PHAN THANH KIẾM

CƠ SỞ TOÁN HỌC

CỦA CÁC PHÉP XỬ LÝ THỐNG KÊ TRONG NGHIÊN CỨU KHOA HỌC NÔNG NGHIỆP

NHÀ XUẤT BẢN NÔNG NGHIỆP

PGS. TS. PHAN THANH KIẾM

CƠ SỞ TOÁN HỌC

CỦA CÁC PHÉP XỬ LÝ THỐNG KÊ TRONG NGHIÊN CỨU KHOA HỌC NÔNG NGHIỆP

NHÀ XUẤT BẢN NÔNG NGHIỆP

Tp. Hồ Chí Minh – 2010

MỤC LỤC

Chương 3

Chương 4

Phần 2

Chương 7

Chương 8

Chương 9

MỘT SỐ THUẬT NGỮ VÀ KÝ HIỆU

Thuật ngữ	Tiếng Anh
Dấu hiệu (đặc trưng) định lượng	Quantitative characteristics
Dấu hiệu (đặc trưng) định tính	Qualitative characteristics
Dung lượng (kích thước) mẫu	Size of sample
Đại lượng (biến) ngẫu nhiên	Random variable
Độ lệch chuẩn	Standard deviation
Độ tin cậy	Degree of confidence
Độ tự do	Degree of freedom
Giả thiết thống kê	Statisticcal hypothesis
Đối thiết	Alternative hypothesis
Hàm phân phối	Distribution function
Hàm mật độ xác suất	Probability density function
Hệ số góc	Slope
Hệ số đường	Path coefficient
Hệ số tương quan	Correlation coefficient
Hiệp phương sai (hiệp sai)	Covariance
Hồi quy tuyến tính	Linear regression
Hồi quy phi tuyến tính	Non - linear regression
Kỳ vọng (kỳ vọng toán)	Mathematical expectation
Mẫu	Sample
Phân tích đường	Path analysis
Phương pháp (nguyên tắc) bình phương tối thiểu	Method (principle) of least squares
Phương sai	Variance (dispersion)
Sai lầm	Risk

Sai số tiêu chuẩn (sai số chuẩn)	Standard error
Tham số (thông số) thống kê	Statistical parameter
Thống kê mô tả	Descriptive statistics
Tổng thể	Population
Tương quan	Correlation
Trung bình (trung bình cộng)	Mean, sample mean, average
Ước lượng điểm	Point estimate
Ước lượng khoảng	Interval estimate

Ký hiệu	**Nghĩa**
AB	Nghiệm thức phối hợp giữa hai yếu tố A với B
X_{ij}	Giá trị nghiệm thức A_iB_j
A×B	Tương tác giữa hai yếu tố A với B
ab_{ij}	Giá trị hiệu quả tương tác $A_i \times B_j$
ABC	Nghiệm thức phối hợp giữa ba yếu tố A, B với C
X_{ijl}	Giá trị nghiệm thức $A_iB_jC_l$
A×B×C	Tương tác giữa ba yếu tố A với B với C
abc_{ijl}	Giá trị hiệu quả tương tác $A_i \times B_j \times C_l$

LỜI NÓI ĐẦU

Thống kê toán học ra đời rất sớm và có mặt ở hầu hết các lĩnh vực hoạt động của con người, từ khoa học tự nhiên, kinh tế học đến khoa học xã hội và nhân văn. A. Ketle (1796 – 1874), F. Galton (1822 – 1911), K. Pearson (1857 – 1936), W. S. Gosset (Student, 1876 – 1937), R. A. Fisher (1890 – 1962), M. Mitrel (1874 – 1948) là những người đặt nền móng cho thống kê sinh học hiện đại.

Trong quá trình phát triển, thống kê sinh học không dừng lại ở việc mô tả, suy đoán mà đã trở thành môn "khoa học về các tiêu chuẩn của việc tính toán". Trong sự lớn mạnh của thống kê sinh học có sự đóng góp đáng kể của các nhà khoa học thực nghiệm.

Năm 1973, khi đề cập đến công tác cải cách giáo dục, UNESCO đã khẳng định rằng Xác suất – Thống kê là một trong 9 vấn đề chủ chốt để xây dựng nền học vấn hiện đại.

Để giúp cho các sinh viên, học viên cao học và những nghiên cứu viên am hiểu cơ sở toán học của các phép xử lý số liệu trong nghiên cứu khoa học nông nghiệp, cuốn sách này được biên soạn. Nội dung của sách gồm hai phần:

- Phần đầu là các phương pháp lấy mẫu, điều tra thu thập và xử lý số liệu, từ thống kê mô tả, ước lượng các tham số thống kê đến việc so sánh và phân tích mối quan hệ giữa các tham số.

- Phần hai là các kiểu bố trí thí nghiệm, các phương pháp xử lý số liệu và cách trình bày báo cáo khoa học.

Để giúp bạn đọc không chuyên ngành thống kê có thể dễ nắm bắt được các nội dung, trong phần đầu tác giả đã trình bày dưới dạng ứng dụng, hạn chế việc lạm dụng các thuật ngữ thống kê. Tuy nhiên các nội dung vẫn đảm bảo tính khoa học, tính logic và tính thực tiễn. Ở phần hai tác giả đã cố gắng để làm rõ

cơ sở lý luận của các kiểu bố trí thí nghiệm, phương pháp phân tích số liệu giúp cho người đọc có thể nắm bắt được và ứng dụng để bố trí và xử lý số liệu các thí nghiệm trong chậu, trong phòng và thí nghiệm đồng ruộng.

Mặc dù ngày càng có nhiều phần mềm tính toán ra đời làm cho việc xử lý các số liệu tiến hành nhanh chóng, nhưng những hiểu biết về cơ sở của các phép tính toán là rất quan trọng, nó giúp cho việc kiểm tra các kết quả tính toán, phân tích và đánh giá đúng các hiện tượng trong nghiên cứu, tránh những sai sót trong sử dụng các phần mềm thống kê.

Tác giả xin chân thành cảm ơn Thầy Nguyễn Đình Hiền Đại học Nông nghiệp Hà Nội, người đã đóng góp nhiều ý kiến quý báu cho nội dung của cuốn sách.

Không thể tránh khỏi những thiếu sót về nội dung và hình thức, rất mong được sự góp ý của bạn đọc. Mọi góp ý xin gửi về:

Bộ môn Di truyền – Chọn giống
Khoa Nông học, Đại học Nông Lâm Tp. HCM.

hoặc E-mail: ptkiem@hotmail.com
ptkiem1@gmail.com

Xin giới thiệu cùng bạn đọc.

Tác giả

Phần 1
XỬ LÝ SỐ LIỆU ĐIỀU TRA KHẢO SÁT

Chương 1
THỐNG KÊ MÔ TẢ - CÁC THAM SỐ THỐNG KÊ

Để nghiên cứu các đối tượng, công việc đầu tiên là điều tra, thu thập số liệu và dùng các tham số thống kê để mô tả đối tượng nghiên cứu. Chương này sẽ đề cập đến các vấn đề:

- Tổng thể và mẫu;

- Các tham số đặc trưng của mẫu và tổng thể.

1.1. TỔNG THỂ VÀ MẪU

1.1.1. Tổng thể

1.1.1.1. Khái niệm

Theo quan điểm thống kê, tổng thể nghiên cứu hay tổng thể là toàn bộ các phần tử hay cá thể có cùng một hay một số đặc trưng (dấu hiệu) định tính hay định lượng nào đó của đối tượng nghiên cứu.

Trong nông học, một tổng thể có thể là một quần thể cây trồng gồm nhiều cá thể. Một tổng thể cũng có thể là một nhân tố cụ thể liên quan đến cây trồng cần được nghiên cứu như một khu đất canh tác khi giả thiết rằng nó bao gồm vô số mẫu đất cần được khảo sát, đánh giá.

Số lượng các phần tử hay cá thể (dưới đây được gọi chung là cá thể) trong tổng thể được gọi là kích thước, cỡ hay dung lượng (dưới đây được gọi là dung lượng) tổng thể,

ký hiệu là N. Thường thì dung lượng tổng thể là một số hữu hạn, nhưng nếu tổng thể quá lớn hoặc không thể nắm được toàn bộ các cá thể, ta có thể coi dung lượng của tổng thể là vô hạn. Điều này dựa trên cơ sở, rằng khi dung lượng của tổng thể tăng lên khá lớn thì ảnh hưởng không đáng kể đến kết quả tính toán cho tổng thể từ số liệu thu được trên từng bộ phận rút ra từ tổng thể đó.

1.1.1.2. Các loại dấu hiệu của tổng thể

Có thể chia các dấu hiệu tổng thể thành hai loại: các dấu hiệu định tính và các dấu hiệu định lượng.

- Các dấu hiệu định tính, còn được gọi là các dấu hiệu về chất (hay dấu hiệu chất lượng) là các dấu hiệu có thể phân biệt sự khác nhau giữa các cá thể hay nhóm cá thể bằng mắt, nếm hay thử. Ví dụ như có lông, râu hoặc không có, màu vàng hay màu xanh, hạt trần hay có màng, tròn hay dài, trơn hay nhăn, nhiễm hay kháng bệnh v.v. Đối với loại dấu hiệu này người ta có phương pháp nghiên cứu riêng biệt.

- Các dấu hiệu định lượng, còn được gọi là các dấu hiệu về lượng (hay dấu hiệu số lượng) là các dấu hiệu không thể phân biệt sự khác nhau giữa các cá thể hay nhóm cá thể bằng mắt, mà phải tiến hành cân, đo, đếm và phân biệt được nhờ sử dụng các phép toán thống kê. Ví dụ như khối lượng hạt, củ, quả, thân, rễ, độ lớn, độ dài của các bộ phận, số lượng hạt, củ, quả, v.v.

Sự phân chia này có tính tương đối vì bất kỳ một dấu hiệu chất lượng nào cũng có thể lượng hóa bằng các mức độ khác nhau, và có nhiều dấu hiệu số lượng cũng có thể phân biệt bằng mắt được như to, trung bình hay nhỏ, cao, trung bình hay thấp, dài hay ngắn, nhiều hay ít.

1.1.1.3. Các phương pháp mô tả tổng thể

- ***Bằng bảng phân bố tần số***

Nếu gọi các trị số x_i nhận được từ phép xác định nào đó và n_i $(i=\overline{1,n})$ là các tần số (n_i là số cá thể của tổng thể có cùng trị số x_i) thì tổng thể có thể mô tả:

Trị số	x_1	x_2	x_3	...	x_i	...	x_n
Tần số	n_1	n_2	n_3	...	n_i	...	n_n

Hiển nhiên

$$\begin{cases} 0 \le n_i \le N, \text{ với } \forall i \\ \sum_{i=1}^{k} n_i = N \end{cases}$$

- ***Bằng liệt kê bảng phân bố tần suất***

Nếu ký hiệu p_i $(i=\overline{1,k})$ là tần suất của x_i, $p_i = \frac{n_i}{N}$ $(i=\overline{1,k})$ thì tổng thể có thể mô tả:

Trị số	x_1	x_2	x_3	...	x_i	...	x_n
Tần suất	p_1	p_2	p_3	...	p_i	...	p_n

với:

$$\begin{cases} 0 \le p_i \le 1, \text{ với } \forall i \\ \sum_{i=1}^{k} p_i = 1 \end{cases}$$

- ***Bằng bảng ghép***

Trị số	x_1	x_2	x_3	...	x_i	...	x_n
Tần số	n_1	n_2	n_3	...	n_i	...	n_n
Tần suất	p_1	p_2	p_3	...	p_i	...	p_n

Đây là những phương pháp mô tả các dấu hiệu lấy các trị số rời rạc.

- ***Bằng tần suất tích lũy***

Nếu w_i $(i = \overline{1,k})$ là tần số tích lũy của các $x_j < x_i$ thì:

$$w_i = \sum_{x_j < x_i} N_j$$

và $f(x_i)$ là tần suất tích lũy của các $x_j < x_i$ thì:

$$f(x_i) = \frac{w_i}{N} = \sum_{x_j < x_i} \frac{N_j}{N}$$

Tần suất tích lũy là một hàm của x_i có tính chất giống như hàm phân phối xác suất của đại lượng ngẫu nhiên rời rạc.

- ***Bằng đồ họa***

Để mô tả tổng thể, từ kết quả điều tra mẫu người ta xây dựng các loại đồ thị, các loại biểu đồ thực nghiệm và tổng thể.

Như vậy, việc mô tả tổng thể bằng bảng phân bố tần số, bảng phân bố tần suất, tần suất tích lũy hay đồ họa cho thấy những dấu hiệu định lượng hoàn toàn có thể mô hình hóa bằng một đại lượng ngẫu nhiên rời rạc. Điều đó cũng đúng cho các tổng thể có dấu hiệu phân phối liên tục.

1.1.2. Mẫu

1.1.2.1. Khái niệm

Mẫu là một bộ phận hữu hạn của tổng thể gồm *n* cá thể ($n < N$) được gọi là dung lượng mẫu, trên đó người ta tiến hành điều tra, khảo sát, đo đếm và thu thập các số liệu.

Từ các số liệu thu thập được, người ta sử dụng các thuật toán theo lý thuyết xác suất để suy đoán những hiện tượng, quy luật của tổng thể. Nội dung chính của sự suy đoán này là:

- Ước lượng các tham số của tổng thể thông qua các tham số của mẫu và kiểm định độ tin cậy của các tham số.

- Tìm hiểu mối quan hệ giữa các dấu hiệu nghiên cứu trong tổng thể thông qua mối quan hệ giữa các dấu hiệu trong mẫu và kiểm định độ tin cậy về mối quan hệ.

1.1.2.2. Các phương pháp chọn mẫu

Để việc suy đoán có độ chính xác cao, các mẫu được rút ra để nghiên cứu phải đại diện được cho toàn bộ các cá thể trong tổng thể.

- ***Với tổng thể thuần nhất***

Với loại tổng thể này, áp dụng các phương pháp rút mẫu sau đây.

Rút ngẫu nhiên trực tiếp từ tổng thể

Đây là cách chọn mẫu một cách ngẫu nhiên có hoàn lại và không hoàn lại. Thông thường, có 4 phương pháp chọn ngẫu nhiên:

- Rút mẫu ngẫu nhiên đơn giản: Mỗi cá thể trong tổng thể đều có cơ hội như nhau trong lựa chọn. Các cá thể được quy định trước theo một thứ tự nào đó (có thể đánh số trực tiếp hay quy ước), sau đó tiến hành bốc thăm.

- Rút ngẫu nhiên hệ thống: Quy định lấy mẫu ở các vị trí nào đó được định trước. Đây cũng coi như là phép lấy mẫu ngẫu nhiên, bởi vì cá thể được chọn đứng ở vị trí đó là ngẫu nhiên, trước khi lấy mẫu điều tra, ta cũng không hề biết tình trạng của cá thể này. Người ta có thể định vị

trí lấy mẫu trên đường chéo góc, trên đường dích dắc hay các kiểu quy định nào đó. Ví dụ: trong quy phạm khảo nghiệm giống ngô, người ta quy định theo dõi 10 cây/1 giống ở mỗi lần nhắc lại, lấy 5 cây liên tiếp nhau từ cây thứ 5 đến cây thứ 9 tính từ đầu hàng thứ 2 và từ cây thứ 5 đến cây thứ 9 tính từ cuối hàng thứ 3 của ô.

- Dùng bảng số ngẫu nhiên: Có thể sử dụng các bảng số ngẫu nhiên sau để chọn mẫu: Bảng Tippett (các số có 4 chữ số), bảng Fisher và Yates, các bảng của Kendall và Babington Smith (các số có 4 chữ số), bảng của Burke Haton.

- Dùng phần mềm Excel (theo cú pháp ghi ở chương 8).

Chọn cá thể điển hình trực tiếp từ tổng thể

Đây là phương pháp chọn mẫu không ngẫu nhiên. Từ quan sát tổng thể, chọn các cá thể điển hình, đại biểu cho tổng thể theo mục tiêu nghiên cứu.

Rút từ các phần của tổng thể (chia nhóm rồi chọn mẫu)

Người ta chia tổng thể thành các nhóm một cách cơ giới theo một quy tắc nào đó, từ mỗi nhóm lấy ra một số cá thể theo một cách thống nhất để nghiên cứu.

- ***Với tổng thể không thuần nhất***

Có những tổng thể không có từng cá thể điển hình mà chỉ có tập hợp mẫu điển hình. Ví dụ, tổng thể là quần thể phân ly được tạo ra từ phép lai hay tác nhân đột biến hoặc là quần thể tạo được từ kỹ thuật di truyền. Để nghiên cứu chúng ta không thể áp dụng phương pháp chọn từng cá thể điển hình. Tốt nhất là theo dõi toàn thể quần thể hoặc lấy một bộ phận liên tục có dung lượng mẫu lớn (nếu quần thể quá lớn), hoặc sử dung một trong 4 phương pháp chọn ngẫu nhiên đã trình bày trong mục 1.2.1 trên đây.

1.2. CÁC THAM SỐ ĐẶC TRƯNG CỦA MẪU VÀ TỔNG THỂ

1.2.1. Các tham số đặc trưng cho sự tập trung

1.2.1.1. Số cực trị:

Số cực trị là số bé nhất và lớn nhất trong mẫu, ký hiệu là X_{min} và X_{max}.

1.2.1.2. Mốt

Mốt là trị số có tần số cao nhất trong một mẫu. Nếu mẫu đã phân tổ thì tổ mốt là tổ có tần số cao nhất và trị số giữa của tổ mốt là trị số mốt của mẫu.

Trong một tổng thể quan sát nhiều mẫu, mỗi mẫu gồm một số cá thể xác định, khi theo dõi một chỉ tiêu nào đấy ta nhận được trị số mốt của các mẫu xấp xỉ bằng nhau thì tổng thể đó đồng nhất theo chỉ tiêu này, ngược lại nếu trị các trị số mốt của các mẫu khác nhau thì tổng thể đó không đồng nhất. Nếu các chỉ tiêu khác cũng cho kết quả tương tự, ta có thể đánh giá được tính đồng nhất hay không đồng nhất của tổng thể. Người ta thường áp dụng tính chất này để đánh giá độ thuần của giống và mức độ đồng đều của đất.

1.2.1.3. Trung bình và kỳ vọng

Trung bình (trung bình mẫu hay trung bình thực nghiệm), thường ký hiệu là $\overline{X}$, là tham số đặc trưng cho sự tập trung của mẫu và kỳ vọng (trung bình tổng thể hay trung bình lý luận), thường ký hiệu là E(X), MX, μ hay m, là tham số đặc trưng cho sự tập trung của tổng thể.

Bản chất của trị trung bình các giá trị quan sát là gần bằng kỳ vọng, nó phản ánh giá trị trung tâm của phân

phối xác suất của đại lượng ngẫu nhiên. Vì vậy, người ta thường sử dụng trị trung bình của mẫu để ước lượng kỳ vọng của tổng thể.

$$E(X) = \mu$$

Khi dung lượng càng lớn, trị trung bình càng gần với kỳ vọng, vì vậy để ước lượng đúng kỳ vọng, dung lượng mẫu phải đủ lớn.

Trong thực nghiệm, khi x_i lấy các trị số rời rạc, $\overline{X}$ được tính theo các công thức sau:

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Ví dụ, nếu có các số đo x_i là:

20	24	24	23	25	14	21	20	31	16
18	21	19	20	19	13	20	24	18	20

thì $\overline{X}$ = (20 + 24 + 24 + 23 + ... + 18 + 20) : 20 = 20,5

Nếu x_i lấy n_i lần với $n = \sum n_i$ thì $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i n_i$

Nếu xác suất bắt gặp của x_i là p_i ($p_i = n_i / n$) và k là số nhóm x_i thì $\overline{X} = \sum_{i=1}^{k} x_i p_i$. Ví dụ, nếu các số đo x_i có n_i lần bắt gặp với xác suất p_i như sau:

x_i:	17	18	19	20	21	22	23	24	25	26
n_i:	2	5	8	10	20	16	15	14	7	3
p_i:	0,02	0,05	0,08	0,10	0,20	0,16	0,15	0,14	0,07	0,03

thì $\overline{X} = \frac{1}{100}$ [(17 × 2) + (18 × 5) + ... + (26 × 0,3)] = 21,82

hoặc $\overline{X}$ = (17 × 0,02) + (18 × 0,05) + ... + (26 × 0,03) =

21,82 Khi biết các x_i và n_i thì tính theo công thức $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i n_i$, còn khi chỉ biết x_i và p_i thì tính theo công thức $\bar{X} = \sum_{i=1}^{k} x_i p_i$.

Với X là đại lượng ngẫu nhiên liên tục:

$$E(X) = \mu = \int_{-\infty}^{\infty} x f(x) dx$$

Các tính chất của kỳ vọng:

1. Kỳ vọng của một hằng số C bằng chính hằng số đó:

$$E(C) = C$$

2. Kỳ vọng của tích giữa một hằng số và một đại lượng ngẫu nhiên bằng tích của hằng số với kỳ vọng của đại lượng ngẫu nhiên đó:

$$E(CX) = CE(X)$$

3. Kỳ vọng của tổng một hằng số C với một đại lượng ngẫu nhiên bằng tổng của hằng số với kỳ vọng của đại lượng ngẫu nhiên đó:

$$E(X + C) = E(X) + C$$

4. Kỳ vọng của tổng các đại lượng ngẫu nhiên bằng tổng các kỳ vọng thành phần:

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

5. Kỳ vọng của tích hai đại lượng ngẫu nhiên độc lập bằng tích của hai kỳ vọng của hai đại lượng ngẫu nhiên đó:

$$E(X_1.X_2) = E(X_1).\ E(X_2)$$

Tất cả các tính chất này đều đúng cho số trung bình thực nghiệm.

1.2.2. Các tham số đặc trưng cho độ phân tán của các dấu hiệu định lượng

1.2.2.1. Khoảng biến thiên

Khoảng biến thiên là khoảng cách giữa hai cực trị:

$$R = X_{max} - X_{min}$$

1.2.2.2. Phương sai mẫu, phương sai tổng thể và độ lệch chuẩn

- ***Phương sai mẫu và phương sai tổng thể***

Trung bình và kỳ vọng chỉ là một số bình quân của đại lượng ngẫu nhiên của mẫu và tổng thể. Do khoảng biến thiên R chỉ đo khoảng cách từ hai trị số lớn nhất và nhỏ nhất, chưa xét đến các giá trị khác, vì vậy khoảng biến thiên không đặc trưng cho độ phân tán của mẫu hay tổng thể xung quanh trị bình quân. Hãy xét hai mẫu sau đây:

Mẫu 1:

20	24	24	23	25	14	21	20	31	16
18	21	19	20	19	13	20	24	18	20

Mẫu 2:

26	25	29	14	23	13	14	22	28	24
15	31	14	13	29	16	28	14	17	15

Hai mẫu này cùng có trị trung bình và khoảng biến thiên bằng nhau ($\overline{X} = 20,5$; $R = 18$) nhưng không thể nói hai mẫu giống nhau do độ đồng đều của hai mẫu khác nhau

rõ ràng, tức là độ phân tán của các số đo so với trị trung bình của từng mẫu khác nhau. Vậy tham số nào đặc trưng cho độ phân tán của các số trong mẫu xung quanh trị trung bình của chúng.

Nếu $(X - \overline{X})$ là độ lệch của mỗi số X với số trung bình $\overline{X}$, theo tính chất 3 và 1 của kỳ vọng, ta có:

$$E[X - E(X)] = E(X) - E[E(X)]$$
$$= E(X) - E(X) = 0$$

tức là: trung bình độ lệch từ mỗi giá trị X với trung bình mẫu luôn bằng không. Nói cách khác: do tổng đại số các độ lệch từ mỗi giá trị của mẫu với trung bình mẫu luôn bằng 0 nên trung bình độ lệch cũng luôn bằng 0. Vì vậy trung bình độ lệch không phản ánh độ phân tán.

Người ta sử dụng tổng bình phương độ lệch và trung bình bình phương để nghiên cứu độ phân tán.

Tổng bình phương độ lệch $\sum_{i=1}^{n}[X - M(X)]^2 = 0$ khi mọi X đều bằng nhau và $\sum_{i=1}^{n}[X - M(X)]^2$ càng tăng khi các giá trị X càng khác nhau.

Trung bình bình phương thực nghiệm, còn gọi là phương sai mẫu hay phương sai, ký hiệu là MS (Mean Square), S^2, s^2 hay V(X) hoặc Var(X), là tham số đặc trưng cho độ phân tán của các cá thể trong mẫu theo dấu hiệu nghiên cứu và trung bình bình phương lý luận, còn gọi phương sai tổng thể, thường ký hiệu là V(X) hoặc Var(X), DX, σ_X^2 hay σ^2 (nói chung), là tham số đặc trưng cho độ phân tán của các cá thể trong tổng thể.

Bản chất của phương sai mẫu là trung bình số học của bình phương các độ lệch giữa các giá trị của đại lượng ngẫu nhiên so với trị trung bình, phản ánh mức độ phân tán của các giá trị quan sát của đại lượng ngẫu nhiên xung quanh giá trị trung bình của chúng. Nếu trị trung bình mẫu dùng để ước lượng kỳ vọng của tổng thể thì phương sai mẫu dùng để ước lượng phương sai tổng thể. Khi dung lượng mẫu càng lớn, phương sai mẫu càng gần với phương sai tổng thể, vì vậy để ước lượng đúng phương sai tổng thể, dung lượng mẫu phải đủ lớn.

$$V(X) = E[X - E(X)]^2 = \sigma_X^2$$

Phương sai có đơn vị đo là bình phương đơn vị đo của đại lượng ngẫu nhiên.

