Hạnh Nguyên

TRƯỜNG ĐẠI HỌC CẦN THƠ KHOA NÔNG NGHIỆP & SHUD

GIÁO TRÌNH

THỐNG KÊ – PHÉP THÍ NGHIỆM

Biên soạn: PHAN THỊ THANH THỦY

LỜI TỰA

Mã số môn học: TT315 - 2 tín chỉ

Trong thời gian qua, sách và tài liệu tham khảo về môn Thống kê - Phép thí nghiệm, đặc biệt là các kiểu bố trí thí nghiệm sử dụng trong nghiên cứu nông nghiệp, còn hạn chế.

Nhằm đáp ứng nhu cầu học tập và làm luận văn tốt nghiệp cho sinh viên ngành Nông nghiệp, đồng thời phục vụ cho công tác nghiên cứu, chúng tôi đã biên soạn giáo trình lý thuyết về môn học này.

Đây là giáo trình tóm tắt một số phương pháp thống kê và các kiểu bố trí thí nghiệm cơ bản được sử dụng phổ biến trong nghiên cứu nông nghiệp.

Mục đích của môn học

Thống kê là công cụ không thể thiếu được trong nghiên cứu khoa học. Một phương pháp thí nghiệm đúng, phân tích hợp lý sẽ dẫn đến kết luận đúng đắn, đáng tin cậy và có sức thuyết phục cao.

Trong nông nghiệp, có các kỹ thuật thống kê khác nhau dành cho nghiên cứu cây trồng và vật nuôi, cho những thí nghiệm ngoài đồng và trong phòng, cho nghiên cứu di truyền và sinh lý,...

Chương trình môn học Thống kê - Phép thí nghiệm được xây dựng để giảng dạy cho sinh viên ngành nông nghiệp. Học phần này trang bị cho sinh viên kỹ thuật ứng dụng thống kê trong nghiên cứu, từ đó giúp sinh viên có khả năng thực hiện thí nghiệm một cách độc lập, biết cách xử lý và phân tích số liệu, giải thích kết quả một cách hợp lý và có khoa học.

Nội dung của môn học

Chương 1 giới thiệu thống kê mô tả. Ôn lại một vài khái niệm về thống kê và các số đo mô tả cho tổng thể và mẫu như các số đo trung tâm, số đo phân tán và số đo bất đối xứng và độ bẹt của đường cong phân bố xác suất.

Chương 2 trình bày các dạng phân bố của biến ngẫu nhiên rời rạc và liên tục: Phân bố xác suất, phân bố nhị thức, phân bố Poisson, phân bố chuẩn, phân bố t của Student, phân bố chi bình phương (χ^2) và phân bố F của Fisher.

Chương 3 giải thích ý nghĩa của kiểm định giả thuyết và các loại sai lầm có liên quan đến việc rút ra kết luận. Chương này cũng cung cấp các phương pháp kiểm định giả thuyết và tìm khoảng tin cậy cho một trung bình hoặc một tỷ lệ của tổng thể và cách xác định cỡ mẫu tối thiểu để ước lượng một trung bình hoặc một tỷ lệ.

Chương 4 bao gồm các phương pháp kiểm định giả thuyết cho hai trung bình độc lập hoặc bắt cặp và phương pháp so sánh hai tỷ lệ. Cách ước lượng khoảng tin cậy cho sai biệt hai trung bình hoặc hai tỷ lệ. Cách xác định cỡ mẫu tối thiểu để so sánh hai trung bình hoặc hai tỷ lệ.

Chương 5 - 7 bao gồm các kiểu bố trí thí nghiệm được sử dụng phổ biến như thí nghiệm một nhân tố, hai nhân tố, ba hay nhiều nhân tố. Với mỗi kiểu bố trí mô tả chi tiết cách làm ngẫu nhiên và phương pháp phân tích phương sai cũng như cách ước lượng số liệu thiếu trong thí nghiệm.

Chương 8 đưa ra một số phương pháp so sánh các trung bình nghiệm thức riêng biệt: LSD, Dunnet, Duncan, Newman & Keuls cho so sánh cặp và phương pháp tương phản một độ tự do *(single d.f. contrast)* cho so sánh nhóm.

Ngoài kỹ thuật phân tích phương sai, chương 9 - 10 còn đưa ra hai phương pháp thống kê được sử dụng phổ biến nhất cho phân tích số liệu trong nghiên cứu nông nghiệp. Đó là phân tích hồi qui và tương quan và chi bình phương.

Chương 11 trình bày cách cải thiện các phương pháp đã được mô tả từ chương 5 đến 7 cần thiết để vận dụng trong các trường hợp đặc biệt sau:

- Thí nghiệm được lặp lại qua nhiều thời gian hoặc địa điểm
- Số liệu trong thí nghiệm vi phạm một hoặc nhiều giả định của phân tích phương sai.

Chương 12 cung cấp những hướng dẫn để trình bày các kết quả nghiên cứu. Hầu hết các hướng dẫn này đều được tham khảo từ quyển sách "Statistical Procedures For Agricultural Research" của Gomez and Gomez, 1984.

Chúng tôi chân thành cám ơn Ts. Dương Ngọc Thành đã có những góp ý về chuyên môn và các đồng nghiệp trong bộ môn Di truyền - Giống Nông nghiệp đã giúp đỡ và tạo điều kiện thuận lợi cho chúng tôi hoàn thành giáo trình này.

Phan Thị Thanh Thủy

Bộ môn Di Truyền Giống Nông Nghiệp
Khoa Nông Nghiệp & Sinh Học Ứng Dụng
Trường Đại Học Cần Thơ.
2007

MỤC LỤC

Chương 1: MỘT VÀI KHÁI NIỆM THỐNG KÊ VÀ CÁC SỐ ĐO MÔ TẢ 1

I. Tổng thể và mẫu, 1

1. Tổng thể, 1
2. Mẫu, 1
3. Biến số ngẫu nhiên, 2

II. Sắp xếp số liệu, 2

1. Số liệu, 2
2. Số liệu gom nhóm, 2

III. Các số đo mô tả, 4

1. Các số đo trung tâm, 4
2. Các số đo phân tán, 10

Chương 2. CÁC DẠNG PHÂN BỐ CỦA BIẾN NGẪU NHIÊN 15

I. Phân bố xác suất, 15

1. Một vài khái niệm về lý thuyết xác suất, 15
2. Phân bố xác suất, 17

II. Phân bố nhị thức, 18

III. Phân bố chuẩn, 20

1. Hàm mật độ của phân bố chuẩn, 21
2. Tính chất của phân bố chuẩn, 21
3. Phát họa đường cong chuẩn, 22
4. Tính tung độ của đường cong chuẩn tắc, 22
5. Tìm diện tích dưới đường cong chuẩn tắc, 23
6. Xác định giá trị của biến chuẩn tắc z, 25
7. Tìm diện tích dưới đường cong chuẩn, 25
8. Xác định giá trị của biến chuẩn X, 26

IV. Phân bố Poisson, 26

V. Phân bố của các số trung bình lấy mẫu, 28

1. Trung bình, phương sai và độ lệch chuẩn của biến ngẫu nhiên $\overline{X}$, 28
2. Phân bố của biến $\overline{X}$ được rút ra từ tổng thể có phân bố chuẩn, 29
3. Phân bố của biến $\overline{X}$ được rút ra từ tổng thể có phân bố bất kỳ, 29
4. Biến chuẩn tắc hóa của biến ngẫu nhiên $\overline{X}$, 29

VI. Phân bố của các số tỷ lệ lấy mẫu, 30

VII. Phân bố t của Student, 32

1. Hàm mật độ của phân bố t, 32
2. Sử dụng bảng t, 33
3. Tìm diện tích dưới đường cong t, 33

VIII. Phân bố chi bình phương (χ^2), 33

1. Hàm mật độ của phân bố χ^2, 33
2. Sử dụng bảng χ^2, 34
3. Tìm giá trị χ^2 cho một diện tích đã định, 34

IX. Phân bố F của Fisher, 36

1. Hàm mật độ của phân bố F, 36
2. Cách sử dụng bảng F, 36

3 Tìm diện tích dưới đường cong F, 37

Chương 3: KIỂM ĐỊNH GIẢ THIẾT CHO MỘT TRUNG BÌNH HOẶC MỘT TỶ LỆ 38

I. Kiểm định giả thuyết, 38
1. Ý nghĩa của kiểm định giả thuyết, 38
2. Sai lầm loại I và loại II, 39
3. Các thuật ngữ được dùng trong kiểm định giả thuyết, 42
4. Các kết luận có thể xảy ra cho một kiểm định giả thuyết, 43
5. Các bước tiến hành cho một kiểm định giả thuyết thống kê, 43

II. Kiểm định giả thuyết cho giá trị trung bình của tổng thể, 43
1. Trường hợp phương sai đã biết hoặc phương sai chưa biết, mẫu lớn, 43
2. Trường hợp phương sai chưa biết, mẫu nhỏ, 45

III. Kiểm định giả thuyết cho giá trị tỷ lệ, 47

IV. Tìm khoảng tin cậy cho trung bình hoặc tỷ lệ, 48
1. Khoảng tin cậy của trung bình tổng thể (μ), 48
2. Khoảng tin cậy cho tỷ lệ (p), 49
3. Kiểm định ý nghĩa đối với khoảng tin cậy, 49

V. Xác định cỡ mẫu, 50
1. Cỡ mẫu cần lấy để ước lượng trung bình (μ), 51
2. Cỡ mẫu cần lấy để ước lượng tỷ lệ (p), 53

Chương 4. SO SÁNH HAI TRUNG BÌNH HOẶC HAI TỈ LỆ 54

I. Phân bố của sự sai biệt giữa hai trung bình (mẫu lớn và độc lập), 54

II. Phân bố của sự sai biệt giữa hai tỷ lệ, 55

III. So sánh hai giá trị trung bình độc lập, 55
1. Hai phương sai đã biết hoặc hai phương sai không biết, mẫu lớn, 55
2. Hai phương sai không biết, mẫu nhỏ, 56

IV. So sánh hai trung bình bắt cặp, 61

V. So sánh hai tỷ lệ, 62

VI. Khoảng tin cậy của sai biệt giữa hai trung bình tập hợp, 63
1. Trường hợp mẫu lớn và độc lập, 63
2. Trường hợp mẫu nhỏ và độc lập, 63
3. Trường hợp hai mẫu bắt cặp, 65

VII. Khoảng tin cậy của sai biệt hai tỷ lệ, 65

VIII. Xác định cỡ mẫu, 66
1. Xác định cỡ mẫu trong trường hợp hai trung bình độc lập, 66
2. Xác định cỡ mẫu trong trường hợp so sánh cặp, 67

Chương 5. THÍ NGHIỆM MỘT NHÂN TỐ 69

A. Một số thuật ngữ dùng trong bố trí thí nghiệm, 69
1. Đơn vị thí nghiệm, 69
2. Nhân tố, 69
3. Nghiệm thức, 69
4. Sai số thí nghiệm, 69
5. Lặp lại, 69

B. Phân tích phương sai, 70
1. Các tiền đề của phân tích phương sai, 70
2. Phân tích phương sai, 70

C. Thí nghiệm một nhân tố, 71
I. Bố trí hoàn toàn ngẫu nhiên, 72
1. Cách bố trí và làm ngẫu nhiên, 72
2. Phân tích phương sai, 74
II. Bố trí khối hoàn toàn ngẫu nhiên, 77
1. Kỹ thuật phân khối, 78
2. Cách bố trí và làm ngẫu nhiên, 79
3. Phân tích phương sai, 80
4. Cách tính số liệu thiếu, 82
III. Bố trí hình vuông latin, 86
1. Cách bố trí và làm ngẫu nhiên, 86
2. Phân tích phương sai, 87
3. Cách tính số liệu thiếu, 90

Chương 6. THÍ NGHIỆM HAI NHÂN TỐ 92
A. Các loại ảnh hưởng của hai nhân tố, 92
1. Ảnh hưởng đơn, 93
2. Ảnh hưởng chính, 93
3. Ảnh hưởng tương tác, 93
B. Thí nghiệm thừa số, 95
I. Bố trí hoàn toàn ngẫu nhiên. 96
II. Bố trí khối hoàn toàn ngẫu nhiên, 99
III. Bố trí lô phụ, 102
1. Cách bố trí và làm ngẫu nhiên, 103
2. Phân tích phương sai, 104
3. Cách tính số liệu thiếu, 107
IV. Bố trí dải phụ, 108
1. Cách bố trí và làm ngẫu nhiên, 109
2. Phân tích phương sai, 109

Chương 7. THÍ NGHIỆM NHIỀU NHÂN TỐ 115
I. Tương tác giữa ba hoặc nhiều nhân tố, 115
II. Các kiểu bố trí thí nghiệm có thể lựa chọn, 117
1. Bố trí thí nghiệm một nhân tố, 117
2. Bố trí thí nghiệm hai nhân tố, 117
3. Bố trí thí nghiệm ba hoặc nhiều nhân tố, 121
4. Bố trí thừa số phân nhỏ, 121
III. Bố trí khối hoàn toàn ngẫu nhiên, 122
IV. Bố trí lô phụ trong lô phụ, 126
1. Cách bố trí và làm ngẫu nhiên, 126
2. Phân tích phương sai, 127

Chương 8 SO SÁNH CÁC TRUNG BÌNH NGHIỆM THỨC 134
I. Kiểm định khác biệt nhỏ nhất có ý nghĩa (LSD), 135
1. Thí nghiệm một nhân tố, 136
2. Thí nghiệm thừa số hai nhân tố, 142
II. Kiểm định Dunnet, 147
III. Kiểm định Duncan, 149

1. Thí nghiệm một nhân tố, 149
2. Thí nghiệm hai nhân tố, 152
IV. Kiểm định Newman và Keuls, 152
V. Phương pháp tương phản *(Contrast Method)*, 152
1. Các nguyên tắc so sánh trực giao, 153
2. Sai số chuẩn của một contrast, 153
3. Tổng bình phương của một contrast, 153

Chương 9 PHÂN TÍCH HỒI QUI VÀ TƯƠNG QUAN 155
A. QUAN HỆ TUYẾN TÍNH, 155
I. Hồi qui tuyến tính đơn, 157
1. Ước lượng α và β, 158
2. Tìm đường hồi qui ước lượng, 160
II. Tương quan tuyến tính đơn, 163
III. Tính đồng nhất của các hệ số hồi qui, 166
1. Hai hệ số hồi qui, 167
2. Nhiều hơn hai hệ số hồi qui, 169
IV. Tính đồng nhất của hệ số tương quan, 171
V. Tương quan và hồi qui tuyến tính bội, 173
B. QUAN HỆ PHI TUYẾN, 177
I. Hồi qui phi tuyến đơn, 178
1. Chuyển đổi biến, 178
2. Tạo biến mới, 180
II. Hồi qui phi tuyến bội, 183

Chương 10. KIỂM ĐỊNH CHI BÌNH PHƯƠNG 184
I. Phân tích số liệu thuộc tính, 184
1. Kiểm định cho giả thiết có một tỉ lệ cố định, 184
2. Kiểm định cho tính độc lập trong bảng liên hợp, 186
3. Kiểm định tính đồng nhất của tỷ lệ, 190
II. Kiểm định cho tính đồng nhất của phương sai, 192
1. Độ tự do bằng nhau, 192
2. Độ tự do không bằng nhau, 193
III. Kiểm định cho độ lệch chuẩn của tổng thể, 194
1. Kiểm định cho độ lệch chuẩn của tổng thể, 194
2. Khoảng tin cậy cho độ lệch chuẩn của tổng thể, 196
IV. Kiểm định cho tính phù hợp, 196

Chương 11. PHÂN TÍCH SỐ LIỆU TỪ MỘT LOẠT THÍ NGHIỆM VÀ BIẾN ĐỔI SỐ LIỆU 201
I. Phân tích số liệu từ một loạt thí nghiệm, 201
1. Phân tích qua nhiều mùa vụ, 201
2. Phân tích qua nhiều năm, 209
3. Phân tích thí nghiệm qua nhiều địa điểm, 213
II. Biến đổi số liệu, 217
1. Đổi sang logarith thập phân, 217
2. Đổi sang căn bậc hai, 220
3. Đổi sang arcsin, 221

Chương 12. TRÌNH BÀY CÁC KẾT QUẢ NGHIÊN CỨU

I.Thí nghiệm một nhân tố, 226

1. Các nghiệm thức riêng biệt, 226
2. Nghiệm thức định lượng: Đồ thị đường, 233

II. Thí nghiệm thừa số, 234

1. Dạng bảng, 236
2. Đồ biểu hình thanh, 239
3. Đồ thị đường, 241

Phụ lục

I. Tung độ của đường cong chuẩn tắc, 226
II. Bảng phân bố chuẩn tắc z, 227
III. Bảng giá trị $e^{-\lambda}$, 228
IV. Bảng phân bố t, 229
V. Bảng phân bố chi bình phương, 230
VI. Bảng phân bố F, 231
VII. Bảng số ngẫu nhiên, 235
VIII. Các bố trí hình vuông latin mẫu, 236
IX. Bảng phân bố Dunnet, 237
X. Bảng DUNCAN, 238
XI. Bảng NEWMAN và KEULS, 240
XII. Bảng hệ số tương quan tuyến tính, 242
XIII. Các giá trị tương ứng của r và z, 243
XIV. Bảng đổi sang Arcsin √phần trăm , 244

TÀI LIỆU THAM KHẢO

Cochran W. G. and G. M. Cox, 1957. Experimental designs. Second edition. New York: Wiley. 611 p.

Collin Jean. 2003. Dispositifs expérimentaux. Université LAVAL. 304 p.

Chu Văn Mẫn. 2001. Ứng dụng tin học trong sinh học. Nhà xuất bản Đại học Quốc gia Hà nội. 262 p.

Dagnélie Pierre. 1984. Théorie et méthodes statistiques. Applications agronomiques. Vol. II. Les Presses Agronomiques de Gembloux.

Diệp Ngọc Liên và Hồ Thị Nguyên. 1992. Giáo trình bố trí thí nghiệm.

Đặng Hùng Thắng. 1999. Thống kê và ứng dụng. Nhà xuất bản giáo dục. 267 p.

Gomez Kwanchai A. & Arturo A. Gomez. 1984. Statistical Procedures for Agricultural Research. Second edition. New York: John Wiley & Sons. 680 p.

Hebert Yves. 1968. Mathématiques Probabilités et Statistique. Cours et Exercices C.B., B.R., Pharmacie, C.P.E.M. VUIBERT.

Hervé Y. 1992. Cours Expérimentation Agronomique - Technique des Essais Agricoles. Ecole Nationale Supéricure agronomique de Rennes. 105 p.

Hoel Paul G. 1971. Introduction to mathematical Statistics. John Wiley & Sons, Inc. New York.

LeCLERG E. L., Leonard W. H., Clark A. G. Field plot technique, second edition.

Mendenhall William and Richard Scheaffer. 1973. Mathematical Statistics With Applications. JMC Press, Incorporated. Philippines.

Montgomery Douglas C. 2001. Design and analysis of experiments.5th edition. New York: John Wiley & Sons. 682 p

Murray R. Spiegel. 1992. Statistics. Second edition. McGraw Hill, Inc.

Neil A. Weiss. 1993. Elementary Statistics. Addison-Wesley, Inc. USA.

Ngô Như Hoa. 1982. Thống kê trong nghiên cứu y học. Tập II. Nhà xuất bản y học.

Nguyễn Đình Huy. 2003. Giáo trình xác suất và thống kê. Nhà xuất bản Đại học Quốc gia TP. Hồ Chí Minh. 137 p

Philippeau G. 1973. Théorie des plans d'expériences - Application à l'agronomie. P.I.T.C.F.

Snedecor, G. W. and W. G. Cochran. 1989. Statistical methods. 8th edition. Ames, Iowa: Iowa State University Press. 503 p.

Steel, R. D. G. and J. H. Torrie. 1980. Principles and procedures of statistics – A biometrical approach. 2nd ed. New York: McGraw-Hill. 633 p.

THỐNG KÊ
PHÉP THÍ NGHIỆM

CHƯƠNG 1

MỘT VÀI KHÁI NIỆM THỐNG KÊ VÀ CÁC SỐ ĐO MÔ TẢ

Thống kê là một ngành khoa học nghiên cứu về các phương pháp thu thập, phân loại, tóm tắt và phân tích các số liệu cũng như rút ra những kết luận khoa học từ các số liệu như thế.

Có hai loại thống kê chính: *thống kê mô tả (descriptive statistics)* và *thống kê suy luận (inferential statistics)*.

* *Thống kê mô tả* bao gồm các phương pháp sắp xếp và tóm tắt thông tin một cách rõ ràng và hiệu quả, chẳng hạn như lập bảng biểu, vẽ đồ biểu và tính toán các số đo mô tả như số trung bình, số phần trăm, số đo phân tán,...

* *Thông kê suy luận* bao gồm các phương pháp lấy mẫu, phân tích và đánh giá độ tin cậy của kết luận về tổng thể dựa trên những thông tin có được từ mẫu.

I. TỔNG THỂ VÀ MẪU

1. Tổng thể = Quần thể *(Population)* là tập hợp các phần tử thuộc đối tượng nghiên cứu.

Mỗi phần tử của tổng thể được gọi là cá thể. Số cá thể của tổng thể được ký hiệu là N. Một biến X là phép đo xác định trên mỗi cá thể của tổng thể.

Tổng thể có thể hữu hạn *(finite)* hoặc vô hạn *(infinite)*. Ví dụ, tập hợp chiều cao của tất cả học sinh trong một lớp là tổng thể hữu hạn, trong khi tập hợp chiều cao lúc chín của tất cả các cá thể ở giống lúa IR-64 trong khu ruộng đang khảo sát là tổng thể vô hạn.

Nhìn chung, hầu hết các nghiên cứu trong lĩnh vực sinh-nông học, tổng thể được giả thuyết lớn vô hạn và nó không phải là một khái niệm dễ định nghĩa theo quan điểm thống kê. Tuy nhiên, một vài ví dụ sau đây sẽ giúp cho định nghĩa được rõ ràng hơn. Giả sử chúng ta đo chiều cao lúc chín của một giống đậu, lúc đó tổng thể là một số lượng lớn vô hạn các số đo chiều cao cây của giống đậu đó; hoặc chúng ta đang theo dõi ngày trổ hoa ở một giống bắp, lúc đó tổng thể sẽ bao gồm một số lượng lớn vô hạn các cây bắp của giống bắp đó mà trên mỗi cây chúng ta ghi nhận ngày trổ hoa. Dĩ nhiên trong một thí nghiệm thực sự chúng ta không thể hy vọng lấy số lượng lớn vô hạn các số đo của tính trạng được quan tâm mà chúng ta chỉ có thể lấy một số giới hạn các số đo như thế do những hạn chế về thời gian, công sức, chi phí,... Vì không thể quan sát mọi cá thể của tổng thể mà chỉ có thể quan sát một số cá thể giới hạn được rút ra từ tổng thể đó nên chúng ta có khái niệm về mẫu.

2. Mẫu *(Sample) bao gồm một số phần tử được rút ra từ tổng thể.*

Số phần tử chứa trong một mẫu gọi là cỡ mẫu hay kích thước mẫu *(sample size)*, ký hiệu là n.

Thường các thí nghiệm được tiến hành trên mẫu. Tuy nhiên, kết quả thu được qua xử lý thống kê sẽ được suy rộng ra cho cả tổng thể. Sự suy luận như thế không thể chắc chắn một cách tuyệt đối nên từ xác suất *(probability)* thường được dùng trong phát biểu các kết luận.

Nhìn chung, do mẫu có kích thước hữu hạn, nên sẽ có một sai số, đó là sai số do lấy mẫu. Vì vậy, muốn cho kết luận của mẫu đại diện cho tổng thể, mẫu cần phải được thu thập một cách ngẫu nhiên và đủ lớn.

3. Biến số ngẫu nhiên = biến ngẫu nhiên *(Random Variable)* là một đại lượng bằng số mà giá trị của nó phụ thuộc vào cách lấy ngẫu nhiên.

Ví dụ: Xét một tổng thể và với mỗi phần tử của tổng thể, chúng ta quan sát (đếm, đo, cân,...) một đặc tính nào đó, được ký hiệu là biến X. Biến X là một đại lượng biến thiên từ phần tử này sang phần tử khác; do đó, được gọi là biến ngẫu nhiên.

Gọi X_i là giá trị của biến X trong lần quan sát thứ i. Nếu có n biến quan sát, giá trị của các biến lần lượt là $X_1, X_2, \cdots, X_n$. Có hai loại biến ngẫu nhiên:

a) ***Biến ngẫu nhiên rời rạc*** *(discrete random variable)* là biến ngẫu nhiên có giá trị là những con số nguyên dương thông qua động tác đếm, như số lứa đẻ của một con heo, số con trong một gia đình, số gié trên một bụi lúa, số cành, số trái trên cây,...

b) ***Biến ngẫu nhiên liên tục*** *(continuous random variable)* là biến ngẫu nhiên mà các giá trị của nó có được thông qua động tác đo lường hay tính toán như năng suất, chiều cao cây, tỉ lệ hạt chắc, lép,... Biến liên tục có thể nhận giá trị bất kỳ trong một khoảng các số thực. Các giá trị này lập thành những khoảng liên tục trên trục số. Ví dụ, chiều cao của một nam sinh viên đo được là 165 cm; thật ra, người đó cao trong khoảng 160 - 170 cm, nếu sai số của dụng cụ đo là 5 cm.

Số liệu có thể được mô tả bằng một biến rời rạc hoặc liên tục được gọi là số liệu rời hoặc liên tục. Chẳng hạn, số con của một trong số 1000 gia đình là số liệu rời, trong khi chiều cao của một trong số 100 sinh viên của một trường đại học là số liệu liên tục.

II. SẮP XẾP SỐ LIỆU

1. Số liệu *(data)*. Số liệu hay dữ liệu là các thông tin được thu thập, sắp xếp và phân tích. Có nhiều loại số liệu khác nhau:

a) ***Số liệu định tính = Số liệu thuộc tính*** *(qualitative data)* là số liệu đối với thông tin không phải bằng số như giống, màu mắt, màu hoa, loại máu,...

b) ***Số liệu đo lường*** *(metric data)* hoặc ***số liệu định lượng*** *(quantitative data)* là số liệu có được thông qua việc đo lường hay tính toán như chiều cao, trọng lượng, năng suất,...

c) ***Số liệu tần số*** *(frequency data)* là số liệu dựa trên số lượng cá thể quan sát trong các nhóm khác nhau.

2. Số liệu gom nhóm *(grouping data)*.

Việc gom nhóm các số liệu nhằm làm cho bộ số liệu phức tạp trở nên dễ hiểu hơn.

a) Các thuật ngữ được sử dụng trong số liệu gom nhóm

• ***Nhóm = Lớp*** *(classes)* còn được gọi là khoảng của nhóm *(class interval)* bao gồm các giá trị nằm trong khoảng giới hạn dưới và giới hạn trên của một nhóm.

• ***Tần số*** *(frequency)* là số lần xuất hiện của một giá trị (hoặc một nhóm) nào đó trong bộ số liệu.

• ***Tần suất*** *(relative frequency = F)* là tỷ số giữa tần số và cỡ mẫu (n).

• ***Phân bố tần số*** *(frequency distribution)* hoặc bảng tần số là sự liệt kê các giá trị (hoặc các nhóm) và tần số (f) của chúng.

• ***Phân bố thực nghiệm*** *(relative-frequency distribution)* là sự liệt kê các giá trị (hoặc các nhóm), tần số (f) và tần suất (F = f / n) của chúng.

Ví dụ: Thu thập số liệu về số con đến tuổi đi học trong mỗi gia đình ở một ấp. Chọn ngẫu nhiên 30 gia đình. Bảng 1.1 ghi nhận số con đến tuổi đi học của một trong 30 gia đình.

Bảng 1.1: Số con đến tuổi đi học của 30 gia đình.

0	3	0	0	3	0
2	2	0	1	2	1
0	0	1	2	4	0
4	2	1	0	1	0
0	2	0	1	3	2

Từ số liệu của Bảng 1.1, chúng ta lập được bảng phân bố thực nghiệm như Bảng 1.2.

Bảng 1.2: Phân bố thực nghiệm về số con đến tuổi đi học của 30 gia đình

Số con đến tuổi đi học	Tần số	Tần suất
0	12	0,400
1	6	0,200
2	7	0,233
3	3	0,100
4	2	0,067
Tổng:	30	1,000

• ***Giới hạn dưới của nhóm*** *(lower class limit)* là giá trị nhỏ nhất của một nhóm.

• ***Giới hạn trên của nhóm*** *(upper class limit)* là giá trị lớn nhất của một nhóm.

• ***Điểm giữa của nhóm*** *(mark class)* là giá trị ở giữa của một nhóm. Đó là giá trị trung bình cộng của giới hạn trên và giới hạn dưới của một nhóm.

• ***Độ rộng của nhóm*** *(class width)* là hiệu số giữa giá trị giới hạn dưới của nhóm đang ghi nhận và giá trị giới hạn dưới của nhóm kế tiếp.

• ***Khoảng*** *(Range)* là hiệu số giữa hai giá trị quan sát: lớn nhất và nhỏ nhất ($R = X_{max} - X_{min}$).

b) Cách trình bày số liệu dưới bảng phân bố thực nghiệm

Ví dụ: Bảng 1.3 trình bày phân bố thực nghiệm trọng lượng (kg) của 100 nam sinh viên ở một trường Đại học.

Bảng 1.3: Trọng lượng (kg) của 100 nam sinh viên ở một trường Đại học.

Trọng lượng	Số sinh viên (Tần số)	Tần suất	Tỷ lệ (%)
40 - 42	5	0,05	5
43 - 45	19	0,19	19
46 - 48	42	0,42	42
49 - 51	26	0,26	26
52 – 54	8	0,08	8
Tổng	100	1,00	100

Nhóm đầu tiên trong bảng bao gồm trọng lượng từ 40 – 42 kg và được ghi bằng ký hiệu khoảng của nhóm '40 – 42'. Tất cả các nhóm đều bằng nhau nên độ rộng của

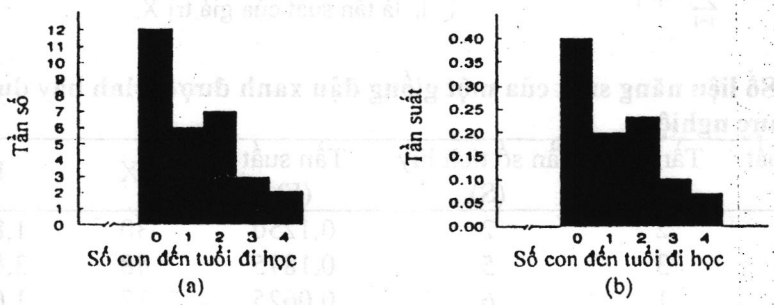
nhóm là 3 (43 - 40). Vì năm sinh viên có trọng lượng thuộc nhóm này nên tần số tương ứng của nhóm bằng 5.

Số liệu được sắp xếp và tóm tắt như trong bảng phân bố thực nghiệm được gọi là số liệu gom nhóm. Nhìn chung, mặc dầu tiến trình gom nhóm đã phá hủy nhiều chi tiết ban đầu (số liệu gốc), nhưng ưu điểm chính là cách trình bày ngắn gọn và rõ ràng.

c) Biểu đồ (histogram) hình thanh của số liệu tần số và tần suất

• ***Biểu đồ tần số*** là đồ thị trình bày các nhóm trên trục hoành và tần số của các nhóm trên trục tung. Tần số của mỗi nhóm được tượng trưng bởi một thanh đứng mà chiều cao của thanh bằng với tần số của nhóm (Hình 1.1a).

• ***Biểu đồ tần suất*** là đồ thị trình bày các nhóm trên trục hoành và tần suất của các nhóm trên trục tung. Tần suất của mỗi nhóm được tiêu biểu bằng một thanh đứng mà chiều cao của thanh bằng với tần suất của nhóm (hình 1.1b).



Hình 1.1: Biểu đồ tần số (a) và tần suất (b) của số con đến tuổi đi học

III. CÁC SỐ ĐO MÔ TẢ *(Descriptive measures)*

Các số đo mô tả là những con số được dùng để mô tả bộ số liệu. Các thuật ngữ thường được sử dụng cho các số đo mô tả:

- *Tham số (parameters)* là những con số đặc trưng cho tổng thể mà ta chưa biết, được ký hiệu bằng chữ La Mã như μ, σ, N,...

- *Số thống kê (statistics)* là những con số đặc trưng cho mẫu, được ký hiệu bằng chữ thường như $\overline{X}$, s, n,...

1. Các số đo trung tâm *(Measures of central tendency)* là giá trị ở giữa, thường được sử dụng làm giá trị đại diện cho bộ số liệu.

a) Số trung bình cộng *= Trung bình số học (arithmetical mean)*

- *Định nghĩa:* Số trung bình cộng là tổng các giá trị quan sát chia cho tổng số quan sát.

Nếu một tổng thể nào đó có N cá thể ($X_1, X_2, \ldots, X_N$) thì trị số trung bình của tổng thể (μ) sẽ là:

tổng thể

$$\mu = \frac{X_1 + X_2 + \cdots + X_N}{N} = \sum_{i=1}^{N} X_i / N$$

Thường trong thực tế chúng ta không biết được μ do N quá lớn mà chúng ta chỉ có thể ước lượng μ qua trung bình của mẫu ($\overline{X}$). Do đó, nếu một mẫu có n cá thể được rút ra từ tổng thể, lúc đó giá trị trung bình của mẫu ($\overline{X}$) được tính như sau:

Mẫu

$$\hat{\mu} = \overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Ví dụ: Năng suất (tạ/ha) của một giống đậu xanh ghi nhận trên 16 lô như sau: 15, 21, 20, 19, 22, 21, 16, 19, 20, 16, 17, 24, 16, 21, 15 và 22

$$\hat{\mu} = \overline{X} = \frac{304}{16} = 19 \text{ tạ/ha}$$

Nếu số liệu được trình bày theo bảng phân bố thực nghiệm (Bảng 1.4), trung bình có thể được tính theo hai cách sau:

- Dựa vào tần số:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i f_i}{n}$$

X_i là giá trị của biến quan sát thứ i
f_i là tần số của giá trị X_i

- Dựa vào tần suất:

$$\overline{X} = \sum_{i=1}^{n} X_i F_i$$

X_i là giá trị của biến quan sát thứ i
f_i là tần suất của giá trị X_i

Bảng 1.4: Số liệu năng suất của một giống đậu xanh được trình bày dưới bảng phân bố thực nghiệm

Năng suất (X)	Tần số (f)	Tần số tích lũy (S)	Tần suất (F)	fX	FX
15	2	2	0,1250	30	1,8750
16	3	5	0,1875	48	3,0000
17	1	6	0,0625	17	1,0625
19	2	8	0,1250	38	2,3750
20	2	10	0,1250	40	2,5000
21	3	13	0,1875	63	3,9375
22	2	15	0,1250	44	2,7500
24	1	16	0,0625	24	1,5000
Tổng:	16		1,0000	304	19,0000

$$\Rightarrow \quad \overline{X} = \sum_{i=1}^{n} X_i f_i / n = 304/16 = 19 \text{ tạ/ha} \qquad (1)$$

hoặc $$\overline{X} = \sum_{i=1}^{n} X_i F_i = 19 \text{ tạ/ha}$$

- *Tính chất của số trung bình cộng:*

* Đây là tham số đặc trưng cho một tính chất nào đó của tổng thể và được sử dụng phổ biến nhất trong các nghiên cứu về nông sinh học. Dựa vào các số trung bình để so sánh giữa hai hay nhiều tổng thể.

* Số trung bình cộng bị ảnh hưởng bởi giá trị của tất cả các số quan sát. Khi giá trị của một số quan sát thay đổi sẽ dẫn đến sự thay đổi giá trị của số trung bình.

* Nó dễ bị lệch bởi sự hiện diện của các giá trị cực (quá nhỏ hoặc quá lớn). Do đó nó rất nhạy với sự hiện diện của các số liệu bất thường.

* Nó trở nên không đại diện cho giá trị ở giữa của mẫu nếu phân bố bị lệch nhiều.

- Đối với số liệu gom nhóm như ở Bảng 1.5, số trung bình có thể được tính theo công thức dựa vào tần số (1) với X_i là giá trị giữa hoặc được tính theo công thức dưới đây:

$$\overline{X} = X_0 + \frac{d}{n} \cdot \sum f_i . h_i$$

với X_0 là giá trị chọn trước bất kỳ. Thông thường chọn giá trị giữa của bộ số liệu.

d là độ rộng của nhóm

f_i là tần số của nhóm có chứa số trung bình

n là tổng số quan sát

$h_i = \frac{X_i - X_0}{d}$ (X_i là giá trị giữa của nhóm thứ i)

Bảng 1.5: Hàm lượng dầu của 100 mẫu đậu nành

HL dầu (%)	Giá trị giữa (X_i)	Tần số (f)	Tần số tích lũy (S)	$h_i = \frac{X_i - X_0}{d}$	f_ih_i	X_if_i
12 - 14	13	7	7	- 3	- 21	91
15 - 17	16	19	26	- 2	- 38	304
18 - 20	19	22	48	- 1	- 22	418
21 - 23	22	31	79	0	0	682
24 - 26	25	12	91	1	12	300
27 - 29	28	9	100	2	18	252
Tổng:		100		- 3	- 51	2047

Chọn $X_0 = 22$

$$\overline{X} = X_0 + \frac{d}{n} \cdot \sum f_i.h_i = 22 + \frac{3}{100} x(-51) = 20,47\%$$

Hoặc có thể áp dụng công thức sau :

$$\overline{X} = \sum_{i=1}^{n} X_i f_i / n = \frac{2047}{100} = 20,47\%$$ (X_i là giá trị giữa của nhóm thứ i)

b) Số trung bình nhân *(Geometrical mean)*: Trung bình nhân hay trung bình hình học của N số dương $X_1, X_2, \ldots, X_N$ là căn bậc N của tích các số đó.

$$G = \overline{X}_{geo} = \sqrt[N]{X_1.X_2 \ldots X_N}$$

Số trung bình nhân lúc nào cũng nhỏ hơn trung bình cộng và nó được dùng thay thế số trung bình cộng trong trường hợp có một hay một vài trị số quá lớn trong dãy số liệu, vì các trị số quá lớn này sẽ che lấp ảnh hưởng của đại đa số các trị số nhỏ.

Thường số trung bình nhân được sử dụng khi xử lý những số liệu về chuẩn độ vi sinh vật, chuẩn độ các độc tố, chuẩn độ sinh hóa trong những thí nghiệm trên động vật. Ví dụ, xác định ngưỡng gây chết 50% (LD_{50})

c) Số trung bình Harmonic *(Harmonic mean)*

Trung bình harmonic của N số quan sát ($X_1, X_2, \ldots, X_N$) là số nghịch đảo của trung bình số học của các số nghịch đảo của số quan sát.

$$H = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{X_i}} = \frac{N}{\sum \frac{1}{X}}$$

Trong thực hành, để dễ nhớ có áp dụng công thức sau:

$$\frac{1}{H} = \frac{\sum \frac{1}{X}}{N} = \frac{1}{N}\sum \frac{1}{X}$$

Ví dụ: Trung bình harmonic của các số 2, 3, 4 và 6

$$H = \frac{4}{\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{6}} = \frac{4}{\frac{15}{12}} = 3{,}2$$

Đối với số liệu gom nhóm, số trung bình harmonic được tính theo công thức sau:

$$\frac{1}{H} = \frac{1}{N}\left(\frac{f_1}{X_1}+\frac{f_2}{X_2}+\frac{f_3}{X_3}+\cdots\right) = \frac{1}{N}\sum\frac{f}{X}$$

với $X_1, X_2, X_3, \cdots$ và $f_1, f_2, f_3, \cdots$ lần lượt là các giá trị giữa và tần số tương ứng của các nhóm.

$N = f_1 + f_2 + \cdots = \sum f$

**** Sự quan hệ giữa trung bình số học, hình học và harmonic***

Số trung bình nhân (G) của N số dương $X_1, X_2, \ldots, X_N$ nhỏ hơn hoặc bằng trung bình cộng ($\overline{X}$) nhưng lớn hơn hoặc bằng trung bình harmonic (H)

$$H \le G \le \overline{X}$$

Dấu bằng (=) chỉ có giá trị nếu tất cả các số $X_1, X_2, \ldots, X_N$ giống hệt nhau.

Ví dụ: Bộ số liệu 2, 3, 4 và 6 có trung bình số học là 3,75; trung bình hình học là 3,46 và trung bình harmonic là 3,2.

d) Căn trung bình bình phương *(Root mean square)* hoặc trung bình bậc hai *(quadratic mean)* của N số quan sát $X_1, X_2, \ldots, X_N$ được định nghĩa như sau:

$$RMS = \sqrt{\overline{X^2}} = \sqrt{\frac{\sum_{i=1}^{N} X_i^2}{N}} = \sqrt{\frac{\sum Xi^2}{N}}$$

Ví dụ: RMS của bộ số liệu 1, 3, 5 và 7 là

$$RMS = \sqrt{\frac{1^2+3^2+5^2+7^2}{4}} = \sqrt{21} = 4{,}58$$

Dạng trung bình này thường được dùng trong những ứng dụng về vật lý.

e) Số trung vị *(Median):*

- *Định nghĩa*: Là trị số ở giữa của một chuỗi số liệu đã được xếp theo thứ tự tăng (hoặc giảm) dần.

Cách xác định số trung vị:

❑ Khi bộ số liệu là các giá trị rời rạc

* Nếu cỡ mẫu (n) là số lẻ, lúc đó số trung vị là số hạng ở giữa của dãy số quan sát; nghĩa là số hạng ở vị trí thứ (n + 1)/2

Ví dụ: 3 7 9 10 12

$M_e = 9$, nghĩa là giá trị ở vị trí thứ (n + 1)/2 = (5 + 1)/2 = 3

* Nếu n là số chẵn, không có số hạng giữa, lúc đó trung vị sẽ là trung bình cộng của hai số hạng ở giữa; nghĩa là các số hạng ở vị trí thứ (n/2) và [(n/2) + 1] trong dãy số.

Ví dụ: 1 3 7 8 12 16

$M_e = \frac{7+8}{2} = 7{,}5$, nghĩa là giá trị trung bình của hai số hạng thứ (n/2) = (6/2) = 3 và thứ [(n/2) + 1] = [(6/2) + 1] = 4.

❑ Khi số liệu được trình bày dưới dạng phân bố tần số; ví dụ số liệu ở Bảng 1.4, số trung vị được xác định dựa vào bảng phân bố tích lũy; nghĩa là số trung vị sẽ là

trung bình của hai giá trị ở vị trí thứ 8 và thứ 9. Vì giá trị thứ 8 là 19 và giá trị thứ 9 là 20, nên số trung vị sẽ bằng 19,5 tạ/ha.

❑ Đối với số liệu gom nhóm, số trung vị có thể được tính theo hai cách:

* *Cách 1:* Lấy lại ví dụ năng suất đậu ở bảng 1.5, trong đó n = 100, nên số trung vị là trung bình của hai giá trị ở vị trí 50 và 51. Tổng của tần số ba nhóm đầu tiên là 7 + 9 + 22 = 48, vì thế để đạt đến số hạng thứ 50, cần có thêm 2 trong số 31 số hạng của nhóm thứ tư. Do khoảng của nhóm thứ tư, 21-23, thật sự tương ứng với năng suất từ 20,5 – 23,5; do đó, trung vị phải nằm 2/31 khoảng cách giữa 20,5 và 23,5; nghĩa là trung vị bằng

$$M_e = 20,5 + \frac{2}{31}(23,5 - 20,5) = 20,69$$

* *Cách 2:* Áp dụng công thức tính trung vị như sau:

$$M_e = X_i + d \cdot \frac{\left(\frac{n}{2} - S_{i-1}\right)}{f_i}$$

với X_i là giới hạn dưới của nhóm có chứa số trung vị = 20,5.

d là độ rộng của nhóm có chứa số trung vị = 3.

S_{i-1} là tần số tích lũy của nhóm đứng kế trước nhóm có chứa số trung vị; nghĩa là, tổng của tất cả các nhóm đứng trước nhóm trung vị = 7 + 9 + 22 = 48.

f_e là tần số của nhóm có chứa số trung vị = 31

n là tổng số cá thể quan sát = 100

Do đó,

$$M_e = 20,5 + 3 \cdot \frac{\frac{100}{2} - 48}{31} = 20,69$$

- *Tính chất của số trung vị:*

* Số trung vị chia đôi chuỗi số liệu thành hai phần bằng nhau; nghĩa là, có một nửa số liệu cao hơn số trung vị và một nửa số liệu thấp hơn trung vị.

* Số trung vị là số đo đại diện cho một nhóm số, nó có tính trội hơn số trung bình là không bị ảnh hưởng bởi các giá trị cực (quá lớn hay quá nhỏ) của chuỗi số liệu. Do đó, trong một số trường hợp số trung vị được xem là một giá trị trung tâm điển hình hơn cho chuỗi số liệu so với số trung bình, nhất là trong trường hợp phân bố bất đối xứng.

* Số trung vị có vai trò mô tả một cách thuần túy nên nó ít được sử dụng trong thống kê suy luận. Mặt khác, trong trường hợp số liệu được gom nhóm thì cách xác định số trung vị đôi khi dài dòng.

f) Mode (M_o):

- *Định nghĩa:* Mode của bộ số liệu là giá trị của biến X ứng với tần số cao nhất; nghĩa là, giá trị của biến X xuất hiện nhiều lần nhất.

Ví dụ: Dãy số liệu: 1, 7, 9, 12, 30, 12, 28 có mode bằng 12 (M_o = 12).

Đối với số liệu gom nhóm, mode được tính như sau:

$$M_o = X_o + d \cdot \frac{f_0 - f_{0-1}}{(f_0 - f_{0-1}) + (f_0 - f_{0+1})}$$

1° Có $M_o \neq M_{o+1}$

Số liệu xấp xỉ = nhau

với X_0 là giới hạn dưới của nhóm có chứa mode (nghĩa là, nhóm có tần số cao nhất).

d là độ rộng của nhóm có chứa mode.

f_0 là tần số của nhóm có chứa mode

f_{0-1} là tần số của nhóm đứng kế trước nhóm có chứa mode

f_{0+1} là tần số của nhóm đứng kế sau nhóm có chứa mode

Ví dụ: Số liệu năng suất đậu trong Bảng 1.5:

$$M_o = 20,5 + 3.\frac{31-22}{(31-22)+(31-12)} = 20,5 + 3.\frac{9}{9+19} = 21,46$$

- Tính chất của mode:

* Đó là số đo hầu như ít được sử dụng. Tuy nhiên, mode có giá trị mô tả quan trọng, nó cho ta biết giá trị của X thường gặp nhất, đó là điều cần biết trong thống kê ứng dụng. Do đó, trong trường hợp phân bố bất đối xứng, chúng ta cần biết mode cùng với số trung bình.

* Khi tất cả các số quan sát trong bộ số liệu có cùng tần số, lúc đó chúng ta nói bộ số liệu không có mode. Như vậy, bộ số liệu có thể không có mode, có một mode, hai mode hoặc nhiều mode.

* Nó không bị ảnh hưởng bởi các giá trị cực (quá lớn hay quá nhỏ) của chuỗi số liệu

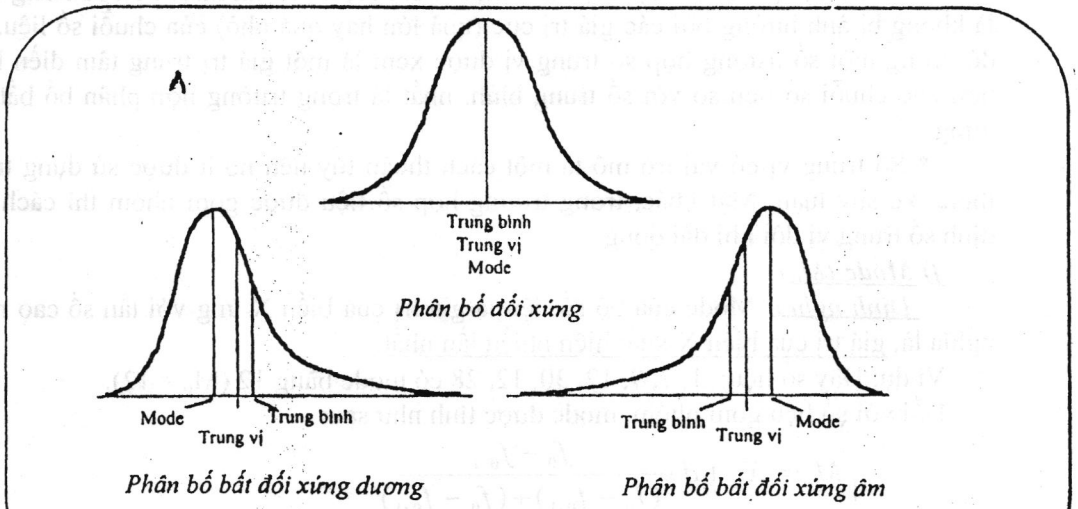
Nhược điểm của mode là rất khó xác định một cách chính xác nếu nó nằm trong một nhóm trong trường hợp số liệu được gom nhóm; do đó, cũng không tiện đưa mode vào tính toán đại số học.

**** Sự liên hệ giữa trung bình, trung vị và mode***

Đối với đường cong tần số bất đối xứng, chúng ta tìm thấy mối liên hệ thực nghiệm như sau:

Trung bình – Mode = 3(trung bình – trung vị)

Hình 1.2 cho thấy các vị trí tương đối của trung bình, trung vị và mode đối với các đường cong tần số bị lệch về bên phải và về bên trái. Đối với đường cong đối xứng, trung bình, trung vị và mode trùng với nhau.



Hình 1.2: Vị trí tương đối của trung bình, trung vị và mode tùy theo mức độ đối xứng

g) Các số tứ phân vị (Quartiles), thập phân vị (Deciles) và bách phân vị (Percentiles)

Nếu bộ số liệu được xếp theo thứ tự tăng (hoặc giảm) dần, các giá trị ở giữa (hoặc các trung bình cộng của hai giá trị ở giữa) chia chuỗi số liệu thành bốn phần bằng nhau được gọi là các số tứ phân vị, ký hiệu lần lượt là Q_1, Q_2 và Q_3. Giá trị Q_2 bằng với số trung vị (Hình 1.3).

Tương tự, các giá trị chia chuỗi số liệu thành 10 phần bằng nhau được gọi là các số thập phân vị, ký hiệu là D_1. D_2,..., D_9. Trong khi. các giá trị chia chuỗi số liệu thành 100 phần bằng nhau được gọi là các số bách phân vị, ký hiệu là P_1, P_2,..., P_{99}. Số thập phân vị thứ 5 và số bách phân vị thứ 50 tương ứng với số trung vị. Các số bách phân vị thứ 25 và 75 lần lượt tương ứng với các số tứ phân vị thứ nhất và thứ ba.

Đối với số liệu gom nhóm, cách tính các số tứ phân vị, thập vị và bách phân vị được suy rộng từ cách tính trung vị và mode. Ví dụ, công thức tính số bách phân vị thứ i

$$P_i = X_i + d \cdot \frac{\left(\frac{n}{100} - S_{i-1}\right)}{f_i}$$

với P_i là giá trị của bách phân vị thứ i

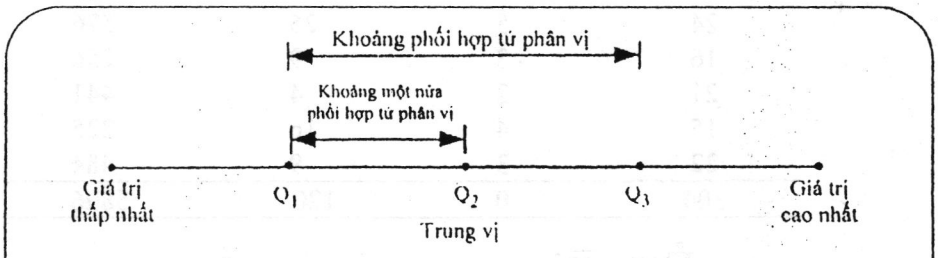
X_i là giới hạn dưới của nhóm có chứa số bách phân vị thứ i.

d là độ rộng của nhóm có chứa số bách phân vị thứ i.

S_{i-1} là tần số tích lũy đến nhóm đứng kế trước nhóm có chứa bách phân vị thứ i

f_i là tần số của nhóm có chứa số bách phân vị thứ i

n là tổng số quan sát



Hình 1.3: Giải thích khoảng một nửa phối hợp tứ phân vị

2. Các số đo phân tán *(Measures of dispersion)*

a) Phương sai (Variance), còn gọi là trung bình bình phương *(Mean quare = MS)* là tham số đặc trưng tiêu biểu nhất cho tính chất phân tán của tổng thể. Giả sử, chúng ta muốn chọn hai kế hoạch đầu tư và chúng ta biết cả hai đều cho lợi nhuận trung bình hằng năm là 10%; như vậy, chúng ta có thể kết luận cả hai kế hoạch đều tốt như nhau. Tuy nhiên, giả sử chúng ta biết kế hoạch A có phương sai lớn gấp đôi kế hoạch B. Tham số này cung cấp cho chúng ta thêm thông tin để chọn. Nếu chúng ta muốn tương đối chắc chắn số lợi nhuận thu được gần 10%, chúng ta sẽ chọn kế hoạch B. Nhưng nếu chúng ta muốn có sự may rủi; nghĩa là, số tiền lời có thể vượt quá 10%, chúng ta sẽ chọn kế hoạch A.

• *Phương sai của tổng thể* là trung bình bình phương của các độ lệch từ giá trị quan sát (X_i) đến giá trị trung bình (μ). Công thức

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$ (N là kích thước của tổng thể)

• *Phương sai của mẫu*. Thường trong thực tế chúng ta không biết được trị số thật của σ^2, vì không biết μ, mà chỉ ước lượng σ^2 bằng cách dùng số thống kê từ một mẫu ngẫu nhiên, gọi là phương sai mẫu.

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

Lấy lại số liệu năng suất của một giống đậu xanh

Năng suất (X)	$X_i - \overline{X}$	$(X_i - \overline{X})^2$	$(X)^2$
15	- 4	16	225
21	2	4	441
20	1	1	400
19	0	0	361
22	3	9	484
21	2	4	441
16	- 3	9	256
19	0	0	361
20	1	1	400
16	- 3	9	256
17	- 2	4	289
24	5	25	576
16	- 3	9	256
21	2	4	441
15	- 4	16	225
22	3	9	484
304	0	120	5896

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = 120 / 15 = 8{,}00$$

hoặc

$$s^2 = \frac{\sum_{i=1}^{n} X_i^2 - [(\sum_{i=1}^{n} X_i)^2 / n]}{n-1}$$

$$s^2 = \frac{5896 - [(304)^2 / 16]}{15} = 8{,}00$$

* $\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} x_i^2 = \sum X_i^2 - [(\sum_{i=1}^{n} X_i)^2 / n]$: Tổng bình phương của các độ lệch, gọi tắt là tổng bình phương *(sum of squares = SS)*

* (n - 1) : Độ tự do *(degree of freedom = df)*

* $\sum_{i=1}^{n} X_i)^2 / n$: Yếu tố hiệu chỉnh *(correction factor = C.F.)*

Đối với số liệu được viết dưới dạng tần số:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 f_i}{n-1}$$

$$s^2 = \frac{\sum_{i=1}^{n} X_i^2 f_i - [(\sum_{i=1}^{n} X_i f_i)^2 / n]}{n-1}$$

Cũng với số liệu năng suất của một giống đậu xanh ghi nhận trên 16 lô được trình bày dưới bảng phân bố thực nghiệm (Bảng 1.6).

