

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/352933891>

Giáo trình Thiết kế thí nghiệm (tái bản lần 1)

Book · May 2017

CITATIONS

0

READS

50

1 author:



Do Duc Luc

Vietnam National University of Agriculture

67 PUBLICATIONS 85 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Ho Chicken in Bac Ninh Province (Vietnam): From an Indigenous Chicken to Local Poultry Breed [View project](#)



Study on natural resistance to African swine fever of surviving pigs in outbreak areas in North Vietnam [View project](#)

HỌC VIỆN NÔNG NGHIỆP VIỆT NAM

ĐỖ ĐỨC LỰC | NGUYỄN ĐÌNH HIỀN | HÀ XUÂN BỘ
ĐỖ ĐỨC LỰC | NGUYỄN ĐÌNH HIỀN (Đồng chủ biên)

GIÁO TRÌNH
THIẾT KẾ THÍ NGHIỆM

(Tái bản lần thứ I)

NHÀ XUẤT BẢN ĐẠI HỌC NÔNG NGHIỆP – 2017

MỤC LỤC

LỜI MỞ ĐẦU	viii
LỜI TỰA	x
PHẦN A - LÝ THUYẾT	1
<i>Chương 1 - MỘT SỐ KHÁI NIỆM TRONG XÁC SUẤT THỐNG KÊ MÔ TẢ</i>	<i>1</i>
<i>1.1. TÓM TẮT VỀ XÁC SUẤT VÀ BIẾN NGẪU NHIÊN</i>	<i>1</i>
1.1.1. Xác suất cơ bản.....	1
1.1.2. Hệ sự kiện đầy đủ	1
1.1.3. Biến ngẫu nhiên, bảng phân phối, hàm phân phối	2
1.1.4. Một số phân phối thường gặp	2
<i>1.2. BIẾN SINH HỌC</i>	<i>4</i>
1.2.1. Khái niệm về biến sinh học	4
1.2.2. Tổng thể và mẫu	5
1.2.3. Sơ lược về cách chọn mẫu	5
1.2.4. Các tham số thống kê của mẫu	6
1.2.5. Biểu diễn số liệu bằng đồ thị	11
<i>1.3. BÀI TẬP</i>	<i>13</i>
<i>Chương 2 - ƯỚC LUỢNG VÀ KIỂM ĐỊNH GIÁ THIẾT</i>	<i>15</i>
<i>2.1. GIÁ THIẾT VÀ ĐÓI THIẾT</i>	<i>15</i>
<i>2.2. ƯỚC LUỢNG GIÁ TRỊ TRUNG BÌNH μ CỦA BIẾN PHÂN PHỐI CHUẨN $N(\mu, \sigma^2)$.....</i>	<i>16</i>
2.2.1. Ước lượng μ khi biết phương sai σ^2	16
2.2.2. Ước lượng μ khi không biết phương sai σ^2	17
<i>2.3. KIỂM ĐỊNH GIÁ TRỊ TRUNG BÌNH μ CỦA BIẾN PHÂN PHỐI CHUẨN $N(\mu, \sigma^2)$.....</i>	<i>17</i>
2.3.1. Kiểm định giả thiết $H_0: \mu = \mu_0$ khi biết σ^2	17
2.3.2. Kiểm định giả thiết $H_0: \mu = \mu_0$ khi không biết σ^2	18
<i>2.4. KIỂM ĐỊNH HAI GIÁ TRỊ TRUNG BÌNH CỦA HAI BIẾN PHÂN PHỐI CHUẨN</i>	<i>19</i>
2.4.1. Chọn mẫu theo cặp	20
2.4.2. Chọn mẫu độc lập	21
<i>2.5. ƯỚC LUỢNG VÀ KIỂM ĐỊNH XÁC SUẤT</i>	<i>24</i>
2.5.1. Ước lượng xác suất P	24
2.5.2. Kiểm định giả thiết $H_0: P = P_0$	25
2.5.3. Kiểm định giả thiết $H_0: P_2 = P_1$	25
<i>2.6. PHÂN TÍCH PHƯƠNG SAI.....</i>	<i>26</i>
<i>2.7. BÀI TẬP</i>	<i>30</i>

<i>Chương 3 - MỘT SỐ KHÁI NIỆM VỀ THIẾT KẾ THÍ NGHIỆM.....</i>	32
3.1. PHÂN LOẠI THÍ NGHIỆM.....	32
3.1.1. Thí nghiệm quan sát	32
3.1.2. Thí nghiệm thực nghiệm.....	32
3.2. MỘT SỐ KHÁI NIỆM TRONG THIẾT KẾ THÍ NGHIỆM.....	33
3.2.1. Yếu tố thí nghiệm	33
3.2.2. Mức	33
3.2.3. Công thức thí nghiệm (công thức thí nghiệm)	33
3.2.4. Đơn vị thí nghiệm.....	33
3.2.5. Dữ liệu (số liệu).....	34
3.2.6. Khối	34
3.2.7. Lặp lại	34
3.2.8. Nhắc lại.....	34
3.2.9. Nhóm đối chứng	34
3.3. CÁC BƯỚC TIẾN HÀNH THÍ NGHIỆM	34
3.4. SAI SÓ THÍ NGHIỆM.....	35
3.5. BỎ TRÍ THÍ NGHIỆM VÀO CÁC CÔNG THỨC THÍ NGHIỆM	36
3.5.1. Sự cần thiết của phân chia ngẫu nhiên.....	36
3.5.2. Các phương pháp phân chia ngẫu nhiên	37
3.6. PHƯƠNG PHÁP LÀM MÙ.....	39
3.7. TĂNG ĐỘ CHÍNH XÁC CỦA UỚC TÍNH.....	39
3.7.1. Lặp lại	39
3.7.2. Kỹ thuật khối	39
3.7.3. Kỹ thuật cặp.....	39
3.8. DUNG LUỢNG MẪU CẦN THIẾT.....	40
3.8.1. Số công thức thí nghiệm	41
3.8.2. Bậc tự do của sai số ngẫu nhiên.....	49
3.8.3. Phương pháp chọn mẫu	50
3.9. BÀI TẬP	55
<i>Chương 4 - THIẾT KẾ THÍ NGHIỆM MỘT YẾU TỐ.....</i>	56
4.1. THÍ NGHIỆM HOÀN TOÀN NGẪU NHIÊN (<i>Completely randomized Design - CRD</i>).....	56
4.1.1. Đặc điểm.....	56
4.1.2. Chất lượng động vật	56
4.1.3. Dung lượng mẫu cần thiết	57
4.1.4. Ưu điểm và nhược điểm	59
4.1.5. Cách thiết kế thí nghiệm	59
4.1.6. Phân tích số liệu.....	60

<i>4.2. THÍ NGHIỆM KHỐI NGĂU NHIÊN ĐÂY ĐỦ (Randomized complete block design - RCBD)</i>	63
4.2.1. Só khói cần thiết	64
4.2.2. Ưu điểm và nhược điểm	65
4.2.3. Cách thiết kế thí nghiệm	65
4.2.4. Phân tích số liệu.....	66
<i>4.3. THÍ NGHIỆM KHỐI NGĂU NHIÊN VỚI NHIỀU ĐƠN VỊ THÍ NGHIỆM TRONG MỘT CÔNG THỨC THÍ NGHIỆM VÀ KHỐI</i>	69
4.3.1. Cách thiết kế thí nghiệm.....	69
4.3.2. Mô hình phân tích.....	70
4.3.3. Cách phân tích	70
<i>4.4. THÍ NGHIỆM Ô VUÔNG LA TINH.....</i>	72
4.4.1. Ưu điểm và nhược điểm của mô hình.....	73
4.4.2. Cách thiết kế thí nghiệm	73
4.4.3. Mô hình phân tích.....	74
4.4.4. Cách phân tích	74
<i>4.5. BÀI TẬP</i>	77
<i>Chuong 5 - THIẾT KẾ THÍ NGHIỆM HAI YẾU TỐ</i>	79
<i>5.1. THÍ NGHIỆM HAI YẾU TỐ CHÉO NHAU (Cross hay Orthogonal)</i>	79
5.1.1. Ưu điểm và nhược điểm	79
5.1.2. Só đơn vị thí nghiệm cần thiết.....	80
5.1.3. Cách thiết kế thí nghiệm	80
5.1.4. Mô hình phân tích.....	81
5.1.5. Cách phân tích	81
<i>5.2. THÍ NGHIỆM HAI YẾU TỐ PHÂN CẤP (Hierachical hay Nested)</i>	84
5.2.1. Ưu và nhược điểm của mô hình.....	85
5.2.2. Cách thiết kế thí nghiệm	85
5.2.3. Mô hình	85
5.2.4. Cách phân tích	85
<i>5.3. THÍ NGHIỆM HAI YẾU TỐ CHIA Ô</i>	87
5.3.1. Ưu và nhược điểm của mô hình.....	88
5.3.2. Cách thiết kế thí nghiệm	88
5.3.3. Mô hình	88
5.3.4. Cách phân tích	89
<i>5.4. THÍ NGHIỆM HAI YẾU TỐ CHIA Ô HOÀN TOÀN NGĂU NHIÊN</i>	92
<i>5.5. BÀI TẬP</i>	93
<i>Chuong 6 - TUỔNG QUAN VÀ HỎI QUY TUYÊN TÍNH</i>	95
<i>6.1. SẮP XẾP SỐ LIỆU.....</i>	95

6.2. HỆ SỐ TƯƠNG QUAN.....	96
6.2.1. Tính hệ số tương quan	96
6.2.2. Tính chất của hệ số tương quan mẫu	96
6.3. HỒI QUY TUYẾN TÍNH.....	98
6.3.1. Đường trung bình của biến ngẫu nhiên Y theo X trong phân phối chuẩn 2 chiều	99
6.3.2. Đường thẳng gần đúng của Y theo X	100
6.4. KIỂM ĐỊNH ĐỐI VỚI HỆ SỐ TƯƠNG QUAN VÀ CÁC HỆ SỐ HỒI QUY.....	103
6.5. DỰ BÁO THEO HỒI QUY TUYẾN TÍNH.....	105
6.6. PHÂN TÍCH PHƯƠNG SAI VÀ HỒI QUY.....	106
<i>Chương 7 - KIỂM ĐỊNH MỘT PHÂN PHỐI VÀ BẢNG TƯƠNG LIÊN</i>	108
7.1. KIỂM ĐỊNH MỘT PHÂN PHỐI.....	108
7.2. BẢNG TƯƠNG LIÊN L x K.....	110
7.3. KIỂM ĐỊNH CHÍNH XÁC CỦA FISHER ĐỐI VỚI BẢNG TƯƠNG LIÊN 2 x 2	115
7.4. THÍ NGHIỆM NGHIÊN CỨU DỊCH TẾ HỌC THÚ Y.....	117
7.4.1. Thiết kế thí nghiệm nghiên cứu cắt ngang (cross sectional studies).....	117
7.4.2. Thiết kế thí nghiệm nghiên cứu bệnh chứng (case – control study)	120
7.4.3. Thiết kế thí nghiệm nghiên cứu thuần tập (cohort study)	122
7.5. BÀI TẬP	124
PHẦN B - THỰC HÀNH	126
Bài 1. TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU	126
1.1. GIỚI THIỆU PHẦN MỀM MINITAB	126
1.2. TÓM TẮT VÀ TRÌNH BÀY ĐỐI VỚI BIẾN ĐỊNH LUỢNG	127
1.3. TÓM TẮT VÀ TRÌNH BÀY ĐỐI VỚI BIẾN ĐỊNH TÍNH	131
Bài 2. ƯỚC LUỢNG, KIỂM ĐỊNH MỘT GIÁ TRỊ TRUNG BÌNH VÀ SO SÁNH HAI GIÁ TRỊ TRUNG BÌNH	136
2.1. ƯỚC LUỢNG VÀ KIỂM ĐỊNH MỘT GIÁ TRỊ TRUNG BÌNH.....	136
2.1.1. Kiểm định phân phối chuẩn	136
2.1.2. Kiểm định Z	137
2.1.3. Kiểm định T	138
2.2. SO SÁNH HAI GIÁ TRỊ TRUNG BÌNH	139
2.2.1. Kiểm định sự đồng nhất của phương sai	139
2.2.2. Kiểm định T	141
2.2.3. Kiểm định T cặp	143
Bài 3. SO SÁNH NHIỀU GIÁ TRỊ TRUNG BÌNH	146
3.1. THÍ NGHIỆM MỘT YẾU TỐ HOÀN TOÀN NGẪU NHIÊN	146
3.2. THÍ NGHIỆM MỘT YẾU TỐ KHỎI NGẪU NHIÊN ĐẦY ĐỦ	153
3.3. THÍ NGHIỆM Ô VUÔNG LA TÍNH	157

3.4. THÍ NGHIỆM HAI YẾU TỐ CHÉO NHAU (TRỰC GIAO).....	161
3.5. THÍ NGHIỆM HAI YẾU TỐ PHÂN CẤP (CHIA Ô).....	163
3.6. THÍ NGHIỆM HAI YẾU TỐ CHIA Ô (<i>Split-Plot</i>).....	166
3.7. PHÂN TÍCH HIỆP PHƯƠNG SAI.....	169
Bài 4. TƯƠNG QUAN VÀ HỒI QUY TUYẾN TÍNH.....	172
4.1. HỆ SỐ TƯƠNG QUAN.....	172
4.2. PHƯƠNG TRÌNH HỒI QUY TUYẾN TÍNH	173
Bài 5. BẢNG TƯƠNG LIÊN.....	177
Bài 6. ƯỚNG TÍNH DUNG LUỢNG MẪU	181
6.1. ƯỚC LUỢNG, KIỂM ĐỊNH MỘT GIÁ TRỊ TRUNG BÌNH.....	181
6.2. ƯỚC LUỢNG, KIỂM ĐỊNH MỘT TỶ LỆ	183
6.3. SO SÁNH HAI GIÁ TRỊ TRUNG BÌNH	184
6.4. SO SÁNH HAI TỶ LỆ	185
6.5. SO SÁNH NHIỀU GIÁ TRỊ TRUNG BÌNH.....	186
BÀI 7. BÀI TẬP	187
7.1. TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU.....	187
7.2. ƯỚC LUỢNG, KIỂM ĐỊNH MỘT GIÁ TRỊ TRUNG BÌNH.....	189
7.3. ƯỚC LUỢNG, KIỂM ĐỊNH XÁC SUẤT P	189
7.4. SO SÁNH HAI GIÁ TRỊ TRUNG BÌNH	190
7.5. SO SÁNH HAI XÁC SUẤT.....	191
7.6. PHÂN TÍCH PHƯƠNG SAI MỘT YẾU TỐ	191
7.7. PHÂN TÍCH PHƯƠNG SAI HAI YẾU TỐ	193
7.8. TƯƠNG QUAN VÀ HỒI QUY TUYẾN TÍNH	196
7.9. KIỂM ĐỊNH MỘT PHÂN PHỐI VÀ BẢNG TƯƠNG LIÊN.....	197
PHỤ LỤC 1 MỘT SỐ THUẬT NGỮ DÙNG TRONG GIÁO TRÌNH	199
PHỤ LỤC 2 BẢNG CÁC KÝ HIỆU TOÁN HỌC	201
PHỤ LỤC 3 HÀM PHÂN PHỐI CHUẨN	202
PHỤ LỤC 4 HÀM PHÂN PHỐI STUDENT (<i>t</i>).....	203
PHỤ LỤC 5 HÀM PHÂN PHỐI KHI BÌNH PHƯƠNG (χ^2).....	205
PHỤ LỤC 6 HÀM PHÂN PHỐI FISHER.....	207
PHỤ LỤC 7 GIÁ TRỊ 2½% PHÍA TRÊN CỦA PHÂN PHỐI FISHER (<i>F</i>).....	210
PHỤ LỤC 8 KHOÁNG Ý NGHĨA ĐỐI VỚI KIỂM ĐỊNH ĐA PHẠM VI DUNCAN.....	211
PHỤ LỤC 9 ĐƯỜNG CONG XÁC ĐỊNH DUNG LUỢNG MẪU TRONG MÔ HÌNH CÓ ĐỊNH.....	213
PHỤ LỤC 10 BẢNG SỐ NGẪU NHIÊN (TABLE OF RANDOM NUMBERS).....	217
PHỤ LỤC 11 SỐ ĐỒ THÍ NGHIỆM Ô VUÔNG LATINH MẪU	218
TÀI LIỆU THAM KHẢO	219

LỜI MỞ ĐẦU

(*Xuất bản lần đầu*)

Trong quá trình nghiên cứu, làm việc trong phòng thí nghiệm, trại thực nghiệm hoặc tại các cơ sở sản xuất, các bạn học viên thường gặp phải hai vấn đề lớn:

+ Thứ nhất, khảo sát, theo dõi các hiện tượng đã lựa chọn trước khi xây dựng đề tài nghiên cứu hoặc các hiện tượng mới xuất hiện nhưng có ảnh hưởng lớn đến đề tài. Khi khảo sát phải ghi chép kỹ mỷ, khoa học các dữ liệu thu được và bảo quản cẩn thận vì đó là các số liệu gốc. Sau đó, trừ các dữ liệu có tính chất mô tả phải phân chia các dữ liệu còn lại thành hai loại biến, biến định tính và biến định lượng. Tiếp theo là khảo sát các biến và nếu cần thì tiến hành các biến đổi thích hợp, sau đó căn cứ vào mục tiêu đặt ra để xử lý số liệu theo các công thức đã trình bày trong lý thuyết xác suất thống kê. Dựa vào kết quả xử lý để đưa ra các kết luận, thường gọi là các kết luận thống kê. Phản tiếp theo cũng là phần quan trọng nhất là căn cứ vào kết luận thống kê để đưa ra các đánh giá, các lý giải về mặt chuyên môn và đưa ra các đề xuất, các kiến nghị cụ thể.

+ Thứ hai, đó là việc thực hiện một thí nghiệm để giải quyết một mục tiêu cụ thể. Công việc này bao gồm nhiều bước như chọn vấn đề, chọn mục tiêu, chọn các biến cần theo dõi, chọn các biến cần điều khiển, các biến cần không chế. Tiếp theo là chọn các mức cụ thể đối với các biến cần điều khiển. Trên cơ sở vật chất hiện có như chuồng trại, vật tư, thời gian, các vật nuôi dùng để thí nghiệm... chọn một thí nghiệm cụ thể. Thí nghiệm này được thực hiện theo một sơ đồ phù hợp với mục tiêu và với cơ sở vật chất hiện có. Việc thí nghiệm theo sơ đồ đã chọn được gọi là bố trí thí nghiệm hay thiết kế thí nghiệm (Experimental design). Sau khi thí nghiệm, các dữ liệu được xử lý theo quy trình phù hợp với kiểu bố trí thí nghiệm đã chọn, tuyệt đối không được xử lý theo quy trình của kiểu bố trí thí nghiệm khác.

Như vậy dù khảo sát, theo dõi, hay bố trí thí nghiệm luôn luôn có sự đóng góp của ba ngành học: Kỹ thuật nông nghiệp, toán học và công nghệ thông tin. Có thể coi kỹ thuật nông nghiệp như đơn vị chủ quản, đơn vị đề xuất vấn đề cần khảo sát, cần nghiên cứu sau đó phối hợp với toán học mà chủ yếu là thống kê để đề ra mục tiêu cụ thể, lựa chọn các biến theo dõi, chọn các mô hình xử lý, giải thích các kết quả và đề xuất các vấn đề mới. Khi xử lý và trình bày kết quả thì không thể thiếu máy tính và các ứng dụng khác của công nghệ thông tin. Như vậy môn thiết kế thí nghiệm là môn học ra đời trên cơ sở ba ngành nói trên.

Cuốn giáo trình **Thiết kế thí nghiệm** này đi sâu vào các khía cạnh chuyên môn của các ngành học để trình bày cách chọn vấn đề nghiên cứu, các điểm cần chú ý khi bố trí thí nghiệm như kích thước, hướng của chuồng trại, cách chọn các vật thí nghiệm, cách tiến hành thí nghiệm, các hoá chất, các loại thuốc, thời gian cách ly, các chỉ tiêu cần đo, các dụng cụ và cách đo... Nhưng do có rất nhiều môn học, nên khó có thể đề

cập đầy đủ tất cả các khía cạnh, do đó nên để các môn học tự trình bày. Giáo trình này chỉ tập trung vào việc xử lý dữ liệu và các kiểu bố trí thí nghiệm thường dùng.

Giáo trình được viết theo đề cương môn học **Thiết kế thí nghiệm** của Khoa Chăn nuôi tương ứng với 3 đơn vị học trình (45 tiết). Đối với những lớp có thời lượng dạy 30 tiết có thể chỉ học một số phần.

Các chương 1, 2, 6, 7 chỉ trình bày cách đặt vấn đề, các công thức, các kết luận thống kê, còn việc tính toán cụ thể được thực hiện khi thực hành ở phòng máy tính. Trước mắt có thể chưa dạy hết chương 4 và chương 5, các phần để lại chắc chắn sẽ được dạy trong vài năm tới.

Đối tượng sử dụng giáo trình này là sinh viên hệ chính quy, hệ vừa học vừa làm các ngành Chăn nuôi, Chăn nuôi thú y, Thú y và Nuôi trồng thuỷ sản; đồng thời là tài liệu tham khảo cho các đối tượng là cán bộ nghiên cứu trong ngành chăn nuôi, thú y.

Để có thêm kiến thức bổ trợ cho môn học này, bạn đọc có thể tham khảo thêm một số tài liệu về toán xác suất thống kê, về tin học và các sách chuyên ngành của chăn nuôi thú y.

Để hoàn thành giáo trình này, nhóm tác giả xin chân thành cảm ơn Ban giám đốc Học viện Nông nghiệp Việt Nam đã giúp đỡ và tạo điều kiện thuận lợi để tái bản cuốn giáo trình này.

Nhóm tác giả cũng xin cảm ơn GS.TS. Đặng Vũ Bình, PGS.TS. Đinh Văn Chính, PGS.TS. Nguyễn Hải Quân, PGS.TS. Nguyễn Xuân Trạch, GS.TS. Pascal Leroy, PGS.TS. Fédéric Farnir, PGS.TS. Peter Thomson, GS.TS. Mick O'Neill đã cung cấp các tư liệu và có nhiều ý kiến đóng góp trong quá trình xây dựng nội dung môn học và viết giáo trình.

Hà Nội, tháng 02, năm 2007

T/M. BAN BIÊN SOẠN

NGUYỄN ĐÌNH HIỀN

LỜI TỰA

(*Cho lần tái bản thứ nhất*)

Giáo trình Thiết kế thí nghiệm được xuất bản lần đầu tiên vào năm 2007 nhằm trang bị cho sinh viên các ngành Chăn nuôi, Thú y và Nuôi trồng thuỷ sản những kiến thức cơ bản về thiết kế thí nghiệm, xử lý dữ liệu từ các mô hình thí nghiệm và rút ra kết luận từ kết quả xử lý dữ liệu. Trong suốt 10 năm qua, nhóm biên soạn giáo trình tác giả Nguyễn Đình Hiền (Chủ biên) và Đỗ Đức Lực (tham gia biên soạn) đã nhận được nhiều phản hồi tích cực từ sinh viên, học viên cao học, nghiên cứu sinh cũng như bạn đọc. Bên cạnh đó, công nghệ thông tin ngày càng được áp dụng rộng rãi trong thiết kế thí nghiệm đòi hỏi phải bổ sung nội dung và kỹ thuật xử lý.

Trong lần tái bản này, giáo trình tiếp tục được biên soạn bởi nhóm tác giả: Nguyễn Đình Hiền, Đỗ Đức Lực và Hà Xuân Bộ. Cấu trúc giáo trình được chia thành 2 phần - Phần A: Lý thuyết và Phần B: Thực hành. Tất cả các chương của lần xuất bản đầu tiên (từ Chương 1 đến 7) được đưa vào Phần A - Lý thuyết. Nội dung các chương trong phần A đều được bổ sung và cập nhật thêm thông tin. Phần B - Thực hành là phần được biên soạn mới bao gồm 6 bài tập thực hành. Các bài thực hành đều có phần giới thiệu các bước tính, cách thực hiện trong phần mềm Minitab 16, kết quả tóm tắt và các kết luận.

Giáo trình Thiết kế thí nghiệm (tái bản lần thứ nhất) được viết theo đề cương môn Thiết kế thí nghiệm của Khoa Chăn nuôi tương ứng với 2 tín chỉ (30 tiết). Đối tượng sử dụng giáo trình là sinh viên các ngành Chăn nuôi, Chăn nuôi thú y, Thú y và Nuôi trồng thuỷ sản; đồng thời là tài liệu tham khảo cho các đối tượng là cán bộ nghiên cứu trong lĩnh vực sinh học.

Nhóm tác giả xin chân thành cảm ơn Ban giám đốc Học viện Nông nghiệp Việt Nam đã giúp đỡ và tạo điều kiện thuận lợi để tái bản cuốn giáo trình; GS.TS. Đặng Vũ Bình, GS.TS. Phạm Tiến Dũng, PGS.TS. Vũ Đình Tôn, PGS.TS. Nguyễn Văn Đức đã có nhiều ý kiến đóng góp cho nội dung của lần tái bản này.

Mặc dù có rất nhiều cố gắng trong quá trình biên soạn, xong không thể tránh được những thiếu sót, nhóm tác giả rất mong sự góp ý của bạn đọc để lần tái bản sau được hoàn thiện hơn.

Hà Nội, tháng 02, năm 2017
T/M. BAN BIÊN SOẠN

ĐỖ ĐỨC LỰC

PHẦN A - LÝ THUYẾT

Chương 1

MỘT SỐ KHÁI NIỆM TRONG XÁC SUẤT THỐNG KÊ MÔ TẢ

Mục đích của chương này là hệ thống lại một số khái niệm về xác suất, các phân phối thường được sử dụng trong sinh học nói chung và trong chăn nuôi, thú y nói riêng; phân loại các biến sinh học và đồng thời khái quát hoá và nêu ý nghĩa của một số tham số thống kê mô tả cơ bản.

1.1. TÓM TẮT VỀ XÁC SUẤT VÀ BIẾN NGẪU NHIÊN

1.1.1. Xác suất cơ bản

Số chỉnh hợp chập k trong n vật: $A_n^k = n(n-1)(n-2)...(n-k+1) = \frac{n!}{(n-k)!}$

Số tổ hợp chập k của n vật: $C_n^k = \frac{A_n^k}{k!} = \frac{n!}{k!(n-k)!}$

Số hoán vị của k vật: $A_k^k = k!$

Số chỉnh hợp lặp chập k của n vật: $\tilde{A}_n^k = n^k$

Nhị thức Niu-ton: $(a+b)^n = \sum_{k=0}^n C_n^k a^{n-k} b^k$

Quy tắc cộng tổng quát: $p(A \cup B) = p(A) + p(B) - p(A \cap B)$.

Quy tắc cộng đơn giản: $p(A \cup B) = p(A) + p(B)$ nếu $A \cap B = \emptyset$.

Quy tắc nhân tổng quát: $p(A \cap B) = p(A) \cdot p(B/A) = p(B) \cdot p(A/B)$.

Quy tắc nhân đơn giản: $p(A \cap B) = p(A) \cdot p(B)$ nếu A, B độc lập.

1.1.2. Hệ sự kiện đầy đủ

Hệ sự kiện đầy đủ hay hệ sự kiện toàn phần nếu:

$$\bigcup_{i=1}^n A_i = \Omega \quad \text{và} \quad A_i \cap A_j = \emptyset \text{ với } i \neq j$$

Công thức xác suất toàn phần: $p(B) = \sum_{k=1}^n p(A_i) \cdot p(B / A_i)$

Công thức Bayes: $p(A_i / B) = \frac{p(A_i) \cdot p(B / A_i)}{p(B)}$

1.1.3. Biến ngẫu nhiên, bảng phân phối, hàm phân phối

Kỳ vọng toán học: $MX = \sum_{i=1}^n x_i p_i$

Phương sai: $DX = \sum_{i=1}^n (x_i - MX)^2 p_i$ hay: $DX = \sum_{i=1}^n x_i^2 p_i - (MX)^2$

Bảng phân phối của biến ngẫu nhiên rời rạc:

X	x_1	x_2	...	x_n	Tổng
p_i	p_1	p_2	...	p_n	1

Hàm phân phối

$$F(x) = p(X < x) = \begin{cases} 0 & x \leq x_1 \\ p_1 & x_1 \leq x < x_2 \\ p_1 + p_2 & x_2 \leq x < x_3 \\ p_1 + p_2 + p_3 & x_3 \leq x < x_4 \\ \dots & \dots \\ 1 & x_n < x \end{cases}$$

1.1.4. Một số phân phối thường gặp

Phân phối Bécnuli

X	0	1	Kỳ vọng $MX = \mu = p$	Phương sai $DX = pq$
p_i	p	q		

Phân phối nhị thức B (n, p)

X	0	1	...	K	...	n	$MX = np$	$DX = npq$
p_i	q^n	$C_1^n p q^{n-1}$...	$C_k^n p^k q^{n-k}$...	p^n		
							ModX là số nguyên	
							$np-q \leq ModX \leq np+p$	

Phân phối siêu bội

Nếu trong N bi có M bi trắng, rút n bi, X là số bi trắng

$$X = 0, n \text{ với } p_k = p(X = k) = \frac{C_M^k C_{N-M}^{n-k}}{C_N^n}$$

$$MX = \frac{nM}{N} \quad DX = n \frac{M}{N} \frac{N-M}{N} \frac{N-n}{N-1}$$

Phân phối hình học

$X = \overline{1, \infty}$ với $p_k = p(X = k) = pq^{k-1}$ (p là xác suất thành công, $q = 1 - p$).

$$MX = \frac{1}{p} \quad DX = \frac{q}{p^2}$$

Phân phối Poisson (Poisson)

$$X = \overline{0, \infty} \text{ với xác suất } p_k = p(X = k) = \frac{e^{-\lambda}}{k!} \lambda^k$$

$$MX = DX = \lambda$$

Phân phối chuẩn N(μ, σ^2)

$$\text{Hàm mật độ xác suất: } f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$p(a < X < b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Với $\Phi(z)$ là hàm phân phối của biến chuẩn tắc.

Phân phối chuẩn tắc N(0, 1)

$$\text{Mật độ xác suất: } \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

$$\text{Hàm phân phối: } \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{x^2}{2}} dx$$

Tính gần đúng phân phối nhị thức bằng phân phối chuẩn khi n lớn:

$$p(k \leq X \leq l) \approx \Phi\left(\frac{l-np}{\sqrt{npq}}\right) - \Phi\left(\frac{k-np}{\sqrt{npq}}\right)$$

$$p(X = k) \approx \frac{1}{\sqrt{npq}} \varphi\left(\frac{k-np}{\sqrt{npq}}\right)$$

Dung lượng mẫu cần thiết để trung bình cộng khác μ không quá ε (độ chính xác) khi có phân phối chuẩn $N(\mu, \sigma^2)$ và mức tin cậy $P = 1 - \alpha$.

$$n \geq \frac{z^2 \sigma^2}{\varepsilon^2} \quad z \text{ là giá trị sao cho } \Phi(z) = 1 - \alpha/2.$$

Dung lượng mẫu cần thiết để tần suất khác xác suất không quá ε trong phân phối nhị thức và mức tin cậy $P = 1 - \alpha$.

$$n \geq \frac{z^2}{4\varepsilon^2} \quad z \text{ là giá trị sao cho } \Phi(z) = 1 - \alpha/2.$$

1.2. BIẾN SINH HỌC

Trong quá trình thực hiện thí nghiệm, chúng ta tiến hành thu thập dữ liệu để sau đó xử lý và đưa ra các kết luận. Các dữ liệu có thể là các giá trị bằng số hoặc bằng chữ đặc trưng cho một cá thể hoặc một nhóm và thay đổi từ cá thể này qua cá thể khác. Các dữ liệu như vậy được gọi là các biến, hay còn được gọi là các biến ngẫu nhiên vì các dữ liệu thu được là kết quả của việc chọn một cách ngẫu nhiên cá thể hay nhóm cá thể trong tổng thể (trong trường hợp với thí nghiệm thực nghiệm) hoặc kết quả quan sát các đặc trưng của các cá thể (trong trường hợp với thí nghiệm quan sát).

1.2.1. Khái niệm về biến sinh học

Đối tượng nghiên cứu trong chăn nuôi là các vật sống, vì vậy các biến như đã nêu trên gọi chung là các biến sinh học. Có thể phân loại các biến sinh học như sau:

Biến định tính (qualitative)

Biến định danh (nominal).

Biến thứ hạng (ranked).

Biến định lượng (quantitative)

Biến liên tục (continuous).

Biến rời rạc (discontinuous).

Biến định tính bao gồm các **biến có hai trạng thái (binary)**: Thí dụ như giới tính (cái hay đực), vật nuôi sau khi được điều trị (sống hay chết, khỏi bệnh hay không khỏi bệnh), tình trạng nhiễm bệnh (có, không), mang thai (có, không)... Tổng quát hơn có các **biến có nhiều trạng thái**, từ đó chia ra các lớp (loại) thí dụ màu lông của các giống lợn (trắng, đen, loang, hung,...) các kiểu gen (đồng hợp tử trội, dị hợp tử, đồng hợp tử lặn...); giống bò (bò Vàng, Jersey, Holstein...). Các biến như thế được gọi là **biến định danh (nominal)** hay biến có thang đo định danh, cũng còn gọi là biến thuộc tính. Trong các biến có nhiều trạng thái, có một số biến có thể sắp thứ tự theo một cách nào đó, ví dụ mức độ mắc bệnh của vật nuôi. Thường dùng số thứ tự để xếp hạng các biến này, thí dụ xếp động vật theo mức độ mắc bệnh (--, -, -+, +, ++), thứ hạng của vật nuôi (đối với bò từ 1-5, 1-rất gầy,..., 5-rất béo). Các biến này gọi là **biến thứ hạng (ranked)** hay biến có thang đo thứ bậc.

Biến định lượng là biến phải dùng một gốc đo, một đơn vị đo để xác định giá trị (số đo) của biến. Biến định lượng bao gồm: **biến rời rạc**, thí dụ số trứng nở khi áp 12 quả ($X = 0, 1, \dots, 12$), số lợn con sinh ra trong một lứa đẻ, số té bào hòng cầu đếm trên đĩa của kính hiển vi và **biến liên tục**, thí dụ khối lượng gà 45 ngày tuổi, sản lượng sữa bò trong một chu kỳ, tăng khối lượng trên ngày của động vật, nồng độ canxi trong máu. Sau khi chọn đơn vị đo thì giá trị cụ thể của X là một số nằm trong một khoảng $[a, b]$ nào đó.

Đối với các biến định lượng có thể phân biệt: 1) **biến khoảng (interval)** hay biến có thang đo khoảng, biến này chỉ chú ý đến mức chênh lệch giữa hai giá trị (giá trị 0

mang tính quy ước, tỷ số hai giá trị không có ý nghĩa). Thí dụ đối với nhiệt độ chỉ nói nhiệt độ tăng thêm hay giảm đi mấy °C (thí dụ cơ thể đang từ 36,5°C tăng lên 38°C là biểu hiện bắt đầu sốt cao) chứ không nói vật thể có nhiệt độ 60°C nóng gấp đôi vật thể có nhiệt độ 30°C. Hướng gió có quy ước 0° là hướng Bắc, 45° là hướng Đông Bắc, 90° là hướng Đông, 180° là hướng Nam..., không thể nói hướng gió Đông gấp đôi hướng gió Đông Bắc; 2) **biến tỷ số** (ratio) hay biến có thang đo tỷ lệ, đối với biến này giá trị 0, mức chênh lệch giữa hai giá trị và tỷ số hai giá trị đều có ý nghĩa. Thí dụ khối lượng bắt đầu thí nghiệm của lợn là 25 kg, khối lượng kết thúc là 90 kg, vậy khối lượng kết thúc thí nghiệm nặng gấp 3,6 lần.

1.2.2. Tổng thể và mẫu

Một đám đông gồm rất nhiều cá thể chung nhau nguồn gốc, hoặc chung nhau nơi sinh sống, hoặc chung nhau nguồn lợi... được gọi là một tổng thể. Lấy từng cá thể ra đo một biến sinh học X, được một biến ngẫu nhiên, có thể định tính hoặc định lượng.

Muốn hiểu biết đầy đủ về biến X phải khảo sát toàn bộ tổng thể, nhưng vì nhiều lý do không thể làm được. Có thể do không đủ tiền tài, vật lực, thời gian..., nên không thể khảo sát toàn bộ, cũng có thể do phải huỷ hoại cá thể khi khảo sát nên không thể khảo sát toàn bộ, cũng có khi cân nhắc giữa mức chính xác thu được và chi phí khảo sát thấy không cần thiết phải khảo sát hết.

Như vậy là có nhiều lý do khiến người ta chỉ khảo sát một bộ phận gọi là mẫu (sample) sau đó xử lý các dữ liệu (số liệu) rồi đưa ra các kết luận chung cho tổng thể. Các kết luận này được gọi là “kết luận thống kê”.

Để các kết luận đưa ra đúng cho tổng thể thì mẫu phải “ phản ánh ” được tổng thể (còn nói là mẫu phải “đại diện”, phải “điển hình” cho tổng thể...), không được thiên về phía “tốt” hay thiên về phía “xấu”.

1.2.3. Sơ lược về cách chọn mẫu

Tùy theo đặc thù của ngành nghề người ta đưa ra rất nhiều cách chọn mẫu khác nhau, thí dụ chọn ruộng để gặt nhằm đánh giá năng suất, chọn các sản phẩm của một máy để đánh giá chất lượng, chọn các hộ để điều tra dân số hoặc điều tra xã hội học, chọn một số sản phẩm ra kiểm tra trước khi xuất khẩu một lô hàng... Cách chọn mẫu phải hợp lý về mặt chuyên môn, phải dễ cho người thực hiện và phải đảm bảo yêu cầu chung về mặt xác suất thống kê là “ngẫu nhiên” không thiên lệch.

Thuần tuý về thống kê cũng có nhiều cách chọn mẫu:

Chọn mẫu hoàn toàn ngẫu nhiên (rút thăm, dùng bảng số ngẫu nhiên để lựa chọn...).

Chia tổng thể thành các lớp đồng đều hơn theo một tiêu chuẩn nào đó thí dụ chia toàn quốc thành các vùng (vùng cao, trung du, đồng bằng), chia theo tầng lớp xã hội, chia theo thu nhập, theo ngành nghề, chia sản phẩm thành các lô hàng theo nguồn vật liệu, theo ngày sản xuất... Sau khi có các lớp thì căn cứ vào mức đồng đều trong từng lớp mà chọn số lượng cá thể (dung lượng mẫu) đại diện cho lớp.

Có thể chia tổng thể thành các lớp, sau đó chọn một số lớp gọi là mẫu cấp một. Mỗi lớp trong mẫu cấp một lại được chia thành nhiều lớp nhỏ hơn, đều hơn. Chọn một số trong đó gọi là mẫu cấp hai. Có thể khảo sát hết các cá thể trong mẫu cấp hai hoặc chỉ khảo sát một bộ phận.

Không đi sâu vào việc chọn mẫu chúng ta chỉ nhấn mạnh mẫu phải ngẫu nhiên, phải chọn mẫu một cách khách quan không được chọn mẫu theo chủ quan người chọn.

1.2.4. Các tham số thống kê của mẫu

Gọi số cá thể được chọn vào mẫu là kích thước (cỡ, dung lượng) mẫu n. Gọi các số liệu đo được trên các cá thể của mẫu là x_1, x_2, \dots, x_n , nếu có nhiều số liệu bằng nhau thì có thể ghi lại dưới dạng có tần số (số lần gấp).

Giá trị x_i	x_1	x_2	...	x_k	$\sum_{i=1}^k m_i = n$
tần số m_i	m_1	m_2	...	m_k	

Các tham số (số đặc trưng) của mẫu, hay còn gọi là các thống kê, được chia thành hai nhóm: 1) các tham số về vị trí và 2) các tham số về độ phân tán của số liệu.

Các **tham số về vị trí** thường gồm: a) trung bình, b) trung vị, c) mode. Các **tham số về độ phân tán** gồm: a) phương sai, b) độ lệch chuẩn, c) sai số chuẩn, d) khoảng biến động và e) hệ số biến động.

TRUNG BÌNH

Trung bình cộng ký hiệu là \bar{x}

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad \text{hay} \quad \bar{x} = \frac{\sum_{i=1}^k x_i m_i}{\sum_{i=1}^k m_i} \quad \text{khi có tần số hoặc tần suất } (m_i).$$

Ví dụ 1.1: Khối lượng (g) của 16 chuột cái tại thời điểm cai sữa như sau:

$$\begin{array}{cccccccc} 54,1 & 49,8 & 24,0 & 46,0 & 44,1 & 34,0 & 52,6 & 54,4 \\ 56,1 & 52,0 & 51,9 & 54,0 & 58,0 & 39,0 & 32,7 & 58,5 \end{array}$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{54,1 + 49,8 + \dots + 58,5}{16} = \frac{761,2}{16} = 47,58 \text{ g}$$

Ví dụ 1.2: Để tính khối lượng trung bình của 4547 lợn Piétrain \times (Yorkshire \times Landrace) nuôi vỗ béo đến 210 ngày tuổi (kg), căn cứ theo khối lượng của lợn để chia thành các nhóm từ thấp đến cao, xác định khối lượng trung bình của từng nhóm và tần số xuất hiện của mỗi nhóm.

Nhóm khối lượng (kg)	Khối lượng trung bình nhóm (kg)	Tần số	Tần suất	Tần suất tích luỹ
60,73 - 66,99	63,86	11	0,24	0,24
67,00 - 74,99	71,00	31	0,68	0,92
75,00 - 82,99	79,00	80	1,76	2,68
83,00 - 90,99	87,00	218	4,79	7,48
91,00 - 98,99	95,00	484	10,64	18,12
99,00 - 106,99	103,00	951	20,91	39,04
107,00 - 114,99	111,00	1083	23,82	62,85
115,00 - 122,99	119,00	907	19,95	82,8
123,00 - 130,99	127,00	512	11,26	94,06
131,00 - 138,99	135,00	203	4,46	98,53
139,00 - 146,99	143,00	55	1,21	99,74
147,00 - 156,10	151,55	12	0,26	100,00

$$\bar{x} = \frac{\sum_{i=1}^k x_i m_i}{\sum_{i=1}^k m_i} = \frac{63,86 \times 11 + 71,00 \times 31 + \dots + 151,55 \times 12}{11 + 31 + \dots + 12} = 110,48 \text{ kg}$$

Giá trị trung bình cộng có bất lợi là bị các giá trị ngoại lai làm ảnh hưởng. Giá trị ngoại lai là giá trị có xu hướng không thích hợp với toàn bộ số liệu thu thập được, thường là các giá trị quá lớn hoặc quá bé so với bình thường. Nếu giá trị ngoại lai quá lớn sẽ làm cho giá trị trung bình có xu hướng tăng quá mức hoặc ngược lại.

Trung bình nhân ký hiệu là G

$$G = \sqrt[n]{x_1 x_2 \dots x_n} \quad G = \sqrt[n]{x_1^{m_1} x_2^{m_2} \dots x_k^{m_k}}$$

Ví dụ 1.3: Bệnh dại đã tăng 10% trong năm thứ nhất, 11% trong năm thứ 2 và 15% trong năm thứ 3. Mức tăng trưởng trung bình của bệnh là bao nhiêu phần trăm?

Không thể tính tăng trưởng trung bình như sau $(10 + 11 + 15)/3 = 12$ mà phải tính mức tăng trưởng trung bình là $G = \sqrt[3]{x_1 x_2 \dots x_n} = \sqrt[3]{1,1 \times 1,11 \times 1,15} = 1,11979$. Nghĩa là mức tăng trưởng trung bình là 0,11979 hay tương đương mức 11,979 %.

Ví dụ 1.4: Một loại mèo bào sinh trưởng sau 3 tháng sẽ tăng gấp đôi khối lượng. Mức tăng trưởng trung bình mỗi tháng là bao nhiêu?

Mức tăng trưởng trung bình mỗi tháng là: $G = \sqrt[3]{2} = 1,26$; nghĩa là 26% mỗi tháng. Có thể minh họa sự tăng trưởng qua 3 tháng như sau:

$$1 \times 1,26 = 1,26.$$

$$1,26 \times 1,26 = 1,5876.$$

$$1,5876 \times 1,26 = 2,00037.$$

Trung bình điều hòa ký hiệu là H.

$$H = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} \quad \text{hoặc} \quad H = \frac{n}{\sum_i \frac{m_i}{x_i}}$$

Ví dụ 1.5: Ba lò mổ mỗi lò mổ 1000 con; lò mổ thứ nhất có năng suất giết mổ 10 con/giờ, lò mổ thứ hai 15 con/giờ và lò mổ thứ ba 30 con/giờ. Trung bình một giờ giết mổ được bao nhiêu con?

Trung bình sẽ không phải là $(10 + 15 + 30)/3 = 55/3$. Đây là trung bình cộng, chính bằng trung bình mỗi giờ nếu cả 3 lò mổ song song với nhau.

$$\text{Giá trị trung bình phải là } H = \frac{n}{\sum_i \frac{1}{x_i}} = \frac{3}{\frac{1}{10} + \frac{1}{15} + \frac{1}{30}} = 15 \text{ con/giờ.}$$

Điều này có thể minh họa như sau: Để giết mổ được 90 con lò thứ nhất phải thực hiện trong 9 giờ, lò thứ hai trong 6 giờ và lò thứ 3 trong 3 giờ; nghĩa là 270 con lợn được giết mổ trong 18 giờ; tức là trung bình 15 con/giờ. Chú ý rằng số lợn giết mổ được cố định khi bắt đầu.

TRUNG VỊ ký hiệu Med

Nếu sắp xếp các giá trị từ nhỏ đến lớn thì giá trị ở vị trí chính giữa được gọi là trung vị (Med). Nói một cách lý thuyết thì Med là giá trị có 50% số giá trị nhỏ hơn và 50% số giá trị lớn hơn. Để tính nhanh giá trị trung vị ta có thể tiến hành các bước sau:

1. Sắp xếp các giá trị theo trình tự tăng dần.
2. Đánh số thứ tự cho các dữ liệu.
3. Tìm trung vị ở vị trí có số thứ tự $(n + 1)/2$.

Nếu n là số lẻ và các giá trị đều khác nhau thì có một giá trị chính ở giữa.

Ví dụ 1.6: Nồng độ vitamin E ($\mu\text{mol/l}$) của 11 bê cái có dấu hiệu lâm sàng của phát triển cơ không bình thường được trình bày như sau:

4,2 3,3 7,0 6,9 5,1 3,4 2,5 8,6 3,5 2,9 4,9

Sau khi sắp xếp theo thứ tự tăng dần sẽ có:

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

Như vậy vị trí trung vị sẽ là $(n + 1)/2 = (11 + 1)/2 = 6$, do 6 là vị trí của trung vị nên giá trị của trung vị sẽ là 4,2.

Nếu n là số chẵn và các giá trị đều khác nhau thì có 2 số đứng giữa, cả hai đều được gọi là trung vị. Khoảng giữa 2 số đứng giữa được gọi là khoảng trung vị. Nếu được phép dùng số thập phân thì lấy điểm giữa của khoảng làm trung vị Med.

Xét ví dụ 1.1: Khối lượng (g) của 16 chuột cái tại thời điểm cai sữa như sau:

54,1 49,8 24,0 46,0 44,1 34,0 52,6 54,4

56,1 52,0 51,9 54,0 58,0 39,0 32,7 58,5

Vị trí của trung vị sẽ là $(16 + 1)/2 = 8,5$; khoảng trung vị sẽ nằm ở vị trí số 8 và số 9, tức là từ 49,8 – 51,9. Như vậy giá trị của trung vị Med = $(49,8 + 51,9)/2 = 50,9$.

Nếu các số liệu chia thành lớp có tần số thì phải chọn lớp trung vị sau đó nội suy để tính gần đúng trung vị.

Ngoài trung vị còn có các phân vị, trong đó hay dùng nhất là tứ phân vị dưới Q_1 mà chúng ta có thể định nghĩa một cách lý thuyết là giá trị có 25% số giá trị nhỏ hơn, tứ phân vị trên Q_2 là giá trị có 25% số giá trị lớn hơn.

MODE ký hiệu Mod

Mode là giá trị có tần suất cao nhất. Thông thường Mode có giá trị khác với giá trị trung bình cộng và trung vị. Ba giá trị này sẽ bằng nhau khi số liệu có phân bố chuẩn. Nhóm Mode hay lớp Mode là nhóm hoặc lớp mà một số lớn các quan sát rơi vào đó. Thông qua tổ chức đồ ta có thể xác định được giá trị của lớp này.

Xét trường hợp ví dụ 2, nhóm Mod được đại diện bằng các giá trị từ 107 đến 115 kg. Từ 4547 lợn quan sát có 1083 con nằm trong khoảng từ 107 đến 115 kg; đây là tần suất cao nhất. Cũng theo ví dụ 1 ta thấy Mod có giá trị khoảng 111 kg.

P (kg)	60,7 66,9	67,0 74,9	75,0 82,9	83,0 90,9	91,0 98,9	99,0 106,9	107,0 114,9	115,0 122,9	123,0 130,9	131,0 138,9	139,0 146,9	147,0 156,1
n	11	31	80	218	484	951	1.083	907	512	203	55	12

Trường hợp có nhiều giá trị có tần số lớn bằng nhau và lớn hơn các tần số khác thì không xác định được Mod.

Trường hợp số liệu chia lớp thì tìm lớp có tần số lớn nhất sau đó dùng cách nội suy để tính gần đúng Mod.

PHƯƠNG SAI MẪU ký hiệu s^2

Phương sai mẫu chưa hiệu chỉnh s_p^2 tính theo công thức:

$$s_p^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \text{ hay } s_p^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 m_i}{n} \text{ khi có tần suất } (m_i).$$

Phương sai mẫu được dùng trong tài liệu này là **phương sai đã hiệu chỉnh**, gọi tắt là phương sai mẫu s^2 :

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \text{ hay } s^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 m_i}{n-1} \text{ khi có tần suất } (m_i).$$

Đối với máy tính bỏ túi, có thể tính phương sai theo công thức sau:

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

Khi có phương sai mẫu chưa hiệu chỉnh s_p^2 có thể tính s^2 theo công thức:

$$s^2 = \frac{n}{(n-1)} s_p^2$$

Xét ví dụ 1.1, khối lượng của 16 chuột cái tại thời điểm cai sữa; giá trị trung bình đã tính là 47,58 g. Như vậy phương sai mẫu hiệu chỉnh sẽ là:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{(54,1-47,58)^2 + (49,8-47,58)^2 + \dots + (58,5-47,58)^2}{16-1} = 103,27 \text{ g}^2$$

ĐỘ LỆCH CHUẨN ký hiệu là **SD**

Căn bậc hai của s^2 gọi là độ lệch chuẩn: $SD = \sqrt{s^2}$.

Xét ví dụ 1, khối lượng của 16 chuột cái tại thời điểm cai sữa. Các số liệu này đã được sử dụng để tính giá trị trung bình (47,58 g) và phương sai ($103,27 \text{ g}^2$) như đã nêu trên. Như vậy độ lệch chuẩn sẽ là: $SD = \sqrt{s^2} = \sqrt{103,27} = 10,16 \text{ g}$.

HỆ SỐ BIẾN ĐỘNG ký hiệu là **CV (%)**

Hệ số biến động được tính theo công thức:

$$CV = \frac{SD}{\bar{x}} \times 100$$

Xét ví dụ 1.1, khối lượng của 16 chuột cái tại thời điểm cai sữa. Vì giá trị trung bình (47,58 g) và độ lệch chuẩn (10,16 g) nên phương sai mẫu hiệu chỉnh sẽ là:

$$CV = \frac{SD}{\bar{x}} \times 100 = \frac{10,16}{47,58} \times 100 = 21,36 \%$$

KHOẢNG BIẾN THIỀN (phạm vi chứa số liệu Range)

Gọi X_{\max} là giá trị lớn nhất, Gọi X_{\min} là giá trị nhỏ nhất, ta có khoảng biến thiên:

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

Với ví dụ 1.1, khối lượng của 16 chuột tại thời điểm cai sữa.

Có $R = X_{\max} - X_{\min} = 58,5 - 24,0 = 34,5 \text{ g}$.

SAI SỐ CHUẨN (sai số của trung bình cộng) ký hiệu là **SE**

$$SE = \frac{SD}{\sqrt{n}}$$

Xét ví dụ 1.1, khối lượng của 16 chuột cái tại thời điểm cai sữa. Vì đã có độ lệch chuẩn (10,16 g) nên sai số chuẩn sẽ là:

$$SE = \frac{SD}{\sqrt{n}} = \frac{10,16}{\sqrt{16}} = 2,54 \text{ g}$$

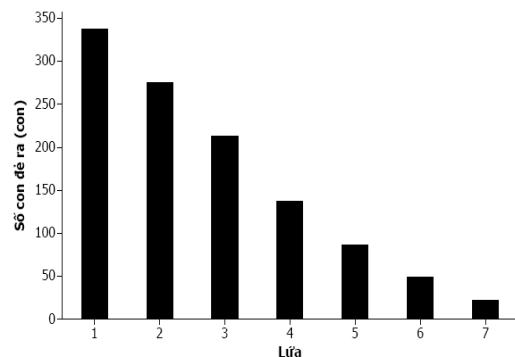
Ngoài các tham số trên, trong thống kê còn dùng độ lệch (độ bát đối xứng), độ nhọn. Hai tham số này được dùng khi xem xét có nên chuyển đổi số liệu không phân phôi chuẩn thành số liệu phân phôi chuẩn hay không.

1.2.5. Biểu diễn số liệu bằng đồ thị

Đồ thị là tóm tắt số liệu ở các dạng hình ảnh khác nhau và cho phép dễ dàng phát hiện những điểm đặc biệt hơn so với tóm tắt bằng số. Đồ thị đặc biệt hiệu quả khi ta muốn biết được các thông tin về số liệu một cách nhanh chóng.

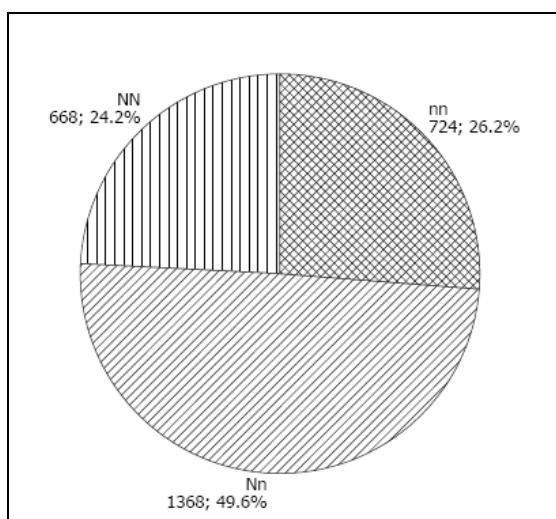
Có nhiều cách biểu diễn số liệu bằng đồ thị: Đồ thị tàn số, đồ thị hình thanh, đồ thị đa giác, chữ nhật (tổ chức đồ).

Đối với biến định tính hoặc biến rời rạc có thể biểu diễn số liệu bằng đồ thị thanh hoặc đồ thị bánh hình tròn.



Lứa	Số con đẻ ra (con)	Tần suất (%)	Tần suất tích luỹ (%)
1	337	30,12	30,12
2	275	24,58	54,69
3	213	19,03	73,73
4	137	12,24	85,97
5	86	7,69	93,66
6	49	4,38	98,03
7	22	1,97	100,00

Biểu đồ thanh biểu diễn số lợn sơ sinh qua 7 lứa ($n = 1.119$)



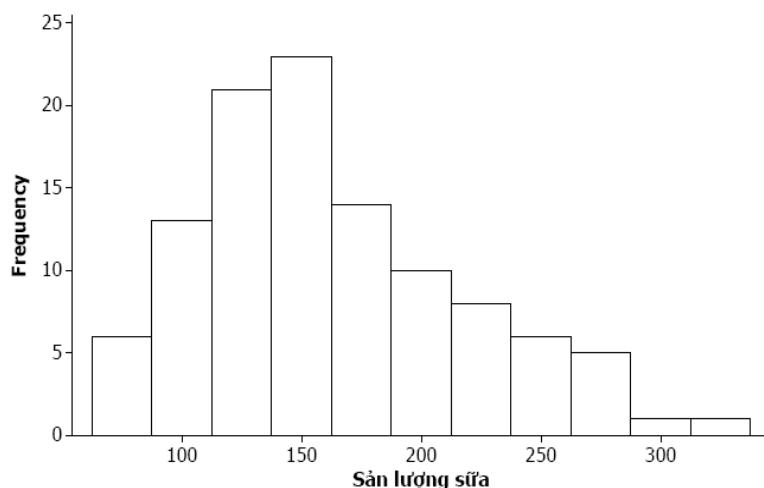
Biểu đồ bánh biểu hiện tần số kiểu gen Halothane của lợn sơ sinh Pietrain ($n = 2.760$)

Đối với biến định lượng có thể sử dụng đồ thị đa giác, đồ thị hộp hay tổ chức đồ để thể hiện.

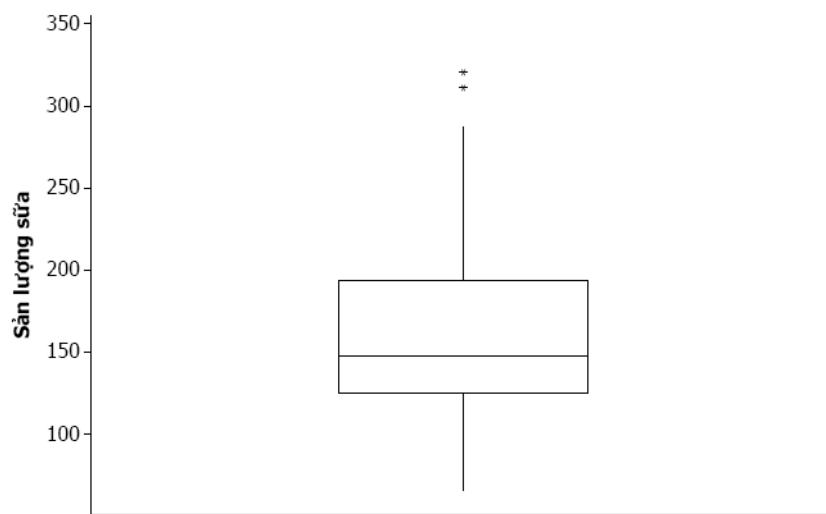
Ví dụ: Sản lượng sữa (kg) của 108 dê Bách Thảo trong một chu kỳ tiết sữa ghi lại như sau:

147,9	125,4	104,1	164,4	193,8	188,4	222,4	287,3	158,1
132,0	224,0	163,8	153,3	100,6	219,5	130,4	114,0	182,1
156,9	66,3	140,6	128,3	193,2	127,1	125,0	129,9	89,7
254,4	240,3	148,2	190,0	176,7	73,8	147,9	222,7	191,6

174,3	211,0	214,5	169,5	115,0	193,6	168,0	196,9	87,3
144,4	138,4	171,6	100,0	125,6	283,9	116,5	71,0	220,1
139,7	140,7	270,5	176,8	155,0	163,5	161,6	152,0	141,0
180,0	202,6	112,8	153,5	77,9	140,7	136,4	272,3	90,0
197,5	96,8	96,8	137,8	150,4	101,5	132,0	146,3	242,3
311,0	118,7	146,6	184,2	243,8	260,7	279,2	135,9	109,5
96,8	119,0	109,3	143,8	102,9	229,3	244,2	137,1	143,6
130,6	72,0	105,1	135,0	320,4	182,2	217,8	172,5	136,4



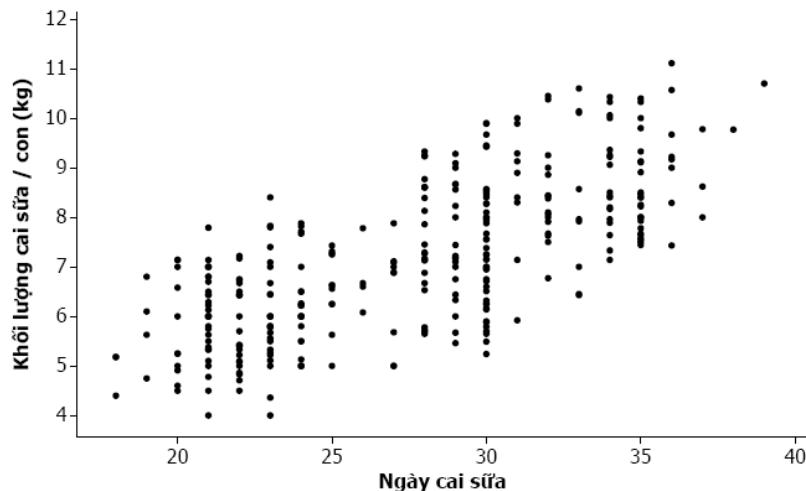
Tổ chức đồ: Phân bố tần suất sản lượng sữa dê Bách Thảo trong chu kỳ tiết sữa



Đồ thị hộp: Phân bố tần suất sản lượng sữa dê Bách Thảo trong chu kỳ tiết sữa

Tóm tắt và biểu diễn dữ liệu của các tính trạng số lượng (dữ liệu 2 chiều).

Đồ thị phân tán được sử dụng một cách rất hữu hiệu khi ta quan tâm đến mối liên hệ giữa 2 biến liên tục. Đồ thị được xây dựng khi ta vẽ n các điểm trên hệ toạ độ, các điểm này có toạ độ là x_i, y_i . Vấn đề này sẽ được đề cập cụ thể trong chương 6.



Đồ thị phân tán thể hiện mối quan hệ giữa thời gian cai sữa (ngày) và khối lượng sơ sinh sinh/con (kg) của lợn Landrace n = 321

Ký hiệu các tham số thống kê của mẫu và quần thể

Tên tham số	Tiếng Anh	Quần thể	Mẫu
Giá trị trung bình	Mean	μ	\bar{x} , Mean
Phương sai	Variance	σ^2	S^2
Độ lệch chuẩn	Standard Deviation	σ	SD
Sai số chuẩn	Standard Error	-	SE, SEM
Trung vị	Median	-	Med
Mode	Mode	-	Mod
Hệ số biến động	Coefficient of Variation	-	CV
Tỷ lệ, xác suất	Proportion, Probability	π, P	p

1.3. BÀI TẬP

1.3.1

Xác suất mắc một bệnh là $P = 0,35$ ($0,35$ là xác suất nhiễm bệnh được tính toán dựa trên một quan sát với dung lượng mẫu lớn). Hãy tính xác suất mắc bệnh của 2 trong số 10 động vật.

1.3.2

Xác suất mắc một bệnh là $0,25$. Hãy tính xác suất không phát hiện được ca nhiễm bệnh trong số 30 động vật kiểm tra.

1.3.3

Bệnh dại xuất hiện với tần suất $0,005$. Cần tiến hành kiểm tra bao nhiêu chó trong vùng để phát hiện bệnh dại với độ chính xác 95% .

1.3.4

Khối lượng (kg) ở 210 ngày tuổi của lợn Pietrain có các kiểu gen Halothane khác nhau được trình bày ở bảng số liệu dưới đây. Vẽ đồ thị và tính các tham số thống kê mô tả của bộ số liệu vừa nêu.

NN	Nn	Nn
118,54	133,90	105,85
123,66	127,07	100,49
97,10	136,34	108,54
96,30	120,10	80,00
112,20	107,60	106,27
124,40	102,68	121,95
109,51	89,50	111,50
110,98	119,02	130,00
128,80	125,61	112,20
119,51	94,70	110,49
120,24	91,33	101,20
114,10	114,60	137,56
100,20	144,88	122,68
114,00	102,89	102,00
104,15	116,80	116,34
101,71	117,56	116,63
86,27	112,44	111,22
106,34	116,34	111,50
110,49	117,11	112,00
128,54	136,10	121,71
112,68	111,57	103,66
107,47	120,00	131,95
103,90	110,98	104,15
101,50	113,20	121,50
114,88	83,90	153,70
102,00	109,76	102,00
82,20	93,73	108,78
109,76	98,07	102,00
129,27	100,00	105,78
100,00	102,89	110,96
118,05	114,94	109,02
111,00	93,01	101,93
120,98	86,02	101,93
110,84	86,51	93,01
125,06	95,85	86,51
145,37	94,70	93,01
88,43	94,70	93,01
130,60	94,70	93,01
120,24	94,70	93,01
113,98	94,70	93,01
117,83	94,70	93,01
104,34	94,70	93,01
131,08	94,70	93,01
102,24	94,70	93,01
90,36	94,70	93,01
108,67	94,70	93,01
105,12	94,70	93,01
129,76	94,70	93,01
108,43	94,70	93,01
115,37	94,70	93,01
113,90	94,70	93,01
119,76	94,70	93,01
113,95	94,70	93,01
111,33	94,70	93,01
120,96	94,70	93,01
118,78	94,70	93,01
126,10	94,70	93,01
105,54	94,70	93,01
104,10	94,70	93,01
110,36	94,70	93,01
133,01	94,70	93,01
118,54	94,70	93,01
109,40	94,70	93,01
104,10	94,70	93,01
111,33	94,70	93,01
102,17	94,70	93,01
120,98	94,70	93,01
110,60	94,70	93,01

Chương 2 **ƯỚC LUỢNG VÀ KIỂM ĐỊNH GIẢ THIẾT**

Mục đích của chương này nhằm cung cấp cho bạn đọc các bước để kiểm ước lượng và kiểm định giả thiết; phân biệt được sự khác biệt giữa kiểm định một phía và hai phía cũng như cách lựa chọn kiểm định phù hợp; việc áp dụng các kiến thức này trong giải quyết các bài toán cụ thể trong chăn nuôi, thú y và thủy sản.

2.1. GIẢ THIẾT VÀ ĐỐI THIẾT

Khi khảo sát một tổng thể (hoặc nhiều tổng thể) và xem xét một (hoặc nhiều) biến ngẫu nhiên có thể đưa ra một giả thiết nào đó liên quan đến phân phối của biến ngẫu nhiên hoặc nếu biết phân phối rồi thì đưa ra giả thiết về tham số của tổng thể. Để có thể đưa ra một kết luận thống kê nào đó đối với giả thiết thì phải chọn mẫu ngẫu nhiên, tính tham số mẫu, chọn mức ý nghĩa α sau đó đưa ra kết luận.

Bài toán kiểm định tham số Θ của phân phối có dạng $H_0: \Theta = \Theta_0$ với Θ_0 là một số đã cho nào đó. Kết luận thống kê có dạng: “chấp nhận H_0 ” hay “bắc bỏ H_0 ”. Nhưng nếu đặt vấn đề như vậy thì cách giải quyết hết sức khó, vì nếu không chấp nhận $H_0: \Theta = \Theta_0$ thì điều đó có nghĩa là có thể chấp nhận một trong vô số Θ khác Θ_0 , do đó thường đưa ra bài toán dưới dạng cụ thể hơn nữa: cho giả thiết H_0 và đối thiết H_1 , khi kết luận thì hoặc chấp nhận H_0 hoặc bắc bỏ H_0 , và trong trường hợp này, tuy không hoàn toàn tương đương, nhưng coi như chấp nhận đối thiết H_1 .

Nếu chấp nhận H_0 trong lúc giả thiết đúng là H_1 thì mắc **sai lầm loại II** và xác suất mắc sai lầm này được gọi là rủi ro loại hai β . Ngược lại nếu bắc bỏ H_0 trong lúc giả thiết đúng chính là H_0 thì mắc **sai lầm loại I** và xác suất mắc sai lầm đó gọi là rủi ro loại một α .

Giả thiết	Quyết định	
	Bắc bỏ H_0	Chấp nhận H_0
H_0 đúng	Sai lầm loại I (α)	Quyết định đúng
H_0 sai	Quyết định đúng	Sai lầm loại II (β)

Như vậy trong bài toán kiểm định giả thiết luôn luôn có hai loại rủi ro, loại I và loại II, tùy vấn đề mà nhấn mạnh loại rủi ro nào. Thông thường người ta hay tập trung chú ý vào **sai lầm loại I** và khi kiểm định phải không chế sao cho **rủi ro loại I** không vượt quá một mức α gọi là **mức ý nghĩa**.

Trước hết xem xét cụ thể bài toán kiểm định giả thiết $H_0: \Theta = \Theta_0$, đối thiết $H_1: \Theta = \Theta_1$ với Θ_1 là một giá trị khác Θ_0 . Đây là bài toán kiểm định giả thiết đơn. Quy tắc kiểm định căn cứ vào hai giá trị cụ thể Θ_1 và Θ_0 , vào mức ý nghĩa α và còn căn cứ vào cả sai lầm loại hai. Việc này về lý thuyết thống kê không gặp khó khăn gì.

Sau đó mở rộng quy tắc sang cho bài toán kiểm định giả thiết kép. $H_1: \Theta \neq \Theta_0; \Theta > \Theta_0$ hoặc $\Theta < \Theta_0$, việc mở rộng này có khó khăn nhưng các nhà nghiên cứu lý thuyết xác suất thống kê đã giải quyết được, do đó về sau khi kiểm định giả thiết $H_0: \Theta = \Theta_0$ có thể chọn một trong 3 đối thiết H_1 sau:

$H_1: \Theta \neq \Theta_0$ gọi là đối thiết hai phía.

$H_1: \Theta > \Theta_0$ gọi là đối thiết phải.

$H_1: \Theta < \Theta_0$ gọi là đối thiết trái.

Hai đối thiết sau gọi là đối thiết một phía. Việc chọn đối thiết nào tuỳ thuộc vào đề khảo sát cụ thể. Nếu chỉ thu thập dữ liệu hoặc thí nghiệm để so sánh 2 giống, 2 phương pháp, không có ưu tiên giống hay phương pháp nào thì chọn $H_0: \mu_1 = \mu_2$ đối thiết $H_1: \mu_1 \neq \mu_2$.

Nếu so sánh giống mới (phương pháp mới: μ_2) với mục đích xem giống mới (phương pháp mới) có hơn giống cũ (phương pháp cũ: μ_1) đang dùng hay không thì chọn đối thiết $H_1: \mu_1 < \mu_2$ (Không quan tâm đến trường hợp $\mu_2 \leq \mu_1$ vì đang tìm giống (phương pháp) tốt hơn giống cũ). Ngược lại nếu quan tâm đến giá thành khi sản xuất giống hay khi sử dụng phương pháp thì lại dùng đối thiết $H_1: \mu_1 > \mu_2$ (vì chỉ chọn giá rẻ hơn).

Trong y học nếu người bệnh có huyết áp cao thì mục đích dùng thuốc là để giảm. Giả sử huyết áp người bệnh đang giao động quanh mức $H_0: \mu = 150$ dùng thuốc phải đạt mục đích giảm do đó $H_1: \mu < 150$. Ngược lại nếu người bệnh huyết áp thấp $H_0: \mu = 90$ thì dùng thuốc để tăng do đó $H_1: \mu > 90$.

Tóm lại kiểm định giả thiết là bài toán do chủ quan người dùng đặt ra trước khi thu thập mẫu và xử lý dữ liệu. Bình thường chọn đối thiết hai phía (chú ý trong thực tế phải hiểu giả thiết $H_0: \mu_1 = \mu_2$ có nghĩa là hai trung bình không khác nhau rõ rệt, còn $H_1: \mu_1 \neq \mu_2$ có nghĩa 2 trung bình khác nhau rõ rệt. Trong phạm vi giáo trình này đề cập chủ yếu đến đối thiết hai phía hay còn gọi là hai đuôi.

2.2. ƯỚC LƯỢNG GIÁ TRỊ TRUNG BÌNH μ CỦA BIẾN PHÂN PHỐI CHUẨN $N(\mu, \sigma^2)$

2.2.1. Ước lượng μ khi biết phương sai σ^2

Dựa vào lý thuyết xác suất có thể đưa ra ước lượng giá trị trung bình quần thể (μ) theo các bước sau đây:

- + Chọn mẫu dung lượng n , tính trung bình cộng \bar{x} .
- + Ở mức tin cậy P đã cho lấy $\alpha = 1 - P$, sau đó tìm giá trị tới hạn $z_{(\alpha/2)}$ trong Phụ lục 3 (hàm $\Phi(z)$ tìm z sao cho $\Phi(z) = 1 - \alpha/2$)
- + Khoảng tin cậy đối xứng ở mức tin cậy P :

$$\bar{x} - z(\alpha / 2) \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z(\alpha / 2) \frac{\sigma}{\sqrt{n}}$$

Ví dụ 2.1: Khối lượng bao thíc ăn gia súc phân phoi chuẩn N (μ, σ^2) với $\sigma = 1,5\text{kg}$. Cân thử 25 bao được khối lượng trung bình $\bar{x} = 49\text{kg}$. Hãy ước lượng kỳ vọng μ với mức tin cậy $P = 0,95$; $z(0,025) = 1,96$.

$$49 - 1,96 \frac{1,5}{\sqrt{25}} \leq \mu \leq 49 + 1,96 \frac{1,5}{\sqrt{25}}$$

$$49 - 0,588 \leq \mu \leq 49 + 0,588$$

$$48,41\text{kg} \leq \mu \leq 49,59\text{kg}$$

2.2.2. Ước lượng μ khi không biết phương sai σ^2

Dựa vào phân phối Student có thể đưa ra ước lượng μ theo các bước sau đây:

- + Chọn mẫu dung lượng n , tính trung bình cộng \bar{x} và độ lệch chuẩn s .
- + Ở mức tin cậy P lấy $\alpha = 1 - P$, tìm giá trị tới hạn $t(\alpha/2, n-1)$ trong bảng 2, cột $\alpha/2$, dòng $n-1$.
- + Khoảng tin cậy đối xứng ở mức tin cậy P :

$$\bar{x} - t(\alpha/2, n-1) \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t(\alpha/2, n-1) \frac{s}{\sqrt{n}}$$

Ví dụ 2.2: Từ một quần thể đồng nhất, chọn và cân ngẫu nhiên 30 con gà được khối lượng trung bình $\bar{x} = 3,03\text{ kg}$; $s = 0,0279\text{ kg}$. Hãy ước lượng μ với mức tin cậy $P = 0,98$; $\alpha = 1 - P = 0,02$; $\alpha/2 = 0,01$; $t(0,01; 29) = 2,756$.

$$3,03 - 2,756 \frac{0,0279}{\sqrt{30}} \leq \mu \leq 3,03 + 2,756 \frac{0,0279}{\sqrt{30}}$$

$$3,03 - 0,014 \leq \mu \leq 3,03 + 0,014$$

$$3,016\text{kg} \leq \mu \leq 3,044\text{ kg}$$

2.3. KIỂM ĐỊNH GIÁ TRỊ TRUNG BÌNH μ CỦA BIẾN PHÂN PHỐI CHUẨN N(μ, σ^2)

2.3.1. Kiểm định giả thiết $H_0: \mu = \mu_0$ khi biết σ^2

Tiến hành kiểm định theo các bước sau:

- + Chọn mẫu dung lượng n , tính trung bình cộng \bar{x} .
- + Chọn mức ý nghĩa α .
- + Tìm giá trị tới hạn $z(\alpha/2)$ nếu kiểm định 2 phía hoặc $z(\alpha)$ nếu kiểm định một phía.

+ Tính giá trị thực nghiệm: $Z_{TN} = \frac{(\bar{x} - \mu_0)}{\frac{\sigma}{\sqrt{n}}} = \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma}$

So sánh Z_{TN} và z tới hạn để rút ra kết luận theo nguyên tắc sau:

Kết luận:

Với $H_1: \mu \neq \mu_0$ (Kiểm định hai phía). Nếu $|Z_{TN}|$ (giá trị tuyệt đối của Z_{TN}) nhỏ hơn hay bằng $z(\alpha/2)$ thì chấp nhận H_0 nếu ngược lại thì bác bỏ H_0 , tức là chấp nhận H_1 .

Với $H_1: \mu > \mu_0$ (Kiểm định một phía). Nếu Z_{TN} nhỏ hơn hay bằng giá trị tới hạn $z(\alpha)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

Với $H_1: \mu < \mu_0$ (Kiểm định một phía). Nếu Z_{TN} lớn hơn hay bằng giá trị tới hạn $-z(\alpha)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

Ví dụ 2.3: Nuôi 100 con cừu theo một chế độ riêng. Mục đích của thí nghiệm là xem chế độ này có làm tăng khối lượng của cừu một năm tuổi hay không. Biết rằng 100 cừu này được lấy mẫu từ một quần thể có khối lượng trung bình một năm tuổi là 30 kg và phuơng sai là 25 kg². Giả thiết tăng khối lượng phân phối chuẩn N ($\mu, 25$), hãy kiểm định giả thiết $H_0: \mu = 30$ đối với $H_1: \mu > 30$ ở mức $\alpha = 0,05$. Biết rằng khối lượng trung bình của 100 cừu thí nghiệm là 32 kg.

$$Z_{TN} = \frac{(32 - 30)\sqrt{100}}{5} = 4; \quad z(0,05) = 1,64$$

Kết luận: Vì $Z_{TN} > Z_{LT}$ nên giả thiết H_0 bị bác bỏ, như vậy tăng khối lượng trung bình không phải là 30 kg. Chế độ nuôi mới đã làm tăng khối lượng cừu một năm tuổi.

Ví dụ 2.4: Một mẫu cho trước gồm 100 bò sữa có sản lượng sữa một chu kỳ tiết sữa trung bình là 3850 kg. Số bò này có xuất phát từ quần thể có giá trị trung bình là 4000 kg và độ lệch chuẩn là 1000 hay không? Giả sử sản lượng sữa của quần thể tuân theo phân phối chuẩn N ($\mu, 1000^2$). Hãy kiểm định giả thiết $H_0: \mu = 4000$ đối với $H_1: \mu \neq 4000$ ở mức $\alpha = 0,05$.

$$Z_{TN} = \frac{(3850 - 4000)\sqrt{100}}{1000} = -1,5 \rightarrow |Z_{TN}| = 1,5; \quad z(0,025) = 1,96$$

Kết luận: Chấp nhận H_0 , số bò sữa nêu trên xuất phát từ một quần thể ban đầu có sản lượng sữa chu kỳ là 4000 kg.

2.3.2. Kiểm định giả thiết $H_0: \mu = \mu_0$ khi không biết σ^2

Đây là trường hợp phổ biến khi kiểm định giá trị trung bình của phân phối chuẩn. Tiến hành các bước sau:

+ Lấy mẫu dung lượng n , tính \bar{x} và s^2

+ Tính giá trị T thực nghiệm $T_{TN} = \frac{(\bar{x} - \mu_0)\sqrt{n}}{s}$

+ Tìm giá trị tới hạn $t(\alpha/2, n-1)$ với kiểm định 2 phía hoặc tìm $t(\alpha, n-1)$ nếu kiểm định 1 phía trong bảng 2.

Kết luận:

Với $H_1: \mu \neq \mu_0$ (Kiểm định hai phía). Nếu $|T_{TN}|$ (giá trị tuyệt đối của T_{tn}) nhỏ hơn hay bằng $t(\alpha/2, n-1)$ thì chấp nhận H_0 nếu ngược lại thì bác bỏ H_0 , tức là chấp nhận H_1 .

Với $H_1: \mu > \mu_0$ (Kiểm định một phía) Nếu $T_{TN} \leq t(\alpha, n-1)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1

Với $H_1: \mu < \mu_0$ (Kiểm định một phía) Nếu $T_{TN} \geq -t(\alpha, n-1)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

Ví dụ 2.5: Thời gian mang thai của bò phân phối chuẩn $N(285, \sigma^2)$. Theo dõi thời gian mang thai (ngày) của 6 bò được các số liệu.

307 293 293 283 294 297

Kiểm định giả thiết $H_0: \mu = 285$ ngày đối thiêt $H_1: \mu \neq 285$ ngày

$$\text{Tính } \bar{x} = \frac{(307 + 293 + 293 + 283 + 294 + 297)}{6} = \frac{1767}{6} = 294,5$$

$$s^2 = \frac{307^2 + 293^2 + \dots + 294^2 + 297^2 - \frac{1767^2}{6}}{5} = 59,9; s = \sqrt{59,9} = 7,7395 \approx 7,74$$

$$T_{TN} = \frac{(294,5 - 285)}{7,74} \times \sqrt{6} = \frac{9,5}{3,16} = 3,007; t(0,025; 5) = 2,571$$

Kết luận: Vì $|T_{TN}| = 3,007 > t(0,025; 5)$ nên bác bỏ H_0 như vậy thời gian mang thai không phải 285 ngày.

Ví dụ 2.6: Trong điều kiện chăn nuôi bình thường, lượng sữa trung bình của một con bò là 19 kg /ngày. Trong một đợt hạn, người ta theo dõi 25 con bò và được lượng sữa trung bình 17,5 kg/ngày, độ lệch chuẩn $s = 2,5$ kg. Giả thiết lượng sữa phân phối chuẩn, hãy kiểm định giả thiết $H_0: \mu = 19$ với đối thiêt $\mu < 19$ ở mức $\alpha = 0,05$.

$$T_{TN} = \frac{(17,5 - 19)\sqrt{25}}{2,5} = -3; t(0,05; 24) = 1,711$$

Kết luận: $T_{TN} < 1,711$ nên giả thiết H_0 bị bác bỏ, như vậy sản lượng sữa trung bình không còn là 19 kg/ngày nữa mà thấp hơn.