Trong thực nghiệm, khi x_i lấy các giá trị rời rạc, V(X) được tính theo các công thức sau:

$$V(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right).$$

Ở mẫu 1:

$$V(X) = \frac{1}{20-1}[(20 - 20{,}5)^2 + (24 - 20{,}5)^2 +$$

$$... + (20 - 20{,}5)^2] = 16{,}37$$

Tương tự, phương sai ở mẫu 2 là: V(x) = 42,789

Khi x_i lấy n_i lần (như ví dụ sau trong mục 1.2.1.3), công thức tính phương sai có dạng:

$$V(x) = \frac{1}{n-1}\left[n\sum_{i=1}^{k} n_i x_i^2 - (\sum_{k=1}^{k} n_i x_i)^2 \right]$$
$$= \frac{1}{n(n-1)}\left[n\sum_{i=1}^{k} x_i^2 p_i - (\sum_{k=1}^{k} x_i p_i)^2 \right] \text{ với } n = \sum n_i .$$

Kết quả tính được: V(x) = 4,452.

Với X là đại lượng ngẫu nhiên liên tục:

$$V(X) = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$$

Các tính chất của phương sai:

1. Phương sai của một hằng số C thì bằng 0:

$$V(C) = 0$$

Thật vậy: $V(C) = E[C - E(C)]^2 = E[C - C]^2 = E(0) = 0$

2. Phương sai của tích một hằng số và một đại lượng ngẫu nhiên bằng tích giữa bình phương hằng số và phương sai của đại lượng ngẫu nhiên đó:

$$V(CX) = C^2V(X)$$

Thật vậy:

$$V(CX) = E[CX - E(CX)]^2 = E[CX - CE(X)]^2$$
$$= E\{C^2[X - E(X)]^2\} = C^2E[X - E(X)]^2$$
$$= C^2V(X)$$

3. Phương sai của tổng một hằng số C với một đại lượng ngẫu nhiên thì bằng chính phương sai của đại lượng ngẫu nhiên đó. Nói cách khác nếu cộng một hằng số C với một đại lượng ngẫu nhiên thì phương sai không đổi:

$$V(X + C) = V(X)$$

Thật vậy:

$$V(X + C) = E[(X + C) - E(X + C)]^2$$
$$= E[(X + C) - E(X) - C]^2$$
$$= E[X - E(X)]^2 = V(X)$$

4. Phương sai của một tổng hai đại lượng ngẫu nhiên độc lập bằng tổng các phương sai thành phần:

$$V(X_1 + X_2) = V(X_1) + V(X_2)$$

Thật vậy:

$$V(X_1 + X_2) = E[(X_1 + X_2) - E(X_1 + X_2)]^2$$
$$= E\{[X_1 - E(X_1)] - [X_2 - E(X_2)]\}^2$$
$$= E\{[X_1 - E(X_1)]^2 + 2[X_1 - E(X_1)].[X_2 - E(X_2)] + [X_2 - E(X_2)]^2\}$$
$$= E[X_1 - E(X_1)]^2 + 2E\{[X_1 - E(X_1)].[X_2 - E(X_2)]\} + E[X_2 - E(X_2)]^2$$
$$= V(X_1) + V(X_2) + 2E[X_1X_2 - X_1E(X_1) - X_2E(X_1) + E(X_1)E(X_2)]$$
$$= V(X_1) + V(X_2) + 2[E(X_1X_2) - E(X_1)E(X_2) - E(X_2)E(X_1)] + E(X_1)E(X_2)]$$
$$= V(X_1) + V(X_2) + 2[E(X_1)E(X_2) - E(X_1)E(X_2)] = V(X_1) + V(X_2)$$

Hệ quả:

1. $$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i)$$

2. $$V(X_1 - X_2) = V(X_1) + V(X_2)$$

Độ lớn nhỏ của phương sai không phụ thuộc vào độ lớn của số trung bình. Các tập hợp mẫu có cùng trị trung bình nhưng phương sai có thể khác nhau (trong so sánh mẫu 1 và mẫu 2 trên đây) và các số trung bình có thể khác nhau nhưng phương sai có thể bằng nhau (tính chất 3 của phương sai).

Như vậy, phương sai đặc trưng cho độ phân tán (hay là độ khác biệt giữa các số). Khi các số càng gần bằng nhau thì phương sai càng nhỏ, mẫu càng đồng đều, ngược lại, khi các số càng khác xa nhau thì phương sai càng lớn, mẫu càng kém đồng đều.

Phương sai cũng như các tính chất của nó được sử dụng như là phương pháp hữu hiệu trong nhiều phép phân tích, đánh giá các số liệu thu thập (sẽ được đề cập ở các phần sau).

Như trên đã nói, μ và σ^2 là kỳ vọng và phương sai tổng thể, người ta đã chứng minh được (không dẫn):

- Kỳ vọng của trung bình mẫu bằng trung bình tổng thể:

$$E(\bar{X}) = \mu$$

- Kỳ vọng phương sai mẫu bằng phương sai tổng thể:

$$E[V(X)] = E(S^2) = \sigma^2$$

Do đó người ta lầy $\bar{X}$ để ước lượng μ và lấy S^2 để ước lượng σ^2 ở mức tin cậy nào đó.

- ***Độ lệch chuẩn***

Độ lệch chuẩn mẫu, ký hiệu là S hay sd (standard deviation) là căn bậc hai của phương sai ($S = \sqrt{S^2}$), còn độ lệch chuẩn tổng thể, ký hiệu là σ_x hay σ (nói chung) là căn

bậc hai của phương sai tổng thể ($\sigma = \sqrt{\sigma^2}$).

Đơn vị tính của độ lệch chuẩn là đơn vị đo của đại lượng ngẫu nhiên.

Độ lệch chuẩn phản ánh mức sai lệch trung bình của các cá thể xung quanh trị trung bình. Mẫu và tổng thể càng đồng đều, S và σ càng bé và ngược lại.

1.2.2.3. Hệ số biến động

Do độ lệch chuẩn là một số tuyệt đối không phụ thuộc vào số trung bình nên không phản ánh mức độ biến động xung quanh trị trung bình. Hai mẫu có cùng độ lệch chuẩn không thể coi chúng biến động như nhau khi chúng có hai trị trung bình khác nhau. Người ta dùng hệ số động (ký hiệu là CV – Coefficient of Variation) để đánh giá mức sai lệch lớn hay nhỏ so với trung bình của nó và được tính bằng %:

$$CV(\%) = \frac{S}{\overline{X}}.100$$

Hệ số biến động được sử dụng trong các trường hợp sau:

- Đánh giá độ biến động của các cá thể trong mẫu và tổng thể theo một chỉ tiêu nào đó, ví dụ chiều cao cây, chiều dài của các bộ phận, khối lượng hạt, củ, quả, số lượng hạt, củ, quả. Để đánh giá độ đồng đều của hạt của một giống, sau khi lấy mẫu phân tích, người ta đếm và cân ít nhất 8 mẫu, mỗi mẫu 100 hạt hay 1.000 hạt (tùy hạt lớn hay nhỏ), rồi tính CV(%). Nếu hạt đồng đều, khối lượng các mẫu ít khác biệt nhau và độ biến động thấp. Nếu CV(%) ≤ 5 là biến động ít – hạt đồng đều, 6 – 10 là biến động vừa phải – hạt tương đối đều và > 10 là biến động nhiều và rất nhiều – hạt không đều.

- Đánh giá sự khác nhau giữa các nhóm cá thể (quần

thể) như: giữa các giống, giữa các nghiệm thức theo đặc trưng nào đó. Giá trị hệ số biến động càng cao chứng tỏ chúng càng khác biệt nhau.

- Chọn ruộng (đất) thí nghiệm. Khi chưa biết được lịch sử canh tác của khu đất, có thể chọn đất thí nghiệm bằng cách lấy mẫu đất, phân tích và đánh giá nhanh sự đồng nhất bằng phương pháp phi tham số hoặc phương pháp so sánh phương sai một số chỉ tiêu chính giữa các lô lấy mẫu trong khu đất. Tuy nhiên có thể chọn đất bằng cách thực hiện một "thí nghiệm trắng". Gọi là "thí nghiệm trắng" vì thí nghiệm không nghiên cứu điều gì ngoài việc chọn đất. Trong thí nghiệm này, người ta sử dụng chỉ dùng một giống để gieo lên các ô đã được thiết kế theo kiểu CRD hoặc RCBD cho các "nghiệm thức" giả định, thu năng suất từng ô như một thí nghiệm thông thường, hoặc là lấy mẫu năng suất từ ruộng đã được gieo trồng sẵn một giống nào đó. Nếu kết quả phân tích số liệu cho thấy giữa các "nghiệm thức" không khác biệt nhau và hệ số biến động của sai số ≤ 10% (càng nhỏ càng tốt) thì đất này sẽ được chọn làm thí nghiệm.

- Đánh giá độ chính xác (ít sai số) của một thí nghiệm. Trong bảng phân tích phương sai (ANOVA), CV phản ánh độ biến động do sai số gây ra:

$$CV(\%) = \frac{\sqrt{MS_e}}{\bar{X}}.100$$

Do $\sqrt{MS_e}$ là độ lệch chuẩn của sai số nên CV(%) càng nhỏ thí nghiệm càng chính xác.

1.2.2.4. Hệ số nhọn của phân phối xác suất

Hệ số nhọn của phân phối xác suất, ký hiệu là α_4, cho

thấy các giá trị x_i của đại lượng biến thiên tập trung nhiều hay ít xung quanh kỳ vọng, tương ứng với phương sai nhỏ hay lớn.

$$\alpha_4 = \frac{\mu^4}{\sigma^4}$$

trong đó: μ^4 là mô men trung tâm bậc 4: $\mu^4 = E[X - E(X)]^4$

σ^4 là bình phương của phương sai

Nếu $\mu^4 = 3$ thì đồ thị phân phối xác suất là bình thường, nếu $\mu^4 > 3$ thì đồ thị nhọn (các x_i tập trung nhiều xung quanh kỳ vọng μ), còn nếu $\mu^4 < 3$ thì đồ thị tù (không nhọn).

Với ví dụ ở mục 2.2.2 trên đây:

mẫu 1: $\mu^4 = 973,80$, $\sigma^4 = 267,91$ và $\mu^4 = 3,6$

mẫu 2: $\mu^4 = 2.436,91$, $\sigma^4 = 1.830,90$ và $\mu^4 = 1,3$

Rõ ràng mẫu 1 số liệu rất tập trung còn mẫu 2 số khá rải rác mặc dù trung bình mẫu đều bằng nhau.

1.2.3. Các tham số đặc trưng cho độ phân tán của các dấu hiệu định tính

Các dấu hiệu định tính, còn gọi là các dấu hiệu chất lượng, thường được biểu thị dưới dạng tần suất. Với loại dấu hiệu này, để biết được độ phân tán người ta đánh giá độ lệch chuẩn của các tần suất phân phối ở các mức chất lượng khác nhau và hệ số biến động biểu thị mức độ sai khác (%) giữa độ lệch chuẩn và độ lệch chuẩn cao nhất.

Hãy xét ví dụ ở Bảng 1.1 sau đây.

Theo mức độ lông của lá, Bảng 1.1 cho thấy có 7,1% số giống không hay rất ít lông, 14,3% - ít lông, 21,4% -

lông vừa, 28,6% - lông nhiều, 28,6% - lông rất nhiều. trong 20 giống kháng rầy có 5% thuộc loại ít lông, 20% thuộc loại lông vừa, 35% thuộc loại nhiều lông, còn 40% thuộc loại rất nhiều lông.

Để đánh giá độ lệch chuẩn biểu thị mức độ khác nhau về các tần suất theo độ lông của các giống ta dùng công thức:

Bảng 1.1: Kết quả điều tra mức độ rầy xanh hại trên 28 giống bông tại Đại học Nông Lâm Tp. HCM, 2009

Mức độ lông của lá	Giống có lông		Giống kháng	
	Số giống (n_i)	Tần suất (p_i)	Số giống kháng (n_i)	Tần suất (p_i)
Không hay rất ít	2	0,071	0	0,000
Ít	4	0,143	1	0,050
Vừa	6	0,214	3	0,150
Nhiều	8	0,286	7	0,350
Rất nhiêu	8	0,286	8	0,400
Tổng	28	1,000	20	1,000

$$S_p = \sqrt[k]{p_1.p_2....p_k} \qquad (1-1)$$

trong đó, S_p là độ lệch chuẩn, p_1, p_2, ..., p_k là tần suất các nhóm chất lượng ($\Sigma p_i = 1$) và k là số nhóm.

Theo Bảng 1.1, về cơ cấu giống có lông:

$$S_p = \sqrt[5]{0,071 \times 0,143 \times 0,214 \times 0,286 \times 0,286}$$

= 0.178 (hay 17,8%)

và hệ số biến động được tính theo công thức:

$$CV(\%) = \frac{S_p}{S_{p\max}}.100 \qquad (1 - 2)$$

Trong ví dụ này là: $CV(\%) = \frac{0,178}{0,200}.100 = 89,0$, tức là biến động rất nhiều.

Cách xác định giá trị $S_{p\ max}$ ở công thức (1 - 2) khá dễ dàng vì: theo công thức (1 - 1), S_p lấy giá trị cao nhất khi các các tần suất p_i của các nhóm bằng nhau. Do $\Sigma p_i = 1$ nên $S_{p\ max} = 1/k$, vì vậy ở ví dụ này $1/5 = S_{p\ max} = 0,20$.

Hãy tính S_p cho cơ cấu tính kháng.

Với số liệu Bảng 1.1, ta chỉ có 4 nhóm chất lượng: ít, vừa, nhiều và rất nhiều, tần suất của 4 nhóm này theo thứ tự là 0,050; 0,150; 0,350 và 0,400.

Một cách tương tự ta có: $S_p = 0,180$ và $CV(\%) = 72,0$.

So sánh với chỉ tiêu cơ cấu giống có lông, độ lệch chuẩn của cơ cấu tính kháng lớn hơn nhưng hệ số biến động thấp hơn vì nó có ít nhóm hơn nên độ lệch chuẩn tối đa lớn hơn (0,25 so 0,20), thành thử CV(%) nhỏ hơn.

Về việc xác định nhóm, ở cột đầu có 5 nhóm theo độ lông khác nhau, chỉ tiêu cơ cấu giống có lông có tần số và tần suất theo 5 nhóm này, nhưng với chỉ tiêu cơ cấu giống kháng tất cả các giống kháng (20 giống) chỉ nằm trong 4 nhóm với $\Sigma p_i = 1$. Do nhóm không và rất ít lông có $n_i = 0$ nên với chỉ tiêu này chỉ xét cho 4 nhóm.

Trong trường hợp chỉ có 2 nhóm ($k = 2$), khi đó $p_1 + p_2 = 1$, tức là $p_1 = 1 - p_2$, $0 < S_p \leq 0,5$ và $S_{p\ max} = 0,5$.

Việc tính toán trên đây khá dễ dàng nhờ sự trợ giúp của phần mềm Excel trên máy điện toán.

Một cách tính toán khác có thể được áp dụng là biến công thức (1– 1) thành: $\log S_p = \frac{1}{k}(\log p_1 + \log p_2 + ... + \log p_k)$, sau khi tính được $\log S_p$ ta sẽ có được S_p.

Cách tính như sau, ví dụ cho chỉ tiêu giống có lông:

$$\log S_p = \frac{1}{5}(\log 0,071 + \log 0,143 + \log 0,214 + \log 0,286 + \log 0,286)$$

$$= \frac{1}{5}(-1,14874 - 0,84466 - 0,66958 - 0,54363 - 0,54363)$$

= - 0,75005. Từ đó: S_p = 0,178 (hay 17,8%)

Như vậy, với các đặc trưng định tính có thể đánh giá được độ phân tán của các tần suất các nhóm chất lượng qua độ lệch chuẩn và và hệ số biến động khi so sánh với độ lệch chuẩn cao nhất.

1.2.4. Các tham số đặc trưng cho mối quan hệ giữa các đại lượng ngẫu nhiên

1.2.4.1. Hiệp phương sai

Hiệp phương sai, thường ký hiệu là Cov(X,Y), Covar(X,Y) hoặc W(X,Y), là kỳ vọng của tích các độ lệch của các đại lượng ngẫu nhiên với kỳ vọng (hay trung bình thực nghiệm) của chúng, biểu thị mức độ quan hệ giữa hai đại lượng ngẫu nhiên và được tính theo công thức:

$$W(X,Y) = E\{[(X - E(X)][(Y - E(Y)]\}$$

Hiệp phương sai có đơn vị đo là tích đơn vị đo của các đại lượng ngẫu nhiên X và Y.

Trong thực nghiệm, công thức tính hiệp phương sai giữa biến X và Y được viết:

$$W_{x,y} = \frac{1}{n-1}\sum\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)$$

hay: $$\mathrm{Cov}\left(X,Y\right) = \frac{1}{n-1}\left[\sum XY - \left(\sum X\right)\left(\sum Y\right)/n\right]$$

Các tính chất của hiệp phương sai:

1. Hiệp phương sai của hai đại lượng ngẫu nhiên lấy các giá trị là hằng số thì bằng 0:

$$W(C_1, C_2) = 0$$

Ví dụ:

c_1:	12	12	12	12	12	12	12	12	12	12
c_2:	15	15	15	15	15	15	15	15	15	15

2. Nếu nhân mỗi đại lượng ngẫu nhiên với mỗi hằng số khác nhau, như C_1 và C_2 thì:

$$W(C_1X, C_2Y) = C_1C_2W(X,Y)$$

3. Nếu cộng mỗi đại lượng ngẫu nhiên với mỗi hằng số khác nhau, như C_1 và C_2 thì hiệp phương sai không đổi:

$$W(X + C_1, Y + C_2) = W(X,Y)$$

4. Hiệp phương sai của hai đại lượng ngẫu nhiên độc lập, như X độc lập với Y, thì bằng 0:

$$W(X,Y) = 0 \qquad (X \neq Y)$$

5. Nếu hai đại lượng ngẫu nhiên X, Y có quan hệ, thì:

$$V(X + Y) = V(X) + V(Y) - 2W(X,Y)$$

1.2.4.2. Hệ số tương quan tuyến tính và hệ số hồi quy

Mục này sẽ được trình bày trong chương 4.

Chương 2
ƯỚC LƯỢNG CÁC THAM SỐ

2.1. KHÁI NIỆM

Các tham số thống kê là những thông tin phản ánh bản chất của tổng thể theo một dấu hiệu (chỉ tiêu) nào đó. Thường thì không thể nghiên cứu toàn bộ số cá thể trong tổng thể. Vậy, để tìm hiểu tổng thể ta phải tìm các phương pháp để suy đoán các tham số thống kê của tổng thể.

Phương pháp tiếp cận thường dùng, như trên đã nói, là phương pháp rút mẫu và từ kết quả nghiên cứu mẫu để suy đoán cho tổng thể bằng phép quy nạp thống kê gọi là ước lượng. Kết quả ước lượng là xác định một cách gần đúng giá trị của các tham số thống kê tổng thể ở độ tin cậy nào đó. Có hai phương pháp sử dụng tham số mẫu để ước lượng cho tham số tổng thể là phương pháp ước lượng điểm và phương pháp ước lượng khoảng.

Ước lượng điểm: là phương pháp dùng trị số của hàm ước lượng được tính toán ở mẫu để thay một cách gần đúng cho tham số tổng thể.

Công thức tổng quát của phương pháp ước lượng điểm như sau:

$$\theta = T_n$$

trong đó: - θ là tham số tổng thể cần ước lượng;
- T_n là hàm ước lượng của tham số θ.

Để ước lượng đúng nhất, phải chọn được hàm ước lượng tốt nhất. Muốn vậy, hàm ước lượng này phải thỏa mãn: không chệch, hội tụ và hiệu nghiệm.

- Ước lượng T_n gọi là ước lượng không chệch cho θ nếu $E(T_n) = \theta$.

Ước lượng không chệch cho biết hàm ước lượng T_n không có sai số hệ thống.

- Ước lượng T_n gọi là ước lượng vững cho θ nếu với mọi $\varepsilon > 0$

$$\lim_{n\to\infty} P\{|T_n - \theta| < \varepsilon\} = 1$$

hay $$\lim_{n\to\infty} P\{\theta - \varepsilon < T_n < \theta + \varepsilon\} = 1$$

Ước lượng vững có xác xuất cao khi dung lượng mẫu đủ lớn.

- Ước lượng T_n gọi là ước lượng hiệu quả cho θ nếu T_n là ước lượng không chệch và có phương sai nhỏ nhất so với mọi ước lượng không chệch khác cho θ.

Ước lượng khoảng: là phương pháp mà tham số ước lượng của tổng thể nằm trong một khoảng với một xác suất (hay độ tin cậy) cho trước. Khoảng này xác định được nhờ những kết quả khi nghiên cứu ở mẫu.

Công thức tổng quát của phương pháp ước lượng khoảng như sau:

$$P(G_1 \le \theta \le G_2) = 1 - \alpha$$

trong đó:

- P là xác suất của sự ước lượng cho tham số θ của tổng thể;
- G_1 và G_2 là giới hạn dưới và giới hạn trên của

khoảng ước lượng được xác định từ kết quả quan sát ở mẫu;

- $1 - \alpha$ là mức tin cậy của ước lượng, α thường chọn là 0,05; 0,01 hay 0,001 (mức sai lầm).

Hiệu số $G_2 - G_1$ được gọi là độ dài khoảng ước lượng. Độ dài khoảng ước lượng càng nhỏ thì độ chính xác của ước lượng càng cao và ngược lại.

Nếu ký hiệu $\varepsilon = \frac{G_2 - G_1}{2}$ thì khoảng tin cậy sẽ được viết $(\theta - \varepsilon; (\theta + \varepsilon)$. ε gọi là sai số tới hạn của ước lượng còn được gọi là độ chính xác của ước lượng và $\varepsilon\% = \frac{\varepsilon}{\overline{X}}.100$ gọi là sai số tương đối hay độ chính xác của ước lượng.

Người ta chia phương pháp ước lượng khoảng ra hai trường hợp:

- Ước lượng khoảng một phía (một chiều - one-tail):

Tham số θ của phân phối lý thuyết được nằm trong một khoảng:

$$P(-\infty < \theta < G_2) = 1 - \alpha \quad \text{(nằm phải)}$$

hay $$P(G_1 < \theta < +\infty) = 1 - \alpha \quad \text{(nằm trái)}$$

- Ước lượng khoảng hai phía (hai chiều - two-tail):

$$P(G_1 \le \theta \le G_2) = 1 - \alpha$$

Đó là khoảng tin cậy cần tìm.

Trong thực tế, người ta thường yêu cầu độ tin cậy $1 - \alpha$, chẳng hạn $1 - \alpha = 0{,}95$ nên theo nguyên lý xác suất số lớn, biến cố $(G_1 < \theta < G_2)$ hầu như chắc chắn xẩy ra. Khi tiến hành rút mẫu quan sát, giá trị của G_1 và G_2 ứng với

mẫu sẽ được viết g_1 và g_2 và $P(g_1 < \theta < g_2) = 1 - \alpha$ hay $P(\bar{X} - \varepsilon \leq \theta \leq \bar{X} + \varepsilon) = 1 - \alpha$.

2.2. ƯỚC LƯỢNG TRUNG BÌNH TỔNG THỂ

2.1. Ước lượng điểm trung bình tổng thể

Giả sử có một tổng thể, để ước lượng trị trung bình tổng thể theo biến X, người ta rút ngẫu nhiên một mẫu độc lập với dung lượng mẫu *n* đủ lớn và quan sát được các số đo x_1, x_2, x_3, ..., x_n.

Người ta chứng minh được rằng trị trung bình mẫu $\bar{X} = \sum_{i=1}^{n} x_i$ là trị ước lượng hiệu nghiệm nhất đối với trị trung bình tổng thể (kỳ vọng μ):

$$\lim_{n\to\infty} P\left\{\left|\bar{X} - \mu\right| > \varepsilon\right\} = 0$$

2.2. Ước lượng khoảng trung bình tổng thể

Giả sử đại lượng ngẫu nhiên X tuân theo quy luật phân phối chuẩn $N(\mu, \sigma^2)$ nhưng chưa biết tham số trung bình μ. Để ước lượng μ ta xét các trường hợp sau.