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 f_i}{n-1} = \frac{120}{15} = 8 \text{ (tạ/ha)}^2$$

Bảng 1.6: Số liệu năng suất của một giống đậu xanh được trình bày dưới bảng phân bố thực nghiệm

Năng suất (X)	Tần số (f)	$(X_i - \overline{X})$	$(X_i - \overline{X})^2 f_i$	f_iX_i	$(X_i)^2 f_i$
15	2	-4	32	30	450
16	3	-3	27	48	768
17	1	-2	4	17	289
19	2	0	0	38	722
20	2	1	2	40	800
21	3	2	12	63	1323
22	2	3	18	44	968
24	1	5	25	24	576
Tổng	**16**		**120**	**304**	**5896**

Hoặc $$s^2 = \frac{\sum_{i=1}^{n} X_i^2 f_i - [(\sum_{i=1}^{n} X_i f_i)^2 / n]}{n-1} = \frac{5896 - (304)^2/16}{15} = \frac{120}{15} = 8 \text{ (tạ/ha)}^2$$

b) Độ lệch chuẩn *(Standard deviation = SD)* là căn bậc hai của phương sai

Nhà toán học người Nga, P.L. Chebychev (1821 - 1894), đã dùng số đo độ lệch chuẩn để đo lường độ phân tán của tổng thể. Số đo này có ưu điểm là cùng đơn vị với số liệu đo, trong khi phương sai là đơn vị bình phương.

$$\sigma = \sqrt{\sigma^2}$$

Tương tự, độ lệch chuẩn của mẫu (s) là:

$$s = \sqrt{s^2}$$

Độ lệch chuẩn của số liệu năng suất của một giống đậu xanh là:

$$s = \sqrt{s^2} = \sqrt{8} = 2{,}83 \text{ tạ/ha}$$

Nguyên tắc của Chebychev cho các tổng thể

Đối với tổng thể bất kỳ

- ***Tính chất 1***: Ít nhất 75% các giá trị quan sát của tổng thể nằm trong hai độ lệch chuẩn về một trong hai phía của trung bình; nghĩa là, giữa $\mu - 2\sigma$ và $\mu + 2\sigma$.

- ***Tính chất 2:*** Ít nhất 89% các giá trị quan sát của tổng thể nằm trong ba độ lệch chuẩn về một trong hai phía của trung bình; nghĩa là, giữa $\mu - 3\sigma$ và $\mu + 3\sigma$.

- *Tính chất 3*: Tổng quát, với số k bất kỳ lớn hơn 1 (k > 1), ít nhất $1 - 1/k^2$ các giá trị quan sát của tổng thể nằm trong k độ lệch chuẩn về một trong hai phía của trung bình; nghĩa là, giữa $\mu - k.\sigma$ và $\mu + k.\sigma$.

Nguyên tắc của Chebychev cũng có thể được áp dụng cho bộ số liệu mẫu:

c) Hệ số biến động = hệ số biến thiên *(Coefficient of Variation)*

Hệ số biến động (cv) cũng được dùng để đo lường độ phân tán của tổng thể. Đó là phần trăm của tỉ số giữa độ lệch chuẩn và trung bình.

$$(\%)cv = \frac{\sigma}{\mu}.100$$

Vì σ và μ không biết, nên hệ số này được ước lượng bằng hệ số biến động của mẫu:

$$(\%)cv = \frac{s}{\overline{X}}.100$$

Với số liệu năng suất của một giống đậu xanh, cv được tính như sau:

$$(\%)cv = \frac{2,83}{19}x100 = 14,9\%$$

Số đo độ lệch chuẩn là một trị số tuyệt đối có cùng đơn vị với trị số trung bình. Tuy nhiên, số đo này không thể dùng để so sánh mức độ biến động của hai hay nhiều chuỗi số liệu có đơn vị đo lường khác nhau. Để khắc phục tình trạng biến động của đơn vị, Pearson đã đưa ra khái niệm hệ số biến động. Đây là một trị số tương đối chỉ độ chính xác của việc so sánh các giá trị trung bình và là chỉ số cho phép đánh giá mức độ tin cậy của thí nghiệm. Giá trị của cv càng cao thì độ tin cậy của thí nghiệm càng thấp.

Hệ số biến động có ưu điểm hơn độ lệch chuẩn ở chỗ giúp ta so sánh mức độ phân tán của hai tổng thể có số trung bình khác nhau và có đơn vị đo lường khác nhau.

Ví dụ : Số liệu về chiều cao và trọng lượng của 200 học sinh lớp 5 như sau:

* Chiều cao $\overline{X}$ = 100 cm, σ = 10 và cv = 10%
* Trọng lượng $\overline{X}$ = 25, σ = 5 và cv = 20%

Như vậy, số liệu về trọng lượng phân tán hơn số liệu về chiều cao

Trong các thí nghiệm nông nghiệp, giá trị cv thay đổi theo từng kiểu thí nghiệm, cây trồng và tính trạng đo lường. Tuy nhiên, một nghiên cứu viên có kinh nghiệm có thể quyết định một cách hợp lý việc chấp nhận giá trị của cv cho kiểu thí nghiệm đang khảo sát. Ví dụ, đối với tính trạng năng suất lúa của các thí nghiệm ngoài đồng, khoảng cv có thể chấp nhận là 6 - 8% cho các thí nghiệm về giống; 10 - 12% cho các thí nghiệm về phân bón và 13 - 15% cho các thí nghiệm về thuốc trừ sâu và thuốc trừ cỏ.

Giá trị cv của các tính trạng khác thường khác với cv của năng suất. Ví dụ, với các thí nghiệm ngoài đồng thì cv của năng suất lúa khoảng 10%, cv của số chồi khoảng 20% và cv của chiều cao cây khoảng 3%.

Bằng cách so sánh cv, chúng ta có thể đánh giá mức độ chính xác khi thực hiện thí nghiệm. Ngoài ra, vì tính chất ổn định của cv nên trong nhiều trường hợp có thể dùng cv để ước lượng độ lệch chuẩn (s) và xác định cỡ mẫu (n) để thu thập.

Ngoài trung bình và phương sai, hai loại số đo khác cũng dùng để mô tả phân bố của tổng thể: Số đo bất đối xứng *(Skewness)* và số đo độ phẳng *(Kurtosis)*.

d) Giá trị SKEW *(Skewness)* phản ánh mức độ bất đối xứng của phân bố số liệu quanh giá trị trung bình.

$$\text{Skewness} = \frac{N}{(N-1)(N-2)}\sum\left(\frac{X_i - \overline{X}}{s}\right)^3$$

$$\text{Skewness} = \frac{\overline{X} - \text{mode}}{s}$$

Để tránh dùng mode, chúng ta có thể áp dụng công thức sau:

$$\text{Skewness} = \frac{3\left(\overline{X} - \text{median}\right)}{s}$$

Giá trị skew dương nếu số liệu phân bố bất đối xứng có đuôi nằm lệch về bên phải; nghĩa là, lệch về phía các giá trị dương. Ngược lại, nó âm nếu số liệu phân bố bất đối xứng có đuôi nằm lệch về bên trái; nghĩa là, lệch về phía các giá trị âm. Trường hợp phân bố đối xứng, trung bình, trung vị và mode ở cùng vị trí và nằm ngay dưới đỉnh của đường cong. Tuy nhiên, với phân bố bất đối xứng các số đo này có giá trị khác nhau (Hình 1.2).

- Giải thích hệ số bất đối xứng

* $-1 \le \text{skew} \le +1$

* Skew chỉ mức độ bất đối xứng và chiều bất đối xứng

+ Đối xứng ⇒ skew = 0

+ Bất đối xứng ⇒ trung bình và trung vị nằm xa nhau

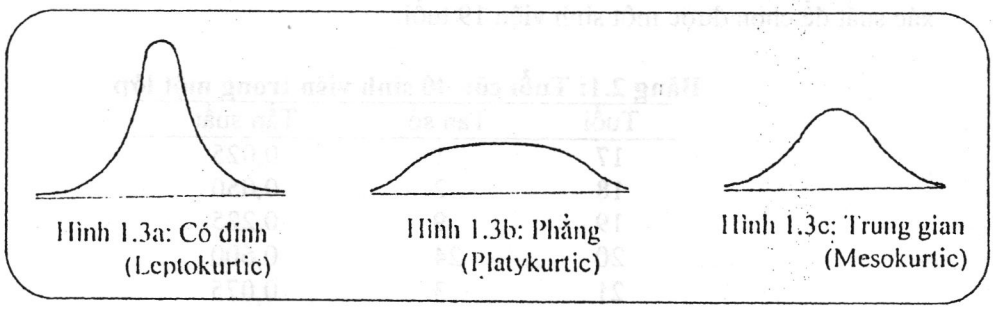
Số dương ⇒ bất đối xứng dương (đuôi dài về bên phải)

Số âm ⇒ bất đối xứng âm (đuôi dài về bên trái)

Ví dụ: skew = 0,05 ⇒ bất đối xứng dương yếu.

e) Giá trị KURT (Kurtosis). Giá trị Kurt phản ánh mức độ có đỉnh của dạng phân bố. Giá trị Kurt dương đối với phân bố có đỉnh *(Leptokurtic)*, hai đuôi rất dài (Hình 1.3a); âm cho số liệu đối với phân bố có đỉnh phẳng *(Platykurtic)*, các đuôi ngắn (Hình 1.3b), và bằng 0, đối với phân bố chuẩn có đỉnh không quá cao hoặc không quá phẳng (Hình 1.3c), được gọi là phân bố có đỉnh trung gian *(Mesokurtic)*.

$$\text{Kurtosis} = \left\{ \frac{N(N+1)}{(N-1)(N-2)(N-3)} \sum \left(\frac{X_i - \overline{X}}{s} \right)^4 \right\} - \frac{3(N-1)^2}{(N-2)(N-3)}$$



Hình 1.3a: Có đỉnh (Leptokurtic) — Hình 1.3b: Phẳng (Platykurtic) — Hình 1.3c: Trung gian (Mesokurtic)

CHƯƠNG 2

CÁC DẠNG PHÂN BỐ CỦA BIẾN NGẪU NHIÊN

I. PHÂN BỐ XÁC SUẤT *(Probability distribution)*

1. Một vài khái niệm về lý thuyết xác suất

a) Phép thử ngẫu nhiên có thể là một thí nghiệm, một quan sát, một cuộc điều tra,... mà kết quả của nó không thể đoán trước được.

b) Biến cố là kết quả của phép thử. Mỗi biến cố tương ứng với một tổng thể.

- *Biến cố ngẫu nhiên (random event)* là biến cố có thể hoặc không thể xảy ra khi tiến hành một phép thử.
- *Biến cố không thể (impossible event)* là biến cố không bao giờ xảy ra.
- *Biến cố chắc chắn (certain event)* là biến cố luôn luôn xảy ra.

c) Định nghĩa xác suất

Xác suất của một biến cố A, p(A), là một số đo khả năng xuất hiện của biến cố A. Số đó luôn nằm giữa 0 và 1. Xác suất của biến cố A càng lớn (càng gần 1) thì biến cố A càng có nhiều khả năng xảy ra. Xác suất của biến cố A càng nhỏ (càng gần 0) thì biến cố A càng ít có khả năng xảy ra.

d) Định nghĩa xác suất cổ điển

Giả sử biến cố A có thể xảy ra f lần trong tổng số n lần thử nghiệm có đồng khả năng xuất hiện. Lúc đó, xác suất của biến cố A (được gọi là xác suất thành công) sẽ là

$$p(A) = \frac{f(A)}{n}$$

Xác suất không xuất hiện biến cố A (gọi là xác suất thất bại) sẽ bằng

q = p(không A) = [n - f(A)]/n = 1 - [f(A)/n] = 1 - p(A)

Do đó, p + q = 1 hoặc p(A) + p(không A) = 1

Ví dụ: Tuổi của 40 sinh viên ở một lớp được trình bày trong Bảng 2.1. Giả sử chọn ngẫu nhiên một sinh viên, nghĩa là mỗi sinh viên có khả năng được chọn như nhau, tìm xác suất để chọn được một sinh viên 19.tuổi.

Bảng 2.1: Tuổi của 40 sinh viên trong một lớp

Tuổi	Tần số	Tần suất
17	1	0,025
18	2	0,050
19	9	0,225
20	24	0,600
21	3	0,075
22	1	0,025
Tổng:	**40**	**1,000**

Vì có 9 trong số 40 sinh viên trong lớp 19 tuổi. Vì vậy, sẽ có 9 trong số 40 cơ hội để chọn một sinh viên có độ tuổi 19. Xác suất sẽ là:

$$\frac{\text{Số sinh viên 19 tuổi}}{\text{Tổng số sinh viên}} = \frac{9}{40} = 0{,}225$$

Cũng với ví dụ trên, hãy tìm tìm xác suất chọn ngẫu nhiên một sinh viên nhỏ hơn 20 tuổi.

Cũng với ví dụ trên, hãy tìm tìm xác suất chọn ngẫu nhiên một sinh viên nhỏ hơn 20 tuổi.

Từ Bảng 2.1, chúng ta thấy có 12 (1 + 2 + 9) sinh viên trong lớp dưới 20 tuổi; vì thế f = 12 và xác suất sẽ bằng

$$\frac{f}{n} = \frac{12}{40} = 0,3$$

Như vậy, sẽ có 30% sinh viên trong lớp có độ tuổi dưới 20.

Đối với số liệu gom nhóm, giả sử chiều cao của 100 nam sinh viên ở một trường Đại học X được trình bày trong Bảng 2.2

Bảng 2.2: Chiều cao của 100 nam sinh viên ở một trường Đại học X.

Chiều cao (cm)	Số sinh viên (Tần số)
150 - 155	5
155 - 160	19
160 - 165	42
165 - 170	26
170 - 175	8
Tổng	**100**

Tìm xác suất để chọn ngẫu nhiên một sinh viên trong trường Đại học X có chiều cao từ 155 cm đến 165 cm.

Từ Bảng 2.2 cho thấy số sinh viên có chiều cao từ 155 cm đến 165 cm là 19 + 42 = 61 (f = 61). Vì vậy, xác suất để chọn được một sinh viên có chiều cao từ 155 cm đến 165 cm sẽ bằng

$$\frac{f}{n} = \frac{61}{100} = 0,61$$

Do đó, sẽ có 61% sinh viên trong lớp có chiều cao từ 155 cm đến 165 cm.

e) Định nghĩa bằng tần suất

Giả sử phép thử có thể lặp lại rất nhiều lần trong những điều kiện giống hệt nhau. Nếu trong n lần thực hiện phép thử, biến cố A xuất hiện f lần thì tỷ số

$$F(A) = f(A)/n$$

được gọi là tần suất của biến cố A trong n phép thử. Khi n gia tăng vô hạn ($n \to \infty$), tần suất này dao động quanh giá trị cố định p nào đó. Hằng số p được gọi là xác suất của A, ký hiệu là p(A).

Tóm lại, nếu A là biến cố có thể xảy ra, f(A) là số lần xuất hiện biến cố A trong n lần thử nghiệm, lúc đó

$$p(A) = f(A)/n$$

chính là tần suất của A trong n phép thử. Khi n đủ lớn, p(A) được gọi là xác suất của biến cố A có thể xảy ra trong thử nghiệm.

f) Xác suất có điều kiện: Biến cố độc lập và phụ thuộc

Hai biến cố A và B được gọi là độc lập với nhau nếu việc xảy ra hoặc không xảy ra của biến cố này không làm ảnh hưởng đến xác suất xuất hiện của biến cố kia. Trường hợp ngược lại, hai biến cố A và B phụ thuộc nhau; nghĩa là, việc xảy ra hoặc không xảy ra biến cố A có ảnh hưởng đến xác suất xảy ra của biến cố B. Xác suất của B được tính trong điều kiện biết rằng A đã xảy ra được gọi là xác suất của B với điều kiện A, ký hiệu là p(B/A) hay còn gọi là xác suất có điều kiện của B với điều kiện A đã xảy ra.

$$p(B/A) = \frac{p(AB)}{p(A)}$$

Nếu chúng ta biểu thị AB là biến cố mà "cả hai A và B cùng xảy ra" hay còn gọi là biến cố phức *(compound event)*, lúc đó

$$p(AB) = p(A).p(B/A)$$

Nếu A và B độc lập thì

$$p(AB) = p(A).p(B)$$

Tổng quát, với ba biến cố bất kỳ A, B và C, ta có:

$$p(ABC) = p(A).p(B/A).p(C/AB)$$

Nếu A, B và C độc lập thì

$$p(ABC) = p(A).p(B).p(C)$$

***g) Các biến cố loại trừ lẫn nhau** (Mutually exclusive events)*

Hai hay nhiều biến cố được gọi là loại trừ lẫn nhau nếu chúng không bao giờ xảy ra đồng thời; nghĩa là, sự xuất hiện của một biến cố bất kỳ sẽ loại trừ sự xuất hiện của các biến cố khác. Do đó, nếu A và B là hai biến cố loại trừ lẫn nhau, lúc đó

$$p(AB) = 0$$

Nếu A + B biểu thị biến cố mà "hoặc A hoặc B hoặc cả hai cùng xảy ra", lúc đó

$$p(A+B) = p(A) + p(B) - p(AB)$$

Trường hợp hai biến cố A và B loại trừ lẫn nhau. Lúc đó

$$p(A+B) = p(A) + p(B)$$

2. Phân bố xác suất

Nếu biến X đến từ bộ số liệu rời rạc có các giá trị $X_1, X_2, \ldots, X_n$ với các xác suất tương ứng $p_1, p_2, \ldots, p_n$ ($p_1 + p_2 + \cdots + p_n = 1$). Lúc đó, phân bố xác suất của X là một bảng trong đó ta ghi các giá trị của X và kèm theo mỗi giá trị của X là xác suất tương ứng với giá trị đó. Như vậy, phân bố xác suất của X sẽ là một bảng có dạng như sau:

X	X_1	$X_2 \ldots\ldots X_n$
p(X)	p_1	$p_2 \ldots\ldots P_n$

Ví dụ: Tìm xác suất của con trai và con gái trong các gia đình có 3 con, Giả định xác suất sinh con trai và con gái bằng nhau.

Đặt A = biến cố "con trai trong gia đình = T" và B = biến cố "con gái trong gia đình = G". Do đó, theo giả định xác suất sinh con trai và con gái bằng nhau, p(T) = p(G) = 1/2. Trong các gia đình có 3 con, các biến cố loại trừ lẫn nhau sau đây có thể xảy ra với các xác suất tương ứng

(a) 3 con trai(TTT)

p(TTT) = p(T).p(T).p(T) = 1/8

Ở đây chúng ta giả định rằng việc sinh con trai lần sau hoàn toàn độc lập với sinh con trai lần trước.

(b) Cách tính tương tự cho trường hợp 3 con gái

p(GGG) = 1/8

(c) 2 trai và 1 gái (TTG + TGT + GTT)

p(TTG + TGT + GTT) = p(TTG) + p(TGT) + p(GTT)

= p(T).p(T).p(G) + p(T).p(G).p(T) + p(G).p(T).p(T)

= 1/8 + 1/8 + 1/8 = 3/8

(d) Tính tương tự cho trường hợp 2 gái và 1 trai, ta có:

p(GGT + GTG + TGG) = 3/8

Nếu chúng ta gọi X là biến ngẫu nhiên chỉ số con trai trong các gia đình có 3 con, phân bố xác suất được trình bày như sau:

Số con trai X	0	1	2	3
Xác suất p(X)	1/8	3/8	3/8	1/8

3. Trung bình và phương sai của biến X

a) Trung bình

Giá trị trung bình (hay kì vọng) của X, ký hiệu là E(X) được định nghĩa như sau:

$$E(X) = \mu = p_1X_1 + p_2X_2 \cdots + p_nX_n = \sum_{i=1}^{n} p_iX_i = \sum pX$$

b) Phương sai

Phương sai của X (σ^2) được định nghĩa như sau:

$$V(X) = \sigma^2 = \sum(X_i - \mu)^2 p(X_i) = \sum X_i^2 p(X_i) - \mu^2$$

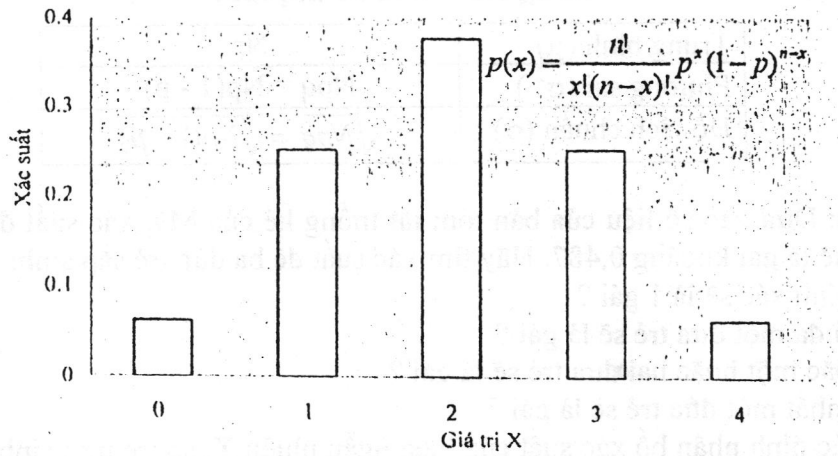
c) Độ lệch chuẩn

Độ lệch chuẩn của X (σ) là căn bậc hai của σ^2

II. PHÂN BỐ NHỊ THỨC *(Binomial distribution)* Co ok ?

Trong thống kê, phân bố nhị thức mô tả số lần có thể xảy ra của một biến cố nào đó trong một chuỗi quan sát. Biến cố được viết bằng mã nhị phân, nó có thể hoặc không thể xảy ra. Phân bố nhị thức được sử dụng khi nhà nghiên cứu quan tâm đến sự xuất hiện của một biến cố, chớ không phải tính chất quan trọng của biến cố đó. Chẳng hạn, trong một thử nghiệm thuốc trừ sâu, một con sâu có thể sống hay chết. Nhà nghiên cứu quan sát số lượng sâu còn sống sót, chớ không phải con sâu sống bao lâu sau khi xử lý. Phân bố nhị thức được ghi rõ bằng số lượng quan sát (n) và xác suất xuất hiện được ký hiệu bằng p.

Một ví dụ kinh điển thường được dùng để giải thích các khái niệm của lý thuyết xác suất là sự tung một đồng xu. Nếu đồng xu được tung 4 lần, lúc đó chúng ta có thể đạt được 0, 1, 2, 3 hay 4 mặt ngửa. Chúng ta cũng có thể đạt được 4, 3, 2, 1, hoặc 0 mặt sấp, nhưng các kết quả này lại tương đương với 0, 1, 2, 3, hoặc 4 mặt ngửa. Khả năng đạt được 0, 1, 2, 3, hoặc 4 mặt ngửa lần lượt là 1/16, 4/16, 6/16, và 1/16. Hình 2.1 chỉ phân bố với p = ½.



Hình 2.1: Biểu đồ phân bố xác suất của biến X với p = ½

Như vậy, nếu p là xác suất để một biến cố nào đó sẽ xảy ra trong một lần thử nghiệm (được gọi là xác suất thành công) và q = 1 - p là xác suất để biến cố đó không xảy ra (gọi là xác suất thất bại), lúc đó xác suất để biến cố đó sẽ xảy ra chính xác X lần trong N thử nghiệm độc lập (nghĩa là, sẽ xảy ra X lần thành công và N - X lần thất bại) được tính như sau:

$$p(x) = C_N^X p^X q^{N-X} = \frac{N!}{X!(N-X)!} p^X q^{N-X} \qquad (1)$$

với X là biến nhị thức (nhị phân); nghĩa là biến rời rạc mà chỉ có thể dẫn đến một trong hai kết quả: thành công hoặc thất bại, (X = 0, 1, 2, . . . , N).

N! = N(N-1)(N-2)· · ·1 và 0! = 1.

Phân bố xác suất rời rạc thường được gọi là phân bố nhị thức, vì đối với X = 0, 1, 2, . . . , N tương ứng với các số hạng liên tiếp của công thức nhị thức hoặc với sự khai triển nhị thức

$$(p+q)^N = q^N + C_n^1 \cdot q^{N-1} p + C_n^2 \cdot q^{N-2} p^2 + \cdots + p^N \qquad (2)$$

với 1, $C_N^1, C_N^2, \ldots$ được gọi hệ số nhị thức *(binomial coefficients)*. Các hệ số này có thể đạt được qua tam giác Pascal. Chú ý các số đầu tiên và cuối cùng trong mỗi hàng là 1 và bất cứ số nào khác trong hàng có thể đạt được bằng cách cộng hai số bên phải và bên trái của nó trong hàng có trước.

P & X cùng nghĩa

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
... ...

Phân bố (1) còn được gọi là phân bố Bernoulli (Jame Bernoulli đã tìm ra công thức trên vào thế kỷ 17). Một vài tính chất của phân bố nhị thức được liệt kê trong Bảng 2.3.

Bảng 2.3: Phân bố nhị thức

Trung bình (μ)	Np
Phương sai (σ^2)	Npq = Np(1 - p)
Độ lệch chuẩn (σ)	$\sqrt{Npq} = \sqrt{Np(1-p)}$

Ví dụ: Dựa vào số liệu của bản tóm tắt thống kê của Mỹ, xác suất để một đứa trẻ mới sinh sẽ là gái khoảng 0,487. Hãy tìm xác suất để ba đứa trẻ sắp sinh:

a) chính xác sẽ là 1 gái ?
b) tối đa một đứa trẻ sẽ là gái ?
c) hoặc một hoặc hai đứa trẻ sẽ là gái ?
d) ít nhất một đứa trẻ sẽ là gái ?
e) Xác định phân bố xác suất của biến ngẫu nhiên X, số trẻ mới sinh trong ba đứa trẻ sắp sinh ra sẽ lá gái
f) Vẽ đồ biểu xác suất của biến X

Giải:

Đặt X là số trẻ mới sinh ra là gái trong ba đứa trẻ sắp sinh
p là xác suất của một đứa trẻ mới sinh ra là gái (p = 0,487)
n là số trẻ sắp sinh (n = 3)

Công thức xác suất nhị thức cho số trẻ sắp sinh sẽ là gái (X) là:

a) $p(X=1) = C_3^1(0,487)(0,513)^2 = \frac{3!}{1!(3-1)!}(0,487)(0,513)^2 = 0,384$

b) $p(X \le 1) = p(X=0) + p(X=1) = C_3^0 p^0 q^3 + C_3^1 pq^2$

$$= \frac{3!}{0!(3-0)!}(0,487)^0(0,513)^3 + \frac{3!}{1!(3-1)!}(0,487)(0,513)^2$$

$$= 0,519$$

c) $p(X = 1 \text{ hoặc } X = 2) = C_3^1(0,487)(0,513)^2 + C_3^2(0,487)^2(0,513)$

$$= \frac{3!}{1!(3-1)!}(0,487)(0,513)^2 + \frac{3!}{2!(3-2)!}(0,487)^2(0,513)$$

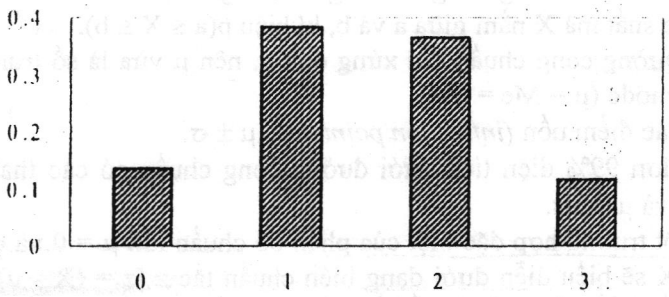
$$= 0,749$$

d) $p(X \ge 1) = 1 - C_3^0 p^0 q^3 = 1 - \frac{3!}{0!(3-0)!}(0,487)^0(0,513)^3 = 0,865$

e) Phân bố xác suất của biến X

X	0	1	2	3
p(X)	0,135	0,384	0,365	0,116

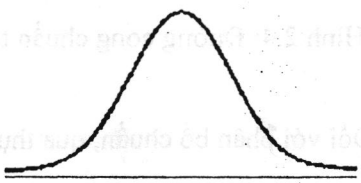
f) Đồ biểu xác suất của biến X:



III. PHÂN BỐ CHUẨN *(Normal distribution)*

Biến liên tục có thể được mô hình hóa bằng một phân bố xác suất lý thuyết có dạng hình chuông, gọi là phân bố chuẩn hay đường cong chuẩn hoặc phân bố Gauss (Carl Friedrich Gauss, 1777 - 1855), Hình 2.2.

Hình 2.2: Đường cong chuẩn



Phân bố chuẩn rất thường gặp (trong lý thuyết lẫn thực hành). Người ta nhận thấy nhiều số đo trong tự nhiên như chiều cao, trọng lượng, năng suất,... đều có phân bố dạng hình chuông. Phân bố chuẩn cũng thường được áp dụng trong thống kê suy luận;

chẳng hạn, trong một số trường hợp, phân bố chuẩn có thể được dùng để suy luận về trung bình tổng thể.

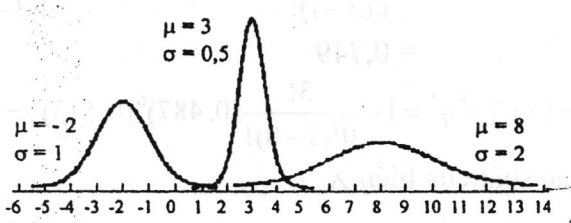
1. Hàm mật độ của phân bố chuẩn

Biến ngẫu nhiên X được gọi là có phân bố chuẩn với hai tham số μ và σ^2, kí hiệu $X \sim N(\mu, \sigma^2)$, nếu hàm mật độ của nó có dạng:

$$f(X) = Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2}$$

với $-\infty < X < +\infty$

μ là trung bình, σ là độ lệch chuẩn, π gần bằng 3,14159 và e là cơ số của logarithm Néper, gần bằng 2,71828. Hai tham số độc lập ở hàm mật độ là μ và σ, với μ là một số thực bất kỳ và σ là một số thực bất kỳ không âm. Ở bất cứ hàm mật độ chuẩn nào, μ và σ đều cố định; do đó, ứng với mỗi cặp μ và σ sẽ có một phân bố chuẩn khác nhau (Hình 2.3).



Hình 2.3: Ba đường cong chuẩn

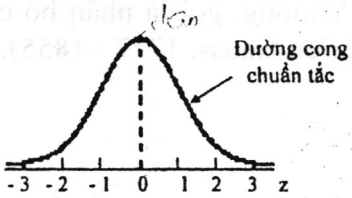
2. Tính chất của phân bố chuẩn

- Tổng diện tích dưới đường cong chuẩn và trục hoành (trục X) bằng 1; vì vậy, diện tích dưới đường cong chuẩn giữa hai giá trị X = a và X = b (với a < b) tiêu biểu cho xác suất mà X nằm giữa a và b, kí hiệu $p(a \le X \le b)$.
- Đường cong chuẩn đối xứng qua μ, nên μ vừa là số trung bình, vừa là trung vị, vừa là mode ($\mu = Me = M_O$).
- Các điểm uốn *(inflection points)* tại $\mu \pm \sigma$.
- Hơn 99% diện tích dưới đường cong chuẩn có các tham số μ và σ nằm giữa $\mu - 3\sigma$ và $\mu + 3\sigma$.

Một trường hợp đặc biệt của phân bố chuẩn khi $\mu = 0$ và $\sigma^2 = 1$, lúc đó biến ngẫu nhiên X sẽ biểu diễn dưới dạng biến chuẩn tắc z $[z = (X - \mu)/\sigma]$ và hàm mật độ của phân bố chuẩn được thay thế bằng dạng chuẩn tắc:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

Đường cong chuẩn tắc (Hình 2.4) đối xứng qua 0, có điểm cực đại tại X = 0. Các điểm uốn tại ± 1.



Hình 2.4: Đường cong chuẩn tắc

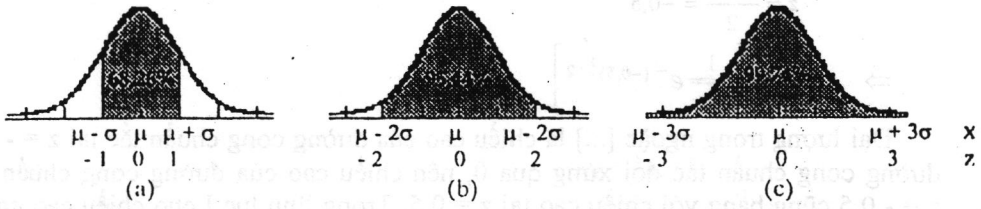
Đối với phân bố chuẩn, qua thực nghiệm cho thấy

(a) 68,26% giá trị của tổng thể nằm trong một độ lệch chuẩn về một trong hai phía của trung bình; nghĩa là, giữa $\mu - \sigma$ và $\mu + \sigma$

(b) 95,44% giá trị của tổng thể nằm trong hai độ lệch chuẩn về một trong hai phía của trung bình; nghĩa là, giữa $\mu - 2\sigma$ và $\mu + 2\sigma$.

(c) 99,74% giá trị của tổng thể nằm trong ba độ lệch chuẩn về một trong hai phía của trung bình; nghĩa là, giữa $\mu - 3\sigma$ và $\mu + 3\sigma$.

Ba tính chất này được trình bày bằng đồ thị ở Hình 2.5



Hình 2.5: Phần trăm giá trị của tổng thể có phân bố chuẩn nằm trong (a) một (b) hai (c) ba độ lệch chuẩn về một trong hai phía của trung bình

3. Phát họa đường cong chuẩn

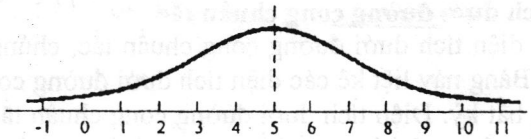
Mặc dầu không cần thiết để vẽ đường cong chuẩn chính xác, nhưng chúng ta có thể phát họa một cách nhanh chóng cho bất kỳ đường cong chuẩn nào theo cách sau:

Ví dụ: Phát họa đường cong chuẩn có các tham số $\mu = 5$ và $\sigma = 2$.

Đường cong chuẩn có tham số $\mu = 5$, $\sigma = 2$ đối xứng qua $\mu = 5$. Do đó, phần lớn diện tích dưới đường cong chuẩn nằm giữa

$\mu - 3\sigma = 5 - (3)(2) = -1$ và $\mu + 3\sigma = 5 + (3)(2) = 11$

Đường cong này được phát họa ở Hình 2.6.



Hình 2.6: Đường cong chuẩn có tham số $\mu = 5$ và $\sigma = 2$

4.Tính tung độ của đường cong chuẩn

Chúng ta có thể dùng công thức để tính tung độ của đường cong chuẩn với tham số μ và σ bất kỳ. Tuy nhiên, trong thực hành thường chúng ta dùng bảng tung độ của đường cong chuẩn tắc ở Phụ lục I nhanh hơn dùng công thức.

Phụ lục I liệt kê giá trị Y (tung độ của đường cong chuẩn tắc) cho các giá trị z dương.

Chẳng hạn, tìm tung độ của đường cong chuẩn tắc tại:

* $z = 0 \Rightarrow Y' = \frac{1}{\sqrt{2\pi}} e^0 = \frac{1}{\sqrt{2\pi}} \approx 0,4$ (tung độ cực đại)

* $z = 1,5 \Rightarrow Y' = 0,1295$

Đối với đường cong chuẩn có μ và σ bất kỳ, để đưa về dạng đường cong chuẩn tắc ta dùng công thức chuẩn tắc hoá như sau:

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

Ví dụ: Tìm chiều cao đường cong chuẩn có μ = 5, σ = 2, tại X = 4.

$$f(X) = Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2}.$$

Đặt: $z = \frac{X-\mu}{\sigma}$

$$\Rightarrow \quad Y = \frac{1}{\sigma}\left(\frac{1}{\sqrt{2\pi}} e^{-z^2/2}\right)$$

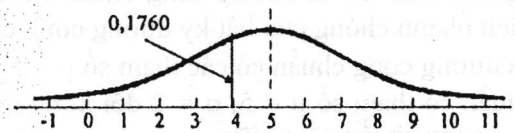
Thế số vào ta được:

$$z = \frac{4-5}{2} = -0,5$$

$$\Rightarrow \quad Y = \frac{1}{2}\left[\frac{1}{\sqrt{2\pi}} e^{-(-0,5)^2/2}\right]$$

Đại lượng trong ngoặc [...] là chiều cao của đường cong chuẩn tắc tại z = - 0,5. Vì đường cong chuẩn tắc đối xứng qua 0, nên chiều cao của đường cong chuẩn tắc tại z = - 0,5 cũng bằng với chiều cao tại z = 0,5. Trong Phụ lục I cho chiều cao này bằng 0,3521.

Vậy: $Y = \frac{0,3521}{2} = 0,1760$ là chiều cao của đường cong chuẩn có μ = 5, σ = 2, tại X = 4 (Hình 2.7)



Hình 2.7: Tung độ của đường cong chuẩn có tham số μ = 5 và σ = 2 tại X = 4

5.Tìm diện tích dưới đường cong chuẩn tắc

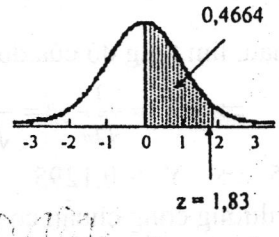
Để tìm nhanh diện tích dưới đường cong chuẩn tắc, chúng ta sử dụng bảng phân bố z ở Phụ lục II. Bảng này liệt kê các diện tích dưới đường cong chuẩn tắc giữa z = 0 và giá trị z dương bất kỳ. Diện tích dưới đường cong chuẩn tắc giữa z = 0 và giá trị z âm bất kỳ có thể tìm được từ bảng này bằng cách sử dụng tính chất đối xứng qua z = 0.

Trong Bảng z, cột đầu tiên bên trái là các giá trị z với một số lẻ và hàng trên cùng là số lẻ thứ hai của z. Các trị số bên trong bảng là diện tích dưới đường cong chuẩn tắc giữa 0 và giá trị z đã định.

Ví dụ 1: Tìm diện tích dưới đường cong chuẩn tắc giữa 0 và 1,83

Tra trong phụ lục II, giá trị ở giao điểm của hàng 1,8 và cột 0,03 là 0,4664 chính là diện tích dưới đường cong chuẩn tắc giữa 0 và 1,83.

Cũng có thể viết: $p(0 \le Z \le 1,83) = 0,4664$ (Hình 2.8).

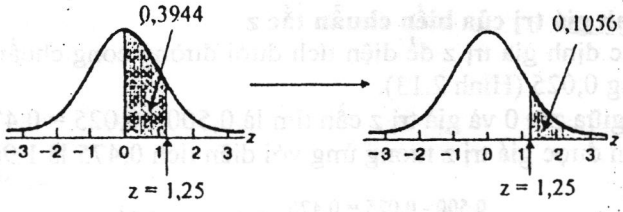


Hình 2.8: Diện tích dưới đường cong chuẩn tắc giữa 0 & 1,83.

Để tìm diện tích dưới đường cong chuẩn tắc giữa 0 và giá trị z âm, chúng ta áp dụng tính chất đối xứng qua 0 của đường cong chuẩn tắc. Chẳng hạn, diện tích của đường cong chuẩn tắc giữa -1,83 và 0 cũng bằng với diện tích dưới đường cong chuẩn tắc giữa 0 và 1,83; nghĩa là bằng 0,4664.

Ví dụ 2: Tìm diện tích dưới đường cong chuẩn tắc về bên phải của 1,25.

Vì tổng diện tích dưới đường cong bằng 1 và đường cong đối xứng qua 0, nên diện tích dưới đường cong về bên phải của z = 0 là 0,5. Tìm trong Phụ lục II diện tích giữa z = 0 và z = 1,25 là 0,3944. Do diện tích về bên phải của z = 0 là 0,5, nên diện tích về bên phải của z = 1,25 sẽ bằng 0,5000 - 0,3944 = 0,1056 (Hình 2.9).

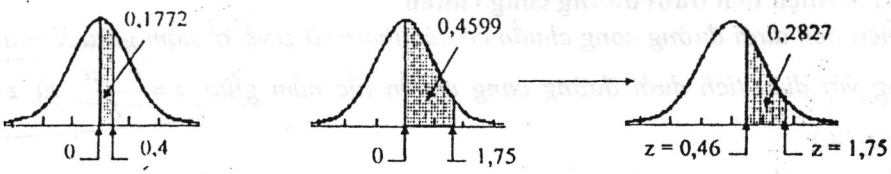


Hình 2.9: Diện tích dưới đường cong chuẩn tắc về bên phải của 1,25

Ví dụ 3: Tìm diện tích dưới đường cong chuẩn tắc giữa 0,46 và 1,75.

Ở Phụ lục II cho diện tích giữa z = 0 và z = 0,46 là 0,1772 và diện tích giữa z = 0 và z = 1,75 là 0,4599 (Hình 2.10). Như vậy, diện tích giữa z = 0,46 và z = 1,75 sẽ bằng hiệu của hai diện tích: z = 0 và z = 0,46 và z = 0 và z = 1,75, nghĩa là:

Diện tích cần tìm = 0,4599 - 0,1772 = 0,2827

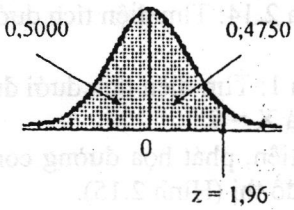


Hình 2.10: Diện tích dưới đường cong chuẩn tắc giữa 0,46 và 1,7

Ví dụ 4: Tìm diện tích dưới đường cong chuẩn tắc về bên trái của 1,96.

Trong Phụ lục II cho diện tích giữa z = 0 và z = 1,96 là 0,4750. Vì diện tích về bên trái của z = 0 là 0,5, nên diện tích về bên trái của z = 1,96 sẽ bằng 0,4750 + 0,5000 = 0,9750 (Hình 2.11).

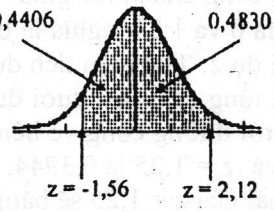
Hình 2.11: Diện tích dưới đường cong chuẩn tắc về bên trái của 1,96



Ví dụ 5: Tìm diện tích dưới đường cong chuẩn tắc giữa -1,56 và 2,12.

Để xác định diện tích cần tìm, chúng ta cộng diện tích giữa z = -1,56 và z = 0 với diện tích giữa z = 0 và z = 2,12. Do tính chất đối xứng nên diện tích giữa z = -1,56 và

z = 0 cũng bằng với diện tích giữa z = 0 và z = 1,56 và bằng 0,4406. Diện tích giữa z = 0 và z = 2,12 là 0,4830. Như vậy, diện tích giữa z = -1,56 và z = 2,12 là 0,4406 + 0,4830 = 0,9236 (Hình 2.12).

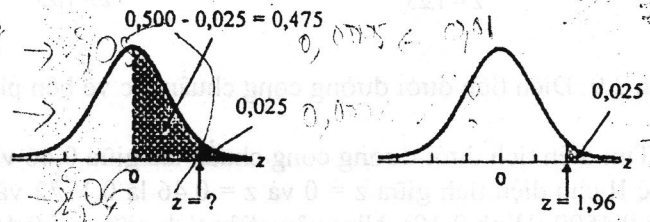


Hình 2.12: Diện tích dưới đường cong chuẩn tắc giữa 1,56 và 2,12

6. Xác định giá trị của biến chuẩn tắc z

Ví dụ: Xác định giá trị z để diện tích dưới đường cong chuẩn tắc về bên phải của giá trị này bằng 0,025 (Hình 2.13).

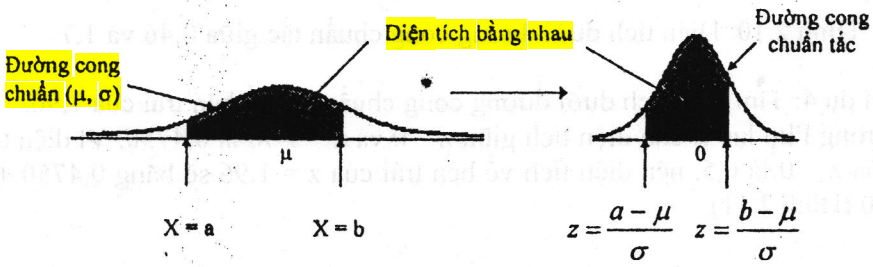
Diện tích giữa z = 0 và giá trị z cần tìm là 0,500 - 0,025 = 0,475. Tra trong Phụ lục II, chúng ta tìm được giá trị z tương ứng với diện tích 0,475 là 1,96.



Hình 2.13: Giá trị z để diện tích dưới đường cong chuẩn tắc về bên phải của giá trị này bằng 0,025

7.Tìm diện tích dưới đường cong chuẩn

Diện tích dưới đường cong chuẩn có các tham số μ và σ² nằm giữa X = a và X = b bằng với diện tích dưới đường cong chuẩn tắc nằm giữa $z = \frac{a-\mu}{\sigma}$ *và* $z = \frac{b-\mu}{\sigma}$ *(Hình 2.14.)*



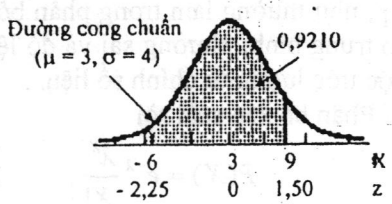
Hình 2.14: Tìm diện tích dưới đường cong chuẩn có các tham số μ và σ²

Ví dụ 1: Tìm diện tích dưới đường cong chuẩn có tham số μ = 3 và σ = 4 nằm giữa X = - 6 và X = 9.

Đầu tiên, phát họa đường cong chuẩn có tham số μ = 3 và σ = 4. Xác định diện tích trên đồ thị (Hình 2.15).

Tính z:		Diện tích giữa 0 và z:
X = -6 ⇒	z = (-6 - 3)/4 = -2,25	0,4878
X = 9 ⇒	z = (9 - 3)/4 = 1,5	0,4332

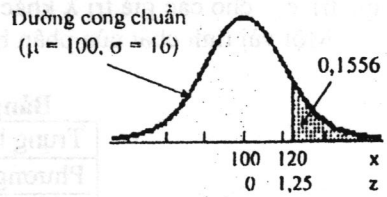
Diện tích cần tìm là 0,4878 + 0,4332 = 0,9210



Hình 2.15: Diện tích dưới đường cong chuẩn có tham số μ = 3 và σ = 4 ở giữa X = - 6 và X = 9

Ví dụ 2: Tìm diện tích dưới đường cong chuẩn có các tham số μ = 100 và σ = 16 về phía bên phải của 120.

Hình 2.16 xác định được diện tích cần tìm là 0,1056



Hình 2.16: Diện tích dưới đường cong chuẩn có μ = 100 và σ = 16 về phía bên phải của 120.

Tính z:

X = 120 ⇒ z = (120 - 100)/16 = 1,25

Diện tích giữa 0 và z:

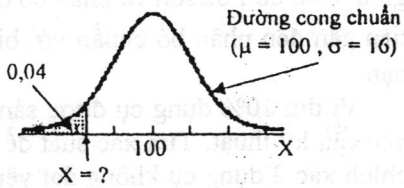
0,3944

Diện tích cần tìm là 0,5000 - 0,3944 = 0,1056

8. Xác định giá trị của biến chuẩn X

Ví dụ: Với đường cong chuẩn có tham số μ = 100 và σ = 16, tìm giá trị X có diện tích bằng 0,04 về bên trái.

Trước tiên, phát họa đường cong chuẩn có tham số μ = 100 và σ = 16 (Hình 2.17). Trên đồ thị, xác định diện tích đã định và giá trị X cần tìm.



Hình 2.17: Giá trị X có diện tích bằng 0,04 về bên trái dưới đường cong chuẩn có tham số μ = 100 và σ = 16.

Tra trong Phụ lục II, giá trị z có diện tích bằng 0,04 về bên trái là z = - 1,75; nghĩa là, giá trị z có diện tích bằng 0,04 về bên trái dưới đường cong chuẩn tắc.

Dùng công thức z = (X - μ)/σ để tìm giá trị X:

z = - 1,75 ⇒ X = μ + z.σ = 100 + (- 1,75)(16) = 100 – (1,75x16) = 72

Do đó, đối với đường cong chuẩn có tham số μ = 100 và σ = 16, giá trị X có diện tích bằng 0,04 về bên trái là 72 (X = 72).

IV. PHÂN BỐ POISSON *(Poisson Distribution)*

Trong phân bố Poisson, chúng ta vẫn còn đề cập đến biến ngẫu nhiên mà chỉ có thể có một trong hai kết quả (thành công hoặc thất bại) có thể loại trừ lẫn nhau. Sự phân biệt giữa phân bố này và phân bố nhị thức là xác suất (p) của biến cố xảy ra rất

nhỏ hoặc biến cố hiếm xảy ra. Do đó, chúng ta không có số mong đợi lý thuyết cho giá trị p, như thường làm trong phân bố nhị thức, nên cũng không thể tính giá trị lý thuyết cho trung bình, phương sai và độ lệch chuẩn. Như vậy, trung bình và phương sai phải được ước lượng từ chính số liệu. .

Phân bố xác suất rời

$$P(X) = e^{-\lambda}\frac{\lambda^X}{X!} \qquad X = 0, 1, 2,.... \qquad (3)$$

với e = 2,71828... và λ là một hằng số đã cho, được gọi là phân bố Poisson theo Siméon-Denis Poisson, người đã tìm ra phân bố này vào đầu thế kỷ thứ 19. Các giá trị của p(X) có thể được tính bằng cách sử dụng bảng trong Phụ lục III (Bảng này ghi các giá trị $e^{-\lambda}$ cho các giá trị λ khác nhau) hoặc bằng cách lấy logarith.

Một vài tính chất của phân bố Poisson được liệt kê trong Bảng 2.4:

Bảng 2.4: Phân bố Poisson

Trung bình	$\mu = \lambda$
Phương sai	$\sigma^2 = \lambda$
Độ lệch chuẩn	$\sigma = \sqrt{\lambda}$

Sự liên hệ giữa các phân bố nhị thức và Poisson

Ở phân bố nhị thức (1), nếu N lớn trong khi xác suất (p) xuất hiện của một biến cố nào đó gần với 0, để q = 1 – p gần với 1, biến cố đó được gọi là biến cố hiếm. Trong thực hành, chúng sẽ khảo sát một biến cố hiếm nếu số lần thử nghiệm ít nhất 50 (N ≥ 50) trong khi Np nhỏ hơn 5 (Np <5). Trong trường hợp như thế phân bố nhị thức (1) xấp xỉ rất gần với phân bố Poisson (3) với λ = Np. Điều này có thể nhận thấy qua việc so sánh các Bảng 2.3 và 2.4 bằng cách đặt λ = Np, q ≈ 1 và p ≈ 0 trong Bảng 2.3, chúng ta được các kết quả trong Bảng 2.4.

Do có sự liên hệ giữa phân bố nhị thức và phân bố Poisson, nên cũng có sự liên hệ giữa phân bố Poisson và phân bố chuẩn. Thực ra, chúng ta có thể thấy phân bố Poisson tiến gần đến phân bố chuẩn với biến chuẩn tắc hóa $z = (X - \lambda)/\sqrt{\lambda}$ khi λ gia tăng vô hạn.

Ví dụ: 10% dụng cụ được sản xuất trong một quy trình chế tạo nào đó không đạt yêu cầu kỹ thuật. Tìm xác suất để một mẫu gồm 10 dụng cụ được chọn ngẫu nhiên sẽ chính xác 2 dụng cụ không đạt yêu cầu bằng cách sử dụng (a) phân bố nhị thức và (b) phân bố Poisson xấp xỉ với phân bố nhị thức.

Giải:

Gọi p là xác suất của dụng cụ không đạt yêu cầu (p = 0,1)

X là số dụng cụ không đạt yêu cầu trong số 10 dụng cụ được chọn ngẫu nhiên

(a) $p(X = 2) = C_{10}^2.(0,1)^2.(0,9)^8 = 0,1937$ hoặc 0,19

(b) Với λ = Np = 10.(0,1) = 1 và sử dụng e = 2,718

$$P(X = 2) = e^{-\lambda}\frac{\lambda^X}{X!} = \frac{(1)^2 e^{-1}}{2!} = \frac{e^{-1}}{2} = \frac{1}{2e} = 0,1839 \text{ hoặc } 0,18$$

Nhìn chung, sự xấp xỉ này khá tốt nếu p ≤ 0,1 và λ = Np ≤ 5

V. PHÂN BỐ CỦA CÁC SỐ TRUNG BÌNH LẤY MẪU

(Sampling distribution of means)

1. Trung bình, phương sai và độ lệch chuẩn của biến ngẫu nhiên $\overline{X}$

Khi lấy mẫu từ tổng thể, trung bình mẫu ($\overline{X}$) là một biến ngẫu nhiên vì giá trị của nó phụ thuộc vào mẫu được rút ngẫu nhiên. Phân bố xác suất của biến ngẫu nhiên $\overline{X}$ được gọi là phân bố của số trung bình lấy mẫu.

Ví dụ: Một tổng thể gồm bốn số hạng: 2, 3, 4 và 7. Từ tổng thể này có thể rút ra tất cả các mẫu có thể với hai số hạng. Hãy tính (a) trung bình và độ lệch chuẩn của tổng thể (b) trung bình và độ lệch chuẩn của các số trung bình mẫu hay còn gọi là sai số chuẩn *(standard error)*.