2.4. KIỂM ĐỊNH HAI GIÁ TRỊ TRUNG BÌNH CỦA HAI BIẾN PHÂN PHỐI CHUẨN

Giả sử có hai tổng thể và theo dõi một biến định lượng X nào đó, ví dụ khối lượng sau 6 tháng nuôi của hai đàn gà, năng suất của hai giống lúa, năng suất của một giống ngô khi bón theo hai công thức phân bón khác nhau, sản lượng một loại quả khi trồng theo hai khoảng cách hàng . . .

Gọi biến X trên tổng thể thứ nhất là X_1 (phân phối chuẩn $N(\mu_1, \sigma_1^2)$) và biến X trên tổng thể thứ hai là X_2 (phân phối chuẩn $N(\mu_2, \sigma_2^2)$). Để so sánh μ_1 và μ_2 chúng ta phải chọn mẫu. Có hai cách chọn mẫu: **Chọn mẫu theo cặp** và **chọn mẫu độc lập**.

2.4.1. Chọn mẫu theo cặp

Từ tổng thể thứ nhất ta chọn một mẫu n cá thể được các giá trị x_1, x_2, \dots, x_n , từ tổng thể thứ hai chọn một mẫu cũng gồm n cá thể được y_1, y_2, \dots, y_n .

Giữa hai mẫu này có mối quan hệ cặp, tức là có n cặp (x_i, y_i) ($i = 1, n$). Các cặp này hình thành do khi chọn mẫu ta đã dùng những quan hệ cặp như quan hệ gia đình (vợ chồng, anh em, thí dụ chọn n tổ chim sau đó bắt chim đực vào mẫu đại diện cho tổng thể chim đực, bắt chim cái vào mẫu đại diện cho tổng thể chim cái), quan hệ trước sau (thí dụ cá thể được đo một chỉ số trước khi dùng thuốc và số liệu này đại diện cho tổng thể trước khi dùng thuốc, một thời gian sau khi dùng thuốc lại đo lại chỉ số và số liệu này đại diện cho tổng thể sau khi dùng thuốc), cũng có khi các cặp này là các cặp số liệu do chúng ta bố trí thí nghiệm theo cặp: chọn 2 ô ruộng, một ô ruộng(hay một chuồng) bố trí giống thử nghiệm, một ô ruộng (một chuồng) bố trí giống đối chứng.

Viết lại số liệu dưới dạng hai cột hay hai hàng rồi tính hiệu số $d_i = y_i - x_i$

X_1	x_1	x_2	...	x_n
X_2	y_1	y_2	...	y_n
d	d_1	d_2	...	d_n

Tiếp theo tính giá trị trung bình \bar{d} và độ lệch chuẩn s_d

Giả thiết $H_0: \mu_2 = \mu_1$ đối với $H_1: \mu_2 \neq \mu_1$ được chuyển thành $H_0: \mu_d = 0$ đối với $H_1: \mu_d \neq 0$ (tương tự $H_1: \mu_2 > \mu_1$ chuyển thành $H_1: \mu_d > 0$ và $H_1: \mu_2 < \mu_1$ chuyển thành $H_1: \mu_d < 0$).

Ở mức ý nghĩa α việc kiểm định gồm các bước sau:

$$+ \text{Tính giá trị thực nghiệm } T_{TN} = \frac{\bar{d}\sqrt{n}}{s_d}$$

+ Tìm giá trị tới hạn $t(\alpha/2, n-1)$ nếu kiểm định 2 phía hoặc $t(\alpha, n-1)$ nếu kiểm định một phía bằng 2.

Kết luận:

+ Kiểm định hai phía $H_1: \mu_2 \neq \mu_1$. Nếu $|T_{TN}| \leq t(\alpha/2, n-1)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

+ Kiểm định một phía $H_1: \mu_2 > \mu_1$. Nếu $T_{TN} \leq t(\alpha, n-1)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

+ Kiểm định một phía $H_1: \mu_2 < \mu_1$. Nếu $T_{TN} \geq -t(\alpha, n-1)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

Ví dụ 2.7: Tăng khối lượng (kg) của 10 cặp bê sinh đôi giống hệt nhau dưới hai chế độ chăm sóc khác nhau (A và B). Bê trong từng cặp được bắt thăm ngẫu nhiên về một trong hai cách chăm sóc. Giả thiết tăng khối lượng có phân phối chuẩn. Hãy kiểm định giả thiết H_0 : Tăng khối lượng trung bình ở hai cách chăm sóc như nhau, đối thiết H_1 : Tăng khối lượng trung bình khác nhau ở hai cách chăm sóc với mức ý nghĩa $\alpha = 0,05$. Số liệu thu được như sau:

Cặp sinh đôi	1	2	3	4	5	6	7	8	9	10
Tăng khối lượng ở cách A	19,50	17,69	17,69	19,05	20,87	19,50	17,24	19,96	23,13	19,50
Tăng khối lượng ở cách B	16,78	15,88	15,42	18,60	17,69	16,78	15,88	18,14	21,77	16,32
Chênh lệch (d)	2,72	1,81	2,27	0,45	3,18	2,72	1,36	1,82	1,36	3,18

$$n = 10; \bar{d} = 2,09; s_d = 0,89; T_{TN} = \frac{2,09\sqrt{10}}{0,89} = 7,43; t(0,025; 9) = 2,262$$

Kết luận: Bác bỏ giả thiết H_0 , chấp nhận H_1 : “Tăng khối lượng trung bình ở hai cách chăm sóc là khác nhau”.

Ví dụ 2.8: Có 15 trại phổi hợp tham gia thử nghiệm khẩu phần ăn bình thường (A) và khẩu phần ăn có bổ sung đồng (B). Mỗi trại lấy 2 khu nuôi lợn tương tự về mọi mặt sau đó chỉ định ngẫu nhiên một khu ăn khẩu phần A, một khu ăn khẩu phần B. Tăng khối lượng trung bình (kg/ngày) của một con lợn được trình bày ở bảng dưới. Kiểm định giả thiết H_0 : “Hai khẩu phần A và B cho kết quả tăng khối lượng trung bình như nhau” với đối thiết H_1 : “Khẩu phần có bổ sung đồng cho tăng khối lượng trung bình cao hơn”.

Trại	Khẩu phần		Trại	Khẩu phần		Trại	Khẩu phần	
	A (x_i)	B (y_i)		A (x_i)	B (y_i)		A (x_i)	B (y_i)
1	0,42	0,53	6	0,50	0,52	11	0,50	0,51
2	0,53	0,47	7	0,44	0,44	12	0,54	0,54
3	0,48	0,56	8	0,45	0,46	13	0,46	0,50
4	0,50	0,59	9	0,30	0,43	14	0,48	0,50
5	0,42	0,47	10	0,52	0,57	15	0,53	0,59

Giá trị trung bình $\bar{d} = 0,0407$; độ lệch chuẩn $s_d = 0,0489$

$$T_{TN} = \frac{0,0407}{0,0489} \times \sqrt{15} = 3,22; t(0,05; 14) = 1,761$$

Kết luận: Vì $T_{TN} > t$ nên bác bỏ H_0 , chấp nhận H_1 . Như vậy khẩu phần bổ sung đồng cho tăng khối lượng trung bình cao hơn khẩu phần ăn thường.

2.4.2. Chọn mẫu độc lập

Tù hai tổng thể chọn ra hai mẫu độc lập, dung lượng có thể bằng nhau hoặc khác nhau. Tính các tham số thống kê $\bar{x}_1; s_1^2$ của mẫu thứ nhất; $\bar{x}_2; s_2^2$ của mẫu thứ hai. Để kiểm định giả thiết H_0 : $\mu_2 = \mu_1$ với các đối thiết H_1 ở mức ý nghĩa α cần phải chia ra 3 trường hợp:

a. Biết phương sai σ_1^2 và σ_2^2

$$+ \text{Tính Z thực nghiệm: } Z_{TN} = \frac{(\bar{x}_2 - \bar{x}_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

+ Tìm giá trị tới hạn $z(\alpha/2)$ nếu kiểm định 2 phía hoặc $z(\alpha)$ nếu kiểm định một phía trong bảng 1.

Kết luận:

+ Kiểm định hai phía $H_1: \mu_2 \neq \mu_1$. Nếu $|Z_{TN}| \leq z(\alpha/2)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

+ Kiểm định một phía $H_1: \mu_2 > \mu_1$. Nếu $Z_{TN} \leq z(\alpha)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

+ Kiểm định một phía $H_1: \mu_2 < \mu_1$. Nếu $Z_{TN} \geq -z(\alpha)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

Ví dụ 2.9: Chiều dài cá trong 2 ao phân phối chuẩn với độ lệch chuẩn $\sigma_1 = 2$ cm và $\sigma_2 = 2,2$ cm. Lấy mẫu 100 con của ao thứ nhất được giá trị trung bình $\bar{x}_1 = 8$ cm; lấy mẫu 120 con của ao thứ hai được giá trị trung bình $\bar{x}_2 = 8,5$ cm. Hãy kiểm định giả thiết $H_0: \mu_1 = \mu_2$ với đối thiết $H_1: \mu_1 \neq \mu_2$ ở mức ý nghĩa $\alpha = 0,05$.

$$Z_{TN} = \frac{(8,5 - 8)}{\sqrt{\frac{2^2}{100} + \frac{2,2^2}{120}}} = 1,764 \quad ; \quad z(0,025) = 1,96$$

Vì $|Z_{TN}| = 1,764 < 1,96$ nên chấp nhận H_0 : “Chiều dài cá trung bình trong 2 ao như nhau”.

b. Không biết phương sai σ_1^2 và σ_2^2 mẫu lớn ($n_1 \geq 30, n_2 \geq 30$)

$$+ \text{Tính giá trị thực nghiệm } Z_{TN} = \frac{(\bar{x}_2 - \bar{x}_1)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

+ Tìm giá trị tới hạn $z(\alpha/2)$ nếu kiểm định 2 phía hoặc $z(\alpha)$ nếu kiểm định một phía trong bảng 1.

Kết luận:

+ Kiểm định hai phía $H_1: \mu_2 \neq \mu_1$. Nếu $|Z_{TN}| \leq z(\alpha/2)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

+ Kiểm định một phía $H_1: \mu_2 > \mu_1$. Nếu $Z_{TN} \leq z(\alpha)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

+ Kiểm định một phía $H_1: \mu_2 < \mu_1$. Nếu $Z_{TN} \geq -z(\alpha)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

Ví dụ 2.10: Để đánh giá tăng khói lượng của lợn ở hai chế độ ăn khác nhau. Khói lượng sau 4 tháng ở hai chế độ nuôi có các số liệu sau. Ở chế độ thứ nhất, tiến hành thí nghiệm 64 con ($n_1 = 64$) được giá trị trung bình $\bar{x}_1 = 73,2$ kg biết $\sigma_1 = 10,9$ kg; tương tự với chế độ thứ 2 sẽ có $n_2 = 68$; $\bar{x}_2 = 76,6$; $\sigma_2 = 11,4$ kg. Giả thiết khói lượng phân phối chuẩn $N(\mu_1, \sigma_1^2)$ và $N(\mu_2, \sigma_2^2)$. Kiểm định giả thiết $H_0: \mu_2 = \mu_1$ với đối thiết $H_1: \mu_2 > \mu_1$.

$$Z_{TN} = \frac{76,6 - 73,2}{\sqrt{\frac{10,9^2}{64} + \frac{11,4^2}{68}}} = 1,75; \quad z(0,05) = 1,645$$

Kết luận:

$Z_{TN} > z(0,05)$ vì vậy chấp nhận H_1 : “chế độ ăn thứ hai cho kết quả trung bình cao hơn chế độ ăn thứ nhất”.

c. Không biết phương sai σ_1^2 và σ_2^2 , mẫu bé (ít nhất một trong 2 số $n_1, n_2 < 30$)

Đây là một bài toán còn rất nhiều vướng mắc về mặt lý thuyết do đó, chỉ trình bày trường hợp có thêm giả thiết phụ: $\sigma_1^2 = \sigma_2^2$

+ Tính phương sai chung: $s_c^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$

+ Tính $T_{TN} = \frac{(\bar{x}_2 - \bar{x}_1)}{\sqrt{s_c^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$

+ Tìm giá trị tới hạn $t(\alpha/2, n_1 + n_2 - 2)$ với kiểm định 2 phía hoặc $t(\alpha, n_1 + n_2 - 2)$ nếu kiểm định một phía.

Kết luận:

+ Kiểm định hai phía $H_1: \mu_2 \neq \mu_1$. Nếu $|T_{TN}| \leq t(\alpha/2, n_1+n_2 -2)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

+ Kiểm định một phía $H_1: \mu_2 > \mu_1$. Nếu $T_{TN} \leq t(\alpha, n_1+n_2 -2)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

+ Kiểm định một phía $H_1: \mu_2 < \mu_1$. Nếu $T_{TN} \geq -t(\alpha, n_1+n_2 -2)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

Ví dụ 2.11: Để so sánh khói lượng của 2 giống bò, chọn ngẫu nhiên 12 bò của giống thứ nhất và 15 bò của giống thứ 2. Khói lượng (kg) của từng bò được xác định và thu được các tham số thống kê sau: $n_1 = 12$; $\bar{x}_1 = 196,2$ kg; $s_1 = 10,62$ kg; $n_2 = 15$; $\bar{x}_2 = 153,70$ kg; $s_2 = 12,30$ kg. Kiểm định giả thiết H_0 : Hai giống bò có khói lượng trung bình

như nhau với đối thiết H_1 : Giống bò thứ nhất có khối lượng trung bình lớn hơn giống bò thứ hai. Giả sử khối lượng của 2 giống bò có phân phối chuẩn và hai phương sai bằng nhau với mức ý nghĩa $\alpha = 0,05$.

$$s_c^2 = \frac{(11 \times 10,62^2 + 14 \times 12,30^2)}{11+14} = 134,33$$

$$T_{TN} = \frac{(196,2 - 153,7)}{\sqrt{134,33 \times \left(\frac{1}{12} + \frac{1}{15} \right)}} = \frac{42,5}{4,489} = 9,46; t(0,05,25) = 1,708$$

Kết luận: Ở mức ý nghĩa $\alpha = 0,05$ vì $T_{TN} > t$ nên bác bỏ H_0 . Như vậy giống thứ nhất có khối lượng trung bình cao hơn giống thứ hai.

Ví dụ 2.12: Hai giống gà có khối lượng phân phối chuẩn, lấy mẫu 10 gà đối với giống thứ nhất và 16 gà của giống thứ 2. Các tham số về khối lượng 45 ngày tuổi của 2 mẫu nêu trên như sau:

Với mẫu thứ nhất $n_1 = 10$; $\bar{x}_1 = 2,8$ kg; $s_1^2 = 0,1111$ kg² với mẫu thứ hai $n_2 = 16$; $\bar{x}_2 = 2,35$ kg; $s_2^2 = 0,0667$ kg². Kiểm định giả thiết H_0 : Hai giống gà có khối lượng trung bình như nhau với đối thiết H_1 : Hai giống gà có khối lượng trung bình khác nhau. Mức ý nghĩa $\alpha = 0,05$.

$$s_c^2 = \frac{9 \times 0,1111 + 15 \times 0,0667}{9+15} = \frac{1,99995}{24} = 0,83331$$

$$T_{TN} = -3,866 \rightarrow |T_{TN}| = 3,866; t(0,025; 24) = 2,064$$

Kết luận: Bác bỏ H_0 , như vậy hai giống gà có khối lượng trung bình khác nhau.

2.5. ƯỚC LUỢNG VÀ KIỂM ĐỊNH XÁC SUẤT

Trường hợp tổng thể có 2 loại cá thể A và A', loại A chiếm tỷ lệ p và A' chiếm tỷ lệ q = 1-p. Sau khi chọn mẫu có thể dùng phân phối chuẩn để tính gần đúng phân phối nhị thức, từ đó suy ra công thức ước lượng p.

2.5.1. Ước lượng xác suất P

Khi dung lượng mẫu lớn ($n \geq 100$) và p không bé quá, cũng không lớn quá ($np > 5$, $nq > 5$). Từ mẫu có dung lượng n, tính số cá thể loại A được lần số m và lần suất f = m/n với mức tin cậy α có khoảng tin cậy đối xứng sau:

$$f - z(\alpha/2) \sqrt{\frac{f(1-f)}{n}} \leq P \leq f + z(\alpha/2) \sqrt{\frac{f(1-f)}{n}}$$

Ví dụ 2.13: Để biết tỷ lệ trứng nở p của một loại trứng; cho vào máy áp 100 quả, kết quả có 80 quả nở.

$f = 80/100 = 0,8$ ở mức tin cậy $\alpha = 0,05$ và $z(0,025) = 1,96$. Ta có thể tính được khoảng tin cậy như sau:

$$0,8 - 1,96\sqrt{\frac{0,8 \times 0,2}{100}} \leq P \leq 0,8 + 1,96\sqrt{\frac{0,8 \times 0,2}{100}}$$

$$0,8 - 0,0784 \leq P \leq 0,8 + 0,0784 \Leftrightarrow 0,72 \leq P \leq 0,88$$

2.5.2. Kiểm định giả thiết $H_0: P = P_0$

Khi dung lượng mẫu lớn ($n \geq 100$) và p không bé quá, cũng không lớn quá ($np > 5$, $nq > 5$). Từ mẫu có dung lượng n , tính số cá thể loại A được tần số m và tần suất $f = m/n$. Ở mức ý nghĩa α tính $z(\alpha/2)$ với kiểm định 2 phía hoặc $z(\alpha)$ nếu kiểm định một phía.

$$\text{Tính: } Z_{TN} = \frac{f - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$$

Kết luận:

Với đối thiết hai phía $H_1: P \neq P_0$. Nếu $|Z_{TN}| \leq z(\alpha/2)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

Với đối thiết một phía $H_1: P > P_0$. Nếu $Z_{TN} \leq z(\alpha)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

Với đối thiết một phía $H_1: P < P_0$. Nếu $Z_{TN} \geq -z(\alpha)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

Ví dụ 2.14: Áp 100 quả trứng có 82 quả nở. Kiểm định giả thiết H_0 : tỷ lệ nở $P = 0,80$, đối thiết $H_1: P \neq 0,8$ với $\alpha = 0,05$.

$$n = 100; m = 82; f = 82/100 = 0,82;$$

$$Z_{TN} = \frac{0,82 - 0,80}{\sqrt{\frac{0,8 \cdot 0,2}{100}}} = 0,5; z(0,025) = 1,96;$$

Kết luận: Chấp nhận H_0 : “Tỷ lệ áp nở là 0,80”.

2.5.3. Kiểm định giả thiết $H_0: P_2 = P_1$

Khi dung lượng cả 2 mẫu đều lớn ($n_1 > 100, n_2 > 100$) và các p_i không bé quá (hoặc lớn quá) có thể kiểm định như sau (ở mức ý nghĩa α).

$$\text{Tính các tần suất: } f_1 = \frac{m_1}{n_1}; f_2 = \frac{m_2}{n_2}$$

$$\text{Tính tần suất chung: } f = \frac{m_1 + m_2}{n_1 + n_2}$$

Tìm giá trị tới hạn $z(\alpha/2)$ nếu kiểm định 2 phía hoặc $z(\alpha)$ nếu kiểm định một phía.

$$\text{Tính giá trị thực nghiệm: } Z_{TN} = \frac{f_2 - f_1}{\sqrt{f(1-f)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Kết luận:

Với đối thiết hai phía $H_1: P_2 \neq P_1$. Nếu $|Z_{TN}| \leq z(\alpha/2)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1

Với đối thiết một phía $H_1: P_2 > P_1$. Nếu $Z_{TN} \leq z(\alpha)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1

Với đối thiết một phía $H_1: P_2 < P_1$. Nếu $Z_{TN} \geq -z(\alpha)$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1

Ví dụ 2.15: Dùng thuốc A điều trị cho 200 bệnh nhân thấy 150 người khỏi bệnh. Tương tự với thuốc B đối với 100 bệnh nhân thì 72 người khỏi bệnh. Hãy kiểm định giả thiết H_0 : Tỷ lệ khỏi bệnh của hai thuốc như nhau với đối thiết H_1 : tỷ lệ khỏi bệnh của hai thuốc khác nhau với mức ý nghĩa $\alpha = 0,05$.

$$n_1 = 200; m_1 = 150; f_1 = 150/200 = 0,75; n_2 = 100; m_2 = 72; f_2 = 72/100 = 0,72;$$

$$f = \frac{150+72}{200+100} = 0,74;$$

$$Z_{TN} = \frac{0,72 - 0,75}{\sqrt{0,74 \times 0,26 \times \left(\frac{1}{200} + \frac{1}{100}\right)}} = -0,5584 \Rightarrow |Z_{TN}| = 0,5584; z(0,025) = 1,96.$$

Kết luận: Chấp nhận H_0 ; tức là tỷ lệ khỏi bệnh ở 2 loại thuốc là như nhau.

2.6. PHÂN TÍCH PHƯƠNG SAI

Mở rộng bài toán so sánh hai trung bình của hai tổng thể ở mục trên khi có nhiều hơn 2 trung bình chúng ta có bài toán phân tích phương sai một yếu tố. Thí dụ có a tổng thể, để khảo sát các biến X_1, X_2, \dots, X_a trên các tổng thể đó chúng ta lấy ở mỗi tổng thể một mẫu các quan sát độc lập:

Mẫu 1 $x_{11}, x_{12}, \dots, x_{1r1}$

Mẫu 2 $x_{21}, x_{22}, \dots, x_{2r2}$

....

Mẫu a $x_{a1}, x_{a2}, \dots, x_{2ra}$

Tất cả có $n = \sum r_i$ quan sát. Viết lại các quan sát x_{ij} dưới dạng

$$x_{ij} = \mu_i + e_{ij} \quad e_{ij} \text{ gọi là sai số hay phần dư} \quad (2.1)$$

Giả thiết các biến X_i độc lập, phân phối chuẩn $N(\mu_i, \sigma^2)$, các quan sát trong mẫu độc lập. Từ giả thiết trên có thể nêu cụ thể 3 giả thiết sau đối với các sai số e_{ij}

- Các biến e_{ij} độc lập với nhau
- Các biến e_{ij} phân phối chuẩn với kỳ vọng bằng 0
- Các biến e_{ij} có phương sai bằng nhau (σ^2)

Bài toán phân tích phương sai một yếu tố chính là bài toán kiểm định giả thiết H_0 : “Các trung bình μ_i bằng nhau” với đối thiêt H_1 : “Có ít nhất một cặp trung bình khác nhau”. Nếu gọi μ là trung bình của các μ_i thì có thể viết (2.1) lại như sau:

$$x_{ij} = \mu + a_i + e_{ij} \quad (2.2)$$

$$\text{Với } a_i = \mu_i - \mu; \quad \sum a_i = 0$$

Giả thiết H_0 bây giờ là: “Các a_i đều bằng 0” còn H_1 là “Không phải tất cả các a_i đều bằng 0”.

Để phân tích phương sai chúng ta gọi các trung bình cộng của các mẫu quan sát là \bar{x}_i . Nếu giả thiết H_0 đúng thì các X_i có cùng phân phối $N(\mu, \sigma^2)$ và có thể coi các mẫu quan sát nói trên được lấy ra từ cùng một tổng thể.

Gọi \bar{x} là trung bình chung của tất cả các mẫu.

Tính tổng bình phương tất cả các sai số (gọi là tổng bình phương toàn bộ SS_{TO})

$$SS_{TO} = \sum_{i=1}^a \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^a \sum_{j=1}^{n_i} x_{ij}^2 - n\bar{x}^2$$

Đem tổng bình phương này chia cho $(n - 1)$ được một ước lượng của σ^2 .

SS_{TO}/σ^2 phân phối χ^2 với $df_{TO} = (n - 1)$ bậc tự do.

Đối với mỗi mẫu quan sát tính tổng bình phương sai số trong mẫu (mà nếu đem chia cho bậc tự do tương ứng $(n_i - 1)$ thì được một ước lượng của σ^2) sau đó gộp lại thành tổng bình phương do sai số SS_E (Giống như cách đã làm khi đi tìm phương sai chung s^2_c trong trường hợp mẫu bé và hai phương sai bằng nhau ở mục 2.4.2.3).

$$SS_E = \sum_{i=1}^a \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

Đem SS_E chia cho $n - a$ được một ước lượng của σ^2

SS_E/σ^2 phân phối χ^2 với $df_E = (n - a)$ bậc tự do.

Có thể chứng minh hệ thức sau:

$$\sum_{i=1}^a \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^a \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + \sum_{i=1}^a \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2$$

Tổng thứ ba gọi là tổng bình phương do yếu tố SS_A .

Nếu x_{ij} phân phối chuẩn $N(\mu_i, \sigma^2)$ thì các trung bình cộng \bar{x}_i phân phối chuẩn $N(\mu_i, \sigma^2/n_i)$. Từ đó suy ra nếu đem SS_A chia cho $(a - 1)$ thì được ước lượng của σ^2 .

Tổng SS_A/σ^2 phân phối χ^2 với $df_A = (a-1)$ bậc tự do.

Như vậy đã tách tổng bình phương toàn bộ ra hai tổng:

$$SS_{TO} = SS_A + SS_E$$

Đồng thời bậc tự do toàn bộ cũng tách thành 2 bậc tự do:

$$df_{TO} = df_A + df_E$$

Mỗi tổng bình phương chia cho bậc tự do tương ứng sẽ cho một ước lượng của phương sai σ^2 và mỗi tổng sau khi chia cho σ^2 sẽ phân phối χ^2 với số bậc tự do tương ứng. Xét tỷ số MS_A / MS_E với $MS_A = SS_A / df_A$ và $MS_E = SS_E / df_E$

Dựa trên lý thuyết về phân phối Khi bình phương (χ^2) và phân phối F có kết luận sau: MS_A / MS_E phân phối Fisher- Snederco (F). Từ đó có cách kiểm định sau đây đối với giả thiết H_0 (đối thiết H_1):

- + Tính giá trị thực nghiệm $F_{TN} = MS_A / MS_E$.
- + Tìm giá trị tới hạn $F_{(\alpha, df_A, df_E)}$.
- + Nếu $F_{TN} \leq F_{(\alpha, df_A, df_E)}$ thì chấp nhận H_0 , ngược lại thì chấp nhận H_1 .

Toàn bộ quy trình phân tích phương sai được tóm tắt trong bảng phân tích phương sai sau:

Nguồn biến động	Bậc tự do	Tổng bình phương	Trung bình bình phương	F_{TN}	F tới hạn
Yếu tố	$df_A = a-1$	SS_A	$MS_A = SS_A / df_A$	MS_A / MS_E	$F_{(\alpha, df_A, df_E)}$
Sai số ngẫu nhiên	$df_E = n-a$	SS_E	$MS_E = SS_E / df_E$		
Tổng biến động	$df_{TO} = n-1$	SS_{TO}			

Để thuận tiện thường kẻ bảng chứa dữ liệu và tính theo thứ tự sau:

- + Tính dung lượng n_i , tổng hàng TH_i , trung bình \bar{x}_i , TH_i^2 / n_i .
- + Tổng các dung lượng $n = \sum n_i$, tổng tất cả các x_{ij} $ST = \sum \sum x_{ij}$
- + Số điều chỉnh $G = ST^2 / n$.
- + $SS_{TO} = \sum \sum x_{ij}^2 - G$ bậc tự do $df_{TO} = n - 1$.
- + $SS_A = \sum TH_i^2 / n_i - G$ bậc tự do $df_A = a - 1$.
- + $SS_E = SS_{TO} - SS_A$ bậc tự do $df_E = df_{TO} - df_A = n - a$.
- + Tính các trung bình $MS_A = SS_A / df_A$ và $MS_E = SS_E / df_E$.
- + Tính $F_{TN} = MS_A / MS_E$.
- + Tìm giá trị $F_{(\alpha, df_A, df_E)}$.
- + So sánh F_{TN} với $F_{(\alpha, df_A, df_E)}$.

Ví dụ 2.16: Khối lượng (kg) của 20 lợn lai Duroc x (Landrace x Yorkshire) được nuôi ở 5 trại (A, B, C, D và E) từ cai sữa (21 ngày tuổi) đến 90 ngày tuổi. Số liệu được

trình bày trong bảng dưới. Giả thiết khói lượng tuân theo phân phối chuẩn. Kiểm định giả thiết H_0 : Khối lượng trung bình của lợn 90 ngày tuổi ở 5 trại bằng nhau với đối thiết H_1 : Khối lượng trung bình của lợn 90 ngày tuổi ở 5 trại không bằng nhau. Mức ý nghĩa $\alpha = 0,05$.

Trại	Khối lượng (kg)				n_i	TH_i	TH_i^2/n_i	\bar{x}_i	
A	32,2	34,9	29,7		3	96,8	3123,413	32,27	
B	28,4	28,0	22,8	28,5	29,4	5	137,1	3759,282	27,42
C	28,8	29,5	23,1	20,1		4	101,5	2575,563	25,38
D	41,5	36,3	31,7	31,0	38,2	5	178,7	6386,738	35,74
E	33,0	26,0	30,6			3	89,6	2676,053	29,87
Tổng					20	603,7	18521,0492		

$$n = 20 \quad ST = 603,7 \quad \Sigma TH_i^2 / n_i = 18521,0492$$

$$\text{Số điều chỉnh } G = 603,7^2 / 20 = 18222,6845$$

$$\text{Tổng các bình phương } \Sigma \Sigma x_{ij}^2 = 18727,6900$$

$$SS_{TO} = 18727,69 - 18222,68 = 505,0055; \quad \text{bậc tự do } df_{TO} = 20 - 1 = 19$$

$$SS_A = 18521,0492 - 18222,6845 = 298,3647; \quad \text{bậc tự do } df_A = 5 - 1 = 4$$

$$SS_E = 505,0055 - 298,3647 = 206,6408; \quad \text{bậc tự do } df_E = 19 - 4 = 15$$

$$MS_A = 298,3647/4 = 74,5912; \quad MS_E = 206,6408/15 = 13,7761$$

$$F_{TN} = 74,5912/ 13,7761 = 5,4145 \quad F(0,05;4;15) = 3,056$$

Có thể tổng hợp các kết quả thu được theo bảng phân tích phương sai (ANOVA) sau:

Nguồn biến động	Bậc tự do	Tổng bình phương	Trung bình bình phương	F_{TN}	F tối hạn
Trại	4	298,3647	74,5912	5,4145	$F_{(0,05;4;15)} = 3,056$
Sai số ngẫu nhiên	15	206,6408	13,7761		
Tổng biến động	19	505,0055			

Kết luận: Bác bỏ H_0 , như vậy là bác bỏ giả thiết “Khối lượng trung bình của lợn 90 ngày tuổi ở 5 trại bằng nhau”.

Sau khi có kết luận như trên thì vấn đề đặt ra là phải so sánh 5 trung bình của 5 trại để tìm ra các trung bình nào bằng nhau, các trung bình nào khác nhau. Vấn đề này sẽ được trình bày kỹ ở phần sau.

Qua cách làm như trên chúng ta thấy để kiểm định giả thiết H_0 : “Các trung bình bằng nhau” với đối thiết H_1 : “Có ít nhất một cặp trung bình khác nhau” phải tìm cách tách tổng bình phương toàn bộ SS_{TO} thành các tổng bình phương SS_A và SS_E căn cứ vào 2 nguồn biến động của số liệu: biến động do sự khác nhau giữa các mẫu và biến động do sự khác nhau giữa các số liệu trong cùng một mẫu. Đồng thời phải tách bậc tự do toàn bộ df_{TO} thành các bậc tự do df_A và df_E tương ứng với các tổng SS_A , SS_E . Từ đó có tên phân tích phương sai.

Trong phần sau khi có nhiều nguồn biến động thì phải tách SS_{TO} thành nhiều tổng ứng với các nguồn biến động và tách bậc tự do df_{TO} thành nhiều bậc tự do, sau đó kiểm định các giả thiết tương ứng với các nguồn biến động nhờ phân phối Fisher- Snederco.

2.7. BÀI TẬP

2.7.1

Tăng khối lượng trung bình (g/ngày) của 36 lợn nuôi vỗ béo giống Landrace được rút ngẫu nhiên từ một trại chăn nuôi. Số liệu thu được như sau:

577 596 594 612 600 584 618 627 588 601 606 559 615 607 608 591 565 586
621 623 598 602 581 631 570 595 603 605 616 574 578 600 596 619 636 589

Cán bộ kỹ thuật trại cho rằng tăng khối lượng trung bình của toàn đàn lợn trong trại là 607 g/ngày. Theo anh chị kết luận đó đúng hay sai, vì sao?

2.7.2

Anh chị hãy kiểm tra kết luận với bài tập tương tự như 2.7.1, biết rằng độ lệch chuẩn của tính trạng này ở Landrace là 24 g/ngày.

2.7.3

Tỷ lệ thụ thai bằng thụ tinh nhân tạo từ tinh trùng của 2 bò đực giống được xác định trên nhóm bò cái gồm 50 con; 18 nhóm bò cái sử dụng tinh trùng của bò đực A và 16 đối với bò đực B. Tỷ lệ thụ thai (%) thu được như sau:

Bò đực A	74,2	62,1	57,7	71,7	62,0	76,1	70,6	68,3	68,4	79,8
		71,1	70,9	65,5	61,2	60,8	73,9	51,9	63,7	
Bò đực B	49,6	49,2	53,2	56,5	69,1	54,2	80,7	62,7	71,5	67,5
		64,6	75,4	79,6	59,8	68,8	60,2			

Hãy cho biết tỷ lệ thụ thai của 2 bò đực nêu trên.

2.7.4

Nồng độ fructoza (mg%) trong tinh dịch bò trước và sau khi ủ được xác định trên 12 mẫu tinh bò đực; các giá trị thu được như sau:

Mẫu số	1	2	3	4	5	6	7	8	9	10	11	12
Trước khi ủ	116	190	570	375	236	505	120	322	429	102	167	299
Sau khi ủ	30	58	100	48	58	153	54	66	67	34	69	82

Kết luận về nồng độ fructoza trong tinh dịch bò trước và sau khi ủ.

2.7.5

Một thí nghiệm được tiến hành nhằm nghiên cứu ảnh hưởng của progesterone lên chu kỳ động dục của cừu Merino. Sử dụng 4 liều khác nhau (0, 10, 25 và 40 mg/ngày) tiêm dưới da liên tục trong 4 ngày tính từ ngày động dục. Chu kỳ động dục (ngày) của 8 cừu trong mỗi nhóm thu được như sau:

Liều 0 mg/ngày 18 14 18 18 18 18 19

Liều 10 mg/ngày	15	14	17	14	12	13	12	13
Liều 25 mg/ngày	11	13	11	11	12	11	11	12
Liều 40 mg/ngày	9	10	12	10	11	11	10	11

Cho biết ảnh hưởng của progesterone lên chu kỳ động dục ở cừu Merino.

Chương 3

MỘT SỐ KHÁI NIỆM VỀ THIẾT KẾ THÍ NGHIỆM

Thiết kế thí nghiệm là lập kế hoạch nghiên cứu nhằm tìm ra những vấn đề mới hoặc khẳng định lại hoặc bác bỏ kết quả của những nghiên cứu trước đó. Thông qua thí nghiệm, người nghiên cứu có thể tìm được câu trả lời cho một số vấn đề đặt ra hoặc rút ra được kết luận về một hiện tượng nào đó. Theo một nghĩa hẹp, thí nghiệm được thiết kế trong một môi trường quản lý nhằm nghiên cứu ảnh hưởng của một hay nhiều yếu tố lên các quan sát. Mục tiêu của chương này nhằm cung cấp cho người học cách phân loại thí nghiệm; các khái niệm thông dụng được sử dụng trong thiết kế thí nghiệm; nguyên tắc bố trí thí nghiệm; phương pháp chọn mẫu và ước tính dung lượng mẫu cần thiết cho một thí nghiệm.

3.1. PHÂN LOẠI THÍ NGHIỆM

Theo bản chất của thí nghiệm, các thí nghiệm có thể chia thành hai loại: 1) thí nghiệm **quan sát**, 2) thí nghiệm **thực nghiệm**. Trong phần thiết kế thí nghiệm của giáo trình này, chúng tôi sẽ tập trung vào các thí nghiệm thực nghiệm.

Trong chăn nuôi, thú y, các thí nghiệm thường tập trung vào 2 lĩnh vực: 1) các nghiên cứu trong chăn nuôi về dinh dưỡng, năng suất và di truyền ở vật nuôi; 2) các nghiên cứu trong thú y về tình hình dịch bệnh và các biện pháp phòng, điều trị bệnh.

3.1.1. Thí nghiệm quan sát

Trong thí nghiệm **quan sát**, chỉ đơn thuần quan sát các động vật thí nghiệm và ghi lại các dữ liệu liên quan đến các tính trạng quan tâm. Không tác động để can thiệp vào sự tồn tại của đối tượng quan sát. Trong loại thí nghiệm quan sát, các động vật không thể bố trí một cách ngẫu nhiên về các công thức thí nghiệm.

Điều tra là một trường hợp đặc biệt của thí nghiệm quan sát. Trong điều tra, cần tiến hành kiểm tra toàn bộ hoặc một nhóm động vật để tìm ra các giá trị của những tham số khác nhau trong quần thể. Điều tra có thể là một trong các trường hợp sau:

- Điều tra quần thể - tiến hành kiểm tra tất cả các động vật trong quần thể.
- Điều tra mẫu - tiến hành kiểm tra những nhóm động vật đại diện và dựa vào kết quả điều tra ta có thể rút ra kết luận cho cả quần thể.

3.1.2. Thí nghiệm thực nghiệm

Trong thí nghiệm thực nghiệm, các nhà khoa học can thiệp vào nghiên cứu bằng cách áp dụng các công thức thí nghiệm khác nhau lên các nhóm động vật nghiên cứu. Sau đó chúng ta tiến hành quan sát ảnh hưởng của các công thức thí nghiệm lên đối

tượng nghiên cứu. Đối với loại thí nghiệm này, các động vật được bố trí một cách ngẫu nhiên đối với các công thức thí nghiệm trong quá trình thiết kế.

3.2. MỘT SỐ KHÁI NIỆM TRONG THIẾT KẾ THÍ NGHIỆM

3.2.1. Yếu tố thí nghiệm

Yếu tố thí nghiệm là một biến độc lập gồm hàng loạt các phần tử có chung một bản chất mà có thể so sánh trong quá trình thực hiện thí nghiệm. Ví dụ như một giống vật nuôi, kiểu gen Halothane ở lợn, hàm lượng protein trong khẩu phần, thuốc kháng sinh, vắc xin trong phòng và điều trị bệnh,...

Một thí nghiệm có thể có một hoặc nhiều yếu tố thí nghiệm và các yếu tố thí nghiệm này có thể là yếu tố cố định hoặc yếu tố ngẫu nhiên.

3.2.2. Mức

Các phần tử riêng biệt khác nhau trong cùng một yếu tố thí nghiệm được gọi là **mức**. Ví dụ có một yếu tố thí nghiệm là kiểu gen Halothane ở lợn thì sẽ có 3 phần tử khác nhau tương ứng với 3 kiểu gen (NN, Nn, nn) hay còn được gọi là **3 mức**. Hoặc khi nghiên cứu ảnh hưởng của protein đến sản lượng sữa bò ta có thể nghiên cứu ở 3 mức protein khác nhau. Trong thú y, các nhà nghiên cứu hiệu quả điều trị bệnh của các loại thuốc khác nhau; có thể coi mỗi loại thuốc tương đương với 1 mức.

3.2.3. Công thức thí nghiệm (công thức thí nghiệm)

Một tổ hợp các mức của các yếu tố được gọi là một **công thức thí nghiệm** hay công thức thí nghiệm. Ví dụ nghiên cứu ảnh hưởng của protein ở 3 mức khác nhau đến sản lượng sữa bò, trong trường hợp này ta sẽ có 3 công thức. Ta xét một hoàn cảnh tương tự nhưng có thêm yếu tố thứ 2 là thức ăn tinh ở 2 mức, lúc này sẽ có tất cả 6 công thức thí nghiệm.

3.2.4. Đơn vị thí nghiệm

Đơn vị thực hiện nhỏ nhất ứng với một công thức được gọi là **đơn vị thí nghiệm**. Đơn vị thí nghiệm trong chăn nuôi, thú y thường là từng động vật nhung đôi khi là một nhóm động vật. Ví dụ: khi nghiên cứu về khả năng sản xuất của vật nuôi, mỗi động vật được đánh số tai (hoặc đeo số cánh), tiến hành cân đo từng cá thể tại từng thời điểm nghiên cứu, mỗi cá thể cho một quan sát do đó đơn vị thí nghiệm là từng động vật. Tuy nhiên, khi nghiên cứu tiêu tốn thức ăn đối với một kg tăng khối lượng, trong thực tế ta không thể theo dõi được lượng thức ăn thu nhận của từng vật nuôi (ngoại trừ từng động vật được gắn chíp điện tử và có hệ thống theo dõi thức ăn thu nhận tự động) mà ta chỉ biết được số thức ăn thu nhận được của một nhóm gồm nhiều cá thể khác nhau. Tức là từ một nhóm cá thể như vậy ta chỉ có một quan sát duy nhất, đây cũng chính là điều mà các nhà nghiên cứu cần phải chú ý.

3.2.5. Dữ liệu (số liệu)

Nếu đơn vị thí nghiệm là một cá thể thì sau khi cân, đo ta được một dữ liệu (data) hay một quan sát (observation). Nếu đơn vị là một nhóm gồm nhiều cá thể thì có thể cân, đo chung cho cả nhóm hoặc lấy một số cá thể nhất định trong nhóm để cân, đo sau đó suy ra một dữ liệu chung cho đơn vị thí nghiệm. Các số liệu của các nhóm có thể lưu trữ để đánh giá sai số của đơn vị thí nghiệm.

3.2.6. Khối

Tập hợp các đơn vị thí nghiệm có chung một hay nhiều đặc tính được gọi là **khối**.

3.2.7. Lặp lại

Mỗi công thức, trừ trường hợp đặc biệt, đều được lặp lại một số lần nhất định. Số lần lặp lại thường chọn bằng nhau vì nhìn chung, đối với nhiều mô hình, khi các lần lặp của các công thức bằng nhau có thể đưa ra các công thức tính khá thuận tiện và đơn giản. Nếu số lần lặp không bằng nhau thì phải sử dụng cách tính theo mô hình hồi quy nhiều biến tổng quát khá phức tạp, kèm theo đó việc kiểm định các giả thiết, đặc biệt việc tính các kỳ vọng của các trung bình bình phương, cũng gặp rất nhiều khó khăn.

Trong thực tế, số lần lặp bằng nhau nhưng trong quá trình thí nghiệm ta ít khi thu thập được đầy đủ dữ liệu vì có một số động vật bị chết hoặc bị loại thải do không đáp ứng được các yêu cầu của thí nghiệm. Số lượng động vật thí nghiệm sống sót đến khi kết thúc thí nghiệm phụ thuộc vào từng loại thí nghiệm và loài vật nuôi khác nhau. Nếu mất ít dữ liệu, có thể tìm cách thay thế dữ liệu bị mất bằng tổ hợp của các dữ liệu còn lại theo một công thức cụ thể, kèm theo sự điều chỉnh của các bậc tự do tương ứng; ngược lại, phải coi như số lần lặp khác nhau và dùng mô hình hồi quy tổng quát.

3.2.8. Nhắc lại

Nhắc lại là làm lại thí nghiệm trong điều kiện tương tự có thể để kết luận đạt mức độ tin cậy.

3.2.9. Nhóm đối chứng

Là nhóm động vật thí nghiệm được tạo ra trong quá trình bố trí thí nghiệm nhưng được nuôi dưỡng, chăm sóc trong điều kiện bình thường hiện có.

3.3. CÁC BƯỚC TIẾN HÀNH THÍ NGHIỆM

Một thí nghiệm thường được bố trí và có thể mô tả qua các bước sau: 1) Đặt vấn đề, 2) Phát biểu giả thiết, 3) Mô tả thiết kế thí nghiệm, 4) Thực hiện thí nghiệm (thu thập số liệu), 5) Phân tích số liệu thu thập được từ thí nghiệm và 6) Giải thích kết quả liên quan đến giả thiết.

Lập kế hoạch cho một thí nghiệm bắt đầu bằng việc nêu lên những vấn đề cần thiết; bên cạnh đó là tập hợp các tài liệu liên quan bao gồm cả những nghiên cứu trước đó; tiếp đến là nêu lên hướng giải quyết vấn đề. Sau những vấn đề vừa nêu, mục đích

nghiên cứu được xác định. Mục đích nghiên cứu phải rõ ràng bởi vì các bước tiếp theo trong quá trình thiết kế thí nghiệm đều phụ thuộc vào mục đích đặt ra.

Bước tiếp theo là xác định nguyên liệu và phương pháp phương pháp nghiên cứu. Thiết kế thí nghiệm phải mô tả số liệu được thu thập như thế nào. Số liệu có thể thu thập từ các nghiên cứu quan sát từ các quá trình tự nhiên hoặc từ các thí nghiệm được bố trí trong môi trường thí nghiệm. Nếu chúng ta biết thông tin nào được thu thập và bằng cách nào sẽ được sử dụng để thu thập các số liệu này, thì việc rút ra kết luận sẽ dễ dàng và hiệu quả hơn rất nhiều. Điều này đúng với cả thí nghiệm quan sát và thí nghiệm thực nghiệm; đồng thời cũng rất quan trọng để phát hiện ra những thông tin bất ngờ dẫn đến những kết luận mới.

Đối với các nhà thống kê, thiết kế thí nghiệm là đặt ra các tiêu chuẩn để sử dụng khi chọn mẫu. Đối với thí nghiệm thực nghiệm việc thiết kế thí nghiệm bao gồm: Xác định các công thức thí nghiệm, xác định các đơn vị thí nghiệm, số lần lặp lại, việc bố trí các đơn vị vào các công thức thí nghiệm, các sai số thí nghiệm có thể mắc phải.

Giả thiết thống kê thường đi theo sau giả thiết nghiên cứu. Chấp nhận hay bác bỏ giả thiết thống kê giúp tìm được câu trả lời cho mục đích nghiên cứu. Trong kiểm định giả thiết các nhà thống kê sử dụng mô hình thống kê. Mô hình thống kê theo sau mô hình thí nghiệm thường được giải thích với các công thức toán học.

Thu thập số liệu được thực hiện theo thiết kế thí nghiệm. Phân tích thống kê được tiến hành sau khi thu thập được số liệu bao gồm phân tích, miêu tả và giả thích kết quả. Mô hình sử dụng trong phân tích được xây dựng dựa trên mục đích và mô hình thí nghiệm. Thông thường cách phân tích số liệu được xác định trước khi thu thập số liệu; đôi khi lại được xác định sau khi thu thập số liệu nếu người nghiên cứu tìm được một cách tốt hơn để rút ra kết luận hoặc xác định được một khía cạnh mới liên quan đến vấn đề nghiên cứu.

Cuối cùng, người nghiên cứu phải có khả năng rút ra kết luận để hoàn thiện mục tiêu nghiên cứu. Kết luận phải rõ ràng và chính xác. Người nghiên cứu phải thảo luận các ứng dụng vào thực tế của nghiên cứu đồng thời nêu ra những khả năng đặt ra trong tương lai liên quan đến vấn đề tương tự.

3.4. SAI SỐ THÍ NGHIỆM

Bản chất của vật liệu sinh học là sự biến động. Toàn bộ sự biến động này có thể phân chia thành phần biến động có thể giải thích được và không giải thích được. Mỗi đơn vị thí nghiệm (y_{ij}) có thể được biểu diễn như sau:

$$y_{ij} = \mu_i + e_{ij}$$

Trong đó, μ là giá trị ước tính miêu tả sự ảnh hưởng giải thích được của nhóm thứ i và e_{ij} ảnh hưởng không giải thích được. Vì vậy, các quan sát (y_{ij}) khác nhau nguyên nhân là do ảnh hưởng giải thích được của các nhóm (i) khác nhau và các ảnh hưởng không giải thích được (e_{ij}) khác nhau. Ước tính μ_i được giải thích do ảnh hưởng của

nhóm i, nhưng sự khác nhau giữa các đơn vị thí nghiệm trong cùng một nhóm thì không thể giải thích được. Biến động này thường được gọi là sai số thí nghiệm.

Sai số thí nghiệm có thể bao gồm 2 dạng sau đây: Sai số ngẫu nhiên và sai số hệ thống. Sai số hệ thống là các ảnh hưởng nhất định làm lệch các giá trị đo được trong một nghiên cứu. Sai số này có thể xuất phát từ sự thiếu đồng nhất trong quá trình thực hiện thí nghiệm, có thể do dụng cụ thí nghiệm không được hiệu chỉnh, do ảnh hưởng của nhiệt độ không ổn định, do thiên lệch trong quá trình sử dụng thiết bị. Nếu sự thiên lệch này được phát hiện thì hiệu chỉnh là biện pháp hiệu quả nhất. Chúng cũng đặc biệt khó giải quyết nếu không phát hiện được vì chúng ảnh hưởng lên các giá trị một cách có hệ thống nhưng không biết theo xu hướng nào.

Sai số ngẫu nhiên xuất hiện do các tác động ngẫu nhiên, không dự đoán được. Chúng tạo ra các biến động không giải thích được. Kỳ vọng của biến động này bằng 0 vì vậy khi có một loạt các quan sát thì các tính toán dựa vào trung bình sẽ không bị thiên lệch về một hướng. Trong sinh học luôn tồn tại sai số ngẫu nhiên ví dụ trong chăn nuôi, các động vật khi đo hay phân tích một chỉ tiêu nào đó, luôn cho các kết quả khác nhau tuy có thể không lớn lắm.

Để giảm được sai số có hệ thống và sự thiên lệch ta xem xét 2 giải pháp sau đây:

- 1 Bố trí động vật vào các công thức thí nghiệm
- 2 Phương pháp làm mù

3.5. BỐ TRÍ THÍ NGHIỆM VÀO CÁC CÔNG THỨC THÍ NGHIỆM

3.5.1. Sự cần thiết của phân chia ngẫu nhiên

Sự thiên lệch có thể xuất hiện trong quá trình phân chia động vật vào các công thức thí nghiệm. Sự thiên lệch này có thể do yếu tố chủ quan. Ví dụ chúng ta phân chia các động vật vào các công thức thí nghiệm theo sở thích chủ quan (thích công thức thí nghiệm nào thì bố trí các động vật ‘tốt’, không thích thì bố trí động xấu’) hoặc có sự khác nhau có hệ thống giữa nhóm đối chứng và nhóm thí nghiệm, lúc đó chúng ta không thể kết luận được sự sai khác sau khi thực hiện thí nghiệm là do ảnh hưởng của công thức thí nghiệm hay do sự khác nhau có hệ thống.

Một phương pháp tiếp cận hay được sử dụng để loại bỏ sự thiên lệch này là bố trí ngẫu nhiên hay còn gọi là ngẫu nhiên hóa các động vật thí nghiệm vào các công thức thí nghiệm. Trong quá trình bố trí chúng ta phân động vật vào các công thức thí nghiệm với các yêu cầu sau:

- Tất cả các động vật thí nghiệm đều có cơ hội nhận được một công thức thí nghiệm bất kỳ;
- Việc bố trí động vật vào công thức thí nghiệm này không ảnh hưởng đến việc bố trí động vật vào công thức thí nghiệm khác;
- Chúng ta không biết trước công thức thí nghiệm mà từng động vật được phân vào.

- Ngẫu nhiên hoá có một số ưu điểm sau:
- Loại bỏ được sự thiên lệch trong quá trình bố trí động vật thí nghiệm;
- Tạo được sự giống nhau giữa các nhóm.

3.5.2. Các phương pháp phân chia ngẫu nhiên

Tốt nhất là tránh sử dụng các phương pháp cơ học như tung đồng xu hoặc ném con xúc xắc để bố trí động vật về các công thức thí nghiệm. Mặc dù các phương pháp này về mặt xác suất vẫn được chấp nhận để tạo ra sự ngẫu nhiên, nhưng nó cồng kềnh và không kiểm tra được. Thông thường, bảng số ngẫu nhiên được sử dụng để phân động vật về với công thức thí nghiệm. Ngoài ra ta có thể sử dụng máy tính để tạo ra các số ngẫu nhiên. Khi thiết kế thí nghiệm, số đơn vị thí nghiệm thường bằng nhau ở các công thức thí nghiệm.

a. Phân chia ngẫu nhiên đơn giản

Đây là cách ngẫu nhiên hoá cơ bản không có sự phân biệt hoặc hạn chế. Ví dụ tiến hành phân 12 động vật thí nghiệm được đánh số từ 1 đến 12 về 2 công thức thí nghiệm (đối chứng - C và thí nghiệm - T). Tiến hành chọn số ngẫu nhiên từ bảng số ngẫu nhiên phần phụ lục. Giả sử ta lấy 10 số có 1 chữ số ở hàng đầu tiên; như vậy ta sẽ được dãy số ngẫu nhiên sau 813766407765. Nếu số ngẫu nhiên là số chẵn động vật sẽ phân về với C và số lẻ về với T.

Đơn vị thí nghiệm số	1	2	3	4	5	6	7	8	9	10	11	12
Số ngẫu nhiên	8	1	3	7	6	6	4	0	7	7	6	5
Công thức	C	T	T	T	C	C	C	C	T	T	C	T

Có thể tiến hành các bước tương tự đối với thí nghiệm có số công thức thí nghiệm nhiều hơn 2. Ví dụ có 3 công thức thí nghiệm A, B và C, chọn các số 1-3, 4-6 và 7-9 tương ứng với các công thức thí nghiệm và bỏ qua số 0. Tương tự như ví dụ trên ta có dãy số ngẫu nhiên 8137664077652 và kết quả thu được CAACBBCCBBA. Trong trường hợp này, sự ngẫu nhiên đã không được tuân thủ vì có 3A, 5B và 4C. Cách *phân chia ngẫu nhiên hạn chế* được đưa ra nhằm khắc phục những hạn chế này.

b. Phân chia ngẫu nhiên theo khối

Phân chia ngẫu nhiên đơn giản dựa trên nguyên tắc tất cả các động vật tương đối đồng đều, mỗi động vật đều có cơ hội như nhau khi sắp vào một công thức thí nghiệm. Tuy nhiên điều này không còn đúng khi dung lượng mẫu lớn. Căn cứ vào một tiêu chí lựa chọn cụ thể thí dụ lựa chọn theo lứa, theo tuổi, theo khối lượng, theo hành vi . . . chúng ta sẽ phân chia các động vật thành một số nhóm sao cho các động vật cùng nhóm tương đối đồng đều, sau đó mới chia ngẫu nhiên các động vật trong từng nhóm vào các công thức thí nghiệm. Đây chính là cách *phân chia ngẫu nhiên theo khối*.

Ví dụ 3.1: Nghiên cứu bệnh viêm khớp ở chó. Tạo ra 3 khối khác nhau tương ứng với 3 nhóm có khối lượng cơ thể **lớn, trung bình và nhỏ**. Như vậy sẽ biết được khối

lượng cơ thể của động vật ảnh hưởng đến mức độ mắc bệnh của từng công thức thí nghiệm. Tức là so sánh các công thức thí nghiệm có đề cập đến khối lượng cơ thể.

c. Phân chia ngẫu nhiên hạn chế

Nhìn chung, ta mong muốn có số đơn vị thí nghiệm bằng nhau ở các công thức thí nghiệm. Kỹ thuật ngẫu nhiên đơn giản đã được sử dụng để đạt được điều này nếu dung lượng mẫu đủ lớn. Tuy nhiên chúng ta có thể gặp sự thiếu cân bằng khi dung lượng mẫu tương đối bé. Điều này đã được minh họa ở ví dụ phần *phân chia ngẫu nhiên đơn giản* với sự phân bố 3A, 5B và 4C. Có thể sử dụng kiểu *phân chia ngẫu nhiên hạn chế* để khắc phục những hạn chế này.

Ví dụ có 16 đơn vị thí nghiệm, cần chia về 4 công thức thí nghiệm A, B, C và D. Ta sẽ chọn các số 1-2, 3-4, 5-6, 7-8 tương ứng với các công thức thí nghiệm A, B và C và bỏ qua số 9 và 0. Tương tự ta có dãy số ngẫu nhiên 8137664~~0~~77652~~99977~~42 và kết quả DABDCCBDD. Như vậy đến số ngẫu nhiên thứ 9 đã có đủ 4 động vật về với công thức thí nghiệm D. Các số ngẫu nhiên 7-8 cũng sẽ bỏ qua vì đã đủ số lượng và đã có 1 động vật thí nghiệm về với A, 2 với B và 2 về với C. Tiếp theo ta sẽ có CC, ở số ngẫu nhiên thứ 11 đã đủ 4 đơn vị cho công thức C. Tương tự như vậy chắc chắn số đơn vị thí nghiệm ở các công thức thí nghiệm bằng nhau.

Phân chia ngẫu nhiên theo khối thường được dùng kết hợp với phân chia ngẫu nhiên giới hạn.

d. Phân chia ngẫu nhiên theo nhóm (Cluster)

Thông thường, một động vật thí nghiệm được coi như một đơn vị thí nghiệm. Tuy nhiên trong chăn nuôi và thú y, thì một nhóm động vật cũng được coi như một đơn vị thí nghiệm. Bởi vì thức ăn, thuốc và vắc xin thường được sử dụng cho một nhóm động vật trong cùng một lứa, nuôi trong cùng một chuồng, một bầy hoặc được sử dụng cho cả đàn hay tất cả cá nuôi trong một bể. Trong trường hợp này, ta tiến hành sử dụng kỹ thuật ngẫu nhiên hóa cho cả nhóm động vật thí nghiệm hay còn gọi là *ngẫu nhiên hóa theo nhóm*. Như vậy tất cả động vật trong nhóm sẽ nhận được cùng một công thức thí nghiệm sau đó cần phải tập hợp kết quả trên các nhóm để đánh giá ảnh hưởng của các công thức thí nghiệm. Lưu ý rằng trong kiểu phân chia này một nhóm động vật chỉ được coi như một đơn vị thí nghiệm.

Ví dụ 3.2: Nghiên cứu tiêu tốn thức ăn trên một kg tăng khối lượng đôi với lợn nuôi vỗ béo. Về lý thuyết có thể tiến hành quan sát lượng thức ăn mà từng con lợn thu nhận hằng ngày; nhưng về thực tế điều này rất khó thực hiện. Ta chỉ có thể quan sát được lượng thức ăn tiêu tốn trong một ô chuồng có nuôi khoảng 30 – 50 con và từ đây có thể tính được tiêu tốn thức ăn cho 1 kg tăng khối lượng. Ở đây 1 ô chuồng nuôi 30 - 50 con được coi như một đơn vị thí nghiệm. Để có thể nghiên cứu được tiêu tốn thức ăn trên 1kg tăng khối lượng ta phải tiến hành thí nghiệm trên nhiều ô chuồng và phải bắt thăm ô chuồng nào áp dụng công thức thí nghiệm nào.

3.6. PHƯƠNG PHÁP LÀM MÙ

Trong phần nêu trên ta đã dùng kỹ thuật bổ trí động vật vào các công thức thí nghiệm bằng kỹ thuật ngẫu nhiên hoá để đảm bảo không có sự sai số có hệ thống. Tuy nhiên sự thiên lệch có thể xuất hiện do những định kiến của người trực tiếp thực hiện và người đánh giá. Để đảm bảo trong thí nghiệm không có sự thiên lệch như đã nêu trên ta sử dụng kỹ thuật làm mù. Có 2 kỹ thuật làm mù:

- 1) Kỹ thuật làm mù đơn và 2) Kỹ thuật làm mù kép.

Kỹ thuật làm mù kép là kỹ thuật mà cả người trực tiếp thực hiện và người và người đánh giá không biết các thông tin về thí nghiệm. Đối với kỹ thuật làm mù đơn, hoặc người trực tiếp thực hiện hoặc người đánh giá không biết các thông tin về thí nghiệm.

Để người trực tiếp thực hiện không thể phân biệt được sự khác nhau giữa nhóm đối chứng và thí nghiệm, có thể sử dụng những vật n้อม, vật giả vờ (placebo). Placebo là những vật mà bê ngoài trong giống hệt vật thí nghiệm, chỉ khác nhau về bản chất. Placebo thường được dùng trong các nghiên cứu về thuốc.

3.7. TĂNG ĐỘ CHÍNH XÁC CỦA ƯỚC TÍNH

3.7.1. Lặp lại

Nhìn chung, số lượng đơn vị thí nghiệm càng lớn thì độ chính xác của ước tính càng cao và càng có nhiều cơ hội để phát hiện được ảnh hưởng của công thức thí nghiệm nếu nó tồn tại. Chi tiết về xác định dung lượng mẫu tối ưu được trình bày ở chương 4 và chương 5.

Lặp lại tức là tiến hành thu thập cùng một kiểu số liệu nhiều lần trên cùng một động vật hay cùng một đơn vị thí nghiệm. Bằng cách này ta có thể phân tách được biến động do sinh học gây ra hay do tác động của công thức thí nghiệm.

3.7.2. Kỹ thuật khôi

Có thể sử dụng kỹ thuật nhóm đơn vị thí nghiệm như một công cụ hỗ trợ để giảm biến động trong quá trình so sánh. Tạo ra các nhóm động vật (khôi) tương đối đồng đều nhau, như vậy sự biến động ngẫu nhiên trong mỗi khôi sẽ bé hơn giữa các khôi. Tiến hành ngẫu nhiên hoá trong từng khôi. Trong quá trình phân tích số liệu, có thể phân tách được sự biến động do công thức thí nghiệm gây ra với biến động do khôi gây ra. Với cách tiếp cận theo kỹ thuật khôi ta sẽ có một ước tính chính xác hơn.

Đối với kỹ thuật khôi có 2 mô hình thiết kế thí nghiệm: 1) khôi ngẫu nhiên đầy đủ, khi trong mỗi khôi bô trí đầy đủ tất cả các công thức thí nghiệm và 2) khôi ngẫu nhiên không đầy đủ, khi trong mỗi khôi không có đầy đủ các công thức thí nghiệm.

3.7.3. Kỹ thuật cặp

Kỹ thuật cặp được đề cập khi ta xem xét trường hợp chỉ có 2 công thức thí nghiệm (2 nhóm) và 2 nhóm này có mối liên hệ với nhau. Nếu các quan sát trong 2 nhóm tạo

thành cặp hoặc một cá thể tham gia ở cả 2 nhóm thì các quan sát ở 2 nhóm phải bằng nhau. Với kỹ thuật cặp, so sánh các công thức thí nghiệm với nhau được thực hiện trong từng cặp. Sự biến động trong từng cặp bao giờ cũng bé hơn giữa các cá thể không cùng cặp, như vậy ước tính sẽ chính xác hơn. Có các kiểu cặp như sau:

- 1 Cặp tự tạo - mỗi động vật tham gia cả 2 công thức thí nghiệm.
- 2 Cặp tự nhiên - động vật sinh đôi hoặc nhân bản.
- 3 Cặp nhân tạo – tạo ra cặp với các tiêu chí lựa chọn tương đối đồng nhất, ví dụ đồng nhất về tuổi, khối lượng, chỉ tiêu sinh lý, sinh hoá...

3.8. DUNG LƯỢNG MẪU CẦN THIẾT

Cần bao nhiêu động vật thí nghiệm, bao nhiêu khối, bao nhiêu ô lớn, bao nhiêu ô nhỏ? Đây là một câu hỏi thực sự khó. Chúng ta xét một số cách tiếp cận sau:

Số động vật thí nghiệm phải đủ sao cho các đặc tính riêng biệt của từng cá thể không làm ảnh hưởng đến kết quả thí nghiệm. Nếu số động vật trong thí nghiệm quá ít thì độ tin cậy của kết quả thu được từ thí nghiệm sẽ không cao. Ngược lại, nếu số động vật quá nhiều thì có thể gây lãng phí. Để đạt được độ chính xác cao không phải lúc nào cũng cần số lượng động vật thí nghiệm quá lớn. Nếu quá nhiều động vật tham gia thí nghiệm thì có thể gây ra nhiều khó khăn trong quá trình theo dõi từng cá thể, khó khăn khi chúng ta muốn tạo ra các điều kiện đồng nhất của thí nghiệm cho mọi cá thể ví dụ như khi cho động vật ăn... những khó khăn đó đã làm giảm độ chính xác về mặt kỹ thuật của thí nghiệm.

Dung lượng mẫu cần thiết còn phụ thuộc vào chất lượng của động vật tham gia thí nghiệm. Động vật tham gia thí nghiệm có độ đồng đều cao thì số lượng giảm xuống và ngược lại. Độ tuổi của vật nuôi cũng đóng vai trò quan trọng trong quá trình chọn dung lượng mẫu. Động vật càng non thì số lượng cần phải tăng lên và ngược lại, bởi vì đối với loại động vật này mức độ biến động rất lớn (cả về mặt sinh lý và ngoại hình). Ngoài ra, dung lượng mẫu còn phụ thuộc vào từng loại vật nuôi; mỗi loại vật nuôi có những đặc điểm riêng vì vậy trong quá trình thiết kế thí nghiệm cũng phải chú ý đến yếu tố này. Cuối cùng, kết quả mong đợi của thí nghiệm (sự chênh lệch giữa các công thức thí nghiệm) cũng ảnh hưởng rất nhiều đến dung lượng mẫu.

Có thể phác sơ qua các yếu tố ảnh hưởng đến dung lượng mẫu như sau:

Yếu tố ảnh hưởng	Dung lượng mẫu		
	bé	trung bình	lớn
Biến động trong đàn	nhỏ	→	lớn
Đối tượng nghiên cứu	đại gia súc	→	gia cầm
Giai đoạn nghiên cứu	đầu	→	cuối
Loại đê tài	thức ăn	giống	phòng bệnh
Phương tiện	bằng tay	→	có máy móc
Nhân lực và vật lực	hạn chế	→	nhiều

Trên đây là các tiêu chí để làm cơ sở quyết định chọn dung lượng mẫu. Bên cạnh đó, để xác định được số lượng động vật thí nghiệm cần thiết có thể dựa vào các tiêu chí sau:

3.8.1. Số công thức thí nghiệm

Cách tiếp cận thứ nhất để xác định được dung lượng mẫu cần thiết đó là dựa vào:

- 1 Số công thức thí nghiệm (a).
- 2 Mức độ đồng đều của tính trạng cần nghiên cứu (σ^2).
- 3 Sai lầm loại I (α)
- 4 Sai lầm loại II (β).

Thông thường một công trình nghiên cứu chấp nhận sai sót loại I khoảng 1% hay 5% (tức $\alpha = 0,01$ hay $0,05$) và xác suất sai sót loại II khoảng $\beta = 0,1$ đến $0,2$ (tức power = $0,8 - 0,9$).

- 5 Sai khác mong đợi (sự khác biệt muốn phát hiện) hoặc chênh lệch bé nhất giữa 2 giá trị trung bình bất kỳ để phát hiện sự sai khác nếu có (Δ).

Đối với trường hợp ước tính một giá trị trung bình

Dung lượng mẫu cần thiết để giá trị trung bình cộng ước tính khác μ không quá Δ khi có phân phối chuẩn $N(\mu, \sigma^2)$ và mức tin cậy $P = 1 - \alpha$ dựa vào công thức sau:

$$n \geq \frac{C \times \sigma^2}{\Delta^2}$$

Trong đó: C là hằng số liên quan giữa α và β ; $C = (Z_{1-\alpha/2} + Z_{1-\beta})^2$; bảng 3.1.

Bảng 3.1. Bảng tham chiếu hằng số C liên quan giữa α và β

α	$\beta = 0,2$ (power = 0,8)	$\beta = 0,1$ (power = 0,9)	$\beta = 0,05$ (power = 0,95)
0,1	6,18	8,56	10,82
0,05	7,85	10,51	12,99
0,01	11,68	14,88	17,81

Ví dụ 3.3: Cần quan sát bao nhiêu bò sữa để ước tính được năng suất trong chu kỳ tiết sữa 305 ngày với mức độ tin cậy 95% nằm trong khoảng ± 75 kg so với giá trị thực của quần thể. Biết rằng sản lượng sữa có phân bố chuẩn $\sigma = 500$ kg.

$$\text{Cần thiết: } n \geq \frac{C \times \sigma^2}{\Delta^2} = \frac{7,85 \times 500^2}{75^2} = 348,88$$

Như vậy cần ít nhất 349 bò sữa để thoả mãn điều kiện bài toán.

Đối với trường hợp ước tính một tỷ lệ

Dung lượng mẫu cần thiết để tỷ lệ ước tính \hat{p} khác không quá d so với tỷ lệ thực π . Nếu biết tỷ lệ hiện hành p (prevalance) và kiểm định ở mức tin cậy $P = 1 - \alpha$ dựa vào công thức sau:

$$n \geq \frac{(z_{1-\alpha/2})^2 \times p(1-p)}{\Delta^2}$$

Lưu ý: Tỷ lệ hiện hành p có thể tìm được thông qua các tài liệu, các nghiên cứu trước hoặc xuất phát từ kinh nghiệm và sự hiểu biết của người nghiên cứu. Nếu khi tiến hành thí nghiệm không có thông tin về tỷ lệ lưu hành, ta sẽ chọn $p = 0,5$. Khi đó:

$$n \geq \frac{(z_{1-\alpha/2})^2}{4\Delta^2}$$

Ví dụ 3.4: Cần dung lượng mẫu bao nhiêu để xác định tỷ lệ hiện nhiễm một loại vi khuẩn trên thân thịt lợn ở một lò mổ với ước tính chênh lệch không quá 5%. Biết rằng tỷ lệ hiện hành $p = 0,2$ và kiểm định ở mức tin cậy 95%.

$$\text{Cần thiết } n \geq \frac{(z_{1-\alpha/2})^2 \times p(1-p)}{\Delta^2} = \frac{1,96^2 \times 0,2 \times (1-0,2)}{0,05^2} = 245,86$$

Như vậy cần khảo sát ít nhất 246 thân thịt.

Đối với trường hợp so sánh 2 giá trị trung bình

Dung lượng mẫu cần thiết (đối với mỗi công thức thí nghiệm) để phát hiện được sự sai khác nếu chênh lệch giữa 2 giá trị trung bình là Δ , sai lầm loại I và loại II ở mức tương ứng là α và β . Giả sử số liệu có phân bố chuẩn. Phương sai của tính trạng nghiên cứu là σ^2 .

$$n \geq \frac{2C\sigma^2}{\Delta^2}$$

Trong đó: C là hằng số liên quan giữa α và β ; $C = (Z_{1-\alpha/2} + Z_{1-\beta})^2$; bảng 3.1.

Ví dụ 3.5: Muốn thiết kế một thí nghiệm để so sánh sản lượng sữa của dê Bách Thảo ở 2 công thức thí nghiệm với yêu cầu $\alpha = 0,05$; $\beta = 0,2$; chênh lệch mong đợi 30 kg sữa biệt $\sigma = 50$ kg.

$$\text{Cần thiết } n \geq \frac{2C\sigma^2}{\Delta^2} = \frac{2(50)^2}{30^2} \approx 43,55$$

Như vậy cần ít nhất 44 dê cho mỗi công thức thí nghiệm.

Đối với trường hợp so sánh hai tỷ lệ

Với các nghiên cứu tiền cứu (Cohort studies), dung lượng mẫu cần thiết để so sánh 2 tỷ lệ là:

$$n_1 \geq \frac{1}{4} \frac{\left(z_{1-\alpha/2} \sqrt{(r+1)pq} + z_{1-\beta} \sqrt{rp_1q_1 + p_1q_1} \right)^2}{r(p_1 - p_2)^2} \times \left(1 + \sqrt{1 + \frac{2(r+1)}{\left(\frac{\left(z_{1-\alpha/2} \sqrt{(r+1)pq} + z_{1-\beta} \sqrt{rp_1q_1 + p_1q_1} \right)^2}{r(p_1 - p_2)^2} \right)^2 r|p_1 - p_2|}} \right)^2$$

Trong đó:

n_1 = dung lượng mẫu tối thiểu cần thiết cho nhóm thứ nhất (không phơi nhiễm).

n_2 = dung lượng mẫu tối thiểu cần thiết cho nhóm thứ hai (có phơi nhiễm).

$$r = n_1 / n_2.$$

p_1 = tỷ lệ mắc bệnh hiện hành ở quần thể thứ 1.

p_2 = tỷ lệ mắc bệnh dự đoán ở quần thể thứ 2.

$$\bar{p} = \frac{p_1 + r p_2}{r+1}; \bar{q} = 1 - \bar{p}; q = 1 - p.$$

$Z_{(1-\alpha/2)}$ = Giá trị z ở mức tương ứng $1-\alpha/2$ (α – xác suất mắc sai lầm loại I).

$Z_{(1-\beta)}$ = Giá trị z ở mức tương ứng $1-\beta$ (β – xác suất mắc sai lầm loại II).

Ví dụ 3.6a: Một tiến cứu được tiến hành để nghiên cứu tỷ lệ tổn thương núm vú ở bò sữa giữa hệ thống vắt sữa tự động (A) và hệ thống bình tay (B). Thời gian nghiên cứu được tiến hành trong 12 tháng với dự đoán tỷ lệ tổn thương ở hệ thống B là 34,5% ($p_1 = 0,345$); $\alpha = 0,05$; $\beta = 0,20$; $n_1 = n_2$. Biết rằng tỷ lệ tổn thương ở hệ thống vắt sữa tự động là 15% ($p_2 = 0,15$). Hãy tính dung lượng mẫu cần thiết đối với một nhóm để thoả mãn điều kiện bài toán.

Cần thiết:

$$n_1 \geq \frac{1}{4} \frac{\left(1,96\sqrt{2 \times 0,2475 \times 0,7525} + 0,84\sqrt{2 \times 0,345 \times 0,655}\right)^2}{(0,345 - 0,15)^2} \times \left(1 + \sqrt{1 + \frac{4}{\left(\left(1,96\sqrt{2 \times 0,2475 \times 0,7525} + 0,84\sqrt{2 \times 0,345 \times 0,655}\right)^2\right)^2 |0,345 - 0,15|}}\right)^2 = 84,82$$

Như vậy cần ít nhất 85 bò sữa cho một nhóm.

Dung lượng mẫu cần thiết để so sánh 2 tỷ lệ (trong các nghiên cứu dịch tễ học thú y) có thể được tính bằng công thức sau:

$$n = \frac{\left(z_{(\alpha/2)}\sqrt{2p(1-p)} + z_{(1-\beta)}\sqrt{p_1(1-p_1) + p_2(1-p_2)}\right)^2}{\Delta^2}$$

n = dung lượng mẫu tối thiểu cần thiết cho một nhóm.

p_1 = tỷ lệ mắc bệnh hiện hành ở quần thể thứ 1.

p_2 = tỷ lệ mắc bệnh dự đoán ở quần thể thứ 2.

$$p = \frac{p_1 + p_2}{2}; q = 1 - p$$

$Z_{(1-\alpha/2)}$ = Giá trị z ở mức tương ứng $1-\alpha/2$ (α – xác suất mắc sai lầm loại I).

$Z_{(1-\beta)}$ = Giá trị z ở mức tương ứng $1-\beta$ (β – xác suất mắc sai lầm loại II).

Δ : Sai khác mong đợi (sự khác biệt muôn phát hiện); $\Delta = p_1 - p_2$.

Ví dụ 3.6b: Một thí nghiệm được thiết kế để đánh giá hiệu quả điều trị của một loại cao mật động vật trên bệnh lợn con phân trắng. Nhóm 1 được điều trị bằng cao mật động vật; nhóm 2 là nhóm đối chứng (không điều trị). Các nhà nghiên cứu giả thiết rằng tỷ lệ bệnh lợn con phân trắng ở nhóm 2 khoảng 10% và cao mật động vật có thể làm giảm tỷ lệ này xuống khoảng 6%. Nếu các nhà nghiên cứu muốn thử nghiệm giả thiết này với sai sót I ($\alpha = 0,01$) và power = 0,9, bao nhiêu lợn con cho nghiên cứu này?

$$n = \frac{(2,57 * \sqrt{2 * 0,08 * 0,92} + 1,28 * \sqrt{0,1 * 0,9 + 0,06 * 0,94})^2}{(0,04)^2} = 1366$$

Như vậy cần ít nhất **1366** lợn con cho một nhóm.

Trường hợp so sánh nhiều giá trị trung bình

Các trường hợp ước tính cỡ mẫu ở trên sử dụng phương pháp ước tính trực tiếp. Tuy nhiên, trường hợp so sánh nhiều giá trị trung bình sử dụng phương pháp ước tính gián tiếp để xác định dung lượng mẫu cần thiết. Cách xác định như sau:

+ Gọi số trung bình của g nhóm là $\mu_1, \mu_2, \dots, \mu_g$.

+ Tính trung bình chung: $\bar{\mu} = \sum_{i=1}^g \frac{\mu_i}{g}$.

+ Tính tổng bình phương: $SS = \sum_{i=1}^g (\mu_i - \bar{\mu})^2$.

+ Tính giá trị Lamda: $\lambda = \frac{SS}{(g-1)\sigma^2}$.

+ Tìm giá trị $F^* = F(\alpha, u, v)$, trong đó: $u = g - 1$ và $v = g(n - 1)$.

Thay các giá trị g, λ, F^* và dung lượng mẫu (n) để sao cho Z_β đáp ứng được yêu cầu độ mạnh của phép thử (Power) đạt tối thiểu 0,8 hoặc 0,9.

$$Z_\beta = \frac{1}{\sqrt{(g-1)(1+n\lambda)F^* + g(n-1)(1+2n\lambda)}} x \\ \left(\sqrt{g(n-1)[2(g-1)(1+n\lambda)^2 - (1+2n\lambda)]} - \sqrt{F^*(g-1)(1+n\lambda)(2g(n-1)-1)} \right)$$

Ví dụ 3.7: Thiết kế một thí nghiệm để so sánh tăng khối lượng (g) của gà ở 4 khẩu phần thức ăn (A, B, C, D). Các giá trị trung bình được chọn lần lượt là: $\mu_A = 79, \mu_B = 71, \mu_C = 80, \mu_D = 102$, với $\alpha = 0,05$ và $1 - \beta = 0,8$; biết $\sigma^2 = 35^2$. Cần bao nhiêu gà tham gia thí nghiệm này?

- Tính trung bình chung: $\mu = \sum_{i=1}^g \frac{\mu_i}{g} = \frac{1}{4}(79+71+80+102) = 83$

- Tính tổng bình phương:

$$SS = \sum_{i=1}^4 (\mu_i - \bar{\mu})^2 = (79 - 83)^2 + (71 - 83)^2 + (80 - 83)^2 + (102 - 83)^2 = 530$$

+ Tìm giá trị $F^* = F(0,05, u, v)$, trong đó: $u = 4 - 1 = 3$ và $v = 4(n - 1)$.

+ Thay lần lượt từng giá trị dung lượng mẫu để thoả mãn điều kiện Z_β đáp ứng được yêu cầu độ mạnh của phép thử (Power) đạt tối thiểu 0,8 hoặc 0,9.

+ $n = 1 \Rightarrow F^* = F(0,05, 3, 0) \Rightarrow$ không xác định được giá trị $F^* \Rightarrow n = 1$ không thoả mãn yêu cầu.

$$+ n = 2 \Rightarrow F^* = F(0,05, 3, 4) = 6,591$$

$$z_\beta = \frac{1}{\sqrt{(4-1)(1+2*0,144)*6,591+4(2-1)(1+2*2*0,144)}} x \\ \left(\sqrt{4(2-1)\left[2(4-1)(1+2*0,144)^2-(1+2*2*0,144)\right]} - \sqrt{6,591(4-1)(1+2*0,144)(2*4(2-1)-1)} \right)$$

$z_\beta = -1,34159 \Rightarrow \text{Power} = 0,0898 (8,98\%) < 0,8 (80\%) \Rightarrow n = 2$ không thoả mãn yêu cầu.

$$+ n = 3, 4, 5, \dots$$

$$+ n = 10 \Rightarrow F^* = F(0,05, 3, 36) = 2,866.$$

$$z_\beta = \frac{1}{\sqrt{3*(1+10*0,144)*2,866+4*9*(1+2*10*0,144)}} x \\ \left(\sqrt{4*9*\left[2*3*(1+10*0,144)^2-(1+2*10*0,144)\right]} - \sqrt{2,866*3*(1+10*0,144)(2*4*9*-1)} \right)$$

$z_\beta = -0,37255 \Rightarrow \text{Power} = 0,3547 (35,47\%) < 0,8 (80\%) \Rightarrow n = 10$ không thoả mãn yêu cầu.

$$+ n = 11, 12, 13, \dots$$

$$+ n = 20 \Rightarrow F^* = F(0,05, 3, 76) = 2,724.$$

$$z_\beta = \frac{1}{\sqrt{3*(1+20*0,144)*2,724+4*19*(1+2*20*0,144)}} x \\ \left(\sqrt{4*19*\left[2*3*(1+20*0,144)^2-(1+2*20*0,144)\right]} - \sqrt{2,724*3*(1+20*0,144)(2*4*19*-1)} \right)$$

$z_\beta = 0,4510 \Rightarrow \text{Power} = 0,6740 (67,40\%) < 0,8 (80\%) \Rightarrow n = 20$ không thoả mãn yêu cầu.

$$+ n = 21, 22, 23, \dots$$

+ $n = 25 \Rightarrow F^* = F(0,05, 3, 96) = 2,699$.

$$z_{\beta} = \frac{1}{\sqrt{3 * (1 + 25 * 0,144) * 2,699 + 4 * 24 * (1 + 2 * 25 * 0,144)}} x \\ \left(\sqrt{4 * 24 * \left[2(4-1)(1+25*0,144)^2 - (1+2*25*0,144) \right]} - \sqrt{2,699 * 3 * (1+25*0,144)(2*4*24-1)} \right)$$

$z_{\beta} = 0,78326 \Rightarrow \text{Power} = 0,78326 (78,32\%) < 0,8 (80\%) \Rightarrow n = 25 \text{ không thoả mãn yêu cầu.}$

+ $n = 26 \Rightarrow F^* = F(0,05, 3, 100) = 2,695$.

$$z_{\beta} = \frac{1}{\sqrt{3 * (1 + 26 * 0,144) * 2,695 + 4 * 25 * (1 + 2 * 26 * 0,144)}} x \\ \left(\sqrt{4 * 25 * \left[2 * 3 * (1+26*0,144)^2 - (1+2*26*0,144) \right]} - \sqrt{2,695 * 3 * (1+26*0,144)(2*4*25*-1)} \right)$$

$z_{\beta} = 0,84564 \Rightarrow \text{Power} = 0,80112 (80,11\%) > 0,8 (80\%) \Rightarrow n = 26 \text{ thoả mãn yêu cầu.}$

Như vậy, dung lượng mẫu cho một nhóm ($n = 26$) để sao cho $Z\beta$ đáp ứng được yêu cầu độ mạnh của phép thử (Power) đạt tối thiểu 0,8. Tổng số gà tham gia thí nghiệm này là: $4 \times 26 = 104$ con.