2.2.2.1. Khi đã biết phương sai σ^2 của tổng thể

Khi đó việc ước lượng khoảng μ được tiến hành theo luật phân phối chuẩn tắc N(0,1):

$$U = \frac{\left(\bar{X} - \mu\right)\sqrt{n}}{\sigma} \qquad (2 - 1)$$

trong đó $\bar{X}$ là trung bình mẫu. Do U có phân phối chuẩn tắc nên với độ tin cậy $1 - \alpha$ cho trước có thể tìm được cặp giá trị α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$. Từ đó tìm được hai giá

trị tới hạn tương ứng $u_{1-\alpha_1}$ và u_{α_2} thỏa mãn điều kiện:

$$P(U < u_{1-\alpha_1}) = \alpha_1$$

và $$P(U > u_{\alpha_2}) = \alpha_2$$

Từ đó $P(u_{1-\alpha_1} < U < u_{\alpha_2}) = 1 - (\alpha_1 + \alpha_2) = 1 - \alpha$

Vì $u_{\alpha_1} = - u_{1-\alpha_1}$ nên có thể viết

$$P(-u_{\alpha_1} < U < u_{\alpha_2}) = 1 - \alpha \qquad (2 - 2)$$

Thay (2 - 1) vào (2 - 2) và giải ra μ ta có:

$$P(\bar{X} - u_{\alpha_2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + u_{\alpha_1}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$(\bar{X} - u_{\alpha_2}\frac{\sigma}{\sqrt{n}}; \bar{X} + u_{\alpha_1}\frac{\sigma}{\sqrt{n}})$ là khoảng ước lượng của μ với độ tin cậy $1 - \alpha$. Do có vô số cặp α_1 và α_2 thỏa mãn $\alpha_1 + \alpha_2 = \alpha$ và vì vậy có vô số khoảng ước lượng. Trong thực tế người ta thường sử dụng một số trường hợp sau để ước lượng:

- Ước lượng khoảng đối xứng:

Nếu lấy $\alpha_1 = \alpha_2 = \alpha/2$ thì:

$$P(\bar{X} - u_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha \quad (2 - 3)$$

$\varepsilon = u_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ được gọi là sai số tới hạn hay độ chính xác của ước lượng và khoảng tin cậy $(\bar{X} - u_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{X} + u_{\alpha/2}\frac{\sigma}{\sqrt{n}})$ có thể viết:

$$I = 2\varepsilon = 2u_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad (2 - 4)$$

Công thức (2 - 4) cho thấy:

- Khi tăng dung lượng mẫu lên và giữ nguyên độ tin cậy 1 – α cho trước thì ε sẽ giảm xuống, độ chính xác của ước lượng tăng lên.

- Khi tăng độ tin cậy 1 – α lên và giữ nguyên dung lượng mẫu thì giá trị tới hạn $u_{\alpha/2}$ tăng lên, do đó sai số tới hạn ε cũng tăng lên làm cho độ chính xác của ước lượng giảm đi.

Trong thực, tùy yêu cầu về độ chính xác của cuộc điều tra để xác định dung lượng mẫu phù hợp.

$$\varepsilon(\%) = \frac{\varepsilon}{\overline{x}}.100 = u_{\alpha/2}\frac{\sigma}{\overline{x}\sqrt{n}}.100$$

$$= u_{\alpha/2}\frac{CV(\%)}{\sqrt{n}}$$ là sai số tương đối biểu thị mức độ chính xác của ước lượng nên còn gọi là độ chính xác tương đối.

Dung lượng mẫu cần thiết để đạt được độ chính xác tương đối cho trước $\varepsilon_0(\%)$ là:

$$n_{\min} = \left(\frac{u_{\alpha/2}CV(\%)}{\varepsilon_0(\%)}\right)^2$$

Nếu α = 0,10 thì $u_{\alpha/2}$ = 1,645; α = 0,05 thì $u_{\alpha/2}$ = 1,960

α = 0,02 thì $u_{\alpha/2}$ = 2,326; α = 0,01 thì $u_{\alpha/2}$ = 2,576

2.2.2.2. Khi chưa biết phương sai σ² của tổng thể nhưng có dung lượng mẫu lớn (n > 30)

Khi dung lượng mẫu đủ lớn thì S ≈ σ nên có thể thay

thế S cho σ, khi đó việc ước lượng được tiến hành theo luật phân phối chuẩn theo công thức (2 - 3).

Với độ tin cậy 95%:

$$P(\bar{X} - 1,96\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + 1,96\frac{S}{\sqrt{n}}) = 0,95$$

Có thể phát biểu: ở độ tin cậy 95%, trung bình tổng thể μ nằm trong khoảng $(\bar{X} - 1,96\frac{S}{\sqrt{n}}; \bar{X} + 1,96\frac{S}{\sqrt{n}})$.

$$\varepsilon = 1,96\frac{S}{\sqrt{n}}$$

$$\varepsilon(\%) = \frac{\varepsilon}{\bar{x}}.100 = 1,96\frac{S}{\bar{x}\sqrt{n}}.100 = 1,96\frac{CV(\%)}{\sqrt{n}}$$

$$n_{min} = \left(\frac{1,96CV(\%)}{\varepsilon_0(\%)}\right)^2$$

Với độ tin cậy 99%:

$$P(\bar{x} - 2,58\frac{S}{\sqrt{n}} \le \mu \le \bar{x} + 2,58\frac{S}{\sqrt{n}}) = 0,99$$

và trung bình tổng thể μ nằm trong khoảng:

$$(\bar{x} - 2,58\frac{S}{\sqrt{n}}; \bar{x} + 2,58\frac{S}{\sqrt{n}})$$

$$\varepsilon = 2,58\frac{S}{\sqrt{n}}$$

$$\varepsilon(\%) = \frac{\varepsilon}{\bar{x}}.100 = 2,58\frac{S}{\bar{x}\sqrt{n}}.100 = 2,58\frac{CV(\%)}{\sqrt{n}}$$

và: $$n_{\min} = \left(\frac{2,58CV(\%)}{\varepsilon_0(\%)} \right)^2$$

Nếu mẫu được chọn từ tổng thể hữu hạn theo cách rút mẫu không lặp (không hoàn lại), khi $n > 0,1N$, để bảo đảm độ chính xác của ước lượng thì sai số tới hạn sẽ nhân thêm hệ số điều chỉnh $\sqrt{\frac{N-n}{N-1}}$, khi đó công thức (2 - 3) trở thành:

$$P(\bar{X} - u_{\alpha/2} \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \le \mu \le \bar{X} + u_{\alpha/2} \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}) = 1-\alpha \quad (2 - 5)$$

và dựa vào đó để tính toán ε, ε(%), từ đó sẽ xác định được dung lượng mẫu cần thiết để đạt được độ chính xác $\varepsilon_0(\%)$ theo yêu cầu.

Ví dụ: Ước lượng trung bình năng suất cá thể và khoảng tin cậy của tổ hợp bông lai F_1 S02-13/TM1 trồng tại Đại học Nông Lâm Tp. HCM, 2008 theo số liệu bảng sau:

Bảng 2.1: Năng suất cá thể 50 cây (g/cây)

84,4	73,5	84,5	93,5	75,2	74,3	93,5	82,1	68,6	85,4
66,8	57,7	79,4	91,0	129,0	74,6	61,4	91,4	99,5	82,7
95,2	88,3	28,4	77,0	81,2	39,7	86,4	51,5	51,2	80,5
101,0	77,7	90,0	92,9	80,8	67,2	57,1	57,3	34,2	79,5
80,3	88,0	61,1	63,8	101,0	70,2	95,1	97,0	50,3	73,7
Trung bình: 76,92, Phương sai: 351,68, Độ lệch chuẩn: 18,75									

Ở trường hợp này $n = N$.

Với độ tin cậy 95%:

$$P(76,92 - 1,96\frac{18,75}{\sqrt{50}} \le \mu \le 76,92 + 1,96\frac{18,75}{\sqrt{50}}) = 0,95$$

$$= P(76,92 - 5,20 \le \mu \le 76,92 + 5,20) = 0,95$$

$$= P(71,72 \le \mu \le 82,12) = 0,95$$

Tức là: ở độ tin cậy 95% trung bình năng suất cá thể của tổ hợp lai S02-13/TM1 nằm trong khoảng 71,72 g/cây đến 82,12 g/cây (76,92 ± 5,20 g/cây) và với mật độ 41.667 cây/ha thì năng suất dao động trong khoảng 29,9 – 34,2 tạ/ha, trung bình là 32,1 tạ /ha.

Với kết quả điều tra này, sai số tới hạn sai ε = 5,20 g/cây, sai số tương đối ε(%) là 6,76.

Để sai số tương đối $\varepsilon_0(\%)$ cho trước không vượt quá 5% thì số mẫu điều tra tối thiểu phải đạt:

$$n_{\min} = \left(\frac{2CV(\%)}{\varepsilon_0(\%)} \right)^2 = \left(\frac{2 \times 24,38}{5} \right)^2 = 95 \text{ cây}$$

Bây giờ, ta hãy giả sử rằng N = 95, n = 50 thì số hiểu chỉnh sai số $\sqrt{\frac{N-n}{N-1}}$ = 0,692, khi đó sai số tới hạn sẽ là: 5,20 × 0,692 = 3,60. Từ đây, ε(%) = (3,60/76,92) × 100 = 4,7. Như vậy nếu tổ hợp lai được gieo 95 cây thì sai số tương đối không vượt quá 5%.

2.2.2.3. Khi chưa biết phương sai σ^2 của tổng thể và có dung lượng mẫu nhỏ (n < 30)

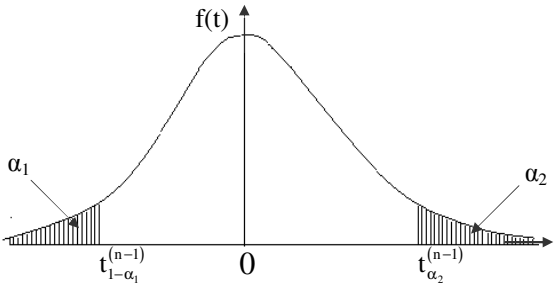
Trong trường hợp mẫu nhỏ việc ước lượng được tiến hành theo luật phân phối Student.

Đại lượng ngẫu nhiên liên tục T phân phối theo luật Student với k bậc tự do, nếu hàm mật độ xác suất của nó được xác định bằng công thức :

$$f(t) = \frac{\Gamma\left(\frac{k}{2}\right)}{\sqrt{\pi(k-1)\Gamma\frac{(k-1)}{2}}}\left(1+\frac{t^2}{n-1}\right)^{-\frac{k}{2}} \text{ với } \forall\ t$$

trong đó Γ(x) là hàm Gamma.

Và, đồ thị hàm mật độ xác suất có dạng như hình 2.1.



Hình 2.1: Đồ thị hàm mật độ xác suất theo luật phân phối Student

Theo luật phân phối Student, với dung lượng mẫu *n*, số bậc tự do là *n* – 1 và độ tin cậy 1 – α cho trước, có thể tìm được cặp α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ và hai giá trị tới hạn Student tương ứng là $t_{1-\alpha_1}^{(n-1)}$ và $t_{\alpha_2}^{(n-1)}$ thỏa mãn điều kiện:

$$P(T \le t_{1-\alpha_1}^{(n-1)}) = \alpha_1 \text{ và}$$

$$P(T \ge t_{\alpha_2}^{(n-1)}) = \alpha_2$$

Từ đó:

$$P(-t_{1-\alpha_1}^{(n-1)} \le T \le t_{\alpha_2}^{(n-1)}) = 1-(\alpha_1+\alpha_2) = 1-\alpha$$

Vì giá trị tới hạn có tính chất $t_{\alpha_1}^{(n-1)} = - t_{1-\alpha_1}^{(n-1)}$ nên có thể viết:

$$P(-t_{\alpha_1}^{(n-1)} \le T \le t_{\alpha_2}^{(n-1)}) = 1-\alpha \qquad (2 - 6)$$

Thay $T = \dfrac{(\overline{x} - \mu)\sqrt{n}}{S}$ vào công thức (2 - 6) thì công thức ước lượng khoảng số trung bình tổng thể theo luật Student là:

$$P(\overline{x} - t_{\alpha_2}^{(n-1)} \frac{S}{\sqrt{n}} \le \mu \le \overline{x} + t_{\alpha_1}^{(n-1)} \frac{S}{\sqrt{n}}) = 1-\alpha \quad (2 - 7)$$

và khoảng tin cậy của μ sẽ là:

$$\left(\overline{x} - t_{\alpha_2}^{(n-1)} \frac{S}{\sqrt{n}}; \overline{x} + t_{\alpha_1}^{(n-1)} \frac{S}{\sqrt{n}} \right)$$

Khi ước lượng khoảng tin cậy đối xứng thì $\alpha_1 = \alpha_2 = \alpha/2$, công thức (2 - 7) trở thành:

$$P(\overline{x} - t_{\alpha/2}^{(n-1)} \frac{S}{\sqrt{n}} \le \mu \le \overline{x} + t_{\alpha/2}^{(n-1)} \frac{S}{\sqrt{n}}) = 1-\alpha \quad (2 - 8)$$

Đây là công thức thường áp dụng nhất để ước lượng khoảng tin cậy của trung bình tổng thể khi $n < 30$.

Từ công thức (2 - 8) sai số tới hạn (hay độ chính xác) là $\varepsilon = t_{\alpha/2}^{(n-1)} \dfrac{S}{\sqrt{n}}$, khoảng tin cậy sẽ là:

$$I = 2\varepsilon = 2t_{\alpha/2}^{(n-1)} \frac{S}{\sqrt{n}},$$

$$\varepsilon(\%) = \frac{\varepsilon}{\overline{x}}.100 = t_{\alpha/2}^{(n-1)} \frac{S}{\overline{x}\sqrt{n}}.100 = t_{\alpha/2}^{(n-1)} \frac{CV(\%)}{\sqrt{n}}$$

Dung lượng mẫu cần thiết để đạt được độ chính xác

$\varepsilon_0(\%)$ là:

$$n_{\min} = \left(t_{\alpha/2}^{(n-1)} \frac{CV(\%)}{\varepsilon_0(\%)} \right)^2$$

Nếu mẫu được chọn từ tổng thể hữu hạn theo cách rút mẫu không lặp (không hoàn lại), khi $n > 0{,}1N$, để bảo đảm độ chính xác của ước lượng thì sai số tới hạn sẽ nhân thêm hệ số điều chỉnh $\sqrt{\frac{N-n}{N-1}}$, khi đó công thức (2 - 8) trở thành:

$$P(\bar{x} - t_{\alpha/2}^{(n-1)} \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \le \mu \le \bar{x} + t_{\alpha/2}^{(n-1)} \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}) = 1-\alpha \quad (2 - 9)$$

Dựa vào đây để tính toán ε, ε(%), từ đó sẽ xác định được dung lượng mẫu cần thiết để đạt được độ chính xác xác tương đối cho trước $\varepsilon_0(\%)$.

Ví dụ: Từ ví dụ ở Bảng 2.1, mục 2.2.2.2 trên đây, nếu chỉ điều tra ngẫu nhiên 10 cây trong tổng số 50 cây. Hãy ước lượng khoảng tin cậy của tổng thể (N = 50).

Bảng 2.2 sau đây là kết quả tính toán các tham số từ 7 cách chọn mẫu có hoàn lại, mỗi mẫu 10 cây.

Bảng 2.2 cho các nhận xét:

- Dù rút mẫu có hoàn lại nhưng với $n < 30$, giá trị trung bình của các mẫu chênh lệch nhau đáng kể, $\bar{x}_{max} - \bar{x}_{min} = 16{,}82$ g/cây, dao động trong khoảng 66,53 g/cây đến 83,35 g/cây.

- Giá trị độ lệch chuẩn S khác nhau không nhiều. Nếu so với trung bình tổng thể thì hệ số biến động tương đối ổn định và khác biệt với hệ số biến động tổng thể không nhiều.

Bảng 2.2: Các tham số của tổng thể và 7 mẫu chọn ngẫu nhiên có hoàn lại

Tham số	Tổng thể	Mẫu						
		1	2	3	4	5	6	7
$\overline{x}$	76,92	83,35	73,76	78,05	82,07	66,53	82,78	69,22
CV(%)	24,38	25,21	27,22	21,87	13,69	33,47	23,81	28,24
S	18,76	21,01	20,08	17,07	11,24	22,27	19,71	19,55
ε	5,20	15,03	14,36	12,21	8,04	15,93	14,10	13,98
I=2ε	10,40	30,06	28,73	24,42	16,08	31,86	28,19	27,96
ε(%)	6,76	18,03	19,47	15,65	9,79	23,94	17,03	20,20
$\varepsilon_{\text{điểm}}$	2,65	6,65	6,35	5,40	3,55	7,04	6,23	6,18
ε_d(%)	3,45	7,97	8,61	6,92	4,33	10,59	7,53	8,93

Điều đó chứng tỏ rằng có thể sử dụng CV(%) của mẫu và độ chính xác ε_0(%) cho trước theo yêu cầu để dự tính số lượng cá thể cần thiết cho cuộc điều tra.

- Khoảng tin cậy của ước lượng (I = 2ε) theo luật Student trong trường hợp mẫu nhỏ lớn và dao động từ 16,08 - 31,86, trong khi đó khoảng tin cậy của ước lượng theo luật chuẩn là 10,40. Tất cả các mẫu đều có sai số tương đối khá lớn, hầu hết đều lớn, từ 16 - 20%, chỉ có mẫu 4 là 9,8%, cao hơn đáng kể so với sai số tổng thể (6,76%).

- Với dung lượng mẫu n = 10, ước lượng điểm mẫu tuy có chênh lệch với ước lượng điểm tổng thể, nhưng có thể áp dụng phương pháp ước lượng điểm để thu thập giá trị trung bình các ô trong thí nghiệm trên đồng ruộng.

Với trung bình tổng thể μ = 76,92, độ lệch chuẩn tổng thể S = 18,76 (Bảng 2.1), Bảng 2.3 sau đây sẽ làm rõ sự khác biệt giữa luật Student và luật chuẩn trong ước lượng

trung bình tổng thể.

Từ Bảng 2.3 có thể thấy rằng:

- Khi dung lượng mẫu n tăng lên, phân phối Student sẽ hội tụ rất nhanh về phân phối chuẩn. Do đó nếu $n > 30$ có thể dùng phân phối chuẩn thay cho phân phối Student.

Bảng 2.3: Ước lượng một số tham số thống kê theo luật Student khi dung lượng mẫu thay đổi với $\alpha/2 = 0,025$.

Độ tự do $(n-1)$	$t_{0,025}^{n-1}$ (*)	Tham số thống kê				
		ε	I = 2ε	ε (%)	ε điểm	ε_d(%)
Theo luật Student với các độ tự do khác nhau						
1	25,452	337,63	675,3	438,93	13,27	17,25
3	4.177	39,18	78,35	50,93	9,38	12,19
5	3,163	24,23	48,46	31,50	7,66	9,96
10	2,634	14,90	29,80	19,37	5,66	7,35
20	2,423	9,92	19,84	12,90	4,09	5,32
30	2,360	7,95	15,90	10,34	3,37	4,38
40	2,329	6,82	13,65	8,87	2,93	3,81
45	2,319	6,41	12,83	8,34	2,77	3,60
49	2,312	6,13	12,27	7,98	2,65	3,45
Theo luật chuẩn						
$u_{0,05}^{n-1} = 1,96$		5,20	10,40	6,76	2,65	3,44

(*) Tra trong phần mềm Excel: = tinv(0.025,$df_{(n-1)}$)

- Khi dung lượng mẫu nhỏ ($n < 30$) việc thay thế luật Student bằng luật chuẩn có thể dẫn đến sai lầm lớn. Chẳng hạn, với hàm phân phối chuẩn với $\alpha = 0,05$, $u_{0,05} = 1,96$, trong khi đó với $n = 4$ (df = 3), giá trị tới hạn Student $t_{0,025}^{(3)} = 4,18$, tức là sai lệch đến 2,22.

- Về ước lượng điểm: Một quần thể của một giống (có kiểu gen đồng nhất), hệ số biến động giữa các cá thể của các chỉ tiêu sinh trưởng như chiều cao cây, số hạt, quả/cây ở trên ruộng phụ thuộc chủ yếu vào sự đồng đều của đất trồng, thông thường là 8 – 15%, có khi lên tới 20 – 25% (như Bảng 2.3) hoặc cao hơn. Với các chỉ tiêu này, để có độ chính xác của ước lượng điểm đạt 6 – 8%, dung lượng mẫu khoảng 10 cây. Tuy nhiên có nhiều chỉ tiêu mà giữa các cá thể trong cùng một giống rất ít khác biệt nhau kể cả khi trồng trên ruộng độ đồng đều về đất không cao. Với các chỉ tiêu này hệ số biến động lại rất thấp, không vượt quá 8%, thậm chí chỉ 1 – 2%. Như vậy chỉ cần theo dõi khoảng 3 cá thể thì có thể đạt được độ chính xác rất cao. Chẳng hạn, hệ số biến động khối lượng 100 hạt của một giống đậu nành là 5%, chỉ cần cân 3 mẫu (n = 3) đã đạt độ chính xác $\varepsilon(\%) = \frac{CV(\%)}{\sqrt{n}} = \frac{5}{\sqrt{3}}$ = 2,9.

Những điều cần lưu ý khi ước lượng trung bình tổng thể:

- Trước hết, cần xem đối tượng nghiên cứu để quyết định phương pháp lấy mẫu. Nếu quần thể đồng nhất (ví dụ như một giống thuần hay giống lai F_1) có thể lấy một mẫu nhỏ hoặc mẫu lớn. Nếu quần thể không đồng nhất (ví dụ như quần thể phân ly các thế hệ lai, quần thể đột biến hay sản phẩm của kỹ thuật di tuyền), khi ước lượng trung bình quần thể cần phải lấy nhiều mẫu, hay lấy mẫu thử theo phép rút mẫu ngẫu nhiên rồi tính toán dung lượng mẫu cần thiết để bảo đảm độ chính xác của ước lượng.

- Căn cứ vào độ lớn tổng thể để quyết định dung lượng mẫu cần thiết. Nếu quần thể không lớn có thể quan sát toàn bộ tổng thể.

- Nên áp dung phương pháp lấy mẫu ngẫu nhiên.

- Chỉ nên áp dụng phương pháp ước lượng điểm khi dung lượng mẫu nhỏ.

2.3. ƯỚC LƯỢNG PHƯƠNG SAI TỔNG THỂ (σ^2)

2.3.1. Ước lượng điểm phương sai tổng thể

Giả sử là đại lượng ngẫu nhiên với phương sai tổng thể σ^2 chưa biết. Kết quả rút mẫu với dung lượng n ta được các giá trị quan sát: $x_1, x_2, ..., x_n$. Giá trị phương sai mẫu chưa hiệu chỉnh sẽ là

$$\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Nếu sử dụng phương sai mẫu chưa hiệu chỉnh để ước lượng σ^2 ta có:

$$E\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right\} = E\left\{\frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\overline{x} + \overline{x}^2\right)\right\}$$

$$= E\left\{\frac{1}{n}\left[\sum_{i=1}^{n}x_i^2 - 2\overline{x}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\overline{x}^2\right]\right\}$$

$$= \frac{1}{n}E\left\{\sum_{i=1}^{n}\left(x_i^2 - 2n\overline{x}^2 + \overline{x}^2\right)\right\}$$

$$= \frac{1}{n}E\left\{\sum_{i=1}^{n}\left(x_i^2 - n\overline{x}\right)^2\right\} = \frac{1}{n}\left\{\sum_{i=1}^{n}E\left(x_i^2\right) - nE\left(\overline{x}^2\right)\right\}$$

Do $E\left(x_i^2\right) = V(x_i) + \left[E(x_i)\right]^2$ và $E\left(\overline{x}^2\right) = V(\overline{x}) + \left[E(\overline{x})\right]^2$.

Cần chứng minh $\hat{S}^2 = V(\overline{x}) = V(X) = \sigma^2$

Theo tính chất 2 và 4 của phương sai ta có:

$$V(\bar{x}) = V\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(x_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(X)$$

$$= \frac{1}{n}\sigma^2$$

Như vậy phương sai chưa hiệu chỉnh là một ước lượng chệch.

Phương sai hiệu chỉnh $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ là ước lượng không chệch cho phương sai tổng thể. Thật vậy:

$$E\left\{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\} = \frac{n}{n-1}E\left\{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}$$

$$= \frac{n}{n-1}.\frac{n-1}{n}V(X) = \sigma^2$$

2.3.2. Ước lượng khoảng phương sai tổng thể

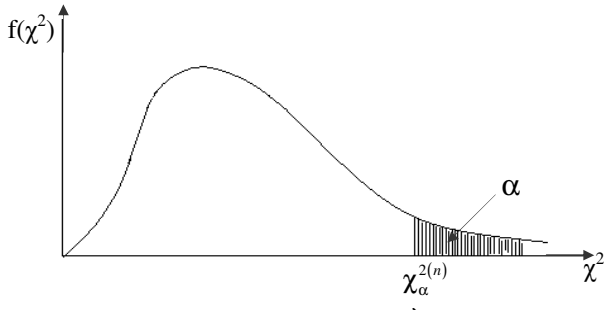
Việc ước lượng khoảng phương sai tổng thể được tiến hành theo luật phân phối "khi bình phương".

Đại lượng ngẫu nhiên liên tục χ^2 phân phối theo luật "khi bình phương" với n bậc tự do, nếu hàm mật độ xác suất của nó được xác định bằng công thức:

$$f(x) = \begin{cases} 0 & \text{với } x \le 0 \\ \dfrac{1}{2^{\frac{n}{2}}.\Gamma\left(\frac{n}{2}\right)} e^{\frac{x}{2}}.x^{\frac{n}{2}-1} & \text{với } x > 0 \end{cases}$$

trong đó $\Gamma(x)$ là hàm Gamma.

Và, đồ thị hàm mật độ xác suất có dạng như hình 2.2.



Hình 2.2: Đồ thị hàm mật độ xác suất
theo luật khi bình phương

Người ta đã chứng minh được rằng nếu đại lượng ngẫu nhiên χ^2 phân phối theo luật khi bình phương với n bậc tự do thì kỳ vọng $E(\chi^2) = n$ và phương sai $V(\chi^2) = 2n$.