Giải:

a) Trung bình, phương sai và độ lệch chuẩn của tổng thể (μ):

$$\mu = \frac{2+3+4+7}{4} = 4$$

$$\sigma^2 = \frac{(2-4)^2+(3-4)^2+(4-4)^2+(7-4)^2}{4} = \frac{4+1+0+9}{4} = 3{,}5$$

$$\sigma = 1{,}87$$

b) Trung bình, phương sai và độ lệch chuẩn của các số trung bình mẫu ($\overline{X}$):

Để có thể rút ra tất cả các mẫu có hai số hạng từ tổng thể, có hai cách lấy mẫu:

(1) Rút ra có hoàn trả lại: Với cách lấy mẫu này sẽ có $4^2 = 16$ mẫu có cỡ mẫu bằng hai (vì bất cứ một trong bốn số ở lần rút đầu tiên có thể kết hợp với bất cứ một trong bốn số ở lần rút thứ hai)

(2, 2)	(2, 3)	(2, 4)	(2, 7)
(3, 2)	(3, 3)	(3, 4)	(3, 7)
(4, 2)	(4, 3)	(4, 4)	(4, 7)
(7, 2)	(7, 3)	(7, 4)	(7, 7)

Trung bình ($\overline{X}$) của 16 mẫu tương ứng là:

2,0	2,5	3,0	4,5
2,5	3,0	3,5	5,0
3,0	3,5	4,0	5,5
4,5	5,0	5,5	7,0

$$\mu_{\overline{X}} = \frac{2{,}0+2{,}5+\cdots+7{,}0}{16} = \frac{64}{16} = 4$$

Như vậy: $\mu_{\overline{X}} = \mu$

$$\sigma^2_{\overline{X}} = \frac{(2{,}0-4)^2+(2{,}5-4)^2+\cdots+(7{,}0-4)^2}{16} = \frac{28}{16} = 1{,}75$$

và $\sigma_{\overline{X}} = \sqrt{1{,}75} = 1{,}32$

Như vậy: $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1{,}87}{\sqrt{2}} = 1{,}32$ (bằng với giá trị trên)

(2) Rút ra không hoàn trả lại: Với cách này sẽ có $C_4^2 = 6$ mẫu có cỡ mẫu bằng hai (nghĩa là, đầu tiên rút một số, sau đó rút một số khác với số đầu). Đó là (2, 3), (2, 4),

(2, 7), (3, 4), (3, 7) và (4, 7). Giá trị trung bình mẫu tương ứng của các mẫu sẽ là 2,5; 3,0; 4,5; 3,5; 5,0 và 5,5. Lúc đó:

$$\mu_{\overline{X}} = \frac{2,5+3,0+4,5+3,5+5,0+5,5}{6} = 4$$

Như vậy: $\mu_{\overline{X}} = \mu$

$$\sigma^2_{\overline{X}} = \frac{(2,5-4)^2+(3,0-4)^2+(4,5-4)^2+(3,5-4)^2+(5,0-4)^2+(5,5-4)^2}{6} = 1,17$$

và $\sigma_{\overline{X}} = \sqrt{1,17} = 1,08$

Như vậy:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{1,87}{\sqrt{2}}\sqrt{\frac{4-2}{4-1}} = 1,08 \text{ (bằng với giá trị trên)}$$

Chú ý: Khi cỡ mẫu gia tăng, các trung bình mẫu hợp thành một nhóm xung quanh giá trị trung bình của tổng thể. Do đó, cỡ mẫu càng lớn, sai số do lấy mẫu càng nhỏ, trung bình mẫu ($\overline{X}$) có khuynh hướng là ước lượng của trung bình tổng thể (μ).

2 Phân bố của biến $\overline{X}$ được rút ra từ tổng thể có phân bố chuẩn

Giả sử một mẫu ngẫu nhiên có cỡ mẫu n được rút ra từ một tổng thể có phân bố chuẩn với trung bình μ và độ lệch chuẩn σ. Lúc đó biến ngẫu nhiên $\overline{X}$ cũng có phân bố chuẩn với trung bình

$$\mu_{\overline{X}} = \mu$$

và độ lệch chuẩn

$$\sigma_{\overline{X}} = \sigma / \sqrt{n}$$

Nghĩa là, mẫu được rút ra từ tổng thể có phân bố chuẩn; lúc đó xác suất của $\overline{X}$ bằng với các diện tích dưới đường cong chuẩn có các tham số μ và $\sigma / \sqrt{n}$.

Như vậy, khi tổng thể có phân bố chuẩn, phân bố của các trung bình mẫu cũng có phân bố chuẩn ngay cả cỡ mẫu nhỏ (n < 30).

3. Phân bố của biến $\overline{X}$ được rút ra từ tổng thể có phân bố bất kỳ

Định lý giới hạn trung tâm *(The central limit theorem)*

Dù biến ngẫu nhiên X có phân bố như thế nào, phân bố của trung bình mẫu ($\overline{X}$) càng tiến gần đến phân bố chuẩn khi n càng lớn.

Giả sử một mẫu ngẫu nhiên với cỡ mẫu **n ≥ 30** được rút ra từ một tổng thể có trung bình μ và độ lệch chuẩn σ. Lúc đó, bất chấp phân bố của tổng thể, biến ngẫu nhiên $\overline{X}$ có phân bố xấp xỉ chuẩn với trung bình

$$\mu_{\overline{X}} = \mu$$

và độ lệch chuẩn

$$\sigma_{\overline{X}} = \sigma / \sqrt{n}$$

Nghĩa là, nếu n ≥ 30, lúc đó xác suất của $\overline{X}$ gần bằng với các diện tích dưới đường cong chuẩn có các tham số μ và $\sigma / \sqrt{n}$.

4. Biến chuẩn tắc hóa của biến ngẫu nhiên $\overline{X}$

Giả sử một mẫu ngẫu nhiên có cỡ mẫu n được rút ra từ một tổng thể có trung bình μ và độ lệch chuẩn σ. Lúc đó biến chuẩn tắc hóa (z) của $\overline{X}$

$$z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

(a) chính xác phân bố chuẩn tắc nếu tổng thể có phân bố chuẩn, bất chấp cỡ mẫu

(b) có phân bố gần chuẩn tắc nếu n ≥ 30, bất chấp tổng thể có phân bố như thế nào

Nói cách khác, nếu tổng thể có phân bố chuẩn hoặc cỡ mẫu ít nhất bằng 30, lúc đó xác suất của biến ngẫu nhiên

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

bằng hoặc ít nhất cũng gần bằng với các diện tích dưới đường cong chuẩn tắc.

Ví dụ: Một quần thể gồm 1500 con gà mới nở có phân bố chuẩn với trung bình là 22,40 g/con và độ lệch chuẩn là 0,048 g/con. Nếu 300 mẫu ngẫu nhiên có cỡ mẫu bằng 36 được rút ra từ quần thể này. Hãy cho biết có bao nhiêu mẫu có trung bình (a) giữa 22,39g/con và 22,41g/con (b) nhỏ hơn 22,38g/con hoặc lớn hơn 22,41g/con.

Trong ví dụ, biến chuẩn tắc z của biến $\overline{X}$ là:

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{\overline{X} - 22{,}40}{0{,}048/\sqrt{36}} = \frac{\overline{X} - 22{,}40}{0{,}008}$$

a) $p(22{,}39 \le \overline{X} \le 22{,}41) = p\left(\frac{22{,}39 - 22{,}40}{0{,}008} \le z \le \frac{22{,}41 - 22{,}40}{0{,}008}\right)$

$= p(-1{,}25 \le z \le 1{,}25)$

$= 2.p(0 \le z \le 1{,}25)$

$= 2.(0{,}3944) = 0{,}7888$

Do đó, số mẫu hy vọng có trung bình giữa 22,39g/con và 22,41g/con là: (300)(0,7888) =236,64 hoặc 237 mẫu

b) $p(\overline{X} \le 22{,}38$ hoặc $\overline{X} \ge 22{,}41) = p\left(z \le \frac{22{,}38 - 22{,}40}{0{,}008}\right.$ hoặc $\left. z \ge \frac{22{,}41 - 22{,}40}{0{,}008}\right)$

$= p(z \le -2{,}5$ hoặc $z \ge 1{,}25)$

$= [0{,}5 - p(0 \le z \le 2{,}5)] + [0{,}5 - p(0 \le z \le 1{,}25)]$

$= (0{,}5 - 0{,}4938) + (0{,}5 - 0{,}3944)$

$= 0{,}0062 + 0{,}1056 = 0{,}1118$

Do đó, số mẫu hy vọng có trung bình nhỏ hơn 22,38g/con và lớn hơn 22,41g/con là: (300)(0,1118) = 33,54 hoặc 34 mẫu

VI. PHÂN BỐ CỦA CÁC SỐ TỶ LỆ LẤY MẪU

(Sampling distribution of proportions)

Giả sử tổng thể vô hạn và xác suất để biến cố A xảy ra bằng p trong khi xác suất để biến cố A không xảy ra bằng q = 1 - p. Nếu có thể rút ra từ tổng thể này tất cả các mẫu có cỡ mẫu bằng n và mỗi mẫu xác định tỷ lệ $\overline{p}$ thành công; lúc đó biến ngẫu nhiên $\overline{p}$ xấp xỉ phân bố chuẩn và có trung bình bằng p và độ lệch chuẩn bằng $\sqrt{\frac{pq}{n}}$

$$\mu_p = p \quad \text{và} \quad \sigma_p = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p \cdot (1-p)}{n}}$$

Nghĩa là xác suất để $\bar{p}$ xấp xỉ bằng diện tích dưới đường cong chuẩn với các tham số p và $\sqrt{p(1-p)/n}$.

Biến chuẩn tắc hóa của $\bar{p}$:

$$z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Tính xấp xỉ phân bố nhị thức bằng phân bố chuẩn

Giả sử X là biến ngẫu nhiên rời rạc có phân bố nhị thức với tham số n và p, lúc đó X có phân bố xấp xỉ chuẩn với giá trị trung bình μ = np và độ lệch chuẩn $\sigma = \sqrt{np(1-p)}$. Biến chuẩn tắc z trở thành

$$z = \frac{X - np}{\sqrt{np(1-p)}}$$

với X là số lần thành công trong một mẫu.

Phân bố nhị thức càng gần phân bố chuẩn khi n càng lớn. Trong thực tế, sự xấp xỉ rất tốt nếu np ≥ 5 và n(1 - p) ≥ 5. Để cho sự xấp xỉ được chính xác hơn, cần có sự hiệu chỉnh liên tục. Cụ thể:

Nếu a là số nguyên dương thì

$$p(X > a) = p\left(z \geq \frac{(a+0{,}5)-\mu}{\sigma}\right)$$

$$p(X < a) = p\left(z \leq \frac{(a-0{,}5)-\mu}{\sigma}\right)$$

$$p(X \geq a) = p\left(z \geq \frac{(a-0{,}5)-\mu}{\sigma}\right)$$

$$p(X \leq a) = p\left(z \leq \frac{(a+0{,}5)-\mu}{\sigma}\right)$$

$$p(X = a) = p\left(\frac{(a-0{,}5)-\mu}{\sigma} \leq z \leq \frac{(a+0{,}5)-\mu}{\sigma}\right)$$

Ví dụ: Một cuộc điều tra cho thấy có 32,6% phụ nữ trong một vùng nào đó hút thuốc lá. Chọn ngẫu nhiên 300 phụ nữ trong vùng này, hãy tính xác suất để có số phụ nữ hút thuốc lá lớn hơn hoặc bằng 100 và nhỏ hơn hoặc bằng 150.

Giải:

Gọi X là số phụ nữ hút thuốc lá trong số 300 phụ nữ. X có phân bố nhị thức với n = 300, p = 0,326.

Vì np = (300)(0,326) = 97,8 ≥ 5

n(1-p) = (300)(1 - 0,326) = (300)(0,674) = 202,2 ≥ 5

nên thỏa điều kiện của phân bố nhị thức xấp xỉ phân bố chuẩn với trung bình,

$$\mu = np = (300).(0{,}326) = 97{,}8$$

và độ lệch chuẩn, $\sigma = \sqrt{np(1-p)} = \sqrt{(300)(0{,}326)(0{,}674)} = 8{,}119$

$$p(100 \leq X \leq 150) = p\left(\frac{(100-0{,}5)-97{,}8}{8{,}119} < z < \frac{(150+0{,}5)-97{,}8}{8{,}119}\right)$$

$$= p(0{,}21 < z < 6{,}49)$$

$$= p(0 < z < 6{,}49) - p(0 < z < 0{,}21)$$

$$= 0{,}5 - 0{,}0832 = 0{,}4168$$

VII. PHÂN BỐ t CỦA STUDENT *(Student's t distribution)*

Trong hầu hết các thí nghiệm, độ lệch chuẩn của tổng thể thường không biết. Do đó, khi cỡ mẫu **n ≥ 30**, độ lệch chuẩn của tổng thể có thể được ước lượng bằng độ lệch chuẩn của mẫu, và lúc đó có thể dùng phân bố chuẩn tắc để kết luận. Nếu cỡ mẫu **n < 30**, phương pháp này sẽ không cho xác suất đáng tin cậy. Do đó, chúng ta cần nhận dạng phân bố xác suất của biến ngẫu nhiên có được bằng cách thay thế độ lệch chuẩn không biết ở phương trình

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

bằng độ lệch chuẩn của mẫu; nghĩa là, chúng ta cần nhận dạng phân bố xác suất của biến ngẫu nhiên

$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

khi mẫu được lấy từ tổng thể có phân bố chuẩn.

Mặc dầu biến ngẫu nhiên z và t có cùng dạng nhưng phân bố xác suất của chúng thì khác nhau (các giá trị z ít biến động hơn các giá trị t). Biến ngẫu nhiên z có phân bố chuẩn tắc, trong khi biến ngẫu nhiên t không có phân bố chuẩn tắc.

1. Hàm mật độ của phân bố t

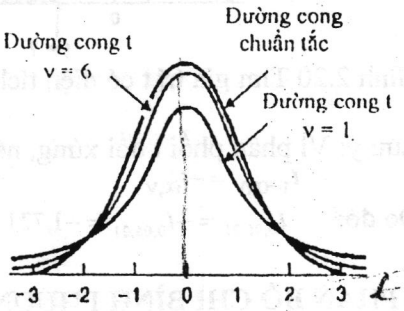
William Sealy Gosset (1908) đã tìm ra phân bố xác suất của biến ngẫu nhiên t và đặt tên là phân bố Student hoặc phân bố t. Hàm mật độ của phân bố t có dạng:

$$f(t) = \frac{C}{\left(1 + \frac{t^2}{n-1}\right)^{\frac{n}{2}}} \quad \text{hoặc} \quad f(t) = \frac{C}{\left(1 + \frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}}$$

trong đó C là một hằng số phụ thuộc vào ν

ν là số độ tự do (n – 1).

Đường cong t phụ thuộc vào độ tự do; nghĩa là, ứng với mỗi độ tự do sẽ có một đường cong t khác nhau, nhưng tất cả các đường cong t đều giống với đường cong chuẩn tắc, cũng đối xứng qua 0 (Hình 2.18).



Hình 2.18: Đường cong chuẩn tắc và hai đường cong t

Khi độ tự do gia tăng thì đường cong t sẽ tiến gần đến đường cong chuẩn tắc. Nói cách khác, đường cong chuẩn tắc là trường hợp đặc biệt của đường cong t khi độ tự do tiến đến vô cực.

Như chúng ta biết, Xác suất của một biến ngẫu nhiên có phân bố chuẩn sẽ bằng với diện tích dưới đường cong chuẩn. Như vậy, xác suất của một biến ngẫu nhiên có phân bố t cũng bằng với diện tích dưới đường cong t.

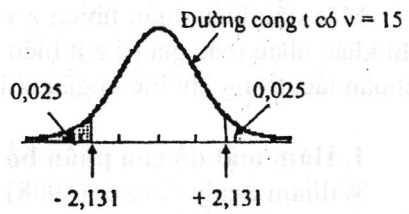
2. Sử dụng bảng t

Ở bảng phân bố t hai đuôi (Phụ lục IV), hàng trên cùng là các tổng diện tích (α) của hai đuôi; nghĩa là, diện tích ở mỗi đuôi tính từ giá trị ± t về bên phải và bên trái bằng $\alpha/2$. Cột đầu tiên ghi số độ tự do, bên trong bảng là các giá trị t. Như vậy, ứng với mỗi độ tự do sẽ tìm được hai giá trị ± t có diện tích của một trong hai đuôi bằng $\alpha/2$.

3. Tìm diện tích dưới đường cong t

Ví dụ 1: Cho đường cong t có 15 độ tự do, hãy xác định hai giá trị t có diện tích tính từ giá trị này về hai phía (phải và trái) bằng 0,025 ($t_{0,025;15}$).

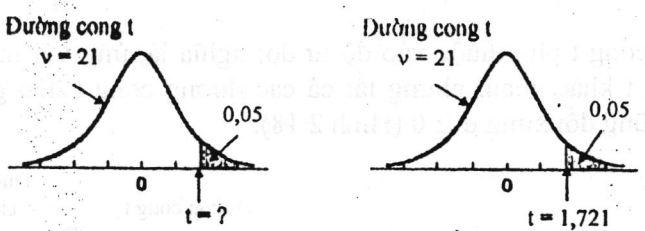
Tìm trong bảng t (Phụ lục IV) ở độ tự do bằng 15 và α = 0,05, các giá trị t tìm được là ± 2,131 (Hình 2.19).



Hình 2.19: Hai giá trị t có diện tích tính từ hai giá trị này về mỗi phía bằng 0,025

Ví dụ 2: Cho đường cong t có 21 độ tự do, hãy xác định giá trị t có diện tích bằng 0,05 ($t_{0,05}$) về bên phải.

Để tìm giá trị t có diện tích về bên phải bằng 0,05; vì sử dụng bảng t hai đuôi trong Phụ lục IV nên ta phải nhân diện tích α cho 2. Lúc này tổng diện tích hai phía sẽ bằng 0,1 (0,05 x 2). Giá trị t tìm được trong bảng ở độ tự do = 21 và α = 0,1 là 1,721 (Hình 2.20).



Hình 2.20 Tìm giá trị t có diện tích từ giá trị này về bên phải bằng 0,05

Lưu ý: Vì phân phối t đối xứng, nên các giá trị ở đuôi nhỏ cũng bằng ở đuôi lớn:

$$t_{1-\alpha,v} = -t_{\alpha,v}$$

Do đó: $t_{0,95;21} = -t_{0,05;21} = -1,721$

VIII. PHÂN BỐ CHI BÌNH PHƯƠNG (χ^2)

1. Hàm mật độ của phân bố χ^2

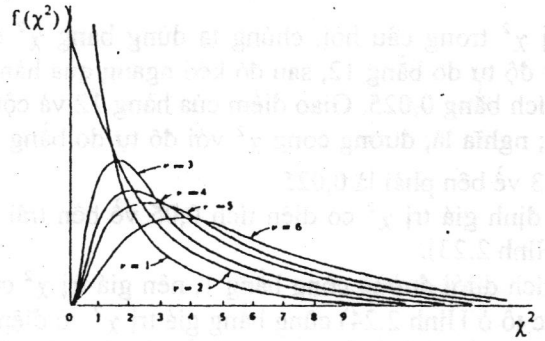
Năm 1876, Frederick R. Helmert đã có một số nghiên cứu đầu tiên về phân bố lý thuyết chi bình phương, sau đó phân bố χ^2 đã được Karl Pearson soạn thảo từ đầu thế

kỹ 20. Phân bố đó có thể được hình dung qua đồ thị của các hàm mật độ xác suất như sau:

$$f(\chi^2) = C(\chi^2)^{\frac{\nu}{2}-1} e^{-\frac{\chi^2}{2}} = C(\chi)^{\nu-2} e^{-\frac{\chi^2}{2}}$$

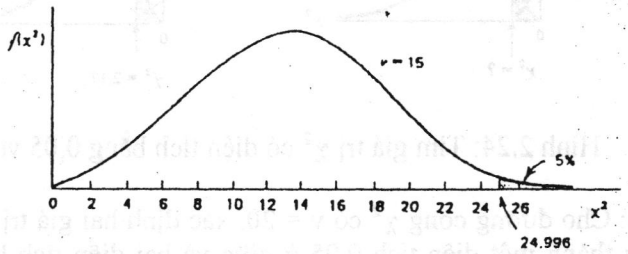
với χ^2 biến thiên từ 0 đến + ∞, ν là số độ tự do (ν), C là một hằng số phụ thuộc vào ν.

Như vậy, sẽ có vô số đường cong χ^2 mà chúng thay đổi theo độ tự do (Hình 2.21).



Hình 2.21: Đường cong chi bình phương thay đổi theo độ tự do.

Giá trị tối đa của $f(\chi^2)$ xảy ra ở $\chi^2 = \nu - 2$ với $\nu \geq 2$ (Hình 2.22)



Hình 2.22: Đường cong χ^2 ở độ tự do, ν = 15 có giá trị tối đa ở $\chi^2 = 13$

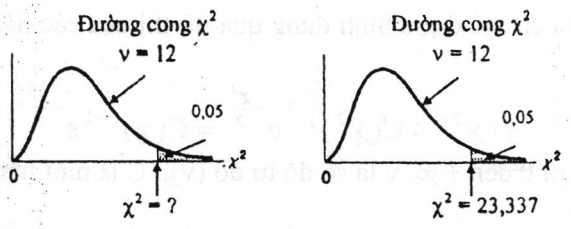
Đường cong χ^2 không đối xứng mà bị lệch về bên phải. Nó bắt đầu ở điểm 0 trên trục hoành và kéo dài vô hạn về bên phải tiệm cận với trục hoành. Khi độ tự do càng tăng các đường cong χ^2 càng trông giống như đường cong chuẩn.

2. Sử dụng bảng χ^2

Trong bảng χ^2 (Phụ lục V), cột đầu tiên chỉ độ tự do. Hàng trên cùng là diện tích α tính từ giá trị χ^2 trở về phía bên phải của đường cong χ^2. Như vậy, ứng với mỗi độ tự do sẽ tìm được giá trị χ^2 có diện tích tính từ giá trị này về bên phải bằng α.

3. Tìm giá trị χ^2 cho một diện tích đã định

Ví dụ 1: Cho đường cong χ^2 có 12 độ tự do, tìm $\chi^2_{0,025}$; nghĩa là, tìm giá trị χ^2 có diện tích tính từ giá trị này về bên phải bằng 0,025 (Hình 2.23)

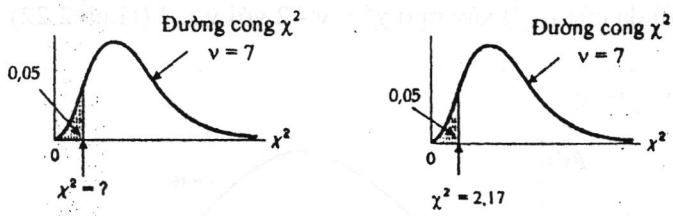


Hình 2.23: Tìm giá trị χ^2 có diện tích 0,025 về bên phải

Để tìm giá trị χ^2 trong câu hỏi, chúng ta dùng bảng χ^2 ở Phụ lục V. Đầu tiên chúng ta tìm ở cột độ tự do bằng 12, sau đó kéo ngang qua hàng cho đến khi xác định được cột có diện tích bằng 0,025. Giao điểm của hàng 12 và cột 0,025 là giá trị χ^2 cần tìm ($\chi^2_{0,025}$ = 23,3); nghĩa là, đường cong χ^2 với độ tự do bằng 12, có diện tích tính từ giá trị χ^2 bằng 23,3 về bên phải là 0,025 .

Ví dụ 2: Xác định giá trị χ^2 có diện tích 0,05 về bên trái của đường cong χ^2 có độ tự do bằng 7 (Hình 2.23).

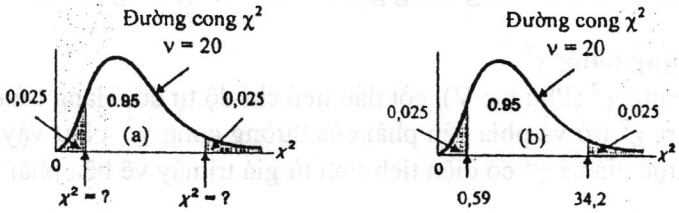
Vì tổng diện tích dưới đường cong bằng 1, nên giá trị χ^2 có diện tích 0,05 về bên trái (diện tích được tô ở Hình 2.24) cũng bằng giá trị χ^2 có diện tích 1 - 0,05 = 0,95 về bên phải. Điều này muốn nói rằng giá trị χ^2 cần tìm là $\chi^2_{0,95}$. Từ Phụ lục V, chúng ta tìm ở ν = 7, giá trị χ^2 có diện tích bằng 0,95 về bên phải là 2,17.



Hình 2.24: Tìm giá trị χ^2 có diện tích bằng 0,05 về bên trái

Ví dụ 3: Cho đường cong χ^2 có ν = 20, xác định hai giá trị χ^2 chia diện tích dưới đường cong thành một diện tích 0,95 ở giữa và hai diện tích hai bên là 0,025 (Hình 2.25a).

Đầu tiên, tìm giá trị χ^2 về bên phải ở Hình 2.25a. Từ hình chúng ta thấy diện tích được tô ở bên phải là 0,025. Tìm trong bảng χ^2 ở ν = 20, giá trị $\chi^2_{0,025}$ về bên phải là 34,2.



Hình 2.25: Hai giá trị χ^2 chia diện tích dưới đường cong thành một diện tích 0,95 ở giữa và hai diện tích 0,025 ở hai bên

Kế tiếp, chúng ta tìm giá trị χ^2 ở bên trái của Hình 2.25a. Vì diện tích ở bên trái của giá trị χ^2 cần tìm là 0,025 tương ứng với diện tích ở bên phải là 1 - 0,025 = 0,975. Do đó, ở $\nu = 20$, giá trị $\chi^2_{0,975}$ về bên phải là 9,59.

Như thế, đối với đường cong χ^2 có $\nu = 20$, hai giá trị χ^2 chia diện tích dưới đường cong thành một diện tích 0,95 ở giữa và hai diện tích 0,025 ở hai bên là 9,59 và 34,2 (Hình 2.25b).

IX. PHÂN BỐ F CỦA FISHER *(Fisher's F distribution)*

1. Hàm mật độ của phân bố F

Biến ngẫu nhiên F được gọi là có phân bố F với (ν_1, ν_2) độ tự do, nếu hàm mật độ của nó có dạng:

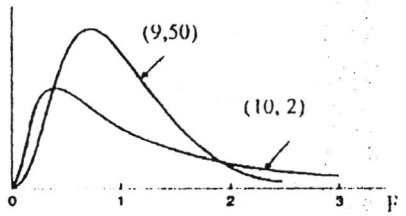
$$f(F) = \frac{CF^{\frac{\nu_1}{2}-1}}{(\nu_1 F + \nu_2)^{(\nu_1+\nu_2)/2}} \qquad \text{nếu } X \geq 0$$

với C là hằng số dương phụ thuộc vào ν_1 và ν_2

ν_1 và ν_2 lần lượt là độ tự do của tử số $(n_1 - 1)$ và mẫu số $(n_2 - 1)$.

Xác suất của một biến ngẫu nhiên có phân phối F sẽ bằng các diện tích dưới đường cong F. Đường cong F phụ thuộc vào một cặp độ tự do, (ν_1, ν_2), ν_1 tương ứng với độ tự do của tử số và ν_2 tương ứng với độ tự do của mẫu số (Hình 2.26).

Hình 2.26: Hai đường cong F khác nhau.



Đường cong F không đối xứng mà lệch về bên phải. Nó bắt đầu ở 0 trên trục hoành và kéo dài vô hạn về bên phải tiệm cận với trục hoành.

Nếu α là diện tích dưới đường cong F về bên phải của trị số F_{α,ν_1,ν_2}, lúc đó:

$$F_{\alpha,\nu_1,\nu_2} = 1/F_{1-\alpha,\nu_2,\nu_1}$$

Phân bố F có quan hệ với phân bố t:

$$F_{\alpha,1,\nu_2} = (t_{\alpha/2,\nu_2})^2$$

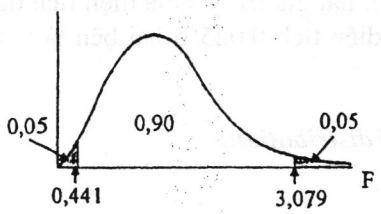
Snedecor đã lập những bảng số cho các giá trị tới hạn của F ở hai ngưỡng xác suất thường dùng: $\alpha = 0,05$ và $\alpha = 0,01$ (Phụ lục VI). Nếu giá trị F tính lớn hơn giá trị F tới hạn ở mức α, chúng ta sẽ kết luận hai phương sai khác biệt có ý nghĩa ở mức α đã định.

2. Cách sử dụng bảng F

Ở bảng F (Phụ lục VI), hàng trên cùng ghi độ tự do của tử số và cột đầu tiên ghi độ tự do của mẫu số. Các giá trị bên trong bảng là các giá trị F. Ứng với một cặp độ tự do, chúng ta sẽ tìm được các giá trị F có diện tích tính từ các giá trị này về bên phải của đường cong bằng 0,05 (in lợt) và 0,01 (in đậm).

3. Tìm diện tích dưới đường cong F

Ví dụ: Cho đường cong F có cộ độ tự do của tử số là 30 và của mẫu số là 8, Hãy xác định hai giá trị F có diện tích về bên phải và về bên trái bằng 0,05 và diện tích ở giữa là 0.90 (Hình 2.27).



Hình 2.27: Hai giá trị F có diện tích về bên phải và về bên trái bằng 0,05 và diện tích ở giữa là 0,90

Tìm trong bảng F (Phụ lục VI) ở diện tích bằng 0,05, độ tự do của tử số bằng 30 và của mẫu số bằng 8 và ngược lại, ta có:

$F_{0,05;30;8} = 3,079$ và $F_{0,05;8;30} = 2,266$

Như vậy, giá trị F có diện tích về bên phải bằng 0,05 là 3,076. Để tìm giá trị F có diện tích về bên trái bằng 0,05, chúng ta dùng công thức liên hệ

$$F_{\alpha,\nu_1,\nu_2} = 1/F_{1-\alpha,\nu_2,\nu_1}$$

Do đó: $F_{0,05;30,8} = 1/F_{0,95;8,30} = 1/2,266 = 0,441$

Vì thế, hai giá trị F cần tìm là 0,441 và 3,079 (Hình 2.27)

CHƯƠNG 3

KIỂM ĐỊNH GIẢ THUYẾT CHO MỘT TRUNG BÌNH HOẶC MỘT TỶ LỆ

I. KIỂM ĐỊNH GIẢ THUYẾT *(Tests of Hypothesis)*

1. Ý nghĩa của kiểm định giả thuyết

Chúng ta thường sử dụng thống kê suy luận để thực hiện việc quyết định về giá trị của một tham số, như trung bình hoặc phương sai của tổng thể dựa trên các thông tin có được từ mẫu. Chẳng hạn, chúng ta muốn kết luận năng suất trung bình của giống lúa mới khác với năng suất của giống lúa đang trồng trong sản suất là 5 tấn/ha, hoặc chúng ta muốn kết luận thuốc mới có hiệu quả hơn thuốc cũ về khả năng chống lại một bệnh nào đó,...

Một trong những phương pháp được sử dụng phổ biến nhất để dẫn đến kết luận như thế là thực hiện ***kiểm định giả thuyết***. Kiểm định giả thuyết, còn được gọi là kiểm định ý nghĩa *(Test of significance)*, là kiểm định tính đúng đắn của một giả thuyết. Giả thuyết thống kê là là một câu phát biểu dựa trên các tham số của một phân bố xác suất. Chẳng hạn, câu: "Năng suất trung bình của giống lúa mới bằng với năng suất của giống lúa đang sử dụng trong sản xuất là 5 tấn/ha" hoặc có thể phát biểu đơn giản hơn "μ = 5 tấn/ha" là một giả thuyết.

Một kiểm định giả thuyết tiêu biểu bao gồm hai giả thuyết: **giả thuyết không** *(null hypothesis)* và **giả thuyết chọn lựa** *(alternative hypothesis)* hay **giả thuyết nghiên cứu** *(research hypothesis)*.

*** Giả thuyết không** là giả thuyết được kiểm định. Đây là giả thuyết mà ta nghi ngờ và muốn bác bỏ. Chúng ta dùng ký hiệu H_0 để tiêu biểu cho giả thuyết không.

*** Giả thuyết chọn lựa** là giả thuyết được xem như là một sự chọn lựa đối với giả thuyết không; nghĩa là, giả thuyết sẽ được chấp nhận khi H_0 bị bác bỏ. Chúng ta thường dùng ký hiệu H_a để tiêu biểu cho giả thuyết chọn lựa (đôi khi còn được ký hiệu là H_1).

Cũng với ví dụ về giống lúa mới, giả thuyết không có thể là "Năng suất trung bình của giống lúa mới bằng với năng suất của giống lúa đang trồng trong sản xuất là 5 tấn/ha" và giả thuyết chọn lựa có thể là "Năng suất trung bình của giống lúa mới khác với năng suất của giống lúa đang trồng trong sản xuất là 5 tấn/ha"

Trước đây chữ "không" trong giả thuyết không được viết tắt cho "sự không khác biệt" hoặc "sự khác biệt bằng không". Tuy nhiên, ngày nay giả thuyết không có nghĩa đơn giản là giả thuyết được kiểm định.

Câu hỏi cần trả lời trong kiểm định giả thuyết là nên bác bỏ giả thuyết không và ủng hộ giả thuyết chọn lựa hoặc ngược lại.

Bước đầu trong việc thực hiện kiểm định giả thuyết là quyết định giả thuyết không và giả thuyết chọn lựa như thế nào?

Giả thuyết không: Luôn luôn nên ở dạng $\mu = \mu_0$, với μ_0 là một số nào đó. Nói khác đi, nên có dấu bằng (=) trong giả thuyết không. Do đó, chúng ta có thể diễn đạt giả thuyết không như sau:

$$H_0 : \mu = \mu_0$$

Giả thuyết chọn lựa: Chọn giả thuyết chọn lựa phụ thuộc vào, và nên phản ánh, mục đích của việc thực hiện kiểm định giả thuyết. Có 3 cách chọn giả thuyết chọn lựa:

Giả thuyết chọn lựa: Chọn giả thuyết chọn lựa phụ thuộc vào, và nên phản ánh, mục đích của việc thực hiện kiểm định giả thuyết. Có 3 cách chọn giả thuyết chọn lựa:

(1) **Kiểm định hai đuôi** *(two-tailed test)*, nếu chúng ta quan tâm đầu tiên đến việc kết luận có phải trung bình của tổng thể (μ) khác ($\neq$) với giá trị đã định (μ_0)? Lúc đó, giả thuyết chọn lựa nên là $\mu \neq \mu_0$. Chúng ta diễn đạt giả thuyết chọn lựa dạng này như sau:

$$H_a : \mu \neq \mu_0$$

Kiểm định giả thuyết có giả thuyết chọn lựa như thế được gọi là kiểm định hai đuôi.

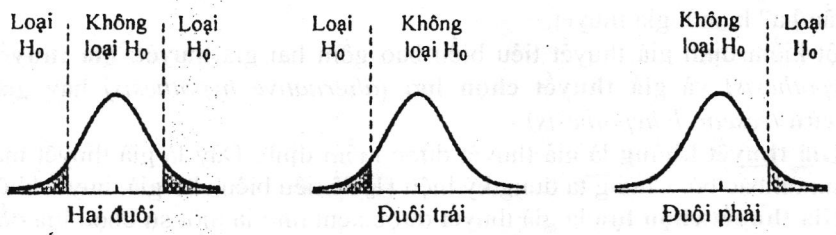
(2) **Kiểm định đuôi trái** *(left-tailed test)*, nếu chúng ta quan tâm đầu tiên đến kết luận có phải trung bình của tổng thể (μ) nhỏ hơn (<) giá trị đã định (μ_0)? Lúc đó, giả thuyết chọn lựa nên là $\mu < \mu_0$

$$H_a: \mu < \mu_0$$

(3) **Kiểm định đuôi phải** *(right-tailed test)*, nếu chúng ta quan tâm đầu tiên đến việc kết luận có phải trung bình của tổng thể (μ) lớn hơn (>) giá trị đã định (μ_0)? Lúc đó giả thuyết chọn lựa nên là $\mu > \mu_0$

$$H_a : \mu > \mu_0$$

Một kiểm định giả thuyết được gọi là kiểm định một đuôi (one-tailed test) nếu nó kiểm định hoặc đuôi trái hoặc đuôi phải (Hình 3.1).



Hình 3.1: Vùng bác bỏ của kiểm định hai đuôi, đuôi trái và đuôi phải.

2. Sai lầm loại I và loại II *(Type I and type II errors)*

a. Sai lầm loại I và loại II

Bất cứ lúc nào sử dụng phương pháp thống kê suy luận, luôn luôn có khả năng kết luận đạt được sẽ không đúng. Điều này do chỉ có một phần thông tin lấy từ mẫu được dùng để rút ra các kết luận cho tổng thể. Việc quyết định lựa chọn một trong hai giả thuyết: H_0 và H_a đều có thể phạm phải hai loại sai lầm chính.

Trong kiểm định giả thuyết bốn khả năng có thể xảy ra: Hai trong bốn sẽ dẫn đến kết luận không đúng (Bảng 3.1). Trong bảng, chúng ta thấy một kết luận không đúng sẽ xảy ra nếu giả thuyết H_0 đúng bị bác bỏ, được gọi là **sai lầm loại I**, hoặc giả thuyết H_0 sai nhưng không bị bác bỏ, được gọi là **sai lầm loại II**.

Như vậy,

* **Sai lầm loại I** là sai lầm bác bỏ giả thuyết 'không' khi nó thật sự đúng. Đây là loại sai lầm nghiêm trọng nhất.

* **Sai lầm loại II** là sai lầm không bác bỏ giả thuyết 'không' khi nó thật sự sai.

Trong kiểm định giả thuyết, người ta xem như giả thuyết không và giả thuyết chọn lựa hoàn toàn đúng. Nói khác đi, nếu giả thuyết không sai, lúc đó giả thuyết chọn lựa đúng và ngược lại.

Bảng 3.1: Bốn khả năng có thể xảy ra cho một kiểm định giả thuyết

		H_0 thật sự	
		Đúng	**Sai**
Kết luận	**Bác bỏ H_0**	Sai lầm loại I $p = \alpha$ = Mức ý nghĩa	Quyết định đúng $p = 1 - \beta$ = Độ mạnh của kiểm định
	Chấp nhận H_0	Quyết định đúng $p = 1 - \alpha$	Sai lầm loại II $p = \beta$

b. Xác suất sai lầm loại I và loại II

*** Xác suất sai lầm loại I**, được ký hiệu là α, là xác suất bác bỏ giả thuyết không (H_0) khi nó thật sự đúng. Xác suất sai lầm loại I còn được gọi là mức ý nghĩa của kiểm định *(level of significance of the test)* được định nghĩa như sau:

p(sai lầm loại I) = p(bác bỏ H_0 khi H_0 đúng) = α

p(chấp nhận H_0 khi H_0 đúng) = 1- α

*** Xác suất sai lầm loại II**, được ký hiệu là β, là xác suất không bác bỏ giả thuyết H_0 khi nó thật sự sai. Xác suất sai lầm loại II tùy thuộc vào giá trị đúng của μ.

p(sai lầm loại II) = p(chấp nhận H_0 khi H_0 sai) = β

Như vậy, độ mạnh của kiểm định *(Power of the test)*

Power = 1 - β

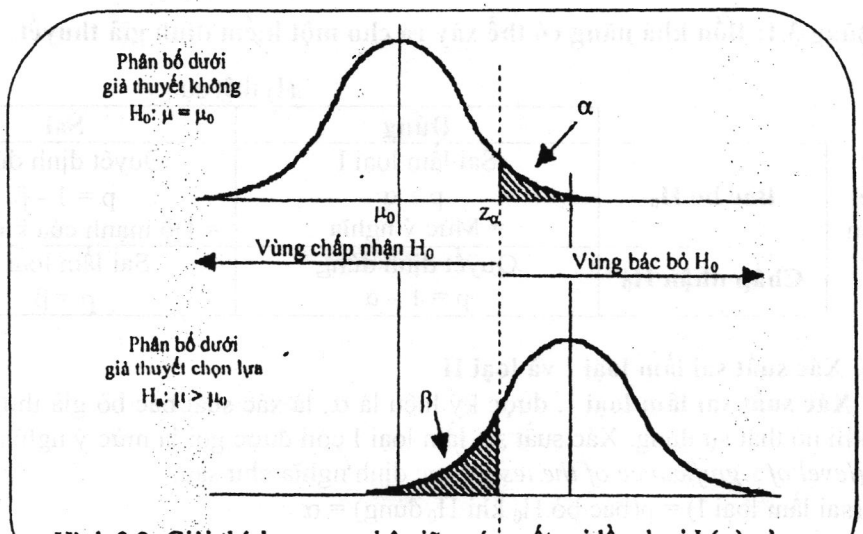
= p(bác bỏ giả thuyết H_0 khi nó thật sự sai)

Một kiểm định thống kê lý tưởng là làm thế nào giảm cực tiểu cả hai loại sai lầm. Lúc đó, xác suất dẫn đến kết luận không đúng sẽ nhỏ bất chấp giả thuyết H_0 có đúng hay không. Tuy nhiên, đối với cỡ mẫu đã định xác suất sai lầm loại I (α) càng nhỏ thì xác suất sai lầm loại II (β) càng lớn và ngược lại. Do đó, chúng ta cần phải ước tính những rủi ro có thể gặp khi phạm phải cả hai loại sai lầm và sử dụng cách ước tính nào đó để cân bằng các xác suất sai lầm loại I và loại II.

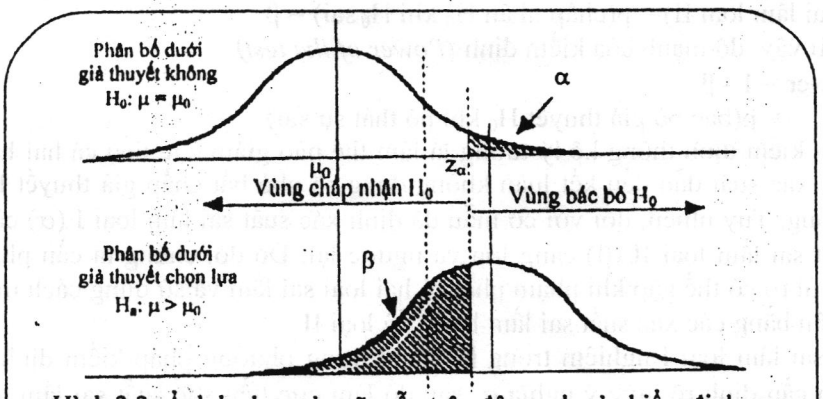
Vì sai lầm loại I nghiêm trọng hơn nên trong phương pháp kiểm định giả thuyết đầu tiên cần định rõ mức ý nghĩa α, sau đó làm cực tiểu xác suất sai lầm loại II bằng cách kiểm soát thích hợp sai số thí nghiệm. Mặt khác, với tầm quan trọng của việc quyết định chấp nhận giả thuyết H_0 khi nó đúng nên định mức ý nghĩa α ở giá trị nhỏ nhất.

Ở Hình 3.2 cho chúng ta thấy mức ý nghĩa của kiểm định càng cao (ví dụ α = 0,01) độ mạnh của kiểm định càng yếu, vì có sự dời ranh giới của vùng bác bỏ H_0 về bên phải (Chúng ta định rõ vùng chấp nhận và vùng bác bỏ H_0, vì giả thuyết H_0 chỉ dự đoán một giá trị duy nhất, trái với giả thuyết chọn lựa, H_a, dự đoán có vô số trường hợp có thể chọn). Như vậy, độ mạnh của kiểm định ở mức 0,05 sẽ cao hơn độ mạnh của kiểm định ở mức 0,01 (diện tích dưới đường cong về bên phải của z_α giảm so với α = 0,05). Trái lại, mức ý nghĩa của kiểm định càng cao, cơ hội phạm phải sai lầm loại I càng thấp.

Hình 3.3 cho thấy ảnh hưởng của cỡ mẫu lên độ mạnh của kiểm định thống kê. Cỡ mẫu càng nhỏ, đường cong sẽ có khuynh hướng càng bẹt. Điều này dẫn đến vùng bác bỏ H_0 sẽ dịch sang bên phải. Ảnh hưởng của độ bẹt sẽ gấp đôi, vì phân bố của mẫu thứ hai sẽ tương tự: diện tích dưới đường cong của mẫu thứ hai ở bên phải của z_α sẽ giảm so với đường cong có được cho cỡ mẫu lớn hơn.



Hình 3.2: Giải thích sự quan hệ giữa xác suất sai lầm loại I (α) và xác suất sai lầm loại II (β) (Phỏng theo Hình 5.2 của Steel and Torrie)



Hình 3.3: Ảnh hưởng của cỡ mẫu trên độ mạnh của kiểm định: sự liên hệ giữa sai lầm loại I (p = α) và sai lầm loại II (p = β) đối với cỡ mẫu nhỏ.

Hình 3.4 chứng tỏ ảnh hưởng của sự khác biệt giữa các trung bình so sánh lên độ mạnh của kiểm định. Các trung bình càng thật sự khác nhau, càng dễ dàng khẳng định ở tất cả độ tin cậy. Nhưng nếu chúng càng thật sự gần nhau, sẽ càng khó quyết định.

Kiểm soát sai lầm loại II đối với cỡ mẫu

Chọn cỡ mẫu và xác suất thực hiện sai lầm loại II bị ràng buộc khá chặt chẽ. Ví dụ, chúng ta kiểm định giả thuyết sau:

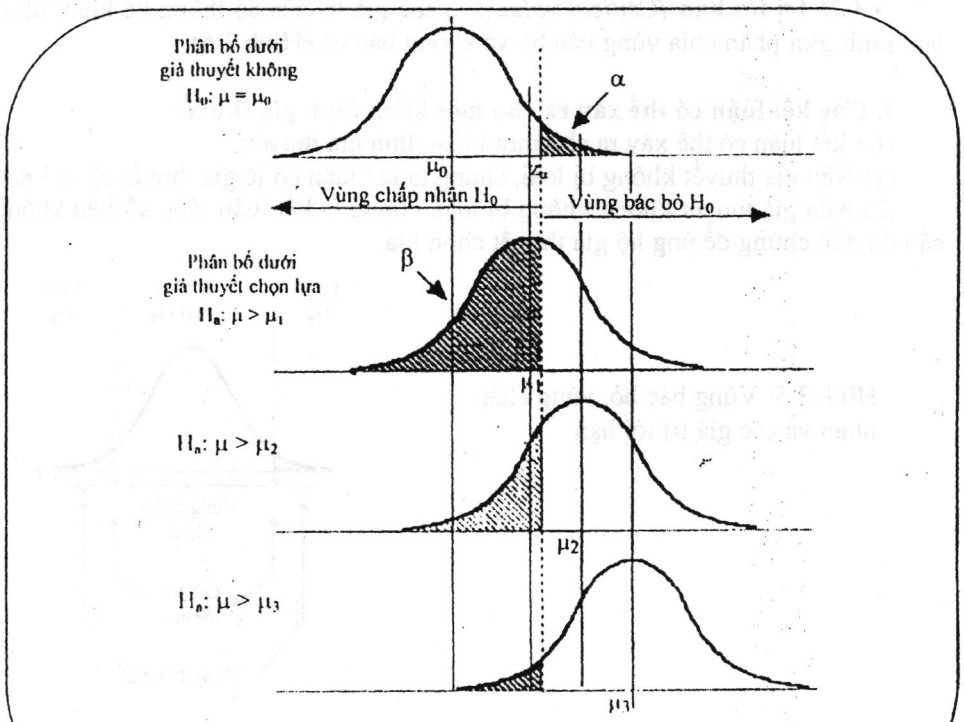
H_0: $\mu_1 = \mu_2$

H_a: $\mu_1 \neq \mu_2$

Chúng ta có thể chứng tỏ rằng:

* Sai lầm β là một hàm số của sự khác biệt thật sự giữa các trung bình: μ_1 và μ_2 khác biệt càng lớn, xác suất thực hiện sai lầm loại II càng yếu, đối với cỡ mẫu và mức ý nghĩa đã định (nghĩa là, kiểm định sẽ phát hiện sự khác biệt mạnh dễ dàng hơn sự khác biệt yếu, xem Hình 3.4).

* Sai lầm β cũng là một hàm số của cỡ mẫu: cỡ mẫu tăng, xác suất thực hiện sai lầm loại II sẽ giảm, đối với sự khác biệt giữa các trung bình và mức ý nghĩa đã định (Hình 3.3).



Hình 3.4: Ảnh hưởng của sự khác biệt giữa trung bình thật và trung bình mẫu trên độ mạnh của kiểm định (Phỏng theo Hình 5.5 của Steel and Torrie)

Cần lưu ý khi kiểm định thống kê dẫn đến việc chấp nhận giả thuyết H_0, xác suất sai lầm loại II là bao nhiêu chúng ta không biết (thường rất khó biết) và có thể lớn. Vì vậy, việc chấp nhận H_0 là một quyết định dè dặt. Khi chấp nhận giả thuyết H_0, ta không nên hiểu rằng H_0 đúng mà chỉ nên hiểu rằng các chứng cớ và số liệu chưa đủ cơ sở để bác bỏ H_0 và cần phải nghiên cứu tiếp.

Trong thực hành, người ta thường sử dụng hai mức ý nghĩa (α): 5% (0,05) và 1% (0,01). Thống kê có ý nghĩa ở 5%, lúc đó sẽ có khoảng 5 trong số 100 cơ hội chúng ta loại bỏ giả thuyết H_0 mà đáng lẽ ra nó nên được chấp nhận; nghĩa là, chúng ta tin cậy 95% quyết định đúng. Như vậy, xác suất chấp nhận giả thuyết H_0 khi nó thật sự đúng sẽ là 1 - α = 95%. Trong trường hợp như thế, chúng ta nói rằng giả thuyết H_0 bị loại ở mức ý nghĩa 0,05 hay phát biểu là "có ý nghĩa" và được đánh dấu sao (*). Tương tự, thống kê có ý nghĩa ở 1%, được phát biểu là "có ý nghĩa cao" và được đánh dấu hai sao (**).

3. Các thuật ngữ được dùng trong kiểm định giả thuyết

• **Số thống kê kiểm định** *(Test statistic)* là số thống kê được sử dụng như nền tảng để quyết định có nên loại giả thuyết không hay không'?

• **Vùng bác bỏ** *(Rejection region)* là vùng chứa bộ giá trị của số thống kê kiểm định dẫn đến việc bác bỏ giả thuyết không.

• **Vùng không bác bỏ** hoặc **vùng chấp nhận** *(non-rejection or acceptance region)* là vùng chứa bộ giá trị của số thống kê kiểm định dẫn đến không bác bỏ giả thuyết không.

• **Giá trị tới hạn** *(Critical values)* là các giá trị của số thống kê kiểm định dùng làm ranh giới phân chia vùng bác bỏ và không bác bỏ (Hình 3.5).

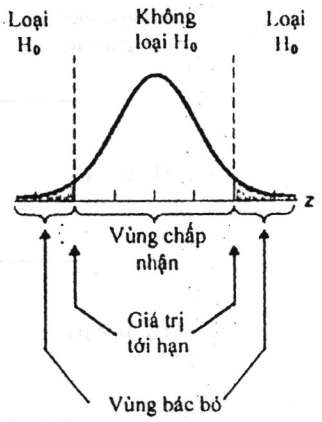
4. Các kết luận có thể xảy ra cho một kiểm định giả thuyết

Hai kết luận có thể xảy ra cho một kiểm định giả thuyết:

(1) Nếu giả thuyết không bị loại, chúng ta kết luận có lẽ giả thuyết chọn lựa đúng.

(2) Nếu giả thuyết không không bị loại, chúng ta kết luận rằng số liệu không cung cấp đủ dẫn chứng để ủng hộ giả thuyết chọn lựa.

Hình 3.5: Vùng bác bỏ, vùng chấp nhận và các giá trị tới hạn



5. Các bước tiến hành cho một kiểm định giả thuyết thống kê

(1) Đặt các giả thuyết H_0 và H_a

(2) Định rõ mức ý nghĩa α

(3) Tính giá trị của số thống kê kiểm định

(4) Tìm giá trị chuẩn ở mức ý nghĩa α đã định

(5) So sánh giá trị của của số thống kê kiểm định và giá trị tới hạn để đi đến quyết định chấp nhận hay bác bỏ giả thuyết H_0, tùy theo giá trị của số thống kê kiểm định rơi vào vùng chấp nhận hay vùng bác bỏ.

(6) Rút ra kết luận

II. KIỂM ĐỊNH GIẢ THUYẾT CHO GIÁ TRỊ TRUNG BÌNH CỦA TỔNG THỂ

1. Trường hợp phương sai đã biết hoặc phương sai chưa biết, mẫu lớn

Để kiểm định giả thuyết cho trung bình tổng thể trong trường hợp phương sai (σ^2) đã biết hoặc σ^2 chưa biết nhưng cỡ mẫu lớn (**n ≥ 30**). Giả thuyết được đặt là:

$$H_0: \mu = \mu_0,$$

với μ_0 là một giá trị nào đó. Vì cỡ mẫu lớn, nên chúng ta sử dụng biến ngẫu nhiên

$$z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

như số thống kê kiểm định để kiểm định giả thuyết. Số thống kê kiểm định này cho chúng ta biết số trung bình mẫu được quan sát cách xa μ_0 bao nhiêu độ lệch chuẩn.

Ví dụ: Một nông trường trồng mía đã sử dụng phân bón của công ty A trong nhiều năm ghi nhận năng suất mía bình quân là 10,5 tấn/ha. Gần đây, một công ty phân bón mới xuất hiện trên thị trường quảng cáo sản phẩm của công ty có thể làm gia tăng năng suất mía. Để kết luận về hiệu quả của phân bón mới, nông trường đã bón thử trên 45 lô. Kết quả năng suất mía trên từng lô được ghi nhận trong Bảng 3.2. Với số liệu thu thập có đủ dẫn chứng để kết luận phân bón mới làm gia tăng năng suất mía trung bình trên đất của nông trường hay không?

Bảng 3.2: Năng suất mía (tấn/ha) của 45 lô được bón phân bón mới

12,0	10,1	10,4	11,6	10,9	8,1	8,2	12,3	13,0
10,1	12,6	11,6	12,7	9,8	11,6	12,8	13,6	11,0
11,3	9,4	9,5	13,3	11,5	12,6	11,5	11,4	12,5
10,6	11,9	12,5	11,6	13,1	12,1	10,7	10,3	12,4
11,0	6,3	10,8	10,5	10,3	10,7	12,7	9,3	11,5

Gọi μ là năng suất mía trung bình khi sử dụng phân bón mới. Đặt các giả thuyết không (H_0) và chọn lựa (H_a) như sau:

$H_0: \mu = 10,5$ tấn/ha đối với $H_a: \mu > 10,5$ tấn/ha

Đây là dạng kiểm định đuôi phải vì có dấu lớn hơn (>) trong giả thuyết chọn lựa.

* Giả sử kiểm định giả thuyết được thực hiện ở mức ý nghĩa 5% ($\alpha = 0,05$). Do đó, giá trị tới hạn cho kiểm định đuôi phải là $z_\alpha = z_{0,05} = 1,645$ (xem Phụ lục II).

* Tính số thống kê kiểm định

Vì cỡ mẫu lớn (n = 45), chúng ta sử dụng biến chuẩn tắc z .

$$z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Chúng ta có $\mu_0 = 10,5$ tấn/ha và n = 45. Từ số liệu ở Bảng 3.2, chúng ta tính được

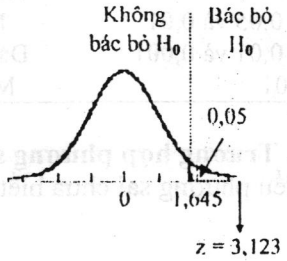
$$\bar{X} = \frac{\sum X}{n} = \frac{503,7}{45} = 11,193 \text{ tấn/ha}$$

và $$s = \sqrt{\frac{(\sum X^2 - (\sum X)^2/n}{n-1}} = \sqrt{\frac{5735,69 - (503,7)^2/45}{45-1}} = 1,489 \text{ tấn/ha}$$

Do đó, giá trị của số thống kê kiểm định là:

$$z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{11,193 - 10,5}{1,489/\sqrt{45}} = 3,123$$

* Vì giá trị của số thống kê kiểm định z lớn hơn giá trị z tới hạn ($z_{0,05}$), nên z nằm trong vùng bác bỏ giả thuyết H_0 (Hình 3.6).



Hình 3.6: Giá trị z của số thống kê kiểm định nằm trong vùng bác bỏ giả thuyết H_0

Như vậy, số liệu cung cấp đủ dẫn chứng để kết luận phân bón mới làm gia tăng năng suất mía trung bình trên đất của nông trường.