Tuy nhiên, yêu cầu độ mạnh của phép thử đạt tối thiểu 0,9, dung lượng mẫu cho một nhóm ($n = 34$) và tổng số gà tham gia thí nghiệm này: $4 \times 34 = 136$ con (Bảng 3.2)

Bảng 3.2. Bảng tham chiếu dung lượng mẫu tương ứng với độ mạnh của phép thử

n	u	v	F*	Zbeta	Power
1	3	0	#NUM!	#NUM!	#NUM!
2	3	4	6,591382	-1,34181	0,089829
3	3	8	4,066181	-1,23865	0,107738
4	3	12	3,490295	-1,09323	0,137147
5	3	16	3,238872	-0,95224	0,170487
6	3	20	3,098391	-0,8207	0,205908
7	3	24	3,008787	-0,69824	0,242513
8	3	28	2,946685	-0,58369	0,279714
9	3	32	2,90112	-0,4759	0,317074
10	3	36	2,866266	-0,37388	0,354246
11	3	40	2,838745	-0,27686	0,390946
12	3	44	2,816466	-0,18417	0,426941
13	3	48	2,798061	-0,09529	0,462042
14	3	52	2,7826	-0,00979	0,496094
15	3	56	2,769431	0,072694	0,528975
16	3	60	2,758078	0,152465	0,56059
17	3	64	2,748191	0,22978	0,590869
18	3	68	2,739502	0,304859	0,619763

n	u	v	F*	Zbeta	Power
19	3	72	2,731807	0,377889	0,647244
20	3	76	2,724944	0,449036	0,673297
21	3	80	2,718785	0,518442	0,697925
22	3	84	2,713227	0,586234	0,721141
23	3	88	2,708186	0,652523	0,742968
24	3	92	2,703594	0,717409	0,763439
25	3	96	2,699393	0,780979	0,782593
26	3	100	2,695534	0,843315	0,800474
27	3	104	2,691979	0,904488	0,817132
28	3	108	2,688691	0,964562	0,832618
29	3	112	2,685643	1,023598	0,846987
30	3	116	2,682809	1,081649	0,860296
31	3	120	2,680168	1,138765	0,8726
32	3	124	2,677699	1,194992	0,883955
33	3	128	2,675387	1,25037	0,894418
34	3	132	2,673218	1,304939	0,904043
35	3	136	2,671178	1,358733	0,912884
36	3	140	2,669256	1,411788	0,920994
37	3	144	2,667443	1,464132	0,928421
38	3	148	2,665729	1,515796	0,935215
39	3	152	2,664107	1,566805	0,94142
40	3	156	2,662569	1,617185	0,947081
41	3	160	2,661108	1,666958	0,952239
42	3	164	2,65972	1,716148	0,956933
43	3	168	2,658399	1,764775	0,961199
44	3	172	2,65714	1,812858	0,965073
45	3	176	2,655939	1,860416	0,968587
46	3	180	2,654792	1,907466	0,97177
47	3	184	2,653695	1,954024	0,974651
48	3	188	2,652646	2,000106	0,977256
49	3	192	2,65164	2,045727	0,979608
50	3	196	2,650677	2,090901	0,981732

Nếu ảnh hưởng của công thức thí nghiệm rất ít, muốn phát hiện được sự ảnh hưởng này đòi hỏi dung lượng mẫu lớn. Bên cạnh đó các giá trị sai lầm loại I và độ mạnh của phép thử tương ứng là α và $1 - \beta$ cũng ảnh hưởng rất nhiều đến dung lượng mẫu cần thiết.

Dung lượng mẫu cần thiết đối với mỗi công thức thí nghiệm (n) để phát hiện sự sai khác nếu có khi chênh lệch bé nhất giữa 2 giá trị trung bình bất kỳ là d , số công thức thí nghiệm là a và phương sai của tính trạng nghiên cứu là σ^2 được dựa trên công thức dưới đây:

$$\phi^2 = \frac{nd^2}{2a\sigma^2}$$

Tham số ϕ^2 sẽ được đề cập chi tiết ở chương 4. Ở đây ta sẽ sử dụng đường cong cho sẵn ở phần phụ lục để tìm dung lượng mẫu cần thiết.

Ví dụ 3.8: Nghiên cứu tăng khối lượng (g/ngày) của lợn nuôi vỗ béo đến 5 tháng tuổi ở 3 công thức thí nghiệm. Hãy xác định dung lượng mẫu (n) cần thiết để phát hiện sự sai khác giữa các công thức thí nghiệm nếu có. Biết rằng sự chênh lệch giữa 2 giá trị trung bình lúc kết thúc thí nghiệm là 40g, tăng khối lượng có phân bố chuẩn với phương sai $\sigma^2 = 480$.

Sử dụng công thức nêu trên cùng với các đường cong ở phần phụ lục ta có thể tìm ra dung lượng mẫu cần thiết ở các mức chính xác tương ứng:

Nếu $\alpha = 0,05$; $1-\beta = 0,80 \rightarrow n = 7$;

Nếu $\alpha = 0,05$; $1-\beta = 0,90 \rightarrow n = 9$;

Nếu $\alpha = 0,01$; $1-\beta = 0,80 \rightarrow n = 10$;

Nếu $\alpha = 0,01$; $1-\beta = 0,90 \rightarrow n = 12$.

Một số vấn đề về ước tính dung lượng mẫu

Điều chỉnh dung lượng mẫu dự phòng rủi ro

Các trường hợp ước tính cỡ mẫu ở trên đều dựa trên giả định thí nghiệm được thực hiện thuận lợi, suôn sẻ và không gặp rủi ro. Tuy nhiên, các thí nghiệm chăn nuôi trong thực tế sau khi chọn được động vật để tiến hành thí nghiệm, một số động vật không đáp ứng được yêu cầu của thí nghiệm (sức khoẻ kém, bị bệnh, thể trạng yếu), hiện tượng này chiếm tỷ lệ khoảng từ 10 – 30% tùy theo đối tượng vật nuôi và độ tuổi tiến hành thí nghiệm. Do vậy, khi ước tính dung lượng mẫu cần phải tính đến trường hợp phải dự phòng khi gặp rủi ro và điều chỉnh dung lượng mẫu cho phù hợp:

- Nếu dung lượng mẫu theo lý thuyết: n .
- Tỷ lệ dự phòng rủi ro: q .
- Dung lượng mẫu thực tế: $n/(1-q)$.

Ví dụ 3.9: Thiết kế một thí nghiệm để so sánh tăng khối lượng (g) của gà ở 4 khẩu phần thức ăn (A, B, C, D). Các giá trị trung bình được chọn lần lượt là: $\mu_A = 79$, $\mu_B = 71$, $\mu_C = 80$, $\mu_D = 102$, với $\alpha = 0,05$ và $1-\beta = 0,8$; biết $\sigma^2 = 35^2$. Theo lý thuyết cần 104 con gà, nếu tỷ lệ dự phòng là 20%, số gà thực tế cần lấy: $104/(1-0,2) = 130$ con.

Điều chỉnh dung lượng mẫu cho hiện tượng mất cân đối giữa hai nhóm

Các trường hợp ước tính cỡ mẫu ở trên đều dựa trên giả định thí nghiệm được thực hiện với hai nhóm, dung lượng mẫu của hai nhóm bằng nhau (ước tính dung lượng mẫu cho một nhóm và nhân đôi để có dung lượng mẫu của nhóm thứ hai). Tuy nhiên, nhiều thí nghiệm chăn nuôi thú y (đặc biệt đối với nghiên cứu bệnh chứng) rất khó để có thể tìm được nhóm bệnh bằng với nhóm đối chứng. Một khác, việc để dung lượng mẫu của nhóm đối chứng (nhóm không được điều trị) bằng với nhóm bệnh (nhóm được điều

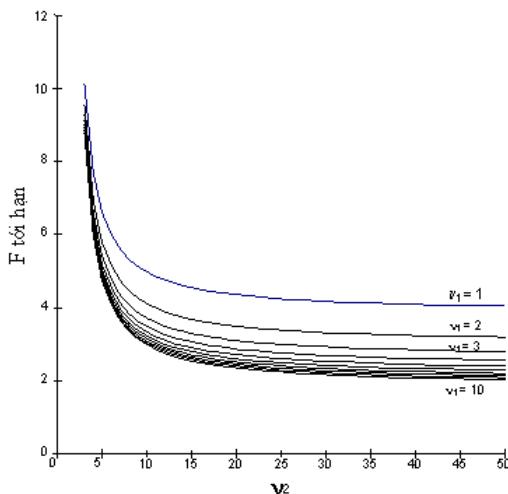
tri) sẽ gây ảnh hưởng đến sản xuất và thiệt hại về kinh tế rất lớn. Do vậy, khi ước tính dung lượng mẫu cần phải tính đến trường hợp mất cân đối giữa hai nhóm và điều chỉnh dung lượng mẫu cho phù hợp:

- Nếu dung lượng mẫu theo lý thuyết: n .
- Tỷ số cỡ mẫu giữa nhóm 1 và nhóm 2 là k .
- + Cỡ mẫu thực tế: $n(1+k)^2/4k$

Ví dụ 3.10: Một thí nghiệm được thiết kế để đánh giá hiệu quả điều trị của một loại cao mật động vật trên bệnh lợn con phân trắng. Nhóm 1 được điều trị bằng cao mật động vật; nhóm 2 là nhóm đối chứng (không điều trị). Dung lượng mẫu theo lý thuyết ($n = 2722$) và dung lượng mẫu nhóm 1 lớn hơn nhóm 2 là 1,5 lần ($k = 1,5$) dung lượng mẫu thực tế: $2722 * (1+1,5)^2 / (4 * 1,5) = 2835$ con.

3.8.2. Bậc tự do của sai số ngẫu nhiên

Giả sử so sánh hai hoặc nhiều công thức thí nghiệm với nhau ta mong muốn bậc tự do của sai số ngẫu nhiên ≥ 20 bởi vì với bậc tự do của sai số ngẫu nhiên bé giá trị tới hạn của F rất lớn nhưng nó sẽ giảm rất nhanh khi bậc tự do này tăng lên. Khi bậc tự do sai số ngẫu nhiên lớn hơn 20 thì giá trị F giảm rất ít. Trong đồ thị dưới đây, v_1 và v_2 tương ứng với bậc tự do của công thức thí nghiệm và bậc tự do của sai số ngẫu nhiên sẽ minh họa điều này.



Đồ thị. Giá trị tối hạn của phân phối F với bậc tự do v_1, v_2 và $\alpha = 0,05$

Sử dụng quy tắc bậc tự do tối thiểu để tính dung lượng mẫu cần thiết cho các ví dụ sau:

Ví dụ 3.11: Thiết kế thí nghiệm kiểu hoàn toàn ngẫu nhiên với số công thức thí nghiệm $a = 5$. Cần bao nhiêu đơn vị thí nghiệm cho một công thức thí nghiệm?

Bậc tự do của sai số ngẫu nhiên $df = (r - 1) \times a$.

Ta cần có $(r - 1) \times a \geq 20$, như vậy $r \geq 5$. Cần ít nhất 5 đơn vị thí nghiệm.

Ví dụ 3.12: Thiết kế thí nghiệm kiểu khói ngẫu nhiên đầy đủ với số công thức thí nghiệm $a = 5$. Cần bao nhiêu khói (b) ?

Bậc tự do của sai số ngẫu nhiên $df = (b - 1) \times (a - 1)$.

Ta cần có $(b - 1) \times 4 \geq 20$, như vậy $b \geq 6$. Vì vậy cần ít nhất 6 khói.

Điều này chứng tỏ rằng khi dung lượng mẫu tăng lên sẽ cho ta có kết luận chính xác hơn. Tuy nhiên, đồ thị trên cho ta thấy khi bậc tự do của sai số ngẫu nhiên lớn hơn 40 thì giá trị F có thay đổi không đáng kể.

Ngoài các cách tiếp cận nêu trên, các nhà nghiên cứu cũng đưa ra các nguyên tắc khác nhau để dựa vào nó mà có thể tìm ra dung lượng mẫu phù hợp:

Trong nghiên cứu về đại gia súc, Preston (1995) cho rằng số động vật trong một công thức thí nghiệm không được ít hơn 3 và bậc tự do của sai số ngẫu nhiên ít nhất là 15.

Trong các nghiên cứu về đại gia súc và lợn, Ovesianhnicov (1976) khuyến cáo số động vật trong một công thức thí nghiệm ít nhất là 6 và thuận lợi hơn nếu con số này là 12.

3.8.3. Phương pháp chọn mẫu

Các phương pháp ước tính dung lượng mẫu ở trên mới chỉ đáp ứng được về mặt số lượng. Bên cạnh về mặt số lượng, việc thực hiện đúng các kỹ thuật lấy mẫu đóng vai trò quan trọng trong thiết kế thí nghiệm để đảm bảo nguyên tắc ngẫu nhiên, đại diện cho quần thể và khách quan. Các phương pháp chọn mẫu được sử dụng như phương pháp chọn mẫu ngẫu nhiên đơn, phương pháp chọn mẫu ngẫu nhiên theo hệ thống, phương pháp phân tầng và phương pháp theo chùm.

Chọn mẫu ngẫu nhiên đơn (simple random sampling)

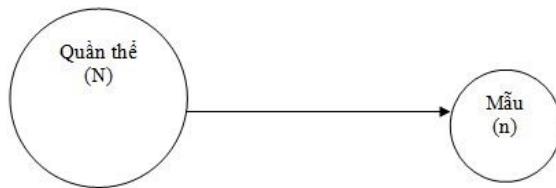
Một tập hợp con các cá thể (mẫu) được lựa chọn từ một tập hợp lớn hơn (quần thể), trong đó mỗi cá thể được chọn ngẫu nhiên hay mỗi một cá thể đều có cùng cơ hội (xác suất) được chọn. Quá trình thực hiện lấy mẫu như vậy được gọi là lấy mẫu ngẫu nhiên đơn. Như vậy, kỹ thuật lấy mẫu ngẫu nhiên đơn là kỹ thuật lấy mẫu mà tất cả các cá thể trong quần thể có xác suất để được chọn vào mẫu bằng nhau. Kỹ thuật chọn mẫu ngẫu nhiên đơn là dạng đơn giản nhất của mẫu xác suất. Đây là loại hình cơ bản của việc lấy mẫu và là một phần của phương pháp lấy mẫu phức tạp hơn.

Cách thực hiện chọn mẫu ngẫu nhiên đơn:

- Chuẩn bị khung mẫu: Danh sách chứa đựng tất cả các đơn vị mẫu và đánh số thứ tự từ 1 đến N.

- Xác định dung lượng mẫu cần thiết (n).

- Sử dụng cách bốc thăm hoặc bảng số ngẫu nhiên hoặc phần mềm máy tính để chọn mẫu.



Sơ đồ chọn mẫu ngẫu nhiên đơn

Ví dụ 3.13: chọn 100 con lợn Piétrain kháng stress trong quần thể gồm 1000 con. Chọn ngẫu nhiên đơn với xác suất: 0,1.

Sử dụng cách bốc thăm: Sử dụng danh sách của 1000 con lợn Piétrain kháng stress, mỗi con lợn nhận một số thứ tự (từ 1 tới 1000). Những số thứ tự này được viết trên một mảnh giấy nhỏ. Toàn bộ những mảnh giấy có số này được gấp lại bỏ vào một cái hộp, lắc kỹ để đảm bảo là ngẫu nhiên. Tiếp theo, 100 mảnh giấy được lấy ra và số của chúng được ghi lại. Những con lợn có những số này nằm trong mẫu nghiên cứu.

Sử dụng bảng số ngẫu nhiên: Sử dụng danh sách của 1000 con lợn Piétrain kháng stress, mỗi con lợn nhận một số thứ tự (từ 1 tới 1000). Lấy 100 số có 3 chữ số kế tiếp nhau trong bảng (3 chữ số đầu hoặc 3 chữ số cuối hoặc 3 chữ số giữa). Dùng bút chì, không nhìn vào bảng, chấm vào một điểm nào đó trong bảng bắt đầu từ điểm đó đọc lần lượt theo chiều từ trên xuống dưới và từ trái qua phải, ví dụ được các số 440, 258, 632, 162, 832, 003, 760, 807, 613, 207,... Chọn ra 100 số có 3 chữ số (không lấy các ký tự 000, chỉ lấy ra một lần, không lấy các ký tự lặp lại); Như vậy ta đã có một mẫu 100 con lợn Piétrain kháng stress.

Chọn mẫu ngẫu nhiên theo hệ thống (systematic samples)

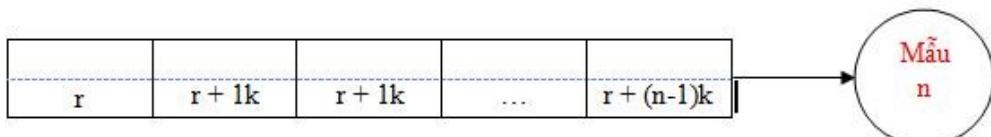
Một tập hợp con các cá thể (mẫu) được lựa chọn từ một tập hợp lớn hơn (quần thể), trong đó mỗi cá thể được chọn theo một trình tự hay một khoảng cách nhất định gọi là khoảng cách mẫu. Quá trình thực hiện lấy mẫu như vậy gọi là chọn mẫu ngẫu nhiên theo hệ thống.

Cách thực hiện chọn mẫu ngẫu nhiên theo hệ thống:

- Chuẩn bị khung mẫu: danh sách chứa đựng tất cả các đơn vị mẫu và đánh số thứ tự từ 1 đến N.

- Xác định dung lượng mẫu cần thiết (n).
- Xác định khoảng cách mẫu $k = N/n$ (N dung lượng quần thể và n dung lượng mẫu).
- Chọn ngẫu nhiên một số r bất kỳ trong khoảng từ 1 đến k .
- Chọn các đơn vị mẫu tương ứng với số thứ tự $r + ik$ ($i = 0$ đến $n-1$).

$k = N/n$ QUẦN THỂ.



Sơ đồ chọn mẫu ngẫu nhiên theo hệ thống

Ví dụ 3.14: lấy 246 mẫu thịt lợn (n) để kiểm tra tỷ lệ hiện nhiễm một loại vi khuẩn trên thân thịt lợn tại một lò mổ với công suất 5000 con/ngày (N).

- Xác định khoảng cách mẫu $k = 5000/246 = 20$.
- Chọn ngẫu nhiên một số r bất kỳ trong khoảng từ 1 đến 20, ví dụ chọn được số 5. Như vậy, mẫu thứ nhất được lấy ở thân thịt mang số 5 và cứ 20 thân thịt sẽ lấy một mẫu cho đến khi đủ 246 mẫu.
- Chọn các đơn vị mẫu tương ứng với số thứ tự $5 + i20$ ($i = 0$ đến 245).

Chọn mẫu phân tầng (simple random sampling)

Một tập hợp con các cá thể (mẫu) được lựa chọn từ một tập hợp lớn hơn gọi là quần thể, trong đó quần thể được phân chia thành các tầng khác nhau (nhóm khác biệt) theo một số tính chất nào đó, trong mỗi tầng chọn một số đơn vị nhất định bằng phương pháp ngẫu nhiên đơn hay chọn ngẫu nhiên theo hệ thống. Quá trình thực hiện lấy mẫu như vậy gọi là chọn mẫu phân tầng.

Chọn mẫu phân tầng được chia thành 2 loại: chọn mẫu phân tầng tỷ lệ và chọn mẫu phân tầng không tỷ lệ.

+ Chọn mẫu phân tầng tỷ lệ: số đơn vị mẫu của mỗi tầng được chọn tỷ lệ với kích thước của tầng.

+ Chọn mẫu phân tầng không tỷ lệ: Số đơn vị mẫu của mỗi tầng được chọn bằng nhau.

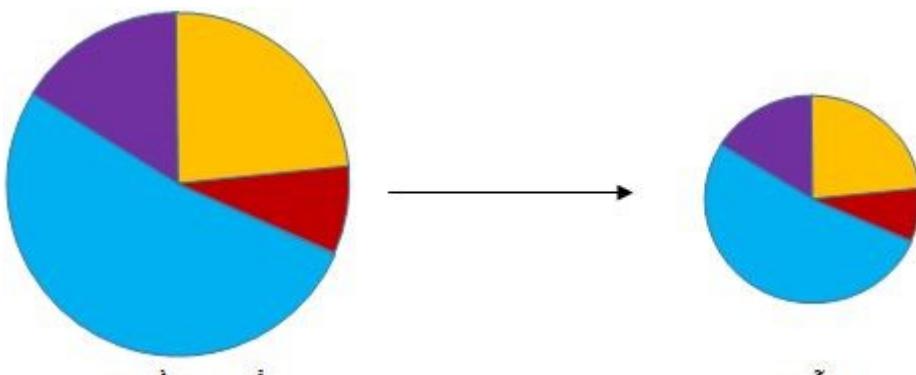
Chọn mẫu phân tầng tỷ lệ đảm bảo tính đại diện cho quần thể tốt hơn so với chọn mẫu phân tầng không tỷ lệ.

Cách thực hiện chọn mẫu phân tầng:

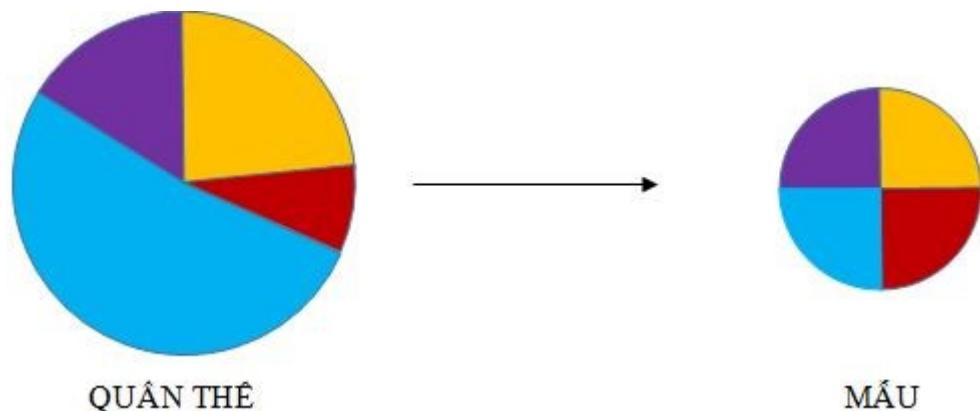
- Chia quần thể thành các tầng khác nhau (nhóm khác biệt) dựa vào một hoặc một vài đặc điểm nào đó như quy mô chăn nuôi, tính biệt, độ tuổi,....

- Xác định dung lượng mẫu cần thiết (n).

- Thực hiện chọn mẫu ngẫu nhiên đơn hoặc hệ thống trong từng tầng.



Sơ đồ chọn mẫu phân tầng tỷ lệ



Sơ đồ chọn mẫu phân tầng không tỷ lệ

Ví dụ 3.15: Một thí nghiệm được thực hiện nhằm đánh giá ảnh hưởng của quy mô đến năng suất sinh sản của lợn nái.

- Dựa trên số lượng lợn nái để chia thành các quy mô bao gồm:
 - + Quy mô: nhỏ hơn 100 nái.
 - + Quy mô: từ 100 – 200 nái.
 - + Quy mô từ 200 – 300 nái.
 - + Quy mô từ 300 – 500 nái.
 - + Quy mô lớn hơn 500 nái.
- Chọn mẫu phân tầng tỷ lệ: số đơn vị mẫu của mỗi quy mô được chọn tỷ lệ với kích thước của từng quy mô.
 - + Chọn mẫu phân tầng không tỷ lệ: số đơn vị mẫu của mỗi quy mô được chọn bằng 20% so với tổng số dung lượng mẫu cần lấy.

Chọn mẫu theo chùm (cluster sampling)

Một tập hợp con các cá thể (mẫu) được lựa chọn từ một tập hợp lớn hơn (quần thể), trong đó quần thể được phân chia thành các nhóm được gọi là các chùm. Việc lựa chọn ngẫu nhiên các chùm từ nhiều chùm trong quần thể nghiên cứu được gọi là chọn mẫu theo chùm. Đơn vị mẫu là các nhóm cá thể (chùm) chứ không phải cá thể.

Chọn mẫu theo chùm được chia thành 2 loại: chọn mẫu chùm một giai đoạn và chọn mẫu chùm hai giai đoạn.

- Chọn mẫu một giai đoạn: sử dụng phương pháp chọn ngẫu nhiên đơn để chọn ra các chùm từ nhiều chùm của quần thể và tất cả các cá thể trong các chùm được chọn ra đó tạo thành mẫu nghiên cứu.

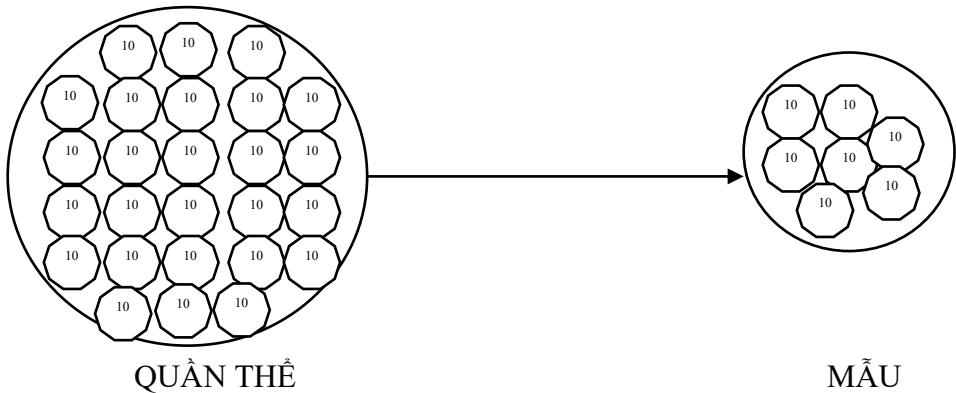
- Chọn mẫu chùm hai giai đoạn:

+ Giai đoạn 1: sử dụng phương pháp chọn ngẫu nhiên đơn để chọn ra các chùm từ nhiều chùm của quần thể nghiên cứu.

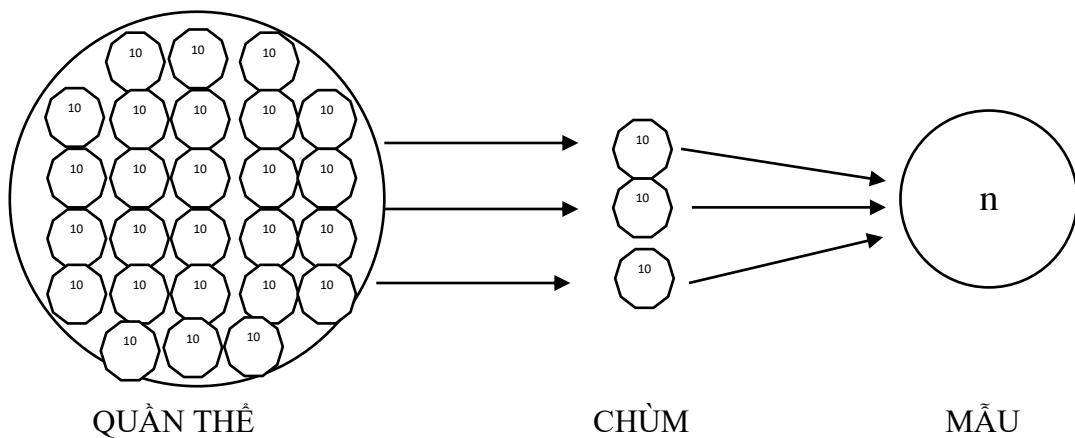
+ Giai đoạn 2: Lựa chọn ngẫu nhiên một số cá thể từ các chùm đã được chọn ở giai đoạn 1 và tập hợp tất cả các cá thể đó tạo thành mẫu nghiên cứu.

Cách thực hiện chọn mẫu chùm:

- Chia quần thể thành các nhóm (chùm) dựa vào một hoặc một vài đặc điểm nào đó như quy trong cùng một ô đẻ, cùng một ô chuồng, được nhốt cùng một lồng....
- Xây dựng khung mẫu: đánh số cho các chùm và lập danh sách tất cả các chùm trong quần thể
- Thực hiện chọn mẫu một giai đoạn hoặc hai giai đoạn để có được mẫu nghiên cứu.



Sơ đồ chọn mẫu chùm 1 giai đoạn



Sơ đồ chọn mẫu chùm 2 giai đoạn

Ví dụ 3.16: Một thí nghiệm được thực hiện nhằm đánh giá hiệu quả điều trị của một loại chế phẩm trên bệnh lợn con phân trắng. Để thực hiện thí nghiệm này cần chọn ra 2700 lợn con để chia về hai nhóm thí nghiệm và đối chứng.

- Xác định các chùm: 1 lợn nái + 10 lợn con = 1 chùm. Như vậy, thí nghiệm cần phải chọn ngẫu nhiên 270 lợn nái từ danh sách 1000 lợn nái.
- Xây dựng khung mẫu: lập danh sách số tai của toàn bộ lợn nái trong trại và đánh số thứ tự tương ứng.

- Thực hiện chọn mẫu chùm 1 giai đoạn: sử dụng phương pháp chọn mẫu ngẫu nhiên đơn chọn ra 270 lợn nái (10 lợn con/nái) và toàn bộ số lợn con của 270 lợn nái này tạo thành mẫu nghiên cứu.

- Thực hiện chọn mẫu chùm 2 giai đoạn:

+ Giai đoạn 1: sử dụng phương pháp chọn mẫu ngẫu nhiên đơn chọn ra 450 lợn nái (10 lợn con/nái) từ quần thể có 1000 lợn nái.

+ Giai đoạn 2: lựa chọn ngẫu nhiên 6 lợn con/chùm x 450 chùm = 2700 lợn con.

3.9. BÀI TẬP

3.9.1

Giả sử anh (chị) muốn ước tính tăng khối lượng trung bình ngày của lợn nuôi vỗ béo giết thịt từ 60 đến 180 ngày tuổi. Từ việc tổng quan tài liệu, anh (chị) tìm được độ lệch chuẩn của tính trạng này là 200 g/ngày. Với khoảng tin cậy 95% và giá trị ước tính nằm trong khoảng ± 50 g/ngày so với giá trị thực của quần thể, hãy tính dung lượng mẫu cần thiết

3.9.2

Một thí nghiệm được tiến hành nhằm nghiên cứu ảnh hưởng của việc bổ sung đồng đến tăng khối lượng của lợn. Chọn ra 20 lợn thí nghiệm giống Yorkshire ở 80 ngày tuổi (bắt đầu thí nghiệm) đồng đều và chia về 2 công thức thí nghiệm (đối chứng và thí nghiệm) hoàn toàn ngẫu nhiên. Khối lượng (kg) ở 210 ngày tuổi (kết thúc thí nghiệm) của 20 lợn nêu trên thu được như sau:

Đối chứng	120	125	130	131	120	115	121	135	115	128
Thí nghiệm	135	131	140	135	130	125	139	119	121	134

Theo phân loại, đây là loại thí nghiệm nào? Cho biết yếu tố và số công thức thí nghiệm. Nếu anh (chị) là người thiết kế thí nghiệm này, số động vật cần thiết là bao nhiêu.

3.9.3

Bệnh East Coast Fever (ECF) gây ra tỷ lệ chết ở vật nuôi là 50%. Sử dụng một loại vắc xin với mong muốn có thể bảo vệ được 95% vật nuôi. Với mức độ tin cậy là 95% và độ mạnh của phép thử là 90%, hãy xác định dung lượng mẫu cần thiết.

3.9.4

Tính số lượng cá thí nghiệm cần thiết cho mỗi bể để có thể phát hiện ra hiệu quả của việc dùng vắc xin. Giả sử rằng tỷ lệ cá nhiễm bệnh trọng trường hợp sử dụng vắc xin và không sử dụng vắc xin tương ứng là 10 và 30%, mức độ tin cậy là 0,95 và độ mạnh của phép thử là 0,80.

Chương 4

THIẾT KẾ THÍ NGHIỆM MỘT YẾU TỐ

Mục tiêu của chương này nhằm giới thiệu cho bạn đọc cách thiết kế thí nghiệm và phân tích kết quả đối với thí nghiệm một yếu tố; các ưu, nhược điểm của mô hình này và các giải pháp hạn chế những nhược điểm này. Ba mô hình được giới thiệu cho bạn đọc gồm: (1) Thí nghiệm hoàn toàn ngẫu nhiên, (2) Thí nghiệm khôi ngẫu nhiên đầy đủ và (3) Thí nghiệm ô vuông La tinh.

4.1. THÍ NGHIỆM HOÀN TOÀN NGÃU NHIÊN (Completely randomized Design - CRD)

4.1.1. Đặc điểm

Đây là phương pháp thiết kế thí nghiệm thông dụng nhất trong các nghiên cứu chăn nuôi - thú y. Thí nghiệm được thiết kế đơn giản và việc phân tích các dữ liệu của thí nghiệm cũng dễ dàng.

Đối với mô hình thí nghiệm này, các đơn vị thí nghiệm được bố trí một cách hoàn toàn ngẫu nhiên vào các công thức thí nghiệm, hay nói một cách khác, mỗi động vật thí nghiệm đều có cơ hội được phân vào một công thức thí nghiệm bất kỳ và chịu ảnh hưởng tác động của công thức thí nghiệm đó. Chính vì vậy, mô hình thí nghiệm này đòi hỏi các động vật thí nghiệm phải đồng đều. Mô hình này chỉ xem xét ảnh hưởng của một yếu tố, ví dụ nghiên cứu ảnh hưởng của thức ăn đến tăng khối lượng, tồn dư thuốc kháng sinh trong cơ thể vật nuôi..., các yếu tố còn lại được cho là không có sai khác, ví dụ tất cả các động vật được chọn có cùng một lứa tuổi, tất cả các trại đều sử dụng các thức ăn như nhau...

Với những yêu cầu nêu trên, trong lĩnh vực chăn nuôi và thú y, mô hình này chỉ thực hiện có hiệu quả khi động vật có tính đồng đều cao và các điều kiện phi thí nghiệm được kiểm soát một cách dễ dàng và có tính ổn định cao.

4.1.2. Chất lượng động vật

Động vật thí nghiệm đòi hỏi phải có sự đồng đều cao, vì vậy trong quá trình chọn động vật thí nghiệm, cần phải lưu ý đến các yếu tố như: giống, nguồn gốc, giới tính, thành tích của bố mẹ...

Chọn động vật cùng một giống. Động vật được chọn ra phải tiêu biểu cho giống đó, không quá khác biệt về ngoại hình và đặc điểm sinh lý. Để đạt được sự đồng đều cao, chọn những động vật là anh em ruột, nửa ruột thịt hoặc những động vật có quan hệ họ hàng trong cùng một dòng, một gia đình. Với thí nghiệm bố trí theo cặp tốt nhất dùng những động vật sinh đôi cùng trứng. Tuy nhiên trong thực tế, xác định được 2

động vật sinh đôi cùng trứng là phúc tạp và tốn kém. Có thể chọn những động vật không cùng dòng, họ nhưng có ngoại hình tương đối đồng đều và đặc tính ổn định.

Để có động vật đồng đều, chỉ chọn những động vật cùng tính biệt, đồng đều theo lứa tuổi, mức độ tăng trưởng, cùng thể chất, tình trạng sức khoẻ... Trong một số trường hợp cần thiết tiến hành những nghiên cứu kiểm tra một số chỉ tiêu hoá sinh, sinh lý.

4.1.3. Dung lượng mẫu cần thiết

Một trong những yếu tố quan trọng trong quá trình thiết kế thí nghiệm là xác định số đơn vị thí nghiệm cần thiết (đơn vị thí nghiệm là một động vật hoặc là một nhóm động vật). Tăng số lượng sẽ làm tăng độ chính xác của ước tính, tuy nhiên khi số lượng tăng sẽ đòi hỏi nhiều không gian, thời gian và nguồn lực. Số lượng có thể bị hạn chế bởi các yếu tố tài chính và điều kiện thực tế.

Sự sai khác, mặc dù có ý nghĩa thống kê, nhưng có thể không có ý nghĩa thực tiễn. Ví dụ, thí nghiệm so sánh tăng khối lượng của lợn ở 2 khẩu phần. Sự chênh lệch về tăng khối lượng trung bình ngày giữa 2 khẩu phần vài g không có ý nghĩa về mặt thực tiễn cũng không có ý nghĩa về kinh tế; mặc dù đây là một thí nghiệm được thiết kế với quy mô lớn và sự sai khác này có ý nghĩa thống kê.

Đối với trường hợp thí nghiệm có nhiều công thức thí nghiệm có thể dùng các đường cong cho sẵn (Phụ lục 9) để xác định dung lượng mẫu cần thiết. Dung lượng mẫu sẽ phụ thuộc vào sự sai khác mong đợi giữa các công thức thí nghiệm, mức sai lầm loại I (α) và mức sai lầm loại II (β). Để có thể sử dụng được các đường cong này ta cần phải xác định được giá trị ϕ^2 . Giá trị này được tính theo công thức:

$$\phi^2 = \frac{n \sum_{i=1}^a d_i^2}{a \sigma^2}$$

Trong đó: n = số đơn vị thí nghiệm cần thiết cho một công thức thí nghiệm.

a = số công thức thí nghiệm.

d_i = sai khác mong đợi của công thức thí nghiệm thứ i với μ .

σ^2 = phương sai của tính trạng cần nghiên cứu.

Để xác định được ϕ cần phải chọn các giá trị trung bình, ví dụ ta có $\mu_1, \mu_2, \dots, \mu_a$ là các giá trị trung bình của từng công thức thí nghiệm. Ta sẽ có $\mu = (1/a) \sum_{i=1}^a \mu_i$ và $d_i = \mu_i - \mu$.

Ví dụ 4.1: muốn thiết kế một thí nghiệm để so sánh tăng khối lượng (g) của gà ở 4 khẩu phần. Các giá trị trung bình được chọn lần lượt là $\mu_1 = 71$ g, $\mu_2 = 79$ g, $\mu_3 = 80$ g và $\mu_4 = 102$ g với $\alpha = 0,05$ và $1 - \beta = 0,80$; biết $\sigma^2 = 35^2$. Cần bao nhiêu đơn vị thí nghiệm?

Ta có:

$$\mu = (71 + 79 + 80 + 102) / 4 = 83.$$

$$d_1 = 71 - 83,00 = -12.$$

$$d_2 = 79 - 83,00 = -4.$$

$$d_3 = 80 - 83,00 = -3.$$

$$d_4 = 102 - 83,00 = +9.$$

$$\sum_{i=1}^4 d_i^2 = 530, \text{ vậy ta có:}$$

$$\phi^2 = \frac{n \sum_{i=1}^4 d_i^2}{a \sigma^2} = \frac{n(530)}{4(35)^2} = 0,11n$$

Sử dụng đường cong với bậc tự do của công thức thí nghiệm là $v_1 = a - 1 = 4 - 1 = 3$, của sai số ngẫu nhiên là $v_2 = N - a = na - a = a(n - 1) = 4(n - 1)$ và $\alpha = 0,05$ ở phần phụ lục 9.

Nếu ta thử với $n = 24$ thì sẽ có các giá trị $\phi^2 = 0,11 \times 6 = 2,64$; $\phi = 1,62$ $v_2 = 4(24 - 1) = 92$. Dựa vào đường cong sẽ có $\beta = 0,23$. Bằng cách tương tự ta có:

n	ϕ^2	ϕ	$4(n - 1)$	β	$1-\beta$
24	2,64	1,62	92	0,23	0,77
25	2,75	1,66	96	0,21	0,79
26	2,86	1,69	100	0,19	0,81
27	2,97	1,72	104	0,17	0,83
28	3,08	1,75	108	0,16	0,84

Để thỏa mãn điều kiện của bài toán, ta cần chọn ít nhất 26 đơn vị thí nghiệm cho một công thức thí nghiệm.

Để có thể sử dụng được đường cong cho sẵn, khó nhất đối với người thiết kế thí nghiệm là phải chọn ra các giá trị trung bình cho từng công thức thí nghiệm để từ đó có thể xác định được dung lượng mẫu cần thiết. Có một cách tiếp cận khác đơn giản hơn để xác định dung lượng mẫu đó là chỉ cần xác định một giá trị d . Sự sai khác của 2 giá trị trung bình bất kỳ nếu vượt quá giá trị d thì giả thiết H_0 bị bác bỏ. Khi đó giá trị ϕ^2 được tính theo công thức rút gọn sau đây (xem mục 3.8.1):

$$\phi^2 = \frac{nd^2}{2a\sigma^2}$$

Để minh họa, ta có thể lấy ví dụ trên. Nếu chọn $d = 33$ g ta sẽ có

$$\phi^2 = \frac{nd^2}{2a\sigma^2} = \frac{n(33)^2}{2(4)(35)^2} = 0,11n$$

Tương tự như trên, ta cần ít nhất 26 đơn vị thí nghiệm cho một công thức thí nghiệm để thỏa mãn điều kiện bài ra.

4.1.4. Ưu điểm và nhược điểm

Ưu điểm của mô hình này là thí nghiệm thiết kế đơn giản, chính vì vậy cho nên hạn chế được nhiều sai sót trong quá trình thu thập dữ liệu. Mô hình phân tích số liệu không phức tạp, kết quả phân tích đơn giản, dễ đọc và dễ hiểu.

Mô hình có lợi thế là thích nghi một cách dễ dàng với trường hợp các đơn vị thí nghiệm không đều nhau vì các nguyên nhân nào đó, ví dụ như số liệu bị khiếm khuyết do tác động của bệnh trong quá trình làm thí nghiệm.

Ngược lại, mô hình thí nghiệm hoàn toàn ngẫu nhiên thường không có hiệu quả cao, hiệu lực của thí nghiệm không lớn do sự không thuận nhất của các vật liệu thí nghiệm.

4.1.5. Cách thiết kế thí nghiệm

Chọn n đơn vị thí nghiệm, bắt thăm n_1 đơn vị để bố trí mức A_1 , bắt thăm n_2 đơn vị để bố trí mức A_2 , ..., bắt thăm n_{k-1} đơn vị để bố trí mức A_{a-1} , n_a đơn vị còn lại bố trí mức A_a . Như vậy là *bắt thăm toàn bộ các đơn vị thí nghiệm để bố trí một cách hoàn toàn ngẫu nhiên các mức của yếu tố*. Cách thiết kế ngẫu nhiên được trình bày chi tiết ở chương 3.

Ví dụ yếu tố thí nghiệm A có 4 công thức thí nghiệm A_1, A_2, A_3 và A_4 với các 5 đơn vị thí nghiệm trong mỗi công thức thí nghiệm. Như vậy toàn bộ số đơn vị thí nghiệm là 20 và giả sử số động vật này được đánh số từ 1 đến 20. Sau khi bố trí một cách ngẫu nhiên ta có thể được mô hình thiết kế thí nghiệm như sau:

A_1	A_2	A_3	A_4
6	11	19	2
1	8	17	18
9	7	13	12
4	14	16	5
20	10	3	15

Khi kết thúc thí nghiệm, số liệu có thể ghi lại để dễ dàng và thuận tiện cho việc tính toán như sau:

A_1	A_2	A_3	A_4
6 x_{11}	11 x_{21}	19 x_{31}	2 x_{41}
1 x_{12}	8 x_{22}	17 x_{32}	18 x_{42}
9 x_{13}	7 x_{23}	13 x_{33}	12 x_{43}
4 x_{14}	14 x_{24}	16 x_{34}	5 x_{44}
20 x_{15}	10 x_{25}	3 x_{35}	15 x_{45}

Dưới dạng tổng quát với a công thức thí nghiệm số lần lặp lại r ta có:

A_1	A_2	...	A_a
x_{11}	x_{21}	...	x_{a1}
x_{12}	x_{22}	...	x_{a2}

A₁	A₂	...	A_a
x_{13}	x_{23}	...	x_{a3}
...
x_{1r}	x_{2r}	...	x_{ar}

4.1.6. Phân tích số liệu

Với các thí nghiệm được bố trí đơn giản với 2 công thức thí nghiệm. Tiến hành so sánh kết quả của 2 công thức thí nghiệm bằng phép thử t . Nếu thí nghiệm bao gồm nhiều công thức thí nghiệm, thì phân tích phương sai (ANOVA) là phù hợp nhất. Phép thử t và phân tích phương sai được trình bày chi tiết ở Chương 2.

a. Mô hình phân tích

$$x_{ij} = \mu + a_i + e_{ij} \quad (i = 1, a; j = 1, r_i)$$

Trong đó: μ : Trung bình chung.

a_i : Chênh lệch do ảnh hưởng của mức i .

e_{ij} : Sai số ngẫu nhiên; các e_{ij} độc lập, phân phối chuẩn $N(0, \sigma^2)$.

b. Cách phân tích

Cách phân tích số liệu được trình bày chi tiết ở Chương 2. Lưu ý rằng, trong mô hình thí nghiệm hoàn toàn ngẫu nhiên có 2 nguồn biến động: 1) biến động giữa các công thức thí nghiệm (SS_A) và 2) biến động do sai số ngẫu nhiên (SS_E); toàn bộ biến động của thí nghiệm (SS_{TO}) bằng tổng số các biến động thành phần (SS_A và SS_E) hợp thành. Các nguồn biến động này có thể được tính như sau:

Tổng bình phương toàn bộ biến động

$$SS_{TO} = \sum_{i=1}^a \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^a \sum_{j=1}^{n_i} x_{ij}^2 - G$$

Tổng bình phương do yếu tố

$$SS_A = \sum_{i=1}^a \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2 = \sum_{i=1}^a \frac{TA_i^2}{r_i} - G$$

Tổng bình phương do sai số

$$SS_E = SS_{TO} - SS_A = \sum_{i=1}^t \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i.} \right)^2$$

Các bậc tự do $df_{TO} = n - 1$; $df_A = a - 1$; $df_E = n - a$.

Các trung bình $MS_A = SS_A / df_A$; $MS_E = SSE / df_E$.

$F_{TN} = MS_A / MS_E$; giá trị tới hạn $F_{(\alpha, dfA, dfE)}$.

Kết luận:

Nếu $F_{TN} \leq F_{(\alpha, dfA, dfE)}$ thì chấp nhận H_0 , ngược lại thì bác bỏ H_0 .

Bảng phân tích phương sai

Nguồn biến động	df	SS	MS	F_{TN}	F
Yếu tố	a - 1	SS _A	MS _A	MS _A / MS _E	$F(\alpha, df_A, df_E)$
Sai số	n - a	SS _E	MS _E		
Toàn bộ	n - 1	SS _{TO}			

Ví dụ 4.2: Một thí nghiệm được tiến hành để so sánh mức độ tăng khối lượng của gà ở 4 khẩu phần ăn khác nhau. Chọn 20 con gà đồng đều nhau và phân một cách ngẫu nhiên vào một trong 4 khẩu phần. Như vậy ta có 4 nhóm động vật thí nghiệm, mỗi nhóm gồm 5 gà; kết quả thí nghiệm được ghi lại ở bảng sau (đơn vị tăng khối lượng tính theo g):

Khẩu phần 1	Khẩu phần 2	Khẩu phần 3	Khẩu phần 4
99	61	42	169
88	112	97	137
76	30	81	169
38	89	95	85
94	63	92	154

Đây là ví dụ về thí nghiệm được bố trí theo mô hình một yếu tố hoàn toàn ngẫu nhiên. Yếu tố thí nghiệm là *Khẩu phần* với 4 công thức thí nghiệm (*Khẩu phần 1, 2, 3 và 4*).

Ta có bảng phân tích phương sai

Nguồn biến động	df	SS	MS	F_{TN}	$F_{(0,05; 3; 16)}$
Khẩu phần	3	16467	5489	6,65	3,24
Sai số ngẫu nhiên	16	13212	826		
Tổng biến động	19	29679			

Kết luận: Bác bỏ H_0 , như vậy tăng khối lượng của gà ở 4 khẩu phần ăn không phải như nhau.

Sự sai khác nhỏ nhất có ý nghĩa (Least Significant Difference - LSD) đối với 2 mức A_i và A_j có số lần lặp n_i và n_j tính theo công thức:

$$LSD_{\alpha} = t_{(\alpha/2, df_E)} \times \sqrt{MS_E \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

Nếu chọn mức ý nghĩa $\alpha = 0,05$ $t(0,025; 16) = 2,12$; $n_i = n_j = 5$ do đó khi so sánh các trung bình có thể dùng $LSD_{0,05} = 2,12 \times \sqrt{826 \times \frac{2}{5}} = 38,54$.

So các trung bình:

$$(A_1) \text{ so với } (A_2) \quad |79 - 71| = 8 < 38,54$$

Sai khác không có ý nghĩa

$$(A_1) \text{ so với } (A_3) \quad |79 - 81,4| = 2,4 < 38,544$$

Sai khác không có ý nghĩa

(A ₁) so với (A ₄)	$ 79 - 142,8 = 63,8 > 38,54$	Sai khác có ý nghĩa
(A ₂) so với (A ₃)	$ 71 - 81,4 = 10,4 < 38,54$	Sai khác không có ý nghĩa
(A ₂) so với (A ₄)	$ 71 - 142,8 = 71,8 > 38,54$	Sai khác có ý nghĩa
(A ₃) so với (A ₄)	$ 81,4 - 142,8 = 61,4 > 38,54$	Sai khác có ý nghĩa

Ta có thể xây dựng một bảng có các chữ cái a, b, c... để thể hiện sự sai khác giữa các công thức thí nghiệm theo các bước sau:

1) Sắp xếp các giá trị trung bình theo thứ tự giảm dần như sau:

Khâu phần	Trung bình	Khâu phần	Trung bình
1	79,00	4	142,80
2	71,00	3	81,40
3	81,40	1	79,00
4	142,80	2	71,00

2) Dựa vào kết quả so sánh để tạo các đường gạch chung cho các khâu phần có giá trị trung bình bằng nhau; cụ thể như sau:

Khâu phần	Trung bình
4	142,80
3	81,40
1	79,00
2	71,00

Mỗi một đường thẳng tương ứng với một chữ cái (a, b, c...).

3) Từ bảng trên, ta có thể đặt các chữ cái bên cạnh các số trung bình và sắp xếp khâu phần theo thứ tự tăng dần như ban đầu ta có như sau:

Khâu phần	Trung bình	Khâu phần	Trung bình
4	142,80 ^a	1	79,00 ^b
3	81,40 ^b	2	71,00 ^b
1	79,00 ^b	3	81,40 ^b
2	71,00 ^b	4	142,80 ^a

Việc so sánh hai trung bình theo LSD thường chỉ dùng để so sánh một số cặp trung bình mà trước khi thí nghiệm chúng ta đã có ý đồ so sánh. Nếu so sánh tất cả các cặp trung bình, hay còn gọi là kiểm định sự bằng nhau của tất cả các cặp trung bình (multiple comparisons) thì mức ý nghĩa không còn là α mà nhỏ đi nhiều, do đó các nhà nghiên cứu thống kê đã đề xuất nhiều cách kiểm định khác để đảm bảo mức ý nghĩa α như kiểm định Scheffé, Tukey, Bonferroni, Dunnett, kiểm định đa phạm vi (multiple range test) Duncan, Student- Newman - Keuls,... Trong các chương trình máy tính chuyên về thống kê còn có nhiều cách so sánh khác.

Thí dụ muốn so sánh theo Duncan (các lần lặp bằng nhau và gọi là r) phải sắp các trung bình từ nhỏ đến lớn. Khi so sánh hiệu số các trung bình thì, tùy theo các trung bình ở kề nhau hay cách nhau một trung bình, cách nhau hai trung bình, mà dùng các phương pháp so sánh khác nhau. Việc so sánh tiến hành như sau:

1) Tính sai số của trung bình $s_{\bar{x}_i} = \sqrt{\frac{MS_E}{r}}$

2) Lấy giá trị r_p trong bảng Duncan ứng với bậc tự do df_E nhân với $s_{\bar{x}_i}$ để có khoảng R_p .

3) So sánh hiệu $x_j - \bar{x}_i$ với R_p .

Nếu hai trung bình liền nhau thì lấy $p = 2$, cách nhau một thì $p = 3$, cách nhau hai thì $p = 4, \dots$

Nếu hiệu bé hơn hay bằng R_p thì sai khác không có ý nghĩa, ngược lại thì sai khác có ý nghĩa.

Trong thí dụ trên

(A ₂)	(A ₁)	(A ₃)	(A ₄)
71,0	79,0	(81,4)	(142,8)

$$s_{\bar{x}_i} \sqrt{\frac{826}{5}} = 12,853 \text{ với bậc tự do } df_E = 16$$

p	2	3	4
r_p	3,0	3,15	3,23
R_p	38,56	40,49	41,52

(A₁) - (A₂) = 79,0 - 71,0 = 8 < R₂ = 38,56 Sai khác không có ý nghĩa.

(A₃) - (A₂) = 81,4 - 71,0 = 10,4 < R₃ = 40,49 Sai khác không có ý nghĩa.

(A₄) - (A₂) = 142,8 - 71 = 71,8 > R₄ = 41,52 Sai khác có ý nghĩa.

(A₃) - (A₁) = 81,4 - 79,0 = 2,4 < R₂ = 38,56 Sai khác không có ý nghĩa.

(A₄) - (A₁) = 142,8 - 79 = 63,8 > R₃ = 40,49 Sai khác có ý nghĩa.

(A₄) - (A₃) = 142,8 - 81,4 = 61,4 > R₂ = 38,56 Sai khác có ý nghĩa.

Trong ví dụ này các kết luận không khác với so sánh theo LSD

4.2. THÍ NGHIỆM KHỐI NGẪU NHIÊN ĐẦY ĐỦ (Randomized complete block design - RCBD)

Như đã nêu trên, mô hình thiết kế thí nghiệm kiểu hoàn toàn ngẫu nhiên chỉ thực sự có hiệu quả khi toàn bộ động vật thí nghiệm có sự đồng đều cao và các điều kiện ngoại cảnh phải được kiểm soát dễ dàng. Trong thực tế, đặc biệt là trong chăn nuôi thú y rất khó có thể thỏa mãn cùng một lúc các điều kiện đã nêu. Mô hình thiết kế thí nghiệm theo kiểu *khối ngẫu nhiên đầy đủ* được đưa ra nhằm hạn chế những khó khăn đó.

Nguyên tắc tạo khối là đạt được sự đồng đều tối đa trong một khối và sự khác nhau lớn nhất giữa các khối. Các khối được gọi là *đầy đủ* khi trong mỗi khối có đầy đủ các đại diện của các công thức thí nghiệm và ngẫu nhiên khi các đơn vị thí nghiệm được

bố trí một cách hoàn toàn *ngẫu nhiên* vào các công thức thí nghiệm. Trong quá trình thí nghiệm, tất cả các đơn vị thí nghiệm trong cùng một khối nhận được tất cả các điều kiện như nhau ngoại trừ yếu tố thí nghiệm.

Trong chăn nuôi - thú y, khói có thể coi là các nhóm động vật cùng một giống, giới tính, tuổi, cùng khối lượng hoặc cũng có thể là nhóm động vật sinh ra cùng một bố, cùng lứa.

Một số lý do để chọn mô hình thí nghiệm khói ngẫu nhiên đầy đủ là:

a) Do không tìm được đủ $n = a \times b$ đơn vị thí nghiệm đồng đều do đó phải chọn b khói, mỗi khói có a đơn vị thí nghiệm để sắp xếp cho a mức của yếu tố. Ví dụ so sánh 6 công thức thí nghiệm, mỗi công thức lặp lại 5 lần. Giả sử ta không tìm được 30 con lợn đồng đều về khối lượng, do đó chọn 5 lô, mỗi lô 6 con đồng đều để bố trí 6 công thức.

b) Có thể có một nguồn biến động theo một hướng, thí dụ hướng nắng, hướng gió, hướng dốc, hướng chảy của nước ngầm, hướng thay đổi của chất đất, . . . khi đó phải bố trí các khói vuông góc với hướng biến động nhằm cân bằng tác động của biến động (vì mỗi công thức đều có mặt ở tất cả các khói, mỗi khói một lần).

4.2.1. Số khói cần thiết

Các kỹ thuật dùng để xác định dung lượng mẫu trong mô hình thiết kế thí nghiệm một yếu tố hoàn toàn ngẫu nhiên có thể được áp dụng trực tiếp đối với mô hình khói ngẫu nhiên đầy đủ. Các đường cong cho sẵn có thể được sử dụng với công thức:

$$\phi^2 = \frac{b \sum_{i=1}^a d_i^2}{a \sigma^2}$$

Hoặc:

$$\phi^2 = \frac{bd^2}{2a\sigma^2}$$

Với b = số khói cần thiết.

Ví dụ ta chọn $d = 0,76$; $\alpha = 0,05$; $1 - \beta = 0,8$; số công thức thí nghiệm $a = 4$; $\sigma = 0,70$; ta sẽ có:

$$\phi^2 = \frac{bd^2}{2a\sigma^2} = \frac{b(1,72)^2}{2(4)(0,68)^2} = 0,8b$$

Với các bậc tự do $v_1 = a - 1 = 4 - 1 = 3$ và $v_2 = (a - 1)(b - 1) = (4 - 1)(b - 1) = 3(b - 1)$.

b	ϕ^2	ϕ	$3(b - 1)$	β	$1-\beta$
3	2,40	1,55	6	0,60	0,40
4	3,20	1,79	9	0,30	0,70
5	4,00	2,00	12	0,20	0,80
6	4,80	2,19	15	0,12	0,88
7	5,60	2,37	18	0,08	0,92

Như vậy cần ít nhất 5 khối để thoả mãn điều kiện bài toán.

4.2.2. Ưu điểm và nhược điểm

Mô hình thí nghiệm kiểu khối ngẫu nhiên đầy đủ được thiết kế đơn giản gần như mô hình thí nghiệm kiểu hoàn toàn ngẫu nhiên. Mô hình thí nghiệm theo khối có thể được thiết kế với số công thức thí nghiệm và với số lần lặp bất kỳ; nhưng đòi hỏi số lần lặp lại phải bằng nhau ở các công thức thí nghiệm. Mô hình thí nghiệm khối ngẫu nhiên đầy đủ chỉ thể hiện đầy đủ các ưu thế cho đến khi có một hay nhiều công thức thí nghiệm hoặc khối bị loại bỏ, ví dụ có những số liệu bị khuyết trong quá trình thu thập hoặc trong quá trình phân tích.

So với mô hình thí nghiệm kiểu hoàn toàn ngẫu nhiên, mô hình khối ngẫu nhiên đầy đủ cho hiệu quả và độ chính xác cao hơn. Điều này được thể hiện rõ, với cùng một nguyên vật liệu thí nghiệm sẽ cho kết quả chính xác hơn hoặc với cùng một độ chính xác có thể giảm được nguyên vật liệu thí nghiệm. Độ chính xác của thí nghiệm tăng lên bởi vì biến động giữa các khối đã được loại bỏ trong quá trình phân tích và khả năng phát hiện được ảnh hưởng của các công thức thí nghiệm tăng lên. Tuy nhiên với một số công thức thí nghiệm tương đối lớn (ví dụ nhiều hơn 20 công thức thí nghiệm) và với các nguyên vật liệu có độ đồng đều thấp thì hiệu quả của mô hình bị giảm một cách đáng kể; khi đó mô hình khối không đầy đủ sẽ được áp dụng.

4.2.3. Cách thiết kế thí nghiệm

Chọn b khối, mỗi khối có a đơn vị thí nghiệm, bắt thăm ngẫu nhiên để xếp a đơn vị thí nghiệm vào a công thức thí nghiệm trong khối 1, sau đó bắt thăm để xếp a công thức vào a ô trong khối 2, . . . , cuối cùng là bắt thăm cho khối b.

Ví dụ bố trí thí nghiệm với 4 công thức thí nghiệm (A1, A2, A3 và A4) với 5 khối khác nhau (1, 2, 3, 4 và 5). Như vậy ta sẽ tạo ra 5 khối khác nhau đảm bảo sự đồng đều tối đa trong từng khối, mỗi khối có 4 đơn vị thí nghiệm (4 lần lặp lại) và kỹ thuật bắt thăm hoàn toàn ngẫu nhiên để phân 4 động vật thí nghiệm trong từng khối về với 4 công thức thí nghiệm.

Nếu động vật thí nghiệm được đánh số theo sơ đồ sau:

		Khối				
		1	2	3	4	5
Động vật thí nghiệm số	1	5	9	13	17	
	2	6	10	14	18	
	3	7	11	15	19	
	4	8	12	16	20	

Sau khi bố trí các đơn vị bằng cách bốc thăm ngẫu nhiên, sơ đồ thiết kế thí nghiệm có thể được trình bày theo sơ đồ:

		Khối				
Công thức	1	2	3	4	5	
A1	1	8	11	14	18	
A2	4	6	9	15	19	
A3	2	7	10	16	17	
A4	3	5	12	13	20	

Số liệu thu được khi kết thúc thí nghiệm có thể được trình bày

Công thức	Khối									
	1	2	3	4	5					
A1	1	x_{11}	8	x_{12}	11	x_{13}	14	x_{14}	18	x_{15}
A2	4	x_{21}	6	x_{22}	9	x_{23}	15	x_{24}	19	x_{25}
A3	2	x_{31}	7	x_{32}	10	x_{33}	16	x_{34}	17	x_{35}
A4	3	x_{41}	5	x_{42}	12	x_{43}	13	x_{44}	20	x_{45}

Hay ở dạng tổng quát với a công thức và b khối

Công thức	Khối									
	1	2	...	b						
A1	x_{11}	x_{12}	...	x_{1b}						
A2	x_{21}	x_{22}	...	x_{2b}						
...						
Aa	x_{a1}	x_{a2}	...	x_{ab}						

4.2.4. Phân tích số liệu

Phân tích phương sai (ANOVA) được sử dụng để phân tích số liệu. Trong mô hình thí nghiệm kiểu khối ngẫu nhiên đầy đủ có 3 nguồn biến động: 1) biến động giữa các khối (SS_K), 2) biến động giữa các công thức thí nghiệm (SS_A) và 3) biến động do sai số ngẫu nhiên (SS_E); toàn bộ biến động của thí nghiệm (SS_{TO}) chính bằng tổng các biến động thành phần (SS_K , SS_A và SS_E). Các nguồn biến động này có thể được trình bày qua mô hình phân tích dưới đây

a. Mô hình phân tích

$$x_{ij} = \mu + a_i + b_j + e_{ij} \quad i = 1, \dots, a; j = 1, \dots, b$$

μ là trung bình chung.

a_i là chênh lệch do ảnh hưởng của mức i của yếu tố, $\sum a_i = 0$.

b_j là chênh lệch do ảnh hưởng của khối j , $\sum b_j = 0$.

e_{ij} là sai số ngẫu nhiên; các e_{ij} độc lập, phân phối chuẩn $N(0, \sigma^2)$.

b. Cách phân tích

Tính tổng bình phương toàn bộ SS_{TO}

$$SS_{TO} = \sum_{i=1}^a \sum_{j=1}^b (x_{ij} - \bar{x})^2$$

Tính tổng bình phương do yếu tố SS_A

$$SS_A = \sum_{i=1}^a \sum_{j=1}^b (\bar{x}_{i\cdot} - \bar{x})^2$$

Tính tổng bình phương do khối SS_K

$$SS_K = \sum_{i=1}^a \sum_{j=1}^b (\bar{x}_{.j} - \bar{x})^2$$

Tính trung bình do sai số SS_E :

$$SS_E = \sum_{i=1}^a \sum_{j=1}^b (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2$$

Cũng có thể tính nhanh các tổng bình phương như sau:

Tính tổng hàng (công thức thí nghiệm) TA_i ($i = 1, a$), trung bình hàng (công thức thí nghiệm) $\bar{x}_{i.}$.

Tổng cột (khỏi) TK_j ($j = 1, r$), trung bình cột $\bar{x}_{.j}$

Tổng số quan sát $n = a \times b$.

Tổng toàn bộ các số liệu $ST = \sum \sum x_{ij}$, trung bình toàn bộ \bar{x}

Tính số hiệu chỉnh $G = ST^2 / n$

$$SS_{TO} = \sum_{i=1}^a \sum_{j=1}^b x_{ij}^2 - G$$

$$SS_A = \frac{1}{b} \sum_{i=1}^a TA_i^2 - G$$

$$SS_K = \frac{1}{a} \sum_{j=1}^b TK_j^2 - G$$

$$SS_E = SS_{TO} - SS_A - SS_K$$

Bậc tự do $df_{TO} = n - 1 = a \times b - 1$; $df_A = a - 1$; $df_K = r - 1$; $df_E = (a-1)(b-1)$

Các trung bình bình phương:

$$MS_A = SS_A / df_A; \quad MS_K = SS_K / df_K; \quad MS_E = SSE / df_E$$

Trong quá trình phân tích thường ít chú ý kiểm định khói mà chỉ tập trung kiểm định yếu tố. Giả thiết H_0 : “Các trung bình của các mức bằng nhau”, đối thiết H_1 : “Có ít nhất một cặp trung bình khác nhau”.

Tính $F_{TN} = MS_A / MS_E$; so với giá trị tới hạn $F(\alpha, df_A, df_E)$

Kết luận:

Nếu $F_{TN} \leq F(\alpha, df_A, df_E)$ thì chấp nhận H_0 , ngược lại thì bác bỏ H_0 .

Dưới dạng tổng hợp ta có bảng phân tích phương sai:

Nguồn biến động	df	SS	MS	F_{TN}	F tới hạn
Yếu tố	a-1	SS_A	MS_A	MS_A / MS_E	$F_{(\alpha, dfA, dfE)}$
Khối	b-1	SS_K	MS_K		
Sai số	(a-1)(b-1)	SS_E	MS_E		
Toàn bộ	ab -1	SS_{TO}			

Ví dụ 4.3: (Mead và cộng sự) Nghiên cứu số lượng tế bào lymphô ở chuột ($\times 1000$ tế bào mm^{-3} máu) được sử dụng 4 loại thuốc khác nhau (A, B, C và D; thuốc D là placebo) qua 5 lứa; số liệu thu được như sau:

	Lứa 1	Lứa 2	Lứa 3	Lứa 4	Lứa 5
Thuốc A	7,1	6,1	6,9	5,6	6,4
Thuốc B	6,7	5,1	5,9	5,1	5,8
Thuốc C	7,1	5,8	6,2	5,0	6,2
Thuốc D	6,7	5,4	5,7	5,2	5,3

Đây là mô hình thí nghiệm kiểu khói ngẫu nhiên đầy đủ với số công thức thí nghiệm $a = 4$, số khói chính bằng số lứa $b = 5$.

$$n = 4 \times 5 = 20; ST = 119,3; G = 119,3^2 / 20 = 711,6245; \sum x_{ij}^2 = 720,51$$

$$(\Sigma TA_i^2) / r = 3567,35 / 5 = 713,47; (\Sigma TK_j^2) / a = 2872,11 / 4 = 718,0275$$

$$SS_{TO} = 720,51 - 711,6245 = 8,8855$$

$$SS_A = 713,47 - 711,6245 = 1,8455$$

$$SS_K = 718,0275 - 711,6245 = 6,4030$$

$$SS_E = 8,8855 - 1,8455 - 6,4030 = 0,6370$$

Bảng phân tích phương sai

Nguồn	df	SS	MS	F_{TN}	$F_{(0,05; 3; 12)}$
Thuốc	3	1,8455	0,6152	11,59	3,49
Lứa	4	6,4030	1,6007		
Sai số	12	0,6370	0,0531		
Tổng số	19	8,8855			

Kết luận: Bác bỏ giả thiết H_0 , điều này chứng tỏ khi sử dụng các loại thuốc khác nhau đã làm cho số lượng tế bào lymphô trong máu thay đổi.

$$Sai số thí nghiệm se = \sqrt{MS_E} = \sqrt{0,0531} = 0,2304$$

Có thể sử dụng sai khác bé nhất có ý nghĩa ở mức 5% (LSD) để xác định sự sai khác có ý nghĩa thống kê của các cặp giá trị trung bình bất kỳ

$$LSD(0,05) = t_{dfE}^{(0,025)} \times \sqrt{\frac{MS_E \times 2}{b}} = 2,179 \times \sqrt{\frac{0,0531 \times 2}{5}} = 0,3176$$

Trung bình

$$(A) = 6,42$$

$$(B) = 5,72$$

$$(C) = 6,06$$

$$(D) = (5,66)$$

- So (A) với (B) $| 6,42 - 5,72 | = 0,70 > LSD$ Khác nhau có ý nghĩa
 So (A) với (C) $| 6,42 - 6,06 | = 0,36 > LSD$ Khác nhau có ý nghĩa
 So (A) với (D) $| 6,42 - 5,66 | = 0,76 > LSD$ Khác nhau có ý nghĩa
 So (B) với (C) $| 5,72 - 6,06 | = 0,34 > LSD$ Khác nhau có ý nghĩa
 So (B) với (D) $| 5,72 - 5,66 | = 0,06 < LSD$ Khác nhau không có ý nghĩa
 So (C) với (D) $| 6,06 - 5,66 | = 0,40 > LSD$ Khác nhau không có ý nghĩa

Sau khi so sánh ta có được các giá trị trung bình cùng với các chữ cái tương ứng thể hiện sự sai khác như sau:

- A 6,42^a
 B 5,72^b
 C 6,06^c
 D 5,66^b

Như vậy, các giá trị trung bình không có chữ giống nhau thì khác nhau ($P < 0,05$).

4.3. THÍ NGHIỆM KHỐI NGẪU NHIÊN VÓI NHIỀU ĐƠN VỊ THÍ NGHIỆM TRONG MỘT CÔNG THỨC THÍ NGHIỆM VÀ KHỐI

4.3.1. Cách thiết kế thí nghiệm

Trong phần trước, đối với thí nghiệm khối ngẫu nhiên đầy đủ chỉ có một đơn vị thí nghiệm trong một tổ hợp (công thức thí nghiệm \times khối) và sai số ngẫu nhiên của mô hình chính bằng tương tác giữa công thức thí nghiệm và khối. Chính vì vậy không thể kiểm tra được tác động tương tác giữa công thức thí nghiệm và khối. Giải pháp duy nhất để kiểm tra tác động tương tác là tăng số đơn vị thí nghiệm trong mỗi tổ hợp (công thức thí nghiệm \times khối) lên ít nhất 2 đơn vị. Một lần nữa xem xét a công thức thí nghiệm và b khối, nhưng trong mỗi tổ hợp (công thức thí nghiệm \times khối) có n đơn vị thí nghiệm. Như vậy số đơn vị thí nghiệm trong mỗi khối sẽ là $(n \times a)$ và được bố trí một cách ngẫu nhiên vào với các công thức thí nghiệm đảm bảo mỗi công thức thí nghiệm trong khối có n đơn vị thí nghiệm.

Ví dụ: Một thí nghiệm có 5 khối, 4 công thức thí nghiệm và 8 đơn vị thí nghiệm trong từng khối; do đó sẽ có 2 đơn vị thí nghiệm trong một tổ hợp (công thức thí nghiệm \times khối). Sơ đồ thiết kế thí nghiệm được thể hiện như sau:

Công thức	Khối				
	1	2	3	4	5
A1	1	12	23	26	39
	7	11	18	31	37
A2	8	9	19	25	36
	6	15	20	32	38
A3	4	10	24	29	33
	5	16	17	27	40
A4	3	13	22	30	35
	2	14	21	28	34

Số liệu khi kết thúc thí nghiệm có thể được trình bày như sau:

		Khối				
Công thức		1	2	3	4	5
A1	X111	X121	X131	X141	X151	
	X112	X122	X132	X142	X152	
A2	X211	X221	X231	X241	X251	
	X212	X222	X232	X242	X252	
A3	X311	X321	X331	X341	X351	
	X312	X322	X332	X342	X352	
A4	X411	X421	X431	X441	X451	
	X412	X422	X432	X442	X452	

4.3.2. Mô hình phân tích

$$x_{ijk} = \mu + a_i + b_j + a \times b_{ij} + e_{ijk} \quad i = 1, \dots, a; j = 1, \dots, b; k = 1, \dots, n$$

x_{ijk} Là quan sát thứ k của khối thứ j và công thức thí nghiệm thứ i

μ Trung bình chung.

a_i Chênh lệch do ảnh hưởng của mức i của yếu tố $\sum a_i = 0$

b_j Chênh lệch do ảnh hưởng của khối j , $\sum b_j = 0$

$a \times b_{ij}$ Chênh lệch do tương tác giữa công thức thí nghiệm và khối

e_{ijk} Sai số ngẫu nhiên; các e_{ijk} độc lập, phân phối chuẩn $N(0, \sigma^2)$

4.3.3. Cách phân tích

Trong mô hình này, các nguồn biến động bao gồm: 1) biến động giữa các khối (SS_K), 2) biến động giữa các công thức thí nghiệm (SS_A), 3) biến động do ảnh hưởng tương tác (SS_{AK}) và 4) biến động do sai số ngẫu nhiên (SS_E); toàn bộ biến động của thí nghiệm (SS_{TO}) chính bằng tổng các biến động thành phần (SS_K, SS_A, SS_{AK} và SS_E). Các nguồn biến động này có thể tính như sau:

Tính tổng bình phương toàn bộ SS_{TO}

$$SS_{TO} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (x_{ijk} - \bar{x})^2$$

Tính tổng bình phương do yếu tố SS_A

$$SS_A = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (\bar{x}_{i..} - \bar{x})^2 = bn \sum_{i=1}^a (\bar{x}_{i..} - \bar{x})^2$$

Tính tổng bình phương do khối SS_K :

$$SS_K = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (\bar{x}_{.j.} - \bar{x})^2 = an \sum_{j=1}^b (\bar{x}_{.j.} - \bar{x})^2$$

Tính tổng bình phương do tương tác yếu tố và khối SS_{AK}

$$SS_{AK} = n \sum_{i=1}^a \sum_{j=1}^b (\bar{x}_{ij.} - \bar{x})^2 - SS_K - SS_A$$

Tổng bình phương do sai số $SS_E = SS_{TO} - SS_A - SS_K$

$$SS_E = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n \left(x_{ijk} - \bar{x}_{ij.} \right)^2$$

Có thể tính nhanh các tổng bình phương như sau:

$$SS_{TO} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n x_{ijk}^2 - G$$

$$SS_A = \frac{1}{bn} \sum_{i=1}^a \left(\sum_{j=1}^b \sum_{k=1}^n x_{ijk} \right)^2 - G$$

$$SS_K = \frac{1}{bn} \sum_{j=1}^b \left(\sum_{i=1}^a \sum_{k=1}^n x_{ijk} \right)^2 - G$$

$$SS_{AK} = \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \left(\sum_{k=1}^n x_{ijk} \right)^2 - SS_K - SS_A - G$$

$$SS_E = SS_{TO} - SS_A - SS_K$$

Bậc tự do $df_{TO} = abn - 1$; $df_A = a - 1$; $df_K = b - 1$; $df_{AK} = (a-1)(b-1)$; $df_E = ab(n-1)$

Các trung bình bình phương:

$$MS_A = SS_A / df_A; MS_K = SS_K / df_K; MS_{AK} = SS_{AK} / df_{AK}; MS_E = se^2 = SSE / df_E$$

Giả thiết đối với tương tác giữa công thức thí nghiệm và khối; H_0 : Không có tương tác giữa công thức thí nghiệm và khối với đối thiết H_1 : Có tương tác giữa công thức thí nghiệm và khối.

Tính $F_{TN} = MS_{AK} / MSE$; so với giá trị tối hạn $F_{(\alpha, df_{AK}, df_E)}$; nếu $F_{TN} \leq F_{(\alpha, df_{AK}, df_E)}$ thì chấp nhận H_0 , ngược lại thì bác bỏ H_0

Giả thiết đối với yếu tố thí nghiệm; H_0 : “Các trung bình của các mức bằng nhau” với đối thiết H_1 : “Có ít nhất một cặp trung bình khác nhau”.

Tính $F_{TN} = MS_A / MSE$; so với giá trị tối hạn $F_{(\alpha, df_A, df_E)}$; nếu $F_{TN} \leq F_{(\alpha, df_A, df_E)}$ thì chấp nhận H_0 , ngược lại thì bác bỏ H_0 .

Dưới dạng tổng hợp ta có bảng phân tích phương sai:

Nguồn biến động	df	SS	MS	F_{TN}	F
Yếu tố	a-1	SS_A	MS_A	MS_A / MSE	$F_{(\alpha, df_A, df_E)}$
Khối	b-1	SS_K	MS_K		
Yếu tố \times Khối	(a-1)(b-1)	SS_{AK}	MS_{AK}	MS_{AK} / MSE	$F_{(\alpha, df_{AK}, df_E)}$
Sai số	$ab(n-1)$	SS_E	MS_E		
Toàn bộ	$abn - 1$	SS_{TO}			

Ví dụ 4.4: Một thí nghiệm được tiến hành để xác định ảnh hưởng của 3 công thức thí nghiệm (A1, A2 và A3) đến tăng khối lượng trung bình trên ngày (g/ngày) của bê đực. Bê đực được cân và chia thành 4 khối dựa theo khối lượng bắt đầu thí nghiệm. Trong mỗi khối có 6 động vật thí nghiệm được chọn ra và được phân ngẫu nhiên về với các công thức thí nghiệm. Như vậy toàn bộ số động vật thí nghiệm tham gia thí nghiệm là $4 \times 3 \times 2 = 24$ bê. Số liệu thu thập sau khi kết thúc thí nghiệm như sau:

	Khối			
	I	II	III	IV
A1	826	864	795	850
	806	834	810	845
A2	827	871	729	860
	800	881	709	840
A3	753	801	736	820
	773	821	740	835

Tổng bình phương do công thức thí nghiệm $SS_A = 8025,58$.

Tổng bình phương do khối $SS_K = 33816,83$.

Tổng bình phương do tương tác giữa khối và công thức thí nghiệm $SS_{AK} = 8087,42$.

Tổng bình phương do sai số $SS_E = 2110,00$.

Bảng phân tích phương sai (ANOVA).

Nguồn biến động	df	SS	MS	F _{TN}	F
Yếu tố	2	8025,58	4012,79	22,82	$F_{(0,05, 2, 12)} = 3,89$
Khối	3	33816,83	11272,28		
Yếu tố \times Khối	6	8087,42	1347,90	7,67	$F_{(0,05, 6, 12)} = 3,00$
Sai số	12	2110,00	175,83		
Toàn bộ	23	52039,83			

Như vậy, ở mức $\alpha = 0,05$; giả thiết H_0 bị bác bỏ đối với cả công thức thí nghiệm và tương tác (công thức thí nghiệm \times khối). Điều này chứng tỏ rằng có ảnh hưởng của công thức thí nghiệm và ảnh hưởng này khác nhau ở từng khối khác nhau. Hay nói một cách khác, ảnh hưởng của công thức thí nghiệm khác nhau tuỳ thuộc vào khối lượng vào thời điểm bắt đầu thí nghiệm.

4.4. THÍ NGHIỆM Ô VUÔNG LA TINH

Ngoài kiểu bố trí hoàn toàn ngẫu nhiên và khối ngẫu nhiên đầy đủ còn hay dùng kiểu ô vuông La tinh trong thí nghiệm một yếu tố. Trong mô hình này công thức thí nghiệm được bố trí vào các khối theo 2 hướng khác nhau, thường gọi là hàng và cột. Mỗi hàng và mỗi cột là một khối đầy đủ chứa tất cả các công thức thí nghiệm.