Theo luật phân phối khi bình phương, với dung lượng mẫu n, số bậc tự do là $n - 1$ và độ tin cậy $1 - \alpha$ cho trước, có thể tìm được cặp α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ và hai giá trị tới hạn khi bình phương tương ứng là $\chi^{2(n-1)}_{1-\alpha_1}$ và $\chi^{2(n-1)}_{\alpha_2}$ thỏa mãn điều kiện:

$$P(\chi^2 \le \chi^{2(n-1)}_{1-\alpha_1}) = \alpha_1 \text{ và}$$

$$P(\chi^2 \ge \chi^{2(n-1)}_{\alpha_2}) = \alpha_2$$

Từ đó:

$$P(\chi^{2(n-1)}_{1-\alpha_1} \le \chi^2 \le \chi^{2(n-1)}_{\alpha_2}) = 1 - (\alpha_1 + \alpha_2) = 1 - \alpha \quad (2 - 10)$$

Thay $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2}$ vào công thức (2 - 10) thì công thức ước lượng phương sai tổng thể theo luật khi bình

phương là:

$$P\left(\frac{(n-1)S^2}{\chi^{2(n-1)}_{\alpha_2}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^{2(n-1)}_{1-\alpha_2}}\right) = 1-\alpha \qquad (2 - 11)$$

và khoảng tin cậy của μ là:

$$\left(\frac{(n-1)S^2}{\chi^{2(n-1)}_{\alpha_2}}; \frac{(n-1)S^2}{\chi^{2(n-1)}_{1-\alpha_2}}\right)$$

Khi ước lượng khoảng tin cậy đối xứng thì $\alpha_1 = \alpha_2 = \alpha/2$, công thức (2 - 11) trở thành:

$$P\left(\frac{(n-1)S^2}{\chi^{2(n-1)}_{\alpha/2}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^{2(n-1)}_{1-\alpha/2}}\right) = 1-\alpha \qquad (2 - 12)$$

Ví dụ: Hãy ước lượng phương sai tổng thể về năng suất (cá thể) của tổ hợp bông lai F_1 S02-13/TM1 trồng tại Đại học Nông Lâm Tp. HCM, 2008 theo số liệu Bảng 2.2 với S = 18,76 với độ tin cậy 0,95.

Ở đây $\alpha = 1 - 0,95 = 0,05$; $\alpha/2 = 0,025$; $1 - \alpha/2 = 0,975$.

Tra giá trị χ^2 trong phần mềm Excel, ta có $\chi^{2(49)}_{0,025} = 70,222$ và $\chi^{2(49)}_{0,975} = 31,555$. Thay các giá trị này vào công thức (2 - 12) ta có:

$$\left(\frac{49\times 18,76^2}{70,222}; \frac{49\times 18,76^2}{31,555}\right) = (546,24 ; 245,46)$$

và: $P(245,46 \le \sigma^2 \le 546,24) = 0,95$

Như vậy ở mức tin cậy 0,95 phương sai tổng thể năng suất cá thể của tổ hợp bông lai F1 S02-13/TM1 nằm trong khoảng từ 245,46 đến 546,24.

2.4. ƯỚC LƯỢNG KHOẢNG XÁC SUẤT CÁC DẤU HIỆU ĐỊNH TÍNH CỦA MỘT TỔNG THỂ

Việc ước lượng khoảng xác suất các dấu hiệu định tính của một tổng thể được tiến hành theo luật "không – một".

- ***Luật "không - một"A(p)***

Giả sử để thử nẩy mầm một lô hạt giống, một hạt giống chỉ có một trong hai khả năng xẩy ra, hoặc là nẩy mầm (biến cố A) hoặc là không nảy mầm (biến cố $\overline{A}$). Vậy xác suất để có m hạt nảy mầm trong n hạt thử là $p = \frac{m}{n}$ còn xác suất hạt không nảy mầm là $q = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - p$.

Một cách tổng quát, giả sử ta làm một phép thử, trong đó biến cố A có thể xảy ra với xác suất bằng p. Gọi X là số lần xuất hiện A. Như vậy X là đại lượng ngẫu nhiên rời rạc với hai giá trị hoặc là bằng 0 (nếu không xuất hiện A) hoặc là bằng 1 (nếu xuất hiện A) với các xác suất tương ứng được biểu thị bằng công thức:

$$P_x = p^x q^{1-x} \text{ với } x = 0; 1, \text{ trong đó } q = 1 - p.$$

Phân phối thỏa mãn với công thức trên đây được gọi là phân phối theo quy luật "không – một" cho đại lượng ngẫu nhiên X với tham số p.

Bảng phân phối xác suất của đại lượng ngẫu nhiên X

theo quy luật "không – một" có dạng:

X	0	1
P	q	p

($q = 1 - p$)

Từ bảng phân phối xác suất ta có:

Kỳ vọng: $E(X) = p$

Phương sai: $V(X) = pq$

Độ lệch chuẩn: $\sigma_x = \sqrt{pq}$

- ***Khoảng ước lượng và độ chính xác của ước lượng***

Với dung lượng mẫu đủ lớn, hoặc nếu $n > 5$ và $\frac{\left|\sqrt{\frac{p_m}{1-p_m}} - \sqrt{\frac{1-p_m}{p_m}}\right|}{\sqrt{n}} < 0{,}3$ (p_m là xác suất mẫu) thì luật phân phối "không – một" sẽ có phân phối gần phân phối chuẩn tắc N(0, 1), do đó vẫn có thể ước lượng khoảng tin cậy tần suất p của tổng thể. Với độ tin cây $1 - \alpha$ cho trước, có thể tìm được cặp α_1 và α_2 sao cho $\alpha_1 + \alpha_2 = \alpha$ và hai giá trị tới hạn chuẩn $u_{1-\alpha_1}$ và u_{α_2} thỏa mãn điều kiện:

$$P(U \le u_{1-\alpha_1}) = \alpha_1 \text{ và } P(U \ge u_{\alpha_2}) = \alpha_2$$

Từ đó:

$$P(u_{1-\alpha_1} \le U \le u_{\alpha_2}) = 1 - (\alpha_1 + \alpha_2) = 1 - \alpha \qquad (2 - 13)$$

Thay $U = \frac{(p_m - p)\sqrt{n}}{\sqrt{p_m(1-p_m)}}$ vào công thức (2 - 13) và áp dụng tính chất $u_{\alpha_1} = -u_{1-\alpha_1}$ thì công thức ước lượng xác suất p

tổng thể là:

$$P\left(p_m - u_{\alpha_2}\sqrt{\frac{p_m(1-p_m)}{n}} \le p \le p_m + u_{\alpha_1}\sqrt{\frac{p_m(1-p_m)}{n}} \right) = 1-\alpha \quad (2 - 14)$$

và khoảng tin cậy của p là:

$$\left(p_m - u_{\alpha_2}\sqrt{\frac{p_m(1-p_m)}{n}}; p_m + u_{\alpha_1}\sqrt{\frac{p_m(1-p_m)}{n}} \right)$$

Khi ước lượng khoảng tin cậy đối xứng thì $\alpha_1 = \alpha_2 = \alpha/2$, công thức (2 - 14) trở thành:

$$P\left(p_m - u_{\alpha/2}\sqrt{\frac{p_m(1-p_m)}{n}} \le p \le p_m + u_{\alpha/2}\sqrt{\frac{p_m(1-p_m)}{n}} \right) = 1-\alpha \quad (2\text{-}15)$$

Đây là công thức thường áp dụng nhất để ước lượng khoảng tin cậy xác suất p. Với $\alpha = 0,05$, $u_{\alpha/2} = 1,96$; $\alpha = 0,01$, $u_{\alpha/2} = 2,58$ và $\alpha = 0,001$, $u_{\alpha/2} = 3,29$.

Từ công thức (2 - 15), sai số tới hạn của ước lượng (độ chính xác) là $\varepsilon = u_{\alpha/2}\sqrt{\frac{p_m(1-p_m)}{n}}$ và khoảng tin cậy là $I = 2\varepsilon$. Dung lượng mẫu cần thiết để đạt được độ chính xác cho trước ε_0 là:

$$n_{\min} = (u_{\alpha/2})^2 \frac{p_m(1-p_m)}{\varepsilon_0^{\,2}}$$

Ví dụ: Để kiểm nghiệm tỷ lệ nảy mầm một giống bắp lai, người ta đã tiến hành thử 4 mẫu, mỗi mẫu 100 hạt. Kết quả như sau:

Mẫu thử	1	2	3	4	Tổng
Số hạt nảy mầm	93	89	87	96	365

Hãy ước lượng khoảng xác suất nảy mầm của lô hạt giống và số lượng hạt cần thử để đạt sai số không vượt quá 3% với độ tin cậy 95%.

Giải: n = 400, p_m = 365/400 = 0,913, α = 1 – 0,95 = 0,05, $u_{0,025}$ = 1,96 và ε_0 = 0,03.

Với dung lượng mẫu đủ lớn, theo công thức (2 - 15) ta có:

$$P\left(0,885 \le p \le 0,940\right) = 0,95$$

Như vậy với mức tin cậy 95% tỷ lệ nảy mầm của hạt giống nằm trong khoảng 88,5% đến 94,0%.

Số hạt cần thử: $n_{\min} = (u_{\alpha/2})^2 \dfrac{p_m(1-p_m)}{\varepsilon_0^{\,2}}$

$$= 1,96^2 \frac{0,913(1-0,913)}{0,03^2} = 340,8 \approx 341$$

Để đạt sai số không vượt quá 3% cần thử tối thiểu 341 hạt.

Chương 3
SO SÁNH CÁC THAM SỐ

Giá trị trung bình, phương sai và các tham số khác theo các dấu hiệu định lượng hoặc xác suất các dấu hiệu định tính của các tổng thể được ước lượng từ những số liệu qua các cuộc điều tra khảo sát mẫu là những kết quả mô tả tổng thể. Để cung cấp những dữ liệu cần thiết phục vụ cho mục tiêu nghiên cứu hoặc ứng dụng trong sản xuất, cần phải đánh giá, so sánh và phân tích mối quan hệ giữa các tham số của các tổng thể. Chương này sẽ đề cập đến việc so sánh các tham số tổng thể từ kết quả các cuộc điều tra khảo sát các mẫu, gồm:

- So sánh hai trung bình và mở rộng (phương pháp tham số và phi tham số);

- So sánh hai phương sai và mở rộng;

- Đánh giá tính độc lập của các dấu hiệu định tính.

Việc so sánh trong các thí nghiệm sẽ được đề cập ở phần 2.

3.1. SO SÁNH HAI TRUNG BÌNH VÀ MỞ RỘNG

3.1.1. Phương pháp tham số

3.1.1.1. Cơ sở lý luận

Để so sánh hai trung bình mẫu $\overline{X}_i$ và $\overline{X}_2$ không đơn giản là so sánh hiệu số $\overline{X}_1$ - $\overline{X}_2$, bởi vì mỗi số trung bình

đều có sai số ε_i: $\varepsilon_i = u_{\alpha/2}\frac{S_i}{\sqrt{n_i}}$ (trường hợp mẫu lớn) hay $\varepsilon_i = t_\alpha \frac{S_i}{\sqrt{n_i}}$ (cho mọi trường hợp, theo chương 2).

Như vậy để so sánh hai hay nhiều trung bình ở độ tin cậy 1 - α nào đó cần phải xác định khoảng khác biệt tối thiểu có ý nghĩa phân biệt (Least Significant Difference - LSD) giữa chúng: $LSD_\alpha = t_\alpha Sd$, trong đó t_α là giá trị tới hạn phân phối Student ở mức α (thường gọi là $t_{bảng}$ hay $t_{lý\ thuyết}$), Sd là sai số thực nghiệm giữa hai trung bình (giá trị Sd sẽ được nêu cụ thể trong từng trường hợp). Khi có được Sd dễ dàng tính được LSD_α.

Giả thuyết H_0: $\bar{X}_1 = \bar{X}_2$ được chấp nhận với độ tin cậy 1 - α (mức sai lầm α) khi $|\bar{X}_1 - \bar{X}_2| < LSD_\alpha$, còn giả thuyết H_1: $\bar{X}_1 \neq \bar{X}_2$ được chấp nhận với độ tin cậy 1 - α khi $|\bar{X}_1 - \bar{X}_2| \geq LSD_\alpha$. Do $T_{TN} = \frac{|\bar{X}_1 - \bar{X}_2|}{Sd}$ và $t_\alpha = \frac{LSD_\alpha}{Sd}$ nên thay vì kiểm định sự chênh lệch giữa hai trung bình $|\bar{X}_1 - \bar{X}_2|$ so với LSD_α, người ta chuyển sang kiểm định T_{TN} so với $t_{bảng}$. Khi $T_{TN} < t_{bảng}$ giả thuyết H_0 được chấp nhận, còn khi $T_{TN} \geq t_{bảng}$ giả thuyết H_1 được chấp nhận.

Trong trường hợp dung lượng mẫu lớn hoặc đã biết phương sai của hai tổng thể, phép nghiệm U sẽ được sử dụng để so sánh hai trung bình: $U_{TN} = \frac{|\bar{X}_1 - \bar{X}_2|}{Sd}$ được so với $u_{\alpha/2}$.

Để phép nghiệm đúng cần lưu ý yếu tố so sánh. Ví dụ: Khi so sánh giống dừa Xiêm được trồng ở Bình Định

và giống dừa Xiêm đó được trồng ở Bến Tre tức là ta muốn đánh giá ảnh hưởng của hai điều kiện trồng đến giống dừa. Yếu tố so sánh ở đây là sinh thái trồng dừa. Còn khi so sánh các giống dừa thì chúng phải được trong một điều kiện như nhau (cùng vùng và cùng một loại đất). Và, để đánh giá giống dừa nào tốt cho vùng nào thì thí nghiệm đó phải được trồng trên nhiều vùng (thí nghiệm hai yếu tố: giống và vùng). Tuy nhiên để đánh giá giống dừa đặc sản của Bình Định được trồng ở Bình Định và giống dừa đặc sản của Bến Tre được trồng ở Bến Tre (hai giống này khác nhau) thì yếu tố so sánh ở đây là dừa của Bình Định và dừa của Bến Tre.

Ở ngay trong một nơi, nếu trồng hai giống ở hai lô khác nhau, mỗi lô lấy một số mẫu (một số cây). Việc so sánh sẽ không chính xác vì đất của hai lô không thể đồng nhất (khác nhau không nhiều thì ít) nên giá trị trung bình của hai giống không chỉ do giống tạo nên, chưa nói đến ngay trong một lô các cá thể trong một giống có sự khác nhau do yếu tố đất trồng. Tuy nhiên có thể chấp nhận khi đã xác định được độ đồng đều của đất ở mức cho phép.

Trong so sánh hai trung bình, nếu so sánh T_{TN} với t_{α} (t tới hạn hai phía – t Critical two-tail), khi đó độ tin cậy sẽ là 1- α. Nếu so sánh T_{TN} với $t_{\alpha/2}$ (t tới hạn một phía – t Critical one-tail), khi đó độ tin cậy sẽ là 1- α/2.

Trong thực nghiệm, người ta thường lấy mức α = 0,05 (độ tin cậy là 95%) khi so sánh T_{TN} với $t_{0,05}$ hoặc α = 0,01 (độ tin cậy sẽ là 99%) khi so sánh T_{TN} với $t_{0,01}$.

3.1.1.2. So sánh hai trung bình khi đã biết phương sai của hai tổng thể σ_1^2 *và* σ_2^2

Khi đã biết phương sai của hai tổng thể, việc so sánh

giữa hai trung bình được thực hiện theo công thức:

$$U_{TN} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (3 - 1)$$

Trong đó:

$\bar{X}_1$ và $\bar{X}_2$ là trung bình của hai mẫu mẫu quan sát

σ_1^2 và σ_2^2 là phương sai của hai mẫu quan sát

n_1 và n_2 là dung lượng của hai mẫu quan sát

$$\text{ở đây } Sd = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Giá trị U_{TN} sẽ được so sánh với giá trị tới hạn $u_{\alpha/2}$.

Nếu $U_{TN} < u_{\alpha/2}$ thì $\bar{X}_1$ không khác $\bar{X}_2$ ở độ tin cậy $1 - \alpha$.

Nếu $U_{TN} \geq u_{\alpha/2}$ thì $\bar{X}_1$ khác $\bar{X}_2$ ở độ tin cậy $1 - \alpha$, hoặc là $\bar{X}_1 > \bar{X}_2$ hoặc là $\bar{X}_1 < \bar{X}_2$.

Giá trị $u_{\alpha/2}$ được ghi ở mục 2.2.1, chương 2.

Thường thì trong sinh học rất hiếm trường hợp biết được phương sai tổng thể.

3.1.1.3. So sánh hai trung bình khi chưa biết phương sai của hai tổng thể nhưng biết rằng chúng bằng nhau ($\sigma_1^2 = \sigma_2^2$)

Việc kiểm tra sự bằng nhau (khác nhau không có ý nghĩa) của σ_1^2 và σ_2^2 thông qua việc kiểm tra hai phương sai mẫu S_1^2 và S_2^2 nhờ phép nghiệm F (sẽ được đề cập ở mục 3.3). Khi phương sai của hai tổng thể không khác nhau có ý

nghĩa, thì việc so sánh giữa hai trung bình được thực hiện theo công thức:

$$T_{TN} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad (3 - 2)$$

ở đây $Sd = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

t_α được tra với $(n_1 + n_2 - 2)$ độ tự do.

Độ tin cậy của phép so sánh khác nhau khi lấy các giá trị tới hạn của t khác nhau (giá trị tới hạn một phía và hai phía).

Ví dụ: So sánh năng suất cá thể của tổ hợp bông lai F_1 C92-52/C118A theo số liệu sau đây và F_1 S02-13/TM1 (ở ví dụ Bảng 2.1, chương 2) với các thông tin sau:

Tổ hợp C92-52/C118A: $n_1 = 45$; $\bar{x} = 63,56$, $S^2 = 387,90$

Tổ hợp S02-13/TM1: $n_2 = 50$; $\bar{x} = 76,92$, $S^2_1 = 351,68$

Giải:

Trước hết phải kiểm tra hai phương sai. Để kiểm tra ta dùng tiêu chuẩn F:

$$F_{TN} = \frac{387,90}{351,68} = 1,10$$

tra bảng F với hai độ tự do 49 và 44 ta có $F_{0,05} = 1,63$. Như vậy có thể coi hai phương sai bằng nhau.

Áp dụng công thức (3 - 2) ta có:

$$T_{TN} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

$$= \frac{76,92 - 63,56}{\sqrt{\frac{(50-1)(351,68) + (45-1)(387,90)}{50+45-2}\left(\frac{1}{50}+\frac{1}{45}\right)}}$$

$$= 3,38$$

t_α được tra với 50 + 45 – 2 = 93 độ tự do ta được:

$$t_{0,05}^{93} = 1,99\,,\ t_{0,01}^{93} = 2,63\,.$$

Như vậy, $T_{TN} = 3,38 > t_{0,01}^{93} = 2,63$, năng suất F_1 tổ hợp S02-13/TM1 cao hơn tổ hợp C92-52/C118A với độ tin cậy 99%.

Kết quả so sánh trung bình F_1 S02-13/TM1 và F_1 C92-52/C118A trên phần mềm Excel:

t-Test: Two-Sample Assuming Equal Variances

	C92-52/C118A	S02-13/TM1
Mean	63.56	76.922
Variance	387.90	351.68
Observations	45	50
Pooled Variance	368.8169	
Hypothesized Mean Difference	0	
df	93	
t Stat	-3.3861	
P(T<=t) one-tail	0.0005	
t Critical one-tail	1.6614	
P(T<=t) two-tail	0.0010	
t Critical two-tail	1.9858	

3.1.1.4. So sánh hai trung bình khi chưa biết phương sai của hai tổng thể nhưng biết rằng chúng khác nhau ($\sigma_1^2 \neq \sigma_2^2$)

Với dung lượng mẫu đủ lớn ($n > 30$), khi phương sai của hai tổng thể khác nhau, việc so sánh giữa hai trung bình được thực hiện theo công thức:

$$T_{TN} = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad (3 - 3)$$

trong đó: $\overline{X}_1$ và $\overline{X}_2$ là trung bình của hai mẫu mẫu quan sát; S_1^2 và S_2^2 là phương sai của hai mẫu quan sát; n_1 và n_2 là dung lượng của hai mẫu quan sát.

Giá trị tới hạn phân phối Student t_α được tra với k độ tự do lấy số nguyên từ công thức sau:

$$k = \frac{(n_1 - 1)(n_2 - 1)\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{(n_1 - 1)\left(\frac{S_1^2}{n_1}\right)^2 + (n_2 - 1)\left(\frac{S_2^2}{n_2}\right)^2} \qquad (3 - 4)$$

Nếu $n_1 = n_2 = n$ thì công thức (3 - 3) trở thành:

$$T_{TN} = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{\frac{S_1^2 + S_2^2}{n}}}.$$

Người ta đã chứng minh được rằng nếu $\overline{X}_1$ khác $\overline{X}_2$ một cách ngẫu nhiên thì cứ 100 lần rút mẫu không quá 5 lần $T_{TN} < t_{0,05}$ (giá trị tới hạn t hai chiều) với độ tin cậy giả thuyết H_1

là 95% và không quá 2,5 lần $T_{TN} < t_{0,025}$ (giá trị tới hạn t một chiều) với độ tin cậy 97,5%. Ngược lại, khi $\overline{X}_1$ không khác $\overline{X}_2$ một cách ngẫu nhiên thì cứ 100 lần rút mẫu không quá 5 lần $T_{TN} > t_{0,05}$ với độ tin cậy giả thuyết H_0 là 95% và không quá 2,5 lần $T_{TN} > t_{0,025}$ với độ tin cậy 97,5%.

Như vậy, nếu $T_{TN} \geq t_{bảng}$ ở mức α thì kết luận rằng $\overline{X}_1$ khác $\overline{X}_2$ ở độ tin cậy 1 - α. Và, nếu $T_{TN} < t_{bảng}$ ở mức α thì kết luận rằng $\overline{X}_1$ không khác với $\overline{X}_2$ ở độ tin cậy 1 - α.

Lưu ý rằng, khi *n* càng lớn t_α càng gần đến $u_{\alpha/2}$, phân phối Student càng gần phân phối chuẩn. Khi đó việc so sánh hai trung bình theo công thức:

$$U_{TN} = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Khi đó ta sử dụng phép nghiệm U.

Ví dụ: So sánh năng suất cá thể thế hệ F_1 và F_2 của tổ hợp bông lai C92-52/C118A trồng tại Đại học Nông Lâm Tp. HCM 2008 theo kết quả theo dõi sau đây.

Năng suất cá thể (g/cây) của 45 cây F_1

50,7	30,0	32,9	78,1	41,3	72,9	57,1	52,0	94,5	87,7
69,7	64,6	72,9	79,6	91,2	46,6	42,9	42,9	29,4	76,4
72,0	65,8	58,1	50,1	53,1	71,0	54,5	52,1	62,3	94,0
59,2	38,5	57,9	66,0	39,6	78,6	37,4	54,8	78,4	48,6
98,0	68,0	96,8	97,8	94,2					

Trung bình: 63,56; Phương sai: 387,90; Độ lệch chuẩn:19,70

Năng suất cá thể (g/cây) của 110 cây F_2

20,9	69,4	42,5	21,3	45,6	21,5	14,9	10,7	11,3	20,3
103,0	10,4	97,0	53,0	57,5	41,8	79,5	91,0	44,5	37,0
42,0	11,7	54,9	41,8	49,2	52,4	55,1	91,0	47,5	43,0
49,6	64,3	132,0	60,7	94,0	4,5	99,0	96,3	89,4	96,0
49,5	59,1	44,9	42,9	62,8	49,7	73,8	46,9	75,8	62,0
40,2	57,9	87,7	53,3	98,5	3,2	98,2	41,9	58,8	79,1
49,5	52,3	63,8	17,4	77,6	69,9	65,5	59,6	79,5	48,5
17,7	38,0	20,5	35,9	47,4	37,0	85,8	45,5	29,0	62,8
28,9	31,6	16,6	34,6	48,5	37,4	64,2	50,4	26,5	94,0
78,0	16,6	37,8	38,1	83,3	86,4	29,6	25,5	33,4	11,3
74,2	19,9	75,8	59,1	33,1	66,4	139	52,4	31,8	98,0
Trung bình: 53,45; Phương sai: 768,61; Độ lệch chuẩn: 27,72									

Rõ ràng hai tập hợp số liệu có phương sai khác nhau. Áp dụng công thức (3 - 3) ta có:

$$T_{TN} = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$= \frac{63,56 - 53,45}{\sqrt{\frac{387,9}{45} + \frac{768,61}{110}}} = 2,56$$

Thay các giá trị vào công thức (3 - 4) ta được k = 136 độ tự do và với độ tự do này tiêu chuẩn T ≈ tiêu chuẩn U, $t_{0,05}^{136} \approx u_{0,025} = 1,98$, còn $t_{0,01}^{136} \approx u_{0,015} = 2,61$.

Như vậy, năng suất F_1 cao hơn năng suất F_2 với độ tin cậy trên 95% gần 99%.