✱ Giá trị P *(Probability)*

Trong phương pháp thống kê cổ điển, để kiểm định giả thuyết, mức ý nghĩa thường được định trước và dựa theo đó kết luận được phát biểu ở dạng bác bỏ hay không bác bỏ giả thuyết không.

Phương pháp này có một vài bất lợi vì nó không cho phép người đọc tự đánh giá kết quả theo mức ý nghĩa mà họ mong muốn và cũng không cung cấp cho đọc giả mức ý nghĩa thật sự của kiểm định. Để khắc phục nhược điểm này, nhiều nhà nghiên cứu và hầu hết các phần mềm thống kê đều ghi nhận giá trị P của kiểm định giả thuyết trong báo cáo của mình. Như vậy

- Giá trị P của kiểm định giả thuyết có thể được xem như ***mức ý nghĩa quan sát***. Đó là xác suất sai lầm loại I tại giá trị quan sát của số thống kê kiểm định. Giá trị P càng nhỏ, dẫn chứng chống lại giả thuyết không càng mạnh.

- Tiêu chuẩn quyết định cho một kiểm định giả thuyết bằng cách sử dụng giá trị P: Nếu giá trị P nhỏ hơn hoặc bằng với mức ý nghĩa đã định (α), lúc đó loại giả thuyết không; ngược lại, không loại giả thuyết không.

Với ví dụ năng suất mía, giá trị của số thống kê kiểm định là 3,12. Giá trị P là xác suất sai lầm loại I tại giá trị $z = 3,12$ chính là $p(z \geq 3,12)$ nếu giả thuyết H_0 thật sự đúng. Xác suất đó bằng 0,5000 - 0,4991 = 0,0009. Vì vậy, giá trị P của kiểm định giả thuyết bằng 0,0009.

Nếu giá trị P nhỏ hơn hoặc bằng giá trị α đã định, lúc đó loại giả thuyết H_0; ngược lại, không loại giả thuyết H_0. Trong ví dụ, vì giá trị P bằng 0,0009 nhỏ hơn mức ý nghĩa đã định ($\alpha = 0,05$), nên chúng ta loại giả thuyết H_0.

Như vậy, mẫu được lấy cung cấp đủ dẫn chứng để kết luận phân bón mới làm gia tăng năng suất mía trung bình trên đất của nông trường.

Tóm lại, một kiểm định giả thuyết được gọi là có ý nghĩa nếu giá trị P đủ nhỏ để bảo đảm loại giả thuyết H_0. Nhà thống kê học, Ronald Fisher, đề nghị giá trị P bằng 0,05 là ranh giới để loại hay không loại giả thuyết H_0; nghĩa là, giữa mức có ý nghĩa và không có ý nghĩa của kiểm định giả thuyết. Ngày nay, hầu hết các nhà nghiên cứu vẫn còn trung thành với đề nghị này.

Ưu điểm của việc sử dụng giá trị P là nó cho phép một khoảng rộng hơn các kết luận có thể có cho một kiểm định giả thuyết vừa đúng mức để không loại hoặc loại giả thuyết H_0. Nghĩa là, nó biểu thị mức độ tin tưởng hoặc không tin tưởng giả thuyết H_0.

Bảng dưới đây hướng dẫn nguyên tắc để giải thích giá trị P của một kiểm định giả thuyết. Giá trị P càng nhỏ, mức ý nghĩa càng mạnh và càng có thể kém tin tưởng giả thuyết H_0.

Giá trị P	Mức quan trọng của dẫn chứng bác bỏ H_0	Giải thích kết quả (Ký hiệu)	Kết luận
> 0,05	Rất yếu	Không ý nghĩa (ns)	Không ý nghĩa
Giữa 0,05 và 0,01	Nhẹ	Có ý nghĩa ở mức 5% (*)	Có ý nghĩa
Giữa 0,01 và 0,001	Đáng kể	Có ý nghĩa ở mức 1% (**)	Có ý nghĩa cao
< 0,001	Mạnh	Có ý nghĩa ở mức 1‰ (***)	Rất có ý nghĩa

2. Trường hợp phương sai chưa biết, mẫu nhỏ (n < 30)

Nếu phương sai chưa biết và cỡ mẫu nhỏ, chúng ta sử dụng biến ngẫu nhiên

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

như số thống kê kiểm định để kiểm định giả thuyết.

Ví dụ: Năng suất trung bình của một giống lúa trồng trên vùng đất bị nhiễm phèn là 4,0 tấn/ha. Sau một thời gian làm thủy lợi để rửa phèn, người ta nhận thấy năng suất lúa đã được cải thiện. Để kiểm chứng điều này, một nhà nghiên cứu đã quan sát 20 lô ruộng thí nghiệm trồng giống lúa này trên vùng đất đã rửa phèn. Kết quả được ghi nhận trong Bảng 3.3.

Bảng 3.3: Năng suất lúa (tấn/ha) của 20 lô thí nghiệm trồng trên vùng đất đã rửa phèn

4,20	4,10	4,27	4,36
4,13	4,18	4,38	4,33
4,24	4,28	4,25	4,35
4,32	4,23	4,31	4,17
4,18	4,27	4,40	4,41

Với số liệu trên có đủ dẫn chứng để kết luận làm thủy lợi rửa phèn có thể cải thiện được năng suất lúa hay không?

Gọi μ là năng suất trung bình của một giống lúa trồng trên vùng đất đã được rửa phèn.

Đặt các giả thuyết:

H_0: μ = 4,0 tấn/ha đối với H_a: μ > 4,0 tấn/ha

Giả sử kiểm định giả thuyết được thực hiện ở mức ý nghĩa 5% (α = 0,05). Tìm giá trị tới hạn cho kiểm định đuôi phải là t_α ở độ tự do (df) bằng n – 1, ký hiệu là $t_{\alpha,(n-1)}$.

Ở đây n = 20, tìm trong Phụ lục IV ở df = 20 - 1 = 19, Giá trị tới hạn $t_{0,05,19}$ = 1,729

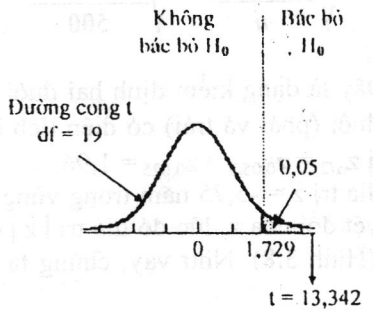
Chúng ta có μ_0 = 4,0 tấn/ha và n = 20. Từ số liệu ở Bảng 3.3, chúng ta tính được

$$\overline{X} = \frac{\sum X}{n} = \frac{85,36}{20} = 4,268 \text{ tấn/ha}$$

và $$s = \sqrt{\frac{(\sum X^2 - (\sum X)^2/n}{n-1}} = \sqrt{\frac{364,4698 - (85,36)^2/20}{20-1}} = 0,0898 \text{ tấn/ha}$$

Tính giá trị của số thống kê kiểm định

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} = \frac{(4,268 - 4,0)\sqrt{20}}{0,0898} = 13,342$$



Hình 3.7: Giá trị t của số thống kê kiểm định nằm trong vùng bác bỏ giả thuyết H_0

Giá trị t = 13,342 nằm trong vùng bác bỏ giả thuyết H_0 (Hình 3.7). Vì vậy, chúng ta bác bỏ giả thuyết H_0 và chấp nhận giả thuyết chọn lựa H_a; nghĩa là, $\mu > 4,0$ tấn/ha

Như vậy, số liệu của mẫu cung cấp đủ dẫn chứng để kết luận làm thủy lợi rửa phèn có thể cải thiện được năng suất trung bình của lúa.

III. KIỂM ĐỊNH GIẢ THUYẾT CHO GIÁ TRỊ TỶ LỆ (P)

Giả sử p là xác suất xuất hiện biến cố A trong một phép thử ngẫu nhiên (p chưa biết). Chúng ta muốn kiểm định giả thuyết $p = p_0$, với p_0 là một số đã cho.

Nếu chúng ta thực hiện n lần phép thử một cách độc lập và quan sát thấy biến cố A xuất hiện X lần. Xác suất của biến cố A, $\overline{p} = \frac{X}{n}$, là giá trị ước lượng của p.

Để thực hiện kiểm định giả thuyết H_0: $p = p_0$ đối với H_a: $p \neq p_0$, người ta dùng biến ngẫu nhiên

$$z = \frac{\overline{p} - p_0}{\sqrt{p_0 \cdot (1 - p_0)/n}} \quad \text{hoặc} \quad z = \frac{X - np_0}{\sqrt{np_0 \cdot (1 - p_0)}}$$

như số thống kê kiểm định, với điều kiện $np_0 \geq 5$ và $n(1 - p_0) \geq 5$.

Ví dụ: Một nhà máy sản xuất tự động cho tỷ lệ chính phẩm là 98%. Sau một thời gian hoạt động, người ta nghi ngờ tỷ lệ trên đã bị giảm. Kiểm tra ngẫu nhiên 500 sản phẩm thấy có 28 phế phẩm. Như vậy, có thể kết luận chất lượng làm việc của máy không còn được như trước hay không?

Gọi $\overline{p}$ là xác suất của chính phẩm ($\overline{p} = 1 - \frac{28}{500} = 0,944$).

Đặt các giả thuyết

H_0: $p_0 = 0,98$ (Chất lượng làm việc của máy vẫn như cũ)

Ha : $p_0 \neq 0,98$ (Chất lượng làm việc của máy đã bị thay đổi)

Tính trung bình và độ lệch chuẩn:

$\mu = np_0 = 500\,(0,98) = 490$

$$\sigma = \sqrt{np_0(1 - p_0)} = \sqrt{(500)(0,98)(0,02)} = 3,13$$

Tính số thống kê kiểm định:

$$z = \frac{(500 - 28) - 490}{3,13} = -5,75$$

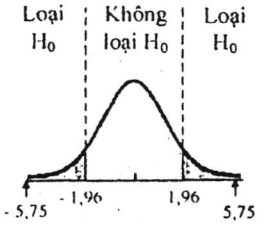
Hoặc có thể sử dụng công thức sau:

$$z = \frac{\overline{p} - p_0}{\sqrt{\frac{p_0(1 - p_9)}{n}}} = \frac{0,944 - 0,98}{\sqrt{\frac{(0,98)(0,2)}{500}}} = -5,75$$

Đây là dạng kiểm định hai đuôi (dấu ≠). Ở mức ý nghĩa 5%, giá trị z tới hạn của mỗi đuôi (phải và trái) có diện tích bằng $\alpha/2 = 0,05/2 = 0,025$. Tìm trong Phụ lục II, giá trị $z_{\alpha/2} = z_{0,05/2} = z_{0,025} = 1,96$.

Giá trị z = -5,75 nằm trong vùng bác bỏ giả thuyết H_0 ở đuôi trái, hoặc nếu lấy giá trị tuyệt đối của z, lúc đó giá trị $|z|$ cũng nằm trong vùng bác bỏ giả thuyết H_0 ở đuôi phải (Hình 3.8). Như vậy, chúng ta có thể kết luận chất lượng làm việc của máy đã

thay đổi hoặc kết luận chất lượng làm việc của máy đã bị giảm vì $\overline{p} = 0{,}944$ nhỏ hơn $p_0 = 0{,}98$.



Hình 3.8: Giá trị | z | nằm trong vùng bác bỏ giả thuyết H_0

IV. TÌM KHOẢNG TIN CẬY CHO TRUNG BÌNH (μ) HOẶC TỈ LỆ (P)

1. Khoảng tin cậy của trung bình tổng thể (μ)

Khoảng tin cậy của μ là một khoảng chứa các số trung bình ước lượng của μ; nghĩa là $\overline{X}$, kèm theo số phần trăm định rõ chúng ta tin cậy như thế nào khi μ nằm trong khoảng đó. Phần trăm tin cậy được gọi là mức độ tin cậy (1 - α).

a) Trường hợp phương sai đã biết hoặc phương sai chưa biết, mẫu lớn (n ≥ 30)

Giả sử một mẫu ngẫu nhiên có n ≥ 30 được rút ra từ một tổng thể có trung bình là μ. Lúc đó biến ngẫu nhiên $\overline{X}$ có phân phối gần chuẩn với trung bình $\mu_{\overline{X}} = \mu$ và độ lệch chuẩn $\sigma_{\overline{X}} = \sigma/\sqrt{n}$. Như vậy, xác suất để một giá trị quan sát của $\overline{X}$ sẽ nằm trong $z_{\alpha/2}$ độ lệch chuẩn về một trong hai phía của trung bình tổng thể gần bằng 1 - α.

$$p[(\mu_{\overline{X}} - z_{\alpha/2}.\sigma_{\overline{X}}) \le \overline{X} \le (\mu_{\overline{X}} + z_{\alpha/2}.\sigma_{\overline{X}})] = 1 - \alpha.$$

Vì $\mu_{\overline{X}} = \mu$ và $\sigma_{\overline{X}} = \sigma/\sqrt{n}$, chúng ta có thể viết lại:

$$p[(\mu - z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}) \le \overline{X} \le (\mu + z_{\alpha/2}.\frac{\sigma}{\sqrt{n}})] = 1 - \alpha.$$

hoặc $p[(\overline{X} - z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}) \le \mu \le (\overline{X} + z_{\alpha/2}.\frac{\sigma}{\sqrt{n}})] = 1 - \alpha.$

Phương trình này cho thấy khoảng tin cậy (1 - α) cho μ ở mỗi lần lấy mẫu là

$$(\overline{X} - z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}) \text{ đến } (\overline{X} + z_{\alpha/2}.\frac{\sigma}{\sqrt{n}})$$

Chú ý: Đối với cỡ mẫu đã định, mức độ tin cậy càng cao, độ rộng của khoảng tin cậy càng lớn.

Ví dụ: Khoảng tin cậy 95% cho năng suất mía trung bình khi sử dụng phân bón mới được tính như sau:

$$\overline{X} \pm z_{\alpha/2}(\sigma/\sqrt{n})$$

Với α = 0,05, tra trong Phụ lục II, giá trị $z_{\alpha/2} = z_{0,05/2} = z_{0,025} = 1{,}96$. Vì n = 45, σ = 1,489 tấn/ha và $\overline{X}$ = 11,19 tấn/ha, khoảng tin cậy 95% cho μ là

$$11{,}193 \pm 1{,}96.(1{,}489/\sqrt{45})$$

hoặc $11{,}193 \pm 0{,}435$

hoặc $10{,}758 \le \mu \le 11{,}628$

Như vậy, với mức độ tin cậy 95% chúng ta có thể kết luận rằng khi sử dụng phân bón mới năng suất mía gia tăng trung bình từ 10,758 đến 11,628 tấn/ha.

b) Trường hợp phương sai chưa biết, mẫu nhỏ (n < 30)

Trường hợp cỡ mẫu nhỏ, n < 30, khoảng tin cậy 1 - α như sau:

$$p[(-t_{\alpha/2} \le \frac{\overline{X}-\mu}{s/\sqrt{n}} \le t_{\alpha/2})] = 1 - \alpha$$

hoặc

$$p\left(\overline{X} - t_{\alpha/2}.\frac{s}{\sqrt{n}} \le \mu \le \overline{X} + t_{\alpha/2}.\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Như vậy, khoảng tin cậy (1 - α) cho μ là

$$\overline{X} - t_{\alpha/2}.\frac{s}{\sqrt{n}} \text{ đến } \overline{X} + t_{\alpha/2}.\frac{s}{\sqrt{n}}$$

Ví dụ: Ước lượng năng suất trung bình của một giống lúa trồng trên vùng đất đã được rửa phèn ở mức độ tin cậy 95% như sau:

Đối với mức độ tin cậy 95%, giá trị t tới hạn tìm trong Phụ lục IV ở độ tự do là 20 - 1 = 19, giá trị $t_{\alpha/2} = t_{0,05/2} = t_{0,025} = 2,093$.

Vì $\overline{X}$ = 4,268 t/ha và s = 0,0898 t/ha. Thế vào công thức, ta được khoảng tin cậy cho μ là:

$$4,268 \pm (2,093)(0,0898)/\sqrt{20}$$

hoặc $4,268 \pm 0,042$

hoặc $4,226 \le \mu \le 4,310$

Như vậy, với mức độ tin cậy 95% chúng ta có thể kết luận lúa được trồng trên vùng đất đã rửa phèn năng suất gia tăng trung bình từ 4,586 đến 4,67 tấn/ha.

2. Khoảng tin cậy cho tỷ lệ (p)

Gọi p là xác suất xuất hiện biến cố A trong tổng thể (p chưa biết). Nhưng tham số này có thể được ước lượng dựa trên số liệu mẫu điều tra. Giả sử trong một mẫu cỡ n có X cá thể xuất hiện biến cố A. Vậy tỷ lệ xuất hiện biến cố A của mẫu bằng $\overline{p} = \frac{X}{n}$ là một ước lượng không chệch của p. Khoảng tin cậy của p là

$$p\left(\overline{P} - z_{\alpha/2}.\sqrt{\frac{p(1-p)}{n}} \le p \le \overline{P} + z_{\alpha/2}.\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

Với ví dụ về chất lượng sản phẩm, khoảng tin cậy 95% cho tỷ lệ chính phẩm được tính như sau:

$$\overline{p} \pm z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}} = 0,944 \pm 1,96\sqrt{\frac{(0,98)(0,02)}{500}}$$

$$= 0,944 \pm 0,012$$

Hoặc $0,932 \le p \le 0,956$

Như vậy, với mức độ tin cậy 95% chúng ta có thể kết luận rằng sau một thời gian hoạt động máy sản xuất tỷ lệ chính phẩm chỉ còn từ 93,2% đến 95,6%

3. Kiểm định ý nghĩa đối với khoảng tin cậy

Mặc dù kiểm định giả thuyết là kiểu phương pháp được áp dụng phổ biến nhất để thực hiện so sánh các trung bình; tuy nhiên, nó cung cấp thông tin tương đối ít về các giá trị có thể có của các tham số đúng. Một phương pháp tương tự với "kiểm định giả thuyết" nhưng cho nhiều thông tin hơn là "ước lượng các tham số". Như vậy, ước

lượng một tham số nhờ vào khoảng tin cậy, đó là khoảng mà bên trong khoảng đó phải tìm thấy giá trị đúng của tham số được ước lượng, cùng với một mức độ tin cậy nào đó.

Khoảng tin cậy cũng có thể được xây dựng để kiểm định giả thuyết liên quan đến việc định vị chính xác cho một tham số (μ_0 hoặc p_0). Với phương pháp này, trước tiên lập khoảng tin cậy, nếu giá trị của giả thuyết H_0 nằm trong khoảng đó, lúc đó chấp nhận giả thuyết H_0. Trái lại, nếu giá trị giả thuyết nằm ngoài khoảng sẽ bác bỏ H_0.

V. XÁC ĐỊNH CỠ MẪU

Để ước lượng các tham số của tổng thể chúng ta dựa vào đặc trưng của mẫu. Mẫu càng nhỏ thì giá trị ước lượng của tổng thể càng bị sai lệch. Trái lại, mẫu càng lớn giá trị ước lượng các tham số của tổng thể càng chính xác.

Do đó, khi lập kế hoạch nghiên cứu, câu hỏi đầu tiên thường được đặt ra là nên chọn cỡ mẫu bằng bao nhiêu để kết quả thí nghiệm có thể tin cậy. Tuy nhiên, nếu n quá lớn sẽ tốn nhiều công sức, thời gian và kinh phí. Vã lại, không cần thiết n lớn mới kết luận được. Cho nên với một độ chính xác nhất định, chúng ta sử dụng n tối thiểu cho phép rút ra kết luận tin cậy.

Ví dụ: Điều tra về mức sử dụng nhiên liệu cho các phương tiện vận chuyển ở từng hộ gia đình, chúng ta chọn ngẫu nhiên 30 hộ và tính được mức sử dụng nhiên liệu trung bình hằng tháng là \$58,56 và σ = \$20,65.

Khoảng tin cậy 95% cho μ tìm được là

$$58{,}56 \pm 1{,}96.(20{,}65/\sqrt{30})$$

hoặc $58{,}56 \pm 7{,}39$

hoặc $51{,}17 \le \mu \le 65{,}95$

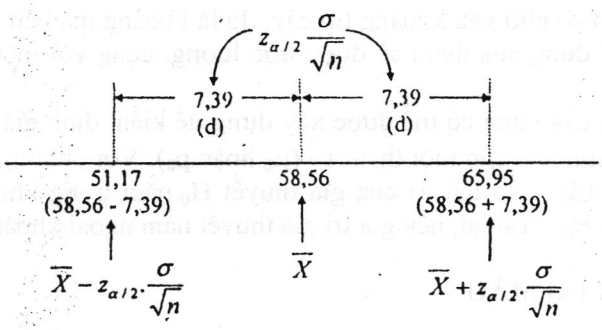
Khoảng tin cậy này tương đối rộng, nên nó cho một khoảng khá lớn các giá trị có thể có của μ. Nói cách khác, độ chính xác của thí nghiệm kém. Để khắc phục, chúng ta cần giảm độ rộng của khoảng tin cậy.

Một cách để làm giảm độ rộng của khoảng tin cậy là hạ thấp mức tin cậy từ 95% đến một mức thấp hơn nào đó. Cách khác, giả sử chúng ta muốn giữ lại cùng mức tin cậy mà vẫn làm hẹp khoảng tin cậy bằng cách gia tăng cỡ mẫu (n). Để có thể thực hiện được điều này, đầu tiên chúng ta hãy quan sát khoảng tin cậy vừa tính được và biểu diễn khoảng tin cậy đó bằng đồ thị (Hình 3.9). Chúng ta thấy độ rộng của khoảng tin cậy được xác định bằng đại lượng

$$d = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

mà d bằng một nửa độ rộng của khoảng tin cậy. Đại lượng d còn được gọi là sai số tối đa của số ước lượng. Trong trường hợp này d = 7,39. Chúng ta dùng thuật ngữ này vì chúng ta tin cậy 95% rằng μ nằm trong khoảng tin cậy. Nói cách khác, sai số tối đa của số ước lượng tính được là d = \$7,39.

Trong bất cứ trường hợp nào, chúng ta biết rằng để làm hẹp khoảng tin cậy và gia tăng độ chính xác của số ước lượng, chỉ cần giảm giá trị của d. Do trong công thức tính giá trị d, n xuất hiện ở mẫu số, nên d sẽ giảm nếu gia tăng cỡ mẫu. Dĩ nhiên, điều này cũng hợp lý vì chúng ta hy vọng có được thông tin chính xác từ mẫu lớn.



Hình 3.9: Khoảng tin cậy 95% cho mức sử dụng nhiên liệu trung bình hàng tháng (μ) cho các phương tiện vận chuyển ở từng hộ.

1. Cỡ mẫu cần lấy để ước lượng trung bình (μ)

Thường độ chính xác (hay sai số tối đa của số ước lượng) và mức tin cậy của một khoảng tin cậy được định trước. Sau đó, chúng ta sẽ xác định cỡ mẫu cần lấy.

Cỡ mẫu cần lấy để ước lượng khoảng tin cậy (1 - α) cho μ, có sai số tối đa của số ước lượng (d) đã định trước được tính theo công thức:

$$n \geq \left(\frac{z_{\alpha/2}.\sigma}{d}\right)^2$$

n được chọn quanh số nguyên gần nhất.

hoặc $$n \geq \frac{z^2_{\alpha/2}.V^2}{d_r^2}$$

với $V = \frac{\sigma}{\overline{X}} \cdot 100$, là **hệ số biến động** (cv) được biểu thị bằng phần trăm của trung bình,

$d_r = \frac{d}{\overline{X}} \cdot 100$, là sai số tương đối tối đa của số ước lượng, cũng được biểu thị bằng phần trăm của trung bình.

Đối với α = 0,05 và α = 0,01, công thức trên có thể tính gần đúng như sau:

$$n \approx \frac{4.V^2}{d_r^2} \quad \text{và} \quad n \approx \frac{7.V^2}{d_r^2}$$

vì: $z^2_{0,025} = (1,96)^2 \approx 4$ và $z^2_{0,005} = (2,58)^2 \approx 7$

Cũng với ví dụ điều tra về mức sử dụng nhiên liệu, nếu d = \$0,50, n được xác định như sau:

$$n = \left(\frac{z_{\alpha/2}.\sigma}{d}\right)^2 = \left(\frac{(1,96).(20,65)}{0,5}\right)^2 = 6552,58$$

Như vậy, chúng ta cần lấy một mẫu gồm 6553 hộ. Dĩ nhiên, với cỡ mẫu quá lớn như thế rất khó thực hiện. Nhưng, nếu chọn một mẫu gồm 6553 hộ, chúng ta có thể tin cậy 95% rằng trung bình của tất cả các hộ sử dụng nhiên liệu cho các phương tiện vận chuyển biến động trong khoảng \$0,50.

Khi n ≤ 10, công thức trên trở thành

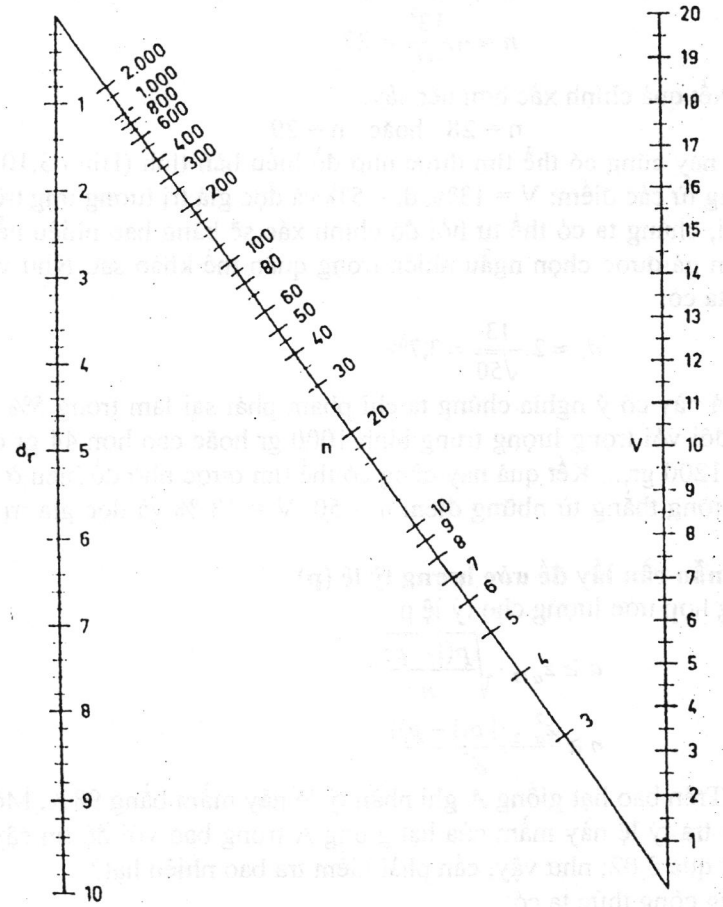
$$n \geq \frac{t^2_{\alpha/2}.V^2}{d_r^2}$$

Trong công thức, $t_{\alpha/2}$ phụ thuộc vào n qua trung gian của độ tự do, nên khi giá trị n nhỏ, mối quan hệ này ít thực tiễn, vì phải áp dụng liên tục phương pháp tính gần đúng. Tuy nhiên, qua trình bày bằng đồ biểu đơn giản, dưới dạng bàn tính (Hình 3.10), chúng ta cũng có thể xác định n. Mặt khác, đồ biểu cũng có thể dùng để xác định giá trị sai số tương đối (nửa độ dài của khoảng tin cậy, d_r) khi biết hệ số biến động V và cỡ mẫu n.

Ví dụ: Xác định số quan sát (n) dùng để ước lượng trọng lượng trung bình của một quần thể gà.

Bảng 3.4 ghi nhận trọng lượng trung bình, độ lệch chuẩn và hệ số biến động quan sát được ở 3 thời kỳ liên tiếp đối với 3 lô, mỗi lô gồm 150 con gà, được nuôi trong điều kiện giống nhau. Qua các kết quả này, chúng ta muốn xác định số quan sát (n) để thực hiện ở các thí nghiệm sau nhằm có được độ chính xác mong muốn trong việc ước lượng trọng lượng trung bình.

Qua Bảng 3.4, chúng ta nhận thấy hệ số biến động trong toàn bộ số liệu khá ổn định; tuy nhiên, trung bình và độ lệch chuẩn hầu như thay đổi từ 1 đến gần gấp đôi. Trung bình của các hệ số biến động (V) ở giữa 12 và 13%.



Hình 3.10: Sự liên hệ giữa số quan sát (n), hệ số biến động (V) và sai số tương đối tối đa (d_r) cho khoảng tin cậy 95%.

Bảng 3.4: Trung bình, độ lệch chuẩn và hệ số biến động được ước lượng từ trọng lượng của 3 lô, mỗi lô gồm 150 con gà, được đo ở 3 thời kỳ khác nhau

Lô	Thời kỳ	$\overline{X}$ (gr)	$\hat{\sigma}$ (gr)	V (%)
1	T_1	1,067	134	12,5
	T_2	1,308	166	12,7
	T_3	1,518	164	10,8
2	T_1	1,022	138	13,5
	T_2	1,242	162	13,0
	T_3	1,421	169	11,9
3	T_1	876	98	11,2
	T_2	1,094	136	12,5
	T_3	1,310	172	13,1

Để thận trọng, giả sử chúng ta gia tăng giá trị V lên đến 13% và chúng ta mong muốn có giá trị ước lượng trung bình của trọng lượng với sai số tương đối tối đa 5% (d_r = 0,05) ở mức tin cậy 95% (nghĩa là sai số tối đa 50 gr cho trọng lượng trung bình 1000 gr hoặc 60 gr cho trọng lượng trung bình 1200 gr,...). Trong điều kiện này:

$$n \approx 4.\frac{13^2}{5^2} = 27$$

Để có kết quả chính xác hơn nên lấy:

n = 28 hoặc n = 29

Giá trị này cũng có thể tìm được nhờ đồ biểu bàn tính (Hình 3.10) bằng cách nối đường thẳng từ các điểm: V = 13%, d_r = 5% và đọc giá trị tương ứng trên trục n.

Trái lại, chúng ta có thể tự hỏi độ chính xác sẽ bằng bao nhiêu nếu cân một mẫu gồm 50 con gà được chọn ngẫu nhiên trong quần thể khảo sát. Như vậy, cùng độ tin cậy chúng ta có:

$$d_r \approx 2.\frac{13}{\sqrt{50}} = 3,7\%$$

Kết quả này có ý nghĩa chúng ta chỉ phạm phải sai lầm trong 5% trường hợp cao hơn 37 gr đối với trọng lượng trung bình 1000 gr hoặc cao hơn 44 gr cho trọng lượng trung bình 1200 gr,... Kết quả này cũng có thể tìm được nhờ đồ biểu ở Hình 3.10 bằng cách nối đường thẳng từ những điểm n = 50, V = 13 % và đọc giá trị tương ứng trên trục d_r.

2. Cỡ mẫu cần lấy để ước lượng tỷ lệ (p)

Trường hợp ước lượng cho tỷ lệ p

$$d \geq z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}$$

$$n \geq \frac{z^2_{\alpha/2} \cdot [p(1-p)]}{d^2}$$

Ví dụ: Trên bao hạt giống A ghi nhận tỷ lệ nảy mầm bằng 98%. Một nhà nông học muốn kiểm tra tỷ lệ nảy mầm của hạt giống A trong bao với độ tin cậy 95% và sai số không vượt quá 0,02; như vậy, cần phải kiểm tra bao nhiêu hạt?

Áp dụng công thức ta có:

$$n \geq \frac{(1,96)^2 \cdot (0,98)(0,02)}{(0,02)^2} = 188,2$$

Vậy cần lấy n = 189 hạt để thử tỷ lệ nảy mầm.

CHƯƠNG 4

SO SÁNH HAI TRUNG BÌNH HOẶC HAI TỶ LỆ

I. PHÂN BỐ CỦA SỰ SAI BIỆT GIỮA HAI TRUNG BÌNH (mẫu lớn và độc lập)

(The sampling distribution of the difference between two means)

Trong phần này, chúng ta sẽ thực hiện kiểm định giả thuyết cho hai trung bình của hai tổng thể. Phương pháp được thảo luận ở đây đòi hỏi các mẫu phải được **chọn ngẫu nhiên** và **độc lập**. Hai mẫu được xem là độc lập nếu việc lấy mẫu từ tổng thể này không làm ảnh hưởng đến việc lấy mẫu từ tổng thể kia.

Chẳng hạn, chúng ta muốn thực hiện kiểm định giả thuyết để đi đến kết luận có sự khác biệt giữa hai biện pháp: dùng thuốc trừ sâu và thả côn trùng đực bất dục để phòng trị dịch sâu hại trên cây bắp. Chúng ta đặt:

Tổng thể 1: Năng suất bắp trung bình (μ_1) đạt được khi phun thuốc trừ sâu

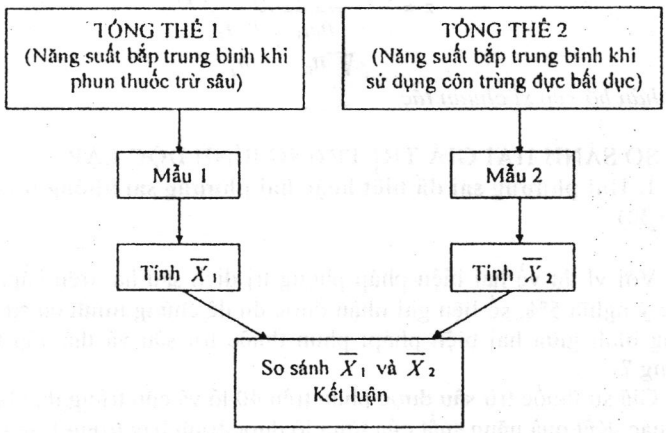
Tổng thể 2: Năng suất bắp trung bình (μ_2) đạt được khi thả côn trùng đực bất dục

Lúc đó chúng ta muốn thực hiện kiểm định giả thuyết:

H_0: $\mu_1 = \mu_2$ (Hai biện pháp cho năng suất trung bình bằng nhau)

H_a: $\mu_1 \neq \mu_2$ (Hai biện pháp cho năng suất trung bình khác nhau)

Để kiểm định giả thuyết, chúng ta lấy một mẫu từ mỗi tổng thể, tính trung bình của mỗi mẫu ($\overline{X}$) và loại giả thuyết không nếu $\overline{X}_1$ và $\overline{X}_2$ quá khác nhau; trái lại, không loại giả thuyết không (Hình 4.1).



Hình 4.1: Thực hiện so sánh hai trung bình tổng thể.

Bây giờ câu hỏi được đặt ra là sự khác biệt giữa hai trung bình mẫu là do sai số lấy mẫu hay sự khác biệt đủ lớn để chứng tỏ hai tổng thể có trung bình khác nhau. Để trả lời câu hỏi này, chúng ta cần biết xác suất để có sự khác biệt như thế ở hai trung bình mẫu bằng bao nhiêu nếu thật sự hai trung bình tổng thể bằng nhau. Muốn thế, chúng ta có thể tính phân bố xác suất của sai biệt giữa hai trung bình mẫu ($\overline{X}_1 - \overline{X}_2$); nghĩa là, phân bố của sai biệt giữa hai trung bình mẫu.

Giả sử một mẫu ngẫu nhiên có cỡ mẫu $n_1 \geq 30$ được rút ra từ một tổng thể với trung bình μ_1 và độ lệch chuẩn σ_1, và một mẫu ngẫu nhiên khác có cỡ mẫu $n_2 \geq 30$ được rút ra từ một tổng thể có trung trung bình μ_2 và độ lệch chuẩn σ_2. Giả sử thêm là

hai mẫu được chọn độc lập, lúc đó biến ngẫu nhiên $\overline{X}_1 - \overline{X}_2$ có phân bố gần chuẩn với trung bình

$$\mu_{\overline{X}_1-\overline{X}_2} = \mu_{\overline{X}_1} - \mu_{\overline{X}_2} = \mu_1 - \mu_2$$

và độ lệch chuẩn

$$\sigma_{\overline{X}_1-\overline{X}_2} = \sqrt{\sigma^2_{\overline{X}_1} + \sigma^2_{\overline{X}_2}} = \sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}$$

Như vậy, biến ngẫu nhiên chuẩn tắc hóa

$$z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

có phân bố gần chuẩn tắc.

II. PHÂN BỐ CỦA SỰ SAI BIỆT GIỮA HAI TỈ LỆ (mẫu lớn và độc lập)

Giả sử một mẫu ngẫu nhiên cỡ mẫu n_1 được rút ra từ một tổng thể có tỉ lệ là p_1 và một mẫu ngẫu nhiên khác có cỡ mẫu n_2 được rút ra từ một tổng thể có tỉ lệ là p_2. Nếu hai mẫu được chọn độc lập, lúc đó biến ngẫu nhiên $\overline{P}_1 - \overline{P}_2$ có phân bố gần chuẩn với trung bình

$$\mu_{\overline{P}_1-\overline{P}_2} = \mu_{\overline{P}_1} - \mu_{\overline{P}_2} = p_1 - p_2$$

và độ lệch chuẩn

$$\sigma_{\overline{P}_1-\overline{P}_2} = \sqrt{\sigma^2_{\overline{P}_1} + \sigma^2_{\overline{P}_2}} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$$

Do đó, biến ngẫu nhiên chuẩn tắc hóa

$$z = \frac{(\overline{P}_1 - \overline{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}}$$

có phân bố xấp xỉ chuẩn tắc.

III. SO SÁNH HAI GIÁ TRỊ TRUNG BÌNH ĐỘC LẬP

1. Hai phương sai đã biết hoặc hai phương sai không biết, mẫu lớn ($n_1 > 30$ và $n_2 > 30$)

Với ví dụ về hai biện pháp phòng trị dịch sâu hại trên bắp, chúng ta muốn biết ở mức ý nghĩa 5%, số liệu ghi nhận được đủ để chứng minh có sự khác biệt về năng suất trung bình giữa hai biện pháp: phun thuốc trừ sâu và thả côn trùng đực bất dục hay không ?

Giả sử thuốc trừ sâu được phun trên 40 lô và côn trùng đực bất dục được thả trên 40 lô khác. Kết quả năng suất của từng lô được trình bày trong Bảng 4.1.

Vì cả hai mẫu đều lớn ($n_1 = n_2 = 40$) và độc lập; do đó, chúng ta có thể áp dụng phân bố của sai biệt giữa hai trung bình lấy mẫu để thực hiện kiểm định giả thiết. Các bước thực hiện như sau:

Đặt các giả thiết:

H_0: $\mu_1 = \mu_2$ (hai biện pháp cho năng suất trung bình bằng nhau)

H_a: $\mu_1 \neq \mu_2$ (hai biện pháp cho năng suất trung bình khác nhau)

Chọn mức ý nghĩa 5% để kiểm định giả thiết ($\alpha = 0,05$).

Vì $\alpha = 0,05$, trong Bảng z (Phụ lục II) tìm được các giá trị tới hạn (hai đuôi) $\pm z_{\alpha/2} = \pm z_{0,05/2} = \pm z_{0,025} = \pm 1,96$.

Bảng 4.1: Năng suất (kg/lô) của 40 lô bắp được phun thuốc trừ sâu và 40 lô khác được thả côn trùng đực bất dục.

Phun thuốc trừ sâu					Thả côn trùng đực bất dục				
109	101	97	89	100	105	109	110	118	109
98	98	94	99	104	113	111	111	99	112
103	88	108	102	106	106	117	99	107	119
97	105	102	104	101	110	111	103	110	108
101	100	105	110	96	104	102	111	114	114
102	95	100	95	109	122	117	101	109	109
91	98	113	91	95	102	109	103	109	106
106	98	101	99	96	107	107	111	128	109

Từ số liệu ở Bảng 4.1, chúng ta tính được trung bình và độ lệch chuẩn của hai mẫu như sau:

$$\overline{X}_1 = \frac{\sum X_1}{n_1} = \frac{4006}{40} = 100,150 \quad \text{và} \quad s_1^2 = \frac{402480 - (4006)^2/40}{39} = 32,797$$

$$\overline{X}_2 = \frac{\sum X_2}{n_2} = \frac{4381}{40} = 109,525 \quad \text{và} \quad s_2^2 = \frac{481261 - (4381)^2/40}{39} = 36,717$$

Do đó, giá trị của số thống kê kiểm định là

$$z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{100,150 - 109,525}{\sqrt{(32,797 + 36,717)/40}} = -7,11$$

So sánh giá trị $z = -7,11$ với các giá trị tới hạn $\pm z_{0,05/2} = \pm 1,96$ cho thấy z nằm trong vùng bác bỏ giả thuyết H_0.

Như vậy, số liệu mẫu đủ dẫn chứng để kết luận có sự khác biệt về năng suất trung bình giữa hai biện pháp: phun thuốc trừ sâu và thả côn trùng đực bất dục. Nói cách khác, biện pháp phun thuốc trừ sâu cho năng suất thấp hơn thả côn trùng đực bất dục.

2. Hai phương sai không biết, mẫu nhỏ

a) Hai phương sai bằng nhau

Khi các mẫu độc lập được rút ra từ hai tổng thể có phân bố chuẩn với các phương sai bằng nhau ($\sigma_1^2 = \sigma_2^2$), lúc đó biến ngẫu nhiên

$$z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

có phân bố chuẩn tắc.

Gọi σ là độ lệch chuẩn chung của hai tổng thể. Thay σ_1 và σ_2 bằng σ, chúng ta có được biến ngẫu nhiên

$$z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{(1/n_1) + (1/n_2)}}$$

Dĩ nhiên, biến ngẫu nhiên này không thể được dùng như số thống kê kiểm định vì độ lệch chuẩn chung (σ) không biết.

Vì vậy, chúng ta cần sử dụng thông tin từ mẫu để ước lượng độ lệch chuẩn không biết của tổng thể (σ). Cách tốt nhất để làm điều này là gom các tổng bình phương và độ tự do của hai mẫu để ước lượng độ lệch chuẩn chung (s_p)

$$s_p = \sqrt{\frac{\sum(X_1 - \overline{X}_1)^2 + \sum(X_2 - \overline{X}_2)^2}{(n_1 - 1) + (n_2 - 1)}}$$

Hoặc có thể ước lượng độ lệch chuẩn chung của tổng thể (s_p) bằng cách tính trung bình có gia trọng (weighed mean) theo độ tự do từ hai phương sai mẫu (s_1^2 và s_2^2). Công thức tính như sau:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Đại lượng s_p được gọi là độ lệch chuẩn mẫu gộp (pooled sample standard deviation) và dùng độ lệch chuẩn này như số ước lượng cho độ lệch chuẩn chung (σ), nhưng không biết, của hai tổng thể.

Thay độ lệch chuẩn tổng thể không biết (σ) bằng số ước lượng của nó (s_p), chúng ta có được biến ngẫu nhiên

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{(1/n_1) + (1/n_2)}}$$

để có thể dùng như số thống kê kiểm định.

Tóm lại, giả sử hai mẫu ngẫu nhiên độc lập, lần lượt có cỡ mẫu n_1 và n_2 được rút ra từ hai tổng thể có phân bố chuẩn với trung trung bình μ_1 và μ_2. Giả sử thêm là độ lệch chuẩn của hai tổng thể bằng nhau. Lúc đó biến ngẫu nhiên.

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{(1/n_1) + (1/n_2)}}$$

có phân bố t với độ tự do, $df = n_1 + n_2 - 2$.

Ví dụ: Một nhà nghiên cứu muốn biết có sự khác biệt về năng suất trung bình của hai giống đậu hay không? Ông ghi nhận năng suất của giống A trên 24 lô ruộng và giống B trên 27 lô ruộng. Kết quả được trình bày trong Bảng 4.2.

Bảng 4.2: Năng suất (tạ/ha) của 24 lô đậu giống A và 27 lô đậu giống B

Giống A				Giống B				
22,1	20,4	23,3	20,9	16,8	21,3	20,6	21,0	19,1
22,5	25,3	26,8	22,6	23,8	18,3	17,7	18,0	20,8
21,0	22,0	21,7	23,0	21,1	19,6	19,0	18,7	20,5
25,4	25,9	24,8	25,3	20,4	17,1	19,5	20,5	
24,7	20,4	22,4	24,5	19,2	19,1	17,2	16,3	
26,6	25,4	25,8	24,2	22,7	20,0	19,2	16,3	

Các bước thực hiện:

(1) Kiểm định hai phương sai:

H_0: $\sigma_1^2 = \sigma_2^2$

H_a: $\sigma_1^2 > \sigma_2^2$

Tính giá trị của số thống kê kiểm định F

$$F = \frac{s_{lớn}^2}{s_{nhỏ}^2} = \frac{4,02}{3,45} = 1,17$$

Vì trong công thức tính giá trị F, s_1^2 nằm ở tử số, nên độ tự do của tử số là $n_1 - 1 = 23$ và s_2^2 nằm ở mẫu số, nên độ tự do của mẫu số là $n_2 - 1 = 26$. Giá trị tới hạn cho kiểm định đuôi phải $F_{0,05;23;26}$ tìm được trong Phụ lục VI là 1,96.

Giá trị của số thống kê $F = 1,17 < F_{0,05;23;26} = 1,96$ nên không thể loại giả thuyết H_0. Vì thế kết luận hai phương sai bằng nhau.

(2) Kiểm định hai trung bình:

Đặt các giả thuyết:

H_0: $\mu_1 = \mu_2$ (Năng suất trung bình của hai giống đậu bằng nhau)

H_a: $\mu_1 \neq \mu_2$ (Năng suất trung bình của hai giống đậu khác nhau)

Kiểm định được thực hiện ở mức ý nghĩa 5% ($\alpha = 0,05$).

Các giá trị tới hạn cho kiểm định hai đuôi là $\pm t_{\alpha/2}$ ở độ tự do (df) là $n_1 + n_2 - 2$.

Trong vì dụ, $n_1 = 24$ và $n_2 = 27$, nên df = 24 + 27 - 2 = 49. Tra trong Bảng t (Phụ lục IV) ở $\alpha = 0,05$, chúng ta tìm được $\pm t_{\alpha/2} = \pm t_{0,05/2} = \pm t_{0,025} = \pm 2,0116$.

Từ số liệu ở Bảng 4.2, chúng ta tính được trung bình và phương sai của hai mẫu :

$\overline{X}_1 = 23,625$ và $s_1^2 = 4,02$; $\overline{X}_2 = 19,4$ và $s_2^2 = 3,45$. Thế vào công thức tính phương sai mẫu gộp:

$$s_p^2 = \frac{23.(4,02) + 26.(3,45)}{49} = 3,72$$

$$s_p = \sqrt{s_p^2} = \sqrt{3,72} = 1,93$$

Tính giá trị của số thống kê kiểm định t:

$$t = \frac{23,625 - 19,4}{1,93\sqrt{(1/24) + (1/27)}} = 7,8$$

So với giá trị $t_{0,05/2} = 2,012$, giá trị của số thống kê t = 7,8 nằm trong vùng bác bỏ giả thuyết H_0, nên số liệu cung cấp đủ dẫn chứng để kết luận năng suất trung bình của hai giống đậu khác nhau, hoặc giống A cho năng suất cao hơn giống B.

*** Trường hợp $n_1 = n_2$**

Trường hợp đặc biệt, khi hai mẫu ngẫu nhiên, độc lập có cỡ mẫu bằng nhau ($n_1 = n_2 = n$), các bước thực hiện cũng tương tự như trên, với

$$s_p^2 = \frac{s_1^2 + s_2^2}{2} \quad \text{và} \quad t = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{2s_p^2/n}}$$

Ví dụ: Thí nghiệm tăng trọng lượng bò thịt (kg) ở hai điều kiện thức ăn: có vitamin A và không có vitamin A. Kết quả ghi nhận được như sau:

Không Vit.A					Có Vit.A				
175	132	218	151	200	142	311	337	262	302
219	206	234	149	187	195	214	253	199	236
123	248	206	179		216	211	176	249	

Từ số liệu trên, chúng ta tính được trung bình và phương sai của hai mẫu :

$\overline{X}_1 = 187,6$ và $s_1^2 = 1451,478$; $\overline{X}_2 = 236,0$ và $s_2^2 = 2946,995$.

(1) Kiểm định hai phương sai:

H_0: $\sigma_1^2 = \sigma_2^2$ đối với H_a: $\sigma_1^2 < \sigma_2^2$

Tính giá trị của số thống kê kiểm định F

$$F = \frac{s^2_{max}}{s^2_{min}} = \frac{2946,995}{1451,478} = 2,03$$

Trong ví dụ, $n_1 = n_2 = 14$ nên độ tự do của tử số và mẫu số bằng nhau ($n_1 - 1 = n_2 - 1 = 13$). Giá trị tới hạn $F_{0,05;,13;13}$ tìm được trong Bảng F (Phụ lục VI) là 2,575.

Do giá trị của số thống kê $F = 2,03 < F_{\alpha,13,13} = 2,575$, nên không thể loại giả thuyết H_0. Vì thế kết luận hai phương sai bằng nhau.

(2) Kiểm định hai trung bình:

Đặt các giả thuyết:

H_0: $\mu_1 = \mu_2$ (Trọng lượng bò thịt ở hai điều kiện thức ăn: có và không có vitamin A bằng nhau)

H_a: $\mu_1 \neq \mu_2$ (Trọng lượng bò thịt ở hai điều kiện thức ăn: có và không có vitamin A khác nhau)

Kiểm định được thực hiện ở mức ý nghĩa 5% ($\alpha = 0,05$).

Các giá trị tới hạn cho kiểm định hai đuôi là $\pm t_{\alpha/2}$ ở $df = 2(n - 1)$.

Trong ví dụ, $n_1 = n_2 = 14$, nên $df = [2.(n - 1)] = 26$. Tra trong Bảng t (Phụ lục IV) ở $\alpha = 0,05$, giá trị $\pm t_{\alpha/2} = \pm t_{0,05/2} = \pm t_{0,025} = \pm 2,056$.

Thế vào các công thức:

$$s_p^2 = \frac{(1451,478 + 2946,995)}{2} = 2199,237$$

$$t = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{2s_p^2/n}} = \frac{(187,6 - 236,0)}{\sqrt{2.(2199,237)/14}} = -2,731$$

So với giá trị $\pm t_{0,05/2} = \pm 2,056$, giá trị của số thống kê $t = -2,731$ nằm trong vùng bác bỏ giả thuyết H_0.

Như vậy, số liệu cung cấp đủ dẫn chứng để kết luận khi thêm vitamin A vào thức ăn sẽ làm tăng trọng lượng bò thịt. Thông thường nếu vitamin A đắt tiền, để tránh sai lầm khi bác bỏ giả thuyết H_0 đúng, ta phải giảm vùng bác bỏ (giảm α). Nếu vitamin A rẽ tiền có thể nới rộng vùng bác bỏ (tăng α).

b) Hai phương sai không bằng nhau

Với phương pháp trên, giả định để áp dụng phương sai gộp là các tổng thể có phương sai bằng nhau. Nếu các tổng thể có phương sai không bằng nhau hoặc chúng ta không biết chúng có bằng nhau hay không, lúc đó không nên sử dụng phương sai gộp. Thay vào đó, chúng ta nên áp dụng phương sai không gộp (phương pháp Behrens-Fisher).

Giả sử hai mẫu ngẫu nhiên độc lập, lần lượt có cỡ mẫu n_1 và n_2 được rút ra từ hai tổng thể có phân bố chuẩn với trung trung bình μ_1 và μ_2 và độ lệch chuẩn của chúng không chắc bằng nhau. Lúc đó biến ngẫu nhiên.

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(s_1^2/n_1) + (s_2^2 + n_2)}}$$

gần bằng phân bố t với độ tự do (df) được tính như sau:

$$df = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

chọn df quanh số nguyên gần nhất

Ví dụ: Một thí nghiệm viên muốn biết có sự khác biệt về sản lượng trung bình giữa hai mức phân đạm (50 và 100 kg N/ha) trên một giống dưa hấu hay không? Ông ghi nhận sản lượng dưa trên 17 lô được bón 50 kgN/ha và 13 lô bón 100 kgN/ha. Kết quả được trình bày trong Bảng 4.3.

Bảng 4.3: Sản lượng dưa hấu của các lô được bón 50 và 100 kgN/ha

Bón 50kg N/ha				Bón 100 kg N/ha		
3,7	1,6	2,4	3,1	2,5	2,5	2,2
2,0	2,5	3,1	3,8	2,4	2,5	3,0
2,3	2,2	2,3		2,6	2,2	2,8
3,0	1,7	3,2		2,6	3,4	
2,4	2,7	1,5		3,3	2,9	

Từ số liệu Bảng 4.3, chúng ta tính được $\overline{x}_1 = 2,56$ và $\overline{x}_2 = 2,68$; $s_1^2 = 0,466$ và $s_2^2 = 0,143$

Các bước thực hiện:

(1) Kiểm định hai phương sai:

H_0: $\sigma_1^2 = \sigma_2^2$ đối với H_a: $\sigma_1^2 > \sigma_2^2$

Vì $s_1^2 > s_2^2$, nên ($n_1 - 1 = 16$) là độ tự do của tử số và ($n_2 - 1 = 12$) là độ tự do của mẫu số. Giá trị tới hạn cho kiểm định đuôi phải $F_{\alpha,16,12}$ tìm được trong Bảng F (Phụ lục VI) là 2,60.

Tính giá trị của số thống kê kiểm định F

$$F = \frac{s_{max}^2}{s_{min}^2} = \frac{0,466}{0,143} = 3,26$$

Giá trị của số thống kê kiểm định là $F = 3,26 > F_{0,05;,16;12} = 2,60$ nên loại giả thuyết H_0 và kết luận hai phương sai không bằng nhau.

(2) Kiểm định hai trung bình:

Đặt các giả thuyết:

H_0: $\mu_1 = \mu_2$ (Hai mức phân đạm cho năng suất trung bình bằng nhau)

Ha: $\mu_1 \neq \mu_2$ (Hai mức phân đạm cho năng suất trung bình khác nhau)

Chọn mức ý nghĩa kiểm định là 5% ($\alpha = 0,05$).

Tính giá trị của số thống kê kiểm định

$$t = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{(s_1^2/n_1) + (s_2^2 + n_2)}} = \frac{2,56 - 2,68}{\sqrt{\left(\frac{0,466}{17} + \frac{0.143}{13}\right)}} = -0,612$$

với độ tự do được điều chỉnh như sau:

$$df = \frac{\left(\frac{0,466}{17} + \frac{0,143}{13}\right)^2}{\frac{\left(\frac{0,466}{17}\right)^2}{16} + \frac{\left(\frac{0,143}{13}\right)^2}{12}} = 25,9 \approx 26$$

Với độ tự do bằng 26, tìm trong Bảng t (Phụ lục IV) ở $\alpha = 0,05$, các giá trị tới hạn $\pm t_{\alpha/2} = \pm t_{0,05/2} = \pm t_{0,025} = \pm 2,056$.

So với giá trị $\pm t_{0,05/2} = \pm 2,056$, $|t| = |-0,612|$ nằm trong vùng chấp nhận giả thuyết H_0. Như vậy, thí nghiệm viên có thể kết luận rằng bón phân đạm ở hai mức 50 và 100 kg N/ha cho sản lượng dưa hấu trung bình khác biệt không ý nghĩa.