Kiểu thí nghiệm này được lựa chọn khi khảo sát yếu tố trong hoàn cảnh có hai hướng biến động mà chúng ta muốn cân bằng, ví dụ theo dõi sản lượng sữa của các bò sữa ở các công thức thí nghiệm khác nhau và trong các giai đoạn tiết sữa khác nhau ở chu kỳ tiết sữa.

Mô hình này đặc biệt hữu ích đối với thí nghiệm có số lượng động vật bị hạn chế và sự đồng đều không cao. Ví dụ nghiên cứu sự biến đổi protein trong dạ cỏ bằng cách

sử kỹ thuật lỗ dò dạ cỏ ở 4 động vật; 4 loại thức ăn (A, B, C và D) được tiến hành nghiên cứu, mỗi loại thức ăn chứa trong các túi nilon được đặt trong dạ cỏ của từng động vật trong các khoảng thời gian khác nhau.

Đặc điểm của cách thiết kế thí nghiệm này là mỗi mức của yếu tố có mặt một lần ở mỗi hàng và một lần ở mỗi cột, sự sắp xếp này là hoàn toàn ngẫu nhiên; ví dụ theo dõi lượng sữa của 4 con bò sữa trong 4 giai đoạn trong chu kỳ tiết sữa, khi cho ăn theo 4 công thức A₁, A₂, A₃, A₄.

Số công thức thí nghiệm chính bằng số hàng và số cột còn só ô vuông cần thiết chính là bình phương của số công thức thí nghiệm. Lưu ý rằng, tất cả các động vật tham gia thí nghiệm phải được giữ lại đến khi kết thúc thí nghiệm, nếu không trong quá trình xử lý số liệu sẽ gặp nhiều khó khăn.

Mô hình ô vuông La tinh thường được sử dụng với số công thức thí nghiệm từ 4 đến 8, hay sử dụng nhất là mô hình 4×4 và ít sử dụng đối với mô hình lớn hơn 8×8.

4.4.1. Ưu điểm và nhược điểm của mô hình

Trong mô hình thí nghiệm này, hai hướng biến động được kiểm soát đồng thời, vì vậy mô hình này về cơ bản cho hiệu quả cao hơn so với mô hình thí nghiệm kiểu hoàn toàn ngẫu nhiên và khối ngẫu nhiên đầy đủ, đồng thời giảm được số động vật tham gia thí nghiệm cũng như khắc phục được sự kém đồng đều của động vật thí nghiệm.

Tuy nhiên, kiểu thí nghiệm này có những nhược điểm là số mức của hai hướng biến động phải chọn bằng nhau và bằng số mức của yếu tố, giả thiết rằng không có tương tác giữa các hướng với nhau và với yếu tố; thêm vào đó, số bậc tự do của sai số ngẫu nhiên tương đối nhỏ, nên các kiểm định F trong phân tích phương sai và các kiểm định về các trung bình kém chính xác.

4.4.2. Cách thiết kế thí nghiệm

Có a mức của yếu tố (A₁, A₂, . . . , A_a). Chọn a mức của hướng biến động thứ nhất, gọi đó là a hàng. Chọn a mức của hướng biến động thứ hai, gọi đó là a cột. Chọn một sơ đồ ô vuông La tinh $a \times a$ để sau đó bắt thăm a mức của yếu tố vào các ô trong sơ đồ. Lưu ý rằng, cần phải tiến hành ngẫu nhiên hóa theo hàng hoặc theo cột cũng như bố trí các công thức thí nghiệm trong các hàng và các cột phải tuân thủ theo nguyên tắc ngẫu nhiên.

Ví dụ bố trí thí nghiệm theo mô hình ô vuông La tinh 4×4 , sơ đồ thiết kế thí nghiệm cơ bản có trong các bảng in sẵn hoặc có thể tự làm một cách đơn giản như sau. Hàng đầu viết các chữ cái a b c d; hàng thứ hai đầy b lên đầu còn a chạy xuống cuối, hàng thứ ba đầy c lên còn b chạy xuống cuối, Cách này gọi tắt là xếp hàng vòng quanh, sau đó ta được:

a	b	c	d
b	c	d	a
c	d	a	b
d	a	b	c

Bát thăm ngẫu nhiên 4 thẻ có ghi các số 1, 2, 3, 4. Thí dụ được 3 4 1 2; như vậy chúng ta có tương ứng: a → A₃, b → A₄, c → A₁, d → A₂

A ₃	A ₄	A ₁	A ₂
A ₄	A ₁	A ₂	A ₃
A ₁	A ₂	A ₃	A ₄
A ₂	A ₃	A ₄	A ₁

Ta có một sơ đồ thiết kế thí nghiệm với 4 công thức thí nghiệm. Các cột và hàng được biểu thị tương ứng với các giai đoạn và các động vật thí nghiệm như sau:

Hàng (Giai đoạn)	Cột (Động vật)			
	1	2	3	4
1	A ₃	A ₄	A ₁	A ₂
2	A ₄	A ₁	A ₂	A ₃
3	A ₁	A ₂	A ₃	A ₄
4	A ₂	A ₃	A ₄	A ₁

Nếu x_{ijk} là giá trị ở hàng thứ i, cột thứ j và ở công thức thí nghiệm k; thì số liệu thu thập được từ mô hình có thể được trình bày dưới dạng tổng quát như sau:

Hàng (Giai đoạn)	Cột (Động vật)			
	1	2	3	4
1	x ₁₁₍₃₎	x ₁₂₍₄₎	x ₁₃₍₁₎	x ₁₃₍₂₎
2	x ₂₁₍₁₄₎	x ₂₂₍₁₎	x ₂₃₍₂₎	x ₂₃₍₃₎
3	x ₃₁₍₁₎	x ₃₂₍₂₎	x ₃₃₍₃₎	x ₃₃₍₄₎
4	x ₄₁₍₂₎	x ₄₂₍₃₎	x ₄₃₍₄₎	x ₄₃₍₁₎

4.4.3. Mô hình phân tích

$$x_{ijk} = \mu + h_i + c_j + a_k + e_{ijk} \quad (i=1, a; j=1, k; k=1, a)$$

x_{ijk} Là quan sát ở hàng thứ i, cột thứ j và ở công thức thí nghiệm k.

μ Trung bình chung.

h_i Chênh lệch do ảnh hưởng của hàng i, $\sum h_i = 0$.

c_j Chênh lệch do ảnh hưởng của cột j, $\sum c_j = 0$.

a_k Chênh lệch do ảnh hưởng của mức k của yếu tố, $\sum a_k = 0$.

e_{ijk} Sai số ngẫu nhiên; giả sử các e_{ijk} độc lập, phân phối chuẩn $N(0, \sigma^2)$.

4.4.4. Cách phân tích

Toàn bộ biến động được hợp thành từ các biến động thành phần hàng, cột, công thức thí nghiệm và sai số ngẫu nhiên.

$$SS_{TO} = SS_H + SS_C + SS_A + SS_E$$

Với các bậc tự do tương ứng:

$$(a^2 - 1) = (a - 1) + (a - 1) + (a - 1) + (a - 2)(a - 1)$$

$$SS_H = a \sum_{i=1}^a \left(\bar{x}_i - \bar{x} \right)^2$$

$$SS_C = a \sum_{j=1}^a \left(\bar{x}_j - \bar{x} \right)^2$$

$$SS_A = a \sum_{k=1}^a \left(\bar{x}_k - \bar{x} \right)^2$$

$$SS_E = a \sum_{i=1}^a \sum_{j=1}^a \left(\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + 2\bar{x} \right)^2$$

$$SS_{TO} = \sum_{i=1}^a \sum_{j=1}^a \left(x_{ijk} - \bar{x} \right)^2$$

Bậc tự do $df_{TO} = a^2 - 1$; $df_H = a - 1$; $df_C = a - 1$; $df_A = a - 1$; $df_E = (a-1)(a-2)$

Các trung bình bình phương:

$$MS_H = SS_A / df_H; MS_C = SS_C / df_C; MS_A = SS_A / df_A; MS_E = SS_E / df_E$$

Giả thiết đối với yếu tố thí nghiệm; H_0 : “Các trung bình của các mức bằng nhau” với đối thiết H_1 : “Có ít nhất một cặp trung bình khác nhau”.

Tính $F_{TN} = MS_A / MS_E$; so với giá trị tới hạn $F_{(\alpha, df_A, df_E)}$; nếu $F_{TN} \leq F_{(\alpha, df_A, df_E)}$ thì chấp nhận H_0 , ngược lại thì bác bỏ H_0 .

Kiểm định đối với hàng và cột thường ít được quan tâm đến vì không mang lại nhiều ý nghĩa, tuy nhiên cũng có thể làm tương tự như kiểm định đối với công thức thí nghiệm.

Có thể tính nhanh các tổng bình phương như sau:

Tính tổng hàng TH_i , tổng cột TC_j , tổng theo từng mức của yếu tố TA_k , sau đó tính $n = a \times a$; Tổng toàn bộ giá trị số liệu trong bảng ST = Σx_{ij} hoặc $ST = \Sigma TH_i$

Tính tổng các giá trị số liệu bình phương SST = $\Sigma \Sigma x_{ij}^2$

Số điều chỉnh $G = ST^2 / n$

Tổng bình phương toàn bộ $SS_{TO} = SST - G$

Tổng bình phương do hàng $SS_H = \Sigma TH_i^2 / a - G$

Tổng bình phương do cột $SS_C = \Sigma TC_j^2 / a - G$

Tổng bình phương do yếu tố $SS_A = \Sigma TA_k^2 / a - G$

Tổng bình phương do sai số $SS_E = SS_{TO} - SS_A - SS_H - SS_C$

Bảng phân tích phuong sai (ANOVA)

Nguồn biến động	df	SS	MS	F_{TN}	F tới hạn
Yếu tố	a-1	SS_A	MS_A	MS_A/MS_E	$F_{(\alpha, df_A, df_E)}$
Hàng	a-1	SS_H	MS_H	MS_H/MS_E	$F_{(\alpha, df_H, df_E)}$
Cột	a-1	SS_C	MS_C	MS_C/MS_E	$F_{(\alpha, df_C, df_E)}$
Sai số	(a-1)(a-2)	SS_E	MS_E		
Toàn bộ	$a^2 - 1$	SS_{TO}			

Ví dụ 4.5: Mead và cộng sự tiến hành nghiên cứu ảnh hưởng của thức ăn mùa đông đến sản lượng sữa theo mô hình ô vuông latin. Có 4 khẩu phần ăn khác nhau (A_1, A_2, A_3, A_4), 4 giai đoạn thí nghiệm (1, 2, 3 và 4) mỗi giai đoạn kéo dài 3 tuần và có 4 động vật thí nghiệm (1, 2, 3 và 4). Mỗi bò ăn từng khẩu phần trong 3 tuần và mỗi bò tham gia ở cả 4 giai đoạn thí nghiệm. Sản lượng sữa chỉ được tính tổng cộng trong tuần thứ 3 của mỗi giai đoạn. Số liệu được ghi lại như sau (đơn vị tính kg)

		Bò (cột)				
		1	2	3	4	Tổng số
Giai đoạn (hàng)	1	A1 192	A2 195	A3 292	A4 249	928
	2	A2 190	A4 203	A1 218	A3 210	821
	3	A3 214	A1 139	A4 245	A2 163	761
	4	A4 221	A3 152	A2 204	A1 134	711
Tổng số		817	869	959	756	3221

Ta có bảng phân tích phương sai:

Nguồn biến động	df	SS	MS	F _{TN}	F
Khẩu phần	3	8608,70	2869,20	21,22	$F_{(0,05; 3; 6)} = 4,76$
Giai đoạn	3	6539,20	2179,20	16,12	
Bò	3	9929,20	3309,70	24,47	
Sai số	6	811,40	135,20		
Toàn bộ	15	25887,50			

Kết luận: Ở mức $\alpha = 0,05$ ta bác bỏ giả thiết H_0 , tức là các khẩu phần ăn khác nhau đã làm ảnh hưởng đến sản lượng sữa.

Có thể dùng phương pháp LSD để so sánh sự khác nhau giữa từng cặp công thức thí nghiệm như sau:

$$LSD = t(0,025; 6) \times \sqrt{\frac{135,20 \times 2}{4}} = 20,12$$

Các giá trị trung bình trước và sau khi so sánh:

Khẩu phần	Trung bình		Khẩu phần	Trung bình
A_1	170,80		A_1	170,80 ^a
A_2	188,00	→	A_2	188,00 ^a
A_3	217,00		A_3	217,00 ^b
A_4	229,50		A_4	229,50 ^b

Ngoài 3 kiểu thiết kế thí nghiệm đã nêu trên (Hoàn toàn ngẫu nhiên, Khối ngẫu nhiên đầy đủ và ô vuông La tinh) còn một số kiểu bố trí thí nghiệm một yếu tố phức tạp hơn như:

Khi mỗi khối không chứa đủ các mức của yếu tố (số ô trong một khối nhỏ hơn số mức a) thì có thể bố trí kiểu khối ngẫu nhiên cân đối không đủ (BIBD).

Khi có 3 hướng biến động thì có thể mở rộng kiểu ô vuông La tinh thành ô vuông La tinh Hy lạp (Greco Latin square).

Khi bố trí ô vuông La tinh với số công thức thí nghiệm ít thì số bậc tự do còn lại cho sai số ngẫu nhiên nhỏ do đó có thể lặp lại ô vuông La tinh để tăng bậc tự do cho sai số.

Trong các thí nghiệm về giống khi khảo sát ban đầu với số lượng các dòng (giống) quá lớn thì có thể chọn kiểu lưới ô vuông (Lattice design).

4.5. BÀI TẬP

4.5.1: Một thí nghiệm được tiến hành nhằm nghiên cứu ảnh hưởng của 4 công thức thức ăn khác nhau (A, B, C và D) đến tăng khối lượng của bò BBB. Chọn 24 bò đồng đều và chia hoàn toàn ngẫu nhiên về với các công thức. Khối lượng (kg) khi kết thúc thí nghiệm của 24 bò nêu trên thu được như sau:

	A	B	C	D
	456	365	502	457
	436	400	476	456
	432	375	487	467
	463	387	499	487
	454	408	476	469
	453	355	453	432

Kết luận ảnh hưởng của các công thức thức ăn đến tăng khối lượng của bò BBB.

4.5.2: Tăng khối lượng của gà ở 16 công thức thí nghiệm, các công thức khác nhau ở các mức axit amin. Mỗi giá trị trong bảng dưới đây là toàn bộ khối lượng (g) của 3 gà cùng một lồng trong giai đoạn từ 10 đến 20 ngày tuổi. Có 6 khu chuồng khác nhau, các công thức thí nghiệm được phân về các lồng một cách hoàn toàn ngẫu nhiên trong cùng một khu chuồng có điều kiện tiêu khí hậu ở mức độ đồng đều cao nhất có thể. Kết luận về ảnh hưởng của axit amin đến tăng khối lượng của gà.

Công thức	Khu chuồng					
	1	2	3	4	5	6
A	125	95	121	92	80	87
B	201	169	152	174	141	128
C	251	216	209	231	226	230
D	332	323	310	317	320	291
E	224	170	176	193	163	153
F	294	290	268	279	274	267
G	206	187	172	180	170	147
H	298	237	281	291	267	184
I	116	101	103	146	94	80
J	135	137	129	138	121	131
K	171	160	156	207	171	144
L	262	277	233	249	213	221
M	165	155	135	165	145	124
N	222	196	184	200	164	167
O	180	156	187	187	162	157
P	247	264	211	247	222	229

4.5.3: Một thí nghiệm được tiến hành nhằm xác định ảnh hưởng của các loại thức ăn bổ sung khác nhau (A, B, C và D) đến lượng cỏ khô mà bê nuôi vỗ béo thu nhận được (kg/ngày). Thí nghiệm được thiết kế theo mô hình ô vuông la tinh với 4 động vật trong 4 giai đoạn, mỗi giai đoạn 20 ngày. Trong mỗi giai đoạn 10 ngày đầu được coi là giai đoạn thích nghi, 10 ngày tiếp theo là giai đoạn thí nghiệm để thu thập số liệu. Số liệu thu được ở bảng bên cạnh là khối lượng cỏ khô trung bình bê thu nhận được ở 10 ngày thí nghiệm. Hãy rút ra kết luận từ thí nghiệm nêu trên.

Giai đoạn	Bê			
	1	2	3	4
1	10,0 (B)	9,0 (D)	11,1 (C)	10,8 (A)
2	10,2 (C)	11,3 (A)	9,5 (D)	11,4 (B)
3	8,5 (D)	11,2 (B)	12,8 (A)	11 (C)
4	11,1 (A)	11,4 (C)	11,7 (B)	9,9 (D)

4.5.4: Giả sử, một thí nghiệm được thiết kế tương tự như ở bài tập 4.5.3, nhưng có có 2 ô vuông la tinh được thiết kế đồng thời và mỗi ô đều có 4 động vật thí nghiệm và 4 công thức thí nghiệm khác nhau. Số liệu ở ô vuông la tinh thứ nhất như trong bài tập 4.5.3, ở ô vuông la tinh thứ 2 như trong bảng bên. Hãy tiến hành phân tích để đưa ra kết luận và đưa ra nhận xét về mô hình thiết kế trong bài tập 4.5.3 và bài tập 4.5.4.

Giai đoạn	Bê			
	1	2	3	4
1	10,9 (C)	11,2 (A)	9,4 (D)	11,2 (B)
2	10,5 (B)	9,6 (D)	11,4 (C)	10,9 (A)
3	11,1 (A)	11,4 (C)	11,7 (B)	9,8 (D)
4	8,8 (D)	12,9 (B)	11,4 (A)	11,2 (C)

Chương 5

THIẾT KẾ THÍ NGHIỆM HAI YẾU TỐ

Mục tiêu của chương này nhằm giới thiệu cho bạn đọc cách thiết kế thí nghiệm và phân tích kết quả đối với thí nghiệm hai yếu tố; các ưu, nhược điểm của mô hình này và các giải pháp hạn chế những nhược điểm này. Trong phần này, bốn kiểu thí nghiệm thường dùng sẽ được giới thiệu cùng bạn đọc gồm: (1) Thí nghiệm hai yếu tố chéo nhau (cross), hay hai yếu tố trực giao (orthogonal); (2) Thí nghiệm hai yếu tố phân cấp (hierachical), hay còn gọi là chia ô (nested); (3) Thí nghiệm hai yếu tố chia ô (split plot) và (4) Thí nghiệm hai yếu tố chia băng hay chia dài (strip plot).

5.1. THÍ NGHIỆM HAI YẾU TỐ CHÉO NHAU (Cross hay Orthogonal)

Trong thí nghiệm kiểu hai yếu tố chéo nhau, chúng ta tiến hành nghiên cứu đồng thời hai yếu tố thí nghiệm và kiểm định tất cả các tổ hợp giữa các mức khác nhau của các yếu tố thí nghiệm. Ngoài ảnh hưởng của từng yếu tố riêng biệt gọi là các yếu tố chính, còn có thể tìm thấy tác động cùng với nhau của 2 yếu tố gọi là tương tác. Mô hình này cũng được thiết kế hoàn toàn ngẫu nhiên vì vậy các đơn vị thí nghiệm được phân về với các tổ hợp của các yếu tố là hoàn toàn ngẫu nhiên. Giả sử yếu tố A có a mức, yếu tố B có b mức, tất cả có $a \times b$ công thức, mỗi công thức $a_i \times b_j$ ($i = 1, a$; $j = 1, b$), lặp lại r lần. Tất cả có $a \times b \times r = n$ đơn vị thí nghiệm.

Xem xét một thí nghiệm nhằm đánh giá ảnh hưởng của hàm lượng protein và các loại thức ăn đến sản lượng sữa của bò. Yếu tố thứ nhất là hàm lượng protein và yếu tố thứ 2 là các loại thức ăn. Protein được xác định ở 3 mức và có 2 loại thức ăn được sử dụng. Mỗi bò có khả năng tham gia vào một trong 6 tổ hợp (protein \times thức ăn). Thí nghiệm này được gọi là mô hình 2 yếu tố trực giao hay bắt chéo 3×2 vì có 3 mức của yếu tố thứ nhất và 2 mức của yếu tố thứ 2 đã được xác định. Mục đích của thí nghiệm là xác định phản ứng của bò khác nhau ở các mức protein khác nhau với các loại thức ăn khác nhau. Mục đích chính của thí nghiệm trực giao là có thể phân tích được tương tác của các yếu tố. Ngoài ra, mô hình này cũng đặc biệt hữu ích khi toàn bộ các yếu tố thí nghiệm và tổ hợp được tiến hành phân tích từ đó có thể kết luận tổ hợp nào là tốt nhất.

5.1.1. Ưu điểm và nhược điểm

Thiết kế thí nghiệm hai yếu tố theo kiểu chéo nhau có hiệu quả cao hơn so với mô hình thiết kế thí nghiệm một yếu tố. Nó có ưu điểm là có thể nghiên cứu đồng thời ảnh hưởng của từng yếu tố độc lập và ảnh hưởng của tương tác giữa các yếu tố. Mô hình này thật sự cần thiết khi tồn tại sự tương tác giữa các mức yếu tố nhằm tránh những kết luận sai lệch.

Trong mô hình thí nghiệm, tất cả các tổ hợp của mức yếu tố được bố trí và thực hiện. Như vậy khi các mức của từng yếu tố tăng lên một cách đáng kể thì số các tổ hợp

sẽ tăng lên một cách nhanh chóng; điều này sẽ kéo theo hàng loạt các vấn đề phức tạp đối với các nguyên vật liệu thí nghiệm. Thậm chí khi có các nguồn vật liệu thí nghiệm thì tổ chức thực hiện cũng gặp khó khăn.

Thiết kế thí nghiệm kiểu chéo nhau được khuyến cáo tối đa ở 4 mức đối với từng yếu tố thí nghiệm. Mô hình này không phải cách tiếp cận phù hợp nhất nếu muốn nghiên cứu rất nhiều mức đối với từng yếu tố.

5.1.2. Số đơn vị thí nghiệm cần thiết

Số đơn vị thí nghiệm cần thiết được chọn theo các tiêu chí đồng đều như đã nêu ở Chương 3. Số lượng cần đơn vị thí nghiệm cần thiết có thể được tính theo công thức sau:

Để loại bỏ giả thiết H_0 khi chênh lệch d giữa 2 giá trị trung bình bất kỳ ở yếu tố thí nghiệm A

$$\phi^2 = \frac{nbd^2}{2a\sigma^2}$$

Để loại bỏ giả thiết H_0 khi chênh lệch d giữa 2 giá trị trung bình bất kỳ ở yếu tố thí nghiệm B

$$\phi^2 = \frac{nad^2}{2b\sigma^2}$$

Để loại bỏ giả thiết H_0 khi chênh lệch d giữa 2 giá trị trung bình bất kỳ của tương tác giữa các mức yếu tố thí nghiệm A và B

$$\phi^2 = \frac{nd^2}{2\sigma^2[(a-1)(b-1)+1]}$$

5.1.3. Cách thiết kế thí nghiệm

Giả sử yếu tố A có a mức, yếu tố B có b mức, tất cả có $a \times b$ công thức, mỗi công thức $a_i \times b_j$ ($i = 1, a$; $j = 1, b$), lặp lại r lần. Tất cả có $a \times b \times r = n$ đơn vị thí nghiệm. Số đơn vị thí nghiệm (n) được phân một cách ngẫu nhiên vào $a \times b$ công thức.

Nếu bố trí thí nghiệm 2 yếu tố theo kiểu khối ngẫu nhiên đầy đủ thì mỗi lần lặp lại là một khối; mỗi khối chia $a \times b$ công thức (khối đầy đủ). Trong phân tích ngoài các tổng bình phương SS_{TO}, SS_A, SS_B, SS_{AB} còn có thêm SS_K (tổng bình phương của khối) sau đó mới đến SS_E.

Trường hợp đơn giản nhất của mô hình chéo nhau là yếu tố A có 2 mức A₁ và A₂, yếu tố B có 2 mức B₁ và B₂. Các tổ hợp có thể của các mức yếu tố là:

Yếu tố A	Yếu tố B	
	B ₁	B ₂
A ₁	A ₁ B ₁	A ₁ B ₂
A ₂	A ₂ B ₁	A ₂ B ₂

Nếu ở mỗi công thức thí nghiệm có 3 đơn vị thí nghiệm ($r = 4$) thì số động vật cần thiết sẽ là $2 \times 2 \times 4$. Giả sử số động vật thí nghiệm này được đánh số từ 1 đến 16; sau khi phân một cách ngẫu nhiên về với 4 tổ hợp có thể như trên ta sẽ có sơ đồ thiết kế thí nghiệm như sau:

	A ₁		A ₂	
	B ₁	B ₂	B ₁	B ₂
Động vật thí nghiệm số	7	12	3	13
	11	8	1	10
	2	6	15	5
	14	4	9	16

Kết thúc thí nghiệm, số liệu có thể ghi lại để dễ dàng và thuận tiện cho việc tính toán như sau:

	A ₁		A ₂	
	B ₁	B ₂	B ₁	B ₂
7	X ₁₁₁	12	X ₁₂₁	3
11	X ₁₁₂	8	X ₁₂₂	1
2	X ₁₁₃	6	X ₁₂₃	15
14	X ₁₁₄	4	X ₁₂₄	9
			X ₂₁₁	13
			X ₂₁₂	10
			X ₂₁₃	5
			X ₂₁₄	16
			X ₂₂₁	
			X ₂₂₂	
			X ₂₂₃	
			X ₂₂₄	

Dưới dạng tổng quát với a công thức thí nghiệm với số lần lặp là r ta có:

	A ₁		A ₂	
	B ₁	B ₂	B ₁	B ₂
X ₁₁₁		X ₁₂₁	X ₂₁₁	X ₂₂₁
X ₁₁₂		X ₁₂₂	X ₂₁₂	X ₂₂₂
...	
X _{11r}		X _{12r}	X _{21r}	X _{22r}

5.1.4. Mô hình phân tích

$$X_{ijk} = \mu + a_i + b_j + (ab)_{ij} + e_{ijk} \quad (i = 1, a; j = 1, b; k = 1, r)$$

μ là trung bình chung.

a_i là chênh lệch so với trung bình chung của mức A_i của yếu tố A, $\sum a_i = 0$.

b_j là chênh lệch so với trung bình chung của mức B_j của yếu tố B, $\sum b_j = 0$.

$(ab)_{ij}$ là chênh lệch so với trung bình chung của công thức $A_i B_j$ sau khi trừ bỏ chênh lệch a_i của mức A_i và chênh lệch b_j của mức B_j .

$$\sum_{i=1}^a ab_{ij} = 0 \text{ với mọi } j \text{ và } \sum_{j=1}^b ab_{ij} = 0 \text{ với mọi } i.$$

e_{ijk} là sai số ngẫu nhiên, giả sử các sai số e_{ijk} độc lập, phân phối chuẩn $N(0, \sigma^2)$.

5.1.5. Cách phân tích

Tính tổng bình phương toàn bộ (SS_{TO}) được cấu thành từ các tổng bình phương thành phần của yếu tố A (SS_A), yếu tố B (SS_B), tương tác giữa các yếu tố (SS_{AB}) và sai số ngẫu nhiên (SS_E)

$$SS_{TO} = SS_A + SS_B + SS_{AB} + SS_E$$

Các tổng bình phương được tính như sau:

$$SS_{TO} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r \left(x_{ijk} - \bar{x} \right)^2$$

$$SS_A = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r \left(\bar{x}_i - \bar{x} \right)^2 = br \sum_{i=1}^a \left(\bar{x}_i - \bar{x} \right)^2$$

$$SS_B = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r \left(\bar{x}_j - \bar{x} \right)^2 = ar \sum_{j=1}^b \left(\bar{x}_j - \bar{x} \right)^2$$

$$SS_{AB} = r \sum_{i=1}^a \sum_{j=1}^b \left(\bar{x}_{ij} - \bar{x} \right)^2 - SS_A - SS_B$$

$$SS_{TO} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r \left(x_{ijk} - \bar{x}_{ij} \right)^2$$

Hoặc có thể tính nhanh các tổng bình phương như sau:

Tính $n = a \times b \times r$; $ST = \sum \sum \sum x_{ijk}$; $SST = \sum \sum \sum x_{ijk}^2$; Số điều chỉnh $G = ST^2 / n$; Sau khi có các tổng $A_i B_j$ (gọi là y_{ij}), sắp xếp lại thành bảng hai chiều; từ bảng đó tính các tổng TA_i , tổng TB_j

$$SS_{TO} = SST - G$$

$$SS_A = \frac{1}{br} \sum_{i=1}^a TA_i^2 - G$$

$$SS_B = \frac{1}{ar} \sum_{j=1}^b TB_j^2 - G$$

$$SS_{AB} = \frac{1}{r} \sum_{i=1}^a \sum_{j=1}^b y_{ij}^2 - G - SS_A - SS_B$$

$$SS_E = SS_{TO} - SS_B - SS_A - SS_{AB}$$

Các bậc tự do $df_{TO} = abr - 1$; $df_A = a - 1$; $df_B = b - 1$; $df_{AB} = (a-1)(b-1)$ và $df_E = ab(r-1)$.

Chia các tổng bình phương cho các bậc tự do tương ứng được các bình phương trung bình.

$$MS_A = SS_A / df_A; MS_B = SS_B / df_B; MS_{AB} = SS_{AB} / df_{AB}; MS_E = SS_E / df_E;$$

Chia MS_A , MS_B , MS_{AB} cho MS_E được các giá trị F thực nghiệm F_{TNA} , F_{TNB} , F_{TNAB} . Các giá trị F tới hạn của yếu tố A là $F_{(a, df_A, df_E)}$; B là $F_{(b, df_B, df_E)}$ và A×B là $F_{(ab, df_{AB}, df_E)}$. So với các giá trị tới hạn có thể kiểm định ba giả thiết theo nguyên tắc $F_{TN} > F_{tới hạn} \Rightarrow$ bác bỏ H_0 và chấp nhận đối thiêt H_1 :

H_0A : “ Các a_i bằng không” đối thiết H_{1A} : “ Có a_i khác 0” .

H_0B : “ Các b_j bằng không” đối thiết H_{1B} : “ Có b_j khác 0” .

H_{0AB} : “ Các ab_{ij} bằng không” đối thiết H_{1AB} : “ Có ab_{ij} khác 0” .

Dưới dạng tổng hợp ta có bảng phân tích phương sai

Nguồn biến động	df	SS	MS	F _{TN}	F tới hạn
Yếu tố A	a-1	SS _A	MS _A	MS _A / MS _E	F _(α, dfA, dfE)
Yếu tố B	b-1	SS _B	MS _B	MS _B / MS _E	F _(α, dfB, dfE)
Tương tác AxB	(a-1)(b-1)	SS _{AB}	MS _{AB}	MS _{AB} / MS _E	F _(α, dfAB, dfE)
Sai số	ab(r-1)	SS _E	MS _E		
Toàn bộ	abr -1	SS _{TO}			

Ví dụ 5.1: Một nghiên cứu được tiến hành để xác định ảnh hưởng của việc bổ sung 2 loại vitamin (A và B) vào thức ăn đến tăng khối lượng (kg/ngày) của lợn. Hai mức đối với vitamin A (0 và 4 mg) và 2 mức đối với vitamin B (0 và 5 mg) được sử dụng trong thí nghiệm này. Tổng số 20 lợn thí nghiệm được phân về 4 tổ hợp (công thức thí nghiệm) một cách ngẫu nhiên. Số liệu thu được khi kết thúc thí nghiệm được trình bày như sau:

Vitamin A	0 mg		4 mg	
	Vitamin B	0 mg	5 mg	0 mg
	0,585	0,567	0,473	0,684
	0,536	0,545	0,450	0,702
	0,458	0,589	0,869	0,900
	0,486	0,536	0,473	0,698
	0,536	0,549	0,464	0,693
Tổng	2,601	2,786	2,729	3,677
Trung bình	0,520	0,557	0,549	0,735

Các tổng bình phương được tính như sau:

$$ST = \sum \sum \sum x_{ijk} = 0,595 + \dots + 0,693 = 11,793$$

$$SST = \sum \sum \sum x_{ijk}^2 = 0,595^2 + \dots + 0,693^2 = 7,275437$$

$$G = ST^2 / n = 11,793^2 / 20 = 6,953742$$

$$TA_0 = 2,601 + 2,786 = 5,387 \text{ và } TA_4 = 2,729 + 3,677 = 6,406$$

$$TB_0 = 2,601 + 2,729 = 5,330 \text{ và } TB_5 = 2,786 + 3,677 = 6,463$$

$$T_{A0B0} = 2,601; T_{A0B5} = 2,786; T_{A4B0} = 2,729; T_{A4B5} = 3,677;$$

$$SS_{TO} = SST - G = 7,275437 - 6,953742 = 0,32169455$$

$$SS_A = \frac{1}{br} \sum_{i=1}^a TA_i^2 - G = (1/10) \times (5,387^2 + 6,406^2) - 6,953742 = 0,05191805$$

$$SS_B = \frac{1}{ar} \sum_{j=1}^b TB_j^2 - G = (1/10) \times (5,330^2 + 6,463^2) - 6,953742 = 0,06418445$$

$$SS_{AB} = \frac{1}{r} \sum_{i=1}^a \sum_{j=1}^b y_{ij}^2 - G - SS_A - SS_B = \frac{1}{5} \times (2,601^2 + 2,786^2 + 2,729^2 + 3,677^2) - \\ 6,953742 - 0,05191805 - 0,06418445 = 0,02910845$$

$$SS_E = SS_{TO} - SS_A - SS_B - SS_{AB} = 0,32169455 - 0,05191805 - 0,06418445 - 0,02910845 = 0,17648360$$

Có thể tổng hợp vào bảng phân tích phương sai sau:

Nguồn biến động	df	SS	MS	F _{TN}	F
Vitamin A	1	0,05191805	0,05191805	4,71	F _(0,05; 1; 16) = 4,49
Vitamin B	1	0,06418445	0,06418445	5,82	F _(0,05; 1; 16) = 4,49
Vit A × Vit B	1	0,02910845	0,02910845	2,64	F _(0,05; 1; 16) = 4,49
Sai số	16	0,17648360	0,01103023		
Toàn bộ	19	0,32169455			

Kết luận: Bổ sung vitamin A và B đã làm cho tăng khói lượng của lợn thay đổi (vì F_{TN} > 4,49 ở mức α = 0,05); tuy nhiên không có tương tác giữa các yếu tố (vì F_{TN} < 4,49 ở mức α = 0,05).

5.2. THÍ NGHIỆM HAI YẾU TỐ PHÂN CẤP (Hierachical hay Nested)

Kiểu thí nghiệm hai yếu tố phân cấp (Hierachical) hay chia ô (Nested) thường được dùng trong các nghiên cứu về di truyền. Trong đó một yếu tố là cấp trên, một yếu tố là cấp dưới, thí nghiệm lặp lại r lần.

Cụ thể xét thí dụ A là bò đực giống, tất cả có 4 con A₁, A₂, A₃, A₄. Mỗi con đực cho phối với 3 con cái gọi tắt là B₁, B₂, B₃. Mỗi con bò cái sinh 4 con. Ta có sơ đồ sau:

A	1			2			3			4		
	1	2	3	1	2	3	1	2	3	1	2	3
B	X111	X121	X131	X211	X221	X131	X311	X321	X331	X411	X421	X431
	X112	X122	X132	X212	X222	X232	X312	X322	X332	X412	X422	X432
	X113	X123	X133	X213	X223	X233	X313	X323	X333	X413	X423	X433
	X114	X124	X134	X214	X224	X234	X314	X324	X334	X414	X424	X434

Cần phải chú ý là 3 con cái cho phối với con đực B₁ khác với 3 con cái cho phối với con đực B₂, khác với 3 con cái cho phối với con đực B₃, khác với 3 con cái cho phối với con đực B₄.

Mỗi cặp bò mẹ sinh được 4 con. Như vậy chúng ta có mô hình phân cấp với con đực là cấp trên, mỗi con đực phối với 3 cái là cấp dưới, mỗi cặp bò mẹ có 4 con là cấp dưới nữa. Cũng có thể coi như có 4 ô, mỗi ô có một con đực và 3 con cái, mỗi cặp vợ chồng có 4 con.

Để thống nhất ký hiệu chúng ta coi yếu tố thứ nhất (A) là cấp trên có a mức, yếu tố thứ 2 (B) là cấp dưới có b mức và mỗi công thức A_iB_j lặp lại r lần.

5.2.1. Ưu và nhược điểm của mô hình

Trong thí nghiệm hai yếu tố phân cấp, các đơn vị thí nghiệm của yếu tố thứ hai trong cùng một mức của yếu tố thứ nhất sẽ độc lập với các đơn vị tương tự nhưng nằm khác mức của yếu tố thứ nhất.

Ta có thể so sánh sự khác nhau giữa các mức của yếu tố thí nghiệm cấp trên và ảnh hưởng giữa các mức khác nhau của yếu tố cấp dưới trong cùng một mức của yếu tố thứ nhất nhưng không thể so sánh sự khác nhau giữa các mức của yếu tố nằm trong các mức khác nhau của yếu tố thứ nhất. Ví dụ ta có thể so sánh 4 con đực với nhau, so sánh các con cái được phối với cùng một đực nhưng không thể so sánh sự khác nhau giữa các con cái được phối với các con đực khác nhau.

5.2.2. Cách thiết kế thí nghiệm

Trong $a \times b$ mức của A phải bắt thăm để xem mức nào gọi là A_1 , mức nào là A_2, \dots, A_a . Trong $a \times b$ cá thể (tương đối đồng đều) phải bắt thăm b cá thể làm cấp dưới cho A_1 , sau đó bắt thăm b cá thể cho A_2, \dots, A_a . Mỗi cặp $A_i B_j$ ($i = 1, a; j = 1, b$) có r lần lặp (tức là thu được r số liệu) ký hiệu là x_{ijk} .

5.2.3. Mô hình

$$x_{ijk} = \mu + a_i + b_{j(i)} + e_{ijk} \quad (i = 1, a; j = 1, b; k = 1, r)$$

μ là trung bình chung.

a_i là chênh lệch do ảnh hưởng của mức A_i của yếu tố A; $\sum a_i = 0$.

$b_{j(i)}$ là chênh lệch do ảnh hưởng của mức B_j (trong \hat{A}_i) của yếu tố B; $\sum b_{j(i)} = 0$ với mọi i.

e_{ijk} là sai số ngẫu nhiên; giả sử các e_{ijk} độc lập phân phối chuẩn $N(0, \sigma^2)$.

5.2.4. Cách phân tích

$$\text{Gọi } n = a \times b \times r; \quad ST = \sum \sum \sum x_{ijk}; \quad SST = \sum \sum \sum x_{ijk}^2$$

$$\text{Số điều chỉnh } G = ST^2 / n; \quad TAB_{ij} = \sum_{k=1}^r x_{ijk}; \quad TA_i = \sum_{j=1}^b \sum_{k=1}^r x_{ijk}$$

Tổng bình phương toàn bộ

$$SS_{TO} = SST - G$$

Tổng bình phương do yếu tố A

$$SS_A = (\sum_{i=1}^a TA_i^2) / (b \times r) - G$$

Tổng bình phương do yếu tố B trong A

$$SS_{B(A)} = \left(\sum_{i=1}^a \sum_{j=1}^b TAB_{ij}^2 \right) / r - \left(\sum_{i=1}^a TA_i^2 \right) / (b \times r) = \left(\sum_{i=1}^a \sum_{j=1}^b TAB_{ij}^2 \right) / r - G - SS_A$$

Tổng bình phương do sai số

$$SS_E = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r x_{ijk}^2 - \left(\sum_{i=1}^a \sum_{j=1}^b TA_{ij}^2 \right) / r = SS_{TO} - SS_A - SS_B$$

Các bậc tự do $df_{TO} = abr - 1$; $df_A = a-1$; $df_B = a(b-1)$ và $df_E = ab(r-1)$. Chia các tổng bình phương cho bậc tự do tương ứng được các bình phương trung bình:

$$MS_A = SS_A / df_A; MS_{B(A)} = SS_{B(A)} / df_{B(A)}; MS_E = SS_E / df_E$$

$$F_{TN_A} = MS_A / MS_{B(A)} \text{ so với giá trị tối hạn } F_{(\alpha, df_A, df_{B(A)})}$$

$$F_{TN_B} = MS_{B(A)} / MS_E \text{ so với giá trị tối hạn } F_{(\alpha, df_{B(A)}, df_E)}$$

Nếu $F_{TN} > F$ tối hạn, H_0 sẽ bị bác bỏ

Dưới dạng tổng hợp ta có bảng phân tích phương sai

Nguồn biến động	df	SS	MS	F_{TN}	F
Yếu tố A	a-1	SS_A	MS_A	$MS_A / MS_{B(A)}$	$F_{(\alpha, df_A, df_{B(A)})}$
Yếu tố B trong A	a(b-1)	$SS_{B(A)}$	$MS_{B(A)}$	$MS_{B(A)} / MS_E$	$F_{(\alpha, df_{B(A)}, df_E)}$
Sai số ngẫu nhiên	$ab(r-1)$	SS_E	MS_E		
Toàn bộ	$abr - 1$	SS_{TO}			

Các ước tính của trung bình bình phương $E(MS)$ được xác định tương ứng khi yếu tố A và B là cố định hay ngẫu nhiên như sau:

$E(MS)$	A và B cố định	A cố định và B ngẫu nhiên	A và B ngẫu nhiên
$E(MS_A)$	$\sigma^2 + Q(A)$	$\sigma^2 + r\sigma^2_B + Q(A)$	$\sigma^2 + r\sigma^2_B + rba\sigma^2_A$
$E(MS_{B(A)})$	$\sigma^2 + Q(B(A))$	$\sigma^2 + r\sigma^2_B$	$\sigma^2 + r\sigma^2_B$
$E(MS_E)$	σ^2	σ^2	σ^2

Trong chăn nuôi, nhân cấp trên được giả thiết là *cố định* nếu tất cả các con đực hiện có là những con cự thể hoặc giả thiết là *ngẫu nhiên* nếu con đực được chọn ngẫu nhiên từ số đực giống trong đàn, yếu tố cấp dưới được giả thiết là *ngẫu nhiên* vì con cái luôn được chọn ngẫu nhiên trong đàn. Từ đó ước lượng được các phương sai thành phần: *phương sai* σ^2 của sai số e_{ijk} , *phương sai* σ^2_B của biến ngẫu nhiên “cái” và *phương sai* σ^2_A của biến ngẫu nhiên “đực”. Từ các phương sai thành phần này có thể tính được hệ số di truyền theo bố hoặc theo mẹ.

Ví dụ 5.2: Mục đích của thí nghiệm là xác định ảnh hưởng của lợn đực giống và lợn nái đến khối lượng sơ sinh của thê hệ con. Mô hình phân cấp 2 yếu tố được sử dụng. Bốn lợn đực giống được chọn ngẫu nhiên ($a = 4$), mỗi đực phối với 3 lợn nái ($b = 3$) và mỗi nái sinh được 2 lợn con ($r = 2$). Khối lượng (kg) sơ sinh của từng lợn con thu được như sau:

Đực	1			2			3			4		
Nái	1	2	3	4	5	6	7	8	9	10	11	12
	1,2	1,2	1,1	1,2	1,1	1,2	1,2	1,3	1,2	1,3	1,4	1,3
	1,2	1,3	1,2	1,2	1,2	1,1	1,2	1,3	1,2	1,3	1,4	1,3

Ta có bảng phân tích phương sai:

Nguồn biến động	df	SS	MS	F _{TN}	F tới hạn
Đực	3	0,093333	0,031111	6,22	F _(0,05; 3; 8) = 4,07
Cái (cùng đực)	8	0,040000	0,005000	3,00	F _(0,05; 8; 12) = 2,85
Sai số ngẫu nhiên (Con cùng bố mẹ)	12	0,020000	0,001667		
Toàn bộ	23	0,153333			

Kết luận: Ta thấy các giá trị F thực nghiệm đều lớn hơn giá trị F tới hạn, chứng tỏ có sự sai khác giữa các con đực và giữa các nái cùng đực.

Theo như ví dụ đã nêu; đực giống và nái là các yếu tố ngẫu nhiên, vì vậy các giá trị của phương sai thành phần được ước tính trong bảng sau:

Nguồn biến động	E(MS)	Phương sai thành phần	Phần trăm so với toàn bộ biến động
Đực	$\sigma^2 + 2\sigma^2_B + 6\sigma^2_A$	0,004352	56,63
Cái cùng đực	$\sigma^2 + 2\sigma^2_B$	0,001667	21,69
Sai số ngẫu nhiên	σ^2	0,001667	21,69
Tổng số	σ^2_T	0,007685	100,00

Từ các phương sai thành phần này ta có thể tính được hệ số di truyền. Tuy nhiên để ước tính hệ số di truyền một cách chính xác thì bậc tự do của các nguồn biến động phải đủ lớn. Tức là thí nghiệm phải bố trí trên nhiều đực, cái và số lượng quan sát ở đời con cũng phải đủ lớn. Trong di truyền số lượng, mô hình này cũng được đặc biệt chú trọng.

5.3. THÍ NGHIỆM HAI YẾU TỐ CHIA Ô

Thí nghiệm hai yếu tố chia ô thích hợp để nghiên cứu ảnh hưởng của 2 yếu tố bố trí theo cách sau. Nguyên vật liệu thí nghiệm chia thành một số các ô lớn và các mức của yếu tố thứ nhất được bố trí ngẫu nhiên vào các ô lớn. Sau đó, mỗi ô lớn lại được chia thành các ô con và các mức của yếu tố thứ 2 được bố trí ngẫu nhiên vào các ô con.

Mô hình thí nghiệm hai yếu tố chia ô được sử dụng khi một yếu tố cần nhiều nguyên vật liệu hơn yếu tố thứ hai. Nếu một yếu tố được áp dụng muộn hơn so với yếu tố còn lại thì yếu tố muộn hơn sẽ được bố trí vào ô con. Ngoài ra, từ kinh nghiệm thực tế ta biết được một yếu tố có mức độ biến động lớn hơn thì yếu tố ngày sẽ được bố trí vào ô lớn. Hoặc ta muốn có một kết luận chính xác đối với một yếu tố thì yếu tố đó được bố trí vào ô nhỏ. Yếu tố trên ô lớn có sai số gọi là sai số ô lớn, yếu tố trên ô nhỏ có sai số gọi là sai số ô nhỏ.

5.3.1. Ưu và nhược điểm của mô hình

Thí nghiệm chia ô có cách phân tích phức tạp hơn hai thí nghiệm giao nhau hay phân cấp. Mức chính xác của hai yếu tố khác nhau, yếu tố trên ô lớn có độ chính xác thấp hơn yếu tố trên ô nhỏ.

Thí nghiệm này rất phù hợp nếu ta chỉ quan tâm đến một trong hai yếu tố và tương tác giữa chúng. Ví dụ, nghiên cứu ảnh hưởng của các loại thức ăn khác nhau đến tăng khối lượng của vật nuôi, đồng thời cũng quan tâm đến tương tác của thức ăn với giới tính.

Trong các nghiên cứu về nông nghiệp mô hình này cũng được sử dụng rộng rãi, trong một khu diện tích lớn đất được coi như một ô lớn và những lô được chia ra được gọi là ô nhỏ.

Mô hình này sẽ gặp khó khăn trong việc ước tính nếu số liệu bị khiếm khuyết. Số bậc tự do của sai số ngẫu nhiên bị giảm rất nhiều do có hai lần tương tác (tương tác giữa hai yếu tố A×B và tương tác giữa yếu tố A với khối hay còn gọi là sai số ô lớn), chính vì vậy cũng làm giảm độ chính xác của các ước lượng và các kết luận.

5.3.2. Cách thiết kế thí nghiệm

Thường bố trí thí nghiệm theo khối, mỗi khối chia thành a ô lớn để bắt thăm cho a mức của yếu tố A. Việc bắt thăm được thực hiện riêng rẽ cho từng khối. Mỗi ô lớn chia thành b ô nhỏ để bắt thăm cho b mức của yếu tố B. Việc bắt thăm thực hiện riêng rẽ cho từng ô lớn.

Thí dụ yếu tố A có 4 mức (A_1, A_2, A_3 và A_4), yếu tố B có 2 mức (B_1 và B_2). Ba mức của yếu tố A được bố trí trên ô lớn trong 3 khối. Mỗi ô lớn chia nhỏ thành 2 ô nhỏ để bố trí ngẫu nhiên các mức của yếu tố B. Sơ bố trí thí nghiệm có thể được trình bày như sau:

Khối 1				Khối 2				Khối 3			
A_4	A_1	A_2	A_3	A_2	A_1	A_4	A_3	A_1	A_2	A_4	A_3
B_2	B_2	B_1	B_2	B_1	B_2	B_1	B_1	B_2	B_1	B_2	B_1
B_1	B_1	B_2	B_1	B_2	B_1	B_2	B_2	B_1	B_2	B_1	B_2

5.3.3. Mô hình

$$x_{ijl} = \mu + a_i + k_1 + (ak)_{il} + b_j + (ab)_{ij} + e_{ijl}; (i = 1, a; j = 1, b; l = 1, r)$$

Trong đó:

μ là trung bình chung.

a_i là chênh lệch do ảnh hưởng của mức i của yếu tố A (trên ô lớn); $\sum a_i = 0$.

b_j là chênh lệch do ảnh hưởng của mức j của yếu tố B (trên ô nhỏ); $\sum b_j = 0$.

k_l là chênh lệch do ảnh hưởng của khối l ; $\sum k_l = 0$.

$(ak)_{il}$ là tương tác giữa yếu tố A và khối i và được dùng làm sai số ô lớn s^2_L .

$(ab)_{ij}$ là tương tác của hai yếu tố A và B.

$$\sum_{j=1}^b (ab)_{ij} = 0 \text{ với mọi } i; \quad \sum_{i=1}^a (ab)_{ij} = 0 \text{ với mọi } j$$

e_{ijk} là sai số độc lập phân phối chuẩn $N(0, \sigma^2)$.

Trong mô hình này khối coi như yếu tố ngẫu nhiên, không tương tác với B. Hai yếu tố A và B coi như yếu tố cố định.

5.3.4. Cách phân tích

$$n = a \times b \times r; \quad ST = \sum \sum \sum x_{ijl}; \quad SST = \sum \sum \sum x_{ijl}^2; \quad G = ST^2 / n;$$

Từ bảng số liệu gốc tính tổng các x_{ijl} theo j được TAC_{ik} sau đó lập bảng hai chiều A x K. Từ bảng số liệu gốc lấy tổng các x_{ijl} theo k được TAB_{ij} sau đó lập bảng hai chiều A x B.

Các tổng bình phương được tính như sau:

Tổng bình phương toàn bộ:

$$SS_{TO} = SST - G$$

Tổng bình phương của khối:

$$SS_K = (\sum TK^2_l) / (a \times b) - G$$

Tổng bình phương của yếu tố A

$$SS_A = (\sum TA^2_i) / (b \times r) - G$$

Tổng bình phương tương tác giữa yếu tố A và khối (sai số ô lớn)

$$SS_{AK} = (\sum \sum TAK^2_{il}) / b - G - SSA - SSK$$

Tổng bình phương của yếu tố B

$$SS_B = (\sum TB^2_j) / (a \times r) - G$$

Tổng bình phương tương tác giữa yếu tố A và B

$$SS_{AB} = (\sum \sum TAB^2_{ij}) / r - G - SSA - SSB$$

Tổng bình phương của sai số ngẫu nhiên (sai số ô nhỏ)

$$SS_E = SS_{TO} - SSA - SSK - SS_{AK} - SS_B - SS_{AB}$$

Với các bậc tự do $df_{TO} = axb \times r - 1$; $df_K = r - 1$; $df_A = a - 1$; $df_{AK} = (a - 1)(r - 1)$; $df_B = b - 1$; $df_{AB} = (a - 1)(b - 1)$; $df_E = a(b - 1)(r - 1)$. Chia các tổng bình phương cho bậc tự do tương ứng được các bình phương trung bình (MS):

$$MS_A = SSA / df_A; \quad MS_B = SSB / df_B; \quad MS_{AB} = SS_{AB} / df_{AB}; \quad MS_E = SS_E / df_E$$

Ta có các giá trị F tương ứng:

$FTNA = MSA / MSAK$ so với giá trị tới hạn $F(\alpha, dfA, dfAK)$.

$FTNB = MSB / MSE$ so với giá trị tới hạn $F(\alpha, dfB, dfE)$.

$FTNAB = MSAB / MSE$ so với giá trị tới hạn $F(\alpha, dfAB, dfE)$.

Nếu $FTN > F$ tới hạn, H_0 sẽ bị bác bỏ.

Kiểm định giả thiết đối với yếu tố trên ô lớn (A)

H_0A : “các ai đều bằng 0” với đối thiết H_1A : “có ai khác 0”.

Kiểm định giả thiết đối với yếu tố trên ô nhỏ (B)

H_0B “Các bj đều bằng 0” với đối thiết H_1B “có bj khác 0”

Kiểm định giả thiết đối với tương tác giữa A và B

H_0AB : “Các $(ab)ij$ đều bằng 0” với đối thiết H_1AB “có $(ab)ij$ khác 0”.

Dưới dạng tổng hợp ta có bảng phân tích phương sai

Nguồn biến động	df	SS	MS	F_{TN}	F
Khối	r - 1	SS_K			
Yếu tố A	a-1	SS_A	MS_A	MS_A / MS_{AK}	$F_{(\alpha, dfA, dfAK)}$
Sai số ô lớn	$(r - 1)(a - 1)$	SS_{AK}	MS_{AK}		
Yếu tố B	(b-1)	SS_B	MS_B	MS_B / MS_E	$F_{(\alpha, dfB, dfE)}$
Tương tác AB	$(a - 1)(b - 1)$	SS_{AB}	MS_{AB}	MS_{AB} / MS_E	$F_{(\alpha, dfAB, dfE)}$
Sai số ô nhỏ	$a(b - 1)(r - 1)$	SS_E	MS_E		
Toàn bộ	$a \times b \times r - 1$	SS_{TO}			

Ví dụ 5.3: Một thí nghiệm được tiến hành để nghiên cứu ảnh hưởng của bãi chăn thả A (1, 2, 3 và 4) và lượng khoáng bổ sung B (1 và 2) đến năng suất sữa. Có tất cả 24 bò tham gia thí nghiệm. Thí nghiệm được thiết kế theo mô hình hai yếu tố kiểu chia ô với yếu tố A được bố trí trên ô lớn và yếu tố B trên ô nhỏ trên 3 khối. Năng suất sữa trung bình được ghi lại như sau (kg /ngày):

Khối 1				Khối 2				Khối 3			
A ₄	A ₁	A ₂	A ₃	A ₂	A ₁	A ₄	A ₃	A ₁	A ₂	A ₄	A ₃
B ₂ 30	B ₂ 27	B ₁ 26	B ₂ 26	B ₁ 32	B ₂ 30	B ₁ 34	B ₁ 33	B ₂ 34	B ₁ 30	B ₂ 36	B ₁ 33
B ₁ 29	B ₁ 25	B ₂ 28	B ₁ 24	B ₂ 37	B ₁ 31	B ₂ 37	B ₂ 32	B ₁ 31	B ₂ 31	B ₁ 38	B ₂ 32

Ta có:

$$n = a \times b \times r = 4 \times 2 \times 3 = 24;$$

$$ST = \sum \sum \sum x_{ijl} = 39 + \dots + 32 = 746;$$

$$SST = \sum \sum \sum x_{ijl}^2 = 30^2 + \dots + 32^2 = 23530;$$

$$G = ST^2 / n = 746^2 / 24 = 23188,167;$$

$$\Sigma T K^2_i = (30 + \dots + 24)^2 + (32 + \dots + 32)^2 + (34 + \dots + 32)^2 = 187206$$

$$\begin{aligned} \Sigma T A^2_i &= (27 + \dots + 31)^2 + (26 + \dots + 31)^2 + (26 + \dots + 32)^2 + (30 + \dots + 38)^2 \\ &= 139556 \end{aligned}$$

$$\Sigma \Sigma TAK^2_{il} = (27 + 25)^2 + (26 + 28)^2 + \dots + (36 + 38)^2 = 46996$$

$$\Sigma TB^2_j = (29 + 25 + \dots + 33)^2 + (30 + 27 + \dots + 32)^2 = 278356$$

$$\Sigma \Sigma TAB^2_{ij} = (25 + 31 + 31)^2 + (27 + 30 + 34)^2 + \dots + (30 + 37 + 36)^2 = 69820$$

Các tổng bình phương được tính như sau:

Tổng bình phương tổng số

$$SS_{TO} = SST - G = 23530 - 23188,167 = 341,833$$

Tổng bình phương của khối

$$SS_K = (\Sigma TK^2_{il}) / (a \times b) - G = 187206 / (4 \times 2) - 23188,167 = 212,583$$

Tổng bình phương của yếu tố A

$$SS_A = (\Sigma TA^2_{ij}) / (b \times r) - G = 139556 / (2 \times 3) - 23188,167 = 71,167$$

Tổng bình phương tương tác giữa yếu tố A và khối (sai số ô lớn)

$$SS_{AK} = (\Sigma \Sigma TAK^2_{il}) / b - G - SSA - SSK = 46996 / 2 - 23188,167 - 71,167 - 212,583 = 26,083$$

Tổng bình phương của yếu tố B

$$SS_B = (\Sigma TB^2_j) / (a \times r) - G = 278356 / (4 \times 3) - 23188,167 = 8,167$$

Tổng bình phương tương tác giữa yếu tố A và B

$$SS_{AB} = (\Sigma \Sigma TAB^2_{ij}) / r - G - SSA - SS_B = 69820 / 3 - 23188,167 - 71,167 - 8,167 = 5,833$$

Tổng bình phương của sai số ngẫu nhiên (sai số ô nhỏ)

$$SS_E = SS_{TO} - SSA - SSK - SS_B - SS_{AB} = 341,833 - 71,167 - 212,583 - 26,083 - 8,167 - 5,833 = 18,000$$

Với các bậc tự do:

$$df_{TO} = a \times b \times r - 1 = 23; df_K = r - 1 = 2; df_A = a - 1 = 3;$$

$$df_{AK} = (a - 1)(r - 1) = 6; df_B = b - 1 = 1;$$

$$df_{AB} = (a - 1)(b - 1) = 3; df_E = a(b - 1)(r - 1) = 8.$$

Bảng phân tích phương sai

Nguồn biến động	df	SS	MS	F _{TN}	F tối hạn
Khối	2	212,583	106,292		
Bài chẵn thả (A)	3	71,167	23,722	5,46	F _(0,05; 3; 6) = 4,76
Sai số ô lớn	6	26,083	4,347		
Khoảng bỗ sung (B)	1	8,167	8,167	3,63	F _(0,05; 1; 8) = 5,32
Tương tác AxB	3	5,833	1,944	0,86	F _(0,05; 3; 8) = 4,07
Sai số ô nhỏ	8	18,000	2,250		
Toàn bộ	23	341,833			

Kết luận: Qua kết quả phân tích được trình bày ở bảng nêu trên ta thấy, năng suất sữa có sự khác nhau giữa các bãі chǎn thǎ (F_{TN} = 5,46 > F_{LT} = 4,76), tuy nhiên việc bổ sung các khoáng chất không làm ảnh hưởng đến năng suất sữa và cũng không có ảnh hưởng tương tác giữa bãі chǎn thǎ và việc bổ sung khoáng.

5.4. THÍ NGHIỆM HAI YẾU TỐ CHIA Ô HOÀN TOÀN NGẪU NHIÊN

Phản trước, ta đã nghiên cứu mô hình kiểu chia ô mà các ô lớn được bố trí trên các khối một cách ngẫu nhiên. Ngoài ra cũng có thể thiết kế để một yếu tố được bố trí ngẫu nhiên trên các ô lớn. Ví dụ yếu tố thứ nhất (A) có 4 mức (A₁, A₂, A₃ và A₄) được bố trí ngẫu nhiên trên 12 ô lớn. Mỗi mức của yếu tố A được lặp lại 3 lần (r = 3). Yếu tố thứ hai (B) có 2 mức (B₁ và B₂). Mỗi ô lớn được chia thành 2 ô con để bố trí ngẫu nhiên các mức của yếu tố B. Đây chính là mô hình thí nghiệm 2 yếu tố kiểu chia ô hoàn toàn ngẫu nhiên. Mô hình bố trí thí nghiệm có thể được trình bày như sau:

A ₄ B ₂ B ₁	A ₁ B ₂ B ₁	A ₂ B ₁ B ₂	A ₃ B ₂ B ₁	A ₂ B ₁ B ₂	A ₁ B ₂ B ₁	A ₄ B ₁ B ₂	A ₃ B ₁ B ₂	A ₁ B ₂ B ₁	A ₂ B ₁ B ₂	A ₄ B ₂ B ₁	A ₃ B ₁ B ₂
--	--	--	--	--	--	--	--	--	--	--	--

Ta sẽ có mô hình phân tích số liệu như sau:

$$x_{ijl} = \mu + a_i + o_{k(i)} + b_j + (ab)_{ij} + e_{ijl}; (i = 1, a; j = 1, b; k = 1, r)$$

μ là trung bình chung.

a_i là chênh lệch do ảnh hưởng của mức i của yếu tố A (trên ô lớn); $\sum a_i = 0$.

b_j là chênh lệch do ảnh hưởng của mức j của yếu tố B (trên ô nhỏ); $\sum b_j = 0$.

$o_{k(i)}$ là chênh lệch do ảnh hưởng của ô lớn k trong mức i của yếu tố A (sai số ô lớn); $\sum o_{k(i)} = 0$.

$(ab)_{ij}$ là tương tác của hai yếu tố A và B

$$\sum_{j=1}^b (ab)_{ij} = 0 \text{ với mọi } i; \quad \sum_{i=1}^a (ab)_{ij} = 0 \text{ với mọi } j$$

e_{ijk} là sai số độc lập phân phối chuẩn $N(0, \sigma^2)$.

Trong mô hình này hai yếu tố A và B coi như yếu tố cố định. Các tổng bình phương của yếu tố A, B, tương tác AB, sai số ngẫu nhiên (sai số ô bé) và các bậc tự do tương ứng được tính tương tự như ở phần 4.3.3. Tổng bình phương của ô lớn nằm trong yếu tố A ($SS_{O(A)}$) được tính theo công thức $SS_{O(A)} = (\Sigma \Sigma TAO_{ik}^2) / b - G - SSA$ và bậc tự do $df_{O(A)} = a(r-1)$.

Tương tự như phần 4.3.3 ta có bảng phân tích phương sai:

Nguồn biến động	df	SS	MS	F _{TN}	F
Yếu tố A	a-1	SS _A	MS _A	MS _A / MS _{O(A)}	F _{(a, dfA, dfO(A))}
Sai số ô lớn	a(r-1)	SS _{O(A)}	MS _{O(A)}		
Yếu tố B	(b-1)	SS _B	MS _B	MS _B / MS _E	F _(b-1, dfB, dfE)
Tương tác AxB	(a-1)(b-1)	SS _{AB}	MS _{AB}	MS _{AB} / MS _E	F _(a-1, b-1, dfAB, dfE)
Sai số ô nhỏ	a(b-1)(r-1)	SS _E	MS _E		
Toàn bộ	axbxr-1	SS _{TO}			

Kết luận cũng tiến hành tương tự như các bước kết luận ở mục 5.3.4.

Ví dụ 5.4: Ta lấy lại ví dụ ở mục 5.3.4. Ảnh hưởng của bãi chăn thả A (1, 2,3 và 4) và lượng khoáng bổ sung B (1 và 2) đến năng suất sữa. Có tất cả 24 bò tham gia thí nghiệm. Tuy nhiên trong thí nghiệm này, khói sẽ không có mà ta có 12 ô lớn để bố trí ngẫu nhiên các mức của yếu tố bãi chăn thả, mỗi mức được lặp lại 3 lần. Năng suất sữa trung bình được ghi lại như sau (kg /ngày):

A ₄ B ₂ 30 B ₁ 29	A ₁ B ₂ 27 B ₁ 25	A ₂ B ₁ 26 B ₂ 28	A ₃ B ₂ 26 B ₁ 24	A ₂ B ₁ 32 B ₂ 37	A ₁ B ₂ 30 B ₁ 31	A ₄ B ₁ 34 B ₂ 37	A ₃ B ₁ 33 B ₂ 32	A ₁ B ₂ 34 B ₁ 31	A ₂ B ₁ 30 B ₂ 31	A ₄ B ₂ 36 B ₁ 38	A ₃ B ₁ 33 B ₂ 32
--	--	--	--	--	--	--	--	--	--	--	--

Ta có bảng phân tích phương sai sau:

Nguồn biến động	df	SS	MS	F _{TN}	F tối hạn
Bãi chăn thả (A)	3	71,167	23,722	0,80	F _(0,05; 3; 8) = 4,07
Sai số ô lớn	8	238,667	29,883		
Khoáng bổ sung (B)	1	8,167	8,167	3,63	F _(0,05; 1; 8) = 5,32
Tương tác AxB	3	5,833	1,944	0,86	F _(0,05; 3; 8) = 4,07
Sai số ô nhỏ	8	18,000	2,250		
Toàn bộ	23	341,833			

Kết luận: Năng suất sữa không có sự sai khác giữa các bãi chăn thả; việc bổ sung khoáng cũng không ảnh hưởng tới năng suất và không có ảnh hưởng của tương tác giữa bãi chăn thả và việc bổ sung khoáng.

So sánh 2 ví dụ ở mô hình hai yếu tố kiểu chia ô, thấy rằng phương pháp ngẫu nhiên hoá các bãi chăn thả khác nhau đã không ảnh hưởng đến năng suất sữa. Tuy nhiên sử dụng khói đã làm tăng độ chính xác của phép thử đối với yếu tố bãi chăn thả. Trên thực tế, những ô liền kề nhau có khuynh hướng giống nhau; chính điều này giải thích tại sao cách tiếp cận theo mô hình khói phù hợp hơn.

5.5. BÀI TẬP

5.5.1

Một thí nghiệm được tiến hành nhằm nghiên cứu ảnh hưởng của progesterone lên chu kỳ động dục của cừu Merino. Sử dụng 4 liều khác nhau (0, 10, 25 và 40 mg/ngày) tiêm dưới da vào ngày động dục hoặc 1 ngày sau đó. Chọn 32 cừu thí nghiệm đồng đều nhau và phân ngẫu nhiên về với các công thức thí nghiệm, mỗi công thức có 4 cừu. Chu kỳ động dục (ngày) của 4 cừu trong mỗi nhóm thu được như sau:

Ngày sử dụng	Liều dùng			
	0	10	25	40
0	17	15	12	8
	18	15	12	9
	17	14	11	11
	17	16	11	6
1	18	16	16	12
	20	14	14	13
	17	16	11	12
	14	16	14	12

Cho biết ảnh hưởng của progesterone lên chu kỳ động dục ở cừu Merino.

5.5.2

Một thí nghiệm được tiến hành nhằm xác định ảnh hưởng của gà trống và gà mái đến khối lượng thê hệ gà con ở 8 tuần tuổi. Chọn ngẫu nhiên 4 gà trống, mỗi gà trống cho phối với 3 gà mái, mỗi gà mái cho 3 gà con. Khối lượng (kg) 8 tuần tuổi của các gà con được trình bày như sau:

Gà trống	Gà mái	Khối lượng gà con (kg)	
1	1	965	813
	2	803	640
	3	644	753
2	1	740	798
	2	701	847
	3	909	800
3	1	696	807
	2	752	863
	3	686	832
4	1	979	798
	2	905	880
	3	797	721
5	1	809	756
	2	887	935
	3	872	811

Hãy cho biết ảnh hưởng của gà trống và gà mái đến khối lượng gà con 8 tuần tuổi

Chương 6

TƯƠNG QUAN VÀ HỒI QUY TUYẾN TÍNH

Chương này sẽ giúp bạn đọc hiểu được mối quan hệ tuyến tính giữa các biến định lượng được khảo sát đồng thời trên một đám đông; ước tính hệ số tương quan và xây dựng phương trình hồi quy từ các biến định lượng; áp dụng các kiến thức này trong chăn nuôi, thú y và thủy sản.

6.1. SẮP XẾP SỐ LIỆU

Khi có ít số liệu có thể để dãy n cặp số dưới dạng cột hay hàng, nếu nhiều hơn thì có thể sắp dưới dạng có tần số, nếu nhiều nữa thì chia khoảng cả X và Y để sắp thành bảng hai chiều.

1) Sắp thành hàng

X	x ₁	x ₂	...	x _n
Y	y ₁	y ₂	...	y _n

2) Sắp thành hàng có tần số

X	x ₁	x ₂	...	x _k
Y	y ₁	y ₂	...	y _k
m	m ₁	m ₂	...	m _k

3) Sắp thành cột hoặc thành cột có tần số

X	Y	X	Y	m
x ₁	y ₁	x ₁	y ₁	m ₁
x ₂	y ₂	x ₂	y ₂	m ₂
...
x _n	y _n	x _k	y _k	m _k
			Tổng	n

4) Sắp thành bảng, X gồm k lớp, Y gồm 1 lớp với các điểm giữa x_i và y_j

	y ₁	y ₂	...	y _l
x ₁	m ₁₁	m ₁₂	...	m _{1l}
x ₂	m ₂₁	m ₂₂	...	m _{2l}
...
x _k	m _{k1}	m _{k2}	...	m _{kl}

Tùy dạng bảng có thể dễ dàng chuyển thành dạng cột hay hàng có tần số và ngược trở lại chuyển từ dạng cột hay hàng có tần số thành bảng.

Ở phần sau các công thức tính toán đưa ra chỉ đúng khi số liệu viết dưới dạng hai cột không có tần số, khi có tần số thì phải thêm tần số vào các công thức.

6.2. HỆ SỐ TƯƠNG QUAN

Trong toán học khi có hai dãy số x_i và y_i người ta có thể khảo sát mối quan hệ giữa X và Y bằng khái niệm hàm số.

Trong thống kê x_i và y_i là các giá trị thu được trong mẫu quan sát của hai biến ngẫu nhiên X, Y và người ta muốn đưa ra một con số để đánh giá hai biến ngẫu nhiên X và Y có quan hệ với nhau hay không.

Có khá nhiều con số được dùng để đánh giá X và Y có quan hệ hay không nhưng không có con số nào thỏa mãn được mọi mong muốn của chúng ta. Trong thực tế, các nhà nghiên cứu thường quan tâm đến mối quan hệ tuyến tính giữa 2 tình trạng. Mức độ quan hệ này được thể hiện bằng hệ số tương quan. Hệ số tương quan được đánh giá là đơn giản, dễ dùng và có nhiều ưu điểm, nhưng chỉ thể hiện được mối quan hệ tuyến tính giữa X và Y chứ không thể dùng để đánh giá mối quan hệ nói chung của hai biến.

6.2.1. Tính hệ số tương quan

Dựa trên lý thuyết xác suất về hệ số tương quan chúng ta có công thức sau để tính hệ số tương quan mẫu r_{XY} giữa hai biến ngẫu nhiên X và Y

$$r_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (6.1)$$

Khai triển công thức này được công thức (6.2) thuận tiện hơn về mặt tính toán

$$r_{XY} = \frac{\sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}\right)\left(\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n}\right)}} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{(x_i^2 - n(\bar{x})^2)(y_i^2 - n(\bar{y})^2)}} \quad (6.2)$$

Nếu tính tuân tự các tham số thì có thể lần lượt tính phương sai mẫu của biến X, phương sai mẫu của biến Y, hiệp phương sai mẫu của X và Y.

$$r_{XY} = \frac{Cov_{XY}}{s_X s_Y} \quad (6.3)$$

$$\text{Trong đó: } s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n-1)}; \quad s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{(n-1)}; \quad \text{cov}_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

6.2.2. Tính chất của hệ số tương quan mẫu

- 1) Là một số nằm giữa -1 và $+1$, nói cách khác $|r_{XY}| \leq 1$

2) Nếu Y và X có quan hệ tuyến tính $Y = a + bX$ thì $|r_{XY}| = 1$ và ngược lại nếu $|r_{XY}| = 1$ thì Y và X có quan hệ tuyến tính $Y = a + bX$

3) Nếu X và Y độc lập về xác suất thì $r_{XY} = 0$ nhưng ngược lại không đúng, nếu $r_{XY} = 0$ (gọi là không tương quan) thì chưa thể kết luận X và Y độc lập về xác suất. (Như vậy độc lập về xác suất suy ra không tương quan nhưng không tương quan không suy ra độc lập về xác suất).

4) Nếu thực hiện hai phép biến đổi tuyến tính

$$U = aX + b; \quad V = cY + d \quad \text{thì } r_{UV} = r_{XY}$$

Tính chất này được phát biểu dưới dạng: Hệ số tương quan bất biến đối với phép biến đổi tuyến tính.

Trong thống kê thường dùng cách chọn gốc đo mới và đơn vị đo mới. Nếu gọi x_0 là gốc mới, h là đơn vị mới, số đo x của biến X bây giờ là u :

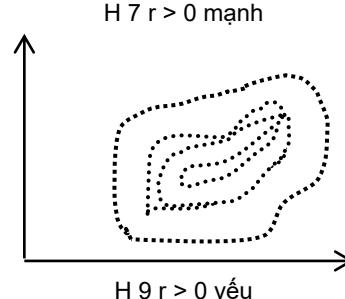
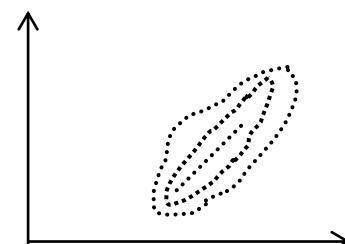
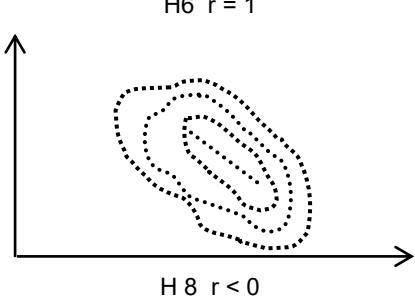
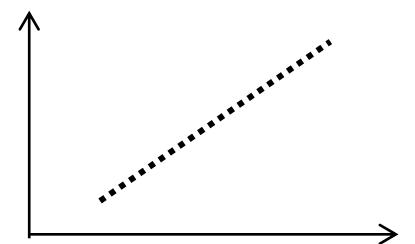
$$u = \frac{(x - x_0)}{h} \quad \text{hay} \quad x = x_0 + hu$$

Như vậy ta đã thực hiện phép biến đổi tuyến tính $X = x_0 + hu$. Tương tự đối với Y ta biến đổi $Y = y_0 + kv$

Bốn tính chất này có thể chứng minh chặt chẽ nhờ các bất đẳng thức toán học đối với 2 dãy số nhưng ở đây chúng ta thừa nhận không chứng minh.

Hệ số tương quan được coi là một số đo mối quan hệ hay liên hệ tuyến tính giữa X và Y vì khi $|r_{XY}|$ gần về phía 1 (thường gọi là tương quan mạnh) thì có thể kết luận X và Y có quan hệ gần với quan hệ tuyến tính, còn nếu $|r_{XY}|$ gần về phía 0 (thường gọi là tương quan yếu) thì không kết luận được gì vì có thể X và Y độc lập hoặc có thể có quan hệ, nhưng nếu có thì quan hệ này không thể là quan hệ tuyến tính.

Về dấu thì nếu $r_{XY} > 0$ ta có tương quan dương, nếu < 0 thì tương quan âm.



Ví dụ 6.1: Nghiên cứu mối quan hệ tuyến tính giữa đường kính lớn x (mm) và khối lượng y (g) của một loại trứng gà. Tiến hành đo đường kính lớn và cân khối lượng của 10 quả trứng. Số liệu thu thập được như sau:

Quả trứng	1	2	3	4	5	6	7	8	9	10
Đường kính lớn (x)	57	54	55	52	55	60	56	56	57	58
Khối lượng (y)	61	59	58	56	57	59	56	58	56	60

Dựa vào công thức 6.1 ta có thể tính được hệ số tương quan như sau:

x	y	(x - \bar{x})	(y - \bar{y})	(x - \bar{x})^2	(y - \bar{y})^2	(x - \bar{x})(y - \bar{y})
57	61	1	3	1	9	3
54	59	-2	1	4	1	-2
55	58	-1	0	1	0	0
52	56	-4	-2	16	4	8
55	57	-1	-1	1	1	1
60	59	4	1	16	1	4
56	56	0	-2	0	4	0
56	58	0	0	0	0	0
57	56	1	-2	1	4	-2
58	60	2	2	4	4	4
560	580	0	0	44	28	16

Ta có: $n = 10$; $\sum x_i = 560$; $\sum y_i = 580$; $\bar{x} = 56$; $\bar{y} = 58$.

Nếu tính theo (6.1)

$$r_{xy} = \frac{16}{\sqrt{44 \times 28}} = 0,4558$$

Nếu tính theo (6.2) thì

$$\sum x_i^2 = 31404; \sum y_i^2 = 33668; (\bar{x})^2 = 3136; \sum x_i^2 - n(\bar{x})^2 = 44$$

$$\sum x_i y_i = 32496; \sum x_i y_i - n \times \bar{x} \times \bar{y} = 16; \sum y_i^2 - n(\bar{y})^2 = 28$$

$$r_{xy} = \frac{16}{\sqrt{44 \times 28}} = 0,4558$$

Nếu tính tuân tự theo (8.3) thì:

$$s_x^2 = \frac{44}{9} = 4,8889; s_y^2 = \frac{28}{9} = 3,1111; \text{cov}_{xy} = \frac{16}{9} = 1,7778$$

$$r_{xy} = \frac{1,7778}{\sqrt{4,8889 \times 3,1111}} = 0,4558$$

6.3. HỎI QUY TUYẾN TÍNH

Về các điểm quan sát $M_i(x_i, y_i)$ trên hệ toạ độ vuông góc, các điểm này họp thành một đám mây quan sát nhìn chung có dạng một elíp (trừ một vài điểm tách ra xa gọi là điểm ngoại lai), nếu $|r_{xy}|$ gần bằng 1 thì elíp rất dẹt, nếu $|r_{xy}|$ vừa phải thì elíp bầu

bình, nếu $|r_{XY}|$ gần bằng không thì có 2 khả năng: Hoặc đám mây quan sát tản漫 trên một phạm vi rộng (không quan hệ), hoặc đám mây quan sát không còn dạng elíp mà tập trung thành một hình cong (phi tuyếnn).

Trường hợp $|r_{XY}|$ gần 1 elíp đám mây quan sát khá dẹt. Để giải thích sự thay đổi của Y khi cho X thay đổi người ta thường đưa ra mô hình hồi quy tuyếnn tính:

$$Y = a + bX$$

Có thể tìm hiểu mô hình hồi quy tuyếnn tính theo hai cách sau đây:

6.3.1. Đường trung bình của biến ngẫu nhiên Y theo X trong phân phôi chuẩn 2 chiều

Khảo sát đồng thời 2 biến ngẫu nhiên định lượng (như đã làm từ đầu chương này). Cặp biến X,Y thường tuân theo luật chuẩn hai chiều, khi ấy nếu theo dõi biến X trước thì ứng với mỗi giá trị x của biến ngẫu nhiên X có vô số giá trị của biến Y, các giá trị này có giá trị trung bình lý thuyết là kỳ vọng $M(Y/x)$.

Khi x thay đổi kỳ vọng $M(Y/x)$ thay đổi theo và các điểm $P(x, M(Y/x))$ chạy trên một đường thẳng gọi là đường hồi quy tuyếnn tính Y theo X.

Nếu theo dõi biến Y trước thì ứng với một giá trị y của Y có vô số giá trị của biến X có trung bình là kỳ vọng $M(X/y)$. Điểm $Q(y, M(X/y))$ chạy trên một đường thẳng gọi là đường hồi quy tuyếnn tính X theo Y.

Như vậy, về mặt lý thuyết, khi có phân phôi chuẩn hai chiều các đường hồi quy tuyếnn tính Y theo X và hồi quy tuyếnn tính X theo Y chính là các đường kỳ vọng có điều kiện $M(Y/x)$ và $M(X/y)$.

Trong trường hợp tổng quát của phân phôi hai chiều các đường kỳ vọng có điều kiện có thể là đường thẳng hoặc đường cong và được gọi là hồi quy Y theo X (hay X theo Y). Trong thực nghiệm chúng ta khảo sát 2 biến định lượng bằng cách lấy mẫu với dung lượng n khá lớn.

