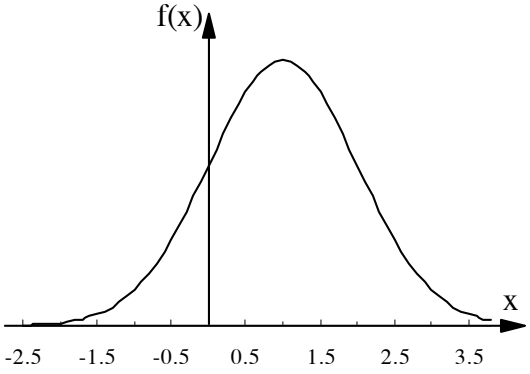
Hàm $f(x) = \dfrac{dF(x)}{dx}$ được gọi là hàm mật độ xác suất của X. Hàm f(x) có các tính chất:

1) $f(x) \geq 0$

2) $\int_{-\infty}^{+\infty} f(x)dx = 1$

3) $\int_{-\infty}^{x} f(x)dx = F(x)$

4) $\int_{\alpha}^{\beta} f(x)dx = P(\alpha \leq X < \beta)$



Hình 1.5 Hàm mật độ xác suất

1.6 Phân bố xác suất thực nghiệm

1.6.1 Xây dựng hàm phân bố thực nghiệm theo công thức kinh nghiệm.

Giả sử có chuỗi số liệu quan trắc $x_t = \{x_1, x_2, \dots, x_n\}$ của biến khí hậu X. Từ chuỗi số liệu này ta sắp xếp thành chuỗi tăng dần hay còn gọi là chuỗi trình tự $x_{(1)} \le \dots \le x_{(n)}$ rồi lập chuỗi xếp hạng $x_t^* = \{x_1^*, x_2^*, \dots, x_{n'}^*\}$, trong đó $x_1^* < x_2^* < \dots < x_{n'}^*$. Vì trong số n thành phần ban đầu của chuỗi $\{x_1, x_2, \dots, x_n\}$ có thể có những trị số bằng nhau nên số thành phần của chuỗi xếp hạng $\{x_1^*, x_2^*, \dots, x_{n'}^*\}$ có thể ít hơn n (n'≤n). Số thứ tự của các thành phần trong chuỗi xếp hạng được gọi là "hạng" và có thể nhận trị số thập phân. Ví dụ, sau khi sắp xếp chuỗi ban đầu theo trình tự tăng dần ta có các thành phần thứ 5 và thứ 6 có trị số bằng nhau, vậy hạng của các thành phần này sẽ là (5+6)/2=5,5 và $x_{5,5}^* = x_{(5)} = x_{(6)}$ (ở đây ký hiệu $x_{(t)}$, t=1..n, là các thành phần của chuỗi sau khi sắp xếp nhưng chưa xếp hạng).

Từ đó hàm phân bố xác suất thực nghiệm của X được xác định bởi:

$$F(x_m^*) = \frac{m}{n+1} \tag{1.6.1}$$

$$F(x_m^*) = \frac{m}{n} \tag{1.6.2}$$

$$F(x_m^*) = \frac{m-0.25}{n+0.55} \tag{1.6.3}$$

$$F(x_m^*) = \frac{m-0.3}{n+0.4} \tag{1.6.4}$$

Trong các công thức trên, x_m^* là giá trị của X ở vị trí thứ m trong chuỗi trình tự, m là số thứ tự (hạng) của x_m^*, n là dung lượng mẫu và $F(x_m^*)$ là tần suất tích luỹ tại x_m^*.

Thực chất công thức (1.6.1) là phép xấp xỉ $F(x_m^*) \approx M[F(x_m^*)]$, trong đó M là toán tử lấy kỳ vọng. Có nghĩa là trên thực tế ta chưa biết được $F(x_m^*)$ nhưng ta có thể xác định được kỳ vọng của nó:

$$M[F(x_m^*)] = \frac{m}{n+1}$$

Bởi vậy (1.6.1) thường được gọi là công thức kỳ vọng.

Công thức (1.6.2) được sử dụng khi biết tất cả các giá trị có thể của X, tức là khi n giá trị quan trắc của chuỗi ban đầu chứa đựng đầy đủ 100% lượng thông tin

của X. Tuy nhiên, trên thực tế dung lượng mẫu *n* của chuỗi là hữu hạn, thậm chí khá bé, do đó thay cho (1.6.2) thông thường người ta sử dụng các công thức (1.6.3) và (1.6.4), trong đó sự sai lệch do dung lượng mẫu bé đã được hiệu chỉnh.

Sau khi lựa chọn được công thức thích hợp ta tiến hành lập bảng tính sau:

m	1	2	...	n'
x^*_m	x^*_1	x^*_2	...	$x^*_{n'}$
$F(x^*_m)$	$F(x^*_1)$	$F(x^*_2)$	....	$F(x^*_{n'})$

Trên cơ sở đó hàm F(x) có thể được xây dựng bằng một trong hai cách sau đây:

1) Từ tập các cặp giá trị (x^*_m, $F(x^*_m)$), m=1,2,...,n', xác định dạng hàm giải tích G(x) biểu diễn mối phụ thuộc hàm giữa $F(x^*_m)$ và x^*_m, sau đó tiến hành xấp xỉ $F(x) \approx G(x)$ bằng phương pháp bình phương tối thiểu.
2) Dựng đồ thị biểu diễn mối phụ thuộc hàm giữa $F(x^*_m)$ và x^*_m bằng cách chọn trục hoành là x^*_m, trục tung là $F(x^*_m)$. Đồ thị đó chính là sự xấp xỉ hàm F(x).

Ngoài việc xác định hàm phân bố thực nghiệm trên đây đôi khi người ta còn xây dựng hàm suất bảo đảm hay đường cong bảo đảm Φ(x). Muốn vậy, thay vì sắp xếp chuỗi ban đầu theo thứ tự tăng dần ta chỉ việc sắp xếp nó theo thứ tự giảm dần và trong các công thức (1.6.1) – (1.6.4) hàm $\Phi(x^*_m)$ sẽ đóng vai trò của hàm $F(x^*_m)$.

Phương pháp trên đây thường được áp dụng trong trường hợp dung lượng mẫu của chuỗi tương đối nhỏ. Khi dung lượng mẫu đủ lớn người ta thường dùng phương pháp phân nhóm.

Ví dụ 1.6.1. Số liệu lịch sử nhiệt độ trung bình năm (X) của một trạm sau khi đã sắp xếp theo thứ tự tăng dần được trình bày trong bảng sau:

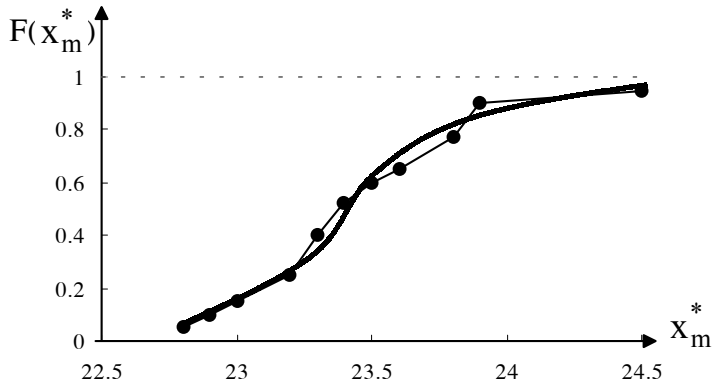
STT	1	2	3	4	5	6	7	8	9	10
X	22.8	22.9	23.0	23.2	23.2	23.2	23.3	23.3	23.3	23.4
STT	11	12	13	14	15	16	17	18	19	
X	23.4	23.5	23.6	23.8	23.8	23.8	23.8	23.9	24.5	

Từ bảng số liệu này, sau khi xếp hạng và sử dụng các công thức (1.6.1) – (1.6.4) để tính toán ta có kết quả được trình bày trong bảng 1.3, trong đó dung lượng mẫu n = 19. Khi so sánh kết quả tính theo các công thức khác nhau có thể thấy trị số của tần suất tích luỹ nói chung chênh lệch nhau không nhiều lắm. Tuy nhiên, nếu dung lượng mẫu n càng giảm thì sự sai khác giữa chúng có thể sẽ lớn đáng kể.

Hình 1.6 dẫn ra đồ thị đường tần suất tích luỹ ứng với công thức (1.6.1).

Bảng 1.3. Tần suất tích luỹ tính theo các công thức khác nhau.

x_m^*	m	Công thức tính			
		(1.6.1)	(1.6.2)	(1.6.3)	(1.6.4)
22.8	1	0.05	0.05	0.04	0.04
22.9	2	0.1	0.11	0.09	0.09
23.0	3	0.15	0.16	0.14	0.14
23.2	5	0.25	0.26	0.24	0.24
23.3	8	0.4	0.42	0.4	0.4
23.4	10.5	0.53	0.55	0.52	0.53
23.5	12	0.6	0.63	0.6	0.6
23.6	13	0.65	0.68	0.65	0.65
23.8	15.5	0.78	0.82	0.78	0.78
23.9	18	0.9	0.95	0.91	0.91
24.5	19	0.95	1	0.96	0.96



Hình 1.6 Đường tần suất tích luỹ nhiệt độ trung bình năm (tính theo công thức kỳ vọng)

1.6.2 Phương pháp phân nhóm xây dựng hàm phân bố thực nghiệm

1.6.2.1 Chỉ tiêu xác định số nhóm

Trong nghiên cứu khí tượng, khí hậu người ta thường sử dụng 3 dạng phân nhóm sau đây:

1) Nhóm định lượng số với cự ly các nhóm bằng nhau.
2) Nhóm định lượng số với cự ly các nhóm không bằng nhau.
3) Nhóm định tính được mô tả bằng lời.

Ví dụ sau đây cho ta thấy rõ ý nghĩa của ba loại nhóm trên:

	Nhóm loại 1	Nhóm loại 2	Nhóm loại 3
STT nhóm	Nhiệt độ TB năm (°C)	Lượng mưa tháng (mm)	Cấp tốc độ gió
1	14.1–16	0–50	Lặng gió
2	16.1–18	50–70	Gió yếu
...	...	...	...
N	28.1–30	300–350	Gió rất mạnh

Tuỳ theo từng đặc trưng yếu tố khí hậu và mục đích cụ thể của vấn đề cần xem xét mà loại nhóm nào sẽ được chọn để sử dụng cho phù hợp. Trong ví dụ trên, nhiệt độ thường được chia theo nhóm loại 1 (khoảng cách các nhóm đều nhau), lượng mưa được chia theo nhóm loại 2 và tốc độ gió có thể được chọn kiểu chia thứ 3. Tuy nhiên trong thực tế có thể xảy ra trường hợp để tiện tính toán trên máy tính điện tử người ta chỉ sử dụng cách chia nhóm loại 1. Khi đó đối với yếu tố tốc độ gió người ta có thể phân khoảng tương ứng với các qui ước "gió yếu", "gió mạnh",...

Số lượng nhóm được chia nói chung phụ thuộc vào dung lượng mẫu. Người ta thường sử dụng các chỉ tiêu sau đây để xác định số nhóm sẽ chia:

1) $$N \approx 5\lg n \tag{1.6.5}$$

2) $$N \approx \frac{x_{max} - x_{min}}{1 + 3.222 \lg n} \tag{1.6.6}$$

Trong đó N là số nhóm, lg$n$ là lôgarit cơ số 10 của n, x_{max}, x_{min} là giá trị lớn nhất và nhỏ nhất của chuỗi số liệu.

Ví dụ 1.6.2. Với các dung lượng mẫu khác nhau khi sử dụng chỉ tiêu (1.6.5) ta nhận được số nhóm tương ứng như sau:

Dung lượng mẫu (n)	50	100	500	1000	10000
Số nhóm được chia (N)	8	10	13	15	20

Nhiều khi thay cho các cách phân nhóm trên đây người ta còn sử dụng một số cách phân nhóm khác:

1) Phân nhóm theo giá trị độ lệch bình phương trung bình σ:

$(-\infty; \bar{x} - 3\sigma)$, $(\bar{x} - 3\sigma; \bar{x} - 2\sigma)$, $(\bar{x} - 2\sigma; \bar{x} - \sigma)$, $(\bar{x} - \sigma; \bar{x})$,

$(\bar{x}; \bar{x} + \sigma)$, $(\bar{x} + \sigma; \bar{x} + 2\sigma)$, $(\bar{x} + 2\sigma; \bar{x} + 3\sigma)$, $(\bar{x} + 3\sigma; +\infty)$.

Theo cách này số nhóm được chia có tất cả là 8 nhóm.

2) Cũng tương tự như trên nhưng khoảng cách nhóm được tính theo 0.5σ. Trong trường hợp này ta có tất cả 14 nhóm:

$(-\infty; \bar{x} - 3\sigma)$, $(\bar{x} - 3\sigma; \bar{x} - 2.5\sigma)$,..., $(\bar{x} + 2.5\sigma; \bar{x} + 3\sigma)$, $(\bar{x} + 3\sigma; +\infty)$

Ngoài ra còn có một số cách phân nhóm khác nhưng không được sử dụng phổ biến.

I.6.2.2 Tần số, tần suất, tần suất tích luỹ

Giả sử ta có chuỗi số liệu $\{x_t,\ t=1,2,...,n\}$. Chuỗi được chia thành N nhóm (N<n):

$$\{(a_1,b_1),\ (a_2,b_2),...,\ (a_N,b_N)\}=\{(a_j,b_j),\ j=1..N\},$$

trong đó $b_j=a_{j+1}$ và $a_1\le\min\{x_t,\ t=1..n\}$, $b_N>\max\{x_t,\ t=1..n\}$. Không mất tính tổng quát ta giả thiết rằng các nhóm có cự ly bằng nhau và bằng $\Delta x=b_j-a_j$.

Ta gọi tần số của nhóm thứ j là *số thành phần* của chuỗi thoả mãn điều kiện $a_j\le x_t<b_j$ và ký hiệu bằng m_j. Khi đó tần suất p_j của nhóm thứ j được xác định bởi:

$$p_j=\frac{m_j}{n} \tag{1.6.7}$$

hoặc dưới dạng %:

$$p_j=\frac{m_j}{n}.100\% \tag{1.6.7'}$$

Tỷ số $\omega_j=\dfrac{p_j}{\Delta x}$ được gọi là mật độ xác suất ứng với nhóm thứ j.

Rõ ràng ta có các quan hệ sau:

$$\sum_{j=1}^{N}m_j=n\ ,\ \sum_{j=1}^{N}p_j=1\quad\text{và}\ \sum_{j=1}^{N}\omega_j\Delta x=1 \tag{1.6.8}$$

Nếu $\Delta x=1$ thì $\omega_j=p_j$ có thể nhận thấy rằng hệ thức cuối cùng trong (1.6.8) tương đương với tính chất 2) của hàm mật độ đã được trình bày trên đây.

Trong ứng dụng thực hành người ta thường biểu diễn bằng đồ thị đường tần số hoặc biểu đồ tần suất lên mặt phẳng toạ độ với trục tung là tần số m_j (hình 1.7) hoặc tần suất p_j (hình 1.8) còn trục hoành là giá trị các nhóm của x. Đường tần suất được xây dựng trên cơ sở biểu đồ tần suất. Đường tần suất được vẽ sao cho trơn tru và phải cố gắng đi sát các trung điểm phía trên của các cột biểu đồ tần suất.

Nếu trục tung là ω_j thì đồ thị nhận được là đường biểu diễn hàm mật độ xác suất thực nghiệm.

Tần suất tích luỹ F_j là đại lượng được xác định bởi:

$$F_j=\sum_{i=1}^{j}p_i\equiv P(x_t<b_j),\ (j=1,2,...,N) \tag{1.6.9}$$

Từ đó ta có: $F_1=p_1$, $F_2=p_1+p_2$,..., $F_N=1$. Hay $F_j=P(x_t<b_j)$, j=1..N.

Trên cơ sở đó, tần suất tích luỹ cũng có thể biểu diễn lên biểu đồ để từ đó xây dựng đồ thị (hình 1.9). Đồ thị tần suất tích luỹ được vẽ sao cho trơn tru và đi qua giới hạn trên các nhóm.

Như vậy, khi $\Delta x=1$ thì đường tần suất chính là ước lượng của hàm mật độ còn đường tần suất tích luỹ là ước lượng của hàm phân bố xác suất. Ta sẽ gọi đường tần suất tích luỹ là phân bố xác suất thực nghiệm và có thể biểu diễn nó dưới dạng:

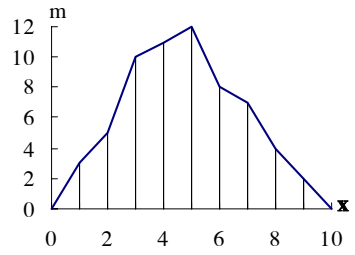
$$F(x) = \begin{cases} 0 & \text{Nếu } x \le \min\{x_t,\ t=1..n\} \\ \dfrac{m}{n} & \text{Nếu có m phần tử trong mẫu bé hơn x} \\ 1 & \text{Nếu } x > \max\{x_t,\ t=1..n\} \end{cases}$$

Từ các giá trị tần số nhóm tính toán theo (1.6.7) hoặc (1.6.7'), suất bảo đảm sẽ được xác định như sau:

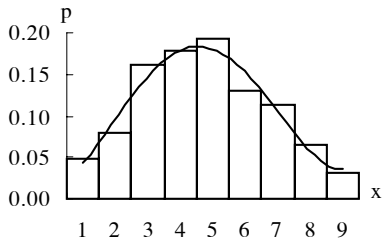
$$\Phi_j = \sum_{i=j}^{N} p_i, (j = \overline{1, N}) \qquad (1.6.10)$$

Căn cứ vào kết quả tính toán này, ta sẽ xây dựng được đường cong bảo đảm (hình 1.10) và qua đó có thể xác định được trị số của đại lượng khí hậu x_{Φ_j} ứng với các suất bảo đảm Φ_j khác nhau:

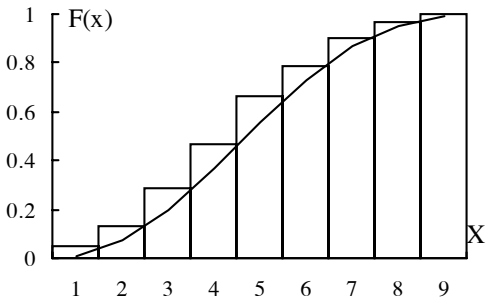
$$P(x \ge x_{\Phi_j}) = \Phi_j \qquad (1.6.11)$$



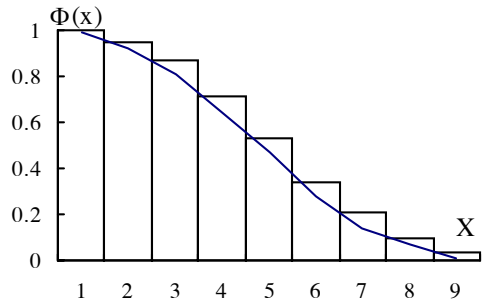
Hình 1.7 Đường tần số



Hình 1.8 Đường tần suất



Hình 1.9 Đường tần suất tích luỹ



Hình 1.10 Đường suất bảo đảm

Ví dụ 1.6.3 Từ chuỗi số liệu lượng mưa tháng 2 của một trạm, sau khi tiến hành khảo sát sơ bộ ta có: n = 97 (năm), x_{min} = 1.4 (mm), x_{max} = 95.0 (mm). Sử dụng

công thức phân nhóm (1.6.5) ta được số nhóm cần chia là N = 5lg(97)≈10 (nhóm). Như vậy có thể chọn giới hạn dưới a_1=0 và b_N=100. Để đơn giản khi tính toán cự ly nhóm được xem là không đổi và bằng 10. Kết quả tính toán trình bày trong bảng 1.4.

Bảng 1.4 Tần suất tích luỹ lượng mưa tính theo phương pháp phân nhóm

j	a_j	b_j	m_j	Σm_j	F_j
0		0	0	0	0
1	0	10	20	20	0.21
2	10	20	26	46	0.47
3	20	30	23	69	0.71
4	30	40	8	77	0.79
5	40	50	8	85	0.88
6	50	60	4	89	0.92
7	60	70	5	94	0.97
8	70	80	1	95	0.98
9	80	90	1	96	0.99
10	90	100	1	97	1

Đồ thị hàm phân bố thực nghiệm được trình bày trên hình 1.11. Từ đó, ta có thể tính:

– Xác suất để lượng mưa tháng 2 (X) nhận giá trị trong khoảng (a_j,b_j)

– Suất bảo đảm mà lượng mưa tháng 2 vượt quá giá trị a_j

– Giá trị của lượng mưa tháng 2 ứng với suất bảo đảm Φ_j cho trước.

Ví dụ: Với a_j = 30 (mm), b_j = 50 (mm), Φ_j = 20 (%), để tính xác suất mà X nhận giá trị trong khoảng (a_j,b_j): $P(a_j \le X < b_j) = P(30 \le X < 50)$, ta tiến hành bằng cách sau đây:

Trên trục hoành lấy các điểm tương ứng với các giá trị của x bằng 30 và 50 (mm); từ các điểm này kẻ song song với trục tung cho đến khi cắt đồ thị F(x) thì kẻ song song với trục hoành. Các đường này cắt trục tung tại các điểm tương ứng với các giá trị của F(30), F(50). Hiệu của chúng chính là xác suất để X nhận giá trị trong khoảng (30;50):

$$P(30;50)=P(30 \le X < 50)=F(50)-F(30) \approx 0.876-0.711=0.165 =16.5\%$$

Việc xác định suất bảo đảm cũng tương tự như vậy. Ta có:

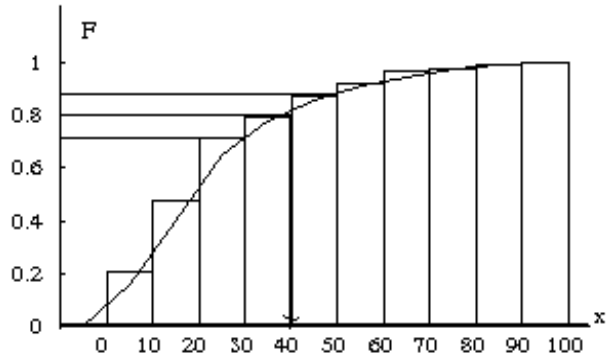
$$\Phi(30) = P(X \ge 30) \approx 1-0.711=0.289 = 28.9\%$$

Trong trường hợp này, nếu thay vì xây dựng đường cong phân bố trên đây ta có đường cong bảo đảm thì vấn đề trở nên đơn giản hơn.

Để tính x_{Φ_j} ta hãy chọn một điểm trên trục tung mà ở đó giá trị của $F(x)=1-\Phi_j$ rồi kẻ song song với trục hoành cho đến khi cắt đồ thị F(x) thì kẻ song song với trục tung. Giá trị đọc được trên trục hoành tại điểm cắt chính là x_{Φ_j} phải tìm. Vậy, lượng mưa ứng với suất bảo đảm Φ=20% sẽ là:

$$x_{80} = x(\Phi=20\%) = x(F=80\%) \approx 40 \text{ mm}$$

Chú ý rằng, nếu hàm phân bố F(x) được xác định dưới dạng giải tích thì các đặc trưng trên đây sẽ được tính theo các công thức (1.5.2), (1.5.3) và (1.5.5) đã nêu.



Hình 1.11 Hàm phân bố thực nghiệm lượng mưa tháng 2

1.6.2.3 Phân bố tần suất các đại lượng vectơ

Các khái niệm tần suất đã được trình bày trên đây nói chung thường được áp dụng đối với những đặc trưng yếu tố khí hậu là những đại lượng vô hướng như nhiệt độ, lượng mưa,... Tuy nhiên, trong nghiên cứu khí hậu, nhiều khi người ta còn xét các đại lượng vectơ như vận tốc gió (gồm tốc độ và hướng), hoặc các đại lượng vô hướng biến đổi theo thời gian (mùa, tháng) được biểu diễn như là những đại lượng vectơ. Trong trường hợp này khái niệm tần suất được mở rộng cho không gian nhiều chiều. Chẳng hạn, tần suất vận tốc gió sẽ được xét trong không gian hai chiều, một chiều là hướng gió và một chiều là tốc độ gió. Để tiện trình bày ta hãy xét ví dụ sau đây.

Giả sử để nghiên cứu đặc trưng vận tốc gió của một địa điểm nào đó ta sẽ tiến hành thống kê số liệu khí hậu về đặc trưng này bằng cách lập bảng phân bố tần số xuất hiện tốc độ gió theo các hướng. Giả sử số trường hợp (dung lượng mẫu) được nghiên cứu bằng 445. Tương ứng với mỗi cấp tốc độ gió và hướng gió ta có một giá trị tần số xuất hiện (bảng 1.5). Chẳng hạn, tần số xuất hiện gió hướng Bắc (N) ở cấp từ 2.1–4 m/s là 16.

Từ bảng số liệu này ta thành lập bảng phân bố tần suất (bảng 1.6) theo nguyên tắc lấy giá trị các ô trong bảng 1.5 (số trường hợp – tần số) chia cho tổng số trường hợp (445).

Bảng 1.5 Phân bố tốc độ gió theo các hướng

Khoảng tốc độ	Hướng gió								Tổng
Gió (m/s)	S	SW	W	NW	N	NE	E	SE	
0–2.0	0	0	2	0	1	0	1	0	4
2.1–4	6	8	2	0	16	13	17	2	64
4.1–6	11	12	5	4	16	8	15	7	78
6.1–8	11	16	10	14	21	7	6	2	87
8.1–10	5	8	9	22	8	1	5	5	63
...	....	....	....	....	....	....	....	....	....
Tổng	35	51	50	138	76	29	47	19	445

Bảng 1.6 Bảng phân bố tần suất (%) tốc độ gió theo các hướng

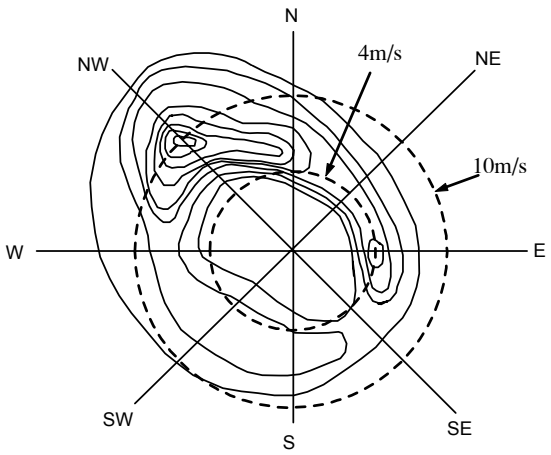
Khoảng tốc độ gió	Hướng gió								Tổng
(m/s)	S	SW	W	NW	N	NE	E	SE	
0–2.0	0.0	0.0	0.4	0.0	0.2	0.0	0.2	0.0	0.9
2.1–4	1.3	1.8	0.4	0.0	3.6	2.9	3.8	0.4	14.4
4.1–6	2.5	2.7	1.1	0.9	3.6	1.8	3.4	1.6	17.5
6.1–8	2.5	3.6	2.2	3.1	4.7	1.6	1.3	0.4	19.6
8.1–10	1.1	1.8	2.0	4.9	1.8	0.2	1.1	1.1	14.2
...									
Tổng	7.9	11.5	11.2	31.0	17.1	6.5	10.6	4.3	100.0

Việc biểu diễn hình học phân bố tần suất có thể được thực hiện theo nhiều cách. Một trong những cách đó là sử dụng mặt toạ độ cực, rất phù hợp đối với đại lượng vận tốc gió:

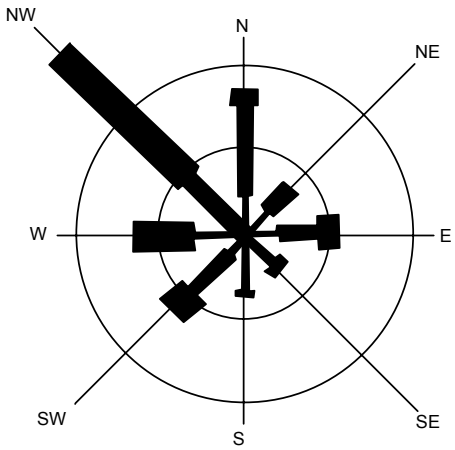
Trên 8 bán trục biểu diễn 8 hướng gió, chia thang độ đo ứng với các cấp tốc độ gió khác nhau, ví dụ các vạch chia ứng với 0–2 m/s, 2.1–4 m/s,... Tại mỗi một điểm đại diện cho các cấp tốc độ trên các hướng ta điền giá trị tần suất tốc độ gió tương ứng được cho trong bảng 1.6. Sau đó vẽ các đường đẳng tần suất (hình 1.12).

Trong trường hợp các cấp tốc độ gió được chia theo kiểu nhóm định tính (gió yếu, gió mạnh,...) ta cũng có thể tiến hành tính tần suất theo cách tương tự như trên. Nhưng khi biểu diễn hình học, để đơn giản người ta chọn ký hiệu cho các cấp tốc độ gió tương ứng. Ví dụ, từ số liệu trong bảng 1.6 ta sẽ chọn biểu diễn nét mảnh cho cấp gió yếu (≤4m/s), nét đậm vừa ứng với cấp gió trung bình (4–10m/s) và nét đậm cho cấp gió mạnh (>10m/s). Mặt khác, trên các bán trục biểu diễn hướng gió, thang chia độ được chia theo tần suất, ví dụ các vòng tròn đồng tâm có bán kính cách nhau một khoảng ứng với mức tần suất 10%. Sau đó căn cứ vào các giá trị tần

suất đã tính toán ta dễ dàng tiến hành vẽ "hoa gió". Hình 1.13 minh hoạ cho trường hợp biểu diễn này, trong đó cấp tốc độ gió được biểu thị bởi các thanh đậm nhạt khác nhau, còn độ dài các thanh chỉ giá trị tần suất tương ứng.



Hình 1.12 Tần suất tốc độ gió



Hình 1.13 Hoa gió

Với các đại lượng khí hậu vô hướng như nhiệt độ, lượng mưa,... xét theo sự biến thiên của thời gian (tháng, mùa), việc biểu diễn sự phân bố tần suất cũng được tiến hành một cách tương tự, tuy nhiên yếu tố thời gian sẽ thay cho hướng gió trong trường hợp trên đây. Hơn nữa, khi biểu diễn hình học, thay cho hệ toạ độ cực, ta sẽ dùng hệ toạ độ vuông góc.

Hình 1.14 dẫn ra một ví dụ về cách biểu diễn tần suất nhiệt độ tối cao tháng trên hệ toạ độ vuông góc. Ở đây, trục hoành biểu thị các tháng trong năm, còn trục tung là giá trị nhiệt độ tối cao tháng. Các đường đẳng trị trên hình vẽ tương ứng với các mức xác suất (tần suất) cách nhau 10%, ngoại trừ đường 5%. Cách biểu diễn này có thể cho ta một bức tranh khái quát về phân bố nhiệt độ tối cao trong năm. Chẳng hạn, với mức xác suất trên 50%, trong các tháng 5,6,7,8 nhiệt độ tối cao trung bình có thể đạt tới 38 – 39^0C.

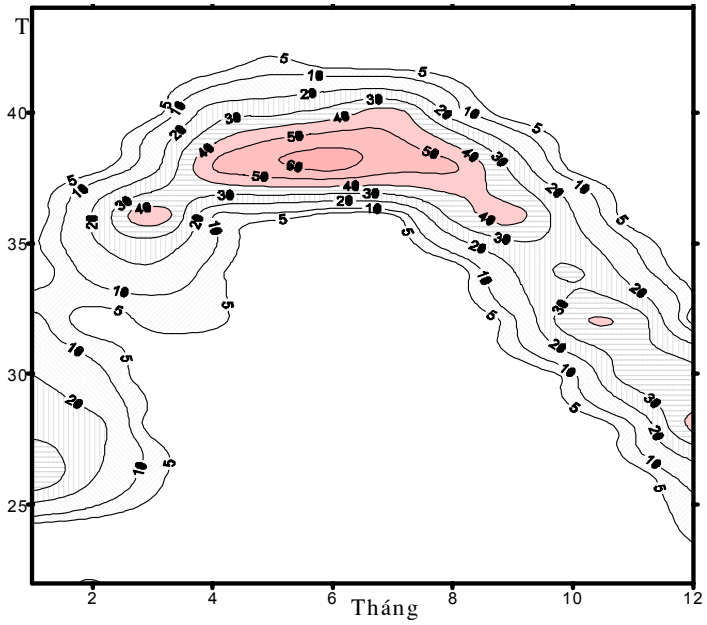
1.7 Phân bố Gumbell và các đại lượng khí hậu cực trị

Trong nghiên cứu khí hậu nhiều khi người ta thường quan tâm đến các đặc trưng của các đại lượng khí hậu cực trị – cực đại và cực tiểu. Khái niệm cực trị ở đây được hiểu theo nghĩa là giá trị lớn nhất hay nhỏ nhất trong một khoảng thời gian quan trắc nào đó. Chẳng hạn nhiệt độ lớn nhất (nhỏ nhất) trong một ngày, một tháng hoặc một năm. Tập hợp các giá trị này lập thành một không gian mẫu. Lý thuyết về phân bố của các đại lượng cực trị đã chỉ ra rằng, nếu các quan trắc là độc lập và có cùng phân bố xác suất $F(x)=P(x_i<x)$, i=1..n, thì phân bố xác suất của các đại lượng cực trị sẽ có dạng sau:

– Đối với các đại lượng cực đại: $P\{x^{(n)}<x\}=[F(x)]^n$

– Đối với các đại lượng cực tiểu: $P\{x_{(n)}<x\}=1-[1-F(x)]^n$,

ở đây, n là dung lượng mẫu.



Hình 1.14 Phân bố tần suất nhiệt độ tối cao tháng

Năm 1928, Fisher và Tippett đã tìm ra 3 dạng phân bố có thể áp dụng cho bất kỳ đại lượng ngẫu nhiên cực trị X nào sau đây:

Dạng I: $$F(y)=\text{Exp}(-e^{-y}), \qquad (1.7.1)$$

Dạng II: $$F(y)= \text{Exp}(-y^{-\delta}) \qquad (1.7.2)$$

Dạng III: $$F(y) = \text{Exp}(-(-y^{\delta})) \qquad (1.7.3)$$

Trong đó dạng I là phân bố giới hạn của dạng II và dạng III và đã được Gumbell (1958) khảo sát một cách tỉ mỉ, nên nó mang tên gọi là phân bố Gumbell.

Phân bố Gumbell (hay phân bố Fisher – Tippett dạng I) được ứng dụng hết sức rộng rãi trong khí hậu. Hàm phấn bố (1.7.1) là hàm của biến chuẩn hoá y:

$$y = \frac{1.283}{s_x}(x - \bar{x}) + 0.577 \qquad (1.7.4)$$

với $$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\ , \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

x là biến ngẫu nhiên nhận giá trị cực trị của X, x_i là những giá trị mẫu của x.

Người ta đã chứng minh được rằng, khi dung lượng mẫu $n \to \infty$ thì:

M[y] = y = 0.57721, và được gọi là hằng số Euler,

$D[y] = \sigma_z^2 = \frac{\pi^2}{6}$, từ đó σ_y = 1.28254.

Như vậy, các hằng số trong (1.7.4) chính là kỳ vọng và độ lệch bình phương trung bình của đại lượng ngẫu nhiên y.

Thông thường phân bố Gumbell được dùng để tính các giá trị cực đại hoặc cực tiểu của đặc trưng yếu tố khí hậu. Xác suất để đại lượng khí hậu cực đại x nhận giá trị vượt quá một giá trị x_0 nào đó được xác định bởi:

$$p_M = P(x \geq x_o) = 1 - Exp(-e^{-y}) \quad (1.7.5)$$

Và xác suất để đại lượng khí hậu cực tiểu x nhận giá trị nhỏ hơn x_o sẽ là:

$$p_m = P(x < x_o) = Exp(-e^{-y}) \quad (1.7.6)$$

1.8 Thời gian lặp lại hiện tượng.

Thời gian lặp lại hiện tượng (ký hiệu bằng T) là khoảng thời gian cần thiết để hiện tượng có thể xuất hiện trở lại sau khi đã xuất hiện. Trong khí hậu, tuỳ theo các khái niệm xác suất được dùng mà thời gian lặp lại T có những ý nghĩa khác nhau:

- Nếu *p* là tần suất xuất hiện hiện tượng khí hậu A nào đó thì:

$$T = \frac{1}{p} \quad (1.8.1)$$

là khoảng thời gian cần thiết để lặp lại hiện tượng A.

- Nếu $P(a_j,b_j)$ là xác suất các khoảng trị số thì T là thời gian cần thiết để đại lượng khí hậu X nhận giá trị trong khoảng (a_j,b_j):

$$T = \frac{1}{P(a_j, b_j)} \quad (1.8.2)$$

- Nếu khái niệm xác suất là suất bảo đảm $\Phi(a_j)$ thì T là thời gian cần thiết để đại lượng khí hậu X nhận giá trị không nhỏ hơn a_j:

$$T = \frac{1}{\phi(a_j)} \quad (1.8.3)$$

Nói chung có thể hiểu T là chu kỳ lặp lại của hiện tượng khí hậu. Cần chú ý rằng:

– Nói chung các thành phần kế cận của chuỗi số liệu khí hậu thường cách nhau một năm nên đơn vị đo của T là năm.

– Giá trị của T được xét trên phương diện thống kê, do đó không nhất thiết là cứ T năm hiện tượng tương ứng đều phải xảy ra.

Ví dụ 1.8 Suất bảo đảm để nhiệt độ trung bình tháng 7 ở một địa điểm vượt quá 32°C là 0.04, tức là $\Phi(32) = P(t \geq 32) = 0.04$. Vậy $T = 1/0.04 = 25$ (năm). Ở đây t là nhiệt độ, T là thời gian lặp lại. Giá trị T = 25 năm có nghĩa là nếu hiện tượng $t \geq 32$ đã xuất hiện thì nó có thể lặp lại sau 25 năm nữa. Tuy nhiên không nhất thiết đúng 25 năm sau hiện tượng mới xuất hiện trở lại.

1.9. Một số bài toán về các đại lượng cực trị

Bài toán 1.9.1. Giả sử x là đại lượng khí hậu cực đại có trung bình số học bằng $\bar{x}$ và độ lệch chuẩn bằng s_x. Hãy xác định giá trị x_o mà x có thể vượt quá nó ứng với chu kỳ lặp lại bằng T cho trước.

Giải:

Xác suất để x nhận giá trị vượt quá x_o được xác định bởi công thức (1.7.5). Tương ứng với xác suất này ta có chu kỳ lặp lại là $T = 1/p_M$. Vì T cho trước nên $p_M = 1/T$. Ta có:

$$\frac{1}{T} = 1 - \text{Exp}(-e^{-y}) \Rightarrow \text{Exp}(-e^{-y}) = \frac{T-1}{T}$$

Lấy lôgarit hai lần và chú ý đến dấu của biểu thức ta nhận được:

$$y = -\ln(\ln(\frac{T}{T-1}))$$

Thay giá trị của biểu thức $y = \frac{1.283}{s_x}(x - \bar{x}) + 0.577$ vào và thực hiện một vài phép biến đổi đơn giản, cuối cùng ta có:

$$x_o = \bar{x} + \frac{s_x}{1.283}(-\ln(\ln(\frac{T}{T-1})) - 0.577) \qquad (1.9.1)$$

Ví dụ 1.9.1 Nhiệt độ tối cao tuyệt đối ở một địa điểm trung bình là t_x=40.0°C và độ lệch chuẩn là s_t=1.0°C. Hãy xác định giá trị nhiệt độ tối cao tuyệt đối T_{xo} ở đây ứng với các chu kỳ lặp lại T bằng 10, 20, 30, 50, 100, 150 năm.

Sử dụng công thức (1.9.1) ta được kết quả tính sau:

Bảng 1.7 Nhiệt độ tối cao tuyệt đối ứng với các chu kỳ lặp lại

T (năm)	10	20	30	50	100	150
T_{xo} (°C)	41.3	41.9	42.2	42.6	43.1	43.5

Có thể nhận thấy rằng, khi chu kỳ lặp lại T tăng lên thì giá trị T_{xo} cũng tăng lên. Tuy nhiên sự tăng lên của T_{xo} là có giới hạn.

Bài toán 1.9.2. Giả sử x là một đại lượng khí hậu cực tiểu có trung bình số học bằng $\bar{x}$ và độ lệch chuẩn bằng s_x. Hãy xác định giá trị x_0 mà x nhỏ hơn nó ứng với chu kỳ lặp lại bằng T cho trước.

Giải:

Xác suất để x nhận giá trị nhỏ hơn x_0 được xác định bởi công thức (1.7.6). Tương ứng với xác suất này ta có chu kỳ lặp lại $T = 1/p_m$. Vì T cho trước nên $p_m = 1/T$. Ta có:

$$\frac{1}{T} = \mathrm{Exp}(-e^{-y})$$

Lấy lôgarit haio lần và chú ý đến dấu của biểu thức ta nhận được:

$$y = -\ln(\ln(T))$$

Thay giá trị của biểu thức $y = \frac{1.283}{s_x}\left(x_0 - \bar{x}\right) + 0.577$ vào và thực hiện một vài phép biến đổi đơn giản, cuối cùng ta có:

$$x_0 = \bar{x} - \frac{s_x}{1.283}(\ln(\ln(T)) + 0.577) \qquad (1.9.2)$$

Ví dụ 1.9.2. Nhiệt độ tối thấp tuyệt đối ở một địa điểm trung bình là $t_m = 15^0$C và độ lệch chuẩn bằng 2^0C. Hãy xác định nhiệt độ tối thấp tuyệt đối t_{m0} của địa điểm này ứng với các chu kỳ lặp lại T bằng 10, 20, 30, 50, 100 và 150 năm.

Sử dụng công thức (1.9.2) ta có kết quả tính toán sau:

Bảng 1.8 Nhiệt độ tối thấp tuyệt đối ứng với các chu kỳ lặp lại

T (năm)	10	20	30	50	100	150
$t_{m0}(^0C)$	12.8	12.4	12.2	12.0	11.7	11.6

Cũng tương tự như trường hợp nhiệt độ cực đại, ở đây khi T càng tăng thì t_{m0} càng giảm, nhưng mức độ suy giảm chậm dần và có giới hạn nhất định.

Bài toán 1.9.3. Giả sử đại lượng khí hậu cực đại x có trung bình số học bằng $\bar{x}$ và độ lệch chuẩn bằng s_x. Hãy xác định thời gian cần thiết T để x nhận giá trị vượt quá giá trị x_o cho trước.

Giải:

Thời gian cần thiết để đại lượng khí hậu cực đại x nhận giá trị vượt quá x_o cho trước chính là chu kỳ lặp lại T. Ta có:

$$T = \frac{1}{p_M} = \frac{1}{1 - \mathrm{Exp}(-e^{-y})} \text{ với } y = \frac{1.283}{s_x}(x_o - \bar{x}) + 0.577 \qquad (1.9.3)$$

Ví dụ 1.9.3. Nhiệt độ không khí tối cao tuyệt đối ở một địa điểm trung bình bằng 40.0^0C và độ lệch chuẩn bằng 1^0C. Hỏi sau bao nhiêu năm mới có một lần xảy

ra hiện tượng nhiệt độ tối cao tuyệt đối ở đây vượt quá 40.5^0C, 41.0^0C, 41.5^0C, 42.0^0C, 42.5^0C, 43.5^0C.

Sử dụng công thức (1.9.3) ta có kết quả tính sau:

Bảng 1.9 Chu kỳ lặp lại ứng với các trị số nhiệt độ cực đại tuyệt đối

$t_{m0}(^0C)$	40	40.5	41	41.5	42	42.5	43	43.5
T(năm)	2	4	7	13	24	45	84	159

Ta lại thấy rằng, cũng với những trị số chênh lệch như nhau (0.50C) của nhiệt độ tối cao tuyệt đối, nhưng khi giá trị nhiệt độ càng thấp thì sự khác biệt của khoảng thời gian lặp lại càng bé hơn rất nhiều lần so với khi giá trị nhiệt độ cao.

Bài toán 1.9.4. Giải sử x là đại lượng khí hậu cực tiểu có trung bình số học bằng $\bar{x}$ và độ lệch chuẩn bằng s_x. Hãy xác định thời gian cần thiết T để x nhận giá trị nhỏ hơn giá trị x_0 cho trước.

Giải:

Thời gian cần thiết để đại lượng khí hậu cực tiểu x nhận giá trị nhỏ hơn x_0 cho trước chính là chu kỳ lặp lại T. Ta có:

$$T = \frac{1}{p_m} = \frac{1}{Exp(-e^{-y})} \text{ với } y = \frac{1.283}{s_x}(x_0 - \bar{x}) + 0.577 \qquad (1.9.4)$$

1.10 Toán đồ xác suất

Một trong những nhiệm vụ quan trọng của nghiên cứu khí hậu là xác định qui luật biến đổi theo không gian và thời gian của các đặc trưng yếu tố khí hậu. Khi nghiên cứu qui luật biến đổi theo không gian chúng ta không chỉ sử dụng số liệu ở một trạm mà ở nhiều trạm khác nhau, tức là ta cần xét đồng thời nhiều chuỗi số liệu. Mặt khác, nếu muốn xem xét sự biến đổi theo thời gian của các đặc trưng yếu tố khí hậu ta có thể cùng một lúc sử dụng chuỗi số liệu các tháng khác nhau tại cùng một trạm. Trên cơ sở những tập số liệu này, sau khi tính toán xử lý, các đặc trưng phân bố sẽ được biểu diễn lên biểu đồ, đồ thị tổng hợp – mà ta gọi là toán đồ – làm cơ sở cho việc phân tích, đánh giá và phán đoán về qui luật khí hậu.

Một trong những toán đồ đơn giản sẽ được giới thiệu ở đây là toán đồ Lebedev, mang tên nhà bác học Xô viết A. N. Lebedev.

Mục đích của toán đồ này là khái quát hoá mối quan hệ giữa đặc trưng trung bình số học của yếu tố khí hậu và giá trị của yếu tố đó ứng với các suất bảo đảm khác nhau tính được theo các chuỗi số liệu tại các trạm trên một vùng không gian nào đó.

Trong ứng dụng thực hành, việc tạo ra được một bức tranh khái quát về mối quan hệ giữa trị số trung bình số học và tập giá trị của yếu tố khí hậu ứng với các

suất bảo đảm khác nhau cho một khu vực là hết sức cần thiết. Nó còn là cơ sở để ước lượng trị số khí hậu ứng với các suất bảo đảm khác nhau cho những trạm có chuỗi số liệu ngắn.

Toán đồ Lebedev được xây dựng theo nguyên tắc sau đây:

Giả sử trong khu vực nghiên cứu có N trạm khác nhau được chọn và yếu tố khí hậu cần xem xét là X. Tương ứng với các trạm ta có thể tính được giá trị trung bình số học của X và các giá trị x_Φ ứng với những giá trị $\Phi(x_\Phi)=\Phi_i=\alpha$ khác nhau. Các giá trị Φ_i thường được chọn là $\Phi_1=95\%$, $\Phi_2=90\%,...,\Phi_m=5\%$.

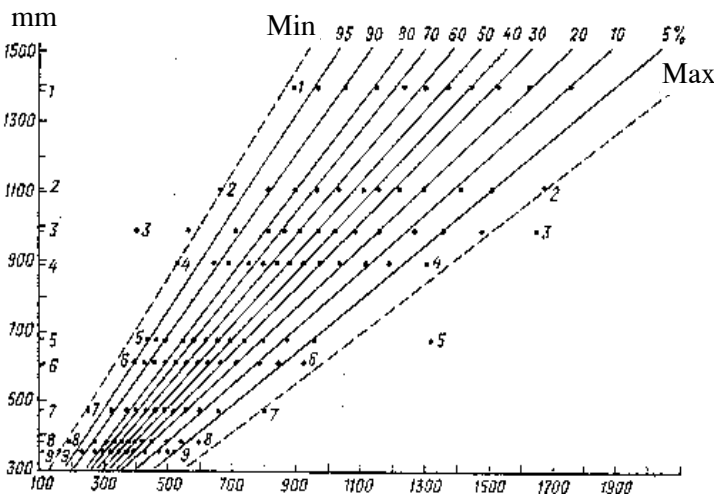
Ký hiệu $\overline{x^j}(j=1...N), x^j_{\Phi_i}(i=1...m)$ theo thứ tự là trung bình số học và trị số của X ứng với suất bảo đảm Φ_i của trạm thứ j, ta chọn các ký hiệu biểu diễn và thành lập bảng tính như trong bảng 1.10.

Từ bảng này, lập hệ trục toạ độ với trục hoành là giá trị x_Φ còn trục tung là $\bar{x}$. Mỗi một bộ N cặp số $(x^j_{\Phi_i}, \overline{x^j})$ tương ứng với một ký hiệu biểu diễn trên mặt phẳng toạ độ. Như vậy có tất cả N đểm biểu diễn cho một giá trị suất bảo đảm Φ_i. Tại mỗi điểm $(x^j_{\Phi_i}, \overline{x^j})$ trên mặt phẳng ta chấm các ký hiệu biểu diễn này và vẽ đường xấp xỉ. Có tất cả m đường biểu diễn tương ứng với m giá trị suất bảo đảm Φ_i khác nhau.

Việc xấp xỉ các đường có thể được thực hiện bằng tay hoặc bằng phương pháp bình phương tối thiểu. Các đường này có thể là đường thẳng hoặc đường cong tuỳ theo mối quan hệ phụ thuộc giữa x_Φ và $\bar{x}$. Trên hình 1.15 dẫn ra một ví dụ về toán đồ Lebedev với trục tung là lượng mưa trung bình tháng còn trục hoành là giá trị lượng mưa tháng ứng với các suất bảo đảm khác nhau.

Bảng 1.10 Giá trị trung bình và trị số ứng với các suất bảo đảm khác nhau

Trạm	Trung bình ($\bar{x}$)	Φ_i (i=1..m)		
		Φ_1	...	Φ_m
1	$\overline{x^1}$	$x^1_{\Phi_1}$	...	$x^1_{\Phi_m}$
2	$\overline{x^2}$	$x^2_{\Phi_1}$	...	$x^2_{\Phi_m}$
...	...	...	...	
N	$\overline{x^N}$	$x^N_{\Phi_1}$	...	$x^N_{\Phi_m}$
Ký hiệu biểu diễn		●	...	⊙



Hình 1.15 Toán đồ biểu diễn quan hệ giữa lượng mưa trung bình tháng và lượng mưa ứng với các suất bảo đảm khác nhau

CHƯƠNG 2
CÁC ĐẶC TRƯNG SỐ CỦA PHÂN BỐ VÀ VẤN ĐỀ PHÂN TÍCH KHẢO SÁT SỐ LIỆU

2.1 Đặt vấn đề

Một trong những ứng dụng rất quan trọng của phương pháp thống kê trong khí tượng, khí hậu là tạo khả năng phán đoán về những tập số liệu mới. Như đã biết, hệ thống quan trắc khí tượng và các sản phẩm tính toán từ những mô hình số trị tạo ra hàng loạt dữ liệu số phản ánh sự biến đổi theo không gian và thời gian của các yếu tố khí tượng. Tuy nhiên, để rút ra được những qui luật biến thiên của chúng cần phải khảo sát phân tích một cách tỷ mỷ. Công cụ thống kê có thể giúp chúng ta nhận biết và phán đoán một tập số liệu mới một cách nhanh chóng để từ đó rút ra bản chất của quá trình khí quyển.

Phương pháp thống kê phân tích khảo sát số liệu yêu cầu phải xử lý một lượng rất lớn số liệu ban đầu. Nó cho phép "nén thông tin", tóm lược số liệu và mô tả chúng thông qua những đặc trưng số hoặc các giản đồ, biểu đồ hay đồ thị.

Trong phân tích khảo sát các trường số liệu khí tượng, đồ thị là một công cụ biểu diễn rất có hiệu quả. Đồ thị có thể biểu diễn một khối lượng số liệu khổng lồ trong một không gian bé, giúp ta phát hiện những đặc điểm không bình thường của tập số liệu. Những chi tiết không bình thường đó có thể hết sức quan trọng, đôi khi chúng chứa đựng sai số quan trắc hoặc truyền số liệu, và cần phải biết càng sớm càng tốt khi phân tích. Cũng có lúc số liệu không bình thường lại là hợp lý và có thể là một bộ phận thông tin lý thú của tập số liệu. Trong lớp các phương pháp đồ thị thông thường nhất người ta sử dụng đồ thị hàm phân bố thực nghiệm (mục 1.6, chương 1). Dựa trên các đường tần suất, tần suất tích lũy, ngoài việc phát hiện những biến đổi đột xuất ta có thể phán đoán một cách nhanh nhất các thuộc tính của phân bố, xác định được các đặc trưng số của nó.

Những đặc trưng thống kê đơn giản và các đặc trưng số của phân bố cũng là những thông tin quan trọng ban đầu, giúp ta phân tích phán đoán có hiệu quả các tập số liệu. Chúng có thể được tính toán một cách nhanh chóng và chính xác bằng những chương trình máy tính đơn giản.

2.2 Các phân vị (Quantiles) và mốt (Mode)

Phân vị mẫu q_p là số có cùng đơn vị đo với số liệu và có giá trị vượt quá những trị số khác của tập số liệu với xác suất bằng p. Có thể hiểu phân vị q_p như là giá trị mà tại đó tần suất tích luỹ bằng p:

$$q_p = x(F(x) = p)$$

Các phân vị mẫu thường được dùng để khảo sát, thăm dò một cách khái quát tập số liệu. Thông thường người ta sử dụng $q_{0.5}$, được gọi là median hay trung vị và ký hiệu là Me. Trung vị Me là giá trị nằm ở vị trí trung tâm của chuỗi số liệu đã sắp xếp theo thứ tự tăng dần (chuỗi trình tự) sao cho số thành phần của chuỗi có trị số nhỏ hơn Me bằng số thành phần lớn hơn Me. Nếu số thành phần của chuỗi là lẻ thì trung vị đơn giản là giá trị nằm ở vị trí giữa của chuỗi trình tự. Tuy nhiên, nếu số thành phần của chuỗi là chẵn thì chuỗi có hai giá trị giữa và trung vị được qui ước lấy bằng trung bình của các giá trị giữa này. Cụ thể, giả sử từ chuỗi ban đầu $\{x_1, x_2,..., x_n\}$ ta sắp xếp thành chuỗi trình tự $\{ x_{(1)}, x_{(2)},..., x_{(n)}\}$ với $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$ (chú ý rằng đây là chuỗi trình tự nhưng chưa xếp hạng). Khi đó ta có:

$$\text{Me} = q_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{với n lẻ} \\ \dfrac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{với n chẵn} \end{cases} \qquad (2.2.1)$$

Ngoài trung vị Me, một số phân vị khác cũng được sử dụng phổ biến là $q_{0.25}$ và $q_{0.75}$. Người ta thường gọi các phân vị này tương ứng là phân vị dưới và phân vị trên hay *tứ vị*, chúng nằm giữa trung vị Me và các cực trị $x_{min} = x_{(1)}$ và $x_{max}=x_{(n)}$. Đôi khi người ta còn gọi $q_{0.25}$ và $q_{0.75}$ bằng những thuật ngữ hình tượng bóng bẩy hơn là *bản lề* hay *khớp nối* hoặc *điểm mấu chốt*. Như vậy các phân vị dưới và trên là hai trung vị của hai nửa tập số liệu giữa Me=$q_{0.5}$ và các cực trị. Nếu n lẻ thì mỗi nửa tập số liệu này bao gồm (n+1)/2 điểm và cả hai đều chứa trung vị. Nếu n chẵn thì mỗi nửa này chứa n/2 điểm và chúng không đè lên nhau (không giao nhau). Một số phân vị khác ít thông dụng hơn đôi khi cũng được xem xét đến là phân vị "tám" hay *bát vị* $q_{0.125}$, $q_{0.325}$, $q_{0.625}$ và $q_{0.825}$, phân vị "mười sáu" $q_{0.0625}$, v.v. và những phân vị "thập phân" $q_{0.1}$, $q_{0.2}$,..., $q_{0.9}$.

Ví dụ 2.2.1 Giả sử tập mẫu gồm n=9 thành phần đã được sắp xếp thành chuỗi trình tự $\{x_{(1)}, x_{(2)},..., x_{(9)}\}$ thì trung vị Me = $q_{0.5} = x_{(5)}$ hoặc giá trị *lớn thứ năm* trong 9 số đã cho. Phân vị dưới là $q_{0.25}=x_{(3)}$ và phân vị trên là $q_{0.75}=x_{(7)}$.

Nếu n=10 thì trung vị là trung bình của hai trị số giữa, nhưng các phân vị dưới và phân vị trên là trị số giữa của nửa dưới và nửa trên của tập số liệu. Có nghĩa là $q_{0.25}= x_{(3)}$, $q_{0.5} =(x_{(5)}+x_{(6)})/2$ và $q_{0.75} = x_{(8)}$.

Nếu n=11, khi đó trung vị Me là trị số giữa duy nhất, còn các phân vị dưới và trên được xác định bởi trung bình của hai trị số giữa của các nửa trên và nửa dưới của tập số liệu: $q_{0.25}=(x_{(3)}+ x_{(4)})/2$, $Me=q_{0.5}=x_{(6)}$ và $q_{0.75}= (x_{(8)}+ x_{(9)})/2$.

Với n=12 thì cả trung vị và hai phân vị dưới và trên đều được xác định bởi trung bình từng cặp trị số giữa: $q_{0.25}=(x_{(3)}+ x_{(4)})/2$, $Me=q_{0.5}=(x_{(6)}+ x_{(7)})/2$ và $q_{0.75}=(x_{(9)}+ x_{(10)})/2$.

Trong khí tượng, khí hậu các phân vị được sử dụng để khảo sát sơ bộ số liệu ban đầu. Ưu điểm chính của việc sử dụng các đặc trưng này là chúng không bị ảnh hưởng đáng kể bởi những số liệu có chứa sai số thô. Có thể lấy ví dụ sau đây để so sánh. Giả sử khi tiến hành nhập số liệu nhiệt độ, các giá trị đúng là {18.9, 19.2, 19.4, 20.3, 20.8, 21.6, 21.9, 22.0, 22.5, 23.9}, khi đó trung bình số học của chuỗi $\bar{x}$=21.1 và trung vị Me=21.2. Nhưng do sơ suất, thay vì trị số cuối cùng bằng *23.9*, người ta đã vào nhầm thành *239* (lớn gấp 10 lần số đúng). Vì vậy, trung bình số học của chuỗi đã bị thay đổi một cách đáng kể: $\bar{x}$=42.3, trong khi đó trung vị Me vẫn không thay đổi. Trong một số trường hợp trung vị làm chức năng thay thế trung bình số học. Chẳng hạn, khi xử lý chuỗi số liệu gió cực đại, tốc độ gió có thể khá lớn và dao động mạnh, nếu sử dụng trung bình số học sẽ thiếu chính xác. Trong trường hợp này người ta dùng trung vị chứ không dùng trung bình số học.

Rõ ràng ta có thể xác định được các phân vị khi đã biết phân bố xác suất F(x) từ phương trình:

$$F(x) = p \tag{2.2.2}$$

Nghiệm của phương trình này chính là q_p. Với p=0.5 ta có:

$$F(x) = 0.5$$

và nghiệm của nó là $x = Me = q_{0.5}$.

Bởi vậy ta còn có biểu thức định nghĩa khác của trung vị là:

$$P(x>Me) = P(x<Me) \tag{2.2.3}$$

Một đặc trưng quan trọng khác cũng thường được ứng dụng trong phân tích khảo sát số liệu là mốt (mode). Mốt được ký hiệu bởi Mo, là giá trị của biến ngẫu nhiên mà tại đó hàm mật độ xác suất đạt cực đại:

$$\left.\frac{df(x)}{dx}\right|_{x=Mo} = 0$$
$$\left.\frac{d^2f(x)}{dx^2}\right|_{x=Mo} < 0 \tag{2.2.4}$$

trong đó f(x) là hàm mật độ xác suất.

Như vậy, về nguyên tắc, tuỳ thuộc vào dạng hàm mật độ xác suất f(x), một phân bố có thể có nhiều mốt hoặc không có mốt nào. Khi xét cụ thể một tập số liệu nào đó, mốt là trị số có tần suất xuất hiện lớn nhất, tức là người ta thường chỉ quan tâm đến mốt quan trọng nhất.

Ví dụ 2.2.2 Xét tập số liệu sau {1, 2, 3, 4, 2, 5, 4, 6, 4, 8} ta thấy xuất hiện hai mốt là Mo_1=4 và Mo_2=2. Nhưng tần số xuất hiện giá trị 4 (3 lần) lớn hơn tần số xuất hiện trị số 2 (2 lần), do đó ta chỉ sử dụng mốt thứ nhất: Mo=Mo_1=4.

Một số phương pháp xác định trung vị và mốt

1) Phương pháp chọn trực tiếp theo công thức (2.2.1).

2) Phương pháp phân nhóm và sử dụng công thức thực nghiệm

Giả sử chuỗi x_t (t=1..n) được chia thành N nhóm với cự ly nhóm Δx=const. Gọi m_j và μ_j là tần số và tần số tích luỹ nhóm thứ j, ta có:

– Trung vị: $$Me = x_M + \Delta x. \frac{\frac{n^*}{2} - \mu_{M-1}}{m_M} \qquad (2.2.5)$$

trong đó:

M là vị trí nhóm trung vị (nhóm chứa $x_{(n/2)}$),

x_M là giới hạn dưới của nhóm thứ M,

m_M là tần số của nhóm thứ M,

μ_{M-1} là tần số tích luỹ của nhóm thứ M–1,

Δx là cự ly nhóm,

$$\frac{n^*}{2} = \begin{cases} \frac{1}{2}(n+1) & \text{nếu n lẻ} \\ \frac{n}{2}+1 & \text{nếu n chẵn} \end{cases}$$

– Mốt: $$Mo = x_M + \Delta x. \frac{m_M - m_{M-1}}{(m_M - m_{M-1}) + (m_M - m_{M+1})} \qquad (2.2.6)$$

trong đó:

M là vị trí nhóm mốt,

x_M là giới hạn dưới của nhóm mốt (nhóm có tần số lớn hơn tần số các nhóm lân cận),

m_M, m_{M-1}, m_{M+1} theo thứ tự là tần số nhóm mốt, nhóm liền trước và liền sau nhóm mốt.

Δx là cự ly nhóm.

– Đối với những phân bố không quá bất đối xứng và có một đỉnh ta có mối liên hệ để tính mốt sau đây:

$$Mo \approx \bar{x} + 3(Me - \bar{x}) \qquad (2.2.7)$$

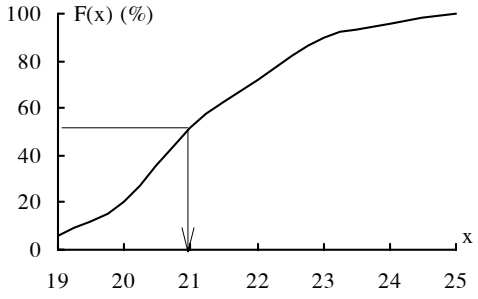
trong đó $\bar{x}$ là trung bình số học của chuỗi:

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t$$

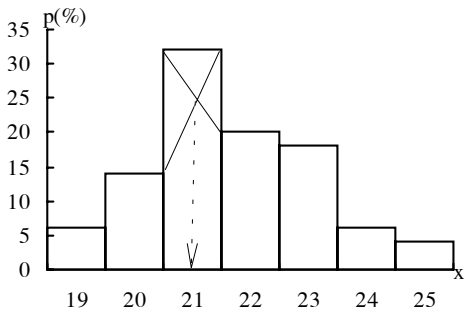
3) Phương pháp đồ thị

– Xác định trung vị: Để xác định trung vị bằng phương pháp đồ thị ta xây dựng đường cong phân bố và chọn điểm trên trục tung ứng với giá trị F(x) = 0.5, sau đó kẻ song song với trục hoành, khi cắt đồ thị F(x) thì kẻ song song với trục tung. Điểm cắt trục hoành chính là Me (hình 2.1).

– Xác định mốt: Muốn xác định mốt bằng phương pháp đồ thị trước hết ta xây dựng biểu đồ phân bố tần suất (hình 2.2). Sau đó, chọn nhóm có tần suất cực đại và kẻ các đoạn thẳng nối các điểm tương ứng với cận trên và cận dưới của nhóm liền trước, nhóm mốt và nhóm liền sau mốt. Từ giao điểm của các đoạn thẳng này kẻ song song với trục tung, cắt trục hoành tại điểm có hoành độ là mốt.



Hình 2.1 Xác định trung vị



Hình 2.2 Xác định mốt

Ví dụ 2.2.3 Từ số liệu lịch sử 50 năm của nhiệt độ không khí ở một trạm ta có bảng thống kê sau:

Nhóm	Khoảng nhiệt độ (°C)	Tần số nhóm	Tần số tích luỹ	Tần suất nhóm (%)	Tần suất tích luỹ (%)
(1)	(2)	(3)	(4)	(5)	(6)
1	18–19	3	3	6	6
2	19–20	7	10	14	20
3	20–21	16	26	32	52
4	21–22	10	36	20	72
5	22–23	9	45	18	90
6	23–24	3	48	6	96
7	24–25	2	50	4	100

Sử dụng công thức (2.2.5) ta có: Với dung lượng mẫu n = 50 thì $n^*/2 = 26$, từ cột (4) suy ra nhóm trung vị là nhóm 3 (M=3), có cận dưới $x_M = 20$. Cự ly nhóm $\Delta x=1$, tần số nhóm trung vị $m_M = 16$, tần số tích luỹ của nhóm trước nhóm trung vị $\mu_{M-1}=10$. Vậy:

$$Me = 20.0 + 1.\frac{(\frac{50}{2}+1)-10}{16} = 21.0$$

Tương tự, đối với công thức (2.2.6), từ cột (3) ta có vị trí nhóm mốt là M=3, cận dưới nhóm mốt $x_M = 20$, tần số các nhóm mốt, liền trước và liền sau nhóm mốt là $m_M = 16$, $m_{M-1} = 7$, $m_{M+1} = 10$, cự ly nhóm $\Delta x=1$. Do đó:

$$Mo = 20.0 + 1.\ \frac{16-7}{(16-7)+(16-10)} = 20.6$$

Bạn đọc có thể nhận thấy các kết quả này trên các hình 2.1 và 2.2.