Chú ý, số thống kê kiểm định (t) trong trường hợp này chỉ là phương pháp gần đúng và thường cho kết luận rất dè dặt (nghĩa là, khó bác bỏ giả thuyết không); do đó, nếu có thể nên tránh sử dụng phương pháp này mà thay vào đó nên dùng cỡ mẫu lớn hơn.

IV. SO SÁNH HAI TRUNG BÌNH BẮT CẶP *(Paired Comparisons)*

Trong phần này, chúng ta sẽ thảo luận phương pháp so sánh các trung bình của hai tổng thể bằng cách dùng các mẫu được bắt cặp.

Chẳng hạn, chúng ta muốn kết luận sai biệt trung bình giữa điểm kiểm tra trước và sau khi tham dự một khóa học, hoặc trung bình sai biệt về thời gian phản ứng với thuốc khi cùng đối tượng nhận hoặc không nhận một loại thuốc nào đó. Trong những trường hợp như thế, người làm thí nghiệm sẽ có hai bộ số liệu mẫu; tuy nhiên, cả hai bộ phải được đo từ những đối tượng giống nhau, mục đích chính là loại bỏ khỏi thí nghiệm nguồn biến động bên ngoài.

Để so sánh hai trung bình được bắt cặp, bộ số liệu mẫu phải có những số hạng tương ứng; nghĩa là, nhất thiết phải có một bộ gồm các cặp số liệu. Sự khác biệt về trung bình giữa hai mẫu bắt cặp có thể được kết luận bằng cách tính toán các sai biệt bên trong các cặp và dùng phân bố t với (n - 1) độ tự do (n là số cặp).

Giả sử một mẫu ngẫu nhiên có n cặp được rút ra từ các tổng thể có trung bình μ_1 và μ_2. Giả sử thêm rằng tổng thể của tất cả sai biệt được bắt cặp có phân bố chuẩn. Lúc đó biến ngẫu nhiên,

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$$

có phân bố t với df = n - 1

Ví dụ: Để so sánh số hạt của các trái trên ngọn và dưới gốc ở đậu xanh. Trên mỗi cây, người ta lấy những trái ngọn và dưới gốc, tính trung bình và chọn ngẫu nhiên ra một cặp và tiếp tục làm như thế ở những cây khác. Kết quả được ghi nhận trong Bảng 4.4.

Gọi μ_d là sai biệt trung bình về số hạt của các trái ở ngọn và ở gốc của cây đậu xanh.

* Đặt các giả thuyết:

H_0: $\mu_d = 0$ (Không có sự khác biệt về số hạt của trái ở ngọn và ở gốc)

H_a: $\mu_d \neq 0$ (Có sự khác biệt về số hạt của trái ở ngọn và ở gốc)

* Giả sử kiểm định được thực hiện ở mức ý nghĩa 5% ($\alpha = 0,05$).

* Vì $\alpha = 0,05$ và mẫu có 10 cặp, nên df = 10 - 1 = 9. Các giá trị tới hạn cho kiểm định hai đuôi tìm được trong Bảng t (Phụ lục IV) là $\pm t_{\alpha/2} = \pm t_{0,05/2} = \pm t_{0,025} = \pm 2,262$.

* Tính trung bình và độ lệch chuẩn của mẫu dựa trên các giá trị d của Bảng 4.4.

$$\bar{d} = \frac{\sum d}{n} = \frac{10,7}{10} = 1,07$$

$$s_d = \sqrt{\frac{\sum d^2 - (\sum d)^2/n}{n-1}} = \sqrt{\frac{11,89 - (10,7)^2/10}{9}} = 0,221$$

Bảng 4.4: Phân phối số hạt của trái ở ngọn và ở gốc của 10 cây đậu xanh và sự sai biệt của chúng.

Số hạt / trái			
Ngọn (X_1)	Gốc (X_2)	d = (X_1 - X_2)	d^2
6,2	5,2	1,0	1,00
6,5	5,4	1,1	1,21
6,7	5,3	1,4	1,96
6,9	5,7	1,2	1,44
6,9	5,6	1,3	1,69
7,0	6,1	0,9	0,81
6,8	5,7	1,1	1,21
6,7	5,6	1,1	1,21
6,8	5,8	1,0	1,00
6,2	5,6	0,6	0,36
$\overline{X}_1 = 6,7$	$\overline{X}_2 = 5,6$	$\sum d = 10,7$	$\sum d^2 = 11,89$

* Tính giá trị của số thống kê kiểm định

$$t = \frac{\overline{d}}{s_d / \sqrt{n}} = \frac{1,07}{0,221/\sqrt{10}} = 15,4$$

* So với giá trị $\pm t_{0,05/2} = \pm 2,262$, giá trị t = 15,4 nằm trong vùng bác bỏ giả thuyết H_0. Như vậy, số liệu cung cấp đủ dẫn chứng để kết luận số hạt của các trái ở ngọn và ở gốc khác nhau. Do $\overline{d}$ có giá trị dương, nên chúng ta có thể kết luận số hạt của các trái ở ngọn nhiều hơn ở gốc.

Kiểu sắp xếp cặp là cách để kiểm soát nguồn biến động bên ngoài. Cũng với thí nghiệm trên, nếu người ta thực hiện trên 20 cây: 10 cây đếm số hạt trong trái ở gốc và 10 cây khác đếm số hạt trong trái ở ngọn; như vậy, sẽ có nguồn biến động khác được đưa vào do hai nhóm số liệu đếm trên hai cây khác nhau. Nếu giữa hai cây khác nhiều về khả năng tăng trưởng, mặc dầu chúng cùng giống, trung bình hai nhóm có thể sẽ khác nhau.

V. SO SÁNH HAI TỶ LỆ (Mẫu độc lập)

Để so sánh hai tỉ lệ, giả thuyết không được đặt ra là H_0: $p_1 = p_2$ (hai tỷ lệ tổng thể bằng nhau). Nếu giả thuyết H_0 đúng, lúc đó $p_1 - p_2 = 0$ và biến chuẩn tắc hóa trở thành

$$z = \frac{(\overline{P}_1 - \overline{P}_2)}{\sqrt{\frac{p_P(1-p_P)}{n_1} + \frac{p_P(1-p_P)}{n_2}}} \quad \text{hoặc} \quad z = \frac{(\overline{P}_1 - \overline{P}_2)}{\sqrt{p_P \cdot (1-p_P) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

với P_p là giá trị chung của p_1 và p_2, được gọi là **tỷ lệ mẫu gộp** (pooled sample proportion)

$$p_p = \frac{X_1 + X_2}{n_1 + n_2}$$

Ví dụ: Để so sánh tỷ lệ nẩy mầm của 2 lô hạt giống A và B, ở mỗi lô người ta lấy 100 hạt đem thử tỷ lệ nảy mầm. Sau 2 tuần người ta ghi nhận kết quả như sau: Lô A có 85 hạt nẩy mầm và nhóm B có 75 hạt nẩy mầm. Hãy kiểm định xem 2 lô hạt có tỷ lệnẩy mầm khác nhau không?

Gọi p_1 và p_2 lần lượt là xác suất nảy mầm của lô hạt giống A và B. Để kiểm định giả thuyết, chúng ta đặt các giả thuyết:

H_0: $p_1 = p_2$ đối với H_a: $p_1 \neq p_2$

Kiểm định giả thuyết được thực hiện ở mức ý nghĩa 5% ($\alpha = 0,05$).

* Tính các xác suất mẫu: $\overline{p}_1$ và $\overline{p}_2$.

$$\overline{p}_1 = \frac{85}{100} = 0,85 \quad \text{và} \quad \overline{p}_2 = \frac{75}{100} = 0,75$$

và xác suất chung

$$P_p = \frac{85 + 75}{200} = 0,8$$

* Tính số thống kê kiểm định

$$z = \frac{0,85 - 0,75}{\sqrt{(0,8)\cdot(0,2)\cdot\left(\frac{1}{100} + \frac{1}{100}\right)}} = \frac{0,1}{0,0566} = 1,77$$

Ở mức ý nghĩa 5%, trong Bảng z (Phụ lục II) giá trị tới hạn tìm được là $z_{0,05/2}$ = 1,96.

Vì giá trị z = 1,77 < giá trị $z_{0,05/2}$ = 1,96; nên không bác bỏ giả thuyết H_0.

Kết luận: Tỷ lệ nảy mầm của hai lô hạt giống khác biệt không ý nghĩa.

VI. KHOẢNG TIN CẬY CỦA SAI BIỆT GIỮA HAI TRUNG BÌNH TẬP HỢP

1. Trường hợp mẫu lớn và độc lập

Khoảng tin cậy 1 - α ($CI_{1-\alpha}$) của sai biệt hai trung bình ($\mu_1 - \mu_2$) là

$$(\overline{X}_1 - \overline{X}_2) \pm z_{\alpha/2}.\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}$$

Ví dụ: Tìm khoảng tin cậy 95% cho sai biệt về năng suất trung bình ($\mu_1 - \mu_2$) giữa hai biện pháp: phun thuốc trừ sâu và thả côn trùng đực bất dục.

Đối với khoảng tin cậy 95%, mức độ tin cậy 1 - α = 0,95 ⇒ α = 0,05. Từ Phụ lục II cho giá trị $z_{\alpha/2} = z_{0,05/2} = z_{0,025} = 1,96$.

Với: $n_1 = n_2 = 40$; $\overline{X}_1 = 100,15$; $s_1^2 = 32,797$ và $\overline{X}_2 = 109,525$; $s_2^2 = 36,717$. Do đó, khoảng tin cậy 95% của ($\mu_1 - \mu_2$) là

$(100,15 - 109,525) \pm 1,96.\sqrt{(32,797 + 36,717)/40)}$

hoặc $-9,375 \pm 2,584$

hoặc $-11,959 \leq (\mu_1 - \mu_2) \leq -6,791$

hoặc $6,791 \leq (\mu_2 - \mu_1) \leq 11,959$

Như vậy, chúng ta có thể tin cậy 95% rằng sự sai biệt về năng suất trung bình ($\mu_2 - \mu_1$) giữa hai biện pháp: phun thuốc trừ sâu và thả côn trùng đực bất dục biến động từ 6,791 đến 11,959 kg/lô.

2. Trường hợp mẫu nhỏ và độc lập

a) Phương sai mẫu gộp

*** Trường hợp $n_1 \neq n_2$**

Khoảng tin cậy (1 - α) của sai biệt hai trung bình ($\mu_1 - \mu_2$) trong trường hợp này là

$$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2}.s_p\sqrt{(1/n_1) + (1/n_2)}$$

với $t_{\alpha/2}$ được tìm trong Bảng t ở df = $n_1 + n_2 - 2$

Ví dụ: Để tìm khoảng tin cậy 95% cho sai biệt về năng suất trung bình (μ_1 - μ_2) giữa hai giống đậu A và B. Các bước thực hiện như sau:

Đối với khoảng tin cậy 95%, mức độ tin cậy 1 - α = 0,95 $\Rightarrow$ α = 0,05. Trong Bảng t (Phụ lục IV) ở df = 24 + 27 - 2 = 49, giá trị $t_{\alpha/2} = t_{0,05/2} = t_{0,025} = 2,0116$.

Thế vào công thức ta được khoảng tin cậy 95%:

$$(23,625 - 19,4) \pm (2,0116)(1,93)\sqrt{(1/24) + (1/27)}$$

hoặc $$4,225 \pm 1,089$$

$$3,136 \le (\mu_1 - \mu_2) \le 5,314$$

Như vậy, chúng ta có thể tin cậy 95% rằng sự sai biệt về năng suất trung bình $(\mu_2 - \mu_1)$ giữa 2 giống đậu A và B biến động giữa 3,136 và 5,314 tạ/ha.

*** Trường hợp $n_1 = n_2 = n$**

Khoảng tin cậy (1 - α) của sai biệt hai trung bình $(\mu_1 - \mu_2)$ trong trường hợp này là

$$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2} \cdot \sqrt{\frac{2s_p}{n}}$$

với $t_{\alpha/2}$ được tìm trong Bảng t ở df = 2(n – 1)

Ví dụ: Tìm khoảng tin cậy 95% cho sai biệt về trọng lượng trung bình (μ_1 - μ_2) giữa hai khẩu phần thức ăn: có và không có vitamin A ở bò thịt. Các bước thực hiện như sau:

Vì $n_1 = n_2 = n = 14$, nên độ tự do df = 2(n – 1) = 26. Tìm trong Bảng t (Phụ lục IV), giá trị $t_{\alpha/2} = t_{0,05/2} = t_{0,025} = 2,056$.

Thế $\overline{X}_1 = 187,6$; $\overline{X}_2 = 236,0$ và $s_p = 2199,237$ vào công thức ta được khoảng tin cậy 95%:

$$(187,6 - 236,0) \pm 2,056 \cdot \sqrt{(2)(2199,237)/14}$$

$$-48,4 \pm 36,44$$

$$-84,84 \le (\mu_1 - \mu_2) \le -11,96$$

hoặc $$11,96 \le (\mu_2 - \mu_1) \le 84,84$$

Như vậy, ở mức độ tin cậy 95%, sự sai biệt về trọng lượng trung bình $(\mu_2 - \mu_1)$ giữa hai khẩu phần thức ăn: có và không có vitamin A ở bò thịt biến động trong khoảng 11,96 và 84,84 kg.

b) Phương sai mẫu không gộp

Khoảng tin cậy (1 - α) của sai biệt hai trung bình $(\mu_1 - \mu_2)$ trong trường hợp này là

$$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

với giá trị $t_{\alpha/2}$ tìm ở độ tự do $$df = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Ví dụ: Tìm khoảng tin cậy 95% cho sai biệt về sản lượng trung bình (μ_1 - μ_2) giữa hai mức phân đạm (50 và 100 kg N/ha).

Ta có: $n_1 = 17$; $n_2 = 13$; $\overline{X}_1 = 2,56$; $\overline{X}_2 = 2,68$; $s_1^2 = 0,466324$ và $s_2^2 = 0,143077$.

$$df = \frac{\left(\frac{0,466}{17} + \frac{0,143}{13}\right)^2}{\frac{\left(\frac{0,466}{17}\right)^2}{16} + \frac{\left(\frac{0,143}{13}\right)^2}{12}} = 25,9 \approx 26$$

Ở độ tự do, df = 26, tìm trong Phụ lục IV, giá trị $t_{\alpha/2} = t_{0,05/2} = t_{0,025} = 2,056$

Thế vào công thức ta được khoảng tin cậy 95%:

$$(2,56 - 2,68) \pm 2,056.\sqrt{(0,466324\ /17) + (0,143077\ /13)}$$

$$-0,12 \pm 0,40$$

$$-0,52 \le \mu_1 - \mu_2 \le 0,28$$

hoặc $-0,28 \le \mu_2 - \mu_1 \le 0,52$

Như vậy, ở mức độ tin cậy 95%, sự sai biệt về sản lượng trung bình ($\mu_1 - \mu_2$) giữa hai mức phân đạm (50 và 100 kg N/ha) biến động từ - 0,28 đến 0,52 tấn/ha.

3. Trường hợp hai mẫu bắt cặp

Khoảng tin cậy (1 - α) của trung bình sai biệt (μ_d) là:

$$\bar{d} \pm z_{\alpha/2}.\frac{\sigma_d}{\sqrt{n}} \quad (\text{với } n \ge 30)$$

hoặc $\bar{d} \pm t_{\alpha/2}.\frac{s_d}{\sqrt{n}}$ (với n < 30)

Ví dụ: Tìm khoảng tin cậy 95% cho sai biệt về số hạt của các trái ở ngọn và ở gốc (μ_d).

Ở df = 10 - 1 = 9 và α = 0,05, tìm trong Phụ lục IV, $t_{\alpha/2} = \pm t_{0,05/2} = \pm t_{0,025} = \pm$ 2,262. Với: n = 10, $\bar{d} = 1,07$ và $s_{\bar{d}} = 0,221$. Ta có khoảng tin cậy của (μ_d) là

$$1,07 \pm 2,262.\frac{0,221}{\sqrt{10}}$$

hoặc $1,07 \pm 0,16.$

$$0,91 \le (\mu_d) \le 1,23.$$

Như vậy, chúng ta có thể tin cậy 95% rằng sự sai biệt giữa số hạt trong trái của các trái ở ngọn và ở gốc (μ_d) khoảng 1 hạt.

VII. KHOẢNG TIN CẬY CỦA SAI BIỆT HAI TỶ LỆ

Khoảng tin cậy (1 - α) của sai biệt hai tỷ lệ(p_1 - p_2) là:

$$(\bar{p}_1 - \bar{p}_2) \pm z_{\alpha/2}.\sqrt{P_p(1 - P_p).(1/n_1) + (1/n_2)}$$

Ví dụ: Tìm khoảng tin cậy 95% cho sai biệt về tỷ lệnảy mầm (p_1 - p_2) của hai lô hạt giống A và B.

Với $n_1 = n_2 = 100$; $\bar{p}_1 = 0,85$; $\bar{p}_2 = 0,75$ và $p_P = 0,8$. Thế vào công thức ta có:

$$(0,85 - 0,75) \pm (1,96).\sqrt{(0,8)\cdot(0,2)\cdot[(1/100) + (1/100)]}$$

hoặc $0,1 \pm 0,11$

$$-0,01 \le (p_1 - p_2) \le 0,21$$

Như vậy, chúng ta có thể tin cậy 95% rằng sự sai biệt về tỷ lệnảy mầm (p_1 - p_2) giữa hai lô hạt giống biến động từ - 1% đến 21%.

VIII. XÁC ĐỊNH CỠ MẪU

1. Xác định cỡ mẫu trong trường hợp hai trung bình độc lập

Ở đây chúng ta chỉ khảo sát trường hợp hai mẫu có cùng cỡ mẫu. Với giả định độ lệch chuẩn đã biết, chúng ta có thể sử dụng công thức sau đây cho $\alpha \leq 0,1$ và $\beta \leq 0,5$:

$$n = \frac{2.(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma^2}{\delta^2}$$

Với $\alpha = 0,05$ và $\beta = 0,5$ và $\beta = 0,1$:

$$n \approx \frac{8.\sigma^2}{\delta^2} \quad \text{và} \quad n \approx \frac{21.\sigma^2}{\delta^2}$$

Khi $n \leq 10$, nếu biết hệ số biến động (V) và sai số tương đối của sự sai biệt δ_r, được biểu thị bằng phần trăm của trung bình chung, công thức trên trở thành:

$$n = \frac{2.(z_{1-\alpha/2} + z_{1-\beta})^2 V^2}{\delta_r^2}$$

Với $\alpha = 0,05$ và $\beta = 0,5$ và $\beta = 0,1$:

$$n \approx \frac{8.V^2}{\delta_r^2} \quad \text{và} \quad n \approx \frac{21.V^2}{\delta_r^2}$$

Sự liên hệ giữa n, V và δ_r, được trình bày qua đồ biểu bàn tính (Hình 4.2) với $\alpha = 0,05$ và $\beta = 0,1$.

Chúng ta cũng có thể xác định gần đúng giá trị quan sát (n) tương ứng với trường hợp $\alpha = 0,05$ và $\beta = 0,5$ bằng cách nhân cho 2 các kết quả suy ra từ đồ biểu (Hình 3.10, chương 3). So sánh sự liên hệ ở trường hợp này với trường hợp một trung bình, chúng ta nhận thấy cần quan sát số cá thể trong mỗi mẫu nhiều hơn khoảng hai lần (nếu chọn $\beta = 0,5$). Tương tự, để thực hiện kiểm định tính bằng nhau của hai trung bình bằng cách giảm sai lầm loại II đến 10% (nghĩa là, $\beta = 0,1$), cần lấy cỡ mẫu nhiều hơn khoảng 5 lần trong trường hợp ước lượng một trung bình.

Ví dụ: Xác định số quan sát (n).trong trường hợp so sánh trọng lượng trung bình của hai quần thể gà (Sử dụng số liệu của Bảng 3.4, chương 3)

Giả sử chúng ta mong muốn so sánh trọng lượng trung bình của hai quần thể gà, với hy vọng sai số tương đối không vượt quá 10%, chọn $\alpha = 0,05$ và $\beta = 0,1$. Trong trường hợp này, chúng ta phải bố trí hai lô có cỡ mẫu:

$$n \approx 21.\frac{13^2}{10^2} = 35$$

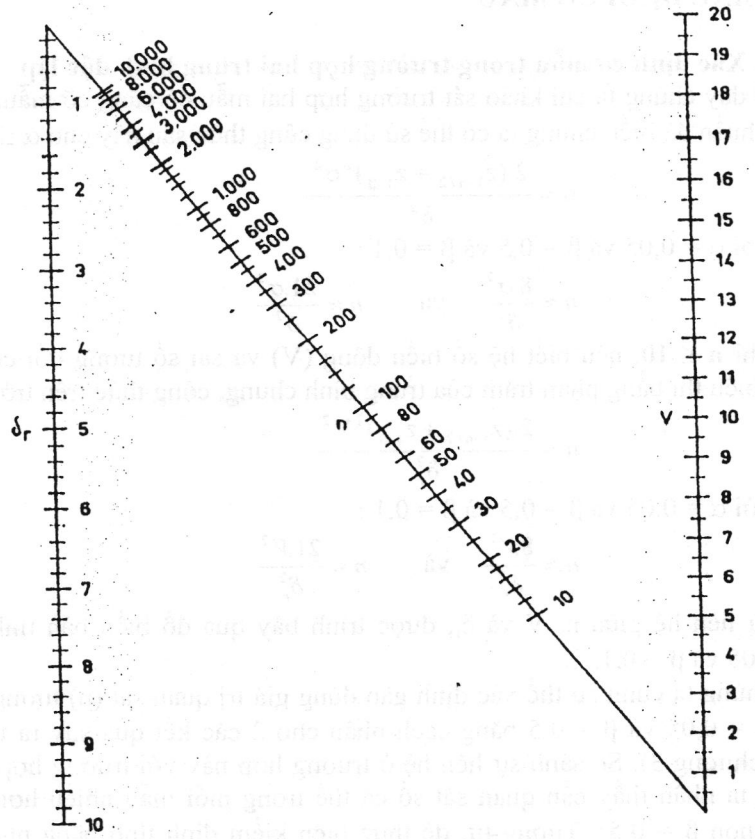
Tương tự, nếu chúng ta mong muốn sai số tương đối không vượt quá 5%, chúng ta phải khảo sát các mẫu mà mỗi mẫu có khoảng 140 con gà.

$$n \approx 21.\frac{13^2}{5^2} \approx 140$$

Nếu chúng ta chỉ bố trí các lô 50 con gà, chúng ta chỉ có cơ hội trên hai của sự khác biệt rõ rệt như thế, vì sự quan hệ đảo cho (đối với $\beta = 0,5$)

$$\delta_r = \frac{13.\sqrt{8}}{\sqrt{50}} = 5,2 \approx 5$$

Tất cả kết quả này có thể tìm được nhờ đồ biểu ở Hình 3.8 ($\beta = 0,5$) và đồ biểu ở Hình 4.2 ($\beta = 0,1$) với điều kiện đừng quên nhân cho 2 các cỡ mẫu suy ra từ đồ biểu ở Hình 3.10. .



Hình 4.2: Sự liên hệ giữa số quan sát (n) của mỗi mẫu, hệ số biến động (V) và sai biệt tương đối tối đa (δ_r) làm nổi bật việc kiểm định tính bằng nhau của hai trung bình đối với xác suất sai lầm $\alpha = 0,05$ và $\beta = 0,1$.

2. Xác định cỡ mẫu trong trường hợp so sánh cặp.

Chúng ta có thể sử dụng công thức sau đây với $n \geq 10$, $\alpha \leq 0,1$ và $\beta \leq 0,5$

$$n = \frac{2.(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma^2}{\delta^2}$$

Khi **n ≥ 10**, $\alpha = 0,05$ và $\beta = 0,5$ và $\beta = 0,1$, công thức tính gần đúng như sau:

$$n \approx \frac{4.\sigma^2}{\delta^2} \quad \text{và} \quad n \approx \frac{11.\sigma^2}{\delta^2}$$

Ví dụ: Lấy lại ví dụ so sánh số hạt của trái ở ngọn và ở gốc. Xác định số quan sát n.

Giả sử độ lệch chuẩn của các sai khác về số hạt trong trái ở ngọn và ở gốc bằng với giá trị quan sát là 0,221

Từ 10 cặp quan sát, chúng ta có cơ hội mạnh để dẫn chứng sự khác biệt về số hạt của trái ở ngọn và ở gốc nhiều hơn hoặc bằng 0,2 hạt. Thật vậy, đối với $\alpha = 0,05$ và $\beta = 0,1$ chúng ta có:

$$\delta = \frac{0{,}221.\sqrt{11}}{\sqrt{10}} = 0{,}232$$

Trái lại, để tìm thấy sự khác biệt bằng 0,1 hạt, chúng ta phải đặt trong điều kiện này khoảng 54 cặp số đo, vì:

$$n \approx 11.\frac{0{,}221^2}{0{,}1^2} = 53{,}7$$

Các công thức được dùng để so sánh hai trung bình tổng thể (μ_1 và μ_2) được tóm tắt trong Bảng 4.3.

Bảng 4.3: Tóm tắt các công thức so sánh hai trung bình tổng thể, μ_1 và μ_2 (độc lập và bắt cặp. Giả thuyết cho tất cả các kiểm định là H_0: $\mu_1 = \mu_2$.

Loại Kiểm định	Các giả định	Số thống kê kiểm định	Các khoảng tin cậy
(2 mẫu lớn) z	1) Mẫu độc lập 2) Mẫu lớn	$z = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{(s_1^2/n_1)+(s_2^2+n_2)}}$	$(\overline{X}_1 - \overline{X}_2) \pm z_{\alpha/2} . \sqrt{(s_1^2/n_1)+(s_2^2/n_2)}$
(2 mẫu nhỏ) (Phương sai gộp) t	1) Mẫu độc lập 2) Tổng thể phân phối chuẩn 3) Độ lệch chuẩn bằng nhau	$t = \frac{(\overline{X}_1 - \overline{X}_2)}{s_p\sqrt{(1/n_1)+(1+n_2)}}$ df = $n_1 + n_2 - 2$	$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2} . \sqrt{s_p(1/n_1)+(1/n_2)}$ ♣ df = $n_1 + n_2 - 2$
(2 mẫu nhỏ) (Phương sai không gộp) t	1) Mẫu độc lập 2) Tổng thể có phân phối chuẩn	$t = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{(s_1^2/n_1)+(s_2^2+n_2)}}$	$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2} . \sqrt{(s_1^2/n_1)+(s_2^2/n_2)}$ *
(Mẫu nhỏ) t (bắt cặp)	1) Mẫu bắt cặp 2) Sai biệt có phân phối chuẩn	$t = \frac{\overline{d}}{s_d/\sqrt{n}}$ df = n - 1	$\overline{d} \pm t_{\alpha/2} . \frac{s_d}{\sqrt{n}}$ df = n - 1
(Mẫu lớn) z (bắt cặp)	1) Mẫu bắt cặp 2) Mẫu lớn	$z = \frac{\overline{d}}{s_d/\sqrt{n}}$	$\overline{d} \pm z_{\alpha/2} . \frac{s_d}{\sqrt{n}}$

♣ $s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$ * $df = \frac{[(s_1^2/n_1)+(s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$

CHƯƠNG 5

THÍ NGHIỆM MỘT NHÂN TỐ
(SINGLE-FACTOR EXPERIMENTS)

A. MỘT SỐ THUẬT NGỮ DÙNG TRONG BỐ TRÍ THÍ NGHIỆM

1. Đơn vị thí nghiệm *(Experimental unit)* hay lô thí nghiệm *(Experimental plot)* là nhóm vật liệu trên đó ta tác động một hoặc một số nhân tố nào đó mà ta muốn đo lường các ảnh hưởng của nó.

Ví dụ: Một lô đất, chậu đất, ống nghiệm, dĩa petrie, cây ăn trái, cây rừng, con thú, con người,...

2. Nhân tố *(Factor)* là nguyên nhân gây ảnh hưởng đến các giá trị quan sát. Nhân tố thí nghiệm là một biến độc lập có kiểm soát. Nó bao gồm các mức độ khác nhau được người làm thí nghiệm đặt ra.

Ví dụ: Giống, phân bón, thuốc sát trùng, thuốc diệt cỏ, phương pháp canh tác, kỹ thuật chăn nuôi, chế biến,...

3. Nghiệm thức *(Treatment)* có thể bao gồm các mức độ khác nhau của một nhân tố hoặc một phối hợp các mức độ của các nhân tố khác nhau mà ta muốn khảo sát ảnh hưởng của nó trên vật liệu thí nghiệm. Như vậy, nhân tố là một loại nghiệm thức thay đổi. Các nghiệm thức khác nhau tạo thành các mức độ khác nhau của một nhân tố.

Ví dụ: Một nhà nghiên cứu muốn biết trong bốn loại phân bón hiện có, loại nào cho năng suất bắp cao nhất. Ông chia cánh đồng trồng bắp thành bốn lô. Mỗi lô được bón với một loại phân bón khác nhau. Như vậy, các lô bắp là các đơn vị thí nghiệm, các loại phân bón (các nghiệm thức), trong đó bốn loại phân khác nhau tạo thành bốn mức độ của nhân tố 'loại phân bón'.

Đối với thí nghiệm nhiều nhân tố, nghiệm thức có thể là một tổ hợp các mức độ của hai nhân tố, chẳng hạn giống và phân đạm hay của ba nhân tố như giống, phân đạm và biện pháp canh tác,...

4. Sai số thí nghiệm *(Experimental error)* là tổng cộng các nguồn biến động không kiểm soát được. Có hai nguồn biến động đưa đến sai số thí nghiệm:

a. Nguồn biến động luôn hiện hữu trong vật liệu thí nghiệm (ngồn biến động khách quan), như độ phì của đất, đặc điểm sinh lý của từng giống, ẩm độ, sự thiệt hại do sâu bệnh,... Để kiểm soát nguồn biến động này, ta phải sắp xếp các vật liệu thí nghiệm như thế nào để sai số thí nghiệm càng nhỏ càng tốt. Chẳng hạn, dùng phương pháp phân khối; hoặc sắp xếp các vật liệu thí nghiệm theo từng cặp (so sánh cặp), nhưng cần chú ý đến độ tự do, vì trong trường hợp này độ tự do sẽ bị giảm đi một nửa,...

b. Do phương pháp thực hiện thí nghiệm hoặc do người làm thí nghiệm (nguồn biến động chủ quan) thiếu thận trọng khi lấy chỉ tiêu hay tính toán,... Để khắc phục cần hoàn thiện phương pháp thí nghiệm.

5. Lặp lại *(Replication)* là tập hợp các đơn vị thí nghiệm được nhận cùng một nghiệm thức. Lặp lại làm giảm sự biến động trong các kết quả thí nghiệm, gia tăng ý

nghĩa và mức độ tin cậy mà nhà nghiên cứu có thể rút ra kết luận về nhân tố thí nghiệm.

- *Chức năng của lặp lại*

* *Ước lượng sai số thí nghiệm.* Thường sai số thí nghiệm biểu hiện qua phương sai. Nếu n = 1 (không có lặp lại), không thể ước lượng được biến động (s^2); do đó, cần phải lặp lại nhiều lần.

* *Tăng tính phổ biến của kết quả thí nghiệm.* Ví dụ, khi trắc nghiệm năng suất của giống, nếu chỉ trắc nghiệm trong một khu vực nào đó, thì kết quả chỉ có giá trị ở khu vực nào có điều kiện đất đai, khí hậu,... giống với khu thực hiện thí nghiệm trên; do đó, kết luận không mang tính phổ biến. Vì vậy, phải lặp lại thí nghiệm tại nhiều nơi và nhiều vụ.

- *Các yêu cầu để xác định số lần lặp lại*

* *Độ chính xác của thí nghiệm.* Muốn thí nghiệm càng chính xác, cần phải lặp lại nhiều lần.

* *Dựa vào sự biến động của vật liệu thí nghiệm.* Trường hợp lý tưởng hoàn toàn không có sự biến động giữa các vật liệu thí nghiệm. Ví dụ, hai lô đất hoàn toàn giống nhau một cách lý tưởng, thì hai giống khác nhau trồng trên hai lô đất đó chỉ cần thực hiện một lần lặp lại. Nhưng trong thực tế, sự biến động này luôn hiện hữu. Vì vậy, nếu vật liệu thí nghiệm biến động càng nhiều, nghiệm thức càng phải được lặp lại nhiều lần.

* *Số nghiệm thức và cách bố trí thí nghiệm.* Nếu số nghiệm thức quá nhiều nên giảm bớt số lần lặp lại. Ngoài ra, có một số kiểu bố trí thí nghiệm yêu cầu phải có số lần lặp lại nhất định; chẳng hạn, đối với kiểu bố trí hình vuông Latin đòi hỏi số lần lặp lại phải bằng số nghiệm thức.

* *Dựa vào ngân sách, thời gian và sức lao động.* Mặc dù càng nhiều lần lặp lại thí nghiệm càng chính xác. Tuy nhiên, nếu gia tăng số lần lặp lại quá nhiều, thí nghiệm sẽ lớn, chi phí thí nghiệm sẽ gia tăng, đồng thời mất nhiều thời gian và công sức để đo đếm các chỉ tiêu.

B. PHÂN TÍCH PHƯƠNG SAI (*ANALYSIS OF VARIANCE = ANOVA*)

1. Các tiền đề của phân tích phương sai

- Tính độc lập của sai số thí nghiệm: Sai số của các nghiệm thức khác nhau phải độc lập. Đây là điều kiện cần thiết để thực hiện phân tích phương sai. Tính độc lập của các sai số thí nghiệm được đảm bảo bằng cách bố trí ngẫu nhiên các nghiệm thức vào các đơn vị thí nghiệm.
- Tính đồng nhất của phương sai sai số thí nghiệm.
- Sai số thí nghiệm có phân bố chuẩn.

2. Phân tích phương sai

Ví dụ: Thí nghiệm gồm có t nghiệm thức. Mỗi nghiệm thức được quan sát r lần (r lần lặp lại). Bảng 5.1 trình bày ký hiệu của các số liệu quan sát cũng như tổng và trung bình của nghiệm thức.

Các công thức tính độ tự do, tổng bình phương, trung bình bình phương và giá trị F cho từng nguồn biến động của phân tích phương sai một chiều *(One-way ANOVA)* được trình bày trong Bảng 5.2.

Đối với mỗi nguồn biến động của phân tích phương sai, trung bình bình phương *(MS = Mean Square)* sẽ bằng tổng bình phương *(SS = Sum Square)* chia cho độ tự do *(df = Degree of fredom)* tương ứng.

Bảng 5.1: Ký hiệu được dùng để mô tả cách tính các tổng bình phương khác nhau

Nghiệm thức	Số quan sát (Lặp lại) 1	2	3	...j...	...r	Tổng (T)	Trung bình ($\overline{X}$)
1	X_{11}	X_{12}	X_{13}	X_{1j}	X_{1r}	T_1	$\overline{X}_1$
2	X_{21}	X_{22}	X_{23}	X_{2j}	X_{2r}	T_2	$\overline{X}_2$
3	X_{31}	X_{32}	X_{33}	X_{3j}	X_{3r}	T_3	$\overline{X}_3$
...i...	X_{i1}	X_{i2}	X_{i3}	X_{ij}	X_{ir}	...T_i...	...$\overline{X}_i$...
...t	X_{t1}	X_{t2}	X_{t3}	X_{tj}	X_{tr}	...T_t	...$\overline{X}_t$
Tổng chung						G	
Trung bình chung							$\overline{x}$

X_{ij} = Nghiệm thức thứ i ở lần lặp lại thứ j

X_{tr} = Nghiệm thức thứ t ở lần lặp lại thứ r

T_i = Tổng các lần lặp lại của nghiệm thức thứ i

$\overline{X}_i$ = Trung bình của các lần lặp lại của nghiệm thức thứ i

G (Grand total) = Tổng chung (Các lần lặp lại và các nghiệm thức)

$$G = \sum_{i=1}^{t} T_i = \sum_{i=1}^{t}\sum_{j=1}^{r} X_{ij}$$

$\overline{x}$ = Trung bình chung $\overline{x} = \frac{G}{tr}$ (nếu r bằng nhau)

Bảng 5.2: Bảng phân tích phương sai một chiều, với số lần lặp lại bằng nhau.

Nguồn biến động	Độ tự do (df)	Tổng bình phương (SS) Định nghĩa	Thực hành	Trung bình bình phương (MS)	F tính
Nghiệm thức	t – 1	$r\sum_{i=1}^{t}(\overline{X}_i - \overline{x})^2$	$\sum_{i=1}^{t}\sum_{j=1}^{r}\frac{T_i^2}{r} - \frac{X_{ij}^2}{tr}$	$\frac{\text{SS nghiệm thức}}{t-1}$	$\frac{\text{MS nghiệm thức}}{\text{MS sai số}}$
Sai số	t(r – 1)	$\sum_{i=1}^{t}\sum_{j=1}^{r}\left(X_{ij} - \overline{X}_i\right)^2$	bằng cách trừ (SS tổng cộng - SS nghiệm thức)	$\frac{\text{SS sai số}}{t(r-1)}$	
Tổng cộng	Tr - 1	$\sum_{i=1}^{t}\sum_{j=1}^{r}\left(X_{ij} - \overline{x}\right)^2$			

C. THÍ NGHIỆM MỘT NHÂN TỐ

Thí nghiệm một nhân tố là kiểu thí nghiệm trong đó chỉ có một nhân tố thay đổi, còn các nhân tố khác được giữ ổn định. Đối với các thí nghiệm như thế, nghiệm thức bao gồm các mức độ khác nhau của một nhân tố biến đổi, còn các nhân tố khác được áp dụng giống nhau trong tất cả các lô ở một mức độ nào đó.

Các dạng thí nghiệm một nhân tố thường gặp:

- Hầu hết các thí nghiệm về giống là thí nghiệm một nhân tố, trong đó nhân tố biến đổi là các giống khác nhau (nghiệm thức). Nghĩa là, chỉ có giống được trồng khác

nhau từ lô này sang lô khác, còn tất cả các biện pháp canh tác như phân bón, kiểm soát sâu bệnh, chế độ tưới nước,... đều được áp dụng giống nhau cho tất cả các lô.

- Trắc nghiệm các mức độ khác nhau của của một loại phân bón nào đó.
- Trắc nghiệm các loại thuốc trừ sâu khác nhau.
- Trắc nghiệm các mật độ gieo, trồng khác nhau.

Có hai nhóm bố trí thí nghiệm có thể áp dụng cho thí nghiệm một nhân tố:

- Nhóm 1: Các bố trí khối hoàn toàn (complete block designs), thích hợp cho các thí nghiệm có số nghiệm thức ít và được đặc trưng bởi khối, mỗi khối chứa ít nhất một bộ nghiệm thức hoàn toàn (nghĩa là, chứa tất cả các nghiệm thức).
- Nhóm 2: Các bố trí khối không hoàn toàn (incomplete block designs), thích hợp cho các thí nghiệm cớ số nghiệm thức lớn và được đặc trưng bởi khối, nhưng trong mỗi khối chỉ chứa một phần nghiệm thức được trắc nghiệm.

Trong chương này chúng tôi chỉ giới thiệu ba kiểu bố trí cơ bản nhất của nhóm 1 là bố trí hoàn toàn ngẫu nhiên, khối hoàn toàn ngẫu nhiên và hình vuông latin.

I. BỐ TRÍ HOÀN TOÀN NGẪU NHIÊN
(Completely Randomized Design = CRD)

Trong bố trí hoàn toàn ngẫu nhiên, các nghiệm thức được sắp xếp hoàn toàn ngẫu nhiên để mỗi đơn vị thí nghiệm có cùng cơ hội nhận bất kỳ một nghiệm thức nào, Đối với kiểu bố trí này, sự khác nhau giữa các đơn vị thí nghiệm của cùng một nghiệm thức được xem là sai số thí nghiệm. Vì vậy, kiểu bố trí hoàn toàn ngẫu nhiên chỉ thích hợp cho các thí nghiệm có các đơn vị thí nghiệm đồng nhất, như các thí nghiệm được thực hiện trong phòng, trong nhà lưới, ở đó ảnh hưởng của môi trường tương đối dễ kiểm soát. Riêng đối với các thí nghiệm ngoài đồng thường có sự biến động lớn giữa các lô thí nghiệm, như đất không đồng đều, nên bố trí CRD không phù hợp.

Kiểu bố trí này có một số ưu điểm như sau:

- Phân tích phương sai dễ dàng ngay cả khi các nghiệm thức có số lần lặp lại không bằng nhau.
- Phương pháp phân tích đơn giản ngay cả khi có số liệu thiếu.
- Có tính linh hoạt cao, được sử dụng với số nghiệm thức và số lần lặp lại bất kỳ, mỗi nghiệm thức có thể có số lần lặp lại không bằng nhau.

1. Cách bố trí và làm ngẫu nhiên

Ví dụ: Có 4 nghiệm thức A, B, C, D. Mỗi nghiệm thức được lặp lại 5 lần. Các bước làm ngẫu nhiên như sau:

Bước 1: Xác định tổng số lô thí nghiệm (n) bằng tích của nghiệm thức (t) và số lần lặp lại (r), nghĩa là, n = rt.

Bước 2: Ghi số thứ tự từ 1 - 20 vào 20 đơn vị thí nghiệm theo thứ tự thích hợp; chẳng hạn, từ trái sang phải hoặc từ trên xuống dưới (xem Hình 5.1).

Bước 3: Ghi các nghiệm thức A, B, C, D vào các lô thí nghiệm theo một trong các phương pháp làm ngẫu nhiên sau đây:

(1) Phương pháp dùng bảng số ngẫu nhiên

Định điểm khởi đầu trong bảng số ngẫu nhiên (Phụ lục A) bằng cách nhắm mắt lại và chỉ ngón tay ở bất kỳ một điểm nào trong bảng. Ví dụ, điểm bắt đầu ở giao điểm của hàng 16, cột 21.

Từ điểm bắt đầu đọc xuống theo chiều dọc để có được 20 số ngẫu nhiên phân biệt có 3 số hạng. Sau đó, xếp hạng 20 số này theo thứ tự từ lớn đến nhỏ (hoặc từ nhỏ đến lớn).

Ví dụ: Từ giao điểm của hàng 16, cột 21 chúng ta có 20 số ngẫu nhiên phân biệt có 3 số hạng. Các số này được xếp hạng từ nhỏ đến lớn như sau:

Số Ngẫu nhiên	Thứ tự	Thứ hạng	Số ngẫu nhiên	Thứ tự	Thứ hạng
568	1	8	879	11	17
836	2	14	949	12	19
202	3	2	322	13	3
745	4	10	080	14	1
797	5	13	502	15	7
845	6	15	960	16	20
785	7	12	436	17	6
396	8	5	767	18	11
856	9	16	325	19	4
664	10	9	890	20	18

Chia n = 20 thứ hạng ra làm t = 4 nhóm, mỗi nhóm có r = 5 số theo số thứ tự như sau:

Số nhóm	Số thứ hạng trong nhóm				
1	8	14	2	10	13
2	15	12	5	16	9
3	17	19	3	1	7
4	20	6	11	4	18

Ghi t nghiệm thức vào n lô thí nghiệm bằng cách dùng số nhóm làm số nghiệm thức và số thứ hạng tương ứng trong mỗi nhóm làm số lô. Ví dụ, nhóm 1 ghi nghiệm thức A được sắp vào các lô: 8, 14, 2, 10, 13; nhóm 2 ghi nghiệm thức B và sắp vào các lô: 15, 12, 5, 16, 9; nhóm 3 ghi nghiệm thức C và sắp vào các lô: 17, 19, 3, 1, 7; nhóm 4 ghi nghiệm thức D được sắp vào các lô: 20, 6, 11, 4, 18. Cuối cùng chúng ta được bố trí thí nghiệm như hình 5.1.

1	C	2	A	3	C	4	D
5	B	6	D	7	C	8	A
9	B	10	A	11	D	12	B
13	A	14	A	15	B	16	B
17	C	18	D	19	C	20	D

Hình 5.1: Sơ đồ thí nghiệm bố trí theo thể thức hoàn toàn ngẫu nhiên với 4 nghiệm thức (A, B, C, D), mỗi nghiệm thức lặp lại 5 lần

(2) Phương pháp rút thăm

Chuẩn bị n mảnh giấy giống nhau và chia thành t nhóm, mỗi nhóm có r mảnh giấy. Tên mỗi nghiệm thức được ghi trên một trong r mảnh giấy; sau đó, xếp n mảnh giấy này lại, trộn đều và đặt vào trong hộp, rồi rút ngẫu nhiên mỗi lần một mảnh giấy (không để trở lại). Ghi nghiệm thức có tên trên mảnh giấy đó vào đơn vị thí nghiệm số 1, tiếp tục rút ngẫu nhiên mảnh giấy thứ hai để ghi nghiệm thức cho đơn vị thí nghiệm số 2,... tiếp tục làm như thế cho đến mảnh giấy cuối cùng (mảnh thứ n) tương ứng với

lô cuối cùng. Lưu ý, mỗi lần rút nên lắc hộp thật đều. Ví dụ, mỗi thăm lần lượt rút được như sau:

Tên nghiệm thức:	D	B	A	B	C	A	D	C	B	D
Số thứ tự lô:	1	2	3	4	5	6	7	8	9	10
Tên nghiệm thức:	D	A	A	B	B	C	D	C	C	A
Số thứ tự lô:	11	12	13	14	15	16	17	18	19	20

2. Phân tích phương sai

Trong kiểu bố trí hoàn toàn ngẫu nhiên, biến động tổng cộng được phân chia thành hai nguồn biến động: nguồn biến động do nghiệm thức và nguồn biến động không giải thích được gọi là sai số thí nghiệm. Độ lớn tương đối của hai nguồn biến động này dùng để kết luận hoặc là các nghiệm thức khác nhau thật sự hoặc là khác nhau do may rủi. Các nghiệm thức được xem là khác nhau thật sự nếu nguồn biến động của nghiệm thức lớn hơn sai số thí nghiệm.

Phân tích phương sai được thực hiện ở hai trường hợp:

a. Số lần lặp lại bằng nhau. Chúng ta sử dụng số liệu của thí nghiệm khảo sát hiệu quả của các loại hóa chất phòng trị rầy nâu và sâu đục thân trên lúa (Bảng 5.3).

Bảng 5.3: Ảnh hưởng của sáu loại thuốc phòng trị rầy nâu và sâu đục thân trên năng suất lúa (t/ha). Bố trí theo thể thức hoàn toàn ngẫu nhiên, với 4 (r) lần lặp lại và 7 (t) nghiệm thức (Trích từ Bảng 2.1 của Gomez and Gomez. 1984. Trang 14).

	Lặp lại				Tổng	Trung bình
Nghiệm thức	I	II	III	IV	N.thức (T)	N.thức ($\overline{X}$)
Dol-Mix (1 kg)	2,537	2,069	2,104	1,797	8,507	2,127
Dol-Mix (2 kg)	3,366	2,591	2,211	2,544	10,712	2,678
DDT + -BHC	2,536	2,459	2,827	2,385	10,207	2,552
Azodrin	2,387	2,453	1,556	2,116	8,512	2,128
Dimecron-Boom	1,997	1,679	1,649	1,859	7,184	1,796
Dimecron-Knap	1,796	1,704	1,904	1,320	6,724	1,681
Đối chứng	1,401	1,516	1,270	1,077	5,264	1,316
Tổng chung (G)					57,110	
T.B. chung ($\overline{x}$)						2,040

Để phân tích phương sai chúng ta tiến hành các bước sau:

Bước 1: Xác định độ tự do (df) của các nguồn biến động:

df tổng cộng = (r)(t) - 1 = 27

df nghiệm thức = t - 1 = 7 - 1 = 6

df sai số = t(r - 1) = 7(4 - 1) = 21, hoặc:

df sai số = df tổng cộng - df nghiệm thức = 27 - 6 = 21

Bước 2: Gọi: X_i là số đo của lô thứ i

T_i là tổng nghiệm thức thứ i

n là tổng số lô thí nghiệm [n = (r)(t)]

Tính yếu tố hiệu chỉnh *(C.F.=correction factor)* và các tổng bình phương (SS) như sau:

$$C.F. = \frac{G^2}{n} = \frac{(57,11)^2}{(4)(7)} = 116,484$$

$$\text{SS tổng cộng} = \sum X^2 - C.F.$$
$$= [(2{,}537)^2 + (2{,}069)^2 + \cdots + (1{,}077)^2] - 116{,}484 = 7{,}577$$

$$\text{SS nghiệm thức} = \frac{\sum T^2}{r} - C.F.$$
$$= \frac{(8{,}507)^2 + (10{,}712)^2 + \cdots + (5{,}264)^2}{4} - 116{,}484 = 5{,}587$$

SS sai số = SS tổng cộng - SS nghiệm thức
= 7,577 – 5,587 = 1,990

Bước 3: Tính các trung bình bình phương (MS = means squares) của mỗi nguồn biến động bằng cách chia mỗi tổng bình phương cho độ tự do tương ứng của nó. Trung bình bình phương chính là phương sai

$$\text{MS nghiệm thức} = \frac{\text{SS nghiệm thức}}{t-1} = \frac{5{,}587}{6} = 0{,}931$$

$$\text{MS sai số} = \frac{\text{SS sai số}}{t(r-1)} = \frac{1{,}99}{(7)(3)} = 0{,}095$$

Bước 4: Tính giá trị F để kiểm định mức độ khác biệt của các nghiệm thức như sau:

$$F = \frac{\text{MS nghiệm thức}}{\text{MS sai số}} = \frac{0{,}931}{0{,}095} = 9{,}8$$

Chú ý, chỉ nên tính trị số F khi độ tự do sai số đủ lớn (df ≥ 6) để ước lượng phương sai sai số đáng tin cậy.

Bước 5: Tìm các trị số F tới hạn (F bảng) trong Phụ lục VI, với df_1 là độ tự do của nghiệm thức, và df_2 là độ tự do của sai số. Trong ví dụ này, giá trị F bảng với $df_1 = 6$, và $df_2 = 21$ là 2,57 ở mức ý nghĩa 5% và 3,81 ở mức ý nghĩa 1%.

Bước 6: Lập bảng phân tích phương sai (ANOVA) bằng cách ghi tất cả các trị số tính được từ bước 2 đến bước 5 vào bảng (xem Bảng 5.4)

Bảng 5.4: Phân tích phương sai (CRD) của số liệu năng suất hạt trong Bảng 5.3

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F(tính)	F(bảng) 5%	F(bảng) 1%
Nghiệm thức	6	5,587	0,931	9,8**	2,57	3,81
Sai số	21	1,990	0,095			
Tổng cộng	27	7,577				

cv = 15,1%
** = khác biệt ở mức ý nghĩa 1%

Bước 7: So sánh giá trị F tính ở bước 4 với giá trị F bảng ở bước 5 để quyết định giữa các nghiệm thức có sự khác biệt ý nghĩa hay không theo nguyên tắc sau:

(1) Nếu giá trị F tính lớn hơn giá trị F bảng ở mức ý nghĩa 1%, chúng ta nói sự khác biệt giữa các nghiệm thức có ý nghĩa cao và đặt ** trên giá trị F tính trong bảng ANOVA.

(2) Nếu giá trị F tính lớn hơn giá trị F bảng ở mức ý nghĩa 5%, nhưng nhỏ hơn hoặc bằng giá trị F bảng ở mức ý nghĩa 1%, chúng ta nói sự khác biệt giữa các nghiệm thức có ý nghĩa và đặt * trên giá trị F tính trong bảng ANOVA.

(3) Nếu giá trị F tính nhỏ hơn hoặc bằng giá trị F bảng ở mức ý nghĩa 5%, chúng ta nói sự khác biệt giữa các nghiệm thức không ý nghĩa và đặt ns (non-significant) trên giá trị F tính trong bảng ANOVA.

Chú ý, kiểm định F không có ý nghĩa trong phân tích phương sai nói lên thí nghiệm không thành công trong việc tìm ra sự khác biệt nào đó giữa các nghiệm thức. Có hai trường hợp dẫn tới kết quả này: Một là mọi nghiệm thức đều giống nhau hoặc khác nhau quá ít; hai là do sai số thí nghiệm quá lớn hoặc do cả hai. Do đó, khi kiểm định F không ý nghĩa, nhà nghiên cứu nên kiểm tra lại độ lớn của sai số thí nghiệm và sự chênh lệch giữa các giá trị trung bình nghiệm thức. Nếu cả hai giá trị đều lớn, có thể lặp lại thí nghiệm và cố gắng làm giảm sai số thí nghiệm để có thể phát hiện ra sự khác biệt giữa các nghiệm thức, nếu có. Mặt khác, nếu cả hai giá trị đều nhỏ chứng tỏ sự khác biệt giữa các nghiệm thức có lẽ quá nhỏ để có thể phát hiện được, vì thế không cần lặp lại thí nghiệm.

Trong ví dụ, F tính bằng 9,83 lớn hơn giá trị F bảng ở mức ý nghĩa 1% là 3,81. Vì vậy, chúng ta nói sự khác biệt của các nghiệm thức có ý nghĩa cao, ở mức $\alpha = 0,01$.

Lưu ý: F tính lớn hơn giá trị F bảng ở mức độ 5% hay 1%, điều này cho chúng ta biết trong các nghiệm thức được so sánh, ít nhất có một cặp nghiệm thức có sự khác biệt. Tuy nhiên, nó chưa cho biết cụ thể là bao nhiêu cặp khác biệt và đó là những cặp nào?

Bước 8: Tính trung bình chung ($\bar{x}$) và hệ số biến động (cv%) như sau:

$$\bar{x} = \frac{G}{n} = \frac{57,11}{28} = 2,04$$

$$cv = \frac{\sqrt{MS\ sai\ số}}{\bar{x}} \times 100 = \frac{\sqrt{0,095}}{2,04} \times 100 = 15,1\%$$

Hệ số biến động chỉ độ chính xác của việc so sánh các nghiệm thức và là chỉ số cho phép đánh giá sự tin cậy của thí nghiệm. Nó ghi nhận sai số thí nghiệm bằng phần trăm của trung bình; do đó, cv càng cao, sự tin cậy của thí nghiệm càng thấp. Giá trị cv thường được đặt phía dưới bảng phân tích phương sai để người đọc có thể đánh giá độ tin cậy của kết quả nghiên cứu trước khi tham khảo bảng.

b. Số lần lặp lại không bằng nhau

Vì phương pháp tính toán đối với kiểu bố trí CRD không quá phức tạp, nên CRD thường được sử dụng cho các nghiên cứu mà vật liệu thí nghiệm khó sử dụng số lần lặp lại bằng nhau cho tất cả nghiệm thức, chẳng hạn như:

* Thí nghiệm so sánh chiều dài thân của các loài côn trùng khác nhau bắt được trong bẫy.

* Thí nghiệm về cách nuôi gia súc, trong đó số gia súc cho mỗi phương pháp nuôi không bằng nhau.

* Thí nghiệm lúc đầu có số lần lặp lại bằng nhau, nhưng một vài đơn vị thí nghiệm có thể bị mất hoặc bị hư trong suốt quá trình thực hiện thí nghiệm.