Kết quả so sánh trung bình F_1 và F_2 trên phần mềm Excel:

t-Test: Two-Sample Assuming Unequal Variances

	F1	F2
Mean	63.56	53.45
Variance	387.90	768.61
Observations	45	110
Hypothesized Mean Difference	0	
df	114	
t Stat	2.5596	
P(T<=t) one-tail	0.0059	
t Critical one-tail	1.6583	
P(T<=t) two-tail	0.0118	
t Critical two-tail	1.9810	

Để so sánh chính xác hai trung bình khi hai phương sai mẫu S_1^2 và S_2^2 khác nhau ở mức tin cậy 95%, đòi hỏi phải có dung lượng mẫu lớn ($n > 30$).

Với dung lượng mẫu nhỏ ($n < 30$), khi hai phương sai mẫu S_1^2 và S_2^2 khác nhau việc so sánh sẽ kém chính xác. Trong trường hợp này có thể áp dụng phương pháp rút mẫu ngẫu nhiên có hoàn lại từ mẫu đã có rất nhiều lần (hàng trăm lần) để ước lượng trung bình mới của hai mẫu và tiến hành so sánh như trường hợp dung lượng mẫu lớn.

3.1.1.5. So sánh hai trung bình lấy mẫu theo cặp (Paired two samples)

Cơ sở của phép so sánh này là: Nếu hai đại lượng ngẫu nhiên X_1, X_2 có quan hệ phân phối theo luật Student thì đại lượng ngẫu nhiên tổng hay hiệu của chúng cũng phân phốitheo luật Student.

Trong trường hợp này, việc so sánh hai trung bình $\overline{X}_1$ và $\overline{X}_2$ được kiểm định theo tiêu chuẩn T sau đây:

$$T_{TN} = \frac{\bar{d}}{S_d}\sqrt{n} \qquad (3 - 5)$$

trong đó:

$\bar{d} = \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - x_{2i})$ là trung bình độ lệch;

$S_d = \sqrt{V(d_i)}$ là độ lệch chuẩn của các độ lệch $d_i = x_{1i} - x_{2i}$

n là dung lượng mẫu ($n = n_{x_1} = n_{x_2}$)

Nếu $T_{TN} \geq t_{0,05}$ với n – 1 bậc tự do thì $\bar{X}_1 \neq \bar{X}_2$, ngược lại $T_{TN} < t_{0,05}$ thì $\bar{X}_1 \approx \bar{X}_2$.

Độ tin cậy của phép so sánh khác nhau khi lấy các giá trị tới hạn của t khác nhau (giá trị tới hạn một phía và hai phía).

Ví dụ: Kết quả học tập của 26 sinh viên năm thứ nhất và năm thứ 2 được ghi ở Bảng 3.1.

Loại trừ những trường hợp may rủi trong thi cử, kết quả học tập ở Bảng 3.1 là do sự cố gắng của các em.

Số liệu Bảng 3.1 cho thấy trong 26 sinh viên, có 14 em có điểm năm 2 cao hơn năm 1 (+), 11 em có điểm năm 2 thấp hơn năm 1 (-) và 1 em điểm 2 năm bằng nhau. Khó có thể so sánh kết quả học tập 2 năm học.

Để áp dụng công thức (3 - 5) ta tính:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - x_{2i}) = (- 0,4 + 0,2 - 0,1 + \ldots$$

$$- 0,4) = 0,073$$

$$S_d = \sqrt{V(d_i)} = 0,346$$

Bảng 3.1: Điểm trung bình các môn của 26 sinh viên

TT	Điểm các năm		$X_2 - X_1$ (d_i)	TT	Điểm các năm		$X_2 - X_1$ (d_i)
	1 (X_1)	2(X_2)			1 (X_1)	2(X_2)	
1	8,3	7,9	– 0,4	14	6,7	7,2	+ 0,5
2	8,4	8,6	+ 0,2	15	6,9	7,1	+ 0,2
3	8,2	8,1	– 0,1	16	9,3	9,4	+ 0,1
4	6,5	7,2	+ 0,7	17	5,8	5,6	– 0,2
5	7,8	7,3	– 0,5	18	8,6	8,8	+ 0,2
6	6,9	7,2	+ 0,3	19	6,2	5,8	– 0,4
7	7,1	7,1	0,0	20	7,9	8,2	+ 0,3
8	8,4	8,7	+ 0,3	21	7,8	7,6	– 0,2
9	7,6	7,9	+ 0,3	22	9,0	8,7	– 0,3
10	7,8	7,7	– 0,1	23	8,2	8,6	+ 0,4
11	7,5	7,7	+ 0,2	24	8,4	8,2	– 0,2
12	6,4	7,2	+ 0,8	25	7,3	7,2	– 0,1
13	6,8	7,1	+ 0,3	26	5,8	5,4	– 0,4

Cuối cùng: $T_{TN} = \frac{\bar{d}}{S_d}\sqrt{n} = \frac{0,073}{0,346}\sqrt{26} = 1,08$

$$T_{TN} = 1,08 < t_{0,05}^{(26-1)} = 2,06.$$

Như vậy kết quả học tập hai năm như nhau ($\bar{X}_1 \approx \bar{X}_2$).

3.1.2. Phương pháp phi tham số

Các phép so sánh hai trung bình bằng phương pháp tham số trên đây được thực hiện với điều kiện là các tổng thể có phân phối theo luật chuẩn hoặc là có dung lượng mẫu lớn để có thể áp dụng định lý giới hạn trung tâm. Nếu điều kiện này bị vi phạm, việc kiểm định theo các tiêu chuẩn trên kém hiệu nghiệm. Trong trường hợp này cần phải sử dụng tiêu chuẩn phi tham số.

Với phương pháp phi tham số, các tiêu chuẩn kiểm

định dựa vào thứ hạng xếp theo độ lớn nhỏ của các giá trị quan sát, không sử dụng tham số trung bình và phương sai.

Do các tiêu chuẩn phi tham số không chính xác bằng các tiêu chuẩn tham số, nên nếu điều kiện kiểm định tham số được thỏa mãn thì không nên sử dụng tiêu chuẩn phi tham số.

Phương pháp phi tham số dùng để so sánh hai hay nhiều trị trung bình của hai hay nhiều mẫu rút từ các tổng thể có nguồn gốc khác nhau (mẫu độc lập) hoặc có cùng nguồn gốc (mẫu phụ thuộc).

3.1.2.1. So sánh các trung bình các mẫu độc lập

- ***So sánh trung bình hai mẫu độc lập***

Tiêu chuẩn U của Mann và Whitney

Để áp dụng tiêu chuẩn U, hãy xét ví dụ sau đây.

Ví dụ: Để đánh giá tính đồng nhất của khu đất thí nghiệm tại Trại thực nghiệm Đại học Nông Lâm Tp. HCM, chúng tôi đã tiến hành đo chiều cao cây của giống bông S02-13 được trồng ở ba vị trí (ba lô), mỗi lô theo dõi chiều cao 10 cây. Kết quả như sau:

Chiều cao cây (cm) của các cây trong ba lô

Lô 1:	72	87	71	70	80	67	80	80	82	66
Lô 2:	97	95	90	81	92	91	95	96	84	72
Lô 3:	62	68	73	85	69	79	77	76	83	84

Để đánh giá lô 1 và 2:

Bước 1: Xếp hạng số liệu

Trước hết hãy sắp xếp hạng từ nhỏ đến lớn các số đo của cả hai lô theo thứ tự 1, 2, 3, ..., 20. Có thể sắp xếp thủ

công hay nhờ phần mềm Excel trên máy vi tính. Trường hợp các số có cùng độ lớn thì thứ hạng được chia đều cho mỗi số. Ở ví dụ này kết quả xếp hạng nhờ Excel như sau:

Lô 1	72	87	71	70	80	67	80	80	82	66
Hạng	*(5)*	*13*	*4*	*3*	*(7)*	*2*	*(7)*	*(7)*	*11*	*1*
Lô 2	97	95	90	81	92	91	95	96	84	72
Hạng	*20*	*17*	*14*	*10*	*16*	*15*	*17*	*19*	*12*	*(5)*

Ở đây có 2 hạng 5 cho số 72 theo thứ tự 5, 6; 3 hạng 7 cho số 80 theo thứ tự 7, 8, 9, vì thế mỗi số 72 có thứ hạng mới là 5,5, tức là (5 + 6)/2 và mỗi số 80 có thứ hạng mới là 8, tức là (7 + 8 + 9)/3. Việc xếp hạng đúng khi: $R_1 + R_2 = \frac{n(n+1)}{2}$; R_1 là tổng thứ hạng của lô 1 và R_2 là tổng thứ hạng của lô 2.

Kết quả xếp hạng lại như sau:

Lô 1	72	87	71	70	80	67	80	80	82	66	
Hạng	*5,5*	*13*	*4*	*3*	*8*	*2*	*8*	*8*	*11*	*1*	R_1=63,5
Lô 2	97	95	90	81	92	91	95	96	84	72	
Hạng	*20*	*17*	*14*	*10*	*16*	*15*	*17*	*19*	*12*	*5,5*	R_2=146,5

Kiểm tra: $R_1 + R_2 = \frac{n(n+1)}{2} = 63,5 + 146,5$

$$= \frac{20(20+1)}{2} = 210$$

Bước 2: Kiểm tra và đánh giá kết quả

Người ta đã chứng minh được rằng khi n_1 và n_2 (n_1 + n_2 = n) các phân phối U_1 (cho tổng thể của mẫu 1) và U_2 (cho tổng thể của mẫu 2) tiệm cận phân phối chuẩn với:

$$E(U) = \frac{n_1 n_2}{2} \text{ và } V(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

Khi đó sử dụng phép thử sau để đánh giá:

$$U_{TN} = \frac{\left| U_1 - \frac{n_1 n_2}{2} \right|}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \qquad (3 - 6)$$

với $$U_1 = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1 \qquad (3 - 7)$$

và $$U_2 = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - R_2$$

hoặc U_2 có thể tính:

$$U_2 = n_1 n_2 - U_1 \qquad (3 - 7')$$

và $U_{TN}(U_1) = - U_{TN}(U_2)$.

Nếu $U_{TN} \geq 1,96$ thì $U_1 \neq U_2$, ngược lại $U_{TN} < 1,96$ thì $U_1 \approx U_2$.

Ở ví dụ này: $R_1 = 63,5$; $R_2 = 146,5$. Thay vào công thức (3 - 7) và (3 - 7') ta có: $U_1 = 91,5$ và $U_2 = 8,5$.

Theo công thức (3 - 6) ta có:

$$U_{TN} = \frac{\left| 91,5 - \frac{10 \times 10}{2} \right|}{\sqrt{\frac{10 \times 10 (10 + 10 + 1)}{12}}} = 3,14 > 1,96$$

$U_1 \neq U_2$ cho thấy hai lô này có đất khác nhau.

Một cách tương tự, kết quả kiểm tra U_1 (lô 1) và U_3 (lô 3):

$R_1 = 104$; $R_3 = 106$; $U_1 = 51,0$; $U_3 = 49,0$; $U_{TN} = 0,08$

và giữa U_2 (lô 2) và U_3 (lô 3):

$R_2 = 143,5$; $R_3 = 66,5$; $U_1 = 11,5$; $U_3 = 88,5$; $U_{TN} = 2,91$

Với các kết quả này thì đất lô 1 và lô 3 đồng nhất và khác với lô 2 về độ tốt xấu.

Bây giờ hãy đánh giá độ đồng đều của các lô qua phép so sánh phương sai (phương pháp tham số) theo số liệu sau:

Lô	$\bar{x}$	S^2	S	CV(%)
1	75,5	51,17	7,15	9,5
2	89,3	64,01	8,00	9,0
3	75,6	57,82	7,60	10,1

So sánh phương sai lô 2 với lô 1:

$$F_{2/1} = \frac{64,01}{51,17} = 1,25 < F_{0,05} = 3,18$$

Hiển nhiên là $F_{3/1}$ và $F_{2/3}$ đều nhỏ hơn 3,18.

Điều đó cho thấy, tuy lô 2 đất có tốt hơn hai lô còn lại nhưng độ đồng đều trong từng lô của cả ba lô như nhau.

Có thể so sánh trung bình lô 1 và 2 theo công thức (3 - 2):

$$T_{TN} = \frac{\left|75,5 - 89,3\right|}{\sqrt{\frac{(10-1)51,17 + (10-1)64,01}{10+10-2}\left(\frac{1}{10} + \frac{1}{10}\right)}}$$

$= 4,07 > t_{0,05}$ với 18 độ tự do là 2,10.

Rõ ràng trung bình lô 2 khác với lô 1 và 3.

Mặc dù độ tốt xấu có khác nhau nhưng cả ba lô đều có hệ số biến động có thể chấp nhận được (CV $\leq$ 10%).

Như vậy, ở mỗi lô có thể bố trí mỗi lần nhắc lại cho thí nghiệm đồng ruộng.

Tiêu chuẩn U của Siegel và Tukey

Để kiểm tra tính đồng nhất của hai mẫu (hai lô) từ hai tổng thể có nguồn gốc khác nhau, Siegel và Tukey cũng xếp hạng chung cho cả hai lô hoàn toàn giống Mann và Whitney nhưng ký hiệu R_1 cho lô có dung lượng mẫu nhỏ, còn R_2 cho lô có dung lượng mẫu lớn. Nếu n_1 và n_2 đều > 9 hoặc $n_1 > 2$ và $n_2 > 20$ thì việc kiểm tra tính đồng nhất của hai lô được thực hiện theo tiêu chuẩn U sau đây:

$$U_{TN} = \frac{\left|2R_1 - n_1\left(n_1 + n_2 + 1\right) + 1\right|}{\sqrt{n_1\left(n_1 + n_2 + 1\right).\frac{n_2}{3}}} \qquad (3 - 8)$$

Nếu $U_{TN} \geq 1{,}96$ thì $U_1 \neq U_2$, ngược lại $U_{TN} < 1{,}96$ thì $U_1 \approx U_2$.

Theo ví dụ trên:

$$U_{TN} = \frac{\left|2\times 63{,}6 - 10\left(10+10+1\right) + 1\right|}{\sqrt{10\left(10+10+1\right).\frac{10}{3}}} = 3{,}10$$

Như vậy: $U_1 \neq U_2$.

Trong trường hợp $n_1 = n_2$ thì có thể thay R_1 hoặc R_2 vào công thức (3 - 8) và kết quả tương đương nhau. Ở đây, nếu thay $R_1 = 63{,}5$ bằng $R_2 = 146.5$, $U_{TN} = 3{,}17$.

- ***So sánh các trung bình nhiều mẫu độc lập***

Tiêu chuẩn H của Kruskal và Wallis

Để kiểm tra tính đồng nhất của nhiều mẫu từ nhiều tổng thể có nguồn gốc khác nhau, việc xây dựng tiêu H

cũng thực hiện sau khi xếp hạng chung cho tập hợp tất cả các số liệu các mẫu. Phương pháp xếp hạng hoàn toàn giống Mann và Whitney.

Người ta đã chứng minh được rằng một đại lượng ngẫu nhiên gồm n biến số được xếp hạng từ 1 đến n, tập hợp từ k mẫu, phân phối theo quy luật "khi bình phương" với $k - 1$ bậc tự do:

$$H = \frac{12}{n(n+1))} \sum_{k=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \qquad (3 - 9)$$

Trong đó: $n = \Sigma n_i$ là tổng dung lượng mẫu

R_i là tổng các hạng trong mẫu i $(i = \overline{1,k})$

k là số mẫu quan sát

Nếu $H \geq \chi^2_{0,05}$ thì các mẫu không thuần nhất.

Nếu $H < \chi^2_{0,05}$ thì các mẫu thuần nhất, tức là các mẫu xuất phát từ một tổng thể.

Ví dụ: Từ kết quả phân tích hàm lượng mùn (%) trong 3 lô thí nghiệm ở bảng ở trang sau, hãy so sánh tính đồng nhất của khu đất.

Lô 1:	12,3	12,5	13,1	13,6	13,8	14,2	14,7	14,9	15,3	n_1=9
Lô 2:	12,8	13,9	14,2	14,7	15,3	15,3	15,8	16,8	17,4	n_2=9
Lô 3:	13,9	14,9	15,7	15,7	15,8	16,5	16,8	17,3	18,5	n_3=9

Áp dụng phương pháp xếp hạng như trên ta có:

Lô 1:	12,3	12,5	13,1	13,6	13,8	14,2	14,7	14,9	15,3	n_1=9
Hạng	*1*	*2*	*4*	*5*	*6*	*9,5*	*11,5*	*13,5*	*16*	R_1=68,5
Lô 2:	12,8	13,9	14,2	14,7	15,3	15,3	15,8	16,8	17,4	n_2=9
Hạng	*3*	*7,5*	*9,5*	*11,5*	*16*	*16*	*20,5*	*23*	*26*	R_2=133
Lô 3:	13,9	14,9	15,7	15,7	15,8	16,5	16,8	17,3	18,5	n_3=9
Hạng	*7,5*	*13,5*	*18,5*	*18,5*	*20,5*	*22*	*24*	*25*	*27*	R_3=176,5

$$\sum_{i=1}^{k} R_i = 68,5 + 133 + 186,5 = 378$$

Kiểm tra lại kết quả xếp hạng:

$$\sum_{i=1}^{k} R_i = \frac{27(27+1)}{2} = 378$$

Như vậy việc xếp hạng là đúng.

Theo công thức (3 - 9) ta có: H = 10,34, $\chi^2_{0,05}$ = 5,99.

Như vậy, không có sự đồng nhất của 3 lô đất thí nghiệm.

3.1.2.2. So sánh trung bình hai mẫu phụ thuộc - Tiêu chuẩn tổng hạng theo dấu của Wilcoxon

Nếu dung lượng mẫu không đủ lớn, các tổng thể lại không theo luật phân phối chuẩn thì việc so sánh được thực hiện bằng phép nghiệm phi tham số bằng tổng hạng của Wilcoxon sau đây trên cơ sở các giả thiết:

- Các đại lượng ngẫu nhiên X_1 và X_2 của hai tổng thể nghiên cứu có thể phân phối theo một quy luật bất kỳ.

- Hai mẫu ngẫu nhiên rút ra từ hai tổng thể phải độc lập và có dung lượng tùy ý.

Khi đó giả thuyết kiểm định hai trung bình sẽ là:

H_0: Hai tổng thể có kỳ vọng bằng nhau với độ tin cậy $1 - \alpha$;

H_1: Hai tổng thể có kỳ vọng khác nhau với độ tin cậy $1 - \alpha$ (hoặc là $X_1 > X_2$ hoặc là $X_2 > X_1$).

Để đánh giá hai trung bình mẫu, trước hết phải xếp hạng từ nhỏ đến lớn các số đo của cả hai mẫu theo thứ tự 1, 2, ..., *n* như phương pháp xếp hạng đã nêu trong mục

2.2.1 trên đây. n là tổng dung lượng của hai mẫu: n_1 của X_1 và n_2 của X_2 ($n = n_1 + n_2$).

Nếu H_0 đúng thì tổng hạng (ký hiệu là T) của mẫu sẽ tỷ lệ thuận với dung lượng mẫu (cũng tức là nếu $n_1 = n_2$ thì T của mẫu 1 bằng T của mẫu 2) và T sẽ có kỳ vọng và phương sai như sau:

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2} \qquad (3 - 10)$$

và

$$\sigma_T^2 = \frac{n_1 n_2}{12}(n_1 + n_2 + 1) \qquad (3 - 11)$$

Trong kiểm định tổng hạng của Wilcoxon, hai đại lượng ngẫu nhiên X_1 và X_2 phân phối liên tục và việc xếp hạng sẽ không có thứ hạng trùng nhau. Thực tế không thể có hai cá thể/phần tử hoàn toàn giống nhau, nhưng khi quan trắc do lấy độ chính xác của đơn vị đo nên có sự trùng nhau. Lúc đó thứ hạng sẽ được chia đều (như đã thực hiện ở 3.1.2.1 trên đây). Giá trị phương sai thực nghiệm sẽ được tính:

$$\sigma_T^2 = \frac{n_1 n_2}{12}[n_1 + n_2 + 1] - \frac{\sum_j t_j(t_j^2 - 1)}{(n_1 + n_2)(n_1 + n_2 - 1)} \qquad (3 - 12)$$

Trong đó t_j là tần số của hạng ghép nhóm thứ *j*. Hiển nhiên, khi không có hạng ghép thì công thức (3 - 12) lại trở thành công thức (3 - 11).

Nếu $n_1 \leq 10$ và $n_2 \leq 10$

Sau khi tính được tổng hạng của mỗi mẫu, tra bảng giá trị tổng hạng Wilcoxon để tìm các giá trị tới hạn T_L và T_u và xác định:

- Nếu kỳ vọng của hai tổng thể giống nhau thì $T < T_u$ hoặc $T > T_L$.

- Nếu kỳ vọng của hai tổng thể khác nhau thì $T > T_u$ hoặc $T < T_L$.

Nếu $n_1 > 10$ và $n_2 > 10$

Trong kiểm định phi tham số tổng hạng, Wilcoxon đã chứng minh được rằng khi cả n_1 và n_2 đều lớn hơn 10 thì phân phối T sẽ tiệm cận với phân phối chuẩn. Khi đó việc so sánh trung bình của hai mẫu theo tiêu chuẩn U:

$$U_{TN} = \frac{T - \mu_T}{\sigma_T} \qquad (3 - 13)$$

Nếu $U_{TN} \geq u_{\alpha/2}$ thì kết luận rằng $\bar{X}_1$ khác $\bar{X}_2$ ở độ tin cậy 1 - α. Và, nếu $U_{TN} < u_{\alpha/2}$ thì kết luận rằng $\bar{X}_1$ không khác với $\bar{X}_2$ ở độ tin cậy 1 - α.

Ví dụ: Để đánh giá một giống bắp mới trong sản xuất một công ty đã hợp đồng với 15 gia đình trong vùng, yêu cầu mỗi gia đình chọn một đám đất đồng đều và gieo 2 giống: giống đang sản xuất phổ biến trong vùng làm đối chứng (ĐC) và giống mới do công ty mang xuống (GM). Họ còn yêu cầu việc thực hiện các công việc gieo trồng, chăm sóc phải thự hiện như nhau cho cả hai giống. Kết quả năng suất và phân hạng (trong ngoặc) được ghi trong Bảng 3.2 sau đây. Hãy so sánh năng suất 2 giống này ?

Giải:

Ở đây, $n_1 = n_2 = 15$; $n = n_1 + n_2 = 30$; T = 465

Thay số liệu vào (3-10) và (3-11) ta được:

$$\mu_T = (15 \times 31)/2 = 232,5$$

Bảng 3.2: Năng suất và phân hạng hai giống bắp

Gia đình	Năng suất (tạ/ha)		Gia đình	Năng suất (tạ/ha)	
	ĐC	GM		ĐC	GM
1	68,5 (*5,5*)	77,5 (*18*)	9	69,0 (*7*)	71,8 (*12*)
2	73,3 (*15*)	83,0 (*22*)	10	72,0 (*13*)	74,8 (*16*)
3	57,5 (*1*)	62,5 (*2*)	11	79,0 (*20*)	88,0 (*29*)
4	81,4 (*21*)	87,6 (*28*)	12	69,5 (*9*)	72,2 (*14*)
5	77,8 (*19*)	69,3 (*8*)	13	87,3 (*27*)	83,8 (*24*)
6	85,5 (*25*)	93,5 (*30*)	14	70,2 (*10*)	63,8 (*3*)
7	68,5 (*5,5*)	71,5 (*11*)	15	68,0 (*4*)	76,5 (*17*)
8	87,2 (*26*)	83,5 (*23*)		T = 465	

$$\sigma_T^2 = \frac{225}{12} \times 31 - \frac{6}{30 \times 29} = 581{,}24 \text{ và } \sigma_T = 24{,}11$$

Từ đó: $U_{TN} = \dfrac{T - \mu_T}{\sigma_T} = \dfrac{465 - 232{,}5}{24{,}11} = 9{,}64$

Với $\alpha = 0{,}05$, $u_{0{,}025} = 1{,}96$ và $\sigma = 0{,}01$, $u_{0{,}005} = 2{,}58$

Như vậy: giống GM > ĐC với độ tin cậy 99%.

3.1.2.3. So sánh các trung bình nhiều mẫu phụ thuộc - Tiêu chuẩn Friedman

Việc kiểm tra tính đồng nhất của các mẫu bằng phép nghiệm χ^2.

- Trước hết, xếp hạng thứ tự 1, 2, 3, ... giữa các phương án trong từng nơi (hoặc từng thời điểm), mỗi nơi một hàng.

- Sau đó tính tổng số hạng cho từng phương án theo từng cột.

- Cuối cùng là kiểm tra sự giống hay khác nhau giữa các phương án theo tiêu chuẩn χ^2:

$$\chi^2_{TN} = \frac{12}{ab(a+1)} \sum R_i^2 - 3b(a+1) \qquad (3 - 14)$$

Trong đó: a là số phương án

b là số nơi (số thời điểm)

R_i là tổng hạng của từng phương án $(i = \overline{1,a})$.

Nếu $\chi^2_{TN} \geq \chi^2_{0,05}$ với $a - 1$ độ tự do thì các phương án cho kết quả khác nhau, còn $\chi^2_{TN} < \chi^2_{0,05}$ thì các phương án khác nhau không đủ tin cậy.

Ví dụ 1: Trong một thử nghiệm 5 giống đậu xanh tại 3 xã Phước Tiến, Phước Thắng và Phước Đại, huyện Bác Ái, tỉnh Ninh Thuận vụ Hè Thu năm 2009, năng suất các giống tại các điểm thí nghiệm được ghi nhận trong Bảng 3.3. Câu hỏi đặt ra: năng suất của các xã này khác nhau không ?