2.3 Các mômen phân bố

Từ quan điểm thống kê, trong hầu hết các bài toán khí tượng, khí hậu người ta xem các tập số liệu quan trắc như là những tập mẫu của các đại lượng ngẫu nhiên hay các biến ngẫu nhiên. Như đã biết, đặc trưng đầy đủ của đại lượng ngẫu nhiên là hàm phân bố xác suất. Tuy nhiên, trong thực tế, nhiều khi không đòi hỏi phải hiểu biết thật đầy đủ về đại lượng ngẫu nhiên mà chỉ cần biết một vài đặc trưng quan trọng có thể mô tả được một cách khái quát về đại lượng ngẫu nhiên là đủ. Các đặc trưng đó được gọi là mômen phân bố.

2.3.1 Mômen gốc

Theo định nghĩa, mômen gốc bậc r của đại ngẫu nhiên X được ký hiệu là α_r và được xác định bởi:

$$\alpha_r = \int_{-\infty}^{+\infty} x^r f(x)dx, r = 1,2.,..$$

trong đó f(x) là hàm mật độ xác suất của X. Trong các mômen gốc của đại lượng ngẫu nhiên X, mômen gốc bậc nhất α_1 có ý nghĩa đặc biệt, nó được gọi là kỳ vọng toán hay giá trị trung bình của đại lượng ngẫu nhiên. Kỳ vọng toán của đại lượng ngẫu nhiên X đặc trưng cho độ lớn của X. Đôi khi người ta còn gọi nó là giá trị nền. Ta sẽ ký hiệu kỳ vọng toán của đại lượng ngẫu nhiên X là M[X] hay m_x và xác định bởi:

$$M[X] = m_x = \int_{-\infty}^{+\infty} x f(x) dx$$

Như vậy, kỳ vọng toán học là kết quả của việc trung bình theo xác suất tất cả các giá trị có thể của đại lượng ngẫu nhiên. Theo định nghĩa đó ta có thể suy rộng ra rằng, mômen gốc bậc r của đại lượng ngẫu nhiên X là kỳ vọng toán học của luỹ thừa bậc r của đại lượng ngẫu nhiên:

$$\alpha_r = M[X^r] \tag{2.3.1}$$

Ở đây M là ký hiệu toán tử lấy kỳ vọng. Từ nay trở đi, nếu không giải thích gì thêm thì ký hiệu này sẽ được giữ nguyên ý nghĩa của nó. Đôi lúc để đơn giản ta còn ký hiệu kỳ vọng toán của X là MX.

Mômen gốc α_r thường được gọi là mômen gốc tổng thể. Giá trị thống kê của mômen gốc α_r ký hiệu a_r và được xác định bởi:

$$a_r = \frac{1}{n}\sum_{t=1}^{n} x_t^r \tag{2.3.2}$$

trong đó x_t, t = 1..n, là các giá trị quan trắc (hay còn gọi là mẫu) của X, *n* là dung lượng mẫu. Bởi vậy người ta thường gọi a_r là mômen gốc mẫu.

Khi r=1 ta có $a_1 = \frac{1}{n}\sum_{t=1}^{n} x_t = \bar{x}$ và được gọi là trung bình số học của X. Trung bình số học là ước lượng thống kê của kỳ vọng toán học m_x. Dấu gạch ngang phía trên ($\bar{x}$) được hiểu là ký hiệu phép lấy trung bình số học hay toán tử lấy kỳ vọng mẫu. Ký hiệu này cũng sẽ được giữ nguyên ý nghĩa của nó trong phạm vi tài liệu này.

2.3.2 Mômen trung tâm

Mômen trung tâm bậc r của đại lượng ngẫu nhiên X được ký hiệu là μ_r và được xác định bởi:

$$\mu_r = M[(X - M[X])^r] = M[(X - m_x)^r] \tag{2.3.3}$$

Khi r =1 ta có $\mu_1 = M[(X - m_x)] = M[X] - m_x = m_x - m_x = 0$. Như vậy mômen trung tâm bậc 1 của đại lượng ngẫu nhiên luôn luôn bằng 0.

Khi r=2: $\mu_2 = M[(X - m_x)^2] = D[X] = D_x$ và được gọi là phương sai của đại lượng ngẫu nhiên, dùng để đặc trưng cho mức độ phân tán của các giá trị của X xung quanh kỳ vọng toán học. Bởi vậy trong nhiều trường hợp người ta còn gọi D_x là độ tán. Ký hiệu D[X] ở đây được hiểu như toán tử lấy phương sai của X. Trong một số trường hợp, để đơn giản, thay cho D[X] ta có ký hiệu DX.

Vì D_x có thứ nguyên bằng bình phương thứ nguyên của X nên việc sử dụng nó để đặc trưng cho độ phân tán nói chung thiếu tính rõ ràng. Do đó trong thực tế thay cho D_x người ta dùng giá trị căn bậc hai của nó.

$$\sigma_x = \sqrt{D_x} \tag{2.3.4}$$

và gọi là độ lệch bình phương trung bình của đại lượng ngẫu nhiên.

Khi r = 3: $$\mu_3 = M[(X-m_x)^3] \tag{2.3.5}$$

Mômen trung tâm bậc ba μ_3 dùng để đặc trưng cho tính bất đối xứng của phân bố.

Khi r=4: $$\mu_4 = M[(X-m_x)^4] \tag{2.3.6}$$

Mômen trung tâm bậc bốn μ_4 dùng để đặc trưng cho mức độ tập trung của phân bố.

Từ (2.3.3) và (2.3.1), khi để ý đến khai triển nhị thức Newton ta có:

$$\mu_r = M[(X - m_x)^r] = M\left[\sum_{k=0}^{r}(-1)^k C_r^k X^{r-k} m_x^k\right] =$$

$$= \sum_{k=0}^{r}(-1)^k C_r^k \alpha_1^k M\left[X^{r-k}\right] = \sum_{k=0}^{r}(-1)^k C_r^k \alpha_1^k \alpha_{r-k}$$

Hay:

$$\mu_r = \sum_{k=0}^{r}(-1)^k C_r^k \alpha_1^k \alpha_{r-k} \tag{2.3.7}$$

Như vậy, mômen trung tâm có thể tính được qua mômen gốc.

Ví dụ: với r=2 ta có $\mu_2=\alpha_2-2(\alpha_1)^2+(\alpha_1)^2=\alpha_2-(\alpha_1)^2$

Ước lượng thống kê của mômen trung tâm μ_r ký hiệu là m_r và được xác định bởi:

$$m_r = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^r \tag{2.3.8}$$

với x_t, t=1...n, là giá trị quan trắc của X, n là dung lượng mẫu. Người ta còn gọi m_r là mômen trung tâm mẫu.

Giữa mômen trung tâm mẫu và mômen gốc mẫu cũng liên hệ với nhau bởi hệ thức:

$$m_r = \sum_{k=0}^{r}(-1)^k C_r^k a_1^k a_{r-k} \tag{2.3.9}$$

Có thể biểu diễn công thức này dưới dạng cụ thể hơn:

$$m_r = \sum_{k=0}^{r} \frac{1}{n} \sum_{t=1}^{n} (-1)^k C_r^k x_t^{r-k} \left(\bar{x}\right)^k \qquad (2.3.9')$$

Khi r=1 ta có $m_1 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x}) = \frac{1}{n}\sum_{t=1}^{n} x_t - \bar{x} = 0$

Khi r =2 ta có $m_2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2 = \tilde{D}_x = \overline{x^2} - \left(\bar{x}\right)^2$ và gọi là phương sai mẫu. Đại lượng $s_x = \sqrt{\tilde{D}_x}$ được gọi là độ lệch tiêu chuẩn hay độ lệch chuẩn của X, nó là ước lượng của độ lệch bình phương trung bình σ_x.

2.3.3 Các phương pháp tính mômen

2.3.3.1 Phương pháp tính trực tiếp

Phương pháp tính trực tiếp là tính các mômen gốc và mômen trung tâm theo các công thức (2.3.2), (2.3.8) và có thể sử dụng cả công thức liên hệ (2.3.9').

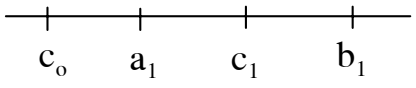
2.3.3.2 Phương pháp phân nhóm

Phương pháp này thường được sử dụng trong trường hợp dung lượng mẫu đủ lớn. Ưu điểm của phương pháp này là số lượng phép tính ít, qui trình tính toán đơn giản; nhược điểm của nó là độ chính xác không cao.

Giả sử tập số liệu ban đầu $\{x_t, t=1..n\}$ được chia thành N nhóm với cự ly các nhóm đều nhau và bằng Δx. Ta có bảng sau:

Nhóm	Giới hạn dưới	Giới hạn trên	Trị số giữa	Tần số
1	a_1	b_1	c_1	m_1
2	a_2	b_2	c_2	m_2
...	...	...	...	...
N	a_N	b_N	c_N	m_N

Trong đó: $a_1 \le \min\{x_t, t = 1..n\}$, $b_N > \max\{x_t, t=1..n\}$, $b_j - a_j = \Delta x = \text{const}$ là cự ly nhóm, $b_j = a_{j+1}$, $c_j = c_o + j\Delta x$ là trị số giữa của nhóm, $c_o = a_1 - \Delta x/2$ (hình 2.3). Tần số m_i là số thành phần của chuỗi rơi vào nhóm thứ j.



Hình 2.3 Sơ đồ chia khoảng

Khi đó các mômen sẽ được tính theo các công thức sau đây:

– Mômen gốc: $a_r \approx a'_r = \frac{1}{n}\sum_{j=1}^{N} m_j c_j^r$ (2.3.10)

– Mômen trung tâm: $m_r \approx m'_r = \frac{1}{n}\sum_{j=1}^{N} m_j (c_j - \bar{c})^r$ (2.3.11)

với $\bar{c} = \frac{1}{n}\sum_{j=1}^{N} m_j c_j$.

Như vậy các mômen a_r và m_r chỉ là giá trị xấp xỉ theo a'_r và m'_r mà chúng được tính khi thừa nhận rằng các thành phần thuộc nhóm thứ j đều lấy cùng một giá trị c_j. Rõ ràng độ chính xác của kết quả tính theo phương pháp này không cao, thậm chí sai lệch nhiều so với kết quả tính trực tiếp. Mặc dù vậy trong nhiều trường hợp người ta vẫn sử dụng phương pháp này, nhất là khi dung lượng mẫu cực lớn hoặc khi cần khảo sát sơ bộ tập số liệu.

Do việc phân nhóm sẽ gây nên sai số khi tính các mômen nên người ta phải tiến hành hiệu chỉnh chúng. Sau đây là một số công thức để hiệu chỉnh giá trị của mômen trung tâm bậc hai và bậc bốn tính bằng phương pháp phân nhóm:

$$m_{2hc} = m_2 - \frac{1}{12}(\Delta x)^2 \quad (2.3.12)$$

$$m_{4hc} = m_1 - \frac{1}{2}m_2 + \frac{7}{240}(\Delta x)^4 \quad (2.3.13)$$

Trong đó m_{2hc} và m_{4hc} là mômen trung tâm bậc hai và bậc bốn đã hiệu chỉnh, Δx là cự ly nhóm.

Ví dụ 2.3.1. Số liệu lịch sử tổng lượng mưa năm của trạm A được cho trong bảng 2.1. Hãy tính mômen gốc bậc 1 và mômen trung tâm bậc 2.

Bảng 2.1 Số liệu tổng lượng mưa năm (mm) của trạm A

1983.8	2325.4	1297.3	1554.3	1931.6	1433.6	1283.1	2246.3
1631.3	1701.9	1736.8	1943.4	1225.5	1249.4	1214.4	1532.1
1719.7	1931.9	1725.7	2128.3	1599.6	1894.4	2115.1	1055.7
1525.9	1829.8	1684.5	1828.9	1315.6	1284.3	1733.7	1760.6
1448.5	1568.8	1256.8	1651.7	1488.2	1390.5	2033.4	1538.1
1884.9	1544.4	1862.8	1806.5	1758.2	1935.2	1726.7	
1405.5	1758.9	1738.8	1744.2	1274.8	1839.6	1766.3	
2061.8	2141.2	1800.0	1954.1	1662.5	1964.5	1646.7	
1995.0	2153.9	2528.2	1561.5	1951.1	1527.2	2225.1	
1147.8	1653.0	2040.3	1623.9	1657.6	1985.9	1596.1	

Ở đây ta có dung lượng mẫu n=105. Áp dụng công thức (2.3.1) với r=1 ta được: $a_1 = \bar{x}$ =1683.9 (mm). Sử dụng công thức (2.3.8) ta được $m_2 = \tilde{D}_x$ =103929.3 (mm^2)

Để tiến hành tính toán bằng phương pháp nhóm theo các công thức (2.3.10) và (2.3.11) ta chia chuỗi số liệu đã cho làm 11 nhóm với cự lý các nhóm bằng bằng nhau và bằng Δx=165. Ta lập bảng thống kê kết quả phân nhóm (bảng 2.2).

Kết quả tính cho ta: $a_1 = \bar{x}$ =1681.2(mm); $m_2 = \tilde{D}_x$ =104366.2(mm^2).

Như vậy kết quả tính theo hai phương pháp trong trường hợp này có sự chênh lệch chút ít. Giá trị hiệu chỉnh của m_2 tính theo công thức (2.3.12) bằng m_{2hc}=102097.5 (mm^2).

2.4 Trung bình số học

Trong thống kê có nhiều khái niệm trung bình khác nhau được sử dụng, như trung bình số học, trung bình điều hoà, trung bình hình học, trung bình bình phương,... Tuy nhiên khái niệm trung bình được sử dụng phổ biến trong khí tượng, khí hậu là trung bình số học. Ý nghĩa cơ bản của trung bình số học là nó chứa đựng thông tin quan trọng nhất về chế độ của đặc trưng yếu tố khí hậu. Chức năng của trung bình số học trong nghiên cứu khí hậu là phản ánh một cách khái quát độ lớn của các thành phần trong chuỗi, dung hoà được các dao động thăng dáng và biểu thị trạng thái trung gian hay giá trị nền của chuỗi.

Bảng 2.2. Kết quả phân nhóm

Nhóm j	a_j	b_j	c_j	m_j	$c_j m_j$	$c_j^2 m_j$
1	835	1000	917.5	1	917.5	841806.3
2	1000	1165	1082.5	4	4330	4687225.0
3	1165	1330	1247.5	10	12475	15562563.5
4	1330	1495	1412.5	15	21187.5	29927343.8
5	1495	1660	1577.5	22	34705	54747137.5
6	1660	1825	1742.5	17	29622.5	51617206.3
7	1825	1990	1907.5	19	36242.5	69132568.8
8	1990	2155	2072.5	11	22797.5	47247818.8
9	2155	2320	2237.5	3	6712.5	15019218.8
10	2320	2485	2402.5	1	2402.5	5772006.3
11	2485	2650	2567.5	2	5135	13184113.5
Tổng				105	176527.5	307739006.3

Giả sử đại lượng khí hậu X có các quan trắc là $\{x_t, t=1..n\}$. Khi đó trung bình số học là ước lượng thống kê của kỳ vọng toán học của X, nên đôi khi nó còn được gọi là kỳ vọng mẫu. Trung bình số học ký hiệu là $\bar{x}$, nó chính là mômen gốc mẫu bậc 1 và được xác định bởi:

$$\bar{x} = a_1 = \frac{1}{n}\sum_{t=1}^{n} x_t \qquad (2.4.1)$$

Trung bình số học có các tính chất sau đây:

1) Tổng độ lệch của các thành phần trong chuỗi so với trung bình số học bằng không: $\sum_{t=1}^{n}(x_t - \bar{x}) = 0$

2) Nếu cộng (trừ) mỗi thành phần của chuỗi với cùng một hằng số C thì trung bình số học sẽ tăng (giảm) một lượng đúng bằng C:

$$\frac{1}{n}\sum_{t=1}^{n}(x_t \pm C) = \bar{x} \pm C \qquad (2.4.2)$$

3) Nếu nhân (chia) mỗi thành phần của chuỗi với cùng một hằng số C khác 0 thì trung bình số học tăng (giảm) C lần:

$$\frac{1}{n}\sum_{t=1}^{n} Cx_t = C\bar{x}\,,\ \frac{1}{n}\sum_{t=1}^{n}\frac{x_t}{C} = \frac{\bar{x}}{C} \qquad (2.4.3)$$

4) Với C là một hằng số bất kỳ ta có $\sum_{t=1}^{n}(x_t - \bar{x})^2 \le \sum_{t=1}^{n}(x_t - C)^2$.

Bên cạnh trung bình số học, để khảo sát mức độ tập trung của các tập số liệu khí tượng, khí hậu người ta còn sử dụng một số đặc trưng đơn giản như trung vị Me hay mốt Mo. Các đặc trưng này nói chung có tính ổn định và không bị ảnh hưởng đáng kể bởi sai số hoặc những giá trị đột xuất. Như đã chỉ ra trong mục 2.2, khi xét tập số liệu {18.9, 19.2, 19.4, 20.3, 20.8, 21.6, 21.9, 22.0, 22.5, 23.9}, trong khi trung vị Me không bị thay đổi thì trung bình số học $\bar{x}$ tăng lên một cách đáng kể, từ 21.1 lên 42.3 nếu số cuối cùng bị thay thế bởi trị số sai 239. Tuy vậy, với những tập số liệu không chứa sai số thì trung bình số học cho độ chính xác cao hơn.

Một số phương pháp tính trung bình số học

1) Phương pháp tính trực tiếp: Tính theo công thức (2.4.1).

2) Phương pháp biến đổi tương đương: Khi giá trị của các thành phần trong chuỗi dao động xung quanh một hằng số C hoặc là bội của một hằng số C nào đó ta có thể áp dụng công thức (2.4.2) hoặc (2.4.3) đã nêu trên đây để biến đổi chuỗi ban đầu về chuỗi mới rồi tiến hành tính toán trên chuỗi mới:

$$x'_t = x_t - C\,,\ \bar{x'} = \frac{1}{n}\sum_{t=1}^{n}(x_t - C) = \bar{x} - C \ \Rightarrow\ \bar{x} = \bar{x'} + C \qquad (2.4.4)$$

Nếu $x'_t = \frac{x_t}{C}$ thì $\bar{x'} = \frac{1}{n}\sum_{t=1}^{n}\frac{x_t}{C}$ và do đó $\bar{x} = \overline{Cx'}$ (2.4.5)

Trong một số trường hợp người ta còn kết hợp cả hai cách biến đổi trên.

Chẳng hạn, khi thực hiện phép biến đổi $x'_t = \frac{x_t - C}{d}$, với C và d là các hằng số, ta được:

$$\overline{x'} = \frac{1}{n}\sum_{t=1}^{n}\frac{x_t - C}{d} = \frac{\frac{1}{n}\sum_{t=1}^{n}x_t - C}{d} = \frac{\bar{x} - C}{d}, \text{ suy ra: } \bar{x} = \overline{x'}d + C \qquad (2.4.5')$$

3) Phương pháp phân nhóm: Tính theo các công thức (2.3.10) trong đó r = 1.

4) Phương pháp điều chỉnh: Giả sử chuỗi mới thành lập từ nhiều chuỗi ban đầu khác nhau mà các chuỗi này đã được tính trung bình thì trung bình chung sẽ được xác định bởi công thức:

$$\bar{x} = \frac{\sum_{i=1}^{K} n_i \overline{x_i}}{\sum_{i=1}^{K} n_i} \qquad (2.4.6)$$

trong đó K là số chuỗi ban đầu, $\overline{x_i} = \frac{1}{n_i}\sum_{t=1}^{n_i} x_{it}$, là trung bình của chuỗi thứ i và n_i là dung lượng mẫu nó.

Ví dụ 2.4.1 Giả sử ta có chuỗi số liệu khí áp

$\{x_t\}$={998.0, 1000.2, 1000.2, 1001.6, 1000.9, 999.1, 999.7, 999.2, 998.8, 998.2}

với độ chính xác ghi đến *mb*. Nếu tính trung bình số học $\bar{x}$ theo các giá trị hiện tại của chuỗi sẽ phải tính toán với những con số khá lớn. Khi xem xét toàn chuỗi ta thấy các giá trị trong chuỗi thường dao động xung quanh trị số 1000. Do đó, để đơn giản ta sử dụng phép biến đổi (2.4.5') với C=1000, d=0.1 và nhận được chuỗi mới:

$$\{x'_t\}=\{-20, 2, 2, 16, 9, -9, -3, -8, -12, -18\}.$$

Rõ ràng với chuỗi này ta dễ dàng nhận được $\overline{x'}$ = –4. Vậy $\bar{x}$=(–4) x (0.1) + 1000 = 999.6

Ví dụ 2.4.2 Giả sử nhiệt độ trung bình năm của 50 năm trước là 23.5°C và của 10 năm tiếp theo là 23.9°C. Sử dụng công thức (2.4.6) ta nhận được nhiệt độ trung bình năm của cả thời kỳ 60 năm là:

$$(23.5x50+23.9x10)/(50+10) = 23.6°C$$

2.5 Phương sai và độ lệch tiêu chuẩn

Như đã biết từ mục 2.3.2, phương sai D_x là đại lượng đặc trưng cho sự phân bố tản mạn của các giá trị của đại lượng ngẫu nhiên X xung quanh kỳ vọng toán học. Phương sai mẫu $\tilde{D}_x$ là ước lượng thống kê của phương sai D_x và được xác định bởi:

$$\tilde{D}_x = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2 \qquad (2.5.1)$$

trong đó x_t, t=1..n, là chuỗi các giá trị quan trắc của X. Căn bậc hai của phương sai mẫu được goi là độ lệch tiêu chuẩn hay độ lệch chuẩn s_x:

$$s_x = \sqrt{\tilde{D}_x} \qquad (2.5.2)$$

Đương nhiên rằng phương sai mẫu $\tilde{D}_x$ là đặc trưng thích hợp cho sự tản mạn của các thành phần trong chuỗi. Song, nó thiếu tính rõ ràng vì thứ nguyên của nó bằng bình phương thứ nguyên của đại lượng được đo. Trong khi đó s_x có cùng thứ nguyên với đại lượng được đo. Do vậy thông thường người ta dùng độ lệch chuẩn s_x làm thước đo mức độ phân tán của các thành phần trong chuỗi xung quanh giá trị trung bình. Độ lệch chuẩn s_x càng lớn thì độ tản mạn của chuỗi càng lớn và ngược lại.

Độ lệch chuẩn có các tính chất sau:

1) Nếu cộng (trừ) các thành phần của chuỗi với cùng một hằng số C bất kỳ thì độ lệch chuẩn vẫn không thay đổi:

$$s_x(X \pm C) = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left[x_t \pm C) - \overline{(x \pm C)}\right]^2} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left[x_t \pm C) - (\bar{x} \pm C)\right]^2}$$

$$s_x(X \pm C) = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2} = s_x(X) \qquad (2.5.3)$$

2) Nếu nhân (chia) các thành phần của chuỗi với cùng một hằng số C khác 0 thì độ lệch chuẩn sẽ tăng (giảm) một số lần tương ứng:

$$s_x(CX) = C.s_x(X) \qquad (2.5.4)$$

3) Độ lệch chuẩn là một ước lượng vững nhưng chệch của độ lệch bình phương trung bình σ_x:

Ký hiệu M[X] và D[X] là kỳ vọng và phương sai của đại lượng ngẫu nhiên X, ta có:

$$\sum(x_t-\bar{x})^2=\sum\left[(x_t-M[X])-(\bar{x}-M[X])\right]^2 =$$

$$=\sum(x_t-M[X])^2-2\sum\left[(x_t-M[X])(\bar{x}-M[X])\right]+\sum(\bar{x}-M[X])^2$$

Vì: $\sum(x_t-M[X])(\bar{x}-M[X])=(\bar{x}-M[X])\sum(x_t-M[X])=$

$=(\bar{x}-M[X])(n\bar{x}-nM[X])=n(\bar{x}-M[X])^2$

Tức là $\sum(\bar{x}-M[X])^2=n(\bar{x}-M[X])^2$

nên: $\sum(x_t-\bar{x})^2=\sum(x_t-M[X])^2-\sum(\bar{x}-M[X])^2$

Suy ra:

$$M\left[s_x^2\right]=M\left[\frac{1}{n}\sum(x_t-\bar{x})^2\right]=$$

$$=M\left[\frac{1}{n}\sum(x_t-M[X])^2\right]-M\left[\frac{1}{n}\sum(\bar{x}-M[X])^2\right]=$$

$$=\frac{1}{n}\sum M\left[(x_t-M[X])^2\right]-\frac{1}{n}\sum M\left[(\bar{x}-M[X])^2\right]=$$

$$=\frac{1}{n}\sum D[X]-\frac{1}{n}\sum D[\bar{x}]=$$

$$=\frac{1}{n}nD[X]-\frac{1}{n}nD[\bar{x}]=D[X]-D[\bar{x}]$$

Mặt khác: $D\left[\bar{x}\right]=D\left[\frac{1}{n}\sum x_t\right]=\frac{1}{n^2}\sum D\left[x_t\right]=\frac{1}{n^2}nD[X]=\frac{1}{n}D[X]$

Do đó: $M\left[s_x^2\right]=D[X]-\frac{1}{n}D[\bar{x}]=\sigma^2-\frac{1}{n}\sigma^2=\frac{n-1}{n}\sigma^2\neq\sigma^2$ (đpcm).

Ký hiệu $s_x^*=\sqrt{\frac{n}{n-1}}s_x$ khi đó $M\left[(s_x^*)^2\right]=\frac{n}{n-1}M\left[s_x^2\right]=\sigma^2$

Như vậy, khác với s_x, s_x^* là một ước lượng vững và không chệch của σ_x. Chính vì lẽ đó, khi dung lượng mẫu n bé thay cho s_x người ta thường sử dụng s_x^*. Tuy nhiên, nếu n đủ lớn thì tỷ số $\frac{n}{n-1}\approx 1$ nên hầu như không có sự khác nhau đáng kể giữa s_x và s_x^*.

2.6 Một số đặc trưng thông dụng khác

2.6.1 Độ bất đối xứng

Độ bất đối xứng được ký hiệu là A_s và được xác định bởi:

$$A=\frac{m_3}{s_x^3}=\frac{\frac{1}{n}\sum_{t=1}^{n}(x_t-\bar{x})^3}{s_x^3} \qquad (2.6.1)$$

trong đó m_3 là mômen trung tâm bậc 3 và s_x độ lệch chuẩn của X.

Hệ số bất đối xứng A là ước lượng thống kê của độ bất đối xứng $A_s=\frac{\mu_3}{\sigma_x^3}$. Nếu đại lượng ngẫu nhiên có phân phối đối xứng thì $\mu_3 = 0$, ngược lại thì $\mu_3 \neq 0$. Do đó độ bất đối xứng A là đại lượng dùng làm thước đo mức độ thiếu cân đối của phân bố thực nghiệm, phản ánh sự phân bố không đồng đều của các thành phần trong chuỗi xung quanh tâm phân phối – giá trị trung bình số học.

Nếu A>0 thì mật độ phân bố có dạng đuôi lệch phải, đặc trưng cho sự tản mản của các thành phần có trị số lớn hơn trung bình số học; nếu A<0 thì mật độ phân bố có dạng đuôi lệch trái, đặc trưng cho sự phân tán của các thành phần có trị số nhỏ hơn trung bình số học.

2.6.2 Hệ số độ nhọn

Độ nhọn $E_s=\frac{\mu_4}{\sigma_x^4}-3$ là đại lượng đặc trưng cho mức độ tập trung của phân phối. Nó phản ánh tình trạng tập trung hay phân tán của các giá trị của đại lượng ngẫu nhiên xung quanh tâm phân phối. Hệ số nhọn là ước lượng của độ nhọn, dùng làm thước đo mức độ tập trung của các thành phần trong chuỗi xung quanh giá trị trung bình.

Ký hiệu hệ số độ nhọn là E, ta có:

$$E=\frac{m_4}{s_x^4}-3 = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t-\bar{x})^4}{s_x^4} - 3 \qquad (2.6.2)$$

trong đó m_4 là mômen trung tâm bậc 4. E càng lớn thì phân phối càng tập trung, hàm mật độ càng có dạng "nhọn", mức độ tản mạn của các thành phần trong chuỗi sẽ nhỏ.

2.6.3 Độ lệch trung bình tuyệt đối.

Một trong những đặc trưng phản ánh mức độ phân tán của các thành phần trong chuỗi là độ lệch trung bình tuyệt đối, hay còn được gọi là độ lệch tuyệt đối. Ký hiệu độ lệch trung bình là v_a, ta có:

$$v_a = \frac{1}{n}\sum_{t=1}^{n}\left|x_t-\bar{x}\right| \qquad (2.6.3)$$

trong đó $\left|x_t-\bar{x}\right|$ là giá trị tuyệt đối của độ lệch của các thành phần trong chuỗi so với trung bình số học.

Đôi khi người ta còn dùng khái niệm độ lệch trung bình tương đối v_r để đặc trưng cho tương quan so sánh giữa mức độ dao động và độ lớn của chuỗi:

$$v_r = \frac{v_a}{\bar{x}} \tag{2.6.4}$$

2.6.4 Hệ số biến thiên

Hệ số biến thiên, còn được gọi là biến suất tương đối hay hệ số biến động, là tỷ số giữa độ lệch tiêu chuẩn và trung bình số học. Hệ số biến thiên là đại lượng phản ánh tương quan so sánh giữa mức độ dao động trung bình s_x và độ lớn của chuỗi $\bar{x}$.

Ký hiệu hệ số biến thiên là C_v ta có:

$$C_v = \frac{s_x}{\bar{x}} \tag{2.6.5}$$

Trong tính toán thực hành người ta thường lấy đơn vị đo C_v là phần trăm (%) nên công thức (2.6.5) có thể được viết dưới dạng khác:

$$C_v = \frac{s_x}{\bar{x}}.100\% \tag{2.6.6}$$

2.6.5 Biên độ

Biên độ của chuỗi là hiệu giữa giá trị lớn nhất và giá trị nhỏ nhất của các thành phần trong chuỗi. Ký hiệu biên độ là Q_A, ta có:

$$Q_A = \max\{x_t, t=1..n\} - \min\{x_t, t=1..n\} = x_{max} - x_{min} \tag{2.6.7}$$

Biên độ là đại lượng đặc trưng cho mức độ dao động tối đa của chuỗi. Để có sự tương quan so sánh giữa mức độ dao động tối đa và độ lớn của chuỗi người ta còn xét tỷ số giữa biên độ và trung bình số học:

$$Q = \frac{Q_A}{\bar{x}} \tag{2.6.8}$$

2.7 Phân tích, khảo sát số liệu dựa trên các đặc trưng số

Khi phân tích khảo sát một tập mẫu bất kỳ nào đó trước hết người ta thường quan tâm đến một số tính chất cơ bản liên quan đến dạng phân bố xác suất của nó. Những tính chất này bao gồm độ tập trung, độ phân tán và tính đối xứng. Độ tập trung đặc trưng cho xu thế dồn vào tâm của các thành phần trong chuỗi, phản ánh độ lớn chung của các giá trị số liệu. Độ phân tán biểu thị mức độ biến động hoặc sự tản mạn của số liệu xung quanh giá trị tâm. Tính đối xứng mô tả mức độ phân bố

đồng đều như thế nào của các giá trị số liệu xung quanh tâm của chúng. Số liệu bất đối xứng có xu thế hoặc tản mạn hơn về bên phải (có đuôi dài về bên phải) hoặc về bên trái (có đuôi dài về bên trái). Ba tính chất nêu trên tương ứng với ba mômen thống kê đầu tiên của tập mẫu.

2.7.1 Độ tập trung

Tính chất tập trung của các thành phần trong chuỗi số liệu thường được đánh giá thông qua đặc trưng trung bình số học. Nhưng nói chung trung bình số học có độ ổn định kém, nhất là trong những trường hợp số liệu biến động mạnh và có thể có những trị số đột xuất hoặc sai số thô. Do đó, mặc dù có độ chính xác kém hơn, trong nhiều trường hợp người ta dùng trung vị thay cho trung bình số học. Ngoài ra, đôi khi người ta còn xem xét thêm cả mốt.

Đặc trưng phức tạp hơn chút ít của độ tập trung là trimean. Trimean được định nghĩa là trung bình có trọng số của trung vị và các phân vị dưới và trên, trong đó trung vị nhận hai lần trọng số lớn hơn trọng số của mỗi phân vị kia:

$$\text{Trimean} = \frac{q_{0.25} + 2q_{0.5} + q_{0.75}}{4} \tag{2.7.1}$$

Trimean thường được xem là đại lượng chứa đựng thông tin về độ lớn của tập số liệu.

Một đặc trưng khác cũng thường được sử dụng để đánh giá độ tập trung của tập số liệu là trung bình hiệu chỉnh, được xác định bởi:

$$\overline{x}_{\alpha} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_{(i)} \tag{2.7.2}$$

trong đó k, là số nguyên làm tròn của tích αn, là số thành phần bị cắt bỏ, tính từ hai đầu mút, của chuỗi trình tự; α là số phần trăm thành phần sẽ bị cắt bỏ ở mỗi đầu mút và được gọi là bậc hiệu chỉnh.

So với trung bình số học, mức độ nhạy cảm đối với các giá trị biên (các giá trị ở hai đầu mút của chuỗi trình tự) của trung bình hiệu chỉnh giảm đi do việc khử bỏ một phần những trị số nhỏ nhất và lớn nhất. Khi α=0 thì trung bình hiệu chỉnh chính là trung bình số học.

2.7.2 Độ phân tán

Đặc trưng đơn giản nhất có thể dùng làm thước đo mức độ phân tán của tập số liệu là biên độ phần tư (Interquartile range – IQR). IQR là hiệu giữa phân vị trên và phân vị dưới:

$$IQR = q_{0.75} - q_{0.25} \tag{2.7.3}$$

Có thể hiểu một cách đơn giản IQR là biên độ của 50% phần trung tâm của tập số liệu. Thực tế là nó bỏ qua 25% phần trên và 25% phần dưới của chuỗi số liệu đã sắp xếp thành chuỗi trình tự với mục đích loại bỏ những giá trị biên. Đôi khi người ta còn gọi IQR là độ tán thứ tư. IRQ phản ánh mức độ dao động cực đại của 50% số thành phần trong chuỗi xung quanh trung vị.

Thông thường, để đánh giá mức độ dao động trung bình của toàn chuỗi người ta dùng độ lệch chuẩn s_x hoặc phương sai mẫu $\widetilde{D}_x$ (công thức (2.5.1) và (2.5.2)). Tuy nhiên, cũng sẽ rất thú vị nếu ta làm phép so sánh giữa s_x và IRQ. Ta biết rằng độ lệch chuẩn là căn bậc hai của phương sai mẫu. Còn phương sai mẫu là trung bình bình phương của hiệu giữa các giá trị thành phần của chuỗi và trung bình số học của chúng. Do đó khi tính toán, thậm chí một giá trị số liệu rất lớn sẽ gây nên sự biến đổi mạnh mẽ kết quả chung, vì nó khác biệt rất lớn so với trung bình, và sự khác biệt này càng được khuyếch đại lên bởi phép tính lấy bình phương. Trong khi đó các giá trị đột xuất như vậy có thể sẽ không làm ảnh hưởng đến IRQ. Ta hãy xét ví dụ sau đây làm minh họa. Giả sử có tập số liệu {11, 12, 13, 14, 15, 16, 17, 18, 19}. Độ lệch chuẩn của chúng là 2.7, nhưng nó sẽ bị phóng đại lên thành 25.6 nếu số "19" được thay bởi số sai "91". Dễ dàng thấy rằng trong cả hai trường hợp trị số IQR không đổi và bằng 4.

Một đặc trưng khác cũng thường được sử dụng để đánh giá mức độ phân tán của tập số liệu là MAD (median absolute deviation – độ lệch trung vị tuyệt đối). Giả sử có chuỗi số liệu $\{x_t, t=1..n\}$. Bằng phép biến đổi:

$$y_t = \left|x_t - q_{0.5}\right| = \left|x_t - M_e\right| \tag{2.7.4}$$

ta nhận được chuỗi mới $\{y_t, t=1..n\}$. Khi đó MAD chính là trung vị của chuỗi y_t.

Còn một đặc trưng phức tạp hơn của độ phân tán là phương sai hiệu chỉnh. Cũng như đối với trung bình hiệu chỉnh (công thức (2.7.2)), phương sai hiệu chỉnh được tính theo công thức:

$$s_\alpha^2 = \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\left(x_{(i)} - \overline{x_\alpha}\right)^2 \tag{2.7.5}$$

trong đó k cũng là số nguyên gần nhất với αn; α là số phần trăm thành phần của chuỗi trình tự sẽ bị cắt bỏ ở mỗi đầu mút và được gọi là bậc hiệu chỉnh. Khi α=0, phương sai hiệu chỉnh đúng bằng phương sai mẫu.

Ngoài những đặc trưng kể trên, trong ứng dụng thực hành người ta còn sử dụng hệ số độ nhọn (E), độ lệch trung bình tuyệt đối (v_a), biên độ (Q_A) và hệ số biến thiên (C_v) để xem xét một cách đầy đủ hơn mức độ phân tán của tập số liệu.

2.7.3 Tính đối xứng

Tính đối xứng thường được đánh giá thông qua hệ số bất đối xứng A (công thức (2.6.1)). Tuy nhiên vẫn có thể nhận thấy đặc trưng này cũng rất nhạy cảm với những giá trị đột xuất (nếu có) của tập mẫu. Bởi vì trong biểu thức tính A, tử số là trung bình lũy thừa ba độ lệch của các thành phần chuỗi so với trung bình số học. Như vậy, so với độ lệch chuẩn, hệ số bất đối xứng thậm chí còn nhạy hơn đối với những giá trị biên. Trung bình mũ ba của độ lệch ở tử số trong (2.6.1) được chia cho luỹ thừa ba của độ lệch chuẩn để chuẩn hoá hệ số bất đối xứng thành đại lượng vô thứ nguyên, tạo cho nó có tính so sánh được khi xét nhiều tập mẫu khác nhau.

Để ý rằng luỹ thừa ba của hiệu giữa các giá trị số liệu và trung bình của chúng bảo toàn dấu của các hiệu này. Vì các hiệu được lấy luỹ thừa ba nên các giá trị số liệu ở xa nhất so với trung bình sẽ chiếm ưu thế so với các thành phần khác trong tổng ở tử số của biểu thức tính A (2.6.1). Nếu có một vài giá trị số liệu rất lớn độ bất đối xứng sẽ có xu hướng dương. Bởi vậy tập số liệu có đuôi kéo dài về bên phải được xem là lệch phải và có độ bất đối xứng dương (A>0). Các đại lượng mà giá trị của chúng bị chặn dưới (như lượng giáng thuỷ hoặc tốc độ gió – giá trị của chúng phải không âm) thường có độ bất đối xứng dương. Ngược lại, với những đại lượng mà giá trị của chúng có thể có một vài trị số rất nhỏ (hoặc âm lớn) thì nhữug giá trị này sẽ cách xa trung bình về phía dưới. Tổng ở tử số trong (2.6.1) khi đó sẽ bị lấn át bởi các hạng tử âm lớn, vì vậy hệ số bất đối xứng sẽ âm (A<0). Trong trường hợp này chuỗi số liệu sẽ có đuôi kéo dài về bên trái (có xu hướng lệch trái). Nếu chuỗi số liệu về cơ bản phân bố đối xứng thì hệ số bất đối xứng sẽ gần bằng 0.

Ngoài hệ số bất đối xứng người ta còn dùng chỉ số Yule–Kendall sau đây:

$$\gamma_{yk} = \frac{(q_{0.75} - q_{0.5}) - (q_{0.5} - q_{0.25})}{IQR} = \frac{q_{0.25} - 2q_{0.5} + q_{0.75}}{IQR} \qquad (2.7.6)$$

Chỉ số Yule–Kendall đánh giá tính đối xứng của chuỗi số liệu trên cơ sở so sánh khoảng cách giữa phân vị trên và trung vị với trung vị và phân vị dưới. Nếu chuỗi số liệu có xu hướng lệch phải, ít nhất trong 50% số liệu ở tâm, khoảng cách từ phân vị trên đến trung vị sẽ lớn hơn khoảng cách từ trung vị đến phân vị dưới. Trong trường hợp này chỉ số Yule–Kendall sẽ dương (γ_{yk}>0), phù hợp với quan niệm thông thường là lệch phải thì dương. Ngược lại, chuỗi lệch trái sẽ được đặc trưng bởi chỉ số Yule–Kendall âm (γ_{yk}<0). Tương tự như khi tính hệ số bất đối xứng, việc chia cho biên độ phần tư IQR trong (2.7.6) nhằm vô thứ nguyên hoá γ_{yk}, tạo khả năng so sánh của nó khi xem xét nhiều tập số liệu khác nhau.

CHƯƠNG 3
MỘT SỐ PHÂN BỐ LÝ THUYẾT

3.1 Khái niệm mở đầu

Trong chương 2 ta đã nghiên cứu một số phương pháp phân tích, khảo sát số liệu dựa trên các đặc trưng thống kê thông thường. Về bản chất, các phương pháp đó cho phép chỉ ra những thuộc tính của các đặc trưng yếu tố khí tượng, khí hậu căn cứ vào những tập số liệu cụ thể thu thập được từ quan trắc thực tế. Tuy nhiên, do hạn chế của dung lượng mẫu, trong nhiều trường hợp những kết quả nhận được có thể sẽ phản ánh không chính xác bản chất của quá trình được xét. Chẳng hạn, khi nghiên cứu nhiệt độ tối cao ở một khu vực nào đó, trong chuỗi số liệu hiện có phạm vi biến đổi của nó là 25°C–39°C. Khi tiến hành xây dựng hàm phân bố thực nghiệm theo phương pháp chia khoảng, tần suất xuất hiện nhiệt độ tối cao trong khoảng từ 27–28°C bằng 0. Xét về mặt vật lý, điều đó là vô lý, vì với khoảng biến thiên của nhiệt độ là 25°C–39°C thì sự kiện nhiệt độ rơi vào khoảng 27–28°C không thể không xảy ra. Rõ ràng ở đây không phải do bản chất của yếu tố nhiệt độ tối cao mà là do chuỗi số liệu của chúng ta chưa đủ để bao quát hết sự biến thiên của nó.

Để khắc phục tình trạng đó, đồng thời với việc nghiên cứu các tập mẫu, chúng ta sẽ sử dụng các phân bố lý thuyết và xấp xỉ các phân bố thực nghiệm bởi những phân bố lý thuyết phù hợp. Việc sử dụng phân bố lý thuyết làm xấp xỉ cho phân bố thực nghiệm cũng có nghĩa là chúng ta đã lý tưởng hóa tập số liệu thực nghiệm, tức là ép buộc các kết quả thực nghiệm vào một lớp hàm toán học cụ thể phù hợp với chúng. Tất nhiên, đây chỉ là sự biểu diễn gần đúng các số liệu thực nghiệm, mặc dù trong rất nhiều trường hợp sự biểu diễn này cho độ chính xác rất cao. Về cơ bản có ba ưu điểm khi sử dụng các phân bố lý thuyết:

– Phân bố lý thuyết cho phép biểu diễn một cách cô đọng, ngắn gọn những thông tin từ tập mẫu thông qua dạng và một vài tham số phân bố. Trong nhiều trường hợp, chúng ta phải lặp đi lặp lại những tính toán thống kê các đặc trưng mẫu cho một địa điểm hoặc một vùng không gian nhất định nào đó. Quá trình tính toán đó có thể rất cồng kềnh, thậm chí xảy ra những sai sót bất thường. Nếu tồn

tại một phân bố lý thuyết phù hợp tốt với tập số liệu, thay cho việc khảo sát đầy đủ *n* bậc thống kê $\{x_1, x_2,...,x_n\}$ ta chỉ cần một vài tham số của phân bố này.

– Phân bố lý thuyết cho phép làm trơn và nội suy các đặc trưng xác suất. Rõ ràng số liệu thực nghiệm phụ thuộc vào dung lượng mẫu. Như đã nêu ở trên, sự hạn chế của dung lượng mẫu có thể dẫn đến sự gián đoạn hoặc đứt quảng trong phân bố thực nghiệm. Việc xấp xỉ phân bố thực nghiệm bởi một phân bố lý thuyết cho tập mẫu tạo khả năng liên tục hóa những khoảng không có số liệu, từ đó cho phép ước lượng xác suất trong những khoảng này.

– Phân bố lý thuyết cho phép tính toán ngoại suy các đặc trưng xác suất. Do sự hạn chế của dung lượng mẫu, phân bố thực nghiệm chỉ có thể phản ánh được sự biến đổi của đặc trưng yếu tố trong phạm vi biến đổi của tập mẫu. Việc ước lượng xác suất cho những sự kiện nằm ngoài phạm vi của tập mẫu đòi hỏi phải chấp nhận những giả thiết về cách xử lý như là chưa có số liệu quan trắc. Hãy trở lại ví dụ trên đây, với khoảng biến thiên của nhiệt độ tối cao là 25°C–39°C, ta sẽ không có cơ sở nào để phán đoán về các sự kiện nhiệt độ tối cao lớn hơn 39°C hoặc nhỏ hơn 25°C (mặc dù trên thực tế chúng có thể xảy ra) nếu chúng ta không xấp xỉ phân bố thực nghiệm bởi một phân bố lý thuyết.

Cũng cần nhấn mạnh rằng, việc xấp xỉ phân bố thực nghiệm bởi một phân bố lý thuyết là một quá trình xử lý tinh tế. Sau khi xây dựng hàm phân bố thực nghiệm, ta cần phải xem xét, khảo sát tỷ mỷ và lựa chọn một trong các lớp hàm lý thuyết sao cho nó phù hợp nhất với phân bố thực nghiệm. Mặt khác, để tránh sự nhầm lẫn đáng tiếc ta cần phân biệt rõ hai khái niệm: các tham số của phân bố và các tham số (hay đặc trưng) thống kê. Các tham số của phân bố là những đại lượng không ngẫu nhiên mà trước đây chúng ta đã chú thích gọi chúng là các đặc trưng tổng thể, còn các tham số thống kê là những đại lượng ngẫu nhiên, chúng được rút ra từ quá trình xử lý tính toán trên tập mẫu.

3.2 Phân bố nhị thức

Ta hãy trở lại bài toán trong mục 1.3, chương 1. Mỗi một phép thử trong *n* phép thử độc lập chỉ có 2 kết cục là A và $\overline{A}$. Xác suất xuất hiện sự kiện A ở mỗi phép thử không đổi, bằng p và không phụ thuộc vào chỉ số phép thử. Nếu ta xét biến ngẫu nhiên X_i liên quan đến kết quả của lần thử thứ *i* như sau:

$$X_i = \begin{cases} 1 & \text{nếu A xuất hiện ở lần thử thứ i} \\ 0 & \text{nếu } \overline{A} \text{ xuất hiện (A không xuất hiện) ở lần thử thứ i} \end{cases} \quad (i=1..n)$$

Vì các lần thử là độc lập nên các X_i là những biến ngẫu nhiên độc lập và có phân bố xác suất được cho bởi:

X_i	0	1
p	q = 1–p	p

Do đó biến ngẫu nhiên $X = \sum_{i=1}^{n} X_i$ chỉ số lần xuất hiện sự kiện A trong loạt *n* phép thử và sẽ có phân bố dạng:

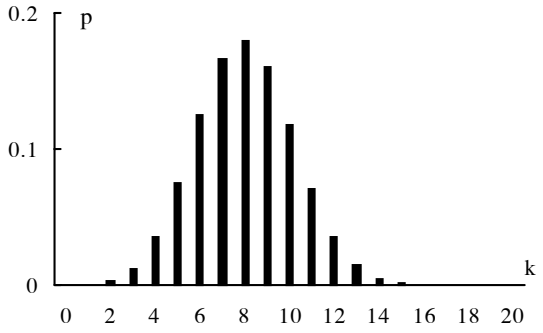
X	0	1	...	n–1	n
p	p_0	p_1	...	p_{n-1}	p_n

trong đó $p_k = C_n^k p^k q^{n-k}$.

Một cách tổng quát, có thể biểu diễn phân bố của X bởi:

$$P(X=k) = P_n(k) = C_n^k p^k q^{n-k},\ k=0,1,...,n \qquad (3.2.1)$$

Phân bố dạng (3.2.1) được gọi là phân bố nhị thức, biến ngẫu nhiên X trong trường hợp này được gọi là biến ngẫu nhiên có phân bố nhị thức. Rõ ràng phân bố nhị thức phụ thuộc vào hai tham số là n và p. Đồ thị hàm mật độ xác suất của X được trình bày trên hình 3.1.



Hình 3.1 Hàm mật độ phân bố nhị thức với n=20, p=0.4

Ví dụ 3.2 Xét sự kiện A là lượng mưa tháng 7 ở một trạm vượt quá 400 mm. Số liệu thống kê trong bảng 3.1 dẫn ra những năm có A xuất hiện trong 105 năm quan trắc. Hãy tính xác suất để trong 10 năm quan trắc: a) Có 1 năm mà lượng mưa tháng 7 vượt quá 400 mm; b) Có ít nhất 1 năm mà lượng mưa tháng 7 vượt quá 400 mm.

Từ bảng 3.1, trong 105 năm quan trắc có tất cả 19 năm xuất hiện sự kiện A. Vậy ước lượng xác suất của A là P(A)=p=19/105=0.181. Theo yêu cầu của bài toán, ta có n=10, p=0.181. Do đó, áp dụng (3.2.1) ta được:

a) Xác suất để trong 10 năm quan trắc có 1 năm mà lượng mưa tháng 7 vượt quá 400 mm sẽ là: $P(X=1) = P_{10}(1) = C_{10}^{1}(0.181)^{1}(1-0.181)^{9} = 0.3001$.

b) Xác suất để trong 10 năm quan trắc có ít nhất 1 năm mà lượng mưa tháng 7 vượt quá 400 mm sẽ là:

$P(X=1)+P(X=2)+...+P(X=10) = P(X\geq 1) = 1-P(X=0) = 0.8642$.

Bảng 3.1 Những năm có lượng mưa tháng 7 trên 400 mm trong thời gian quan trắc 105 năm

1892	1904	1928	1935	1960
1894	1914	1929	1939	1965
1899	1926	1933	1942	1967
1902	1927	1934	1943	

3.3 Phân bố Poisson

Phân bố Poisson được dùng để mô tả số sự kiện xuất hiện trong một chuỗi liên tiếp các sự kiện rời rạc cùng loại độc lập nhau. Thông thường sự liên tiếp của chuỗi các sự kiện được hiểu theo nghĩa thời gian, như sự xuất hiện các cơn bão trên một vùng biển nào đó trong mùa bão, hoặc sự xảy ra những năm hạn hán hay rét đậm. Tuy nhiên phân bố Poisson cũng có thể được áp dụng để tính xác suất xuất hiện sự kiện trong một hoặc một số vùng không gian nhất định, chẳng hạn, xác định sự phân bố của các cây xăng dọc theo một con đường cao tốc hay phân bố của những cục mưa đá trên một vùng nhỏ hẹp nào đó.

Khi xét chuỗi các sự kiện theo thời gian phân bố Poisson được áp dụng nếu thỏa mãn các điều kiện sau:

– Xác suất xuất hiện sự kiện vào khoảng thời gian đang xét phụ thuộc vào số các sự kiện và độ dài khoảng thời gian nhưng không phụ thuộc vào thời điểm đầu của khoảng.

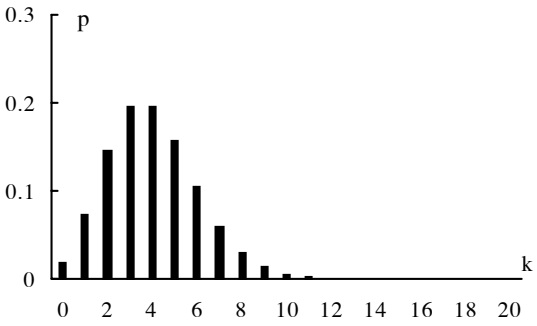
– Xác suất của số lần xuất hiện sự kiện trong khoảng thời gian đang xét không phụ thuộc vào sự xuất hiện sự kiện trước thời điểm ban đầu.

– Xác suất xuất hiện hai hay nhiều sự kiện vào một khoảng thời gian vô cùng bé nhỏ hơn rất nhiều so với xác suất xuất hiện một sự kiện trong khoảng đó.

Nếu giả thiết rằng, trong phân bố nhị thức (3.2.1) xác suất xuất hiện sự kiện A phụ thuộc vào số lần thử n sao cho khi $n\to\infty$ mà $P(A)=p\to 0$ và $np\to\lambda=\text{const}$, thì phân bố nhị thức sẽ tiệm cận đến phân bố Poisson:

$$P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!},\ k=0,1,2,... \qquad (3.3.1)$$

Rõ ràng phân bố Poisson chỉ phụ thuộc vào một tham số λ, nó có thứ nguyên là số lần xuất hiện trên một đơn vị thời gian. Đồ thị hàm mật độ xác suất của phân bố Poisson được dẫn ra trên hình 3.2.

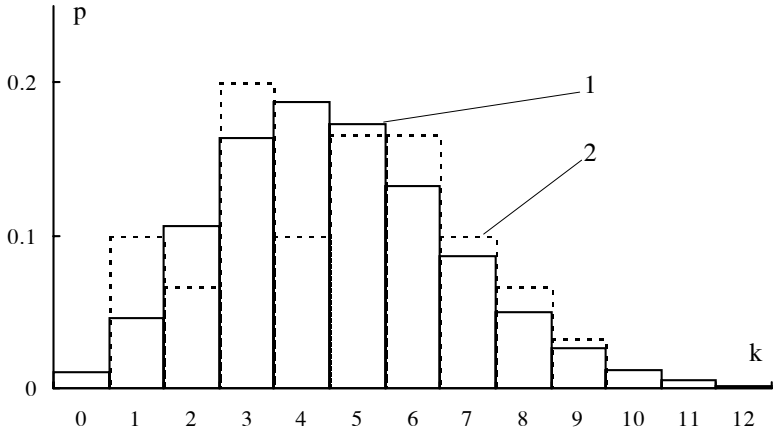


Hình 3.2 Hàm mật độ phân bố Poisson với λ=4

Ví dụ 3.3 Bảng 3.2 dẫn ra số liệu về số lần xuất hiện lốc hàng năm ở một địa phương trong vòng 30 năm quan trắc, từ 1959 đến 1988. Gọi X là biến ngẫu nhiên chỉ số lần xuất hiện lốc hàng năm ở đây và giả thiết rằng X có phân bố Poisson. Ta thấy, tổng số có 138 lần xuất hiện lốc trong 30 năm, vậy trung bình hàng năm có 138/30 = 4.6 (lần/năm). Nếu lấy giá trị này làm ước lượng của tham số λ trong phân bố Poisson, ta có thể sử dụng công thức (3.3.1) để tính xác suất số lần xuất hiện lốc hàng năm cho địa phương nói trên. Hình 3.3 biểu diễn đồ thị hàm mật độ xác suất lý thuyết của phân bố Poisson với λ=4.6 và mật độ xác suất thực nghiệm tính theo số liệu ở bảng 3.2.

Bảng 3.2 Số lần xuất hiện lốc hàng năm

Năm	Số lần	Năm	Số lần	Năm	Số lần
1959	3	1969	7	1979	3
1960	4	1970	4	1980	4
1961	5	1971	5	1981	3
1962	1	1972	6	1982	3
1963	3	1973	6	1983	8
1964	1	1974	6	1984	6
1965	5	1975	3	1985	7
1966	1	1976	7	1986	9
1967	2	1977	5	1987	6
1968	2	1978	8	1988	5



Hình 3.3 Biểu đồ biểu diễn mật độ xác suất xuất hiện lốc

1. Lý thuyết; 2. Thực nghiệm

Từ hình 3.3 có thể nhận thấy rằng mật độ xác suất lý thuyết đạt giá trị lớn nhất khi k=4 (hàng năm có 4 lần xuất hiện lốc). Trong khi đó, theo kết quả thực nghiệm, xác suất để hàng năm có 3 lần xuất hiện lốc đạt giá trị lớn nhất. Hơn nữa, cũng theo phân bố thực nghiệm, xác suất khi k=4 nhỏ hơn rất nhiều so với khi k=3 và k=5. Xét về ý nghĩa vật lý, điều đó hoàn toàn khó lý giải. Tình huống xảy ra tương tự khi so sánh k=2 với k=1 và k=3. Rõ ràng, trong trường hợp này việc xấp xỉ phân bố thực nghiệm bởi phân bố lý thuyết đã tạo cho ta khả năng phán đoán và nhận định tốt hơn mà không lệ thuộc vào kết quả thực nghiệm.

3.4 Phân bố chuẩn và phân bố chuẩn chuẩn hoá

Phân bố chuẩn, hay còn gọi là phân bố Gauss, đóng vai trò hết sức quan trọng trong thống kê cổ điển, nó được ứng dụng rộng rãi và hiệu quả trong khí tượng, khí hậu.

Biến ngẫu nhiên X được gọi là có phân bố chuẩn nếu hàm mật độ xác suất của nó có dạng:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad (3.4.1)$$

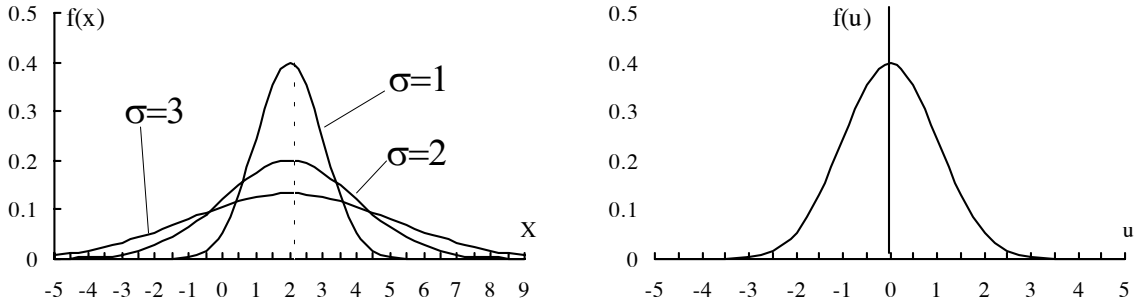
Như vậy, phân bố chuẩn phụ thuộc vào hai tham số μ và σ (nên người ta thường ký hiệu $X \in N(\mu,\sigma)$ để chỉ biến ngẫu nhiên X có phân bố chuẩn với hai tham số μ, σ). Có thể chứng minh được rằng các tham số này chính là kỳ vọng toán học và độ lệch bình phương trung bình (căn bậc hai của phương sai) của X:

$$M[X] = \int_{-\infty}^{+\infty} xf(x)dx = \mu \qquad (3.4.2)$$

$$D[X] = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx = \sigma^2 \qquad (3.4.3)$$

Từ (3.4.1) suy ra rằng mật độ phân bố chuẩn được xác định trên toàn miền của trục số và đồ thị của nó nhận đường x=μ làm trục đối xứng (hình 3.4a).

Để sử dụng phân bố chuẩn biểu diễn một tập số liệu ta cần ước lượng chính xác hai tham số μ và σ. Như đã được biết trong chương 2, các ước lượng này là mômen gốc mẫu bậc nhất $\bar{x}$ và độ lệch chuẩn s^*. Ta hãy xét thêm một vài đặc trưng khác của phân bố chuẩn.



(a) (b)

Hình 3.4 Hàm mật độ phân bố chuẩn với μ=2 và các giá trị σ khác nhau (a) và phân bố chuẩn chuẩn hóa (b)

Mômen trung tâm bậc lẻ của phân bố chuẩn được xác định bởi:

$$\mu_{2r+1} = \int_{-\infty}^{+\infty} (x-\mu)^{2r+1} f(x)dx = 0 \qquad (3.4.4)$$

Từ đó thấy rằng, do tính chất đối xứng của hàm mật độ, các mômen trung tâm bậc lẻ đều bằng 0. Đương nhiên ta có độ bất đối xứng $A_s=\mu_3/\sigma^3=0$.

Mômen trung tâm bậc chẵn:

$$\mu_{2r} = \int_{-\infty}^{+\infty} (x-\mu)^{2r} f(x)dx = \frac{1}{\sqrt{\pi}} 2^r \sigma^{2r} \Gamma(r+\frac{1}{2}) \qquad (3.4.5)$$

Hay $$\mu_{2r}=1.3.5...(2r-1)\sigma^{2r}=(2r-1)!!\sigma^{2r} \qquad (3.4.5')$$

Khi $$r=1: \mu_{2r} = \mu_2 = \sigma^2 = D[X]$$

$$r=2: \mu_{2r} = \mu_4 = 3\sigma^4$$

Ta nhận thấy độ nhọn của phân bố chuẩn $E_s = \mu^4/\sigma^4 - 3 = 0$. Và như vậy, hệ số độ nhọn được chỉ ra trong mục 2.6.2 sẽ còn mang ý nghĩa so sánh một phân bố nào đó "nhọn" hơn hay "tù" hơn so với phân bố chuẩn.

Tương ứng với hàm mật độ (3.4.1) ta có hàm phân bố xác suất:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2} dt \qquad (3.4.6)$$

Xác suất để đại lượng ngẫu nhiên X nhận giá trị trong khoảng (α;β) được xác định bởi:

$$P(\alpha < X < \beta) = \frac{1}{\sigma\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \Phi\left(\frac{\beta-\mu}{\sigma}\right) - \Phi\left(\frac{\alpha-\mu}{\sigma}\right)$$

Hay
$$P(\alpha<X<\beta) = \Phi\left(\frac{\beta-\mu}{\sigma}\right) - \Phi\left(\frac{\alpha-\mu}{\sigma}\right) \qquad (3.4.7)$$

trong đó
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-\frac{1}{2}t^2} dt \qquad (3.4.8)$$

là hàm Laplas.

Dễ thấy rằng hàm Laplas là một hàm lẻ, $\Phi(x) = -\Phi(-x)$ và khi $x \to \infty$ thì $\Phi(x) \to \frac{1}{2}$. Do đó ta có thể biểu diễn hàm phân bố (3.4.6) qua hàm Laplas:

$$F(x) = \frac{1}{2} + \Phi\left(\frac{x-\mu}{\sigma}\right) \qquad (3.4.9)$$

Từ (3.4.7) suy ra xác suất để đại lượng ngẫu nhiên X nhận giá trị trong khoảng đối xứng đối với kỳ vọng toán học (μ−ε; μ+ε) là:

$$P(|X-\mu| < \varepsilon) = \Phi(\frac{\varepsilon}{\sigma}) - \Phi(-\frac{\varepsilon}{\sigma}) = 2\Phi(\frac{\varepsilon}{\sigma}) \qquad (3.4.10)$$

Hay
$$P(|X-\mu| > \varepsilon) = 1 - 2\Phi(\frac{\varepsilon}{\sigma}) \qquad (3.4.10')$$

Trong ứng dụng thực hành người ta thường lập bảng tính sẵn giá trị của hàm Φ(x).

Nếu $X \in N(\mu,\sigma)$ thì biến ngẫu nhiên U nhận được qua phép biến đổi:

$$U = \frac{X-\mu}{\sigma}$$

cũng sẽ có phân bố chuẩn với hai tham số μ=0 và σ =1 và được ký hiệu là U∈N(0,1). Hàm mật độ phân bố của U nhận được từ biểu thức (3.4.1) bằng cách thay $\frac{x-\mu}{\sigma}$=u:

$$f(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \tag{3.4.11}$$

Khi đó hàm phân bố (3.4.6) sẽ có dạng:

$$F(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{1}{2}t^2} dt \tag{3.4.12}$$

Các hệ thức (3.4.11) và (3.4.12) được gọi là hàm mật độ và hàm phân bố chuẩn chuẩn hóa. Hàm (3.4.11) là một hàm chẵn, đồ thị của nó có dạng đối xứng với trục đối xứng là trục tung (hình 3.4b).

Trong thực tế để áp dụng phân phối chuẩn người ta thường thực hiện phép biến đổi chuỗi số liệu ban đầu về dạng chuẩn hóa:

$$u = \frac{x - \bar{x}}{\sigma}$$

Khi đó chuỗi mới nhận được sẽ có trung bình bằng 0 và phương sai bằng 1. Phép biến đổi này trong nhiều trường hợp có thể làm cho một biến nào đó từ chỗ không tuân theo luật phân bố chuẩn trở thành có phân bố chuẩn hoặc gần chuẩn.

Phân bố chuẩn là một trong những phân bố được ứng dụng hết sức phổ biến. Trong khí tượng, khí hậu phân bố chuẩn và phân bố chuẩn chuẩn hoá thường được dùng trong xử lý số liệu, trong kiểm nghiệm sự bằng của các tham số và làm công cụ trung gian để kiểm nghiệm sự phù hợp giữa phân bố thực nghiệm và phân bố lý thuyết.

Phân bố chuẩn được Moivre [4] tìm thấy lần đầu tiên vào năm 1733 khi ông nghiên cứu giới hạn của phân bố nhị thức. Sau đó nó lại được phát hiện bởi Gauss (1809) và Laplace (1812).

3.5 Phân bố Gamma

Nhiều biến khí quyển có tính bất đối xứng khác nhau và thường phân bố lệch phải. Thông thường sự lệch phải xuất hiện đối với những biến mà giá trị của chúng bị chặn trái, chẳng hạn lượng mưa và tốc độ gió là những yếu tố không âm. Trong những trường hợp này việc xấp xỉ phân bố của chúng bởi luật chuẩn sẽ không có hiệu quả. Hãy lấy ví dụ sau đây làm minh họa. Xét yếu tố tổng lượng mưa tháng 1 ở một trạm cho ở bảng 3.3.