Ví dụ: Một nghiên cứu viên muốn biết sự tăng trưởng của loài thỏ hoang trong một khu rừng nào đó có giống nhau hay không. Ông đặt 5 cái bẫy để bắt thỏ ở 5 địa điểm khác nhau trong một khu rừng. Trọng lượng (oz.) của từng con được ghi nhận trong Bảng 5.5.

Các bước phân tích phương sai như sau:

Bước 1: Gọi t là số nghiệm thức

n là tổng số liệu quan sát, ($n = \sum_{j=1}^{t} r_j$).

Xác định độ tự do cho mỗi nguồn biến động như sau:

df tổng cộng = n - 1 = 17 - 1 = 16

df nghiệm thức = t - 1 = 5 - 1 = 4

df sai số = $\sum_{j=1}^{t} r_j$ - t , hoặc

= df tổng cộng - df nghiệm thức = 16 - 4 = 12

Bảng 5.5: Trọng lượng của thỏ ở năm địa điểm. Bố trí theo thể thức hoàn toàn ngẫu nhiên có số lặn lặp lại không bằng nhau.

Nghiệm thức	Trọng lượng (oz./con)					Tổng N.thức (T)	Trung bình N.thức ($\bar{X}$)
1	37	40	46			123	41,00
2	29	33	34	31		127	31,75
3	49	47				96	48,00
4	40	38	42	39	41	200	40,00
5	50	46	49			145	48,33
Tổng chung (G)						691	
TB chung ($\bar{x}$)							40,65

Bước 2: Tính yếu tố hiệu chỉnh và các tổng bình phương như sau:

$$C.F. = \frac{G^2}{n} = \frac{(691)^2}{17} = 28087,118$$

SS tổng cộng = $\sum_{i,j=1}^{t,r} X_{ij} - C.F.$

$$= [(37)^2 + (40)^2 + \cdots + (49)^2] - 28087,118 = 681,882$$

SS nghiệm thức = $\frac{\sum_{i=1}^{t} T_i^2}{r_i} - C.F.$

$$= \left[\frac{(123)^2}{3} + \frac{(127)^2}{4} + \cdots + \frac{(145)^2}{3}\right] - 28087,118 = 604,466$$

SS sai số = SS tổng cộng - SS nghiệm thức

= 681,883 - 604,466 = 77,417

Từ bước 3 đến bước 8 tính tương tự như số lần lặp lại bằng nhau. Kết quả phân tích phương sai được trình bày trong Bảng 5.6.

Bảng 5.6: Phân tích phương sai (CRD) của số liệu trọng lượng thỏ trong Bảng 5.5

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F (tính)	F(bảng) 5%	F(bảng) 1%
Nghiệm thức	4	604,466	151,116	23,42**	3,26	5,41
Sai số	12	77,417	6,451			
Tổng cộng	16	681,883				

cv = 6,25%

** = khác biệt ở mức ý nghĩa 1%

II. BỐ TRÍ KHỐI HOÀN TOÀN NGẪU NHIÊN

(Randomized Complete Block Design = RCBD)

Bố trí khối hoàn toàn ngẫu nhiên là một trong những kiểu bố trí được dùng phổ biến nhất trong nghiên cứu nông nghiệp. Đặc biệt kiểu bố trí này thích hợp với các thí

nghiệm ngoài đồng có số nghiệm thức không nhiều và khu thí nghiệm có chiều biến động về độ phì có thể đoán trước được; chẳng hạn, đất sườn đồi, có hàng rào (hoặc mương), ảnh hưởng của vụ trồng trước,.... Điểm phân biệt đầu tiên của bố trí RCBD là các khối phải có kích thước bằng nhau và **mỗi khối** (tương ứng với **một lần lặp lại**) phải chứa tất cả các nghiệm thức.

1. Kỹ thuật phân khối

Mục đích chủ yếu của việc phân khối là làm giảm sai số thí nghiệm bằng cách loại bỏ các nguồn biến động đã biết giữa các đơn vị thí nghiệm. Việc gom nhóm các đơn vị thí nghiệm lại thành những khối sao cho biến động bên trong mỗi khối nhỏ nhất và biến động giữa các khối lớn nhất, vì chỉ có biến động bên trong khối trở thành một phần của sai số thí nghiệm. Sự phân khối có hiệu quả nhất khi khu thí nghiệm có một chiều biến động có thể đoán trước được. Với dự đoán này, dạng lô và hướng phân khối được chọn sao cho phần lớn sự biến động được giải thích bởi sự khác biệt giữa các khối, còn các lô thí nghiệm trong cùng một khối càng đồng đều càng tốt.

Có hai quyết định quan trọng cần phải thực hiện để việc phân khối đạt hiệu quả là:

- Chọn nguồn biến động để làm cơ sở cho việc phân khối.
- Chọn dạng khối và định hướng của khối.

Nguồn biến động lý tưởng dùng làm cơ sở cho việc phân khối là nguồn biến động mà chúng ta có thể đoán trước được về độ mạnh và độ lớn, chẳng hạn:

* Đối với các thí nghiệm về giống và phân bón, trong đó số liệu năng suất là tính trạng đầu tiên cần quan tâm thì tính không đồng nhất của đất được chọn làm cơ sở cho việc phân khối.

* Đối với các thí nghiệm về thuốc trừ sâu, trong đó mức độ phá hoại của côn trùng là đặc tính đầu tiên cần quan tâm thì hướng di trú của côn trùng được chọn làm cơ sở cho việc phân khối.

* Trong thí nghiệm nghiên cứu về phản ứng của cây trồng đối với stress nước thì độ dốc của cánh đồng được chọn làm cơ sở cho việc phân khối.

Khi đã xác định nguồn biến động chính để dùng làm cơ sở cho việc phân khối, sau đó chọn cỡ và dạng khối sao cho biến động giữa các khối lớn nhất. Một vài hướng dẫn để thực hiện quyết định là:

(1) Khi khu thí nghiệm chỉ có một hướng biến động, nên dùng khối dài và hẹp. Chiều dài của khối thẳng góc với chiều biến động.

(2) Khi khu thí nghiệm có hai chiều biến động, trong đó có một chiều mạnh hơn chiều còn lại, nên lờ đi chiều yếu và phân khối giống như trường hợp (1).

(3) Nếu khu thí nghiệm có hai chiều biến động mạnh như nhau và thẳng góc với nhau, nên chọn một trong các cách sau:

* Dùng khối càng vuông càng tốt.

* Dùng khối dài và hẹp, với chiều dài của khối thẳng góc với một chiều biến động, như trường hợp (1) và dùng phương pháp phân tích hiệp phương sai để tách ra chiều biến động kia.

* Dùng kiểu bố trí hình vuông latin.

(4) Khi không đoán trước được chiều biến động, nên phân khối càng vuông càng tốt.

Sau khi đã phân khối thì tính đồng nhất của khối và mục đích sử dụng các khối này phải được giữ vững xuyên suốt thí nghiệm; nghĩa là, khi nguồn biến động đang có khuynh hướng vượt quá sự kiểm soát của người làm thí nghiệm, ông phải bảo đảm nguồn biến động như thế chỉ xảy ra giữa các khối hơn là xảy ra bên trong khối.

Ví dụ, nếu các thao tác nào đó như gieo, phun thuốc trừ sâu, hoặc thu thập số liệu không thể làm hoàn tất cho cả thí nghiệm trong một ngày thì công việc nên hoàn tất cho tất cả các lô của cùng một khối trong cùng một ngày. Làm như thế, biến động giữa các ngày (có thể tăng do yếu tố thời tiết) sẽ trở thành một phần biến động của khối; do đó, nó được loại ra khỏi sai số thí nghiệm. Hoặc, nếu có nhiều người theo dõi để lấy chỉ tiêu trong cùng một thí nghiệm thì tốt nhất mỗi người nên theo dõi tất cả các lô trong cùng một khối; như vậy, nguồn biến động giữa các người quan sát khác nhau, nếu có, cũng sẽ được loại khỏi sai số thí nghiệm.

2. Cách bố trí và làm ngẫu nhiên

Tiến trình làm ngẫu nhiên đối với kiểu bố trí RCBD được thực hiện lần lượt cho từng khối. Ví dụ, xét một thí nghiệm với 5 nghiệm thức A, B, C, D, E và 4 lần lặp lại. Các bước làm ngẫu nhiên như sau:

Bước 1: Chia khu thí nghiệm thành r khối bằng nhau, với r là số lần lặp lại theo kỹ thuật phân khối đã mô tả ở phần trên. Theo ví dụ, khu thí nghiệm được chia thành 4 khối như Hình 5.2. Giả sử có một chiều biến động về độ phì của đất dọc theo chiều dài của khu thí nghiệm, dạng khối sẽ là hình chữ nhật và thẳng góc với chiều biến động.

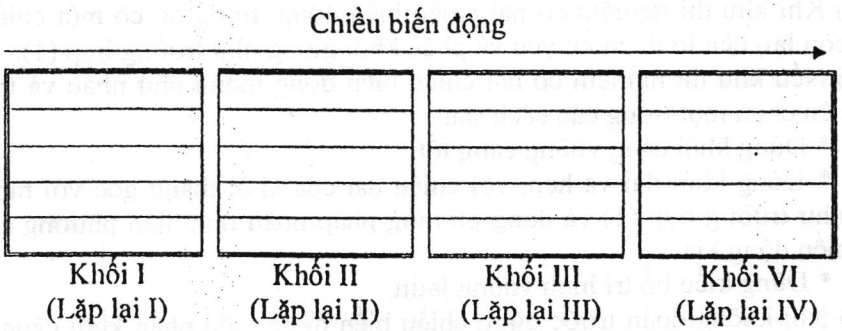
Bước 2: Chia nhỏ khối đầu tiên thành t lô thí nghiệm, với t là số nghiệm thức. Đánh số thứ tự t lô này từ 1 → t và ghi ngẫu nhiên t nghiệm thức vào t lô như cách làm ngẫu nhiên đã mô tả trong bố trí CRD. Trong ví dụ này, khối I được chia nhỏ thành 5 lô có kích thước bằng nhau, đánh số thứ tự từ trên xuống dưới (Hình 5.3) và 5 nghiệm thức được xếp ngẫu nhiên vào 5 lô bằng cách dùng bảng số ngẫu nhiên như sau:

* Chọn 5 số ngẫu nhiên có 3 số hạng. Chúng ta bắt đầu ở giao điểm của hàng thứ 17 và cột thứ 18 của Phụ lục A và đọc dọc xuống:

Số ngẫu nhiên	Số thứ tự	Thứ hạng
584	1	3
965	2	5
072	3	1
695	4	4
192	5	2

* Xếp hạng các số ngẫu nhiên từ nhỏ đến lớn.

* Ghi 5 nghiệm thức vào 5 lô bằng cách dùng số thứ tự làm số nghiệm thức và số hạng tương ứng làm số lô. Do đó, nghiệm thức A được ghi vào lô số 3, B vào lô 5, C lô 1, D lô 4 và E lô 2. Bố trí của khối đầu tiên được trình bày ở hình 5.3



Hình 5.2: Chia khu thí nghiệm thành 4 khối, mỗi khối có 5 lô cho kiểu bố trí khối hoàn toàn ngẫu nhiên với 5 nghiệm thức và 4 lần lặp lại. Việc phân khối được thực hiện như trên, các khối hình chữ nhật và thẳng góc với hướng của một chiều biến động (hướng của mũi tên)

1	C
2	E
3	A
4	D
5	B

Hình 5.3: Đánh số lô và ghi ngẫu nhiên 5 nghiệm thức (A, B. C, D và E) vào trong khối đầu tiên của bố trí ở Hình 5.2

Bước 3: Lặp lại bước 2 cho mỗi khối còn lại. Giả sử cuối cùng chúng ta được bố trí như Hình 5.4.

1	C	6	A	11	D	16	E
2	E	7	E	12	C	17	C
3	A	8	C	13	A	18	D
4	D	9	D	14	B	19	A
5	B	10	B	15	E	20	B
Khối I		Khối II		Khối III		Khối VI	

Hình 5.4: Cách bố trí mẫu của kiểu thí nghiệm khối hoàn toàn ngẫu nhiên, với 5 nghiệm thức (A, B, C, D và E) và 4 lần lặp lại.

Chú ý, sự khác biệt chủ yếu giữa bố trí CRD và RCBD là cách làm ngẫu nhiên. Trong CRD thì không có bất cứ sự ràng buộc nào, nhưng trong RCBD thì khối phải chứa tất cả các nghiệm thức. Sự khác biệt này có thể giải thích bằng cách so sánh kiểu bố trí RCBD của Hình 5.4 với cùng thí nghiệm dựa trên kiểu bố trí CRD như Hình 5.5. Mặt khác, mỗi nghiệm thức trong bố trí CRD có thể xuất hiện ở bất cứ lô nào trong số 20 lô ở ngoài đồng. Ví dụ, trong bố trí CRD, nghiệm thức A xuất hiện ở 3 lô kề nhau (lô 1, 5 và 11), Hình 5.5. Điều này không thể xảy ra trong bố trí RCBD.

1	A	5	A	11	A	13	C	17	D
2	E	6	B	12	D	14	B	18	C
3	C	7	D	9	E	15	A	19	E
4	B	8	E	10	C	16	D	20	B

Hình 5.5: Cách bố trí giả định của kiểu thí nghiệm hoàn toàn ngẫu nhiên với 5 nghiệm thức (A, B, C, D và E) và 4 lần lặp lại.

3. Phân tích phương sai

Có 3 nguồn biến động trong kiểu bố trí RCBD: nghiệm thức, lặp lại (hay khối) và sai số thí nghiệm. So với kiểu CRD, ở đây có thêm một nguồn biến động là lặp lại, tương ứng với biến động giữa các khối.

Để giải thích các bước trong phân tích phương sai của kiểu RCBD, chúng ta dùng số liệu của thí nghiệm bón phân đạm vào giai đoạn đẻ nhánh và làm đòng của một giống lúa mùa (Bảng 5.7).

Các bước phân tích phương sai như sau:

Bước 1: Gọi: r là số lần lặp lại

t là số nghiệm thức

Xác định độ tự do của mỗi nguồn biến động như sau:

df tổng cộng = rt - 1 = 20 - 1 = 19
df lặp lại = r - 1 = 4 - 1 = 3
df nghiệm thức = t - 1 = 5 - 1 = 4
df sai số = (r - 1)(t - 1) = (3)(4) = 12, hoặc
df sai số = df tổng cộng - df lặp lại - df nghiệm thức
= 19 - 3 - 4 = 12

Bước 2: Tính yếu tố hiệu chỉnh và các tổng bình phương như sau:

$$C.F. = \frac{G^2}{rt} = \frac{(967,2)^2}{(4)(5)} = 46773,792$$

$$\text{SS tổng cộng} = \sum Xi^2 - C.F.$$

$$= [(50,7)^2 + (49,5)^2 + \cdots + (35,9)^2] - 46773,792 = 594,868$$

$$\text{SS lặp lại} = \frac{\sum R_j^2}{t} - C.F.$$

$$= \frac{(243,2)^2 + (243,0)^2 + (238,9)^2 + (242,1)^2}{5} - 46773,792 = 2,38$$

$$\text{SS nghiệm thức} = \frac{\sum T_i^2}{r} - C.F.$$

$$= \frac{(199,6)^2 + (199,2)^2 + \cdots + (152,0)^2}{4} - 46773,792 = 558,448$$

SS sai số = SS tổng cộng - SS lặp lại - SS nghiệm thức
= 594,868 - 2,38 - 558,448 = 34,04

Bảng 5.7: Năng suất (tạ/ha) của giống lúa mùa ở năm cách bón phân đạm khác nhau. Bố trí theo thể thức khối hoàn toàn ngẫu nhiên (RCBD) với bốn lần lặp lại

Nghiệm thức*	Lặp lại				Tổng N.thức (T)	Trung bình N.thức ($\bar{X}$)
	I	II	III	IV		
A	50,7	49,5	48,6	50,8	199,6	49,9
B	49,2	49,8	48,7	51,5	199,2	49,8
C	51,4	50,9	54,3	53,8	210,4	52,6
D	52,1	54,0	49,8	50,1	206,0	51,5
E	39,8	38,8	37,5	35,9	152,0	38,0
Tổng lặp lại (R)	243,2	243,0	238,9	242,1		
Tổng chung (G)					967,2	
T.B. chung ($\bar{x}$)						48,36

* *A = Bón 60 N lúc bắt đầu đẻ nhánh + 30 N lúc bắt đầu làm đòng*
B = Bón 60 N lúc bắt đầu đẻ nhánh + 30 N vào cuối giai đoạn làm đòng
C = Bón 60 N vào cuối giai đoạn đẻ nhánh + 30 N lúc bắt đầu làm đòng
D = Bón 60 N vào cuối giai đoạn đẻ nhánh + 30 N vào cuối giai đoạn làm đòng
E = Đối chứng không bón đạm

Bước 3: Tính trung bình bình phương (MS) của mỗi nguồn biến động bằng cách chia mỗi tổng bình phương với độ tự do tương ứng.

$$\text{MS lặp lại} = \frac{\text{SS lặp lại}}{r-1} = \frac{2,38}{3} = 0,793$$

$$\text{MS nghiệm thức} = \frac{\text{SS nghiệm thức}}{t-1} = \frac{558,448}{4} = 139,612$$

$$\text{MS sai số} = \frac{\text{SS sai số}}{(t-1)(r-1)} = \frac{34,04}{12} = 2,837$$

Tính F để kiểm định sự khác biệt của nghiệm thức:

$$\text{F nghiệm thức} = \frac{\text{MS nghiệm thức}}{\text{MS sai số}} = \frac{139,612}{2,837} = 49,217$$

$$\text{F lặp lại} = \frac{\text{MS lặp lại}}{\text{MS sai số}} = \frac{0,793}{2,837} = 0,280$$

So sánh giá trị F nghiệm thức với các giá trị F bảng (Phụ lục VI) ở độ tự do của nghiệm thức bằng 4 và độ tự do của sai số bằng 12. Trong ví dụ, giá trị F bảng tìm được là 3,26 ở mức ý nghĩa 5% và 5,41 ở mức ý nghĩa 1%. Vì giá trị F tính lớn hơn giá trị F bảng ở mức ý nghĩa 1%. Chúng ta kết luận có sự khác biệt rất ý nghĩa giữa năm nghiệm thức.

Bước 4: Tính hệ số biến động

$$cv = \frac{\sqrt{\text{MS sai số}}}{\bar{x}} \times 100 = \frac{\sqrt{2,837}}{48,36} \times 100 = 3,48\%$$

Bước 5: Đưa tất cả giá trị tính được từ bước 1 đến bước 4 vào bảng phân tích phương sai (Bảng 5.8)

Bảng 5.8: Phân tích phương sai (RCBD) của số liệu năng suất hạt trong Bảng 5.7

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F(tính)	F(bảng) 5%	F(bảng) 1%
Lặp lại	3	2,38	0,793	0,280^{ns}	3,49	5,95
Nghiệm thức	4	558,448	139,612	49,217**	3,26	5,41
Sai số	12	34,04	2,837			
Tổng cộng	19	594,868				

cv = 6.7%

** = khác biệt ở mức ý nghĩa 1%

4. Cách tính số liệu thiếu *(Missing data)*

a. Thiếu một số liệu

Thỉnh thoảng khi thực hiện thí nghiệm kiểu bố trí khối, có một số liệu quan sát bị mất. Ví dụ, cây chết do bệnh hoặc chuột ăn gần hết số hạt/cây,... Trong trường hợp này, có hai cách giải quyết:

* Cách thứ nhất là loại bỏ hoàn toàn khối có chứa số liệu thiếu; lúc đó, phân tích phương sai được thực hiện với (r - 1) khối; hoặc loại bỏ nghiệm thức có lô bị mất. Tuy nhiên, cả hai phương pháp này không được khuyến khích vì thông tin bị mất quá nhiều so với chi phí thí nghiệm đã bỏ ra.

* Cách khác là ước lượng số liệu thiếu X_{ij} theo công thức của Yates:

$$X = \frac{rBo + tTo - Go}{(r-1)(t-1)}$$

với: X = số ước lượng của số liệu thiếu

t = số nghiệm thức

r = số lần lặp lại

B_0 = Tổng các trị số quan sát của lần lặp lại có chứa số liệu thiếu

T_0 = Tổng các trị số quan sát của nghiệm thức có chứa số liệu thiếu

G_0 = Tổng các trị số quan sát.

Thay trị số này vào chỗ bị mất trong bảng số liệu và phân tích phương sai bình thường như trường hợp có đầy đủ số liệu; ngoại trừ:

* Độ tự do tổng cộng và độ tự do sai số phải trừ đi 1.
* Tính yếu tố hiệu chỉnh cho B lệch như sau:

$$B = \frac{[Bo-(t-1)X]^2}{t(t-1)}$$

và trừ trị số B này với tổng bình phương tổng cộng và tổng bình phương nghiệm thức để được tổng bình phương tổng cộng và tổng bình phương nghiệm thức đã hiệu chỉnh.

Ví dụ: Số liệu trong Bảng 5.5 có một trị số của nghiệm thức C ở lần lặp lại thứ hai (trị số 50,9 tạ/ha) giả sử bị mất, như ở Bảng 5.9.

Các bước phân tích phương sai được thực hiện như sau:

Bước 1: Dùng công thức để ước lượng số liệu thiếu:

$$X = \frac{4.(192,1)+5.(159,5)-916,3}{(4-1)(5-1)} = 54,1 \text{ t/ha}$$

Bảng 5.9: Số liệu của bố trí RCBD có một trị số quan sát bị mất

Nghiệm thức	Lặp lại I	II	III	IV	Tổng N.thức (T)
A	50,7	49,5	48,6	50,8	199,6
B	49,2	49,8	48,7	51,5	199,2
C	51,4	m^*	54,3	53,8	159,5 = T_0
D	52,1	54,0	49,8	50,1	206,0
E	39,8	38,8	37,5	35,9	152,0
Tổng lặp lại (R)	243,2	(192,1 = B_0)	238,9	242,1	
Tổng chung (G)					916,3 = G_0

m^* = số liệu thiếu

Bước 2: Thay trị số ước lượng (X) vào chỗ bị mất trong Bảng (xem Bảng 5.10) và phân tích phương sai như trường hợp có đầy đủ số liệu.

Bảng 5.10: Số liệu ở Bảng 5.9, có một số liệu thiếu được thay thế bằng trị số ước lượng theo công thức ước lượng số liệu thiếu.

Nghiệm thức	Lặp lại I	II	III	IV	Tổng N.thức (T)
A	50,7	49,5	48,6	50,8	199,6
B	49,2	49,8	48,7	51,5	199,2
C	51,4	**54,1***	54,3	53,8	213,6
D	52,1	54,0	49,8	50,1	206,0
E	39,8	38,8	37,5	35,9	152,0
Tổng Rep. (R)	243,2	246,2	238,9	242,1	
Tổng chung (G)					970,4

* ước lượng số liệu thiếu theo công thức

Bước 3: Sửa đổi kết quả phân tích phương sai ở bước 2 như sau:

* Độ tự do của tổng cộng và sai số bị trừ đi một. Ví dụ, độ tự do tổng cộng là 19 sẽ còn 18 và độ tự do sai số là 12 sẽ còn lại 11.

* Tính yếu tố hiệu chỉnh (C.F.) cho B lệch như sau:

$$B = \frac{\left[192,1-(5-1).(54,1)\right]}{5.(5-1)} = 29,82$$

Lấy tổng bình phương nghiệm thức và tổng bình phương tổng cộng chưa điều chỉnh trừ cho trị số B = 29,82.

SS(nghiệm thức đã điều chỉnh) = 587,632 - 29,82 = 557,812

SS(tổng cộng đã điều chỉnh) = 620,852 - 29,82 = 591,032

Kết quả phân tích phương sai đã điều chỉnh được trình bày trong Bảng 5.11.

Bảng 5.11: Phân tích phương sai (RCBD) của số liệu ở Bảng 5.10 với một số liệu thiếu được ước lượng bằng công thức tính số liệu thiếu

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F(tính)	F(bảng) 5%	F(bảng) 1%
Lặp lại	**3**	**5,452**	**1,817**			
Nghiệm thức	4	557,812	139,453	55,25**	3,36	5,67
Sai số	11	27,768	2,524			
Tổng cộng	18	591,032				

** = khác biệt ở mức ý nghĩa 1%

b. Thiếu hai số liệu

Nếu có nhiều số liệu bị mất chúng ta không thể áp dụng trực tiếp công thức tính số liệu thiếu giống như trên, mà phải dùng phương pháp lặp đi lặp lại. Nguyên tắc cơ bản của phương pháp này thì tương tự cho tất cả các kiểu bố trí.

Ví dụ: Dùng số liệu của bố trí RCBD trong Bảng 5.7. Chúng ta giả định trị số của nghiệm thức C, ở lần lặp lại thứ hai (50,9) và của nghiệm thức A, ở lần lặp lại thứ nhất (50.7) bị mất (Bảng 5.12).

Bảng 5.12: Số liệu của bố trí RCBD thiếu hai trị số, định giá trị ban đầu cho một trong hai trị số như ở bước 1 theo phương pháp lặp lại

	Lặp lại				Tổng
Nghiệm thức	I	II	III	IV	N.thức (T)
A	m^a	49,5	48,6	50,8	199,6
B	49,2	49,8	48,7	51,5	199,2
C	51,4	m^b	54,3	53,8	159,5 = T_0
D	52,1	54,0	49,8	50,1	206,0
E	39,8	38,8	37,5	35,9	152,0
Tổng Rep. (R)	243,2	(192,1 = B_0)	238,9	242,1	
Tổng chung (G)					916,3 = G_0

m^a Định giá trị đầu tiên cho một trong hai số liệu

m^b Ước lượng số liệu thiếu theo công thức

Để ước lượng hai số liệu thiếu này, ta thực hiện các bước sau:

Bước 1: Mặc dầu chúng ta có thể dùng bất cứ trị số nào làm trị số ban đầu; tuy nhiên, nếu trị số này lớn hoặc nhỏ quá có thể làm kéo dài cách tính. Do đó, trị số ban

đầu được dùng phổ biến nhất là trị trung bình của các số liệu còn lại của cùng một nghiệm thức và cùng lần lặp lại.

$$Xij = \frac{ti + bj}{2}$$

với: X_{ij} là trị số ban đầu của nghiệm thức thứ i và lần lặp lại thứ j.

t_i là trung bình của tất cả trị số quan sát của nghiệm thức thứ i.

b_j là trung bình của tất cả trị số quan sát của lần lặp lại thứ j.

Ví dụ: Chúng ta định trị số ban đầu là nghiệm thức A, ở lần lặp lại thứ nhất. Số trung bình X_{11} sẽ là:

* Tính t_1 và b_1, dùng các trị số từ Bảng 5.5:

$$t_1 = \frac{49,5 + 48,6 + 50,8}{3} = 49,6 \text{ tạ/ha}$$

$$b_1 = \frac{49,2 + 51,4 + 52,1 + 39,8}{4} = 48,1 \text{ tạ/ha}$$

* Trung bình X_{11}:

$$X_{11} = \frac{t_1 + b_1}{2}$$

$$X_{11} = \frac{49,6 + 48,1}{2} = 48,9$$

Bước 2: Thế X_{11} = 48,9 tạ/ha mới tìm được vào bảng số liệu, rồi tiếp tục ước lượng X_{32} theo công thức ước lượng một số liệu thiếu.

$$X_{32} = \frac{4.(192,1) + 6.(159,5) - 914,5}{(3)(4)} = 54,3 \text{ tạ/ha}$$

Bước 3: Đưa trị số X_{32} = 54,3 tạ/ha vào trong bảng, sau đó ước lượng lại số liệu thiếu thứ nhất, X_{11}, theo công thức ước lượng một số liệu thiếu như ở bước 2.

$$X'_{11} = \frac{4.(192,5) + 5.(148,9) - 919,9}{(3).(4)} = 49,6$$

Thay trị số X'_{11} vào bảng và ước lượng lại trị số X'_{32} theo công thức như trên:

$$X'_{32} = \frac{4.(192,1) + 5.(159,5) - 915,2}{12} = 54,2$$

Bước 4: Lặp lại bước 3 cho chu kỳ lặp lại thứ hai theo cùng thứ tự của số liệu thiếu đã sử dụng trước. So sánh bộ ước lượng mới với bộ ước lượng tính ở chu kỳ 1. Nếu sự khác biệt không đáng kể, có thể chấp nhận bộ ước lượng mới và chấm dứt tiến trình lặp lại. Nếu hai bộ ước lượng ở chu kỳ 1 và 2 còn khác biệt nhiều nên bắt đầu chu kỳ lặp lại thứ ba và tiến trình nên tiếp tục cho đến khi nào có sự khác biệt không đáng kể giữa hai bộ ước lượng sau cùng (nghĩa là, từ hai chu kỳ lặp lại sau cùng). Chú ý, càng thực hiện nhiều chu kỳ sự khác biệt sẽ càng nhỏ. Do đó, quyết định khi nào ngừng tiến trình lặp lại sẽ tùy thuộc vào sự cân bằng giữa mức độ chính xác mong muốn và phương tiện tính toán sẵn có.

Trong ví dụ, lặp lại bước 3 và bắt đầu chu kỳ lặp lại thứ hai. Hai số liệu thiếu được ước lượng lần nữa là:

$$X''_{11} = \frac{4.(192,5) + 5.(148,9) - 919,8}{12} = 49,6$$

$$X''_{32} = \frac{4.(192,1) + 5.(159,5) - 915,2}{12} = 54,2$$

Vì hai số ước lượng mới của chu kỳ 2 bằng với chu kỳ đầu nên có thể ngừng lại. Như vậy, hai trị số thiếu được ước lượng là $X_{11} = 49,6$ tạ/ha và $X_{32} = 54,2$ tạ/ha.

Chú ý, thường không nhất thiết phải thực hiện tiến trình lặp lại cho đến khi các trị số của hai chu kỳ liên tiếp giống nhau gần như chính xác như ở ví dụ này. Tiến trình có thể chấm dứt ngay khi sự khác biệt giữa các số ước lượng ở hai chu kỳ liên tiếp khá nhỏ.

Bước 5: Thay hai số ước lượng của chu kỳ lặp lại sau cùng vào bảng số liệu và phân tích phương sai như thường, nhưng phải trừ m (m là tổng số liệu thiếu) cho cả hai độ tự do: tổng cộng và sai số.

Với ví dụ trên, hai trị số ước lượng của số liệu thiếu được đưa vào bảng và phân tích phương sai theo cách thông thường. Trừ 2 cho độ tự do tổng cộng và độ tự do sai số. Kết quả phân tích phương sai sau cùng được trình bày trong Bảng 5.13.

Lưu ý, việc ước lượng số liệu thiếu không có nghĩa là tìm lại số liệu đã mất. Thực chất là do kiểu bố trí RCBD đòi hỏi số lần lặp lại của mỗi nghiệm thức phải bằng nhau; do đó, khi có số liệu bị mất thì "bộ" số liệu trở nên mất "cân bằng", lúc đó các phương pháp phân tích chuẩn bị kém hiệu quả. Việc ước lượng số liệu thiếu chỉ làm cho số liệu trở nên "cân bằng" để phân tích có hiệu quả hơn, chớ không phải "tìm lại số liệu đã mất".

Bảng 5.13: Phân tích phương sai của số liệu ở Bảng 5.12 với hai số liệu thiếu được ước lượng theo phương pháp lặp lại.

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F(tính)	F(bảng) 5%	F(bảng) 1%
Lặp lại	3	5,526	1,842	0,82		
Nghiệm thức	4	585,827	146,457	54,26**	3,71	6,55
Sai số	10	26,989	2,699			
Tổng cộng	17	618,342				

** = khác biệt ở mức ý nghĩa 1% .

III. BỐ TRÍ HÌNH VUÔNG LATIN *(Latin Square Design = LS)*

Đặc điểm chính của bố trí hình vuông latin là loại khỏi sai số thí nghiệm hai nguồn biến động đồng thời đã biết giữa các đơn vị thí nghiệm. Việc phân khối hai chiều trong bố trí hình vuông latin, phân khối theo hàng và phân khối theo cột,được thực hiện sao cho mỗi nghiệm thức chỉ xuất hiện một lần trong mỗi khối hàng và một lần trong mỗi khối cột. Với điều kiện này thì số lần lặp lại phải bằng với số nghiệm thức. Kiểu bố trí hình vuông latin có thể ước lươûng biến động giữa các khối hàng cũng như các khối cột và loại chúng ra khỏi sai số thí nghiệm.

1. Cách bố trí và làm ngẫu nhiên

Cách bố trí và làm ngẫu nhiên đối với kiểu bố trí này được tiến hành như sau. Ví dụ, với thí nghiệm có 4 nghiệm thức: A, B, C, D.

Bước 1: Chọn một mô hình LS mẫu với 4 nghiệm thức ở Phụ lục VIII. Trong ví dụ này, mô hình LS 4 x 4 chọn từ Phụ lục F là:

A	B	C	D
B	A	D	C
C	D	B	A
D	C	A	B

Bước 2: Làm ngẫu nhiên theo hàng của mô hình đã chọn ở bước 1 theo phương pháp làm ngẫu nhiên đã được mô tả trong bố trí CRD. Trong thí nghiệm này, phương pháp dùng bảng số ngẫu nhiên được áp dụng như sau:

- Chọn 4 số ngẫu nhiên có 3 số hạng từ Phụ lục A. Ví dụ, bắt đầu từ giao điểm của hàng thứ 17, cột thứ 22, chúng ta có 4 số là: 364, 024, 457 và 979.
- Xếp hạng các số ngẫu nhiên từ nhỏ đến lớn.

Số ngẫu nhiên	Số thứ tự	Thứ hạng
364	1	2
024	2	1
457	3	3
979	4	4

Số thứ hạng sẽ là số hàng của mô hình chọn ở bước 1 và số thứ tự sẽ là số hàng của mô hình mới. Ví dụ, hàng thứ hai (hạng 2) của mô hình bước 1 sẽ trở thành hàng thứ nhất của mô hình mới; hàng thứ nhất của mô hình đầu sẽ là hàng thứ hai của mô hình mới,v.v...

B	A	D	C
A	B	C	D
C	D	B	A
D	C	A	B

Bước 3: Tiến hành tương tự như bước 2 để làm ngẫu nhiên theo cột. Ví dụ, 4 số ngẫu nhiên được chọn và xếp hạng như sau:

Số ngẫu nhiên	Số thứ tự	Thứ hạng
792	1	3
032	2	1
947	3	4
293	4	2

Bây giờ số thứ hạng sẽ là số cột của mô hình ở bước 2 và số thứ tự là số cột của mô hình cuối. Ví dụ, cột thứ ba của mô hình ở bước 2 sẽ trở thành cột thứ nhất của mô hình cuối; cột thứ nhất của mô hình ở bước 2 trở thành cột thứ 2 của mô hình cuối, v.v.... Cuối cùng ta được bố trí thí nghiệm như sau:

D	B	C	A
C	A	D	B
B	C	A	D
A	D	B	C

2. Phân tích phương sai

Có 4 nguồn biến động trong kiểu bố trí LS, đó là: hàng, cột, nghiệm thức và sai số thí nghiệm.

Để giải thích phương pháp tính toán trong phân tích phương sai của bố trí LS, chúng ta dùng số liệu năng suất hạt của 3 giống bắp lai F_1 triển vọng (A, B và D) và giống đối chứng (C) của một thí nghiệm được bố trí kiểu hình vuông latin (Bảng 5.14)

Bảng 5.14 : Năng suất hạt của ba giống bắp lai triển vọng (A,B,D) và giống đối chứng (C) trong thí nghiệm kiểu bố trí hình vuông latin (Trích từ bảng 2.7 của Gomez and Gomez. 1984. Trang 33)

	Năng suất hạt (t/ha)				
Số hàng	Cột 1	Cột 2	Cột 3	Cột 4	Tổng hàng (R)
1	1,640(B)	1,210(D)	1,425(C)	1,345(A)	5,620
2	1,475(C)	1,185(A)	1,400(D)	1,290(B)	5,350
3	1,670(A)	0,710(C)	1,665(B)	1,180(D)	5,225
4	1,565(D)	1,290(B)	1,655(A)	0,660(C)	5,170
Tổng cột (C)	6,350	4,395	6,145	4,475	
Tổng chung (G)					21,365

Bước 1: Tính tổng hàng (R), tổng cột (C), tổng chung (G) như ở Bảng 5.14. Tính tổng nghiệm thức và trung bình nghiệm thức như sau:

Nghiệm thức	Tổng	Trung bình
A	5,855	1,464
B	5,885	1,471
C	4,270	1,068
D	5,355	1,339

Bước 2: Tính độ tự do của các nguồn biến động như sau:

df tổng cộng = t^2 - 1 = 16 - 1 = 15

df hàng = df cột = df nghiệm thức = t - 1 = 4 - 1 = 3

df sai số = (t - 1)(t - 2) = (4 - 1)(4 - 2) = 6 , hoặc :

df sai số = df tổng cộng - df hàng - df cột - df nghiệm thức

= 15 - 3 - 3 - 3 = 6

Bước 3: Tính yếu tố hiệu chỉnh (C.F.) và các tổng bình phương (SS) của các nguồn biến động:

$$C.F. = \frac{G^2}{t^2} = \frac{(21,365)^2}{16} = 28,529$$

SS tổng cộng = $\sum X^2 - C.F.$

$$= [(1,640)^2 + (1,210)^2 + \cdots + (0,660)^2] - 28,529 = 1,414$$

SS hàng = $\frac{\sum R^2}{t} - C.F.$

$$= \frac{(5,620)^2 + (5,350)^2 + (5,225)^2 + (5,170)^2}{4} - 28,529 = 0,030$$

SS cột = $\frac{\sum C^2}{t} - C.F.$.

$$= \frac{(6,350)^2 + (4,395)^2 + (6,145)^2 + (4,475)^2}{4} - 28,529 = 0,827$$

SS nghiệm thức = $\frac{\sum T^2}{t} - C.F.$

$$= \frac{(5,855)^2 + (5,885)^2 + (4,270)^2 + (5,355)^2}{4} - 28,529 = 0,427$$

SS sai số = SS tổng cộng - SS hàng - SS cột - SS nghiệm thức

= 1,414 - 0,030 - 0,827 - 0,427 = 0,130

Bước 4: Tính trung bình bình phương của mỗi nguồn biến động bằng cách chia mỗi tổng bình phương cho độ tự do tương ứng.

$$\text{MS hàng} = \frac{\text{SS hàng}}{t-1} = \frac{0{,}030}{3} = 0{,}010$$

$$\text{MS cột} = \frac{\text{SS cột}}{t-1} = \frac{0{,}827}{3} = 0{,}276$$

$$\text{MS nghiệm thức} = \frac{\text{SS nghiệm thức}}{t-1} = \frac{0{,}427}{3} = 0{,}142$$

$$\text{MS sai số} = \frac{\text{SS sai số}}{(t-1)(t-2)} = \frac{0{,}130}{(3)(2)} = 0{,}022$$

Bước 5: Tính F để kiểm định ảnh hưởng của nghiệm thức:

$$\text{F hàng} = \frac{\text{MS hàng}}{\text{MS sai số}} = \frac{0{,}010}{0{,}022} = 0{,}05$$

$$\text{F cột} = \frac{\text{MS cột}}{\text{MS sai số}} = \frac{0{,}276}{0{,}022} = 12{,}77$$

$$\text{F nghiệm thức} = \frac{\text{MS nghiệm thức}}{\text{MS sai số}} = \frac{0{,}142}{0{,}022} = 6{,}59$$

So sánh các giá trị F tính của nghiệm thức, hàng và cột với giá trị F bảng trong Phụ lục VI ở độ tự do $df_1 = t - 1 = 3$ và $df_2 = (t - 1)(t - 2) = 6$, ta được 2 giá trị F bảng là 4,76 ở mức ý nghĩa 5% và 9,78 ở mức ý nghĩa 1%. Kết luận:

- Vì giá trị F nghiệm thức lớn hơn giá trị F bảng ở mức ý nghĩa 5%, nhưng thấp hơn giá trị F bảng ở mức ý nghĩa 1%; do đó, giữa các nghiệm thức khác biệt ở mức ý nghĩa 5%.

- Giá trị F hàng nhỏ hơn F bảng ở mức ý nghĩa 5%; như vậy, giữa các hàng khác biệt không ý nghĩa. Riêng giá trị F cột lớn hơn F bảng ở mức ý nghĩa 1%; vì thế, sự khác biệt giữa các khối cột có ý nghĩa cao. Điều này chứng tỏ sưü thành công của việc phân khối theo cột vì đã làm giảm sai số thí nghiệm, trong khi phân khối theo hàng lại không có hiệu quả. Như vậy, đối với thí nghiệm này, nếu bố trí theo kiểu RCBD và phân khối theo cột sẽ cho hiệu quả tương đương với kiểu bố trí LS.

Bước 6: Tính hệ số biến động:

$$cv = \frac{\sqrt{\text{MS sai số}}}{\bar{x}} \times 100 = \frac{\sqrt{0{,}021598}}{1{,}335} \times 100 = 11{,}0\%$$

Bước 7: Đưa tất cả các trị số tính được từ bước 1 đến bước 5 vào bảng phân tích phương sai (Bảng 5.15).

Bảng 5.15: Phân tích phương sai (bố trí LS) của số liệu năng suất hạt ở bảng 6.12

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F(tính)	F(bảng) 5%	F(bảng) 1%
Hàng	3	0,030	0,010	$0{,}05^{ns}$	4,76	9,78
Cột	3	0,827	0,276	12,77**	4,76	9,78
Nghiệm thức	3	0,427	0,142	6,59*	4,76	9,78
Sai số	6	0,130	0,022			
Tổng cộng	15	1,414				

cv = 11,0%

** = khác biệt ở mức ý nghĩa 5%*

3. Cách tính số liệu thiếu

Trường hợp có một số liệu bị thiếu, chúng ta có thể dùng công thức của Yates để ước lượng số liệu thiếu như sau:

$$X = \frac{t(Ro + Co + To) - 2Go}{(t-1)(t-2)}$$

với: t = Số nghiệm thức

R_0 = Tổng các trị số quan sát của hàng có chứa số liệu thiếu.

C_0 = Tổng các trị số quan sát của cột có chứa số liệu thiếu.

T_0 = Tổng các trị số quan sát của nghiệm thức có chứa số liệu thiếu.

G_0 = Tổng chung của tất cả số liệu quan sát

Ví dụ: Số liệu trong Bảng 5.16. Chúng ta giả sử trị số năng suất của hàng thứ tư và cột thứ ba bị mất.

Bảng 5.16: Số liệu của bố trí hình vuông latin; có một trị số quan sát bị mất

Số hàng	Năng suất hạt (t/ha)				Tổng hàng
	Cột 1	Cột 2	Cột 3	Cột 4	(R)
1	1,640(B)	1,210(D)	1,425(C)	1,345(A)	5,620
2	1,475(C)	1,185(A)	1,400(D)	1,290(B)	5,350
3	1,670(A)	0,710(C)	1,665(B)	1,180(D)	5,225
4	1,565(D)	1,290(B)	m^a (A)	0,660(C)	(3,515 = R_0)
Tổng cột (C)	6,350	4,395	(4,490 = C_0)	4,475	
Tổng chung (G)					(19,710 = G_0)

Nghiệm thức	Tổng
A	(4,200 = T_0)
B	5,885
C	4,270
D	5,355

Phương pháp phân tích phương sai được thực hiện như sau:

Bước 1: Dùng công thức để ước lượng số liệu thiếu:

$$X = \frac{4(3{,}515 + 4{,}490 + 4{,}200) - 2(19{,}710)}{(3)(2)} = 1{,}567(t/ha)$$

Bước 2: thay trị số ước lượng vào chỗ bị thiếu trong Bảng 6.14 và phân tích phương sai như trường hợp có đầy đủ số liệu. Sau đó, hiệu chỉnh kết quả phân tích phương sai như sau:

- Độ tự do của tổng cộng và sai số bị trừ đi một. Ví dụ, độ tự do tổng cộng bằng 15 sẽ còn 14; và độ tự do sai số bằng 6 sẽ còn 5.

- Tính yếu tố hiệu chỉnh cho B lệch như sau:

$$B = \frac{[G_0 - R_0 - C_0 - (t-1).T_0]^2}{[(t-1).(t-2)]^2}$$

$$B = \frac{[19{,}710 - 3{,}515 - 4{,}490 - (4-1).(4{,}200)]^2}{[(4-1).(4-2)]^2} = 0{,}022$$

Lấy tổng bình phương tổng cộng và tổng bình phương nghiệm thức trừ với trị số B. Kết quả phân tích phương sai cuối cùng được trình bày trong Bảng 5.17.

Bảng 5.17: Phân tích phương sai của số liệu ở Bảng 5.14, có một số liệu thiếu được ước lượng bằng công thức tính số liệu thiếu.

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F(tính)	F(bảng) 5%	F(bảng) 1%
Hàng	3	0,039	0,013	< 1	-	-
Cột	3	0,793	0,264	10,44*	5,41	12,06
Nghiệm thức	3	0,383	0,128	$5,05^{ns}$	5,41	12,06
Sai số	5	0,127	0,025			
Tổng cộng	14	1,343				

** = khác biệt ở mức ý nghĩa 5%.*

ns = khác biệt không ý nghĩa.

CHƯƠNG 6

THÍ NGHIỆM HAI NHÂN TỐ
(TWO-FACTOR EXPERIMENTS)

Thường người làm thí nghiệm cần quan tâm đến ảnh hưởng phối hợp của hai hay nhiều nhân tố. Ví dụ, nghiên cứu về ảnh hưởng của năng suất trên các giống khác nhau phối hợp với các mức độ đạm khác nhau.

Khác với thí nghiệm một nhân tố, chỉ kết luận được ảnh hưởng các mức độ khác nhau của một nhân tố, trong khi thí nghiệm nhiều nhân tố cho phép kết luận ảnh hưởng phối hợp các mức độ khác nhau của nhiều nhân tố cũng như ảnh hưởng của từng nhân tố riêng biệt. Do đó kiểu thí nghiệm này rất kinh tế vì cùng một lúc có thể kết luận được nhiều vấn đề.

Ví dụ: giống là nhân tố A, có 4 mức độ (4 giống) và đạm là nhân tố B, có 5 mức độ; như vậy, sẽ có 20 tổ hợp nghiệm thức. Kiểu thí nghiệm này được gọi là thí nghiệm hai nhân tố hay chính xác hơn là thí nghiệm thừa số 4 x 5.

Với thí nghiệm thừa số, ngoài khảo sát ảnh hưởng của từng nhân tố, chúng ta có thể khảo sát thêm ảnh hưởng tương tác lẫn nhau của hai nhân tố khác nhau.

A. CÁC LOẠI ẢNH HƯỞNG CỦA HAI NHÂN TỐ

Đối với thí nghiệm hai hay nhiều nhân tố thường gặp ba loại ảnh hưởng: ảnh hưởng đơn, ảnh hưởng chính và ảnh hưởng tương tác.

Hai nhân tố được xem là có tương tác với nhau khi mức độ của nhân tố này thay đổi thì ảnh hưởng của nhân tố còn lại cũng thay đổi. Chúng ta sẽ định nghĩa và mô tả cách tính ảnh hưởng tương tác dựa vào thí nghiệm với hai nhân tố A và B, mỗi nhân tố có hai mức độ (a_0 và a_1 đối với nhân tố A; b_0 và b_1 đối với nhân tố B). Bốn tổ hợp nghiệm thức được ký hiệu là a_0b_0 , a_1b_0, a_0b_1 và a_1b_1. Ngoài ra, chúng ta cũng định nghĩa và mô tả cách tính ảnh hưởng đơn và ảnh hưởng chính của mỗi nhân tố A và B, vì các ảnh hưởng này có liên quan đến ảnh hưởng tương tác.

Để giải thích cách tính của ba loại ảnh hưởng này, chúng ta khảo sát hai bộ số liệu được trình bày ở Bảng 6.1 đối với hai giống X và Y và hai mức độ đạm N_0 và N_1: Một bộ không có ảnh hưởng tương tác và một bộ có ảnh hưởng tương tác.

Bảng 6.1: Hai bộ số liệu giả định của thí nghiệm thừa số 2 x 2 = 2^2. Một bộ có ảnh hưởng tương tác giữa hai nhân tố (giống và mức độ đạm) và một bộ không có ảnh hưởng tương tác (Trích từ bảng 3.1 của Gomez and Gomez. 1984. Trang 85)

Giống	Năng suất lúa, tấn/ha		
	0 kg N/ha	60 kg N/ha	Trung bình
	N_0	N_1	
	Không tương tác		
X	1,0	3,0	**2,0**
Y	2,0	4,0	**3,0**
Trung bình	**1,5**	**3,5**	
	Có tương tác		
X	1,0	1,0	**1,0**
Y	2,0	4,0	**3,0**
Trung bình	**1,5**	**2,5**	

1. Ảnh hưởng đơn *(Simple effect)*

Ảnh hưởng đơn của nhân tố A, đó là sự sai biệt giữa hai mức độ của nhân tố này ở cùng mức độ của nhân tố B, nghĩa là:

- Ảnh hưởng đơn của A ở $b_0 = a_1b_0 - a_0 b_0$
- Ảnh hưởng đơn của A ở $b_1 = a_1b_1 - a_0 b_1$

Tương tự, tính ảnh hưởng đơn của nhân tố B ở mỗi mức độ của nhân tố A như sau:

- Ảnh hưởng đơn của B ở $a_0 = a_0 b_1 - a_0 b_0$
- Ảnh hưởng đơn của B ở $a_1 = a_1b_1 - a_1b_0$

Trong ví dụ (Bảng 6.1), cách tính dựa vào bộ số liệu có tương tác như sau:

- Ảnh hưởng đơn của giống ở N_0 = 2,0 - 1,0 = 1,0 t/ha
- Ảnh hưởng đơn của giống ở N_1 = 4,0 - 1,0 = 3,0 t/ha
- Ảnh hưởng đơn của đạm ở giống X = 1,0 - 1,0 = 0,0 t/ha
- Ảnh hưởng đơn của đạm ở giống Y = 4,0 - 2,0 = 2,0 t/ha

Tính dựa trên bộ số liệu không có tương tác:

- Ảnh hưởng đơn của giống ở N_0 = 2,0 - 1,0 = 1,0 t/ha
- Ảnh hưởng đơn của giống ở N_1 = 4,0 - 3,0 = 1,0 t/ha
- Ảnh hưởng đơn của đạm ở giống X = 3,0 - 1,0 = 2,0 t/ha
- Ảnh hưởng đơn của đạm ở giống Y = 4,0 - 2,0 = 2,0 t/ha

2. Ảnh hưởng chính *(Main effect)*

Ảnh hưởng chính của nhân tố A, đó là trung bình của các ảnh hưởng đơn của nhân tố A trên tất cả mức độ của nhân tố B:

- Ảnh hưởng chính của A

  = (1/2)(ảnh hưởng đơn của A ở b_0 + ảnh hưởng đơn của A ở b_1)

  = (1/2)[($a_1 b_0 - a_0 b_0$) + ($a_1b_1 - a_0 b_1$)]

Tương tự, tính ảnh hưởng chính của nhân tố B như sau:

- Ảnh hưởng chính của B

  = (1/2)(ảnh hưởng đơn của B ở a_0 + ảnh hưởng đơn của B ở a_1)

  = (1/2)[($a_0 b_1 - a_0 b_0$) + ($a_1b_1 - a_1b_0$)]

Trong ví dụ, cách tính dựa vào bộ số liệu có tương tác như sau:

- Ảnh hưởng chính của giống = (1/2)(1,0 + 3,0) = 2,0 t/ha
- Ảnh hưởng chính của đạm = (1/2)(0,0 + 2,0) = 1,0 t/ha

Tính dựa trên bộ số liệu không có tương tác:

- Ảnh hưởng chính của giống = (1/2)(1,0 + 1,0) = 1,0 t/ha
- Ảnh hưởng chính của đạm = (1/2)(2,0 + 2,0) = 2,0 t/ha

3. Ảnh hưởng tương tác *(Interaction effect)*

Ảnh hưởng tương tác giữa hai nhân tố A và B, đó là hàm số của sự sai biệt giữa các ảnh hưởng đơn của A ở 2 mức độ của B hoặc của sự sai biệt giữa các ảnh hưởng đơn của B ở hai mức độ của A:

A x B = (1/2)(ảnh hưởng đơn của A ở b_1 - ảnh hưởng đơn của A ở b_0)

= (1/2)[($a_1b_1 - a_0 b_1$) - ($a_1b_0 - a_0 b_0$)

hoặc:

A x B = (1/2)(ảnh hưởng đơn của B ở a_1 - ảnh hưởng đơn của B ở a_0)

= (1/2)[($a_1b_1 - a_1b_0$) - ($a_0b_1 - a_0 b_0$)]

Trong ví dụ, tính ảnh hưởng tương tác giống x đạm dựa trên bộ số liệu có tương tác:

V x N = (1/2)(ảnh hưởng đơn của giống ở N_1 - ảnh hưởng đơn của giống ở N_0)

= (1/2)(3,0 - 1,0) = 1,0 t/ha

hoặc:

V x N = (1/2)(ảnh hưởng đơn của đạm ở giống Y - ảnh hưởng đơn của đạm ở giống X)

= (1/2)(2,0 - 0,0) = 1,0 t/ha

Tính dựa trên bộ số liệu không có tương tác:

V x N = (1/2)(1,0 - 1,0) = 0,0 t/ha

hoặc:

V x N = (1/2)(2,0 - 2,0) = 0,0 t/ha

Trình bày bằng đồ thị về phản ứng phân đạm của hai giống ở Hình 6.1a đối với số liệu không có tương tác và ở Hình 6.1c đối với số liệu có tương tác, trong đó ảnh hưởng tương tác bằng 1,0 t/ha. Hình 6.1b và 6.1d giải thích các ảnh hưởng tương tác thấp hơn và cao hơn 1,0 t/ha. Hình 6.1b cho thấy sự đáp ứng của đạm dương cho cả hai giống, nhưng phản ứng của giống Y (2,0 t/ha) cao hơn giống X (1,0 t/ha), với ảnh hưởng tương tác là 0,5 t/ha. Hình 7.1d cho thấy với giống X khi bón đạm năng suất tăng 2,0 t/ha, nhưng với giống Y khi bón đạm năng suất lại giảm 2,0 t/ha và ảnh hưởng tương tác là 2,0 t/ha. Qua cách tính toán và trình bày bằng đồ thị các ảnh hưởng tương tác, có ba điểm cần lưu ý:

(1) Ảnh hưởng tương tác giữa hai nhân tố chỉ có thể đo được nếu hai nhân tố ấy được trắc nghiệm chung trong cùng một thí nghiệm (nghĩa là, thực hiện thí nghiệm thừa số).