Bảng 3.3: Năng suất (tạ/ha) của 5 giống đậu xanh và kết quả xếp hạng từng xã cho mỗi giống

Giống\Xã	Phước Tiến /hạng		Phước Thắng /hạng		Phước Đại /hạng	
NP 305	13,9 b	*3*	12,3 c	*1*	13,8 bc	*2*
ĐX208	17,8 a	*3*	16,4 a	*2*	15,6 ab	*1*
HL 89-E3	16,2 a	*3*	15,1 b	*1*	16,1 a	*2*
V 99-1	12,3 b	*2*	12,2 c	*1*	12,5 c	*3*
Agredec-01	13,6 b	*2*	12,9 c	*1*	14,6 ab	*3*
P	< 0,05		< 0,05		< 0,05	
ΣR_i		13		6		11

Giải:

Ở đây: $a = 3$, $b = 5$, $\Sigma R_1 = 13$, $\Sigma R_2 = 6$, $\Sigma R_3 = 11$

Thay giá trị vào công thức (3-14) ta được:

$$\chi^2_{TN} = \frac{12}{3\times5(3+1)}\left(13^2+6^2+11^2\right)-3\times5\times4$$

$= 5{,}20 < \chi^{2(2)}_{0,05} = 6{,}0$. Như vậy năng suất đậu xanh của 3 xã này không có sự khác nhau.

Ví dụ 2: Kết quả cân khối lượng 100 cây mầm một giống cải bẹ xanh *Brassica juncea* L. được gieo trên 4 loại giá thể khác nhau (TN1, NT2, NT3 và TN4) tại Bảo Lộc, Lâm Đồng được ghi ở bảng 3.4. Theo kết quả này, có sự khác nhau hay không về khối lượng cây mầm trên các loại giá thể?

Bảng 3.4: Khối lượng trung bình 100 cây mầm (g) của 4 nghiệm thức (NT) trên các lần lặp lại.

Lặp lại	Khối lượng trung bình 100 cây mầm (g)			
	NT1	NT2	NT3	NT4
I	4,5 (*1*)	5,5 (*4*)	5,2 (*2,5*)	5,2 (*2,5*)
II	4,7 (*1*)	5,3 (*4*)	4,8 (*2*)	5,0 (*3*)
III	4,7 (*1*)	5,1 (*4*)	5,0 (*3*)	4,8 (*2*)
ΣR_i	3	12	7,5	7,5

Ghi chú: Số liệu trong dấu ( ) là hạng từ nhỏ đến lớn của các NT cho mỗi lần lặp lại.

Giải:

Ở đây: $a = 4$, $b = 3$, $\Sigma R_1 = 3$, $\Sigma R_2 = 12$, $\Sigma R_3 = \Sigma R_5 = 7{,}5$

Thay giá trị vào công thức (3-14) ta được:

$$\chi^2_{TN} = \frac{12}{4\times 3(4+1)}\left(3^2 + 12^2 + 2\times 7{,}5^2\right) - 3\times 3\times 5$$

$$= 8{,}1 > \chi^{2(3)}_{0{,}05} = 7{,}8.$$

Như vậy, khối lượng cây mầm trên các loại giá thể có khác nhau.

Để biết được sự tốt xấu của 4 loại giá thể ta chỉ cần kiểm tra sự khác nhau giữa NT1 và NT3, NT2 và NT3 theo tiêu chuẩn U của Mann và Whitney hay của Siegel và Tukey đã nêu trên đây.

Kết quả kiểm tra theo tiêu chuẩn U của Mann và Whitney cho biết:

U_1 và U_3: $R_1 = 6$; $R_3 = 15$; $U_1 = 9$; $U_3 = 0$; $U_{TN} = 1{,}96$

U_2 và U_3: $R_2 = 14$; $R_3 = 7$; $U_1 = 1$; $U_3 = 8$; $U_{TN} = 1{,}53 < 1{,}96$

Như vậy $U_1 \neq U_3$ và $U_2 \approx U_3$

Từ đây, có thể nói NT2 là tốt nhất, NT1 là xấu nhất, NT2 hơn NT3 và NT4 không đủ mức tin cậy 95%.

3.2. SO SÁNH HAI PHƯƠNG SAI VÀ MỞ RỘNG

3.2.1. Cơ sở lý luận

Như đã đề cập ở chương 1, phương sai là tham số đặc trưng cho độ phân tán của đại lượng ngẫu nhiên, nói cách khác là sự khác biệt giữa các giá trị x_i của trong một tập hợp số liệu quan sát so với số trung bình. Nếu các x_i là số đo của một tổng thể thuần nhất thì phương sai phản ánh độ đồng đều của tổng thể. Việc so sánh hai phương sai của hai tổng thể loại này là so sánh sự đồng nhất của tổng thể này với sự đồng nhất của tổng thể khác, ví dụ như giữa

giống này với giống khác. Nếu các x_i là các giá trị của một tổng thể không thuần nhất này (ví dụ như giữa các giống), còn các y_i là giá trị của một tổng thể không thuần nhất khác (ví dụ như giữa các mức bón), thì việc so sánh phương sai của hai tổng thể này cho biết sự khác nhau giữa các mức của yếu tố này (giữa các giống) giống hay là khác với sự khác nhau giữa các mức của yếu tố khác (giữa các mức phân), tức là yếu tố nào khác nhau nhiều hơn. Nếu so sánh phương sai gây ra do sự khác nhau giữa các nghiệm thức với một loại phương sai khác do sai số (ngẫu nhiên) gây ra, thì việc so sánh đó cho biết giữa các nghiệm thức có khác nhau hay không. Trong trường hợp này, nếu như phương sai do các nghiệm thức khác với phương sai ngẫu nhiên không đủ tin cậy ở mức nào đó, thì ta nói rằng sự khác nhau giữa các nghiệm thức là do sai số (hay, cũng như sai số), thực sự là chúng không khác biệt nhau. Đây là phép phân tích phương sai (ANOVA) các nghiệm thức trong các loại thí nghiệm, sẽ được nói tới ở phần 2.

Việc so sánh phương sai σ_1^2 và σ_2^2 của hai tổng thể có hai trung bình tổng thể μ_1 và μ_2 không thể đo trực tiếp bằng khoảng hiệu số $\sigma_1^2 - \sigma_2^2$ như so sánh hai trung bình, bởi vì: phương sai đặc trưng cho độ phân tán, không là đặc trưng về vị trí, có đơn vị tính là bình phương của số đo theo dấu hiệu nào đó.

Người ta đã dùng phép tỷ số để so sánh hai phương sai mẫu, từ đó suy đoán cho tổng thể. Nếu hai phương sai bằng nhau, thương số sẽ bằng 1. Tuy nhiên do trong ước lượng, các phương sai tổng thể nằm trong những khoảng khác nhau, tỷ lệ giữa hai phương sai tuân theo luật phân phối Fisher – Snedecor $F(n_1, n_2)$ nên được trắc nghiệm bằng tiêu chuẩn F.

3.2.2. So sánh hai phương sai

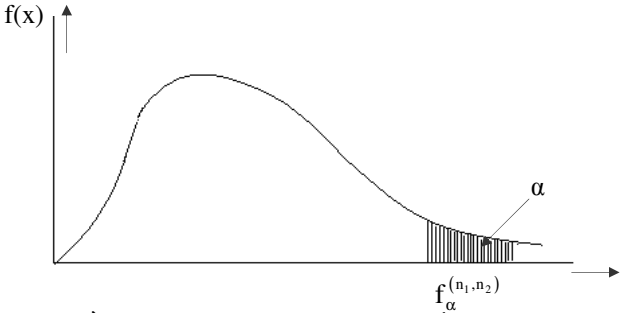
3.2.2.1. Luật phân phối Fisher – Snedecor F(n_1, n_2)

Đại lượng ngẫu nhiên liên tục F phân phối theo luật Fisher – Snedecor với n_1 và n_2 bậc tự do, hàm mật độ xác suất của nó được xác định bằng công thức :

$$f(x)=\begin{cases} 0 & \text{với } x \le 0 \\ \dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right) n_1^{\frac{n_1}{2}} n_2^{\frac{n_2}{2}}}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)} \cdot \dfrac{x^{\frac{n_1-n_2}{2}}}{\left(n_2+n_1 x\right)^{\frac{n_1+n_2}{2}}} & \text{với } x > 0 \end{cases}$$

trong đó $\Gamma(n)$ là hàm Gamma.

Và, đồ thị hàm mật độ xác suất có dạng như hình 3.1.



Hình 3.1: Đồ thị hàm mật độ xác suất theo luật phân phối Fisher – Snedecor

Người ta đã chứng minh được rằng nếu đại lượng ngẫu nhiên F phân phối theo luật phân phối Fisher – Snedecor với n_1 và n_2 bậc tự do thì kỳ vọng:

$$E(F)=\frac{n_2}{n_2-2}$$

và phương sai $V(F) = \frac{2n_2^2\left(n_1 + n_2^2 - 2\right)}{n_1\left(n_2 - 2\right)^2\left(n_2 - 4\right)}$.

3.2.2.2. Phép nghiệm Fisher – Snedecor F(n_1 -1, n_2 - 1)

Theo luật phân phối Fisher – Snedecor, với dung lượng mẫu của hai đại lượng ngẫu nhiên n_1 và n_2, số bậc tự do là (n_1 – 1) và (n_2 – 1), thỏa mãn điều kiện:

$$P(F \geq f_{\alpha}^{\left[(n_1-1)(n_2-1)\right]}) = \alpha$$

và có tính chất: $f_{\alpha}^{\left[(n_1-1)(n_2-1)\right]} = \frac{1}{f_{1-\alpha}^{\left[(n_1-1)(n_2-1)\right]}}$.

Với độ tin cậy 1 – α cho trước, có thể tìm được giá trị tới hạn Fisher – Snedecor f_α với (n_1 – 1) và (n_2 – 1) bậc tự do để kiểm định sự khác nhau của hai phương sai hai đại lượng ngẫu nhiên:

$$F_{TN} = \frac{S_1^2}{S_2^2} \qquad (S_1^2 > S_2^2) \qquad (3 - 15)$$

Nếu $F_{TN} \geq f_\alpha$ thì $S_1^2 > S_2^2$ ở độ tin cậy 1 - α, còn $F_{TN} < f_\alpha$ thì $S_1^2 \approx S_2^2$.

Ví dụ: Hãy so sánh phương sai F_1 (S_1^2 = 387,9, n_1 = 45) và phương sai F_2 (S_2^2 =768,61, n_2 = 110) của tổ hợp bông lai C92-52/C118A với số liệu trong ví dụ ở mục 3.2.1.4.

Giải: Ở đây giá trị $S_2^2 > S_1^2$, công thức (3 - 15) trở thành: $F_{TN} = \frac{S_2^2}{S_1^2}$. Thay S_1^2 và S_2^2 vào ta có: $F_{TN} = \frac{768,61}{387,90} =$ 1,98. Giá trị $f_{0,05}$ = 1,55; $f_{0,01}$ = 1,88 và $f_{0,001}$ = 2,34 (giá trị $f_{bảng}$ được tra trên phần mềm Excel với độ tự do tử số 109 và mẫu số 44).

Như vậy, phương sai F_2 lớn hơn phương sai F_1 với độ tin cậy 99%.

Đó là điều đương nhiên vì phương sai F_1 do ngẫu nhiên (môi trường đất) gây ra. Nếu đất hoàn toàn đồng nhất thì $S_1^2 = 0$ vì các cá thể F1 có cùng kiểu gen, còn phương sai S_2^2 của F_2 là vừa do sự phân ly về kiểu gen vừa do môi trường đất gây ra. Chênh lệch phương sai do sự khác nhau kiểu gen gây ra là: $S_2^2 - S_1^2 = 768,61 - 387,90 = 380,71$ và hệ số di truyền năng suất trong F_2 là: $H^2 = 380,71/768,61 = 0,495$ (49,5%). Đây là hệ số di truyền nghĩa rộng.

3.2.3. Đánh giá sự đồng nhất các phương sai của nhiều tổng thể

3.2.3.1. Khi dung lượng mẫu rút ra từ các tổng thể khác nhau

Nếu dung lượng mẫu của k phương sai mẫu S_1^2, S_2^2, ...S_k^2 là n_1, n_2, ... n_k $(i = \overline{1,k})$, $n_1 \neq n_2 \neq ... \neq n_k$, $h_i = (n_i - 1)$, $h = \Sigma h_i$ và $\overline{S}^2$ là trung bình số học của k phương sai:

$$\overline{S}^2 = \frac{\sum h_i S_i^2}{h}$$

Tiêu chuẩn Bartlett dùng để kiểm định sự đồng nhất của các phương sai là:

$$B = \frac{V}{C}, \text{ trong đó:}$$

$$V = 2,303\left[h \lg \overline{S}^2 - \sum h_i \lg S_i^2 \right]$$

$$C = 1 + \frac{1}{3(k-1)}\left[\sum \frac{1}{h_i} - \frac{1}{h}\right]$$

Các phương sai mẫu đồng nhất khi $B < \chi^{2(k-1)}_{0,05}$, ngược lại khi $B \geq \chi^{2(k-1)}_{0,05}$ thì các phương sai không đồng nhất.

***Ví dụ*:** Kết quả điều tra biến động năng suất cá thể (g/cây) của 5 giống bông thuần với n_1 = 26, n_2 = 32, n_3 = 29, n_4 = 30, n_5 = 19 và các phương sai tương ứng S_1^2 = 623,8, S_2^2 = 420,4, S_3^2 = 630,6, S_4^2 = 461,0 và S_5^2 = 586,6. Hãy kiểm định tính đồng nhất của các phương sai với độ tin cậy 95%.

Để tính B ta lập bảng sau

Mẫu	h_i	S_i^2	$h_i S_i^2$	$\lg S_i^2$	$h_i \lg S_i^2$	$1/h_i$
1	25	623,8	15.595,0	2,795	69,9	0,040
2	31	420,4	13.032,4	2,624	81,3	0,032
3	28	630,6	17.656,8	2,800	78,4	0,036
4	29	461,0	13.369,0	2,664	77,2	0,034
5	18	586,6	10.558,8	2,768	49,8	0,056
Σ	131		70.212,0		356,7	0,198

Từ đó:

$$\overline{S}^2 = \frac{\sum h_i S_i^2}{h} = \frac{70.212,0}{131} = 536,0$$

$$\lg \overline{S}^2 = 2,729$$

$$V = 2,303\left[h \lg \overline{S}^2 - \sum h_i \lg S_i^2\right]$$

$$= 2,303[(131)(2,729) - 356,7] = 1,982$$

$$C = 1 + \frac{1}{3(k-1)}\left[\sum \frac{1}{h_i} - \frac{1}{h}\right]$$

$$= 1 + \frac{1}{3(5-1)}\left[0,198 - \frac{1}{131}\right] = 1,016$$

$$B = \frac{V}{C} = \frac{1,982}{1,016} = 1,898 \; < \; \chi^{2(4)}_{0,05} = 9,488$$

Như vậy, các phương sai được xem là đồng nhất, tức là các giống đều thuần chủng.

3.2.3.2. Khi dung lượng mẫu rút ra từ các tổng thể bằng nhau

Trong trường hợp này, có thể dùng tiêu chuẩn Bartlett để kiểm tra. Tuy nhiên do phân phốixác suất theo tiêu chuẩn Bartlett chỉ là xấp xỉ nên kém chính xác. Khi kích thước mẫu bằng nhau, ta dùng tiêu chuẩn Cochran (G).

$$G = \frac{S^2_{max}}{S^2_1 + S^2_2 + ... + S^2_k}$$

Các phương sai mẫu đồng nhất khi $G < g_\alpha^{(n-1,k)}$, ngược lại khi $G \geq g_\alpha^{(n-1,k)}$ thì các phương sai không đồng nhất.

$g_\alpha^{(n-1,k)}$ là giá trị tới hạn của phân phối Cochran được tra trong bảng phụ lục 7.

Ví dụ: Cũng với các giống trên, nếu lấy dung lượng mẫu bằng nhau và bằng 19, các phương sai mẫu sẽ là 653,2; 466,1; 671,3; 581,4 và 586,6. Hãy kiểm định tính đồng nhất của các phương sai với độ tin cậy 95% và ước lượng phương sai tổng thể nếu các phương sai đồng nhất.

Giá trị Cochran từ mẫu quan sát:

$$G = \frac{671,3}{653,2+466,1+671,3+581,4+586,6} = 0,227$$

Với $\alpha = 0,05$, số bậc tự do là 19 – 1 = 18, số lượng mẫu là 5, giá trị tới hạn tra được là $g_{\alpha}^{(n-1,k)} = g_{0,05}^{(18,5)} = 0,3645$

$G < g_{0,05}^{(18,5)}$ cho thấy các phương sai là đồng nhất, độ đồng đều các giống là như nhau.

Phương sai tổng thể được ước lượng:

$$\sigma^2 = \overline{S}^2 = \frac{653,2+466,1+671,3+581,4+586,6}{5}$$

$$= 591,7$$

3.3. ĐÁNH GIÁ TÍNH ĐỘC LẬP CỦA CÁC DẤU HIỆU ĐỊNH TÍNH

Người ta sử dụng tiêu chuẩn khi bình phương (χ^2) để xác định mối quan hệ giữa hai dấu hiệu định tính.

Giả sử có một tổng thể có hai dấu hiệu định tính A và B. Giả thiết H_0 : A và B độc lập và H_1: A và B phụ thuộc.

Để kiểm tra các giả thiết này, từ tổng thể có dung lượng mẫu n, lập bảng trình bày các đặc trưng A, B và tần số tương ứng:

A \ B	B_1	B_2	...	B_j	Tổng A_i
A_1	n_{11}	n_{12}	...	n_{1j}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2j}	$n_{2.}$
...	...	...	...	...	...
A_i	n_{i1}	n_{i2}	...	n_{ij}	$n_{i.}$
Tổng B_j	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	$\Sigma n_{ij} = n$

Trong bảng:

n là dung lượng mẫu.

n_{ij} là là tần số ứng với các mức độ của A_i $(i = \overline{1,i})$ và B_j $(j = \overline{1,j})$.

$n_{i.}$ là là tần số ứng với các mức độ của dấu hiệu A.

$n_{.j}$ là là tần số ứng với các mức độ của dấu hiệu B.

Tính độc lập của hai dấu hiệu A và B được kiểm tra theo tiêu chuẩn khi bình phương (χ^2):

$$\chi^2_{TN} = n\left(\sum_{i=1}^{i}\sum_{j=1}^{j}\frac{n_{ij}^2}{n_{i.}n_{.j}} - 1\right) \qquad (3 - 16)$$

Nếu $\chi^2_{TN} \geq \chi^2_{\alpha}$ với $(i - 1)(j - 1)$ độ tự do thì bác bỏ H_0 và chấp nhận H_1, còn $\chi^2_{TN} < \chi^2_{\alpha}$ thì chấp nhận H_0 và bác bỏ H_1 với độ tin cậy $1 - \alpha$.

Ví dụ: Kết quả điều tra mức độ lông của lá bông và mức độ kháng rầy xanh được ghi ở Bảng 3.5. Vậy, tính có lông có quan hệ với mức độ kháng rầy không ?

Bảng 3.5: Kết quả điều tra tính kháng rầy 50 giống bông

Mức độ lông của lá	KTB	K	RB	Tổng, A_i
Ít	3	0	0	3
Vừa	7	3	0	10
Nhiều	5	8	5	18
Rất nhiêu	3	5	11	19
Tổng, B_j	18	16	16	$\Sigma n_{ij} = 50$

Ghi chú : KTB – kháng trung bình; K – kháng ; RK – rất kháng.

Giải:

Ở đây: i = 4; j = 3; n = 50; $n_{i.}$ = 3, 10, 18 và 19; $n_{.j}$ = 18, 16 và 16.

Thay giá trị vào công thức (3 - 16) ta được:

$\chi^2_{TN} = 50 \times \{[3^2/(18 \times 3) + 7^2/(18 \times 10) + 5^2/(18 \times 18) + ...$

$+ 5^2/(16 \times 18) + 11^2/(16 \times 19)] - 1\} = 19{,}40$

$\chi^2_{TN} = 19{,}40 > \chi^{2(6)}_{0,01} = 16{,}81.$

Như vậy, tính có lông có quan hệ chặt chẽ với mức độ kháng rầy với độ tin cậy 99%.

Để giải bài này có thể sử dụng phương pháp sau:

Trước hết tính tần số lý thuyết n'_{ij} cho các n_{ij}: $n'_{ij} = \frac{n_{i.} n_{.j}}{n}$. Ví dụ cho n'_{32} là: (16 × 18)/50 = 5,76 trong khi đó n_{32} = 8. Sau khi tính xong n'_{ij}, kiểm tra sự khác nhau giữa tần số lý thuyết với tần số thực nghiệm theo công thức:

$$\chi^2_{TN} = \sum \frac{\left(n_{ij} - n'_{ij}\right)^2}{n'_{ij}} \qquad (3 - 17)$$

Kết quả tính các tần số n'_{ij} được ghi ở bảng sau (số trong ngoặc):

Độ lông	KTB	K	RK	Tổng, A_i
Ít	3 (1,08)	0 (0,96)	0 (0,96)	3
Vừa	7 (3,60)	3 (3,20)	0 (3,20)	10
Nhiều	5 (6,48)	8 (5,76)	5 (5,76)	18
Rất nhiêu	3 (6,84)	5 (6,08)	11 (6,08)	19
Tổng, B_j	18	16	16	T = 50

Thay giá trị vào công thức (3 - 17) ta được:

χ^2_{TN} = [(3 – 1,08)^2/1,08 + (7 – 3,6)^2/3,6 + (5 – 6,48)^2/6,48 + ... + (5 – 5,76)^2/5,76 + (11 – 6,08)^2/6,08)] = 19,40 > $\chi^{2(6)}_{0,01}$ = 16,81.

Tiêu chuẩn χ^2 có thể sử dụng rộng rãi để kiểm tra các tần số thực nghiệm khi đã biết các tần số lý thuyết.

Chương 4
PHÂN TÍCH MỐI QUAN HỆ

4.1. CÁC LOẠI QUAN HỆ

Phép ước lượng các tham số thống kê là phép mô tả tổng thể từ các mẫu điều tra khảo sát theo một chỉ tiêu đó, còn phép so sánh là phép phân định sự khác nhau hay giống nhau giữa các tổng thể theo một hay một số tham số nhất định. Chương này sẽ đề cập đến mối quan hệ giữa các đặc trưng trong một tổng thể và giữa các tổng thể.

Nếu coi các số đo nhận được từ điều tra khảo sát về một chỉ tiêu nào đó của mẫu rút ra từ tổng thể là các biến số của một đại lượng ngẫu nhiên thì mối quan hệ giữa hai hay nhiều tổng thể, giữa hai hay nhiều đặc trưng trong một tổng thể là mối quan hệ giữa hai hay nhiều đại lượng ngẫu nhiên.

Các mối quan hệ được biểu thị bởi các hàm số phụ thuộc gọi là phương trình hồi quy. Tùy các mối quan hệ khác nhau người ta chia ra:

1. Tương quan và hồi quy tuyến tính (đường thẳng), gồm:

- Tương quan và hồi quy và tuyến tính một biến
- Tương quan và hồi quy và tuyến tính đa biến

2. Tương quan và hồi quy phi tuyến tính (đường cong), gồm:

- Tương quan và hồi quy và phi tuyến tính một biến

- Tương quan và hồi quy và phi tuyến tính đa biến.

4.2. QUAN HỆ TUYẾN TÍNH

4.2.1. Các dạng quan hệ tuyến tính

Phương trình biểu thị mối quan hệ tuyến tính giữa X và Y có dạng:

$$y = f(x) = a + bx \qquad (4 - 1)$$

hoặc $$y = f(x_i) = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n \qquad (4 - 2)$$

Phương trình (4 - 1) gọi là phương trình hồi quy tuyến tính một biến, trong đó *y* là hàm số (số phụ thuộc), *x* là đối số (số độc lập). Ứng với mỗi giá trị của *x* ta có một giá trị xác định tương ứng của *y*, b là hệ số góc, còn gọi là hệ số hồi quy và a là hằng số. Trên đồ thị hai chiều (trục tung *y* và trục hoành *x*), đồ thị hàm số cắt trục tung tại điểm a. Nếu a > 0 đồ thị đi qua phía trên góc tọa độ, ngược lại, nếu a < 0 đồ thị đi qua phía dưới góc tọa độ và đồ thị đi qua góc tọa độ khi a = 0. Về chiều hướng và mức độ của mối quan hệ phụ thuộc vào hệ số tương quan *r* (sẽ xét cụ thể ở mục hệ số tương quan).

Người ta gọi là tương quan tuyến tính vì đường biểu diễn của phương trình hồi quy $\bar{y} = a + b\bar{x}$ từ các mẫu số quan sát là một đường thẳng và $\hat{y} = \hat{a} + b\hat{x}$ là phương trình của đường hồi quy lý luận.