Bảng 3.3 Số liệu tổng lượng mưa R tháng 1 (mm)

Năm	R	Năm	R	Năm	R	Năm	R	Năm	R
1933	11.2	1943	34.3	1953	64.3	1963	33.3	1973	36.6
1934	30.0	1944	13.7	1954	50.8	1964	44.7	1974	46.7
1935	68.3	1945	69.6	1955	28.4	1965	55.1	1975	42.9
1936	52.8	1946	28.7	1956	54.1	1966	60.5	1976	76.2
1937	93.0	1947	63.5	1957	34.5	1967	29.5	1977	34.5
1938	43.7	1948	43.7	1958	124.5	1968	35.3	1978	161.8
1939	71.6	1949	57.7	1959	74.7	1969	34.5	1979	115.6
1940	18.3	1950	71.6	1960	44.5	1970	26.2	1980	13.2
1941	37.1	1951	50.3	1961	42.9	1971	28.2	1981	22.1
1942	33.0	1952	62.0	1962	47.8	1972	34.3	1982	38.4

Từ tập số liệu này ta tính được $\bar{x}$ = 49.8 và s^* = 28.3. Nếu sử dụng phân bố chuẩn làm xấp xỉ phân bố lý thuyết ta dễ dàng tính được xác suất sự kiện lượng mưa tháng 1 nhỏ hơn 0:

$$P(X<0) = F(0) = \frac{1}{28.3\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{1}{2}(\frac{t-49.8}{\sigma})^2} dt = 0.04$$

Mặc dù xác suất này rất nhỏ nhưng vẫn khác không, điều đó có nghĩa là sự kiện đang xét vẫn có thể xảy ra! Sự vô lý này đương nhiên là không chấp nhận được, tức là không thể sử dụng phân bố chuẩn trong trường hợp này.

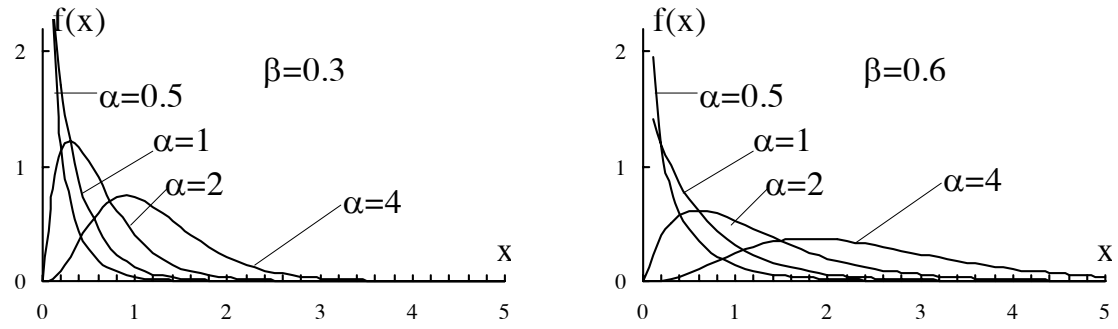
Để giải quyết những vấn đề tương tự trên đây, người ta thường chọn phân bố Gamma, đặc biệt trong nghiên cứu các chuỗi số liệu lượng mưa. Hàm mật độ xác suất của phân bố Gamma có dạng:

$$f(x) = \frac{(x/\beta)^{\alpha-1}\exp(-x/\beta)}{\beta\Gamma(\alpha)} \quad \text{với } x, \alpha, \beta>0 \qquad (3.5.1)$$

Hoặc dưới dạng khác:

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1}\exp(-x/\beta) \qquad (3.5.1')$$

Phân bố Gamma phụ thuộc vào hai tham số α và β. Tham số α đặc trưng cho dáng điệu (hình dạng) của đường cong đồ thị hàm mật độ, còn tham số β phản ánh mức độ "co, duỗi" của đồ thị. Hình 3.5 dẫn ra đồ thị của mật độ phân bố Gamma ứng với các trường hợp α và β khác nhau.



Hình 3.5 Hàm mật độ phân bố Gamma

Từ hình 3.5 ta nhận thấy rằng, khi $\alpha<1$ phân bố Gamma lệch rất mạnh và $f(x)\to\infty$ khi $x\to 0$. Khi $\alpha=1$ đồ thị sẽ cắt trục tung tại điểm $1/\beta$ (khi x=0). Với những giá trị $\alpha>1$ đồ thị hàm mật độ xuất phát từ gốc toạ độ (0; 0) và phân bố Gamma sẽ tiệm cận đến phân bố chuẩn khi α nhận giá trị rất lớn.

Phân bố Gamma có kỳ vọng toán học bằng tích $\alpha.\beta$ và phương sai bằng $\alpha.\beta^2$. Các ước lượng của tham số α và β được xác định bởi các hệ thức sau đây:

$$\tilde{\alpha}=\frac{\bar{x}^{2}}{(s^{*})^{2}} \quad \text{và} \quad \tilde{\beta}=\frac{(s^{*})^{2}}{\bar{x}} \tag{3.5.2}$$

Hoặc:
$$\tilde{\alpha}=\frac{1+\sqrt{1+4D/3}}{4D} \quad \text{và} \quad \tilde{\beta}=\frac{\bar{x}}{\tilde{\alpha}} \tag{3.5.3}$$

Với $D=\ln(\bar{x}-\frac{1}{n}\sum_{i=1}^{n}\ln(x_i))$

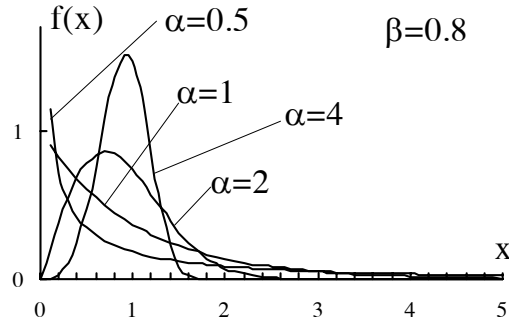
3.6 Phân bố Weibull

Một dạng phân bố khác cũng thường được sử dụng trong khí tượng, khí hậu là phân bố Weibull. Phân bố Weibull được ứng dụng nhiều nhất trong nghiên cứu sự biến đổi của tốc độ gió, đặc biệt là gió mặt đất. Hàm mật độ phân bố Weibull có dạng:

$$f(x)=\left(\frac{\alpha}{\beta}\right)\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right], \text{ với } x,\ \alpha,\ \beta>0 \tag{3.6.1}$$

Hoặc:
$$f(x)=\left(\frac{\alpha}{\beta^{\alpha}}\right)x^{\alpha-1}\exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right] \tag{3.6.1'}$$

Đồ thị hàm mật độ xác suất của phân bố Weibull được dẫn ra trên hình 3.6. Kỳ vọng toán học của phân bố Weibull bằng $\beta\Gamma(1+1/\alpha)$ và phương sai bằng $\beta^2(\Gamma(1+2/\alpha)-\Gamma^2(1+1/\alpha))$.



Hình 3.6 Hàm mật độ phân bố Weibull với các tham số khác nhau

3.7. Phân bố χ^2 (khi bình phương).

Trong lớp các bài toán kiểm nghiệm giả thiết thống kê phân bố χ^2 đóng một vai trò hết sức quan trọng, nó được dùng để kiểm nghiệm sự phù hợp hay không phù hợp giữa phân bố thực nghiệm và phân bố lý thuyết.

Phân bố χ^2 được xây dựng trên cơ sở nghiên cứu tổng các biến ngẫu nhiên độc lập $X_1,X_2,...,X_n$ có cùng phân bố chuẩn, $X_i \in N(\mu;\sigma)$:

$$\chi^2(n) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \qquad (3.7.1)$$

và gọi là biến ngẫu nhiên χ^2 với n tham số.

Hàm mật độ xác suất của χ^2 có dạng:

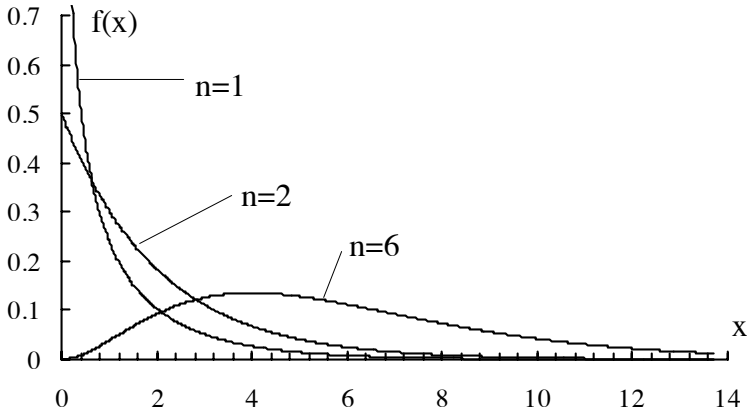
$$f_n(x) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{khi } x > 0 \\ 0 & \text{khi } x \le 0 \end{cases} \qquad (3.7.2)$$

Hàm mật độ xác suất của biến ngẫu nhiên χ^2 xác định với mọi x>0 và với mọi số nguyên dương n.

Hàm phân bố xác suất của χ^2 tương ứng với mật độ xác suất (3.7.2) sẽ bằng 0 khi x≤0, còn khi x>0 thì:

$$F_n(x) = P(\chi^2 < x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\int_0^x t^{\frac{n}{2}-1} e^{-\frac{t}{2}} dt \qquad (3.7.3)$$

Như vậy phân bố χ^2 phụ thuộc vào chỉ một tham số *n* và được gọi là bậc tự do của phân bố. Khi n≤2 hàm mật độ xác suất $f_n(x)$ luôn luôn giảm với mọi x>0, khi n>2 hàm $f_n(x)$ có cực đại duy nhất tại x=n–2. Trên hình 3.7 dẫn ra đồ thị của hàm $f_n(x)$ với 3 trường hợp n=1, n=2 và n=6.



Hình 3.7 Hàm mật độ phân bố χ^2 với các bậc tự do khác nhau

Về khái niệm *số bậc tự do n* bạn đọc có thể tìm hiểu kỹ hơn, chẳng hạn, trong [4]. Thuật ngữ này do Fisher đặt ra và nó cũng sẽ được dùng với cùng ý nghĩa đó khi xét đến một số phân bố khác sau này.

Kỳ vọng và phương sai của χ^2 bằng:

$$M[\chi^2(n)]=n \text{ va } D[\chi^2(n)]=2n \tag{3.7.4}$$

Nếu $\chi^2(n_1)$ và $\chi^2(n_2)$ là hai biến ngẫu nhiên độc lập có phân bố χ^2 với n_1 và n_2 bậc tự do thì tổng của chúng cũng là một biến ngẫu nhiên có phân bố χ^2 với (n_1+n_2) bậc tự do:

$$\chi^2(n_1) + \chi^2(n_2) = \chi^2(n_1+n_2) \tag{3.7.5}$$

Xác suất $\chi^2(n)$ nhận giá trị vượt quá một giá trị χ_0^2 cho trước được xác dịnh bởi:

$$p=P(\chi^2>\chi_0^2) = \int_{x_0^2}^{\infty} f_n(x)dx = 1-F_n(\chi_0^2) \tag{3.7.6}$$

Xác suất này chính bằng diện tích giới hạn bởi nhánh đường cong mật độ ở bên phải trục thẳng đứng đi qua điểm x= χ_0^2 và trục hoành. Do ý nghĩa sử dụng của các xác suất này nên trong thực tế người ta thường lập bảng tính sẵn giá trị của χ_p^2 ứng với các mức xác suất p và số bậc tự do *n* khác nhau.

3.8 Phân bố Student (t)

Phân bố Student thường được gọi là một cách đơn giản và quen thuộc là phân bố *t,* được xác định trên cơ sở xét biến ngẫu nhiên là tỷ số giữa hai biến ngẫu nhiên độc lập $X_1 \in N(0,1)$ và $X_2 \in \frac{\chi(n)}{\sqrt{n}}$: $t=X_1/X_2$. Biến ngẫu nhiên *t* trong trường hợp này được gọi là có phân bố Student với *n* bậc tự do và ký hiệu $t \in St(n)$ hay gọn hơn t(n).

Mật độ xác suất của phân bố Student có dạng:

$$f_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1+\frac{x^2}{n})^{\frac{n+1}{2}} \qquad (3.8.1)$$

Hoặc:

$$f_n(x) = \frac{1}{B\left(\frac{n}{2},\frac{1}{2}\right)\sqrt{n}}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \qquad (3.8.1')$$

Phân bố Student hay phân bố *t* được W.S.Gosset sử dụng lần đầu tiên trong một bài toán thống kê quan trọng [4] và được tác giả lấy biệt hiệu là Student. Hàm mật độ của biến *t* cũng chỉ phụ thuộc vào một tham số duy nhất *n* là số bậc tự do. Từ (3.8.1) hoặc (3.8.1') có thể suy ra rằng phân bố Student là một phân bố đối xứng đối với x=0. Trên hình 3.8 dẫn ra đồ thị mật độ xác suất của phân bố Student tương ứng với số bậc tự do n=3, 6 và 50.

Do tính đối xứng của phân bố, tất cả các mômen trung tâm bậc lẻ (nếu có) đều bằng 0, còn các mômen bậc chẵn được xác định bởi:

$$\mu_{2r} = \frac{1.3...(2r-1)n^r}{(n-2)(n-4)...(n-2r)} \qquad (3.8.2)$$

Khi r=1 và n>2 ta có phương sai của *t*(n) bằng:

$$D[t(n)] = D_t = \frac{n}{n-2} \qquad (3.8.3)$$

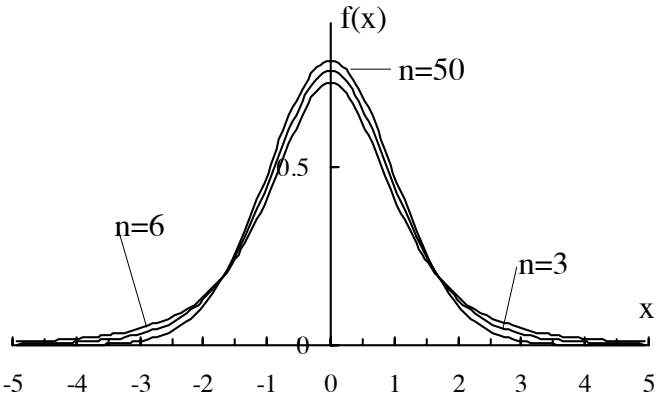
Dĩ nhiên kỳ vọng của phân bố Student bằng 0. Người ta cũng đã chứng minh được rằng khi $n \to \infty$ thì phân bố Student tiện cận đến phân bố chuẩn chuẩn hoá.

Xác suất để biến ngẫu nhiên có phân bố Student với *n* bậc tự do nhận giá trị nằm ngoài khoảng đối xứng $(-t_0; t_0)$ được tính theo công thức:

$$P(|t| > t_0) = 2\int_{t_0}^{\infty} f_n(x)dx \qquad (3.8.4)$$

trong đó $f_n(x)$ là mật độ xác suất dược cho bởi (3.8.1) hoặc (3.8.1').

Phân bố Student là một trong những phân bố được dùng để kiểm nghiệm giả thiết thống kê trong khí hậu.



Hình 3.8 Hàm mật độ phân bố Student với các bậc tự do khác nhau

3.9 Phân bố Fisher (F)

Phân bố Fisher đóng vai trò rất quan trọng trong khí tượng, khí hậu, nó thường được sử dụng để kiểm nghiệm giả thiết thống kê trong phân tích phương sai. Biến ngẫu nhiên F được gọi là có phân bố Fisher nếu hàm mật độ xác suất của nó có dạng:

$$f(x) = \frac{n_1^{\frac{n_1}{2}} n_2^{\frac{n_2}{2}} \Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} \frac{x^{\frac{n_1}{2}-1}}{(n_1 x + n_2)^{\frac{n_1+n_2}{2}}} \qquad (3.9.1)$$

Như vậy, mật độ xác suất của phân bố Fisher phụ thuộc vào hai tham số n_1 và n_2, chúng được gọi là các bậc tự do. Do đó thông thường người ta ký hiệu hàm mật độ phân bố Fisher là $f_{n1,n2}(x)$ hay $f(x,n_1,n_2)$.

Khi $n_2>2$ kỳ vọng của biến F được xác định bởi $M[F]=\frac{n_2}{n_2-1}$.

Đồ thị hàm mật độ phân bố Fisher có dạng như trên hình 3.9.

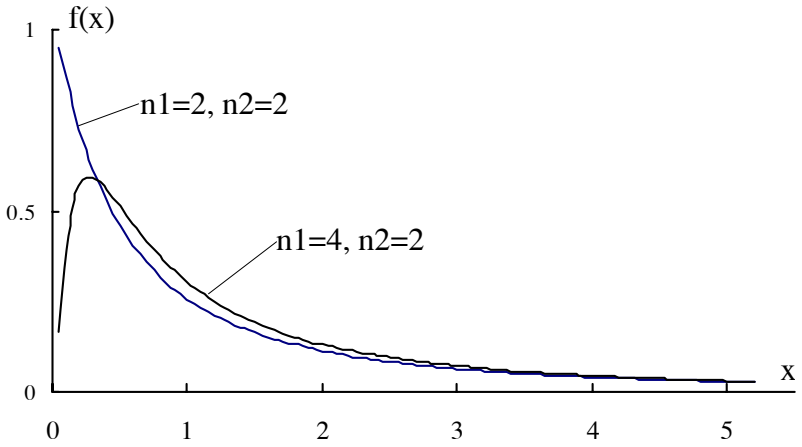
3.10 Một số phân bố khác

Những luật phân bố trên đây, trong ứng dụng thực hành, người ta còn sử dụng một số phân bố khác cho những nghiên cứu cấu trúc thống kê các chuỗi số liệu. Nói chung những yếu tố khí tượng, khí hậu mà khoảng biến thiên giá trị của chúng không thực sự rõ ràng, như nhiệt độ không khí, nhiệt độ đất, các đặc trưng

độ ẩm tuyệt đối,... thì tính bất đối xứng của phân bố thường không lớn. Chúng thường được mô tả một cách gần đúng bởi phân bố chuẩn hoặc phân bố Sarle sau đây:

$$f_s(x) = f_0(x) + \frac{1}{\sigma}\left[\frac{A_s(x)}{6} f(t)(t^3 - 3t) + \frac{E(x)}{24} f(t)(t^4 - 6t^2 + 3)\right] \qquad (3.10.1)$$

trong đó $f_s(x)$ là mật độ phân bố Sarle; $f_0(x)$ – mật độ phân bố chuẩn $t = \frac{x - \bar{x}}{\sigma}$; f(t) – mật độ phân bố chuẩn chuẩn hoá; $A_s(x)$ – độ bất đối xứng; E(x) – độ nhọn.



Hình 3.9 Hàm mật độ phân bố Fisher

Có thể nhận thấy rằng, hạng thứ hai trong (3.10.1) chính là phần hiệu chỉnh cho phân bố chuẩn. Nếu $A_s(x)=0$ và E(x) = 0 thì phần bố Sarle trùng với phân bố chuẩn.

Sử dụng phép thay thế $t=\frac{x - \bar{x}}{\sigma}$ ta có thể viết $f_0(x)=\frac{1}{\sigma} f(t)$ và khi đó phân bố Sarle sẽ có dạng:

$$f_s(t) = \frac{k}{\sigma} f(t)\left[1 + \frac{A_s(x)}{6}(t^3 - 3t) + \frac{E(x)}{24}(t^4 - 6t^2 - 3\right] \qquad (3.10.2)$$

Đối với các đặc trưng yếu tố mà khoảng biến thiên giá trị của chúng bị chặn một phía hoặc cả hai phía, như lượng mưa, độ ẩm tương đối, tầm nhìn xa, tốc độ gió,... thì qui luật phân bố của chúng thường được mô tả bởi các phân bố Gamma, Weibull, Beta, chuẩn lôga.

Các phân bố Gamma và Weibull đã xét trong các mục 3.5 và 3.6 trên đây. Sau đây ta sẽ xét phân bố chuẩn lôga và phân bố Beta.

Phân bố chuẩn lôga là một phân bố được sử dụng cho những trường hợp bất đối xứng dương (lệch phải) và có miền biến thiên dương (x>0). Thông thường nhất, phân bố chuẩn lôga được dùng để biểu diễn sự biến đổi của các đặc trưng về mây và nó cũng thường được ứng dụng rộng rãi trong thủy văn. Nếu biến ngẫu nhiên Y nhận được từ biến ngẫu nhiên X bằng phép biến đổi Y=ln(X) tuân theo luật phân bố chuẩn (phân bố Gauss) thì biến X được gọi là có phân bố chuẩn lôga với hàm mật độ xác suất có dạng:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2} \right] \tag{3.10.3}$$

trong đó hai tham số μ và σ tương ứng là kỳ vọng và độ lệch bình phương trung bình của biến đã được biến đổi Y (tức $\mu \equiv \mu_y$ và $\sigma \equiv \sigma_y$).

Giữa các tham số trong (3.10.3) và kỳ vọng và độ lệch bình phương trung bình của biến ban đầu μ_x và σ_x tồn tại mối liên hệ sau:

$$\mu_x = \exp\left[\mu_y + \frac{\sigma_y^2}{2} \right] \tag{3.10.4}$$

và

$$\sigma_x^2 = \left[\exp\left(\sigma_y^2\right) - 1 \right] \exp\left(2\mu_y + \sigma_y^2\right) \tag{3.10.5}$$

Phân bố Beta thường được áp dụng đối với những yếu tố mà miền biến thiên bị chặn cả hai phía và thường là bị giới hạn trong đoạn [0; 1]. Chẳng hạn, lượng mây được đo bằng phần mười bầu trời, hay độ ẩm tương đối. Hàm mật độ xác suất của phân bố Beta có dạng:

$$f(x) = \frac{\Gamma(p+q)}{\Gamma(p).\Gamma(q)} . x^{p-1} . (1-x)^{q-1}, \text{ với } 0 \le x \le 1 \text{ và } p, q > 0 \tag{3.10.6}$$

Như vậy, phân bố Beta cũng phụ thuộc vào hai tham số p và q. Kỳ vọng và phương sai của phân bố được xác định bởi:

$$\mu = \frac{p}{p+q} \tag{3.10.7}$$

và

$$\sigma^2 = \frac{pq}{(p+q)^2 (p+q+1)} \tag{3.10.8}$$

Trên cơ sở đó, có thể nhận được ước lượng của các tham số p và q:

$$\tilde{p} = \frac{\bar{x}^2 (1-\bar{x})}{\left(s^*\right)^2} - \bar{x} \quad \text{và} \quad \tilde{q} = \frac{\tilde{p}(1-\bar{x})}{\bar{x}} \tag{3.10.9}$$

CHƯƠNG 4
KIỂM NGHIỆM GIẢ THIẾT THỐNG KÊ TRONG KHÍ HẬU

4.1 Khái niệm về kiểm nghiệm giả thiết thống kê

4.1.1 Giả thiết thống kê và bài toán kiểm nghiệm giả thiết thống kê

Trong thực tế, khi nghiên cứu một hiện tượng nào đó thường nảy sinh vấn đề nghi hoặc giữa cái "thật" và cái "giả", giữa "đúng" và "sai", giữa cái "ngẫu nhiên" và "bản chất" của hiện tượng. Chẳng hạn, sau khi xem xét chuỗi số liệu lượng mưa ta phát hiện ra rằng "hình như kể từ khi thay đổi vị trí trạm, lượng mưa có dấu hiệu tăng lên so với trước?". Điều nghi ngờ đó có đúng hay không? Dấu hiệu lượng mưa tăng lên sau khi thay đổi vị trí trạm là bản chất hay chỉ là ngẫu nhiên? v.v... Một loạt câu hỏi tương tự được đặt ra buộc ta phải kiểm tra lại sự nghi ngờ đó. Muốn vậy ta nêu ra giả thiết "lượng mưa tăng lên kể từ khi thay đổi vị trí trạm" và tiến hành kiểm nghiệm nó. Ngược lại với giả thiết này là đối thiết "lượng mưa không tăng lên".

Từ đó bài toán kiểm nghiệm giả thiết thống kê được đặt ra dưới dạng tổng quát sau:

"Cho đại lượng ngẫu nhiên X và một giả thiết H_o về phân bố xác suất của X. Một mệnh đề khác với H_o được gọi là đối thiết H_1. Cần kiểm nghiệm xem H_o đúng hay H_1 đúng trên cơ sở tập mẫu có được $x_t=\{x_1, x_2,..., x_n\}$".

Thông thường đối thiết H_1 là phủ định của giả thiết H_o. Giả thiết H_o có thể là giả thiết đơn giản hoặc giả thiết phức tạp. Giả thiết đơn giản là giả thiết chỉ chứa một giả định. Ví dụ, H_o: $a_1=a_2$. Giả thiết phức tạp là giả thiết chứa nhiều giả định. Ví dụ, H_o: $a_1<a<a_2$.

4.1.2 Các loại sai lầm

Khi kiểm nghiệm giả thiết thống kê, việc phán đoán nói chung chỉ dựa vào một lần thực nghiệm là tập mẫu có được $\{x_1, x_2,..., x_n\}$, do đó những kết luận đưa ra có thể phạm phải sai lầm. Có hai loại sai lầm:

– Sai lầm loại I: Là sai lầm bác bỏ giả thiết H_o khi giả thiết này đúng. Chẳng hạn, giả thiết H_o: $\theta_1 = \theta_2$. Sự kiện chân thật là $\theta_1 = \theta_2$ (H_o đúng). Nhưng khi kiểm nghiệm, kết quả ta nhận được là $\theta_1 \neq \theta_2$ và đưa ra kết luận H_o sai. Như vậy ta đã phạm phải sai lầm là phủ nhận giả thiết nêu ra khi nó đúng.

– Sai lầm loại II: Là sai lầm chấp nhận giả thiết H_o khi giả thiết này sai. Ví dụ, giả thiết đưa ra là H_o: $\theta_1 = \theta_2$. Sự kiện chân thật là $\theta_1 \neq \theta_2$ (H_o sai). Nhưng khi kiểm nghiệm, kết quả ta nhận được là $\theta_1 = \theta_2$ và đưa ra kết luận H_o đúng. Sai lầm phạm phải ở đây là chấp nhận giả thiết nêu ra khi nó sai.

Ký hiệu xác suất phạm sai lầm loại I là α và xác suất phạm sai lầm loại II là β ta có thể biểu diễn chúng dưới dạng sau:

$$\alpha = P(\overline{H_o}/H_o) \text{ (Bác bỏ } H_o \text{ khi } H_o \text{ đúng)}$$

$$\beta = P(H_o/\overline{H_o}) \text{ (Chấp nhận } H_o \text{ khi } H_o \text{ sai)}$$

Nói chung quan hệ giữa α và β là ngược nhau: nếu α giảm thì β tăng và ngược lại. Khi dung lượng mẫu n càng lớn thì giá trị của α và β càng nhỏ. Bởi vậy với dung lượng mẫu *n* cố định khi tiến hành kiểm nghiệm người ta cố gắng lựa chọn được một chỉ tiêu thích hợp sao cho có thể loại trừ được cả hai loại sai lầm càng nhiều càng tốt. Tuy nhiên ta không thể cực tiểu hoá đồng thời cả α và β, vì chúng liên hệ với nhau bởi các hệ thức:

$$P(H_o/\overline{H_o}) + P(\overline{H_o}/\overline{H_o}) = 1$$

và $$P(H_o/H_o) + P(\overline{H_o}/H_o) = 1$$

Hoặc có thể biểu diễn một cách rõ ràng hơn:

Kết quả kiểm nghiệm	Thực tế H_0 đúng (H_1 sai)	Thực tế H_0 sai (H_1 đúng)
Bác bỏ H_0	Phạm sai làm loại I với xác suất $P(\overline{H_o}/H_0)=\alpha$	Quyết định đúng với xác suất $P(\overline{H_o}/H_0)=1-\alpha$
Chấp nhận H_0	Quyết định đúng với xác suất $P(\overline{H_o}/H_0)=1-\beta$	Phạm sai lầm loại II với xác suất $P(\overline{H_o}/H_0)=\beta$

4.1.3 Kiểm nghiệm tham số và kiểm nghiệm phi tham số

Người ta chia lớp các bài toán kiểm nghiệm giả thiết thống kê ra làm hai loại: kiểm nghiệm tham số và kiểm nghiệm phi tham số. Kiểm nghiệm tham số là kiểm nghiệm được hình thành khi đã biết hoặc đã chấp nhận rằng tồn tại một phân bố lý thuyết cụ thể nào đó phù hợp với phân bố của tập mẫu hiện có. Như vậy, khái niệm

kiểm nghiệm tham số có thể hiểu là kiểm nghiệm lý thuyết hay, phổ biến hơn, kiểm nghinệm các tham số của phân bố lý thuyết. Ngược lại, kiểm nghiệm phi tham số hoàn toàn không bị lệ thuộc vào giả thiết về dạng phân bố lý thuyết. Người ta còn gọi kiểm nghiệm phi tham số là kiểm nghiệm phân bố tự do (distribution–free), nó không cần biết phân bố lý thuyết nào phù hợp với tập mẫu hiện có.

4.1.4 Các bước tiến hành một bài toán kiểm nghiệm giả thiết thống kê

Thông thường một bài toán kiểm nghiệm giả thiết được tiến hành theo các bước sau đây:

1) Căn cứ vào tập mẫu hiện có và yêu cầu của bài toán, xác định loại kiểm nghiệm nào sẽ được tiến hành: tham số hay phi tham số và quyết định các đặc trưng định lượng sẽ được tính toán từ tập mẫu.

2) Xác định giả thiết H_o. Thông thường giả thiết H_o được chọn sao cho đó chỉ là một "hình nộm" mà người ta hy vọng nó sẽ bị loại bỏ.

3) Xác định đối thiết H_1. Trong nhiều trường hợp H_1 là phủ định của H_o. Tuy nhiên ứng với một H_o có thể lựa chọn nhiều H_1 khác nhau.

4) Tương ứng với giả thiết H_o đúng ta sẽ nhận được phân bố "không" là một phân bố mẫu. Chú ý rằng đây là phân bố mẫu, tức phân bố của các tham số thống kê, nó có thể khác với những phân bố được dùng để biểu diễn gần đúng luật phân bố của một tập số liệu.

5) So sánh các đặc trưng xác suất nhận được từ tính toán trên tập mẫu và từ phân bố "không" để rút ra kết luận thống kê.

4.1.5 Miền thừa nhận và miền loại bỏ

Xét biến ngẫu nhiên X. Để tiến hành bài toán kiểm nghiệm ta lập không gian mẫu $(X_1, X_2,..., X_n)$ của X và trên không gian đó xác định một miền D_1 gọi là miền loại bỏ H_o. Phần bù của miền D_1 là miền D_o, miền thừa nhận H_o. Tập mẫu đã có $(x_1, x_2,..., x_n)$ tương ứng với một điểm X^* trong không gian mẫu.

- Nếu điểm $X^* \in D_o$ thì giả thiết H_o được coi là đúng và ta chấp nhận H_o.
- Nếu điểm $X^* \in D_1$ thì giả thiết H_o được coi là sai và ta bác bỏ H_o.

Khi đó:

$$P(D_1/H_o) = P(X \in D_1/H_o) = \int_{D_1} f(s)ds = \alpha \qquad (4.1.1)$$

Hay: $$P(D_o/H_o) = P(X \in D_o/H_o) = 1 - \int_{D_1} f(s)ds = 1 - \alpha \quad (4.1.2)$$

trong đó f(s) là mật độ xác suất của X.

Người ta gọi ranh giới giữa D_o và D_1 là điểm tới hạn d. Trong trường hợp một chiều, nếu $f(x/H_o)$ là mật độ xác suất có điều kiện của X thì có thể biểu diễn (4.1.1) dưới dạng:

$$P(X \in D_1/H_o) = \int_{-\infty}^{-d} f(x/H_o)dx + \int_{d}^{+\infty} f(x/H_o)dx = \alpha \quad (4.1.3)$$

Hay: $$P(X \in D_o/H_o) = \int_{-d}^{d} f(x/H_o)dx = 1 - \alpha \quad (4.1.4)$$

Thông thường trong các bài toán kiểm nghiệm ta cố định xác suất phạm sai lầm loại I để xác định các miền D_o và D_1. Từ các công thức (4.1.3) và (4.1.4), khi cho trước α, giải phương trình tích phân ta tìm được nghiệm là cận tích phân d. Trong đa số trường hợp ta có:

$$D_1 = \{-\infty; -d\} \cup \{d; +\infty\}$$

Nói chung các giá trị của X được xác định từ thực nghiệm, nghĩa là từ tập mẫu $(x_1, x_2, ..., x_n)$ ta có thể tính được X^* gọi là giá trị quan sát của X. Mặt khác, ứng với mức xác suất phạm sai lầm loại I bằng α ta sẽ xác định được các miền D_o và D_1.

Trong thực tế, do cách chọn giả thiết H_o của chúng ta thường với mục đích muốn loại bỏ nó, nên nếu $X^* \in D_1$ ta sẽ đưa ra kết luận ngay là H_o sai và ta bác bỏ nó. Trường hợp ngược lại, nếu $X^* \in D_o$ thì nói chung chỉ nên đưa ra kết luận một cách thận trọng "thực nghiệm chưa cho ta cơ sở để bác bỏ H_o" chứ không khẳng định một cách chắc chắn rằng H_o đúng.

4.2. Những vấn đề thực tế và việc hình thành giả thiết thống kê

4.2.1.Tính đồng nhất của các chuỗi

Khảo sát về tính đồng nhất chuỗi là một trong những vấn đề quan trọng của bài toán kiểm nghiệm giả thiết thống kê trong khí tượng, khí hậu. Có hai khái niệm đồng nhất được xét đến ở đây là sự đồng nhất giữa các chuỗi khác nhau trên cùng một khu vực (các chuỗi số liệu của các trạm khác nhau) và sự đồng nhất giữa các thời đoạn khác nhau của cùng một chuỗi. Tuỳ theo nội dung cụ thể của từng bài toán mà vấn đề nào sẽ được nêu ra để giải quyết.

Việc xác định về sự đồng nhất của các chuỗi số liệu được gọi là kiểm nghiệm tính đồng nhất. Tính đồng nhất ở đây được hiểu là sự đồng nhất tập thể: giữa tập thể các thành phần của chuỗi này (hoặc thời đoạn này) với tập thể các thành phần của chuỗi kia (hoặc thời đoạn kia). Ngoài ra, tính đồng nhất của các chuỗi cũng có thể được xét trên nhiều phương diện khác nhau, như đồng nhất về phân bố, đồng nhất về tham số, đồng nhất về độ lớn,...

Tính bất đồng nhất giữa các thời đoạn khác nhau của cùng một chuỗi thông thường xuất hiện do tác động của những nhân tố khách quan, như việc dời trạm, sự xuất hiện những công trình xây dựng mới gần trạm quan trắc,...

Chú ý rằng có sự phân biệt giữa khái niệm đồng nhất về mặt thống kê và đồng nhất về khía cạnh khí hậu.

Trong khí hậu, một chuỗi có thể được xem là đồng nhất nếu sự biến đổi hàng năm (từ năm nay qua năm khác) của các thành phần trong chuỗi được qui định bởi sự biến đổi tự nhiên của các quá trình qui mô lớn cấu thành điều kiện thời tiết và khí hậu của khu vực nghiên cứu. Sự phá huỷ tính đồng nhất khí hậu được xác định bởi rất nhiều nguyên nhân, như do ảnh hưởng của các công trình xây dựng, sự di chuyển địa điểm đặt trạm, sự thay đổi của lớp phủ thực vật và cảnh quan, sự thay đổi qui trình qui phạm quan trắc hoặc thay đổi dụng cụ, phương pháp quan trắc,...Có những nguyên nhân có thể gây nên sự bất đồng nhất trên toàn mạng lưới trạm, như thay đổi qui trình qui phạm hoặc phương pháp quan trắc, nhưng cũng có những nguyên nhân chỉ gây nên sự bất đồng nhất cục bộ (trong một số chuỗi nào đó).

Trong thống kê, chuỗi được xem là đồng nhất nếu, với một mức ý nghĩa cho trước nào đó, tất cả các thành phần của nó thuộc cùng một tập hợp. Sự bất đồng nhất thống kê xuất hiện do biến đổi khí hậu qui mô lớn gây nên bởi nhân tố thiên nhiên và con người. Nó xảy ra trên một mạng lưới trạm rộng lớn. Phát hiện được sự bất đồng nhất thống kê của chuỗi cho phép ta phán đoán về xu thế biến đổi khí hậu. Điều này có ý nghĩa rất quan trọng trong nghiên cứu sự dao động và biến đổi khí hậu.

Đồng nhất (bất đồng nhất) về mặt khí hậu không có ý nghĩa là đồng nhất (bất đồng nhất) về mặt thống kê. Nhưng nếu chuỗi đồng nhất thống kê thì luôn kéo theo sự đồng nhất khí hậu.

4.2.2 Một số bài toán điển hình

Nội dung kiểm nghiệm giả thiết thống kê về tính đồng nhất của các chuỗi số liệu khí hậu có thể đưa về một số bài toán cơ bản sau đây:

1) Giả sử, vì một lý do nào đó, trạm A phải di chuyển địa điểm vào năm YYYY. Khi xem xét chuỗi số liệu lượng mưa người ta thấy từ năm đó trở đi lượng

mưa có dấu hiệu tăng lên. Vậy, dấu hiệu "lượng mưa tăng lên kể từ khi dời trạm" có đúng không ?

Việc di chuyển địa điểm trạm có thể là nguyên nhân gây nên sự bất đồng nhất của chuỗi số liệu. Tính bất đồng nhất đó có thể biểu hiện qua dấu hiệu lượng mưa tăng lên hay giảm đi và có thể được đánh giá bằng việc so sánh trị số trung bình của hai giai đoạn. Bài toán đặt ra là kiểm nghiệm giả thiết về sự bằng nhau của trị số trung bình lượng mưa trước và sau khi dời trạm.

2) Xem xét chuỗi số liệu nhiệt độ trung bình tháng 7 của trạm B người ta nhận thấy rằng, kể từ khi thay đổi thiết bị đo vào năm YYYY hình như mức độ dao động thăng giáng của nhiệt độ có tăng lên so với trước. Hãy xác minh nhận định đó.

Số liệu quan trắc của nhiệt độ nói chung liên quan đến sai số đo, độ nhạy của thiết bị đo,... Việc thay đổi thiết bị đo có thể là nguyên nhân dẫn đến sự bất đồng nhất trong toàn chuỗi. Xác minh nhận định nêu trên có nghĩa là cần xem xét độ lệch chuẩn của chuỗi số liệu nhiệt độ trước và sau khi thay đổi dụng cụ đo sai khác nhau có đáng kể không. Điều đó đưa đến bài toán kiểm nghiệm sự bằng nhau của hai phương sai mẫu tính được từ số liệu của hai giai đoạn.

3) Khảo sát sơ bộ số liệu nhiệt độ tháng 1 của trạm C người ta nhận thấy hình như nó không tuân theo luật phân bố chuẩn như một số trạm lân cân. Điều nhận định đó đúng hay sai?

Trả lời câu hỏi này có nghĩa là cần tiến hành kiểm nghiệm giả thiết về sự phù hợp giữa phân bố thực nghiệm được xây dựng trên cơ sở tập số liệu trạm C và phân bố lý thuyết là phân bố chuẩn. Khái niệm đồng nhất được xét ở đây là tính đồng nhất về phân bố giữa các chuỗi khác nhau trên phạm vi một vùng không gian nhất định. Hiển nhiên vẫn có thể áp dụng bài toán này cho các thời đoạn khác nhau của cùng một chuỗi.

Ngoài ra, trong nghiên cứu khí tượng, khí hậu còn có nhiều vấn đề gắn liền với bài toán kiểm nghiệm giả thiết thống kê. Sau đây là một số dạng bài toán khác.

1) Như đã biết, ngoài hệ thống các trạm quan trắc khí tượng mà nhiệm vụ của nó là cung cấp số liệu phục vụ công tác dự báo thời tiết và tạo lập các chuỗi số liệu khí hậu, còn có những trạm quan trắc chuyên dụng. Các trạm quan trắc chuyện dụng thông thường được thành lập và duy trì hoạt động nhằm phục vụ cho các mục đích khác nhau. Vấn đề nảy sinh khi thành lập trạm loại này là phải trả lời được câu hỏi "Cần duy trì hoạt dộng của trạm trong thời gian bao lâu?", hay nói cách khác, "độ dài chuỗi số liệu quan trắc mà trạm cung cấp ít nhất là bao nhiêu năm".

Ví dụ: Cho biết phương sai của nhiệt độ tháng 1 của trạm X. Hãy xác định xem trạm X cần duy trì thời gian quan trắc ít nhất bao nhiêu năm để, với một giới

hạn tin cậy cho trước, trung bình số học của nhiệt độ tháng 1 trạm X sai khác không quá 0.1^0C so với chuẩn khí hậu.

2) Khi khảo sát mối quan hệ giữa hai đại lượng khí hậu người ta thấy rằng, hệ số tương quan thực nghiệm của chúng khá bé. Vậy, trên thực tế giữa hai đại lượng này có tồn tại mối quan hệ tuyến tính hay không?

Đây là bài toán kiểm nghiệm độ tin cậy của hệ số tương quan mẫu.

3) Sau khi xây dựng phương trình hồi qui tuyến tính giữa biến khí quyển Y và các biến $X_1,X_2,...,X_m$, người ta thấy sai số ước lượng khá lớn. Hỏi phương trình trình hồi qui tìm được có ý nghĩa sử dụng không?

Giải quyết vấn đề này có nghĩa là thực hiện bài toán đánh giá chất lượng phương trình hồi qui.

Cũng cần lưu ý rằng, các bài toán được nêu ra trên đây có thể xem như là những ví dụ cụ thể. Trong thực tế những vấn đề cần giải quyết chắc chắn còn chứa đựng nhiều sắc thái khác nhau, muôn hình muôn vẻ và là tổ hợp của nhiều bài toán. Do đó, để vận dụng nội dung của các bài toán này đòi hỏi ta phải phân tích vấn đề một cách kỹ lưỡng.

4.3 Kiểm nghiệm U

Kiểm nghiệm U được dùng để kiểm nghiệm các tham số khí hậu. Luật phân bố được sử dụng là phân bố chuẩn chuẩn hoá. Yêu cầu của bài toán kiểm nghiệm là dung lượng mẫu phải đủ lớn, trừ trường hợp biến khí hậu đang xét có phân bố chuẩn.

4.3.1 So sánh kỳ vọng với một số cho trước

Bài toán: Cho biến ngẫu nhiên phân bố chuẩn X có phương sai σ^2 (σ có thể đã được biết hoặc đã được chấp nhận) với n trị số quan sát $\{x_1, x_2,..., x_n\}$. Hãy kiểm nghiệm sự bằng nhau của kỳ vọng μ của X với một số cho trước μ_o.

Giải:

Trên thực tế số cho trước μ_o có thể là chuẩn khí hậu hoặc ở mức độ nào đó nó được chấp nhận là kỳ vọng của phân bố lý thuyết. Mục đích ứng dụng của kiểm nghiệm này là xác minh về sự bằng nhau của trung bình số học tính được từ tập mẫu với số cho trước μ_o.

Ta đặt giả thiết kiểm nghiệm là:

$$H_o: \mu = \mu_o \qquad (4.3.1)$$

Vì chưa có giá trị của μ nên thay cho μ ta sử dụng ước lượng của nó:

$$\mu \approx \bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t \qquad (4.3.1')$$

và đưa (4.3.1) về giả thiết tương đương:

$$H_0: \bar{x} = \mu_0 \quad \text{hay} \quad H_0: \bar{x} - \mu_0 = 0 \qquad (4.3.1'')$$

Thực chất của việc kiểm nghiệm giả thiết này là xét xem trị số $|\bar{x} - \mu_o|$ có lớn đến mức đáng kể không. Nếu $|\bar{x} - \mu_0|$ lớn đáng kể, tức là $\bar{x} \neq \mu_o$ quá nhiều, thì ta bác bỏ giả thiết H_o. Ngược lại ta sẽ chấp nhận H_o. Muốn vậy ta cần chọn giới hạn ban đầu d và đưa ra chỉ tiêu kiểm nghiệm:

Nếu $|\bar{x} - \mu_0| < d$ thì chấp nhận H_o

Ngược lại, nếu $|\bar{x} - \mu_0| \geq d$ thì bác bỏ H_o.

Với xác suất phạm sai lầm $\alpha = P(\text{Bỏ } H_o/H_o)$ cho trước thì giới hạn ban đầu d sẽ được xác định bởi:

$$P(|\bar{x} - \mu_0| \geq d) = \alpha, \text{ hay } P\left(\frac{|\bar{x} - \mu_o|}{\frac{\sigma}{\sqrt{n}}} \geq \frac{d}{\frac{\sigma}{\sqrt{n}}}\right) = \alpha. \qquad (4.3.2)$$

Đặt:
$$u = \frac{\bar{x} - \mu_o}{\frac{\sigma}{\sqrt{n}}}, \quad u_\alpha = \frac{d}{\frac{\sigma}{\sqrt{n}}} \qquad (4.3.3)$$

ta có $P(|u| \geq u_\alpha) = \alpha$. Từ đó chỉ tiêu kiểm nghiệm sẽ trở thành:

Nếu $|u| \geq u_\alpha$ thì bác bỏ H_o

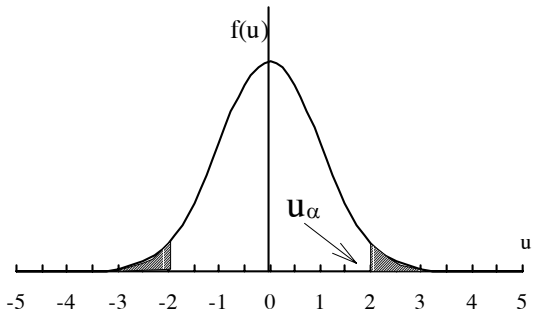
Ngược lại $|u| < u_\alpha$ thì chấp nhận H_o

Vấn đề còn lại là xác định u_α. Dễ dàng chứng minh được rằng biến *u* trong (4.3.3) có phân bố chuẩn chuẩn hoá với hai tham số 0 và 1: $u \in N(0,1)$. Từ đó ta nhận được:

$$P(|u| \geq u_\alpha) = 2\frac{1}{\sqrt{2\pi}}\int_{u_\alpha}^{+\infty} e^{-\frac{1}{2}t^2} dt = \alpha$$

Hay
$$\frac{1}{\sqrt{2\pi}}\int_{0}^{u_\alpha} e^{-\frac{1}{2}t^2} dt = 0.5 - \frac{\alpha}{2} \qquad (4.3.4)$$

Phương pháp xác định u_α được chỉ ra trên hình 4.1, trong đó toàn bộ diện tích miền giới hạn bởi đường cong phân bố và trục hoành bằng 1, còn tổng diện tích hai miền gạch chéo bằng α. Giá trị u_α cần tìm là cận tích phân trong công thức (4.3.4).



Hình 4.1 Xác định u_α

Trong các tài liệu về thống kê toán học người ta thường cung cấp bảng tính sẵn giá trị của u_α ứng với các α khác nhau (Bảng giá trị hàm Laplas $\Phi(u)$). Ta có thể tra bảng để xác định nó. Tuy nhiên, việc tra bảng như vậy vừa mang tính thủ công, mất thời gian lại vừa không thuận tiện. Hiện nay nhờ có phương tiện tính toán bằng máy tính điện tử, trị số của u_α thường được xác định một cách trực tiếp nhờ những phần mềm thông dụng hoặc bằng chương trình giải phương trình (4.3.4).

Tóm lại, ta có các bước thực hiện bài toán như sau:

1) Từ tập số liệu ban đầu $\{x_1, x_2,..., x_n\}$, tính các đại lượng $\bar{x}$, u theo các công thức (4.3.1') và (4.3.3).
2) Chọn giá trị xác suất phạm sai lầm loại I (α) thích hợp và xác định u_α bằng cách tra bảng tính sẵn hoặc giải phương trình (4.3.4).
3) So sánh $|u|$ và u_α để rút ra kết luận:

   Nếu $|u| \geq u_\alpha$ thì bác bỏ H_o và đưa ra kết luận $\mu \neq \mu_o$.

   Nếu $|u| < u_\alpha$ thì chấp nhận H_o, tức là chấp nhận giả thiết $\mu = \mu_o$.

Ví dụ 4.3.1 Số liệu nhiệt độ trung bình 100 năm của trạm A là T_{tb100}=25°C và độ lệch chuẩn s_{100} = 1°C. Vì mục đích sử dụng người ta muốn lấy nhiệt độ trung bình trong thời kỳ 10 năm gần đây thay cho trung bình dài năm kể trên. Sau khi tính toán người ta nhận được trị số trung bình của chuỗi 10 năm là T_{tb10}=24°C, khác biệt đáng kể so với trung bình dài năm. Hỏi nếu lấy T_{tb10} làm giá trị trung bình của nhiệt độ đại diện cho trạm A thì có đủ tiêu chuẩn không?

Giải: Nếu ta coi số liệu nhiệt độ trung bình 100 năm tương đương với chuẩn khí hậu, tức là μ_o=25°C và σ=1°C, thì bài toán dẫn đến việc kiểm nghiệm giả thiết:

$$H_o: T_{tb10}=T_{tb100}$$

Giả thiết rằng nhiệt độ trung bình năm có phân bố chuẩn ta có thể áp dụng kiểm nghiệm U trên đây để giải bài toán này. Ta có: n=10, đặt $u=(T_{tb10}-T_{tb100})/(1/\sqrt{10})$ và thay số vào rồi tính ra ta nhận được:

$$|u| = \frac{|24-25|}{1/\sqrt{10}} = 3.162$$

Nếu chọn $\alpha=0.05$ ta xác định được $u_\alpha=1.96$. Ta thấy $|u|>u_\alpha$, vậy H_o bị bác bỏ và ta kết luận rằng số liệu trung bình 10 năm không đủ tiêu chuẩn đại diện cho trung bình khí hậu của trạm A.

4.3.2 So sánh hai kỳ vọng

Bài toán: Cho hai biến ngẫu nhiên X, Y có phân bố chuẩn với n_1 và n_2 trị số quan sát tương ứng là $\{x_1, x_2, ..., x_{n_1}\}$ và $\{y_1, y_2, ..., y_{n_2}\}$, trong đó n_1, n_2 đủ lớn. Biết phương sai của X và Y tương ứng là σ_x^2, σ_y^2, hơn nữa $\sigma_x^2=\sigma_y^2=\sigma^2$. Hãy kiểm nghiệm sự bằng nhau của các kỳ vọng μ_x và μ_y của X và Y.

Giải:

Đặt giả thiết kiểm nghiệm là:

$$H_o: \mu_x = \mu_y$$

Trên thực tế ta không có các giá trị μ_x và μ_y, nên thay vào đó ta sử dụng các ước lượng thống kê của chúng là trung bình số học $\bar{x}$ và $\bar{y}$.

Ta có
$$\bar{x} = \frac{1}{n_1}\sum_{t=1}^{n_1} x_t, \bar{y} = \frac{1}{n}\sum_{t=1}^{n_2} y_t \qquad (4.3.5)$$

Khi đó giả thiết kiểm nghiệm được đưa về dạng:

$$H_o: \bar{x} = \bar{y}$$

Hay
$$H_o: \bar{x} - \bar{y} = 0$$

Với giới hạn tin cậy ban đầu *d* được chọn ta có chỉ tiêu kiểm nghiệm là:

Nếu $|\bar{x}-\bar{y}| \geq d$ thì bác bỏ H_o

Ngược lại, nếu $|\bar{x}-\bar{y}| < d$ thì chấp nhận H_o.

Tương tự như trước đây, *d* được chọn sao cho khi H_o đúng thì với xác suất phạm sai lầm loại I bằng α cho trước ta có:

$$P(|\bar{x}-\bar{y}| \geq d) = \alpha \qquad (4.3.6)$$

Đặt $$u = \frac{\bar{x} - \bar{y}}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad u_\alpha = \frac{d}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{4.3.7}$$

ta có thể đưa (4.3.6) về hệ thức tương đương: $P(|u| \geq u_\alpha) = \alpha$

Và chỉ tiêu kiểm nghiệm sẽ là:

Nếu $|u| \geq u_\alpha$ thì bác bỏ H_o

Nếu $|u| < u_\alpha$ thì chấp nhận H_o

Để xác định u_α cần phải biết luật phân bố của biến u. Người ta đã chứng minh được rằng biến u trong (4.3.7) có phân bố chuẩn chuẩn hóa $u \in N(0,1)$. Như vậy u_α hoàn toàn được xác định tương tự như đã xét trên đây (công thức 4.3.4).

Từ đó ta có các bước thực hiện bài toán như sau:

1) Từ các tập mẫu $\{x_1, x_2, ..., x_{n_1}\}$ và $\{y_1, y_2, ..., y_{n_2}\}$ tính $\bar{x}$, $\bar{y}$ và u theo công thức (4.3.5) và (4.3.7)
2) Chọn xác suất phạm sai lầm loại I (α) thích hợp và xác định u_α bằng cách tra bảng hoặc giải phương trình (4.3.4)
3) So sánh $|u|$ và u_α để rút ra kết luận theo chỉ tiêu kiểm nghiệm đã nêu.

Ghi chú: Hai chuỗi quan trắc $\{x_1, x_2, ... x_{n_1}\}$ và $\{y_1, y_2, ..., y_{n_2}\}$ tương ứng của các biến ngẫu nhiên X và Y có thể hiểu là hai thời đoạn của cùng một chuỗi hoặc hai chuỗi khác nhau.

Ví dụ 4.3.2 Từ chuỗi quan trắc 50 năm trước khi dời trạm đến địa điểm mới người ta tính được trung bình lượng mưa năm trạm A là X_{tb50}=1859.0 mm. Sau khi di chuyển được 42 năm thì trung bình lượng mưa năm ở đây là X_{tb42}=2031.3mm. Sự chênh lệch này có vẻ khá lớn. Phải chăng do di chuyển địa điểm mà lượng mưa tăng lên? Sự tăng lên này có đến mức đáng kể không? Biết rằng, kết quả kiểm nghiệm đã khẳng định phương sai của hai giai đoạn bằng nhau và bằng 179776mm^2, hay σ = 424,0mm.

Giải: Có thể nêu giả thiết: "lượng mưa tăng lên không đáng kể" và đặt giả thiết kiểm nghệm là H_0: $X_{tb50} = X_{tb42}$. Từ (4.3.7) ta có:

$$u = \frac{X_{tb50} - X_{tb42}}{\sigma\sqrt{\frac{1}{50} + \frac{1}{42}}} = \frac{1859.0 - 2031.3}{424\sqrt{\frac{1}{50} + \frac{1}{42}}} \approx -1.9416$$

Hay $|u| = 1.9416$

Chọn xác suất phạm sai lầm loại I là α = 0.05 ta được u_α=1.96. Vậy $|u| < u_\alpha$. Do đó giả thiết được chấp nhận, tức "lượng mưa tăng lên không đáng kể".

4.4 Kiểm nghiệm t

4.4.1 So sánh kỳ vọng với một số cho trước

Bài toán: Cho biến khí hậu X có phân bố chuẩn, $X \in N(\mu,\sigma)$ với n trị số quan sát $\{x_1, x_2,..., x_n\}$, nhưng chưa cho biết σ. Yêu cầu hãy kiểm nghiệm sự bằng nhau của kỳ vọng μ và số μ_0 cho trước.

Giải:

Có thể nhận thấy nội dung bài toán này gần với bái toán 4.3.1 nhưng ở đây chưa cho biết σ.

Đặt giả thiết kiểm nghiệm là: $H_o: \mu = \mu_o$

Vì chưa biết giá trị của μ nên ta thay μ bằng ước lượng của nó:

$$\mu \approx \bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t \qquad (4.4.1)$$

và đưa giả thiết về dạng tương:

$$H_o: \bar{x} = \mu_0 \quad \text{hay} \quad H_o: \bar{x} - \mu_o = 0$$

Chọn giới hạn tin cậy ban đầu *d* sao cho khi H_o đúng thì xác suất phạm sai lầm loại I là:

$$P(|\bar{x} - \mu_o| \geq d) = \alpha \qquad (4.4.2)$$

ta có thể lập được chỉ tiêu kiểm nghiệm là:

Nếu $|\bar{x} - \mu_0| \geq d$ thì bác bỏ H_o

Nếu $|\bar{x} - \mu_0| < d$ thì chấp nhận H_o

Đặt $$t = \frac{\bar{x} - \mu_0}{\frac{s^*}{\sqrt{n}}}, \qquad t_\alpha = \frac{d}{\frac{s^*}{\sqrt{n}}} \qquad (4.4.3)$$

trong đó $s^* = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(x_t - \bar{x})^2}$ là độ lệch chuẩn của X.

Ta có thể chuyển (4.4.2) về dạng tương đương: $P(|t| \geq t_\alpha) = \alpha$, và chỉ tiêu kiểm nghiệm sẽ là:

Nếu $|t| \geq t_\alpha$ thì bác bỏ H_o

Nếu $|t| < t_\alpha$ thì chấp nhận H_o

Vấn đề còn lại là xác định t_α. Muốn vậy cần phải biết luật phân bố của t. Người ta đã chứng minh được rằng biến t trong (4.4.3) có phân bố Student với (n–1) bậc tự do $t \in St(n-1)$. Từ đó ta có thể xác định được t_α ứng với xác suất α cho trước. Thông thường trong các tài liệu thống kê người ta cũng dẫn ra bảng tính sẵn các giá trị $t_\alpha(n)$ ứng với từng mức α và số bậc tự do n. Ta có thể tra bảng để nhận được t_α cho bài toán của mình. Tuy nhiên, t_α cũng có thể được xác định bằng việc giải phương trình:

$$\int_{-t_\alpha}^{t_\alpha} f(x,n-1)dx = 1-\alpha \qquad (4.4.4)$$

trong đó f(x,n–1) là hàm mật độ phân bố Student với n–1 bậc tự do. Do tính đối xứng của phân bố Student nên có thể viết (4.4.4) dưới dạng khác:

$$\int_{0}^{t_\alpha} f(x,n-1)dx = 0.5-\frac{\alpha}{2} \qquad (4.4.5)$$

Tóm lại ta có các bước giải bài toán như sau:

1) Từ tập mẫu $\{x_1, x_2,..., x_n\}$ ta tính $\bar{x}$, s^*, rồi tính *t* theo công thức (4.4.3)
2) Chọn α thích hợp và xác định t_α bằng cách tra bảng hoặc giải phương trình (4.4.5)
3) So sánh $|t|$ và t_α để rút ra kết luận.

Ví dụ 4.4.1 Cũng với nội dung như *ví dụ 4.3.1*, ta có $T_{tb100} = 25^0C$, $T_{tb10}=24^0C$, nhưng chưa cho biết độ lệch tiêu chuẩn s_{100}, thay vào đó từ tập số liệu 10 năm ta tính được $s_{10}^* = 1.2^0C$. Yêu cầu kiểm nghiệm giả thiết H_0: $T_{tb10}=T_{tb100}$.

Theo (4.4.3) ta tính được: $|t| = \frac{24-25}{1.2/\sqrt{10}} = 2.635$. Nếu chọn xác suất α=0.05 ta có t_α=2.262. Vậy $|t| > t_\alpha$, tức là giả thiết bị bác bỏ.

4.4.2 So sánh hai kỳ vọng

Bài toán: Cho hai biến ngẫu nhiên X, Y có phân bố chuẩn với n_1 và n_2 trị số quan sát tương ứng là $\{x_1,x_2,... x_{n_1}\}$ và $\{y_1,y_2,..., y_{n_2}\}$, (nếu chưa biết phân bố của X và Y thì n_1, n_2 phải đủ lớn). Các phương sai tương ứng σ_x^2, σ_y^2 chưa được biết, nhưng bằng kiểm nghiệm F người ta đã xác minh được $\sigma_x^2 = \sigma_y^2 = \sigma^2$. Yêu cầu hãy kiểm nghiệm sự bằng nhau của hai kỳ vọng μ_x và μ_y của X và Y.

Giải:

Giả thiết cần kiểm nghiệm là: H_o: $\mu_x = \mu_y$. Vì không có μ_x và μ_y nên ta thay chúng bằng các ước lượng thống kê:

$$\mu_x = \bar{x} = \frac{1}{n_1}\sum_{t=1}^{n_1} x_t \quad \text{và} \quad \mu_y = \bar{y} = \frac{1}{n_2}\sum_{t=1}^{n_2} y_t \tag{4.4.6}$$

Từ đó ta có: $H_o: \bar{x} = \bar{y}$

Hay $H_o: \bar{x} - \bar{y} = 0$

Chọn giới hạn tin cậy ban đầu d sao cho với xác suất phạm sai lầm loại I (α) cho trước ta có $P(|\bar{x} - \bar{y}| \geq d) = \alpha$. Khi đó chỉ tiêu kiểm nghiệm sẽ là:

Nếu $|\bar{x} - \bar{y}| \geq d$ thì bác bỏ H_o

Ngược lại, nếu $|\bar{x} - \bar{y}| < d$ thì chấp nhận H_o.

Đặt $$t = A(|\bar{x} - \bar{y}|), \quad t_\alpha = d.A \tag{4.4.7}$$

trong đó:
$$A = \frac{\dfrac{1}{\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}}{\dfrac{\sqrt{(n_1-1)s_x^{*2} + (n_2-1)s_y^{*2}}}{\sqrt{n_1+n_2-2}}}$$

$$s_x^* = \sqrt{\frac{1}{n_1-1}\sum_{t=1}^{n_1}(x_t - \bar{x})^2}, \quad s_y^* = \sqrt{\frac{1}{n_2-1}\sum_{t=1}^{n_1}(y_t - \bar{y})^2}$$

Khi đó nếu H_o đúng thì $P(|t| \geq t_\alpha) = \alpha$ và chỉ tiêu kiểm nghiệm sẽ là:

Nếu $|t| \geq t_\alpha$ thì bác bỏ H_o

Nếu $|t| < t_\alpha$ thì chấp nhận H_o

Để xác định giá trị chưa biết t_α cần phải biết phân bố xác suất của t. Có thể chứng minh được rằng $t \in St(n_1+n_2-2)$. Từ đó ta dễ dàng xác định được t_α bằng cách tra bảng tính sẵn hoặc giải phương trình:

$$\int_0^{t_\alpha} f(x, n_1+n_2-2)dx = 0.5 - \frac{\alpha}{2}$$

Như vậy, các bước để giải bài toán sẽ là:

1) Từ các tập số liệu $\{x_1, x_2, ..., x_{n_1}\}$ và $\{y_1, y_2, ..., y_{n_2}\}$, tính $\bar{x}$, $\bar{y}$, s_x^*, s_y^*, rồi tính t theo (4.4.7).
2) Chọn α thích hợp rồi xác định t_α với $t \in St(n_1+n_2-2)$.
3) So sánh $|t|$ và t_α để rút ra kết luận.

Ví dụ 4.4.2 Hãy kiểm nghiệm sự bằng nhau của tổng lượng mưa trung bình trạm A thời kỳ 30 năm trước và 20 năm sau, biết rằng từ số liệu thực tế người ta đã tính được R_{tb30}=1602.9, R_{tb20}=1770.7, s_{30}=367.0, s_{20}=293.1. Cho xác suất phạm sai lầm loại I là α=0.05.

Giả thiết cần kiểm nghiệm là H_o: $R_{tb30}=R_{tb20}$. Ta có $n_1=30$, $n_2=20$. Vậy:

$$t = \frac{\frac{1602.9-1770.7}{\sqrt{\frac{1}{30}+\frac{1}{20}}}}{\frac{\sqrt{(30-1)367.0^2+(20-1)293.1^2}}{\sqrt{30+20-2}}} = -1.7113,$$

$$t_{0.05}(30+20-2) = 1.6772$$

Vì $|t|=1.7113 > t_\alpha=1.6772$ do đó ta bác bỏ giả thiết H_o, tức là tổng lượng mưa trung bình trạm A của hai thời kỳ không bằng nhau.

4.5 Kiểm nghiệm F

Bài toán: Cho hai biến ngẫu nhiên có phân bố chuẩn $X\in N(\mu_1,\sigma_1)$, $Y\in N(\mu_y,\sigma_y)$ với n_1 và n_2 trị số quan sát tương ứng là $\{x_1,x_2,...,x_{n_1}\}$ và $\{y_1,y_2,...,y_{n_1}\}$. Yêu cầu hãy kiểm nghiệm sự bằng nhau của σ_x^2 và σ_y^2.

Giải:

Đặt giả thiết kiểm nghiệm là H_o: $\sigma_x^2=\sigma_y^2$

Vì chưa biết σ_x^2 và σ_y^2 nên ta thay chúng bằng các ước lượng tương ứng:

$$\sigma_x^2 \approx s_x^{*2}=\frac{1}{n_1-1}\sum_{t=1}^{n_1}(x_t-\bar{x})^2, \quad \sigma_y^2 \approx s_y^{*2}=\frac{1}{n_2-1}\sum_{t=1}^{n_2}(y_t-\bar{y})^2 \qquad (4.5.1)$$

trong đó $$\bar{x}=\frac{1}{n}\sum_{t=1}^{n_1}x_t, \bar{y}=\frac{1}{n_2}\sum_{t=1}^{n_2}y_t,$$

và đưa giả thiết kiểm nghiệm về dạng tương đương: H_o: $S_x^{*2}=S_y^{*2}$.

Giả sử $S_x^{*2} > s_y^{*2}$, ta lập biến mới

$$f = S_x^{*2}/S_y^{*2} \qquad (4.5.2)$$

và xây dựng chỉ tiêu kiểm nghiệm là:

Nếu $f \ge f_\alpha$ thì bác bỏ H_o (Hai phương sai không bằng nhau)

Nếu $f < f_\alpha$ thì chấp nhận H_o

Trong đó f_α là giới hạn tin cậy của f ứng với xác suất phạm sai lầm loại I bằng α: $P(f \ge f_\alpha) = \alpha$.

Để xác định f_α ta cần thiết phân bố của f. Bằng một số phép biến đổi ta có thể chứng minh được khi H_o đúng thì biến f có phân bố Fisher với n_1-1 bà n_2-1 bậc tự do: $f \in F(n_1-1,n_2-1)$.