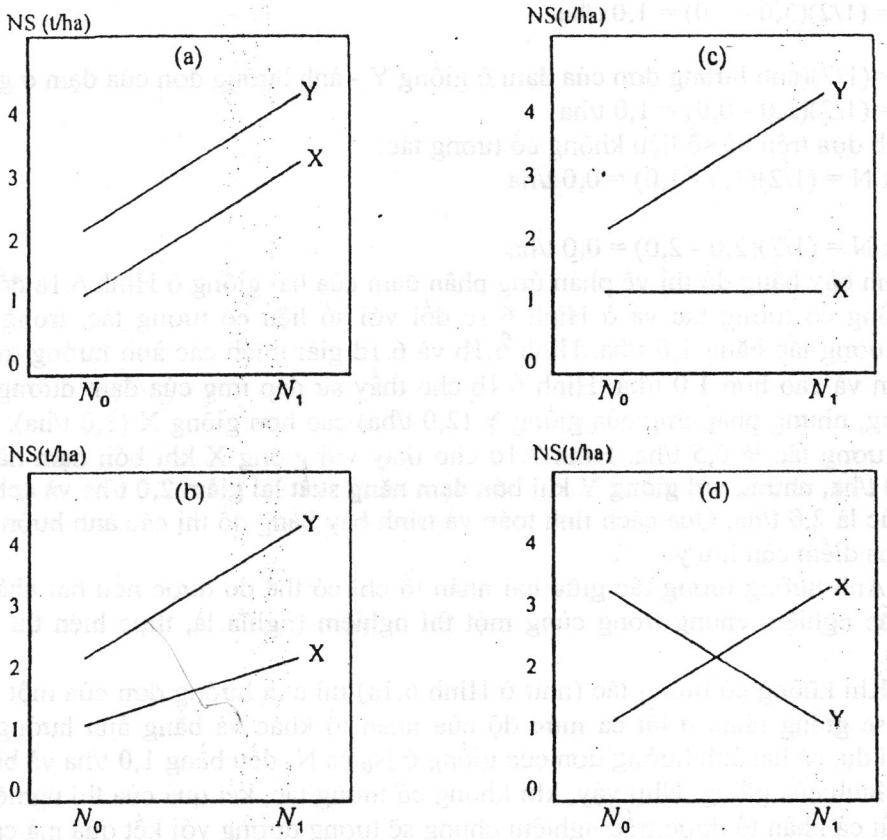
(2) Khi không có tương tác (như ở Hình 6.1a) thì ảnh hưởng đơn của một nhân tố nào đó sẽ giống nhau ở tất cả mức độ của nhân tố khác và bằng ảnh hưởng chính. Trong ví dụ, cả hai ảnh hưởng đơn của giống ở N_0 và N_1 đều bằng 1,0 t/ha và bằng ảnh hưởng chính của giống. Như vậy, khi không có tương tác, kết quả của thí nghiệm thừa số với tất cả nhân tố được trắc nghiệm chung sẽ tương đương với kết quả mà các nhân tố ấy được trắc nghiệm riêng biệt trong từng thí nghiệm (thí nghiệm một nhân tố). Trong ví dụ, ảnh hưởng của giống được ước lượng là 1,0 t/ha, bất chấp

* Hai giống được trắc nghiệm ở mức độ N_0 trong thí nghiệm một nhân tố, hoặc
* Hai giống được trắc nghiệm ở mức độ N_1 trong thí nghiệm một nhân tố, hoặc
* Hai giống được trắc nghiệm ở tổ hợp hai mức độ đạm (N_0 và N_1) trong thí nghiệm hai nhân tố.

(3) Khi có ảnh hưởng của tương tác (như ở Hình 6.1b, 6.1c và 6.1d) thì ảnh hưởng đơn của nhân tố này sẽ thay đổi khi mức độ của nhân tố kia thay đổi. Do đó, ảnh hưởng chính sẽ khác với các ảnh hưởng đơn. Trong ví dụ, ở Hình 6.1c các ảnh hưởng đơn của đạm là 0,0 t/ha đối với giống X và 2,0 t/ha đối với giống Y và ảnh hưởng chính của đạm là 1,0 t/ha. Nói cách khác, mặc dầu ở giống Y có sự đáp ứng mạnh đối với phân đạm, nhưng ở giống X thì không. Hoặc ở Hình 6.1d giống Y cho năng suất cao hơn giống X là 2,0 t/ha ở mức N_0 nhưng lại cho năng suất thấp hơn giống X là 2,0 t/ha ở mức N_1. Nếu tính năng suất trung bình của hai giống ở hai mức độ đạm thì hai số trung bình của giống sẽ bằng nhau (2,0 t/ha). Do đó, nếu nhìn sự khác biệt giữa hai số trung bình của giống (nghĩa là, ảnh hưởng chính của giống), chúng ta sẽ kết luận không có sự khác biệt giữa các giống. Như vậy, khi có ảnh hưởng tương tác giữa các nhân tố thì:

* Chỉ tính ảnh hưởng đơn, không tính ảnh hưởng chính.

* Kết quả của thí nghiệm một nhân tố chỉ có thể áp dụng đối với từng mức độ riêng biệt của nhân tố khác trong thí nghiệm và không thể khái quát hóa kết quả bao hàm mọi mức độ.



Hình 6.1: Trình bày bằng đồ thị các mức độ tương tác khác nhau giữa hai giống (X và Y) và hai mức độ đạm (N_0 và N_1), với (a) không có tương tác, (b) và (c) tương tác trung bình, (d) tương tác mạnh.

B. THÍ NGHIỆM THỪA SỐ *(Factorial experiment)*

Thí nghiệm, trong đó các nghiệm thức bao gồm tất cả tổ hợp các mức độ của hai hay nhiều nhân tố đang khảo sát được gọi là thí nghiệm thừa số. Ví dụ, một thí nghiệm gồm hai nhân tố, mỗi nhân tố có hai mức độ, như hai giống và hai mức độ đạm, được xem như thí nghiệm thừa số 2 x 2 hoặc 2^2. Các nghiệm thức bao gồm bốn tổ hợp của hai mức độ trong mỗi nhân tố:

	Tổ hợp nghiệm thức	
Số nghiệm thức	Giống	Mức độ N, kg/ha
1	X	0
2	X	60
3	Y	0
4	Y	60

Cũng với thí nghiệm trên, nếu chúng ta muốn khảo sát thêm một nhân tố thứ ba, như hai cách kiểm soát cỏ dại. Lúc đó thí nghiệm trở thành thí nghiệm thừa số 2 x 2 x 2 hoặc 2^3, với tám tổ hợp nghiệm thức như sau:

Số nghiệm thức	Tổ hợp nghiệm thức		
	Giống	Mức độ N, kg/ha	kiểm soát cỏ dại
1	X	0	có
2	X	0	không
3	X	60	có
4	X	60	không
5	Y	0	có
6	Y	0	không
7	Y	60	có
8	Y	60	không

Chú ý: Từ thừa số chỉ mô tả cách tạo thành nghiệm thức và không liên quan gì đến kiểu bố trí thí nghiệm được sử dụng. Do đó, khi chọn kiểu bố trí nào cần nói rõ trong thí nghiệm. Ví dụ, thí nghiệm thừa số 2^3 bố trí theo thể thức khối hoàn toàn ngẫu nhiên.

Trong thí nghiệm thừa số, tổng số nghiệm thức sẽ bằng tích của các mức độ của mỗi nhân tố. Ví dụ, thí nghiệm thừa số 2 x 4 = 8 hoặc 2 x 3 x 2 = 12. Do đó, khi số nhân tố hoặc số mức độ của mỗi nhân tố gia tăng thì số nghiệm thức cũng gia tăng nhanh chóng. Ví dụ, thí nghiệm thừa số gồm năm giống, bốn mức độ đạm và ba phương pháp kiểm soát cỏ dại thì tổng số nghiệm thức sẽ là 5 x 4 x 3 = 60. Đối với những thí nghiệm có số nghiệm thức lớn như thế rất khó thực hiện vì tốn nhiều chi phí, công sức, đồng thời khó tìm được khu đất thí nghiệm đồng đều. Khi số nghiệm thức lớn, các kiểu bố trí dùng trong thí nghiệm một nhân tố không còn thích hợp cho các thí nghiệm thừa số; do đó, các nhà thống kê đã phát triển các kiểu bố trí khác hiệu quả hơn cho loại thí nghiệm thừa số.

I. BỐ TRÍ HOÀN TOÀN NGẪU NHIÊN

Các kiểu bố trí mô tả trong thí nghiệm một nhân tố đều có thể được áp dụng cho thí nghiệm thừa số. Phương pháp bố trí và làm ngẫu nhiên của từng kiểu bố trí có thể áp dụng trực tiếp bằng cách *xem mỗi tổ hợp của các nhân tố là một nghiệm thức* và coi như tất cả các nghiệm thức không liên hệ với nhau. Tuy nhiên, khi phân tích phương sai sẽ có thêm những bước tính toán để chia tổng bình phương của nghiệm thức thành các thành phần tương ứng đối với ảnh hưởng chính của từng nhân tố và ảnh hưởng tương tác giữa chúng.

Ví dụ: So sánh năng suất của 2 giống lúa (A và B) được bón ở 3 mức phân đạm (0N, 60N và 120N). Một thí nghiệm viên chọn một khu đất thí nghiệm tương đối đồng nhất và chia ra làm 18 lô. Mỗi phối hợp nghiệm thức được bố trí ngẫu nhiên vào 3 lô. Sáu phối hợp nghiệm thức được trình bày trong Bảng 6.2, bố trí thí nghiệm ở Hình 6.2 và số liệu trong Bảng 6.3.

Các bước phân tích phương sai của thí nghiệm hai nhân tố kiểu bố trí hoàn toàn ngẫu nhiên được thực hiện như sau:

Bước 1: Gọi a là số mức độ của nhân tố A (Giống), a = 2

b là số mức độ của nhân tố B (Đạm), b = 3

r là số lần lặp lại, r = 3

Bảng 6.2: Sáu phối hợp nghiệm thức của thí nghiệm thừa số 2 x 3 (Hai giống lúa và ba mức phân đạm)

Giống	Mức độ đạm (kg/ha)		
	0 (N_0)	60 (N_1)	120 (N_2)
V_1	V_1N_0	V_1N_1	V_1N_2
V_2	V_2N_0	V_2N_1	V_2N_2

V_2N_1	V_2N_2	V_1N_1	V_1N_1	V_1N_2	V_2N_1
V_1N_0	V_1N_2	V_1N_0	V_2N_2	V_1N_0	V_2N_0
V_2N_0	V_2N_2	V_2N_0	V_1N_1	V_2N_2	V_2N_1

Hình 6.2: Bố trí mẫu của thí nghiệm thừa số 2 x 3 bao gồm hai giống lúa (V_1 và V_2) và ba mức phân đạm (N_0, N_1 và N_2). Bố trí theo thể thức hoàn toàn ngẫu nhiên với ba lần lặp lại.

Bảng 6.3: Năng suất hạt (t/ha) của hai giống lúa được trắc nghiệm với ba mức phân đạm theo kiểu bố trí CRD.

Nghiệm thức	Lặp lại I	Lặp lại II	Lặp lại III	Tổng nghiệm thức (T)	Trung bình nghiệm thức ($\bar{X}$)
Đạm		***Giống A***			
0 (N_0)	5,91	6,41	6,82	19,14	6,38
60 (N_1)	9,81	10,26	10,68	30,75	10,25
120 (N_2)	10,23	10,67	11,11	32,01	10,67
		Giống B			
0 (N_0)	7,32	7,66	7,61	22,59	7,53
60 (N_1)	8,98	9,50	9,41	27,89	9,30
120 (N_2)	9,76	10,11	10,22	30,09	10,03
Tổng (G)				**162,47**	
T.B. ($\bar{x}$)					**9,026**

Tính tổng nghiệm thức (T) và tổng chung (G) như trong Bảng 6.3 và tính các tổng bình phương (SS): tổng cộng, nghiệm thức và sai số giống như cách tính của bố trí CRD trong thí nghiệm một nhân tố.

$$C.F = \frac{G^2}{rab} = \frac{(162,47)^2}{18} = 1466,47$$

$$\text{SS tổng cộng} = \sum X^2 - C.F.$$

$$= [(5,91)^2 + (9,81)^2 + \cdots + (10,22)^2] - 1466,47 = 45,083$$

$$\text{SS nghiệm thức} = \frac{\sum T^2}{r} - C.F = \frac{(19,147)^2 + (30,75)^2 + \cdots + (30,09)^2}{3} - 1466,47 = 43,564$$

SS sai số = SS tổng cộng - SS nghiệm thức

= 45,083 - 43,564 = 1,519

Bước 2: Lập bảng tổng của tương tác hai nhân tố A x B.

Trong ví dụ, bảng tổng năng suất giống x đạm (A x B), với tổng giống (A) và tổng đạm (B) được tính như trong Bảng 6.4.

Bước 3: Tính ba thành phần thừa số của tổng bình phương nghiệm thức:

* Tổng bình phương nhân tố A:

Bảng 6.4: Tổng năng suất giống x đạm (A x B) từ số liệu trong bảng 6.3

Giống	Đạm N_0	N_1	N_2	Tổng giống (A)
V_1	19,14	30,75	32,01	81,90
V_2	22,59	27,89	30,09	80,57
Tổng đạm (B)	41,73	58,64	62,10	

$$SSgiống(A)=\frac{\sum A^2}{rb}-C.F.=\frac{(81,9)^2+(80,57)^2}{(3)(3)}-1466,47=0,098$$

* Tổng bình phương nhân tố B:

$$SSđạm(B)=\frac{\sum B^2}{ra}-C.F.=\frac{(41,73)^2+(58,64)^2+(62,1)^2}{(3)(2)}-1466,47=39,603$$

* Tổng bình phương tương tác A x B:

SS(AxB) = SS nghiệm thức - SS(A) - SS(B)

= 43,564 - 0,098 - 39,603 = 3,863

Bước 4: Tính trung bình bình phương (MS) cho từng nguồn biến động bằng cách chia mỗi tổng bình phương cho độ tự do tương ứng:

$$MS(A)=\frac{SS(A)}{a-1}=\frac{0,098}{1}=0,098$$

$$MS(B)=\frac{SS(B)}{b-1}=\frac{39,603}{2}=19,802$$

$$MS(AxB)=\frac{SS(AxB)}{(a-1)(b-1)}=\frac{3,863}{(1)(2)}=1,932$$

$$MS\ sai\ số=\frac{SS\ sai\ số}{ab(r-1)}=\frac{1,519}{(6)(2)}=0,127$$

Bước 5: Tính các giá trị của mỗi thành phần thừa số:

$$F(A)=\frac{MS\ (A)}{MS\ sai\ số}=\frac{0,098}{0,127}=0,777$$

$$F(B)=\frac{MS\ (B)}{MS\ sai\ số}=\frac{19,802}{0,127}=156,486$$

$$F(AxB)=\frac{MS\ (AxB)}{MS\ sai\ số}=\frac{1,932}{0,127}=15,265$$

Bước 6: So sánh mỗi giá trị F tính với giá trị F bảng ở Phụ lục VI, với df_1 = độ tự do của MS tử số và df_2 = độ tự do của MS mẫu số ở mức ý nghĩa α.

Trong ví dụ, so sánh các trị số F(A), F(B) và F(AxB) tính được với các trị số F bảng với độ tự do của MS tử số df_1 lần lượt là (a - 1 = 1), (b - 1 = 2) và [(a-1)(b-1) = 2] và độ tự do của MS mẫu số là 12 (độ tự do của sai số).

Bước 7: Tính hệ số biến động cv

$$cv=\frac{\sqrt{MS\ sai\ số}}{\bar{x}}\times 100=\frac{\sqrt{0,127}}{9,026}\times 100=3,94\ \%$$

Bước 8: Lập bảng phân tích phương sai và điền tất cả các trị số tính được từ bước 1 đến bước 7 vào Bảng 6.5. Kết quả chứng tỏ ảnh hưởng chính của nhân tố A không có ý nghĩa; nghĩa là, giữa các giống khác biệt không ý nghĩa. Trong khi ảnh hưởng chính của nhân tố B có ý nghĩa; nghĩa là, giữa các mức độ đạm khác biệt có ý nghĩa.

Có ảnh hưởng tương tác giữa giống và đạm, điều này chứng tỏ giữa các giống có sự đáp ứng khác nhau với các mức độ đạm.

Bảng 6.5: Phân tích phương sai của số liệu trong bảng 6.3 từ thí nghiệm thừa số 2x3, với bố trí hoàn toàn ngẫu nhiên (CRD).

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F(tính)	F(bảng) 5%	F(bảng) 1%
Giống (A)	1	0,098	0,098	< 1	-	-
Đạm (B)	2	39,603	19,802	156,486**	3,88	6,93
A X B	2	3,863	1,932	15,265**	3,88	6,93
Sai số	12	1,519	0,127			
Tổng cộng	17	45,083				

cv = 3,94 %

** = Khác biệt ở mức ý nghĩa 1%

II. BỐ TRÍ KHỐI HOÀN TOÀN NGẪU NHIÊN

Các bước phân tích phương sai của thí nghiệm hai nhân tố kiểu bố trí khối hoàn toàn ngẫu nhiên (RCBD) sẽ được minh họa trong thí nghiệm gồm năm mức độ đạm, ba giống lúa và bốn lần lặp lại. 15 tổ hợp nghiệm thức được trình bày trong Bảng 6.6, bố trí thí nghiệm ở Hình 6.3 và số liệu trong Bảng 6.7.

Bảng 6.6: Các tổ hợp nghiệm thức của thí nghiệm thừa số 3 x 5 (ba giống lúa và năm mức độ đạm)

Mức độ đạm kg/ha	Tổ hợp nghiệm thức thừa số		
	6966(V_1)	P1215936(V_2)	Milfor 6(2)(V_3)
0 (N_0)	N_0V_1	N_0V_2	N_0V_3
40 (N_1)	N_1V_1	N_1V_2	N_1V_3
70 (N_2)	N_2V_1	N_2V_2	N_2V_3
100 (N_3)	N_3V_1	N_3V_2	N_3V_3
130 (N_4)	N_4V_1	N_4V_2	N_4V_3

V_3N_2	V_2N_1	V_1N_4	V_1N_1	V_2N_3
V_3N_0	V_1N_3	V_3N_4	V_2N_2	V_3N_3
V_2N_4	V_3N_1	V_2N_0	V_1N_0	V_2N_2

Lặp lại I

V_2N_3	V_3N_3	V_1N_1	V_2N_0	V_2N_1
V_1N_3	V_3N_2	V_1N_2	V_1N_4	V_2N_4
V_1N_0	V_3N_4	V_2N_2	V_3N_1	V_3N_0

Lặp lại II

V_1N_1	V_3N_0	V_1N_0	V_3N_1	V_1N_4
V_2N_2	V_1N_2	V_1N_3	V_2N_4	V_3N_4
V_2N_0	V_3N_2	V_2N_1	V_2N_3	V_3N_3

Lặp lại III

V_1N_2	V_2N_2	V_2N_4	V_1N_0	V_2N_0
V_1N_3	V_3N_1	V_1N_4	V_1N_1	V_2N_3
V_3N_0	V_2N_1	V_3N_2	V_3N_3	V_3N_4

Lặp lại IV

Hình 6.3: Bố trí mẫu của thí nghiệm thừa số 3 x 5 bao gồm ba giống lúa (V_1, V_2 và V_3) và năm mức độ đạm (N_0, N_1, N_2, N_3 và N_4). Bố trí theo thể thức khối hoàn toàn ngẫu nhiên với bốn lần lặp lại.

Bảng 6.7: Năng suất hạt (t/ha) của ba giống lúa được trắc nghiệm với năm mức độ đạm theo kiểu bố trí RCBD (Trích từ Bảng 3.3 của Gomez and Gomez. 1984. Trang 92).

Mức độ đạm Kg/ha	Lặp lại I	II	III	VI	Tổng nghiệm thức (T)
			V_1		
N_0	3,852	2,606	3,144	2,894	12,496
N_1	4,788	4,936	4,562	4,608	18,894
N_2	4,576	4,454	4,884	3,924	17,838
N_3	6,034	5,276	5,906	5,652	22,868
N_4	5,874	5,916	5,984	5,518	23,292
			V_2		
N_0	2,846	3,794	4,108	3,444	14,192
N_1	4,956	5,128	4,150	4,990	19,224
N_2	5,928	5,698	5,810	4,308	21,744
N_3	5,664	5,362	6,458	5,474	22,958
N_4	5,458	5,546	5,786	5,932	22,722
			V_3		
N_0	4,192	3,754	3,738	3,428	15,112
N_1	5,250	4,582	4,896	4,286	19,014
N_2	5,822	4,848	5,678	4,932	21,280
N_3	5,888	5,524	6,042	4,756	22,210
N_4	5,864	6,264	6,056	5,362	23,546
Tổng lặp lại (R)	76,992	73,688	77,202	69,508	
Tổng chung(G)					297,390

Bước 1: Tính tổng nghiệm thức (T), tổng lặp lại (R) và tổng chung (G) như trong Bảng 6.3 và tính các tổng bình phương (SS): tổng cộng, lặp lại, nghiệm thức và sai số giống như cách tính của bố trí RCB trong thí nghiệm một nhân tố.

$$C.F. = \frac{G^2}{rab} = \frac{(297,390)^2}{(4)(3)(5)} = 1.474,014$$

$$\text{SS tổng cộng} = \sum X^2 - C.F.$$

$$= [(3,852)^2 + (2,606)^2 + \cdots + (5,362)^2] - 1.474,014 = 53,530$$

$$\text{SS lặp lại} = \frac{\sum R^2}{ab} - C.F. = \frac{(76,992)^2 + \cdots + (69,508)^2}{(3)(5)} - 1.474,014 = 2,599$$

$$\text{SSnghiệm thức} = \frac{\sum T^2}{r} - C.F. = \frac{(12,496)^2 + \cdots + (23,546)^2}{4} - 1.474,014 = 44,578$$

SS sai số = SS tổng cộng - SS lặp lại - SS nghiệm thức

= 53,530 - 2,599 - 44,578 = 6,353

Bước 2: Lập bảng tổng của tương tác hai nhân tố A x B.

Trong ví dụ, bảng tổng năng suất giống x đạm (AB), với tổng giống (A) và tổng đạm (B) được tính như trong Bảng 6.8.

Bước 3: Tính ba thành phần thừa số của tổng bình phương nghiệm thức:

* Tổng bình phương nhân tố A (Giống):

$$SS(A) = \frac{\sum A^2}{rb} - C.F. = \frac{(95,388)^2 + (100,840)^2 + (101,162)^2}{(4)(5)} - 1.474,014 = 1,052$$

* Tổng bình phương nhân tố B (Đạm):

$$SS(B) = \frac{\Sigma B^2}{ra} - C.F. = \frac{(41,800)^2 + \cdots + (69,560)^2}{(4)(3)} - 1.474,014 = 41,234$$

* Tổng bình phương tương tác A x B:

SS(AxB) = SS nghiệm thức - SS(A) - SS(B)

= 44,578 - 1,052 - 41,234 = 2,292

Bảng 6.8: Tổng tương tác giống x đạm từ số liệu trong Bảng 6.7

Đạm	Tổng năng suất (AB)			Tổng đạm (B)
	V_1	V_2	V_3	
N_0	12,496	14,192	15,112	41,800
N_1	18,894	19,224	19,014	57,132
N_2	17,838	21,744	21,280	60,862
N_3	22,868	22,958	22,210	68,036
N_4	23,292	22,722	23,546	69,560
Tổng giống (A)	95,388	100,840	101,162	297,390

Bước 4: Tính trung bình bình phương (MS) cho từng nguồn biến động bằng cách chia mỗi tổng bình phương cho độ tự do tương ứng:

$$\text{MS lặp lại} = \frac{\text{SS lặp lại}}{(r-1)} = \frac{2,599}{3} = 0,866$$

$$MS(A) = \frac{SS(A)}{a-1} = \frac{1,052}{2} = 0,526$$

$$MS(B) = \frac{SS(B)}{b-1} = \frac{41,234}{4} = 10,308$$

$$MS(AxB) = \frac{SS(AxB)}{(a-1)(b-1)} = \frac{2,292}{(3)(4)} = 0,286$$

$$\text{MS sai số} = \frac{\text{SS sai số}}{(r-1)(ab-1)} = \frac{6,353}{3.[(3x5)-1]} = 0,151$$

Bước 5: Tính trị số F của khối và mỗi thành phần thừa số:

$$\text{F lặp lại} = \frac{\text{MS lặp lại}}{\text{MS sai số}} = \frac{0,866}{0,151} = 5,74$$

$$F(A) = \frac{\text{MS (A)}}{\text{MS sai số}} = \frac{0,526}{0,151} = 3,48$$

$$F(B) = \frac{\text{MS (B)}}{\text{MS sai số}} = \frac{10,308}{0,151} = 68,26$$

$$F(AxB) = \frac{\text{MS (AxB)}}{\text{MS sai số}} = \frac{0,286}{0,151} = 1,89$$

Bước 6: So sánh mỗi giá trị F tính với giá trị F bảng ở Phụ lục VI, với df_1 = độ tự do của MS tử số và df_2 = độ tự do của MS mẫu số ở mức ý nghĩa α.

Chẳng hạn, so sánh trị số F(A) tính được với trị số F bảng (df_1 = 2 và df_2 = 42) là 3,22 ở mức ý nghĩa 5% và 5,15 ở mức ý nghĩa 1%. Kết quả chứng tỏ ảnh hưởng chính của nhân tố A có ý nghĩa ở mức 5%; nghĩa là, giữa các giống khác biệt có ý nghĩa.

Bước 7: Tính hệ số biến động cv

$$cv = \frac{\sqrt{\text{MS sai số}}}{\bar{x}} \times 100 = \frac{\sqrt{0,151}}{4,956} \times 100 = 7,8\ \%$$

Bước 8: Lập bảng phân tích phương sai và điền tất cả các trị số tính được từ bước 1 đến bước 7 vào Bảng 6.9. Kết quả cho thấy tương tác giữa giống và đạm không ý nghĩa, điều này chứng tỏ khi áp dụng các mức độ đạm khác nhau đã không làm ảnh hưởng đáng kể đến sự khác biệt của giống cũng như ảnh hưởng của đạm tương đương nhau trên tất cả các giống được trắc nghiệm. Như vậy, trong thí nghiệm này không có ảnh hưởng tương tác giữa giống và đạm. Tuy nhiên, cả hai ảnh hưởng chính của giống và đạm đều có ý nghĩa.

Bảng 6.9: Phân tích phương sai của số liệu trong Bảng 6.7 từ thí nghiệm thừa số 3x5, với bố trí khối hoàn toàn ngẫu nhiên.

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F(tính)	F(bảng) 5%	F(bảng) 1%
Lặp lại	3	2,599	0,866	5,74**	2,83	4,29
Giống (A)	2	1,052	0,526	3,48*	3,22	5,15
Đạm (B)	4	41,234	10,308	68,26**	2,59	3,80
A X B	8	2,292	0,286	$1,89^{ns}$	2,17	2,96
Sai số	42	6,353	0,151			
Tổng cộng	59	33,530				

cv = 7,8 %

** = Khác biệt ở mức ý nghĩa 1%

* = Khác biệt ở mức ý nghĩa 5%

ns = Khác biệt không ý nghĩa.

Đối với thí nghiệm thừa số, kiểu bố trí CRD và RCBD cho mức độ chính xác của các nhân tố đều như nhau. Tuy nhiên, do kiểu bố trí CRD đòi hỏi tất cả các đơn vị thí nghiệm phải đồng nhất và kiểu bố trí RCBD đòi hỏi các lô thí nghiệm trong cùng một khối phải đồng nhất; vì thế, khi số nghiệm thức gia tăng nhiều, điều kiện này khó thỏa. Trong trường hợp đó, nên chọn kiểu bố trí khác thích hợp hơn, như bố trí lô phụ hoặc bố trí dải phụ.

III. BỐ TRÍ LÔ PHỤ *(Split-Plot Design)*

Đây là kiểu bố trí thích hợp cho thí nghiệm hai nhân tố, trong đó cho phép tăng số nghiệm thức nhiều hơn so với kiểu bố trí khối hoàn toàn ngẫu nhiên. Trong bố trí lô phụ, một nhân tố được phân vào lô chính, được gọi là nhân tố lô chính. Lô chính được chia thành các lô phụ, trong đó nhân tố thứ hai được bố trí vào, được gọi là nhân tố lô phụ. Như vậy, mỗi lô chính trở thành một khối của các nghiệm thức lô phụ.

Với bố trí lô phụ, độ chính xác của nhân tố lô chính sẽ thấp hơn của nhân tố lô phụ. Trái lại, việc đánh giá ảnh hưởng chính của nhân tố lô phụ (nghĩa là, các mức độ của nhân tố lô phụ) và ảnh hưởng tương tác của nhân tố này với nhân tố lô chính thì chính xác hơn so với kiểu bố trí CRD và RCBD; trong khi, việc đánh giá ảnh hưởng của các nghiệm thức lô chính (nghĩa là, các mức độ của nhân tố lô chính) thì kém chính xác hơn.

Đối với kiểu bố trí lô phụ, do kích thước lô và độ chính xác trong việc đo lường các ảnh hưởng tương tác không giống nhau cho cả hai nhân tố, nên việc chọn nhân tố nào đặt vào lô chính hoặc lô phụ rất quan trọng. Để thực hiện việc lựa chọn như thế, cần tuân theo các chỉ dẫn được đề nghị như sau:

(1) Mức độ chính xác: Người làm thí nghiệm muốn khảo sát nhân tố nào chính xác hơn thì nhân tố đó sẽ được bố trí vào lô phụ.

(2) Độ lớn tương đối của các ảnh hưởng chính: Nếu ảnh hưởng chính của một nhân tố (chẳng hạn nhân tố B) có hy vọng lớn hơn nhiều và dễ phát hiện hơn so với nhân tố kia (nhân tố A), lúc đó nhân tố B sẽ được đặt vào lô chính và nhân tố A vào lô phụ. Điều này làm gia tăng cơ hội tìm thấy sự khác biệt giữa các mức độ của nhân tố A. Ví dụ, thí nghiệm về phân bón và giống, nhà nghiên cứu có thể đặt giống vào lô phụ và phân bón vào lô chính, vì ông hy vọng ảnh hưởng của phân bón lớn hơn nhiều so với ảnh hưởng của giống.

(3) Phương pháp canh tác: Do điều kiện canh tác đòi hỏi một nhân tố nào đó phải sử dụng lô lớn (lô chính). Ví dụ, thí nghiệm đánh giá về phương pháp tưới và giống; lúc đó, các phương pháp tưới sẽ được đặt vào lô chính để tối thiểu hóa sự di chuyển nước giữa các lô tiếp giáp nhau và làm giảm ảnh hưởng của hàng bìa.

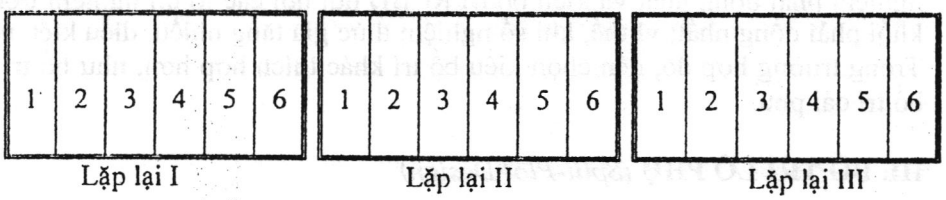
Trong kiểu bố trí lô phụ, cả hai phương pháp làm ngẫu nhiên và phân tích phương sai được thực hiện qua hai giai đoạn: Giai đoạn 1 trên lô chính và giai đoạn 2 trên lô phụ.

1. Cách bố trí và làm ngẫu nhiên

Với bố trí lô phụ, có hai tiến trình làm ngẫu nhiên riêng biệt: Một cho lô chính và một cho lô phụ. Trong mỗi lần lặp lại, đầu tiên các nghiệm thức lô chính được đặt ngẫu nhiên vào lô chính giống như bố trí RCBD; sau đó, các nghiệm thức lô phụ sẽ được bố trí ngẫu nhiên vào mỗi lô chính.

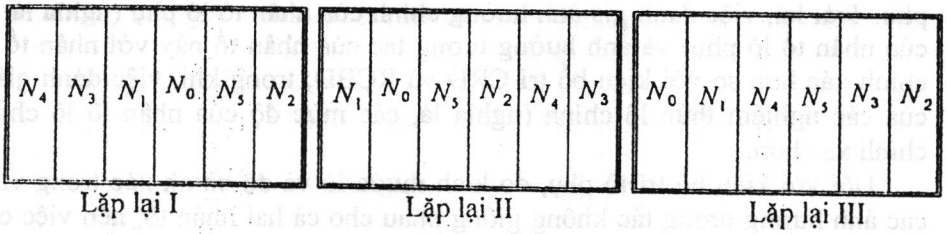
Gọi a là số nghiệm thức lô chính, b là số nghiệm thức lô phụ và r là số lần lặp lại. Để giải thích, chúng ta lấy ví dụ thí nghiệm hai nhân tố gồm sáu mức độ đạm (nghiệm thức lô chính) và bốn giống lúa (nghiệm thức lô phụ) với ba lần lặp lại.

Bước 1: Chia khu thí nghiệm thành r = 3 khối, mỗi khối chia thành a = 6 lô chính như Hình 6.4.



Hình 6.4: Chia khu thí nghiệm ra làm ba khối (Lặp lại), mỗi khối có sáu lô chính, như mô tả ở bước 1 theo cách bố trí của thí nghiệm lô phụ gồm ba lần lặp lại và sáu nghiệm thức lô chính.

Bước 2: Lần lượt bố trí ngẫu nhiên sáu nghiệm thức vào mỗi khối tương tự như cách làm ngẫu nhiên của bố trí RCB (xem Hình 6.5).



Hình 6.5: Bố trí ngẫu nhiên sáu mức độ đạm (N_0, N_1, N_2, N_3, N_4 và N_5) vào sáu lô chính ở một trong ba lần lặp lại của Hình 6.4.

Bước 3: Chia mỗi lô chính thành bốn lô phụ (b = 4) và bố trí ngẫu nhiên bốn giống vào bốn lô phụ của từng lô chính như Hình 6.6.

V_2	V_1	V_1	V_2	V_4	V_3
V_1	V_4	V_2	V_3	V_3	V_2
V_3	V_2	V_4	V_1	V_2	V_1
V_4	V_3	V_3	V_4	V_1	V_4

Lặp lại I

V_1	V_4	V_3	V_1	V_1	V_3
V_3	V_1	V_4	V_2	V_4	V_2
V_2	V_2	V_1	V_4	V_2	V_4
V_4	V_3	V_2	V_3	V_3	V_1

Lặp lại II

V_4	V_3	V_3	V_1	V_2	V_1
V_2	V_4	V_2	V_3	V_3	V_4
V_1	V_1	V_4	V_2	V_4	V_2
V_3	V_2	V_1	V_4	V_1	V_3

Lặp lại III

Hình 6.6: Bố trí mẫu của thí nghiệm lô phụ gồm bốn giống lúa (V_1, V_2, V_3 và V_4) là bốn nghiệm thức lô phụ và sáu mức độ đạm (N_0, N_1, N_2, N_3, N_4 và N_5) là sáu nghiệm thức lô chính, với ba lần lặp lại.

Chú ý: Cách bố trí ngoài đồng của lô phụ có một số đặc điểm quan trọng như sau:

* Kích thước của lô chính lớn gấp b lần kích thước của lô phụ.

* Mỗi nghiệm thức lô chính chỉ được trắc nghiệm r lần; trong khi, mỗi nghiệm thức lô phụ được trắc nghiệm (r)(a) lần. Như vậy, số lần trắc nghiệm của nghiệm thức lô phụ luôn luôn lớn hơn so với nghiệm thức lô chính, đó là lý do chính giải thích tại sao nghiệm thức lô phụ chính xác hơn nghiệm thức lô chính. Trong ví dụ, mỗi mức độ đạm chỉ được trắc nghiệm ba lần, nhưng mỗi giống lại được trắc nghiệm 18 lần.

2. Phân tích phương sai:

Phân tích phương sai của bố trí lô phụ được thực hiện qua hai giai đoạn: Phân tích lô chính và phân tích lô phụ.

Ví dụ: Số liệu của thí nghiệm hai nhân tố: sáu mức độ đạm và bốn giống lúa (Hình 6.6). Số liệu năng suất hạt được trình bày trong Bảng 6.10.

Gọi A là nhân tố lô chính (đạm), B là nhân tố lô phụ (giống). Các bước tính trong phân tích phương sai như sau:

Bước 1: Lập hai bảng tổng:

* Bảng tổng của lặp lại x đạm (RxA), tính tổng lặp lại (R), tổng đạm (A) và tổng chung (G), xem Bảng 6.11.

* Bảng tổng của giống x đạm = AB, tính tổng giống (B), xem Bảng 6.12

Bước 2: Tính yếu tố hiệu chỉnh (C.F.) và các tổng bình phương (SS) của lô chính như sau:

$$C.F. = \frac{G^2}{rab} = \frac{(394,481)^2}{(3)(4)(6)} = 2161,323$$

$$\text{SS tổng cộng} = \Sigma X^2 - C.F. = (4,430)^2 + \cdots + (2,014)^2 - 2161,323 = 204.748$$

$$\text{SS lặp lại} = \frac{\Sigma R^2}{ab} - C.F. = \frac{(128,873)^2 + (135,604)^2 + (130,004)^2}{(6)(4)} - 2161,323$$

$$= 1,083$$

$$\text{SS A (đạm)} = \frac{\Sigma A^2}{rb} - C.F.$$

$$= \frac{(48,67)^2 + \cdots + (69,561)^2}{(3)(4)} - 2161,323$$

$$= 30,429$$

$$\text{SS sai số (a)} = \frac{\Sigma(RA)^2}{b} - C.F. - \text{SS lặp lại} - \text{SS A}$$

$$= \frac{(15{,}964)^2 + \cdots + (22{,}318)^2}{4} - 2161{,}323 - 1{,}083 - 30{,}429$$

$$= 1{,}420$$

Bảng 6.10: Số liệu năng suất hạt (t/ha) của bốn giống lúa trồng ở sáu mức độ đạm. Bố trí lô phụ với ba lần lặp lại (Trích từ Bảng 3.7 của Gomez and Gomez, 1984. Trang 102).

Giống	Lặp lại I	II	III
		N_0 (0 kgN/ha)	
V_1 (IR8)	4,430	4,478	3,850
V_2 (IR5)	3,944	5,314	3,660
V_3 (C4-63)	3,464	2,944	3,142
V_4 (Peta)	4,126	4,482	4,836
		N_1 (60 kgN/ha)	
V_1	5,418	5,166	6,432
V_2	6,502	5,858	5,586
V_3	4,768	6,004	5,556
V_4	5,192	4,604	4,652
		N_2 (90 kgN/ha)	
V_1	6,076	6,420	6,704
V_2	6,008	6,127	6,642
V_3	6,244	5,724	6,014
V_4	4,546	5,744	4,146
		N_3 (120 kgN/ha)	
V_1	6,462	7,056	6,680
V_2	7,139	6,982	6,564
V_3	5,792	5,880	6,370
V_4	2,774	5,036	3,638
		N_4 (150 kgN/ha)	
V_1	7,290	7,848	7,552
V_2	7,682	6,594	6,576
V_3	7,080	6,662	6,320
V_4	1,414	1,960	2,766
		N_5 (180 kgN/ha)	
V_1	8,452	8,832	8,818
V_2	6,228	7,387	6,006
V_3	5,594	7,122	5,480
V_4	2,248	1,380	2,014

Bước 3: Tính các tổng bình phương của lô phụ:

$$\text{SS B (giống)} = \frac{\Sigma B^2}{ra} - C.F.$$

$$= \frac{(117,964)^2 + \cdots + (65,558)^2}{(3)(6)} - 2161,323 = 89,888$$

$$SS\ AxB = \frac{\Sigma(AB)^2}{r} - C.F. - SS\ A - SS\ B$$

$$= \frac{(12,758)^2 + \cdots + (5,642)^2}{3} - 2161,323 - 89,888 - 30,429 = 69,343$$

SS sai số(b) = SS tổng cộng - (tất cả SS đã tính)

= 204,748 - (1,083 + 30,429 + 1,420 + 89,888 + 69.343 = 12.585

Bảng 6.11: Tổng năng suất của lặp lại x giống (RA) được tính từ số liệu ở Bảng 6.10

Đạm (A)	Lặp lại I	II	III	Tổng đạm (A)
N_0	15,964	17,218	15,488	48,670
N_1	21,880	21,632	22,226	65,738
N_2	22,874	24,015	23,506	70,395
N_3	22,167	24,954	23,252	70,373
N_4	23,466	23,064	23,214	69,744
N_5	22,522	24,721	22,318	69,561
Tổng Rep.(R)	128,873	135,604	130,004	
Tổng chung(G)				394,481

Bảng 6.12: Tổng năng suất đạm x giống (AB) được tính từ số liệu ở Bảng 6.10

Đạm (A)	Tổng năng suất (AB) V_1	V_2	V_3	V_4
N_0	12,758	12,918	9,550	13,444
N_1	17,016	17,946	16,328	14,448
N_2	19,200	18,777	17,982	14,436
N_3	20,198	20,685	18,042	11,448
N_4	22,690	20,852	20,062	6,140
N_5	26,102	19,621	18,196	5,642
Tổng giống (B)	117,964	110,799	100,160	65,558

Bước 4: Tính trung bình bình phương cho từng nguồn biến động bằng cách chia mỗi tổng bình phương cho độ tự do tương ứng.

$$MS\ lặp\ lại = \frac{SS\ lặp\ lại}{r-1} = \frac{1,083}{2} = 0,542$$

$$MS\ (A) = \frac{SS\ (A)}{a-1} = \frac{30,429}{5} = 6,086$$

$$MS\ sai\ số\ (a) = \frac{SS\ sai\ số\ (a)}{(r-1)(a-1)} = \frac{1,420}{(2)(5)} = 0,142$$

$$MS\ (B) = \frac{SS\ (B)}{b-1} = \frac{89,888}{3} = 29,963$$

$$MS\ (A \times B) = \frac{SS\ (AxB)}{(a-1)(b-1)} = \frac{69,343}{(5)(3)} = 4,623$$

$$\text{MS sai số (b)} = \frac{\text{SS sai số (b)}}{a(r-1)(b-1)} = \frac{12,585}{(6)(2)(3)} = 0,350$$

Bước 5: Tính các giá trị F lặp lại và F của từng ảnh hưởng cần được kiểm định, bằng cách chia mỗi trung bình bình phương cho trung bình bình phương sai số tương ứng:

$$\text{F lặp lại} = \frac{\text{MS lặp lại}}{\text{MS sai số (a)}} = \frac{0,541}{0,142} = 3,82$$

$$F(A) = \frac{\text{MS (A)}}{\text{MS sai số (a)}} = \frac{6,086}{0,142} = 42,86$$

$$F(B) = \frac{\text{MS (B)}}{\text{MS sai số (b)}} = \frac{29,963}{0,350} = 85,61$$

$$F(A \times B) = \frac{\text{MS (AxB)}}{\text{MS sai số (b)}} = \frac{4.,623}{0,350} = 13,21$$

Bước 6: Các giá trị F của lặp lại và của nhân tố lô chính (nhân tố A) được so với giá trị F bảng tương ứng trong Phụ lục VI, với độ tự do của tử số là độ tự do của lặp lại hoặc của nhân tố A và độ tự do của mẫu số là độ tự do của sai số lô chính (sai số a) ở mức ý nghĩa α. Trong khi các giá trị F của nhân tố lô phụ (nhân tố B) và của tương tác AxB được so với giá trị F bảng với độ tự do của tử số là độ tự do của nhân tố B hoặc của tương tác AxB và độ tự do của mẫu số là độ tự do của sai số lô phụ (sai số b). Ví dụ, giá trị F bảng tương ứng với F(AxB) có $df_1 = 15$ và $df_2 = 36$ là 1,96 ở $\alpha = 0,05$ và 2,58 ở $\alpha = 0,01$.

Bước 7: Tính hai hệ số biến động: Một tương ứng với lô chính và một tương ứng với lô phụ.

$$cv\,(a) = \frac{\sqrt{\text{MS sai số (a)}}}{\bar{x}} \times 100 = \frac{\sqrt{0,142}}{5,479} \times 100 = 6,9\ \%$$

$$cv\,(b) = \frac{\sqrt{\text{MS sai số (b)}}}{\bar{x}} \times 100 = \frac{\sqrt{0,350}}{5,479} \times 100 = 10,8\ \%$$

Trị số cv(a) chỉ độ chính xác của nhân tố lô chính và trị số cv(b) chỉ độ chính xác của nhân tố lô phụ và sự tương tác của nhân tố này với nhân tố lô chính. Giá trị của cv(b) được dự đoán nhỏ hơn giá trị cv(a), vì nhân tố bố trí vào lô chính thường kém chính xác hơn nhân tố bố trí vào lô phụ. Tuy nhiên, điều này không luôn luôn xảy ra; như trong ví dụ, trị số cv(b) lớn hơn cv(a). Nếu các kết quả như thế thường xảy ra, cần thảo luận với những người chuyên về thống kê để tìm ra nguyên nhân và biện pháp khắc phục. Kết quả phân tích phương sai được trình bày ở Bảng 6.13

3. Cách tính số liệu thiếu

Số liệu thiếu trong lô phụ được ước lượng như sau:

$$X = \frac{rM_0 + bT_0 - P_0}{(b-1)(r-1)}$$

với: b = Mức độ của nhân tố lô phụ

r = Số lần lặp lại

M_0 = Tổng các giá trị quan sát của lô chính có chứa số liệu thiếu

T_0 = Tổng các giá trị quan sát của nghiệm thức có chứa số liệu thiếu

P_0 = Tổng các giá trị quan sát của nghiệm thức lô chính có chứa số liệu thiếu

Chú ý: Công thức tính số liệu thiếu của bố trí lô phụ cũng giống với công thức của bố trí khối hoàn toàn ngẫu nhiên, với lô chính thay thế cho lần lặp lại.

Bảng 6.13: Phân tích phương sai của số liệu trong Bảng 6.6 từ thí nghiệm thừa số 4 x 6, với kiểu bố trí lô phụ.

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F(tính)	F(bảng) 5%	F(bảng) 1%
Lặp lại	2	1,083	0,542	$3,82^{ns}$	4,10	7,56
Đạm (A)	5	30,429	6,086	42,86*	3,33	5,64
Sai số (a)	10	1,420	0,142			
Giống (B)	3	89,888	29,963	85,61**	2,86	4,38
A X B	15	69,343	4,623	13,21**	1,96	2,58
Sai số (b)	36	12,585	0,350			
Tổng cộng	71	204,748				

cv(a) = 6,9% cv(b) = 10,8%

* = khác biiệt ở mức ý nghĩa 5%

** = khác biiệt ở mức ý nghĩa 1%

^{ns}= khác biệt không ý nghĩa

Ví dụ: năng suất của nghiệm thức N_2V_1 ở lần lặp lại thứ hai của Bảng 6.10 bị thiếu. Phương pháp ước lượng số liệu thiếu như sau:

$$X = \frac{rM_0 + bT_0 - P_0}{(b-1)(r-1)}$$

với b = 4, mức độ của nhân tố lô phụ (giống)

r = 3, số lần lặp lại.

M_0 = 6,127 + 5,724 + 5,744 = 17,595 (tổng các giá trị quan sát của lô chính N_2 ở lần lặp lại thứ hai).

T_0 = 6,076 + 6,704 = 12,780 (tổng các trị số quan sát của N_2V_1)

P_0 = 6,076 + 6,704 + 6,088 + 6,127 + ... + 4,146 = 63,975 (tổng các giá trị quan sát của N_2)

$$X = \frac{3(17,595) + 4(12,780) - 63,975}{(3)(2)} = 6,655 \text{ t/ha}$$

Thay trị số ước lượng vào chỗ bị mất trong bảng số liệu và phân tích phương sai như cách thông thường, nhưng lưu ý trừ 1 cho độ tự do tổng cộng và độ tự do sai số (b). Với ví dụ trên, độ tự do tổng cộng bằng 70 và độ tự do sai số (b) bằng 35. Kết quả phân tích phương sai sau cùng được trình bày trong Bảng 6.14.

IV. BỐ TRÍ DẢI PHỤ *(Strip-plot design)*

Bố trí dải phụ áp dụng đặc biệt cho thí nghiệm 2 nhân tố, trong đó độ chính xác mong muốn để đo ảnh hưởng tương tác giữa hai nhân tố cao hơn là để đo ảnh hưởng chính của một trong hai nhân tố. Điều này được thực hiện với việc sử dụng 3 cỡ lô:

(1) Lô dải đất dọc cho nhân tố đầu tiên – Nhân tố dọc

(2) Lô dải đất ngang cho nhân tố thứ hai – Nhân tố ngang

(3) Lô tương tác cho sự tương tác giữa hai nhân tố

Lô dải dọc và lô dải ngang luôn luôn thẳng góc với nhau. Tuy nhiên, không có sự quan hệ giữa các kích thước của chúng, không giống như trường hợp lô chính và lô phụ của bố trí lô phụ. Dĩ nhiên, lô giao nhau là nhỏ nhất. Do đó, trong bố trí dải phụ,

các mức độ chính xác có liên quan đến các ảnh hưởng chính của cả hai nhân tố được hy sinh để cải thiện độ chính xác của ảnh hưởng tương tác.

Bảng 6.14: Phân tích phương sai (bố trí lô phụ) của số liệu trong Bảng 6.10, có một số liệu thiếu (N_2V_1) được ước lượng theo công thức tính số liệu thiếu

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F(tính)	F(bảng) 5%	F(bảng) 1%
Lặp lại	2	1,165	0,582	4,13*	4,10	7,56
Đạm (N)	5	30,615	6,123	43,43**	3,33	5,64
Sai số (a)	10	1,411	0,141			
Giống (V)	3	90,395	30,132	83,93**	2,87	4,40
N X V	15	69,101	4,607	12,83**	1,96	2,60
Sai số (b)	35	12,557	0,359			
Tổng cộng	70	205,245				

* = khác biệt ở mức ý nghĩa 5%

** = khác biệt ở mức ý nghĩa 1%

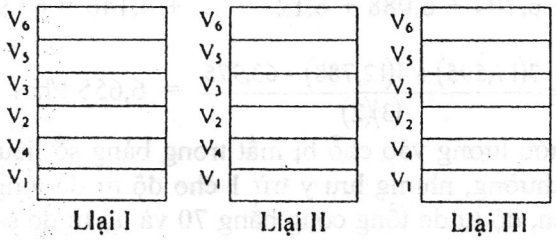
1. Cách bố trí và làm ngẫu nhiên

Phương pháp bố trí và làm ngẫu nhiên của bố trí dải phụ gồm có hai tiến trình làm ngẫu nhiên độc lập – Một cho nhân tố ngang và một cho nhân tố dọc.

Đặt A tiêu biểu cho nhân tố ngang, B nhân tố dọc, và a và b tiêu biểu cho các mức độ của chúng và r là số lần lặp lại.

Chúng ta lấy ví dụ cho thí nghiệm hai nhân tố bao gồm 6 giống lúa (các nghiệm thức ngang) và 3 mức độ đạm (các nghiệm thức dọc) được trắc nghiệm theo kiểu bố trí dải phụ với ba lần lặp lại.

Bước 1: Chia khu thí nghiệm ra làm r = 3 khối và chia mỗi khối thành a = 6 dải ngang. Áp dụng cách làm ngẫu nhiên giống như bố trí khối hoàn toàn ngẫu nhiên với a = 6 nghiệm thức và r = 3 lần lặp lại và bố trí ngẫu nhiên 6 giống vào 6 dải ngang ở một trong ba khối, riêng biệt và độc lập. Kết quả được trình bày ở Hình 6.7.



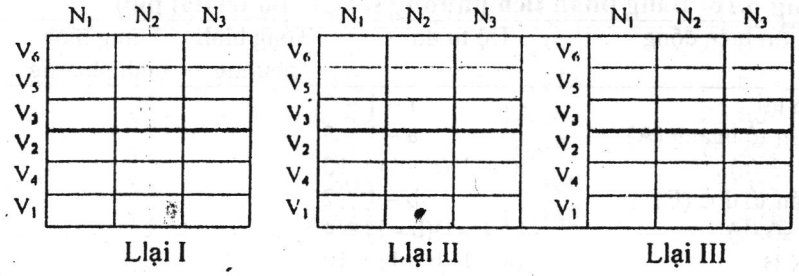
Hình 6.7: Bố trí ngẫu nhiên sáu giống (V_1, V_2, V_3, V_4, V_5 và V_6) vào sáu dải ngang trong bố trí dải phụ với ba lần lặp lại.

Bước 2: Chia mỗi khối thành b = 3 dải dọc. Tương tự như bước 1, ấn định ngẫu nhiên 3 mức độ đạm vào 3 dải dọc ở một trong ba khối, riêng biệt và độc lập. Bố trí sau cùng được trình bày ở Hình 6.8.

2. Phân tích phương sai:

Phân tích phương sai của bố trí dải phụ được chia thành ba phần: Phân tích nhân tố ngang, phân tích tích nhân tố dọc, và phân tích tương tác. Chúng ta lấy ví dụ cho thí nghiệm hai nhân tố bao gồm 6 giống lúa (các nghiệm thức ngang) và 3 mức độ đạm

(các nghiệm thức dọc) được trắc nghiệm theo kiểu bố trí dải phụ với ba lần lặp lại. Bố trí ngoài đồng được trình bày ở Hình 6.8, số liệu ở Bảng 6.15.



Hình 6.8: Sơ đồ bố trí mẫu của bố trí dải phụ với sáu giống (V_1, V_2, V_3, V_4, V_5 và V_6) là các nghiệm thức nằm ngang và ba mức độ đạm (N_1, V_2 và N_3) ở ba lần lặp lại.

Bảng 6.15: Số liệu năng suất hạt của sáu giống lúa trồng ở ba mức độ đạm. Bố trí dải phụ với ba lần lặp lại (Trích của Gomez và Gomez. 1984. Trang 110)

Mức độ đạm kgN/ha	Năng suất hạt, kg/ha		
	Lặp lại I	Lặp lại II	Lặp lại III
		IR8 (V_1)	
0 (N_1)	2,373	3,958	4,384
60 (N_2)	4,076	6,431	4,889
120 (N_3)	7,254	6,808	8,582
		IR127-80 (V_2)	
0 (N_1)	4,007	5,795	5,001
60 (N_2)	5,630	7,334	7,177
120 (N_3)	7,053	8,284	6,297
		IR305-4-12 (V_3)	
0 (N_1)	2,620	4,508	5,621
60 (N_2)	4,676	6,672	7,019
120 (N_3)	7,666	7,328	8,611
		IR400-2-5 (V_4)	
0 (N_1)	2,726	5,630	3,821
60 (N_2)	4,838	7,007	4,816
120 (N_3)	6,881	7,735	6,667
		IR665-58 (V_5)	
0 (N_1)	4,447	3,276	4,582
60 (N_2)	5,549	5,340	6,011
120 (N_3)	6,880	5,080	6,076
		Peta (V_6)	
0 (N_1)	2,572	3,724	3,326
60 (N_2)	3,896	2,822	4,425
120 (N_3)	1,556	2,706	3,214

Bước 1: Phát thảo bảng phân tích phương sai cho kiểu bố trí dải phụ trong Bảng 6.16.

Bảng 6.16: Bảng phân tích phương sai của bố trí dải phụ

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F(tính)	F(bảng) 5%	F(bảng) 1%
Lặp lại	r – 1 = 2					
Nhân tố ngang (A)	a – 1 = 5					
Sai số (a)	(r – 1)(a – 1) = 10					
Nhân tố dọc (B)	b – 1 = 2					
Sai số (b)	(r -1)(b - 1) = 4					
A X B	(a – 1)(b – 1) = 10					
Sai số (c)	(r – 1)(a – 1)(b – 1) = 20					
Tổng cộng	rab – 1 = 53					

Bước 2: Lập ba bảng tổng

(1) Bảng tổng lặp lại x nhân tố ngang với tổng lặp lại, tổng nhân tố ngang và tổng chung. Trong ví dụ, bảng tổng lặp lại x giống (RA) với tổng lặp lại (R), tổng giống (A) và tổng chung (G) được tính trong Bảng 6.17.