Thay cho đường hồi quy tuyếnn lý thuyết có đường hồi quy thực nghiệm. Gọi (x, y) là toạ độ của một điểm chạy trên đường thẳng hồi quy, \bar{x} và \bar{y} là trung bình cộng của X và Y, s_x và s_y là độ lệch chuẩn của X và Y, phương trình hồi quy tuyếnn thực nghiệm có dạng:

$$y - \bar{y} = r_{XY} \frac{s_y}{s_x} (x - \bar{x}) \quad (6.4)$$

Nếu viết phương trình đường thẳng dưới dạng $y = a + bx$ thì:

$$\text{Hệ số góc: } b = r_{XY} \frac{s_y}{s_x} \quad \text{tung độ gốc } a = \bar{y} - b\bar{x} \quad (6.5)$$

Nếu dùng công thức (6.2) để tính hệ số tương quan thì:

$$\text{Hệ số góc } b = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \quad \text{tung độ gốc } a = \frac{\sum y_i - b \sum x_i}{n} \quad (6.6)$$

Nếu dùng công thức (8.1) để tính hệ số tương quan thì:

$$\text{hệ số góc } b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad \text{tung độ gốc } a = \bar{y} - b\bar{x} \quad (6.7)$$

Đường hồi quy tuyến tính thực nghiệm X theo Y có phương trình:

$$x - \bar{x} = d(y - \bar{y}) \quad \text{với hệ số góc } d = r_{XY} \frac{s_x}{s_y}$$

Nếu viết dưới dạng $x = c + dy$ thì hoành độ gốc $c = \bar{x} - d\bar{y}$

Nếu nhân hệ số góc b của hồi quy tuyến tính Y theo X với hệ số góc d của hồi quy tuyến tính X theo Y thì được r_{xy}^2 :

$$b \times d = r_{XY}^2$$

Với ví dụ 6.1: Nghiên cứu mối quan hệ tuyến tính giữa đường kính lớn x (mm) và khối lượng y (g) của một loại trứng gà. Tiến hành đo đường kính lớn và cân khối lượng của 10 quả trứng. Số liệu thu thập được như sau:

Ta đã có: $\bar{x} = 56$; $\bar{y} = 58$; $s_x^2 = 4,8889$; $s_y^2 = 3,1111$; $r_{XY} = 0,4558$

Hồi quy tuyến tính Y theo X

$$y - 58 = 0,4558 \frac{\sqrt{3,1111}}{\sqrt{4,8889}} (x - 56)$$

Viết dưới dạng $y = a + bx$ thì

(1) Nếu tính theo (5.5) ta có:

$$\text{Hệ số góc } b = 0,4558 \frac{\sqrt{3,1111}}{\sqrt{4,8889}} = 0,3636 \quad \text{và tung độ gốc } a = 58 - 0,3636 \cdot 56 = 37,6384$$

Nếu tính theo (5.6) ta có:

$$\text{Hệ số góc } b = \frac{16}{44} = 0,3636 \quad \text{và tung độ gốc } a = \frac{580 - 0,3636 \times 560}{10} = 37,6384$$

6.3.2. Đường thẳng gần đúng của Y theo X

Xét bài toán thường gặp trong các thí nghiệm nông nghiệp và sinh học sau:

Một biến X định lượng có các giá trị $x_i (i = 1, n)$, biến này hoặc do chúng ta chủ động điều khiển ví dụ thời gian cai sữa, mức protein trong khẩu phần, mật độ nuôi trong

chuồng, liều lượng thuốc,..., hoặc quan sát trong tự nhiên như tuổi của vật nuôi, thời gian tiết sữa, số con đẻ ra trên lứa, số con cai sữa, tiêu tốn thức ăn . . .

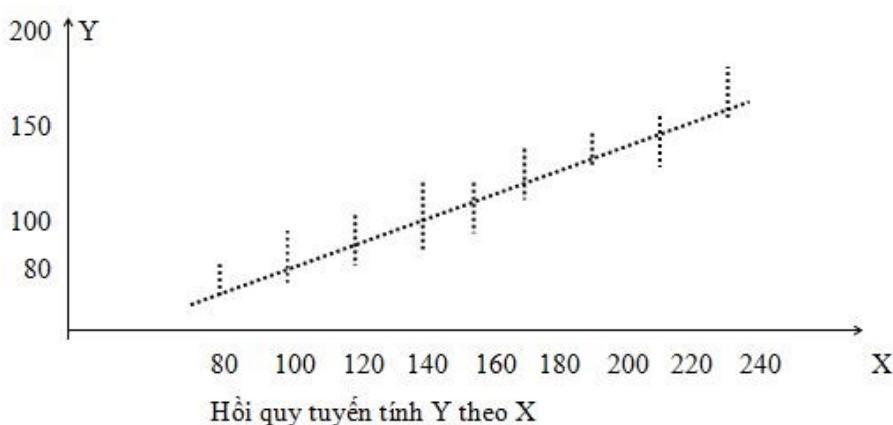
Biến thứ hai là một biến Y mà qua quan sát thấy thay đổi theo X, ví dụ khối lượng vật nuôi thay đổi theo tuổi, năng suất sữa trong một chu kỳ thay đổi theo thời gian tiết sữa, chỉ tiêu Y về phản xạ của chuột thay đổi theo lượng thuốc X đã tiêm ...

Vấn đề đặt ra là tìm một hàm của X để tính gần đúng các giá trị của Y.

Hàm này thường chọn trong các lớp hàm: bậc nhất (tuyến tính), bậc hai, lôgarít, mũ... hàm phải đơn giản và dễ lý giải về mặt chuyên môn.

Nếu dùng x_i làm hoành độ, y_i làm tung độ thì có n điểm quan sát $M_i(x_i, y_i)$ và bài toán ở đây là dùng một đường thẳng, đường parabol, đường lôgarít, đường mũ, . . . để lý giải sự thay đổi của Y theo X, đường này không buộc phải đi qua tất cả các điểm mà chỉ cần đi “sát”, đi “gần” các điểm quan sát M_i .

Trong phần hàm nhiều biến của toán học cao cấp sau khi tính đạo hàm riêng có đề cập đến đường thẳng “tốt” nhất theo nguyên tắc (hay phương pháp) bình phương bé nhất.



Giả sử chọn đường gần đúng là đường thẳng $z = a + bx$, ta sẽ có mô hình tuyến tính sau:

$$y_i = z_i + e_i = a + bx_i + e_i \quad (6.8)$$

e_i là độ chênh lệch giữa giá trị thực y_i và giá trị tương ứng z_i trên đường thẳng (thường gọi e_i là sai số hay phần dư).

Theo nguyên tắc bình phương bé nhất thì đường “tốt” nhất trong các đường thẳng dùng làm đường gần đúng là đường có tổng bình phương các phần dư $\sum e_i^2$ nhỏ nhất.

Dùng cách tính cực trị của hàm hai biến để tìm min $\sum e_i^2$ thu được hệ hai phương trình (gọi là hệ phương trình chuẩn) để tìm a và b.

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

Có nhiều cách giải hệ hai phương trình bậc nhất với hai ẩn số. Nếu dùng định thức để giải thì có ngay kết quả sau:

$$b = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \quad a = \frac{\sum y_i - b \sum x_i}{n} \quad (6.9)$$

Trùng với công thức (5.6) đã dùng để tính các hệ số hồi quy a và b ở phần a.

Nếu các biến ngẫu nhiên e_i trong mô hình tuyến tính (5.8) phân phối chuẩn thoả mãn 3 điều kiện:

- a/ Kỳ vọng bằng
 - b/ Phương sai bằng nhau
 - c/ Độc lập với nhau
- (6.10)

Thì sau khi tính các hệ số theo (5.9) có thể tính được sai số của các hệ số, phân tích và đánh giá các nguồn biến động, phân tích sai số dự báo.

Đường thẳng gần đúng tốt nhất vừa tìm được theo (8.9) trong trường hợp này cũng được gọi là đường hồi quy tuyến tính Y theo X.

(Để phân biệt có khi người ta gọi đường này là đường hồi quy tuyến tính dạng I, còn đường trung bình trong mô hình phân phối chuẩn hai chiều ở a/ là đường hồi quy tuyến tính dạng II).

Trong mô hình hồi quy tuyến tính dạng I biến X (không ngẫu nhiên) được gọi là biến độc lập, biến giải thích hay biến điều khiển còn biến Y (ngẫu nhiên) thay đổi theo X được gọi là biến phụ thuộc, biến kết quả hay biến đáp.

Trở lại đường hồi quy tuyến tính ở phần a/, nếu chọn trước biến ngẫu nhiên X và coi như biến độc lập thì biến thay đổi theo Y trong phân phối chuẩn hai chiều thoả mãn các điều kiện vừa nêu ở (5.10). Như vậy đường hồi quy tuyến tính dạng II, theo nghĩa đường trung bình của biến Y theo biến X, cũng chính là đường hồi quy tuyến tính theo nghĩa vừa trình bày: “đường thẳng gần đúng tốt nhất đối với biến Y”, tức là đường hồi quy tuyến tính dạng I.

Tóm lại khi cần tính hồi quy tuyến tính theo nghĩa “Đường thẳng gần đúng tốt nhất đối với biến Y thì dù X là biến không ngẫu nhiên với các sai số e_i của mô hình thoả mãn điều kiện (5.10), hay X là biến ngẫu nhiên trong mô hình phân phối chuẩn hai chiều ta đều có thể tính các hệ số a và b bằng cách dùng các công thức (5.5), (5.6), (5.7) hoặc giải hệ 2 phương trình chuẩn.

Việc tính sai số của a và b , việc phân tích biến động chung thành biến động do hồi quy và biến động do sai số, việc tính và đánh giá dự báo hoàn toàn giống nhau.

Với ví dụ 6.1: Nghiên cứu mối quan hệ tuyến tính giữa đường kính lớn x (mm) và khối lượng y (g) của một loại trứng gà. Tiến hành đo đường kính lớn và cân khối lượng của 10 quả trứng. Số liệu thu thập được như sau:

Ta đã có: $n = 10$; $\sum x_i = 560$; $\sum y_i = 580$; $\sum x_i^2 = 31404$; $\sum x_i y_i = 32496$

$$10a + 560b = 580$$

$$560a + 31404b = 32496$$

Giải hệ phương trình ta được $a = 37,6$; $b = 0,364$. Như vậy hồi quy tuyến tính khối lượng theo đường kính lớn của trứng là:

$$y = 37,6 + 0,364x$$

6.4. KIỂM ĐỊNH ĐỐI VỚI HỆ SỐ TƯƠNG QUAN VÀ CÁC HỆ SỐ HỒI QUY

Trong mô hình phân phối chuẩn hai chiều thì hệ số tương quan mẫu là một thống kê có kỳ vọng là hệ số tương quan lý thuyết ρ . Để kiểm định giả thiết $H_0: \rho = 0$ với đối thiết $H_1: \rho \neq 0$ phải tính giá trị T_{TN} theo công thức:

$$T_{TN} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \text{ rồi so với giá trị tới hạn } t(\alpha/2, n-2) \quad (6.11)$$

Kết luận:

Nếu $|T_{TN}| \leq t(\alpha/2, n-2)$ thì chấp nhận H_0 , ngược lại thì bác bỏ H_0

Với ví dụ 6.1: Nghiên cứu mối quan hệ tuyến tính giữa đường kính lớn x (mm) và khối lượng y (g) của một loại trứng gà.

Ta đã có: $n = 10$; $r = 0,4558$

$$T_{TN} = \frac{0,4558}{\sqrt{\frac{1-0,4558^2}{10-2}}} = 1,448; \quad t(0,025; 8) = 2,306$$

Kết luận: chấp nhận $H_0: \rho=0$

Để kiểm định giả thiết $H_0: \rho = \rho_0$ với đối thiết $H_1: \rho \neq \rho_0$ thường thực hiện phép biến đổi

$$z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$$

Biến này phân phối chuẩn với kỳ vọng $\frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right)$ và phương sai $1/(n-3)$

Từ đó có quy tắc kiểm định:

$$Z_{TN} = \frac{\sqrt{n-3}}{2} \left(\ln\left(\frac{1+r}{1-r}\right) - \ln\left(\frac{1+\rho_0}{1-\rho_0}\right) \right) = \frac{\sqrt{n-3}}{2} \ln\left(\frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0)}\right)$$

So với giá trị tới hạn $z(\alpha/2)$ của phân phối chuẩn tắc

Kết luận: Nếu $|Z_{TN}| \leq z(\alpha/2)$ thì chấp nhận H_0 , ngược lại thì bác bỏ H_0

Trong mô hình hồi quy tuyến tính $y = a + bx$ các sai số được giả thiết phân phối chuẩn $N(0, \sigma^2)$.

Sau khi tính các hệ số a và b của đường hồi quy có thể tính được chênh lệch giữa giá trị quan sát (y_i) và giá trị tương ứng trên đường hồi quy (y^H_i)

$$y^H_i = a + bx_i \rightarrow e_i = y_i - y^H_i = y_i - (a + bx_i)$$

Phương sai σ^2 được ước lượng bởi se^2

$$SE^2 = \frac{\sum_{i=1}^n (y_i - (a + bx_i))^2}{(n - 2)} \quad (6.12)$$

SE được gọi là sai số của một quan sát trong mô hình hồi quy tuyến tính.

Tung độ gốc a có sai số:

$$SE(a) = SE \sqrt{\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}} \quad (6.13)$$

Hệ số gốc b có sai số:

$$SE(b) = \frac{se}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (6.14)$$

Với ví dụ 6.1: Nghiên cứu mối quan hệ tuyến tính giữa đường kính lớn x (mm) và khối lượng y (g) của một loại trứng gà.

x	y	$y^H_i = 37,6 + 0,364x_i$	$e_i = y_i - y^H_i$	e_i^2
57	61	58,36	2,64	6,95
54	59	57,27	1,73	2,98
55	58	57,64	0,36	0,13
52	56	56,55	-0,55	0,30
55	57	57,64	-0,64	0,40
60	59	59,45	-0,45	0,21
56	56	58,00	-2,00	4,00
56	58	58,00	0,00	0,00
57	56	58,36	-2,36	5,59
58	60	58,73	1,27	1,62
560	580	580	0,00	22,18

Ta có: $\Sigma e_i^2 = 22,182$; $SE^2 = 22,182 / (10-2) = 2,773$; $se = 1,664$;

$$\Sigma x_i^2 = 31404; (x_i - \bar{x})^2 = 44$$

$$SE(a) = 1,664 \sqrt{\frac{31404}{10 \times 44}} = 14,07 \quad \text{và} \quad SE(b) = \frac{1,664}{\sqrt{44}} = 0,251$$

Từ đó có quy tắc kiểm định đối với các hệ số a và b:

Giả thiết H_{0A} : $a = 0$ đối với H_{1A} : $a \neq 0$

Tính $T_{TNA} = \frac{a}{s(a)}$ so với giá trị tới hạn $t(\alpha/2, n-2)$

Kết luận:

Nếu $|T_{TNA}| \leq t(\alpha/2, n-2)$ thì chấp nhận H_{0A} , nếu ngược lại thì bác bỏ H_{0A}

Giả thiết H_{0B} : $b = 0$ đối với H_{1B} : $b \neq 0$

Tính $T_{TNB} = \frac{b}{s(b)}$ và so với giá trị tới hạn $t(\alpha/2, n-2)$

Kết luận:

Nếu $|T_{TNB}| \leq t(\alpha/2, n-2)$ thì chấp nhận H_{0B} , nếu ngược lại thì bác bỏ H_{0B}

Với ví dụ 6.1: Nghiên cứu mối quan hệ tuyến tính giữa đường kính lớn x (mm) và khối lượng y (g) của một loại trứng gà.

$$T_{TNA} = 37,6 / 14,07 = 2,672 \quad t(0,025;8) = 2,306 \quad \text{Kết luận: } a \neq 0$$

$$T_{TNB} = 0,364 / 0,251 = 1,450 \quad t(0,025,5) = 2,306 \quad \text{Kết luận: } b = 0$$

6.5. DỰ BÁO THEO HỒI QUY TUYẾN TÍNH

Khi có đường hồi quy tuyến tính thì có thể dùng đường đó để dự báo giá trị Y_M ứng với giá trị x_M ngoài các giá trị x_i đã có của mẫu quan sát:

$$y_M = a + b x_M \quad (6.15)$$

Trong ví dụ 6.1 hồi quy khói lượng theo đường kính lớn của trứng là

$$y = 37,6 + 0,364x$$

Dùng đường hồi quy để dự báo khói lượng một quả trứng có đường kính lớn là 59 mm

$$y_{59} = 37,6 + 0,364 \times 59 = 59,076 \text{ g}$$

Các dự báo này cho ta một giá trị dự báo y_M và có thể tính được sai số dự báo, sai số này lớn dần nếu điểm dự báo x_M ở xa giá trị \bar{x} , như vậy dự báo xa \bar{x} không tốt vì sai số quá lớn.

Sai số dự báo: $SE_M = SE \sqrt{1 + \frac{1}{n} + \frac{(x_M - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$ (6.16)

Với ví dụ 1 ta có sai số dự báo là $SE_{59} = 1,664 \sqrt{1 + \frac{1}{10} + \frac{(59 - 56)^2}{44}} = 1,834$

6.6. PHÂN TÍCH PHƯƠNG SAI VÀ HỒI QUY

Dựa theo ý tưởng của phương pháp phân tích phương sai có thể khảo sát tổng bình phương toàn bộ (biến động toàn bộ của y)

$$SS_{TO} = \sum_{i=1}^n (y_i - \bar{y})^2$$

Có thể tách SS_{TO} thành hai tổng bình phương: 1) tổng bình phương do hồi quy SS_R và 2) tổng bình phương do sai số SS_E

$$SS_R = \sum_{i=1}^n (y_i^H - \bar{y})^2 \text{ với } y_i^H = a + bx_i \text{ (giá trị trên đường hồi quy)}$$

$$SS_E = \sum_{i=1}^n (y_i - y_i^H)^2 = \sum_{i=1}^n e_i^2$$

Từ đó có bảng phân tích phương sai sau:

Nguồn biến động	df	SS	MS	F _{TN}	F tối hạn
Hồi quy	1	SS _R	MS _R = SS _R /df _R	MS _R / MS _E	F(α, df _R , df _E)
Sai số	n-2	SS _E	MS _E = SS _E /df _E = se ²		
Toàn bộ	n-1	SS _{TO}			

Giả thiết H_0 : Không có hồi quy (hệ số hồi quy $b = 0$) với đối thiết H_1 : hệ số $b \neq 0$

Nếu $F_{TN} \leq F(\alpha, df_R, df_E)$ thì chấp nhận H_0 ngược lại thì chấp nhận H_1

Chia SS_R cho SS_{TO} được $\frac{SS_R}{SS_{TO}} = r^2$ và SS_E cho SS_{TO} được $\frac{SS_E}{SS_{TO}} = 1 - r^2$

r^2 được gọi là hệ số xác định (6.16)

Ta còn có $F_{TN} = \frac{msR}{msE} = \frac{r^2}{1-r^2} = T^2_{tnR}$ (6.17)

Như vậy kiểm định F tương đương với kiểm định T đối với hệ số tương quan r và tương đương với kiểm định T đối với hệ số góc b.

Với ví dụ 6.1: Nghiên cứu mối quan hệ tuyến tính giữa đường kính lớn x (mm) và khối lượng y (g) của một loại trứng gà.

Từ đó có bảng phân tích phương sai sau:

Nguồn biến động	df	SS	MS	F _{TN}	F tới hạn
Hồi quy	1	5,818	5,818	2,10	0,185
Sai số	8	22,182	2,773		
Toàn bộ	9	28,000			

Kết luận: Vì $F_{TN} > F$ tới hạn cho nên giả thiết H_0 bị bác bỏ

$$F_{TN} = 5,818 / 2,773 = 2,10 = (1,449)^2 = (T_{TNB})^2 = (T_{TNR})^2$$

6.7. BÀI TẬP

6.7.1

Xác định mối liên hệ giữa khối lượng của gà mái (kg) và thu nhận thức ăn trong một năm (kg). Tiến hành quan sát trên 10 gà mái và thu được kết quả như sau:

Khối lượng gà mái	2,3	2,6	2,4	2,2	2,8	2,3	2,6	2,6	2,4	2,5
Khối lượng thức ăn	43	46	45	46	50	46	48	49	46	47

Xây dựng phương trình hồi quy tuyến tính và tính hệ số tương quan.

6.7.2

Một thí nghiệm được tiến hành để xác định mối liên hệ giữa khối lượng thân thịt lợn (kg) và độ dày mỡ lưng (mm). Tiến hành xác định các chỉ tiêu vừa nêu trên 8 thân thịt lợn, kết quả thu được như sau:

Khối lượng thân thịt	100	130	140	110	105	95	130	120
Độ dày mỡ lưng	42	38	53	34	35	31	45	43

Xây dựng phương trình hồi quy tuyến tính và tính hệ số tương quan.

6.7.3

Để xác định khối lượng của cùu (kg) thông qua chu vi lồng ngực, tiến hành cân đo trên 66 cùu. Số liệu thu được như sau:

Khối lượng (Y) và chu vi lồng ngực (X) của cùu

Y	X	Y	X	Y	X	Y	X	Y	X	Y	X
30	76	20	63	28	77	29	73	18	62	19	67
24	71	28	70	25	71	30	74	28	70	27	69
20	63	22	65	27	72	21	64	27	71	31	74
25	69	28	72	28	74	28	74	30	73	23	67
25	67	25	67	25	65	48	89	28	72	22	63
19	62	20	62	20	64	17	60	22	69	35	75
35	77	35	78	35	78	46	86	48	90	44	84
37	84	43	81	32	73	43	84	31	73	31	73
39	78	36	81	33	80	44	82	39	80	45	86
43	88	41	87	36	82	43	80	33	79	35	78
38	78	36	76	35	74	39	81	34	74	39	76

Xây dựng phương trình hồi quy tuyến tính.

Chương 7

KIỂM ĐỊNH MỘT PHÂN PHỐI VÀ BẢNG TƯƠNG LIÊN

Mục tiêu của chương này nhằm giới thiệu cùng bạn đọc cách phân tích và kiểm định đối với biến định tính; sử dụng phép thử Khi bình phương (χ^2) và phép thử chính xác của Fisher đối với bảng tương liên 2 x 2; đặc biệt phân tích các số liệu đối với các nghiên cứu về dịch tễ học.

7.1. KIỂM ĐỊNH MỘT PHÂN PHỐI

Để khảo sát một biến định tính X ta lấy mẫu quan sát gồm N cá thể và căn cứ vào sự thể hiện của biến X để phân chia thành k lớp như bảng sau:

(L_i là lớp thứ i, O_i là số lần quan sát thấy X thuộc lớp i).

Biến X	L_1	L_2	...	L_k	Tổng
Tần số O_i	O_1	O_2	...	O_k	$N=\sum O_i$

Từ một lý thuyết nào đó, có thể là một lý thuyết đã được xây dựng chặt chẽ, có giải thích cơ chế, cũng có thể chỉ là một lý thuyết mang tính kinh nghiệm, đúc kết từ những quan sát trước đây về biến X, người ta đưa ra một giả thiết H_0 thể hiện ở dãy các tần suất lý thuyết f_1, f_2, \dots, f_k của biến X (có nghĩa là dãy tần suất này được tính từ lý thuyết đã nêu trên). Căn cứ vào tần suất lý thuyết f_i và tần số thực tế m_i chúng ta phải đưa ra một trong hai kết luận:

1) Chấp nhận H_0 tức là coi tần số thực tế m_i phù hợp với lý thuyết đã nêu thể hiện ở tần suất f_i .

2) Bác bỏ H_0 tức là dãy tần số thực tế m_i không phù hợp với lý thuyết đã nêu.

Việc kiểm định được thực hiện với mức ý nghĩa α , tức là nếu giả thiết H_0 đúng thì xác suất để bác bỏ một cách sai làm H_0 bằng α .

Các bước thực hiện:

1) Tính các tần số lý thuyết theo công thức: $E_i = N \cdot f_i$ (7.1)

2) Tính khoảng cách giữa hai số O_i và E_i theo cách tính khoảng cách

$$\chi^2 = \frac{(O_i - E_i)^2}{E_i}$$

3) Tính khoảng cách giữa hai dãy tần số thực tế m_i và tần số lý thuyết t_i theo công thức:

$$\chi^2_{TN} = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (7.2)$$

4) Tìm giá trị tới hạn trong bảng 4 (cột α , dòng $k-1$, ký hiệu là $\chi^2(\alpha, k-1)$)

5) Nếu $\chi^2_{\text{tn}} \leq \chi^2(\alpha, k-1)$ thì chấp nhận H_0 : “Tần số thực tế O_i phù hợp với lý thuyết đã nêu”. Nếu $\chi^2_{\text{tn}} > \chi^2(\alpha, k-1)$ thì bác bỏ H_0 , tức là “Tần số thực tế O_i không phù hợp với lý thuyết đã nêu”.

Để sử dụng phép thử χ^2 , cần thoả mãn các điều kiện sau:

- 1) Các O_i là các quan sát độc lập.
- 2) Tất cả các E_i đều phải lớn hơn hoặc bằng 5.
- 3) Các O_i và E_i không phải là các tỷ lệ phần trăm.

Ví dụ 7.1: Số liệu thống kê năm 1995 cho thấy, tỷ lệ màu lông (fi) *trắng*, *nâu* và *đen trắng* của thỏ trong một quần thể tương ứng là 0,36; 0,48 và 0,16. Năm 2005, từ 400 con thỏ rút một cách ngẫu nhiên từ quần thể nêu trên có 140 con màu lông *trắng*, 240 con màu *nâu* và 20 con màu *đen trắng*. Câu hỏi đặt ra: Sau 10 năm (từ 1995 đến 2005) tỷ lệ màu lông của thỏ trong quần thể có thay đổi hay không?

Giả thiết H_0 : Tỷ lệ màu lông của thỏ trong quần thể sau 10 năm không thay đổi

Ta có thể tóm tắt số liệu quan sát thu được năm 2005 như sau:

Màu lông	Trắng	Nâu	Đen trắng	Tổng số
Tần số (O_i)	140	240	20	400

Dựa vào tỷ lệ ban đầu (năm 1995) ta có các tần suất lý thuyết (ti)

Màu lông	Trắng	Nâu	Đen trắng	Tổng số
fi	0,36	0,48	0,16	1
Ei	$400 \times 0,36 = 144$	$400 \times 0,48 = 192$	$400 \times 0,16 = 64$	400

$$\chi^2_{\text{TN}} = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \frac{(140 - 144)^2}{144} + \frac{(240 - 192)^2}{192} + \frac{(20 - 64)^2}{64} = 42,361$$

Bậc tự do $df = (3 - 1) = 2$; giá trị tới hạn $\chi^2(0,05; 2) = 5,991$

Kết luận: $\chi^2_{\text{TN}} < \chi^2(0,05, 2)$ nên bác bỏ giả thiết H_0 . Chứng tỏ tỷ lệ màu lông thỏ trong quần thể sau 10 năm có sự thay đổi.

Ví dụ 7.2: Trong trường hợp điều tra giới tính của một quần thể cho trước. Trong một mùa nhất định trong năm người ta thấy tỷ lệ giới tính lúc sinh ra có xu hướng con cái cao hơn. Để giải đáp câu hỏi trên tiến hành chọn ngẫu nhiên 297 con chim mới sinh thì thấy có 167 con cái. Liệu yếu tố mùa có làm ảnh hưởng đến tỷ lệ giới tính hay không?

Đối với trường hợp giới tính, ta luôn thừa nhận tỷ lệ đực cái là 1:1 hay 0,5:0,5. Nếu mùa không làm ảnh hưởng đến tỷ lệ giới tính thì theo ước tính lý thuyết từ 297 con chim quan sát ta sẽ có số chim đực và số chim cái bằng nhau và bằng $297 \times 0,5 = 148,5$.

Ta có bảng tổng hợp sau:

	Đực	Cái	Tổng số
Tần số quan sát (O _i)	130	167	297
Tần số lý thuyết (E _i)	148,5	148,5	297

$$\chi^2_{TN} = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2_{TN} = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \frac{(130 - 148,5)^2}{148,5} + \frac{(167 - 148,5)^2}{148,5} = 4,61$$

Bậc tự do df = (2 - 1) = 1; giá trị tới hạn $\chi^2(0,05; 1) = 3,84$

Kết luận: $\chi^2_{TN} < \chi^2(0,05, 1)$ nên bác bỏ giả thiết H_0 . Chứng tỏ tỷ lệ giới tính không tuân theo tỷ lệ đực cái 1:1. Điều kiện khí hậu đã làm thay đổi tỷ lệ này.

Hiệu chỉnh Yate

$$\chi^2 = \sum_{i=1}^k \frac{(|O_i - E_i| - 0,5)^2}{E_i}$$

Hệ số 0,5 trong công thức nêu trên gọi là hệ số hiệu chỉnh Yate hay còn gọi là hiệu chỉnh tính liên tục để loại bỏ sự thiên lệch. Hiệu chỉnh Yate sẽ được trình bày chi tiết ở phần tiếp theo

Theo ví dụ trên ta có giá trị χ^2 hiệu chỉnh là:

$$\chi^2_{TN} = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \frac{(|130 - 148,5| - 0,5)^2}{148,5} + \frac{(|167 - 148,5| - 0,5)^2}{148,5} = 4,36$$

Giá trị χ^2 hiệu chỉnh (4,36) bé hơn giá trị χ^2 trước khi hiệu chỉnh (4,61), tuy nhiên giá trị hiệu chỉnh vẫn lớn hơn giá trị tới hạn (3,84) cho nên ta vẫn có kết luận tương tự như trên.

7.2. BẢNG TƯƠNG LIÊN L x K

Có 2 biến định tính, biến X chia ra k lớp, biến Y chia ra l lớp, qua khảo sát thu được bảng hai chiều chứa các số quan sát được của các ô O_{ij} (gọi là bảng tương liên):

Bảng các tần số O_{ij}

		Y					
		X	Y ₁	Y ₂	...	Y _l	TH _i
X ₁		O ₁₁	O ₁₂	...	O _{1l}		TH ₁
X ₂		O ₂₁	O ₂₂	...	O _{2l}		TH ₂
...	
X _k		O _{k1}	O _{k2}	...	O _{kl}		TH _k
TC _j		TC ₁	TC ₂	...	TC _l		N

Các số O_{ij} thường được gọi là các tần số thực tế. Bài toán đặt ra ở đây là biến X (hàng) và biến Y (cột) có quan hệ hay không?

Giả thiết H_0 : “hàng và cột không quan hệ” với đối thuyết H_1 : “hàng và cột có quan hệ”. Để kiểm tra giả thiết này phải thực hiện các bước sau:

1) Từ giả thiết hàng và cột không quan hệ suy ra các số ở trong ô về lý thuyết phải bằng tổng hàng (TH_i) nhân với tổng cột (TC_j) chia cho tổng số quan sát N (trong thí dụ 7.4 chúng ta sẽ lý giải vấn đề này). Gọi tần số lý thuyết là E_{ij} ta có:

$$E_{ij} = \frac{TH_i \times TC_j}{N} \quad (7.3)$$

2) Tính khoảng cách giữa 2 tần số O_{ij} và E_{ij} theo cách tính khoảng cách χ^2

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

3) Tính khoảng cách giữa 2 dãy m_{ij} và t_{ij} bằng χ^2_{tn} :

$$\chi^2_m = \sum_1^k \sum_1^l \frac{(m_{ij} - t_{ij})^2}{t_{ij}} \quad (7.4)$$

4) Chọn mức ý nghĩa α và tìm giá trị tới hạn trong bảng 4 $\chi^2(\alpha, (k-1)(l-1))$ tương ứng với cột α và bậc tự do $(k-1)(l-1)$

5) Kết luận: Ở mức ý nghĩa α nếu $\chi^2_{\text{tn}} \leq \chi^2(\alpha, (k-1)(l-1))$ chấp nhận H_0 , ngược lại thì bác bỏ H_0

Bài toán về bảng tương liên thường thể hiện dưới hai dạng:

1) X và Y là hai tính trạng, giả thiết H_0 : “Hai biến X, Y không có quan hệ” hay còn phát biểu một cách khác là “X và Y độc lập”. Thường gọi bài toán này là bài toán kiểm định tính độc lập của hai biến định tính, hay kiểm định tính độc lập của hai tính trạng.

2) Hàng X là các đám đông, cột Y là các nhóm, việc phân chia đám đông thành các nhóm căn cứ vào một tiêu chuẩn nào đó. Bài toán này thường được gọi là bài toán kiểm định tính thuần nhất của các đám đông (tức là các đám đông có cùng tỷ lệ phân chia), hay còn gọi là kiểm định các tỷ lệ.

Ví dụ 7.3: Từ một đàn trước khi cho tiêm vắc xin, chọn ra 295 động vật thí nghiệm (tiêm vắc xin) và 55 động vật đối chứng (không tiêm vắc xin). Số động vật này sau khi cho tiêm vắc xin có làm giảm tỷ lệ chết hay không ?

Thuốc	Kết quả		
	Sống	Chết	Tổng hàng
Vắc xin	120	175	295
Đối chứng	30	25	55
Tổng cột	150	200	350

Ở đây có thể coi hàng là các lớp của biến thuộc X (có 2 lớp A, B), cột là các lớp của biến kết quả Y (có 2 lớp: sống và chết). Cũng có thể coi hàng là các đám đông: “những động vật tiêm vắc xin” và “những động vật không tiêm vắc xin”. Cột là sự phân chia mỗi đám đông thành 2 nhóm sống và chết.

Bảng tần số lý thuyết:

Thuốc	Kết quả		Tổng hàng
	Sống	Chết	
Vắc xin	$\frac{295 \times 150}{350} = 126,4$	$\frac{295 \times 200}{350} = 168,6$	295
Đối chứng	$\frac{55 \times 150}{350} = 23,6$	$\frac{55 \times 200}{350} = 31,4$	55
Tổng cột	150	200	350

$$\chi^2_{TN} = \frac{(120 - 126,4)^2}{126,4} + \frac{(175 - 168,6)^2}{168,6} + \frac{(30 - 23,6)^2}{23,6} + \frac{(25 - 31,4)^2}{31,4} = 3,64$$

Bậc tự do $df = (2-1)(2-1) = 1$. Giá trị tới hạn $\chi^2(0,05,1) = 3,84$

Kết luận: Vì $\chi^2_{TN} = 3,64 < \chi^2(0,05,1) = 3,84$, ta chưa có đủ bằng chứng để bác bỏ H_0 . Hay nói một cách khác vắc xin đã không làm giảm được tỷ lệ chết.

Ví dụ 7.4: Nghiên cứu ảnh hưởng của việc thiến đến sự xuất hiện bệnh tiểu đường ở chuột. Từ 100 chuột thí nghiệm, chia ngẫu nhiên về 1 trong 2 cách xử lý thiến và không thiến. Số chuột ở 2 lô thí nghiệm được theo dõi cho đến 140 ngày tuổi và tiến hành lấy mẫu nghiên cứu từ 42 ngày tuổi. Bệnh tiểu đường được xác định đối với chuột có hàm lượng đường trong máu lớn hơn 200 mg/dl. Kết quả thí nghiệm được ghi lại ở bảng sau:

Cách xử lý	Kết quả		Tổng
	Mắc bệnh	Không mắc bệnh	
Thiến	26	24	50
Không thiến	12	38	50
Tổng số	38	62	100

Tần suất lý thuyết

Cách xử lý	Kết quả		Tổng
	Mắc bệnh	Không mắc bệnh	
Thiến	$\frac{50 \times 38}{100} = 19$	$\frac{50 \times 62}{100} = 31$	50
Không thiến	$\frac{50 \times 38}{100} = 19$	$\frac{50 \times 62}{100} = 31$	50
Tổng số	38	62	100

$$\chi^2_{TN} = \frac{(26 - 19)^2}{19} + \frac{(12 - 19)^2}{19} + \frac{(24 - 31)^2}{31} + \frac{(38 - 31)^2}{31} = 8,32$$

Đối với trường hợp bảng tương liên 4 ô

a	b
c	d

Có thể tính χ^2_{TN} theo công thức

$$\chi^2_{TN} = n \times \frac{(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} = 100 \times \frac{(26 \times 38 - 12 \times 24)^2}{50 \times 50 \times 38 \times 62} = 8,32$$

Bậc tự do $df = (2-1)(2-1) = 1$. Giá trị tới hạn $\chi^2(0,05;1) = 3,84$

Kết luận: Vì $\chi^2_{TN} = 8,32 > \chi^2(0,05;1) = 3,84$ nên giả thiết H_0 bị bác bỏ. Chứng tỏ, tỷ lệ chuột sau khi thiến mắc bệnh đái đường cao hơn so với chuột không bị thiến.

Hiệu chỉnh Yates

$$\chi^2 = \frac{\left(|ad - bc| - \frac{n}{2} \right)^2 n}{(a+b)(a+c)(b+d)(c+d)}$$

Với ví dụ trên ta có giá trị χ^2 hiệu chỉnh là:

$$\chi^2 = \frac{\left(|26 \times 38 - 24 \times 12| - \frac{100}{2} \right)^2 \times 100}{(26+24)(26+12)(24+38)(12+38)} = 7,17$$

Kết luận: Với hiệu chỉnh Yate, giá trị χ^2 thực nghiệm bé hơn ($\chi^2 = 7,17$) so với trước khi hiệu chỉnh ($\chi^2 = 8,32$). Tuy nhiên giá trị χ^2 thực nghiệm vẫn lớn hơn giá trị tới hạn, nên ta có kết luận tương tự về bệnh tiêu đường của chuột như đã nêu ở phần trên.

Lưu ý:

Hệ số điều chỉnh của Yate trong kiểm định một phân phối có 2 lớp và trong bảng tương liên 2×2 .

a) Kiểm định một phân phối có 2 lớp

Tính trạng nghiên cứu	Loại 1	Loại 2	Tổng
Tần số thực tế	m_1	m_2	N
Tần số lý thuyết	$t_1 = N \times p_1 / (p_1 + p_2)$	$t_2 = N \times p_2 / (p_1 + p_2)$	N

Để kiểm định giả thiết H_0 : “Hai lớp nói trên phân phối theo tỷ lệ $p_1:p_2$ “có thể sử dụng phương pháp χ^2 với nội dung:

$$\text{Tính} \quad \chi^2_m = \frac{(m_1 - t_1)^2}{t_1} + \frac{(m_2 - t_2)^2}{t_2}$$

So χ^2_{TN} với giá trị tới hạn χ^2 với mức ý nghĩa α và bậc tự do bằng 1. Nếu $\chi^2_{TN} \leq \chi^2_{(\alpha,1)}$ thì chấp nhận H_0 , nếu $\chi^2_{TN} > \chi^2_{(\alpha,1)}$ thì bác bỏ H_0 .

Bài toán kiểm định này tương đương với bài toán kiểm định một xác suất, việc tính toán dựa trên cách tính xấp xỉ phân phối nhị thức bằng phân phối chuẩn, từ đó suy ra χ^2_{TN} xấp xỉ phân phối χ^2 (là một phân phối liên tục suy ra từ phân phối chuẩn). Trường hợp $N < 100$ phép xấp xỉ không thật tốt, thường cho χ^2_{TN} hơi to do đó Yate đề nghị điều chỉnh lại χ^2_{TN} theo hướng làm nhỏ bớt χ^2_{TN} , điều chỉnh này thường gọi là điều chỉnh do tính liên tục.

Công thức tính χ^2_{TN} điều chỉnh như sau:

$$\chi^2_m = \frac{(|m_1 - t_1| - 0,5)^2}{t_1} + \frac{(|m_2 - t_2| - 0,5)^2}{t_2}$$

b) Bảng tương liên 4 ô (2 x 2)

Tính trạng A	Tính trạng B		Tổng hàng
	Lớp B1	Lớp B2	
Loại A1	a	b	a+b
Loại A2	c	d	c+d
Tổng cột	a+c	b+d	N=a+b+c+d

Để kiểm định giả thiết H_0 : “Hai tính trạng A và B độc lập” có thể dùng phương pháp χ^2 với các nội dung sau:

+ Tính các số lý thuyết

$$\hat{a} = \frac{(a+b)(a+c)}{N} \quad \hat{b} = \frac{(a+b)(b+d)}{N} \quad \hat{c} = \frac{(c+d)(a+c)}{N} \quad \hat{d} = \frac{(c+d)(b+d)}{N}$$

$$+ \text{Tính } \chi^2_{TN} = \frac{(a-\hat{a})^2}{\hat{a}} + \frac{(b-\hat{b})^2}{\hat{b}} + \frac{(c-\hat{c})^2}{\hat{c}} + \frac{(d-\hat{d})^2}{\hat{d}}$$

Có thể tính χ^2_{TN} bằng công thức sau:

$$\chi^2_m = \frac{(ad-bc)^2 \times N}{(a+b)(a+c)(c+d)(b+d)}$$

+ So với giá trị tới hạn χ^2 với mức ý nghĩa α và bậc tự do bằng 1. Nếu $\chi^2_{TN} \leq \chi^2(\alpha, 1)$ thì chấp nhận H_0 , nếu $\chi^2_{TN} > \chi^2(\alpha, 1)$ thì bác bỏ H_0 .

Bài toán này tương đương với bài toán so sánh hai xác suất, việc tính toán dựa trên cách tính xấp xỉ phân phối nhị thức bằng phân phối chuẩn, từ đó suy ra χ^2_{TN} xấp xỉ phân phối χ^2 .

Khi N nhỏ việc xấp xỉ không tốt do đó có một số hướng dẫn như sau:

+ Nếu $N \leq 20$ thì không nên dùng phương pháp χ^2_{TN}

+ Nếu $20 < N \leq 40$ và có ô có số lý thuyết bé < 5 thì cũng không nên dùng phương pháp χ^2_{TN}

Cả hai trường hợp này nên dùng phương pháp chính xác Fisher (xem phần 7.3)

Nếu $N \geq 100$ thì có thể dùng phương pháp χ^2 .

Nếu $N < 100$ và không rơi vào 2 trường hợp đầu thì nên đưa thêm điều chỉnh do tính liên tục Yate nhằm làm nhỏ bớt χ^2_{TN} như sau:

$$\chi^2_m = \frac{(|ad - bc| - 0,5N)^2 \times N}{(a+b)(a+c)(c+d)(b+d)}$$

7.3. KIỂM ĐỊNH CHÍNH XÁC CỦA FISHER ĐỐI VỚI BẢNG TƯƠNG LIÊN 2x2

Khi các giá trị ước tính (Ei) trong bảng tương liên 2×2 rất bé ($Ei < 5$) thì việc sử dụng phép kiểm định χ^2 không còn đảm bảo được độ chính xác. Trường hợp này hay gặp trong nghiên cứu dịch tễ học và phép kiểm định chính xác của Fisher được sử dụng. Phép kiểm định này cho ta một xác suất trực tiếp và chính xác thay vì đi tìm giá trị xác suất từ bảng.

Nếu ta có bảng tương liên 2×2

a	b	a + b
c	d	c + d
a + c	b + d	n

Fisher dựa trên phân phối siêu hình học (hypergeometric distribution) để tính xác suất của phép thử theo công thức.

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!}$$

Các bước thực hiện:

- 1) Tính p_1 với bảng số liệu đã cho
- 2) Tính $ad - bc$.
 - + Nếu $ad - bc > 0$ thì tăng a và d, giảm b và c bằng 1 đơn vị rồi tính xác suất p_2 ; làm tương tự cho đến khi a bằng min của (a+b) hoặc (a+c).
 - + Nếu $ad - bc < 0$ thì giảm a và d, tăng b và c rồi tính xác suất p_2 ; làm tương tự cho đến khi a bằng 0.
- 3) Tính $P = 2 \times (p_1 + p_2 + \dots + p_n)$.
- 4) Nếu xác suất $P < 0,05$ thì kết luận bác bỏ H_0 .

Ví dụ 7.5: Từ một đàn trước khi cho tiếp xúc với nguồn bệnh, chọn ra 10 động vật thí nghiệm (tiêm vắc xin) và 10 động vật đối chứng (không tiêm vắc xin). Số động vật này sau khi cho tiếp xúc với nguồn bệnh ta thu được kết quả như trong bảng sau. Liệu vắc xin có làm giảm tỷ lệ chết hay không?

Thuốc	Kết quả		Tổng hàng
	Sống	Chết	
Vắc xin	9	1	10
Đối chứng	2	8	10
Tổng cột	11	9	20

$$1) p_1 = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} = \frac{10!10!11!9!}{9!1!2!8!20!} = 0,002679$$

$$2) ad - bc = 9 \times 8 - 1 \times 2 = 70 > 0$$

Tăng a, d và giảm b, c bằng 1 đơn vị ta có

9 + 1	2 - 1	11	→	10	1	11
1 - 1	8 + 1	9		0	9	9
10	10	20		10	10	20

$$p_2 = \frac{10!10!11!9!}{10!0!1!9!20!} = 0,000059537985$$

$$3) P = 2 \times (p_1 + p_2 + \dots + p_n) = 2 \times (0,002679 + 0,000059537985) = 0,005477076$$

4) Với xác suất này, giả thiết H_0 bị bác bỏ. Điều này chứng tỏ vắc xin đã làm giảm tỷ lệ chết.

Ví dụ 7.6: Tương tự như ví dụ 7.5 từ 15 động vật thí nghiệm (tiêm vắc xin) có 2 động vật mắc bệnh và từ 13 động vật đối chứng (không tiêm vắc xin) có 10 động vật mắc bệnh. Liệu vắc xin có làm giảm tỷ lệ mắc bệnh hay không?

Thuốc	Kết quả		Tổng hàng
	Mắc bệnh	Không	
Vắc xin	2	13	15
Đối chứng	10	3	13
Tổng cột	12	16	28

$$1) p_1 = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} = \frac{15!13!12!16!}{2!13!0!3!28!} = 0,00098712$$

$$2) ad - bc = 2 \times 3 - 13 \times 10 = -124 < 0$$

Giảm a, d và tăng b, c bằng 1 đơn vị ta có

2 - 1	13 + 1	15	→	1	14	15
10 + 1	3 - 1	13		11	2	13
12	16	28		12	16	28

$$p_2 = \frac{15!13!12!16!}{1!14!11!2!28!} = 0,00003846$$

Giảm a, d và tăng b, c bằng 1 đơn vị ta có

1 - 1	14 + 1	15	→	0	15	15
11 + 1	2 - 1	13		12	1	13
12	16	28		12	16	28

$$p_3 = \frac{15!13!12!16!}{0!15!2!12!8!} = 0,0000004273$$

3) $P = 2 \times (p_1 + p_2 + \dots + p_n) = 2 \times (0,00098712 + 0,00003846 + 0,0000004273) = 0,00205202$

4) Với xác suất này, giả thiết H_0 bị bác bỏ. Điều này chứng tỏ vắc xin đã làm giảm tỷ lệ mắc bệnh.

Cochran khuyến cáo nên sử dụng phép thử chính xác của Fisher nếu trong thí nghiệm $n < 20$ hoặc $20 < n < 40$ và dự đoán bé nhất nhỏ hơn 5.

7.4. THÍ NGHIỆM NGHIÊN CỨU DỊCH TỄ HỌC THÚ Y

Một vấn đề về sức khoẻ động vật được nghiên cứu một cách đầy đủ thông thường phải trải qua các giai đoạn: mô tả, phân tích, tiến hành thí nghiệm, phân tích dữ liệu và trình bày kết quả.

Hai loại thiết kế nghiên cứu cơ bản được thực hiện trong nghiên cứu dịch tỨ học thú y: thí nghiệm quan sát và thí nghiệm thực nghiệm.

- Thí nghiệm quan sát: Dựa trên những cái có sẵn trong thực tế sản xuất và không sử dụng bất kỳ yếu tố thí nghiệm để tác động lên đối tượng nghiên cứu. Thí nghiệm quan sát được chia thành hai loại: thí nghiệm quan sát mô tả (descriptive study) và thí nghiệm quan sát phân tích (analytic study).

+ Thí nghiệm quan sát mô tả: quan tâm đến việc mô tả bệnh với một hoặc một số yếu tố nguy cơ để tìm ra mối liên hệ giữa yếu tố nguy cơ và bệnh tại một thời điểm. Thí nghiệm quan sát mô tả được chia thành các loại: thí nghiệm nghiên cứu trường hợp (case study), nghiên cứu tương quan và nghiên cứu cắt ngang (cross sectional studies).

+ Thí nghiệm quan sát phân tích: quan tâm đến quá trình diễn biến và tập trung đi sâu vào phân tích mối liên hệ giữa yếu tố nguy cơ và bệnh. Thí nghiệm quan sát phân tích được chia thành các loại: thí nghiệm nghiên cứu bệnh chứng hay hồi cứu (case-control studies) và nghiên cứu thuần tập hay tiến cứu (cohort study).

- Thí nghiệm thực nghiệm: sử dụng một hoặc một số yếu tố thí nghiệm (yếu tố nguy cơ) tác động lên đối tượng nghiên cứu, sau đó theo dõi, thu thập dữ liệu và phân tích mối liên hệ giữa yếu tố nguy cơ và bệnh.

Trong khuôn khổ của giáo trình này chỉ đề cập đến cách thiết kế thí nghiệm đối với các nghiên cứu cắt ngang, bệnh chứng hay hồi cứu và thuần tập hay tiến cứu:

7.4.1. Thiết kế thí nghiệm nghiên cứu cắt ngang (cross sectional studies)

a. Đặc điểm

Thí nghiệm nghiên cứu cắt ngang được tiến hành trên mẫu nhằm đánh giá tỷ lệ lưu hành. Đây là loại nghiên cứu thu thập dữ liệu trên từng cá thể về tình trạng mắc bệnh và tình trạng phơi nhiễm với yếu tố nguy cơ. Thí nghiệm này được thực hiện để mô tả mối liên quan giữa yếu tố nguy cơ với hiện tượng sức khoẻ của quần thể vật nuôi tại một thời điểm nhất định, hay nhằm đánh giá sự khác biệt về tần suất mắc bệnh giữa nhóm có tiếp xúc và nhóm không tiếp xúc với yếu tố nguy cơ.

b. Lựa chọn đối tượng trong nghiên cứu cắt ngang

Đối tượng trong nghiên cứu cắt ngang là những cá thể trong quần thể được quan tâm, cá thể đó có thể bị bệnh, có thể không bị bệnh, có thể phơi nhiễm với yếu tố nguy cơ, có thể không phơi nhiễm với yếu tố nguy cơ.

Dung lượng mẫu cần thiết cho nghiên cứu cắt ngang: Mục đích của nghiên cứu cắt ngang nhằm đánh giá sự khác biệt về tần suất mắc bệnh giữa nhóm có tiếp xúc và nhóm không tiếp xúc với yếu tố nguy cơ, nên dung lượng mẫu được ước tính cho trường hợp so sánh hai tỷ lệ (xem mục 3.8 chương 3).

Dung lượng mẫu cần thiết để so sánh 2 tỷ lệ trong các nghiên cứu cắt ngang được tính bằng công thức sau:

$$n = \frac{\left(z_{(\alpha/2)} \sqrt{2p(1-p)} + z_{(1-\beta)} \sqrt{p_1(1-p_1) + p_2(1-p_2)} \right)^2}{\Delta^2}$$

n = dung lượng mẫu tối thiểu cần thiết cho một nhóm.

p₁ = tỷ lệ mắc bệnh hiện hành ở quần thể thứ 1.

p₂ = tỷ lệ mắc bệnh dự đoán ở quần thể thứ 2.

$$p = \frac{p_1 + p_2}{2}; q = 1 - p$$

Z_(1-α/2) = Giá trị z ở mức tương ứng 1-α/2 (α – xác suất mắc sai lầm loại I).

Z_(1-β) = Giá trị z ở mức tương ứng 1-β (β – xác suất mắc sai lầm loại II).

Δ: Sai khác mong đợi (sự khác biệt muốn phát hiện); Δ = p₁ – p₂.

c. Cách tiến hành thí nghiệm nghiên cứu cắt ngang:

Chọn ngẫu nhiên n đối tượng của quần thể quan tâm và tiến hành đánh giá các tần suất quan sát:

- Tình trạng có phơi nhiễm với yếu tố nguy cơ và tình trạng có mắc bệnh: a.
- Tình trạng có phơi nhiễm với yếu tố nguy cơ và tình trạng không mắc bệnh: b.
- Tình trạng không phơi nhiễm với yếu tố nguy cơ và tình trạng có mắc bệnh: c.
- Tình trạng không phơi nhiễm với yếu tố nguy cơ và tình trạng không mắc bệnh: d.

		Tình trạng mắc bệnh		Tổng số
Tình trạng phơi nhiễm	Có	Không		
Có	a	b	a + b	
Không	c	d	c + d	
Tổng số	a + c	b + d	n	

d. Phân tích số liệu

Với các thí nghiệm bố trí đơn giản với bảng tương liên 2 x 2, tiến hành so sánh tần suất mắc bệnh giữa nhóm có tiếp xúc và nhóm không tiếp xúc với yếu tố nguy cơ bằng

phép thử Khi bình phương (trường hợp dung lượng mẫu lớn) và phép thử chính xác của Fisher (trường hợp dung lượng mẫu bé). Phép thử Khi bình phương và phép thử chính xác của Fisher được trình bày chi tiết ở mục 7.2 và 7.3.

Bên cạnh việc đánh giá tỷ lệ lưu hành và so sánh tần suất mắc bệnh giữa nhóm có tiếp xúc và nhóm không tiếp xúc với yếu tố nguy cơ, thí nghiệm nghiên cứu cắt ngang còn có thể tính được tỷ suất chênh OR (odds ratio). Khi số liệu được trình bày theo bảng tương liên 2 x 2, OR được tính như sau:

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}$$

e. Ưu điểm và nhược điểm

Ưu điểm của thí nghiệm nghiên cứu cắt ngang là thí nghiệm thiết kế đơn giản, dễ dàng tổ chức thực hiện, cho kết quả nhanh và chi phí cho loại thí nghiệm này không lớn. Tuy nhiên, giá trị của loại thí nghiệm nghiên cứu cắt ngang thường không cao do số lần khảo sát trên mỗi nhóm đối tượng trong quá trình nghiên cứu thường là một lần và tại một thời điểm nhất định.

Ví dụ 7.7: Tỷ lệ bò mắc bệnh viêm vú giữa 2 trại (A và B) có sự sai khác có ý nghĩa hay không? Biết rằng sau khi kiểm tra 96 bò ở trại A và 72 bò ở trại B trong 1 ngày thấy số lượng bò mắc bệnh viêm vú tương ứng là 36 và 10.

Dung lượng mẫu cần thiết để thực hiện nghiên cứu trên:

$$p_1 = \text{tỷ lệ mắc bệnh của trại A: } 0,38 (38\%)$$

$$p_2 = \text{tỷ lệ mắc bệnh của trại B: } 0,14 (14\%)$$

$$p = \frac{p_1 + p_2}{2} = \frac{0,38 + 0,14}{2} = 0,26; q = 1 - p = 1 - 0,26 = 0,74$$

$$Z_{(1-0,05/2)} = 1,96$$

$$Z_{(1-0,2)} = 0,84$$

$$\Delta: \text{Sai khác mong đợi (sự khác biệt muốn phát hiện); } \Delta = 0,24$$

$$n = \frac{\left(1,96 * \sqrt{2 * 0,26 * 0,74} + 0,84 * \sqrt{0,38 * 0,62 + 0,14 * 0,86}\right)^2}{(0,24)^2} = 53$$

Như vậy cần ít nhất **53** con bò cho một nhóm.

Giả thiết H_0 : Tỷ lệ bò mắc bệnh viêm vú ở hai trại là như nhau với đối thiết H_1 : Tỷ lệ bò mắc bệnh viêm vú ở 2 trại là khác nhau.

Nếu sử dụng phép thử χ^2 ta được giá trị $\chi^2_{TN} = 11,535$; giá trị $\chi^2_{(0,05; 1)} = 3,841$.

Kết luận:

Vì $\chi^2_{TN} > \chi^2$ tới hạn nên có thể kết luận rằng tỷ lệ bò mắc bệnh viêm vú ở hai trại là khác nhau. Mặt khác ta có tỷ suất chênh OR = $(36 \times 62) / (60 \times 10) = 3,72$; tức là số bò mắc bệnh viêm vú ở trại A cao gấp 3,72 lần so với số bò mắc bệnh ở trại B.

7.4.2. Thiết kế thí nghiệm nghiên cứu bệnh chứng (case – control study)

a. Đặc điểm

Thí nghiệm nghiên cứu bệnh chứng là loại nghiên cứu dọc hồi cứu. Nghiên cứu này thu thập dữ liệu trên từng cá thể về tình trạng mắc bệnh và không mắc bệnh với tình trạng phơi nhiễm yếu tố nguy cơ. Thí nghiệm này được thiết kế để phân tích, so sánh nhằm tìm ra sự khác biệt về tần suất quan sát giữa hai nhóm bệnh và nhóm chứng (không bệnh). Đặc trưng nổi bật của thí nghiệm nghiên cứu bệnh chứng là xuất phát điểm bắt đầu từ bệnh. Hai nhóm động vật có bệnh và không có bệnh được chọn ra và hồi cứu xem các động vật trong hai nhóm này có hay không phơi nhiễm với yếu tố nguy cơ.

b. Lựa chọn đối tượng trong nghiên cứu bệnh chứng

Đối tượng trong nghiên cứu bệnh chứng là những cá thể trong quần thể quan tâm được chọn vào hai nhóm bệnh và không bệnh (nhóm chứng).

Dung lượng mẫu cần thiết cho nghiên cứu bệnh chứng: Mục đích của nghiên cứu chứng nhằm so sánh nhằm tìm ra sự khác biệt về tần suất quan sát giữa hai nhóm bệnh và nhóm chứng (không bệnh), nên dung lượng mẫu được ước tính cho trường hợp so sánh hai tỷ lệ (xem mục 3.8 chương 3).

Dung lượng mẫu cần thiết để so sánh 2 tỷ lệ trong các nghiên cứu bệnh chứng được tính bằng công thức sau:

$$n = \frac{\left(z_{(\alpha/2)} \sqrt{2p(1-p)} + z_{(1-\beta)} \sqrt{p_1(1-p_1) + p_2(1-p_2)} \right)^2}{\Delta^2}$$

n = dung lượng mẫu tối thiểu cần thiết cho một nhóm

p₁ = tỷ lệ mắc bệnh hiện hành ở quần thể thứ 1

p₂ = tỷ lệ mắc bệnh dự đoán ở quần thể thứ 2

$$p = \frac{p_1 + p_2}{2}; q = 1 - p$$

Z_(1-α/2) = Giá trị z ở mức tương ứng 1-α/2 (α – xác suất mắc sai lầm loại I)

Z_(1-β) = Giá trị z ở mức tương ứng 1-β (β – xác suất mắc sai lầm loại II)

Δ: Sai khác mong đợi (sự khác biệt muốn phát hiện); Δ = p₁ – p₂

c. Cách tiến hành thí nghiệm nghiên cứu bệnh chứng:

Chọn ngẫu nhiên hai nhóm: bệnh và không bệnh và tiến hành đánh giá các tần suất quan sát:

- Nhóm có bệnh và có phơi nhiễm với yếu tố nguy cơ: a.
- Nhóm không có bệnh và có phơi nhiễm với yếu tố nguy cơ: b.
- Nhóm có bệnh và không phơi nhiễm với yếu tố nguy cơ: c.
- Nhóm không có bệnh và không phơi nhiễm với yếu tố nguy cơ: d.

		Tình trạng mắc bệnh		Tổng số
Tình trạng phơi nhiễm		Có	Không	
Có	a	b	$a + b$	$c + d$
	c	d		
Tổng số	$a + c$		$b + d$	n

d. Phân tích số liệu

Với các thí nghiệm bố trí đơn giản với bảng tương liên 2 x 2, so sánh nhằm tìm ra sự khác biệt về tần suất quan sát giữa hai nhóm bệnh và nhóm chứng (không bệnh) bằng phép thử Khi bình phương (trường hợp dung lượng mẫu lớn) và phép thử chính xác của Fisher (trường hợp dung lượng mẫu bé). Phép thử Khi bình phương và phép thử chính xác của Fisher được trình bày chi tiết ở mục 7.2 và 7.3.

Số đo quan trọng trong nghiên cứu bệnh chứng là tỷ suất chênh OR (odds ratio). Khi số liệu được trình bày theo bảng tương liên 2 x 2, OR được tính như sau:

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}$$

e. Ưu điểm và nhược điểm

Ưu điểm của thí nghiệm nghiên cứu bệnh chứng là thí nghiệm dễ dàng tổ chức thực hiện, cho kết quả nhanh và chi phí cho loại thí nghiệm này không lớn. Tuy nhiên, sai số của loại thí nghiệm nghiên cứu bệnh chứng thường cao.

Ví dụ 7.8: Trong một nghiên cứu, có 62 bò sữa được chẩn đoán ung thư biểu mô mắt và 124 không mắc được chọn ngẫu nhiên từ quần thể. Có mối liên hệ nào giữa giống bò và tỷ lệ mắc bệnh ung thư biểu mô mắt hay không? Nếu số liệu thu thập được như sau:

Giống	Mắc bệnh	Không mắc bệnh	Tổng số
Hereford	44	63	107
Giống khác	18	61	79
Tổng số	62	124	186

Dung lượng mẫu cần thiết để thực hiện nghiên cứu trên:

p_1 = tỷ lệ mắc bệnh của giống bò Hereford: 0,41 (41%)

p_2 = tỷ lệ mắc bệnh của giống bò khác: 0,23 (23%)

$$p = \frac{p_1 + p_2}{2} = \frac{0,41 + 0,23}{2} = 0,32; q = 1 - p = 1 - 0,32 = 0,68$$

$$Z_{(1-0,05/2)} = 1,96$$

$$Z_{(1-0,2)} = 0,84$$

Δ : Sai khác mong đợi (sự khác biệt muốn phát hiện); $\Delta = 0,18$

$$n = \frac{(1,96 * \sqrt{2 * 0,32 * 0,68} + 0,84 * \sqrt{0,41 * 0,59 + 0,23 * 0,77})^2}{(0,18)^2} = 100$$

Như vậy cần ít nhất **100** con bò cho một nhóm.

Giả thiết H_0 : Không có mối liên hệ giữa giống và tỷ lệ mắc bệnh với đối thiết H_1 : Có mối liên hệ giữa bệnh và giống.

Sử dụng phép thử χ^2 , ta có $\chi^2_{TN} = 6,876$ và $\chi^2(0,05;1) = 3,841$.

Kết luận:

Vì $\chi^2_{TN} > \chi^2$ tới hạn nên ta bác bỏ H_0 chấp nhận H_1 ; chứng tỏ có mối liên hệ giữa giống và bệnh. Tỷ suất chênh OR = $(44 \times 61) / (18 \times 63) = 2,37$. Hay nói cách khác giống Hereford mắc bệnh ung thư biểu mô mắt cao hơn 2,37 lần so với các giống khác.

7.4.3. Thiết kế thí nghiệm nghiên cứu thuần tập (cohort study)

a. Đặc điểm

Thí nghiệm nghiên cứu thuần tập hay tiến cứu là loại nghiên cứu dọc mang tính theo dõi. Nghiên cứu này thu thập dữ liệu trên từng cá thể về tình trạng có phơi nhiễm hoặc không có phơi nhiễm với yếu tố nguy cơ. Thí nghiệm này được thiết kế để phân tích, so sánh nhằm tìm ra sự khác biệt về tần suất mắc bệnh giữa hai nhóm có phơi nhiễm và không có phơi nhiễm với yếu tố nguy cơ. Đặc trưng nổi bật của thí nghiệm nghiên cứu thuần tập là xuất phát điểm bắt đầu từ phơi nhiễm. Hai nhóm động vật có phơi nhiễm và không có phơi nhiễm được chọn ra và theo dõi trong tương lai xem các động vật trong hai nhóm này có hay không xuất hiện bệnh.

b. Lựa chọn đối tượng trong nghiên cứu thuần tập

Đối tượng trong nghiên cứu thuần tập là những cá thể trong quần thể quan tâm được chọn theo hai cách:

+ Cách 1: chọn một mẫu ngẫu nhiên; trong mẫu chọn ra đó có nhóm phơi nhiễm và nhóm không phơi nhiễm với yếu tố nghiên cứu.

+ Cách 2: Chọn riêng hai nhóm mẫu khác nhau, một nhóm mẫu phơi nhiễm với yếu tố nghiên cứu và một nhóm mẫu không phơi nhiễm với yếu tố nghiên cứu.

Dung lượng mẫu cần thiết cho nghiên cứu thuần tập: Mục đích của nghiên cứu chứng nhằm so sánh nhằm tìm ra sự khác biệt về tần suất mắc bệnh giữa hai nhóm phơi nhiễm và nhóm không phơi nhiễm, nên dung lượng mẫu được ước tính cho trường hợp so sánh hai tỷ lệ (xem mục 3.8 chương 3).

Dung lượng mẫu cần thiết để so sánh 2 tỷ lệ trong các nghiên cứu thuần tập được tính bằng công thức sau:

$$n = \frac{(z_{(\alpha/2)}\sqrt{2p(1-p)} + z_{(1-\beta)}\sqrt{p_1(1-p_1) + p_2(1-p_2)})^2}{\Delta^2}$$

n = dung lượng mẫu tối thiểu cần thiết cho một nhóm

p_1 = tỷ lệ mắc bệnh hiện hành ở quần thể thứ 1

p_2 = tỷ lệ mắc bệnh dự đoán ở quần thể thứ 2

$$p = \frac{p_1 + p_2}{2}; q = 1 - p$$

$Z_{(1-\alpha/2)}$ = Giá trị z ở mức tương ứng $1-\alpha/2$ (α – xác suất mắc sai lầm loại I)

$Z_{(1-\beta)}$ = Giá trị z ở mức tương ứng $1-\beta$ (β – xác suất mắc sai lầm loại II)

Δ : Sai khác mong đợi (sự khác biệt muôn phát hiện); $\Delta = p_1 - p_2$

c. Cách tiến hành thí nghiệm nghiên cứu thuần tập:

Chọn ngẫu nhiên hai nhóm: có phơi nhiễm và không phơi nhiễm và tiến hành theo dõi, xác định các tần suất quan sát:

- Nhóm có phơi nhiễm với yếu tố nguy cơ và có bệnh: a
- Nhóm có phơi nhiễm với yếu tố nguy cơ và không có bệnh: b
- Nhóm không phơi nhiễm với yếu tố nguy cơ và có bệnh: c
- Nhóm không phơi nhiễm với yếu tố nguy cơ và không có bệnh: d

		Tình trạng mắc bệnh		Tổng số
Tình trạng phơi nhiễm		Có	Không	
Có	a	b	$a + b$	
	c	d	$c + d$	
Tổng số	$a + c$	$b + d$	n	

d. Phân tích số liệu

Với các thí nghiệm bố trí đơn giản với bảng tương liên 2 x 2, so sánh nhằm tìm ra sự khác biệt về tần suất quan sát giữa hai nhóm phơi nhiễm và nhóm không phơi nhiễm bằng phép thử Khi bình phương (trường hợp dung lượng mẫu lớn) và phép thử chính xác của Fisher (trường hợp dung lượng mẫu bé). Phép thử Khi bình phương và phép thử chính xác của Fisher được trình bày chi tiết ở mục 7.2 và 7.3.

Số đo quan trọng trong nghiên cứu bệnh chứng là nguy cơ tương đối RR (relative risk). Khi số liệu được trình bày theo bảng tương liên 2 x 2, RR được tính như sau:

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

e. Ưu điểm và nhược điểm

Ưu điểm của thí nghiệm nghiên cứu thuần tập là thí nghiệm cho kết quả chính xác hơn do sai số của loại thí nghiệm này thường thấp. Tuy nhiên, loại nghiên cứu này tổ chức thực hiện phức tạp, chi phí lớn, thời gian cần thiết kéo dài.

Ví dụ 7.9: Xem xét ví dụ 7.5, từ một đàn trước khi cho tiệp xúc với nguồn bệnh, chọn ra 10 động vật thí nghiệm (tiêm vắc xin) và 10 động vật đối chứng (không tiêm vắc xin). Số động vật này sau khi cho tiệp xúc với nguồn bệnh ta thu được kết quả như trong bảng sau. Liệu vắc xin có làm giảm tỷ lệ chết hay không?

Thuốc	Kết quả		Tổng hàng
	Sống	Chết	
Vắc xin	9	1	10
Đối chứng	2	8	10
Tổng cột	11	9	20

Dung lượng mẫu cần thiết để thực hiện nghiên cứu trên:

$$p_1 = \text{tỷ lệ chết của nhóm tiêm vắc xin: } 0,10 (10\%)$$

$$p_2 = \text{tỷ lệ nhóm đối chứng: } 0,80 (80\%)$$

$$p = \frac{p_1 + p_2}{2} = \frac{0,10 + 0,80}{2} = 0,45; q = 1 - p = 1 - 0,45 = 0,55$$

$$Z_{(1-0,05/2)} = 1,96$$

$$Z_{(1-0,2)} = 0,84$$

$$\Delta: \text{Sai khác mong đợi (sự khác biệt muốn phát hiện); } \Delta = 0,70$$

$$n = \frac{(1,96 * \sqrt{2 * 0,45 * 0,55} + 0,84 * \sqrt{0,10 * 0,90 + 0,80 * 0,20})^2}{(0,70)^2} = 7$$

Như vậy cần ít nhất 7 động vật cho một nhóm.

Nếu sử dụng phép thử chính xác của Fisher ta có xác suất $P = 0,005477076$.

Kết luận: Với xác suất này, giả thiết H_0 bị bác bỏ. Điều này chứng tỏ vắc xin đã làm giảm tỷ lệ chết. Bên cạnh đó, nguy cơ tương đối $RR = (9/10)/(2/10) = 4,5$. Hay nói một cách khác động vật sử dụng vắc xin mức độ sống sót gấp 4,5 lần so với động vật không dùng vắc xin.

7.5. BÀI TẬP

7.5.1

Một trung tâm thu tinh nhân tạo tiến hành thử nghiệm 3 phương pháp thụ tinh nhân tạo khác nhau. Tỷ lệ phôi có chửa ở 3 phương pháp thu được như sau: ở phương pháp I, có 275 bò có chửa từ 353 bò tham gia thí nghiệm; tương tự ở phương pháp II,

các con số này lần lượt là 192 và 256 con, phương pháp III là 261 và 384 con. Tỷ lệ thụ tinh thành công ở 3 phương pháp này có khác nhau hay không?

7.5.2

Chọn mẫu ngẫu nhiên thế hệ con của bò lang Shorthorn thu được kết quả sau đây: 82 con màu lông đỏ, 209 con lông và 89 con trắng. Phân bố màu lông của bò có tuân theo giả thiết rằng màu lông được xác định bởi một cặp allele trội không hoàn toàn? Biết rằng trội không hoàn toàn là trường hợp có một allele trội và dị hợp tử thể hiện sự ảnh hưởng của đồng thời cả 2 allele.

7.5.3

Một thí nghiệm được tiến hành nhằm đánh giá sự liên hệ giữa tỷ lệ viêm nội mạc tử cung và giống. Trong tổng số 700 bò sữa trong nghiên cứu thuần tập (cohort studies), có 500 con giống Holstein Friesian và 200 con giống Jersey. Kết quả nghiên cứu thu được như sau:

Giống		Viêm nội mạc tử cung		Tổng số
		Có	Không	
Holstein	100	400	500	
Jersey	10	190	200	
Tổng số	110		590	700

Có sự liên hệ giữa tỷ lệ viêm nội mạc tử cung và các giống hay không?

PHẦN B - THỰC HÀNH

Phần này bạn đọc sẽ được hướng dẫn thực hiện các bài thực hành trên phần mềm Minitab 16. Đối với từng bài tập thực hành đều có cấu trúc gồm các phần sau: (1) đề bài, (2) hình minh họa nhập dữ liệu và phân tích trên phần mềm Minitab, (3) kết quả xử lý từ phần mềm và (4) trình bày và giải thích kết quả.

Bài 1. TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU

1.1. GIỚI THIỆU PHẦN MỀM MINITAB

Minitab là phần mềm thống kê ứng dụng được phát triển ở Đại học Pennsylvania (Mỹ) từ năm 1972. Minitab là sản phẩm có bản quyền của công ty Minitab Inc với các chức năng quản lý dữ liệu, tính toán, phân tích dữ liệu, vẽ các biểu đồ, đồ thị, một cách hoàn toàn tự động.

Minitab 16 for Windows được sử dụng để minh họa cho các bài tập trong phần giáo trình này. Nếu bạn đọc sử dụng các phiên bản khác của Minitab có thể sẽ không hỗ trợ một số các công cụ và giao diện sẽ khác so với giáo trình này.

Khởi động Minitab

Nếu cài đặt Minitab 16 for Windows theo mặc định ta có thể khởi động phần mềm bằng 3 cách sau đây:



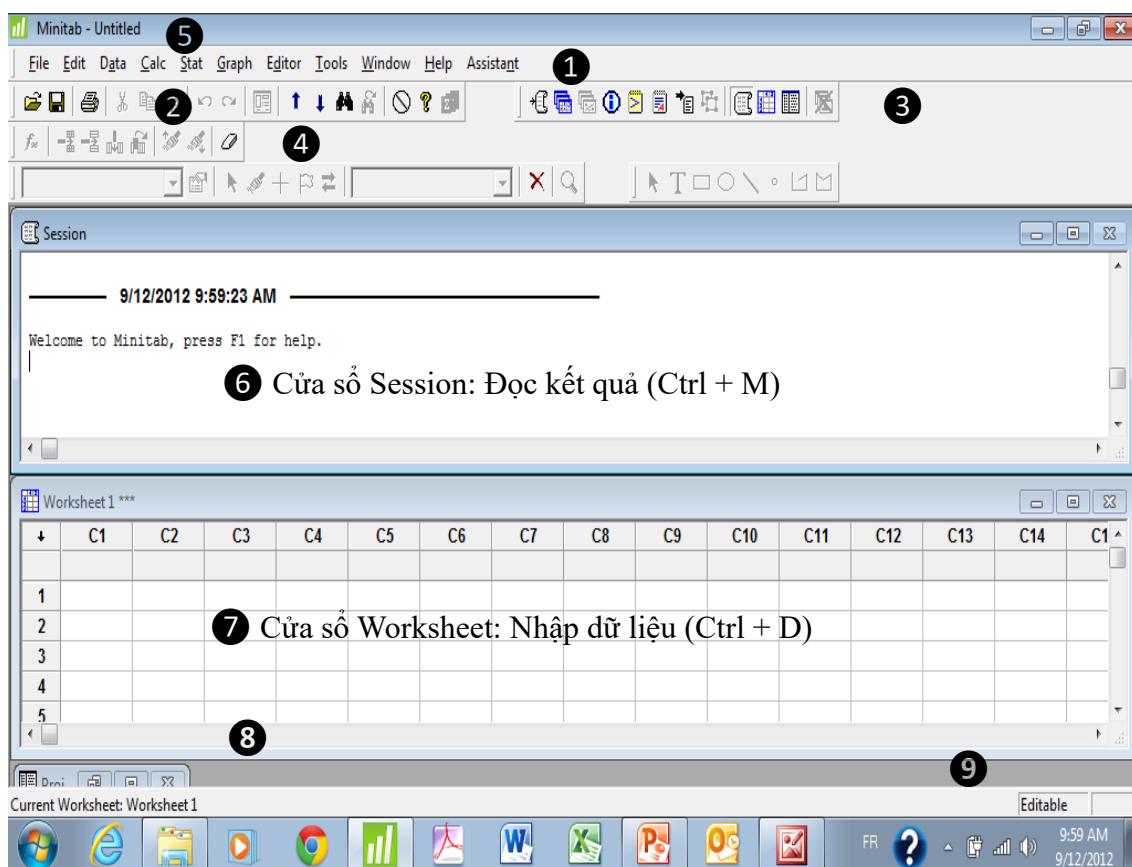
- a) Chọn biểu tượng Minitab trên Desktop của màn hình .
- b) Theo đường dẫn Start > Prog > MINITAB > Minitab 16 statistical software.
- c) C:\Prog Files\Minitab\Minitab 16\Mtb.exe .

Giao diện phần mềm Minitab

Giao diện của phần mềm Minitab 16 gồm những thành phần chính sau (hình 1):

- ❶ **Menu Bar** gồm các lệnh để điều khiển phần mềm Minitab như File, Edit, Data, Calc, Stat, Graph, Editor, Tools, Windows, Help, Assistant);
- ❷ **Standard toolbar** gồm các lệnh tắt như mở tệp đã ghi, ghi tệp, in, cắt, copy, dán,...;
- ❸ **Project Manager Toolbar** gồm các lệnh tắt điều khiển cửa sổ Project Manager;
- ❹ **Worksheet Toolbar** (gồm các lệnh tắt điều khiển cửa sổ Worksheet);

- 5 Title;
- 6 Session Window để đọc kết quả phân tích;
- 7 Data Window chứa nhiều ô (cell) được tạo ra bởi sự kết hợp giữa hàng và cột. Mỗi Worksheet bao gồm 10.000.000 hàng và 4.000 cột (từ C1 đến C4000);
- 8 Project Manager Window (quản lý các lệnh làm việc);
- 9 Status bar.



Hình 1. Cửa sổ làm việc của Minitab 16

1.2.TÓM TẮT VÀ TRÌNH BÀY ĐỐI VỚI BIẾN ĐỊNH LƯỢNG

Ví dụ M-1.1: Khối lượng (g) của 16 chuột cái tại thời điểm cai sữa như sau:

54,1	49,8	24,0	46,0	44,1	34,0	52,6	54,4
56,1	52,0	51,9	54,0	58,0	39,0	32,7	58,5

Đối với các biến định lượng có thể tính các giá trị thống kê mô tả bằng các tham số như trung bình, phương sai, độ lệch chuẩn, hệ số biến động... Bên cạnh đó có thể trình bày ở dạng biểu đồ, đồ thị.

Số liệu được nhập vào một cột trong Windows **Worksheet** như sau:

Minitab - ViduM1.1b.mpj - [Worksheet 1 ***]

	C1	C2	C3	C4	C5
	P	Lua			
1	54.1	1			
2	49.8	1			
3	24.0	1			
4	46.0	1			
5	44.1	1			
6	34.0	1			
7	52.6	1			
8	54.4	1			
9	56.1	2			
10	52.0	2			

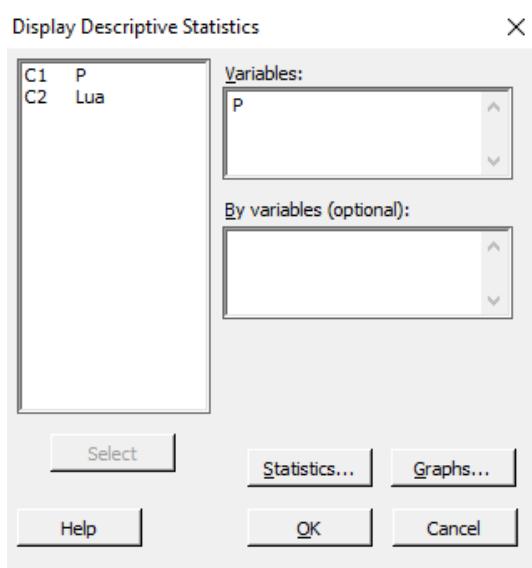
Thay thế dấu phẩy (,) bằng dấu chấm (.) trong phần thập phân. Ô số liệu khuyết được thay thế bằng dấu sao (*), không được để trống.

Cột số liệu phải ở dạng số.

Đối với một chỉ tiêu nghiên cứu, số liệu được nhập dưới dạng cột.

Tên cột số liệu luôn nằm ở trên hàng thứ 1. Đặt tên cột ngắn gọn, không nên dùng các ký tự đặc biệt (:, /...) hoặc các ký tự tiếng Việt (ô, ă...). Trong cùng một **worksheet** không đặt tên cột trùng nhau. Phần mềm Minitab không phân biệt các ký tự viết hoa và viết thường (ví dụ: MINITAB = Minitab = minitab).

Chọn Stat → Basic Statistics → Display Descriptive Statistics



Phần ô bên trái hộp thoại hiển thị cột (C1) và tên của cột số liệu (P).

Kích đúp chuột trái vào biến P hoặc chọn biến P và nhấn **Select** để hiển thị cột cần tính các tham số thống kê mô tả vào ô **Variables**.

Chọn **OK** để hiển thị kết quả.

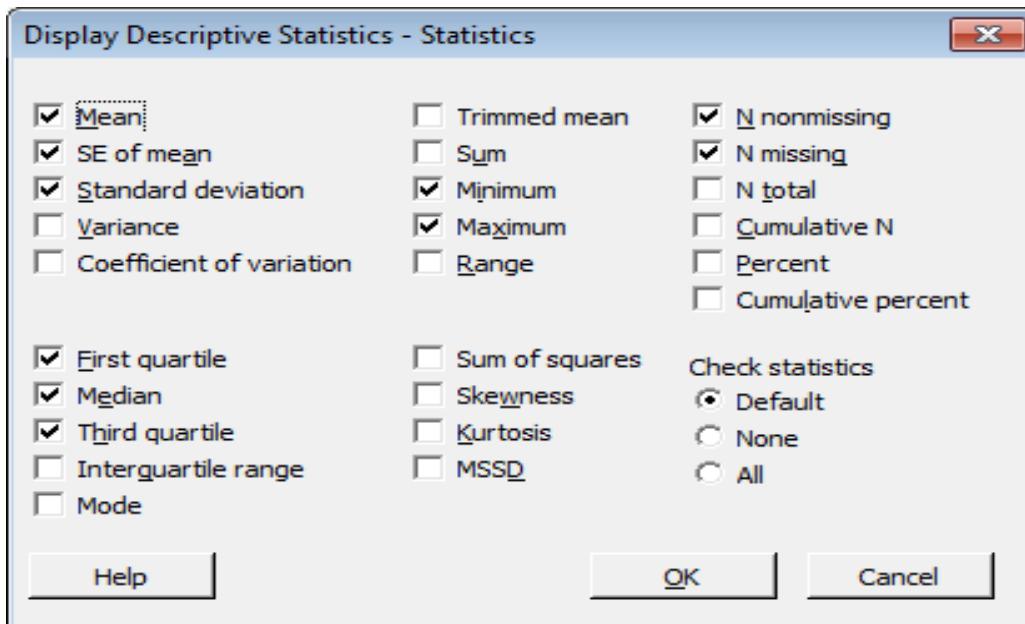
Kết quả thu được từ Minitab như sau.