Từ đó, f_α sẽ được xác định bởi:

$$\int_0^{f_\alpha} f(t,n_1-1,n_2-1)dt = 1-\alpha, \qquad (4.5.3)$$

trong đó $f(t,n_1-1,n_2-1)$ là mật độ xác suất của phân bố Fisher với (n_1-1) và (n_2-1) bậc tự do.

Như vậy, ta có các bước giải bài toán sau đây:

1) Từ các tập số liệu $\{x_1,x_2,...,x_{n_1}\}$ và $\{y_1,y_2,...,y_{n_2}\}$, tính s_x^{*2} và s_y^{*2} theo (4.5.1). Sau đó lập tỉ số $f = s_x^{*2}/s_y^{*2}$ nếu $s_x^{*2} > s_y^{*2}$. Trong trường hợp ngược lại ta đổi vai trò của s_x^{*2} và s_y^{*2} cho nhau.
2) Chọn α thích hợp rồi xác định f_α bằng cách tra bảng tính sẵn hoặc giải phương trình (4.5.3).
3) So sánh f và f_α để rút ra kết luận.

Ví dụ 4.5 Giả sử nhiệt độ tháng 1 của trạm A và B đều tuân theo luật phân bố chuẩn. Từ số liệu lịch sử 34 năm của trạm A và 30 năm của trạm B người ta tính được độ lệch chuẩn của chúng tương ứng là $s_A^*=1.95$, $s_B^*=1.50$. Hỏi sự khác biệt của độ lệch chuẩn nhiệt độ tháng 1 giữa hai trạm có đáng kể không?

Giải: Bài toán đặt ra là kiểm nghiệm giả thiết H_0: $s_A^{*2} = s_B^{*2}$ – không có sự khác biệt đáng kể giữa độ lệch chuẩn của hai trạm.

Ta có $f = s_A^{*2}/s_B^{*2} = 1.68$, $n_1=34$, $n_2= 30$, nên biến $f \in F(33,29)$. Chọn xác suất phạm sai lầm loại I là $\alpha = 0.05$ ta tính được $f_\alpha=1.84$. Vậy $f<f_\alpha$, nên giả thiết H_0 được chấp nhận, tức độ lệch chuẩn của nhiệt độ tháng 1 ở hai trạm không có sự khác nhau đáng kể. Nói cách khác, với mức ý nghĩa 5% có thể xem rằng độ lệch chuẩn của nhiệt độ hai trạm bằng nhau.

4.6 Kiểm nghiệm χ^2

Kiểm nghiệm χ^2 được dùng để kiểm nghiệm sự phù hợp giữa phân bố thực nghiệm và phân bố lý thuyết.

Bài toán: Cho biến khí hậu X với n trị số quan sát $\{x_1, x_2,..., x_n\}$ (n đủ lớn). Từ tập mẫu này ta xây dựng được hàm phân bố thực nghiệm với K tham số $\theta_1, \theta_2,...,\theta_K$: $F(x; \theta_1, \theta_2,...,\theta_K)$. Yêu cầu xác minh:

$$F(x; \theta_1, \theta_2,...,\theta_K) = G(x; \theta_1, \theta_2,...,\theta_K),$$

trong đó $G(x; \theta_1, \theta_2,...,\theta_K)$ là một phân bố lý thuyết đã biết.

Giải:

Đặt giả thiết kiểm nghiệm H_o: $F(x; \theta_1, \theta_2,...,\theta_K) = G(x; \theta_1, \theta_2,...,\theta_K)$.

Với n đủ lớn, ta chia tập mẫu $\{x_1, x_2,..., x_n\}$ thành N nhóm (a_j, b_j), j=1..N, trong đó, $b_j = a_{j+1}$, $a_1 \le \min[x_t, t=1..n\}$, $b_N > \max\{x_t, t=1..n\}$.

Vì xác suất để X nhận giá trị trong khoảng (a_j,b_j) tính theo phân bố thực nghiệm bằng $P(a_j \le X < b_j) = F(b_j) - F(a_j)$ nên tần số thực nghiệm:

$$m_j = n[F(b_j) - F(a_j)] = n[F(a_{j+1}) - F(a_j)].$$

Mặt khác, xác suất này tính theo phân bố lý thuyết bằng:

$$p_j = P(a_j \le X < b_j) = G(a_{j+1}) - G(a_j)$$

nên tần số lý thuyết của nhóm (a_j, b_j) sẽ là np_j. Ta có bảng sau:

Nhóm	Giới hạn dưới	Giới hạn trên	Tần số thực nghiệm	Xác suất lý thuyết	Tần số lý thuyết
1	a_1	b_1	m_1	p_1	np_1
2	a_2	b_2	m_2	p_2	np_2
...	...	...	...	...	...
N	a_N	b_N	m_N	p_N	np_N

Hiệu $Q_j = np_j - m_j$ được dùng làm thước đo mức độ khác biệt giữa phân bố thực nghiệm $F(x; \theta_j)$ và phân bố lý thuyết $G(x;\theta_j)$.

Ta lập biến mới:
$$\eta = \sum_{j=1}^{N} \frac{Q_j^2}{np_j} = \sum_{j=1}^{N} \frac{(np_j - m_j)^2}{np_j} \qquad (4.6.1)$$

và đưa ra tiêu chuẩn kiểm nghiệm là:

Nếu $\eta \ge \eta_\alpha$ thì bác bỏ H_o (phân bố thực nghiệm không phù hợp với phân bố lý thuyết)

Nếu $\eta < \eta_\alpha$ thì chấp nhận H_o.

Trong đó η_α là giới hạn tin cậy, được xác định sao cho khi H_o đúng thì:

$$P(\eta \ge \eta_\alpha) = \alpha \qquad (4.6.2)$$

Vấn đề còn lại là phải xác định η_α, tức là phải xác định luật phân bố của biến η. Người ta đã chứng minh được rằng, khi n đủ lớn thì η có phân bố χ^2 với (N–K–1) bậc tự do: $\eta \in \chi^2(N - K - 1)$ (Bạn đọc có thể xem thêm quá trình chứng minh này trong [4,5]). Vậy giá trị của η_α có thể được xác định từ các bảng tính sẵn hoặc giải phương trình:

$$\int_{\eta_\alpha}^{\infty} f_{N-K-1}(x)dx = \alpha \qquad (4.6.3)$$

Hay:
$$\int_{0}^{\eta_\alpha} f_{N-K-1}(x)dx = 1-\alpha \qquad (4.6.3')$$

Trong đó $f_{N-K-1}(x)$ là mật độ xác suất $\chi^2(N-K-1)$ với N–K–1 bậc tự do. Từ đó ta có các bước tiến hành sau:

1) Phân chia tập số liệu thành N nhóm và xác định tần số các nhóm m_j.
2) Từ phân bố lý thuyết đã biết, xác định tần số lý thuyết các nhóm np_j.
3) Tính giá trị của η theo công thức (4.6.1)
4) Chọn giá trị α thích hợp, xác định η_α theo phân bố χ^2 với N–K–1 bậc tự do.
5) So sánh η và η_α để rút ra kết luận.

Ví dụ 4.6 Hãy kiểm tra tính phân bố chuẩn của chuỗi số liệu nhiệt độ trung bình tháng 1 trạm A cho trong bảng 4.1.

Bảng 4.1 Nhiệt độ trung bình tháng 1 của trạm A (^{0}C)

17.0	16.4	18.2	18.1	15.0	13.1	19.2
17.9	17.4	16.3	15.5	17.6	16.2	17.8
17.1	17.2	15.5	15.0	17.0	17.3	15.2
12.3	16.7	19.6	17.2	15.2	17.4	17.3
17.6	20.1	15.2	15.7	14.7	17.2	
17.3	17.5	17.4	14.3	16.8	18.1	
12.7	15.0	16.6	14.8	16.2	14.5	
13.0	18.8	19.8	16.8	15.9	13.7	
17.1	15.4	14.5	18.0	16.3	14.1	
13.6	18.9	15.8	18.2	16.1	16.7	

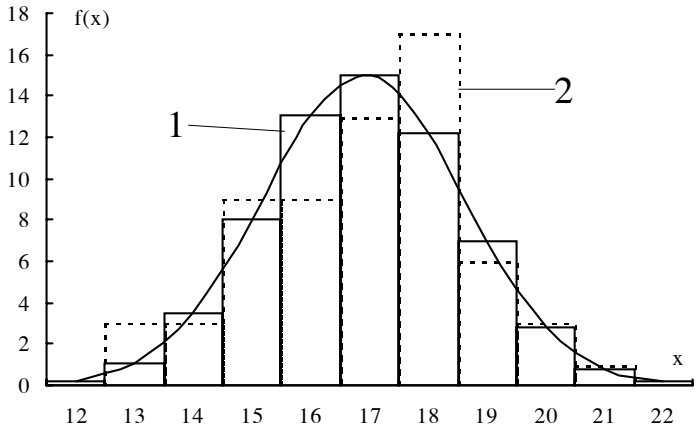
Giải: Với n=64, như vậy dung lượng mẫu đủ lớn để ta có thể tiến hành phân nhóm. Số nhóm được lấy bằng N=5lg64≈9 (nhóm). Cự lý các nhóm được chọn đều nhau và bằng 1(^{0}C). Kết quả tính toán trung gian được trình bày trong bảng 4.2. Từ đó ta nhận được η=4.337 ≈ 4.34.

Mặt khác, vì phân bố lý thuyết là phân bố chuẩn nên nó phụ thuộc vào hai tham số là kỳ vọng (μ) và độ lệch bình phương trung bình (σ). Từ bảng 4.1 ta nhận được ước lượng của các đại lượng này tương ứng là $\mu \approx \bar{x}$=16.4 và $\sigma \approx s^*$=1.7. Hơn nữa ta có K=2 và số bậc tự do bằng N–K–1=6. Nếu chọn α=0.05 ta xác định được η_α theo phân bố $\chi^2(6)$: η_α=12.59. Kết quả so sánh ta có $\eta < \eta_\alpha$ nên giả thiết H_0 được chấp nhận, nghĩa là nhiệt độ trung bình tháng 1 trạm A tuân theo luật phân bố chuẩn.

Trên hình 4.2 biểu diễn đồ thị hàm mật độ phân bố chuẩn lý thuyết (đường liền nét) và phân thực nghiệm (đường gạch nối) theo kết quả tính toán trong bảng 4.2

Bảng 4.2 Kết quả tính trung gian

Nhóm	a_j	b_j	m_j	p_j	np_j	$\frac{(np_j - m_j)^2}{np_j}$
1	12	13	3	0.0255	1.6328	1.1448
2	13	14	3	0.0584	3.7404	0.1466
3	14	15	8	0.1260	8.0631	0.0005
4	15	16	10	0.1974	12.636	0.5498
5	16	17	13	0.2250	14.397	0.1355
6	17	18	17	0.1864	11.9266	2.1582
7	18	19	6	0.1122	7.1832	0.1949
8	19	20	3	0.0491	3.1450	0.0067
9	20	21	1	0.0156	1.0007	0.0000
Tổng			64	0.9956		η=4.337



Hình 4.2 Kết quả xấp xỉ phân bố nhiệt độ tháng 1 trạm A bởi phân bố chuẩn

1) Phân bố lý thuyết; 2) Phân bố thực nghiệm

4.7. Kiểm nghiệm U phi tham số

Kiểm nghiệm U phi tham số còn được gọi là kiểm nghiệm Wilcoxon, hay kiểm nghiệm Mann–Whiteney, vì nó được Wilcoxon phát minh vào năm 1945, sau đó được Mann–Whitney triển khai ứng dụng. Đây là một trong những kiểm nghiệm phi tham số, được ứng dụng phổ biến trong trường hợp dung lượng mẫu *n* bé, hơn nữa không yêu cầu biết trước dạng phân bố của chuỗi. Thông thường trong khí tượng, khí hậu kiểm nghiệm U phi tham số dùng để xác minh tính đồng nhất tương

đối về độ lớn giữa các thành phần trong hai chuỗi số liệu khí hậu độc lập hoặc hai thời đoạn khác nhau của cùng một chuỗi.

Bài toán: Xét biến khí quyển X. Giả sử $\{x_1,x_2,..,x_m\}$ và $\{y_1,y_2,...y_n\}$ là hai chuỗi số liệu quan trắc của X (có thể là hai chuỗi của hai trạm khác nhau hoặc hai thời đoạn quan trắc của cùng một trạm). Hãy xác minh sự đồng nhất tương đối về độ lớn giữa m thành phần của chuỗi $\{x_t, t=1..m\}$ và n thành phần của chuỗi $\{y_t, t=1..n\}$.

Giải:

Trước hết ta đánh dấu số liệu của một trong hai chuỗi, chẳng hạn chuỗi $\{y_t\}$, rồi gộp hai chuỗi lại thành một và lập chuỗi trình tự $\{z_{(t)}, t=1..m + n\}$, với $z_{(1)} \le z_{(2)} \le ... \le z_{(m+n)}$. Từ chuỗi này ta lập hai chuỗi mới $\{u_i\}$ và $\{v_i\}$ theo nguyên tắc sau đây:

u_i= Số thành phần của chuỗi $\{y_t\}$ đứng trước x_i trong chuỗi $\{z_{(t)}\}$,i=1..m

v_i= Số thành phần của chuỗi $\{x_t\}$ đứng trước y_i trong chuỗi $\{z_{(t)}\}$,i=1..n

Sau đó lập các biến mới:

$$U = \sum_{i=1}^{m} u_i, V = \sum_{i=1}^{n} v_i \qquad (4.7.1)$$

Vì có m thành phần của chuỗi $\{x_t\}$, n thành phần của chuỗi $\{y_t\}$ nên:

$$U + V = mn = \text{Tổng số lần so sánh.}$$

Các biến U và V có thể nhận giá trị từ 0 (tất cả các x_t đều nhỏ hơn hoặc lớn hơn y_t) đến mn (tất cả các x_t đều lớn hơn hoặc nhỏ hơn y_t). Hai chuỗi được gọi là đồng nhất nếu giả thiết H_0:U=V được chấp nhận.

Để rõ hơn ta xét ví dụ sau đây. Giả sử ta có m=6, n=4 và sau khi sắp xếp theo thứ tự tăng dần ta được chuỗi sau:

$$\{z_{(t)}\} = \{y_1,x_1,x_2,y_2,y_3,y_4,x_3,y_5,y_6,x_4\}$$

Từ đó: $\{u_i\}$={1,1,4,6} và $\{v_i\}$={0,2,2,2,3,3}.

Vậy: U=1+1+4+6=12 và V=0+2+2+2+3+3=12.

Phương pháp trên đây thường chỉ áp dụng cho những trường hợp dung lượng mẫu khá bé. Việc kiểm nghiệm giả thiết nêu trên được thực hiện bằng cách so sánh giá trị nhỏ nhất trong hai giá trị U và V với bảng giá trị sẵn ứng với từng mức xác suất cho trước.

Khi dung lượng mẫu tương đối lớn người ta tiến hành tính toán theo phương thức sau đây. Từ chuỗi trình tự $\{z_{(t)}\}$ ta lập các biến mới:

$$U=mn+\frac{m(m+1)}{2}-T \tag{4.7.2}$$

trong đó $$T=\sum_{t=1}^{m+n} t\Big|_{z_{(t)}\in\{y_t\}}$$

và $$V=mn+\frac{n(n+1)}{2}-T' \tag{4.7.3}$$

với $$T'=\sum_{t=1}^{m+n} t\Big|_{z_{(t)}\in\{x_t\}}$$

Người ta đã chứng minh được rằng, khi m, n≥8 thì U, V∈N(μ,σ), trong đó:

$$\mu=M[U]=M[V]=\frac{mn}{2} \text{ và } \sigma=\sqrt{\frac{mn}{12}(m+n+1)} \tag{4.7.4}$$

Trong mọi trường hợp sau khi tính U và V ta sẽ chọn giá trị nhỏ nhất trong chúng. Giả sử U≤V, khi đó sự bất đồng nhất giữa hai chuỗi có thể được đánh giá bằng hiệu $|U-M[U]|$. Hiệu $|U-M[U]|=0$ ứng với trường hợp hai chuỗi đồng nhất thực sự. Hiệu này càng lớn thì sự bất đồng nhất giữa hai chuỗi càng lớn.

Do đó ta đặt giả thiết kiểm nghiệm là H_0: $|U-M[U]|=0$. Nếu H_0 đúng ta kết luận hai chuỗi đồng nhất và ngược lại. Thực chất điều kiện $|U-M[U]|=0$ tương đương với việc chọn giới hạn tin cậy d sao cho khi H_0 đúng, với xác suất $P(|U-M[U]|\geq d)=\alpha$, thì:

$|U-M[U]|\geq d$: giả thiết H_0 bị bác bỏ (hai dãy không đồng nhất).

$|U-M[U]|<d$: giả thiết H_0 được chấp nhận (hai dãy đồng nhất)

Ta có: $P(|U-M[U]|\geq d)=P(\frac{|U-M[U]|}{\sqrt{D[U]}}\geq\frac{d}{\sqrt{D[U]}})=\alpha$

Đặt $$u=\frac{U-M[U]}{\sqrt{D[U]}} \text{ và } u_\alpha=\frac{d}{\sqrt{D[U]}} \tag{4.7.5}$$

khi đó nếu H_0 đúng thì $P(|u|\geq u_\alpha)=\alpha$.

Vì U∈N(μ,σ) nên u∈N(0,1). Từ đây ta dễ dàng xác định được u_α.

Tóm lại, ta có các bước thực hiện sau:

1) Từ hai chuỗi số liệu ban đầu, gộp lại và lập chuỗi trình tự $\{z_{(t)},t=1..m+n\}$
2) Tính U, V theo (4.7.2) và (4.7.3). Giả sử U ≤ V, tính $M[U]=\mu=\frac{mn}{2}$ và $\sigma=\sqrt{D[U]}=\sqrt{\frac{mn}{12}(m+n+1)}$, sau đó tính u theo (4.7.5).

3) Chọn α thích hợp rồi xác định u_α từ phân bố chuẩn châu hóa.

4) So sánh $|u|$ và u_α để phán đoán về sự đồng nhất của hai chuỗi.

Ví dụ 4.7 Tổng lượng mưa năm trước và sau khi dời trạm của trạm A được cho trong bảng 4.3. Hãy xác minh tính đồng nhất của số liệu hai thời đoạn đó. Cho xác suất phạm sai lầm loại I là α=0.05.

Bảng 4.3 Số liệu lượng mưa năm trạm A trước và sau khi dời trạm (mm)

Trước khi dời trạm (x)					Sau khi dời trạm (y)				
1076.0	1373.9	1435.1	1583.1	1838.8	1256.8	1568.8	1736.8	1829.8	2040.3
1120.2	1375.4	1464.1	1605.9	1847.9	1297.3	1653.0	1738.8	1862.8	2141.2
1200.4	1376.6	1493.0	1622.0	1860.8	1544.4	1684.5	1758.9	1931.9	2153.9
1342.1	1390.9	1540.4	1637.5	1864.8	1554.3	1725.7	1800.0	1943.4	2528.2
1346.4	1394.6	1542.0	1690.8	1983.8					
2063.6	2071.0	2149.8	2200.5	2617.0					

Nếu gọi chuỗi số liệu trước khi dời trạm là $\{x_t, t=1..m\}$, và sau khi dời trạm là $\{y_t, t=1..n\}$ thì m=30 và n=20. Từ hai chuỗi này ta lập chuỗi trình tự $\{z_t, t=1..m+n\}$ trong đó ta đánh dấu các thành phần của chuỗi $\{y_t\}$. Kết quả của bước này được trình bày trong bảng 4.4.

Bảng 4.4 Chuỗi lượng mưa đã sắp xếp

t	z	t	z	t	z	t	z	t	z
1	1076.0	11	1390.9	21	1583.1	31	1758.9	41	1983.8
2	1120.2	12	1394.6	22	1605.9	32	1800.0	42	2040.3
3	1200.4	13	1435.1	23	1622.0	33	1829.8	43	2063.6
4	1256.8	14	1464.1	24	1637.5	34	1838.8	44	2071.0
5	1297.3	15	1493.0	25	1653.0	35	1847.9	45	2141.2
6	1342.1	16	1540.4	26	1684.5	36	1860.8	46	2149.8
7	1346.4	17	1542.0	27	1690.8	37	1862.8	47	2153.9
8	1373.9	18	1544.4	28	1725.7	38	1864.8	48	2200.5
9	1375.4	19	1554.3	29	1736.8	39	1931.9	49	2528.2
10	1376.6	20	1568.8	30	1738.8	40	1943.4	50	2617.0

Từ bảng 4.4 ta nhận được:

$t(z_t \in y)$					$t(z_t \in x)$				
4	20	29	33	42	1	9	15	24	41
5	25	30	37	45	2	10	16	27	43
18	26	31	39	47	3	11	17	34	44
19	28	32	40	49	6	12	21	35	46
					7	13	22	36	48
					8	14	23	38	50
$T = \sum t\|_{z_{(t)} \in \{y_t\}} = 599$					$T' = \sum t\|_{z_{(t)} \in \{x_t\}} = 676$				

Vậy, theo (4.7.2) và (4.7.3) ta có: U = 446, V = 134. Vì U>V nên để tiến hành kiểm nghiệm ta sẽ sử dụng V.

Theo (4.7.4), $\mu = M[V] = 30.20/2 = 300$; $\sigma = \sqrt{\frac{30.20}{12}(30+20+1)} = 50.5$. Đổi vai trò của U trong (4.7.5) thành V ta tính được:

$$u = (V-\mu)/\sigma = (134-300)/50.5 = -3.29$$

Với $\alpha = 0.05$ ta có $u_\alpha = 1.96$. Vậy, $|u| = 3.29 > u_\alpha = 1.96$. Do đó ta kết luận hai chuỗi không đồng nhất.

CHƯƠNG 5
PHÂN TÍCH TƯƠNG QUAN VÀ HỒI QUI

5.1 Những khái niệm mở đầu

Trong thực tế nghiên cứu khí tượng, khí hậu có không ít những vấn đề được đặt ra trong đó cần phải xác định được qui luật biến đổi của các hiện tượng khí quyển. Tuy nhiên, hiện tượng khí quyển lại được phản ánh thông qua các đặc trưng yếu tố khí quyển mà chúng, đến lượt mình, lại phụ thuộc vào sự biến đổi của các nhân tố bên ngoài. Muốn nắm được qui luật biến đổi của các hiện tượng khí quyển cần thiết phải xác định sự liên hệ giữa các đặc trưng yếu tố khí quyển (được xem là biến phụ thuộc) với tập hợp các nhân tố ảnh hưởng mà người ta gọi là các biến độc lập. Điều đó cũng có nghĩa là, về phương diện thống kê, thông thường ta cần phải giải quyết một số vấn đề sau đây:

1) Xác định sự phân bố không gian của các đặc trưng yếu tố khí tượng, khí hậu, tức là nghiên cứu qui luật phụ thuộc vào toạ độ không gian của các biến khí quyển.
2) Xác định qui luật, tính chất diễn biến theo thời gian của các đặc trưng yếu tố khí quyển.
3) Xác định mối quan hệ ràng buộc để từ đó tìm qui luật liên hệ giữa các đặc trưng yếu tố khí quyển với nhau theo không gian và thời gian.

Một trong những phương pháp giải quyết các vấn đề đó là phương pháp phân tích tương quan và hồi qui mà nội dung của nó có thể được chia thành:

1) Tương quan và hồi qui theo không gian: Là xét mối quan hệ giữa hai hay nhiều biến khí quyển với nhau của cùng một yếu tố, cùng thời gian (đồng thời) nhưng khác nhau về vị trí không gian.
2) Tương quan và hồi qui theo thời gian: Là xét mối quan hệ giữa hai hay nhiều biến khí quyển với nhau của cùng một yếu tố, cùng một địa điểm nhưng khác nhau về thời gian.
3) Tương quan và hồi qui phổ biến: Là xét mối quan hệ giữa hay nhiều biến khí quyển của một hoặc nhiều yếu tố, có thể khác nhau về không gian, thời gian hoặc cả không–thời gian.

Về phương diện toán học, căn cứ vào dạng thức của biểu thức biểu diễn, người ta chia sự quan hệ tương quan làm bốn dạng:

1) Tương quan và hồi qui tuyến tính một biến: Xét mối quan hệ tương quan và hồi qui tuyến tính giữa một bên là biến phụ thuộc với một bên là một biến độc lập.
2) Tương quan và hồi qui phi tuyến một biến: Xét mối quan hệ tương quan và hồi qui phi tuyến giữa một bên là biến phụ thuộc với một bên là một biến độc lập.
3) Tương quan và hồi qui tuyến tính nhiều biến: Xét mối quan hệ tương quan và hồi qui tuyến tính giữa một bên là biến phụ thuộc với một bên là tập hợp nhiều biến độc lập.
4) Tương quan và hồi qui phi tuyến nhiều biến: Xét mối quan hệ tương quan và hồi qui phi tuyến giữa một bên là biến phụ thuộc với một bên là tập hợp nhiều biến độc lập.

Thông thường để giải quyết các bài toán tương quan và hồi qui trong khí tượng, khí hậu cần phải tiến hành các bước sau:

1) Xác lập được dạng thức của mối liên hệ tương quan, tức là tìm ra dạng hồi qui thích hợp: Tuyến tính hay phi tuyến, nếu là phi tuyến thì cụ thể là dạng nào.
2) Đánh giá được mức độ chặt chẽ của các mối liên hệ theo nghĩa quan hệ tương quan.
3) Bằng phương pháp nào đó, xác lập biểu thức giải tích của phương trình hồi qui xấp xỉ mối liên hệ tương quan, tức là xây dựng hàm hồi qui. Trong khí tượng, khí hậu phương pháp phổ biến để xây dựng hàm hồi qui là phương pháp bình phương tối thiểu.
4) Đánh giá độ chính xác và khả năng sử dụng của phương trình hồi qui.

5.2 Tương quan tuyến tính

5.2.1 Hệ số tương quan tổng thể

Xét hai biến ngẫu nhiên X_1 và X_2. Khi đó phương sai của tổng (hiệu) hai biến được xác định bởi:

$$\begin{aligned} D[X_1 \pm X_2] &= M[(X_1 \pm X_2) - M(X_1 \pm X_2)]^2 = M[(X_1 - MX_1) \pm (X_2 - MX_2)]^2 = \\ &= M[(X_1 - MX_1)^2] + M[(X_2 - MX_2)^2] \pm 2M[(X_1 - MX_1)(X_2 - MX_2)] = \\ &= D[X_1] + D[X_2] \pm 2\, M[(X_1 - MX_1)(X_2 - MX_2)] = \\ &= \mu_{11} + \mu_{22} + \pm 2\mu_{12} \end{aligned}$$

trong đó μ_{12} là mômen tương quan giữa X_1 và X_2, μ_{11} và μ_{22} tương ứng là phương sai của X_1 và X_2. Nếu X_1 và X_2 không tương quan với nhau thì:

$$D[X_1 \pm X_2] = D[X_1] + D[X_2], \text{ suy ra } \mu_{12} = 0.$$

Do vậy, người ta dùng μ_{12} làm thước đo mức độ tương quan giữa X_1 và X_2. Vì μ_{12} là một đại lượng có thứ nguyên (bằng tích thứ nguyên của X_1 và X_2) nên để thuận tiện trong việc so sánh, phân tích thay cho μ_{12} người ta dùng đại lượng vô thứ nguyên:

$$\rho_{12} = \frac{\mu_{12}}{\sqrt{\mu_{11}\mu_{22}}} \tag{5.2.1}$$

và được gọi là hệ số tương quan giữa hai biến X_1 và X_2. Người ta gọi ρ_{12} là hệ số tương quan tổng thể hay hệ số tương quan lý thuyết và là một hằng số.

Hệ số tương quan có các tính chất sau đây:

1) Hệ số tương quan nhận giá trị trên đoạn [–1;1]: $-1 \le \rho_{12} \le 1$.

Thật vậy, ta có:

$$D\left[\frac{X_1}{\sqrt{DX_1}} \pm \frac{X_2}{\sqrt{DX_2}}\right] = \left[\left(\frac{X_1}{\sqrt{DX_1}} - M\left[\frac{X_1}{\sqrt{DX_1}}\right]\right) \pm \left(\frac{X_2}{\sqrt{DX_2}} - M\left[\frac{X_2}{\sqrt{DX_2}}\right]\right)\right]^2 =$$

$$= D\left[\frac{X_1}{\sqrt{DX_1}}\right] + D\left[\frac{X_2}{\sqrt{DX_2}}\right] \pm 2M\left[\left(\frac{X_1}{\sqrt{DX_1}} - M\left[\frac{X_1}{\sqrt{DX_1}}\right]\right)\left(\frac{X_2}{\sqrt{DX_2}} - M\left[\frac{X_2}{\sqrt{DX_2}}\right]\right)\right]$$

$$= \frac{1}{DX_1}DX_1 + \frac{1}{DX_2}DX_2 \pm 2\frac{1}{\sqrt{DX_1DX_2}}\mu_{12} = 2 \pm 2\frac{\mu_{12}}{\sqrt{\mu_{11}\mu_{22}}} = 2(1 \pm \rho_{12}) \ge 0$$

Hay $1 \pm \rho_{12} \ge 0 \Rightarrow$ đpcm

2) Điều kiện cần và đủ để $|\rho_{12}|=1$ là X_1 và X_2 có quan hệ hàm tuyến tính.

Điều kiện đủ:

Giả sử ta có quan hệ hàm tuyến tính giữa X_1 và X_2: $X_2 = a + bX_1$, với a, b là các hệ số hằng số. Khi đó:

$$\mu_{12} = M[(X_1 - MX_1)(X_2 - MX_2)] = M[(X_1 - MX_1)(a + bX_1 - a - bMX_1)] =$$

$$= M[b(X_1 - MX_1)^2] = b\mu_{11}$$

$$\mu_{22} = M[(X_2 - MX_2)^2] = M[(a + bX_1 - a - bMX_1)^2] = b^2M[(X_1 - MX_1)^2] = b^2\mu_{11}$$

$$\text{Vậy } \rho_{12} = \frac{\mu_{12}}{\sqrt{\mu_{11}\mu_{22}}} = \frac{b\mu_{11}}{\sqrt{b^2\mu_{11}^2}} = \frac{b}{|b|} = \begin{cases} 1 & \text{khi } b > 0 \\ -1 & \text{khi } b < 0 \end{cases}$$

Điều kiện cần:

Từ hệ thức $D\left[\frac{X_1}{\sqrt{DX_1}} \pm \frac{X_2}{\sqrt{DX_2}}\right] = 2(1 \pm \rho_{12})$ ta có:

Nếu $(1 \pm \rho_{12}) = 0$ thì $\left[\frac{X_1}{\sqrt{DX_1}} \pm \frac{X_2}{\sqrt{DX_2}}\right] = C = \text{Const}$

Từ đó suy ra $X_2 = \pm\sqrt{\frac{\mu_{22}}{\mu_{11}}} X_1 + C\sqrt{\mu_{22}}$, tức là giữa X_2 và X_1 tồn tại quan hệ hàm tuyến tính.

Do tính chất này nên hệ số tương quan được xem là đại lượng đặc trưng cho mức độ tương quan tuyến tính giữa hai biến.

5.2.2 Hệ số tương quan mẫu

Cho hai biến khí quyển X_1, X_2 với *n* cặp trị số quan sát:

$$\{x_{t1}, x_{t2}\} = \{(x_{11}, x_{12}), (x_{21}, x_{22}),..., (x_{n1}, x_{n2})\}$$

Khi đó mômen tương quan mẫu – ước lượng của mômen tương quan tổng thể μ_{12} – giữa X_1 và X_2 được xác định bởi:

$$R_{12} = \frac{1}{n}\sum_{t=1}^{n}(x_{t1} - \overline{x_1})(x_{t2} - \overline{x_2}) = \overline{(x_1 - \overline{x_1})(x_2 - \overline{x_2})} \qquad (5.2.2)$$

và hệ số tương quan mẫu:

$$r_{12} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_{t1} - \overline{x_1})(x_{t2} - \overline{x_2})}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_{t1} - \overline{x_1})^2}\sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_{t2} - \overline{x_2})^2}} = \frac{l_{12}}{\sqrt{l_{11}l_{22}}} \qquad (5.2.3)$$

trong đó:

$l_{12} = \sum_{t=1}^{n}(x_{t1} - \overline{x_1})(x_{t2} - \overline{x_2}) = nR_{12}$ là tổng của tích các độ lệch của X_1 và X_2 so với trung bình của chúng.

$l_{11} = \sum_{t=1}^{n}\left(x_{t1} - \overline{x_1}\right)^2 = n s_1^2$ – tổng bình phương các độ lệch của X_1 so với trung bình của nó.

$l_{22} = \sum_{t=1}^{n}\left(x_{t2} - \overline{x_2}\right)^2 = n s_2^2$ – tông bình phương các độ lệch của X_2 so với trung bình của nó.

$\overline{x_1} = \frac{1}{n}\sum_{t=1}^{n}x_{t1}$, $\overline{x_2} = \frac{1}{n}\sum_{t=1}^{n}x_{t2}$ – trung bình của X_1 và X_2

Hệ số tương quan mẫu r_{12} là ước lượng của hệ số tương quan tổng thể ρ_{12}. Nếu ρ_{12} là một hằng số thì trái lại r_{12} là một đại lượng ngẫu nhiên. Năm 1915

R.A.Fisher [3,5,6] đã tìm ra biểu thức chính xác của hàm mật độ xác suất của hệ số tương quan mẫu r_{12} trong trường hợp phân bố đồng thời của X_1 và X_2 là chuẩn:

$$f_n(r)=\frac{2^{n-3}}{\pi\Gamma(n-2)}(1-\rho^2)^{\frac{n-1}{2}}(1-r^2)^{\frac{n-4}{2}}\sum_{i=0}^{\infty}(\Gamma(\frac{n+i-1}{2}))^2\frac{(2\rho r)^i}{i!}, \qquad (5.2.4)$$

với $-1 \le r \le 1$. Ở đây, để tiện biểu diễn ta đã thay ký hiệu r_{12} bằng ký hiệu r.

Bằng phép biến đổi chuỗi luỹ thừa vế phải của biểu thức $f_n(r)$ người ta đã thu được dạng khác đối với mật độ xác suất của r:

$$f_n(r) = \frac{n-2}{\pi}(1-\rho^2)^{\frac{n-1}{2}}(1-r^2)^{\frac{n-4}{2}}\int_0^1\frac{x^{n-2}}{(1-\rho rx)^{n-1}}\frac{dx}{\sqrt{1-x^2}} \qquad (5.2.5)$$

Ta thấy rằng phân bố của r chỉ phụ thuộc vào dung lượng mẫu n và hệ số tương quan tổng thể ρ. Khi n = 2 thì $f_n(r) = 0$, điều đó phù hợp với sự kiện hệ số tương quan được tính từ tập mẫu chỉ có 2 quan trắc phải bằng ±1.

Kỳ vọng của hệ số tương quan mẫu r: M[r] = ρ

Phương sai của hệ số tương quan mẫu r:

$$D[r] = \frac{\rho^2}{4n}(\frac{\mu_{40}}{\mu_{20}^2}+\frac{\mu_{04}}{\mu_{02}^2}+\frac{2\mu_{22}}{\mu_{20}\mu_{20}}+\frac{4\mu_{22}}{\mu_{11}^2}-\frac{4\mu_{31}}{\mu_{11}\mu_{20}}-\frac{4\mu_{13}}{\mu_{11}\mu_{02}})$$

trong đó $\mu_{ij}= M\left[(X_1-MX_1)^i(X_2-MX_2)^j\right]$– các mômen trung tâm bậc i+j.

Để thuận tiện trong tính toán thực hành, nhất là việc ước lượng khoảng cho ρ, người ta thường dùng phép biến đổi sau đây của Fisher:

$$z = \frac{1}{2}\log\frac{1+r}{1-r},\ \zeta = \frac{1}{2}\log\frac{1+\rho}{1-\rho} \qquad (5.2.6)$$

Fisher đã chứng minh được rằng ngay cả với những giá trị *n* không lớn lắm biến z cũng phân bố xấp xỉ chuẩn với giá trị trung bình và phương sai được cho bởi biểu thức gần đúng sau:

$$M[z] = \zeta + \frac{\rho}{2(n-1)},\ D[z] = \frac{1}{n-3} \qquad (5.2.7)$$

Vì vậy khoảng tin cậy của ζ với độ tin cậy 1–α là:

$$(z-\frac{r}{2(n-1)}-u_\alpha\frac{1}{\sqrt{n-3}}, z-\frac{r}{2(n-1)}+u_\alpha\frac{1}{\sqrt{n-3}}) \qquad (5.2.8)$$

trong đó u_α nhận được từ phân bố chuẩn N(0,1) bởi hệ thức: $P(|u| \ge u_\alpha) = \alpha$. Từ đó ta nhận được khoảng tin cậy của ρ.

Trong trường hợp ρ = 0 thì biến $t = r\sqrt{\frac{n-2}{1-r^2}}$ có phân bố Student với n–2 bậc tự do. Hệ số tương quan mẫu r là ước lượng vững nhưng chệch của hệ số tương quan tổng thể ρ với độ chệch bằng $\frac{-\rho(1-\rho^2)}{2n}$. Do đó khi tính toán thực hành nếu nhận được r = 0 thì điều đó không có nghĩa là ρ bằng 0. Và ngược lại, nếu r≠0 thì cũng không hẳn là ρ khác 0. Nếu dung lượng mẫu nhỏ thì mặc dù ρ = 0 nhưng giá trị của r lại có thể có ý nghĩa. Vì vậy ta cần kiểm tra xem độ lớn của r có ý nghĩa thực sự hay không, hay nói cách khác cần kiểm nghiệm độ rõ rệt của r.

Để kiểm nghiệm, ta đặt giả thiết H_o: ρ = 0. Thay ρ ≈ r, với giới hạn tin cậy ban đầu d thì khi H_o đúng ta có $P(|r| \geq d) = \alpha$.

Đặt $$t = \frac{r}{\sqrt{1-r^2}/\sqrt{n-2}},\ t_\alpha = \frac{d}{\sqrt{1-r^2}/\sqrt{n-2}} \qquad (5.2.9)$$

Khi đó nếu H_o đúng thì: $P(|t| \geq t_\alpha) = \alpha$. Biến *t* trong (5.2.9) có phân bố Student (t) với n–2 bậc tự do. Từ đó ta xác định được t_α. Và chỉ tiêu kiểm nghiệm sẽ là:

Nếu $|t| \geq t_\alpha$ thì bác bỏ H_o và đưa ra kết luật r lớn rõ rệt

Nếu $|t| < t_\alpha$ thì chấp nhận H_o và kết luận r không lớn rõ rệt.

Ví dụ 5.2.1 Từ tập mẫu $\{x_t, y_t, t=1..11\}$ ta tính được hệ số tương quan r_{xy}=0.76. Hãy cho biết với giá trị nhận được như vậy thì hệ số tương quan có lớn rõ rệt không nếu lấy mức ý nghĩa α=0.01?

Để trả lời câu hỏi đặt ra ta cần kiểm nghiệm giả thiết: H_o: r_{xy}=0. Muốn vậy, ta tính đại lượng $t=\frac{r_{xy}}{\sqrt{1-r^2}/\sqrt{n-2}}=\frac{0.76}{\sqrt{1-0.76^2}/\sqrt{11-2}}$=3.51. Từ α=0.01 ta xác định được t_α từ phân bố Student: t_α=St(11–2,0.01) = 3.25.

Vì $|t|$=3.51> 3.25=t_α do đó ta bác bỏ giả thiết H_o và đưa ra kết luận r_{xy} lớn rõ rệt.

Ngoài việc kiểm tra độ rõ rệt của hệ số tương quan, trong thực tế người ta còn đánh giá sự có nghĩa của nó. Để xác định sự có nghĩa của *r* trước hết ta tính giá trị $H=|r|\sqrt{n-1} \equiv H(n, r)$. Tương ứng với các giá trị dung lượng mẫu n khác nhau, khi cho trước độ tin cậy p, tra bảng ta sẽ tính được trị số tới hạn H_o của H: $H_o = H(p,n)$. Trong bảng 5.1 đã cho các giá trị tới hạn H_0 ứng với các độ tin cậy *p* và dung lượng mẫu *n* khác nhau.

Từ đó chỉ tiêu kiểm nghiệm sự có nghĩa của r sẽ là:

Nếu $H(n,r) > H_o(p,n)$ thì kết luận r có nghĩa với độ tin cậy *i*

Nếu $H(n,r) \leq H_o(p,n)$ thì kết luận r không có nghĩa với độ tin cậy p.

Bảng 5.1 Giá trị tới hạn $H_0(p,n)$

	p					p		
n	0.90	0.95	0.99	0.999	n	0.95	0.99	0.999
10	1.65	1.90	2.29	2.62	25	1.941	2.475	3.026
11	1.65	1.90	2.32	2.68	26	1.941	2.479	3.037
12	1.65	1.92	2.35	2.73	27	1.492	2.483	3.047
13	1.65	1.92	2.37	2.77	28	1.943	2.487	3.056
14	1.65	1.92	2.39	2.81	29	1.493	2.490	3.064
15	1.65	1.92	2.40	2.85	30	1.944	2.492	3.071
16	1.65	1.93	2.41	2.87	35	1.947	2.505	3.102
17	1.65	1.93	2.42	2.90	40	1.949	2.514	3.126
18	1.65	1.93	2.43	2.92	45	1.950	2.521	3.145
19	1.65	1.93	2.44	2.94	50	1.951	2.527	3.161
20	1.65	1.94	2.45	2.96	60	1.953	2.535	3.830
21	1.65	1.94	2.45	2.98	70	1.954	2.541	3.190
22	1.65	1.94	2.46	2.99	80	1.955	2.546	3.209
23	1.65	1.94	2.47	3.00	90	1.956	2.550	3.219
24	1.65	1.94	2.47	3.02	100	1.956	2.553	3.226
					∞	1.960	2.576	3.291

5.2.3 Cách tính hệ số tương quan mẫu

Cho hai biến ngẫu nhiên X_1, X_2 với *n* cặp trị số quan sát:

$$\{x_{t1}, x_{t2}\} = \{(x_{11}, x_{12}), (x_{21}, x_{22}),..., (x_{n1}, x_{n2})\}$$

Từ tập mẫu này có thể tính hệ số tương quan giữa X_1, X_2 theo các phương pháp sau đây.

5.2.3.1 Phương pháp tính trực tiếp

Phương pháp trực tiếp tính hệ số tương quan mẫu là tính theo công thức (5.2.3). Thế nhưng, trong thực hành người ta thường biến đổi và đưa nó về dạng khác.

$$R_{12} = \overline{(x_1 - \overline{x_1})(x_2 - \overline{x_2})} = \overline{x_1x_2 - \overline{x_1}x_2 + \overline{x_2}x_1 - \overline{x_1}\,\overline{x_2}} = \overline{x_1x_2 - \overline{x_1}\,\overline{x_2}}$$

$$= \overline{x_1x_2} - \overline{x_1}.\overline{x_2} = \frac{1}{n}\sum_{t=1}^{n} x_{t1}x_{t2} - \frac{1}{n}\sum_{t=1}^{n} x_{t1}\frac{1}{n}\sum_{t=1}^{n} x_{t2} \qquad (5.2.10)$$

$$s_1^2 = \overline{(x_1 - \overline{x_1})^2} = \overline{(x_1)^2 - 2x_1\overline{x_1} + (\overline{x_1})^2} = \overline{(x_1)^2} - (\overline{x_1})^2$$

$$= \frac{1}{n}\sum_{t=1}^{n}(x_{t1})^2 - (\frac{1}{n}\sum_{t=1}^{n} x_{t1})^2 \qquad (5.2.11)$$

Tương tự ta có:

$$s_2^2 = \frac{1}{n}\sum_{t=1}^{n}(x_{t2})^2 - (\frac{1}{n}\sum_{t=1}^{n}x_{t2})^2 \qquad (5.2.12)$$

Kết hợp (5.2.10)–(5.2.12) ta nhận được: $r_{12} = \frac{R_{12}}{s_1 s_2}$ (5.2.13)

Hoặc có thể tính theo công thức:

$$r_{12} = \frac{\sum_{t=1}^{n}x_{t1}x_{t2} - \frac{1}{n}\sum_{t=1}^{n}x_{t1}\sum_{t=1}^{n}x_{t2}}{\sqrt{\sum_{t=1}^{n}(x_{t1})^2 - \frac{1}{n}(\sum_{t=1}^{n}x_{t1})^2}\sqrt{\sum_{t=1}^{n}(x_{t2})^2 - \frac{1}{n}(\sum_{t=1}^{n}x_{t2})^2}} \qquad (5.2.14)$$

Ví dụ 5.2.2 Trong bảng 5.2 dẫn ra số liệu quan trắc tổng lượng mưa tháng 1 của hai trạm mà ta đặt chúng là hai biến X_1, X_2 và kết quả các bước tính trung gian theo công thức (5.2.14). Cột thứ nhất chỉ số thứ tự năm (t). Hai cột tiếp theo của bảng chứa số liệu hai chuỗi $\{x_{t1}\}$ và $\{x_{t2}\}$. Cột thứ tư là tích từng cặp (x_{t1},x_{t2}), hai cột cuối cùng chứa bình phương các giá trị x_{t1} và x_{t2}. Dòng cuối cùng của bảng là tổng theo từng cột.

Bảng 5.2 Số liệu lượng mưa tháng 1 và những kết quả tính trung gian

t	x_{t1}	x_{t2}	$x_{t1}x_{t2}$	$(x_{t1})^2$	$(x_{t2})^2$
1	10.6	19.1	202.46	112.36	364.81
2	0.9	11.8	10.62	0.81	139.24
3	9.6	86.9	834.24	92.16	7551.61
4	2.0	16.4	32.80	4.00	268.96
5	38.3	12.4	474.92	1466.89	153.76
6	0.9	9.6	8.64	0.81	92.16
7	46.7	26.8	1251.56	2180.89	718.24
8	142.5	48.7	6939.75	20306.25	2371.69
9	68.2	28.9	1970.98	4651.24	835.21
10	54.1	87.4	4728.34	2926.81	7638.76
11	25.9	66.1	1711.99	670.81	4369.21
12	41.3	42.7	1763.51	1705.69	1823.29
13	11.8	37.7	444.86	139.24	1421.29
14	5.0	55.1	275.50	25.00	3036.01
15	30.0	104.1	3123.00	900.00	10836.81
16	21.8	33.9	739.02	475.24	1149.21
17	26.0	39.0	1014.00	676.00	1521.00
18	6.0	38.0	228.00	36.00	1444.00
19	15.0	116.0	1740.00	225.00	13456.00
Tổng	556.6	880.6	27494.19	36595.20	59191.26

Đối sánh với từng thành phần trong (5.2.14) ta có: n=19

$\sum_{t=1}^{n} x_{t1}x_{t2} = 27494.19$, $\frac{1}{n}\sum_{t=1}^{n} x_{t1}\sum_{t=1}^{n} x_{t2}$ =556.6*880.6/19=25796,

$\sum_{t=1}^{n} (x_{t1})^2$ =36595.20, $\frac{1}{n}(\sum_{t=1}^{n} x_{t1})^2$ =16305.45

$\sum_{t=1}^{n} (x_{t2})^2$ =59191.26, $\frac{1}{n}(\sum_{t=1}^{n} x_{t2})^2$ =40813.49

Sau khi thay vào và tính ra ta được r_{12}=0.087894.

5.2.3.2 Phương pháp biến đổi tương đương.

Khi giá trị của các thành phần trong chuỗi khá lớn việc tính toán trực tiếp theo các công thức (5.2.10)–(5.2.14) thường gặp trở ngại, phức tạp và dễ gây sai số, nhất là quá trình tính toán được tiến hành thủ công. Do đó, trong nhiều trường hợp, để đơn giản ta sử dụng phép biến đổi sau đây:

$$y_{t1} = d_1 x_{t1} - C_1 \qquad (*)$$

$$y_{t2} = d_2 x_{t2} - C_2 \qquad (**)$$

trong đó d_1, d_2, C_1, C_2 là những hằng số nào đó, mà trong những trường hợp cụ thể, sẽ được chọn sao cho thích hợp. Chẳng hạn, khi xử lý chuỗi số liệu nhiệt độ ta thấy chúng thường dao động xung quanh trị số 20 (^{0}C), vậy có thể chọn C=20; các giá trị khí áp thường lên xuống quanh giá trị 1000 (mb) thì chọn C=1000,...

Với phép biến đổi (*), (**) ta có:

$$x_{t1} = \frac{y_{t1} + C_1}{d_1}, \quad x_{t2} = \frac{y_{t2} + C_2}{d_2}$$

Hay

$$\overline{x_1} = \frac{\overline{y_1} + C_1}{d_1}, \quad \overline{x_2} = \frac{\overline{y_2} + C_2}{d_2}$$

Suy ra

$$l_{12} = \sum\left(\frac{y_{t1} + C_1}{d_1} - \frac{\overline{y_1} + C_1}{d_1}\right)\left(\frac{y_{t2} + C_2}{d_2} - \frac{\overline{y_2} + C_2}{d_2}\right)$$

$$= \frac{1}{d_1 d_2}\sum(y_{t1} - \overline{y_1})(y_{t2} - \overline{y_2}) = \frac{l'_{12}}{d_1 d_2}$$

Tương tự ta được: $l_{11} = \frac{l'_{11}}{d_1^2}$, $l_{22} = \frac{l'_{22}}{d_2^2}$

Do đó:

$$r_{12} = \frac{l_{12}}{\sqrt{l_{11}l_{22}}} = \frac{\frac{1}{d_1 d_2} l'_{12}}{\frac{1}{d_1 d_2}\sqrt{l_{11}l_{22}}} = \frac{l'_{12}}{\sqrt{l'_{11}l'_{22}}} = r'_{12} \qquad (5.2.15)$$

Như vậy, qua phép biến đổi (*) và (**) hệ số tương quan vẫn không bị thay đổi.

5.2.4 Ma trận tương quan

Trong thực tế ta thường gặp những bài toán mà ở đó đòi hỏi phải khảo sát mối quan hệ tương quan giữa các biến khác nhau của một tập nhiều hơn hai biến. Khi đó ta không chỉ có một hệ số tương quan mà là một ma trận tương quan.

Xét tập hợp *m* biến ngẫu nhiên X_1, X_2,..., X_m. Hệ số tương quan tổng thể giữa các biến X_j và X_k được xác định bởi hệ thức:

$$\rho_{jk} = \frac{\mu_{jk}}{\sqrt{\mu_{jj}\mu_{kk}}}, \; j,k=1..m \tag{5.2.16}$$

trong đó μ_{jk} là mômen tương quan giữa X_j và X_k, μ_{jj} là phương sai của X_j. Tập hợp các hệ số tương quan ρ_{jk} lập thành ma trận tương quan:

$$(\rho_{jk}) = \begin{pmatrix} \rho_{11} & \cdots & \rho_{1m} \\ \cdots & \cdots & \cdots \\ \rho_{m1} & \cdots & \rho_{mm} \end{pmatrix} \tag{5.2.16'}$$

Ma trận tương quan là một ma trận đối xứng có các phần tử trên đường chéo chính bằng 1.

Nếu X_{tj}, j=1..m, t=1..n là số liệu thực nghiệm của các biến X_j thì ước lượng r_{jk} của ρ_{jk} được xác định bởi:

$$r_{jk} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_{tj} - \overline{x_j})(x_{tk} - \overline{x_k})}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_{tj} - \overline{x_j})^2}\sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_{tk} - \overline{x_k})^2}} \tag{5.2.17}$$

trong đó $\overline{x_j} = \frac{1}{n}\sum_{t=1}^{n} x_{tj}$ là trung bình của biến X_j, j=1..m.

Tập hợp các hệ số tương quan r_{jk} cũng lập thành một ma trận đối xứng:

$$(r_{jk}) = \begin{pmatrix} r_{11} & \cdots & r_{1m} \\ \cdots & \cdots & \cdots \\ r_{m1} & \cdots & r_{mm} \end{pmatrix} \tag{5.2.17'}$$

5.2.5 Khảo sát mối quan hệ tương quan giữa hai biến

Việc đánh giá mối quan hệ tương quan giữa hai biến có thể được tiến hành thông qua việc xem xét hệ số tương quan giữa chúng tính được từ tập mẫu. Giá trị tuyệt đối của hệ số tương quan càng lớn thì mối quan hệ tuyến tính giữa hai biến càng chặt chẽ. Hệ số tương quan dương phản ánh mối quan hệ cùng chiều (đồng biến), ngược lại, hệ số tương quan âm biểu thị mối quan hệ ngược (nghịch biến)

giữa hai biến. Tuy nhiên, như đã chỉ ra trong mục 5.2.1, khái niệm hệ số tương quan được trình bày trên đây mới chỉ cho phép ta đánh giá được mối quan hệ tuyến tính giữa hai tập mẫu.

Thực tế trong nhiều trường hợp, khi khảo sát mối quan hệ giữa hai biến, người ta chưa cần hoặc thậm chí không cần những kết quả tính toán chính xác của hệ số tương quan, mà trước hết muốn biết bức tranh khái quát về quan hệ giữa hai tập mẫu để từ đó đưa ra quyết định cho những bước xử lý tiếp theo. Đa số trong những trường hợp như vậy người ta thường quan tâm đến khả năng tồn tại mối quan hệ tương quan tuyến tính giữa các biến khảo sát. Khi đó thay cho việc tính hệ số tương quan trên đây, người ta có thể xây dựng các đồ thị điểm biểu diễn sự phụ thuộc hoặc tính các hệ số tương quan giản lược.

Ngày nay nhờ có phương tiện máy tính, việc biểu diễn đồ thị điểm để khảo sát sơ bộ sự phụ thuộc tương quan giữa các biến đã trở nên phổ biến và rất có hiệu quả. Đồ thị điểm thông thường được biểu diễn trên hệ tọa độ vuông góc trong mặt phẳng, với hai trục tọa độ biểu thị sự biến thiên của hai biến X, Y (hay X_1, X_2). Mỗi một cặp quan trắc $\{x_t, y_t\}$ được biểu diễn bởi một điểm trên mặt phẳng. Căn cứ vào sự phân bố của tập hợp các điểm này ta có thể đánh giá được quan hệ giữa các biến.

Hình 5.1 dẫn ra một ví dụ đồ thị điểm biểu diễn mối quan hệ giữa nhiệt độ tối cao (T_x) và nhiệt độ tối thấp (T_m) trong những ngày tháng 1 ở một trạm. Từ đồ thị ta có thể thấy sự phân bố "hỗn loạn" của tập hợp các điểm trên mặt phẳng. Có những chỗ các điểm qui tụ khá dày đặc nhưng cũng có những chỗ chỉ rải rác 1–2 điểm. Sự phân bố tản mạn đó của các điểm biểu thị mối quan hệ "kém chặt chẽ" giữa hai yếu tố T_x và T_m. Tuy vậy, xét một cách tổng thể ta thấy giữa hai yếu tố này tồn tại sự phụ thuộc lẫn nhau: Dường như nhiệt độ tối thấp bé có liên quan tới giá trị của nhiệt độ tối cao bé, và nhiệt độ tối thấp lớn có xu hướng kéo theo nhiệt độ tối cao lớn. Ngoài ra, đồ thị còn cho thấy trong khoảng nhiệt độ T_m từ 12–18°C mối liên hệ giữa T_m và T_x có vẻ yếu hơn nhiều so với trường hợp giá trị T_m nằm ngoài khoảng đó.

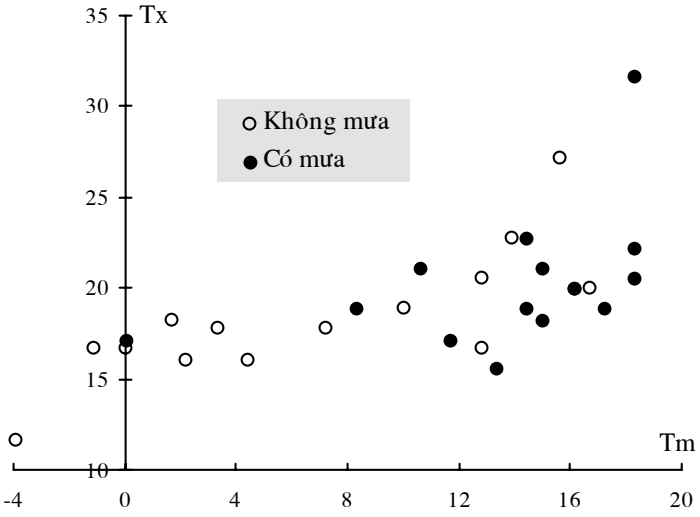
Việc chia tập số liệu ra làm hai trường hợp có mưa và không mưa sẽ làm đa dạng hóa đồ thị, cho phép khảo sát tỷ mỷ hơn mối quan hệ giữa hai biến. Hiện tượng các điểm ứng với trường hợp có mưa qui tụ vào khoảng nhiệt độ tối thấp từ 12–18°C gợi cho ta một nhận định rằng trong những ngày có mưa mối quan hệ giữa hai biến trở nên "kém chặt chẽ" hơn. Mặt khác, điều đó làm cho ta liên tưởng đến xác suất có điều kiện đã xét trước đây.

Với mục đích đánh giá mức độ tương quan tuyến tính giữa hai biến một cách nhanh chóng nhưng không cần độ chính xác cao ngoài việc sử dụng phương pháp đồ thị điểm đôi khi người ta còn tính hệ số tương quan hạng (*range correlation*

coefficient). Khác với hệ số tương quan mà ta đã xét, hệ số tương quan hạng được tính không phải với chính các giá trị của số liệu mà với *thứ hạng lớn bé* của chúng trong toàn tập mẫu. Nghĩa là từ tập mẫu ban đầu $\{x_t, y_t, t=1..n\}$ ta biến đổi thành tập mới $\{u_t, v_t, t=1..n\}$ trong đó u_t, v_t tương ứng chỉ các thành phần x_t, y_t *được xếp thứ bao nhiêu* trong bảng xếp hạng từ nhỏ nhất đến lớn nhất của mỗi chuỗi. Rõ ràng, các tập các thành phần của tập mới phải thỏa mãn $1 \le u_t, v_t \le n$. Hệ số tương quan hạng được tính bởi công thức:

$$r_{range} = 1 - \frac{6\sum_{t=1}^{n} D_t^2}{n(n-1)(n+1)} \qquad (5.2.18)$$

trong đó $D_t = u_t - v_t$ là hiệu giữa các thứ hạng của x_t và y_t trong từng chuỗi.



Hình 5.1 Đồ thị điểm biểu diễn sự phụ thuộc giữa T_x và T_m

Ví dụ 5.2.3 Bảng 5.3 dẫn ra kết quả tính hệ số tương quan hạng cho tập mẫu nhiệt độ tối thấp (T_m) và nhiệt độ tối cao (T_x). Cột thứ nhất và cột thứ hai chứa số liệu ban đầu. Cột 3, 4, 5 chứa các giá trị tương ứng của T_m, T_x trong tập ban đầu và kết quả xếp hạng chúng. Cột 6 và cột 7 chứa giá trị hạng của từng thành phần tương ứng trong cột 1 và cột 2. Cột cuối cùng là hiệu giữa các hạng. Chẳng hạn, u_1=4 có nghĩa là ứng với T_{m1}=12.8 ở cột 1, khi đối chiếu giá trị này ở kết quả xếp hạng (cột 3 và cột 5) ta nhận được hạng của T_{m1} bằng 4. Tương tự như vậy với v_1=8 (giá trị T_{x1}=20.6, tìm giá trị này ở cột 4 rồi đối chiếu sang cột 5 ta có hạng bằng 8). Hiệu D_1 = 4–8=–4.

Sử dụng kết quả tính trung gian ở bảng 5.3 kết hợp với công thức (5.2.18) với n=10 ta nhận được r_{range} = 0.4546.

Bảng 5.3 Tính hệ số tương quan hạng

Số liệu ban đầu		Kết quả xếp hạng			Số liệu xếp hạng		
T_m	T_x	T_m	T_x	Hạng	u_t	v_t	D_t
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
12.8	20.6	1.7	16.1	1	4	8	–4
16.1	20.0	4.4	18.0	2	9	7	2
14.4	18.6	10.0	18.3	3	6	5	1
1.7	18.0	12.8	18.4	4	1	2	–1
4.4	16.1	13.9	18.6	5	2	1	1
10.0	18.4	14.4	18.9	6	3	4	–1
13.9	22.8	14.8	20.0	7	5	9	–4
14.8	23.0	15.0	20.6	8	7	10	–3
15.0	18.3	16.1	22.8	9	8	3	5
17.2	18.9	17.2	23.0	10	10	6	4

5.3 Hồi qui tuyến tính một biến

5.3.1 Khái niệm về hồi qui

Xét mối quan hệ giữa hai biến ngẫu nhiên X và Y. Khi đó có thể xảy ra hai trường hợp sau đây:

- Giữa chúng có mối quan hệ phụ thuộc hàm nếu tồn tại một hàm f nào đó sao cho có thể biểu diễn được X = f(Y).
- Giữa chúng có mối quan hệ phụ thuộc thống kê nếu mỗi giá trị x của X tương ứng với một hàm phân bố (hoặc hàm mật độ) có điều kiện F(y/x) (hoặc f(y/x)) của Y. Ta gọi mối quan hệ phụ thuộc này là sự phụ thuộc tương quan giữa hai biến ngẫu nhiên.

Để nghiên cứu mối phụ thuộc tương quan giữa hai biến X và Y trên cơ sở tập mẫu quan trắc $\{(x_t,y_t), t=1..n\}$ ta cần phải chọn dạng lý thuyết của phân bố đồng thời F(x,y), hoặc dạng hàm mật độ đồng thời f(x,y), sau đó phải ước lượng các tham số này. Từ đó ta tìm được mật độ phân bố có điều kiện:

$$f(y/x) = \frac{f(x,y)}{f_1(x)}, \qquad f(x/y) = \frac{f(x,y)}{f_2(y)} \tag{5.3.1}$$

trong đó $f_1(x)$, $f_2(y)$ là các hàm mật độ riêng của X và Y.

(Chú ý rằng, trong mục này và một số mục tiếp theo ta đã thay đổi một cách tự nhiên ký hiệu các biến ngẫu nhiên X, Y thay cho ký hiệu trước đây vẫn dùng là X_1, X_2. Sự thay đổi này hoàn toàn không ảnh hưởng tới bản chất của vấn đề. Tuy

nhiên, do thói quen cố hữu trong toán học, nếu ta dùng ký hiệu mới này thì khái niệm hàm (Y) và đối số (X) tỏ ra dễ chấp nhận khi trình bày ?!. Sau này, ta sẽ quay lại ký hiệu trước đây).

Như vậy việc nghiên cứu sự phụ thuộc tương quan như trên là hết sức cồng kềnh và phức tạp. Do đó trong thực tế người ta chỉ giới hạn xét mối quan hệ phụ thuộc giữa X và một số đặc trưng có điều kiện của Y, như kỳ vọng, trung vị, mốt,... trong đó phổ biến hơn cả là nghiên cứu mối quan hệ giữa X và kỳ vọng có điều kiện M[Y/X]:

$$m_y(x) = M[X/Y = x] = \int_{-\infty}^{+\infty} y f(y/x)dy \qquad (5.3.2)$$

Và người ta gọi sự phụ thuộc này là phụ thuộc hồi qui: Hồi qui của Y lên X. Hệ thức (5.3.2) thông thường được biểu diễn dưới dạng:

$$y = m_y(x) \qquad (5.3.3)$$

Quan hệ (5.3.3) được gọi là phương trình hồi qui I hay đường hồi qui I. Nếu quan hệ này là một hàm tuyến tính thì hồi qui được gọi là hồi qui tuyến tính. Tuy nhiên, trong trường hợp tổng quát (5.3.3) là một hàm bất kỳ.

Một tính chất quan trọng của hồi qui I là tính cực tiểu:

Nếu ta tìm được một hàm g(X) sao cho $M[Y - g(X)]^2 \text{ — } \min$

thì $$g(X) = M[Y/X], \text{ hay } g(x) = m_y(x). \qquad (5.3.4)$$

Vì quan hệ (5.3.3) là một đường bất kỳ mà việc biểu diễn giải tích nó nói chung rất khó khăn, thậm chí không thể được cho nên trong thực tế thay cho (5.3.3) người ta xấp xỉ nó trong một lớp hàm *f* xác định nào đó đã biết:

$$y \approx \hat{y} = f(x) \qquad (5.3.5)$$

Trong trường hợp này hàm hồi qui tìm được gọi là hồi qui II. Nếu hàm hồi qui II được xác định bằng phương pháp bình phương tối thiểu thì nó được gọi là hồi qui bình phương trung bình. Trường hợp đơn giản nhất của hồi qui bình phương trung bình là hồi qui bình phương trung bình tuyến tính–f(x) là hàm bậc nhất.

Từ nay trở đi, nếu không nói gì thêm, ta sẽ hiểu hồi qui II là hồi qui bình phương trung bình và được gọi một cách đơn giản là hồi qui II.

Nếu hồi qui II (5.3.5) là tuyến tính, khi đó ta có thể viết:

$$Y = f(X) = \alpha + \beta X$$

Hay $$\hat{y} = f(x) = \alpha + \beta x$$

Ta có thể chứng minh được rằng để f(x) xấp xỉ tốt nhất theo nghĩa bình phương tối thiểu của hồi qui I thì các hệ số α và β sẽ được xác định bởi:

$$\alpha = M[Y] - \beta M[X], \quad \beta = \mu_{12}/\mu_{11}$$

trong đó μ_{12} là mômen tương quan giữa X và Y còn $\mu_{11} = D[X]$. Ta sẽ quay trở lại vấn đề này khi trình bày cách xác định các hệ số hồi qui thực nghiệm mà chúng là ước lượng của α và β trong mục sau.