- Phương trình (4 - 2) gọi là phương trình hồi quy tuyến tính bội (đa biến), trong đó y là hàm số - số phụ thuộc, còn x_i $\left(i = \overline{1,n}\right)$ là các đối số - các số độc lập. Phương trình của đường hồi quy lý luận được ước lượng từ các mẫu số quan sát là: $\hat{y} = \hat{b}_0 + \hat{b}_1\hat{x}_1 + \hat{b}_1\hat{x}_2 + ... + \hat{b}_n\hat{x}_n$

4.2.2. Mô hình tuyến tính đơn các đặc trưng định lượng

4.2.2.1. Hệ số tương quan đơn và đánh giá sự tồn tại của hệ số tương quan

- ***Hệ số tương quan đơn***

Để đo mức độ quan hệ tuyến tính giữa hai đại lượng ngẫu nhiên X và Y, người ta đưa ra khái niệm *hệ số tương quan*. *Hệ số tương quan lý luận* giữa X và Y, ký hiệu là ρ, được định nghĩa bởi công thức:

$$\rho = \frac{E(X-\mu_X)(Y-\mu_Y)}{\sigma_X \sigma_Y}$$

trong đó: μ_X, σ_X là kỳ vọng và độ lệch chuẩn lý luận của X và μ_Y, σ_Y là là kỳ vọng và độ lệch chuẩn lý luận của Y.

Người ta đã chứng minh được $-1 \le \rho \le 1$.

Trong thực nghiệm, hệ số tương quan giữa X và Y, ký hiệu là r hoặc r_{XY}, được tính bằng công thức:

$$r_{XY} = \frac{Cov(X,Y)}{S_x S_y}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

$$= \frac{\overline{xy} - \overline{x}.\overline{y}}{\sqrt{\left[\overline{x}^2 - (\overline{x})^2\right]\left[\overline{y}^2 - (\overline{y})^2\right]}} = \frac{\overline{xy} - \overline{x}.\overline{y}}{\sqrt{Var(x)Var(y)}}$$

Tính chất của hệ số tương quan

1. $-1 \le r_{XY} \le 1$

2. Nếu S_x và S_y đều khác 0 và r_{XY} càng gần 1 hay càng gần -1 thì X và Y có quan hệ hàm số chặt chẽ.

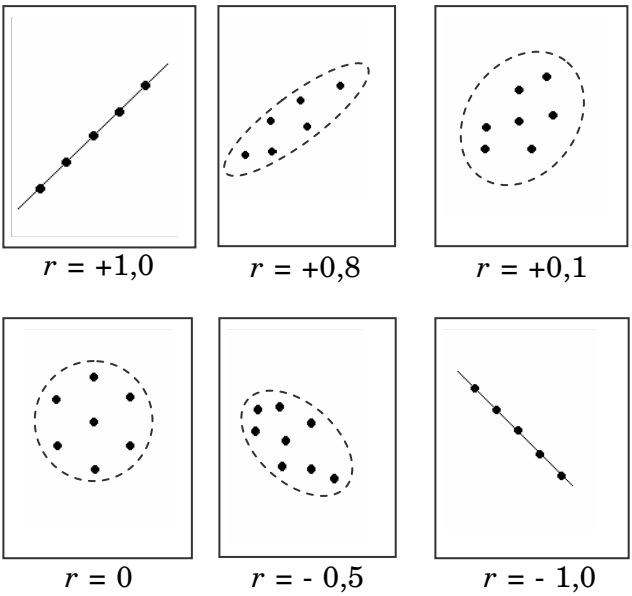
3. Nếu S_x và S_y đều khác 0 và r_{XY} càng gần 0 thì X và Y không có quan hệ tuyến tính hoặc độc lập với nhau.

4. Nếu r_{XY} càng gần 0 và:

- $S_x \approx 0$ thì X độc lập với Y
- $S_y \approx 0$ thì Y độc lập với X

5. Nếu $r_{XY} > 0$ thì X và Y đồng biến

$r_{XY} < 0$ thì X và Y nghịch biến



Hình 4.1: Sơ đồ biểu thị hệ số tương quan *r* đơn với sự biến động các giá trị

Có thể dựa vào độ lớn của hệ số tương quan để xác định giá mức độ quan hệ giữa hai đại lượng.

$|r| \simeq 0$: X và Y độc lập hoặc có quan hệ phi tuyến tính

$0 < |r| \leq 0{,}3$: X và Y có quan hệ yếu

$0{,}3 < |r| \leq 0{,}5$: X và Y có quan hệ vừa

$0{,}5 < |r| \leq 0{,}7$: X và Y có quan hệ tương đối chặt

$0{,}7 < |r| \leq 0{,}9$: X và Y có quan hệ chặt

$0{,}9 < |r| \leq 1$: X và Y có quan hệ rất chặt

- ***Đánh giá sự tồn tại của hệ số tương quan***

Hệ số tương quan biểu thị mức độ quan hệ giữa hai đại lượng. Tuy nhiên, mối quan hệ có thực (tồn tại) hay không phụ thuộc vào độ tin cậy (nói cách khác là mức ý nghĩa) của *r*. Có những mối quan hệ "chặt" nhưng chưa đủ tin cậy (do dung lượng mẫu quá ít), nhưng có những mối quan hệ yếu, thậm chí hai đại lượng không quan hệ với nhau ($|r| \simeq 0$) nhưng đáng tin cậy (dung lượng mẫu lớn).

Trong thực nghiệm, có hai cách để đánh giá độ tin cậy của hệ số tương quan *r*:

1. So sánh hệ số tương quan thực nghiệm r_{xy} với giá trị *r* lý luận:

Hệ số tương quan lý luận được tính sẵn với độ tự do $n - 2$ ở các mức tin cậy khác nhau (phụ lục 4). Nếu $r_{xy} \geq r_{\alpha}$ với $n - 2$ bậc tự do, ta nói r_{xy} tồn tại với độ tin cậy $1 - \alpha$.

Ví dụ: với n = 20, độ tự do là 18, nếu $r_{xy} \geq 0{,}444$ thì *r* tồn tại với độ tin cậy 95% (sai lầm < 5%); nếu $r_{xy} \geq 0{,}561$ thì *r* tồn tại với độ tin cậy 99% (sai lầm < 1%) và nếu $r_{xy} \geq 0{,}679$ thì *r* tồn tại với độ tin cậy 99,9% (sai lầm < 0,1%).

2. Tra độ tin cậy r qua giá trị tới hạn t trong bảng Student:

Đặt: $$T = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}}$$

Tra bảng T với n - 2 bậc tự do ta được các trị tới hạn t với các mức á khác nhau. Nếu $T \geq t_{\alpha}^{(n-2)}$, ta nói r_{xy} tồn tại với độ tin cậy $1 - \alpha$.

Ví dụ: với n = 20, độ tự do là 18, $t_{0,05}^{(18)} = 2,101; t_{0,01}^{(18)} = 2,878; t_{0,001}^{(18)} = 3,922$. Nếu $r_{xy} \geq 0,486$, tính theo công thức trên ta có: T = 2,359 > 2,101, ta nói r tồn tại với độ tin cậy 95%. Nếu $r_{xy} \geq 0,582$ ta có: T = 3,036 > 2,878, ta nói r tồn tại với độ tin cậy 99% và nếu $r_{xy} \geq 0,701$ ta có T = 4,170 > 3,922, ta nói r tồn tại với độ tin cậy 99,9%.

Thông thường, người ta ghi sau hệ số tương quan thực nghiệm các ký hiệu : (*), (**) và (***) tương ứng với các độ tin cậy 95%, 99% và 99,9% của r.

4.2.2.2. Xác định các hệ số a, b và ý nghĩa của a, b

- ***Xác định các hệ số* â, b**

Để dự đoán mô hình tuyến tính đơn $\hat{y} = \hat{a} + b\hat{x}$, phải xác định giá trị tối thích â và b. Để tìm các giá trị này ta sử dụng phương pháp bình phương tối thiểu do Karl Pearson (1857 – 1936) đề xuất.

Theo phương pháp bình phương tối thiểu, â và b tối thích khi tổng bình phương độ lệch từ các giá trị thực nghiệm y_i và giá trị lý luận $\hat{y}$ là nhỏ nhất, tức là:

$$Q = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \text{ là nhỏ nhất.}$$

Do $Q \geq 0$ nên nhất định sẽ tìm được $\hat{a}$ và b khi Q nhỏ nhất.

Tại điểm $\hat{a}$ và điểm b: $\frac{\partial Q}{\partial \hat{a}} = 0$ và $\frac{\partial Q}{\partial b} = 0$, tức là:

$$\frac{\partial \sum_{i=1}^{n} (y_i - \hat{a}x_i - b)^2}{\partial \hat{a}} = 0$$

$$\frac{\partial \sum_{i=1}^{n} (y_i - \hat{a}x_i - b)^2}{\partial b} = 0$$

hay

$$\begin{cases} nb + \hat{a}\sum x_i = \sum y_i \\ b\sum x_i + \hat{a}\sum x_i^2 = \sum x_i y_i \end{cases}$$

Giải hệ phương trình ta được:

$$\hat{a} = \frac{\sum x_i y_i - \sum x_i \sum y_i}{\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x}^2 - \left(\overline{x}\right)^2}$$

$$b = \frac{\sum y_i}{n} - \hat{a}\frac{\sum x_i}{n} = \overline{y} - \hat{a}\overline{x}$$

- ***Kiểm định ý nghĩa của* $\hat{a}$, b**

Để kiểm định $\hat{a}$ ta sử dụng:

$$T = \frac{\hat{a} - a_0}{s_e(\hat{a})} \text{ so với } t_{\alpha/2}^{(n-2)}$$

trong đó: $$s_e(\hat{a}) = \sqrt{\frac{\sum(y_i - \hat{y})^2}{(n-2)\sum(x_i - \bar{x})^6}}$$

Để kiểm định b ta sử dụng:

$$T = \frac{b - b_0}{s_e(b)} \text{ so với } t_{\alpha/2}^{(n-2)}$$

trong đó: $$s_e(b) = \sqrt{\frac{\sum(y_i - \hat{y})^2 \sum x_i^2}{n(n-2)\sum(x_i - \bar{x})^2}}$$

4.2.2.3. Kiểm định độ tin cậy của hàm hồi quy tuyến tính

Để kiểm định sự tồn tại của hàm hồi quy ta sử dụng phép nghiệm F:

$$F = \frac{R^2}{\frac{(1-R^2)}{(n-2)}} \text{ so với } f_\alpha(1, n\text{-}2)$$

trong đó: $R^2 = r_{xy}^2 = \frac{SS_{Regression}}{SS_T}$ hay $R^2 = 1 - \frac{SS_{Residual}}{SS_T}$

$$SS_T = \sum(y_i - \bar{y})^2 = \sum y_i^2 - n(\bar{y})^2$$

$$SS_{Regr.} = SS_T - SS_{Residual}$$

$$SS_{Residual} = (\hat{a})^2 \left[\sum x_i^2 - n(\bar{x})^2\right]$$

R^2 được gọi là *hệ số xác định* (determinant coefficient). R^2 cho biết tỷ lệ phần biến động do X gây nên trong quan hệ tuyến tính ($SS_{Regression}$) so với tổng số biến động (SS_T) của Y. Tỷ lệ này được tính bằng phần trăm (%)

khi nhân R^2 với 100. Phần trăm còn lại là do các yếu tố khác gây nên trong quan hệ đa biến (nếu có) và do sai số ngẫu nhiên.

R^2 càng lớn thì Y càng phụ thuộc vào X.

Trong phép nghiệm F, nếu $F > f_{\alpha}(1, n\text{-}2)$ thì hàm hồi quy tuyến tính có độ tin cậy $1 - \alpha$.

Ví dụ:

Số quả /cây (x)	9,0	10,7	10,7	13,1	9,8	14,4	10,6	11,4	8,6	12,2
NS, tạ/ha (y)	23,6	31,2	28,8	37,3	28,5	37,1	30,8	28,1	18,3	26,2

Kết quả xử lý trên phần mềm Excel

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.831
R Square	0.691
Adjusted R Square	0.652
Standard Error	3.377
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regression	1	203.93	203.93	17.88	0.0029
Residual	8	91.24	11.41		
Total	9	295.17			

	Coefficients	*SE*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.27	7.001	-0.039	0.9702	-16.41	15.88
x	2.65	0.626	4.229	0.0029	1.20	4.09

Ở ví dụ này, hệ số tương quan $r = 0{,}831^{**}$

Hàm hồi quy tuyến tính $y = 2{,}65x - 0{,}27$ tồn tại với độ tin cậy 99% (mức ý nghĩa của F = 0,0029).

4.2.3. Mô hình tuyến tính đa biến

4.2.3.1. Hệ số tương quan riêng

Hệ số tương quan riêng được tính toán trong nghiên cứu quan hệ đa biến.

Có m đại lượng ngẫu nhiên X_1, X_2. ... X_m có quan hệ với nhau. Hệ số tương quan riêng của X_1 với X_2 khi cố định X_3, ký hiệu $r_{12,3}$ sẽ là:

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1-r_{13}^2\right)\left(1-r_{23}^2\right)}} \qquad (4 - 3)$$

Tương tự, ta có hệ số tương quan riêng của X_i với X_j khi cố định X_k ($r_{ij,k}$) sẽ là:

$$r_{ij,k} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{\left(1-r_{ik}^2\right)\left(1-r_{jk}^2\right)}}$$

Hệ số tương quan riêng của X_1 với X_2 khi cố định X_3 và X_4 ($r_{12,34}$) là:

$$r_{12,34} = \frac{r_{12,3} - r_{14,3}r_{24,3}}{\sqrt{\left(1-r_{14,3}^2\right)\left(1-r_{24,3}^2\right)}} \qquad (4 - 4)$$

Một cách tổng quát, hệ số tương quan riêng của X_1 với X_2 khi cố định X_3, X_4, ... X_m ($r_{12,34...m}$) là:

$$r_{12,34...m} = \frac{r_{12,34...(m-1)} - r_{1m,34...(m-1)}r_{2n,34...(m-1)}}{\sqrt{\left(1-r_{1m,34...(m-1)}^2\right)\left(1-r_{2m,34...(m-1)}^2\right)}}$$

Ví dụ: Tính $r_{12,3}$ và $r_{13,2}$ theo Bảng 4.1 sau.

Thay số liệu vào công thức (4-3 và (4-4) ta có:

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1-r_{13}^2\right)\left(1-r_{23}^2\right)}}$$

Bảng 4.1: Hệ số tương quan đơn giữa 7 đại lượng

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	1						
x_2	0,244	1					
x_3	0,627	-0,228	1				
x_4	0,609	-0,363	0,794	1			
x_5	0,461	-0,249	0,894	0,689	1		
x_6	0,636	-0,069	0,701	0,628	0,687	1	
x_7	0,638	0,705	0,524	0,240	0,438	0,438	1

$$= \frac{0,244 - 0,627 \times (-0,228)}{\sqrt{\left(1-0,627^2\right)\left[1-(-0,228)^2\right]}} = 0,510$$

và:
$$r_{13,2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{\left(1-r_{12}^2\right)\left(1-r_{32}^2\right)}}$$

$$= \frac{0,627 - 0,244 \times (-0,228)}{\sqrt{\left(1-0,244^2\right)\left[1-(-0,228)^2\right]}} = 0,723$$

4.2.3.2. Xác định phương trình hồi quy tuyến tính

- ***Cách tính hệ số hồi quy***

Như đã nêu ở trên, phương trình hồi quy tuyến tính lý luận có dạng:

$$\hat{y} = \hat{b}_0 + \hat{b}_1\hat{x}_1 + \hat{b}_1\hat{x}_2 + ... + \hat{b}_n\hat{x}_n$$

Theo nguyên lý của phương pháp bình phương tối

thiểu, có thể tìm được các hệ số $\hat{b}_0$, $\hat{b}_1$, $\hat{b}_2$, ..., $\hat{b}_n$ sao cho:

$$Q = \sum_i (y_i - \hat{y}_i)^2 = \sum \left(y_i - \hat{b}_0 - \hat{b}_1 x_{i1} - ... - \hat{b}_n x_{in} \right)^2$$

là tối thiểu.

Ví dụ, với phương trình $\hat{y} = \hat{b}_0 + \hat{b}_1 \hat{x}_1 + \hat{b}_2 \hat{x}_2$

Ta tìm $\hat{b}_0$, $\hat{b}_1$, $\hat{b}_2$ sao cho $Q = \sum \left(y_i - \hat{b}_0 - \hat{b}_1 x_{i1} - \hat{b}_2 x_{i2} \right)^2$ nhỏ nhất.

Tại ba điểm $\hat{b}_0$, $\hat{b}_1$ và $\hat{b}_2$:

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = \sum \left(y_i - \hat{b}_0 - \hat{b}_1 x_{i1} - \hat{b}_2 x_{i2} \right) = 0 \\ \dfrac{\partial Q}{\partial b_1} = \sum \left(y_i - \hat{b}_0 - \hat{b}_1 x_{i1} - \hat{b}_2 x_{i2} \right) x_{i1} = 0 \\ \dfrac{\partial Q}{\partial b_2} = \sum \left(y_i - \hat{b}_0 - \hat{b}_1 x_{i1} - \hat{b}_2 x_{i2} \right) x_{i2} = 0 \end{cases}$$

hay

$$\begin{cases} nb_0 + \left(\sum_{i=1}^{n} x_{i1} \right) b_1 + \left(\sum_{i=1}^{n} x_{i2} \right) b_2 = \sum_{i=1}^{n} y \\ \left(\sum_{i=1}^{n} x_{i1} \right) b_0 + \left(\sum_{i=1}^{n} x_{i1}^2 \right) b_1 + \left(\sum_{i=1}^{n} x_{i1} x_{i2} \right) b_2 = \sum_{i=1}^{n} x_{i1} y \\ \left(\sum_{i=1}^{n} x_{i2} \right) b_0 + \left(\sum_{i=1}^{n} x_{i2} x_{i1} \right) b_1 + \left(\sum_{i=1}^{n} x_{i2}^2 \right) b_2 \\ \quad = \sum_{i=1}^{n} x_{i2} y \end{cases}$$

Giải hệ phương trình này ta được $\hat{b}_0$, $\hat{b}_1$ và $\hat{b}_2$

• ***Kiểm định độ tin cậy của hàm hồi quy tuyến tính***

Để kiểm định sự tồn tại của hàm hồi quy ta sử dụng phép nghiệm F:

$$F = \frac{R^2(n-p)}{(1-R^2)(p-1)} \text{ so với } f_\alpha(p-1,\ n-p)$$

Trong đó: $R^2 = 1 - \frac{SS_{Residual}}{SS_T}$

p là số đại lượng ngẫu nhiên được nghiên cứu trong quan hệ với y.

$$SS_T = \|y_i - \bar{y}\|^2$$

$$SS_{Regr.} = SS_T - SS_{Residual}$$

$$SS_{Residual} = \|\hat{y} - \bar{y}\|^2$$

• ***Kiểm định ý nghĩa của các hệ số*** $\hat{b}_i \left(i = \overline{0,1,n}\right)$

Để kiểm định sự tồn tại $\hat{b}_i \neq 0$ ta sử dụng:

$$T = \frac{\hat{b}_i}{s_e\left(\hat{b}_i\right)} \text{ so với } t_{\alpha/2}^{(n-p)}$$

Trong đó: $s_e\left(\hat{b}_i\right) = \sqrt{MS_{(\hat{b}_i)}}$

Ví dụ: Kết quả theo dõi 7 chỉ tiêu trên 36 giống lúa được ghi ở Bảng 4.2. Hãy xác định phương trình quan hệ với năng suất (g/bụi).

Bảng 4.2: Kết quả theo dõi 7 chỉ tiêu trên 36 giống lúa

Giống	Cao cây (cm)	TS nhánh	Nhánh HH	Dài bông (cm)	Số hạt /bông	M100 hạt (g)	NSTT (g/bụi)	NS hồi quy (g/bụi)
1	109,5	46,1	41,9	20,3	52,5	3,9	85,7	86,2
2	119,1	36,2	33,4	20,0	63,2	4,0	83,9	85,4
3	114,4	47,6	44,6	21,7	48,6	4,4	96,4	94,3
4	119,7	57,2	55,2	20,0	52,0	4,1	117,0	114,1
5	117,0	44,0	42,1	21,2	64,1	4,3	115,8	108,4
6	121,5	35,9	34,0	22,1	47,6	4,3	68,7	69,8
7	120,9	53,1	50,4	22,1	57,1	4,8	108,5	122,4
8	119,5	33,3	30,8	20,0	40,9	4,0	50,7	50,7
9	126,8	46,7	43,8	19,8	61,1	3,7	98,3	98,2
10	111,7	43,7	42,3	20,2	77,4	3,3	108,1	111,8
11	120,1	41,5	40,5	21,7	51,1	4,3	88,7	87,8
12	121,2	52,5	50,7	22,0	51,2	3,9	101,4	101,5
13	127,8	30,6	28,3	22,5	66,8	3,8	72,3	78,2
14	127,7	42,9	40,1	21,9	52,6	4,4	92,7	90,7
15	131,8	49,7	47,5	20,1	41,5	4,3	84,4	88,4
16	111,8	42,1	37,3	18,7	43,2	4,6	74,1	75,6
17	117,5	40,2	38,1	19,4	56,7	4,7	101,7	97,1
18	109,8	47,6	45,9	20,1	45,5	4,4	91,3	92,1
19	122,4	57,0	54,8	19,0	43,5	4,2	100,2	103,7
20	110,2	50,0	47,4	23,1	52,8	4,4	110,0	105,2
21	129,3	55,7	52,8	20,3	52,8	4,2	116,6	112,4
22	126,1	45,4	41,2	20,3	52,2	4,3	91,6	90,5
23	120,7	46,9	44,2	20,6	48,2	4,0	84,3	85,7
24	111,9	49,8	45,9	19,4	55,5	4,2	107,2	103,0
25	100,0	36,9	34,2	21,3	39,8	4,5	61,1	63,1
26	116,4	29,9	26,7	20,8	41,7	4,2	56,8	46,8
27	130,7	34,9	32,5	20,0	40,5	4,1	54,1	55,2
28	117,1	45,1	42,7	20,8	44,0	4,6	86,1	86,8
29	110,4	34,6	33,0	20,1	53,2	4,6	80,3	80,6
30	123,4	39,9	38,2	21,9	53,6	4,6	94,1	91,5
31	133,5	57,2	52,8	19,7	43,6	4,4	100,4	102,8
32	120,4	58,5	56,1	21,0	50,3	4,2	118,6	115,5
33	116,2	30,2	27,8	20,9	41,4	5,0	57,0	60,1
34	126,1	49,0	43,9	20,2	48,3	4,3	91,0	90,5
35	130,1	53,5	51,2	21,9	52,8	4,5	121,7	114,1
36	126,3	49,8	46,7	20,0	42,3	4,2	82,0	85,6

Ghi chú: TS – tổng số, HH – hữu hiệu, M – khối lượng, NSTT- năng suất thực thu

Kết quả tính tốn nhờ phần mềm Excel 5.0 cho thấy:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.974
R Square	0.949
Adjusted R Square	0.938
Standard Error	4.794
Observations	36

ANOVA

	df	*SS*	*MS*	*F*	*Sig.F*
Regression	6	12412.126	2068.688	90.02	2.1E-17
Residual	29	666.403	22.979		
Total	35	13078.529			

	Coefficients	*St-dard Error*	*t Stat*	*P-value*
Intercept	-125.236	25.808	-4.85	3.82E-05
X Variable 1	-0.046	0.112	-0.41	6.82E-01
X Variable 2	0.130	0.936	0.14	8.91E-01
X Variable 3	1.822	0.943	1.93	6.32E-02
X Variable 4	0.244	0.865	0.28	7.80E-01
X Variable 5	1.354	0.116	11.67	1.78E-12
X Variable 6	15.129	2.943	5.14	1.72E-05

Phương trình hồi quy theo tính toán trên đây :

$$y = -125{,}24 - 0{,}05\ x_1 + 0{,}13\ x_2 + 1{,}82x_3 + 0{,}24\ x_4 + 1{,}35x_5 + 15{,}13x_6$$

Kết quả trắc nghiệm sự tồn tại của các hệ số hồi quy ở độ tin cậy 95% cho thấy, chiều cao cây (x_1), tổng số nhánh (x_2), chiều dài bông (x_4) không quan hệ với năng suất ($P > 0{,}68$ và $P > 0{,}89$). Các tính trạng số hạt/bông (x_5) và khối lượng 100 hạt (x_6) có quan hệ chắc chắn với năng suất với độ tin cậy của b_5 và b_6 đều lớn hơn 99%. Riêng số nhánh hữu hiệu (x_3) có thể có quan hệ với năng suất. Do sự có mặt của chiều cao cây, tổng số nhánh và chiều dài bông trong mô hình nghiên cứu có thể đã ảnh hưởng đến độ tin cậy của hệ số hồi quy b_3 (gần 94%).

Sau khi loại trừ ba tính trạng không quan hệ với năng suất, kết quả xử lý lại với các tính trạng số nhánh hữu hiệu (x_1), số hạt/bông (x_2) và khối lượng 100 hạt (x_3) cho thấy:

	Coefficients	*Std Error*	*t Stat*	*P-value*
Intercept	-126.90	15.495	-8.189	2.4E-09
X Variable 1	1.94	0.096	20.293	7.7E-20
X Variable 2	1.36	0.105	13.028	2.4E-14
X Variable 3	15.47	2.701	5.728	2.4E-06

Rõ ràng, cả ba tính trạng này đều quan hệ với năng suất. Phương trình hồi quy được xác định là:

$$y = -126{,}90 + 1{,}94x_1 + 1{,}36x_2 + 15{,}47x_3 \quad (4 - 5)$$

Thay các giá trị x_1, x_2, x_3 của ba tính trạng số nhánh hữu hiệu, số hạt/bông và khối lượng 100 hạt vào phương trình (4 - 5) ta được năng suất hồi quy ở bảng 4.1.