Descriptive Statistics: P

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
P	16	0	47.58	2.54	10.16	24.00	40.28	51.95	54.33	58.50

Đây là các tham số Minitab cho kết quả theo mặc định. Có thể sử dụng một trong các Option sau đây để cho ra kết quả theo lựa chọn phù hợp với từng trường hợp cụ thể

➔ Chọn **Statistics...** có thể lựa chọn các tham số sau đây



Một số thuật ngữ trong options Minitab của thống kê mô tả

Minitab	Tiếng Việt	Minitab	Tiếng Việt
Mean	Trung bình	Trimmed mean	Trung bình thu gọn
SE of Mean	Sai số chuẩn	Sum	Tổng số
Standard deviation	Độ lệch chuẩn	Minimum	Giá trị bé nhất
Variance	Phương sai	Maximum	Giá trị lớn nhất
Coefficient of variation	Hệ số biến động	Range	Khoảng biến động
First quartile	Tứ vị thứ nhất	Sum of squares	Tổng bình phương
Median	Trung vị	Skewness	Độ lệch
Third quartile	Tứ vị thứ 3	Kurtosis	Độ nhọn
Interquartile	Tứ vị thứ 2	MSSD	
N nonmissing	N không khuyết	Cumulative N	N cộng gộp
N missing	N khuyết	Percent	Phần trăm
N total	N tổng số	Cumulative percent	Phần trăm cộng gộp

➔ Chọn **Graphs...** để hiển thị đồ thị sau đây (click √ vào).

Histogram of data tô chúc đồ.

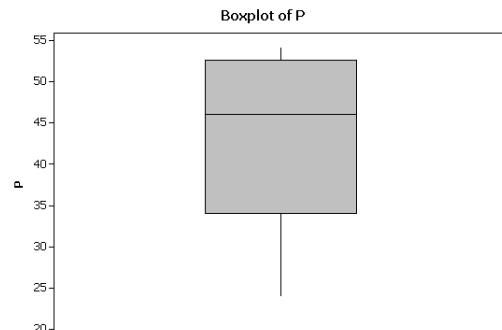
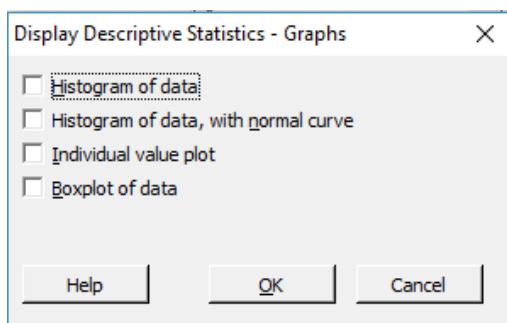
Histogram of data, with normal curve tô chúc đồ với đường cong chuẩn.

Individual value plot thể hiện các điểm của từng giá trị.

Boxplot of data đồ thị hộp.

Chọn **OK**.

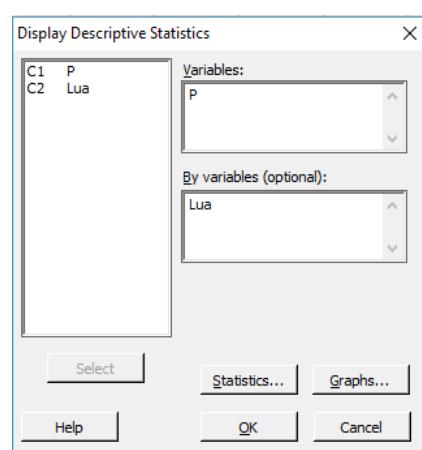
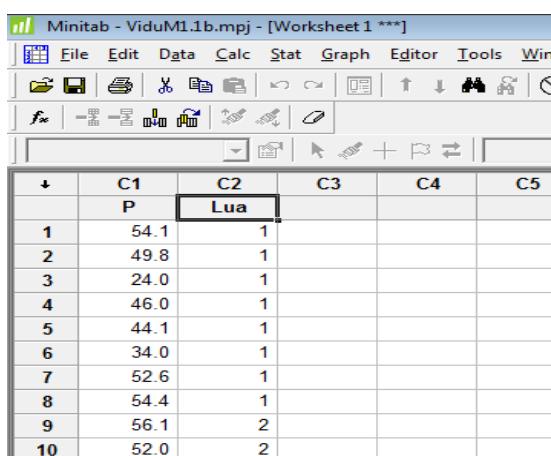
Ví dụ chọn **Boxplot of data** để được đồ thị hộp dưới đây.



➔ Vào **By variables (Optional)** để tính các tham số thống kê theo phân loại nhóm.

Trong trường hợp muốn tính các tham số thống kê theo từng công thức thí nghiệm hay mức của yếu tố thí nghiệm, cần cấu trúc số liệu theo từng yếu tố thí nghiệm.

Xét **Ví dụ M-1.1**, giả sử 8 chuột cái đầu tiên sinh ra ở lứa thứ nhất và 8 chuột tiếp theo sinh ra ở lứa thứ 2. Ta có thể bố trí cấu trúc số liệu thành 2 cột, cột C1 (P) và cột C2 (LUA).



Kết quả từ Minitab

Descriptive Statistics: P

Variable	LUA	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
P	1	8	0	44.88	3.82	10.79	24.00	36.53	47.90	53.73
	2	8	0	50.28	3.32	9.39	32.70	42.23	53.00	57.53

Sử dụng kết quả này được sử dụng để trình bày theo từng lứa. Trong trường hợp có hai hoặc nhiều yếu tố, có thể khai báo hai hoặc nhiều yếu tố vào cửa sổ **By variables**.

1.3. TÓM TẮT VÀ TRÌNH BÀY ĐỐI VỚI BIẾN ĐỊNH TÍNH

Đối với biến định tính số liệu thu thập được từ thí nghiệm có thể được trình bày theo một trong 2 cách sau đây:

Ví dụ M-1.2: Số bò sữa ở ba trại A, B, C lần lượt là 106, 132 và 122 con. Chọn ngẫu nhiên và kiểm tra bệnh viêm nội mạc tử cung ở 3 trại, kết quả như sau:

Cách 1:

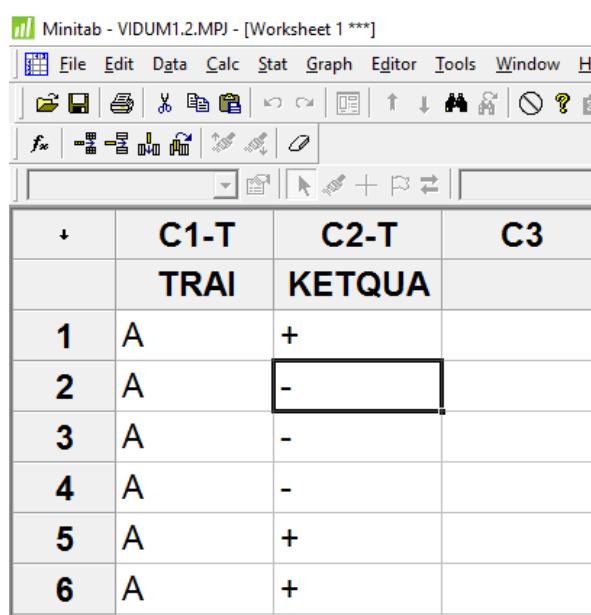
Trại	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
Bò số	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Kết quả	+	-	-	-	+	+	+	-	-	-	+	-	-	-	-	-	+
Trại	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
Bò số	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Kết quả	-	-	+	-	-	-	-	+	+	-	-	-	-	+	-	+	-
Trại	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
Bò số	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Kết quả	+	+	-	-	-	-	+	-	-	+	-	-	-	-	+	+	+

Cách 2:

Trại	Viêm nội mạc tử cung		Tổng số
	Có	Không	
A	6	11	17
B	6	16	22
C	8	12	20

Số liệu có thể được nhập và phân tích bằng phần mềm Minitab theo 2 cách tương ứng với 2 cách trình trình bày số liệu thô đã nêu trên.

Cách 1: Số liệu được nhập vào cột trong Windows Worksheet



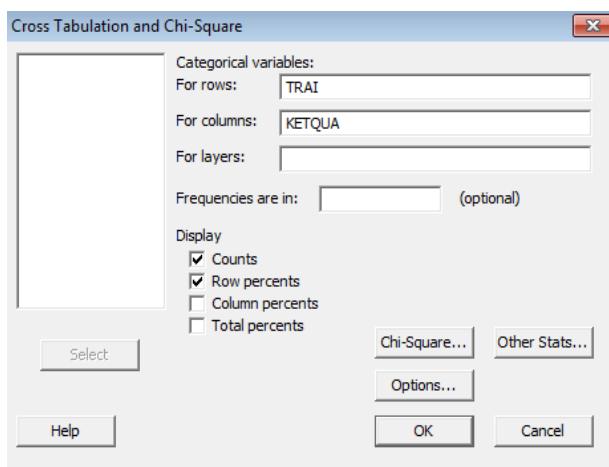
Nhập dữ liệu vào 2 cột, **Trại** vào cột C1 (TRAI) và cột **Kết quả xét nghiệm** vào cột C2 (KETQUA).

Lưu ý: Sau khi nhập thông tin vào cột C1 và C2 ký hiệu cột thay đổi thành C1-T và C2-T. (Minitab thông báo các thông tin trong cột không phải dạng số mà dạng ký tự (Text)).

Với số liệu ở dạng thô (cách 1) có thể tạo thành bảng tóm tắt như ở cách 2 bằng các lệnh sau:

Stat → Tables → Cross Tabulation and Chi-Square...

Vào ô **For rows** và **For columns**:



Options **Display** hiển thị:

Count tần số đối với từng trường hợp;

Row percents tỷ lệ (phần trăm) theo hàng;

Column percents tỷ lệ (phần trăm) theo cột;

Total percents tỷ lệ (phần trăm) theo hàng/cột tổng số.

Chọn **OK** để có kết quả

Tabulated statistics: TRAI, KETQUA

	Rows: TRAI	Columns: KETQUA	
	-	+	All
A	11	6	17
	64.71	35.29	100.00
B	16	6	22
	72.73	27.27	100.00
C	12	8	20
	60.00	40.00	100.00
All	39	20	59
	66.10	33.90	100.00

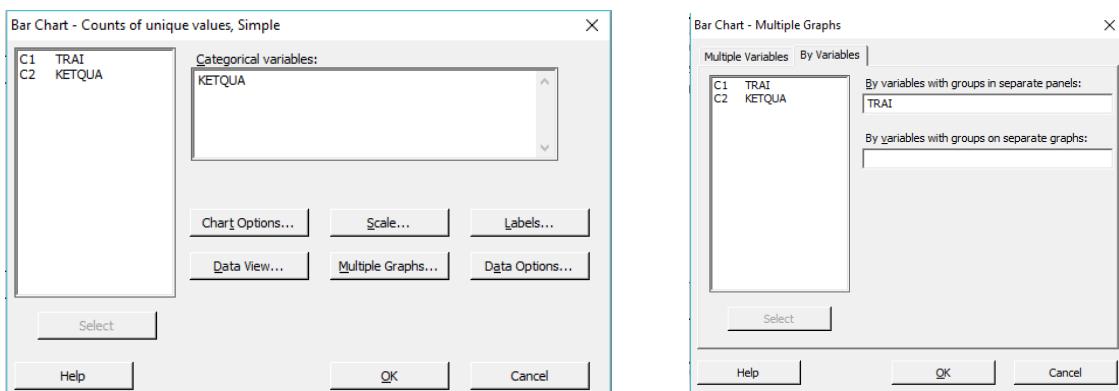
Cell Contents: Count
% of Row

Đối với biến định tính có thể mô tả bằng biểu đồ thanh (Bar Chart), biểu đồ bánh (Pie Chart).

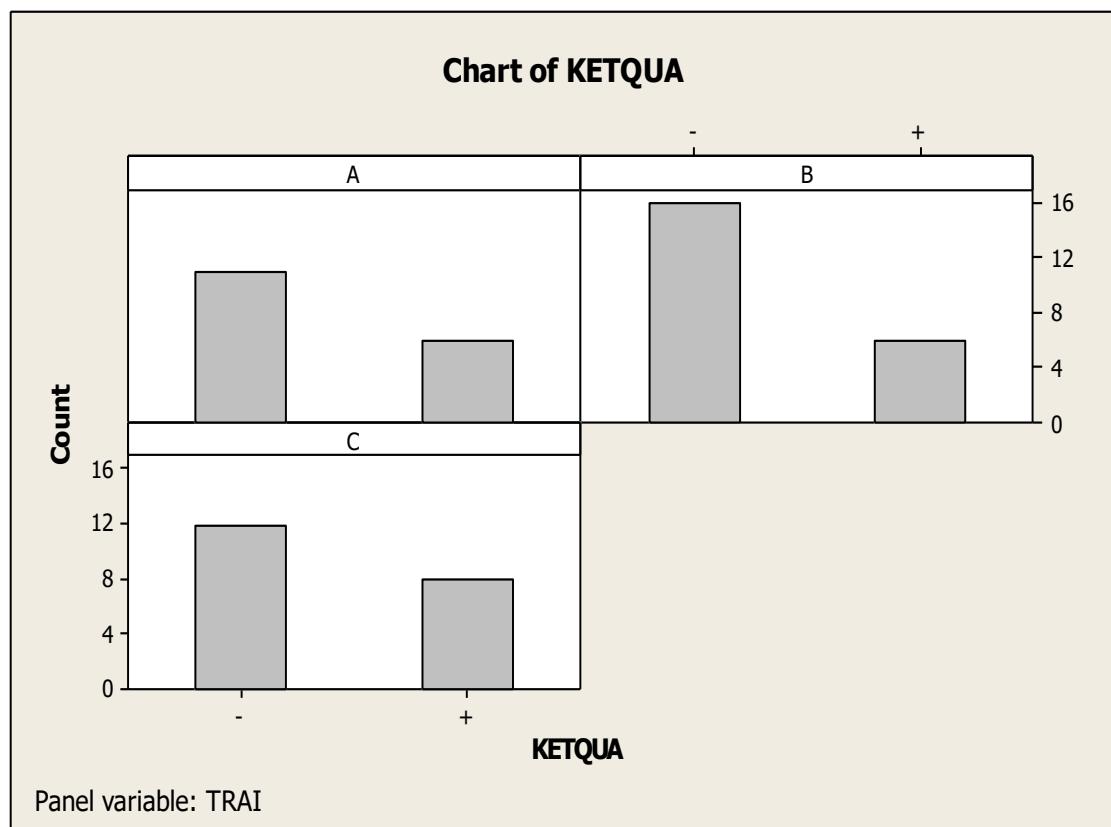
Graph → Bar Chart...Counts of unique values

Chọn OK

Chọn Multiple Graphs...



Chọn OK để có biểu đồ thanh



Cách 2: Số liệu được nhập vào cột trong Windows Worksheet

Minitab - VIDUM1.2.MPJ - [Worksheet 2 ***]

	C1-T	C2-T	C3	C4
	TRAI	KQ	TS	
1	A	+	6	
2	A	-	11	
3	B	+	6	
4	B	-	16	
5	C	+	8	
6	C	-	12	
7				
8				

Nhập dữ liệu vào 3 cột, **Trại** vào cột C1 (TRAI), cột **Kết quả xét nghiệm** vào cột C2 (KETQUA) và **Tần suất** vào cột C3 (TANSUAT).

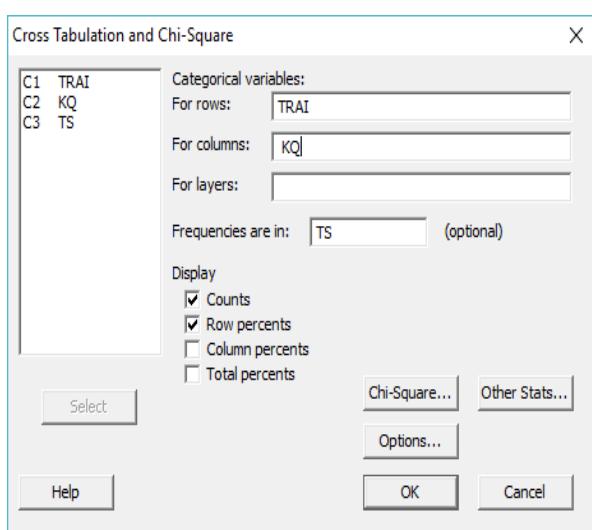
Dạng ký tự (Text).

Với số liệu ở dạng thô (cách 1) có thể tạo thành bảng tóm tắt như ở cách 2 bằng các lệnh sau:

Stat → Tables → Cross Tabulation and Chi-Square...

Khai báo vào ô **For rows, For columns** và **Frequencies are in**

Chọn **Counts** và **Row percents** trong
Display để có kết quả



Tabulated statistics: TRAI, KQ

Using frequencies in TS

Rows: TRAI Columns: KQ

- + All

A	11	6	17
64.71	35.29	100.00	

B	16	6	22
72.73	27.27	100.00	

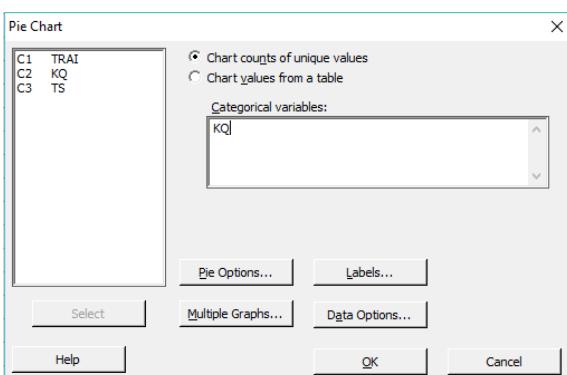
C	12	8	20
60.00	40.00	100.00	

All	39	20	59
66.10	33.90	100.00	

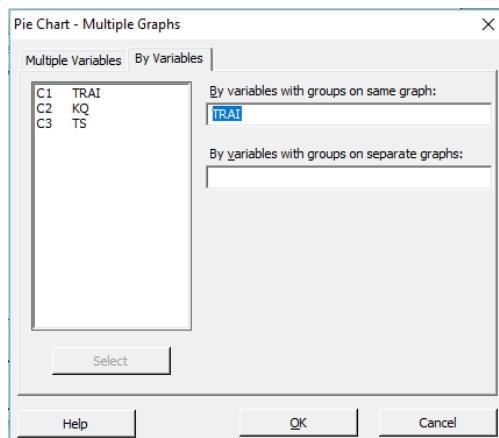
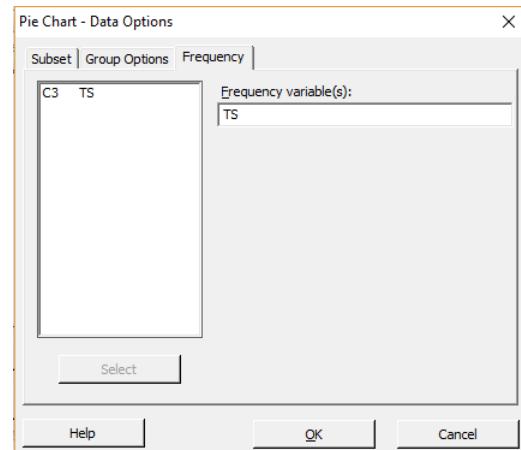
Cell Contents: Count
% of Row

Graph → Pie Chart... Counts of unique values

Chọn OK



Chọn Multiple Graphs...



Chọn OK để có biểu đồ bánh

BÀI 2. ƯỚC LUỢNG, KIỂM ĐỊNH MỘT GIÁ TRỊ TRUNG BÌNH VÀ SO SÁNH HAI GIÁ TRỊ TRUNG BÌNH

2.1. ƯỚC LUỢNG VÀ KIỂM ĐỊNH MỘT GIÁ TRỊ TRUNG BÌNH

2.1.1. Kiểm định phân phối chuẩn

Đối với tất cả các phép thử đối với biến định lượng, đều giả thiết rằng số liệu thu thập được (số liệu thô) tuân theo phân phối chuẩn. Nếu số liệu không tuân theo phân phối chuẩn thì các phép thử dưới đây sẽ không có hiệu lực. Trong trường hợp này cần biến đổi số liệu về phân phối chuẩn hoặc sử dụng kiểm định phi tham số. Giả thiết của phép thử:

H₀: Số liệu phân bố chuẩn và **H₁:** Số liệu không phân bố chuẩn

Ví dụ M-1.3: Tăng khối lượng trung bình (g/ngày) của 36 lợn nuôi vỗ béo giống Landrace được rút ngẫu nhiên từ một trại chăn nuôi. Số liệu thu được như sau:

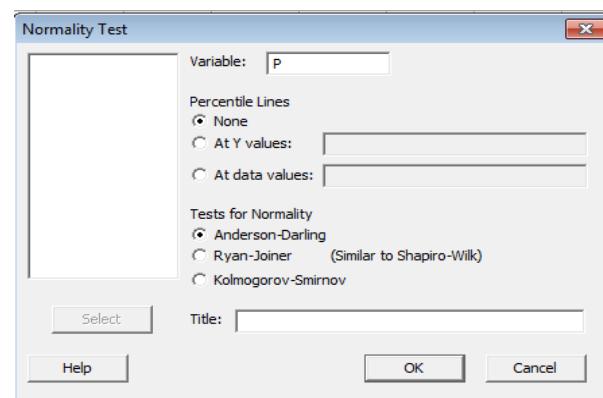
577 596 594 612 600 584 618 627 588 601 606 559 615 607 608
591 565 586
621 623 598 602 581 631 570 595 603 605 616 574 578 600 596
619 636 589

Cán bộ kỹ thuật trại cho rằng tăng khối lượng trung bình của toàn đàn lợn trong trại là 607 g/ngày. Theo anh (chị) kết luận đó đúng hay sai, vì sao? Biết rằng độ lệch chuẩn của tính trạng này là 21,75 g.

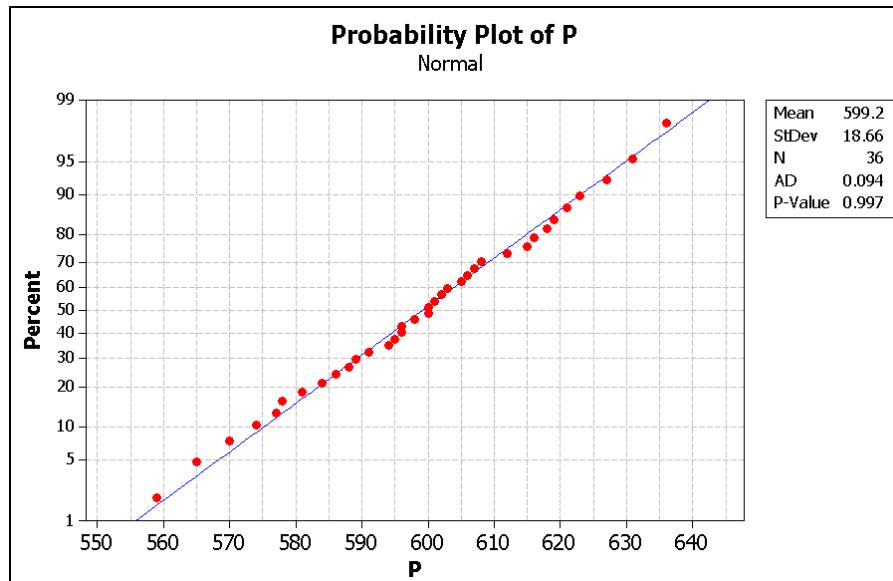
Nhập số liệu vào Worksheet

	C1	C2	C3	C4
1	577			
2	596			
3	594			
4	612			
5	600			
6	584			
7	618			
8	627			
9	588			
10	601			

Stat → Basic Statistics → Normality Test...



Chọn OK để có kết quả.

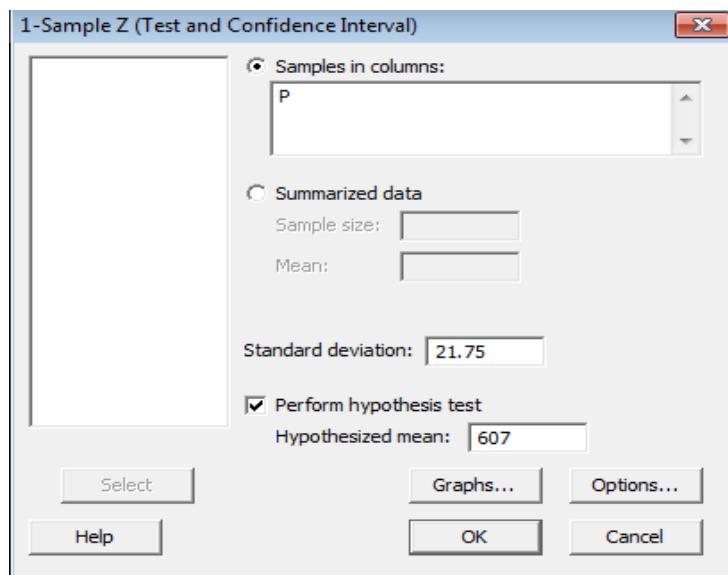


Giá trị **P-Value = 0,997** trong đồ thị trên lớn hơn 0,05 (α), như vậy H_0 được chấp nhận. Kết luận số liệu tuân theo phân phối chuẩn.

2.1.2. Kiểm định Z

Sử dụng kiểm định Z để kiểm định giá trị trung bình khi biết độ lệch chuẩn của quân thê (σ). Minitab sẽ tính khoảng tin cậy (CI 95%) và thực hiện phép kiểm định. Đối với kiểm định 2 phía ta có giả thiết: $H_0: \mu = \mu_0$ với đối thiêt $\mu \neq \mu_0$; trong đó μ là giá trị trung bình của quân thê và μ_0 là giá trị trung bình của quân thê kiểm định. Sử dụng số liệu ở ví dụ M-1.3.

Stat → Basic Statistics → 1-sample Z...



Trong **Samples in columns** khai báo cột số liệu (P).

Trong **Standard deviation** điền giá trị **21,75** (độ lệch chuẩn của quần thể σ).

Kích chuột vào ô **Perform hypothesis test** và nhập giá trị **607** (giá trị quần thể kiểm định μ_0) vào ô **Hypothesized mean**.

Chọn **OK** để có kết quả.

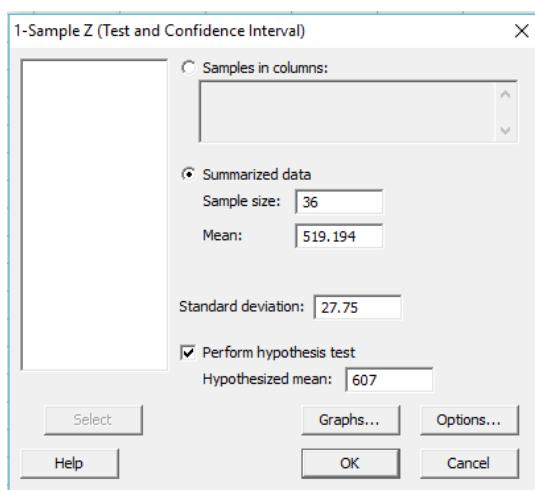
One-Sample Z: P

Test of mu = 607 vs not = 607

The assumed standard deviation = 21.75

Variable	N	Mean	StDev	SE Mean	95% CI	Z	P
P	36	599.194	18.656	3.625	(592.090; 606.299)	-2.15	0.031

Với xác suất của kiểm định $P = 0,031 < 0,05 (\alpha)$, bác bỏ H_0 và chấp nhận đối thiết H_1 . Kết luận: Tăng khối lượng của lợn Landrace ở trại nêu trên không đạt được 607 g/ngày ($P < 0,05$). Như vậy, cán bộ kỹ thuật trại kết luận sai và khoảng tin cậy 95% là 592,090 – 606,299 g/ngày.



Lưu ý: Trong một số trường hợp, số liệu đã được tóm tắt (số liệu đã đúc kết) dưới dạng các thống kê mô tả. Như ở **ví dụ 1.3** ta có $n = 36$; $\bar{x} = 599,194$ g. Vì vậy các giá trị này có thể sử dụng để khai báo trong lựa chọn **Summarized data**, các giá trị khác (σ và μ) được khai báo tương tự để có kết quả sau.

One-Sample Z

Test of mu = 607 vs not = 607

The assumed standard deviation = 21.75

N	Mean	SE Mean	95% CI	Z	P
36	599.194	3.625	(592.089; 606.299)	-2.15	0.031

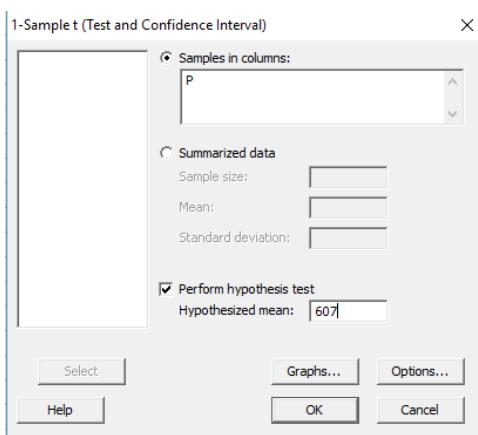
2.1.3. Kiểm định T

Trong trường hợp không có độ lệch chuẩn của quần thể (σ), kiểm định T được sử dụng để kiểm định giá trị trung bình và độ lệch chuẩn của mẫu (s) được sử dụng thay

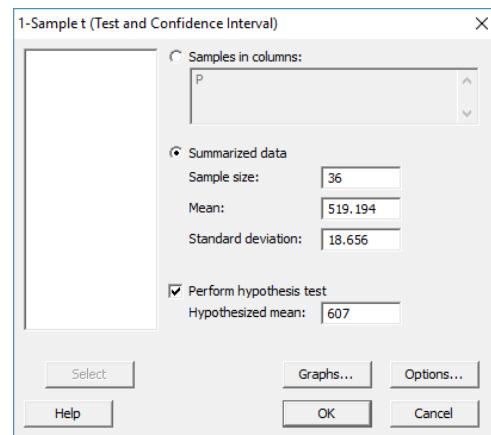
thé độ lệch chuẩn σ quan thê. Giả thiết của kiểm định, cấu trúc số liệu tương tự như ở kiểm định Z.

Stat → Basic Statistics → 1-sample t...

Khai báo đối với số liệu thô



... với số liệu đã tóm tắt (số liệu tinh)



Chọn OK để có kết quả

One-Sample T: P

Test of mu = 607 vs not = 607

Variable	N	Mean	StDev	SE Mean	95% CI	T	P
P	36	599.194	18.656	3.109	(592.882; 605.507)	-2.51	0.017

Với $P = 0,017$ ta cũng có kết luận tương tự như đối với khi sử dụng kiểm định Z.

2.2. SO SÁNH HAI GIÁ TRỊ TRUNG BÌNH

Khi tiến hành thí nghiệm để so sánh sự khác nhau giữa 2 công thức thí nghiệm, có 2 trường hợp chọn mẫu có thể xảy ra: 1) Chọn mẫu độc lập và 2) chọn mẫu theo cặp (xem 2.4, tr.23, Giáo trình Thiết kế thí nghiệm 2007). Tuỳ thuộc vào cách chọn mẫu bô trí thí nghiệm ta có thể sử dụng phép thử T hay T cặp cho phù hợp.

2.2.1. Kiểm định sự đồng nhất của phương sai

Đối với kiểm định 2 giá trị trung bình, ngoài giả thiết số liệu tuân theo phân phối chuẩn còn một vấn đề thứ 2 đặt ra là *Hai phương sai có bằng nhau hay không?*

Đối với kiểm định hai phía ta có giả thiết H_0 : Hai phương sai bằng nhau ($\sigma^2_1 = \sigma^2_2$) và H_1 : Hai phương sai không bằng nhau ($\sigma^2_1 \neq \sigma^2_2$). Khi chấp nhận giả thiết H_0 , phương sai chung (σ^2) sẽ được sử dụng để tiến hành kiểm định trong phép thử T; ngược lại (bác bỏ H_0) thì phép thử T **gần chính xác** sẽ được thực hiện.

Ví dụ M-1.4: Để so sánh khối lượng của 2 giống bò, tiến hành chọn ngẫu nhiên và cân 12 con đực với giống thứ nhất và 15 con đực với giống thứ 2. Khối lượng (kg) thu được như sau:

Giống bò thứ nhất	187,6	180,3	198,6	190,7	196,3	203,8	190,2	201,0
	194,7	221,1	186,7	203,1				
Giống bò thứ hai	148,1	146,2	152,8	135,3	151,2	146,3	163,5	146,6
	162,4	140,2	159,4	181,8	165,1	165,0	141,6	

Theo anh (chi), khối lượng của 2 giống bò có sự sai khác không?

Cấu trúc số liệu của bài toán kiểm định 2 giá trị trung bình có thể được trình bày bằng một trong 2 cách sau đây:

Cách 1: Số liệu của 2 công thức thí nghiệm được nhập vào một cột và cột thứ 2 để xác định giá trị của từng công thức.

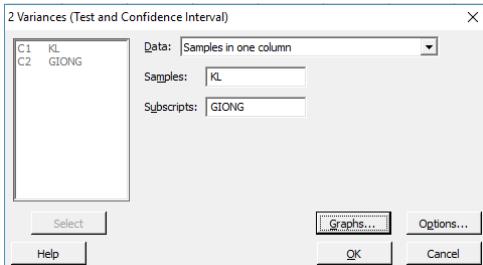
	C1	C2	C3	C4
	KL	GIONG		
1	187.6	1		
2	180.3	1		
3	198.6	1		
4	190.7	1		
5	196.3	1		

Cách 2: Số liệu được nhập vào 2 cột riêng biệt theo từng công thức thí nghiệm. Tên cột thể hiện giá trị trong mỗi công thức.

	C1	C2	C3	C4
	GIONG1	GIONG2		
1	187.6	148.1		
2	180.3	146.2		
3	198.6	152.8		
4	190.7	135.3		
5	196.3	151.2		

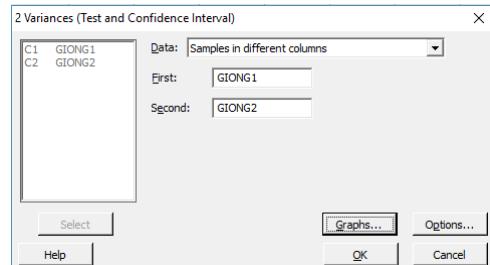
Stat → Basic Statistics → 2 Variances...

Cấu trúc số liệu thô cách 1

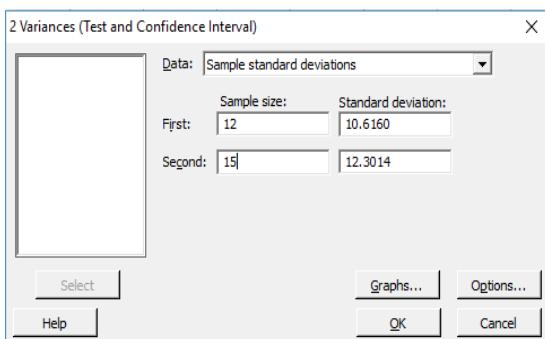


Đối với số liệu đã tóm tắt
ở dạng độ lệch chuẩn

... cách 2



Đối với số liệu đã tóm tắt
ở dạng phuông sai



Chọn OK để có kết quả

Test for Equal Variances: KL versus GIONG

95% Bonferroni confidence intervals for standard deviations

	GIONG	N	Lower	StDev	Upper
1	12	7.17875	10.6160	19.6238	
2	15	8.63359	12.3014	20.8502	

F-Test (normal distribution)

Test statistic = 0.74; p-value = 0.631

Levene's Test (any continuous distribution)

Test statistic = 0.46; p-value = 0.503

Xác suất p-value = 0,631 > 0,05 (α) vì vậy H_0 được chấp nhận. Kết luận *Hai phuông sai bằng nhau (P > 0,05)*.

2.2.2. Kiểm định T

Sử dụng kiểm định T để kiểm định 2 giá trị trung bình khi không biết độ lệch chuẩn của quần thể (σ). Minitab sẽ tính khoảng tin cậy (CI 95%) sự chênh lệch giữa 2 giá trị trung bình quần thể và thực hiện phép kiểm định. Đối với kiểm định 2 phía ta có giả thiết: $H_0: \mu_1 = \mu_2$ với đối thiết $H_1: \mu_1 \neq \mu_2$; trong đó μ_1 và μ_2 là giá trị trung bình của quần thể thứ nhất và thứ 2.

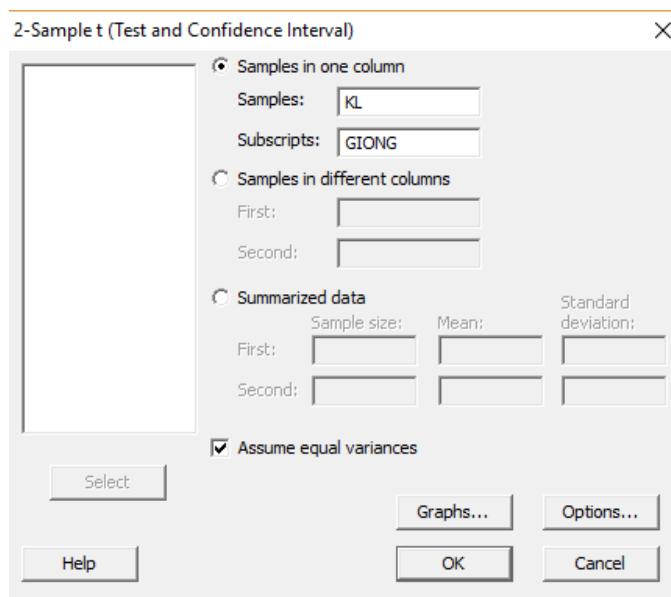
Stat → Basic Statistics → 2-Sample t...

Có thể sử dụng **Summarized data** khi đã tính các thống kê. Đối với trường hợp này cần khai báo dung lượng mẫu (**Sample size**), giá trị trung bình (**Mean**) và độ lệch chuẩn (**Standard deviation**) đối với từng công thức thí nghiệm tương ứng (**First** hoặc **Second**).

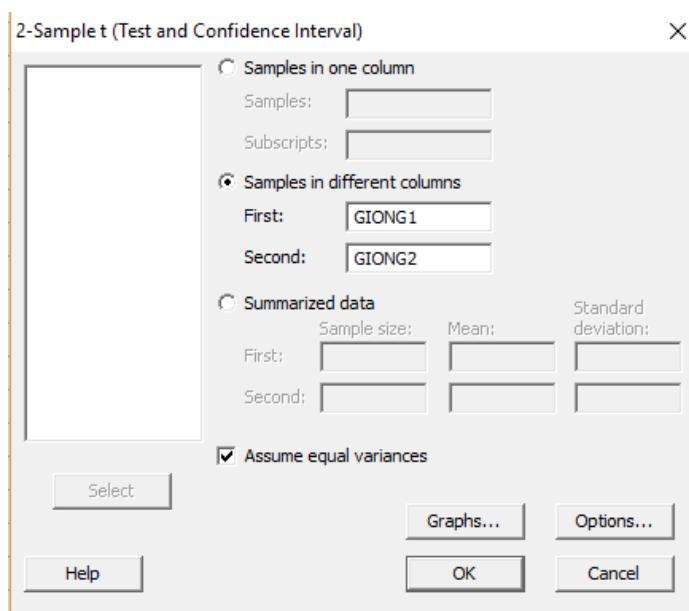
Chọn **Assume equal variances** nếu 2 phương sai bằng nhau và ngược lại nếu 2 phương sai không bằng nhau.

Chọn hiển thị đồ thị trong **Graphs...** và mức tin cậy trong **Options...**, theo mặc định Minitab tính khoảng tin cậy 95%.

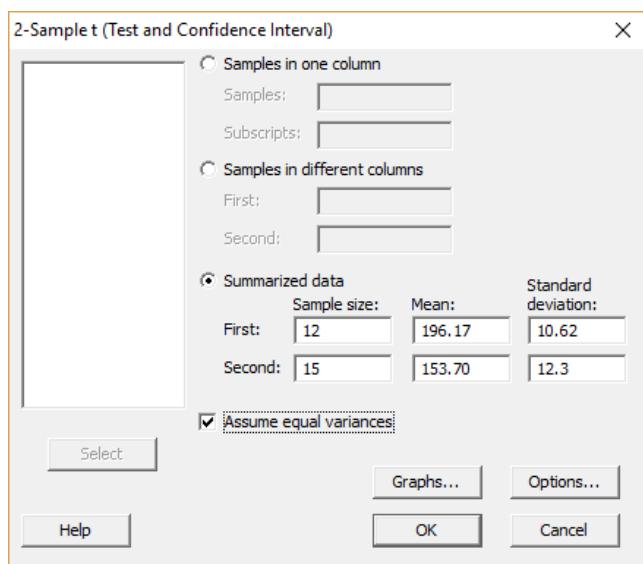
Với cấu trúc số liệu thô cách 1:



Cách 2:



Với số liệu đã được tóm tắt:



Chọn **OK** để có kết quả.

Two-Sample T-Test and CI: KL; GIONG

```
Two-sample T for KL
GIONG   N    Mean    StDev    SE Mean
1        12  196.2    10.6      3.1
2        15  153.7    12.3      3.2
Difference = mu (1) - mu (2)
Estimate for difference:  42.4750
95% CI for difference:  (33.2301; 51.7199)
T-Test of difference = 0 (vs not =: T-Value = 9.46 P-Value = 0.000 DF = 25
Both use Pooled StDev = 11.5901
```

Xác suất **P-value = 0,000 < 0,05 (α)** vì vậy H_0 bị bác bỏ và H_1 được chấp nhận. Kết luận rằng *Khối lượng của hai giống bò có sự sai khác ($P < 0,05$).*

2.2.3. Kiểm định T cặp

Đối với các thí nghiệm chọn mẫu theo cặp, điều kiện duy nhất của bài toán là kiểm tra phân phối chuẩn của chênh lệch (d) của cặp số liệu trong 2 công thức thí nghiệm.

Với kiểm định 2 phía ta có giả thiết $H_0: \mu_d = 0$ đối với $H_1: \mu_d \neq 0$ (μ_d là trung bình của sự chênh lệch giữa 2 trung bình μ_1 và μ_2).

Ví dụ M-1.5: Tăng khối lượng (kg) của 10 cặp bê sinh đôi giống hệt nhau dưới hai chế độ chăm sóc khác nhau (A và B). Bê trong từng cặp được bắt thăm ngẫu nhiên về một trong hai cách chăm sóc. Hãy kiểm định giả thiết H_0 : Tăng khối lượng trung bình ở hai cách chăm sóc như nhau, đối với H_1 : Tăng khối lượng trung bình khác nhau ở hai cách chăm sóc với mức ý nghĩa $\alpha = 0,05$. Số liệu thu được như sau:

	1	2	3	4	5	6	7	8	9	10
Tăng khối lượng ở cách A	19,50	17,69	17,69	19,05	20,87	19,50	17,24	19,96	23,13	19,50
Tăng khối lượng ở cách B	16,78	15,88	15,42	18,60	17,69	16,78	15,88	18,14	21,77	16,32
Chênh lệch (d)	2,72	1,81	2,27	0,45	3,18	2,72	1,36	1,82	1,36	3,18

Nhập số liệu vào Worksheet

	C1 A	C2 B
1	19.50	16.78
2	17.69	15.88
3	17.69	15.42
4	19.05	18.60
5	20.87	17.69
6	19.50	16.78
7	17.24	15.88
8	19.96	18.14
9	23.13	21.77
10	19.50	16.32

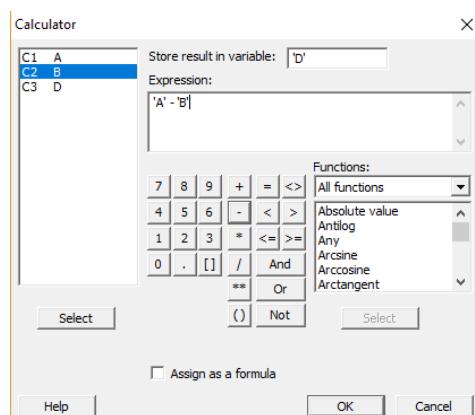
Lưu ý:

Số liệu được nhập vào Worksheet theo một cách duy nhất vào 2 cột theo từng cặp số liệu tương ứng.

Thứ tự các cặp số liệu không đóng vai trò quan trọng nhưng từng 1 cặp một luôn phải cùng nhau.

Sự thay đổi vị trí trong 1 cặp có thể đưa ta đến các kết luận thiếu chính xác.

Calc → Calculator...

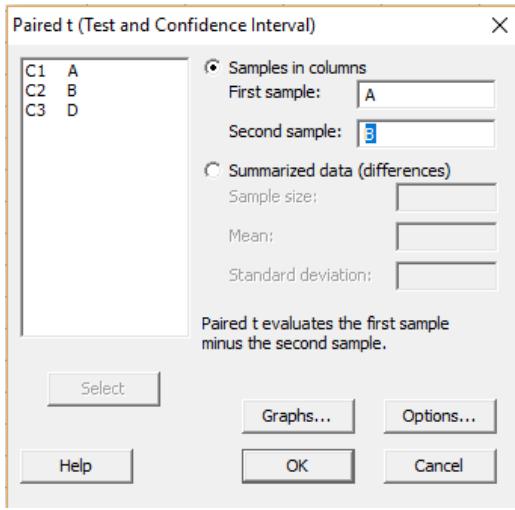


...chọn OK để có được phần chênh lệch

	C1 A	C2 B	C3 D
1	19.50	16.78	2.72
2	17.69	15.88	1.81
3	17.69	15.42	2.27
4	19.05	18.60	0.45
5	20.87	17.69	3.18
6	19.50	16.78	2.72
7	17.24	15.88	1.36
8	19.96	18.14	1.82
9	23.13	21.77	1.36
10	19.50	16.32	3.18

Tiến hành kiểm định phân bố chuẩn của phần chênh lệch **D**

Stat → Basic Statistics → Paired t...



Có thể sử dụng **Summarized data (differences)** khi sử dụng các thông tin của cột chênh lệch **D** để kiểm định.

Đối với trường hợp này cần khai báo dung lượng mẫu (**Sample size**), giá trị trung bình (**Mean**) và độ lệch chuẩn (**Standard deviation**) của cột **D**.

Chọn hiển thị đồ thị trong **Graphs...** và mức tin cậy trong **Options...**, theo mặc định Minitab tính khoảng tin cậy 95%.

Chọn **OK** để có kết quả

Paired T-Test and CI: A; B

Paired T for A - B

	N	Mean	StDev	SE Mean
A	10	19.413	1.734	0.548
B	10	17.326	1.874	0.592
Difference	10	2.087	0.889	0.281

95% CI for mean difference: (1.451, 2.723)
T-Test of mean difference = 0 (vs not = 0: T-Value = 7.43 P-Value = 0.000

Xác suất **P-value = 0,000 < 0,05** (α) vì vậy H_0 bị bác bỏ và H_1 được chấp nhận.
Kết luận *Tăng khối lượng trung bình ở hai cách chăm sóc có sự sai khác ($P < 0,05$).*

Bài 3. SO SÁNH NHIỀU GIÁ TRỊ TRUNG BÌNH

Phân tích phương sai (Analysis of Variance - ANOVA) là công cụ hữu ích để so sánh nhiều giá trị trung bình. Điều kiện của bài toán phân tích phương sai 1) số liệu tuân theo phân bố chuẩn và 2) phương sai đồng nhất. Trong khuôn khổ giáo trình này chúng tôi chỉ đề cập đến việc kiểm tra điều kiện của bài toán đối với các mô hình thiết kế thí nghiệm đơn giản (Thí nghiệm một yếu tố hoàn toàn ngẫu nhiên).

Giả thiết cần kiểm định $H_0: \mu_1 = \mu_2 = \dots = \mu_a$ đối với $H_1: \mu_1 \neq \mu_2 \neq \dots \neq \mu_a$ (μ là trung bình của quần thể ở công thức thí nghiệm thứ 1, 2, ...a).

3.1. THÍ NGHIỆM MỘT YẾU TỐ HOÀN TOÀN NGẪU NHIÊN

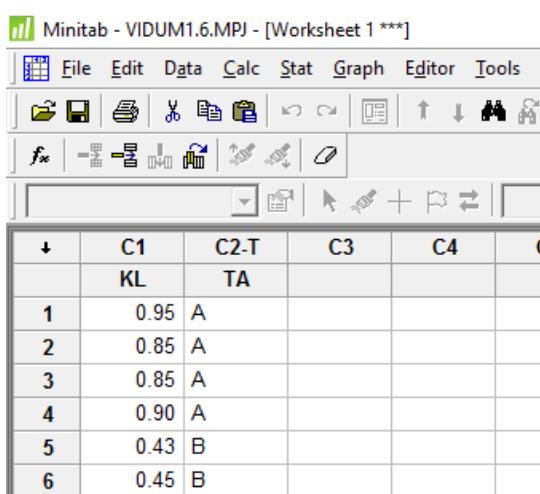
Xét trường hợp đơn giản nhất đối với bài toán phân tích phương sai. Chỉ có một yếu tố duy nhất trong thí nghiệm, các yếu tố phi thí nghiệm còn lại được coi là có tác động như nhau đến đối tượng thí nghiệm.

Ví dụ M-1.6: Theo dõi tăng khối lượng của cá (kg) trong thí nghiệm với 5 công thức nuôi (A, B, C, D và E). Hãy cho biết tăng khối lượng của cá ở các công thức nuôi. Nếu có sự khác nhau, tiến hành so sánh sự sai khác của từng cặp giá trị trung bình có thể bằng các chữ cái.

A	B	C	D	E
0,95	0,43	0,70	1,00	0,90
0,85	0,45	0,90	0,95	1,00
0,85	0,40	0,75	0,90	0,95
0,90	0,42	0,70	0,90	0,95

Cấu trúc số liệu của bài toán kiểm định nhiều giá trị trung bình có thể được trình bày bằng một trong 2 cách sau:

Cách 1: Số liệu của các công thức thí nghiệm được nhập vào một cột và cột thứ 2 để xác định giá trị của từng công thức



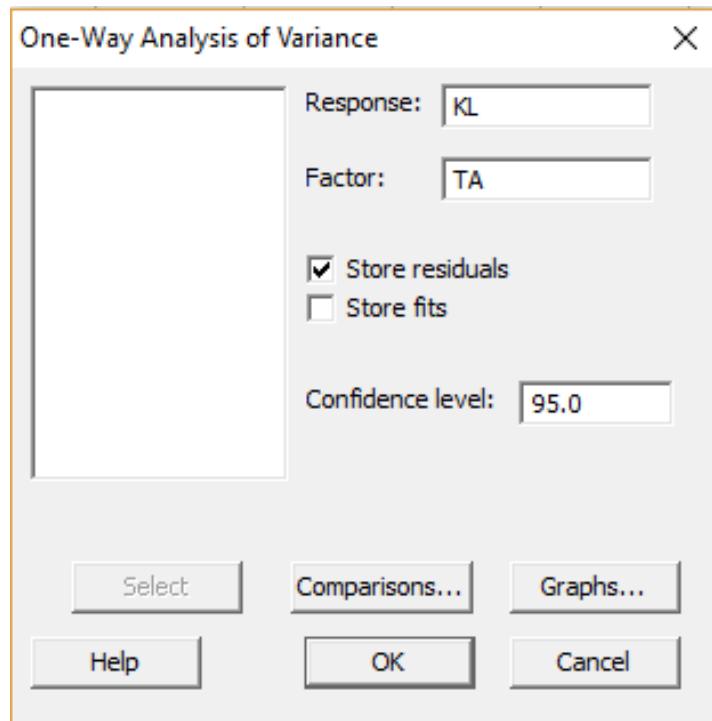
Cách 2: Số liệu được nhập vào các cột riêng biệt theo công thức thí nghiệm. Tên cột thể hiện giá trị trong mỗi công thức

	C1	C2	C3	C4	C5
	A	B	C	D	E
1	0.95	0.43	0.70	1.00	0.90
2	0.85	0.45	0.90	0.95	1.00
3	0.85	0.40	0.75	0.90	0.95
4	0.90	0.42	0.70	0.90	0.95
5					
6					

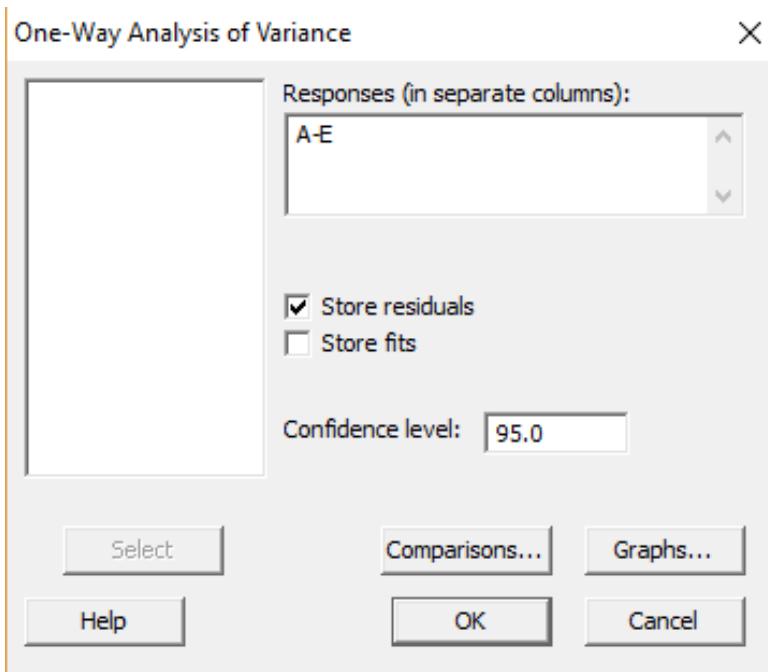
Kiểm tra điều kiện của bài toán (sự đồng nhất của phương sai và phân phối chuẩn của số liệu) sẽ được trình bày sau. Tiến hành so sánh các giá trị trung bình bằng phép phân tích phương sai (ANOVA) đối với cấu trúc số liệu cách 1 và cách 2.

Với các bài toán sử dụng phép phân tích phương sai để so sánh, cấu trúc số liệu cách 1 sẽ phù hợp và thuận lợi hơn trong quá trình xử lý số liệu. Trong các ví dụ tiếp theo chúng tôi chỉ đề cập đến việc xử lý số liệu có cấu trúc cách 1.

Stat → ANOVA → One-Way...



Stat → ANOVA → One-Way (Unstacked)...



Chọn OK để có kết quả

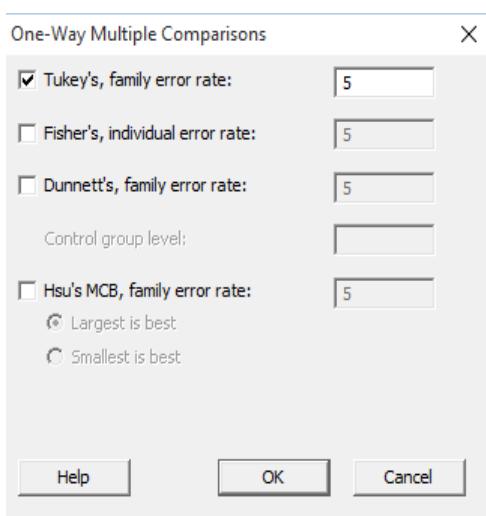
One-way ANOVA: KL versus TA

Source	DF	SS	MS	F	P
TA	4	0.76325	0.19081	60.99	0.000
Error	15	0.04693	0.00313		
Total	19	0.81018			
S = 0.05593 R-Sq = 94.21% R-Sq(adj) = 92.66%					
Individual 95% CIs For Mean Based on Pooled StDev					
Level	N	Mean	StDev	-----+-----+-----+-----	
A	4	0.8875	0.0479	(--*--)	
B	4	0.4250	0.0208	(--*--)	
C	4	0.7625	0.0946	(--*--)	
D	4	0.9375	0.0479	(--*--)	
E	4	0.9500	0.0408	(-*--)	
-----+-----+-----+-----					
0.40 0.60 0.80 1.00					
Pooled StDev = 0.0559					

Xác suất P-value = 0,000 < 0,05 (α) vì vậy H_0 bị bác bỏ và H_1 được chấp nhận.
Kết luận rằng Tăng khói lượng trung bình của cá ở các công thức thức ăn có sự sai khác ($P < 0,05$).

So sánh cặp khi bác bỏ giả thiết H_0 chấp nhận giả thiết H_1

Chọn Comparisons... trong hộp thoại One-Way Analysis of Variances



Các lựa chọn:

Tukey's, family error rate: với sai số của toàn bộ các cặp so sánh là 5%.

Fisher's, individual error rate: với sai số của từng cặp so sánh là 5%.

Dunnett's, family error rate: so sánh với nhóm đối chứng, sai số của toàn bộ các cặp so sánh là 5%.

Hsu's MCB, family error rate: với sai số của toàn bộ các cặp so sánh là 5%.

Chọn OK để có kết quả.

Mỗi phương pháp so sánh cặp đều có những ưu điểm và hạn chế. Để lựa chọn được phương pháp so sánh cặp phù hợp, bạn đọc có thể tham khảo Giáo trình Phân tích số liệu thí nghiệm và công bố kết quả nghiên cứu chăn nuôi (Nguyễn Xuân Trạch và Đỗ Đức Lực, 2016).

Tukey 95% Simultaneous Confidence Intervals
 All Pairwise Comparisons among Levels of TA
 Individual confidence level = 99.25%

TA = A subtracted from:

TA	Lower	Center	Upper	
B	-0.58471	-0.46250	-0.34029	(---*---)
C	-0.24721	-0.12500	-0.00279	(---*---)
D	-0.07221	0.05000	0.17221	(---*---)
E	-0.05971	0.06250	0.18471	(---*---)

TA = B subtracted from:

TA	Lower	Center	Upper	
C	0.21529	0.33750	0.45971	(---*---)
D	0.39029	0.51250	0.63471	(---*---)
E	0.40279	0.52500	0.64721	(---*---)

TA = C subtracted from:

TA	Lower	Center	Upper	
D	0.05279	0.17500	0.29721	(---*---)

	E	0.06529	0.18750	0.30971	(--*--)
					-0.35 0.00 0.35 0.70
TA = D subtracted from:					
TA	Lower	Center	Upper		(--*--)
E	-0.10971	0.01250	0.13471		-0.35 0.00 0.35 0.70

Ngoài kết quả phân tích phuơng sai như phần trên, Minitab đã cung cấp kết quả so sánh từng cặp. Sự sai khác có ý nghĩa ($P < 0,05$) giữa các công thức thí nghiệm dựa trên khoảng tin cậy của từng cặp. **Không có sự sai** khác giữa các công thức thí nghiệm nếu khoảng tin cậy có chứa số 0 (giá trị cận dưới và cận trên cùng dấu) và ngược lại có **sự sai khác** nếu không chứa số 0 (giá trị cận dưới và cận trên khác dấu). Ví dụ trong kết quả nêu trên nếu so sánh giữa A-B ta có khoảng tin cậy (-0,58471; -0,34029) không chứa số 0 (cùng dấu -) nên kết luận không có sự sai khác giữa A và B ($P > 0,05$). Nếu so sánh A và D ta có khoảng tin cậy (-0,07221; +17221) có chứa số 0 (khác dấu) nên kết luận có sự sai khác giữa A và D ($P < 0,05$). Để có thể trình bày kết quả so sánh cặp, bạn đọc có thể tham khảo chương 4 của giáo trình này.

Bên cạnh việc cung cấp kết quả so sánh từng cặp giữa các công thức thí nghiệm dựa trên khoảng tin cậy của từng cặp, Minitab 16 còn thể hiện sự sai khác bằng cách tạo thành các nhóm với các chữ cái a, b, c, ...

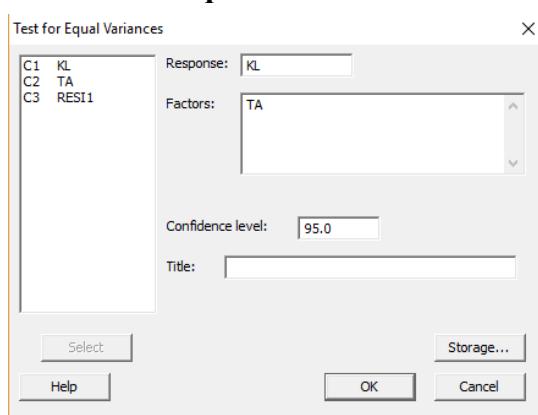
Grouping Information Using Tukey Method

TA	N	Mean	Grouping
E	4	0.95000	A
D	4	0.93750	A
A	4	0.88750	A
C	4	0.76250	B
B	4	0.42500	C

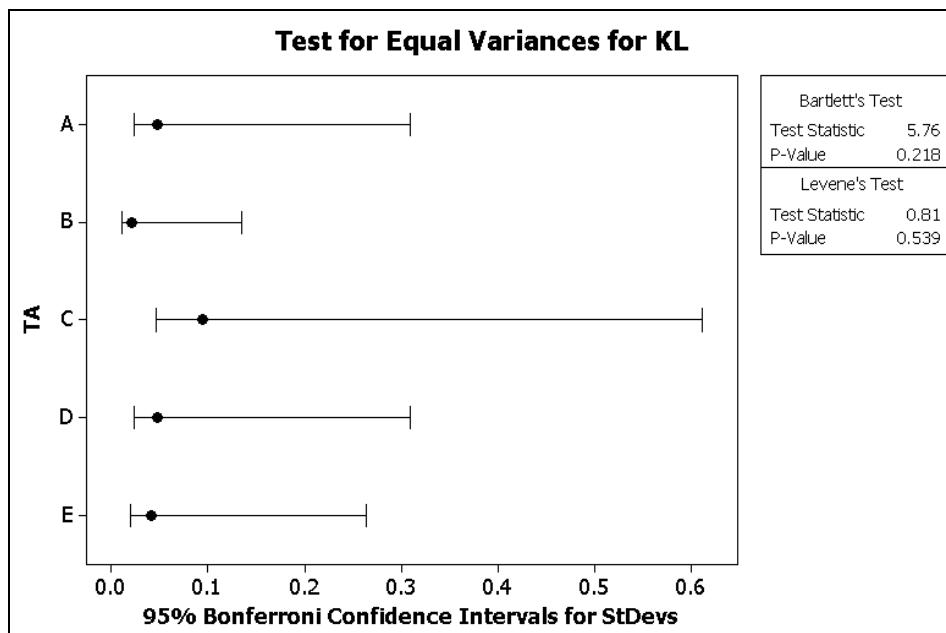
Means that do not share a letter are significantly different.

Kiểm tra sự đồng nhất của phuơng sai với cấu trúc số liệu cách 1

Stat ➔ ANOVA ➔ Test for Equal Variances...



Chọn **OK** để hiển thị đồ thị và kết quả trong cửa sổ Session:



Test for Equal Variances: KL versus TA

95% Bonferroni confidence intervals for standard deviations

TA	N	Lower	StDev	Upper
A	4	0.0231412	0.0478714	0.309607
B	4	0.0100628	0.0208167	0.134631
C	4	0.0457534	0.0946485	0.612137
D	4	0.0231412	0.0478714	0.309607
E	4	0.0197348	0.0408248	0.264034

Bartlett's Test (normal distribution)

Test statistic = 5.76; p-value = 0.218

Levene's Test (any continuous distribution)

Test statistic = 0.81; p-value = 0.539

Xác suất **p-value = 0,539 > 0,05 (α)** vì vậy H_0 được chấp nhận. Kết luận rằng *Các Phuong sai bằng nhau (P > 0,05)*.

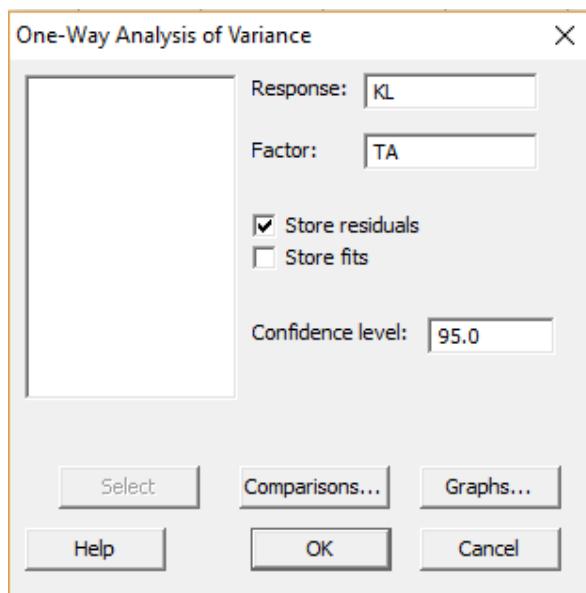
Kiểm tra phân phối chuẩn với cấu trúc số liệu cách 1

Không tiến hành kiểm tra phân bố chuẩn của cột số thô (**KL**) mà tiến hành kiểm tra phần sai **số ngẫu nhiên ε_{ij}** theo mô hình:

$$y_{ij} = \mu + a_i + \varepsilon_{ij} \quad (i = 1, a; j = 1, r_i)$$

Trong đó y_{ij} = quan sát thứ j ở công thức i , μ = trung bình chung, a_i = chênh lệch do ảnh hưởng của công thức i và ε_{ij} = sai số ngẫu nhiên; các ε_{ij} độc lập, phân phối chuẩn $N \sim (0, \sigma^2)$. Nếu phần sai số ngẫu nhiên tuân theo phân phối chuẩn thì số liệu bài toán cũng có phân phối chuẩn.

Stat → ANOVA → One-Way...



Chọn **Store residuals** và **OK** để có **RESI1** (ε_{ij})

Minitab - VIDUM1.6.MPJ - [Worksheet 1 ***]

	C1	C2-T	C3	C4
	KL	TA	RESI1	
1	0.95	A	0.0625	
2	0.85	A	-0.0375	
3	0.85	A	-0.0375	
4	0.90	A	0.0125	
5	0.43	B	0.0050	
6	0.45	B	0.0250	

Tiến hành kiểm tra phân phối chuẩn của cột số liệu RESI1 (xem 3.1 Kiểm định phân phối chuẩn). Phép kiểm định sẽ cho ta **P-Value = 0,159 > 0,05** (α) nên có thể kết luận *Số liệu tuân theo phân phối chuẩn* ($P > 0,05$).

Lưu ý:

Với cấu trúc số liệu cách 2, có thể kiểm định phân phối chuẩn của số liệu với từng công thức thí nghiệm riêng biệt. Kết quả kiểm định, xác suất để số liệu ở các công thức thí nghiệm A, B, C, D và E có phân phối chuẩn lần lượt là 0,255; 0,845; 0,092; 0,255 và 0,410. Ta cũng có kết luận tương tự.

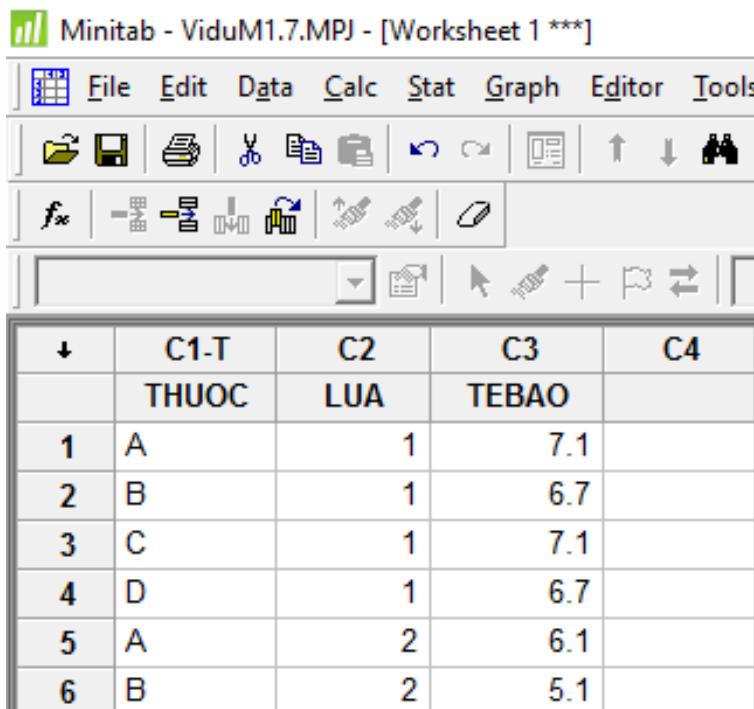
3.2. THÍ NGHIỆM MỘT YẾU TỐ KHỐI NGẦU NHIÊN ĐÀY ĐỦ

Xem xét một thí nghiệm mà đối tượng thí nghiệm chịu tác động đồng thời của một yếu tố chính (yếu tố thí nghiệm) và yếu tố phụ (khối).

Ví dụ M-1.7: Nghiên cứu số lượng tế bào lymphô ở chuột ($\times 1000$ tế bào mm^{-3} máu) được sử dụng 4 loại thuốc khác nhau (A, B, C và D; thuốc D là placebo) qua 5 lứa; số liệu thu được trình bày ở bảng dưới. Cho biết ảnh hưởng của thuốc đến tế bào lymphô?

	Lứa 1	Lứa 2	Lứa 3	Lứa 4	Lứa 5
Thuốc A	7,1	6,1	6,9	5,6	6,4
Thuốc B	6,7	5,1	5,9	5,1	5,8
Thuốc C	7,1	5,8	6,2	5,0	6,2
Thuốc D	6,7	5,4	5,7	5,2	5,3

Cấu trúc số liệu



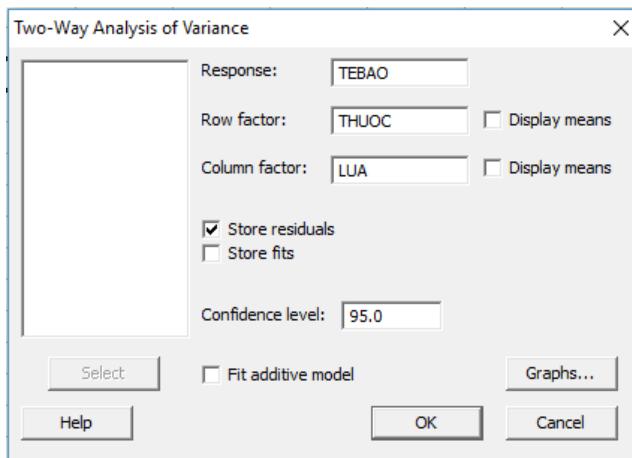
Số liệu của bài toán này có một cấu trúc duy nhất trong Minitab; bao gồm 3 cột:

- 1) cột Số lượng tế bào C1 (TEBAO),
- 2) cột Thuốc C2 (THUOC) và
- 3) cột Lứa C3 (LUA)

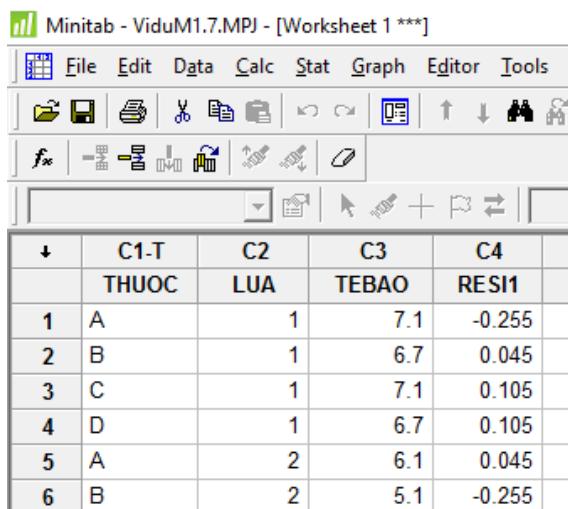
Trong thí nghiệm này đối tượng thí nghiệm bị tác động bởi yếu tố chính (yếu tố thí nghiệm) và yếu tố phụ (khối)

So sánh sai khác giữa công thức thí nghiệm bằng Phân tích phương sai (ANOVA).

Stat → ANOVA → Two-Way...



Chọn **Store residuals** để có RESI1:



Chọn **OK** để có kết quả.

Two-way ANOVA: TEBAO versus THUOC; LUA

Source	DF	SS	MS	F	P
THUOC	3	1.8455	0.61517	11.59	0.001
LUA	4	6.4030	1.60075	30.16	0.000
Error	12	0.6370	0.05308		
Total	19	8.8855			

S = 0.2304 R-Sq = 92.83% R-Sq(adj) = 88.65%

Xác suất của phép thử đối với *Thuốc* **P = 0,001 < 0,05** (α), bác bỏ giả thiết H₀ và chấp nhận đối thiêt H₁. Kết luận thuốc có ảnh hưởng khác nhau lên tế bào lymphô của chuột ($P < 0,05$).

So sánh sự sai khác giữa các công thức thí nghiệm bằng cách tạo thành các nhóm với các chữ cái a, b, c, ...

Grouping Information Using Tukey Method and 95.0% Confidence

THUOC	N	Mean	Grouping
A	5	6.4	A
C	5	6.1	A B
B	5	5.7	B
D	5	5.7	B

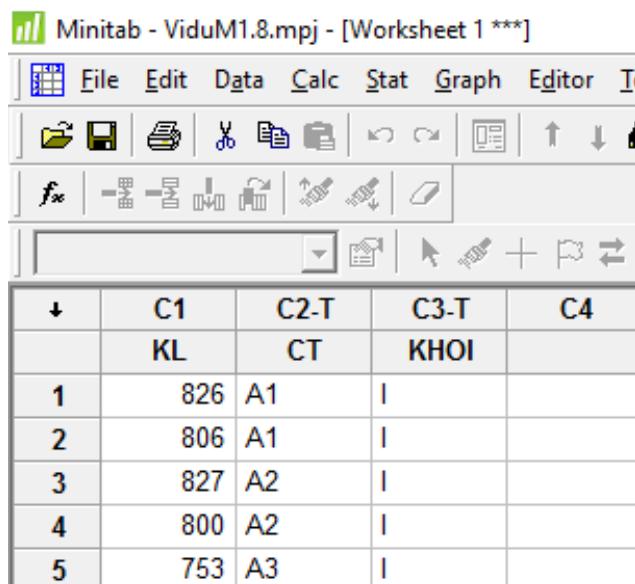
Means that do not share a letter are significantly different.

Ví dụ M-1.8: Một thí nghiệm được tiến hành để xác định ảnh hưởng của 3 công thức thức ăn (A1, A2 và A3) đến tăng khối lượng trung bình trên ngày (g/ngày) của bê đực. Bê đực được cân và chia thành 4 khối dựa theo khối lượng bắt đầu thí nghiệm. Trong mỗi khối có 6 động vật thí nghiệm được chọn ra và được phân ngẫu nhiên về với các công thức thí nghiệm. Số liệu thu thập sau khi kết thúc thí nghiệm như sau:

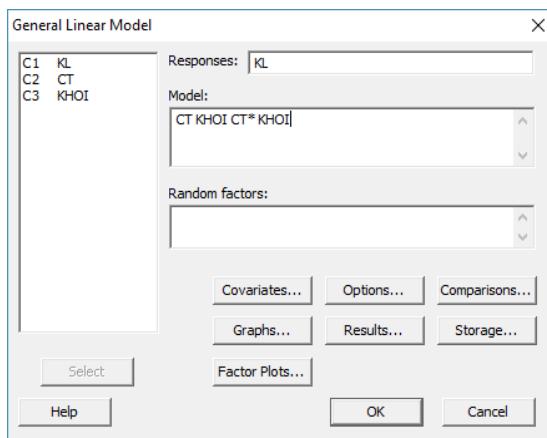
	Khối			
	I	II	III	IV
A1	826	864	795	850
	806	834	810	845
A2	827	871	729	860
	800	881	709	840
A3	753	801	736	820
	773	821	740	835

Cấu trúc số liệu mô hình thí nghiệm trong ví dụ 1.8 tương tự như ở ví dụ 1.7.

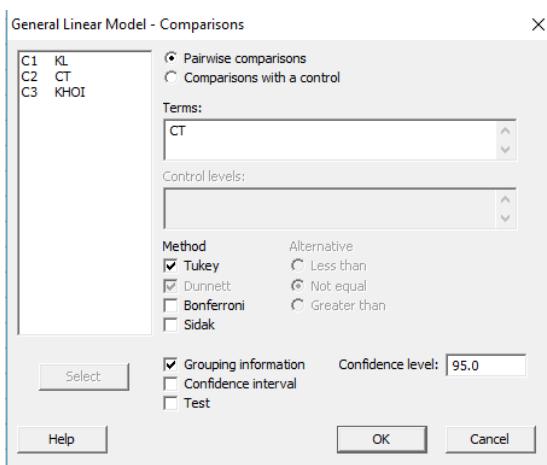
Trong ví dụ 1.8 có 2 đơn vị thí nghiệm ở một công thức thí nghiệm và khối vì vậy ngoài tác động của **khối** và **công thức thí nghiệm** còn tồn tại sự **tương tác** giữa khối và công thức thí nghiệm.



Stat → ANOVA → General Linear Model...



Chọn Comparisons để so sánh cặp đôi



Chọn OK để có kết quả.

General Linear Model: KL versus CT; KHOI

Factor	Type	Levels	Values			
CT	fixed	3	A1; A2; A3			
KHOI	fixed	4	I; II; III; IV			
Analysis of Variance for KL, using Adjusted SS for Tests						
Source	DF	Seq SS	Adj SS	Adj MS	F	P
CT	2	8025.6	8025.6	4012.8	22.82	0.000
KHOI	3	33816.8	33816.8	11272.3	64.11	0.000
CT*KHOI	6	8087.4	8087.4	1347.9	7.67	0.001
Error	12	2110.0	2110.0	175.8		
Total	23	52039.8				
S = 13.2602	R-Sq = 95.95%	R-Sq(adj) = 92.23%				

Xác suất của phép thử đối với yếu tố *Thức ăn* **P = 0,000** và tương tác (*CT*KHOI*) **P = 0,001 < 0,05**, bác bỏ giả thiết H_0 và chấp nhận đối thiết H_1 . Kết luận công thức ăn

có ảnh hưởng đến tăng khói lượng của bê và có tương tác giữa công thức thức ăn và khói lượng bê vỗ béo ($P < 0,05$).

So sánh sự sai khác giữa các công thức thí nghiệm bằng cách tạo thành các nhóm với các chữ cái a, b, c, ...

```
Grouping Information Using Tukey Method and 95.0% Confidence
CT N Mean Grouping
A1 8 828.8 A
A2 8 814.5 A
A3 8 784.9 B
Means that do not share a letter are significantly different.
```

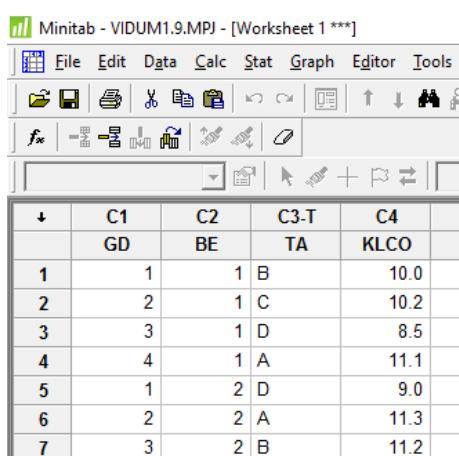
3.3. THÍ NGHIỆM Ô VUÔNG LA TINH

Đối với mô hình thí nghiệm Ô vuông latin, ngoài yếu tố thí nghiệm ta còn 2 yếu tố khác (yếu tố hàng và cột) tác động lên đối tượng thí nghiệm.

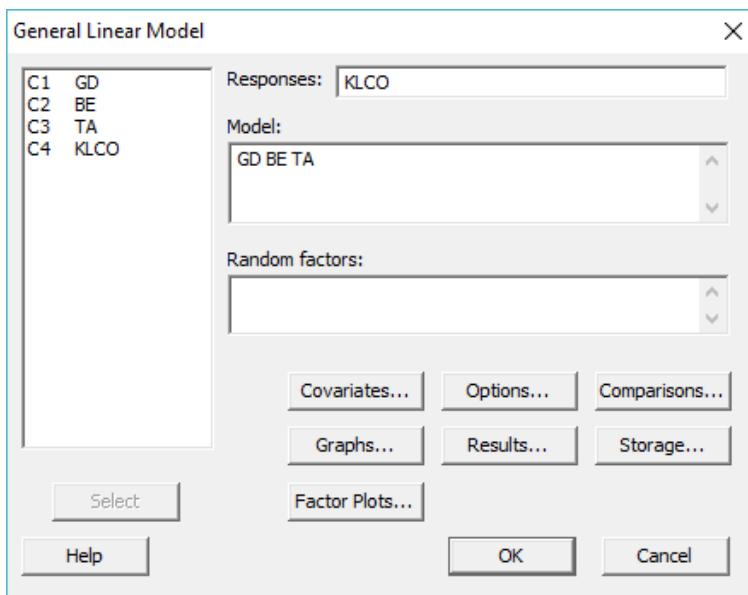
Ví dụ M-1.9a: Một thí nghiệm được tiến hành nhằm xác định ảnh hưởng của các loại thức ăn bổ sung khác nhau (A, B, C và D) đến lượng cỏ khô mà bê nuôivỗ béo thu nhận được (kg/ngày). Thí nghiệm được thiết kế theo mô hình ô vuông la tinh với 4 động vật trong 4 giai đoạn, mỗi giai đoạn 20 ngày. Trong mỗi giai đoạn 10 ngày đầu được coi là giai đoạn thích nghi, 10 ngày tiếp theo là giai đoạn thí nghiệm để thu thập số liệu. Số liệu thu được là khói lượng cỏ khô trung bình bê thu nhận được ở 10 ngày thí nghiệm. Hãy rút ra kết luận từ thí nghiệm nêu trên.