5.3.2 Xây dựng phương trình hồi qui tuyến tính một biến từ số liệu thực nghiệm

Cho hai biến khí quyển X và Y với *n* cặp trị số quan sát $\{(x_t, y_t), t=1..n\}$. Xét sự phụ thuộc hồi qui II của Y lên X là hồi qui tuyến tính, tức là:

$$y \approx \hat{y} = a_o + a_1 x \tag{5.3.6}$$

trong đó a_o và a_1 là các hệ số phải tìm. Chúng là các giá trị ước lượng của tham số lý thuyết α và β trong phương trình $\hat{y} = \alpha + \beta x$.

Với các trị số quan sát x_t của X ta có các giá trị của Y tính được theo (5.3.6) là:

$$\hat{y}_t = a_o + a_1 x_t, \ (t=1..n) \tag{5.3.6'}$$

Các trị số quan trắc thực nghiệm y_t và giá trị tính toán (ước lượng) của Y theo (5.3.6') sai khác nhau một lượng bằng $\delta_t = y_t - \hat{y}_t$, chúng được gọi là sai số của phép xấp xỉ $y = m_y(x)$ bởi (5.3.6). Để phép xấp xỉ này là tốt nhất theo nghĩa bình phương tối thiểu các hệ số a_o và a_1 phải được xác định sao cho tổng bình phương các sai số δ_t phải đạt nhỏ nhất:

$$\sum_{t=1}^{n} \delta_t^2 = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2 \qquad \to \min$$

Xem rằng tổng các bình phương sai số như là hàm của các hệ số a_o, a_1, khi đó chúng phải thỏa mãn điều kiện:

$$R(a_o,a_1) = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2 \ \to \min \tag{5.3.7}$$

Người ta đã chứng minh được rằng, để $R(a_o,a_1)$ đạt cực tiểu thì các đạo hàm riêng của $R(a_o,a_1)$ theo a_o và a_1 phải đồng thời triệt tiêu:

$$\frac{\partial R(a_o,a_1)}{\partial a_o} = \frac{\partial R(a_o,a_1)}{\partial a_1} = 0$$

Từ đó ta nhận được hệ phương trình với các ẩn số a_o và a_1:

$$\begin{cases} \dfrac{\partial R(a_o, a_1)}{\partial a_0} = -2\sum_{t=1}^{n}(y_t - a_o - a_1 x_t) = 0 \\ \dfrac{\partial R(a_o, a_1)}{\partial a_1} = -2\sum_{t=1}^{n}(y_t - a_o - a_1 x_t)x_t = 0 \end{cases}$$

Hay:
$$\begin{cases} \sum_{t=1}^{n}(y_t - a_o - a_1 x_t) = 0 \\ \sum_{t=1}^{n}(y_t - a_o - a_1 x_t)x_t = 0 \end{cases} \qquad (5.3.8)$$

Từ phương trình thứ nhất trong hệ (5.3.8) ta có:

$$\sum_{t=1}^{n}(y_t - a_o - a_1 x_t) = 0.$$

Suy ra:
$$a_o = \bar{y} - a_1 \bar{x} \qquad (5.3.9)$$

Thay (5.3.9) vào phương trình thứ hai của (5.3.8) ta nhận được:

$$\sum_{t=1}^{n}(y_t - a_o - a_1 x_t)x_t = \sum_{t=1}^{n}(y_t - \bar{y} + a_1\bar{x} - a_1 x_t)x_t = 0$$

Hay
$$\sum_{t=1}^{n}(y_t - \bar{y})x_t - a_1\sum_{t=1}^{n}(x_t - \bar{x})x_t = 0$$

Do đó:
$$a_1 = \frac{\sum_{t=1}^{n}(y_t - \bar{y})x_t}{\sum_{t=1}^{n}(x_t - \bar{x})x_t}$$

Vì $\sum_{t=1}^{n}(y_t - \bar{y})\bar{x} = 0$ và $\sum_{t=1}^{n}(x_t - \bar{x})\bar{x} = 0$ nên ta có:

$$a_1 = \frac{\sum_{t=1}^{n}(y_t - \bar{y})x_t - \sum_{t=1}^{n}(y_t - \bar{y})\bar{x}}{\sum_{t=1}^{n}(x_t - \bar{x})x_t - \sum_{t=1}^{n}(x_t - \bar{x})\bar{x}} = \frac{\sum_{t=1}^{n}(y_t - \bar{y})(x_t - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} = \frac{l_{xy}}{l_{xx}} \qquad (5.3.10)$$

Hay:
$$a_1 = \frac{l_{xy}}{l_{xx}} = \frac{l_{xy}\sqrt{l_{yy}}}{\sqrt{l_{xx}l_{yy}}\sqrt{l_{xx}}} = \frac{r_{xy}\sqrt{l_{yy}}}{\sqrt{l_{xx}}} = r_{xy}\frac{s_y}{s_x} \qquad (5.3.11)$$

Như vậy, phương trình (5.3.6) với các hệ số a_o và a_1 được tính theo (5.3.9) và (5.3.10) hoặc (5.3.11) xác định mối quan hệ hồi qui II của Y lên X. Nó được gọi là phương trình hồi qui tuyến tính một biến (một biến độc lập). Người ta gọi Y (hay y) là biến phụ thuộc, còn X (hay x) là biến độc lập.

Nếu không xét trực tiếp tập số liệu $\{(x_t,y_t),t=1..n\}$ mà thay cho nó ta sử dụng tập số liệu chuẩn hoá $\{(x'_t, y'_t), t=1..n\}$:

$$x'_t = \frac{x_t - \bar{x}}{s_x},\ y'_t = \frac{y_t - \bar{y}}{s_y}$$

thì, bằng các phép biến đổi tương tự trên đây ta nhận được:

$$a'_0 = 0 \text{ và } a'_1 = r_{xy}$$

Ví dụ 5.3.1: Từ số liệu nhiệt độ tháng 5 trạm A (biến Y – cột 1) và trạm B (biến X – cột 2) cho trong bảng 5.4, sau khi tiến hành các bước tính trung gian (ở các cột tiếp theo) ta nhận được:

$\bar{x}$ = 25,9; $\bar{y}$ =22,9; l_{xy} = 7,588; l_{xx} = 18,624;

$a_1 = l_{xy}/l_{xx}$= 7,588/18,624 = 0,407;

$a_0 = \bar{y} - a_1.\ \bar{x}$ = 22,9 – 0,407 x 25,9 = 12,361;

Vậy phương trình hồi qui tuyến tính giữa Y và X có dạng:

$$y = 12{,}361 + 0{,}407.\mathbf{X}$$

Bảng 5.4 Các bước tính hệ số hồi qui giữa y và x

y	x	$y-\bar{y}$	$x-\bar{x}$	$(y-\bar{y})(x-\bar{x})$	$(x-\bar{x})^2$
22,7	27,7	–0,2	1,8	–0,4048	3,0976
23,8	26,0	0,9	0,1	0,0522	0,0036
23,7	26,5	0,8	0,6	0,4312	0,3136
21,3	24,3	–1,6	–1,6	2,6732	2,6896
22,5	28,0	–0,4	2,1	–0,8858	4,2436
25,1	27,4	2,2	1,5	3,1682	2,1316
23,3	25,9	0,4	0,0	–0,0148	0,0016
23,8	24,4	0,9	–1,5	–1,3398	2,3716
21,2	24,3	–1,7	–1,6	2,8372	2,6896
21,9	24,9	–1,0	–1,0	1,0712	1,0816
$\bar{y}$ =22,9	$\bar{x}$ =25,9			l_{xy}=7,5880	l_{xx}=18,6240

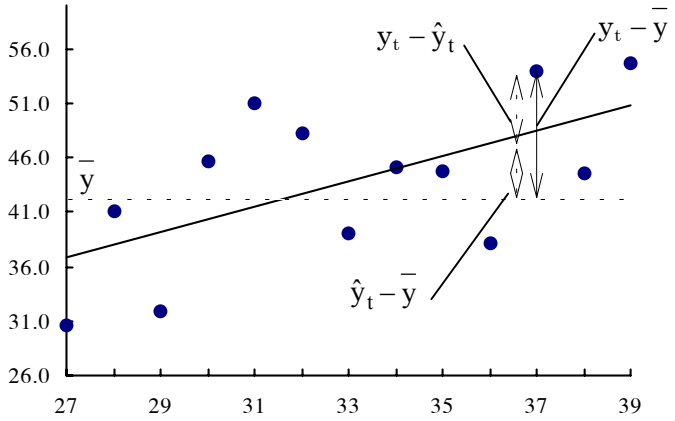
5.3.3 Phân tích phương sai phương trình hồi qui tuyến tính một biến

Phương trình hồi qui $\hat{y}=a_o+a_1x$ là hệ thức biểu thị mối quan hệ tuyến tính giữa hai biến Y và X. Tuy nhiên, do những dao động ngẫu nhiên mà các điểm thực

nghiệm (x_t, y_t) nói chung thường phân bố xoay quanh đường thẳng hồi qui, tức là có sự sai khác giữa y_t và $\hat{y}_t$. Mặt khác, các giá trị quan trắc y_t của Y cũng dao động biến đổi xung quanh giá trị trung bình $\bar{y}$ (hình 5.2). Những dao động của y_t xung quanh $\bar{y}$ thường do nhiều nguyên nhân gây nên. Phân tích phương sai là xem xét vai trò của các nguyên nhân tạo nên những biến đổi của Y.

Mức độ biến động của Y được đánh giá thông qua tổng bình phương các độ lệch của y_t khỏi giá trị trung bình của nó:

$$l_{yy} = \sum_{t=1}^{n}(y_t - \bar{y})^2 .$$



Hình 5.2 Sơ đồ phân tích phương sai

Từ hình 5.2 ta thấy, mỗi một thành phần $y_t - \bar{y}$ có thể được tách thành tổng 2 thành phần: Sự sai lệch của y_t so với đường hồi qui và sự sai lệch của giá trị hồi qui $\hat{y}_t$ so với trung bình $\bar{y}$:

$$y_t - \bar{y} = (y_t - \hat{y}_t) + (\hat{y}_t - \bar{y})$$

Do đó:
$$l_{yy} = \sum_{t=1}^{n}\left[(y_t - \hat{y}_t) + (\hat{y}_t - \bar{y})\right]^2 =$$

$$= \sum_{t=1}^{n}(y_t - \hat{y}_t)^2 + \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2 + 2\sum_{t=1}^{n}(y_t - \hat{y}_t)(\hat{y}_t - \bar{y})$$

Vì
$$\sum_{t=1}^{n}(y_t - \hat{y}_t)(\hat{y}_t - \bar{y}) = \sum_{t=1}^{n}(y_t - a_o - a_1x_t)(a_o + a_1x_t - \bar{y}) =$$

$$= \sum_{t=1}^{n}(y_t - \bar{y} - a_1\bar{x} - a_1x_t)(\bar{y} + a_1\bar{x} + a_1x_t - \bar{y}) =$$

$$= n(a_1(\overline{xy} - \bar{x}\bar{y}) - a_1^2(\overline{x^2} - \bar{x}^2)) = a_1r_{xy}s_xs_y - a_1^2s_x^2 = 0$$

Nên $$l_{yy} = \sum_{t=1}^{n}(y_t - \hat{y}_t)^2 + \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2 = Q + U \qquad (5.3.12)$$

trong đó $$Q = \sum_{t=1}^{n}(y_t - \hat{y}_t)^2, \quad U = \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2 \qquad (5.3.13)$$

Người ta gọi U là tổng bình phương các biến sai hồi qui, còn Q là tổng bình phương các biến sai thặng dư. Như vậy tổng bình phương các độ lệch của y khỏi giá trị trung bình là sự đóng góp của tổng bình phương các biến sai hồi qui và tổng bình phương các biến sai thặng dư.

Ta thấy đối với một tập mẫu thì $\bar{y}$ không đổi, do đó sự biến đổi $\hat{y}_t$ là nguyên nhân gây nên sự biến đổi của U. Đại lượng U đặc trưng cho mức đóng góp của nhân tố hồi qui trong độ phân tán của Y. Còn Q đặc trưng cho sự đóng góp ngoài hồi qui.

Ta có:

$$U = \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2 = \sum_{t=1}^{n}(a_o + a_1 x_t - a_o - a_1\bar{x})^2 = a_1^2\sum_{t=1}^{n}(x_t - \bar{x})^2 =$$

$$= a_1^2 l_{xx} = a_1\frac{l_{xy}}{l_{xx}}l_{xx} = a_1 l_{xy}$$

$$Q = l_{yy} - U = l_{yy} - a_1 l_{xy}$$

Do đó $$\frac{U}{l_{yy}} = \frac{a_1 l_{xy}}{l_{yy}} = \frac{l_{xy}^2}{l_{xx}l_{yy}} = r_{xy}^2. \qquad (5.3.14)$$

Như vậy, U càng lớn khi r_{xy} càng lớn. Tức là U càng lớn thì mức độ tương quan tuyến tính giữa X và Y càng chặt chẽ.

$$\frac{Q}{l_{yy}} = \frac{l_{yy} - U}{l_{yy}} = 1 - \frac{U}{l_{yy}} = 1 - r_{xy}^2 \qquad (5.3.15)$$

Từ đó suy ra rằng, r_{xy} càng lớn thì Q càng bé. Hồi qui được gọi là tốt nhất (lý tưởng) nếu tổng bình phương các biến sai thặng dư Q = 0. Khi đó r_{xy}^2=1, tất cả các điểm thực nghiệm đều nằm trên đường hồi qui. Nếu Q càng bé thì hồi qui càng tốt, điều đó cũng có nghĩa là nếu U càng lớn thì hồi qui càng có hiệu quả.

5.3.4 Sự dao động của các điểm thực nghiệm xung quanh đường hồi qui

Từ (5.3.15) ta thấy rằng khi r_{xy}^2=1 thì Q = 0. Như vậy ta có thể dùng đại lượng Q để đo mức độ dao động của các điểm thực nghiệm xung quanh đường hồi qui. Tuy nhiên, theo (5.3.13) thứ nguyên của Q bằng bình phương thứ nguyên của Y. Hơn

nữa, số bậc tự do của l_{yy} là n–1, của U là 1 (1 nhân tố), do đó số bậc tự do của Q là n–2. Chính vì vậy thay cho Q, trong thực tế người ta sử dụng đại lượng:

$$s = \sqrt{\frac{Q}{n-2}} \tag{5.3.16}$$

làm thước đo mức độ dao động của các giá trị thực nghiệm xung quanh trị số hồi qui. Giá trị của *s* càng nhỏ thì các điểm thực nghiệm càng nằm sát đường hồi qui. Đại lượng *s* được gọi là chuẩn sai thặng dư. Vậy chuẩn sai thặng dư là thước đo phần đóng góp trung bình của nhân tố ngoài hồi qui đối với sai số của phép hồi qui. Nói cách khác, *s* là chỉ tiêu phản ánh độ chính xác của hồi qui.

Khi $|r_{xy}| \neq 1$ thì các điểm thực nghiệm không nằm trùng hoàn toàn trên đường hồi qui $\hat{y} = a_o + a_1 x$ và sự tản mạn này có thể thấy được thông qua số liệu thực tế (hình 5.2). Vậy một vấn đề đặt ra là ứng với mỗi giá trị x_t xác định, quan hệ giữa y_t và $\hat{y}_t$ sẽ như thế nào?

Theo (5.3.16), nói chung các trị số y_t của Y dao động xung quanh $\hat{y}_t$ với mức trung bình là *s* và người ta đã xác định được rằng sự phân bố của y_t xung quanh $\hat{y}_t$ gần với phân bố chuẩn. Tức là:

$$y_t \in N(\hat{y}_t, s)$$

Hay
$$y'_t = \frac{y_t - \hat{y}_t}{s} \in N(0,1)$$

Từ đó ta có: $P(|y_t - \hat{y}_t| < s) = P\left(\left|\frac{y_t - \hat{y}_t}{s}\right| < 1\right) = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-\frac{1}{2}t^2} dt \approx 0.68$

Như vậy, xác suất để các giá trị y_t dao động xung quanh $\hat{y}_t$ trong khoảng 1s bằng 68%. Hay nói cách khác, có khoảng 68% số điểm thực nghiệm nằm trong phạm vi ±1s kể từ đường hồi qui.

Bằng cách tính tương tự, ta có:

$$P(|y_t - \hat{y}_t| < 2s) \approx 0.95 \text{ và } P(|y_t - \hat{y}_t| < 3s) \approx 0.997$$

Tức là có khoảng 95% số điểm thực nghiệm rơi vào miền $\hat{y}_t \pm 2s$ và 99.7% số điểm rơi vào miền $\hat{y}_t \pm 3s$. Vậy hầu như tất cả các giá trị y_t đều nằm trong khoảng $\hat{y}_t \pm 3s$.

5.3.5 Đánh giá chất lượng phương trình hồi qui

Có thể nhận thấy rằng, việc đánh giá chất lượng phương trình hồi qui (5.3.6) là "tốt" hay "không tốt" hoặc "xấu" căn cứ vào hệ số tương quan r_{xy} hoặc theo giá trị

chuẩn sai thặng dư *s*, dù sao vẫn mang dáng dấp định tính. Trong thực tế ta cần khẳng định rằng phương trình hồi qui $\hat{y} = a_o + a_1x$ có dùng được hay không.

Như đã biết, phương trình hồi qui $\hat{y} = a_o + a_1x$ được xây dựng trên cơ sở tập các số liệu thực nghiệm. Nó là ước lượng tốt nhất của phương trình hồi qui lý thuyết. Tuy nhiên chất lượng của nó lại phụ thuộc vào mức độ quan hệ tuyến tính giữa X và Y. Để khẳng định khả năng dùng được của phương trình này ta cần xác định xem Y có thực sự phụ thuộc tuyến tính vào X hay không, tức cần kiểm nghiệm giả thiết:

$$H_o: a_1 = 0$$

Nếu H_0 đúng thì phương trình hồi qui không dùng được. Muốn vậy ta lập biến mới:

$$f = \frac{U(n-2)}{Q} \tag{5.3.17}$$

trong đó: $$U = \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2 = a_1 l_{xy} = \frac{\left[\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})\right]^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$$

$$Q = l_{yy} - U$$

Người ta đã chứng minh được rằng nếu giả thiết H_o đúng thì f có phân bố Fisher với (1, n–2) bậc tự do: $f \in F(1, n-2)$. Từ đó, với xác suất phạm sai lầm loại I (α) cho trước ta có:

$$P(f \geq F_\alpha) = \alpha$$

Và chỉ tiêu kiểm nghiệm là:

Nếu $f \geq F_\alpha$ thì bác bỏ H_o, tức là phương trình hồi qui có thể dùng được.

Nếu $f < F_\alpha$ thì chấp nhận H_o, tức là không thể sử dụng phương trình hồi qui để mô tả quan hệ tuyến tính giữa X và Y.

Ví dụ 5.3.2: Từ hai dãy số liệu $\{x_t, y_t, t=1..62\}$ ta xây dựng được phương trình hồi qui tuyến tính dạng $y = 312.9 - 0.565x$ (a_o=312.9, a_1=–0.565). Với hệ số tương quan r_{xy}=0.1298 ta thấy mối quan hệ tương quan giữa X và Y rất yếu. Vậy phương trình hồi qui tìm được có ý nghĩa sử dụng hay không, nếu lấy mức ý nghĩa α=0.01?

Bài toán được đưa về việc kiểm nghiệm giả thiết H_o: a_1=0. Muốn vậy, trước hết ta tính các đại lượng Q và U, sau đó tính f theo công thức (5.3.17). Kết quả nhận được f=1.767.

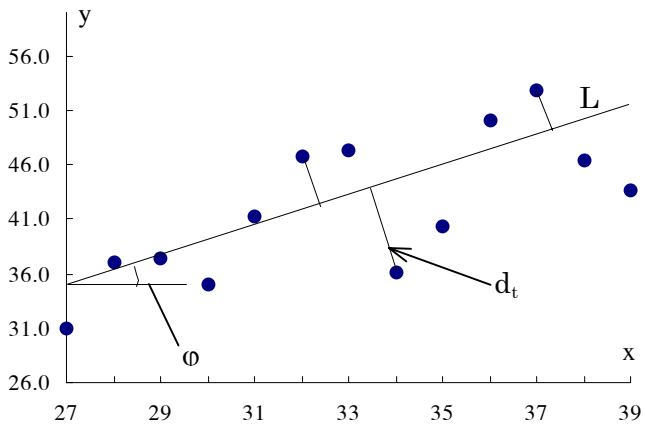
Mặt khác ta có n=62, α=0.01 khi tra bảng hoặc tính trực tiếp ta nhận được $F_\alpha = F_{0.01}(1,60) = 7.08$. So sánh f và F_α ta có: $f=1.767<7.08=F_\alpha$, tức là giả thiết H_o được chấp nhận ($a_1=0$). Vậy ta kết luận phương trình hồi qui tìm được không có ý nghĩa sử dụng.

5.3.6 Hồi qui bình phương trung bình trực giao

Hồi qui chúng ta vừa xét trên đây là hồi qui bình phương trung bình, trong đó nguyên lý bình phương tối thiểu được áp dụng cho tổng bình phương các khoảng cách từ các điểm thực nghiệm đến đường hồi qui theo phương song song với trục toạ độ (Oy) (hình 5.2).

Trong nhiều trường hợp, thay cho việc xét đường hồi qui kiểu đó, người ta xây dựng một đường hồi qui khác dựa trên nguyên tắc: trung bình bình phương các khoảng cách (ngắn nhất) từ các điểm thực nghiệm đến đường thẳng hồi qui là nhỏ nhất. Hay nói cách khác, nếu gọi d_t là khoảng cách từ điểm (x_t,y_t) đến đường thẳng hồi qui L (Hình 5.3) thì L phải thoả mãn điều kiện:

$$M[d^2] \approx \frac{1}{n}\sum_{t=1}^{n} d_t \to \min$$



Hình 5.3 Hồi qui bình phương trung bình trực giao

Khi đó phương trình đường hồi qui sẽ được xác định bởi:

$$(x - m_x)\sin\varphi - (y - m_y)\cos\varphi = 0 \qquad (5.3.18)$$

Với: $m_x = M[X]$, $m_y = M[Y]$, φ là góc giữa trục Ox và đường L, nhận giá trị dương khi quay ngược chiều kim đồng hồ.

Khi $x = m_x$ thì $y = m_y$, và đường L đi qua tâm phân phối chung của X và Y. Đó cũng là điểm cắt nhau của hai đường hồi qui.

Đại lượng $M[d^2]$ được xác định sao cho đạt cực tiểu đối với L có thể được xem như là mômen quán tính và bằng:

$$M[d^2] = M[(x - m_x)\sin\varphi - (y - m_y)\cos\varphi]^2 =$$
$$= \sigma_x^2 \sin^2\varphi + \sigma_y^2 \cos^2\varphi - \mu_{xy}\sin 2\varphi$$

5. 4 Tương quan phi tuyến. Tỷ số tương quan

5.4.1 Tỷ số tương quan tổng thể

Xét hai biến ngẫu nhiên X và Y. Như đã thấy trong mục 5.2, hệ số tương quan ρ_{12} chỉ đo mức độ quan hệ tương quan tuyến tính giữa chúng. Vì vậy nếu chỉ dùng $|\rho_{12}|$ để đánh giá mức độ tương quan nói chung giữa X và Y thì chưa đầy đủ, bởi có thể giữa chúng vẫn có thể tồn tại mối quan hệ tương quan không tuyến tuyến tính nào đó mà ta gọi là tương quan phi tuyến. Do đó, bên cạnh hệ số tương quan ta sẽ xét một đại lượng khác gọi là tỷ số tương quan.

Ta có phương sai của Y:

$$D[Y] = M\left[(Y - M[Y])^2\right] = M\left[((Y - m_y(x)) + (m_y(x) - M[Y]))^2\right] =$$
$$= M\left[(Y - m_y(x))^2\right] + M\left[(m_y(x) - M[Y])^2\right] + 2M\left[(Y - m_y(x))(m_y(x) - M[Y])\right]$$

Trong đó $m_y(x)$ là kỳ vọng có điều kiện của Y với điều kiện X=x. Hạng thứ ba trong vế phải của hệ thức này bằng không, nên:

$$D[Y] = M\left[(Y - m_y(x))^2\right] + M\left[(m_y(x) - M[Y])^2\right] \qquad (5.4.1)$$

Chia hai vế của biểu thức này cho D[Y] ta được:

$$1 = \frac{M\left[(Y - m_y(x))^2\right]}{D[Y]} + \frac{M\left[(m_y(x) - M[Y])^2\right]}{D[Y]}$$

Hay

$$1 - \frac{M\left[(Y - m_y(x))^2\right]}{D[Y]} = \frac{M\left[(m_y(x) - M[Y])^2\right]}{D[Y]}$$

Đặt

$$Q' = M\left[(Y - m_y(x))^2\right],\ U' = M\left[(m_y(x) - M[Y])^2\right],$$

$$\eta^2 = 1 - \frac{M\left[(Y - m_y(x))^2\right]}{D[Y]} = 1 - \frac{Q'}{D[Y]} \qquad (5.4.2)$$

ta có:

$$\eta^2 = \frac{M\left[(m_y(x) - M[Y])^2\right]}{D[Y]} = \frac{U'}{D[Y]} \qquad (5.4.3)$$

Đại lượng η được gọi là tỷ số tương quan giữa X và Y. Vì η≥0 nên thay cho η người ta thường dùng η^2.

Từ (5.4.1), (5.4.2) và (5.4.3) rõ ràng $0 \le \eta^2 \le 1$. Trị số $\eta^2 = 1$ khi và chỉ khi $M\left[(Y - m_y(x))^2\right] = 0$ còn $\eta^2 = 0$ khi $M\left[(m_y(x) - M[Y])^2\right] = 0$. Như vậy η^2 đặc trưng cho mức độ quan hệ phụ thuộc hàm giữa X và Y. Nếu η^2 càng lớn thì sự phụ thuộc hàm giữa hai biến càng chặt chẽ.

Theo (5.42) ta có: $Q' = M\left[(Y - m_y(x))^2\right]$. Nếu xấp xỉ $m_y(x)$ bởi đường hồi qui tuyến tính $m_y(x) \approx y = \alpha + \beta x$ thì $Q' \approx Q'' = M\left[(Y - \alpha - \beta X)^2\right]$.

Vậy nên $Q'' = M\left[(Y - m_y(x) + m_y(x) - \alpha - \beta X)^2\right] =$

$$= M\left[(Y - m_y(x))^2 + (m_y(x) - \alpha - \beta X)^2 + 2(Y - m_y(x))(m_y(x) - \alpha - \beta X)\right]$$

$$= M\left[(Y - m_y(x))^2\right] + M\left[(m_y(x) - \alpha - \beta X)^2\right] + 2M\left[(Y - m_y(x))(m_y(x) - \alpha - \beta X)\right]$$

Hạng thứ ba vế phải bằng không nên:

$$Q'' = M\left[(Y - m_y(x))^2\right] + M\left[(m_y(x) - \alpha - \beta X)^2\right] =$$

$$= Q' + M\left[(m_y(x) - \alpha - \beta X)^2\right] \qquad (5.4.4)$$

Vì hạng thứ nhất vế phải không phụ thuộc vào α, β do đó Q'' đạt cực tiểu khi các hệ số α, β làm cho hạng thứ hai đạt cực tiểu. Tức là:

$$Q'' = Q''_{\min} \quad \text{khi } \alpha = M[Y] - \beta.M[X], \quad \beta = \rho_{12}\frac{\sigma_2}{\sigma_1}. \qquad (5.4.5)$$

trong đó $(\sigma_1)^2 = D[X]$, $(\sigma_2)^2 = D[Y]$.

Từ đó ta có:

$$Q''_{\min} = M\left[(Y - (M[Y] - \rho_{12}\frac{\sigma_2}{\sigma_1}M[X]) - \rho_{12}\frac{\sigma_2}{\sigma_1}[X])^2\right]$$

$$= M\left[((Y - M[Y]) - \rho_{12}\frac{\sigma_2}{\sigma_1}(X - M[X]))^2\right] =$$

$$= M\left[(Y - M[Y])^2\right] + \rho_{12}^2\frac{\sigma_2^2}{\sigma_1^2}M\left[(X - M[X])^2\right] - 2\rho_{12}\frac{\sigma_2}{\sigma_1}M\left[(Y - M[Y])(X - M[X])\right] =$$

$$= \sigma_2^2 + \rho_{12}^2\sigma_2^2 - 2\rho_{12}^2\sigma_2^2 = \sigma_2^2(1 - \rho_{12}^2) \qquad (5.4.7)$$

Kết hợp (5.4.4), (5.4.2) và (5.4.6) ta nhận được:

$$Q' = Q''_{\min} - M\left[(m_y(x) - \alpha - \beta X)^2\right] =$$

$$= \sigma_2^2(1 - \rho_{12}^2) - M\left[(m_y(x) - \alpha - \beta X)^2\right] \qquad (5.4.7)$$

Thay (5.4.7) vào (5.4.2) và để ý rằng D[Y] = σ_2^2 ta được:

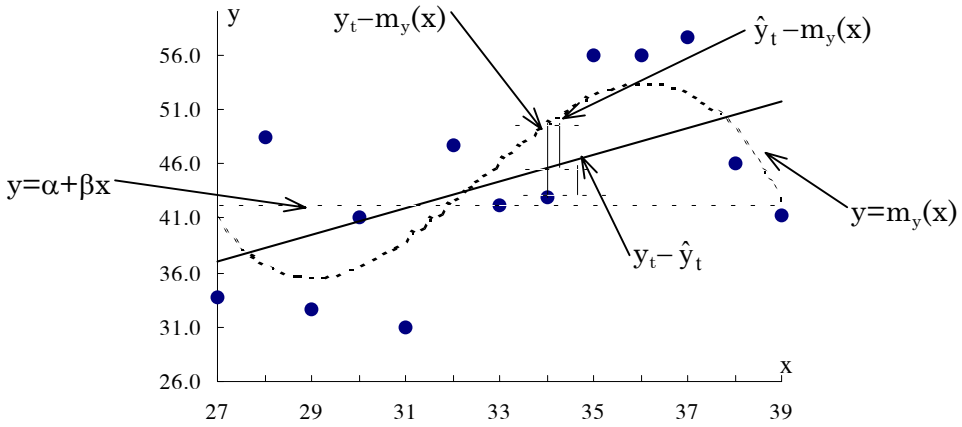
$$\eta^2 = 1 - \frac{\sigma_2^2(1-\rho_{12}^2) - M\left[(m_y(x)-\alpha-\beta X)^2\right]}{\sigma_2^2}$$

Hay

$$\eta^2 = \rho_{12}^2 + \frac{1}{\sigma_2^2} M\left[(m_y(x)-\alpha-\beta X)^2\right] \qquad (5.4.8)$$

Từ đó ta thấy rằng:

1) $\eta^2 = 0$ khi và chỉ khi $m_y(x)$=const, tức là khi đường hồi qui là đường thẳng nằm ngang, do đó $\rho_{12} = \beta = 0$ và hạng thứ hai triệt tiêu.
2) $\eta^2 = 1$ khi tất cả các điểm thực nghiệm đều nằm trên đường $y=m_y(x)$, điều này xảy ra khi giữa X và Y tồn tại quan hệ hàm thực sự.
3) Với những giá trị trung gian của η^2, hệ thức (5.4.2) cho thấy η^2 đặc trưng cho mức độ tập trung của các điểm thực nghiệm xung quanh đường hồi qui.
4) Khi $y = m_y(x)$ là đường thẳng thì hạng thứ hai trong (5.4.8) triệt tiêu, do đó $\eta^2=\rho_{12}^2$.
5) Vì hạng thứ hai của (5.4.8) không âm nên trong trường hợp $y = m_y(x)$ là một đường bất kỳ nhưng không phải là đường thẳng thì η^2 luôn luôn lớn hơn ρ_{12}^2 một lượng; lượng đó đặc trưng cho độ lệch của đường $y = m_y(x)$ so với đường thẳng $y = \alpha + \beta x$ (hình 5.4).



Hình 5.4 Đường hồi qui I và đường hồi qui II

5.4.2 Tỷ số tương quan mẫu

Cũng như hệ số tương quan, để phân biệt với tỷ số tương quan tổng thể η^2 ta sẽ ký hiệu tỷ số tương quan mẫu là ζ^2. Mặc dù ký hiệu này không phổ biến, nhưng dù sao nó sẽ giúp chúng ta đỡ nhầm lẫn trong khi trình bày.

Xét hai biến ngẫu nhiên X và Y với *n* cặp trị số quan sát $\{(x_t, y_t), t=1..n\}$. Từ các hệ thức (5.3.12), (5.3.13), (5.4.2) và (5.4.3), tỷ số tương quan mẫu giữa X và Y được xác định bởi:

$$\zeta^2 = \frac{U}{l_{yy}} \tag{5.4.9}$$

trong đó $U = \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2$, với $\hat{y} = m_y(x)$.

Tỷ số tương quan mẫu là một ước lượng của tỷ số tương quan tổng thể. Từ hệ thức (5.4.8) ta cũng có thể nhận được một dạng khác của tỷ số tương quan mẫu bằng cách thay các đại lượng trong đó bởi các ước lượng tương ứng của chúng:

$$\zeta^2 = r_{xy}^2 + \frac{1}{s_y^2}\frac{1}{n}\sum_{t=1}^{n}\left[(m_y(x_t) - a_o - a_1 x_t)^2\right] \tag{5.4.10}$$

Hệ thức (5.4.10) tự bản thân nó đã phản ánh đầy đủ mọi nội dung vừa trình bày trong mục trước. Chúng ta sẽ không nhắc lại ở đây nữa. Sau đây chúng ta sẽ đưa ra cách tính tỷ số tương quan mẫu.

Có thể nhận thấy rằng các biểu thức (5.4.9) và (5.4.10) đều chứa đại lượng $m_y(x)$ là kỳ vọng có điều kiện của Y theo X mà trên thực tế chúng ta không biết. Do vậy để tính ζ^2 từ số liệu thực nghiệm ta sẽ tiến hành theo các bước sau đây [3]:

1) Căn cứ vào tập mẫu ta tiến hành phân nhóm các chuỗi. Giả sử chuỗi $\{x_t\}$ được phân thành K nhóm, chuỗi $\{y_t\}$ được phân thành L nhóm, ta đi xác định các tần số thực nghiệm n_{ij}:

n_{ij} = Số trường hợp thoả mãn điều kiện *x_t thuộc nhóm i và y_t thuộc nhóm j*,

(i=1..K, j=1..L)

X	Y				Tổng
	Nhóm 1 $y_{(1)}$	Nhóm 2 $y_{(2)}$	...	Nhóm L $y_{(L)}$	
Nhóm 1: $x_{(1)}$	n_{11}	n_{12}	...	n_{1L}	$n_{1.}$
Nhóm 2: $x_{(2)}$	n_{21}	n_{22}	...	n_{2L}	$n_{2.}$
...	...	...	...	...	...
Nhóm K: $x_{(K)}$	n_{K1}	n_{K1}	...	n_{KL}	$n_{K.}$
Tổng	$n_{.1}$	$n_{.2}$		$n_{.L}$	n

trong đó: $n_{i.} = \sum_{j=1}^{L} n_{ij}$, $n_{.j} = \sum_{i=1}^{K} n_{ij}$, $x_{(i)}$, $y_{(j)}$ là trị số giữa của nhóm *i* (đối với x) hoặc nhóm j (đối với y).

2) Tính các đại lượng:

$$A_i = \sum_{j=1}^{L} n_{ij} y_{(j)}, \quad B = \sum_{j=1}^{L} n_{.j} y_{(j)}, \quad C = \sum_{j=1}^{L} n_{.j} (y_{(j)})^2$$

3) Tính các đại lượng:

$$\frac{A_i^2}{n_{i.}} = \frac{1}{n_{i.}} \left(\sum_{j=1}^{L} n_{ij} y_{(j)} \right)^2, \quad D = \sum_{i=1}^{K} \frac{1}{n_{i.}} \left(\sum_{j=1}^{L} n_{ij} y_{(j)} \right)^2 = \sum_{i=1}^{K} \frac{A_1^2}{n_{i.}}$$

4) Tính tỷ số tương quan theo công thức:

$$\zeta^2 = \frac{\sum_{i=1}^{K} \frac{1}{n_{i.}} \left(\sum_{j=1}^{L} n_{ij} y_{(j)} \right)^2 - \frac{1}{n} \left(\sum_{j=1}^{L} n_{.j} y_{(j)} \right)^2}{\sum_{j=1}^{L} n_{.j} \left(y_{(j)} \right)^2 - \frac{1}{n} \left(\sum_{j=1}^{L} n_{.j} y_{(j)} \right)^2} = \frac{D - \frac{B^2}{n}}{C - \frac{B^2}{n}}$$

5.4.3 Hồi qui phi tuyến một biến

Xét hồi qui II giữa hai biến khí hậu X và Y. Khi quan hệ giữa chúng không phải là quan hệ tuyến tính ta có thể xây dựng phương trình hồi qui phi tuyến. Muốn vậy trước hết cần xem xét sự phụ thuộc tương quan giữa chúng để từ đó xác lập dạng hàm có thể xấp xỉ.

Sau khi đã quyết định chọn lớp hàm phụ thuộc biểu thị mối liên hệ giữa Y và X, nguyên tắc chung để xây dựng đường hồi qui là tuyến tính hoá các thành phần phi tuyến bằng cách đặt biến mới rồi đưa về trường hợp hồi qui tuyến tính đã xét ở mục 5.3.

Trong khí tượng, khí hậu chúng ta thường gặp các dạng hồi qui phi tuyến sau đây:

1) Dạng hyperbol

$$y = a_o + \frac{a_1}{x}$$

Phép tuyến tính hoá được thực hiện bằng cách đặt $\frac{1}{x} = x'$ và phương trình đã cho được đưa về dạng mới: $y = a_o + a_1 x'$.

2) Dạng luỹ thừa

$$y = a_o x^{a_1}$$

Bằng cách lôgarit hoá hai vế ta được: $\log y = \log a_o + a_1 \log x$ và đặt $\log y = y'$, $\log x = x'$, $\log a_o = a'_o$, khi đó phương trình được đưa về dạng:

$$y' = a'_o + a_1 x'$$

3) Dạng hàm mũ

$$y = a_o e^{a_1 x}$$

Lấy lôgarit tự nhiên hai vế rồi đặt lny = y', $\ln a_o = a'_o$ ta được:

$$y' = a'_o + a_1 x$$

4) Dạng loga

$$y = a_o + a_1 \log x$$

Đặt logx = x' ta được dạng mới:

$$y = a_o + a_1 x'$$

5) Dạng đa thức bậc cao

$$y = a_o + a_1 x + a_2 x^2 + a_3 x^3 + ... + a_m x^m$$

Điều cần lưu lý ở đây là để xác định các hệ số a_i (i=0..m) thì bậc của đa thức phải nhỏ hơn số cặp quan trắc (x_t,y_t), t=1..n, ít nhất một đơn vị, tức là m≤n−1. Trong trường hợp đó, các hệ số a_i sẽ được tìm thông qua việc giải hệ phương trình đại số tuyến tính sau đây:

$$\begin{cases} a_o n + a_1 \sum x + a_2 \sum x^2 + ... + a_m \sum x^m = \sum y \\ a_0 \sum x + a_1 \sum x^2 + a_2 \sum x^3 + ... + a_m \sum x^{m+1} = \sum xy \\ a_0 \sum x^2 + a_1 \sum x^3 + a_2 \sum x^4 + ... + a_m \sum x^{m+2} = \sum x^2 y \\ ... \\ a_0 \sum x^m + a_1 \sum x^{m+1} + a_2 \sum x^{m+2} + ... + a_m \sum x^{2m} = \sum x^m y \end{cases} \qquad (5.4.11)$$

trong đó ẩn số là các a_i (i=0..m).

Khi m=1 hệ (5.4.11) tương đương với hệ (5.3.8) và phương trình hồi qui trở về dạng tuyến tính quen thuộc.

Như vậy, để tìm m+1 hệ số a_i (i=0..m) ta cần phải giải hệ m+1 phương trình đại số tuyến tính, trong đó bậc cao nhất của x bằng 2m. Bậc đa thức càng cao thì đường hồi qui càng đi sát các điểm thực nghiệm. Khi bậc đa thức bằng n−1 thì đường cong đa thức sẽ mô tả chính xác các điểm thực nghiệm. Trong trường hợp này phương trình hồi qui sẽ không có tác dụng mô tả qui luật phụ thuộc của Y vào X, vì lúc đó tất cả các nhiễu loạn của thực nghiệm chứa đựng trong số liệu quan trắc ban đầu được bảo toàn. Mặt khác khi bậc đa thức càng cao thì việc tính toán càng trở nên phức tạp và sai số mắc phải do tính toán càng lớn. Chính vì vậy, trước khi tính toán chúng ta cần tiến hành chọn bậc tối ưu của đa thức xấp xỉ.

Điều cần lưu ý là trong quan hệ hồi qui phi tuyến nói chung hiệu quả của phương trình hồi qui giảm đi. Mặt khác, việc xác định dạng hàm hồi qui thông thường phải tiến hành bằng thực nghiệm. Bởi vậy để chọn được dạng hàm thích hợp, không còn cách nào khác là phải thử đi thử lại nhiều lần.

5.5 Hồi qui tuyến tính nhiều biến

5.5.1 Mặt hồi qui

Trong mục 5.3.1 chúng ta đã xét hồi qui I giữa Y lên X mà phương trình hồi qui được mô tả bởi hệ thức (5.3.2) và (5.3.3), trong đó ta xem Y là biến phụ thuộc còn X là biến độc lập. Phương trình hồi qui I giữa Y lên X là một đường cong $y = m_y(x)$. Ta cũng có thể mở rộng khái niệm này cho trường hợp nhiều biến.

Ta xét *m* biến ngẫu nhiên $X_1, X_2,...,X_m$ với phân bố đồng thời $f(x_1, x_2,..., x_m)$. Khi đó hồi qui I giữa X_1 lên $X_2, X_3,..., X_m$ được định nghĩa bởi hệ thức:

$$x_1 = m_1(x_2,..., x_m) = M[X_1/X_2=x_2,...,X_m=x_m]=$$

$$= \int_{-\infty}^{+\infty} x_1 f(x_1 / x_2,...,x_m)dx_1 \qquad (5.5.1)$$

trong đó $f(x_1/x_2,..., x_m)$ là mật độ có điều kiện của X_1 khi $X_2=x_2,...,X_m=x_m$.

Phương trình (5.5.1) là quĩ tích của những điểm $(m_1, x_2,...,x_m)$ với mọi giá trị có thể có của $x_2,...,x_m$ và người ta gọi nó là mặt hồi qui I:

$$x_1 = m_1(x_2,..., x_m) \qquad (5.5.2)$$

Trong trường hợp tổng quát (5.5.2) là một mặt bất kỳ và trên thực tế ta khó có thể biết được dạng thức giải tích của nó. Bởi vậy trong tính toán thực hành người ta thường xấp xỉ nó vào một lớp hàm *f* nào đó đã biết:

$$m_1(x_2,..., x_m) \approx \hat{x}_1 = f(x_2,..., x_m) \qquad (5.5.3)$$

và được gọi là hồi qui II của X_1 lên $X_2,...,X_m$. Nếu hàm f thuộc lớp hàm tuyến tính thì mặt hồi qui được gọi là một siêu phẳng. Khi đó ta có phương trình hồi qui tuyến tính nhiều biến:

$$\hat{x}_1 = \beta_1 + \beta_2 x_2 +...+ \beta_m x_m \qquad (5.5.4)$$

Hay

$$X_1 = \beta_1 + \beta_2 X_2 +...+ \beta_m X_m \qquad (5.5.5)$$

Ở đây β_i (i=1..m) là các hệ số hồi qui mà chúng được xác định sao cho xấp xỉ (5.5.3) là tốt nhất theo nghĩa bình phương tối thiểu. Như vậy vế phải biểu thức (5.5.4) là

ước lượng tuyến tính tốt nhất của X_1 theo các $X_2,...,X_m$ với nghĩa cực tiểu hoá đại lượng $M\left[X_1 - \left(\beta_1 + \sum_{i=2}^{m}\beta_i X_i\right)\right]^2$:

$$M\left[X_1 - \left(\beta_1 + \sum_{i=2}^{m}\beta_i X_i\right)\right]^2 \to \min$$

5.5.2 Xây dựng phương trình hồi qui tuyến tính nhiều biến thực nghiệm

Cho m biến ngẫu nhiên $X_1,...,X_m$ với n bộ trị số quan sát $\{x_{t1}, x_{t2},...,x_{tm}\}$, (t=1..n). Xét hồi qui II là hồi qui tuyến tính giữa X_1 lên các $X_2,...,X_m$. Ta cần xây dựng phương trình:

$$\hat{x}_1 = \beta_1 + \beta_2 x_2 + ... + \beta_m x_m$$

biểu thị mối phụ thuộc tuyến tính của biến phụ thuộc x_1 vào m−1 biến độc lập $x_2,...,x_m$ trên cơ sở tập số liệu thực nghiệm $\{x_{t1}, x_{t2},...,x_{tm}\}$. Tức là ta đi xác định các ước lượng của các β_i (i=1..m). Ký hiệu a_i (i=1..m) là các ước lượng đó, bài toán được đưa về việc xác định các giá trị a_i (i=1..m) sao cho phương trình:

$$\hat{x}_1 = a_1 + a_2 x_2 + ... + a_m x_m = a_1 + \sum_{i=2}^{m} a_i x_i \qquad (5.5.6)$$

là ước lượng tuyến tính tốt nhất của xấp xỉ (5.5.3) trên cơ sở tập số liệu có được. Điều đó tương đương với việc cần phải xác định các a_i (i=1..m) sao cho:

$$R(a_1,...,a_m) = \sum_{t=1}^{n}(x_{t1} - \hat{x}_{t1})^2 = \sum_{t=1}^{n}\left(x_{t1} - \left(a_1 + \sum_{i=2}^{m} a_i x_i\right)\right)^2 \longrightarrow \min$$

trong đó $\hat{x}_{t1} = a_1 + \sum_{i=2}^{m} a_i x_{ti}$ là các giá trị hồi qui của x_1 lên x_{ti}, (i=2..m), $R(a_1,...,a_m)$ là hàm của m biến $a_1,...,a_m$.

Người ta đã chứng minh được rằng điều kiện cần và đủ để $R(a_1,...,a_m)$ đạt cực tiểu là các đạo hàm riêng của R theo các a_i đồng thời phải triệt tiêu:

$$\frac{\partial R(a_1,...,a_m)}{\partial a_i} = 0, \text{ (i=1..m)}$$

Điều đó cũng có nghĩa là các a_i phải thoả mãn hệ:

$$\begin{cases} \sum_{t=1}^{n}\left(x_{t1} - a_1 - \sum_{i=2}^{m} a_i x_{ti}\right) = 0 \\ \sum_{t=1}^{n}\left(x_{t1} - a_1 - \sum_{i=2}^{m} a_i x_{ti}\right)x_{tk} = 0, (k = 2..m) \end{cases} \qquad (5.5.7)$$

Đây là hệ m phương trình đại số tuyến tính với m ẩn là $a_1,...,a_m$. Đẳng thức thứ nhất trong (5.5.7) cho ta:

$$\frac{1}{n}\sum_{t=1}^{n}\left(x_{t1}-a_1-\sum_{i=2}^{m}a_i x_{ti}\right)=\frac{1}{n}\sum_{t=1}^{n}x_{t1}-a_1-\sum_{i=2}^{m}\frac{1}{n}\sum_{t=1}^{n}a_i x_{ti}=0$$

Suy ra $$\overline{x_1}-a_1-\sum_{i=2}^{m}a_i\overline{x_i}=0$$

hay $$a_1=\overline{x_1}-\sum_{i=2}^{m}a_i\overline{x_i} \qquad (5.5.8)$$

Thay (5.5.8) vào đẳng thức thứ hai trong (5.5.7) ta được:

$$\sum_{t=1}^{n}\left(x_{t1}-\left(\overline{x_1}-\sum_{i=2}^{m}a_i\overline{x_i}\right)-\sum_{i=2}^{m}a_i x_{ti}\right)x_{tk}=0$$

Suy ra $$\frac{1}{n}\sum_{t=1}^{n}\left(x_{t1}x_{tk}-\overline{x_1}x_{tk}+\sum_{i=2}^{m}a_i\overline{x_i}x_{tk}-\sum_{i=2}^{m}a_i x_{ti}x_{tk}\right)=0$$

Từ đó: $$\frac{1}{n}\sum_{t=1}^{n}x_{t1}x_{tk}-\frac{1}{n}\sum_{t=1}^{n}\overline{x_1}x_{tk}+\frac{1}{n}\sum_{t=1}^{n}\sum_{i=2}^{m}a_i\overline{x_i}x_{tk}-\frac{1}{n}\sum_{t=1}^{n}\sum_{i=2}^{m}a_i x_{ti}x_{tk}=0$$

hay $$(\overline{x_1x_k}-\overline{x_1}\,\overline{x_k})-\sum_{i=2}^{m}a_i(\overline{x_ix_k}-\overline{x_i}\,\overline{x_k})=0 \qquad (5.5.9)$$

Vì $(\overline{x_1x_k}-\overline{x_1}\,\overline{x_k})=R_{1k}$ là mômen tương quan giữa X_1 và X_k (k=2..m)

$(\overline{x_ix_k}-\overline{x_i}\,\overline{x_k})=R_{ik}=R_{ki}$ là mômen tương quan giữa X_i và X_k (i,k=2..m)

Do đó (5.5.9) có thể được viết lại dưới dạng:

$$\sum_{i=2}^{m}R_{ki}a_i=R_{1k}, \quad (k=2..m) \qquad (5.5.10)$$

Đây là hệ m−1 phương trình với m−1 ẩn số a_i (i=2..m). Giải hệ này ta sẽ được các a_i. Có nhiều phương pháp giải hệ (5.5.10), như phương pháp gần đúng, phương pháp ma trận nghịch đảo, phương pháp khử Gauss,... Ở đây ta đưa ra nghiệm của (5.5.10) theo phương pháp Cramer:

$$a_i=\frac{D_i}{D}, \ (i=2..m) \qquad (5.5.11)$$

trong đó D là định thức của ma trận (R^c_{ki}), (i, k=2..m), là ma trận nhận được từ ma trận (R_{ki}) bằng cách bỏ đi hàng 1, cột 1, còn D_i là định thức của ma trận nhận được từ ma trận (R^c_{ki}) bằng cách thay cột thứ i bởi vectơ (R_{1k}), (k=2..m):

$$D = \begin{vmatrix} R_{22} & R_{23} & \dots & R_{2m} \\ R_{32} & R_{33} & \dots & R_{3m} \\ \dots & \dots & \dots & \dots \\ R_{m2} & R_{m3} & \dots & R_{mm} \end{vmatrix},$$

$$D_i = \begin{vmatrix} R_{22} & \dots & R_{2,i-1} & R_{12} & R_{2,i+1} & \dots & R_{2m} \\ R_{32} & \dots & R_{3,i-1} & R_{13} & R_{3,i+1} & \dots & R_{3m} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ R_{m2} & \dots & R_{m,i-1} & R_{1m} & R_{m,i+1} & \dots & R_{mm} \end{vmatrix}$$

Từ các hệ thức (5.5.8) và (5.5.11) ta hoàn toàn xác định được các hệ số a_i (i=1..m). Thay chúng vào (5.5.6) ta được phương trình hồi qui phải tìm.

Thay cho tập số liệu ban đầu $\{x_{t1}, x_{t2},...,x_{tm}\}$, (t=1..n), nếu ta xét tập số liệu chuẩn hoá qua phép biến đổi:

$$x'_{ti} = \frac{x_{ti} - \overline{x_i}}{s_i}, \text{ (i=1..m)} \qquad (5.5.12)$$

trong đó $s_i = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(x_{ti} - \overline{x_i}\right)^2}$, thì phương trình (5.5.6) sẽ có dạng:

$$\hat{x}'_1 = a'_1 + \sum_{i=2}^{m} a'_i x'_i \qquad (5.5.13)$$

Và các hệ số a'_i, (i=1..m) sẽ được xác định bởi $a'_1 = \overline{x'_1} - \sum_{i=2}^{m} a'_i \overline{x'_i}$. Nhưng, do $\overline{x'_i} = 0$ (i=1..m)

nên: $$a'_1 = 0\,,\ \sum_{i=2}^{m} R_{ki} a'_i = R_{1k}\,,\ \text{(k=2..m)}, \qquad (5.5.14)$$

với: $R_{1k} = (\overline{x'_1 x'_k} - \overline{x'_1}\,\overline{x'_k}) = r_{1k}$, (k=2..m), $R_{ik} = (\overline{x'_i x'_k} - \overline{x'_i}\,\overline{x'_k}) = r_{ik} = r_{ki}$, (i,k=2..m) trong đó r_{ik} là hệ số tương quan giữa X_i và X_k. Do đó:

$$a'_i = \frac{D'_i}{D'} \qquad (5.5.15)$$

Ở đây D' là định thức của ma trận (r^c_{ki}), (i,k=2..m), là ma trận nhận được từ ma trận (r_{ki}) bằng cách bỏ đi hàng 1 cột 1, còn D'_i là định thức của ma trận được tạo nên từ ma trận (r^c_{ki}) bằng cách thay cột thứ i bởi vectơ (r_{1k}), (k=2..m):

$$D' = \begin{vmatrix} r_{22} & r_{23} & \dots & r_{2m} \\ r_{32} & r_{33} & \dots & r_{3m} \\ \dots & \dots & \dots & \dots \\ r_{m2} & r_{m3} & \dots & r_{mm} \end{vmatrix},$$

$$D'_i = \begin{vmatrix} r_{22} & \dots & r_{2,i-1} & r_{12} & r_{2,i+1} & \dots & r_{2m} \\ r_{32} & \dots & r_{3,i-1} & r_{13} & r_{3,i+1} & \dots & r_{3m} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{m2} & \dots & r_{m,i-1} & r_{1m} & r_{m,i+1} & \dots & r_{mm} \end{vmatrix}$$

Từ (5.5.14) và (5.5.15) ta có phương trình hồi qui cho các biến chuẩn hoá (5.5.12) là:

$$\hat{x}'_1 = \sum_{i=2}^{m} a'_i x'_i \qquad (5.5.16)$$

5.5.3 Thặng dư và phương sai thặng dư

Bây giờ ta hãy ký hiệu:

Δ là định thức ma trận tương quan (R_{ki}), (i,k=1..m)

M_{ki} là định thức của ma trận con của ma trận (R_{ki}) sau khi đã bỏ đi hàng thứ k cột thứ i

Δ_{ki} là phần phụ đại số của phần tử R_{ki} của ma trận (R_{ki}), $\Delta_{ki} = (-1)^{k+i}$. M_{ki} ($R'_{ki.1i}$) là ma trận (R_{ki}) sau khi đã đổi vị trí của cột thứ nhất cho cột thứ i.

D_i là định thức của ma trận con của ma trận ($R'_{ki.1i}$) sau khi đã bỏ đi hàng thứ nhất, cột thứ nhất.

$$\Delta = \begin{vmatrix} R_{11} & R_{12} & \dots & R_{1m} \\ R_{21} & R_{22} & \dots & R_{2m} \\ \dots & \dots & \dots & \dots \\ R_{m1} & R_{m2} & \dots & R_{mm} \end{vmatrix},$$

$$M_{ki} = \begin{vmatrix} R_{11} & \dots & R_{1,i-1} & R_{1,i+1} & \dots & R_{1m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ R_{k-1,1} & \dots & R_{k-1,i-1} & R_{k-1,i+1} & \dots & R_{k-1,m} \\ R_{k+1,1} & \dots & R_{k+1,i-1} & R_{k+1,i+1} & \dots & R_{k+1,m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ R_{m1} & \dots & R_{m,i-1} & R_{m,i+1} & \dots & R_{mm} \end{vmatrix}$$

$$(R'_{ki.1i}) = \begin{pmatrix} R_{1i} & R_{12} & \dots & R_{1,i-1} & R_{11} & R_{1,i+1} & \dots & R_{1m} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ R_{mi} & R_{m2} & \dots & R_{m,i-1} & R_{m1} & R_{m,i+1} & \dots & R_{mm} \end{pmatrix}$$

$$D_i = \begin{vmatrix} R_{22} & \dots & R_{2,i-1} & R_{21} & R_{2,i+1} & \dots & R_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ R_{m2} & \dots & R_{m,i-1} & R_{m1} & R_{m,i+1} & \dots & R_{mm} \end{vmatrix}$$

Ta thấy:

$$\Delta_{11} = M_{11} = D,\ D_i = (-1)^i.M_{1i},\ \Delta_{1i} = (-1)^{1+i}.M_{1i} \text{ nên } D_i = -\Delta_{1i} \qquad (5.5.17)$$

Từ (5.5.11) và (5.5.17) có thể nhận cách biểu diễn khác của các hệ số hồi qui a_i:

$$a_i = -\frac{\Delta_{1i}}{\Delta_{11}} \qquad (5.5.18)$$

Trở lại công thức (5.5.6) và giả thiết rằng các $\overline{x_i} = 0$, (i=1..m). Điều này hoàn toàn có thể thực hiện được, bởi vì nếu không như vậy thì thay cho các biến x_i ta sẽ xét các biến qui tâm. Khi đó hệ số $a_1 = 0$ và phương trình (5.5.6) trở thành:

$$\hat{x}_1 = a_2x_2 + ... + a_mx_m = \sum_{i=2}^{m} a_ix_i \qquad (5.5.19)$$

Xét đại lượng $$q = x_1 - \hat{x}_1 = x_1 - \sum_{i=2}^{m} a_ix_i \qquad (5.5.20)$$

được gọi là thặng dư của x_1 đối với các $x_2,..., x_m$, khi đó $q_t = x_{t1} - \sum_{i=2}^{m} a_ix_{ti}$ là độ lệch của các trị số quan trắc thực nghiệm khỏi giá trị hồi qui.

Thay (5.5.18) vào (5.5.20) ta được:

$$q = x_1 - \sum_{i=2}^{m}(-\frac{\Delta_{1i}}{\Delta_{11}}x_i) = \frac{\Delta_{11}}{\Delta_{11}}x_1 + \sum_{i=2}^{m}\frac{\Delta_{1i}}{\Delta_{11}}x_i = \sum_{i=1}^{m}\frac{\Delta_{1i}}{\Delta_{11}}x_i \qquad (5.5.21)$$

Ta thấy $$\bar{q} = \overline{\sum_{i=1}^{m}\frac{\Delta_{1i}}{\Delta_{11}}x_i} = \frac{1}{\Delta_{11}}\overline{\sum_{i=1}^{m}\Delta_{1i}x_i} = \frac{1}{\Delta_{11}}\sum_{i=1}^{m}\Delta_{1i}\overline{x_i} = 0$$

Do đó: $$\overline{x_kq} = \frac{1}{\Delta_{11}}\sum_{i=1}^{m}\Delta_{1i}\overline{x_kx_i} = \frac{1}{\Delta_{11}}\sum_{i=1}^{m}\Delta_{1i}R_{ki} = \begin{cases}\dfrac{\Delta}{\Delta_{11}} & \text{khi } k=1\\ 0 & \text{khi } k\neq 1\end{cases}$$

(vì khi k=1 thì $\sum_{i=1}^{m}\Delta_{1i}R_{1i} = |R_{ki}| = \Delta$)

Như vậy thặng dư của x_1 đối với $x_2,..., x_m$ không tương quan với các $x_2,...,x_m$. Từ đó ta có hệ thức:

$$\overline{x_1q} = \frac{\Delta}{\Delta_{11}} \qquad (5.5.22)$$

Đại lượng $s_q^2 = \overline{q^2}$ được gọi là phương sai thặng dư. Từ (5.5.21) ta có:

$$s_q^2 = \overline{\frac{1}{\Delta_{11}^2}\sum_{i=1}^{m}\Delta_{1i}x_i\sum_{k=1}^{m}\Delta_{1k}x_k} = \frac{1}{\Delta_{11}^2}\sum_{i=1}^{m}\sum_{k=1}^{m}\Delta_{1i}\Delta_{1k}\overline{x_i x_k} = \frac{1}{\Delta_{11}^2}\sum_{i=1}^{m}\Delta_{1i}\left(\sum_{k=1}^{m}\Delta_{1k}R_{ki}\right)$$

Vì $$\sum_{k=1}^{m}\Delta_{1k}R_{ki} = \begin{cases}\Delta & \text{khi } i=1\\ 0 & \text{khi } i\neq 1\end{cases}$$

Do đó: $$s_q^2 = \frac{1}{\Delta_{11}^2}\Delta_{11}\Delta = \frac{\Delta}{\Delta_{11}} = \overline{x_1 q}$$

Suy ra: $$s_q^2 = \overline{x_1 q} = \frac{\Delta}{\Delta_{11}} \qquad (5.5.23)$$

Vì q là đại lượng đặc trưng cho độ lệch của các giá trị x_1 khỏi mặt hồi qui nên có thể xem phương sai thặng dư là thước đo mức độ xấp xỉ tốt nhất thu được khi biểu diễn x_1 bằng tổ hợp tuyến tính của các x_2,..., x_m.

Khi m = 2 ta có:

$$(R_{ki}) = \begin{pmatrix} R_{11} & R_{12}\\ R_{21} & R_{22}\end{pmatrix} \text{ nên: } \Delta = R_{11}R_{22} - R_{12}R_{21}, \quad \Delta_{11} = R_{22}.$$

Mặt khác $R_{11}=s_1^2$, $R_{22}=s_2^2$, $R_{12}= R_{21}=r_{12}s_1s_2$, nên:

$$s_q^2 = \frac{s_1^2s_2^2 - s_1^2s_2^2r_{12}^2}{s_2^2} = s_1^2(1-r_{12}^2)$$

Để ý rằng ký hiệu x_1 tương đương với biến y, nên $s_1\equiv s_y$, $r_{12}=r_{xy}$. Vậy nên: $s_q^2 = s_1^2(1-r_{12}^2) = \frac{Q}{n-2} = s^2$ chính là bình phương của chuẩn sai thặng dư mà ta đã đề cập đến trong mục (5.3.4).

Khi m>2, từ (5.3.12) ta có:

$$\sum_{t=1}^{n}\left(x_{t1}-\overline{x_1}\right)^2 = \sum_{t=1}^{n}\left(x_{t1}-\hat{x}_{t1}+\hat{x}_{t1}-\overline{x_1}\right)^2 = \sum_{t=1}^{n}\left(x_{t1}-\hat{x}_{t1}\right)^2 + \sum_{t=1}^{n}\left(\hat{x}_{t1}-\overline{x_1}\right)^2$$

Đặt $$\sum_{t=1}^{n}\left(x_{t1}-\overline{x_1}\right)^2 = l_{x1x1},$$

$$\sum_{t=1}^{n}\left(x_{t1}-\hat{x}_{t1}\right)^2 = Q,$$

$$\sum_{t=1}^{n}\left(\hat{x}_{t1}-\overline{x_1}\right)^2 = U,$$

Ta được:

$$l_{x1x1} = Q + U$$

trong đó l_{x1x1}, Q và U có ý nghĩa tương tự như đã nêu trong mục (5.3.3). Từ đây ta thấy $Q = q^2$ nên Q chính là tổng bình phương các thặng dư.

Ta có: Số bậc tự do của l_{x1x1} là n−1

Số bậc tự do của U là m−1 (phụ thuộc vào m−1 biến)

Do đó: Số bậc tự do của Q là n−m, tức số nhân tố tạo nên sai số hồi qui bằng n−m.

Bởi vậy, thay cho phương sai thặng dư s_q^2, trong thực hành tính toán người ta dùng đại lượng:

$$s = \sqrt{\frac{Q}{n-m}} \tag{5.5.24}$$

làm thước đo sai số hồi qui. Nó được gọi là chuẩn sai thặng dư.

5.5.4 Tương quan riêng

Các hệ số tương quan r_{ki} được dùng để đo mức độ tương quan tuyến tính giữa hai biến X_k và X_i và được gọi là hệ số tương quan toàn phần. Nếu ta xét mối quan hệ giữa X_k và X_i đồng thời với việc xét m−2 biến còn lại thì, ở một mức độ nào đó, có thể xem độ biến thiên của X_k và X_i là do sự biến thiên của các biến khác gây nên. Để loại trừ ảnh hưởng của sự biến thiên của các biến khác đến sự biến thiên của X_k và X_i ta sẽ xét mối quan hệ tương quan riêng giữa chúng.

Giả sử có m biến $X_1,...,X_m$. Xét hai thặng dư:

$$q_1 = x_1 - \sum_{i=3}^{m} a_i x_i , \qquad q_2 = x_2 - \sum_{i=3}^{m} b_i x_i \tag{5.5.25}$$

Từ mục (5.5.3) ta thấy q_1 và q_2 biểu thị phần còn lại của X_1 và X_2 sau khi trừ các biến đó cho ước lượng tuyến tính tốt nhất của chúng theo các biến $X_3,...,X_m$. Vậy có thể xem hệ số tương quan giữa hai thặng dư này là độ đo mức độ tương quan giữa X_1 và X_2 sau khi đã khử đi phần của độ biến thiên do ảnh hưởng của các biến $X_3,...,X_m$ gây nên. Ta sẽ gọi đó là hệ số tương quan riêng của X_1 và X_2 đối với $X_3,...,X_m$ và ký hiệu là $r_{12.34...m}$. Ta có:

$$r_{12.34...m} = \frac{M[q_1 q_2]}{\sqrt{M[q_1^2]M[q_2^2]}} \tag{5.5.26}$$

Bây giờ ta hãy xây dựng công thức tính $r_{12.34...m}$. Muốn vậy, ta đưa vào một số ký hiệu qui ước ngoài những ký hiệu đã có ở mục (5.5.3):

(M_{jj}) là ma trận con của ma trận (R_{ki}) sau khi đã bỏ đi hàng thứ j, cột thứ j

$\Delta_{jj.ki}$ là phần phụ đại số của phần tử $M_{jj.ki}$ (hàng k cột i của M_{jj}).

Lưu ý rằng các chỉ số i, j, k được lấy tương ứng theo các biến X_i, X_j, X_k.