4.2.4. Vai trò của từng biến trong quan hệ đa biến

Phương trình hồi quy đa biến: $y = b_0 + b_1x_1 + ... + b_nx_n$ cho biết các biến X_1, X_2, ... X_n có quan hệ với Y. Do đơn vị tính của các biến khác nhau nên độ lớn nhỏ của hệ số hồi quy b_i không biểu thị vai trò của từng biến X_i đối với Y. Ví dụ, xét vai trò của biến x_1 (số nhánh hữu hiệu) và x_3 (khối lượng 100 hạt) với năng suất (y), ta thấy mặc dù hệ số hồi quy của x_1 nhỏ (1,94) nhưng giá trị của x_1 lớn (trên dưới 40), còn hệ số hồi quy của x_3 lớn (15,47) nhưng giá trị của x_3 nhỏ (khoảng 4g). Vì thế, không thể nói khối lượng 100 hạt có vai trò quan trọng hơn số nhánh hữu hiệu trong quan hệ với năng suất.

Để xác định vai trò của từng biến trong quan hệ đa biến ta sử dụng phép phân tích đường (Path Analysis).

Phép phân tích đường do S.Wright đề xuất năm 1921

là phương pháp phân tích mối quan hệ nhân quả giữa các đại lượng ngẫu nhiên. Theo đó, trong mối quan hệ của một cặp đại lượng có thể là kết quả ảnh hưởng của một số mối quan hệ thành phần với chiều hướng và mức độ khác nhau.

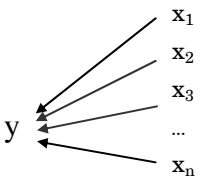
4.2.4.1. Hệ số đường

- ***Các mối quan hệ***

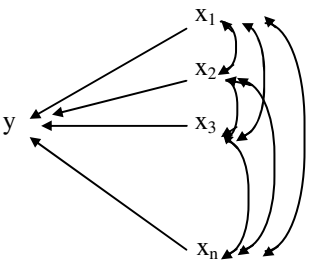
Quan hệ giữa hai hay nhiều đại lượng, nói chung có hai loại:

- Quan hệ nhân quả: một bên là nguyên nhân còn một bên là kết quả và không thể đảo ngược. Ví dụ, quan hệ giữa số quả, số hạt với năng suất.

- Quan hệ ngang bằng: có ảnh hưởng qua lại, không phân biệt nguyên nhân và kết quả. Ví dụ, chiều dài quả và to quả, số quả và khối lượng quả.



Hình 4.1: Quan hệ đồng thời x_i lên y



Hình 4.2: Quan hệ giữa nhiều đại lượng

- ***Khái niệm hệ số đường***

Dùng y, x_1, x_2 biểu thị giá trị thực của đại lượng kết quả và các đại lượng nguyên nhân. Giả sử y có quan hệ với x_1, x_2 theo phương trình:

$$y = b_0 + b_1x_1 + b_2x_2$$

b_1 và b_2 là hệ số hồi quy riêng của x_1, x_2. Hệ số b_1 phản ánh ảnh hưởng của x_1 tới y khi cố định x_2, tương tự, hệ số b_2 phản ánh ảnh hưởng của x_2 tới y khi cố định x_1. Như vậy khi cố định x_1, x_2 là đơn vị của y, còn khi cố định x_2, x_1 là đơn vị của y. Do x_1 và x_2 có đơn vị tính khác nhau và khác với y nên không thể dùng b_1 và b_2 để đánh giá vai trò của x_1 và x_2 đối với y.

Để đánh giá vai trò của x_1 và x_2 đối với y phải chuyển đổi đơn vị của y, x_1, x_2 thành số tương đối.

Nếu gọi $\overline{y}$, $\overline{x}_1$ và $\overline{x}_2$ là trị số trung bình và s_y, s_1 và s_2 là độ lệch chuẩn của y, x_1 và x_2, ta có:

$$(y - \overline{y}) = b_1(x_1 - \overline{x}_1) + b_2 (x_2 - \overline{x}_2)$$

Chia hai vế cho s_y và sự dung s_1, s_2 trong thuật toán:

$$\frac{(y-\overline{y})}{s_y} = b_1.\frac{s_1}{s_y}.\frac{(x_1-\overline{x}_1)}{s_1} + b_2.\frac{s_2}{s_y}.\frac{(x_2-\overline{x}_2)}{s_2}$$

Đặt $y' = \frac{y-\overline{y}}{s_y}$, $x_1' = \frac{x_1-\overline{x}_1}{s_1}$ và $x_2' = \frac{x_2-\overline{x}_2}{s_2}$ ta có:

$$y' = P_{y.1}.x_1' + P_{y.2}.x_2'$$

trong đó: $P_{y.1} = b_1.\frac{s_1}{s_y}$ và $P_{y.2} = b_2.\frac{s_2}{s_y}$. $P_{y.1}$ và $P_{y.2}$ được gọi là *hệ số đường* biểu thị ảnh hưởng của nguyên nhân x_1 và

x_2 đến kết quả y.

Một cách tổng quát, khi đại lượng y có *n* đại lượng nguyên nhân x_i (i = 1, 2, ...n) thì hệ số đường ảnh hưởng của x_i đến y là: $P_{y.i} = b_i.\frac{s_i}{s_y}$

Từ công thức $y' = P_{y.1}.x_1' + P_{y.2}.x_2'$ không khó để nhận ra hệ số đường $P_{y.i}$ là hệ số hồi quy riêng của x_i' đến y'. Đó cũng là số hồi quy riêng của x_i đến y sau khi đã chuyển hóa đơn vị.

4.2.4.2. Xác định hệ số đường

Mối quan hệ giữa các x_i với y và giữa các x_i với nhau tác động đến y là hết sức phức tạp. Trong tài liệu này ta chỉ xét vai trò của từng đại lượng x_i khi chúng tác động đồng thời đến y theo sơ đồ hình 4.1.

Theo tính chất 1 của hệ số đường, nếu quan hệ giữa các đại lượng x_1, x_2, ... x_n với y là quan hệ nhân quả theo hình 4.1 thì hệ số tương quan giữa đại lượng nguyên nhân x_i với đại lượng kết quả y sẽ là:

$$r_{yi} = \sum_{j=1}^{n} P_{y.j} r_{ij} \qquad (4 - 6)$$

Công thức này là phương trình thứ *i* trong hệ phương trình tuyến tính. Theo tính chất này, hệ số tương quan giữa x_i và y có thể chia thành *n* số hạng, mỗi số hạng là $P_{y.j} r_{ji}$.

Nếu đại lượng x_i và x_j độc lập, tức là $r_{ij} = 0$ thì công thức (4 - 6) trở thành:

$$r_{yi} = P_{y.j}$$

Để xác định các hệ số đường ($P_{y.x}$) biểu thị mức độ ảnh hưởng của các đại lượng nguyên nhân x_i đến đại lượng y phải giải hệ phương trình $r_{yx} = [A]P_{y.x}$, trong đó $[A]$ là ma trận của các hệ số tương quan giữa các đại lượng:

$$[A] = \begin{pmatrix} r_{11} & r_{12} & r_{13} & \dots & r_{1n} \\ r_{12} & r_{22} & r_{23} & \dots & r_{2n} \\ r_{13} & r_{32} & r_{33} & \dots & r_{3n} \\ & \dots & \dots & \dots & \dots \dots \\ r_{1n} & r_{2n} & r_{3n} & \dots & r_{nn} \end{pmatrix}$$

và r_{yx} là véctơ hệ số tương quan trực tiếp giữa các đại lượng x_1, x_2, ... với y:

$$r_{y.x} = \begin{pmatrix} r_{y1} \\ r_{y2} \\ r_{y3} \\ \dots \\ r_{yn} \end{pmatrix}$$

Nghiệm sẽ là: $P_{y.x} = [A]^{-1} r_{y.x}$, $[A]^{-1}$ là ma trận đảo của $[A]$. Đại lượng nào có $P_{y.x}$ càng cao, vai trò đối với y càng lớn.

4.2.4.3. Kiểm tra mức độ quan hệ của các đại lượng nguyên nhân với đại lượng kết quả

Để kiểm tra mức độ ảnh hưởng của các đại lượng nguyên nhân X_i lên đại lượng kết quả Y, ta xét giá trị Bx: $Bx = \sum r_{y.x} P_{y.x}$ hoặc giá trị R: $R = (1 - Bx)^{0,5}$.

Bx gọi là *hệ số xác định,* cho biết tỷ lệ (mức độ) biến động của các đại lượng X_i gây nên so với tổng số biến động của Y. Tỷ lệ này được tính bằng phần trăm (%) khi nhân

Bx với 100. Giá trị R cho biết mức độ của phần còn lại do các đại lượng khác chưa được nghiên cứu (nếu có) hoặc là sai số ngẫu nhiên.

Nếu $0,8 \leq Bx \leq 1$ thì các đại lượng ảnh hưởng đến Y, về cơ bản, đã được nghiên cứu đầy đủ, khi đó $R \leq 0,4$;

Nếu $Bx < 0,7$ thì còn có một hoặc một số đại lượng khác ảnh hưởng đến Y mà chưa được nghiên cứu, khi đó $R \geq 0,6$.

Ví dụ ứng dụng

Hãy tìm hiểu vai trò của sáu chỉ tiêu x_1, x_2, ..., x_6 đến năng suất lúa từ số liệu Bảng 4.1.

Xử lý số liệu trên Excel ta xác định được ma trận tương quan giữa các x_i $([A])$ và vector r_{yx} như sau :

Hệ số tương quan giữa các x_i ([A])							r_{yx}
	x_1	x_2	x_3	x_4	x_5	x_6	
x_1	1	0,252	0,244	0,001	-0,037	-0,108	0,134
x_2	0,252	1	0,993	-0,113	0,014	-0,060	0,802
x_3	0,244	0,993	1	-0,072	0,042	-0,068	0,822
x_4	0,001	-0,113	-0,072	1	0,185	0,123	0,091
x_5	-0,037	0,014	0,042	0,185	1	-0,468	0,503
x_6	-0,108	-0,060	-0,068	0,123	-0,468	1	-0,072

Xác định $[A]^{-1}$ bằng hàm MINVERSE trong Excel:

$$[A]^{-1} = \begin{pmatrix} 1,098 & -0,875 & 0,600 & -0,103 & 0,132 & 0,182 \\ -0,875 & 90,219 & -89,246 & 3,509 & 1,575 & -0,446 \\ 0,600 & -89,246 & 89,359 & -3,380 & -1,631 & 0,440 \\ -0,103 & 3,509 & -3,380 & 1,246 & -0,293 & -0,321 \\ 0,132 & 1,575 & -1,631 & -0,293 & 1,436 & 0,707 \\ 0,182 & -0,446 & 0,440 & -0,321 & 0,707 & 1,394 \end{pmatrix}$$

Nhân $[A]^{-1}$ với r_{yx} bằng hàm MMULT ta được vector các hệ số đường $P_{y.x}$:

$$\vec{P}_{y.x} = \begin{pmatrix} -0,018 \\ 0,055 \\ 0,765 \\ 0,013 \\ 0,586 \\ 0,254 \end{pmatrix}$$

Ta thấy, các hệ số đường của chiều cao cây (x_1), tổng số nhánh (x_2), chiều dài bông (x_4) là bằng 0 cho thấy các tính trạng này không quan hệ với năng suất. Trong các tính trạng còn lại, số nhánh hữu hiệu (x_3) có hệ số đường cao nhất, cho thấy số bông có vai trò quan trong nhất đến năng suất, kế đến là số hạt/bông (x_5) sau cùng là khối lượng 100 hạt (x_6).

Xét hệ số tương quan đơn giữa các tính trạng với năng suất (vector r_{yx}) thì chỉ có tổng số nhánh (x_2) và số nhánh hữu hiệu (x_3) có quan hệ chặt chẽ với năng suất, số hạt/ bông (x_5) quan hệ trung bình còn khối lượng hạt (x_6) không có quan hệ với năng suất. Điều đó cho thấy hệ số tương quan đơn phản ánh không đầy đủ và đúng đắn các mối quan hệ.

Để đánh giá mức độ quan hệ giữa ba tính trạng số nhánh hữu hiệu, số hạt/ bông và khối lượng hạt đến năng suất, ta tính Bx. Ở ví dụ này:

$$Bx = P_{y.x}\, r_{yx} = (- 0,018 \times 0,134) + (0,055 \times 0,802) + \ldots$$

$$+ (0,586 \times 0,503) + (0,254 \times - 0,072 = 0,95$$

Trong Excel, dùng hàm SUMPRODUCT để nhân vector $P_{y.x}$ với vector r_{yx}. Bx = 0,95 cho thấy trong nghiên cứu này ba tính trạng số nhánh hữu hiệu, số hạt/ bông và khối lượng hạt kiểm soát toàn bộ năng suất lúa.

4.3. QUAN HỆ PHI TUYẾN TÍNH

4.3.1. Tỷ số tương quan

Hệ số tương quan *r* biểu thị mức độ quan hệ tuyến tính giữa hai đại lượng ngẫu nhiên. Tuy nhiên, trong nhiều trường hợp, hệ số tương quan *r* rất nhỏ, thậm chí bằng 0, tức là không có quan hệ chặt chẽ hay độc lập với nhau theo quan hệ tuyến tính nhưng lại có quan hệ chặt chẽ theo một mối quan hệ khác – quan hệ phi tuyến tính.

Để biểu thị mức độ quan hệ khác giữa hai đại lượng Y theo X, người ta sử dụng tỷ số tương quan, ký hiệu là η_{yx}:

$$\eta_{yx} = \sqrt{1 - \frac{E\left[Y - E(Y_X)\right]^2}{DY}}$$

hay

$$= \sqrt{\frac{\sum(Y - \bar{Y})^2 - \sum(Y - \bar{Y}_X)^2}{\sum(Y - \bar{Y})^2}}$$

trong đó: $E(Y_X)$ hay $\bar{Y}_X$ là kỳ vọng của Y khi cố định một giá trị X, còn gọi là kỳ vọng của Y với điều kiện X.

Người ta đã chứng minh được:

$0 \le \eta_{yx} \le 1$ và $\rho \le \eta_{yx}$, ρ là hệ số tương quan lý luận và $0 \le r^2 \le \eta^2$.

Nếu r^2 là hệ số xác định biểu thị mức độ biến động của X trong quan hệ tuyến tính so với biến động của Y, thì

η^2 biểu thị mức độ biến động của X trong quan hệ phi tuyến tính so với biến động của Y.

$(\eta_{yx} - \rho)$ biểu thị mức độ quan hệ phi tuyến giữa Y và X. $(\eta_{yx} - \rho)$ càng lớn thì Y và X có tương quan phi tuyến càng chặt.

Để xác định tỷ tương quan ta thực hiện các bước sau đây:

1. Lập bảng số liệu:

Y \ X	x_1	x_2	...	x_k	
y_1	y_{11}	y_{12}	...	y_{1k}	
y_2	y_{21}	y_{22}	...	y_{2k}	
...	...	...	...	...	
y_n	$y_{n_1 1}$	$y_{n_1 2}$	...	$y_{n_1 k}$	
	n_1	n_2		n_k	$n = \Sigma n_i$
	T_1	T_2		T_k	$T = \Sigma T_i$

Trong bảng: $T_i = \sum_{i=1}^{n_i} y_{ij}$; n_i là số các số liệu các cột x_j

2. Tính tổng bình phương tổng số (SS_T):

$$SS_T = \sum_{1}^{n} y_{ij}^2 - \frac{T^2}{n}$$

3 . Tính tổng bình phương Y do sai khác giữa các x_j (SS_X):

$$SS_{Y_X} = \sum_{1}^{n} \frac{T_i^2}{n_i} - \frac{T^2}{n}$$

4. Tỷ số tương quan:

$$\eta_{yx} = \eta = \sqrt{\frac{SS_{Y_X}}{SS_T}}$$

4.3.2. Đánh giá sự tồn tại của tỷ số tương quan

Để kiểm định sự tồn tại của tỷ số tương quan ta sử dụng phép nghiệm F:

$$F_{TN} = \frac{(\eta^2 - r^2)(n-k)}{(1-\eta^2)(k-2)} \qquad (4-7)$$

Trong đó: n là tổng số số liệu

k là số nhóm của đại lượng X

F_{TN} được so sánh với so với $F_\alpha (k-2, n-k)$.

Nếu $F_{TN} \geq F_\alpha$ thì tồn tại quan hệ phi tuyến tính, ngược lại, $F_{TN} < F_\alpha$ thì tồn tại quan hệ tuyến tính.

Ví dụ ứng dụng

Kết quả theo dõi giữa thời gian sinh trưởng (TGST, ngày) và năng suất bông (g/cây) được ghi ở Bảng 4.3. Hãy đánh giá quan hệ giữa TGST với năng suất.

Bảng 4.3: TGST và năng suất (NS) cá thể

	Thời gian sinh trưởng, X (ngày)					
	98	99	100	101	102	103
NS cá thể, Y (g)	64,4 39,7 59,0	59,2 51,7 54,1 60,5	50,0 42,4 51,2	57,4 50,3	60,9 50,4 41,3 67,1 51,9 54,1	38,2 37,4 61,7 64,0
n_i	3	4	3	2	6	4
T_i	163,1	225,5	143,6	107,7	325,7	201,3

(tiếp)	Thời gian sinh trưởng, X (ngày)					
	104	105	106	107	108	109
NS cá thể, Y (g)	66,0 67,2 72,2 73,8	59,8 48,1	66,7 53,5 56,7	63,3 63,0 75,1	64,9 68,2	45,8 63,5 51,3
n_i	4	2	3	3	2	3
T_i	279,2	107,9	176,9	201,4	133,1	160,6
$\Sigma n_i = 39$; $\Sigma T_i = 2.226,0$						

Theo số liệu bảng 4.3:

1. Tổng bình phương tổng số:

$$SS_T = \sum_1^n y_{ij}^2 - \frac{T^2}{n} = [(64,4)^2 + (29,7)^2 +$$

$$... + (51,3)^2] - (2.226,0)^2/39 = 3.685,769$$

2 . Tính tổng bình phương Y do sai khác giữa các x_j (SS_X):

$$SS_{Y_X} = \sum_1^n \frac{T_i^2}{n_i} - \frac{T^2}{n} = [(163,1)^2/3 + (225,5)^2/4 +$$

$$... + (160,6)^2/3] - (2.226,0)^2/39 = 1.726,818$$

3. Tỷ số tương quan:

$$\eta_{yx} = \eta = \sqrt{\frac{SS_{Y_X}}{SS_T}} = \sqrt{\frac{1.726,818}{3.685,769}} = 0,684$$

4. Kiểm định η

Tính hệ số tương quan theo mục 4.2.2.1 ta được:

$r = 0,281$

Thay η và r vào công thức (4 – 7), ta có:

$$F_{TN} = \frac{(\eta^2 - r^2)(n-k)}{(1-\eta^2)(k-2)}$$

$$= \frac{\left[(0,684)^2 - (0,281)^2\right](39-12)}{\left[1-(0,684)^2\right](12-2)} = 1,98$$

$F_{TN} = 1,98 < F_{0,05}(10, 27) = 2,20$

Như vậy, với dung lượng mẫu n = 39, giữa thời gian sinh trưởng và năng suất bông chưa cho thấy có quan hệ phi tuyến tính.

Kiểm tra mức ý nghĩa của hệ số tương quan tuyến tính:

$$T = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}} = 1,78 < t_{0,05} = 2,03$$

Kết quả này cũng cho thấy chưa đủ cơ sở để kết luận về sự tồn tại mối quan hệ tuyến tính giữa thời gian sinh trưởng với năng suất.

Như vậy, để xác định đúng mối quan hệ này cần phải có dung lượng mẫu lớn hơn.

4.3.4. Chuyển hàm hồi quy phi tuyến tính về dạng tuyến tính

Để ước lượng các tham số của mô hình phi tuyến tính, một số hàm có thể chuyển thành dạng tuyến tính và thực hiện như mô hình tuyến tính, cuối cùng lại chuyển về dạng ban đầu.

1. Hàm lũy thừa

$$y = a x_1^{b_1} x_2^{b_2} ... x_n^{b_n}$$

Logarit hai vế:

$$\ln y = \ln a + b_1 \ln x_1 + b_2 \ln x_2 + ... + b_n \ln x_n$$

Đặt $\ln y = z$, $\ln a = b_0$, $\ln x_i = k_i$ và chuyển thành hàm tuyến tính:

$$z = b_0 + b_1 k_1 + b_2 k_2 + ... + b_n k_n$$

2. Hàm mũ

$$y = e^{b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n}$$

Logarit hai vế:

$$\ln y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n$$

Đặt $\ln y = z$ ta có:

$$z = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n$$

3. Hàm parabol

$$y = ax^2 + bx + c$$

Đặt $x^2 = k_1$, $x = k_2$ ta có:

$$y = ak_1 + bk_2 + c$$

4. Hàm lorarit

$$\ln y = b_0 + b_1 \ln x_1 + b_2 \ln x_2 + ... + b_n \ln x_n$$

Đặt $\ln y = z$, $\ln x_i = k_i$ ta có :

$$z = z = b_0 + b_1 k_1 + b_2 k_2 + ... + b_n k_n$$

4.4. QUAN HỆ GIỮA CÁC DẤU HIỆU ĐỊNH TÍNH

4.4.1. Hai dấu hiệu phân phối số liệu hai chiều

Dạng phân phối số liệu như sau:

A \ B	b_1	b_2	Tổng
a_1	n_1	n_2	$n_1 + n_2$
a_2	n_3	n_4	$n_3 + n_4$
Tổng	$n_1 + n_3$	$n_2 + n_4$	$n_1 + n_2 + n_3 + n_4$

Mối quan hệ giữa A và B được xác định bởi hệ số F:

$$F = \frac{n_1 n_4 - n_2 n_3}{\sqrt{(n_1 + n_2)(n_1 + n_3)(n_2 + n_4)(n_3 + n_4)}}$$

Ví dụ: Nghiên cứu mối quan hệ giữa độ lông của lá bông với mức độ kháng rầy xanh theo kết quả điều tra ghi ở Bảng 4.4 dưới đây.

Bảng 4.4: Kết quả điều tra mức độ kháng rầy trên các giống bông có mức độ lông của lá khác nhau

Đơn vị: số giống

Mức độ lông của lá (A)	Mức độ kháng rầy (B)		
	Kháng vừa	Kháng	Rất kháng
Ít	3	0	0
Vừa	7	3	0
Nhiều	5	8	5
Rất nhiều	11	5	3

Với số liệu trong bảng có thể nghiên cứu mối quan hệ giữa mức độ lông của lá bông với mức độ kháng rầy qua rất nhiều mối quan hệ. Chẳng hạn:

- Giữa lông ít và lông vừa với kháng vừa và kháng
- Giữa lông ít và lông nhiều với kháng vừa và kháng
- Giữa lông ít và lông rất nhiều với kháng vừa và kháng

- Giữa lông vừa và lông nhiều với kháng vừa và kháng

...

Giữa lông ít và lông nhiều với rất kháng và kháng vừa ta có bảng sau:

Mức độ lông của lá (A)	Mức độ kháng rầy (B)		
	Kháng vừa	Rất kháng	Tổng
Ít	3	0	3
Nhiều	5	5	10
Tổng	8	5	13

Ta có: $F = \frac{(3)(5)-(0)(5)}{\sqrt{(3)(8)(5)(10)}} = 0,43$

Giữa lông ít và lông rất nhiều với rất kháng và kháng vừa ta có bảng sau:

Mức độ lông của lá (A)	Mức độ kháng rầy (B)		
	Kháng vừa	Rất kháng	Tổng
Ít	3	0	3
Rất nhiều	3	11	14
Tổng	6	11	17

Ta có: $F = \frac{(3)(11)-(0)(3)}{\sqrt{(3)(6)(11)(14)}} = 0,63$

Kết quả cho thấy, lá càng nhiều lông thì càng kháng rầy.

4.4.2. Tương quan theo thứ hạng

Kết quả theo dõi lượng bón phân tổng hợp (tạ/ha) và năng suất lúa (tấn/ha) như sau:

Lượng bón	0,8	0,9	1,0	1,1	1,2	1,3	1,4	1,5	1,6
Năng suất	5,0	5,1	5,3	5,4	5,4	5,5	5,6	5,6	5,7

Với cách tính hệ số tương quan theo dấu hiệu định lượng, từ số liệu ở bảng ta được: $r = 0,97$.

Để tính hệ số tương quan theo thứ hạng, năng suất được xếp hạng theo thứ tự hạng lượng phân bón từ nhỏ đến lớn như sau:

Lượng bón	1	2	3	4	5	6	7	8	9
Năng suất	1	2	3	4	4	6	7	7	9

Hệ số tương quan thứ hạng được tính theo công thức của Spearman (1904):

$$r_s = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)}$$

trong đó: d_i là hiệu số của hai thứ hạng theo cặp, n là số cặp hạng. Ví dụ: $d_1 = 1 - 1 = 0$, $d_5 = 5 - 4 = 1$.

Từ số liệu phân hạng ta có:

$$r_s = 1 - \frac{6\left[(0)^2 + (0)^2 + \ldots (1)^2 + (0)^2\right]}{9\left[(9)^2 - 1\right]} = 0,98$$

Hai kết quả khác nhau không nhiều và đều được chấp nhận.