Giai đoạn	Bê			
	1	2	3	4
1	10,0 (B)	9,0 (D)	11,1 (C)	10,8 (A)
	10,2 (C)	11,3 (A)	9,5 (D)	11,4 (B)
2	8,5 (D)	11,2 (B)	12,8 (A)	11 (C)
	11,1 (A)	11,4 (C)	11,7 (B)	9,9 (D)

Cấu trúc số liệu:



Stat → ANOVA → General Linear Model...



Chọn OK để có kết quả

General Linear Model: KLCO versus GD; BE; TA

Factor	Type	Levels	Values
GD	fixed	4	1; 2; 3; 4
BE	fixed	4	1; 2; 3; 4
TA	fixed	4	A; B; C; D

Analysis of Variance for KLCO, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
GD	3	1.4819	1.4819	0.4940	3.41	0.094
BE	3	3.5919	3.5919	1.1973	8.27	0.015
TA	3	12.0219	12.0219	4.0073	27.68	0.001
Error	6	0.8688	0.8688	0.1448		
Total	15	17.9644				

S = 0.380515 R-Sq = 95.16% R-Sq(adj) = 87.91%

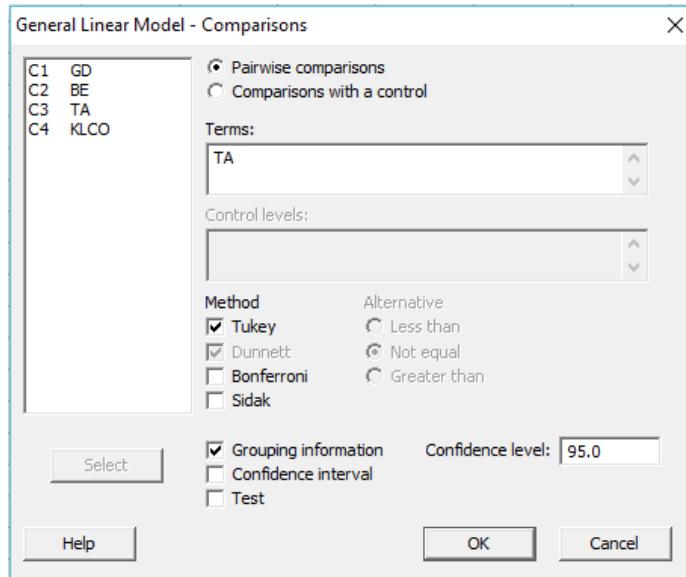
Unusual Observations for KLCO

Obs	KLCO	Fit	SE Fit	Residual	St Resid
11	12.8000	12.2875	0.3008	0.5125	2.20 R

R denotes an observation with a large standardized residual.

Kết quả phân tích cho thấy xác suất của kiểm định đối với yếu tố thí nghiệm (TA) P = 0,001, vì vậy giả thiết H₀ bị bác bỏ, kết luận *Có ảnh hưởng của thức ăn bổ sung đến lượng cỗ khô mà bê thu nhận được* (P < 0,05).

Ngoài ra Minitab cũng đã hiển thị giá trị bất thường (Unusual Observation) trong bộ số liệu nêu trên đối với mô hình xử lý thống kê đã lựa chọn. Giá trị này là **12,8000** nằm ở hàng thứ 11 của cột KLCO trong phần cửa sổ số liệu.



Khi giả thiết H_0 bị bác bỏ, ta có thể tiến hành so sánh cặp đôi giữa các công thức thí nghiệm của yếu tố thí nghiệm.

Trong hộp thoại General Linear Model, chọn **Comparisons...**

Khai báo yếu tố (TA) cần so sánh cặp trong **Terms:**

Chọn phương pháp so sánh (**Method**): **Tukey**, **Bonferroni** hoặc **Sidak**,

Chọn **OK** để có kết quả.

So sánh sự sai khác giữa các công thức thí nghiệm bằng cách tạo thành các nhóm với các chữ cái a, b, c, ...

Grouping Information Using Tukey Method and 95.0% Confidence

TA	N	Mean	Grouping
A	4	11.5	A
B	4	11.1	A
C	4	10.9	A
D	4	9.2	B

Means that do not share a letter are significantly different.

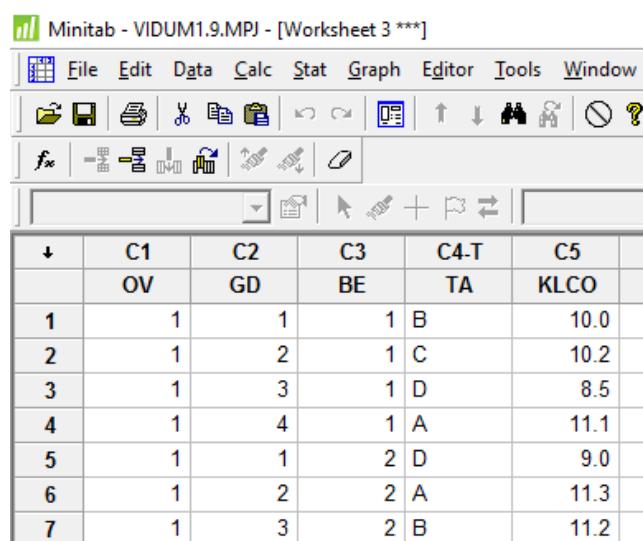
Nếu thí nghiệm được tiến hành trên nhiều ô vuông latinhh khác nhau việc phân tích số liệu sẽ bao gồm ảnh hưởng của 3 yếu tố trong một ô vuông (hàng, cột, yếu tố thí nghiệm) và ảnh hưởng giữa các ô.

Ví dụ M-1.9b: Giả sử, một thí nghiệm được thiết kế tương tự như ở ví dụ M-1.9a, nhưng có 2 ô vuông la tinh được thiết kế đồng thời và mỗi ô đều có 4 động vật thí nghiệm và 4 công thức thí nghiệm khác nhau. Số liệu ở ô vuông la tinh thứ nhất như trong ví dụ M-1.9a, ô vuông la tinh thứ 2 như trong bảng bên.

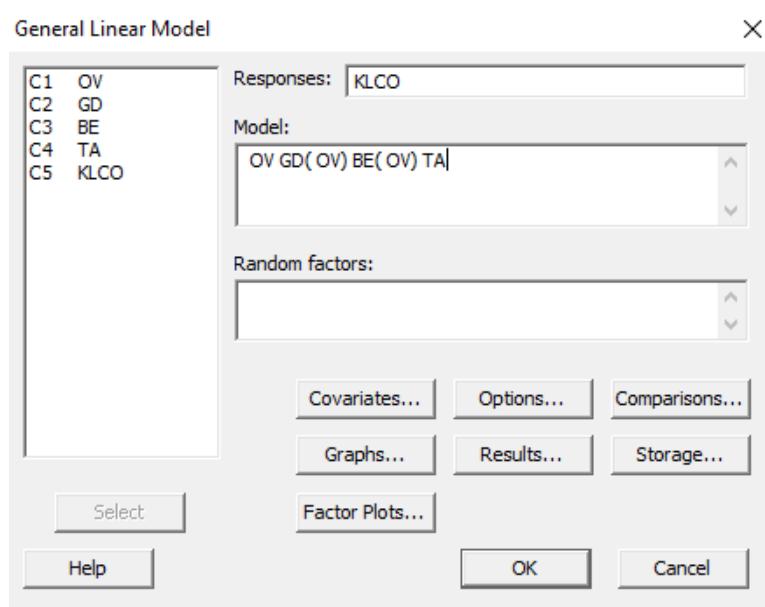
Hãy tiến hành phân tích để đưa ra kết luận và đưa ra nhận xét về mô hình thiết kế trong ví dụ M-1.9a và M-1.9b.

Giai đoạn	Bê			
	1	2	3	4
1	10,9 (C)	11,2 (A)	9,4 (D)	11,2 (B)
2	10,5 (B)	9,6 (D)	11,4 (C)	10,9 (A)
3	11,1 (A)	11,4 (C)	11,7 (B)	9,8 (D)
4	8,8 (D)	12,9 (B)	11,4 (A)	11,2 (C)

Cáu trúc số liệu:



Stat → ANOVA → General Linear Model...



Lưu ý: trong Model: đã khai báo dòng lệnh **OV GD(OV) BE(OV) TA**. Chi tiết về kiến thức của mô hình xem Chương 5 (tr.77) GTTKTN-2007.

Chọn OK để có kết quả.

General Linear Model: KLCO versus OV; TA; GD; BE

```

Factor      Type    Levels   Values
OV         fixed     2   1; 2
GD(OV)    fixed     8   1; 2; 3; 4; 1; 2; 3; 4
BE(OV)    fixed     8   1; 2; 3; 4; 1; 2; 3; 4
TA         fixed     4   A; B; C; D

Analysis of Variance for KLCO, using Adjusted SS for Tests
Source   DF   Seq SS   Adj SS   Adj MS      F       P
OV        1   0.1953   0.1953   0.1953   0.97   0.340
GD(OV)    6   2.1444   2.1444   0.3574   1.78   0.171
BE(OV)    6   5.4994   5.4994   0.9166   4.56   0.008
TA        3   22.6634  22.6634  7.5545  37.59   0.000
Error     15   3.0147   3.0147   0.2010
Total     31   33.5172

S = 0.448307   R-Sq = 91.01%   R-Sq(adj) = 81.41%
Unusual Observations for KLCO
Obs      KLCO      Fit   SE Fit  Residual   St Resid
11   12.8000  12.0344  0.3268   0.7656     2.49 R
24   12.9000  12.0781  0.3268   0.8219     2.68 R
R denotes an observation with a large standardized residual.

```

3.4. THÍ NGHIỆM HAI YẾU TỐ CHÉO NHAU (TRỰC GIAO)

Với mô hình thí nghiệm 2 yếu tố chéo nhau, ngoài nghiên cứu tác động của từng yếu tố thí nghiệm ta còn nghiên cứu mối tương tác giữa 2 yếu tố. Về mặt cấu trúc dữ liệu trong Minitab hoàn toàn tương tự như với thiết kế thí nghiệm theo khái **mục 5.2** của tài liệu nhưng khai báo câu lệnh có thêm phần tương tác.

Ví dụ M-1.10: Một nghiên cứu được tiến hành để xác định ảnh hưởng của việc bổ sung 2 loại vitamin (A và B) vào thức ăn đến tăng khối lượng (kg/ngày) của lợn. Hai mức đối với vitamin A (0 và 4 mg) và 2 mức đối với vitamin B (0 và 5 mg) được sử dụng trong thí nghiệm này. Tổng số 20 lợn thí nghiệm được phân về 4 tổ hợp (công thức thí nghiệm) một cách ngẫu nhiên. Số liệu thu được khi kết thúc thí nghiệm được trình bày như sau:

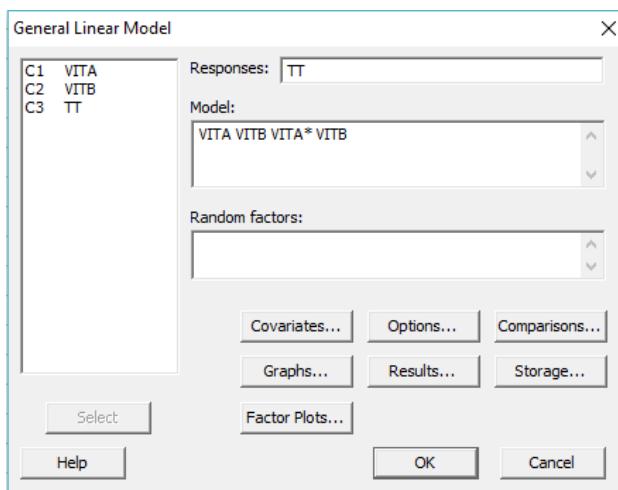
A	0 mg		4 mg	
	B	0 mg	5 mg	0 mg
	0,585	0,567	0,473	0,684
	0,536	0,545	0,450	0,702
	0,458	0,589	0,869	0,900
	0,486	0,536	0,473	0,698
	0,536	0,549	0,464	0,693

Cáu trúc số liệu:

The screenshot shows the Minitab software interface with the title bar "Minitab - VIDUM1.10.mpj - [Worksheet 1 ***]". The menu bar includes File, Edit, Data, Calc, Stat, Graph, Editor, Tools, and Window. Below the menu is a toolbar with various icons. The main area displays a data table with columns labeled C1, C2, C3, C4, and C5. The first row contains headers: C1, C2, C3, C4, and C5. Subsequent rows contain data: Row 1 (VITA, VITB, TT) has values 0, 0, 0.585; Row 2 has values 0, 0, 0.536; Row 3 has values 0, 0, 0.458; Row 4 has values 0, 0, 0.486; Row 5 has values 0, 0, 0.536; Row 6 has values 0, 5, 0.567; Row 7 has values 0, 5, 0.545; Row 8 has values 0, 5, 0.589; Row 9 has values 0, 5, 0.536; Row 10 has values 0, 5, 0.549.

	C1	C2	C3	C4	C5
	VITA	VITB	TT		
1	0	0	0.585		
2	0	0	0.536		
3	0	0	0.458		
4	0	0	0.486		
5	0	0	0.536		
6	0	5	0.567		
7	0	5	0.545		
8	0	5	0.589		
9	0	5	0.536		
10	0	5	0.549		

Stat → ANOVA → General Linear Model...



Chọn OK để có kết quả.

General Linear Model: TT versus VITA; VITB

Factor	Type	Levels	Values
VITA	fixed	2	0; 4
VITB	fixed	2	0; 5

Analysis of Variance for TT, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
VITA	1	0.05192	0.05192	0.05192	4.71	0.045
VITB	1	0.06418	0.06418	0.06418	5.82	0.028
VITA*VITB	1	0.02911	0.02911	0.02911	2.64	0.124
Error	16	0.17648	0.17648	0.01103		
Total	19	0.32169				

Các giá trị xác suất 0,045; 0,028 và 0,124 đều được xem xét để đưa ra quyết định với từng yếu tố (Vitamin A và Vitamin B) và tương tác giữa (Vitamin A × Vitamin B). Với xác suất 0,045 và $0,028 < 0,05$ có thể kết luận Vitamin A và B có ảnh hưởng đến tăng khói lượng của lợn nhưng không có tương tác ($P = 0,14 > 0,05$).

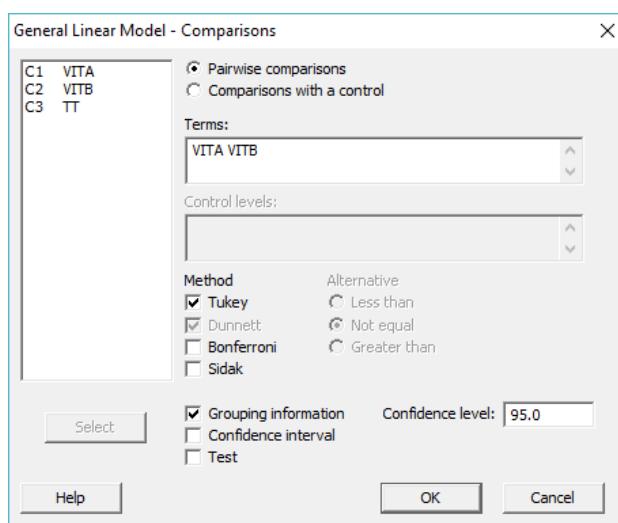
Khi giả thiết H_0 bị bác bỏ, ta có thể tiến hành so sánh cặp đôi giữa các công thức thí nghiệm của yếu tố thí nghiệm.

Trong hộp thoại General Linear Model, chọn **Comparisons...**

Khai báo yếu tố (**VITA**, **VITB**) cần so sánh cặp trong **Terms:**

Chọn phương pháp so sánh (**Method**): **Tukey**, **Bonferroni** hoặc **Sidak**,

Chọn **OK** để có kết quả.



So sánh sự sai khác giữa các công thức thí nghiệm bằng cách tạo thành các nhóm với các chữ cái a, b, c, ...

Grouping Information Using Tukey Method and 95.0% Confidence

	N	Mean	Grouping
4	10	0.6	A
0	10	0.5	B

Means that do not share a letter are significantly different.

Grouping Information Using Tukey Method and 95.0% Confidence

	N	Mean	Grouping
5	10	0.6	A
0	10	0.5	B

Means that do not share a letter are significantly different.

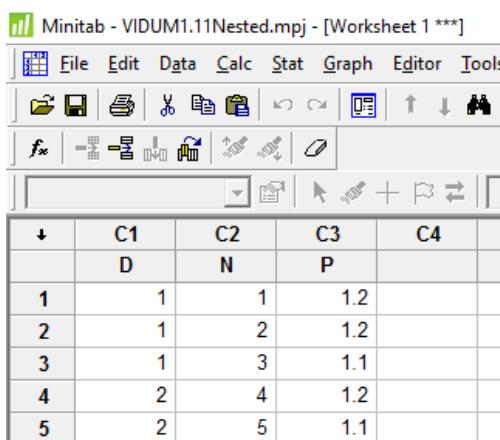
3.5. THÍ NGHIỆM HAI YẾU TỐ PHÂN CẤP (CHIA Ô)

Với mô hình phân cấp, yếu tố cấp trên (A) là cố định và cấp dưới (B) là ngẫu nhiên. Như vậy B sẽ làm ô (nested) trong A.

Ví dụ M-11: Mục đích của thí nghiệm là xác định ảnh hưởng của lợn đực giống và lợn nái đến khối lượng sơ sinh của thê hê con. Mô hình phân cấp 2 yếu tố được sử dụng. Bốn lợn đực giống được chọn ngẫu nhiên ($a = 4$), mỗi đực phối với 3 lợn nái ($b = 3$) và mỗi nái sinh được 2 lợn con ($r = 2$). Khối lượng (kg) sơ sinh của từng lợn con thu được như sau:

Đực	1			2			3			4		
	1	2	3	4	5	6	7	8	9	10	11	12
Nái	1,2	1,2	1,1	1,2	1,1	1,2	1,2	1,3	1,2	1,3	1,4	1,3
	1,2	1,3	1,2	1,2	1,2	1,1	1,2	1,3	1,2	1,3	1,4	1,3

Cấu trúc số liệu:

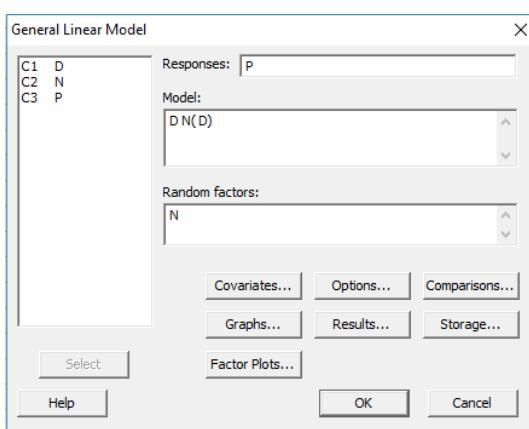


Lưu ý: Có thể chọn một trong hai cách khai báo trong Minitab có để kết quả phân tích phương sai.

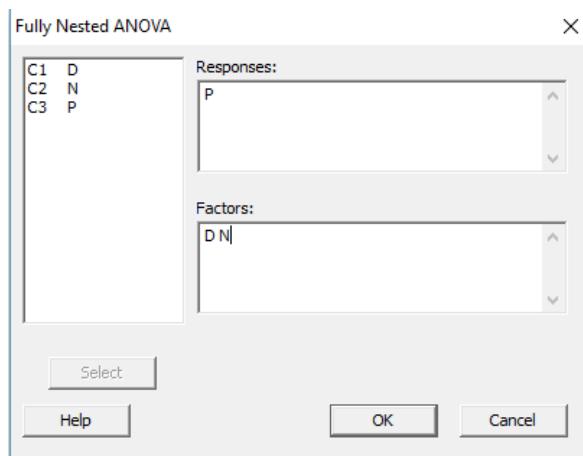
Với cách lựa chọn thứ nhất **Stat → ANOVA → General Linear Model...** Minitab chỉ hiển thị kết quả phân tích phương sai.

Với cách lựa chọn thứ hai **Stat → ANOVA → Fully Nested ANOVA...** ngoài kết quả phân tích phương sai Minitab còn cung cấp bảng các thành phần phương sai và ước tính các giá trị trung bình bình phương.

Stat → ANOVA → General Linear Model...



Hoặc Stat → ANOVA → Fully Nested ANOVA...



Trong ô **Model:** đã khai báo D N(D) thể hiện Nái (N) làm ô (nested) trong Đực (D) và ô Random factors: N thể hiện yếu tố Nái (N) là ngẫu nhiên.

Chọn **OK** để có kết quả:

General Linear Model: P versus D; N

Factor	Type	Levels	Values
D	fixed	4	1; 2; 3; 4
N(D)	random	12	1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12

Analysis of Variance for P, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
D	3	0.093333	0.093333	0.031111	6.22	0.017
N(D)	8	0.040000	0.040000	0.005000	3.00	0.042
Error	12	0.020000	0.020000	0.001667		
Total	23	0.153333				

S = 0.0408248 R-Sq = 86.96% R-Sq(adj) = 75.00%

Hoặc:

Nested ANOVA: P versus D; N

Analysis of Variance for P

Source	DF	SS	MS	F	P
D	3	0.0933	0.0311	6.222	0.017
N	8	0.0400	0.0050	3.000	0.042
Error	12	0.0200	0.0017		
Total	23	0.1533			

Variance Components

Source	Var Comp.	Total	StDev	% of
D	0.004	56.63	0.066	
N	0.002	21.69	0.041	

Error	0.002	21.69	0.041
Total	0.008		0.088
Expected Mean Squares			
1 D	1.00 (3)	+	2.00 (2)
2 N	1.00 (3)	+	2.00 (2)
3 Error	1.00 (3)		

Trong cả 2 trường hợp ta đều có giá trị xác suất đối với yếu tố **Đực** và yếu tố **Nái** tương ứng là 0,017 và 0,042, như vậy giả thiết H_0 bị bác bỏ. Kết luận có sự sai khác giữa các đực và giữa các cái trong cùng một đực ($P < 0,05$).

Cách tính các phuong sai thành phần ước tính dựa vào bảng **Expected Mean Squares** được triển khai cụ thể như sau:

$$\text{MSE}_{\text{Error}} = 0,0017 = 1,00(\sigma^2_{\varepsilon}) \rightarrow \sigma^2_{\varepsilon} = 0,0017/1,00 \approx 0,002;$$

$$\text{MSN} = 0,0050 = 1,00(\sigma^2_{\varepsilon}) + 2,00(\sigma^2_N) \rightarrow \sigma^2_N = [0,0050 - 1,00(0,0017)] / 2,00 \approx 0,002;$$

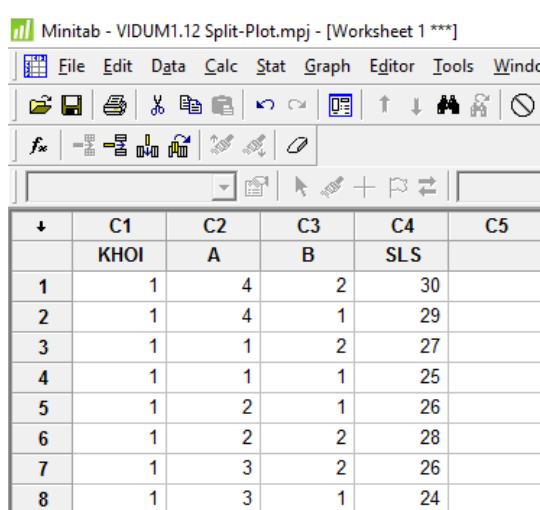
$$\text{MSD} = 0,0311 = 1,00(\sigma^2_{\varepsilon}) + 2,00(\sigma^2_N) + 6,00(\sigma^2_D) \rightarrow \sigma^2_D \approx 0,004.$$

3.6. THÍ NGHIỆM HAI YẾU TỐ CHIA Ô (Split-Plot)

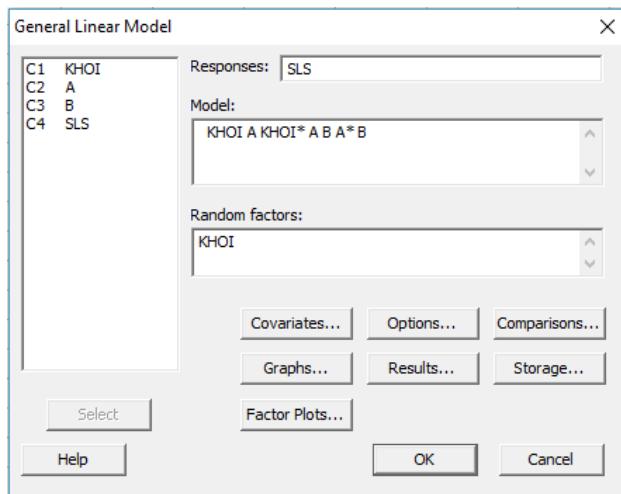
Ví dụ M-1.12a: Một thí nghiệm được tiến hành để nghiên cứu ảnh hưởng của bối chăn thả A (1, 2,3 và 4) và lượng khoáng bổ sung B (1 và 2) đến năng suất sữa. Có tất cả 24 bò tham gia thí nghiệm. Thí nghiệm được thiết kế theo mô hình hai yếu tố kiểu chia ô với yếu tố A được bố trí trên ô lớn và yếu tố B trên ô nhỏ trên 3 khối. Năng suất sữa trung bình được ghi lại như sau (kg /ngày):

Khối 1				Khối 2				Khối 3			
A ₄	A ₁	A ₂	A ₃	A ₂	A ₁	A ₄	A ₃	A ₁	A ₂	A ₄	A ₃
B ₂ 30	B ₂ 27	B ₁ 26	B ₂ 26	B ₁ 32	B ₂ 30	B ₁ 34	B ₁ 33	B ₂ 34	B ₁ 30	B ₂ 36	B ₁ 33
B ₁ 29	B ₁ 25	B ₂ 28	B ₁ 24	B ₂ 37	B ₁ 31	B ₂ 37	B ₂ 32	B ₁ 31	B ₂ 31	B ₁ 38	B ₂ 32

Cấu trúc số liệu:



Stat → ANOVA → General Linear Model...



Chọn OK để có kết quả

General Linear Model: SLS versus KHOI, A, B

Factor	Type	Levels	Values			
KHOI	random	3	1, 2, 3			
A	fixed	4	1, 2, 3, 4			
B	fixed	2	1, 2			
Analysis of Variance for SLS, using Adjusted SS for Tests						
Source	DF	Seq SS	Adj SS	Adj MS	F	P
KHOI	2	212.583	212.583	106.292	24.45	0.001
A	3	71.167	71.167	23.722	5.46	0.038
KHOI*A	6	26.083	26.083	4.347	1.93	0.191
B	1	8.167	8.167	8.167	3.63	0.093
A*B	3	5.833	5.833	1.944	0.86	0.498
Error	8	18.000	18.000	2.250		
Total	23	341.833				
S = 1.5 R-Sq = 94.73% R-Sq(adj) = 84.86%						

Lưu ý:

Trong ví dụ nêu trên khói được coi là yếu tố ngẫu nhiên vì vậy trong ô **Random factor** đã khai báo yếu tố **khói** (KHOI).

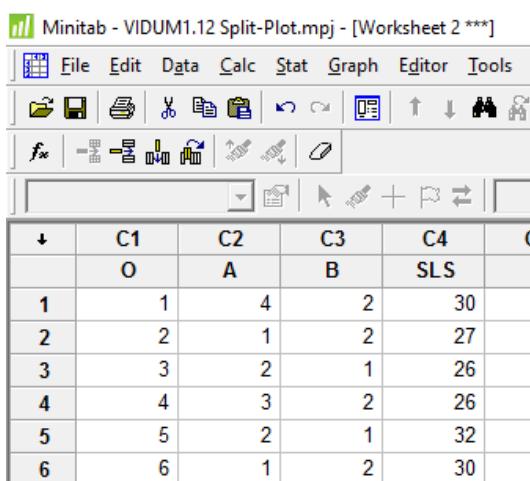
Tương tác giữa khói và yếu tố A (KHOI*A) chính là sai số ô lớn. Chính vì vậy giá trị F = 5,46 là thương của MS(A) / MS(KHOI*A) = 23,722/4,347.

Trong kết quả phân tích ta chỉ quan tâm đến xác suất đối với yếu tố A, B và tương tác A*B. Các giá trị này lần lượt là 0,038; 0,093 và 0,498. Với các giá trị này ta có thể kết luận năng suất sữa có sự khác nhau giữa các bối chẵn thả ($P < 0,05$), tuy nhiên việc bổ sung các khoáng chất không làm ảnh hưởng đến năng suất sữa và cũng không có ảnh hưởng tương tác giữa bối chẵn thả và việc bổ sung khoáng ($P > 0,05$).

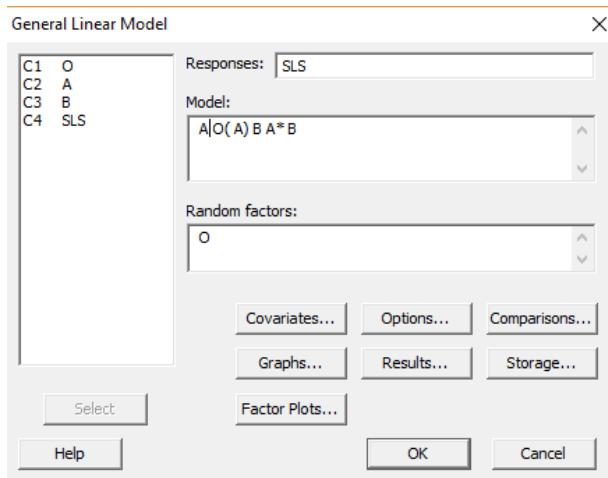
Ví dụ M-1.12b: Xem xét ví dụ M-1.12a, giả sử rằng thí nghiệm được thực hiện không có khói và chỉ có yếu tố A và B được thiết kế trên 12 ô lớn. Năng suất sữ trung bình được ghi lại như sau (kg/ngày):

1 A ₄	2 A ₁	3 A ₂	4 A ₃	5 A ₂	6 A ₁	7 A ₄	8 A ₃	9 A ₁	10 A ₂	11 A ₄	12 A ₃
B ₂ 30	B ₂ 27	B ₁ 26	B ₂ 26	B ₁ 32	B ₂ 30	B ₁ 34	B ₁ 33	B ₂ 34	B ₁ 30	B ₂ 36	B ₁ 33
B ₁ 29	B ₁ 25	B ₂ 28	B ₁ 24	B ₂ 37	B ₁ 31	B ₂ 37	B ₂ 32	B ₁ 31	B ₂ 31	B ₁ 38	B ₂ 32

Cấu trúc số liệu:



Stat → ANOVA → General Linear Model...



Chọn OK để có kết quả:

General Linear Model: SLS versus A, B, O

Factor	Type	Levels	Values
A	fixed	4	1, 2, 3, 4
O (A)	random	12	2, 6, 9, 3, 5, 10, 4, 8, 12, 1, 7, 11
B	fixed	2	1, 2

Analysis of Variance for SLS, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
A	3	71.167	71.167	23.722	0.80	0.530
O(A)	8	238.667	238.667	29.833	13.26	0.001
B	1	8.167	8.167	8.167	3.63	0.093
A*B	3	5.833	5.833	1.944	0.86	0.498
Error	8	18.000	18.000	2.250		
Total	23	341.833				

S = 1.5 R-Sq = 94.73% R-Sq(adj) = 84.86%

Trong ví dụ này, yếu tố ô (O) được coi là ngẫu nhiên và Ô nested trong yếu tố A. Sai số ô lớn chính là O(A). Chính vì vậy mà giá trị F của yếu tố A được tính $23,722/29,833 = 0,80$. Ba giá trị xác suất quan tâm đến bao gồm 0,530; 0,093 và 0,498 tương ứng với yếu tố A, B và tương tác A*B. Với cách thiết kế thí nghiệm theo mô hình thứ 2 (M-1.12b) ta đã không tìm thấy ảnh hưởng của bất kỳ một yếu tố nào ($P > 0,05$).

3.7. PHÂN TÍCH HIỆP PHƯƠNG SAI

Ví dụ M-1.13:

Một thí nghiệm được tiến hành nhằm nghiên cứu khối lượng (kg) kết thúc vỗ béo và tăng khối lượng (g/ngày) trong giai đoạn nuôi vỗ béo từ 21 đến 24 tháng tuổi của ba nhóm bò lai LaiSind, F1(Brahman × LaiSind) và F1(Charolais × LaiSind). Kết quả thu được như sau:

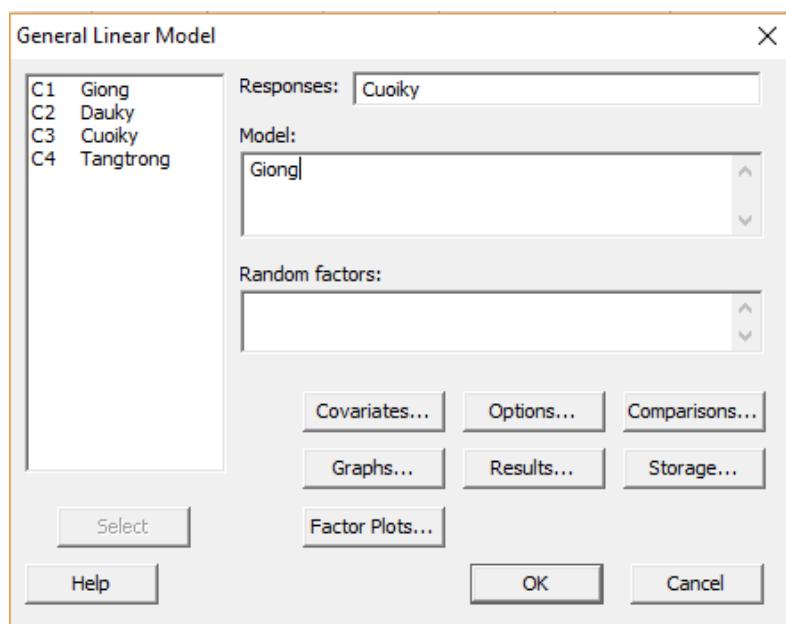
Bò số	Khối lượng đầu kỳ	Khối lượng cuối kỳ	Tăng khối lượng
LaiSind			
1	196	259	700,00
2	187	247	666,67
3	210	266	622,22
4	196	255	655,56
5	185	249	711,11
F1(Brahman × LaiSind)			
6	215	285	777,78
7	210	295	944,44
8	198	280	911,11
9	217	298	900,00
10	209	285	844,44
F1(Charolais × LaiSind)			
11	216	310	1044,44
12	235	321	955,56
13	230	320	1000,00
14	215	305	1000,00
15	210	295	944,44

Hãy cho biết ảnh hưởng của khối lượng đầu kỳ đến khối lượng cuối kỳ và tăng khối lượng. So sánh khối lượng cuối kỳ và tăng khối lượng giữa 3 nhóm bò lai trên. Phần minh họa dưới đây tiến hành phân tích ảnh hưởng của khối lượng đầu kỳ (hiệp phương sai) và nhóm bò đến khối lượng cuối kỳ.

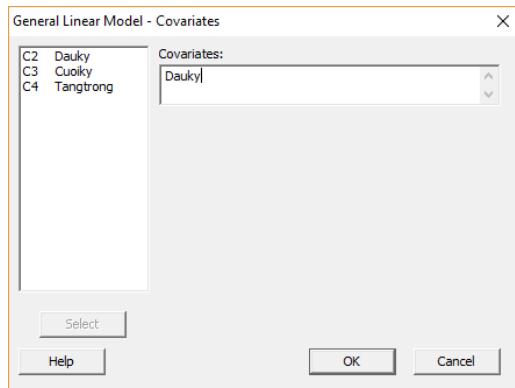
Cấu trúc số liệu:

	C1-T	C2	C3	C4
	Giong	Dauky	Cuoiky	Tangtrong
1	LS	196	259	700.00
2	LS	187	247	666.67
3	LS	210	266	622.22
4	LS	196	255	655.56
5	LS	185	249	711.11
6	BR	215	285	777.78

Stat → ANOVA → General Linear Model...



Vào Covariates để khai báo biến "Dau ky" với vai trò *hiệp phương sai*



Chọn **OK** để có kết quả.

General Linear Model: Cuoiky versus Giong

```

Factor   Type    Levels   Values
Giong    fixed      3   BR, CH, LS

Analysis of Variance for Cuoiky, using Adjusted SS for Tests

Source   DF   Seq SS   Adj SS   Adj MS      F       P
Dauky     1   7343.6    735.7    735.7   40.25   0.000
Giong     2   1070.7   1070.7    535.3   29.29   0.000
Error    11   201.1    201.1     18.3
Total    14   8615.3

S = 4.27535   R-Sq = 97.67%   R-Sq(adj) = 97.03%
Term       Coef   SE Coef      T       P
Constant  111.56   27.31   4.09   0.002
Dauky     0.8298   0.1308   6.34   0.000

Unusual Observations for Cuoiky

Obs   Cuoiky      Fit   SE Fit  Residual  St Resid
  6   285.000   292.915   2.029    -7.915    -2.10  R

R denotes an observation with a large standardized residual.

```

Với xác suất của "**Dauky**" và "**Giong**" (nhóm bò) lần lượt là 0,000 và 0,000 đều bé hơn 0,001 nên có thể kết luận: Khối lượng đầu kỳ và nhóm bò lai có ảnh hưởng rõ rệt đến khối lượng cuối kỳ ($P < 0,001$).

Khi H_0 bị bác bỏ (khối lượng cuối kỳ của 3 nhóm bò lai khác nhau), ta có thể tiến hành so sánh cặp giữa các công thức thí nghiệm tương tự như ví dụ M-1.9a mục 5.3 để có kết quả như sau:

```

Grouping Information Using Tukey Method and 95.0% Confidence

Giong   N   Mean   Grouping
CH      5   299.7   A
BR      5   287.6   B
LS      5   266.7   C

Means that do not share a letter are significantly different.

```

BÀI 4. TƯƠNG QUAN VÀ HỒI QUY TUYẾN TÍNH

Để tính hệ số tương quan và xây dựng phương trình hồi quy, số liệu luôn phải tạo thành từng cặp và được nhập vào từng cột đôi với từng chỉ tiêu.

4.1. HỆ SỐ TƯƠNG QUAN

Ví dụ M-1.14: Tiến hành cân khói lượng (P), đo đường kính lớn (D) và đường kính bé (d) của 22 quả trứng gà. Số liệu thu được trình bày ở bảng dưới đây.

P (g)	66,80	60,10	71,20	61,60	61,20	59,00	67,90	59,00	51,50	62,60	64,20
D (mm)	58,37	54,95	60,58	56,73	57,36	53,26	57,07	58,17	52,28	55,62	56,82
d (mm)	45,12	44,35	45,56	44,34	43,57	44,86	46,27	42,82	41,91	44,95	44,79
P (g)	71,20	54,20	54,50	69,10	55,90	66,00	68,00	62,00	56,70	67,00	53,80
D (mm)	61,15	54,24	54,99	60,99	54,41	58,19	59,93	56,80	55,66	58,49	52,44
d (mm)	46,00	42,58	42,32	44,85	42,62	45,69	45,50	44,20	42,41	45,56	43,38

Cấu trúc số liệu trong Worksheet

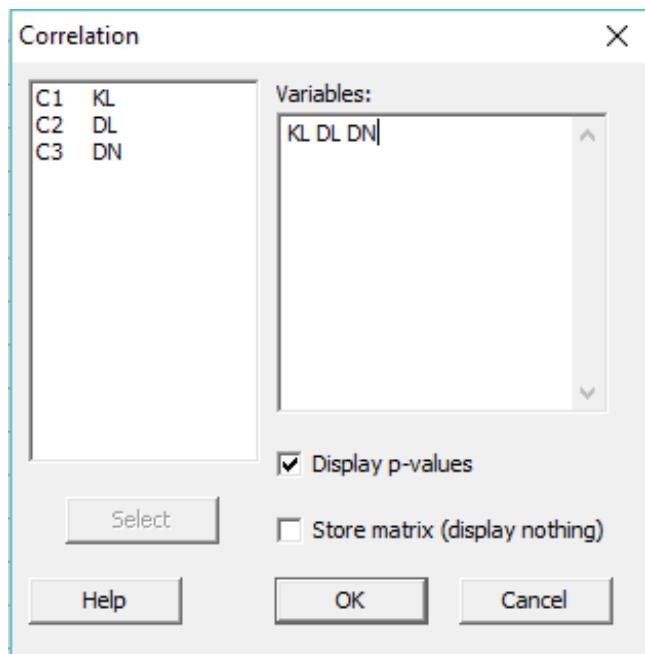
Lưu ý:

Để tính hệ số tương quan và xây dựng phương trình hồi quy, số liệu luôn phải tạo thành từng cặp và được nhập vào từng cột đôi với từng chỉ tiêu.

- 1) Cột Khối lượng **C1 (KL)**
- 2) Cột Đường kính lớn **C2 (DL)**
- 3) Cột Đường kính bé **C3 (DB)**

Giả thiết đối với kiểm định hai phía $H_0: \rho = 0$ và đối thiết $H_1: \rho \neq 0$, trong đó ρ là tương quan giữa 2 biến nghiên cứu.

Stat → Basic Statistics → Correlation...



Chọn **OK** để có kết quả:

Correlations: KL; DL; DN

	KL	DL
DL	0.897	
	0.000	
DN	0.905	0.648
	0.000	0.001

Cell Contents: Pearson correlation

P-Value

Chọn **Display p-values** để hiển thị xác suất đối với từng hệ số tương quan.

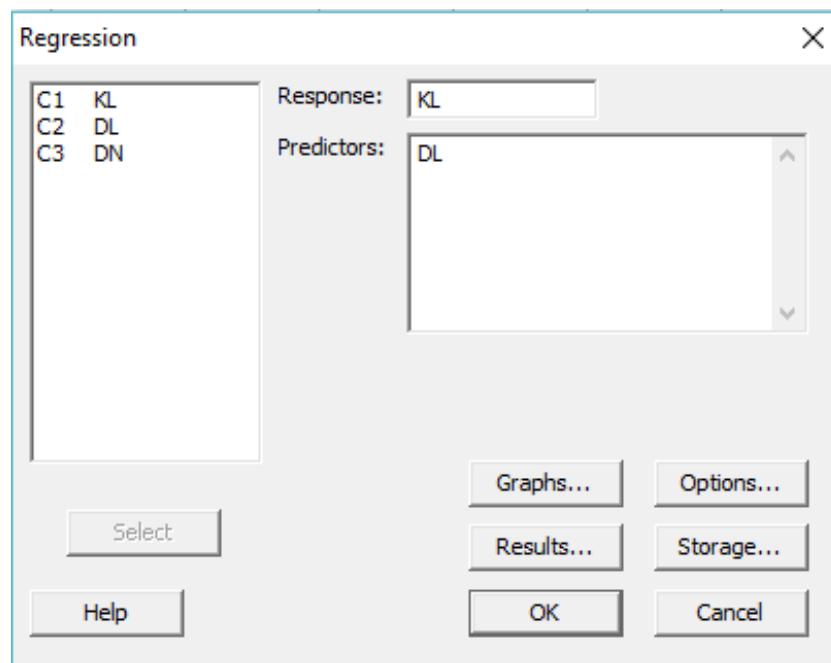
Chọn **Store matrix (display nothing)** để nhớ ma trận hệ số tương quan vào bộ nhớ đệm và không hiển thị kết quả ra màn hình.

Hệ số tương quan giữa **khối lượng** và **đường kính lớn** là 0,897; **khối lượng** và **đường kính bé** là 0,905; **đường kính lớn** và **đường kính bé** là 0,648. Xác suất đối với từng hệ số tương quan (**P-values**) đều bé hơn 0,05 (α) vì vậy kết luận mối quan hệ giữa các chỉ tiêu này khác 0.

4.2. PHƯƠNG TRÌNH HỒI QUY TUYẾN TÍNH

Có thể xây dựng hồi quy đơn biến $y = a + bx$ hoặc đa biến $y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n$. Có thể xây dựng phương trình hồi quy tuyến tính đơn biến quy ước tính khối lượng trung thông qua đường kính lớn/đường kính bé hoặc đa biến thông qua đường kính lớn và đường kính bé.

Stat → Regression → Regression...



Response: Khai báo cột C1 (KL) biến phụ thuộc.

Predictors: Khai báo cột C2 (DL) biến độc lập.

Chọn **OK** để có kết quả.

Regression Analysis: KL versus DL

The regression equation is

$$KL = -53.7 + 2.04 \text{ DL}$$

Predictor	Coef	SE Coef	T	P
-----------	------	---------	---	---

Constant	-53.67	12.78	-4.20	0.000
----------	--------	-------	-------	-------

DL	2.0379	0.2250	9.06	0.000
----	--------	--------	------	-------

S = 2.69651 R-Sq = 80.4% R-Sq(adj) = 79.4%

Analysis of Variance

Source	DF	SS	MS	F	P
--------	----	----	----	---	---

Regression	1	596.60	596.60	82.05	0.000
------------	---	--------	--------	-------	-------

Residual Error	20	145.42	7.27		
----------------	----	--------	------	--	--

Total	21	742.02			
-------	----	--------	--	--	--

Unusual Observations

Obs	DL	KL	Fit	SE Fit	Residual	St Resid
-----	----	----	-----	--------	----------	----------

7	57.1	67.900	62.629	0.579	5.271	2.00R
---	------	--------	--------	-------	-------	-------

8	58.2	59.000	64.871	0.658	-5.871	-2.25R
---	------	--------	--------	-------	--------	--------

R denotes an observation with a large standardized residual.

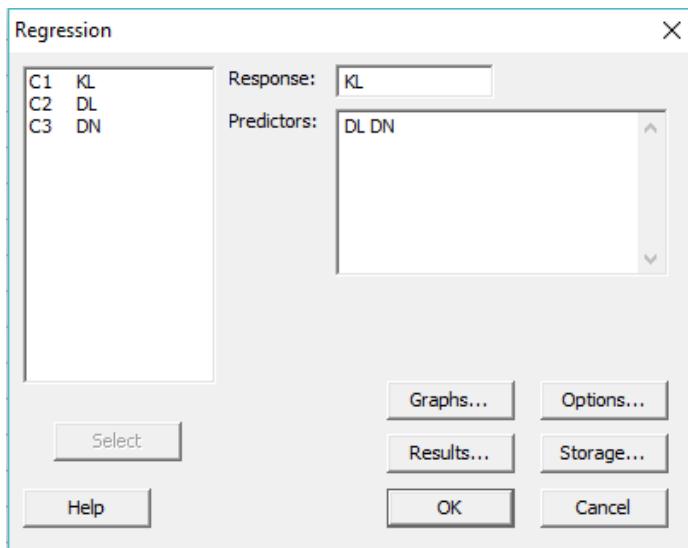
Phương trình hồi quy ước tính khối lượng (Y) thông qua đường kính lớn (X)
Y = -53,7 + 2,04X.

Bảng thứ nhát trong phần kết quả kiểm định các hệ số của phương trình hồi quy. Với xác suất $P = 0,000$ ta có thể kết luận các hệ số trong phương trình hồi quy khác 0 ($P < 0,05$).

Hệ số xác định của phương trình $R^2 = 80,4\%$, hiệu chỉnh $R^2 = 79,4\%$.

Các quan sát ngoại lai (**Unusual Observations**) trong mô hình và ví dụ nêu trên. Các giá trị ở hàng thứ 7 và 8 trong ví dụ trên được coi là ngoại lai.

Stat → Regression → Regression...



Predictors: Khai báo cột C2 (DL) và C3 (DN) biến độc lập.

Để xây dựng phương trình hồi quy đa biến, biến độc lập bao gồm từ 2 biến.

Chọn **OK** để có kết quả.

Regression Analysis: KL versus DL; DN

The regression equation is

$$KL = -117 + 1.21 \cdot DL + 2.48 \cdot DN$$

Predictor	Coef	SE Coef	T	P
Constant	-116.555	5.472	-21.30	0.000
DL	1.21473	0.08323	14.60	0.000
DN	2.4764	0.1623	15.26	0.000

$$S = 0.759757 \quad R-Sq = 98.5\% \quad R-Sq(\text{adj}) = 98.4\%$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	731.05	365.53	633.24	0.000
Residual Error	19	10.97	0.58		
Total	21	742.02			
Source	DF	Seq SS			
DL	1	596.60			

Ta có kết quả hoàn toàn tương tự như việc xây dựng phương trình hồi quy đơn giản.

$$Y = -117 + 1.21X_1 + 2.48X_2$$

Trong đó:

Y = khối lượng trứng;

X_1 = đường kính lớn;

X_2 = đường kính bé.

Điều khác biệt trong trường hợp này là hệ số xác định $R^2 = 98,5\%$ lớn hơn nhiều so với trường hợp đơn biến $R^2 = 80,4\%$.

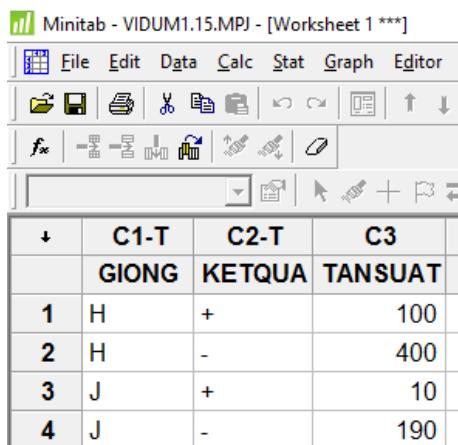
Bài 5. BẢNG TƯƠNG LIÊN

Khi so sánh các tỷ lệ hoặc nghiên cứu mối liên hệ giữa các yếu tố đối với biến định tính ta luôn đặt giả thiết H_0 : Không có sự sai khác có ý nghĩa thống kê giữa các tỷ lệ hoặc Không có mối liên hệ giữa các yếu tố (tuỳ theo mục tiêu của bài toán đặt ra).

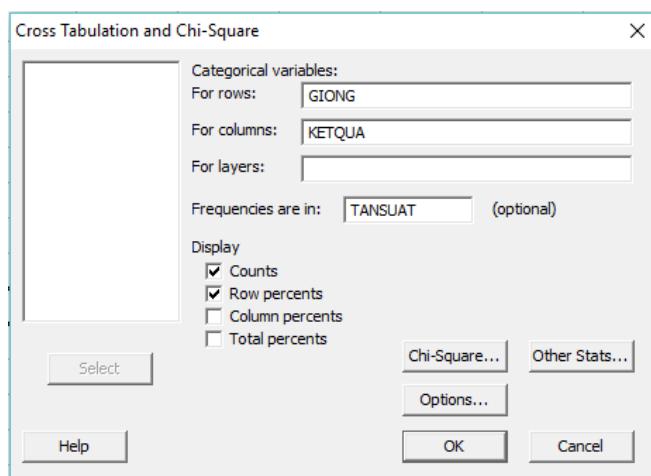
Ví dụ M-1.15: Một thí nghiệm được tiến hành nhằm đánh giá sự liên hệ giữa tỷ lệ viêm nội mạc tử cung và giống. Trong tổng số 700 bò sữa trong nghiên cứu thuần tập (cohort studies), có 500 con giống Holstein Friesian và 200 con giống Jersey. Kết quả nghiên cứu thu được như sau:

Giống	Viêm nội mạc tử cung		Tổng số
	Có	Không	
Holstein	100	400	500
Jersey	10	190	200
Tổng số	110	590	700

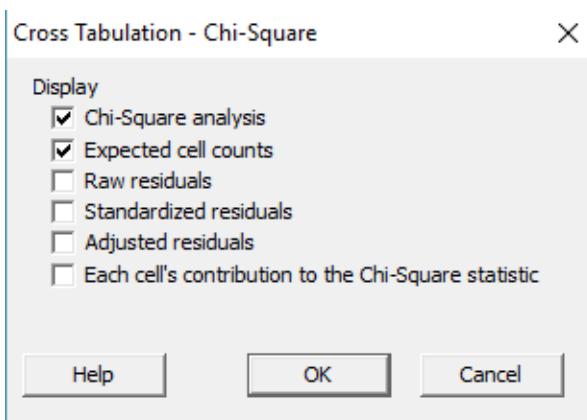
Cấu trúc và nhập số liệu vào Minitab như sau:



Stat → Tables → Cross Tabulation and Chi-Square...



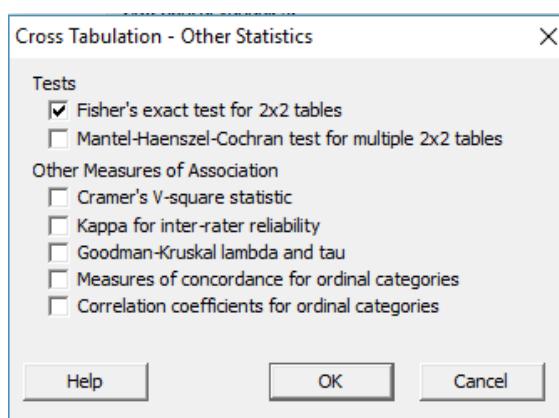
Chọn **Chi-Square...** (χ^2) nếu tần suất ước tính ≥ 5



Chi-Square Analysis – Phân tích χ^2

Expected cell count - Tần suất ước tính lý thuyết

Hoặc **Other Stats...** nếu tần suất ước tính trong một ô bất kỳ < 5



Fisher's exact test for 2x2 table – phép thử chính xác của Fisher đối với bảng 2x2

Khi sử dụng **Chi-Square...**, nếu giá trị trong một ô ước tính bé hơn 5, Minitab sẽ có dòng lệnh cảnh báo:

* NOTE * 1 cells with expected counts less than 5

Trong trường hợp này ta cần sử dụng **Other Stats...**. Với lựa chọn này ta cũng có phần Output tương tự như trên.

Chọn **OK** để có kết quả

Tabulated statistics: GIONG; KETQUA

Using frequencies in TANSUAT
Rows: GIONG Columns: KETQUA
- + All
H 400 100 500
80.00 20.00 100.00

	421.4	78.6	500.0
J	190	10	200
	95.00	5.00	100.00
	168.6	31.4	200.0
All	590	110	700
	84.29	15.71	100.00
	590.0	110.0	700.0

Cell Contents: Count
% of Row
Expected count

Pearson Chi-Square = 24.268; DF = 1; P-Value = 0.000
Likelihood Ratio Chi-Square = 29.054; DF = 1; P-Value = 0.000

Trong phần kết quả đối với từng ô ta có 3 giá trị. Ví dụ đối với ô thứ nhất lần lượt là: 1) Tần suất quan sát (400), 2) Phần trăm theo hàng (80%) và 3) tần suất ước tính (421,4).

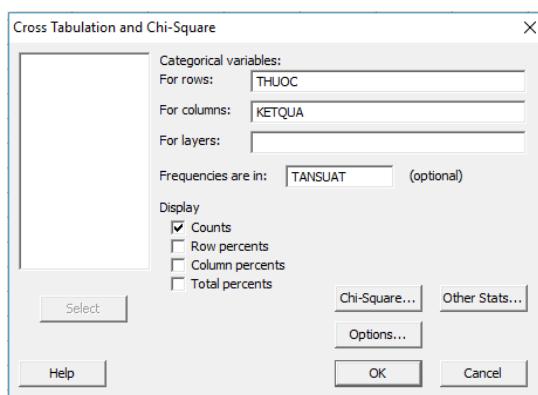
Giá trị Chi-Square $\chi^2 = 24,268$, bậc tự do DF = 1 và xác suất P-value = 0,000. Với xác suất này giả thiết H_0 bị bác bỏ và kết luận *Có mối liên hệ giữa bệnh viêm nội mạc tử và giống bò (P< 0,001)*.

Ví dụ M-1.16: Từ một đòn trước khi cho tiêm xúc với nguồn bệnh, chọn ra 10 động vật thí nghiệm (tiêm vắc xin) và 10 động vật đối chứng (không tiêm vắc xin). Số động vật này sau khi cho tiêm xúc với nguồn bệnh ta thu được kết quả như trong bảng sau. Liệu vắc xin có làm giảm tỷ lệ chết hay không?

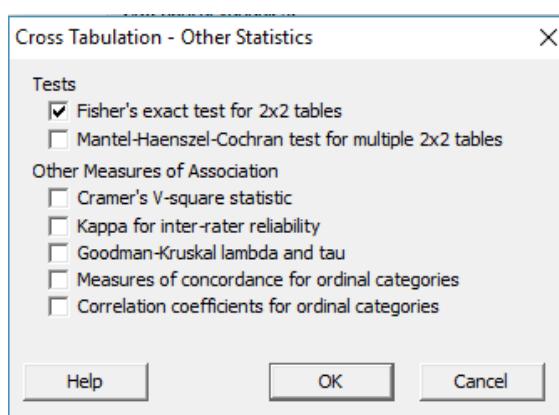
Thuốc	Kết quả		Tổng hàng
	Sống	Chết	
Vắc xin	9	1	10
Đối chứng	2	8	10
Tổng cộng	11	9	20

Cấu trúc và nhập số liệu vào Minitab như sau:

Stat → Tables → Cross Tabulation and Chi-Square...



Chọn **Other Stats...** khi tần suất ước tính trong một ô bất kỳ < 5



Fisher's exact test for 2×2 table – Phép thử chính xác của Fisher đối với bảng 2×2.

Chọn **OK** để có kết quả.

Tabulated statistics: THUOC, KETQUA

Using frequencies in TANSUAT

Rows: THUOC Columns: KETQUA

	-	+	All
DC	8	2	10
	80	20	100
VAC	1	9	10
	10	90	100
All	9	11	20
	45	55	100

Cell Contents: Count

% of Row

Fisher's exact test: P-Value = 0.0054775

Xác suất P-value = 0,0054775. Với xác suất này giả thiết H_0 bị bác bỏ và kết luận *Có mối liên hệ giữa tiêm hay không tiêm vắc xin với tỷ lệ ché* ($P < 0,01$).

Bài 6. ƯỚNG TÍNH DUNG LƯỢNG MẪU

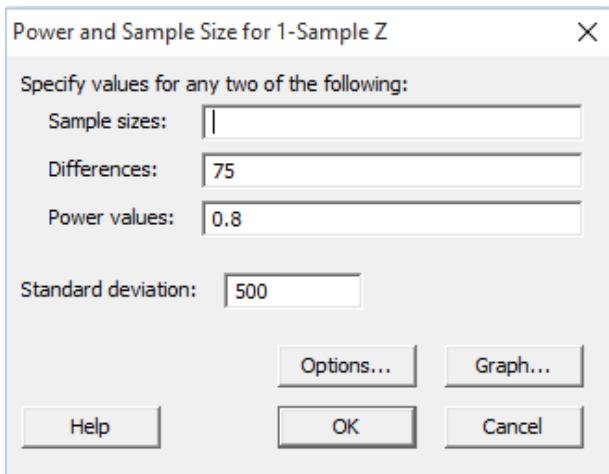
Các ví dụ và minh họa ở phần này giúp bạn đọc xác định dung lượng mẫu cần thiết bằng phần mềm Minitab khi muốn thiết kế một thí nghiệm. Một số thuật ngữ trong Minitab (phần xác định dung lượng mẫu cần thiết):

Thuật ngữ trong Minitab	Thuật ngữ tiếng Việt
Sample sizes	Dung lượng mẫu
Differences	Sự chênh lệch dự kiến
Power Values	Độ mạnh của phép thử
Standard deviation	Độ lệch chuẩn
Significance level	Mức ý nghĩa
Hypothesized p	Giá xác suất cần kiểm định
Alternative value of p	Giá trị xác suất lựa chọn
Proportion 1 values	Xác suất của sự kiện ở quần thể 1
Proportion 2	Xác suất của sự kiện ở quần thể 2

6.1. ƯỚC LUỢNG, KIỂM ĐỊNH MỘT GIÁ TRỊ TRUNG BÌNH

Ví dụ M-1.17a: Cần quan sát bao nhiêu bò sữa để ước tính được năng suất trong chu kỳ tiết sữa 305 ngày nằm trong khoảng $\pm 75\text{kg}$ so với giá trị thực của quần thể. Biết rằng sản lượng sữa có phân phối chuẩn $\sigma = 500\text{kg}$, độ mạnh của phép thử 0,8 mức ý nghĩa 0,05.

Stat → Power and Sample Size → 1-Sample Z...



- Để tính dung lượng mẫu, ô Sample sizes cần để trống.
- Differences: 75 - sự chênh lệch mong đợi so với quần thể cho trước.
- Power values: 0,8 - Độ mạnh của phép thử.
- Standard deviation: 500 - Độ lệch chuẩn của quần thể cho trước.
- Theo mặc định mức ý nghĩa là 0,05, tuy nhiên bạn đọc có thể thay đổi giá trị bằng cách vào Options... trong phần hộp thoại.

Chọn **OK** để có kết quả:

Power and Sample Size

1-Sample Z Test

Testing mean = null (versus not = null)

Calculating power for mean = null + difference

Alpha = 0.05 Assumed standard deviation = 500

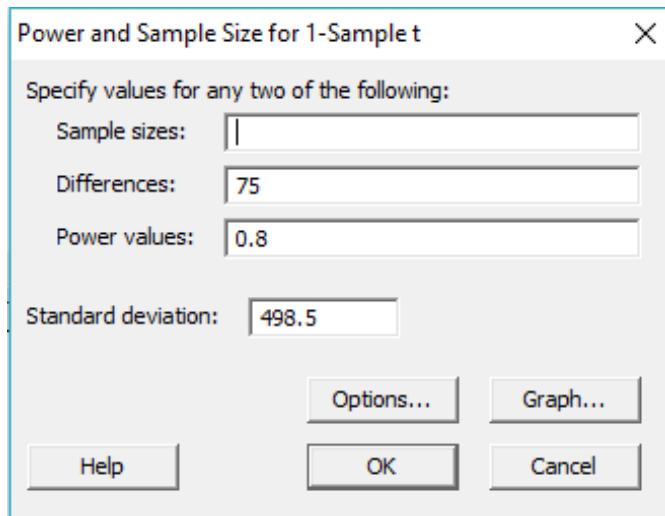
	Sample	Target	
Difference	Size	Power	Actual Power
75	349	0.8	0.800182

Như vậy, để thoả mãn điều kiện bài toán đặt ra cần quan sát ít nhất 349 bò.

Có thể thay thế độ lệch chuẩn của quần thể (σ) nếu không có bằng độ lệch chuẩn của mẫu (SD). Trong trường hợp cần chọn menu **Stat→Power and Sample Size→1-Sample t...** trong Minitab.

Ví dụ M-1.17b: Cần quan sát bao nhiêu bò sữa để ước tính được năng suất trong chu kỳ tiết sữa 305 ngày nằm trong khoảng $\pm 75\text{kg}$ so với giá trị thực của quần thể. Biết rằng sản lượng sữa có phân phối chuẩn $SD = 498,5 \text{ kg}$, độ mạnh của phép thử 0,8 mức ý nghĩa 0,05.

Stat→Power and Sample Size→1-Sample t...



- Để tính dung lượng mẫu, ô Sample sizes cần để trống;
- Differences: 75 - sự chênh lệch mong đợi so với quần thể cho trước;
- Power values: 0,8 - Độ mạnh của phép thử;
- Standard deviation: 498,5 - Độ lệch chuẩn (SD);
- Theo mặc định mức ý nghĩa là 0,05, tuy nhiên bạn đọc có thể thay đổi giá trị bằng cách vào Options... trong phần hộp thoại.

Chọn **OK** để có kết quả:

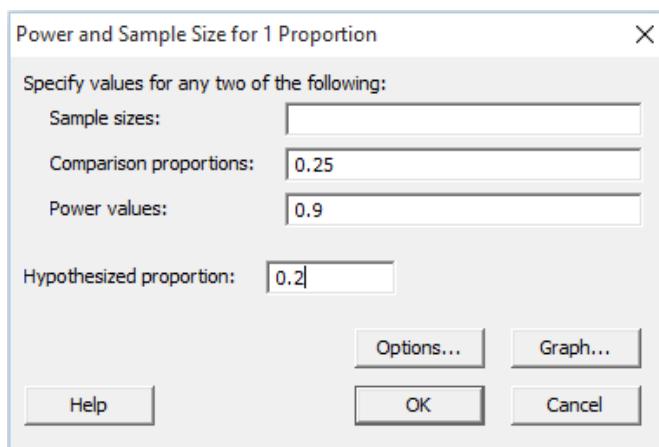
Power and Sample Size

```
1-Sample t Test  
Testing mean = null (versus not = null)  
Calculating power for mean = null + difference  
Alpha = 0.05 Assumed standard deviation = 498.5  
Sample Target  
Difference Size Power Actual Power  
75      349     0.8      0.800368
```

6.2. UỐC LUỢNG, KIỂM ĐỊNH MỘT TỶ LỆ

Ví dụ M-1.18: Cần dung lượng mẫu bao nhiêu để xác định tỷ lệ hiện nhiễm một loại vi khuẩn trên thân thịt lợn ở một lò mổ với ước tính chênh lệch không quá 5%. Biết rằng tỷ lệ hiện hành $p = 0,2$ và kiểm định ở mức ý nghĩa 0,05 và độ mạnh của phép thử 0,9.

Stat → Power and Sample Size → 1-Proportion...



- Để tính dung lượng mẫu, ô Sample sizes cần để trống
- Alternative values of p : $0,25 = 0,20 + 0,05$ chính bằng tỷ lệ hiện hành + sự chênh lệch mong đợi so với quần thể cho trước
 - Power values: 0,9 - Độ mạnh của phép thử
 - Hypothesized p: 0,2 - Tỷ lệ hiện hành
 - Theo mặc định mức ý nghĩa là 0,05, tuy nhiên bạn đọc có thể thay đổi giá trị bằng cách vào Options... trong phần hộp thoại

Chọn OK để có kết quả:

Power and Sample Size

```
Test for One Proportion  
Testing proportion = 0.2 (versus not = 0.2)  
Alpha = 0.05
```

Alternative	Sample	Target	
Proportion	Size	Power	Actual Power
0.25	718	0.9	0.900350

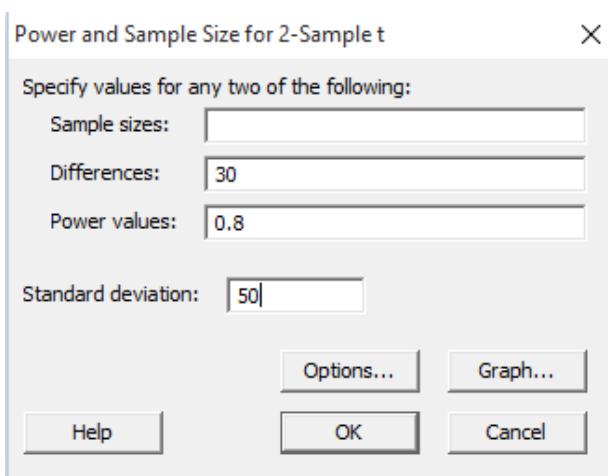
Kết luận: Cần xét nghiệm ít nhất 718 thân thịt để thoả mãn điều kiện của bài toán.

6.3. SO SÁNH HAI GIÁ TRỊ TRUNG BÌNH

Ví dụ M-1.19: Muốn thiết kế một thí nghiệm để so sánh sản lượng sữa của dê Bách Thảo ở 2 công thức thí nghiệm với yêu cầu $\alpha = 0,05$; $\beta = 0,2$; chênh lệch mong đợi 30 kg sữa biết $\sigma = 50$ kg.

Trong ví dụ này chỉ đề cập đến trường hợp 2 phương sai bằng nhau và chọn mẫu độc lập, các trường hợp khác không trình bày trong nội dung phần tài liệu này;

Stat→Power and Sample Size→2-Sample t...



- Để tính dung lượng mẫu, ô Sample sizes cần để trống
- Differences: 30 - sự chênh lệch mong đợi giữa 2 quần thể cho trước
- Power values: 0,8 = 1-0,2 Độ mạnh của phép thử
- Standard deviation: 50 - Độ lệch chuẩn của 2 quần thể cho trước (giả sử rằng độ lệch chuẩn bằng nhau)
 - Theo mặc định mức ý nghĩa là 0,05, tuy nhiên bạn đọc có thể thay đổi giá trị bằng cách vào **Options...** trong phần hộp thoại

Chọn **OK** để có kết quả:

Power and Sample Size

```
2-Sample t Test
Testing mean 1 = mean 2 (versus not =)
Calculating power for mean 1 = mean 2 + difference
Alpha = 0.05 Assumed standard deviation = 50
          Sample Target
```

Difference	Size	Power	Actual Power
30	45	0.8	0.803697

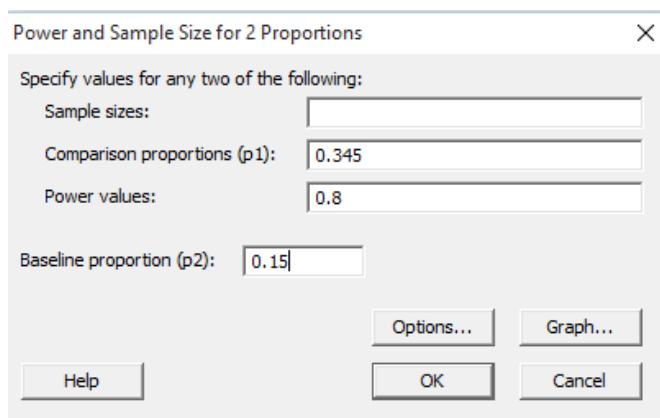
The sample size is for each group.

Kết luận: Cần ít nhất 45 để cho **mối công thức**, hay trong thí nghiệm này cần ít nhất 90 động vật thí nghiệm để thỏa mãn điều kiện bài toán

6.4. SO SÁNH HAI TỶ LỆ

Ví dụ M-1.20: Một thí nghiệm được tiến hành nhằm nghiên cứu tỷ lệ tồn thương num vú ở bò sữa giữa hệ thống vắt sữa A và B. Thời gian nghiên cứu được tiến hành trong 12 tháng với dự đoán tỷ lệ tồn thương ở hệ thống B là 34,5% và hệ thống A là 15%. Hãy tính dung lượng mẫu cần thiết đối với một nhóm để thỏa mãn điều kiện bài toán.

Stat → Power and Sample Size → 2-Proportions...



- Để tính dung lượng mẫu, ô Sample sizes cần để trống
- Proportion 1 values: 0,345 - Tỷ lệ mắc bệnh dự đoán ở quần thể 1
- Power values: 0,8 - Độ mạnh của phép thử
- Proportion 2: 0,15 - Tỷ lệ mắc bệnh dự đoán ở quần thể 2
- Theo mặc định mức ý nghĩa là 0,05, tuy nhiên bạn đọc có thể thay đổi giá trị bằng cách vào Options... trong phần hộp thoại

Chọn OK để có kết quả:

Power and Sample Size

```

Test for Two Proportions
Testing proportion 1 = proportion 2 (versus not =)
Calculating power for proportion 2 = 0.15
Alpha = 0.05
      Sample   Target
      Proportion 1    Size    Power   Actual Power
                  0.345     76       0.8       0.801595
  
```

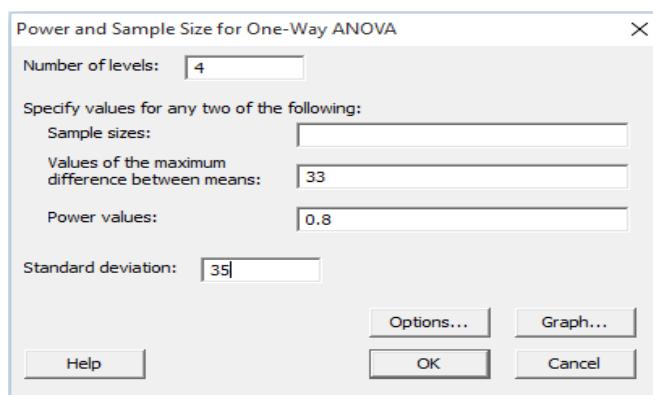
The sample size is for each group.

Kết luận: Cần ít nhất 76 bò sữa cho **mỗi công thức**, hay trong thí nghiệm này cần ít nhất 152 động vật thí nghiệm để thoả mãn điều kiện bài toán.

6.5. SO SÁNH NHIỀU GIÁ TRỊ TRUNG BÌNH

Ví dụ M-1.21: Thiết kế một thí nghiệm để so sánh tăng khối lượng (g) của gà ở 4 khẩu phần thức ăn (A, B, C, D). Các giá trị trung bình được chọn lần lượt là: $\mu_A = 79$, $\mu_B = 71$, $\mu_C = 80$, $\mu_D = 102$, với $\alpha = 0,05$ và $1 - \beta = 0,8$; biết $\sigma^2 = 35^2$. Cần bao nhiêu gà tham gia thí nghiệm này?

Stat → Power and Sample Size → One – Way ANOVA



- Để tính dung lượng mẫu, ô Sample sizes cần để trống
- Number of levels: 4 – số công thức thí nghiệm
- Power values: 0,8 - Độ mạnh của phép thử
- Values of the maximum difference between menas: 33
- Standard deviation: 35 – độ lệch chuẩn
- Theo mặc định mức ý nghĩa là 0,05, tuy nhiên bạn đọc có thể thay đổi giá trị bằng cách vào Options... trong phần hộp thoại

Chọn OK để có kết quả:

Power and Sample Size

One-way ANOVA

Alpha = 0.05 Assumed standard deviation = 35

Factors: 1 Number of levels: 4

Maximum Difference	Sample Size	Power	Actual Power
33	26	0.8	0.808303

The sample size is for each level.

Kết luận: Cần ít nhất 26 con gà cho **mỗi công thức**, hay trong thí nghiệm này cần ít nhất 104 động vật thí nghiệm để thoả mãn điều kiện bài toán.

BÀI 7. BÀI TẬP

7.1. TÓM TẮT VÀ TRÌNH BÀY DỮ LIỆU

Bài tập 7.1.1

Tóm tắt và trình bày đối với số liệu về chiều dài của 30 con cá (cm)

14	21	27	38	22	26
25	27	34	40	49	26
17	16	37	25	30	34
20	11	35	32	24	19
35	19	40	31	29	20

Bài tập 7.1.2a

Khối lượng trung bình (g) của cá hồi bảy sắc trước và sau 30 ngày nuôi thí nghiệm ở 3 công thức thí nghiệm (LP = 25%, OP = 45% và HP = 65%) và 6 mức lysin (1 = 1%, 2 = 1,2%, 3 = 1,4%, 4 = 1,6%, 5 = 1,8% và 6 = 2%) được trình bày như sau:

Lysin	Trước thí nghiệm			Sau thí nghiệm		
	Protein			Protein		
	LP	OP	HP	LP	OP	HP
1	0,91	0,91	0,92	1,66	2,42	1,93
2	0,95	0,91	0,90	2,22	2,89	2,35
3	0,93	0,92	0,92	2,51	3,60	2,84
4	0,93	0,92	0,93	3,06	4,37	3,22
5	0,91	0,91	0,91	3,49	4,89	3,76
6	0,94	0,92	0,92	3,37	4,71	4,15

Tóm tắt và trình bày đối với số liệu trên.

Bài tập 7.1.2b

Tương tự như ví dụ 7.2a, ta có lượng thức ăn thu nhận (g/cá) và protein thu nhận (g/cá) ở các mức protein và lysin tương ứng như sau:

Lysin	Thức ăn thu nhận			Protein thu nhận		
	Protein			Protein		
	LP	OP	HP	LP	OP	HP
1	1,20	1,24	1,01	0,30	0,56	0,65
2	1,51	1,60	1,14	0,38	0,72	0,74
3	1,80	1,98	1,44	0,45	0,89	0,93
4	2,07	2,25	1,73	0,52	1,01	1,12
5	2,19	2,53	1,95	0,55	1,14	1,27
6	2,29	2,55	2,08	0,57	1,15	1,35

Tóm tắt và trình bày đối với số liệu trên.

Bài tập 7.1.3

Tóm tắt và trình bày đối với số liệu về khối lượng (g) của 160 quả trứng gà cân được tại trại Quang Trung, ĐH Nông nghiệp I Hà Nội (nay là Học viện Nông nghiệp Việt Nam)

54,9	54,0	55,8	50,4	55,3	50,3	53,1	50,9	50,9	53,8
54,5	52,2	54,3	55,5	51,8	53,6	52,5	48,5	52,8	55,0
52,3	52,0	52,0	53,1	55,8	53,4	51,2	49,5	52,6	54,7
56,4	56,1	55,4	53,5	44,7	64,4	55,4	54,8	55,5	58,7
65,6	59,9	65,5	48,0	65,5	55,0	55,0	55,0	62,2	61,6
46,1	50,0	53,5	53,0	61,5	62,0	61,1	58,6	59,7	52,6
50,6	54,2	63,1	53,6	61,0	58,2	53,9	50,6	55,5	57,5
65,2	61,0	61,6	63,0	58,0	58,6	58,4	58,7	65,2	61,8
60,7	63,7	62,2	63,4	64,1	63,7	73,4	62,7	61,5	59,9
58,2	54,2	53,8	49,4	60,3	64,6	61,5	59,0	70,4	61,8
64,2	59,8	56,2	62,9	56,5	37,9	43,3	39,4	41,3	41,3
41,6	43,8	39,4	42,3	40,8	40,0	41,3	37,9	45,8	41,4
40,6	40,4	45,4	38,4	37,5	42,0	38,6	37,8	40,3	41,3
38,5	43,3	42,6	38,2	43,7	41,6	38,8	39,0	39,4	51,7
49,7	51,7	50,7	47,6	54,8	52,9	52,9	54,0	41,6	50,3
52,1	47,9	49,1	47,0	49,8	51,9	48,6	48,6	60,0	52,9

Nguồn: Đề tài nhóm sinh viên nghiên cứu khoa học năm học 2002 - 2003

Bài tập 7.1.4

Tóm tắt và trình bày đối với số liệu P - khối lượng (g), D- đường kính lớn (cm) và d - đường kính bé (cm) trứng của giống gà Coqard (Bỉ)

P	D	d	P	D	d	P	D	d	P	D	d
57,70	55,23	43,76	62,90	56,68	44,79	60,90	56,42	43,73	66,80	58,37	45,12
55,20	55,39	42,33	66,50	57,57	42,27	60,30	56,93	43,63	60,10	54,95	44,35
62,10	54,43	42,17	56,10	58,15	45,10	62,20	55,20	44,97	71,20	60,58	45,56
53,90	58,04	44,39	60,00	55,76	43,64	55,30	53,67	42,85	61,60	56,73	44,34
63,00	55,69	44,84	60,50	57,71	43,37	63,60	57,19	44,30	61,20	57,36	43,57
69,30	56,64	46,66	63,30	58,82	43,80	63,60	57,05	44,98	59,00	53,26	44,86
64,80	55,40	45,64	57,90	60,51	41,79	65,90	56,94	45,96	67,90	57,07	46,27
60,10	57,25	43,09	62,30	56,67	44,72	56,30	54,65	43,03	59,00	58,17	42,82
51,90	52,88	41,91	57,90	54,63	44,04	58,80	57,06	45,90	51,50	52,28	41,91
56,80	54,64	42,70	55,50	53,68	43,26	59,30	55,46	43,81	62,60	55,62	44,95
59,10	54,24	44,24	67,40	56,65	45,98	53,40	52,71	42,45	64,20	56,82	44,79
58,60	53,83	44,35	60,90	56,68	43,89	58,90	55,05	43,77	71,20	61,15	46,00
58,90	58,31	42,78	63,50	56,88	44,86	61,60	58,59	45,43	54,20	54,24	42,58
60,70	57,26	43,54	57,60	57,65	42,53	64,40	57,78	45,20	54,50	54,99	42,32
65,50	58,10	44,54	61,20	56,28	44,11	66,30	55,12	45,17	69,10	60,99	44,85
68,50	58,51	45,64	56,60	53,94	43,65	67,40	55,89	43,34	55,90	54,41	42,62
59,10	55,71	43,98	56,70	55,88	42,82	60,10	55,24	44,08	66,00	58,19	45,69
52,50	53,07	42,12	65,70	57,13	45,74	59,40	55,68	43,69	68,00	59,93	45,50
61,20	56,94	44,13	56,30	56,03	42,57	53,90	52,69	42,72	62,00	56,80	44,20
67,30	59,19	44,97	57,10	57,39	42,21	61,30	54,99	44,38	56,70	55,66	42,41
63,90	54,77	45,65	70,30	58,99	46,38	64,10	57,28	45,08	67,00	58,49	45,56
54,70	56,57	44,69	61,90	56,82	44,08	59,00	55,73	43,23	53,80	52,44	43,38

7.2. ƯỚC LUỢNG, KIỂM ĐỊNH MỘT GIÁ TRỊ TRUNG BÌNH

Bài tập 7.2.1

Sản lượng cá đánh được trong 1 giờ (tấn) có phân phối chuẩn $X \sim N(\mu, \sigma^2)$. Ước lượng μ và kiểm định giả thiết $H_0: \mu = 2$ đối với $H_1: \mu \neq 2$.