Bảng 6.17: Tổng năng suất của lặp lại x giống (R x A) được tính từ số liệu ở Bảng 6.15

Giống	Tổng năng suất (R x A)			Tổng giống (A)
	L.lại I	L.lại II	L.lại III	
V_1	13,703	17,197	17,855	48,755
V_2	16,690	21,413	18,475	56,578
V_3	14,962	18,508	21,251	54,721
V_4	14,445	20,372	15,304	50,121
V_5	16,876	13,696	16,669	47,241
V_6	8,024	9,252	10,965	28,241
Tổng Rep.(R)	84,700	100,438	100,519	
Tổng chung(G)				285,657

(2) Bảng tổng lặp lại x nhân tố dọc với việc tính tổng nhân tố dọc. Trong ví dụ, bảng tổng lặp lại x đạm (RB) với tổng đạm (B) được tính trong Bảng 6.18.

Bảng 6.18: Tổng năng suất của lặp lại x đạm (R x B) được tính từ số liệu ở Bảng 6.15

Đạm	Tổng năng suất (R x B)			Tổng đạm (B)
	L.lại I	L.lại II	L.lại III	
N_1	18,745	26,891	26,735	72,371
N_2	28,665	35,606	34,337	98,608
N_3	37,290	37,941	39,447	114,678

(3) Bảng tổng nhân tố ngang x nhân tố dọc. Trong ví dụ, bảng tổng giống x đạm (A x B) được trình bày trong Bảng 6.19.

Bảng 6.19: Tổng năng suất giống x đạm (Axx B) được tính từ số liệu ở Bảng 6.15

Giống	Tổng năng suất (AB)		
	N_1	N_2	N_3
V_1	10,715	15,396	22,644
V_2	14,803	20,141	21,634
V_3	12,749	18,367	23,605
V_4	12,177	16,661	21,283
V_5	12,305	16,900	18,036
V_6	9,622	11,143	7,476

Bước 3: Tính yếu tố hiệu chỉnh (CF) và tổng bình phương tổng cộng như sau:

$$C.F. = \frac{G^2}{rab} = \frac{(285,657)^2}{(3)(6)(3)} = 1511,110$$

$$\text{SS tổng cộng} = \sum X^2 - C.F.$$

$$= (2,373)^2 + \cdots + (3,214)^2 - 1511,110 = 167,006$$

Bước 4: Tính các tổng bình phương cho phân tích nhân tố ngang như sau:

$$\text{SS lặp lại} = \frac{\sum R^2}{ab} - C.F.$$

$$= \frac{(84,700)^2 + (100,438)^2 + (100,519)^2}{(6)(3)} - 1511,110 = 9,221$$

$$\text{SS A (giống)} = \frac{\sum A^2}{rb} - C.F.$$

$$= \frac{(48,755)^2 + \cdots + (28,241)^2}{(3)(3)} - 1511,110 = 57,100$$

$$\text{SS sai số (a)} = \frac{\sum (RA)^2}{b} - C.F. - \text{SS lặp lại} - \text{SS A}$$

$$= \frac{(13,703)^2 + \cdots + (10,965)^2}{3} - 1511,110 - 9,221 - 57,100$$

$$= 14922620$$

Bước 5: Tính các tổng bình phương cho phân tích nhân tố dọc như sau:

$$\text{SS B (đạm)} = \frac{\sum B^2}{ra} - C.F.$$

$$= \frac{(72,371)^2 + (98,608)^2 + (114,678)^2}{(3)(6)} - 1511,110 = 50,676$$

$$\text{SS sai số (b)} = \frac{\sum (RB)^2}{a} - C.F. - \text{SS lập lại} - \text{SS B}$$

$$= \frac{(18,745)^2 + \cdots + (39,447)^2}{6} - 1511,110 - 9,221 - 50,676 = 2,975$$

Bước 6: Tính các tổng bình phương cho phân tích tương tác như sau:

$$\text{SS AxB} = \frac{\sum (AB)^2}{r} - C.F. - \text{SS A} - \text{SS B}$$

$$= \frac{(10,715)^2 + \cdots + (7,476)^2}{3} - 1511,110 - 57,100 - 50,676$$

$$= 23,878$$

SS sai số(c) = SS tổng cộng - (tất cả SS đã tính)
= 167,006 - (9,221 + 57,100 + 14,923 + 50,676 + 2,975 + 23,878)
= 8,233

Bước 7: Tính trung bình bình phương cho từng nguồn biến động bằng cách chia mỗi tổng bình phương cho độ tự do tương ứng.

$$\text{MS lặp lại} = \frac{\text{SS lặp lại}}{r-1} = \frac{9,221}{2} = 4,61$$

$$\text{MS (A)} = \frac{\text{SS (A)}}{a-1} = \frac{57,100}{5} = 11,42$$

$$\text{MS sai số (a)} = \frac{\text{SS sai số (a)}}{(r-1)(a-1)} = \frac{14,923}{10} = 1,492$$

$$\text{MS (B)} = \frac{\text{SS (B)}}{b-1} = \frac{50,676}{2} = 25,338$$

$$\text{MS sai số (b)} = \frac{\text{SS sai số (b)}}{(r-1)(b-1)} = \frac{2,975}{4} = 0,744$$

$$\text{MS (A x B)} = \frac{\text{SS (AxB)}}{(a-1)(b-1)} = \frac{23,878}{10} = 2,388$$

$$\text{MS sai số (c)} = \frac{\text{SS sai số (c)}}{(r-1)(a-1)(b-1)} = \frac{8,233}{20} = 0,412$$

Bước 5: Tính các giá trị F như sau:

$$F(A) = \frac{\text{MS (A)}}{\text{MS sai số (a)}} = \frac{11,42}{1,492} = 7,65$$

$$F(B) = \frac{\text{MS (B)}}{\text{MS sai số (b)}} = \frac{25,338}{0,744} = 34,07$$

$$F(A \times B) = \frac{\text{MS (AxB)}}{\text{MS sai số (c)}} = \frac{2,388}{0,412} = 5,80$$

Bước 9: So sánh giá trị F tính của mỗi ảnh hưởng với giá trị F bảng tương ứng, ở Phụ lục VI, với df_1 là độ tự do của MS của mỗi ảnh hưởng và df_2 là độ tự do của MS của sai số ở mức ý nghĩa α. Ví dụ, giá trị F bảng tương ứng với F(AxB) có $df_1 = 10$ và $df_2 = 20$ là 2,35 ở $\alpha = 0,05$ và 3,37 ở $\alpha = 0,01$.

Bước 10: Tính ba hệ số biến động tương ứng ba trung bình bình phương sai số.

$$cv\,(a) = \frac{\sqrt{\text{MS sai số (a)}}}{\bar{x}} \times 100 = \frac{\sqrt{1,492}}{5,29} \times 100 = 23,1\ \%$$

$$cv\,(b) = \frac{\sqrt{\text{MS sai số (b)}}}{\bar{x}} \times 100 = \frac{\sqrt{0,744}}{5,29} \times 100 = 16,3\ \%$$

$$cv\,(b) = \frac{\sqrt{\text{MS sai số (c)}}}{\bar{x}} \times 100 = \frac{\sqrt{0,412}}{5,29} \times 100 = 12,1\ \%$$

Giá trị cv(a) chỉ độ chính xác liên quan đến nhân tố ngang, cv(b) nhân tố dọc và cv(c) liên quan đến sự tương tác giữa hai nhân tố. Giá trị của cv(c) được mong đợi nhỏ nhất; do đó, chỉ độ chính xác để đo ảnh hưởng tương tác cao nhất. Tuy nhiên, đối với cv(a) và cv(b) không có cơ sở để mong đợi cv(a) nhỏ hơn hoặc lớn hơn cv(b) hay ngược lại. Kết quả phân tích phương sai sau cùng được trình bày trong Bảng 6.20.

Bảng 6.20: Bảng phân tích phương sai của số liệu trong Bảng 6.15 của thí nghiệm thừa số 6 x 3 ở bố trí dải phụ[a]

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F(tính)	F(bảng) 5%	F(bảng) 1%
Lặp lại	2	9,221	4,610	3,09[ns]	4,10	7,56
Nhân tố ngang (A)	5	57,100	11,420	7,65**	3,33	5,64
Sai số (a)	10	14,923	1,492			
Nhân tố dọc (B)	2	50,676	25,338	[b]	4,10	7,56
Sai số (b)	4	2,975	0,744			
A X B	10	23,878	2,388	5,80**	2,35	3,37
Sai số (c)	20	8,233	0,412			
Tổng cộng	53	167,006				

[a]cv(a) = 23,1%, cv(b) = 16,3% và cv(c) = 12,1%

[b] df sai số (b) không đủ để kiểm định ý nghĩa có giá trị.

[ns] = không ý nghĩa

** = có ý nghĩa ở mức 1%.

CHƯƠNG 7

THÍ NGHIỆM NHIỀU NHÂN TỐ
(MULTI-FACTOR EXPERIMENTS)

Thí nghiệm hai nhân tố có thể được khai triển thêm nhân tố thứ ba, thứ tư,...Tuy nhiên, có hai hậu quả quan trọng khi thêm vào các nhân tố trong một thí nghiệm.

- Có sự gia tăng nhanh chóng số nghiệm thức được trắc nghiệm
- Có sự gia tăng số và loại ảnh hưởng tương tác. Ví dụ thí nghiệm ba nhân tố có bốn ảnh hưởng tương tác. Thí nghiệm bốn nhân tố có 10 ảnh hưởng tương tác.

Mặc dù thường chúng ta không mong muốn thí nghiệm lớn do chi phí cao và phức tạp, nhưng thông tin được bổ sung mang lại từ các ảnh hưởng tương tác giữa các nhân tố có thể rất có giá trị. Do đó, quyết định của nhà nghiên cứu về số nhân tố nên đưa vào trong thí nghiệm thừa số được dựa trên sự dàn xếp giữa việc mong muốn đánh giá càng nhiều ảnh hưởng tương tác càng tốt và cần giữ kích thước thí nghiệm trong giới hạn của nguồn sẵn có.

I. TƯƠNG TÁC GIỮA BA HOẶC NHIỀU NHÂN TỐ

Dựa vào định nghĩa tương tác hai nhân tố, tương tác k nhân tố (với k > 2) có thể được định nghĩa như sự chênh lệch giữa các ảnh hưởng tương tác (k – 1) nhân tố riêng biệt trên các mức độ khác nhau của nhân tố thứ k. Ví dụ, ảnh hưởng tương tác ba nhân tố giữa các nhân tố A, B và C (tương tác A x B x C) có thể được định nghĩa theo một trong ba cách sau đây:

- Sự chênh lệch giữa các ảnh hưởng tương tác A x B trên các mức độ của nhân tố C.
- Sự chênh lệch giữa các ảnh hưởng tương tác A x C trên các mức độ của nhân tố B.
- Sự chênh lệch giữa các ảnh hưởng tương tác B x C trên các mức độ của nhân tố A.

Để giải thích, chúng ta khảo sát thí nghiệm thừa số 2 x 2 x 2 = 2^3 bao gồm ba nhân tố, mỗi nhân tố có hai mức độ. Hai bộ số liệu giả thuyết, bộ (a) cho thấy có sự hiện diện của ảnh hưởng tương tác A x B x C, và bộ (b) vắng mặt ảnh hưởng tương tác A x B x C (Bảng 7.1)

Ảnh hưởng tương tác A x B x C có thể được đánh giá theo một trong ba phương pháp sau đây:

↳ **Phương pháp 1:** Dựa trên sự chênh lệch về các ảnh hưởng tương tác A x B

□ **Bước 1:** Ở mỗi mức độ của nhân tố C, tính ảnh hưởng tương tác A x B

Đối với số liệu bộ (a)

Ở c_0: tương tác A x B = ½ (0,5 - 2,0) = - 0,75 t/ha

Ở c_1: tương tác A x B = ½ (4,0 - 2,5) = 0,75 t/ha

Đối với số liệu bộ (b)

Ở c_0: tương tác A x B = ½ (3,5 - 2,0) = 0,75 t/ha

Ở c_1: tương tác A x B = ½ (3,5 - 2,5) = 0,75 t/ha

□ **Bước 2:** Tính ảnh hưởng tương tác A x B x C như sự chênh lệch giữa các ảnh hưởng tương tác A x B ở hai mức độ của nhân tố C được tính ở bước 1.

Đối với bộ (a): tương tác A x B x C = ½ [(0,75 – (- 0,75)] = 0,75 t/ha

Đối với bộ (b): tương tác A x B x C = ½ (0,75 - 0,75) = 0,00 t/ha

Bảng 7.1: Hai bộ số liệu giả thuyết của thí nghiệm thừa số 2 x 2 x 2 = 2^3[a]. Bộ (a) cho thấy có sự hiện diện của ảnh hưởng tương tác ba nhân tố và bộ (b) vắng mặt ảnh hưởng tương tác ba nhân tố.

Mức độ của nhân tố A	Năng suất hạt (t/ha)					
	c_0			c_1		
	b_0	b_1	$b_1 - b_0$	b_0	b_1	$b_1 - b_0$
	(a) hiện diện tương tác A x B x C					
a_0	2,0	3,0	1,0	2,5	5,0	2,5
a_1	4,0	3,5	- 0,5	5,0	9,0	4,0
$a_1 - a_0$	2,0	0,5	- 1,5	2,5	4,0	1,5
	(b) vắng mặt tương tác A x B x C					
a_0	2,0	2,5	0,5	3,0	3,5	0,5
a_1	4,0	6,0	2,0	5,0	7,0	2,0
$a_1 - a_0$	2,0	3,5	1,5	2,0	3,5	1,5

[a] Ba nhân tố A, B và C. mỗi nhân tố có 2 mức độ; a_0 và a_1 cho nhân tố A, b_0 và b_1 cho nhân tố B và c_0 và c_1 cho nhân tố C.

Phương pháp 2: Dựa trên sự chênh lệch về các ảnh hưởng tương tác A x C

□ **Bước 1:** Đối với mỗi mức độ của nhân tố B, tính ảnh hưởng tương tác A x C theo phương pháp ở chương 6, mục A-3.

Đối với số liệu bộ (a)

Ở b_0: tương tác A x C = ½ (2,5 - 2,0) = 0,25 t/ha

Ở b_1: tương tác A x C = ½ (4,0 - 0,5) = 1,75 t/ha

Đối với số liệu bộ (b)

Ở b_0: tương tác A x C = ½ (2,0 - 2,0) = 0,00 t/ha

Ở b_1: tương tác A x C = ½ (3,5 - 3,5) = 0,00 t/ha

□ **Bước 2:** Tính ảnh hưởng tương tác A x B x C như sự chênh lệch giữa các ảnh hưởng tương tác A x C ở hai mức độ của nhân tố B được tính ở bước 1.

Đối với bộ (a): Tương tác A x B x C = ½ [(1,75 – 0,25)] = 0,75 t/ha

Đối với bộ (b): Tương tác A x B x C = ½ (0,00 - 0,00) = 0,00 t/ha

Phương pháp 3: Dựa trên sự chênh lệch về các ảnh hưởng tương tác B x C

□ **Bước 1:** Đối với mỗi mức độ của nhân tố A, tính ảnh hưởng tương tác B x C theo phương pháp ở chương 6, mục A-3.

Đối với số liệu bộ (a)

Ở a_0: Tương tác B x C = ½ (2,5 - 1,0) = 0,75 t/ha

Ở a_1: Tương tác B x C = ½ [(4,0 – (-0,5)] = 2,25 t/ha

Đối với số liệu bộ (b)

Ở a_0: Tương tác B x C = ½ (0,5 - 0,5) = 0,0 t/ha

Ở a_1: Tương tác B x C = ½ (2,0 - 2,0) = 0,0 t/ha

□ **Bước 2:** Tính ảnh hưởng tương tác A x B x C như sự chênh lệch giữa các ảnh hưởng tương tác B x C ở hai mức độ của nhân tố A được tính ở bước 1.

Đối với bộ (a): Tương tác A x B x C = ½ (2,25 – 0,75) = 0,75 t/ha

Đối với bộ (b): Tương tác A x B x C = ½ (0,0 - 0,0) = 0,00 t/ha

Do đó, dù áp dụng phương pháp tính nào, ảnh hưởng tương tác A x B x C của số liệu bộ (a) là 0,75 t/ha và của bộ (b) là 0,0 t/ha.

Phương pháp tính ảnh hưởng tương tác ba nhân tố đã đề cập ở phần trên có thể dễ dàng mở rộng để bao hàm tương tác bốn nhân tố, tương tác năm nhân tố,... Chẳng hạn, tương tác bốn nhân tố A x B x C x D có thể tính theo một trong bốn cách sau đây:

- Bằng sự chênh lệch giữa các ảnh hưởng tương tác A x B x C trên các mức độ của nhân tố D.
- Bằng sự chênh lệch giữa các ảnh hưởng tương tác A x B x D trên các mức độ của nhân tố C.
- Bằng sự chênh lệch giữa các ảnh hưởng tương tác A x C x D trên các mức độ của nhân tố B.
- Bằng sự chênh lệch giữa các ảnh hưởng tương tác B x C x D trên các mức độ của nhân tố A.

II. CÁC KIỂU BỐ TRÍ THÍ NGHIỆM CÓ THỂ LỰA CHỌN

Có nhiều kiểu bố trí thí nghiệm có thể được sử dụng trong thí nghiệm ba hay nhiều nhân tố. Các bố trí này có thể phân thành bốn loại, cụ thể là bố trí thí nghiệm một nhân tố, bố trí thí nghiệm hai nhân tố, bố trí thí nghiệm ba hay nhiều nhân tố và bố trí thừa số phân nhỏ *(Fractional Factorial Designs)*.

1. Bố trí thí nghiệm một nhân tố

Tất cả kiểu bố trí đối với thí nghiệm một nhân tố được mô tả ở chương 5 đều có thể áp dụng cho các thí nghiệm ba hay nhiều nhân tố. Điều này được thực hiện bằng cách xem như tất cả các phối hợp nghiệm thức thừa số là các mức độ của một nhân tố.

Ví dụ, lấy trường hợp của thí nghiệm ba nhân tố bao gồm hai giống (V), bốn mức đạm (N) và ba phương pháp kiểm soát cỏ dại (W) được trắc nghiệm ba lần lặp lại. Nếu sử dụng bố trí khối hoàn toàn ngẫu nhiên (RCBD), 2 x 4 x 3 = 24 phối hợp nghiệm thức thừa số sẽ được ấn định hoàn toàn ngẫu nhiên vào 24 lô thí nghiệm ở một trong ba lần lặp lại. Bố trí ngoài đồng của kiểu bố trí như thế có thể được trình bày ở Hình 7.1 và phát thảo phân tích phương sai tương ứng được chỉ ở Bảng 7.2. Chú ý rằng đối với kiểu bố trí RCBD chỉ có một cỡ lô và chỉ có một phương sai sai số để kiểm định ý nghĩa của tất cả các ảnh hưởng (nghĩa là ba ảnh hưởng chính, ba ảnh hưởng tương tác hai nhân tố và một ảnh hưởng tương tác ba nhân tố) để tất cả các ảnh hưởng được đánh giá với cùng mức độ chính xác. Do đó, bố trí khối hoàn toàn (RCBD) chỉ nên được sử dụng với điều kiện là:

* Tất cả các ảnh hưởng (nghĩa là ảnh hưởng chính và các ảnh hưởng tương tác) đều có tầm quan trọng như nhau, và vì thế nên được đánh giá với cùng mức độ chính xác.

* Các đơn vị thí nghiệm đủ đồng nhất để đạt được mức độ đồng đều cao trong cùng một khối.

Do thí nghiệm có ba hoặc nhiều nhân tố thường kéo theo số nghiệm thức khá lớn, nên tính đồng đều ở các đơn vị thí nghiệm trong cùng một khối khó đạt được; do đó, bố trí khối hoàn toàn không được sử dụng phổ biến.

2. Bố trí thí nghiệm hai nhân tố

Tất cả các kiểu bố trí cho thí nghiệm hai nhân tố được mô tả ở chương 6 đều có thể áp dụng cho các thí nghiệm ba hoặc nhiều nhân tố. Lấy lại ví dụ thí nghiệm thừa số 2 x 4 x 3 đã được mô tả ở phần II-1.

$V_2N_0W_3$	$V_2N_3W_1$	$V_1N_3W_2$	$V_1N_2W_3$	$V_1N_3W_3$	$\mathbf{V_2N_2W_3}$
$V_1N_1W_1$	$V_2N_1W_1$	$V_1N_0W_3$	$V_2N_0W_2$	$V_1N_0W_1$	$\mathbf{V_2N_2W_1}$
$V_1N_1W_2$	$V_1N_0W_2$	$V_1N_1W_3$	$V_1N_2W_1$	$V_2N_3W_3$	$\mathbf{V_2N_1W_3}$
$V_1N_2W_2$	$V_1N_3W_1$	$V_2N_3W_2$	$V_2N_2W_2$	$V_2N_0W_1$	$\mathbf{V_2N_1W_2}$

Lặp lại I

$V_1N_3W_1$	$V_2N_2W_1$	$V_2N_0W_2$	$V_2N_2\mathbf{W_3}$	$V_1N_0W_2$	$V_1N_3W_2$
$V_2N_1W_1$	$V_2N_3W_3$	$V_2N_1W_2$	$V_1N_0\mathbf{W_3}$	$V_1N_2W_3$	$V_2N_2W_1$
$V_2N_3W_2$	$V_2N_1W_3$	$V_2N_0W_3$	$V_1N_1\mathbf{W_1}$	$V_1N_1W_2$	$V_1N_1W_3$
$V_1N_0W_1$	$V_2N_0W_1$	$V_1N_3W_3$	$V_2N_3\mathbf{W_1}$	$V_1N_2W_2$	$V_2N_2W_2$

Lặp lại II

$V_1N_2W_1$	$V_2N_3W_1$	$\mathbf{V_1N_1W_1}$	$V_2N_3W_3$	$V_1N_1W_2$	$V_2N_2W_3$
$V_2N_3W_2$	$V_1N_0W_2$	$\mathbf{V_2N_2W_2}$	$V_2N_1W_3$	$\mathbf{V_2N_0W_2}$	$V_2N_0W_3$
$V_2N_0W_1$	$V_1N_1W_3$	$\mathbf{V_2N_2W_1}$	$V_1N_2W_2$	$\mathbf{V_1N_0W_3}$	$V_1N_3W_3$
$V_2N_1W_1$	$V_1N_3W_2$	$\mathbf{V_1N_3W_1}$	$V_1N_2W_3$	$\mathbf{V_2N_1W_2}$	$V_1N_0W_1$

Lặp lại III

Hình 7.1: Cách bố trí mẫu của thí nghiệm thừa số 2 x 4 x 3 gồm có hai giống (V_1 và V_2), bốn mức độ đạm (N_0, N_1, N_2 và N_3), và ba phương pháp kiểm soát cỏ dại (W_1, W_2 và W_3) được sắp xếp theo kiểu bố trí khối hoàn toàn ngẫu nhiên với ba lần lặp lại.

Bảng 7.2: Phát thảo phân tích phương sai cho thí nghiệm thừa số 2 x 4 x 3 theo bố trí khối hoàn toàn ngẫu nhiên với ba lần lặp lại.

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F(tính)	F(bảng) 5%	F(bảng) 1%
Lặp lại	r – 1 = 2					
Giống (V)	(v – 1) = 1					
Đạm (N)	(n – 1) = 3					
Kiểm soát cỏ dại (W)	(w - 1) = 2					
V x N	(v – 1)(n – 1) = 1					
V x W	(v – 1)(w – 1) = 2					
N x W	(n – 1)(w – 1) = 6					
V x N x W	(v – 1)(n – 1)(w – 1) = 6					
Sai số	(r – 1)(vnw – 1) = 46					
Tổng cộng	rvnw – 1 = 71					

Chú thích: r = số lần lặp lại; v, n và w lần lượt là các mức độ của ba nhân tố V, N và W

Bước 1: Chia k nhân tố được trắc nghiệm thành hai nhóm, với k_1 nhân tố ở một nhóm và k_2 nhân tố ở nhóm khác (với $k_1 + k_2 = k$), bằng cách đặt các nhân tố được đánh giá có cùng mức độ chính xác vào chung nhóm. Do đó, nhóm I gồm có hai nhân tố ($k_1 = 2$) và nhóm II có một nhân tố ($k_2 = 1$)

Bước 2: Xem các phối hợp nghiệm thức thừa số của k_1 nhân tố ở nhóm I như các mức độ của một nhân tố được gọi là nhân tố A, và các phối hợp nghiệm thức thừa số của các nhân tố k_2 ở nhóm II như các mức độ của một nhân tố được gọi là nhân tố B. Do đó, thí nghiệm k nhân tố bây giờ được đổi thành thí nghiệm hai nhân tố bao gồm hai nhân tố A và B mới tạo ra.

Ví dụ, 3 x 2 = 6 phối hợp nghiệm thức thừa số giữa giống và kiểm soát cỏ dại ở nhóm I được xử lý như sáu mức độ của nhân tố A mới tạo ra, và bốn mức phân đạm như các mức độ của nhân tố B mới tạo ra. Vì vậy, thí nghiệm thừa số 2 x 4 x 3 bây giờ có thể được xem như thí nghiệm thừa số 6 x 4 gồm có hai nhân tố mới A và B.

Bước 3: Chọn kiểu bố trí thí nghiệm thích hợp ở chương 6 và áp dụng kiểu bố trí này cho hai nhân tố mới được tạo ra ở bước 2 theo phương pháp tương ứng đã được mô tả ở chương 6.

Ví dụ, nếu áp dụng bố trí lô phụ với nhân tố B (đạm) là nhân tố lô phụ, sơ đồ của kiểu bố trí như thế được trình bày ở Hình 7.2 và mẫu phân tích phương sai tương ứng được chỉ ở Bảng 7.3.

Chú ý rằng với bố trí lô phụ, có hai cỡ lô và hai trung bình bình phương sai số để kiểm định ý nghĩa của các ảnh hưởng khác nhau.

* Trung bình bình phương sai số (a) cho ảnh hưởng chính của giống, ảnh hưởng chính của phương pháp kiểm soát cỏ dại, và ảnh hưởng tương tác giữa chúng.

* Trung bình bình phương sai số (b) cho ảnh hưởng chính của phân đạm và ảnh hưởng tương tác của nó với hai nhân tố khác nhau.

Vì trung bình bình phương sai số (b) được mong đợi nhỏ hơn trung bình bình phương sai số (a) nên mức độ chính xác để đánh giá tất cả các ảnh hưởng liên quan đến phân đạm mong đợi cao hơn mức độ chính xác có liên quan đến giống hoặc kiểm soát cỏ dại.

Do đó, bố trí lô phụ (hoặc bố trí dải phụ) thích hợp cho thí nghiệm ba hoặc nhiều nhân tố nếu nắm vững cả hai điều kiện sau đây:

* Tổng số phối hợp nghiệm thức thừa số thì quá lớn đối với bố trí khối hoàn toàn.

* k nhân tố có thể được chia dễ dàng thành hai nhóm với sự khác biệt về mức độ chính xác mong muốn gắn liền với mỗi nhóm có thể nhận ra.

V_1W_2	V_2W_1	V_1W_3	V_2W_3	V_1W_1	V_2W_2
N_2	N_3	N_1	N_2	N_3	N_1
N_0	N_2	N_0	N_3	N_0	N_2
N_3	N_1	N_2	N_0	N_2	N_0
N_1	N_0	N_3	N_1	N_1	N_3

Lặp lại I

V_1W_2	V_2W_2	V_2W_3	V_1W_1	V_1W_3	V_2W_1
N_2	N_3	N_0	N_0	N_2	N_3
N_3	N_1	N_1	N_2	N_1	N_1
N_1	N_0	N_3	N_1	N_3	N_0
N_0	N_2	N_2	N_3	N_0	N_2

Lặp lại II

V_2W_1	V_2W_2	V_1W_2	V_1W_1	V_2W_3	V_1W_3
N_2	N_3	N_0	N_1	N_3	N_2
N_3	N_1	N_3	N_2	N_1	N_0
N_0	N_2	N_1	N_3	N_0	N_1
N_1	N_0	N_2	N_0	N_2	N_3

Lặp lại III

Hình 7.2: Cách bố trí mẫu của thí nghiệm thừa số 2 x 4 x 3 gồm có hai giống (V_1 và V_2), bốn mức độ đạm (N_0, N_1, N_2 và N_3), và ba phương pháp kiểm soát cỏ dại (W_1, W_2 và W_3) được sắp xếp theo kiểu bố trí lô phụ với các mức độ đạm là các nghiệm thức lô phụ, ba lần lặp lại.

Bảng 7.3: Phát thảo phân tích phương sai cho thí nghiệm thừa số 2 x 4 x 3 theo kiểu bố trí lô phụ với ba lần lặp lại.

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F(tính)	F(bảng) 5%	F(bảng) 1%
Lặp lại	r – 1 = 2					
Nhân tố lô chính (VW)	vw – 1 = 5					
Giống (V)	(v – 1) = 1					
Kiểm soát cỏ dại (W)	(w – 1) = 2					
V x W	(v – 1)(w - 1) = 2					
Sai số (a)	(r – 1)(vw – 1) = 10					
Nhân tố lô phụ (N)	(n – 1) = 3					
Nhân tố lô chính x nhân tố lô phụ	(vw – 1)(n – 1) = 15					
N x V	(n – 1)(v - 1) = 3					
N x W	(n – 1)(w - 1) = 6					
N x V x W	(n – 1)(v – 1)(w – 1) = 6					
Sai số (b)	vw(r – 1)(n – 1) = 36					
Tổng cộng	rvnw – 1 = 71					

Áp dụng cho thí nghiệm hai nhân tố mới tạo ra, với nhân tố lô chính là phối hợp của hai nhân tố gốc V và W, và nhân tố lô phụ đại diện cho nhân tố gốc thứ ba N
r = Số lần lặp lại; v, n và w lần lượt là các mức độ của ba nhân tố V, N và W

3. Bố trí thí nghiệm ba hoặc nhiều nhân tố

Các bố trí thí nghiệm được phát triển riêng biệt cho thí nghiệm ba hay nhiều nhân tố được sử dụng phổ biến trong nghiên cứu nông nghiệp chủ yếu là sự mở rộng của bố trí lô phụ hoặc dải phụ.

Ví dụ, bố trí lô phụ có thể được mở rộng để đưa vào nhân tố thứ ba qua việc chia nhỏ thêm mỗi lô phụ thành các lô phụ-trong lô phụ *(sub-subplots)* và mở rộng thêm nữa để đưa vào nhân tố thứ tư qua việc chia nhỏ thêm mỗi lô phụ-trong lô phụ thành các lô phụ-trong lô phụ-trong lô phụ *(sub-sub-subplots)*,... Các kiểu bố trí được kể đến là bố trí lô phụ-trong lô phụ *(Split-split-plot design)*, bố trí lô phụ-trong lô phụ-trong lô phụ *(Split-split-split-plot design)*,... Bố trí lô phụ-trong lô phụ áp dụng cho thí nghiệm ba nhân tố có nhân tố đầu tiên được ấn định cho lô chính, nhân tố thứ hai cho lô phụ và nhân tố thứ ba cho lô phụ-trong lô phụ. Theo cách này, không cần phối hợp ba nhân tố thành hai nhóm để tái tạo thí nghiệm hai nhân tố giống như áp dụng bố trí lô phụ cho thí nghiệm ba nhân tố.

Tương tự, bố trí dải phụ cũng có thể mở rộng để sáp nhập nhân tố thứ ba, nhân tố thứ tư,... qua việc chia nhỏ mỗi lô giao nhau thành các lô phụ và chia nhỏ thêm nữa của mỗi lô phụ thành lô phụ-trong lô phụ. Kết quả các bố trí được kể đến là bố trí dải phụ-trong dải phụ *(Strip-split-plot design)*, bố trí dải phụ-trong dải phụ-trong dải phụ *(Strip-split-split-plot design)*,...

4. Bố trí thừa số phân nhỏ *(Fractional Factorial Designs = FFD)*

Không giống như các kiểu bố trí áp dụng cho thí nghiệm một, hai hay nhiều nhân tố, ở đó đòi hỏi trong thí nghiệm phải chứa đầy đủ các phối hợp nghiệm thức thừa số. Bố trí thừa số phân nhỏ, như tên gọi, chỉ chứa một phần của bộ đầy đủ các phối hợp nghiệm thức thừa số. Ưu điểm hiển nhiên của FFD là giảm được độ lớn của thí nghiệm, mà điều này có lẽ mong muốn bất cứ lúc nào bộ phối hợp nghiệm thức thừa số quá lớn đến nỗi trên thực tế không thể thực hiện đầy đủ. Tuy nhiên, ưu điểm quan trọng đạt được qua việc làm giảm một số ảnh hưởng có thể được ước lượng.

III. BỐ TRÍ KHỐI HOÀN TOÀN NGẪU NHIÊN

Ví dụ: Khảo sát ảnh hưởng của 3 nhân tố: lysine, methionine và protein lên sự tăng trọng của heo, một nhà nghiên cứu thực hiện thí nghiệm thừa số 4 x 3 x 2, bao gồm 4 mức lysine, 3 mức methionine và 2 mức protein, theo kiểu bố trí khối hoàn toàn ngẫu nhiên (RCBD) với 2 lần lặp lại tương ứng với hai cách nuôi khác nhau. Kết quả tăng trọng trung bình mỗi ngày (kg) ghi nhận được trong Bảng 7.4.

Bảng 7.4: Sự tăng trọng trung bình mỗi ngày (kg) của heo được nuôi với khẩu phần ăn có bổ sung tỷ lệ phần trăm khác nhau của lysine, methionine và protein (Trích từ Bảng 12.10.1 của Jean Collin. 2003. Trang 192)

Lysine L (g.kg^{-1})	Methionine M (g.kg^{-1})	Protein P (%)	Lặp lại (Khối) I	Lặp lại (Khối) II	Tổng nghiệm thức	T.B. N.thức
0	0	12	1,11	0,97	2,08	1,040
		14	1,52	1,45	2,97	1,485
	0,025	12	1,09	0,99	2,08	1,040
		14	1,27	1,22	2,49	1,245
	0,050	12	0,85	1,21	2,06	1,030
		14	1,67	1,24	2,91	1,544
0,05	0	12	1,30	1,00	2,30	1,150
		14	1,55	1,53	3,08	1,540
	0,025	12	1,03	1,21	2,24	1,120
		14	1,24	1,34	2,58	1,290
	0,050	12	1,12	0,96	2,08	1,040
		14	1,76	1,27	3,03	1,515
0,10	0	12	1,22	1,13	2,35	1,175
		14	1,38	1,08	2,46	1,230
	0,025	12	1,34	1,41	2,75	1,375
		14	1,40	1,21	2,61	1,305
	0,050	12	1,34	1,19	2,53	1,265
		14	1,46	1,39	2,85	1,425
0,15	0	12	1,19	1,03	2,22	1,110
		14	0,80	1,29	2,09	1,045
	0,025	12	1,36	1,16	2,52	1,260
		14	1,42	1,39	2,81	1,405
	0,050	12	1,46	1,03	2,49	1,245
		14	1,62	1,27	2,89	1,445
Tổng			**31,50**	**28,97**	**60,47**	**1,260**

2. Phân tích phương sai:

Số liệu tăng trọng của heo (Bảng 7.4) của thí nghiệm thừa số 4 x 3 x 2 được thực hiện theo kiểu bố trí khối hoàn toàn ngẫu nhiên với hai lần lặp lại để sử dụng giải thích phân tích phương sai.

Bước 1: Phát thảo bảng phân tích phương sai cho thí nghiệm thừa số 4 x 3 x 2, bố trí khối hoàn toàn ngẫu nhiên với hai lần lặp lại (Bảng 7.5)

Bước 2: Tính tổng nghiệm thức (T), tổng lặp lại (R) và tổng chung (G) như trong Bảng 7.4. Tính yếu tố hiệu chỉnh (CF) và các tổng bình phương (SS): tổng cộng, lặp

lại, nghiệm thức và sai số giống như cách tính của bố trí RCBD trong thí nghiệm một nhân tố.

Bảng 7.5: Phát thảo phân tích phương sai cho thí nghiệm thừa số 4 x 3 x 2 theo bố trí khối hoàn toàn ngẫu nhiên với hai lần lặp lại.

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F(tính)	F(bảng) 5%	1%
Lặp lại	r – 1 = 1					
Lysine (L)	(l – 1) = 3					
Methionine (M)	(m – 1) = 2					
L x M	(l – 1)(m – 1) = 6					
Protein (P)	(p – 1) = 1					
L x M	(l – 1)(m – 1) = 3					
L x P	(l – 1)(p – 1) = 2					
L x M x P	(l – 1)(m – 1)(p – 1) = 6					
Sai số	(r – 1)(vnw – 1) = 23					
Tổng cộng	rlmp – 1 = 47					

Chú thích: r = số lần lặp lại; l, m và p lần lượt là các mức độ của ba nhân tố L, M và P

Gọi l là số mức độ của nhân tố lysine (l = 4)
m là số mức độ của nhân tố methionine (m = 3)
p là số mức độ của nhân tố protein (p = 2)
r là số lần lặp lại (r = 2)

$$C.F. = \frac{G^2}{rlmp} = \frac{(60,47)^2}{(2)(4)(3)(2)} = 76,1796$$

$$SS\ tổng\ cộng = \sum X^2 - C.F.$$
$$= [(1,11)^2 + (1,52)^2 + \cdots + (1,27)^2] - 76,1796 = 2,041$$

$$SS\ lặp\ lại = \frac{\sum R^2}{lmp} - C.F. = \frac{(31,50)^2 + (28,97)^2}{(4)(3)(2)} - 76,1796 = 0,133$$

$$SSnghiệm\ thức = \frac{\sum T^2}{r} - C.F. = \frac{(2,08)^2 + \cdots + (2,89)^2}{2} - 76,1796 = 1,276$$

SS sai số = SS tổng cộng - SS lặp lại - SS nghiệm thức
= 2,041 - 0,133 - 1,276 = 0,632

Bước 2: Lập các bảng tổng của hai nhân tố và ba nhân tố

- Lặp bảng tổng lysine x methionine (L x M). Tính tổng lysine (L), tổng methionine (M) và các tổng bình phương (Bảng 7.6).

Bảng 7.6: Tổng của lysine x methionine (L x M) được tính từ số liệu ở bảng 7.5

Methionine	Lysine 0	0,05	0,10	0,15	Tổng methionine (M)
0	5.05	5.38	4.81	4.31	19.55
0,025	4.57	4.82	5.36	5.33	20.08
0,050	4.97	5.11	5.38	5.38	20.84
Tổng lysine (L)	14.59	15.31	15.55	15.02	60,47

$$SS\ L(lysine) = \frac{\sum L^2}{rmp} - C.F.$$

$$= \frac{(14,59)^2 + (15,31)^2 + (15,55)^2 + (15,02)^2}{(2)(3)(2)} - 76,1796 = 0,043$$

$$\text{SS M(methionine)} = \frac{\sum M^2}{rlp} - \text{C.F.}$$

$$= \frac{(19,55)^2 + (20,08)^2 + (20,84)^2}{(2)(4)(2)} - 76,1796 = 0,053$$

$$\text{SS LxM} = \frac{\sum (LM)^2}{rp} - \text{C.F.} - \text{SS(L)} - \text{SS(M)}$$

$$= \frac{(5,05)^2 + \cdots + (5,38)^2}{(2)(2)} - 76,1796 - 0,043 - 0,053 = 0,254$$

- Lập bảng tổng lysine x protein (L x P). Tính tổng protein (P) và các tổng bình phương (Bảng 7.7).

Bảng 7.7: Tổng của lysine x protein (L x P) được tính từ số liệu ở bảng 7.5

Protein	Lysine				Tổng protein
	0	0,05	0,10	0,15	(P)
12	6,22	6,62	7,63	7,23	27,70
14	8,37	8,69	7,92	7,79	32,77

$$\text{SS P (protein)} = \frac{\sum P^2}{rlm} - \text{C.F.}$$

$$= \frac{(27,70)^2 + (32,77)^2}{(2)(2)(3)} - 76,1796 = 0,536$$

$$\text{SS L x P} = \frac{\sum (LP)^2}{rm} - \text{C.F.} - \text{SS(L)} - \text{SS(P)}$$

$$= \frac{(6,22)^2 + \cdots + (7,79)^2}{(2)(3)} - 76,1796 - 0,043 - 0,536 = 0,240$$

- Lập bảng tổng methionine x protein (M x P). Tính SS tương tác M x P (Bảng 7.8)

Bảng 7.8: Tổng methionine x protein (M x P) được tính từ số liệu ở bảng 7.5

Protein	Methionine		
	0	0,025	0,050
12	8.95	9.59	9.16
14	10.60	10.49	11.68

$$\text{SS MxP} = \frac{\sum (MP)^2}{rl} - C.F. - SS(M) - SS(P)$$

$$= \frac{(8,95)^2 + \cdots + (11,68)^2}{(2)(4)} - 76,1796 - 0,053 - 0,536 = 0,082$$

SS LxMxP = SSnghiệm thức – SS L – SS M – SS P

= 1,276 – 0,043 – 0,053 – 0,536 = 0,068

Bước 3: Tính trung bình bình phương (MS) cho từng nguồn biến động bằng cách chia mỗi tổng bình phương (SS) cho độ tự do tương ứng (df).

$$\text{MS lặp lại} = \frac{\text{SS lặp lại}}{r-1} = \frac{0,133}{1} = 0,133$$

$$MS(L) = \frac{SS(L)}{l-1} = \frac{0,043}{3} = 0,014$$

$$MS(M) = \frac{SS(M)}{m-1} = \frac{0,053}{2} = 0,026$$

$$MS(LxM) = \frac{SS(LxM)}{(l-1)(m-1)} = \frac{0,254}{6} = 0,042$$

$$MS(P) = \frac{SS(P)}{p-1} = \frac{0,536}{1} = 0,536$$

$$MS(LxP) = \frac{SS(LxP)}{(l-1)(p-1)} = \frac{0,240}{3} = 0,080$$

$$MS(MxP) = \frac{SS(MxP)}{(m-1)(p-1)} = \frac{0,082}{2} = 0,041$$

$$MS(LxMxP) = \frac{SS(LxMxP)}{(l-1)(m-1)(p-1)} = \frac{0,068}{6} = 0,011$$

$$\text{MS sai số} = \frac{\text{SS sai số}}{lmp(r-1)} = \frac{0,632}{23} = 0,027$$

Bước 4: Tính các giá trị F như sau:

$$F(L) = \frac{MS(L)}{\text{MS sai số}} = \frac{0,014}{0,027} = 0,517$$

$$F(M) = \frac{MS(M)}{\text{MS sai số}} = \frac{0,026}{0,027} = 0,956$$

$$F(LxM) = \frac{MS(LxM)}{\text{MS sai số}} = \frac{0,042}{0,027} = 1,543$$

$$F(P) = \frac{MS(P)}{\text{MS sai số}} = \frac{0,536}{0,027} = 19,492$$

$$F(LxP) = \frac{MS(LxP)}{\text{MS sai số}} = \frac{0,080}{0,027} = 2,911$$

$$F(MxP) = \frac{MS(MxP)}{\text{MS sai số}} = \frac{0,041}{0,027} = 1,495$$

$$F(LxMxP) = \frac{MS(LxMxP)}{\text{MS sai số}} = \frac{0,011}{0,027} = 0,416$$

Bước 5: So sánh giá trị F (tính) của mỗi ảnh hưởng với giá trị F (bảng) tương ứng, trong Phụ lục VI, với df_1 là độ tự do của MS của từng ảnh hưởng và df_2 là độ tự do của MS sai số ở mức ý nghĩa α. Ví dụ, giá trị F (bảng) tương ứng với F(L) có $df_1 = 3$ và $df_2 = 23$ là 3,03 ở $\alpha = 0,05$ và 4,76 ở $\alpha = 0,01$ (Bảng 7.9).

Bước 6: Tính hệ số biến động

$$cv = \frac{\sqrt{MSsai\ số}}{\bar{x}} x\,100 = \frac{\sqrt{0,027}}{1,26} x\,100 = 13,16\,\%$$

Bảng 7.9: Phân tích phương sai cho thí nghiệm thừa số 4 x 3 x 2 theo kiểu bố trí khối hoàn toàn ngẫu nhiên với hai lần lặp lại.

Nguồn biến động	Độ tự do	Tổng bình phương	Trung bình bình phương	F(tính)	F(bảng) 5%	F(bảng) 1%
Lặp lại	1	0,133	0,133	4,854*	4,28	7,88
Lysine (L)	3	0,043	0,014	< 1	-	-
Methionine (M)	2	0,053	0,026	< 1	-	-
L x M	6	0,254	0,042	$1,543^{ns}$	2,53	3,71
Protein (P)	1	0,536	0,536	19,492**	4,28	7,88
L x P	3	0,240	0,080	$2,911^{ns}$	3,03	4,76
M x P	2	0,082	0,041	$1,495^{ns}$	3,42	5,66
L x M x P	6	0,068	0,011	< 1	-	-
Sai số	23	0,632	0,027			
Tổng cộng	47	2,041				

cv = 13,16%

* = Ý nghĩa ở mức 5%, ** = Ý nghĩa ở mức 1% và ^{ns} = không ý nghĩa

IV. BỐ TRÍ LÔ PHỤ TRONG LÔ PHỤ *(Split-split-plot design)*

Bố trí lô phụ trong lô phụ là sự khai triển của bố trí lô phụ để đưa vào thêm nhân tố thứ ba. Đây là kiểu bố trí duy nhất thích hợp với thí nghiệm ba nhân tố, ở đó chúng ta mong muốn ba mức độ chính xác khác nhau đối với các ảnh hưởng khác nhau. Mỗi mức độ chính xác ấn định cho các ảnh hưởng liên quan đến một trong ba nhân tố. Bố trí này được tiêu biểu bởi hai tính chất quan trọng.

(1) Có ba cỡ lô tương ứng với ba nhân tố, cụ thể là lô lớn nhất *(lô chính)* đối với nhân tố lô chính, lô trung gian *(lô phụ)* cho nhân tố lô phụ, và lô nhỏ nhất *(lô phụ-trong lô phụ)* cho nhân tố lô phụ-trong lô phụ.

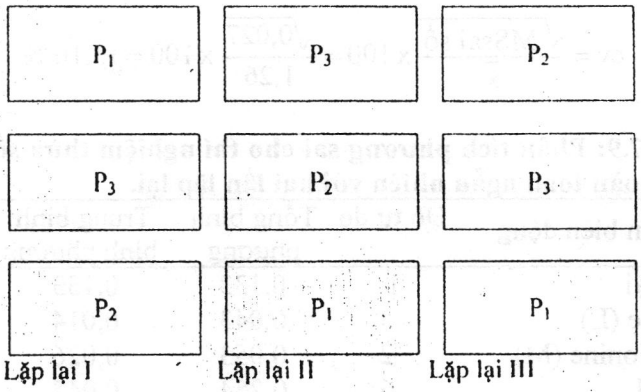
(2) Có ba mức độ chính xác, với nhân tố lô chính nhận mức độ chính xác thấp nhất và nhân tố lô phụ-trong lô phụ nhận mức độ chính xác cao nhất.

Phương pháp làm ngẫu nhiên, cách bố trí và phân tích phương sai sẽ được giải thích trong thí nghiệm thừa số 3 x 2 x 3 với ba lần lặp lại. Các nghiệm thức là ba thời điểm cấy khác nhau ở cây củ cải *(nhân tố lô chính)*, hai biện pháp kiểm soát rầy mềm *(nhân tố lô phụ)* và ba thời điểm thu hoạch khác nhau *(nhân tố lô phụ-trong lô phụ)*. Sử dụng ký hiệu r cho số lần lặp lại; A, B và C lần lượt là nhân tố lô chính, lô phụ và lô phụ-trong lô phụ; a, b và c là các mức độ nghiệm thức tương ứng với các nhân tố A, B và C.

1. Cách bố trí và làm ngẫu nhiên:

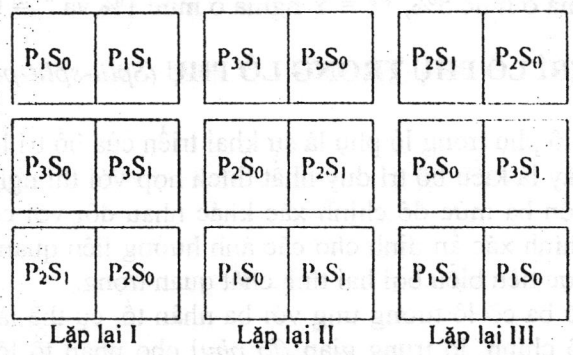
Đối với kiểu bố trí lô phụ-trong lô phụ có ba bước thực hiện cách bố trí và làm ngẫu nhiên:

Bước 1: Chia khu thí nghiệm thành r = 3 lần lặp lại và mỗi lần lặp lại thành a = 3 lô chính tương ứng với ba thời điểm cấy (P_1, P_2 và P_3). Ở mỗi lần lặp lại, ấn định ngẫu nhiên ba nghiệm thức lô chính vào ba lô chính tách biệt và độc lập. Kết quả được trình bày ở Hình 7.3.



Hình 7.3: Ấn định ngẫu nhiên ba thời điểm cấy (P_1, P_2 và P_3) vào các lô chính ở một trong ba lần lặp lại

Bước 2: Chia mỗi lô chính thành b = 2 lô phụ, trong đó hai nghiệm thức lô phụ tương ứng với hai biện pháp kiểm soát rầy mềm (S_0 và S_1) được ấn định ngẫu nhiên, tách biệt và độc lập vào một trong (r)(a) = (3)(3) = 9 lô chính. Kết quả được trình bày ở Hình 7.4.



Hình 7.4: Ấn định ngẫu nhiên hai biện pháp kiểm soát rầy mềm (S_0 và S_1) vào hai lô phụ ở một trong 9 lô chính

Bước 3: Chia mỗi lô phụ thành c = 3 lô phụ-trong lô phụ. Ấn định ngẫu nhiên vào đó ba thời điểm thu hoạch (H_1, H_2 và H_3). Tiến trình làm ngẫu nhiên này được lặp lại (r)(a)(b) = (3)(3)(2) = 18 lần. Bố trí sau cùng được trình bày ở Hình 7.5.

2. Phân tích phương sai:

Số liệu năng suất củ cải (Bảng 7.10) của thí nghiệm thừa số 3 x 2 x 3 được thực hiện theo kiểu bố trí lô phụ-trong lô phụ được sử dụng để giải thích phương pháp phân tích phương sai.

Bước 1: Phát thảo bảng phân tích phương sai cho bố trí lô phụ-trong lô phụ (Bảng 7.11)

Bước 2: Thực hiện phân tích lô chính.

- Lập bảng tổng lặp lại x nhân tố A (Bảng 7.12). Tính tổng lặp lại (R), tổng nhân tố A và tổng chung (G).

$P_1S_0H_2$	$P_1S_0H_1$	$P_1S_0H_3$
$P_1S_1H_3$	$P_1S_1H_1$	$P_1S_1H_2$
$P_3S_0H_1$	$P_3S_0H_3$	$P_3S_0H_2$
$P_3S_1H_2$	$P_3S_1H_3$	$P_3S_1H_1$
$P_2S_1H_3$	$P_2S_1H_2$	$P_2S_1H_1$
$P_2S_0H_1$	$P_2S_0H_2$	$P_2S_0H_3$

Lặp lại I

$P_3S_1H_1$	$P_3S_1H_3$	$P_3S_1H_2$
$P_3S_0H_3$	$P_3S_0H_2$	$P_3S_0H_1$
$P_2S_0H_2$	$P_2S_0H_1$	$P_2S_0H_3$
$P_2S_1H_1$	$P_2S_1H_3$	$P_2S_1H_2$
$P_1S_0H_1$	$P_1S_0H_2$	$P_1S_0H_3$
$P_1S_1H_2$	$P_1S_1H_1$	$P_1S_1H_3$

Lặp lại II

$P_2S_1H_1$	$P_2S_1H_2$	$P_2S_1H_3$
$P_2S_0H_2$	$P_2S_0H_3$	$P_2S_0H_1$
$P_3S_0H_2$	$P_3S_0H_3$	$P_3S_0H_1$
$P_3S_1H_3$	$P_3S_1H_1$	$P_3S_1H_2$
$P_1S_1H_3$	$P_1S_1H_1$	$P_1S_1H_2$
$P_1S_0H_1$	$P_1S_0H_2$	$P_1S_0H_3$

Lặp lại III

Hình 7.5: Bố trí mẫu của thí nghiệm thừa số 3 x 2 x 3 được sắp xếp trong bố trí theo kiểu lô phụ-trong lô phụ với ba thời điểm cấy (P_1, P_2 và P_3) là nghiệm thức lô chính, hai biện pháp kiểm soát rầy mềm (S_0 và S_1) là nghiệm thức lô phụ và thời điểm thu hoạch (H_1, H_2 và H_3) là nghiệm thức lô phụ-trong lô phụ ở ba lần lặp lại.

Bảng 7.10: Năng suất củ cải (t/ha) của thí nghiệm kiểm soát rầy mềm. Bố trí lô phụ trong lô phụ với 3 lần lặp lại (Trích từ Bảng 9.1 của Little and Hills, 1978. Trang 103).

Nghiệm thức			Lặp lại			Tổng	Trung bình
P	S	H	I	II	III	nghiệm thức	nghiệm thức
1	1	1	25,7	25,4	23,8	74,9	24,97
		2	31,8	29,5	28,7	90,0	30,00
		3	34,6	37,2	29,1	100,9	33,63
	2	1	27,7	30,3	30,2	88,2	29,40
		2	**38,0**	**40,6**	**34,6**	**113,2**	**37,73**
		3	42,1	43,6	44,6	130,3	43,43
2	1	1	28,9	24,7	27,8	81,4	27,13
		2	37,5	31,5	31,0	100,0	33,33
		3	38,4	32,5	31,2	102,1	34,03
	2	1	38,0	31,0	29,5	98,5	32,83
		2	36,9	31,9	31,5	100,3	33,43
		3	44,2	41,6	38,9	124,7	41,57
3	1	1	23,4	24,2	21,2	68,8	22,93
		2	25,3	27,7	23,7	76,7	25,57
		3	29,8	29,9	24,3	84,0	28,00
	2	1	20,8	23,0	25,2	69,0	23,00
		2	29,0	32,0	26,5	87,5	29,17
		3	36,6	37,8	34,8	109,2	36,40

Gọi a là số mức độ của nhân tố lô chính (Thời điểm cấy), a = 3.

b là số mức độ của nhân tố lô phụ (biện pháp kiểm soát rầy mềm), b = 2.

c là số mức độ của nhân tố lô phụ trong lô phụ (Thời điểm thu hoạch), c = 3.

r là số lần lặp lại, r = 3.

- Tính yếu tố hiệu chỉnh (CF) và các tổng bình phương.

$$C.F. = \frac{G^2}{rabc} = \frac{(1699,7)^2}{(3)(3)(2)(3)} = 53499,631$$