Khi đó, theo (5.5.23) ta có:

$$\sigma_{q1}^2 = M\left[q_1^2\right] = \frac{\Delta_{22}}{\Delta_{22.11}} = \frac{\Delta_{22}}{\Delta_{11.22}},\ \sigma_{q2}^2 = \overline{q_2^2} = \frac{\Delta_{11}}{\Delta_{11.22}} \tag{5.5.27}$$

$$\overline{q_1 q_2} = \overline{\left(x_1 - \sum_{i=3}^{m} a_i x_i\right) q_2} = \overline{x_1 q_2} - \overline{\left(\sum_{i=3}^{m} a_i x_i\right) q_2} = \overline{x_1 q_2} - \sum_{i=3}^{m} a_i \overline{x_i q_2}$$

Hạng thứ hai vế phải bằng 0 khi i≠2. Do đó:

$$\overline{q_1 q_2} = \overline{x_1 q_2} = \overline{x_1 \left(x_2 - \sum_{i=3}^{m} b_i x_i\right)}$$

Vì $b_i = -\frac{\Delta_{11.2j}}{\Delta_{11.22}}$, (i=3..m) nên suy ra:

$$\overline{q_1 q_2} = \overline{x_1 \sum_{i=2}^{m} \frac{1}{\Delta_{11.22}} \Delta_{11.2i} x_i} = \frac{1}{\Delta_{11.22}} \sum_{i=2}^{m} \Delta_{11.2i} \overline{x_1 x_i} =$$

$$= \frac{1}{\Delta_{11.22}} \sum_{i=2}^{m} \Delta_{11.2i} R_{1i} = -\frac{\Delta_{12}}{\Delta_{11.22}} \tag{5.5.28}$$

Thay (5.5.27) và (5.5.28) vào (5.5.26) ta được:

$$r_{12.34\ldots m} = \frac{-\frac{\Delta_{12}}{\Delta_{11.22}}}{\sqrt{\frac{\Delta_{11}}{\Delta_{11.22}} \frac{\Delta_{22}}{\Delta_{11.22}}}} = -\frac{\Delta_{12}}{\sqrt{\Delta_{11}\Delta_{22}}} \tag{5.5.29}$$

Một cách tổng quát, hệ số tương quan riêng của hai biến X_k, X_i đối với m−2 biến còn lại được tính bởi công thức:

$$r_{ki.12\ldots k-1,k+1\ldots i-1,i+1\ldots m} = -\frac{\Delta_{ki}}{\sqrt{\Delta_{kk}\Delta_{ii}}} \tag{5.5.30}$$

Trường hợp riêng, khi m = 3 ta có:

$$(R_{ki}) = \begin{pmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{pmatrix}$$

$$\Delta_{12} = (-1)^{(1+2)} \begin{vmatrix} R_{21} & R_{23} \\ R_{31} & R_{33} \end{vmatrix} = -(R_{21}R_{33} - R_{23}R_{31}) = -s_1 s_2\ s_3^2\,(r_{12} - r_{23}r_{31})$$

$$\Delta_{11} = (-1)^{(1+1)} \begin{vmatrix} R_{22} & R_{23} \\ R_{32} & R_{33} \end{vmatrix} = R_{22}R_{33} - R_{23}R_{32} = s_2^2 s_3^2 (1 - r_{23}^2)$$

$$\Delta_{22} = (-1)^{(2+2)} \begin{vmatrix} R_{11} & R_{13} \\ R_{31} & R_{33} \end{vmatrix} = R_{11}R_{33} - R_{13}R_{31} = s_1^2 s_3^2 (1 - r_{13}^2)$$

Do đó:

$$r_{12.3} = -\frac{-(R_{21}R_{33} - R_{23}R_{31}}{\sqrt{(R_{22}R_{33} - R_{23}R_{32})(R_{11}R_{33} - R_{31}R_{13})}} = \frac{r_{12-} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}} \qquad (5.5.31)$$

Khi các biến không tương quan với nhau thì hệ số tương quan riêng bằng 0, còn nếu giữa các biến có tương quan với nhau thì nói chung hệ số tương quan riêng khác hệ số tương quan toàn phần: $r_{12.3...m} \neq r_{12}$ và thậm chí chúng còn ngược dấu nhau.

Công thức (5.5.31) cho thấy có thể tính hệ số tương quan riêng từ các hệ số tương quan toàn phần. Trong trường hợp m > 3 ta có công thức truy hồi sau:

$$r_{12.34...m} = \frac{r_{12.34...m-1} - r_{1m.34...m-1}r_{2m.34...m-1}}{\sqrt{(1-r_{1m.34...m-1}^2)(1-r_{2m.34...m-1}^2)}} \qquad (5.5.32)$$

Hay viết dưới dạng ma trận:

$$r_{12.34...m} = -\frac{P_{12}}{\sqrt{P_{11}P_{22}}} \qquad (5.5.33)$$

trong đó ký hiệu P_{ki} có ý nghĩa như Δ_{ki} nhưng ma trận xuất phát là ma trận tương quan chuẩn hoá (r_{ki}).

Cũng như hệ số tương quan toàn phần, ta có thể kiểm nghiệm độ rõ rệt của các hệ số tương quan riêng bằng kiểm nghiệm t, với $t = \frac{r}{\sqrt{1-r^2}/\sqrt{n-m}}$, trong đó t có phân bố Student với (n–m) bậc tự do.

5.5.5 Tương quan bội

Bây giờ ta hãy xét mối quan hệ giữa X_1 và $\hat{X}_1$ là hồi qui tuyến tính của X_1 lên các biến $X_2,...,X_m$. Ta có phương trình hồi qui được xác lập trên cơ sở tập số liệu quan trắc $\{x_{ti}, t=1..n, i=1..m\}$:

$$\hat{x}_1 = a_2x_2 + ... + a_mx_m = \sum_{i=2}^{m} a_ix_i$$

Ở đây ta đã giả thiết các $\overline{x_i} = 0$, (i=1..m). Người ta đã chứng minh được rằng trong tất cả các tổ hợp tuyến tính của X_1 theo các X_i (i=2..m) thì $\hat{X}_1$ có tương quan tốt nhất với X_1. Như vậy có thể xem hệ số tương quan giữa X_1 và $\hat{X}_1$ là đặc trưng tương quan giữa một bên là biến phụ thuộc X_1 và một bên là tập hợp m–1 biến độc lập $X_2,..,X_m$. Ta sẽ gọi đó là hệ số tương quan bội hay hệ số tương quan tập hợp và ký hiệu là $r_{1.23...m}$:

$$r_{1.23...m} = \frac{M\left[X_1\hat{X}_1\right]}{\sqrt{M\left[X_1^2\right]M\left[\hat{X}_1^2\right]}}$$

Hay:
$$r_{1.23...m} = \frac{\overline{x_1\hat{x}_1}}{\sqrt{\overline{x_1^2}\,\overline{\hat{x}_1^2}}} \tag{5.5.34}$$

Ta có: $\overline{x_1\hat{x}_1} = \overline{x_1(x_1 - q_1)} = \overline{x_1^2 - x_1q_1} = \overline{x_1^2} - \overline{x_1q_1}$

trong đó q_1 là thặng dư của x_1 đối với $x_2,..., x_m$. Theo (5.5.23) thì

$$\overline{x_1q_1} = \overline{q_1^2}$$

Do đó:
$$\overline{x_1\hat{x}_1} = \overline{x_1^2} - \overline{q_1^2} = s_1^2 - \frac{\Delta}{\Delta_{11}} = R_{11} - \frac{\Delta}{\Delta_{11}} \tag{5.5.35}$$

$$\overline{\hat{x}_1^2} = \overline{(x_1 - q_1)^2} = \overline{x_1^2 + q_1^2 - 2x_1q_1} = \overline{x_1^2} + \overline{q_1^2} - 2\,\overline{x_1q_1} =$$

$$= s_1^2 + \frac{\Delta}{\Delta_{11}} - 2\frac{\Delta}{\Delta_{11}} = s_1^2 - \frac{\Delta}{\Delta_{11}} = R_{11} - \frac{\Delta}{\Delta_{11}} \tag{5.5.36}$$

Thay (5.5.35) và (5.5.36) vào (5.5.34) ta được:

$$r_{1.23...m} = \frac{R_{11} - \frac{\Delta}{\Delta_{11}}}{\sqrt{R_{11}(R_{11} - \frac{\Delta}{\Delta_{11}})}} = \frac{R_{11} - \frac{\Delta}{\Delta_{11}}}{R_{11}\sqrt{(1 - \frac{\Delta}{\Delta_{11}R_{11}})}} = \sqrt{1 - \frac{\Delta}{R_{11}\Delta_{11}}} \tag{5.5.37}$$

Hoặc dưới dạng khác:

$$r_{1.23...m} = \sqrt{1 - \frac{P}{P_{11}}} \tag{5.5.38}$$

trong đó P là định thức của ma trận tương quan chuẩn hoá, P_{11} là phần phụ đại số của phần tử r_{11} của ma trận tương quan chuẩn hoá.

Theo [4] thì với ma trận đối xứng và xác định dương (R_{ki}) ta có:

$$0 < \Delta \le R_{11}\Delta_{11}$$

nên hệ thức trong dấu căn của (5.5.37) không âm. Vậy $r_{1.23...m} \ge 0$.

Với $r_{1.23...m}$ =1 có thể nói hầu như chắc chắn biến X_1 bằng một tổ hợp tuyến tính nào đó của các biến $X_2,...,X_m$. Khi đó toàn bộ các điểm thực nghiệm đều nằm trên siêu phẳng hồi qui. Hệ số tương quan bội $r_{1.23...m}$ bằng 0 khi và chỉ khi tất cả các r_{1i} (i=2..m) đều bằng 0, tức là khi biến X_1 không tương quan với bất kỳ một biến $X_2,...,X_m$ nào cả.

Ví dụ 5.5 Từ những số liệu quan trắc về sản lượng lúa vụ thu (x_1 – kg/ha) và nhiệt độ không khí trung bình mùa đông năm trước (x_2 – °C), nhiệt độ không khí trung bình mùa hè (thời gian gieo trồng) (x_3 – °C), tổng lượng mưa trong suốt thời gian gieo trồng (x_4 – mm) của khu vực A trong 30 năm, ta cần nghiên cứu mối liên hệ giữa x_1 và các x_2, x_3, x_4 để từ đó tiến hành xây dựng phương trình hồi qui dự báo sản lượng lúa. Muốn vậy ta tính các hệ số tương quan cặp, hệ số tương quan riêng, hệ số tương quan bội và các hệ số hồi qui.

Kết quả tính toán cho ta:

1) Ma trận tương quan chuẩn hoá:

$$(r_{ki}) = \begin{pmatrix} 1 & 0.59107 & 0.41082 & 0.46120 \\ 0.59107 & 1 & 0.67028 & 0.31838 \\ 0.41082 & 0.67028 & 1 & 0.10720 \\ 0.46120 & 0.31838 & 0.10720 & 1 \end{pmatrix}$$

trong đó các r_{ki} là các hệ số tương quan giữa x_k và x_i. Bằng kiểm nghiệm độ rõ rệt của hệ số tương quan, ta thấy tất cả các r_{ki}, trừ r_{24} và r_{34}, đều lớn rõ rệt với mức ý nghĩa 5%. Điều đó cho phép ta bác bỏ giả thiết về sự không phụ thuộc của x_1 vào các x_2, x_3, x_4. Hơn nữa, từ ma trận tương quan ta có r_{12} lớn hơn r_{13} một cách đáng kể. Như vậy, hình như nhiệt độ mùa đông năm trước có ảnh hưởng đến sản lượng lúa hơn là nhiệt độ mùa hè.

2) Các hệ số tương quan riêng:

$r_{12.3}$=0.4666 $r_{13.2}$=0.0244 $r_{14.2}$=0.3570

$r_{12.4}$=0.5281 $r_{13.4}$=0.4096 $r_{14.3}$=0.4602

Nếu lấy mức ý nghĩa α=1% để kiểm nghiệm độ rõ rệt của các hệ số tương quan riêng trên đây thì chỉ có $r_{12.4}$ là đạt tiêu chuẩn. Các hệ số $r_{12.3}$ và $r_{14.3}$ cũng rất gần với tiêu chuẩn này. So sánh r_{13} (0.41082) với các $r_{13.2}$ và $r_{13.4}$ ta thấy nếu bỏ qua ảnh hưởng của nhiệt độ mùa đông năm trước (x_2) sẽ làm giảm tương quan giữa sản lượng x_1 và nhiệt độ mùa hè x_3 xuống đến mức không đáng kể ($r_{13.2}$=0.0244); còn nếu bỏ qua tác động của lượng mưa (x_4) thì hầu như không ảnh hưởng tới tương quan này. Cũng bằng cách so sánh tương tự ta thấy tương quan giữa sản lượng và nhiệt độ mùa đông sẽ không giảm nhiều lắm khi bỏ qua ảnh hưởng của nhiệt độ mùa hè và lượng mưa. Từ đó suy ra rằng, nhiệt độ mùa đông năm trước và lượng mưa là hai nhân tố quan trọng thực sự.

Trên đây ta chỉ tính hệ số tương quan riêng khi xét đến một nhân tố ảnh hưởng. Ta cũng có thể làm tương tự cho các hệ số tương quan riêng $r_{12.34}$,...

3) Đối với các hệ số tương quan bội:

$r_{1.23}$=0.5914 $r_{1.24}$=0.6575 $r_{1.34}$=0.5872 $r_{1.234}$=0.6606

Việc so sánh r_{12} và $r_{1.23}$ cho ta nhận xét rằng, khi đã biết x_2 thì việc biết thêm x_3 hầu như không có ý nghĩa. Hay nói cách khác, nhân tố x_3 hầu như không cung cấp thêm một lượng thông tin mới nào cho x_1 khi đã có nhân tố x_2. Tương tự như vậy, giá trị của $r_{1.24}$ cũng không nhỏ hơn $r_{1.234}$ bao nhiêu.

5.5.6 Đánh giá chất lượng của phương trình hồi qui tuyến tính nhiều biến

Phương trình (5.5.6) là ước lượng tuyến tính của biến X_1 theo các biến $X_2,...,X_m$ trong đó các hệ số a_i (i=1..m) được tìm trên cơ sở tập số liệu ban đầu. Nó là thông tin duy nhất phản ánh mối quan hệ tuyến tính giữa X_1 và các biến $X_2,...,X_m$. Tuy nhiên để sử dụng nó cho mục đích xác định giá trị của biến phụ thuộc X_1 thì cần thiết phải đánh giá được mức độ tin cậy đến đâu. Điều đó cũng có nghĩa là ta cần phải trả lời câu hỏi mối liên hệ phụ thuộc giữa X_1 và các biến $X_2,...,X_m$ là có ý nghĩa hay không. Nói cách khác, ta cần kiểm tra giả thiết hệ số tương quan bội $r_{1.23..m}=0$ (mà thực chất giả thiết này tương đương với giả thiết $a_2=a_3=...=a_m=0$).

Để kiểm nghiệm giả thiết $r_{1.23..m} = 0$ ta lập biến mới:

$$f = \frac{U/(m-1)}{Q/(n-m)} \tag{5.5.39}$$

trong đó $U = \sum_{t=1}^{n}\left(\hat{x}_{t1} - \overline{x_1}\right)^2$ là tổng bình phương các biến sai hồi qui,

$Q = \sum_{t=1}^{n}\left(x_{t1} - \hat{x}_{t1}\right)^2$ là tổng bình phương các biến sai thặng dư

$\hat{x}_{t1} = a_1 + \sum_{i=2}^{m} a_i x_{ti}$ là giá trị hồi qui của x_1 theo các $x_2,...,x_m$

(m–1) và (n–m) theo thứ tự là bậc tự do của U và Q.

Biến f trong (5.5.39) có phân bố Fisher với (m–1,n–m) bậc tự do. Vậy, ứng với xác suất phạm sai lầm loại I (α) ta hoàn toàn xác định được giá trị F_α, và chỉ tiêu kiểm nghiệm sẽ là:

Nếu $f \geq F_\alpha$ thì bác bỏ giả thiết và đưa ra kết luận sự phụ thuộc tuyến tính giữa X_1 và các $X_2,...,X_m$ là có ý nghĩa.

Nếu $f < F_\alpha$ thì chấp nhận giả thiết, tức là chấp nhận sự không tồn tại quan hệ tuyến tính giữa X_1 và các $X_2,...,X_m$.

Trong tính toán thực hành, thay cho (5.5.39) người ta thường tính f theo công thức:

$$f = \frac{r^2_{1.23...m}}{1-r^2_{1.23...m}}\frac{n-m}{m-1} \tag{5.5.40}$$

5.6 Tương quan và hồi qui phi tuyến nhiều biến

Sự phụ thuộc tương quan giữa biến phụ thuộc X_1 và m−1 biến độc lập $X_2,...,X_m$ không phải khi nào cũng tuyến tính. Và do đó trong nhiều trường hợp quan hệ (5.5.6) không phản ánh đúng thực chất của mối quan hệ giữa X_1 và các biến $X_2,...,X_m$. Bởi vì giữa X_1 và một số biến nào đó trong các X_i (i=2..m) có thể tồn tại quan hệ phi tuyến mà việc xấp xỉ nó bằng một hàm tuyến tính là không chấp nhận được. Vậy vấn đề đặt ra là làm thế nào để xây dựng được một quan hệ hàm:

$$x_1 = m_1(x_2,...,x_m) \qquad (*)$$

có thể biểu diễn được sự phụ thuộc giữa X_1 và các $X_2,...,X_m$? Giải quyết vấn đề này là nhiệm vụ của bài toán hồi qui phi tuyến nhiều biến. Sau đây ta sẽ xét một số phương pháp nghiên cứu sự phụ thuộc phi tuyến nhiều biến.

Cũng cần nhấn mạnh rằng, khi số biến độc lập lớn hơn hoặc bằng 2 nói chung ta không hy vọng tìm được một hàm dạng (*) theo đúng nghĩa là kỳ vọng có điều kiện của X_1 với điều kiện $X_2=x_2,...X_m=x_m$. Do đó, sau đây ta sẽ xét một số phương pháp khả dĩ nghiên cứu sự phụ thuộc phi tuyến nhiều biến thường được ứng dụng trong khí tượng, khí hậu.

5.6.1 Liên kết các mối quan hệ riêng rẽ

Phương pháp này được tiến hành theo các bước:

1) Nghiên cứu mối tương quan riêng biệt (cặp) giữa X_1 và từng biến X_i để tìm ra qui luật phụ thuộc của chúng. Kết quả của bước này cho ta m−1 quan hệ hàm:

$$x_1 = f_2(x_2)$$
$$x_1 = f_3(x_3)$$
$$...$$
$$x_1 = f_m(x_m)$$

Ở đây $f_i(x_i)$ là hồi qui II của x_1 lên x_i: $m_1(x_i) \approx f_i(x_i)$

2) Xác lập qui tắc liên kết các $f_i(x_i)$, (i=2..m), sao cho qui tắc đó có thể biểu diễn được sự phụ thuộc của X_1 vào các X_i. Tức là cần tìm được một toán tử tác dụng L nào đó sao cho:

$$x_1 = m_1(x_2,...,x_m) \approx f(x_2,...,x_m)$$

trong đó $f(x_2,...,x_m) = L\{ f_1(x_1), f_2(x_2),..., f_m(x_m)\}$

Trong trường hợp L là toán tử tuyến tính ta có thể biểu diễn:

$$f(x_2,...,x_m) = \sum_{i=2}^{m} \alpha_i f_i(x_i) \qquad (5.6.1)$$

3) Xây dựng phương trình hồi qui $\hat{x}_1 = f(x_2, x_3, \ldots, x_m)$, tức là tìm các hệ số hồi qui thực nghiệm, theo nguyên lý bình phương tối thiểu:

$$R = \sum_{t=1}^{n} (x_{t1} - f(x_{t2}, x_{t3}, \ldots, x_{tm})^2 \longrightarrow \min$$

Bước này có thể được thực hiện bằng cách tuyến tính hoá các thành phần phi tuyến thông qua việc đặt biến mới để đưa hàm phi tuyến $f(x_2,\ldots,x_m)$ về dạng tuyến tính. Người ta gọi cách làm này là tuyến tính bên trong nhưng phi tuyến bên ngoài.

Để dễ hình dung ta hãy lấy ví dụ sau làm minh hoạ. Giả sử cần xây dựng phương trình hồi qui giữa Y với X_1 và X_2:

$$Y = f(X_1, X_2)$$

Các bước cần tiến hành là:

1) Lập ma trận số liệu ban đầu $\{Y_t, X_{t1}, X_{t2}, t=1..n\}$ và xây dựng các quan hệ $Y=f_1(X_1)$, $Y = f_2(X_2)$ từ tập số liệu thực nghiệm. Có thể tiến hành bước này bằng phương pháp đồ thị kết hợp tính toán liên tiếp. Để đơn giản, ta giả thiết rằng các quan hệ này có dạng sau:

$$f_1(x_1) = a\,x_1^2 + bx_1 + c$$

$$f_2(x_2) = \alpha e^{-x2}$$

2) Chọn qui tắc liên kết là tổ hợp tuyến tính:

$$f(x_1,x_2) = A\,f_1(x_1) + B\,f_2(x_2)$$

Khi đó, khai triển ra ta được:

$$f(x_1,x_2) = Aa\,x_1^2 + Abx_1 + Ac + B\alpha e^{-x2}$$

Hay $$f(x_1,x_2) = \beta_o + \beta_1\,x_1^2 + \beta_2 x_1 + \beta_3\,e^{-x2}$$

3) Xây dựng phương trình hồi qui: Đây là bước xác định các hệ sô β_i (i=0..3) trong phương trình trên đây. Muốn vậy, ta đặt biến mới:

$$u = x_1^2,\ v = x_1 \text{ và } w = e^{-x2}$$

rồi đưa phương trình về dạng $f(x_1,x_2) = g(u,v,w) = \beta_o + \beta_1 u + \beta_2 v + \beta_3 w$. Rõ ràng đây là một phương trình tuyến tính đối với u, v, w. Sử dụng các công thức đã trình bày trong các mục trước ta dễ dàng tìm được:

$$y = \beta_o + \beta_1\,x_1^2 + \beta_2 x_1 + \beta_3\,e^{-x2} \qquad (5.6.2)$$

5.6.2 Dạng phụ thuộc bậc hai (dạng toàn phương)

Trong nhiều trường hợp, thay cho việc nghiên cứu tỷ mỉ mà thậm chi đôi khi kém hiệu quả như phương pháp trên đây, người ta giả thiết sự phụ thuộc của X_1 vào các X_i được biểu diễn dưới dạng đa thức bậc hai:

$$x_1 = a_1 + \sum_{i=2}^{m} a_i x_i + \sum_{i=2}^{m} b_i x_i^2 + \sum_{i,j=2,i<j}^{m} c_{ij} x_i x_j$$

Như vậy, vế phải có tất cả M =(m−1)+(m−1)+$\frac{(m-1)(m-2)}{2}$ hạng tử chứa các x_i.

Muốn xây dựng được phương trình hồi qui ta phải tìm được tất cả (m−1)+(m−1) + $\frac{(m-1)(m-2)}{2}$+1 = (M+1) hệ số a_i, b_i, c_{ij}. Tuy nhiên cũng có thể nhận thấy rằng, so với số biến độc lập ban đầu (m−1) thì số thành phần trong phương trình hồi qui đã tăng lên một cách đáng kể. Chẳng hạn, nếu m=4, thì phương trình hồi qui sẽ chứa tất cả 3+3+3=9 hạng tử vế phải chứa các biến và ta phải tìm được 10 hệ số hồi qui.

Để tìm các hệ số hồi qui thông thường ta đặt biến mới:

$$z_{i-1} = x_i,\ i=2..m,$$

$$z_{m-1+i-1} = x_i^2,\ i=2..m,$$

$$z_{2m-2+k} = x_i x_j,\ i,j=2..m,\ i<j$$

và chuyển phương trình ban đầu về dạng: $x_1 = \alpha_o + \sum_{i=1}^{M} \alpha_i z_i$

Rõ ràng đây là dạng phương trình hồi qui tuyến tính nhiều biến quen thuộc mà việc tìm hệ số α_i đã được trình bày trong mục 5.5.2.

5.6.3 Dạng luỹ thừa

Dạng phụ thuộc luỹ thừa được biểu diễn bởi:

$$x_1 = a_o \prod_{i=2}^{m} x_i^{a_i} \qquad (5.6.3)$$

Để tìm các hệ số hồi qui trong trường hợp này ta lấy lôgarit hai vế:

$$\log x_1 = \log a_o + \sum_{i=2}^{m} a_i \log x_i$$

và đặt biến mới: $z_1=\log x_1$, $b_o=\log a_o$, $z_i=\log x_i$, i=2..m, ta được phương trình hồi qui tuyến tính với biến phụ thuộc là z_1, các biến độc lập là z_i:

$$z_1 = b_o + \sum_{i=2}^{m} a_i z_i$$

Sau khi xác định được các hệ số b_o, a_i, thay vào (5.6.3) ta được phương trình phải tìm.

5.7 Hồi qui từng bước

5.7.1 Đặt vấn đề

Trong nghiên cứu khí tượng thuỷ văn nói chung, ta thường gặp bài toán hồi qui nhiều biến, tức là nghiên cứu mối phụ thuộc giữa một bên là biến phụ thuộc X_1 với một bên là một loạt các biến độc lập $X_2,...,X_m$. Tuy nhiên các yếu tố khí tượng thuỷ văn thường có những tác động qua lại và ảnh hưởng lẫn nhau, bởi vậy khái niệm biến độc lập chỉ mang nghĩa hình thức. Điều đó có nghĩa là giữa các biến độc lập thường có mối quan hệ tương quan nào đó. Mặt khác, giữa các biến độc lập và biến phụ thuộc cũng tồn tại những mối quan hệ ràng buộc. Do đó có thể xảy ra tình trạng các biến độc lập được chọn đều tương quan tốt với nhau và tương quan tốt cả với biến phụ thuộc, ý nghĩa cung cấp thông tin của các biến độc lập vì thế mà giảm đi. Trong nhiều trường hợp, điều đó dẫn đến hậu quả là mặc dù phương trình hồi qui khá phức tạp do sự có mặt của nhiều biến độc lập nhưng độ chính xác của nó lại kém hơn do sai số quan trắc, do dao động ngẫu nhiên, sai số tính toán,... mang lại.

Vậy vấn đề đặt ra là cần phải xác định xem những biến nào trong các biến độc lập có ảnh hưởng đáng kể đến biến phụ thuộc, có nhất thiết tất cả các biến được chọn đều phải có mặt trong phương trình hồi qui hay chỉ là một bộ phận nào đó. Đó là mục tiêu của bài toán hồi qui từng bước.

5.7.2 Các bước thực hiện

Xét hồi qui tuyến tính giữa biến phụ thuộc X_1 và m−1 biến độc lập $X_2,..., X_m$. Để giải bài toán này bằng phương pháp hồi qui từng bước ta tiến hành theo thứ tự sau:

Bước 1: Tính các hệ số tương quan toàn phần r_{1i} giữa X_1 và các X_i (i=2..m) và chọn trong chúng hệ số có giá trị tuyệt đối lớn nhất. Giả sử:

$$|r_{12}| = \max_{2\le i\le m}\{|r_{1i}|\} \qquad (5.7.1)$$

khi đó, biến X_2 là biến có tác động chính lên X_1 và ta xây dựng phương trình hồi qui:

$$x_1^{(1)} = a_1^{(1)} + a_2^{(1)} x_2 . \qquad (5.7.2)$$

Tương ứng với phương trình (5.7.2) ta tính được chuẩn sai thặng dư $s^{(1)}$.

Bước 2: Tính các hệ số tương quan riêng $r_{1i.2}$ (i=3..m) và chọn hệ số có giá trị lớn nhất trong chúng. Giả sử:

$$|r_{13.2}| = \max_{3 \le i \le m} \{|r_{1i.2}|\} \qquad (5.7.3)$$

Khi đó ta chọn tiếp biến X_3 và xây dựng phương trình hồi qui:

$$x_1^{(2)} = a_1^{(2)} + a_2^{(2)} x_2 + a_3^{(2)} x_3 . \qquad (5.7.4)$$

Tương ứng với nó ta cũng tính được chuẩn sai thặng dư $s^{(2)}$. Kết thúc bước này ta có phương trình hồi qui hai biến (5.7.4) mà độ chính xác của nó được đánh giá bởi $s^{(2)}$.

Bước 3: Bây giờ ta đem so sánh giá trị chuẩn sai thặng dư $s^{(2)}$ với $s^{(1)}$:

Nếu $$\left| \frac{s^{(2)} - s^{(1)}}{s^{(2)}} \right| < \varepsilon \qquad (5.7.5)$$

thì biến X_3 sẽ bị bỏ qua và quá trình hồi qui sẽ kết thúc. Ở đây, ε là một số dương tuỳ ý ta đưa vào để đánh giá xem nếu tăng thêm biến cho phương trình hồi qui thì độ chính xác có tăng lên đáng kể hay không. Hay nói cách khác, khi thêm vào phương trình hồi qui một biến mới thì đóng góp thông tin của nó làm giảm sai số được bao nhiêu phần trăm; nếu mức độ giảm không vượt quá ε thì có thể bỏ qua nó.

Nếu $\left| \frac{s^{(2)} - s^{(1)}}{s^{(2)}} \right| \ge \varepsilon$ thì biến X_3 sẽ được chọn. Ta lại tính tiếp các hệ số tương quan riêng $r_{1i.23}$ với i=4..m và qui trình được lặp lại bắt đầu như bước 2. Quá trình cứ tiếp tục như vậy cho đến khi hết tất cả các biến hoặc tự kết thúc như đã trình bày.

Như vậy ở bước thứ k ta có chuẩn sai thặng dư $s^{(k)}$ tương ứng với phương trình hồi qui:

$$x_1^{(k)} = a_1^{(k)} + a_2^{(k)} x_2 + ... + a_{k+1}^{(k)} x_{k+1}$$

và điều kiện lựa chọn:

$$\left| \frac{s^{(k)} - s^{(k-1)}}{s^{(k)}} \right| < \varepsilon$$

Trong thực tế, nhiều khi thay cho việc dùng hệ số tương quan riêng để lựa chọn biến ở bước 2 người ta có thể dùng hệ số tương quan bội. Chẳng hạn, thay cho (5.7.3) ta có thể dùng:

$$r_{1.23} = \max_{3\le i\le m}\{r_{1.2i}\} \qquad (5.7.6)$$

Tuy vậy, việc lựa chọn biến theo (5.7.3) hoặc (5.7.6) có cho cùng một kết quả hay không cho đến nay vẫn chưa được xác minh chính xác. Thông thường ta tiến hành theo cả hai cách và so sánh chúng. Kết quả giống nhau hay khác nhau đều cho ta những thông tin cần thiết để phân tích về mối quan hệ giữa biến phụ thuộc và các biến độc lập.

Một điều đáng lưu ý nữa là chất lượng của các phương trình hồi qui sau mỗi một lần tính. Nói chung bao giờ chúng ta cũng phải tiến hành kiểm nghiệm độ tin cậy của kết quả nhận được trong quá trình tính toán.

CHƯƠNG 6
CHỈNH LÝ SỐ LIỆU KHÍ HẬU

6.1 Đặt vấn đề

Như đã biết, số liệu là bộ phận quan trọng nhất mà từ đó ta có thể tiến hành tính toán, thống kê, thực hiện những vấn đề trong nghiên cứu khí hậu bằng phương pháp thống kê. Ngoài việc lựa chọn đúng phương pháp nghiên cứu, chất lượng số liệu là yếu tố quyết định đến sự chính xác của kết quả.

Nói đến chất lượng số liệu trước hết cần xem xét đến độ chính xác của chúng. Có nhiều nguyên nhân gây nên sự thiếu chính xác, hay nói đúng hơn là sai số, trong bản thân các chuỗi được sử dụng để tính toán, như sai sót do quan trắc, nhầm lẫn trong quá trình xử lý ban đầu hoặc khi tiến hành lấy mẫu, do tác động ngẫu nhiên của những nhân tố bên ngoài,...Bởi vậy, bài toán đặt ra ở đây là cần loại bỏ sai số chứa đựng trong chuỗi số liệu ban đầu trước khi đưa vào xử lý, tính toán.

Mặt khác, trong thực tế, nhất là ở nước ta, vì nhiều lý do khác nhau, chuỗi số liệu khí tượng thuỷ văn nói chung, số liệu khí hậu nói riêng, ít khi đảm bảo tính liên tục. Điều đó gây không ít khó khăn cho việc triển khai nghiên cứu ứng dụng trong một loạt bài toán. Chẳng hạn, do điều kiện chiến tranh, chuỗi số liệu của trạm A bị khuyết đi một số tháng của các năm nào đó; hoặc do điều kiện lưu trữ không tốt, số liệu của trạm B bị phai mờ hoặc mất lẻ tẻ một số điểm,... Vấn đề đặt ra là bằng cách nào đó hãy phục hồi lại những số liệu khuyết thiếu để chuỗi trở thành liên tục.

Một vấn đề khác cũng được đặt ra khi tiến hành xử lý số liệu. Đó là sự duy trì, thành lập các trạm phụ thuộc vào nhiều điều kiện khách quan cũng như chủ quan mà kết quả là chuỗi thời gian quan trắc của các trạm dài ngắn khác nhau. Điều này làm nảy sinh hai vấn đề: Khi độ dài của chuỗi ngắn thì số liệu của trạm không mang đầy đủ tính tiêu biểu; và khi độ dài các chuỗi khác nhau thì số liệu của toàn mạng lưới trạm sẽ không bảo đảm tính so sánh. Vậy vấn đề cần giải quyết ở đây là bổ khuyết số liệu cho những trạm có độ dài chuỗi ngắn, tạo cơ sở để tính toán các đặc trưng thống kê trên những chuỗi này.

6.2 Khử sai số trong số liệu ban đầu

Thực tế khẳng định rằng, trong các chuỗi số liệu quan trắc luôn luôn chứa đựng những sai số tiềm ẩn nào đó và người ta chia những sai số này ra làm 3 loại: Sai số thô, sai số hệ thống và sai số ngẫu nhiên.

Sai số thô sinh ra chủ yếu bởi những thao tác nhầm lẫn, sơ suất trong quá trình đo đạc hoặc lấy mẫu. Chẳng hạn, trong qui ước ban đầu, số liệu nhiệt độ được lấy chính xác đến phần mười độ và không ghi dấu phẩy thập phân, nhưng khi tiến hành thu thập số liệu từ các báo biểu quan trắc, do thói quen người ta ghi lẫn lộn một vài số nào đó có dấu phẩy thập phân (tách phần nguyên và phần mười độ – ví dụ, trị số 240 bị ghi sai thành 24). Như vậy, vô tình những giá trị này đã bị giảm đi mười lần so với trị số thực. Trong nhiều trường hợp những giá trị có chứa sai số kiểu này rất khó phát hiện do chúng bị ẩn dấu trên nền chuỗi số liệu. Ví dụ, cũng với kiểu xảy ra sai sót nói trên nhưng không phải đối với nhiệt độ mà là lượng mưa, thì hầu như không thể chỉ ra được số liệu nghi ngờ.

Sai số hệ thống gây nên bởi rất nhiều nguyên nhân khác nhau, mỗi nguyên nhân mang một dáng vẻ. Đây là loại sai số rất khó phát hiện nếu không có sự khảo sát tỷ mỷ. Ví dụ, khi xem xét các báo biểu quan trắc người ta nhận thấy rằng do hiệu đính dụng cụ không đúng nên số liệu nhiệt độ đã bị lệch đi một lượng nào đó, hoặc do thói quen, khi đọc nhiệt biểu quan trắc viên thường đọc giá trị nhiệt độ trên nhiệt kế thấp hơn so với qui định chung. v.v.

Sai số ngẫu nhiên là sai số còn lại sau khi đã khử bỏ sai số thô và sai số hệ thống. Sai số ngẫu nhiên gây nên bởi một lượng vô cùng lớn các nguyên nhân mà ảnh hưởng của mỗi một trong chúng bé đến mức ta không thể phân định nổi mức đóng góp của từng nguyên nhân, chúng luôn luôn tồn tại trong mọi chuỗi số liệu quan trắc.

Trong ba loại sai số nêu trên, sai số ngẫu nhiên không thể khử bỏ được trong từng thành phần của chuỗi quan trắc. Tuy vậy, bằng các phương pháp của lý thuyết xác suất ta có thể tính được ảnh hưởng của chúng đến việc xác định các ước lượng thống kê. Đối với sai số hệ thống, nếu phát hiện được và biết nguyên nhân gây nên sai số ta hoàn toàn có thể loại trừ chúng. Song, nói chung việc phát hiện sai số hệ thống đòi hỏi phải khảo sát hết sức công phu. Sau đây ta sẽ đề cập đến phương pháp phát hiện và loại bỏ sai số thô.

1) *Cách phát hiện sai số thô*

Giả sử ta có chuỗi quan trắc $\{x_t\}=\{x_1,x_2,...,x_n\}$ của đại lượng khí hậu X. Khi đó sai số thô (nếu có) thường ẩn chứa trong những giá trị nằm ở các vị trí đầu hoặc cuối chuỗi trình tự $\{x_{(t)}\}=\{x_{(1)},...,x_{(n)}\}$, ($x_{(1)}<...<x_{(n)}$). Do dó muốn phát hiện chúng, ta

sắp xếp chuỗi ban đầu thành chuỗi trình tự và xem xét các giá trị đầu và cuối của chuỗi này. Các giá trị bị nghi ngờ có chứa sai số thường là quá lớn hoặc quá bé so với trị số nền của chuỗi. Khái niệm quá lớn hoặc quá bé được đánh giá định lượng theo qui tắc "ba xinma"" (3σ): $x_{(t)} >> \bar{x} + 3s$ hoặc $x_{(t)} << \bar{x} - 3s$, trong đó $\bar{x}$ và s là trung bình độ lệch chuẩn của X – ước lượng của μ và σ. Như vậy, trước hết ta tính giá trị trung bình $\bar{x}$ và độ lệch chuẩn s của chuỗi. Sau đó xác định những giá trị $x_{(t)}$ quá lớn hoặc quá bé và đánh dấu chúng, xem đó là những giá trị nghi ngờ có chứa sai số thô, hay gọi một cách ngắn gọn hơn là giá trị đột xuất. Điều đáng chú ý ở đây là, những giá trị được xem là có chứa sai số thô hay giá trị đột xuất nhiều khi là những giá trị số liệu đúng, nó ẩn chứa những thông tin lý thú về sự biến đổi bất thường của tự nhiên và ta cần quan tâm đến chúng.

2) *Cách khử bỏ sai số thô*

Ký hiệu giá trị đột xuất là x^* và tách chúng ra khỏi chuỗi ban đầu. Giả sử chuỗi còn lại *m* thành phần $\{x_1,...,x_m\}$, ta tính trung bình của chuỗi này:

$$\overline{x_*} = \frac{1}{n}\sum_{t=1}^{m} x_t$$

– Trường hợp đã biết độ lệch bình phương trung bình σ của X, ta tính đại lượng:

$$u = \frac{x^* - \overline{x^*}}{\sigma\sqrt{\frac{m+1}{m}}} \qquad (6.2.1)$$

Đại lượng *u* trong (6.2.1) có phân bố chuẩn chuẩn hoá: $u \in N(0,1)$. Với σ và *m* cố định, rõ ràng trị tuyệt đối của hiệu $x^* - \overline{x^*}$ càng lớn thì $|u|$ càng lớn. Kết quả đánh giá x^* có chứa sai số hay không tuỳ thuộc vào độ lớn của $|u|$. Đặt giả thiết "x^* không chứa sai số", khi đó với xác suất sai phạm sai lầm loại I (α) cho trước ta có:

$$P(|u| \geq u_\alpha) = \alpha \qquad (6.2.2)$$

Từ đó tính được u_α. Và chỉ tiêu để kiểm nghiệm giả thiết là:

1) Nếu $|u| \geq u_\alpha$ thì x^* có chứa sai số thô và ta loại bỏ nó với xác suất phạm sai lầm loại I bằng α.
2) Nếu $|u| < u_\alpha$ thì x^* không chứa sai số thô, có nghĩa là ta chấp nhận x^* với độ tin cậy $1-\alpha$.

– Trường hợp chưa biết độ lệch bình phương trung bình σ của X, ta tính đại lượng:

$$t = \frac{x^* - \overline{x^*}}{s^*} \qquad (6.2.3)$$

trong đó $$s^* = \sqrt{\frac{1}{m-1}\sum_{t=1}^{m}\left(x_t - \overline{x^*}\right)^2}$$

Trị số t trong (6.2.3) sẽ được so sánh với một giá trị tới hạn t(p,m):

Nếu $|t| \geq t(p,m)$ thì x^* có chứa sai số thô và nó sẽ bị khử bỏ

Nếu $|t| < t(p,m)$ thì x^* không chứa sai sô thô, tức là ta chấp nhận nó với độ tin cậy p.

Bảng 6.1 dẫn ra các giá trị tới hạn t(p,n) ứng với các giá trị của độ tin cậy p và dung lượng mẫu m khác nhau. Để quyết định xem có nên khử bỏ giá trị đột xuất x^* hay không ta tính t theo (6.2.3), sau đó chọn độ tin cậy p rồi căn cứ vào dung lượng mẫu m, tra bảng 6.1 ta tìm được t(p,m); kết luận cuối cùng được dựa trên cơ sở so sánh $|t|$ và t(p,n).

Bảng 6.1 Giá trị tới hạn t(p,m) để loại bỏ sai số thô

	p					p			
m	0.950	0.980	0.990	0.999	m	0.950	0.980	0.990	0.999
5	3.04	4.11	5.04	9.430	20	2.145	2.602	2.932	3.979
6	2.78	3.64	4.36	7.41	25	2.105	2.541	2.852	3.819
7	2.62	3.36	3.96	6.37	30	2.079	2.503	2.802	3.719
8	2.51	3.18	3.71	5.73	35	2.061	2.476	2.768	3.652
9	2.43	3.05	3.54	5.31	40	2.048	2.456	2.742	3.602
10	2.37	2.96	3.41	5.01	45	2.038	2.441	2.722	3.565
11	2.33	2.89	3.31	4.79	50	2.030	2.429	2.707	3.532
12	2.29	2.83	3.23	4.62	60	2.018	2.411	2.683	3.492
13	2.26	2.78	3.17	4.48	70	2.009	2.399	2.667	3.462
14	2.24	2.74	3.12	4.37	80	2.003	2.389	2.655	3.439
15	2.22	2.71	3.08	4.28	90	1.998	2.382	2.646	3.423
16	2.20	2.68	3.04	4.20	100	1.994	2.377	2.639	3.409
17	2.18	2.66	3.01	4.13					
18	2.17	2.64	2.98	4.07	∞	1.960	2.326	2.576	3.291

Ghi chú: Những trường hợp 20<m<100 không có trong bảng tính trên đây ta có thể sử dụng phép nội suy tuyến tính. Khi n>100 giá trị t(p,m) được xác định theo công thức::

$$t(p,m) = t(p,\infty) + \frac{t(p,100) - t(p,\infty)}{m}100$$

Ví dụ 6.2 Giả sử số liệu nhiệt độ trung bình tháng 2 trạm A (ghi đến phần mười độ) được cho trong bảng 6.2. Sau khi xem xét ta thấy giá trị 275 đáng nghi ngờ, rất có thể mắc sai số thô. Vậy có nên loại bỏ giá trị này không?

Muốn xác định điều này, ta đánh dấu và để riêng giá trị 275 ra, sau đó tính trung bình và độ lệch chuẩn tập số liệu còn lại. Ta có, m=18, $\overline{x}^{*}$=171, s^{*}=12, do đó, theo (6.2.3) ta tính được t=8.95. Mặt khác, nếu chọn p=0.999 thì t(0.999,18)=4.07. Ta thấy $|t|$=8.59>4.07=t(0.999,18). Do đó, với độ tin cậy 99.9% ta khẳng định số 275 có chứa sai số thô và ta loại bỏ nó ra khỏi chuỗi ban đầu.

Bảng 6.2 Số liệu nhiệt độ trung bình tháng 2 trạm A

161	182	170	172	176
161	181	145	191	190
151	173	171	178	275
162	164	176	166	

Ghi chú: Như đã nói ở trên, việc phát hiện và loại bỏ sai số thô không phải lúc nào cũng thực hiện được. Mặt khác, khi xem xét chuỗi số liệu của một số đặc trưng yếu tố khí hậu ta có thể chỉ ra được những giá trị đột xuất và bằng phương pháp nêu trên ta có đủ cơ sở để loại bỏ chúng. Tuy vậy, thực tế chúng không chứa sai số thô. Trong trường hợp này nếu ta loại bỏ những giá trị đột xuất được phát hiện sẽ vấp phải sai lầm. Bởi vậy trước khi quyết định loại bỏ những giá trị đột xuất được xem là có chứa sai số thô phải cân nhắc, suy xét một cách kỹ lưỡng.

6.3. Bổ khuyết số liệu và kéo dài chuỗi

6.3.1 Đặt bài toán

Giả sử trên một khu vực nào đó có M trạm quan trắc. Khi tiến hành xử lý số liệu cho mục đích nghiên cứu, người ta thấy rằng chỉ có K trong số M trạm đó có độ dài chuỗi đủ lớn, còn M–K trạm khác độ dài chuỗi khá bé. Điều này dẫn đến việc các đặc trưng tính toán được trên M–K chuỗi dung lượng bé không bảo đảm tính ổn định thống kê của điều kiện khí hậu, và do đó chúng không có ý nghĩa sử dụng trong việc so sánh, phân tích.

Vậy, vấn đề đặt ra là, từ lượng thông tin của K trạm dài năm, hãy bổ sung số liệu cho M–K trạm ngắn năm để những đặc trưng thống kê của chúng trở nên có ý nghĩa.

Giải quyết vấn đề này là nội dung của bài toán bổ khuyết số liệu. Ở đây chúng ta sẽ hiểu khái niệm bổ khuyết bao hàm cả việc kéo dài chuỗi số liệu. Cơ sở lý luận của việc giải bài toán này như sau:

Đối với các trường khí tượng giả thiết cơ bản mà trên thực tế thường được chấp nhận là tính đồng nhất và đẳng hướng địa phương. Tức là trong cùng một khu vực có nhiều trạm phân bố tại những địa điểm khác nhau, nhưng nhìn chung các

trạm đều nằm trong cùng một phạm vi tác động của các nhân tố khí hậu. Như vậy hai trạm kế cận trong khu vực sẽ cùng chịu những tác động đồng thời của các nhân tố khí hậu. Và do đó từ những thông tin có được về mức độ tác động của trạm này ta có thể suy ra được mức độ tác động của trạm kia.

Mặt khác, xét các chuỗi số liệu của hai trạm kế cận A và B, giả sử rằng trạm A có chuỗi dài hơn, khi đó dù số liệu của cả hai trạm có tản mạn (các chuỗi đứt quãng) đi chẳng nữa ta vẫn có thể qui chúng vào ba nhóm: Nhóm *n* năm bao gồm những khoảng thời gian mà cả hai trạm đồng thời có số liệu; nhóm *m* năm trong đó chỉ có trạm A có số liệu còn trạm B không có; nhóm *p* năm trong đó trạm B có số liệu còn trạm A không có. Như vậy độ dài thực của chuỗi trạm A là N=n+m, trạm B là n+p. Tuy vậy, vì mục đích của bài toán chúng ta sẽ không đề cập đến *p* năm có số liệu của trạm B. Trên cơ sở qui luật phụ thuộc thống kê giữa hai chuỗi được xây dựng từ nhóm *n* năm mà cả hai trạm cùng có số liệu, ta sẽ bổ khuyết cho trạm B.

Phép suy diễn sẽ được tiến hành tương tự khi sử dụng số liệu của nhiều trạm để bổ khuyết cho một trạm.

6.3.2 Các phương pháp bổ khuyết số liệu

Xét các chuỗi số liệu của hai trạm A và B, trong đó chuỗi trạm A có N thành phần $\{x_t\}=\{x_1,x_2,...,x_n,x_{n+1},...,x_N\}$, chuỗi trạm B có n thành phần $\{y_t\}=\{y_1,y_2,...,y_n\}$, hơn nữa n thành phần $\{y_t,\ t=1..n\}$ của chuỗi trạm B tương ứng cùng thời gian với n thành phần $\{x_t,\ t=1..n\}$ của chuỗi trạm A. Tức là ta có n năm cả hai chuỗi đồng thời có số liệu. Từ tập $\{(x_t,y_t),\ t=1..n\}$ ta tiến hành xây dựng phương trình hồi qui tuyến tính (xem mục 5.3.2):

$$\hat{y} = a_o + a_1 x$$

Hay

$$\hat{y}_t = a_0 + a_1 x_t,\ t=1..n \qquad (6.3.1)$$

trong đó:

$$a_o = \overline{y^{(n)}} - a_1 \overline{x^{(n)}},\quad a_1 = r_{xy}\frac{s_y}{s_x}$$

$$\overline{x^{(n)}} = \frac{1}{n}\sum_{t=1}^{n} x_t,\ \overline{y^{(n)}} = \frac{1}{n}\sum_{t=1}^{n} y_t,\ s_x = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(x_t - \overline{x^{(n)}}\right)^2},$$

$$s_y = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_t - \overline{y^{(n)}}\right)^2},\ r_{xy} = \left[\frac{1}{n}\sum_{t=1}^{n}\left(x_t - \overline{x^{(n)}}\right)\left(y_t - \overline{y^{(n)}}\right)\right]/(s_x.s_y)$$

(Trong chương này, ký hiệu chỉ số phía trên nằm trong ngoặc đơn chỉ độ dài chuỗi được sử dụng để tính toán. Ví dụ, đại lượng $\overline{y^{(n)}}$ là giá trị trung bình của chuỗi $\{y_t,t=1..n\}$, còn $\overline{y^{(N)}}$ là trung bình của chuỗi $\{y_t,\ t=1..N\}$.

Hệ thức (6.3.1) có thể được viết thành:

$$\hat{y}_t = \overline{y_n} + r_{xy}\frac{s_y}{s_x}(x_t - \overline{x_n}),\ (t=1..n) \qquad (6.3.2)$$

Phương trình (6.3.2) mô tả qui luật phụ thuộc tuyến tính của chuỗi $\{y_t\}$ vào chuỗi $\{x_t\}$ trong thời gian *n* năm. Nếu giả thiết rằng qui luật này vẫn phù hợp với thời đoạn N–n năm mà trạm B bị khuyết, ta có công thức bổ khuyết sau:

$$y_{n+i} = \overline{y^{(n)}} + r_{xy}\frac{s_y}{s_x}(x_{n+i} - \overline{x^{(n)}}),\ (i=1..N-n) \qquad (6.3.3)$$

Công thức (6.3.3) được gọi là phương pháp hồi qui bổ khuyết số liệu. Nếu cả hai trạm A và B có chung nhịp điệu dao động về trị số khí hậu, khi đó một cách gần đúng có thể xem $r_{xy}\approx 1$ và (6.3.2) trở thành:

$$\hat{y}_t = \overline{y^{(n)}} + \frac{s_y}{s_x}(x_t - \overline{x^{(n)}}),\ (t=1..n) \qquad (6.3.4)$$

Người ta gọi đây là phương pháp Wild. Tương ứng với (6.3.3) và (6.3.4) ta có công thức bổ khuyết cho trạm B là:

$$y_{n+i} = \overline{y^{(n)}} + \frac{s_y}{s_x}(x_{n+i} - \overline{x^{(n)}}),\ (i=1..N-n) \qquad (6.3.5)$$

Nếu giả thiết số liệu hai chuỗi đồng thời có cùng nhịp điệu dao động và mức độ dao động, tức là xem $r_{xy}=1$ và $s_x=s_y$ thì công thức bổ khuyết được gọi là công thức hiệu số (hay phương pháp hiệu số)

$$y_{n+i} = \overline{y^{(n)}} + (x_{n+i} - \overline{x^{(n)}}),\ (i=1..N-n) \qquad (6.3.6)$$

Trong trường hợp các chuỗi số liệu của hai trạm A và B quan hệ với nhau theo qui luật tỷ lệ thuận:

$$y_t = kx_t,\ (t=1..n) \qquad (6.3.7)$$

Ta có:
$$\sum_{t=1}^{n} y_t = k\sum_{t=1}^{n} x_t \text{, hay: } k = \frac{\overline{y^{(n)}}}{\overline{x^{(n)}}} \qquad (6.3.8)$$

Với giả thiết qui luật này vẫn đúng cho N–n năm còn lại, ta có công thức bổ khuyết:

$$y_{n+i} = \frac{\overline{y^{(n)}}}{\overline{x^{(n)}}}x_i,\ (i=1..N-n) \qquad (6.3.9)$$

Người ta gọi công thức bổ khuyết này là phương pháp tỷ số.

Ta nhận thấy rằng, các công thức bổ khuyết theo phương pháp Wild và phương pháp hiệu số chỉ là những trường hợp riêng của phương pháp hồi qui tuyến tính. Trong trường hợp hai chuỗi quan hệ với nhau theo qui luật phi tuyến tính ta cũng có thể tiến hành tương tự.

Đặc biệt, nếu lân cận trạm cần bổ khuyết (trạm B) có nhiều hơn một trạm có chuỗi số liệu dài (chẳng hạn có K trạm) ta cũng có thể phân các chuỗi số liệu thành hai nhóm: Nhóm *n* năm trong đó tất cả các trạm đồng thời có số liệu và nhóm N–n năm trong đó các trạm khác có số liệu, trừ trạm cần bổ khuyết:

Trạm A_1	Trạm A_2	...	Trạm A_k	Trạm B
x_{11}	x_{12}	...	x_{ik}	y_1
x_{21}	x_{22}	...	x_{2k}	y_2
...	...	...	...	...
x_{n1}	x_{n2}	...	x_{nk}	y_n
$x_{n+1,1}$	$x_{n+1,2}$	...	$x_{n+1,k}$	
...	...	...	...	
x_{N1}	x_{N2}	...	x_{NK}	

Từ bộ số liệu $\{y_1,x_{t1},x_{t2},...x_{tk}\}$ (t=1..n) ta tiến hành xây dựng phương trình hồi qui tuyến tính (xem mục 5.5.2):

$$\hat{y} = a_0+a_1x_1+a_2x_2+...+a_kx_k \qquad (6.3.10)$$

Hay $$\hat{y}_t =a_0+a_1x_{t1}+a_2x_{t2}+...+a_Kx_{tK},\ (t=1..n) \qquad (6.3.11)$$

trong đó a_i, i=0..K là các hệ số hồi qui.

Phương trình (6.310) biểu thị sự phụ thuộc hàm tuyến tính của số liệu trạm B vào số liệu của K trạm $A_1,...A_k$. Với giả thiết rằng qui luật này vẫn phù hợp đối với thời gian N–n năm mà trạm B không có số liệu ta có công thức bổ khuyết là:

$$\hat{y}_{n+i} =a_0+a_1x_{n+i,1}+a_2x_{n+i,2}+...+a_Kx_{n+i,K},\ (i=1..N–n) \qquad (6.3.12)$$

Đây là công thức bố khuyết bằng hồi qui tuyến tính nhiều biến (hay còn gọi là hồi qui nhiều trạm).

6.4 Qui số liệu trung bình về cùng thời kỳ dài

Trong ứng dụng thực hành người ta thường quan tâm đến các đặc trưng có tính ổn định của điều kiện khí hậu. Một trong những đặc trưng hết sức quan trọng thường được chú ý đến là trị số trung bình.

Đối với những trạm có chuỗi số liệu ngắn trị số trung bình tính được nhiều khi không đảm bảo độ ổn định và vì thế nó không có tác dụng so sánh. Bởi vậy, vấn

đề đặt ra là cần phải qui trị số trung bình của những trạm ngắn năm về thời kỳ dài trên cơ sở những mối quan hệ thống kê giữa nó và các trạm dài năm.

Giả sử cần qui số liệu trung bình của trạm ngắn năm B về thời kỳ dài căn cứ vào mối quan hệ tương quan giữa nó với trạm dài A. Ta nhận thấy rằng trong thời kỳ *n* năm (mà cả hai trạm đồng thời có số liệu), ta có thể xác định được các đặc trưng thống kê như trung bình, hệ số tương quan, độ lệch chuẩn. Mặt khác đối với trạm A ta tính được giá trị trung bình trong thời kỳ N năm (thời kỳ dài). Vấn đề ở đây là cần xác định được giá trị trung bình của chuỗi B cũng trong thời kỳ N năm đó. Việc tính trung bình của chuỗi B như vậy được gọi là qui số liệu trung bình về thời kỳ dài.

Nếu chuỗi số liệu trạm A đủ dài và được coi là trạm chuẩn thì phép qui trung bình của trạm B về thời kỳ dài theo trạm A được gọi là phép qui về chuẩn.

Trong quá trình tiến hành phép qui ta có thể sử dụng phép qui nhiều bước. Chẳng hạn, nếu số liệu trạm B có thể qui được về thời kỳ dài theo trạm A nhưng ta không thể thực hiện được phép qui từ trạm C về thời kỳ dài theo trạm A do phép qui không đạt tiêu chuẩn, khi đó ta có thể tiến hành qui số liệu của trạm C về thời kỳ dài theo trạm B là trạm đã qui theo A, với điều kiện phép qui đạt tiêu chuẩn.

Sau đây ta sẽ xét một số phương pháp qui dựa trên cơ sở các phương pháp bổ khuyết số liệu đã trình bày ở trên.

Ký hiệu $\overline{y^{(N)}}$ là giá trị trung bình đã qui của trạm B (trung bình thời kỳ dài), $\overline{y^{(n)}}$ là trung bình của B tính trên số liệu thực có, $\overline{x^{(N)}}$ và $\overline{x^{(n)}}$ tương ứng là trung bình trạm A trong thời kỳ dài (N năm) và thời kỳ ngắn (n năm). Từ các công thức (6.3.2) và (6.3.3) ta có:

$$\overline{y^{(N)}} = a_0 + a_1\overline{x^{(N)}} = \overline{y^{(n)}} - a_1\overline{x^{(n)}} + a_1\overline{x^{(N)}} = \overline{y^{(n)}} + a_1(\overline{x^{(N)}} - \overline{x^{(n)}})$$

Hay
$$\overline{y^{(N)}} = \overline{y^{(n)}} + r_{xy}\frac{s_y}{s_x}(\overline{x^{(N)}} - \overline{x^{(n)}}) \qquad (6.4.1)$$

Công thức (6.4.1) được gọi là phép qui theo phương pháp hồi qui. Bằng cách tương tự ta có thể nhận được:

– Phép qui theo phương pháp Wild: $\overline{y^{(N)}} = \overline{y^{(n)}} + \frac{s_y}{s_x}(\overline{x^{(N)}} - \overline{x^{(n)}})$ (6.4.2)

– Phép qui theo phương pháp hiệu số: $\overline{y^{(N)}} = \overline{y^{(n)}} + (\overline{x^{(N)}} - \overline{x^{(n)}})$ (6.4.3)

– Phép qui theo phương pháp tỷ số: $\overline{y^{(N)}} = \frac{\overline{y^{(n)}}}{\overline{x^{(n)}}}\overline{x^{(N)}}$ (6.4.4)

– Phép qui theo hồi qui nhiều trạm:

$$\overline{y^{(N)}} = \overline{y^{(n)}} + \sum_{i=1}^{K} a_i \left(\overline{x_i^{(N)}} - \overline{x_i^{(n)}} \right) \qquad (6.4.5)$$

trong đó $\overline{x_i^{(N)}}$ và $\overline{x_i^{(n)}}$ là trung bình thời kỳ N năm và *n* năm của trạm A_i, còn a_i là các hệ số hồi qui (i=1..K).

Một số nhận xét

Việc bổ khuyết số liệu cũng như qui số liệu trung bình về thời kỳ dài được trình bày trên đây nói chung khá thuận tiện cho quá trình tính toán thủ công hoặc tính toán bằng những công cụ thô sơ. Khi xử lý với những tập số liệu dài hoặc cần xử lý với nhiều tập số liệu mà khối lượng tính toán lớn thì các phương pháp trên đây cho phép làm giảm thời gian tính toán một cách đáng kể.

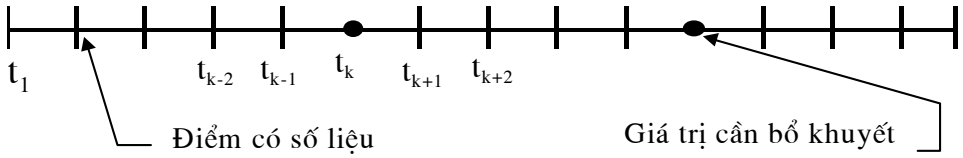
Tuy nhiên, cùng với sự phát triển mạnh mẽ của công nghệ tin học và máy tính, thời gian tính toán cũng như khối lượng tính toán nhiều khi không còn là vấn đề lo ngại. Do đó cái mà người ta quan tâm hiện nay là độ chính xác của phương pháp. Bởi vậy trong các phương pháp bổ khuyết số liệu và qui số liệu trung bình về thời kỳ dài được xét trên đây thì phương pháp hồi qui được áp dụng nhiều nhất.

6.5 Liên tục hoá chuỗi số liệu

6.5.1 Đặt bài toán

Liên tục hoá (hay còn gọi là lấp đầy) chuỗi số liệu là thực hiện việc bổ sung vào những vị trí khuyết số liệu của chuỗi để biến chuỗi ban đầu thành chuỗi có bước thời gian đều nhau. Hình 6.1 đưa ra sơ đồ ví dụ minh họa về yêu cầu của bài toán liên tục hoá chuỗi số liệu.

Ta có thể thực hiện việc liên tục hoá bằng các phương pháp bổ khuyết được trình bày trên đây. Người ta gọi đó là phương pháp sử dụng trạm tựa. Nó là một trong những phương pháp có hiệu quả vì nó được dựa trên giả thiết về tính đồng nhất, đẳng hướng địa phương của các trường khí tượng. Tuy nhiên trong một vài trường hợp phương pháp này tỏ ra không hiệu lực bởi các chuỗi đều bị gián đoạn vào cùng một thời điểm hoặc các trạm cách nhau quá xa, làm cho giả thiết về tính đồng nhất đẳng hướng địa phương bị vi phạm; mối liên hệ tương quan giữa các chuỗi vì thế mà quá yếu, không đảm bảo độ chính xác. Trong trường hợp này phương pháp nội suy trên chính chuỗi cần bổ khuyết tỏ ra có ưu thế hơn.



Hình 6.1 Sơ đồ chuỗi số liệu cần liên tục hoá

Về cơ bản bài toán liên tục hoá chuỗi số liệu được đặt ra như sau:

Cho chuỗi thời gian $x(t_i)$, (i=1,2,...,n) từ t_1 đến t_n, trong đó t_i chỉ thời điểm có số liệu. Về nguyên tắc các thời điểm t_i cách đều nhau. Nhưng trên thực tế chuỗi bị khuyết đi một số giá trị $x(t_o)$ nào đó ($t_1 < t_o < t_n$– hình 6.1). Yêu cầu cần tính được giá trị $x(t_o)$ bị khuyết thiếu này.

6.5.2 Phương pháp nội suy tuyến tính tối ưu lấp đầy chuỗi

Phương pháp nội suy tuyến tính tối ưu được áp dụng trên cơ sở giả thiết rằng chuỗi $x(t_i)$, (i=1,2,...,n) là các giá trị của một thể hiện của quá trình ngẫu nhiên dừng X(t) tại *n* lát cắt t_i. Giá trị cần nội suy $x(t_o)$ được xem như là kết quả của việc tác dụng toán tử tuyến tính lên tập hợp các giá trị $x(t_k)$, với $t_k \neq t_o$ và k=1,2,...,m là các lát cắt được sử dụng để nội suy giá trị $x(t_o)$:

$$x(t_o) = \sum_{k=1}^{m} \alpha_k x(t_k) \tag{6.5.1}$$

trong đó α_k (k=1..m) được gọi là các trọng số nội suy, đó là những hệ số phải tìm. Bài toán dẫn đến việc xác định các hằng số α_k (k=1..m) để cho sai số bình phương trung bình của phép nội suy đạt cực tiểu:

$$\sigma_m^2(\alpha_1,\alpha_2,...,\alpha_m) = \overline{\left[\left(X(t_o) - \sum_{k=1}^{m} \alpha_k X(t_k)\right)^2\right]} \longrightarrow \min \tag{6.5.2}$$

Điều kiện cần và đủ để thoả mãn (6.5.2) là tất cả các đạo hàm riêng của $\sigma_m^2(\alpha_1,\alpha_2,...,\alpha_m)$ theo các α_k đều phải triệt tiêu:

$$\frac{\partial \sigma_m^2(\alpha_1,\alpha_2,...,\alpha_m)}{\partial \alpha_k} = 0, \text{ (k=1..m)} \tag{6.5.3}$$

Không làm mất tính tổng quát, ta giả thiết rằng kỳ vọng toán học M[X(t)] = 0, điều này cũng có nghĩa là chuỗi ban đầu đã được qui tâm, khi đó, từ (6.5.2) ta có:

$$\sigma_m^2(\alpha_1,..,\alpha_m) = \overline{X^2(t_o) - 2\sum_{k=1}^{m} \alpha_k X(t_o)X(t_k) + \sum_{k=1}^{m}\sum_{j=1}^{m} \alpha_k \alpha_j X(t_k)X(t_j)} =$$

$$= R_x(0) - 2\sum_{k=1}^{m} \alpha_k R_x(t_o - t_k) + \sum_{k=1}^{m}\sum_{j=1}^{m} \alpha_k \alpha_j R_x(t_j - t_k) \tag{6.5.4}$$

Trong đó $R_x(t_j–t_k)$ và $R_x(t_o–t_k)$ là các giá trị của hàm tương quan của quá trình ngẫu nhiên X(t). Thay (6.5.4) vào (6.5.3) ta nhận được:

$$\frac{\partial\sigma_m^2(\alpha_1,..,\alpha_m)}{\partial\alpha_k} = -2\,R_x(t_o - t_k) + 2\sum_{j=1}^{m}\alpha_j R_x(t_j - t_k) = 0,\ (k=1..m)$$

Hay
$$\sum_{j=1}^{m}\alpha_j R_x(t_j - t_k) = R_x(t_o - t_k),\ (k=1..m) \qquad (6.5.5)$$

Đây là một hệ phương trình đại số tuyến tính có *m* phương trình và *m* ẩn số. Trong đó hàm tương quan $R_x(\tau)$ được xác định theo công thức sau:

$$R_x(\tau_k) = R_x(k\Delta\tau) = \frac{1}{n-k}\sum_{i=1}^{n-k}x(t_i)x(t_{i+k}) \qquad (6.5.6)$$

với $\Delta\tau$ là bước thời gian của chuỗi. Thông thường trong khí hậu $\Delta\tau$ không đổi và bằng 1 năm.