Lưới	1	2	3	4	5	6	7	8	9	10
Sản lượng /giờ	1,2	2,5	1,0	4,0	3,0	2,8	0,6	3,4	2,0	2,5

Bài tập 7.2.2

Thời gian mang thai của bò phân phối chuẩn $X \sim N(\mu, \sigma^2)$. Theo dõi 6 con bò thấy thời gian mang thai (ngày) là: 307, 293, 293, 283, 294, 297. Ước lượng giá trị trung bình của thời gian mang thai và kiểm định sự khác biệt về thời gian mang thai của giống bò này với 285 ngày?

Bài tập 7.2.3a

Để xác định khối lượng tôm trong ao nuôi, tiến hành bắt và cân khối lượng của 30 con. Số liệu thu được như sau, đơn vị tính g:

34,14	32,41	25,90	33,43	34,38	38,25	28,94	27,80	26,67	32,37
35,83	23,47	25,69	34,08	27,73	35,00	26,81	32,01	23,61	29,22
31,98	36,09	35,43	23,37	29,63	30,69	34,01	32,27	28,01	31,32

Hãy ước lượng khối lượng trung bình của tôm trong ao nuôi và kiểm định giả thiết $H_0: \mu = 32$ g đối với $H_1: \mu \neq 32$ g.

Bài tập 7.2.3b

Với số liệu ở bài tập 7.2.3a, hãy kiểm định khối lượng trung bình của tôm ở ao nuôi $\mu = 29$ g, biết rằng độ lệch của tính trạng tại thời điểm tương ứng là 5 g.

7.3. ƯỚC LUỢNG, KIỂM ĐỊNH XÁC SUẤT P

Bài tập 7.3.1

Thả 1500 tôm bột vào lưới, sau 5 ngày thấy số còn sống trong lưới là 1250 con. Ước tính tỷ lệ sống của tôm bột sau khi thả trong ao nuôi.

Bài tập 7.3.2

Tỷ lệ tôm sống sau 1 tháng thả ở vụ trước là 80%. Để so sánh tỷ lệ tôm sống tại thời điểm tương ứng của vụ hiện tại, tiến hành đặt 5 sàng ăn trong ao. Số tôm đếm được lần lượt ở các sàng sau 3 giờ cho ăn là 26, 25, 32, 31 và 30 con. Anh (chị) cho biết tỷ lệ tôm sống ở 2 vụ có sự sai khác hay không? Nếu có sự sai khác, hãy ước tính tỷ lệ tôm sống ở vụ hiện tại. Giả sử mật độ thả ở hai vụ đều là 40 con/m² và số tôm trong sàng được ước tính bằng con/m².

Bài tập 7.3.3

Tại một trại lợn, chọn ngẫu nhiên 200 lợn thấy có 40 con bị viêm phổi.

- 1 Hãy ước tính tỷ lệ viêm phổi ở trại lợn nêu trên.
- 2 Nếu tỷ lệ mắc bệnh này ở trong vùng là 25%, theo anh (chị), tỷ lệ mắc bệnh viêm phổi của trại nêu trên có sai khác so với toàn vùng hay không?

Bài tập 7.3.4

Trong một mùa nhất định trong năm người ta thấy hình như trong số mới sinh của quần thể chim có nhiều con cái hơn. Chọn ngẫu nhiên 297 con chim mới sinh thì thấy có 167 con cái. Liệu có thể coi tỷ lệ giới tính đực /cái của quần thể nêu trên khác với tỷ lệ 1 /1 hay không?

Bài tập 7.3.5

Tỷ lệ gà lông nâu /gà lông trắng của một giống gà ở thế hệ F₁ là 3/ 1. Trong một trại nuôi giống gà trên thấy trong tổng số 400 con ở thế hệ F₁ có 320 con lông nâu. Có thể coi tỷ lệ gà lông nâu / gà lông trắng của trại cao hơn tỷ lệ 3/ 1 hay không?

7.4. SO SÁNH HAI GIÁ TRỊ TRUNG BÌNH

Bài tập 7.4.1

Tiến hành kiểm tra ảnh hưởng của việc điều trị đến sản lượng của bò sữa. Bò thí nghiệm được chọn có cùng chu kỳ và giai đoạn tiết sữa. Sản lượng sữa (kg) thu được trước và sau khi điều trị như sau:

Bò số	1	2	3	4	5	6	7	8	9
Trước khi điều trị	27	45	38	20	22	50	40	33	18
Sau khi điều trị	31	54	43	28	21	49	41	34	20

Liệu việc điều trị có ảnh hưởng đến sản lượng trung bình sữa bò hay không?

Bài tập 7.4.2

Để so sánh khối lượng của 2 giống bò, tiến hành chọn ngẫu nhiên và cân 12 con đực với giống thứ nhất và 15 con đực với giống thứ 2. Khối lượng (kg) thu được như sau:

Giống bò thứ nhất	187,6 194,7	180,3 221,1	198,6 186,7	190,7 203,1	196,3	203,8	190,2	201,0
Giống bò thứ hai	148,1 162,4	146,2 140,2	152,8 159,4	135,3 181,8	151,2 165,1	146,3 165,0	163,5 141,6	146,6

Theo anh (chị), khối lượng trung bình của 2 giống bò có khác nhau không?

Bài tập 7.4.3

Một thí nghiệm sinh lý học động vật nhằm nghiên cứu khả năng hấp thụ của động vật lưỡng cư. Phần trăm tăng khối lượng cơ thể sau khi ngâm mình trong nước hai giờ được ghi lại đối với ếch và cóc như sau:

Cóc	2,31	25,23	28,37	14,16	28,39	27,94	17,68
Éch	0,85	2,90	2,47	17,72	3,82	2,86	13,71

Theo anh (chị), cóc hay éch hấp thụ nước có khác nhau không?

Bài tập 7.4.4

Để so sánh khối lượng sơ sinh giữa 2 giống lợn Landrace và Yorksire nuôi tại trại Mỹ Văn; tiến hành cân khối lượng sơ sinh ngẫu nhiên của 100 lợn Landrace và 180 con của Yorkshire. Khối lượng sơ sinh trung bình của 100 lợn Landrace là 1,21 kg và độ lệch chuẩn là 0,15 kg; đối với 180 lợn giống Yorkshire có các giá trị tương ứng là 1,30 kg và 0,11 kg. Anh (chị) cho biết kết luận của mình về khối lượng lợn sơ sinh của 2 giống nêu trên.

7.5. SO SÁNH HAI XÁC SUẤT

Bài tập 7.5.1

Để thử hiệu lực của một loại vắc xin mới người ta tiêm vắc xin cho 400 cá thể, kết quả còn sống 350. Trong nhóm đối chứng (không tiêm) gồm 300 cá thể có 250 còn sống. Hãy kiểm định giả thiết H_0 : tỷ lệ sống không có sự sai khác khi sử dụng vắc xin và không sử dụng với đối thiết H_1 : sử dụng vắc xin có tỷ lệ sống cao hơn.

7.6. PHÂN TÍCH PHƯƠNG SAI MỘT YẾU TỐ

Bài tập 7.6.1

Theo dõi tăng khối lượng của cá (kg) trong thí nghiệm với 5 công thức nuôi (A, B, C, D và E). Hãy cho biết tăng khối lượng của cá ở các công thức nuôi có khác nhau không? Nếu khác hãy tiến hành so sánh cặp các giá trị trung bình và thể hiện sự khác nhau bằng các chữ cái.

A	B	C	D	E
0,95	0,43	0,70	1,00	0,90
0,85	0,45	0,90	0,95	1,00
0,85	0,40	0,75	0,90	0,95
0,90	0,42	0,70	0,90	0,95

Bài tập 7.6.2

So sánh tăng khối lượng của chuột ở 4 khẩu phần ăn khác nhau (1, 2, 3 và 4). Số chuột tham gia vào thí nghiệm vào từng khẩu phần lần lượt là 7, 8, 6 và 8. Số liệu thu được trình bày ở bảng sau (% tăng khối lượng so với khối lượng cơ thể):

Khẩu phần 1	3,42	3,96	3,87	4,19	3,58	3,76	3,84	
Khẩu phần 2	3,17	3,63	3,38	3,47	3,39	3,41	3,55	3,44
Khẩu phần 3	3,34	3,72	3,81	3,66	3,55	3,51		
Khẩu phần 4	3,64	3,93	3,77	4,18	4,21	3,88	3,96	3,91

Cho biết ảnh hưởng của khẩu phần ăn đến tăng khối lượng của chuột. Nếu có sự khác nhau, tiến hành so sánh các trung bình và thể hiện sự sai khác giữa từng cặp bằng các chữ cái.

Bài tập 7.6.3

Theo dõi trọng lượng xuất chuồng (kg) của lợn ở 5 công thức lai khác nhau. Lợn nuôi thí nghiệm được bố trí ở 4 khu chuồng khác nhau. Kết quả thu được trình bày như sau:

Công thức	Khu 1	Khu 2	Khu 3	Khu 4
1	77,8	76,9	75,4	74,1
2	83,7	80,3	80,6	78,0
3	76,7	72,0	72,4	70,7
4	78,0	77,0	75,9	75,7
5	71,8	70,0	73,0	71,6

Hãy so sánh khối lượng xuất chuồng của lợn ở các công thức lai nêu trên. Nếu có sự khác nhau, tiến hành so sánh các trung bình và thể hiện sự sai khác giữa từng cặp bằng các chữ cái.

Bài tập 7.6.4

Nghiên cứu số lượng tế bào lymphô ở chuột ($\times 1000$ tế bào mm^{-3} máu) được sử dụng 4 loại thuốc khác nhau qua 5 lứa; kết quả như sau:

	Lứa 1	Lứa 2	Lứa 3	Lứa 4	Lứa 5
Thuốc A	7,1	6,1	6,9	5,6	6,4
Thuốc B	6,7	5,1	5,9	5,1	5,8
Thuốc C	7,1	5,8	6,2	5,0	6,2
Thuốc D	6,7	5,4	5,7	5,2	5,3

Hãy so sánh ảnh hưởng của thuốc đến số lượng tế bào lymphô trong máu. Nếu có sự khác nhau, tiến hành so sánh các trung bình và thể hiện sự sai khác giữa từng cặp bằng các chữ cái.

Bài tập 7.6.5

Nghiên cứu ảnh hưởng của mật độ nuôi cá đến sản lượng trong ao nuôi. Có tất cả 6 mật độ khác nhau (1, 2, 3, 4, 5 và 6) được thử nghiệm trên 4 khu vực (1, 2, 3 và 4). Năng suất cá (tấn) ở thí nghiệm thu được như sau:

Mật độ nuôi	Khu 1	Khu 2	Khu 3	Khu 4
1	3,6	2,8	3,0	4,0
2	4,8	4,2	4,0	5,6
3	6,0	5,7	5,2	6,2
4	6,6	6,4	5,4	6,5
5	7,0	6,5	5,9	7,0
6	7,1	6,8	6,0	7,2

Cho biết ảnh hưởng của mật độ nuôi đến năng suất cá.

Bài tập 7.6.6

Nuôi 4 giống tảo trong 4 ao (cột) ở khu vực có nắng gắt do đó mỗi ao được chia thành 4 băng (hàng) vuông góc với hướng nắng sau đó chọn cách che phủ. Theo dõi năng suất của tảo (kg / m^2) ở thí nghiệm. Hãy cho biết kết luận của anh (chị) về năng

suất của 4 giống tảo nêu trên, nếu có sự sai khác hãy thể hiện sự sai khác của từng cặp giá trị trung bình bằng các chữ cái.

Che phủ	Ao			
	1	2	3	4
1	1,640 (B)	1,210 (D)	1,425 (C)	1,345 (A)
	1,475 (C)	1,185 (A)	1,400 (D)	1,290 (B)
3	1,670 (A)	0,710 (C)	1,665 (B)	1,180 (D)
	1,565 (D)	1,290 (B)	1,655 (A)	0,660 (C)

Bài tập 7.6.7

Một thí nghiệm được nghiên cứu trên bò nhằm xác định sự phân giải protein (%) trong dạ cỏ bằng kỹ thuật đặt túi nylon qua lỗ rò dạ cỏ trên 4 động vật thí nghiệm. Có 4 loại thức ăn được nghiên cứu và mỗi loại thức ăn được đặt lần lượt trong dạ cỏ của từng động vật thí nghiệm. Số liệu thu được trình bày trong bảng bên. Cho biết kết luận của anh (chị) về mức độ phân giải protein ở 4 loại thức ăn nêu trên, nếu có sự sai khác hãy thể hiện sự sai khác của từng cặp giá trị trung bình bằng các chữ cái.

Giai đoạn	Bò			
	1	2	3	4
1	42,80 (B)	42,90 (A)	69,10 (D)	49,40 (C)
	47,40 (A)	53,50 (D)	47,10 (C)	56,80 (B)
3	52,30 (C)	61,20 (B)	40,50 (A)	51,80 (D)
	61,30 (D)	51,20 (C)	54,00 (B)	39,70 (A)

7.7. PHÂN TÍCH PHƯƠNG SAI HAI HAI YẾU TỐ

Bài tập 7.7.1. Phân tích phương sai 2 nhân tố bối trí kiểu chéo nhau (crossed design)

Một thí nghiệm được tiến hành nhằm nghiên cứu ảnh hưởng của 3 mức (0, 100 và 200 ppm) axit sorbic và 6 mức (0,98; 0,94; 0,90; 0,86; 0,82 và 0,78) của hoạt hóa nước (a_w) đến tỷ lệ sống của *Salmonella typhimurium*. Mô hình thiết kế thí nghiệm theo kiểu khối ngẫu nhiên được sử dụng với 3 khối (I, II và III) với 18 tổ hợp công thức. Số liệu thu được (mật độ/ml) sau 7 ngày làm thí nghiệm như sau:

Axit	a_w	I	II	III
0 ppm	0,98	8,19	8,37	8,83
	0,94	6,65	6,70	6,25
	0,90	5,87	5,98	6,14
	0,86	5,06	5,35	5,01
	0,82	4,85	4,31	4,52
	0,78	4,31	4,34	4,20
	100 ppm	7,64	7,79	7,59
	0,98	6,52	6,19	6,51
	0,94			

Axit	a_w	I	II	III
200 ppm	0,90	5,01	5,28	5,78
	0,86	4,85	4,95	4,29
	0,82	4,29	4,43	4,18
	0,78	4,13	4,39	4,18
	0,98	7,14	6,92	7,19
	0,94	6,33	6,18	6,43
	0,90	5,20	5,10	5,43
	0,86	4,41	4,40	4,79
	0,82	4,26	4,27	4,37
	0,78	3,93	4,12	4,15

Cho biết kết luận của anh (chị) về tỷ lệ sống của *Salmonella typhimurium* ở các mức khác nhau của axit và a_w khác nhau. Nếu có sự sai khác, tiến hành so sánh và thể hiện sự khác nhau giữa các nghiệm thức bằng các chữ cái.

Bài tập 7.7.2:

Một thí nghiệm được tiến hành trên 4 khu chuồng (I, II, III và IV) nhằm xác định ảnh hưởng của việc bổ sung kẽm (Zn_0 và Zn_1) và đồng (Cu_0 và Cu_1) vào khẩu phần ăn đến tăng khối lượng của gà con. Tăng khối lượng (g) trung bình tuần của thí nghiệm như sau:

Kẽm	Đồng	I	II	III	IV
Zn_0	Cu_0	21,6	22,4	25,7	22,3
		16,3	18,9	24,8	22,2
	Cu_1	23,7	22,0	26,7	25,0
		19,2	11,2	15,5	19,9
Zn_1	Cu_0	21,2	26,3	28,1	23,0
		22,6	17,1	24,4	18,8
	Cu_1	20,9	22,3	26,7	21,5
		14,2	12,5	13,9	15,7

Kết luận của anh chị về ảnh hưởng của việc bổ sung đồng kẽm vào khẩu phần ăn đến tăng khối lượng của gà. Nếu có ảnh hưởng của việc bổ sung, thể hiện sự sai khác của từng cặp bằng các chữ cái.

Bài tập 7.7.3. Phân tích phương sai 2 nhân tố bố trí kiểu chia ô (split plot design)

Khảo sát năng suất của 4 giống cỏ (S23, New Zealand, Kent và X) được bón phân ở 2 mức khác nhau, mức nhiều (A) và mức trung bình (B). Thí nghiệm được tiến hành trên 4 khối (I, II, III và IV) của 4 ô lớn (giống cỏ) và 2 ô nhỏ (phân bón). Năng suất vật chất khô (tấn /ha) của vụ hè thu được như sau:

Giống cỏ	Mức bón phân	I	II	III	IV
S23	A	3,36	3,58	3,20	3,14
	B	2,78	2,27	1,92	2,06
New Zealand	A	3,54	2,78	3,25	3,45
	B	2,89	1,97	2,12	1,96
X	A	4,53	4,94	3,99	3,65
	B	2,50	1,91	2,16	1,98
Kent	A	4,30	3,97	4,31	3,49
	B	2,62	2,43	2,25	1,61

Kết luận về năng suất của 4 giống cỏ nêu trên. Nếu có ảnh hưởng của việc bỗ sung, thể hiện sự sai khác của từng cặp bằng các chữ cái.

Bài tập 7.7.4

Bốn trại (khối) nuôi 3 giống lợn tại 3 dãy (ô lớn), mỗi dãy ngăn làm 4 chuồng (ô nhỏ) để theo dõi 4 khẩu phần ăn.

Giống	Khẩu phần	I	II	III	IV
A	1	42,9	41,6	28,9	30,8
	2	53,8	58,5	43,9	46,3
	3	49,5	53,8	40,7	39,4
	4	44,4	41,8	28,3	34,7
B	1	53,3	69,6	45,4	35,1
	2	57,6	69,6	42,4	51,9
	3	59,8	65,8	41,4	45,4
	4	64,1	57,4	44,1	51,6
C	1	62,3	58,5	44,6	50,3
	2	63,4	50,4	45,0	46,7
	3	64,5	46,1	62,6	50,3
	4	63,6	56,1	52,7	51,8

Hãy phân tích phương sai và đưa ra kết luận đối với giống, khẩu phần ăn. Nếu có sự khác nhau giữa các giống, các khẩu phần ăn thì so sánh trung bình và thể hiện sự sai khác giữa các cặp bằng các chữ cái.

Bài tập 7.7.5: Phân tích hiệp phương sai

Một thí nghiệm được tiến hành nhằm nghiên cứu khối lượng (kg) kết thúc vỗ béo và tăng khối lượng (g/ngày) trong giai đoạn nuôi vỗ béo từ 21 đến 24 tháng tuổi của ba nhóm bò lai LaiSind, F1(Brahman × LaiSind) và F1(Charolais × LaiSind). Kết quả thu được như sau:

Bò số	Khối lượng đầu kỳ	Khối lượng cuối kỳ	Tăng khối lượng
LaiSind			
1	196	259	700,00
2	187	247	666,67
3	210	266	622,22
4	196	255	655,56
5	185	249	711,11
F1(Brahman × LaiSind)			
6	215	285	777,78
7	210	295	944,44
8	198	280	911,11
9	217	298	900,00
10	209	285	844,44
F1(Charolais × LaiSind)			
11	216	310	1044,44
12	235	321	955,56
13	230	320	1000,00
14	215	305	1000,00
15	210	295	944,44

Hãy cho biết ảnh hưởng của khối lượng đầu kỳ đến khối lượng cuối kỳ và tăng khối lượng. So sánh khối lượng cuối kỳ và tăng khối lượng giữa 3 nhóm bò lai trên.

7.8. TƯƠNG QUAN VÀ HỒI QUY TUYẾN TÍNH

Bài tập 7.8.1

Theo dõi khối lượng (kg) của bê ở các thời điểm khác nhau (tháng). Kết quả thu được như sau

Tuổi (x)	0	2	3	4	6	8	12
Khối lượng (y)	18	32	64	70	91	127	164

Tính hệ số tương quan r và kiểm định giả thiết $H_0: \rho = 0$ đối với $\rho \neq 0$. Tìm đường hồi quy tuyến tính để ước tính khối lượng bê tại các thời điểm khác nhau và kiểm định giả thiết đối với các hệ số của phương trình.

Bài tập 7.8.2

Khối lượng cá mẹ ($\times 100$ g) và số lượng trứng ($\times 1000$ quả) của một loài cá như sau:

Khối lượng	14	17	24	25	27	33	34	37	40	41	42
Số trứng	61	37	65	69	54	93	87	89	100	90	97

Tìm đường hồi quy tuyến tính để ước tính số lượng trứng thông qua khối lượng của cá mẹ và kiểm định giả thiết đối với các hệ số của phương trình. Dự báo số lượng trứng khi cá mẹ có khối lượng 4,5 kg.

Bài tập 7.8.3

Để xác định hệ số tương quan giữa các tính trạng năng suất sinh sản của lợn nái ngoại Landrace nuôi tại các cơ sở giống ở miền Bắc Việt Nam; tiến hành rút một cách ngẫu nhiên thành tích của 25 nái được bảng số liệu như sau:

scdr	sc21	poss	pssc	po21	p21c	scdr	sc21	poss	pssc	po21	p21c
13	10	13,5	1,23	32,0	3,20	10	10	14,0	1,40	51,0	5,10
8	8	11,2	1,40	41,6	5,20	12	10	15,5	1,29	42,0	4,20
13	12	18,2	1,40	72,0	6,00	11	10	12,9	1,29	63,8	6,38
14	8	14,0	1,40	42,0	5,25	10	10	13,0	1,30	46,0	4,60
8	8	11,5	1,44	36,0	4,50	12	9	14,0	1,27	43,0	4,78
10	10	14,5	1,45	60,0	6,00	12	10	14,6	1,22	52,0	5,20
9	8	10,8	1,20	40,0	5,00	9	8	12,5	1,39	34,0	4,25
12	12	19,0	1,58	65,0	5,42	10	9	17,6	1,76	50,0	5,56
10	10	10,5	1,05	38,2	3,82	10	9	13,7	1,37	40,3	4,48
11	10	12,0	1,09	62,0	6,20	11	10	14,0	1,27	51,0	5,10
10	8	14,0	1,40	47,0	5,88	10	8	13,0	1,30	46,3	5,79
10	10	16,0	1,60	56,0	5,60	8	8	9,0	1,13	36,6	4,58

Ghi chú:

scdr - Số con đẻ ra/lứa

sc21 - Số con còn sống đến 21 ngày tuổi

poss - Khối lượng sơ sinh/lứa (kg)

pssc - Khối lượng sơ sinh/con (kg)

po21 - Khối lượng 21 ngày tuổi/lứa (kg)

p21c - Khối lượng 21 ngày tuổi/con (kg)

Tính và kiểm định hệ số tương quan giữa các tính trạng, và điền các giá trị vào bảng sau:

Chỉ tiêu	(1)	(2)	(3)	(4)	(5)	(6)
Số con đẻ ra/lứa (1)						
Số con sống đến 21 ngày tuổi (2)						
Khối lượng sơ sinh/lứa (3)						
Khối lượng sơ sinh/con (4)						
Khối lượng 21 ngày tuổi/lứa (5)						
Khối lượng 21 ngày tuổi/con (6)						

Bài tập 7.8.4

Trong giai đoạn từ năm 1990 đến 2001, số lượng ngựa trong một trại được ghi lại như sau:

Năm	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Ngựa	110	110	105	104	90	95	92	90	88	85	78	80

Xây dựng phương trình hồi quy để mô tả diễn biến số lượng ngựa qua các năm. Hãy ước tính số lượng ngựa của trại năm 2007, giả sử rằng số lượng có xu hướng tuyến tính.

7.9. KIỂM ĐỊNH MỘT PHÂN PHỐI VÀ BẢNG TƯƠNG LIÊN

Bài tập 7.9.1

Theo dõi tỷ lệ chim đực và chim cái của một loài chim được kết quả: trong 500 con chim non có 230 con đực và 270 con cái. Kiểm định giả thiết tỷ lệ chim đực /chim cái là 1/1.

Bài tập 7.9.2

Với số liệu ở bài tập B-2.17 (tr.8) của BTTKTN-2010. Xử lý số liệu này bằng phép thử χ^2 . Anh (chị) cho biết có tồn tại mối liên hệ giữa vắc xin và tỷ lệ chết. Nếu tồn tại mối liên hệ, hãy tính nguy cơ tương đối (RR).

Bài tập 7.9.3a

Kết quả điều tra về sinh lý đường tiết niệu ở chó được như bảng bên. Liệu có tồn tại mối liên hệ giữa việc thiến và rối loạn bài tiết hay không? Nếu tồn tại, hay tính tỷ suất chênh và bình luận về tỷ suất này.

	Rối loạn	Bình thường
Thiến	34	757
Không	7	2427

Bài tập 7.9.3b

Hãy giải quyết các câu hỏi tương tự như ở bài tập B-4.4a, trong trường hợp dung lượng mẫu hạn chế và số liệu thu được như bảng sau:

	Rối loạn	Bình thường
Thiến	7	121
Không	2	642

Bài tập 7.9.4

Kết quả theo dõi 3 nhóm (A, B và O) máu và 2 nhóm bệnh (viêm loét dạ dày - I và ung thư vú ở nữ - II) trong thời gian 12 tháng thấy số bệnh nhân ở các nhóm máu A, B và O mắc bệnh I lần lượt là 679, 134 và 983; bệnh II lần lượt là 416, 84 và 383. Liệu có tồn tại mối liên hệ giữa giữa nhóm máu và tỷ lệ mắc bệnh hay không? Nếu có, hãy tính nguy cơ tương đối (RR) và giải thích hệ số này.

Bài tập 7.9.5

Nghiên cứu tác dụng của 3 loại vắc xin (A, B và C) và đói chứng (ĐC) đến mức độ bệnh sau 24 tháng theo dõi:

Vắc xin	Mức độ bệnh		
	Không	Trung bình	Năng
ĐC	100	71	29
A	146	32	17
B	149	28	16
C	146	37	17

Hãy cho biết các biện pháp xử lý có cho kết quả phòng bệnh như nhau không?

PHỤ LỤC 1

MỘT SỐ THUẬT NGỮ DÙNG TRONG GIÁO TRÌNH

Thuật ngữ	Tiếng Anh
Bảng tương liên	Contingency table
Bậc tự do	Degree of freedom
Biến ngẫu nhiên	Random variable
Các số đặc trưng của mẫu	Statistics, Statistical measures, Characteristics of a sample
Công thức Bayes	Bayes rule
Công thức xác suất toàn phần	Total probability formula
Chỉnh hợp	Arrangement
Chỉnh hợp lặp	Arrangement with repetition
Chấp nhận hay bác bỏ giả thiết	Accept and reject hypothesis
Phân phối xác suất của biến rời rạc, bảng (dãy) phân phối	Discrete probability distribution, frequency array
Dung lượng mẫu (kích thước mẫu)	Size of sample
Dự báo	Prediction, forecasting
Dữ liệu định lượng	Quantitative data
Dữ liệu định tính	Qualitative data
Độ lệch chuẩn	Standard deviation
Độc lập	Independent
Độ nhọn	Kurtosis
Độ lệch, độ bất đối xứng	Skewness
Độ tin cậy	Degree of confidence
Giả thiết thống kê	Statistical hypothesis
Giả thiết và đối thiết	Hypothesis and alternative hypothesis
Giả thiết không (H_0)	Null hypothesis
Hàm phân phối	Distribution function
Hàm mật độ xác suất	Probability density function
Hiệp phương sai	Covariance
Thuật ngữ	Tiếng Anh
Hệ số góc	Slope
Hoán vị	Permutation
Hồi quy tuyến tính	Linear regression
Kiểm định giả thiết	Tests of hypotheses, Testing hypothesis
Kiểm định một phân phối	Goodness of fit test
Kiểm định hai phia	Two tailed test
Đối thiết hai phia	Two side alternative
Kỳ vọng toán học	Mathematical expectation
Mẫu quan sát	Sample
Mod	Mode
Nguyên tắc bình phương bé nhất	Method (principle) of least squares
Nhật đồ, tổ chức đồ	Histogram

Phân phối χ^2	Chi-square distribution
Phân phối chuẩn	Normal distribution Gaussian distribution
Phân phối chuẩn tắc	Standard normal distribution
Phân phối hình học	Geometric distribution
Phân phối Fisher Snedecor	Fisher Snedecor distribution F distribution
Phân phối liên tục	Continuos distribution
Phân phối rời rạc	Discrete distribution
Phân phối nhị thức	Binomial distribution
Phân phối Poátxông	Poisson distribution
Phân phối siêu bội	Hypergeometric distribution
Phân phối Student	Student distribution t distribution
Phần trăm	Percentage
Phép thử	Experiment
Phương sai	Variance (dispersion)
Quy tắc cộng xác suất	Additive rule of probability
Quy tắc nhân xác suất	Multiplicative rule
Rủi so (Sai lầm) loại I và II	Type I and II risk (error)
Thuật ngữ	Tiếng Anh
Sai số chuẩn	Standard error
So sánh trung bình lấy mẫu theo cặp	Paired comparaison for means
Sự kiện	Event
Sự kiện cơ bản	Element
Tương quan	Correlation
Hệ số tương quan	Correlation coefficient
Tần số	Frequency
Thống kê mô tả	Descriptive Statistics
Thiết kế hoàn toàn ngẫu nhiên	Completely randomized design
Thiết kế khối ngẫu nhiên đầy đủ	Radomized completely block design
Thiết kế kiểu chéo nhau	Crossed design
Thiết kế kiểu phân cấp hay chia ô	Hierachical Nested design
Thiết kế kiểu chia ô	Split plot design
Tổ hợp	Combination
Tổng thể	Population
Tứ phân vị	Quartile
Trung bình cộng	Mean, sample mean, arithmetic mean, average
Trung vị	Median
Tung độ gốc	Intercept
Xác suất	Probability
Ước lượng, ước lượng tham số	Estimate, estimation of parameters
Ước lượng điểm	Point estimate
Ước lượng khoảng	Interval estimate
Ước lượng khoảng của kỳ vọng (Khoảng tin cậy của kỳ vọng)	Interval estimation of mean (Confidence interval for mean)
Ước lượng khoảng của xác suất (khoảng tin cậy của xác suất)	Interval estimation of Probability (Confidence interval for p)

PHỤ LỤC 2

BẢNG CÁC KÝ HIỆU TOÁN HỌC

Tên đầy đủ	Viết tắt
Phương sai của tổng thể	σ^2
Độ lệch chuẩn của mẫu quan sát	SD
Hệ số biến động	CV%
Hệ số góc của đường hồi quy tuyến tính	b
Hệ số tương quan của mẫu	r
Hệ số tương quan của tổng thể	p
Khoảng tin cậy	CI
Mức sai cho phép, mức ý nghĩa	α ($\alpha = 1 - P$)
Mức tin cậy	P
Mode	Mod
Nguồng χ^2 ở mức α , bậc tự do df	$\chi^2(\alpha, df)$
Nguồng F ở mức α , bậc tự do dft, dfm	F(α , dft, dfm)
Nguồng t ở mức α , bậc tự do df	t(α, df)
Nguồng Z của phân phối chuẩn ở mức α	Z(α)
Phân phối chuẩn	$N(\mu, \sigma^2); X \sim N(\mu, \sigma^2)$
Phương sai của sai số trong phân tích phương sai	msE se ²
Phương sai của tổng thể	σ^2
Phương sai mẫu đã điều chỉnh	S ² _P
Phương sai	S ²
Sai số chuẩn của hiệu số	SED SE(D)
Sai số chuẩn	SE, se(\bar{x}), S _{\bar{x}} , s _m , SE mean SEM
Sai số của một quan sát trong phân tích phương sai và trong phân tích hồi quy	SE
Trung bình cộng	\bar{x} xtb
Trung bình của tổng thể	μ
Trung vị Median	Med
Tung độ gốc của đường hồi quy tuyến tính	a
Xác suất của tổng thể	p
Tần suất trong mẫu	f hay k

PHỤ LỤC 3

HÀM PHÂN PHỐI CHUẨN

Các giá trị trong bảng là của phân bố chuẩn với trung bình bằng 0 và độ lệch chuẩn là 1. Ứng với mỗi giá trị z trong là giá trị P , $P(Z < z)$.

z	P	z	P	z	P	z	P
-4,00	0,00003	-1,50	0,0668	0,00	0,5000	1,55	0,9394
-3,50	0,00023	-1,45	0,0735	0,05	0,5199	1,60	0,9452
-3,00	0,0013	-1,40	0,0808	0,10	0,5398	1,65	0,9505
-2,95	0,0016	-1,35	0,0885	0,15	0,5596	1,70	0,9554
-2,90	0,0019	-1,30	0,0968	0,20	0,5793	1,75	0,9599
-2,85	0,0022	-1,25	0,1056	0,25	0,5987	1,80	0,9641
-2,80	0,0026	-1,20	0,1151	0,30	0,6179	1,85	0,9678
-2,75	0,0030	-1,15	0,1251	0,35	0,6368	1,90	0,9713
-2,70	0,0035	-1,10	0,1357	0,40	0,6554	1,95	0,9744
-2,65	0,0040	-1,05	0,1469	0,45	0,6736	2,00	0,9772
-2,60	0,0047	-1,00	0,1587	0,50	0,6915	2,05	0,9798
-2,55	0,0054	-0,95	0,1711	0,55	0,7088	2,10	0,9821
-2,50	0,0062	-0,90	0,1841	0,60	0,7257	2,15	0,9842
-2,45	0,0071	-0,85	0,1977	0,65	0,7422	2,20	0,9861
-2,40	0,0082	-0,80	0,2119	0,70	0,7580	2,25	0,9878
-2,35	0,0094	-0,75	0,2266	0,75	0,7734	2,30	0,9893
-2,30	0,0107	-0,70	0,2420	0,80	0,7881	2,35	0,9906
-2,25	0,0122	-0,65	0,2578	0,85	0,8023	2,40	0,9918
-2,20	0,0139	-0,60	0,2743	0,90	0,8159	2,45	0,9929
-2,15	0,0158	-0,55	0,2912	0,95	0,8289	2,50	0,9938
-2,10	0,0179	-0,50	0,3085	1,00	0,8413	2,55	0,9946
-2,05	0,0202	-0,45	0,3264	1,05	0,8531	2,60	0,9953
-2,00	0,0228	-0,40	0,3446	1,10	0,8643	2,65	0,9960
-1,95	0,0256	-0,35	0,3632	1,15	0,8749	2,70	0,9965
-1,90	0,0287	-0,30	0,3821	1,20	0,8849	2,75	0,9970
-1,85	0,0322	-0,25	0,4013	1,25	0,8944	2,80	0,9974
-1,80	0,0359	-0,20	0,4207	1,30	0,9032	2,85	0,9978
-1,75	0,0401	-0,15	0,4404	1,35	0,9115	2,90	0,9981
-1,70	0,0446	-0,10	0,4602	1,40	0,9192	2,95	0,9984
-1,65	0,0495	-0,05	0,4801	1,45	0,9265	3,00	0,9987
-1,60	0,0548	0,00	0,5000	1,50	0,9332	3,50	0,99977
-1,55	0,0606					4,00	0,99997

Một vài giá trị tối hạn của z :

P	0,80	0,90	0,95	0,975	0,99	0,995	0,999
z	0,842	1,282	1,645	1,960	2,326	2,576	3,090

PHỤ LỤC 4

HÀM PHÂN PHỐI STUDENT (t)

Các giá trị trong bảng là của phân bố t . Cột thứ nhất là bậc tự do (df). Các cột còn lại cho ta các giá trị lý thuyết về kiểm định một hướng (phần trên); $P(T_{df} > t) = P$, hoặc 2 hướng; $P(T_{df} > t \text{ hoặc } T_{df} < -t) = P$ trong đó P là mức xác suất được thể hiện ở đầu cột.

df	<i>P</i>					
	0,10	0,05	0,025	0,01	0,005	0,001 (1 hướng)
	0,20	0,10	0,05	0,02	0,01	0,002 (2 hướng)
1	3,078	6,314	12,706	31,821	63,657	318,313
2	1,886	2,920	4,303	6,965	9,925	22,327
3	1,638	2,353	3,182	4,541	5,841	10,215
4	1,533	2,132	2,776	3,747	4,604	7,173
5	1,476	2,015	2,571	3,365	4,032	5,893
6	1,440	1,943	2,447	3,143	3,707	5,208
7	1,415	1,895	2,365	2,998	3,499	4,785
8	1,397	1,860	2,306	2,896	3,355	4,501
9	1,383	1,833	2,262	2,821	3,250	4,297
10	1,372	1,812	2,228	2,764	3,169	4,144
11	1,363	1,796	2,201	2,718	3,106	4,025
12	1,356	1,782	2,179	2,681	3,055	3,930
13	1,350	1,771	2,160	2,650	3,012	3,852
14	1,345	1,761	2,145	2,624	2,977	3,787
15	1,341	1,753	2,131	2,602	2,947	3,733
16	1,337	1,746	2,120	2,583	2,921	3,686
17	1,333	1,740	2,110	2,567	2,898	3,646
18	1,330	1,734	2,101	2,552	2,878	3,611
19	1,328	1,729	2,093	2,539	2,861	3,579
20	1,325	1,725	2,086	2,528	2,845	3,552
21	1,323	1,721	2,080	2,518	2,831	3,527
22	1,321	1,717	2,074	2,508	2,819	3,505
23	1,319	1,714	2,069	2,500	2,807	3,485
24	1,318	1,711	2,064	2,492	2,797	3,467
25	1,316	1,708	2,060	2,485	2,787	3,450
26	1,315	1,706	2,056	2,479	2,779	3,435
27	1,314	1,703	2,052	2,473	2,771	3,421

28	1,313	1,701	2,048	2,467	2,763	3,408
29	1,311	1,699	2,045	2,462	2,756	3,396
30	1,310	1,697	2,042	2,457	2,750	3,385
40	1,303	1,684	2,021	2,423	2,704	3,307
60	1,296	1,671	2,000	2,390	2,660	3,232
120	1,289	1,658	1,980	2,358	2,617	3,160
∞	1,282	1,645	1,960	2,326	2,576	3,090

PHỤ LỤC 5

HÀM PHÂN PHỐI KHI BÌNH PHƯƠNG (χ^2)

Giá trị trong bảng là của phân bố χ^2 . Cột thứ nhất là bậc tự do (df). Các cột còn lại cho ta các giá trị lý thuyết ở phần đuôi; $P(\chi^2_{df} > x^2) = P$, trong đó P là mức xác suất thể hiện ở đầu cột.

df	<i>P</i>					
	0,10	0,05	0,025	0,01	0,005	0,001
1	2,71	3,84	5,02	6,63	7,88	10,83
2	4,61	5,99	7,38	9,21	10,60	13,82
3	6,25	7,81	9,35	11,34	12,84	16,27
4	7,78	9,49	11,14	13,28	14,86	18,47
5	9,24	11,07	12,83	15,09	16,75	20,51
6	10,64	12,59	14,45	16,81	18,55	22,46
7	12,02	14,07	16,01	18,48	20,28	24,32
8	13,36	15,51	17,53	20,09	21,95	26,12
9	14,68	16,92	19,02	21,67	23,59	27,88
10	15,99	18,31	20,48	23,21	25,19	29,59
11	17,28	19,68	21,92	24,73	26,76	31,26
12	18,55	21,03	23,34	26,22	28,30	32,91
13	19,81	22,36	24,74	27,69	29,82	34,53
14	21,06	23,68	26,12	29,14	31,32	36,12
15	22,31	25,00	27,49	30,58	32,80	37,70
16	23,54	26,30	28,85	32,00	34,27	39,25
17	24,77	27,59	30,19	33,41	35,72	40,79
18	25,99	28,87	31,53	34,81	37,16	42,31
19	27,20	30,14	32,85	36,19	38,58	43,82
20	28,41	31,41	34,17	37,57	40,00	45,31
21	29,62	32,67	35,48	38,93	41,40	46,80
22	30,81	33,92	36,78	40,29	42,80	48,27
23	32,01	35,17	38,08	41,64	44,18	49,73
24	33,20	36,42	39,36	42,98	45,56	51,18
25	34,38	37,65	40,65	44,31	46,93	52,62
26	35,56	38,89	41,92	45,64	48,29	54,05
27	36,74	40,11	43,19	46,96	49,65	55,48
28	37,92	41,34	44,46	48,28	50,99	56,89
29	39,09	42,56	45,72	49,59	52,34	58,30
30	40,26	43,77	46,98	50,89	53,67	59,70
40	51,81	55,76	59,34	63,69	66,77	73,40
50	63,17	67,50	71,42	76,15	79,49	86,66
60	74,40	79,08	83,30	88,38	91,95	99,61
80	96,58	101,88	106,63	112,33	116,32	124,84

100 118,50 124,34 129,56 135,81 140,17 149,45

Đối với trường hợp bậc tự do lớn ta có thể tính toán như sau, áp dụng phân bố chuẩn cho χ^2 , $z = \sqrt{2\chi^2} - \sqrt{2 \times df - 1}$, và so sánh giá trị z với “Bảng xác suất của phân bố tiêu chuẩn hóa”

PHỤ LỤC 6

HÀM PHÂN PHỐI FISHER

Trong bảng là giá trị của phân bố Fisher F . Bậc tự do (v_1) xác định vị trí của cột và bậc tự do (v_2) xác định vị trí của hàng. Các giá trị trong bảng là giá trị lý thuyết của phần đuôi trên; $P = (F_{v_1, v_2} > f) = P$, trong đó P là xác suất (0,10; 0,05; 0,01).

v_2	P	v_1																			
		1	2	3	4	5	6	7	8	9	10	11	12	15	20	24	30	40	60	120	∞
1	0,10	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19	60,47	60,71	61,22	61,74	62,00	62,26	62,53	62,79	63,06	63,33
	0,05	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,0	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
	0,01	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6083	6107	6157	6209	6234	6260	6286	6313	6340	6366
2	0,10	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,40	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,48	9,49
	0,05	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
	0,01	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40	99,41	99,42	99,43	99,45	99,46	99,47	99,48	99,48	99,49	99,50
3	0,10	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,22	5,20	5,18	5,18	5,17	5,16	5,15	5,14	5,13
	0,05	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
	0,01	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,13	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	0,10	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,91	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78	3,76
	0,05	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
	0,01	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,45	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	0,10	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,28	3,27	3,24	3,21	3,19	3,17	3,16	3,14	3,12	3,10
	0,05	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
	0,01	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,96	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	0,10	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,92	2,90	2,87	2,84	2,82	2,80	2,78	2,76	2,74	2,72
	0,05	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
	0,01	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	0,10	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,68	2,67	2,63	2,59	2,58	2,56	2,54	2,51	2,49	2,47

	0,05	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
	0,01	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	0,10	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,52	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32	2,29
	0,05	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
	0,01	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	0,10	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,40	2,38	2,34	2,30	2,28	2,25	2,23	2,21	2,18	2,16
	0,05	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
	0,01	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31

v_2	P	1	2	3	4	5	6	7	8	9	10	11	12	15	20	24	30	40	60	120	∞
10	0,10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,30	2,28	2,24	2,20	2,18	2,16	2,13	2,11	2,08	2,06
	0,05	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
	0,01	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,77	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	0,10	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,23	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00	1,97
	0,05	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
	0,01	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	0,10	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,17	2,15	2,10	2,06	2,04	2,01	1,99	1,96	1,93	1,90
	0,05	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
	0,01	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
15	0,10	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06	2,04	2,02	1,97	1,92	1,90	1,87	1,85	1,82	1,79	1,76
	0,05	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
	0,01	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
20	0,10	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,91	1,89	1,84	1,79	1,77	1,74	1,71	1,68	1,64	1,61
	0,05	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
	0,01	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,29	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
24	0,10	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88	1,85	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57	1,53
	0,05	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,22	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73

	0,01	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,09	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
30	0,10	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,79	1,77	1,72	1,67	1,64	1,61	1,57	1,54	1,50	1,46
	0,05	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,13	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
	0,01	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,91	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	0,10	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,74	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42	1,38
	0,05	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
	0,01	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,73	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	0,10	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,68	1,66	1,60	1,54	1,51	1,48	1,44	1,40	1,35	1,29
	0,05	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,95	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
	0,01	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,56	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	0,10	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,63	1,60	1,55	1,48	1,45	1,41	1,37	1,32	1,26	1,19
	0,05	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,87	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
	0,01	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,40	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	0,10	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63	1,60	1,57	1,55	1,49	1,42	1,38	1,34	1,30	1,24	1,17	1,00
	0,05	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,79	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00
	0,01	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,25	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

PHỤ LỤC 7

GIÁ TRỊ 2½% PHÍA TRÊN CỦA PHÂN PHỐI FISHER (F)

Giá trị trong bảng là của phân bố Fisher F . Bậc tự do (v_1) xác định vị trí của cột và bậc tự do (v_2) xác định vị trí của hàng. Các giá trị trong bảng là giá trị lý thuyết tại điểm 2,5%; $P(F_{v_1, v_2} > f) = 0,025$

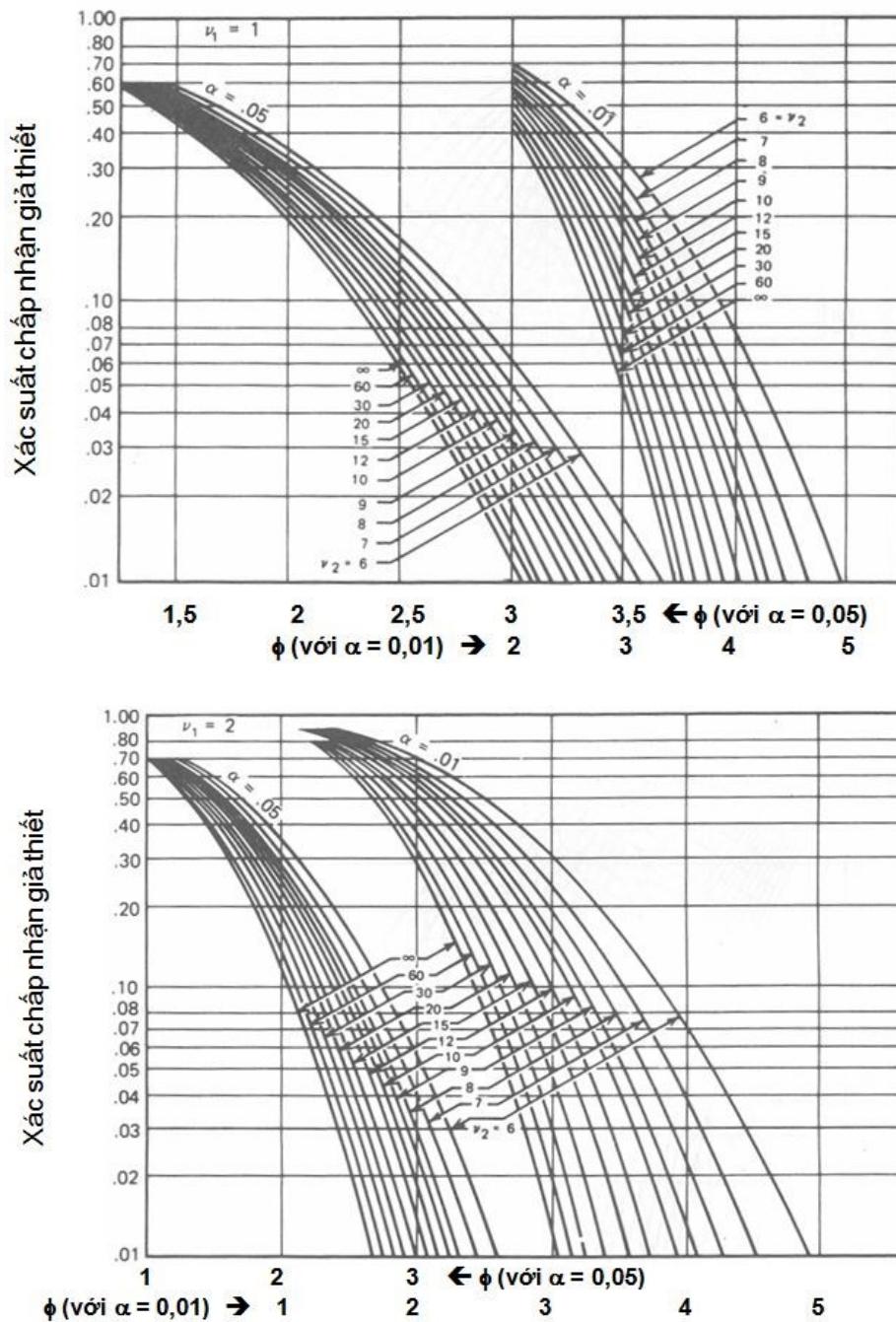
v_2	1	2	3	4	5	6	7	8	9	10	11	12	15	20	24	v_1	30	40	60	120	∞
1	647,8	799,5	864,2	899,6	921,8	937,1	948,2	956,6	963,3	968,6	973,0	976,7	984,9	993,1	997,3	1001	1006	1010	1014	1018	
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,41	39,41	39,43	39,45	39,46	39,46	39,47	39,48	39,49	39,50	
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,37	14,34	14,25	14,17	14,12	14,08	14,04	13,99	13,95	13,90	
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,79	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26	
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,57	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02	
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,41	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85	
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,71	4,67	4,57	4,47	4,41	4,36	4,31	4,25	4,20	4,14	
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,24	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67	
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,91	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33	
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,66	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08	
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,47	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88	
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,32	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,72	
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	3,01	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40	
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,72	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09	
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,59	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94	
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,46	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79	
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,33	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64	
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,22	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48	
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16	2,10	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31	
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05	1,99	1,94	1,83	1,71	1,64	1,57	1,48	1,39	1,27	1,00	

PHỤ LỤC 8
KHOẢNG Ý NGHĨA ĐỐI VỚI
KIỂM ĐỊNH ĐA PHẠM VI DUNCAN

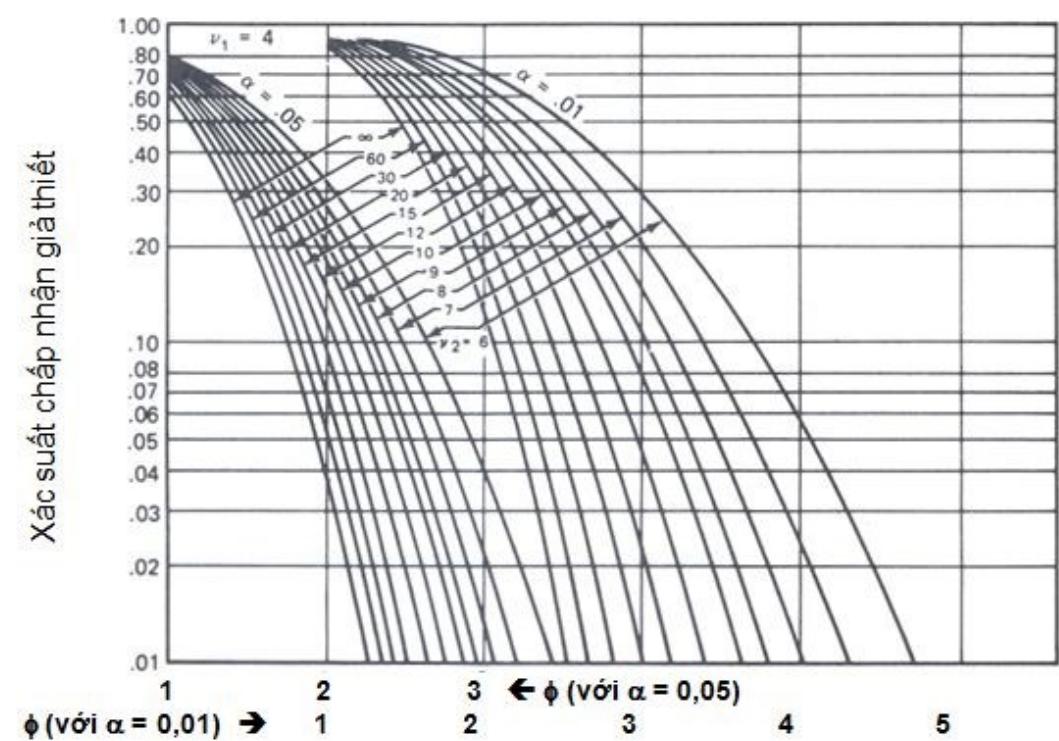
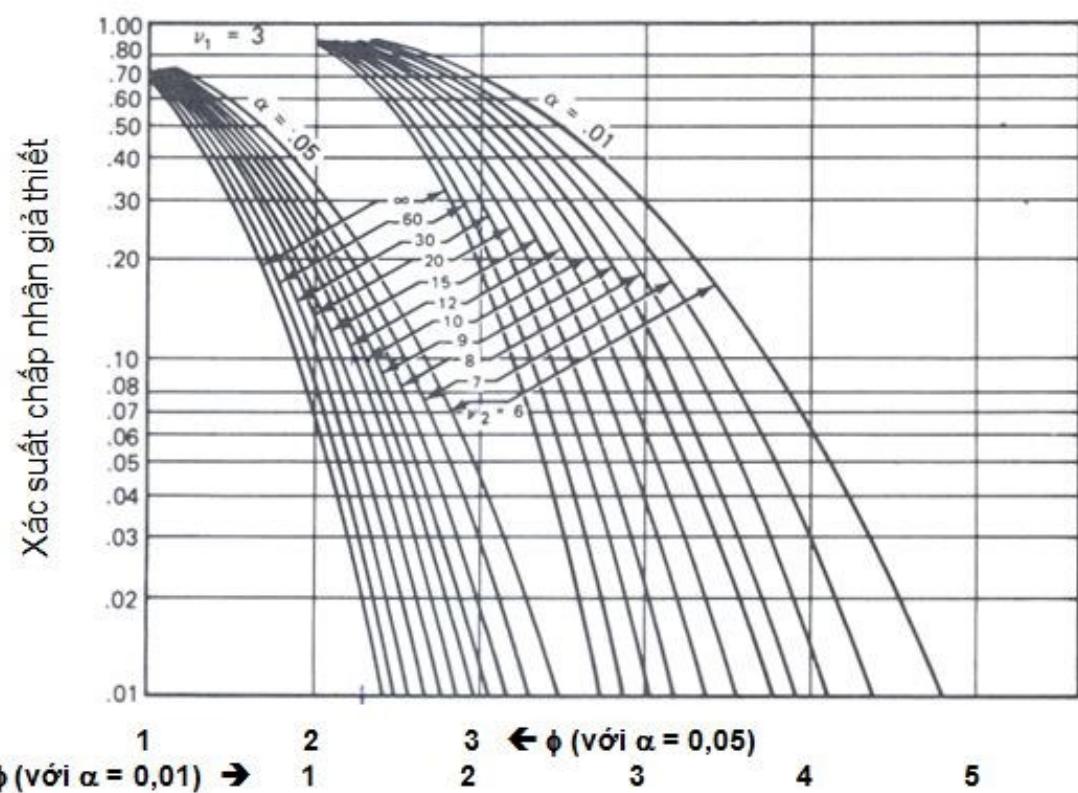
df	P	Số công thức thí nghiệm so sánh									
		2	3	4	5	6	7	8	9	10	
5	0,05	3,64	3,74	3,79	3,83	3,83	3,83	3,83	3,83	3,83	
	0,01	5,70	5,96	6,11	6,18	6,26	6,33	6,40	6,44	6,50	
6	0,05	3,46	3,58	3,64	3,68	3,68	3,68	3,68	3,68	3,68	
	0,01	5,24	5,51	5,65	5,73	5,81	5,88	5,95	6,00	6,00	
7	0,05	3,35	3,47	3,54	3,58	3,60	3,61	3,61	3,61	3,61	
	0,01	4,95	5,22	5,37	5,45	5,53	5,61	5,69	5,73	5,80	
8	0,05	3,26	3,39	3,47	3,52	3,55	3,56	3,56	3,56	3,56	
	0,01	4,74	5,00	5,14	5,23	5,32	5,40	5,47	5,51	5,50	
9	0,05	3,20	3,34	3,41	3,47	3,50	3,52	3,52	3,52	3,52	
	0,01	4,60	4,86	4,99	5,08	5,17	5,25	5,32	5,36	5,40	
10	0,05	3,15	3,30	3,37	3,43	3,46	3,47	3,47	3,47	3,47	
	0,01	4,48	4,73	4,88	4,96	5,06	5,13	5,20	5,24	5,28	
11	0,05	3,11	3,27	3,35	3,39	3,43	3,44	3,45	3,46	3,46	
	0,01	4,39	4,63	4,77	4,86	4,94	5,01	5,06	5,12	5,15	
12	0,05	3,08	3,27	3,33	3,36	3,40	3,42	3,44	3,44	3,46	
	0,01	3,42	4,55	4,68	4,76	4,84	4,92	4,96	5,02	5,07	
13	0,05	3,06	3,21	3,30	3,35	3,38	3,41	3,42	3,44	3,45	
	0,01	4,26	4,48	4,62	4,69	4,74	4,84	4,88	4,94	4,98	
14	0,05	3,03	3,18	3,27	3,33	3,37	3,39	3,41	3,42	3,44	
	0,01	4,21	4,42	4,55	4,63	4,70	4,78	4,83	4,87	4,91	
15	0,05	3,01	3,16	3,25	3,31	3,36	3,38	3,40	3,42	3,43	
	0,01	4,17	4,37	4,50	4,58	4,64	4,72	4,77	4,81	4,84	
16	0,05	3,00	3,15	3,23	3,30	3,34	3,37	3,39	3,41	3,43	
	0,01	4,13	4,34	4,45	4,53	4,60	4,67	4,72	4,76	4,79	
17	0,05	2,98	3,13	3,22	3,28	3,33	3,36	3,38	3,40	3,42	
	0,01	4,10	4,30	4,41	4,50	4,56	4,63	4,68	4,72	4,75	
18	0,05	2,97	3,12	3,21	3,27	3,32	3,35	3,37	3,39	2,41	
	0,01	4,07	4,27	4,38	4,46	4,53	4,59	4,64	4,69	4,71	
19	0,05	2,96	3,11	3,19	3,26	3,31	3,35	3,37	3,39	3,41	
	0,01	4,05	4,24	4,35	4,43	4,50	4,56	4,61	4,64	4,67	
20	0,05	2,95	3,10	3,18	3,25	3,30	3,34	3,36	3,38	3,40	
	0,01	4,02	4,22	4,33	4,40	4,47	4,53	4,58	4,61	4,65	
24	0,05	2,92	3,07	3,15	3,22	3,28	3,31	3,34	3,37	3,38	
	0,01	3,96	4,14	4,24	4,43	4,39	4,44	4,49	4,53	4,57	
30	0,05	2,89	3,04	3,12	3,20	3,25	3,29	3,32	3,35	3,37	
	0,01	3,89	4,06	4,16	4,22	4,32	4,36	4,41	4,45	4,48	
40	0,05	2,86	3,01	3,10	3,17	3,22	3,27	3,30	3,33	3,35	

df	P	Số công thức thí nghiệm so sánh								
		2	3	4	5	6	7	8	9	10
60	0,01	3,82	3,99	4,10	4,17	4,24	4,30	4,34	4,37	4,41
	0,05	2,83	2,98	3,08	3,14	3,20	3,24	3,28	3,31	3,33
	0,01	3,76	3,92	4,03	4,12	4,17	4,23	4,27	4,31	4,34
120	0,05	2,80	2,95	3,05	3,12	3,18	3,22	3,26	3,29	3,32
	0,01	3,71	3,86	3,98	4,06	4,11	4,17	4,21	4,25	4,29
	∞	0,05	2,77	2,92	3,02	3,09	3,15	3,19	3,23	3,26
		0,01	3,64	3,80	3,90	3,98	4,04	4,09	4,14	4,17
										4,20

PHỤ LỤC 9
ĐƯỜNG CONG XÁC ĐỊNH DUNG LƯỢNG MẪU
TRONG MÔ HÌNH CỐ ĐỊNH

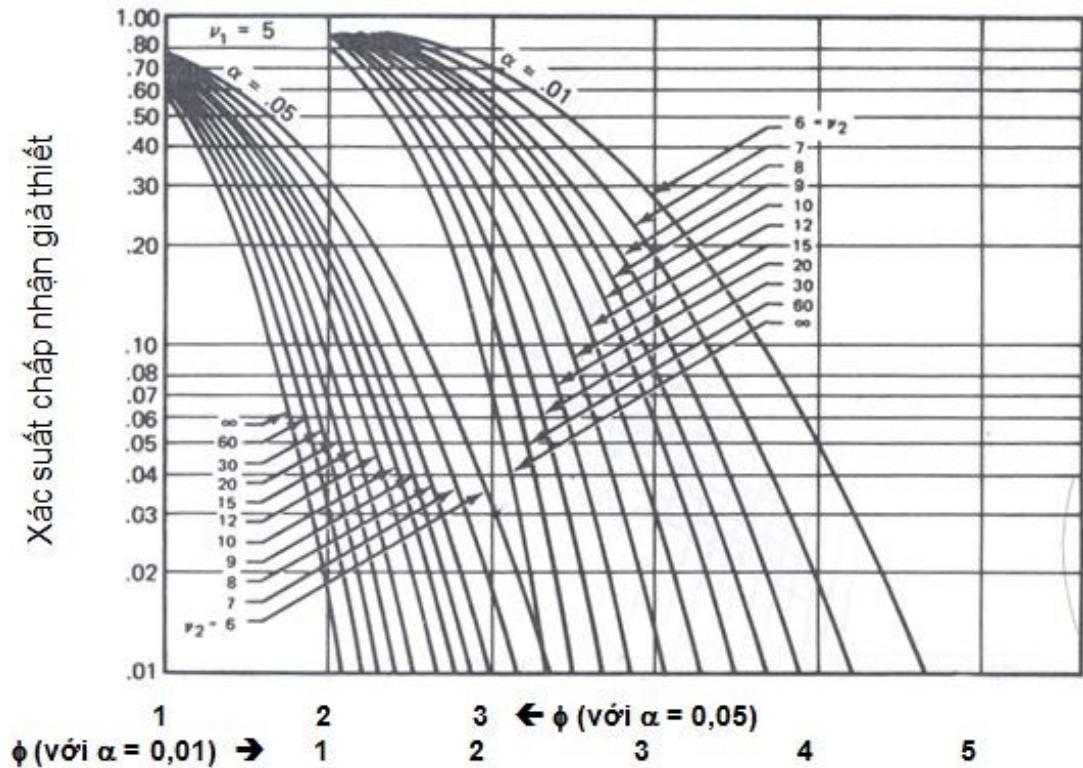


Với v_1 = bậc tự do ở tử số, v_2 = bậc tự do ở mẫu số.

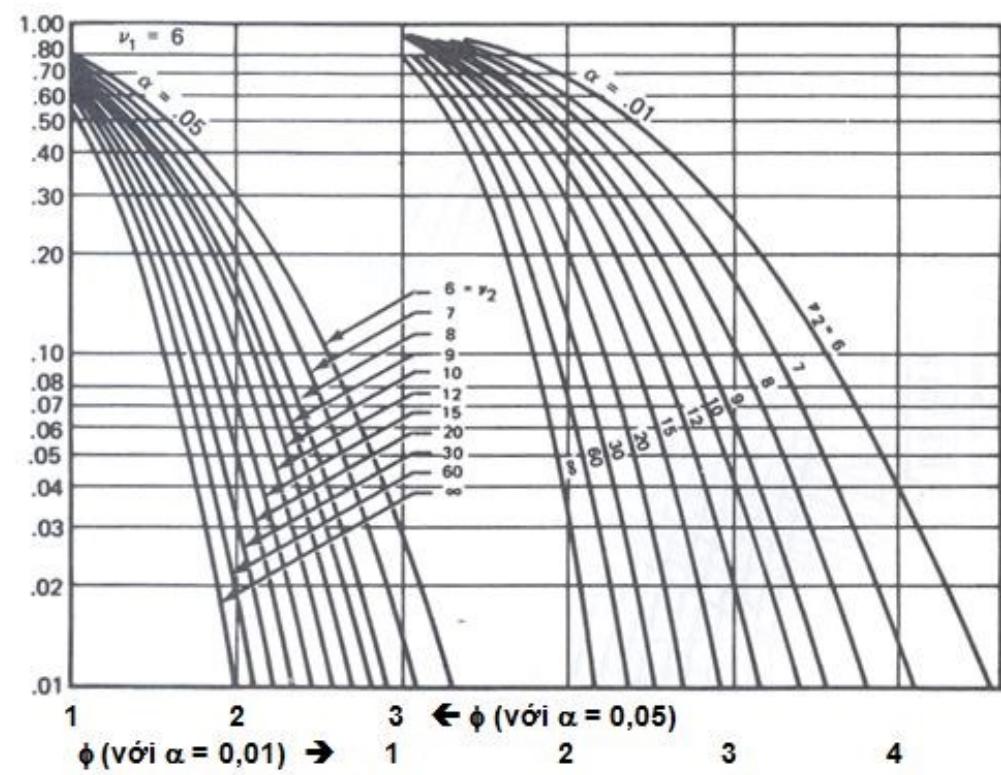


Với ν_1 = bậc tự do ở tử số, ν_2 = bậc tự do ở mẫu số.

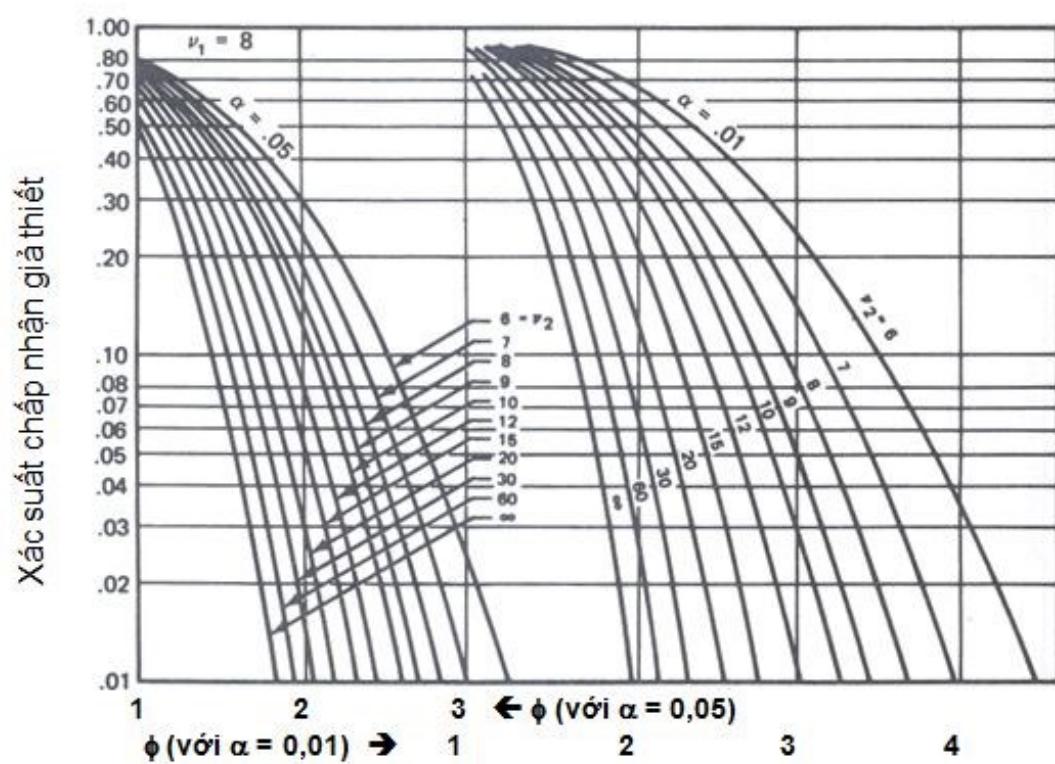
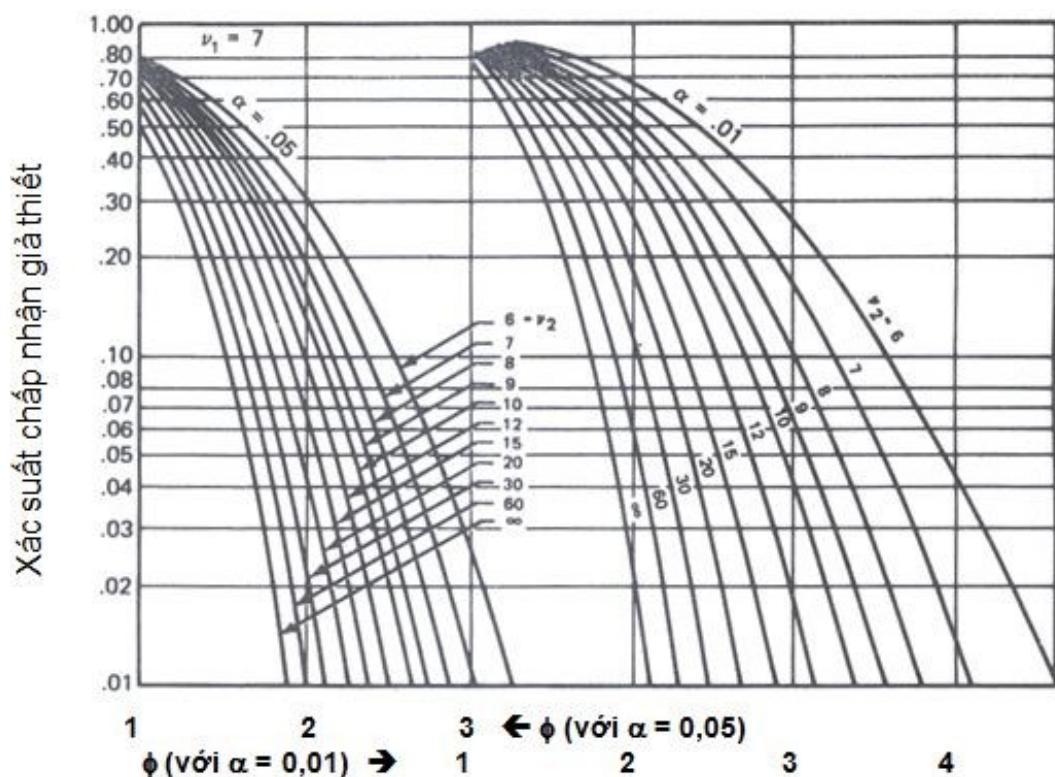
Xác suất chấp nhận giả thiết



Xác suất chấp nhận giả thiết



Với ν_1 = bậc tự do ở tử số, ν_2 = bậc tự do ở mẫu số.



Với v_1 = bậc tự do ở tử số, v_2 = bậc tự do ở mẫu số.

PHỤ LỤC 10
BẢNG SỐ NGẪU NHIÊN (TABLE OF RANDOM NUMBERS)

81	37	66	40	77	65	29	99	77	42	92	78	15	25	07	76	79	24	21	84
48	03	48	91	03	57	56	56	42	76	57	27	60	60	16	30	76	96	94	49
86	49	52	63	66	70	80	71	09	64	84	36	03	54	53	39	36	30	69	27
73	59	16	61	43	18	86	80	19	42	23	78	86	08	44	08	55	51	12	97
10	46	82	01	40	55	50	91	24	12	34	43	20	37	71	52	13	25	67	31
63	34	98	49	54	23	60	36	10	40	08	12	34	46	59	82	91	74	60	92
18	40	40	07	42	21	10	22	39	57	86	80	03	29	64	96	73	84	72	47
59	86	66	45	91	17	29	15	92	05	97	60	76	48	44	58	89	64	01	26
30	99	69	70	16	08	76	29	74	90	18	42	43	71	47	22	10	21	08	69
14	49	02	64	25	44	27	12	36	82	67	84	58	21	61	72	45	23	63	43
99	76	35	87	72	35	14	61	70	33	94	30	18	23	70	30	80	72	72	04
50	42	77	64	94	44	17	80	67	98	72	15	00	52	41	76	16	85	33	23
10	38	18	55	57	31	38	12	97	80	91	47	94	45	67	92	31	55	16	91
46	52	61	13	33	04	30	47	97	11	30	03	87	98	33	06	29	77	56	41
29	21	02	78	61	84	33	50	43	75	42	28	40	16	12	42	03	44	10	28
83	59	26	14	81	77	04	94	98	12	33	71	07	29	35	25	86	82	52	43
87	22	31	54	76	04	80	79	92	37	97	31	53	34	10	57	19	48	32	86
73	53	23	83	40	45	57	33	18	29	13	61	64	03	38	09	01	88	13	14
29	32	83	46	27	05	18	31	46	93	59	83	90	79	53	91	47	02	26	90
70	71	37	04	12	71	30	23	31	51	92	96	09	93	08	52	94	79	45	34
87	29	28	54	53	54	33	39	22	61	46	98	84	24	28	71	42	75	98	07
83	78	88	92	75	35	07	41	70	05	83	13	45	06	24	89	75	66	06	27
69	26	97	35	72	95	58	30	84	12	70	41	36	92	05	62	89	01	62	31
07	82	88	94	99	80	07	37	94	52	15	26	90	39	39	51	53	40	98	78
55	80	29	81	32	27	28	59	29	74	27	46	15	47	00	47	94	04	03	43
80	73	03	69	35	68	22	77	82	26	83	58	62	71	77	88	00	70	45	58
45	69	97	79	98	33	45	64	83	62	20	36	34	64	67	29	08	47	56	72
25	15	57	13	07	95	01	02	02	70	86	74	56	14	94	33	49	73	62	71
82	87	56	32	99	86	35	13	22	12	25	90	89	20	82	87	46	23	14	27
00	98	13	94	00	85	09	30	97	98	72	40	81	87	33	96	58	28	08	64
61	99	16	38	11	08	28	65	70	71	79	51	31	38	27	99	64	57	99	98
79	93	50	34	41	50	21	49	74	52	03	52	53	24	89	53	96	19	31	06
36	19	99	62	65	08	46	68	44	96	73	98	65	41	72	37	46	27	11	41
88	27	35	22	39	59	19	39	65	55	59	20	25	48	23	61	78	35	48	89
24	20	27	94	31	17	47	50	37	11	15	19	46	34	23	80	37	60	30	50
54	55	44	08	73	05	63	52	47	43	82	40	98	97	92	13	46	31	02	67
83	93	99	35	06	85	63	39	04	12	93	91	86	88	63	68	62	75	91	38
64	64	87	77	53	05	29	76	06	23	88	81	10	33	02	86	86	93	12	00
74	72	31	23	20	17	06	56	26	91	86	60	48	28	08	93	56	03	26	44
81	76	68	15	22	70	38	56	71	59	69	38	45	64	79	98	69	02	11	90

PHỤ LỤC 11
SƠ ĐỒ THÍ NGHIỆM Ô VUÔNG LATINH MẪU

3 x 3

A	B	C
B	C	A
C	A	B

4 x 4

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

4 x 4

A	B	C	D
B	A	D	C
C	D	B	A
D	C	A	B

4 x 4

A	B	C	D
B	D	A	C
C	A	D	B
D	C	B	A

4 x 4

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

5 x 5

A	B	C	D	E
B	A	E	C	D
C	D	A	E	B
D	E	B	A	C
E	C	D	B	A

A	B	C	D	E	F
B	F	D	C	A	E
C	D	E	F	B	A
D	A	F	E	C	B
E	C	A	B	F	D
F	E	B	A	D	C

A	B	C	D	E	F	G
B	C	D	E	F	G	A
C	D	E	F	G	A	B
D	E	F	G	A	B	C
E	F	G	A	B	C	D
F	G	A	B	C	D	E
G	A	B	C	D	E	F

8 x 8

A	B	C	D	E	F	G	H
B	C	D	E	F	G	H	A
C	D	E	F	G	H	A	B
D	E	F	G	H	A	B	C
E	F	G	H	A	B	C	D
F	G	H	A	B	C	D	E
G	H	A	B	C	D	E	F
H	A	B	C	D	E	F	G

10 x 10

A	B	C	D	E	F	G	H	I	J
B	C	D	E	F	G	H	I	J	A
C	D	E	F	G	H	I	J	A	B
D	E	F	G	H	I	J	A	B	C
E	F	G	H	I	J	A	B	C	D
F	G	H	I	J	A	B	C	D	E
G	H	I	J	A	B	C	D	E	F
H	I	J	A	B	C	D	E	F	G
I	J	A	B	C	D	E	F	G	H
J	A	B	C	D	E	F	G	H	I

11 x 11

A	B	C	D	E	F	G	H	I	J	K
B	C	D	E	F	G	H	I	J	K	A
C	D	E	F	G	H	I	J	K	A	B
D	E	F	G	H	I	J	K	A	B	C
E	F	G	H	I	J	K	A	B	C	D
F	G	H	I	J	K	A	B	C	D	E
G	H	I	J	K	A	B	C	D	E	F
H	I	J	K	A	B	C	D	E	F	G
I	J	K	A	B	C	D	E	F	G	H
J	K	A	B	C	D	E	F	G	H	I
K	A	B	C	D	E	F	G	H	I	J

TÀI LIỆU THAM KHẢO

Tiếng Việt

1. Chu Văn Mẫn, Đào Hữu Hò (1999). Thống kê sinh học. Nhà xuất bản Khoa học và Kỹ thuật.
2. Đặng Vũ Bình (2002). Di truyền số lượng. Nhà xuất bản Nông nghiệp.
3. Nguyễn Văn Đức (2002). Mô hình thí nghiệm trong nông nghiệp. Nhà xuất bản Nông nghiệp.
4. Nguyễn Văn Thiện (1997). Phương pháp nghiên cứu trong chăn nuôi. Nhà xuất bản Nông nghiệp.
5. Nguyễn Xuân Trạch và Đỗ Đức Lực (2016). Giáo trình Phân tích số liệu thí nghiệm và Công bố kết quả nghiên cứu chăn nuôi. Nhà xuất bản Đại học Nông nghiệp.
6. Pascal Leroy và Frederic Farnir (1999). Thống kê sinh học. Tài liệu dịch từ bản tiếng Pháp, dịch giả: Đặng Vũ Bình. Đại học Nông nghiệp I Hà Nội.
7. Phạm Chí Thành (1988). Phương pháp thí nghiệm đồng ruộng. Đại học Nông nghiệp I Hà Nội.
8. Phan Hiếu Hiền (2001). Phương pháp bố trí thí nghiệm. Nhà xuất bản Nông Nghiệp.

Tiếng nước ngoài

9. Aviva Petrie and Paul Watson (2001). Statistics for veterinary and animal science. Blackwell Science.
10. Campbell R. C. (2000). Statistics for Biologists. Cambridge University Press.
11. Claustraux J. J. (2002). Expérimentation, concevoir pour analyser. Gembloux, faculté universitaire des sciences agronomique.
12. Cochran W. G. and Cox G. M. (1966). Experimental Designs. Wiley International Edition.
13. Cox D. R. (1958). Planning of experiments. Wiley International Edition.
14. Доспехов Б. А. (1985). Методика полевого опыта. Агропромиздат.
15. Douglas C. Montgomery (1996). Design and analysis of experiments. Wiley International Edition.
16. Hans Houe, Annette Kjær Ersbøll and Nils Toft (2004). Introduction to Veterinary Epidemiology. Narayana Press.
17. Harold R. Lindman (1991). Analysis of variance in experimental design. Springer-Verlag
18. Kaps M. and Lamberson W. R. (2004). Biostatistics for animal science. CABI Publishing
19. Mead R., Curnow R. N. and Hasted A. M. (1993). Statistical methods in agriculture and experimental biology. Chapman & Hall/Crc.
20. Mick O'Neill, Peter Thomson (2002). Third year biometry: Experimental design, Statistical modelling. The University of Sydney.

21. Овсянников А. И. (1976). Основы опытного дела в животноводстве. Колос.
22. Peter Thomson, Frank Nicholas and Cris Moran (2002). *Genetics and biometry*. The University of Sydney.
23. Pierre Dagnelie (1993). *Principes d'expérimentation*. Les Presses Agronomiques de Gembloux.
24. Robert R. Sokal and James Rohlf F. (2000). *Biometry*. W.H. Freeman and Company.
25. Preston T. R. (1995). Tropical animal feeding. Food and Agriculture Organization of the United Nations.

NHÀ XUẤT BẢN ĐẠI HỌC NÔNG NGHIỆP

Trâu Quỳ - Gia Lâm - Hà Nội

Điện thoại: 043. 876. 0325 – 04. 6261. 7649

Email: nxbdhnn@vnua.edu.vn

www.vnua.edu.vn/nxb

Chịu trách nhiệm xuất bản:

NGUYỄN QUỐC OÁNH

Biên tập: **BÙI TÙNG LÂM**

Thiết kế bìa: **ĐỖ LÊ ANH**

Chế bản vi tính: **BÙI TÙNG LÂM**

ISBN: 978 – 604 – 924 – 286 – 1

NXBĐHNN - 2017

In 300 cuốn, khổ 19 x 27 cm, tại Công ty TNHHMTV NXB Nông nghiệp.

Địa chỉ: Số 6 ngõ 167 Phương Mai, Đống Đa, Hà Nội.

Số đăng ký kế hoạch xuất bản: 1174 – 2017/CXBIPH/04 – 01/ĐHNN.

Số quyết định xuất bản: 14/QĐ-NXB-HVN ký ngày 16/5/2017.

In xong và nộp lưu chiểu quý II - 2017.