Giải hệ (6.5.5) ta nhận được các trong số nội suy α_k phải tìm. Sau khi đã có được các α_k, thay vào công thức (6.5.1) ta tính được giá trị cần nội suy $x(t_o)$.

Thay (6.5.5) vào (6.5.4) ta có biểu thức để đánh giá sai số của phép nội suy:

$$\sigma_m^2(\alpha_1,..,\alpha_m) = R_x(0) - \sum_{k=1}^{m}\sum_{j=1}^{m}\alpha_k\alpha_j R_x(t_j - t_k) \qquad (6.5.7)$$

Vì hàm tương quan là xác định dương nên hạng thứ hai vế phải không âm:

$$\sum_{k=1}^{m}\sum_{j=1}^{m}\alpha_k\alpha_j R_x(t_j - t_k) \ge 0$$

Từ đó ta có: $\sigma_m^2(\alpha_1,..,\alpha_m) \le R_x(0) = D_x$.

Tức là sai số của phép nội suy không vượt quá phương sai của quá trình ngẫu nhiên X(t).

Ta hãy xét một số trường hợp đặc biệt:

1) Giả sử $R_x(t_o–t_k) = 0$, tức là giá trị cần nội suy không tương quan với các điểm được chọn để nội suy, khi đó:

$$\sum_{j=1}^{m}\alpha_j R_x(t_j - t_k) = 0,\ (k=1..m) \qquad (6.5.8)$$

Từ đó suy ra $\alpha_1=\alpha_2=...=\alpha_m=0$, tức là giá trị nội suy chính bằng kỳ vọng (trung bình) của chuỗi. Đây là một tính chất quan trọng nhưng được áp dụng trong thực tế: nhiều khi để đơn giản người ta gán giá trị khuyết thiếu (giá trị cần nội suy) bằng chính trung bình của chuỗi. Sai số nội suy trong trường hợp này bằng phương sai của chuỗi.

2) Giả sử $R_x(t_j - t_k) = 0$ khi j≠k, tức là các giá trị được chọn làm nội suy không tương quan với nhau nhưng có tương quan với giá trị cần nội suy, khi đó ta có:

$$\alpha_k R_x(0) = R_x(t_o - t_k),\ (k=1..m)$$

Suy ra:
$$\alpha_k = \frac{R_x(t_o - t_k)}{R_x(0)} = r_x(t_o - t_k) \qquad (6.5.9)$$

Trong trường hợp này, các trọng số nội suy α_k bằng giá trị của hệ số tương quan giữa điểm cần nội suy và các điểm được chọn để nội suy.

6.5.3 Nội suy parabol

Nội suy parabol dựa trên cơ sở xem chuỗi ban đầu như là một hàm của thời gian:

$$x(t) = f(t) \qquad (6.5.10)$$

còn $x(t_0)$, $t_0 \neq t_i$, là điểm cần nội suy.

ta sẽ gọi các điểm t_ii=1..n, l các nút nội suy. Đa thức P(t) dược xác định duy nhất bằng các nút và bằng giá trị của chuỗi tại các nút đó. Yêu cầu của phép nội suy là giữ nguyên giá trị của chuỗi các nút nội suy, nên sai số quan trắc, nếu có, vẫn được bảo toàn.

Đa thức nội suy P(t) được thiết lập theo công thức Lagrange:

$$P(t) = \sum_{i=1}^{n} L_i(t) x_i \qquad (6.5.11)$$

trong đó $x_i = x(t_i)$ là các giá trị của chuỗi. Đa thức $L_i(t)$ được xác định bởi các nút nội suy:

$$L_i(t) = \frac{(t-t_1)...(t-t_{i-1})(t-t_{i+1})...(t-t_n)}{(t_i-t_1)...(t_i-t_{i-1})(t_i-t_{i+1})...(t_i-t_n)},\ (i=1..n) \qquad (6.512)$$

và lấy giá trị tại các nút đó: $L_i(t_j) = \delta_{ij} = \begin{cases} 1 & \text{khi} i = j \\ 0 & \text{khi} i \neq j \end{cases}$ (6.513)

Như vậy ta dễ dàng xác định được giá trị nội suy $x(t_o)$:

$$x(t_o) = P(t_o) = \sum_{i=1}^{n} L_i(t_o) x(t_i) \qquad (6.5.14)$$

với
$$L_i(t_o) = \frac{(t_o-t_1)...(t_o-t_{i-1})(t_o-t_{i+1})...(t_o-t_n)}{(t_i-t_1)...(t_i-t_{i-1})(t_i-t_{i+1})...(t_i-t_n)},\ (i=1..n)$$

Khi n=2 ta có công thức nội suy tuyến tính quen thuộc:

$$\frac{x(t_o)-x(t_1)}{x(t_2)-x(t_1)} = \frac{t_o-t_1}{t_2-t_1} \text{ hay } x(t_o) = \frac{t_o-t_2}{t_1-t_2}x(t_1) + \frac{t_o-t_1}{t_2-t_1}x(t_2).$$

CHƯƠNG 7
PHÂN TÍCH CHUỖI THỜI GIAN

7.1 Cấu trúc chuỗi thời gian

Chuỗi thời gian là chuỗi số liệu được sắp xếp theo trình tự thời gian. Phân tích chuỗi thời gian là nghiên cứu cấu trúc bên trong của chuỗi với mục đích tìm kiếm và phát hiện những qui luật biến đổi theo thời gian. Nói chung các chuỗi thời gian thường ẩn chứa nhiều thành phần khác nhau. Đối với các quá trình khí tượng, khí hậu chuỗi thời gian thường chứa đựng các thành phần sau đây:

- Dao động ngẫu nhiên: Là những biến đổi thăng giáng không phụ thuộc vào thời gian của các thành phần trong chuỗi
- Nhiễu động: Là những biến đổi bất thường mang tính ngẫu nhiên, tuy vậy giữa chúng vẫn tồn tại những mối quan hệ nào đó và chúng có thể xuất hiện sau những khoảng thời gian nhất định
- Dao động tuần hoàn: Là những biến đổi biểu hiện tính chất thẳng giáng có nhịp điệu đều đặn, vì vậy người ta còn gọi đó là thành phần dao động nhịp điệu
- Dao động có chu kỳ: Là những dao động biến đổi có tính lặp lại tương đối thường xuyên sau những khoảng thời gian khá đều đặn
- Thành phần xu thế: Biểu hiện xu hướng tăng hoặc giảm theo thời gian của các thành phần trong chuỗi

Trong thực tế nghiên cứu người ta thường đồng nhất thành phần dao động ngẫu nhiên với thành phần nhiễu động và thành phần tuần hoàn với thành phần dao động có chu kỳ, mặc dù sự đồng nhất này chắc chắn không thoả đáng. Tuy nhiên, có sự phân biệt đáng kể giữa khái niệm chuỗi thời gian trong khí tượng và chuỗi thời gian trong khí hậu. Theo quan điểm khí tượng, hai trị số kế cận trong chuỗi thời gian có thể cách nhau một giờ, một kỳ quan trắc (3 hoặc 6 giờ), một ngày, một tháng và thậm chí dưới một giờ, nhưng không nhất thiết phải là một năm. Vì vậy, có thể xem chuỗi thời gian trong khí tượng bao gồm các thành phần:

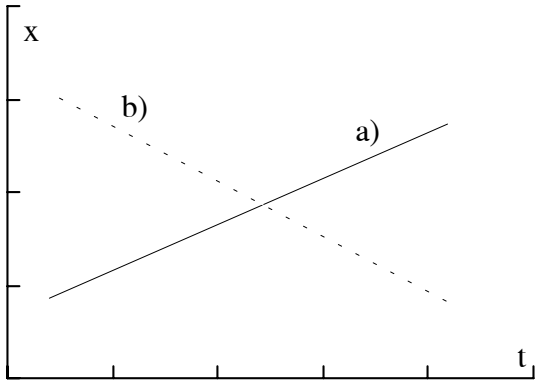
- Dao động tuần hoàn ngày, tức là những biến đổi theo chu kỳ ngày

- Dao động tuần hoàn năm, tức là những biến đổi theo chu kỳ năm
- Xu thế dài năm
- Chu kỳ dài năm
- Dao động ngẫu nhiên

Còn cơ cấu chuỗi thời gian trong khí hậu chỉ chứa 3 thành phần cơ bản:

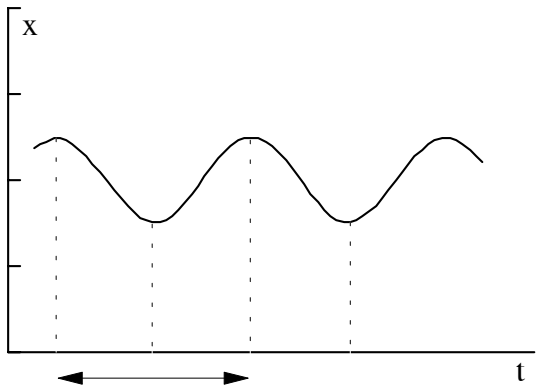
- Xu thế dài năm
- Chu kỳ dài năm
- Thành phần ngẫu nhiên

1) Xu thế dài năm: Minh hoạ về xu thế dài năm được dẫn ra trên hình 7.1. Đó là những biến đổi của chuỗi số liệu có tính chất đơn điệu và tương đối thường xuyên. Tốc độ biến đổi của chuỗi gần như đồng đều. Các trị số của chuỗi có xu thế tăng dần hoặc giảm dần đến giá trị lớn nhất hoặc nhỏ nhất. Tuy vậy không nhất thiết đó là xu thế tuyến tính.
2) Chu kỳ dài năm: Chu kỳ dài năm là những biến đổi của chuỗi mang tính chất lặp lại giá trị sau những khoảng thời gian nhất định nào đó (hình 7.2). Mối tương quan giữa các thành phần trong chuỗi thường đạt trị số lớn nhất khi xét tới hai thành phần cách nhau một số năm xấp xỉ với độ dài chu kỳ.



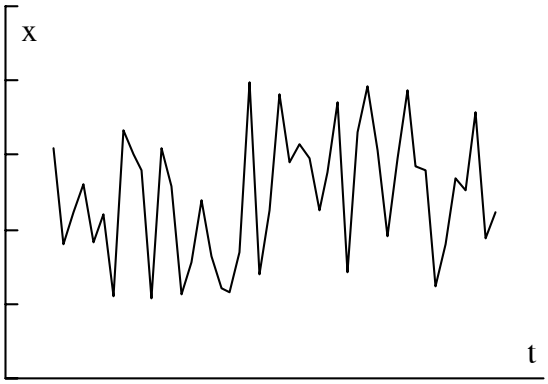
Hình 7.1 Biến đổi xu thế dài năm

a) Xu thế tăng; b) Xu thế giảm

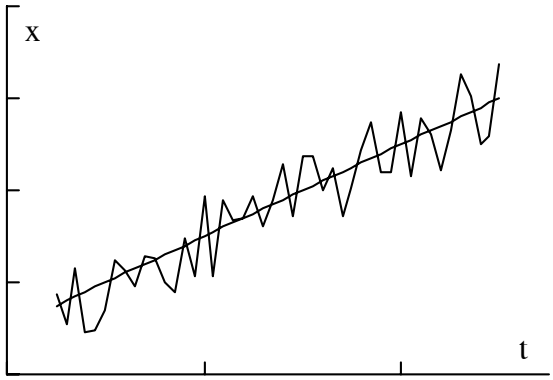


Hình 7.2 Biến đổi chu kỳ dài năm

3) Dao động ngẫu nhiên: Hình 7.3 minh hoạ về tính dao động ngẫu nhiên của chuỗi. Đó là những biến đổi thường xuyên không ổn định. Dấu chuẩn sai của một vài thành phần kế cận thường khác nhau. Biên độ động thường không quá lớn và nói chung xoay quanh giá trị trung bình. Bởi vậy giá trị trung bình được coi là chuẩn mực thăng bằng của các dao động ngẫu nhiên.

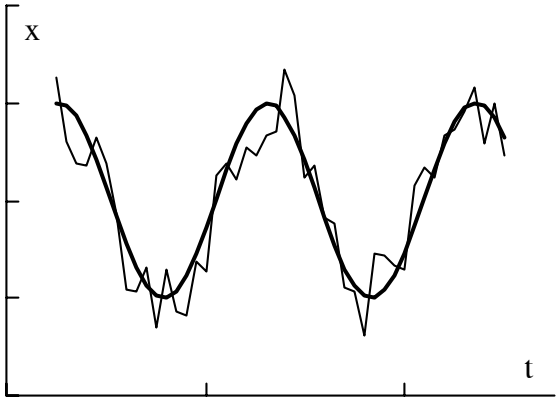


Hình 7.3 Dao động ngẫu nhiên

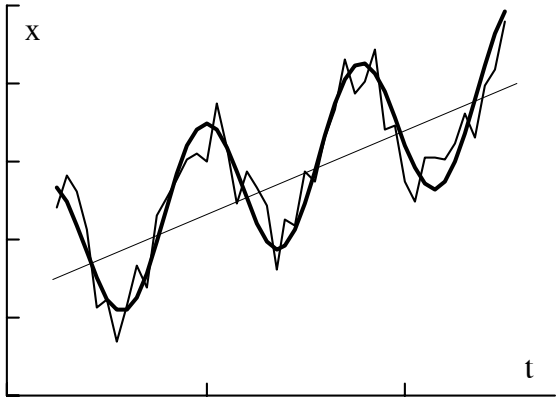


Hình 7.4 Kết hợp xu thế và ngẫu nhiên

Trong thực tế các chuỗi thường tồn tại kết hợp hai (hình 7.4, 7.5) hoặc ba (hình 7.6) thành phần nói trên, trong đó thành phần ngẫu nhiên luôn xuất hiện.



Hình 7.5 Kết hợp chu kỳ và ngẫu nhiên



Hình 7.6 Kết hợp cả 3 thành phần

Nội dung bài toán phân tích chuỗi thời gian bao gồm hai vấn đề chính là phân tích xu thế và phân tích chu kỳ. Đó cũng là những nội dung cơ bản của bài toán nghiên cứu biến đổi khí hậu mà ta có thể nêu lên dưới dạng bài toán sau:

Cho chuỗi thời gian $\{x_t, t=1..n\}$ của đặc trưng yếu tố khi hậu nào đó. Trên cơ sở phân tích cấu trúc thống kê của chuỗi hãy xác định xu thế biến đổi dài năm và tính dao động có chu kỳ của đặc trưng yếu tố đó.

Tuy nhiên, như đã thấy, chuỗi thời gian luôn luôn chứa đựng thành phần dao động ngẫu nhiên. Để có thể phát hiện được xu thế biến đổi và các chu kỳ dao động, cần thiết phải lọc bỏ những dao động ngẫu nhiên trong chuỗi. Và như vậy, xuất hiện một nhiệm vụ quan trọng trong bài toán phân tích chuỗi thời gian là lọc chuỗi hay làm trơn chuỗi.

7.2 Vài nét về phân tích chuỗi thời gian trong khí tượng, khí hậu

Việc phân tích chuỗi thời gian bằng công cụ thống kê buộc phải chấp nhận một giả thiết hết sức cơ bản là tính dừng của các quá trình khí quyển. Tính dừng ở

đây có nghĩa là mọi tính chất thống kê của quá trình trong quá khứ vẫn được bảo toàn cho cả trong tương lai. Khái niệm này được ứng dụng khá phổ biến trong các mô hình thống kê dự báo thời tiết, khí hậu. Đương nhiên rằng ta không nên tin tưởng tuyệt đối vào những trị số dự báo được trong tương lai thông qua chuỗi số liệu quan trắc hiện có của quá trình đang xét. Chẳng hạn, từ việc phân tích chuỗi số liệu nhiệt độ (và chỉ có nhiệt độ mà thôi!) ta có thể đưa ra được giá trị dự báo của nó trong tương lai, nhưng hãy cảnh giác với độ chính xác của dự báo. Tuy nhiên, trong rất nhiều trường hợp giả thiết về tính dừng lại tỏ ra rất hợp lý.

Có hai phương pháp tiếp cận cơ bản khi phân tích chuỗi thời gian, là phân tích chuỗi trên miền thời gian và phân tích chuỗi trên miền tần số. Về bản chất, xuất phát điểm của các phương pháp này rất khác nhau, nhưng chúng không hoàn toàn độc lập với nhau mà bù trừ cho nhau về mặt biểu diễn toán học.

Phương pháp phân tích trên miền thời gian tìm các đặc trưng của chuỗi số liệu dựa vào công cụ cơ bản là hàm tự tương quan (autocorrelation function). Phương pháp phân tích trên miền tần số biểu diễn sự biến đổi của chuỗi số liệu như là hàm của những tần số dao động, qua đó làm xuất hiện sự đóng góp hay tích luỹ năng lượng của quá trình tại những quy mô thời gian hoặc những tần số đặc trưng khác nhau.

Đối với những chuỗi số liệu mà có thể xem chúng như tập các giá trị có thể của biến ngẫu nhiên rời rạc, phân tích miền thời gian được thực hiện trên cơ sở khái niệm xích Markov. Có thể hình dung xích Markov như là hệ thống các trạng thái xảy ra liên tiếp theo thời gian. Chuỗi các trạng thái này cần phải thoả mãn những thuộc tính nào đó, được gọi là thuộc tính Markov. Chẳng hạn, thuộc tính của xích Markov bậc nhất có thể được biểu diễn bởi:

$$P(X_{t+1}/X_t,X_{t-1},...,X_1) = P(X_{t+1}/X_t) \qquad (7.2.1)$$

trong đó X_i, i=1, 2,... là các trạng thái của hệ thống tại các thời điểm i=1, i=2,..., i=t, còn t là thời điểm hiện tại.

Biểu thức (7.2.1) hàm ý rằng xác suất để hệ nhận trạng thái X_{t+1} tại thời điểm *t+1* chỉ phụ thuộc vào trạng thái của hệ tại thời điểm *t* (X_t). Hay nói cách khác, xác suất của trạng thái tương lai chỉ phụ thuộc vào trạng thái hiện tại mà không phụ thuộc vào quá khứ. Ví dụ, giá trị dự báo nhiệt độ tối thấp ngày mai chỉ phụ thuộc vào số liệu quan trắc ngày hôm nay, còn những số liệu của các quan trắc trước đó không có ý nghĩa cung cấp thông tin thêm cho việc dự báo này. Người ta gọi xác suất biểu diễn bởi (7.2.1) là xác suất chuyển trạng thái của xích Markov, nó là xác suất có điều kiện.

Mô hình xích Markov cho các biến rời rạc có thể được xét trên nhiều phương diện khác nhau, như xích Markov bậc nhất hay bậc cao, xích Markov hai hay nhiều trạng thái. Ví dụ, có thể ứng dụng xích Markov bậc nhất hai trạng thái để khảo sát chuỗi các sự kiện "có mưa" hay "không mưa". Các sự kiện này diễn ra liên tiếp theo thời gian và chúng có thể được mã hoá bởi các trị số 0 (không có mưa xuất hiện) và 1 (có mưa xuất hiện). Biến trạng thái của hệ trong trường hợp này là một biến nhị phân X={0, 1}. Như vậy, theo tiến trình thời gian giá trị của X là một chuỗi các số 0 hoặc 1. Tức là ta có, chẳng hạn, x_1=0, x_2=0, x_3=1, x_4=1, x_5=0,...,x_t=1. Với mô hình bậc nhất ta cần quan tâm đến xác suất để hệ nhận trạng thái tại thời điểm *t+1* trong tương lai khi đã biết trạng thái hiện tại của hệ (xác suất chuyển trạng thái): $P(X_{t+1}/X_t)$. Các xác suất chuyển trạng thái đó là:

$$p_{00} = P(X_{t+1} = 0/\ X_t = 0)$$
$$p_{01} = P(X_{t+1} = 1/\ X_t = 0)$$
$$p_{10} = P(X_{t+1} = 0/\ X_t = 1)$$
$$p_{11} = P(X_{t+1} = 1/\ X_t = 1)$$

Đối với những biến liên tục, như nhiệt độ, áp suất, lượng mưa,... mô hình xích Markov trên đây không phù hợp, bởi ta không thể liệt kê tất cả các giá trị có thể của chúng. Trong trường hợp này, thay cho xích Markov người ta sử dụng khái niệm mô hình tự hồi qui, hay mô hình Box–Jenkins. Mô hình đơn giản nhất loại này là mô hình tự hồi qui bậc nhất (First order Autoregression – AR(1)). Đôi khi người ta còn gọi mô hình AR(1) là quá trình Markov hay sơ đồ Markov. Thuộc tính Markov (7.2.1) trong trường hợp này có thể được biểu diễn dưới dạng:

$$P(X_{t+1} \le x_{t+1} / X_t \le x_t, X_{t-1} \le x_{t-1},..., X_1 \le x_1) = P(X_{t+1} \le x_{t+1} / X_t \le x_t) \qquad (7.2.2)$$

trong đó x_t là giá trị của X tại thời điểm t.

Mô hình tự hồi qui bậc nhất đối với chuỗi thời gian $\{x_t\}$ của biến liên tục X có thể được biểu diễn dưới dạng:

$$x_{t+1} - \mu = \phi(x_t - \mu) + \varepsilon_{t+1}$$

trong đó x_t và x_{t+1} tương ứng là giá trị của chuỗi tại thời điểm t và t+1, μ là trung bình của chuỗi, ϕ là tham số tự hồi qui và ε là phần dư hay sai số.

Có thể hiểu mô hình AR(1) như là phương trình hồi qui tuyến tính dự báo giá trị của biến ngẫu nhiên X với yếu tố dự báo là giá trị trong tương lai (thời điểm t+1) và nhân tố dự báo là giá trị hiện tại của X. Giá trị tại thời điểm tương lai x_{t+1} của X được xác định bởi hai thành phần: thành phần thứ nhất là hàm của x_t, thành phần thứ hai, ε_{t+1}, là một biến ngẫu nhiên mà thường được giả thiết là có phân bố chuẩn với kỳ vọng bằng 0 và phương sai bằng σ_ε^2. Trong thực tế, do giả thiết tính

dừng của chuỗi thời gian, trung bình μ được lấy bằng trung bình số học của chuỗi và xem nó không đổi theo thời gian. Ước lượng thống kê của tham số tự hồi qui ϕ là trị số của hàm tự tương quan tại đối số bằng khoảng thời gian giữa hai thời điểm.

7.3 Các phép biến đổi và lọc chuỗi

Trong nhiều trường hợp việc biến đổi chuỗi số liệu ban đầu về chuỗi mới để từ đó tiến hành tính toán, phân tích sẽ mang lại hiệu quả hết sức lý thú. Chẳng hạn, khi giữ nguyên số liệu ban đầu thì biến đang xét có tính bất đối xứng lớn, nhưng nếu ta lấy lôgarit tất cả các giá trị số liệu để nhận được chuỗi số liệu mới thì chuỗi này không những thoả nãm tính đối xứng mà còn tuân theo luật chuẩn. Thông thường trong khí tượng, khí hậu người ta sử dụng các phép biến đổi sau đây.

7.3.1 Phép biến đổi luỹ thừa

Phép biến đổi luỹ thừa thường được áp dụng cho những chuỗi số liệu bất đối xứng, nhận giá trị dương. Ký hiệu số liệu ban đầu là x, chuỗi sẽ được biến đổi theo một trong các dạng thức:

$$y = \begin{cases} x^{\lambda} & \lambda > 0 \\ \ln(x) & \lambda = 0 \\ -x^{\lambda} & \lambda < 0 \end{cases} \qquad (7.3.1)$$

$$y = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases} \qquad (7.3.2)$$

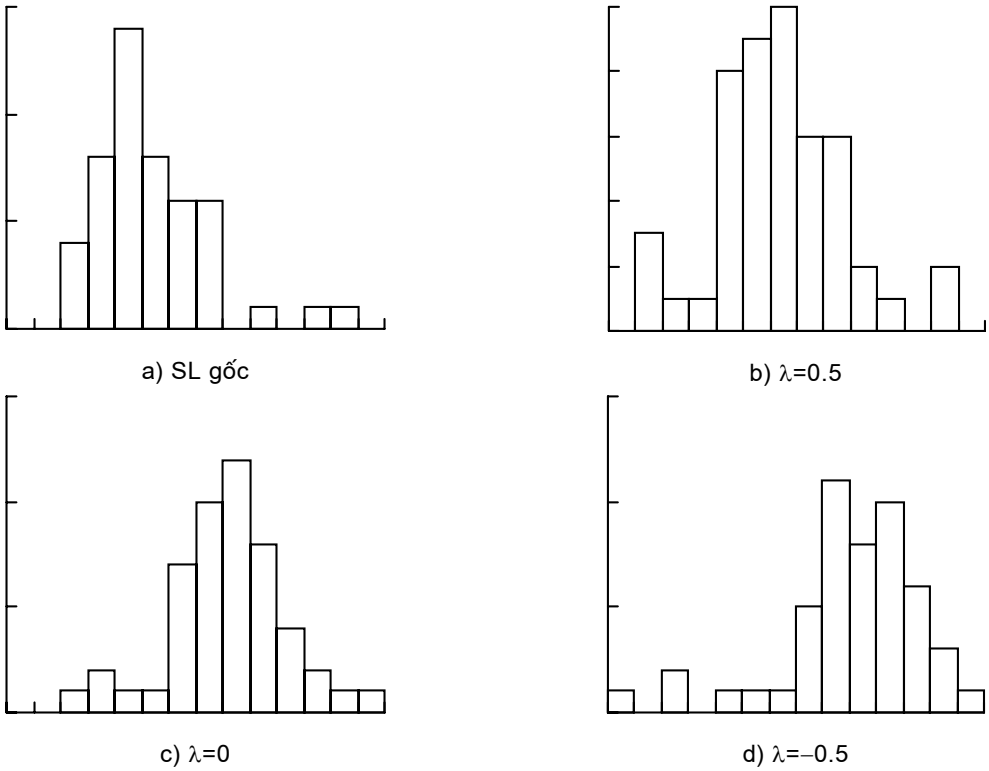
trong đó λ là một tham số được chọn tuỳ ý sao cho chuỗi đã biến đổi trở nên phù hợp hơn theo nghĩa nào đó.

Ví dụ 7.3 Từ chuỗi số liệu lượng mưa tháng 1 trong thời gian 50 năm của trạm A, sử dụng phép biến đổi (7.3.2) với các giá trị λ khác nhau ta nhận được kết quả trình bày trong bảng 7.1. Từ đó ta tính được độ bất đối xứng ứng với từng chuỗi:

	SL gốc	λ=1	λ=0.5	λ=0	λ=−0.5
Độ bất đối xứng	1.83	1.83	0.84	−0.18	−1.20

Rõ ràng sau khi thực hiện phép biến đổi tính bất đối xứng của chuỗi thay đổi rất đáng kể. Với giá trị λ=1, chuỗi mới chỉ khác chuỗi ban đầu một hằng số cộng (y=x−1), do đó tính bất đối xứng vẫn được bảo toàn. Khi λ==0.5, so với chuỗi ban đầu tính bất đối xứng đã giảm đi nhưng vẫn còn lệch phải. Nếu λ giảm xuống đến 0, bất đối xứng của chuỗi đã biến đổi từ lệch phải sang lệch trái. Nếu λ càng giảm

tính lệch trái càng tăng. Trong trường hợp trên, độ bất đối xứng nhỏ nhất khi λ=0. Điều này còn được thể hiện rõ trên hình 7.7.



Hình 7.7 Phân bố tần suất chuỗi lượng mưa trạm A qua các phép biến đổi

7.3.2 Biến đổi qui tâm và chuẩn hoá số liệu

Như đã nói trên đây, giả thiết về tình dừng của chuỗi có ý nghĩa rất quan trọng khi sử dụng công cụ thống kê nghiên cứu chuỗi thời gian. Tuy nhiên, hầu hết các quá trình khí quyển hoặc không thoả mãn tính dừng hoặc thoả mãn với mức độ yếu ớt. Với mục đích làm "tăng" tính dừng của quá trình người ta thường thực hiện phép biến đổi qui tâm và chuẩn hoá chuỗi. Qua phép biến đổi qui tâm chuỗi trở thành có trung bình bằng 0, còn phép chuẩn hoá làm cho chuỗi vừa có trung bình bằng 0 vừa có phương sai bằng đơn vị. Ký hiệu chuỗi qui tâm bởi x' còn chuỗi chuẩn hoá bởi z, ta có:

$$x' = x - \bar{x} \tag{7.3.3}$$

$$z = \frac{x - \bar{x}}{s_x} = \frac{x'}{s_x} \tag{7.3.4}$$

trong đó $\bar{x}$ và s_x tương ứng là trung bình và độ lệch chuẩn của chuỗi.

Như vậy, phép biến đổi qui tâm không làm thay đổi thứ nguyên của chuỗi trong khi phép chuẩn hoá biến chuỗi trở thành vô thứ nguyên.

7.3.3 Lọc chuỗi bằng phương pháp trung bình trượt

Phương pháp trung bình trượt là một trong những phương pháp được ứng dụng phổ biến trong khí hậu. Mục đích của phương pháp là loại trừ vai trò của tính ngẫu nhiên trong chuỗi, loại trừ ảnh hưởng của những chu kỳ ngắn và tạo cơ sở để phân tích xu thế và dao động có chu kỳ dài.

Có thể hiểu phương pháp trung bình trượt như là một phép biến đổi tuyến tính, biến chuỗi số liệu ban đầu $\{x_t, t=1..n\}$ thành chuỗi mới, trong đó các dao động ngẫu nhiên và chu kỳ ngắn đã được khử bỏ. Bởi vậy cũng có thể xem phương pháp trung bình trượt như là một toán tử lọc mà sau khi tác dụng nó lên chuỗi ban đầu ta được một chuỗi mới.

Giả sử có chuỗi số liệu ban đầu $\{x_t, t=1..n\}$. Với một trị số *m* nguyên dương xác định (thông thường *m* lẻ) ta có công thức biến đổi sau, được gọi là trung bình trượt với bước trượt m:

$$y_i = \frac{1}{m}\sum_{t=i}^{m+i-1} x_t \text{ , } (i=1,2,...n-m+1) \qquad (7.3.5)$$

Hay: $y_1 = \frac{1}{m}\sum_{t=1}^{m} x_t$, $y_2 = \frac{1}{m}\sum_{t=2}^{m+1} x_t$, $y_3 = \frac{1}{m}\sum_{t=3}^{m+2} x_t$,..., $y_{n-m+1} = \frac{1}{m}\sum_{t=n-m+1}^{n} x_t$

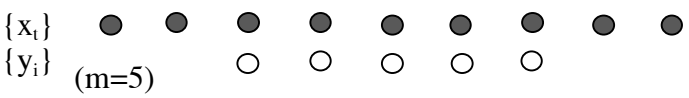
Như vậy mỗi thành phần của chuỗi mới $\{y_i\}$ là trung bình cộng của *m* thành phần $x_i,...,x_{m+i-1}$ của chuỗi ban đầu $\{x_t\}$. Thành phần thứ *i* của chuỗi mới $\{y_i\}$ không tiêu biểu cho thời gian t=i mà tiêu biểu cho cả khoảng thời gian từ t=i đến t=i+m−1. Hay nói cách khác, thành phần thứ *i* của chuỗi $\{y_i\}$ tiêu biểu cho thời gian t=(m+1)/2−1+i:

$$\{x_t, t=1..n\} \longrightarrow \{y_{(t)}, t=(m+1)/2..(n-(m+1)/2-1)\}$$

Chẳng hạn,

y_1	tương ứng với	$y_{((m+1)/2)}$
y_2	tương ứng với	$y_{((m+1)/2+1)}$
...		
y_{n-m+1}	tương ứng với	$y_{(n-(m+1)/2-1)}$

Tức là so với chuỗi $\{x_t\}$ số thành phần của chuỗi $\{y_i\}$ bị giảm đi (m−1)/2 thành phần đầu và (m−1)/2 thành phần cuối. Nếu chuỗi $\{x_t\}$ có *n* thành phần thì chuỗi $\{y_i\}$ có (n−m+1) thành phần. Trên hình 7.8 minh họa sơ đồ các thành phần của hai chuỗi trước và sau khi thực hiện phép trượt. Rõ ràng, khi chọn m=5 thì số thành phần bị mất đi sau khi trượt là m−1=4.



Hình 7.8 Sơ đồ trung bình trượt

Bảng 7.1 Số liệu lượng mưa trạm A trước và sau khi biến đổi

TT	SL gốc	λ=1	λ=0.5	λ=0	λ=–0.5	TT	SL gốc	λ=1	λ=0.5	λ=0	λ=–0.5
1	11.2	10.20	4.69	2.42	1.40	26	43.7	42.70	11.22	3.78	1.70
2	13.2	12.20	5.27	2.58	1.45	27	44.5	43.50	11.34	3.80	1.70
3	13.7	12.70	5.40	2.62	1.46	28	44.7	43.70	11.37	3.80	1.70
4	18.3	17.30	6.56	2.91	1.53	29	46.7	45.70	11.67	3.84	1.71
5	22.1	21.10	7.40	3.10	1.57	30	47.8	46.80	11.83	3.87	1.71
6	26.2	25.20	8.24	3.27	1.61	31	50.3	49.30	12.18	3.92	1.72
7	28.2	27.20	8.62	3.34	1.62	32	50.8	49.80	12.25	3.93	1.72
8	28.4	27.40	8.66	3.35	1.62	33	52.8	51.80	12.53	3.97	1.72
9	28.7	27.70	8.71	3.36	1.63	34	54.1	53.10	12.71	3.99	1.73
10	29.5	28.50	8.86	3.38	1.63	35	55.1	54.10	12.85	4.01	1.73
11	30.0	29.00	8.95	3.40	1.63	36	57.7	56.70	13.19	4.06	1.74
12	33.0	32.00	9.49	3.50	1.65	37	60.5	59.50	13.56	4.10	1.74
13	33.3	32.30	9.54	3.51	1.65	38	62.0	61.00	13.75	4.13	1.75
14	34.3	33.30	9.71	3.54	1.66	39	63.5	62.50	13.94	4.15	1.75
15	34.3	33.30	9.71	3.54	1.66	40	64.3	63.30	14.04	4.16	1.75
16	34.5	33.50	9.75	3.54	1.66	41	68.3	67.30	14.53	4.22	1.76
17	34.5	33.50	9.75	3.54	1.66	42	69.6	68.60	14.69	4.24	1.76
18	34.5	33.50	9.75	3.54	1.66	43	71.6	70.60	14.92	4.27	1.76
19	35.3	34.30	9.88	3.56	1.66	44	71.6	70.60	14.92	4.27	1.76
20	36.6	35.60	10.10	3.60	1.67	45	74.7	73.70	15.29	4.31	1.77
21	37.1	36.10	10.18	3.61	1.67	46	76.2	75.20	15.46	4.33	1.77
22	38.4	37.40	10.39	3.65	1.68	47	93.0	92.00	17.29	4.53	1.79
23	42.9	41.90	11.10	3.76	1.69	48	115.6	114.60	19.50	4.75	1.81
24	42.9	41.90	11.10	3.76	1.69	49	124.5	123.50	20.32	4.82	1.82
25	43.7	42.70	11.22	3.78	1.70	50	161.8	160.80	23.44	5.09	1.84

Tính chất của trung bình trượt:

Giả sử chuỗi $\{x_t\}$ có chu kỳ là p, khi đó ta có thể viết:

$$x_t \equiv x = x(t) = A\cos\frac{2\pi}{p}t \qquad (7.3.6)$$

trong đó A là biên độ dao động ngẫu nhiên ứng với chu kỳ p. Từ (7.3.6) các thành phần của chuỗi $\{x_t\}$ có thể được biểu diễn bởi:

$$x_1 = A\cos\frac{2\pi}{p}1,\ x_2 = A\cos\frac{2\pi}{p}2,\ldots,\ x_m = A\cos\frac{2\pi}{p}m \qquad (7.3.7)$$

Mặt khác, đối với chuỗi đã trượt $\{y_i\}$ ta cũng có:

$$y_1 = \frac{1}{m}\sum_{t=1}^{m} x_t = \frac{1}{m}\sum_{t=1}^{m} A\cos\frac{2\pi}{p}t \qquad (7.3.8)$$

Sử dụng công thức Euler $\sum_{t=1}^{m}\cos\varphi t = \frac{\sin\frac{m}{2}\varphi\cos\frac{m+1}{2}\varphi}{\sin\frac{\varphi}{2}}$ cho (7.3.8) ta nhận được:

$$y_1 \equiv y_{((m+1)/2)} = \frac{A}{m}\sum_{t=1}^{m}\cos\frac{2\pi}{p}t = \frac{A}{m}\left(\frac{\sin\frac{m}{2}\frac{2\pi}{p}\cos\frac{m+1}{2}\frac{2\pi}{p}}{\sin\frac{2\pi}{2p}}\right) =$$

$$= \frac{A}{m}\frac{\sin\frac{\pi}{p}m}{\sin\frac{\pi}{p}}\cos\frac{\pi}{p}(m+1) = A_1\cos\frac{\pi}{p}(m+1) \qquad (7.3.9)$$

với $A_1 = \frac{A}{m}\frac{\sin\frac{\pi}{p}m}{\sin\frac{\pi}{p}}$ là biên độ dao động ngẫu nhiên.

Từ (7.3.7) ta có thành phần thứ (m+1)/2 của chuỗi $\{x_t\}$:

$$x_{(m+1)/2} = A\cos\frac{2\pi}{p}\left(\frac{m+1}{2}\right) = A\cos\frac{\pi}{p}(m+1) \qquad (7.3.10)$$

So sánh (7.3.9) và (7.3.10) ta thấy sau khi thực hiện phép trượt, biên độ của $y_{((m+1)/2)}$ giảm đi chỉ còn bằng $k = \frac{A_1}{A}$ lần biên độ của $x_{(m+1)/2}$:

$$k = \frac{A_1}{A} = \frac{\frac{A}{m}\frac{\sin\frac{\pi}{p}m}{\sin\frac{\pi}{p}}}{A} = m\frac{\sin\frac{\pi}{p}m}{\sin\frac{\pi}{p}} \qquad (7.3.11)$$

Như vậy, nếu p=m, m/2, m/3,... thì $\frac{\pi}{p}m = \pi, 2\pi, 3\pi,...$ và $\sin\frac{\pi}{p}m=0$, hay k=0. Từ đó suy ra rằng với bước trượt *m*, biên độ của những dao động có chu kỳ bằng m, m/2, m/3,... của chuỗi ban đầu sẽ giảm đến 0. Điều đó có nghĩa là nếu thực hiện phép trượt bước *m* ta sẽ biến chuỗi ban đầu thành chuỗi mới trong đó các dao động có chu kỳ bằng m, m/2, m/3,... (các chu kỳ nhận *m* làm bội số) đã được khử bỏ, chuỗi đã trượt trở nên trơn tru, dễ phân tích hơn.

Trong tính toán thực hành việc chọn *m* hoàn toàn tuỳ thuộc vào mục đích của bài toán. Tuy vậy ta cố gắng chọn nhiều trị số *m* khác nhau và so sánh các kết quả nhận được để rút ra kết luận. Mặt khác cũng cần lưu ý rằng, sau khi trượt, độ dài của chuỗi mới bị mất đi (m–1) thành phần so với chuỗi ban đầu. Do vậy nếu chọn *m* quá lớn sẽ làm cho số thành phần bị mất đi quá nhiều.

Chẳng hạn, để phân tích những biến đổi có chu kỳ của chuỗi số liệu lượng mưa tháng, nếu cần quan tâm đến những chu kỳ trên một năm ta có thể chọn m=13. Trong trường hợp này những dao động ngẫu nhiên có các chu kỳ 13 tháng, 13/2=6.5 tháng,... sẽ được khử bỏ. Sau khi thực hiện phép trượt ta được chuỗi mới thể hiện những dao động rõ nét hơn.

Hình 7.9 dẫn ra ví dụ về làm trơn chuỗi lượng mưa năm của một trạm bằng trung bình trượt với bước trượt m=5. Từ hình vẽ có thể nhận thấy sau khi lọc chuỗi đã được làm trơn một cách đáng kể. Những dao động ngẫu nhiên đã được loại bỏ bớt và qui luật dao động dài năm đươc thể hiện khá rõ nét.

7.3.4 Lọc chuỗi bằng phép lọc có trọng lượng

Lọc có trọng lượng là thực hiện phép biến đổi chuỗi ban đầu $\{x_i\}$ về chuỗi mới $\{y_i\}$ bằng cách tác dụng một toán tử tuyến tính – tổng có trọng lượng, lên chuỗi đã cho:

$$y_i = \sum_{k=1}^{m} \omega_k x_{i+k-1}, (i = 1,2,..n-m+1) \qquad (7.3.12)$$

Hay

$$y_1 = \sum_{k=1}^{m} \omega_k x_k, y_2 = \sum_{k=1}^{m} \omega_k x_{1+k}$$

$$y_3 = \sum_{k=1}^{m} \omega_k x_{2+k}, ..., y_{n-m+1} = \sum_{k=1}^{m} \omega_k x_{n-m+k}$$

trong đó ω_k, k=1..m, là các trọng số của toán tử lọc. Các trọng số này phải thoả mãn hệ thức:

$$\sum_{k=1}^{m} \omega_k = 1 \qquad (7.3.13)$$

Ta thấy mỗi thành phần của chuỗi mới $\{y_i\}$ bằng trung bình có trọng lượng của *m* thành phần $x_i,...,x_{m+i-1}$ của chuỗi ban đầu $\{x_i\}$. Tương tự như trung bình trượt, thành phần thứ *i* của chuỗi mới $\{y_i\}$ không tiêu biểu cho thời gian t=i mà tiêu biểu cho cả khoảng thời gian từ t=i đến t=i+m–1. Hay nói cách khác, thành phần thứ *i* của chuỗi $\{y_i\}$ tiêu biểu cho thời gian t=(m+1)/2–1+i.

$$\{x_t,\ t=1..n\} \longrightarrow \{y_{(t)},\ t=(m+1)/2...(n-(m+1)/2-1)\}$$

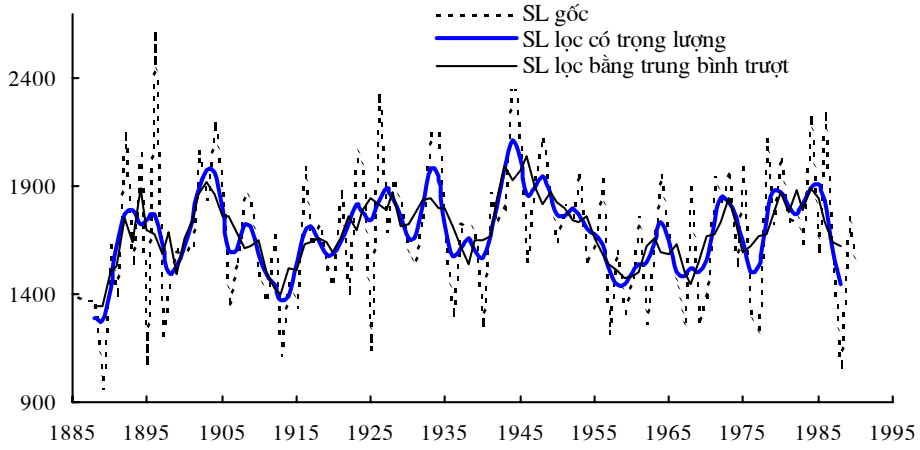
So sánh (7.3.5) và (7.3.12) ta thấy trung bình trượt là một trường hợp riêng của phép lọc có trọng lượng khi cho các trọng số ω_k bằng nhau và bằng $\frac{1}{m}$. Như vậy, sự khác nhau giữa phương pháp lọc chuỗi theo công thức (7.3.12) và phương pháp trung bình trượt là ở chỗ, nếu trong (7.3.12) những thành phần càng cách xa trị số lọc (i) sẽ có trọng lượng càng nhỏ, thì ở phương pháp trung bình trượt các trọng lượng lọc được lấy bằng nhau đối với mọi thành phần tham gia lọc.

Điều quan trọng ở đây là các trọng số lọc ω_k, k=1..m, cần dược chọn sao cho thích hợp với bản chất của quá trình đang xét. Thông thường người ta chọn số trọng số *m* lẻ và giá trị của chúng đối xứng nhau qua $\omega_{(m+1)/2}$. Ví dụ, một trong những toán tử lọc dạng này đã được tổ chức Khí tượng thế giới (WMO) công bố và nó đã được sử dụng để khảo sát các chuỗi lượng mưa là:

$$\omega_k=\{0.06,\ 0.25,\ 0.38,\ 0.25,\ 0.06\} \qquad (7.3.14)$$

Hình 7.9 minh hoạ kết quả áp dụng toán tử lọc (7.3.14) cho chuỗi lượng mưa đã nêu ở mục trên.

Từ đó ta thấy, về cơ bản kết quả của hai phương pháp lọc tương tự nhau, những dao động dài năm đều được thể hiện ở cả hai chuỗi đã lọc. Tuy vậy, nếu xem xét chi tiết cũng có thể phân biệt được biên độ dao động của chuỗi lọc bằng phép lọc có trọng lượng nhỏ hơn chút ít so với chuỗi lọc bằng trung bình trượt.



Hình 7.9 Chuỗi lượng mưa năm trước và sau khi lọc

7.4 Sử dụng hàm tự tương quan xác định chu kỳ dao động

Nghiên cứu tính dao động có chu kỳ của chuỗi bằng hàm tự tương quan – tức hàm tương quan chuẩn hoá – dựa trên giả thiết cho rằng, các thành phần của chuỗi thời gian $\{x_t, t=1..n\}$ là những trị số quan trắc của thể hiện x(t) tại *n* lát cắt $t_1, t_2,..., t_n$ của quá trình ngẫu nhiên dừng X(t). Thực chất của phương pháp là xem xét sự biến thiên của hàm tương quan chuẩn hoá tính được từ chuỗi đã cho. Nếu chuỗi có chu kỳ bằng *k* (đơn vị thời gian) thì giá trị của hệ số tương quan giữa hai lát cắt t_j và t_{j+k} sẽ gần bằng 1 hoặc khá lớn (Chú ý rằng đối với các chuỗi số liệu khí hậu khoảng thời gian giữa hai lát cắt liên tiếp thường là một năm).

Giả sử xét chuỗi $\{x_t, t=1..n\}$. Khi đó hàm tương quan chuẩn hoá (hay hàm tự tương quan) $r_x(k)=r_x(t_{j+k}-t_j)$ được xác định bởi:

$$r_x(k) = \frac{1}{n-k}\sum_{t=1}^{n-k}\frac{\left(x_t-\overline{x_o}\right)\left(x_{t+k}-\overline{x_k}\right)}{s_o s_k} = \frac{1}{n-k}\sum_{t=1}^{n-k}\frac{x_t x_{t+k}-\overline{x_o}\,\overline{x_k}}{s_o s_k} \qquad (7.4.1)$$

trong đó:

$$\overline{x_o}=\frac{1}{n-k}\sum_{t=1}^{n-k}x_t\,, \quad \overline{x_k}=\frac{1}{n-k}\sum_{t=k+1}^{n}x_t \qquad (7.4.2)$$

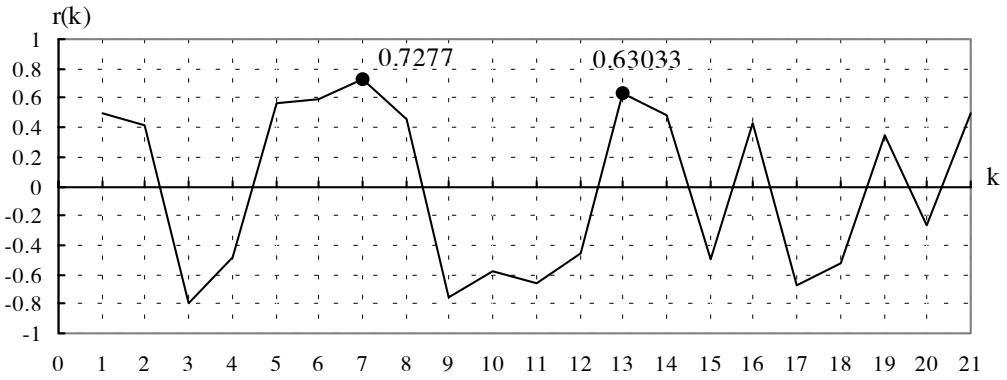
$$s_o=\sqrt{\frac{1}{n-k}\sum_{t=1}^{n-k}\left(x_t-\overline{x_o}\right)^2}\,,\ s_k=\sqrt{\frac{1}{n-k}\sum_{t=k+1}^{n}\left(x_t-\overline{x_k}\right)^2} \qquad (7.4.3)$$

k = 1,2,...,m (đơn vị thời gian).

Để dễ dàng nhận biết được các chu kỳ, thông thường sau khi tính, người ta biểu diễn hàm tự tương quan lên hệ trục toạ độ với trục tung là $r_x(k)$ còn trục hoành là k. Các giá trị *k* ứng với $r_x(k)$ khá lớn hoặc gần bằng 1 sẽ được xem là các chu kỳ dao động của chuỗi.

Hình 7.10 dẫn ra đồ thị hàm tự tương quan của chuỗi số liệu nhiệt độ trung bình năm của một trạm như một ví dụ về khảo sát tính dao động của chuỗi. Ta thấy trị số hàm tự tương quan biến đổi theo *k* khá rõ. Xu thế $r_x(k)$ giảm khi *k* tăng thể hiện tính dao động tắt dần của hàm tự tương quan. Với trị số $r_x(k)>0.6$ có thể xem các giá trị k=7 và k=13 tương ứng với những chu kỳ dao động của chuỗi.

Từ (7.4.1) có thể thấy rằng, nếu k càng lớn thì n–k càng giảm, tức dung lượng mẫu trong công thức tính các hệ số tương quan càng bé. Khi k quá lớn so với dung lượng mẫu n, giá trị tính được $r_x(k)$ sẽ không đảm bảo độ ổn định thống kê. Bởi vậy, số lượng giá trị của hàm tự tương quan $r_x(k)$ không thể vượt quá một trị số k_{max} nào đó mà người ta gọi là điểm cắt (hay độ dịch chuyển cực đại) của hàm tự tương quan. Nói chung k_{max} phụ thuộc vào dung lượng mẫu *n*. Thông thường đối với các quá trình khí tượng thuỷ văn k_{max} được chọn trong khoảng n/10 đến n/4.



Hình 7.10 Hàm tự tương quan chuỗi số liệu nhiệt độ trung bình năm

7.5 Phương pháp phân tích điều hoà biểu diễn chuỗi thời gian

7.5.1 Khái niệm

Một trong những phương pháp phổ biến được áp dụng để phân tích sự biến đổi chu kỳ của các chuỗi số liệu khí tượng, khí hậu là phương pháp phân tích điều hoà. Phân tích điều hoà là biểu diễn những dao động biến đổi của chuỗi thời gian dưới dạng tổng các thành phần dao động điều hoà (dao động hình sin). Việc phân tích như vậy cho phép hiểu được bản chất vật lý của những dao động biến đổi thông thường. Nguyên lý cơ bản của phương pháp này dựa trên cơ sở xem biến khí quyển đang xét biến đổi liên tục theo thời gian, và chuỗi số liệu chính là giá trị của biến đo được tại n điểm hữu hạn, rời rạc. Giả thiết rằng khoảng cách thời gian giữa hai thành phần kế cận của chuỗi không đổi, bằng đơn vị thời gian, thì độ dài chuỗi n sẽ là chu kỳ dao động cơ bản của chuỗi.

Tuy nhiên, việc thực hiện bài toán này dẫn đến một số vấn đề nảy sinh. Đó là, đối số của các hàm lượng giác (sin và cosin) là góc (độ hoặc radian), trong khi chuỗi số liệu có thể được xem như là hàm của thời gian. Mặt khác, các hàm sin và cosin chỉ nhận giá trị trên đoạn [–1; 1], trong khi chuỗi thời gian thường dao động với những biên độ rất khác nhau.

Để giải quyết vấn đề thứ nhất ta xem độ dài chuỗi n phủ đầy một chu kỳ cơ bản của hàm sin, tức là ta sẽ thực hiện phép biến đổi đối số thời gian thành đối số góc theo công thức sau:

$$t \longrightarrow \frac{2\pi}{n} t \qquad (7.5.1)$$

Hay

$$t \longrightarrow \frac{360^o}{n} t \qquad (7.5.1')$$

Như vậy, khi t biến đổi từ 0 đến n thì góc ($\frac{2\pi}{n}t$) biến đổi từ 0 đến 2π (hay 360°). Đại lượng

$$\omega_1 = \frac{2\pi}{n} \tag{7.5.2}$$

được gọi là tần số cơ bản, có thứ nguyên bằng Radian/đơn vị thời gian. Nó là tỷ số giữa chu kỳ cơ bản của hàm sin và độ dài chuỗi n. Chỉ số "1" trong (7.5.2) cũng có nghĩa là sóng có tần số ω_1 thực hiện một chu kỳ dao động mất một khoảng thời gian bằng n đơn vị.

Vấn đề thứ hai được giải quyết một cách đơn giản bằng việc nhân thêm một hệ số tỷ lệ C_1 vào thành phần dao động và cộng thêm một hằng số cộng là giá trị trung bình của chuỗi sao cho có thể biển diễn chuỗi dưới dạng:

$$x_t = \bar{x} + C_1 \cos\left(\frac{2\pi}{n}t - \varphi_1\right) \tag{7.5.3}$$

trong đó, hệ số C_1 được gọi là biên độ của dao động điều hoà cơ bản và φ_1 được gọi là góc pha hay pha dao động.

Từ (7.5.3) suy ra rằng x_t đạt cực đại khi $\cos\left(\frac{2\pi}{n}t - \varphi_1\right) = 1$ hay $\frac{2\pi}{n}t = \varphi_1$.

7.5.2 Ước lượng biên độ và pha của dao động điều hoà đơn

Để biểu diễn chuỗi số liệu theo (7.5.3) ta cần phải xác định được hai tham số C_1 và φ_1. Hạng thứ hai trong (7.5.3) có thể được viết lại dưới dạng:

$$C_1 \cos\left(\frac{2\pi}{n}t - \varphi_1\right) = A_1 \cos\frac{2\pi}{n}t + B_1 \sin\frac{2\pi}{n}t \tag{7.5.4}$$

trong đó
$$A_1 = C_1\cos(\varphi_1),\ B_1 = C_1\sin(\varphi_1) \tag{7.5.5}$$

Từ đó có thể xác định được hệ số C_1 và góc pha φ_1:

$$C_1 = \sqrt{A_1^2 + B_1^2} \tag{7.5.6}$$

$$\varphi_1 = \begin{cases} \text{arctg}\left(\frac{B_1}{A_1}\right) & \text{nếu } A_1 > 0 \\ \text{arctg}\left(\frac{B_1}{A_1}\right) \pm \pi & \text{nếu } A_1 < 0 \\ \frac{\pi}{2} & \text{nếu } A_1 = 0 \end{cases} \tag{7.5.7}$$

Vấn đề còn lại là phải xác định được A_1 và B_1. Kết hợp (7.5.3) và (7.5.4) ta có:

$$x_t = \bar{x} + A_1\cos\frac{2\pi}{n}t + B_1\sin\frac{2\pi}{n}t \qquad (7.5.8)$$

Nếu tuyến tính hoá các thành phần sin và cos trong (7.5.8) bằng cách đặt biến mới $u=\cos\frac{2\pi}{n}t$, $v=\sin\frac{2\pi}{n}t$ ta có thể đưa (7.5.8) về dạng phương trình hồi qui tuyến tính quen thuộc $x=a_o + a_1u + a_2v$. Và từ đó dễ dàng xác định được: $A_1=a_1$, $B_1=a_2$, còn hệ số tự do a_o chính là giá trị trung bình $\bar{x}$. Tuỳ theo giá trị nhận được của A_1 và B_1 mà khi tính φ_1 theo (7.5.7), trường hợp thứ hai ($A_1<0$) sẽ chọn dấu (+) hay dấu (–) sao cho thoả mãn điều kiện $0<\varphi_1<2\pi$.

Trong thực tế, nếu khoảng cách thời gian giữa các thành phần kế cận của chuỗi đều nhau ta có thể tính các hệ số A_1 và B_1 theo các công thức sau:

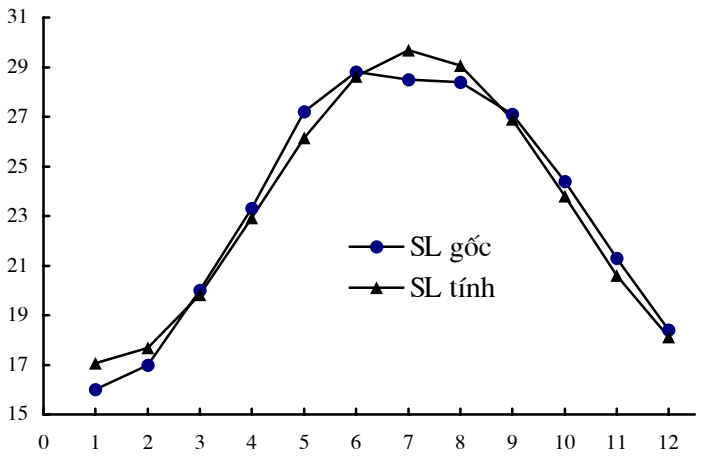
$$A_1 = \frac{2}{n}\sum_{t=1}^{n} x_t \cos\frac{2\pi}{n}t\ ;\ B_1 = \frac{2}{n}\sum_{t=1}^{n} x_t \sin\frac{2\pi}{n}t \qquad (7.5.9)$$

Ví dụ 7.5.1 Bảng 7.2 dẫn ra số liệu nhiệt độ trung bình tháng nhiều năm của một trạm và những kết quả tính trung gian để xác định các hệ số theo các công thức trong (7.5.9).

Từ bảng 7.2 ta nhận được n=12 (tháng), $\bar{x}$=23.37 (°C), A_1= –31.485/6 = –5.25, B_1=–21.225/6=–3.54. Do đó C_1=6.329 và φ_1=3.735. Ta cũng có thể nhận được kết quả tương tự bằng phương pháp hồi qui tuyến tính khi xem cột thứ hai là biến phụ thuộc và hai cột tiếp theo là các biến độc lập. Sử dụng kết quả tính này để biểu diễn lại chuỗi số liệu ban đầu theo (7.5.3) ta nhận được cột cuối cùng của bảng.

Bảng 7.2 Kết quả tính trung gian cho nhiệt độ trung bình tháng nhiều năm

t	x_t	$\cos\frac{2\pi}{n}t$	$\sin\frac{2\pi}{n}t$	$x_t\cos\frac{2\pi}{n}t$	$x_t\sin\frac{2\pi}{n}t$	$\hat{x}_t$
1	16.0	0.866	0.500	13.856	8.000	17.1
2	17.0	0.500	0.866	8.500	14.722	17.7
3	20.0	0.000	1.000	0.000	20.000	19.8
4	23.3	–0.500	0.866	–11.650	20.178	22.9
5	27.2	–0.866	0.500	–23.556	13.600	26.1
6	28.8	–1.000	0.000	–28.800	0.000	28.6
7	28.5	–0.866	–0.500	–24.682	–14.250	29.7
8	28.4	–0.500	–0.866	–14.200	–24.595	29.1
9	27.1	0.000	–1.000	0.000	–27.100	26.9
10	24.4	0.500	–0.866	12.200	–21.131	23.8
11	21.3	0.866	–0.500	18.446	–10.650	20.6
12	18.4	1.000	0.000	18.400	0.000	18.1
Tổng		0.000	0.000	–31.485	–21.225	



Hình 7.11 Kết quả biểu diễn chuỗi số liệu trung bình tháng bằng hàm điều hoà đơn

Hình 7.11 dẫn ra đồ thị của số liệu gốc (cột 2) và số liệu tính toán xấp xỉ theo (7.5.3). Từ đó nhận thấy rằng, mặc dù có sự khác biệt giữa số liệu thực và số liệu tính toán, song mức độ sai lệch không đáng kể.

7.5.3 Phân tích điều hoà xác định chu kỳ dao động

Phân tích điều hoà đơn trên đây cho phép biểu diễn chuỗi số liệu chỉ có một chu kỳ dao động. Nhưng nhiều bài toán trong thực tế yêu cầu xác định được những chu kỳ dao động khác còn tiềm ẩn trong chuỗi mà bằng phương pháp khảo sát thông thường ta không thể phát hiện được. Trong trường hợp này thay cho (7.5.3) ta sẽ biểu diễn chuỗi dưới dạng tổng của nhiều dao động điều hoà với các biên độ và pha khác nhau:

$$x_t = \bar{x} + \sum_{k=1}^{n/2}\left[C_k \cos\left(\frac{2\pi k}{n}t - \varphi_k\right)\right] = \bar{x} + \sum_{k=1}^{n/2}\left[A_k \cos\left(\frac{2\pi k}{n}t\right) + B_k \sin\left(\frac{2\pi k}{n}t\right)\right] \qquad (7.5.10)$$

trong đó $\omega_k = \frac{2\pi k}{n}$ là các tần số dao động, bằng bội số nguyên của tần số cơ bản ω_1.

Như vậy, chuỗi x_t được xem là sự chồng chất của n/2 dao động với các tần số khác nhau. Dao động ứng với k=1 là tần số $\omega_1=2\pi/n$ có chu kỳ bằng độ dài chuỗi, ứng với k=2 là tần số $\omega_2=4\pi/n$ có chu kỳ bằng 1/2 chuỗi,...

Tương tự như trên, các hệ số A_k và B_k trong (7.5.10) có thể nhận được bằng phương pháp hồi qui tuyến tính thông qua việc đặt biến phụ $u_1=\cos\frac{2\pi}{n}t$, $u_2=\sin\frac{2\pi}{n}t$, $u_3=\cos\frac{2\pi 2}{n}t$, $u_4=\sin\frac{2\pi 2}{n}t$, v.v. Trong trường hợp các thành phần của chuỗi cách đều nhau (khoảng cách thời gian đều nhau) ta có thể sử dụng các công thức sau đây để tính:

$$A_k = \frac{2}{n}\sum_{t=1}^{n} x_t \cos\left(\frac{2\pi k}{n}t\right),\ B_k = \frac{2}{n}\sum_{t=1}^{n} x_t \sin\left(\frac{2\pi k}{n}t\right) \qquad (7.5.11)$$

$$(k=1,2,\ldots,(n/2)-1)$$

Và
$$A_{n/2} = \begin{cases} \frac{1}{2}.\frac{2}{n}\sum_{t=1}^{n} x_t \cos\left[\frac{2\pi(n/2)t}{n}\right] = \frac{1}{n}\sum_{t=1}^{n} x_t \cos(\pi t) & \text{khi n chẵn} \\ 0 & \text{khi n lẻ} \end{cases}$$

$$B_{n/2} = 0 \qquad (7.5.11')$$

Từ đó ta nhận được các biên độ và pha dao động:

$$C_k = \sqrt{A_k^2 + B_k^2} \qquad (7.5.12)$$

$$\varphi_k = \begin{cases} \text{arctg}\left(\frac{B_k}{A_k}\right) & \text{nếu} A_k > 0 \\ \text{arctg}\left(\frac{B_k}{A_k}\right) \pm \pi & \text{nếu} A_k < 0 \\ \frac{\pi}{2} & \text{nếu} A_k = 0 \end{cases} \qquad (7.5.13)$$

Biểu diễn chuỗi thời gian x_t theo (7.5.10) được gọi là phép biến đổi Fourier rời rạc. Như vậy, *n* thành phần ban đầu của chuỗi có thể được biểu diễn bởi các hệ số C_k và φ_k. Vì các hệ số A_k và B_k đều là những hàm của tần số ω_k nên C_k và φ_k cũng là hàm của tần số ω_k. Tức là, thay cho việc xét chuỗi trên miền thời gian, phân tích điều hoà cho phép biểu diễn chuỗi trên miền tần số. Điều đó giúp ta tách được những đóng góp của các loại dao động khác nhau lên sự biến đổi của chuỗi.

Với độ dài chuỗi bằng *n* ta sẽ có n/2 (nếu *n* chẵn) hoặc (n−1)/2 (nếu *n* lẻ) bộ các giá trị C_k, φ_k và ω_k. Thông thường sau khi tính toán người ta biểu diễn chúng lên đồ thị với trục hoành là ω_k còn trục tung là C_k^2 hoặc φ_k. Nói chung trong thực tế người ta quan tâm nhiều đến sự biến đổi của C_k^2 theo ω_k và đồ thị của chúng được gọi là đồ thị phổ năng lượng hay đơn giản là phổ. Giá trị nhỏ nhất của ω_k (tần số thấp nhất) là $\omega_1=2\pi/n$ (tần số cơ bản) ứng với sóng hình sin thực hiện một chu kỳ bằng độ dài chuỗi *n*, và tần số cao nhất là $\omega_{n/2}=\pi$, được gọi là tần số Nyquist, ứng với sóng hình sin thực hiện một chu kỳ bằng hai khoảng thời gian giữa các thành phần của chuỗi và thực hiện n/2 chu kỳ bằng độ dài chuỗi. Tần số Nyquist phụ thuộc vào độ phân giải thời gian của chuỗi ban đầu x_t.

Tần số góc ω_k có thứ nguyên là radian/thời gian, nhưng trong ứng dụng thực hành người ta thường sử dụng khái niệm tần số dài:

$$f_k = \frac{k}{n} = \frac{\omega_k}{2\pi} \qquad (7.5.14)$$

Tần số f_k có thứ nguyên là 1/thời gian. Tương ứng với khoảng biến thiên của ω_k, f_k biến đổi từ tần số cơ bản $f_1 = \frac{1}{n}$ đến tần số Nyquist $f_{n/2} = \frac{1}{2}$. Trị số nghịch đảo của f_k được gọi là chu kỳ điều hoà:

$$\tau_k = \frac{1}{f_k} = \frac{n}{k} = \frac{2\pi}{\omega_k} \qquad (7.5.15)$$

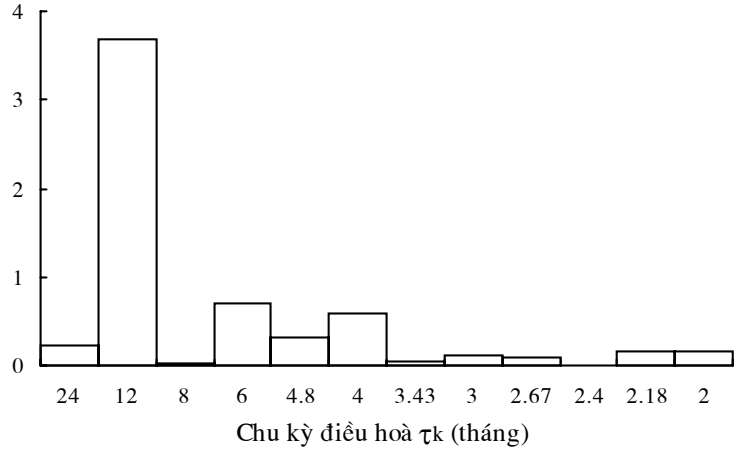
Chu kỳ τ_k là khoảng thời cần thiết để sóng có tần số ω_k thực hiện trọn ven một chu kỳ dao động.

Ví dụ 7.5.2 Bảng 7.3 dẫn ra chuỗi số liệu nhiệt độ trung bình tháng hai năm liên tục của một trạm và những kết quả tính toán theo (7.5.12) và (7.5.15). Số liệu ban đầu của chuỗi x_t ở cột thứ hai và thứ ba. Hai cột tiếp theo chứa chỉ số (k) và tần số (τ_k) điều hoà. Cột cuối cùng là bình phương biên độ của các dao động điều hoà mà ta sẽ gọi là phổ.

Bảng 7.3 Phân tích điều hoà chuỗi nhiệt độ trung bình tháng

Tháng	1995	1996	k	τ_k	C_k^2
1	17.0	17.9	1	24.00	0.241
2	16.1	15.1	2	12.00	38.533
3	17.5	20.6	3	8.00	0.029
4	22.5	23.9	4	6.00	1.003
5	27.7	27.5	5	4.80	0.380
6	29.2	27.7	6	4.00	0.781
7	28.4	28.2	7	3.43	0.040
8	28.5	27.9	8	3.00	0.111
9	27.0	28.0	9	2.67	0.087
10	24.0	23.0	10	2.40	0.009
11	22.1	20.0	11	2.18	0.169
12	19.7	19.2	12	2.00	0.181

Dĩ nhiên, không cần phân tích ta cũng có thể khẳng định rằng chuỗi sẽ gồm các chu kỳ năm (12 tháng), nửa năm (6 tháng), v.v. Với độ dài chuỗi n = 12 x 2 = 24 ta có k=24/2=12 và τ_k nhận giá trị lớn nhất bằng τ_{12}=24/1=24 (tháng), nhỏ nhất bằng τ_1=24/12=2 (tháng). Giá trị phổ đạt cực đại tại các chu kỳ τ_2=12 (tháng) và τ_4=6 (tháng),... nói lên rằng, đóng góp vào sự dao động biến đổi của chuỗi là những sóng có chu kỳ 12 tháng, 6 tháng,... trong đó mức đóng góp của sóng có chu kỳ 12 tháng chiếm một tỷ trọng lớn gấp nhiều lần so với các sóng khác. Kết quả phân tích này cũng được minh hoạ trên hình 7.12, trong đó trục hoành biểu thị các chu kỳ điều hoà τ_k còn trục tung biểu thị giá trị phổ đã được biến đổi thành thang độ đo lôgarit theo công thức: $\left(C_k^2\right)' = \mathrm{Ln}\left(C_k^2 + 1\right)$



Hình 7.12 Biểu diễn phổ chuỗi số liệu nhiệt độ trung bình tháng

7.5.4 Vài nét về phương pháp FFT (Fast Fourier Transforms)

Mặc dù việc tính toán các hệ số biến đổi Fourier rời rạc chuỗi thời gian theo (7.5.11) hết sức rõ ràng, cụ thể, chúng vẫn có những tồn tại nhất định, trong đó tồn tại lớn nhất là thuật toán làm giảm tốc độ tính một cách đáng kể. Cho mãi đến giữa những năm sáu mươi người ta mới tìm được một thuật toán cho phép đẩy nhanh tốc độ tính lên rất nhiều lần. Đó là thuật toán hay phương pháp biến đổi Fourier nhanh (FFT). Ngày nay phương pháp này đã được ứng dụng rộng rãi trong mọi lĩnh vực và nó đã được chương trình hoá trong nhiều phần mềm khác nhau. Tuỳ thuộc vào độ dài chuỗi, so với phương pháp tính đã trình bày ở mục trên, phương pháp FFT làm tăng tốc độ tính lên $n\log_2 n$ lần. Chẳng hạn, với n=100 tốc độ tính của FFT nhanh hơn $100\log_2(100)\approx 15$ lần, với n=10000, số lần nhanh hơn sẽ là $10000\log_2(10000)\approx 752$ lần.

Phương pháp FFT tính các hệ số Fourier trên cơ sở biểu diễn chuỗi thời gian dưới dạng:

$$x_t = \bar{x} + \sum_{k=1}^{n/2} H_k e^{i(2\pi k/n)t} \qquad (7.5.16)$$

trong đó H_k là các hệ số Fourier phức:

$$H_k = A_k + iB_k \qquad (7.5.17)$$

với A_k và B_k là phần thực và phần ảo của H_k và $i = \sqrt{-1}$.

7.6 Phổ của các quá trình liên tục

7.6.1 Mật độ phổ của quá trình ngẫu nhiên

Như đã biết, ta có thể sử dụng hàm tương quan chuẩn hoá để phân tích tính dao động có chu kỳ của chuỗi trên miền thời gian. Ta cũng có thể sử dụng phương pháp phân tích điều hoà biểu diễn chuỗi thời gian trên miền tần số. Tuy vậy, trên thực tế bằng các phương pháp này nhiều khi một số chu kỳ dao động không thể hiện rõ và do đó ta không thể phát hiện được. Sau đây ta sẽ khảo sát một phương pháp khác, đó là phổ phương sai. Phương pháp này dựa trên cơ sở thừa nhận tính dừng của quá trình ngẫu nhiên X(t) mà nội dung của nó có thể trình bày tóm tắt như sau.

Giả sử X(t) là một quá trình ngẫu nhiên dừng xác định trên đoạn [–T,T] có kỳ vọng toán học $m_x=0$. Hiển nhiên điều này có thể thực hiện được, bởi nếu $m_x\neq 0$ ta có thể xét quá trình qui tâm của nó. Và giả sử rằng ta có thể biểu diễn X(t) dưới dạng tổng vô hạn các dao động điều hoà với các tần số $\omega_k=\frac{\pi k}{T}$ và biên độ ngẫu nhiên X_k khác nhau:

$$X(t)=\sum_{k=-\infty}^{\infty} X_k e^{i\omega_k t} \tag{7.6.1}$$

Do giả thiết kỳ vọng $m_x=0$ suy ra $M[X_k]=0$, và hàm tương quan của quá trình ngẫu nhiên X(t) được viết dưới dạng:

$$R_x(t+\tau,t) = M[X(t+\tau)X^*(t)] \tag{7.6.2}$$

trong đó

$$X(t+\tau) = \sum_k X_k e^{i\omega_k (t+\tau)} \tag{7.6.3}$$

$$X^*(t)=\sum_l X_l^* e^{-i\omega_l t} \tag{7.6.4}$$

Từ đó:

$$R_x(t+\tau,t) = \left[\sum_k X_k e^{i\omega_k (t+\tau)} \sum_l X_l^* e^{-i\omega_l t}\right] =$$

$$= M\left\{\sum_k \sum_l X_k X_l^* e^{i[\omega_k (t+\tau)-\omega_l t]}\right\} = \sum_k \sum_l M\left[X_k X_l^*\right] e^{i[\omega_k (t+\tau)-\omega_l t]} \tag{7.6.5}$$

Từ điều kiện dừng của X(t) suy ra hàm tương quan chỉ phụ thuộc vào một đối số là hiệu giữa hai lát cắt, nên biểu thức (7.6.5) trở thành:

$$R_x(\tau)=\sum_k M\left[X_k X_k^*\right] e^{i\omega_k \tau} \tag{7.6.6}$$

Rõ ràng, $M[X_k, X_k^*]=D_k$ là phương sai của các biên độ ngẫu nhiên X_k. Từ đó ta nhận được:

$$R_x(\tau) = \sum_{k=-\infty}^{\infty} D_k e^{i\omega_k \tau} \tag{7.6.7}$$

Biểu thức này được gọi là biểu diễn hàm tương quan dưới dạng chuỗi Fourier. Cần chú ý rằng, do X(t) xác định trên đoạn [–T,T] nên đối số $\tau = t_2 - t_1$ của $R_x(\tau)$ sẽ nhận giá trị trên đoạn [–2T,2T], và do đó các hệ số D_k được xác định bởi:

$$D_k = \frac{1}{4T}\int_{-2T}^{2T} R_x(\tau)e^{-ie_k t}d\tau, \omega_k = \frac{\pi k}{2T} \tag{7.6.8}$$

Nếu đặt $\tau=0$ vào (7.6.7) ta được: $D_x=R_x(0) = \sum_{k=-\infty}^{\infty} D_k$ (7.6.9)

Như vậy, khi khai triển quá trình ngẫu nhiên dừng X(t) thành tổng vô hạn các dao động điều hoà với các biên độ ngẫu nhiên X_k thì phương sai D_x của quá trình X(t) sẽ dược biểu diễn dưới dạng tổng vô hạn các phương sai của những biên độ ngẫu nhiên X_k tương ứng.

So sánh (7.6.1) và (7.6.7) ta thấy rằng, việc xét bài toán khai triển quá trình ngẫu nhiên tương đương với việc xét bài toán khai triển hàm tương quan của nó.

Bây giờ thay cho (7.6.1) ta xét biểu thức:

$$X(t) = \int_{\infty}^{\infty} e^{i\omega t} d\Phi(\omega) \tag{7.6.10}$$

được gọi là khai triển phổ quá trình ngẫu nhiên X(t), trong đố $\Phi(\omega)$ là một hàm ngẫu nhiên của đối số ω. Tích phân ở vế phải (7.6.10) là tích phân Fourier – Stiltex.

Nếu đặt:

$$S_x^T(\omega_k) = \frac{1}{2\pi}\int_{-2T}^{2T} R_x(\tau)e^{-i\omega_k t}d\tau \tag{7.6.11}$$

$$\Delta\omega_k = \omega_k - \omega_{k-1} = \frac{\pi k}{2T} - \frac{\pi(k-1)}{2T} = \frac{\pi}{2T} \tag{7.6.12}$$

thì từ (7.6.8) ta có: $$S_X^T(\omega_k) = \frac{D_k}{\Delta\omega_k} \tag{7.6.13}$$

Tức $S_x^T(\omega_k)$ là mật độ trung bình của phương sai trên đoạn tần số $\Delta\omega_k$.

Kết hợp (7.6.7) và (7.6.13) ta nhận được:

$$R_x(\tau) = \sum_{k=-\infty}^{\infty} S_x^T(\omega_k) e^{i\omega_k t} \Delta\omega_k \qquad (7.6.14)$$

Chuyển qua giới hạn biểu thức (7.6.14) khi $T \to \infty$, còn $\Delta\omega_k \to 0$ tổng tích phân sẽ trở thành tích phân:

$$R_x(\tau) = \int_{-\infty}^{\infty} S_x(\omega_k) e^{i\omega\tau} d\omega \qquad (7.6.15)$$

Như vậy, hàm $S_x(\omega)$ là giới hạn của mật độ phương sai trung bình $S_x^T(\omega_k)$ khi $\Delta\omega_k$ dần đến 0, nó biểu thị mật độ phương sai của hàm ngẫu nhiên X(t) ứng với tần số ω và được gọi là mật độ phổ của quá trình ngẫu nhiên dừng X(t).

Giữa hàm mật độ phổ $S_x(\omega)$ và hàm tương quan $R_x(\tau)$ của quá trình ngẫu nhiên dừng X(t) liên hệ với nhau qua phép biến đổi Fourier:

$$S_x(\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} R_x(\tau) e^{-i\omega\tau} d\tau \qquad (7.6.16)$$

trong đó ω là tần số góc của dao động.

Hàm mật độ phổ $S_x(\omega)$ là một hàm không âm. Khi đặt τ=0 vào (7.6.15) ta được:

$$R_x(0) = D_x = \int_{-\infty}^{+\infty} S_x(\omega) d\omega \qquad (7.6.17)$$

Như vậy tổng diện tích giới hạn bởi đường cong mật độ phổ $S_x(\omega)$ và trục hoành bằng phương sai của quá trình ngẫu nhiên. Do thứ nguyên của D_x bằng bình phương thứ nguyên của X(t) nên người ta xem D_x như là năng lượng trung bình trong một đơn vị thời gian của quá trình hay còn gọi là công suất. Chính vì vậy $S_x(\omega)$ mang nhiều tên gọi khác nhau: Phổ phương sai, phổ năng lượng hay phổ công suất.

Nếu quá trình ngẫu nhiên X(t) có phương sai D_x hữu hạn, thì theo (7.6.17) hàm $S_x(\omega)$ khả tích. Khi đó hàm

$$F_x(\omega) = \int_{-\infty}^{\omega} S_x(\omega) d\omega \qquad (7.6.18)$$

được gọi là hàm phổ hay phổ tích phân của quá trình ngẫu nhiên dừng X(t).

Từ đây có thể nói về ý nghĩa vật lý của hàm phổ và mật độ phổ như sau:

Hàm phổ là hàm phân bố mô tả phân bố năng lượng của dao động ngẫu nhiên theo các tần số khác nhau; mật độ phổ là mật độ phân bố của năng lượng theo tần số.

Hàm $s_x(\omega)$ xác định bởi: $$s_X(\omega) = \frac{S_X(\omega)}{D_X} \tag{7.6.19}$$

được gọi là mật độ phổ chuẩn hoá.

Giữa hàm mật độ phổ chuẩn hoá $s_x(\omega)$ và hàm tương quan chuẩn hoá $r_x(\tau)$ cũng liên hệ với nhau qua phép biến đổi Fourier:

$$s_x(\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} r_X(\tau)e^{-i\omega\tau}d\tau \tag{7.6.20}$$

$$r_x(\tau) = \int_{-\infty}^{+\infty} s_X(\omega)e^{i\omega\tau}d\omega \tag{7.6.21}$$

Đối với quá trình ngẫu nhiên thực, do tính chẵn của hàm tương quan suy ra tính chẵn của mật độ phổ, nên có thể viết:

$$S_x(\omega) = \frac{1}{\pi}\int_{0}^{+\infty} R_X(\tau)\cos\omega\tau d\tau \tag{7.6.22}$$

$$R_x(\tau) = 2\int_{0}^{+\infty} S_X(\omega)\cos\omega\tau d\omega \tag{7.6.23}$$

$$s_x(\omega) = \frac{1}{\pi}\int_{0}^{+\infty} r_X(\tau)\cos\omega d\tau \tag{7.6.24}$$

$$r_x(\tau) = 2\int_{0}^{+\infty} s_X(\omega)\cos\omega d\omega \tag{7.6.25}$$

Trong thực tế, việc nghiên cứu chuỗi thời gian $\{x_t, t=1..n\}$ có thể được coi như xét quá trình ngẫu nhiên X(t) trên một thể hiện quan sát được x(t) của nó tại *n* lát cắt khác nhau. Hàm tương quan thực nghiệm $\widetilde{R}_X(\tau)$ tính được từ chuỗi $\{x_t\}$ là ước lượng của hàm tương quan thực $R_x(\tau)$, trong đó τ nhận giá trị trong một khoảng hữu hạn $|\tau| \le \tau_{max}$ với τ_{max} được gọi là độ dịch chuyển cực đại (hay còn gọi là điểm cắt) của hàm tương quan. Bởi vậy để tránh sai số khi sử dụng $\widetilde{R}_X(\tau)$ thay cho $R_x(\tau)$, người ta đưa vào khái niệm hàm cửa sổ $\lambda(\tau)$ trong biểu thức tính phổ:

$$\widetilde{S}_X(\omega) = \frac{1}{\pi}\int_{0}^{\tau_{max}} \lambda(\tau)\widetilde{R}_X(\tau)\cos\omega\tau d\tau \tag{7.6.26}$$

và $$\widetilde{s}_X(\omega) = \frac{1}{\pi}\int_{0}^{\tau_{max}} \lambda(\tau)\widetilde{r}_X(\tau)\cos\omega\tau d\tau \tag{7.6.27}$$

Ý nghĩa của hàm $\lambda(\tau)$ là ở chỗ nó có chức năng làm trơn hàm tương quan thực nghiệm và hợp lý hoá tích phân trên khoảng hữu hạn thay cho tích phân trên khoảng vô hạn trong các biểu thức tính phổ.

Độ chính xác của hàm mật độ phổ thực nghiệm phụ thuộc nhiều vào việc chọn hàm cửa sổ. Nhưng dạng hàm cửa sổ lại phụ thuộc vào dạng hàm tương quan thực nghiệm, tức là không có một dạng hàm cửa sổ nào chung cho tất cả các dạng hàm tương quan. Vì thế đã có nhiều tác giả khác nhau đưa ra các dạng hàm cửa sổ riêng biệt không giống nhau. Một trong những hàm được ứng dụng nhiều trong khí tượng, khí hậu là hàm Hamming:

$$\lambda(\tau) = \begin{cases} 0.54 + 0.46\cos\dfrac{\pi\tau}{\tau_{max}} & \text{khi} |\tau| \le \tau_{max} \\ 0 & \text{khi} |\tau| > \tau_{max} \end{cases} \qquad (7.6.28)$$

Trong tính toán thực hành thay cho (7.6.26) và (7.6.27) ta sử dụng các công thức sau:

$$\widetilde{S}_x(\omega_k) = \frac{1}{\pi}\sum_{i=0}^{m}\lambda(i\Delta\tau)\widetilde{R}_x(i\Delta\tau)\cos\frac{\pi k}{\tau_{max}}i\Delta\tau \qquad (7.6.29)$$

$$\widetilde{s}_x(\omega_k) = \frac{1}{\pi}\sum_{i=0}^{m}\lambda(i\Delta\tau)\widetilde{r}_x(i\Delta\tau)\cos\frac{\pi k}{\tau_{max}}i\Delta\tau \qquad (7.6.30)$$

Ở đây $m = \tau_{max}/\Delta\tau$, $\Delta\tau$ là bước thời gian của chuỗi $\{x_t\}$, ω_k là tần số góc của dao động, $\omega_k = \dfrac{\pi k}{\tau_{max}}$, k = 1,2,..,m là các giá trị mật độ phổ.

Nếu bước thời gian giữa hai thành phần kế cận của chuỗi bằng đơn vị ($\Delta\tau$=1) ta có các công thức tính sau đây (với giả thiết trung bình của chuỗi bằng 0):

1) Hàm tương quan và hàm tương quan chuẩn hoá:

$$\widetilde{R}_x(k) = \frac{1}{n-k}\sum_{t=1}^{n-k}x_t x_{t+k}\,,\ k=0,1,...,m \qquad (7.6.31)$$

$$\widetilde{r}_x(k) = \frac{\widetilde{R}_x(k)}{\widetilde{R}_x(0)} = \frac{\widetilde{R}_x(k)}{D_x}\,,\ k=0,1,...,m \qquad (7.6.31')$$

2) Hàm mật độ phổ và mật độ phổ chuẩn hoá:

$$\widetilde{S}_x(\omega_k) = \frac{1}{\pi}\sum_{i=0}^{m}\lambda(i)\widetilde{R}_x(i)\cos\frac{\pi k}{m}i\,,\ k=0,1,...,m \qquad (7.6.32)$$

$$\widetilde{S}_x(\omega_k) = \frac{1}{\pi}\sum_{i=0}^{m}\lambda(i)\widetilde{r}_x(i)\cos\frac{\pi k}{m}i\,,\ k=0,1,...,m \qquad (7.6.32')$$

Trong đó: $\omega_k = \dfrac{\pi k}{m}$, k = 1,2,..,m

$$\lambda(k) = \begin{cases} 0.54 + 0.46\cos\dfrac{\pi k}{m} & \text{khi } k \le m \\ 0 & \text{khi } k > m \end{cases}$$

Khi đã tính được hàm mật độ phổ, thông thường để nhận biết các đỉnh phổ mà tương ứng với nó là các chu kỳ dao động của chuỗi, người ta biểu diễn nó trên đồ thị với trục hoành là ω_k còn trục tung là $S_x(\omega_k)$ hoặc $s_x(\omega_k)$.

7.6.2 Đánh giá độ tin cậy của đặc trưng phổ.

Ta biết rằng, mật độ phổ được xác định từ chuỗi số liệu $\{x_t\}$ với dung lượng mẫu *n* nhất định. Do đó, mối quan hệ giữa dung lượng mẫu *n*, độ dịch chuyển cực đại τ_{max} và bước thời gian $\Delta\tau$ là một trong những vấn đề cần được xem xét. Thông thường, các chuỗi thời gian trong khí tượng thuỷ văn đều có $\Delta\tau$ nhận giá trị bằng 1 năm, 1 tháng,... tức là ta có thể xem $\Delta\tau$ =1 (đơn vị thời gian). Với dung lượng mẫu *n* cố định, nếu ta chọn τ_{max} lớn sẽ làm tăng độ phân giải của mật độ phổ, tức là cho phép tách được nhiều đỉnh phổ. Nhưng khi đó độ ổn định thống kê của hàm tương quan $R_x(\tau)$ sẽ không đảm bảo, dẫn đến sai số tiềm ẩn trong mật độ phổ nhận được. Nếu chọn τ_{max} bé để đảm bảo độ tin cậy của hàm tương quan thì độ phân giải của mật độ phổ tính được giảm đi, thậm chí quá nhỏ, làm cho các đỉnh phổ bị mờ chìm và ta sẽ không thể phát hiện được chúng.

Để giải quyết mâu thuẫn nói trên người ta đưa ra các công thức xác định τ_{max} ứng với các mức sai số cho phép khi tính mật độ phổ sau đây:

Mức sai số	2%	5%	10%
τ_{max}	2πn/50	2πn/20	2πn/10

trong đó n độ dài chuỗi.

Từ đó thấy rằng, muốn có kết quả nhận được vừa bảo đảm độ ổn định thống kê vừa phản ánh đúng những chu kỳ dao động của chuỗi thì dung lượng mẫu *n* phải đủ lớn. Dĩ nhiên *n* càng lớn thì độ tin cậy của kết quả càng cao.

Mặt khác, vì chuỗi thời gian được xem là những giá trị quan trắc thực nghiệm của một thể hiện của quá trình ngẫu nhiên, nên khi khảo sát quá trình X(t) trên cơ sở chuỗi $\{x_t, t=1..n\}$ sẽ có sự khác biệt giữa mật độ phổ thực nghiệm $\widetilde{S}_x(\omega_k)$ và mật độ phổ lý thuyết $S_x(\omega)$. Và do đó những dao động nhận được qua mật độ phổ thực nghiệm có thể chứa đựng cả các thành phần "nhiễu" mà người ta gọi là "ồn". Đối với các quá trình khí tượng thuỷ văn người ta phân biệt hai loại ồn là ồn trắng và ồn màu (hay ồn đỏ). Quá trình được gọi là ồn trắng nếu mật dộ phổ của nó không đổi trên toàn khoảng biến thiên của tần số. Quá trình được gọi là ồn màu nếu mật độ phổ của nó giảm theo qui luật hàm mũ khi tần số tăng.

Vậy, vấn đề đặt ra là cần kiểm tra xem các đỉnh phổ nhận được có thực sự phản ánh đúng những chu kỳ dao động tương ứng của chuỗi không, hay nói cách khác, mức độ tin cậy của các đỉnh phổ bằng bao nhiêu. Điều đó có nghĩa là cần phải kiểm nghiệm giả thiết "trong phổ của quá trình đang xét không tồn tại dao động điều hoà". Giả thiết được kiểm nghiệm bằng việc so sánh mật độ phổ tính toán $\widetilde{S}_x(\omega)$ với một giới hạn tin cậy $I_\alpha(S_\omega)$ nào đó. Một cách gần đúng, nếu thừa nhận rằng mật độ phổ thực nghiệm tuân theo luật phân bố χ_2/L, trong đo L là số bậc tự do được xác định bởi:

$$L = \frac{2n - 0.5m}{m} \tag{7.6.33}$$

thì, đối với ồn trắng, $I_\alpha(S_\omega)$ được tính theo công thức:

$$I_\alpha(S_x) = \overline{S_x}\frac{\chi^2}{I} \tag{7.6.34}$$

với $\overline{S_x}$ là mức trung bình của *m* giá trị mật độ phổ thực nghiệm.

Từ đó ta có các bước kiểm nghiệm sau đây:

1) Tính giá trị trung bình của mật độ phổ thực nghiệm $\overline{S_x}$

2) Chọn mức xác suất α (thường là 0.05, 0.10, 0.20) sau đó xác định giá trị $\chi^2/L = \chi^2(\alpha, L)/L$

3) Tính $I_\alpha(S_\omega)$ theo (7.6.32)

4) So sánh $\widetilde{S}_x(\omega_k)$ tính được với $I_\alpha(S_\omega)$:

– Nếu $\widetilde{S}_x(\omega_k) < I_\alpha(S_\omega)$ thì trong phổ không chứa dao động điều hoà ứng với tần số ω_k, giả thiết đặt ra đúng và ta chấp nhận nó.

– Nếu $\widetilde{S}_x(\omega_k) \geq I_\alpha(S_\omega)$ thì trong chuỗi tồn tại dao động điều hoà với tần số dao động là ω_k.

Ví dụ 7.6 Từ chuỗi số liệu tổng lượng mưa 3 tháng chính mùa mưa của 78 năm ở một trạm, ta tính mật độ phổ thực nghiệm $\widetilde{S}_x(\omega)$ với độ dịch chuyển cực đại của hàm tương quan được chọn m=19. Kết quả tính được biểu diễn lên đồ thị (hình 7.13), trong đó thay cho tần số ω trên trục hoành là các chu kỳ T tương ứng. Nhận xét sơ bộ ta thấy tồn tại 5 đỉnh phổ, trong đó có hai đỉnh khá mờ. Vậy câu hỏi đặt ra ở đây là những đỉnh phổ nào trong số đó phản ánh đúng các chu kỳ dao động của chuỗi.

Để giải quyết vấn đề này ta tính $I_\alpha(S_\omega)$ với các mức α khác nhau. Kết quả nhận được: $I_{0.01}(S_\omega)$=2.53, $I_{0.05}(S_\omega)$=1.93, $I_{0.10}(S_\omega)$=1.65, $I_{0.20}(S_\omega)$=1.34, $I_{0.30}(S_\omega)$=1.15. So sánh các $I_\alpha(S_\omega)$ tính được với trị số mật độ phổ tại các đỉnh ta thấy:

Nếu chọn α=0.01 thì chỉ có hai đỉnh phổ ứng với các chu kỳ T=2.5 năm và T=3.4 năm thoả mãn điều kiện $\widetilde{S}_x(\omega_k) \geq I_\alpha(S_\omega)$; nếu α được chọn bằng 0.05, 0.10,

0.20 ta nhận được ba đỉnh T=2.5, T=3.4 và T=5.7; nhưng khi α tăng lên bằng 0.30 (tức chấp nhận xác suất phạm sai lầm tới 30%) thì ngoài các đỉnh phổ trên ta còn nhận được thêm một đỉnh ứng với chu kỳ T=3.1 năm.

7.7 Ước lượng phổ năng lượng bằng phương pháp entropy cực đại

Như đã đề cập ở trên, khái niệm phổ phương sai cũng có thể hiểu là phổ năng lượng hay phổ công suất. Ở đây ta sẽ xét khái niệm phổ năng lượng dưới một góc độ hơi khác một chút, mặc dù, như sau này sẽ thấy, về bản chất có thể xem chúng là một.

Xét thể hiện x(t) của quá trình X(t). Khi đó tổng năng lượng của quá trình được xác định bởi:

$$\text{Power} = \int_{-\infty}^{\infty} |x(t)|^2 dt \tag{7.7.1}$$

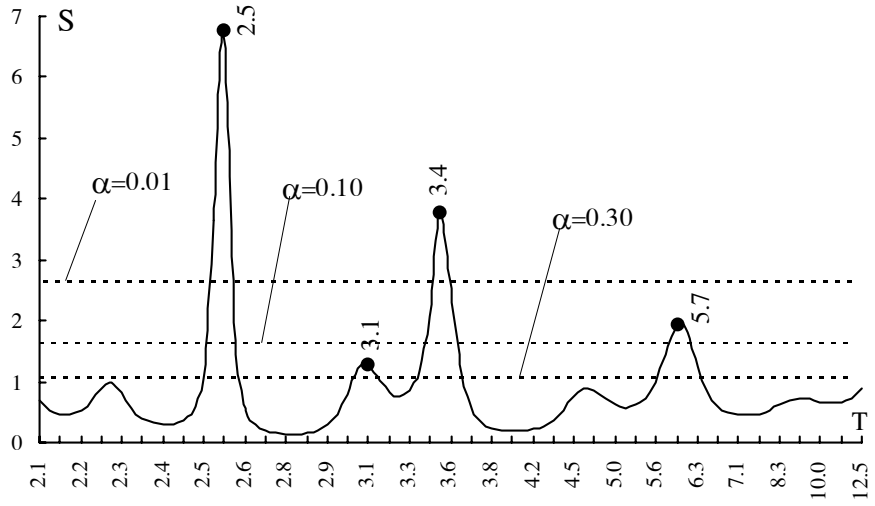
Mặt khác, theo định lý Parseval, năng lượng này cũng có thể được xác định bởi:

$$\text{Power} = \int_{-\infty}^{\infty} |x(t)|^2 dt = \int_{-\infty}^{\infty} |H(f)|^2 df \tag{7.7.1'}$$

Trong đó

$$H(f) = \int_{-\infty}^{\infty} x(t) e^{2\pi i f t} dt \tag{7.7.1''}$$

$$x(t) = \int_{-\infty}^{\infty} H(f) e^{-2\pi i f t} df$$



Hình 7.13 Đồ thị hàm mật độ phổ tổng lượng mưa 3 tháng mùa mưa

Có thể hình dung H(f) như là biên độ dao động điều hoà ứng với tần số dài f. Giữa tần số góc và tần số dài liên hệ với nhau bởi hệ thức $\omega=2\pi f$. Hệ thức (7.7.1') biểu diễn sự phân bố năng lượng của quá trình phân bố theo tần số. Sự phân bố này phản ánh cấu trúc bên trong của quá trình. Vấn đề ở chỗ ta cần xác định được "có bao nhiêu năng lượng của quá trình chứa trong khoảng tần số từ f đến f+df".

Trong thực tế người ta thường chỉ xét miền biến đổi của tần số f là những giá trị dương, mặt khác giá trị của H(f) trên miền tần số âm và tần số dương thường không khác nhau, nên thay cho H(f), người ta sử dụng hàm:

$$S(f) = |H(f)|^2 + |H(-f)|^2,\ 0 \le f < \infty$$

và được gọi là phổ năng lượng của quá trình.

Nếu thể hiện x(t) được cho tại *n* điểm t_i, i=1..n: $x_i=x(t_i)$ với $t_i=i\Delta\tau$, khi đó tại các điểm tần số rời rạc $f_k=\frac{k}{n\Delta\tau}$, k=0,1,2,... biểu thức (7.7.1'') sẽ có dạng:

$$H(f_k) = \int_{-\infty}^{\infty} x(t)e^{2\pi i f_k t}dt \approx \sum_{t=1}^{n} x_t e^{2\pi i f_k t\Delta\tau}\Delta\tau = \Delta\tau\sum_{t=1}^{n} x_t e^{2\pi i\frac{k.t}{n}} \qquad (7.7.2)$$

Nếu $\Delta\tau=1$ (đơn vị thời gian), từ (7.7.2) ta có công thức biến đổi ngược Fourier rời rạc để nhận các giá trị của chuỗi ban đầu:

$$x_t = \frac{1}{n}\sum_{k=1}^{n} H(f_k)e^{-i\frac{2\pi}{n}k.t} \qquad (7.7.3)$$

Và định lý Parseval có thể được viết lại dưới dạng:

$$\sum_{t=1}^{n}|x_t|^2 = \frac{1}{n}\sum_{k=1}^{n}|H(f_k)|^2 \qquad (7.7.4)$$

Nếu X(t) dừng có kỳ vọng bằng 0 thì khi chia vế trái của (7.7.4) cho *n* ta nhận được ước lượng phương sai của quá trình. Như vậy có thể hiểu phổ phương sai như là hàm biểu thị sự phân bố năng lượng trung bình của quá trình theo tần số, trong khi phổ năng lượng xét sự phân bố của tổng năng lượng của quá trình. Trên cơ sở biểu thức (7.7.4) ta có thể sử dụng phương pháp FFT để tính mật độ phổ của quá trình. Tuy nhiên, mặc dù phương pháp FFT cho phép tính toán nhanh, nó vẫn chứa đựng những hạn chế nhất định. Sau đây ta sẽ xét một phương pháp khác là phương pháp entropy cực đại (MEM – Maximum Entropy Method).

Ký hiệu tần số Nyquist là f_c, ta có:

$$f_c = \frac{1}{2\Delta\tau} \text{ và } f_k = 2f_c\frac{k}{n},\ k=0,1,2,...,n/2 \qquad (7.7.5)$$

Các tần số f_k chỉ nhận giá trị trên đoạn $[-f_c; f_c]$. Nếu ta thực hiện phép biến đổi:

$$z = e^{2\pi i f \Delta\tau} \tag{7.7.6}$$

khi đó có thể biểu diễn phổ năng lượng dưới dạng:

$$S(f) = \left| \sum_{k=-n/2}^{n/2-1} x_k z^k \right|^2 \tag{7.7.7}$$

Nếu x(t) xác định trên toàn miền vô hạn thì biểu thức (7.7.7) sẽ có dạng:

$$S(f) = \left| \sum_{k=-\infty}^{\infty} x_k z^k \right|^2 \tag{7.7.8}$$

Có thể biểu diễn hệ thức (7.7.7) dưới dạng gần đúng sau:

$$S(f) \approx \frac{1}{\left| \sum_{k=-m/2}^{m/2} b_k z^k \right|^2} = \frac{a_o}{\left| 1 + \sum_{k=-1}^{m} a_k z^k \right|^2} \tag{7.7.9}$$

Người ta gọi phép xấp xỉ (7.7.9) là phương pháp entropy cực đại (MEM) hay mô hình tất cả các cực (all–poles model) hoặc mô hình tự hồi qui (auto–regressive model – AR). Kết hợp (7.7.7) và (7.7.9) ta nhận được:

$$S(f) = \left| \sum_{k=-n/2}^{n/2-1} x_k z^k \right|^2 \approx \frac{a_o}{\left| 1 + \sum_{k=-1}^{m} a_k z^k \right|^2} \tag{7.7.10}$$

Điều đó có nghĩa là để xác định mật độ phổ S(f) cần phải tính được *m+1* hệ số a_o, a_1,...,a_m. Người ta đã chứng minh được rằng, để nhận được các hệ số a_k (k=0..m) cần phải giải phương trình sau:

$$\frac{a_o}{\left| 1 + \sum_{k=-1}^{m} a_k z^k \right|^2} \approx \sum_{j=-m}^{m} R_j z^j \tag{7.7.11}$$

trong đó các R_j được xác định bởi:

$$R_j \approx \frac{1}{n-j} \sum_{t=1}^{n-j} x_t x_{t+j}, \quad j = 0,1,2,...,n-1 \tag{7.7.12}$$

và

$$R_j = R_{-j} \tag{7.7.12'}$$

Số *m* được gọi là bậc xấp xỉ hay số cực xấp xỉ. Về nguyên tắc *m* có thể nhận bất kỳ một số nguyên dương nào cho đến *n–1* là tổng số các mômen tự tương quan R_j có

thể có. Thậm chí *m* có thể nhận giá trị lớn hơn dung lượng mẫu *n*, nhưng trong trường hợp này cần phải ngoại suy hàm tự tương quan. Trên thực tế thường người ta chọn *m* nhỏ hơn *n* nhiều.

Có nhiều thủ thuật để giải phương trình (7.7.11) mà một trong những phương pháp đó là đưa phương trình về dạng ma trận:

$$\begin{pmatrix} R_1 & R_2 & ... & R_{m+1} \\ R_2 & R_1 & ... & R_m \\ ... & ... & ... & ... \\ R_{m+1} & R_m & ... & R_1 \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ ... \\ a_m \end{pmatrix} = \begin{pmatrix} a_0 \\ 0 \\ ... \\ 0 \end{pmatrix} \tag{7.7.13}$$

Sau khi tính được các hệ số a_k, k=0..m, ứng với các giá trị tần số cho trước f_k, thay vào (7.7.10) ta sẽ xác đinh được mật độ phổ. Chú ý rằng các giá trị tần số phải thoả mãn điều kiện:

$$-\frac{1}{2} \le f_k \Delta\tau \le \frac{1}{2} \tag{7.7.14}$$

7.8 Phương pháp chuẩn sai tích luỹ phân tích xu thế

Chuẩn sai là hiệu giữa các thành phần của chuỗi và giá trị trung bình $d_t=x_t-\overline{x}$. Do đó, từ chuỗi ban đầu ta có thể thành lập được chuỗi chuẩn sai {d_t, t=1..n} Dấu chuẩn sai cho biết trị số của chuỗi tăng hay giảm, còn giá trị chuẩn sai đánh giá mức độ tăng hay giảm của các thành phần trong chuỗi so với trung bình.

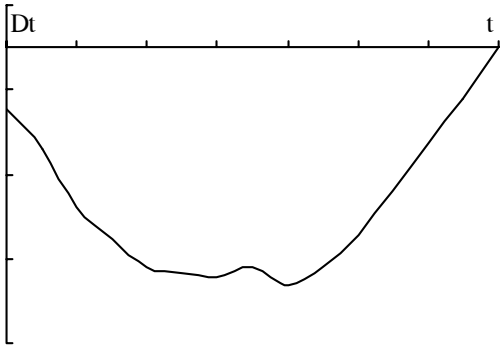
Chuẩn sai tích luỹ D_t là tổng các chuẩn sai d_t, được xác định bởi:

$$D_t = \sum_{i=1}^{t} d_i$$

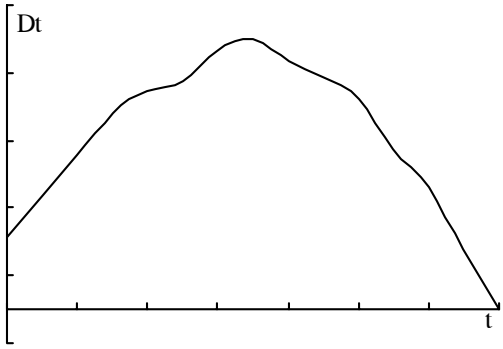
Như vậy, $D_1 = d_1$, $D_2 = d_1+d_2$, ... , $D_n = d_1+ d_2+...+ d_n = 0$, tức là từ chuỗi {$d_t$, t=1..n} ta sẽ thành lập được chuỗi {D_t, t=1..n}. Ta thấy giá trị của D_t phụ thuộc vào trạng thái dao động và xu thế của chuỗi. Nếu chuỗi có xu thế tăng thuần tuý thì giá trị trung bình của chuỗi sẽ nằm ở vị trí khoảng giữa chuỗi, do đó các thành phần đầu chuỗi chuẩn sai d_t sẽ mang dấu âm còn các thành phần cuối chuỗi mang dấu dương. Các giá trị chuẩn sai tích luỹ D_t vì thế sẽ càng âm khi *t* tăng cho đến khoảng giữa chuỗi, sau đó sẽ giảm dần về giá trị tuyệt đối (nhưng vẫn mang dấu âm) cho đến 0. Tình huống ngược lại sẽ xảy ra nếu chuỗi có xu thế giảm thuần tuý. Các hình 7.14a và 7.14b đã dẫn ra ví dụ mô phỏng minh hoạ cho các trường hợp này, trong đó xu thế tăng, giảm được cho theo qui luật gần tuyến tính.

Nếu chuỗi có xu thế tăng rồi giảm hoặc giảm rồi tăng, tức là theo biến trình thời gian chuỗi có một cực đại hoặc một cực tiểu, trị số trung bình của chuỗi nằm ở

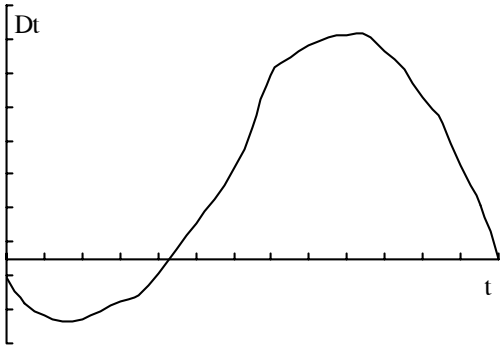
hai ví trí khoảng giữa đoạn đầu và khoảng giữa đoạn cuối của chuỗi. Các chuẩn sai d_t sẽ bắt đầu bằng những giá trị âm (nếu chuỗi tăng rồi giảm) hoặc dương (nếu chuỗi giảm rồi tăng), cho đến khoảng giữa đoạn đầu thì mang dấu ngược lại và giữ nguyên dấu đến khoảng giữa đoạn cuối lại đổi dấu cho đến hết chuỗi. Minh hoạ cho các trường hợp này được dẫn ra trên các hình 7.15a và 7.15b.



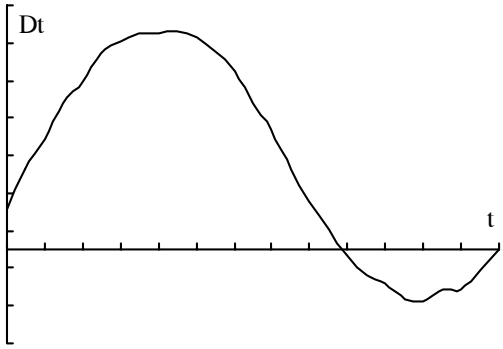
Hình 7.14a Xu thế tăng thuần tuý



Hình 7.14b Xu thế giảm thuần tuý



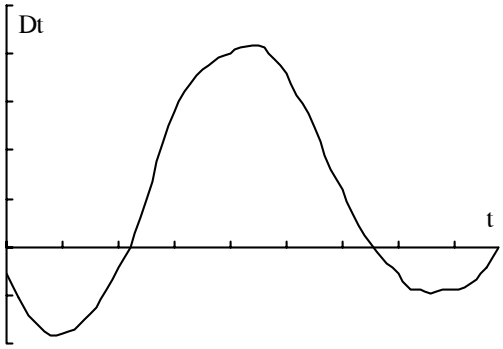
Hình 7.15a Xu thế tăng rồi giảm



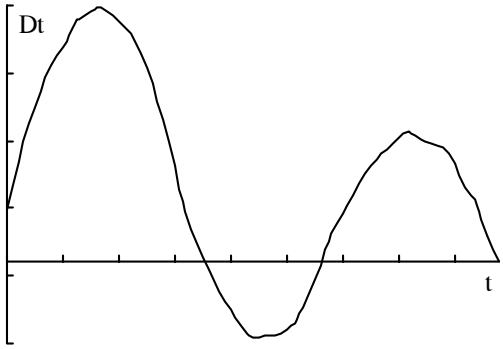
Hình 7.15b Xu thế giảm rồi tăng

Trên các hình 7.16a và 7.16b dẫn ra đồ thị biểu diễn sự biến đổi của D_t theo *t* ứng với các tình huống chuỗi có xu thế tăng rồi giảm sau đó tăng và giảm rồi tăng sau đó giảm.

Từ các hình 7.14–7.16 có thể nhận thấy một tính chất chung của chuẩn sai tích luỹ D_t là, nhìn chung với D_t âm thì chuỗi có xu thế tăng còn với D_t dương thì chuỗi có xu thế giảm.



Hình 7.16a Xu thế tăng rồi giảm sau đó lại tăng



Hình 7.16b Xu thế giảm rồi tăng sau đó lại giảm

Trường hợp các thành phần của chuỗi dao động ngẫu nhiên xung quanh giá trị trung bình, các chuẩn sai d_t có dấu âm dương đan xen nhau, hơn nữa giá trị tuyệt đối của chúng không khác nhau nhiều, thì có thể nói chuỗi không có xu thế. Đồ thị của chuẩn sai tích luỹ D_t theo t vì vậy cũng sẽ dao động ngẫu nhiên xung quanh trị số 0 và có không thể hiện rõ ràng qui luật biến đổi.

Ví dụ 7.8 Bảng 7.4 dẫn ra trích đoạn chuỗi số liệu tổng lượng mưa năm của một trạm và những kết quả tính chuẩn sai d_t và chuẩn sai tích luỹ D_t của chuỗi. Trị số trung bình toàn chuỗi là 1899.2 (mm). Kết quả tính toán được biểu diễn trên hình 7.17.

Từ hình này ta thấy sự biến đổi của chuỗi lương mưa năm về cơ bản được chia làm hai giai đoạn. Giai đoạn từ đầu thế kỷ đến khoảng những năm đầu thập kỷ bốn mươi lượng mưa có xu thế tăng dần theo thời gian. Giai đoạn từ đầu thập kỷ bốn mươi đến cuối những năm tám mươi lượng mưa có xu thế giảm dần. Tuy vậy, sự nhấp nhô của đồ thị đã phản ánh xu thế biến thiên của chuỗi đan xen nhau giữa các thời kỳ có tổng lượng mưa vượt chuẩn và dưới chuẩn. Hay nói cách khác, tồn tại trong chuỗi một sự luân phiên của các thời đoạn có dấu chuẩn sai dương và âm.

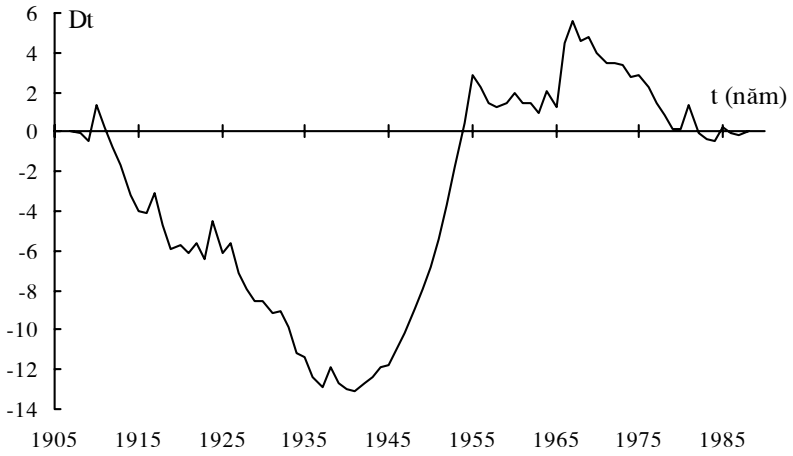
Bảng 7.4 Chuẩn sai và chuẩn sai tích luỹ tổng lượng mưa năm (mm)

Năm	Tổng lượng mưa năm	Chuẩn sai d_t	Chuẩn sai tích luỹ D_t
1907	1906.2	7.0	7.0
1908	1837.4	–61.8	–54.7
1909	1644.0	–255.2	–309.9
1910	3029.6	1130.4	820.6
1911	1233.8	–665.4	155.2
1912	1239.9	–659.3	–504.0
...	...	...	...
1985	2346.4	447.2	166.5
1986	1669.9	–229.3	–62.8
1987	1838.8	–60.4	–123.1
1988	2022.3	123.1	0.0

7.9 Phương pháp hồi qui phân tích xu thế

Một trong những phương pháp phân tích xu thế thường được xét đến trong việc nghiên cứu dao động khí hậu là phương pháp hồi qui.

Phương pháp hồi qui được đề cập ở đây là hồi qui giữa biến khí hậu x và thời gian t, tức là sự biến đổi của x theo t: $x = f(t)$. Nếu f(t) là một hàm tuyến tính ta có xu thế biến đổi tuyến tính. Trong những trường hợp khác ta sẽ gọi nó là xu thế không tuyến tính.



Hình 7.17 Đường chuẩn sai tích luỹ tổng lượng mưa năm

Để nghiên cứu xu thế biến đổi tuyến tính ta thành lập phương trình hồi qui:

$$x(t) = at + b \qquad (*)$$

trong đó a và b là các hệ số hồi qui được xác định bởi:

$$a = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(t - \bar{t})}{\sqrt{\sum_{t=1}^{n}(x_t - \bar{x})^2 \sum_{t=1}^{n}(t - \bar{t})^2}}, \quad b = \bar{x} - a\bar{t}$$

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t, \quad \bar{t} = \frac{1}{n}\sum_{t=1}^{n} t$$

Từ phương trình (*) ta có thể nhận biết được xu thế biến đổi của chuỗi thông qua phân tích hệ số góc a. Dấu của hệ số a xác định xu thế tăng (khi a>0) hoặc giảm (khi a<0), còn trị tuyệt đối của a cho biết mức độ tăng giảm của chuỗi.

Trong ứng dụng thực hành, ta có thể chia chuỗi thành các đoạn khác nhau để phân tích xu thế. Khi đó căn cứ vào các hệ số góc a ta có thể biết được xu thế của chuỗi qua các thời kỳ khác nhau.

PHẦN PHỤ LỤC

Phụ lục 1. Một số kiến thức về ma trận và định thức

1. Ma trận

Ma trận A cấp (hay bậc m×n, ký hiệu là A(m×n) là một bảng hình chữ nhật gồm các số hay các phần tử a_{ik} được sắp xếp thành m hàng và n cột:

$$A = \begin{pmatrix} a_{11} & a_{12} & .... & a_{1n} \\ a_{21} & a_{22} & .... & a_{2n} \\ .... & .... & .... & .... \\ a_{m1} & a_{m2} & .... & a_{mn} \end{pmatrix}$$

Đôi khi để cho gọn ta thường viết A = (a_{ik}), hoặc đơn giản hơn: (a_{ik}). Trong phạm vi tài liệu này ta sẽ luôn giả thiết rằng các phần tử a_{ik} là những số thực.

Khi m = 1, ma trận A chỉ gồm một phần tử a_{11}, và nó được đồng nhất với một số. Khi m = n, ma trận A được gọi là ma trận vuông.

– Ma trận A'(n×m) tạo thành từ ma trận A(m×n) bằng cách đổi chỗ vị trí hàng thành cột được gọi là ma trận chuyển vị của A: $a'_{ik} = a_{ki}$

– Ma trận A'(nxm) tạo thành từ ma trận A(mxn) bằng cách đổi chỗ vị trí hàng thành cột được gọi là ma trận chuyển vị của A: $a'_{ik} = a_{ki}$ với k=1..m, i=1..n.

– Tổng của hai ma trận cùng cấp A và B là một ma trận cùng cấp C mà các phần tử của nó bằng tổng các phần tử tương ứng của các ma trận A và B:

$$c_{ik} = a_{ik} + b_{ik}$$

– Tích của hai ma trận A(mxn) và B(nxp) sẽ là một ma trận C cấp (mxp), trong đó các phần tử c_{ik} được xác định bởi:

$$c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk}$$

Cần lưu ý rằng tích hai ma trận chỉ thực hiện được khi số cột của ma trận thứ nhất bằng số hàng của ma trận thứ hai. Hơn nữa phép nhân ma trận không có tính giao hoán, tức là nói chung AB≠BA trong trường hợp phép nhân thực hiện được.

– Nếu A là một ma trận vuông thì các phần tử a_{ii} (chỉ số hàng bằng chỉ số cột) lập nên đường chéo chính của A và các phần tử đó được gọi là các phần tử đường chéo. Nếu A có các phần tử đối xứng qua đường chéo chính thì A được gọi là ma trận đối xứng.

– Ma trận đối xứng A có các phần tử bằng 0 trừ đường chéo chính được gọi là ma trận đường chéo. Ma trận đường chéo mà các phần tử đều bằng 1 được gọi là ma trận đơn vị.

– Ma trận chỉ gồm một hàng hoặc một cột được gọi là vectơ hàng hoặc vectơ cột.

2. Định thức

Mỗi một ma trận vuông A(nxn) tương ứng với một số D nào đó được gọi là định thức của A. Số D được ký hiệu bằng:

$$D = |A| = |a_{ik}| = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{vmatrix}$$

và được xác định bởi tổng:

$$D = \sum \pm a_{1r_1} a_{2r_2} \dots a_{nr_n}$$

trong đó các chỉ số r_1, r_2,..., r_n chạy qua tất cả n! hoán vị có thể có của các số 1,2,...,n, còn dấu của mỗi số hạng là (+) hay (–) tuỳ theo hoán vị tương ứng là chẵn hay lẻ. Số n được gọi là cấp của định thức.

– Định thức của ma trận vuông A và chuyển vị A' của nó bằng nhau.

– Nếu đổi chỗ hai hàng hay hai cột của ma trận cho nhau thì định thức đổi dấu. Do đó nếu ma trận có hai hàng hoặc hai cột giống nhau thì định thức của nó bằng không.

– Định thức của tích hai ma trận vuông bằng tích các định thức của chúng.

– Nếu A là một ma trận bất kỳ thì ma trận A_1 nhận được từ ma trận A bằng cách bỏ đi một số hàng và một số cột được gọi là ma trận con của A. Nếu A là ma trận vuông thì ma trận con vuông của A có các phần tử đường chéo chính là những phần tử đường chéo chính của A.

– Định thức của một ma trận con vuông của ma trận A được gọi là một tử thức của A. Ta gọi phần phụ đại số A_{ik} của phần tử a_{ik} của ma trận vuông A là tích của tử thức nhận được bằng cách bỏ đi hàng thứ i cột thứ k, nhân với $(-1)^{i+k}$.

– Ta có một số đồng nhất thức quan trọng:

$$\sum_{j=1}^{n} a_{ij}A_{kj} = \begin{cases} D & \text{khi } i = k \\ 0 & \text{khi } i \neq k \end{cases}$$

$$\sum_{j=1}^{n} a_{ji}A_{jk} = \begin{cases} D & \text{khi } i = k \\ 0 & \text{khi } i \neq k \end{cases}$$

Phụ lục 2. Một số hàm đặc biệt

1. Hàm Gamma

Hàm Gamma Γ(p) được xác định bởi hệ thức sau với mọi số thực p>0:

$$\Gamma(p) = \int_0^{+\infty} x^{p-1}e^{-x}dx$$

Nếu p dần tới 0 hay tới +∞ thì Γ(p) dần tới +∞. Hàm Γ(p) có cực tiểu duy nhất tại $p_o \in (0; +\infty)$ và người ta đã xác định được giá trị gần đúng của p_o là p_o=1.4616. Giá trị hàm Γ(p) tại điểm cực tiểu bằng $\Gamma(p_o) = 0.8856$.

Một số tính chất của hàm Γ(p):

1) Với mọi p>0 ta có Γ(p+1) = pΓ(p)

2) Nếu p là số nguyên dương n thì Γ(1) = 1 và Γ(n+1) = n!

3) Với mọi α>0, λ>0 ta có: $\int_0^{+\infty} x^{\lambda-1}e^{-\alpha x}dx = \frac{\Gamma(\lambda)}{\alpha^{\lambda}}$

4) $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, $\Gamma^2\left(\frac{1}{2}\right) = \pi$

2. Hàm Bêta

Hàm Bêta B(p,q) được xác định bởi hệ thức sau với mọi p>0 và q>0:

$$B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx$$

Giữa hàm Γ(p) và hàm B(p,q) liên hệ với nhau bởi hệ thức:

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$$

Phụ lục 3. Một số bảng tính sẵn

1. Bảng giá trị hàm Laplas Φ(x) = $\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{1}{2}t^2}dt$

x	Φ(x)	x	Φ(x)	x	Φ(x)	x	Φ(x)
0.10	0.0398	0.60	0.2257	1.10	0.3643	1.60	0.4452
0.15	0.0596	0.65	0.2422	1.15	0.3749	1.65	0.4505
0.20	0.0793	0.70	0.2580	1.20	0.3849	1.70	0.4554
0.25	0.0987	0.75	0.2734	1.25	0.3944	1.75	0.4599
0.30	0.1179	0.80	0.2881	1.30	0.4032	1.80	0.4641
0.35	0.1368	0.85	0.3023	1.35	0.4115	1.85	0.4678
0.40	0.1554	0.90	0.3159	1.40	0.4192	1.90	0.4713
0.45	0.1736	0.95	0.3289	1.45	0.4265	1.95	0.4744
0.50	0.1915	1.00	0.3413	1.50	0.4332	2.00	0.4772
0.55	0.2088	1.05	0.3531	1.55	0.4394	2.05	0.4798

x	Φ(x)	x	Φ(x)	x	Φ(x)	x	Φ(x)
2.10	0.4821	2.60	0.4953	3.10	0.4990	3.60	0.4998
2.15	0.4842	2.65	0.4960	3.15	0.4992	3.65	0.4999
2.20	0.4861	2.70	0.4965	3.20	0.4993	3.70	0.4999
2.25	0.4878	2.75	0.4970	3.25	0.4994	3.75	0.4999
2.30	0.4893	2.80	0.4974	3.30	0.4995	3.80	0.4999
2.35	0.4906	2.85	0.4978	3.35	0.4996	3.85	0.4999
2.40	0.4918	2.90	0.4981	3.40	0.4997	3.90	0.5000
2.45	0.4929	2.95	0.4984	3.45	0.4997	3.95	0.5000
2.50	0.4938	3.00	0.4987	3.50	0.4998	4.00	0.5000
2.55	0.4946	3.05	0.4989	3.55	0.4998	4.05	0.5000

2. Phân bố χ^2

Bảng tính giá trị χ^2_p ứng với xác suất $p=P(\chi^2>\chi^2_p)=\int_{\chi^2_p}^{\infty} f(x,n)dx$, trong đó f(x,n) là hàm mật độ xác suất χ^2 với n bậc tự do.

	p													
n	0.99	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.000	0.001	0.004	0.016	0.064	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	0.020	0.040	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	0.115	0.185	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.345	16.266
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.466
5	0.554	0.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.515
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.321
8	1.647	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.124
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.041	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.471	27.688	34.527
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.124
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.698
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000	39.252
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409	40.791
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805	42.312
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191	43.819
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566	45.314
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932	46.796
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289	48.268
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638	49.728
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980	51.179
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314	52.619
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642	54.051
27	12.878	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963	55.475
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278	56.892
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588	58.301
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892	59.702

3. Phân bố Student (t)

Bảng tính giá trị t_p ứng với xác suất $p=P\left(|t|>t_p\right)=2\int_{t_p}^{\infty} f(x,n)dx$, trong đó f(x,n) là hàm mật độ phân bố *t* với *n* bậc tự do.

	p											
n	0.900	0.800	0.700	0.600	0.500	0.400	0.300	0.200	0.100	0.050	0.010	0.001
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	63.656	636.58
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	9.925	31.600
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	5.841	12.924
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.533	2.132	2.776	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	4.032	6.869
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	3.499	5.408
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	3.355	5.041
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	3.250	4.781
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	3.169	4.587
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	3.106	4.437
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	3.055	4.318
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	3.012	4.221
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.977	4.140
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.947	4.073
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.921	4.015
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.898	3.965
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.878	3.922
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.861	3.883
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.845	3.850
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.831	3.819
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.819	3.792
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.807	3.768
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.797	3.745
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.787	3.725
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.779	3.707
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.771	3.689
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.763	3.674
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.756	3.660
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.750	3.646
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.704	3.551
60	0.126	0.254	0.387	0.527	0.679	0.848	1.045	1.296	1.671	2.000	2.660	3.460
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.617	3.373
∞	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.576	3.290

TÀI LIỆU THAM KHẢO

1. *Đào Hữu Hồ, Nguyễn Văn Hữu, Hoàng Hữu Như:* **Thống kê toán học.** NXB Đại học và Trung học chuyên nghiệp, Hà Nội, 1984, 505 tr.
2. *Ngô Như Hoà:* **Thống kê trong nghiên cứu y học** (Tập II). NXB Y học, 1982, 416 tr.
3. *Doerffel:* **Thống kê trong hoá học phân tích** (Trần Bính và Nguyễn Văn Ngạc dịch). NXB Đại học và Trung học chuyên nghiệp, Hà Nội, 1983, 272 tr.
4. *Harald Cramér:* **Phương pháp toán học trong thống kê** (Tập I) (Nguyễn Khắc Phúc, Nguyễn Duy Tiến, Đào Hữu Hồ dịch). NXB Khoa học, Hà Nội, 1969, 474 tr.
5. *Harald Cramér:* **Phương pháp toán học trong thống kê** (Tập II) (Nguyễn Khắc Phúc, Nguyễn Duy Tiến, Đào Hữu Hồ dịch). NXB Khoa học, Hà Nội, 1969, 326 tr.
6. *Khac Kevit A. A.:* **Phổ và phân tích phổ** (Nguyễn Văn Ngọ và Phương Xuân Nhàn dịch). NXB Đại học và Trung học chuyên nghiệp, Hà Nội, 1977, 260 tr.
7. *Pugatrep V.S.:* **Lý thuyến hàm ngẫu nhiên và ứng dụng vào các vấn đề điều khiển tự động** (Tập I) (Huỳnh Sum và Nguyễn Văn Hữu dịch). NXB Đại học và Trung học chuyên nghiệp, Hà Nội, 1978, 558 tr.
8. *Pugatrep V.S.:* **Lý thuyến hàm ngẫu nhiên và ứng dụng vào các vấn đề điều khiển tự động** (Tập II) (Huỳnh Sum và Nguyễn Văn Hữu dịch). NXB Đại học và Trung học chuyên nghiệp, Hà Nội, 1978, 380 tr.
9. *Rumsixki L. Z.:* **Phương pháp toán học xử lý các kết quả thực nghiệm**. NXB Khoa học và Kỹ thuật, Hà Nội, 1972, 283 tr.
10. *Ventxel A.D.:* **Giáo trình lý thuyết quá trình ngẫu nhiên** (Nguyễn Viết Phú, Nguyễn Duy Tiến dịch). NXB "MIR" Maxcơva, 1975; NXB Đại học và THCN, Hà Nội, 1987, 461 tr.
11. *Anderson T. W.:* **An introduction to multivariate statistical analysis**. Copyright (C) 1958 by John Wiley & Sons, Inc. Canada, 353 p.
12. *Chang C.P., Kirshnamurti T.N.:* **Mosoon mteorology**. Oxford University Press. New York, Clarendon Press. Oxford, 1987, 544 p.
13. *Daniel S. Wilks:* **Statistical methods in the Atmospheric Sciences** - An Introduction. Academic Press, 1995, 465 p.
14. *Hans A. Panofsky, Glenn W. Brier:* **Some applications of statistics to meteorology**. University Park, Pennsylvania, 1965, 223 p.

15. *Palul G. Hoel:* **Introduction Mathematical Statistics**. New York John Wiley & Sons, Inc. London, 1961, 331 p.

16. *Thiébaux H.J., Pedder M.A.:* **Spatial objective analysis: with applications in atmospheric science**. Academic Press, 1987, 297 p.

17. *William H. Press, Brian P. Flannerg, Saul A. Teukolsky, William T. Verrerling:* **Numerical recipes**. Cambridge University Press Inc., 1990, 681 p.

18. *Ben@at §j., Pircon A.:* Idmerenie i analid xluqa`n h procexxov. Id@atel]xtvo MIR, Moxkva, 1974, 464 x.

19. *German §j. R, Gol@berg R.A.:* Xolnce, pogo@a i klimat. Leningra@ gi@rometeoid@at, 1981, 318 x.

20. *Gruda G.V., Ran]kova D.¡.:* Vero¨tnoxtn e meteorologiqexkie prognod . Gi@rometeoid@at Leningra@, 1983, 271 x.

21. *Gumbel] /.:* Xtatixtika \kxtremal]n h dnaqeni`. Id@atel]xtvo MIR, Moxkva, 1965, 452 x.

22. *Ivahnenko A.G., Lapa V.G.:* Pre@xkadanie xluqa`n h procexxov. Id@atel]xtvo Naukova, Kiev, 1971,446 x.

23. *Ixaev A.A.:* Xtatixtika v meteorologii i klimatologii. Id@atel]xtvo Moxkovxkogo Univerxiteta, 1988, 244 c.

24. *Kadakeviq §.I.:* Oxnov teorii xluqa`n h funkci` i eª primenenie v gi@rometeorologii. Gi@rometeorologiqexkoe id@atel]xtvo, Leningra@, 1971, 267 x.

25. *Kov seva N.V., Gol]berg M.A.:* Meto@iqexkie ukadani¨ po xtatixtiqexko` obrabotke meteorologiqexkih r¨@ov. Leningra@, gi@rometeoid@at, 1990, 85 x.

26. *L]vovxki` E.N.:* Xtatixtiqexkie meto@ poxtroeni¨ \mpiriqexkih formul. Moxkva, V xsa¨ skola, 1982,224 x.

27. *Otnex R., /nokxon L.:* Prikla@no` analid vremenn h r¨@ov - Oxnovn e meto@ . Id@atel]xtvo MIR, Moxkva, 1982, 428 x.

28. *Pugaqev V.X.:* Teori¨ vero¨tnoxte` i matematiqexka¨ xtatixtika. Moxkva Nauka glavna¨ re@akci¨ fidiko-matematiqexko` literatur , 1979, 495 x.

29. *Roj@extvenxki` A.V., Qebotaerev A.I.:* Xtatixtiqexkie meto@ v gi@rologii. Gi@rometeoid@at, Leningra@, 1974, 424 x.